The Journal of Chemical Physics

Message-passing neural networks for high-throughput polymer screening

Cite as: J. Chem. Phys. **150**, 234111 (2019); https://doi.org/10.1063/1.5099132
Submitted: 07 April 2019 • Accepted: 29 May 2019 • Published Online: 19 June 2019

Peter C. St. John, Caleb Phillips, Travis W. Kemper, et al.

AIP Publishing

Message-passing neural networks for high-throughput polymer screening

Cite as: J. Chem. Phys. 150, 234111 (2019); doi: 10.1063/1.5099132
Submitted: 7 April 2019 • Accepted: 29 May 2019 •
Published Online: 19 June 2019

Peter C. St. John,[1,a)] Caleb Phillips,[2] Travis W. Kemper,[2] A. Nolan Wilson,[3] Yanfei Guan,[4] Michael F. Crowley,[1] Mark R. Nimlos,[3] and Ross E. Larsen[2]

AFFILIATIONS

[1] Biosciences Center, National Renewable Energy Laboratory, Golden, Colorado 80401-3393, USA
[2] Computational Science Center, National Renewable Energy Laboratory, Golden, Colorado 80401-3393, USA
[3] National Biaoenergy Center, National Renewable Energy Laboratory, Golden, Colorado 80401-3393, USA
[4] Department of Chemistry, Colorado State University, Fort Collins, Colorado 80523-1872, USA

[a)] Electronic mail: peter.stjohn@nrel.gov

ABSTRACT

Machine learning methods have shown promise in predicting molecular properties, and given sufficient training data, machine learning approaches can enable rapid high-throughput virtual screening of large libraries of compounds. Graph-based neural network architectures have emerged in recent years as the most successful approach for predictions based on molecular structure and have consistently achieved the best performance on benchmark quantum chemical datasets. However, these models have typically required optimized 3D structural information for the molecule to achieve the highest accuracy. These 3D geometries are costly to compute for high levels of theory, limiting the applicability and practicality of machine learning methods in high-throughput screening applications. In this study, we present a new database of candidate molecules for organic photovoltaic applications, comprising approximately 91 000 unique chemical structures. Compared to existing datasets, this dataset contains substantially larger molecules (up to 200 atoms) as well as extrapolated properties for long polymer chains. We show that message-passing neural networks trained with and without 3D structural information for these molecules achieve similar accuracy, comparable to state-of-the-art methods on existing benchmark datasets. These results therefore emphasize that for larger molecules with practical applications, near-optimal prediction results can be obtained without using optimized 3D geometry as an input. We further show that learned molecular representations can be leveraged to reduce the training data required to transfer predictions to a new density functional theory functional.

I. INTRODUCTION

High-throughput computational screening offers the ability to explore large regions of chemical space for particular functionality, greatly enhancing the efficiency of material development efforts.[1–3] Due to its favorable balance between computational cost and chemical accuracy, density functional theory (DFT) has served as the workhorse of high-throughput computational material design. However, while DFT sacrifices chemical accuracy for numerical efficiency, DFT calculations are still too slow to screen the vast combinatorial landscape of potential chemical structures.[4,5] As an alternative to detailed quantum chemistry calculations, fully empirical machine learning (ML) predictions offer calculation times nearly six orders of magnitude faster than DFT [$O(10^{-3}$s) for ML and $O(10^{3}$s) for DFT on approximately 30 heavy atom molecules]. Machine learning approaches have recently been effective in reproducing DFT results given sufficient training data[6] and therefore offer an opportunity to efficiently screen much larger libraries of compounds without further reduction in chemical fidelity.

Developing ML pipelines for molecular property prediction often involves encoding variable-sized molecules as a finite-dimensional vector. Traditional approaches use group contribution methods, molecular fingerprints, and molecular descriptors to convert molecular structures into a suitable input for dense neural networks or other ML models.[7–13] However, hand-engineered molecular features may not sufficiently capture all the variability present in the space of chemically feasible compounds. Neural network architectures that operate directly on graph-valued inputs have been developed,[14] allowing "end-to-end" learning on molecular space. In this approach, models simultaneously learn both how to extract appropriate features as well as how to use these features to make accurate predictions. End-to-end learning techniques have supplanted traditional methods in image recognition and computer translation, similar applications where determining a suitable fixed-size numerical representation of the input data is difficult.

A number of approaches for end-to-end learning on molecules have recently been unified into a single theoretical framework known as Message Passing Neural Networks (MPNNs) and even more recently as graph networks.[15,16] In MPNNs, predictions are generated from input graphs with node and edge features. The network comprises a sequence of layers, including a number of *message passing* layers and a *readout* layer. In the message passing layers, node-level state vectors are updated according to the graph's connectivity and the current states of neighboring nodes. Following a number of message passing layers, the readout layer generates a single graph-level vector from node-level states. These networks have demonstrated best-in-class performance on all properties in the QM9 computational dataset, a benchmark dataset for molecular property prediction consisting of DFT-optimized 3D coordinates and energies for 134 000 molecules with nine or fewer heavy atoms.[17] Further modifications of the MPNN framework have demonstrated even higher accuracies.[18–21] However, both Gilmer *et al.*[15] and more recent studies have noted that optimized, equilibrium 3D molecular geometries were required to achieve optimal accuracy on the QM9 dataset. Since obtaining minimum-energy atomic coordinates is a numerically intensive task, this requirement is limiting for applications in high-throughput chemical screening—particularly for molecules with a large number of atoms.

While effective, deep learning requires large amounts of data in order to learn appropriate feature representations.[22] However, many applications of deep learning have benefited from *transfer learning*, where weights from a neural network trained on a large dataset are used to initialize weights for a related task with limited data.[23] In this way, the model's ability to extract useful features from inputs—learned from the larger dataset—is transferred to the new regression task, improving predictive accuracy with fewer training samples. In the molecular space, previous studies have shown that models are able to successfully predict molecules outside their training set,[24,25] improve their predictive accuracy with additional training on molecules from a different distribution than the prediction target,[26] and estimate nonequilibrium atomic energies at a higher level of theory by pretraining networks on lower-level calculations.[27]

In this study, we apply a MPNN to a newly developed computational dataset of 91 000 molecules with optoelectronic calculations for organic photovoltaic (OPV) applications. For OPV applications, single-molecule electronic properties play a role in determining overall device efficiency,[28–30] and the search space of molecular structures is sufficiently large that experimental exploration is impractical.[31] Machine learning approaches have previously been used to predict the properties of candidate OPV materials,[32–34] and a recent study demonstrated that a gap still exists between models that consider XYZ coordinates and those based only on simplified molecular-input line-entry system (SMILES) strings.[34] While chemical structures of candidate molecules can be rapidly enumerated (referred to as a molecule's 2D geometry), calculating atomic positions at a high level of theory is computationally prohibitive when screening millions of possible molecules. We therefore design a ML pipeline to predict optoelectronic properties (e.g., $\varepsilon_{\mathrm{HOMO}}$, $\varepsilon_{\mathrm{LUMO}}$, optical excitation energy) directly from a molecule's 2D structure, without requiring 3D optimization using DFT. We demonstrate that for the types of molecules considered in this study, MPNNs trained without explicit spatial information are capable of approaching chemical accuracy and show nearly equivalent performance to models trained with spatial information. Moreover, we show that weights from models trained on one DFT functional are able to improve performance on an alternative DFT functional with limited training data, even when the two target properties are poorly correlated. This application demonstrates that high-throughput screening of molecular libraries (in the millions of molecules) can be accomplished at chemical accuracy quickly with machine learning methods without the computational burden of DFT structure optimization. Additionally, these results indicate that the best neural network architectures trained on existing small-molecule quantum chemical datasets may not be optimal when molecular sizes increase. We therefore make the newly developed OPV dataset considered in this work (with both 2D and 3D structures) publicly available for future graph network architecture development.

II. METHODS

A. Dataset preparation

The database considered in this study contains calculations performed with several DFT functionals and basis sets (denoted functional/basis below) using the Gaussian 09 electronic structure package with default settings.[35] A web interface to the database is available.[36] The structures consist of combinations of building blocks, largely single and multiring heterocycles commonly found in OPV applications.[2,28,29] The database is primarily focused on quantifying the behavior of polymer systems, and therefore, calculations were performed at a range of oligomer lengths to extrapolate to behavior at the polymer limit.[37] Two datasets were extracted from the database by selecting entries performed with the two functional/basis combinations with the greatest number of calculations, B3LYP/6-31g(d) and CAM-B3LYP/6-31g. Each dataset consists of monomer structures, with or without 3D structural information, and associated DFT-calculated optoelectronic properties. Molecular structures were encoded using SMILES strings,[38] and optimized 3D coordinates (when used) were stored in SDF files. The specific electronic properties we predict are the energy of highest occupied molecular orbital for the monomer ($\varepsilon_{\mathrm{HOMO}}$), the lowest unoccupied molecular orbital of the monomer ($\varepsilon_{\mathrm{LUMO}}$), the first excitation energy of the monomer calculated with time-dependent DFT (gap), and the spectral overlap (integrated overlap between the optical absorption spectrum of a dimer and the AM1.5 solar spectrum). In addition to these properties, we also predict electronic properties that have been extrapolated to the polymer limit, including the polymer $\varepsilon_{\text{HOMO}}$, polymer $\varepsilon_{\text{LUMO}}$, polymer gap, and optical $\varepsilon_{\text{LUMO}}$ (sum of the polymer $\varepsilon_{\text{HOMO}}$ and polymer gap). In addition to polymers, the database also contains soluble small molecules for solution-processable OPV devices.[39,40] As these molecules are not polymerized, these entries lack information on extrapolated polymer electronics. These entries were included in the training set but excluded from the validation and test sets.

In order to screen a larger number of molecules, conformational sampling of each molecule was not performed; instead, a single optimization was performed for each molecule or oligomer. The primary B3LYP/6-31g(d) dataset consists of approximately 91 000 molecules with unique SMILES strings, approximately 54 000 of which contain polymer properties. Of these 54 000 with polymer properties, 5000 were randomly selected for each of the validation and test sets. Transfer learning was examined with a secondary dataset consisting of results from the CAM-B3LYP/6-31g functional. This dataset consists of approximately 32 000 unique molecules, 17 000 of which contain polymer results. From the 17 000 with polymer properties, 2000 were selected for the validation and test sets. The remainder of the calculations served as the training set. When only a subset of training data was considered (i.e., in generating learning curves), these calculations were randomly selected from the remainder small molecule and monomer results. Prior to prediction, each property is scaled to have zero median and unit inner quartile range (followed by an inverse transformation after prediction).

Determining an appropriate optimal (or target) error rate that is representative of a best-case validation loss is an important step in optimizing the hyperparameters of a ML pipeline. In previous studies, target errors were determined based on estimated experimental chemical accuracies for each of the regression tasks.[6,15] However, since many of these parameters are not directly measurable experimentally, we sought to determine a target error directly from the data. We therefore used calculation results from conformational isomers: molecules with identical connectivity but different 3D structure. Due to the size of the considered molecules, energy minimization routines can often converge to different lowest-energy states, with slightly altered optoelectronic properties. Since our model only considers atomic connectivity, it cannot distinguish between conformational isomers, and predictions for molecules with identical SMILES strings will yield identical predictions. By iterating over all pairs of conformers in the dataset, we calculate a mean absolute error (MAE) to establish a representative lower limit for predictive accuracy for a model that does not consider 3D atom positions. For the B3LYP/6-31g(d) dataset, 3225 molecules were present with at least two DFT calculations for the same 2D structure. These optimal errors are presented in Table I.

B. Message passing architecture

The molecules considered in this study and used as building blocks for OPV polymers are relatively large, with a maximum size of 201 atoms and 424 bonds (including explicit hydrogen atoms). Inputs to the neural network are generated from the molecules' SMILES strings and consist of discrete node types, edge types, and connectivity matrix. Atoms are categorized into discrete types based on their atomic symbol, degree of bonding, and whether or not they are present in an aromatic ring. Bonds are similarly categorized into discrete types based on their type (single, double, triple, or aromatic), conjugation, presence in a ring, and the atom symbols of the two participating atoms.

TABLE I. Mean absolute errors (MAEs) for test set predictions for models trained on B3LYP/6-31g(d) results. The conformers column (italicized values) reports MAE between calculations for pairs of conformational isomers, representing an optimal error rate for models trained on 2D coordinates. Distributions of prediction errors are shown in Fig. 2.

		2D		3D	
B3LYP/6-31g(d)	*Conformers*	Single-task	Multitask	DFT	UFF
Gap	*28.0 meV*	36.9	35.4	32.7	45.1
$\varepsilon_{\text{HOMO}}$	*22.0 meV*	32.1	29.4	27.0	33.1
$\varepsilon_{\text{LUMO}}$	*25.5 meV*	27.9	29.2	24.8	33.9
Spectral overlap	*81.3 W/mol*	149.3	149.2	96.6	170.0
Polymer $\varepsilon_{\text{HOMO}}$	*37.4 meV*	49.1	47.4	56.9	64.8
Polymer $\varepsilon_{\text{LUMO}}$	*45.0 meV*	47.8	46.8	56.8	63.0
Polymer gap	*46.3 meV*	57.1	56.3	69.8	74.3
Pol. optical $\varepsilon_{\text{LUMO}}$	*42.6 meV*	47.8	43.9	57.2	60.2

A schematic of the neural network is shown in Fig. 1. The *message passing* step was implemented using the matrix multiplication method,[15,41] where messages m are passed between neighboring atoms,

$$m_v^{t+1} = \sum_{w \in N(v)} A_{e_{vw}} h_w^t,$$

where v is the node index, $N(v)$ are the neighboring nodes, e_{vw} is the bond type, h_v^t is the feature vector for node v at step t, and $A_{e_{vw}}$ is a learned weight matrix for each bond type.

The *update* step was implemented as a gated recurrent unit block,[15]

$$h_v^{t+1} = \text{GRU}(h_v^t, m_v^{t+1}).$$

Initial atom embeddings, h_v^0, are initialized randomly for each atom class and learned as additional model parameters. The dimension of the atom state was chosen to be 128, with $M = 3$ message-passing layers. The readout function used was similar to the one used by Duvenaud *et al.*[14] but uses only the final hidden state of the recurrent atom unit to generate a whole-graph feature vector $\hat{y}$,

$$\hat{y} = \sum_{v \in G} \sigma(W h_v^M),$$

where W is a learned weight matrix. The dimension of $\hat{y}$ was chosen to be 1024. This summed fingerprint is then passed through a series of two fully connected layers with batch normalization and ReLU activation functions (dimensions 512 and 256, respectively), before being passed to an output layer corresponding to each property prediction.

When 3D molecular geometries were considered, the SchNet structure with edge updates from Jørgensen, Jacobsen, and Schmidt[20] was used. A nearest-neighbor cutoff of 48 was used to determine the connectivity matrix of passed messages. The dimension of the atom hidden state was chosen as $C = 64$, and separate models were trained for each of the eight target properties. As the

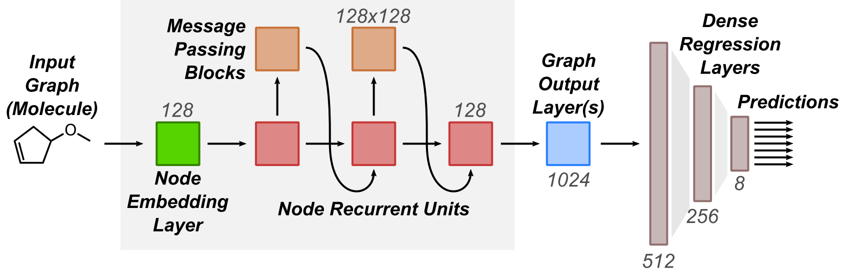

FIG. 1. Schematic of the message passing framework for 2D structures. Input molecules are labeled according to their atom and bond types. Atom embedding layers are used to initialize the weights of the message passing layers. Molecule-level feature vectors are generated through an output layer which pools all atoms through summation, which is then passed to a series of dense layers to generate a final prediction. Dimensions of each layer for the multitask model are shown in gray. For single-task models, all dimensions are identical except for the final output layer, which has dimension 1.

targets are mainly orbital energies, we similarly use an average in the readout function. SchNet-like models were trained with the Adam optimizer[42] with an initial learning rate of 1×10^{-4} and a decay rate of 1×10^{-5} per epoch. The models used a batch size of 32 and were trained for 500 epochs. The proposed model architecture was also benchmarked against the QM9 computational dataset (Table S1), demonstrating that the QM9 prediction task is more difficult without the availability of optimized 3D coordinates. Orbital energies are predicted to a MAE of 78 and 70 meV for the HOMO and LUMO energies, compared with 37 and 31 meV for models that consider 3D coordinates. We also note that the proposed model architecture is optimized for predicting orbital energy and not extrinsic properties such as total molecule enthalpy and therefore performs poorly in these regression tasks.

C. Software

Message passing operations were implemented using Keras and Tensorflow. Scikit-learn was used to scale the prediction targets, and rdkit was used to encode the atoms and bonds as integer classes. A Python library used to implement the MPNNs described in this study is available on Github (github.com/nrel/nfp) and installable via `pip`. All datasets, model scripts, and trained model weights for the models described in Table I are available at https://cscdata.nrel.gov/#/datasets/ad5d2c9a-af0a-4d72-b943-1e433d5750d6.

D. Hyperparameter optimization

For the 2D model, model sizes (atom vector dimension, molecule vector dimension, and number and size of dense layers) were increased until training errors fell below the target optimal error rate while the model still fit on a single GPU (Tesla K80) with a batch size of 100. Models were optimized using the ADAM optimizer. Learning rates were varied between 1×10^{-2} and 1×10^{-5}, with 1×10^{-3} yielding the best result. Explicit learning rate decay was also noticed to improve optimization; a decay value of 2×10^{-6} each epoch was used. Models were trained for 500 epochs. Methods for explicit regularization, including dropout and l_2 schemes, were tried but did not decrease the validation loss. All models (including refitting weights during transfer learning) used early stopping by evaluating the validation loss every 10 epochs and using the model which yielded the lowest validation loss.

III. RESULTS

A. Prediction performance on B3LYP/6-31g(d) results

The largest database consists of calculations performed at the B3LYP/6-31g(d) level of theory. By comparing calculation results for molecules with identical SMILES strings but different 3D geometries, a baseline error rate was established for models that only considered SMILES strings (2D features) as inputs. This error rate was relatively low: for $\varepsilon_{\mathrm{HOMO}}$, the mean absolute error (MAE) between pairs of conformers was 28.0 meV. This value is similar to the test-set prediction error reached by a machine learning study for similar molecules using Morgan fingerprints (28 meV)[33] and is also lower than both the target "chemical accuracy" of 43 meV used in the work of Faber *et al.*[6] and the MAE reached by the current best-performing model on the QM9 dataset, 36.7 meV.[20]

Two strategies were used to train models using only 2D coordinates. First, a series of models were trained for each property (Table I, "2D, single-task"). These models were capable of closely matching DFT results, with MAEs in orbital energies approximately 10 meV higher than the calculated optimal error. These errors, 32.1 meV for $\varepsilon_{\mathrm{HOMO}}$, are lower than state-of-the-art models on the QM9 dataset, suggesting 2D connectivity is sufficient to specify molecular properties for these types of molecules. Next, a single model was trained to simultaneously predict all eight target properties ("2D, multitask"). This model greatly improves prediction speed while demonstrating similar error rates to the single-task models.

For comparison, models were also trained using DFT-optimized 3D coordinates. The MPNN structure of these models was adapted from that of Jørgensen, Jacobsen, and Schmidt,[20] and a single model was trained for each target property. Resulting error distributions were similar to those of models trained on only 2D coordinates (Table I, "3D, DFT"; Fig. 2). The similarity in error distributions between models which consider 3D and 2D further indicates that for the molecules considered in this database, 2D structural information is sufficient to specify optoelectronic properties. Errors for the 3D model were smaller for monomer and dimer properties (gap, $\varepsilon_{\mathrm{HOMO}}$, $\varepsilon_{\mathrm{LUMO}}$, spectral overlap) while slightly larger for extrapolated polymer properties. This effect may suggest that polymer properties are less dependent on the monomer's precise 3D configuration.

Approximate 3D coordinates can be computed rapidly using empirical force fields, for instance, the UFF force field.[43] Molecules

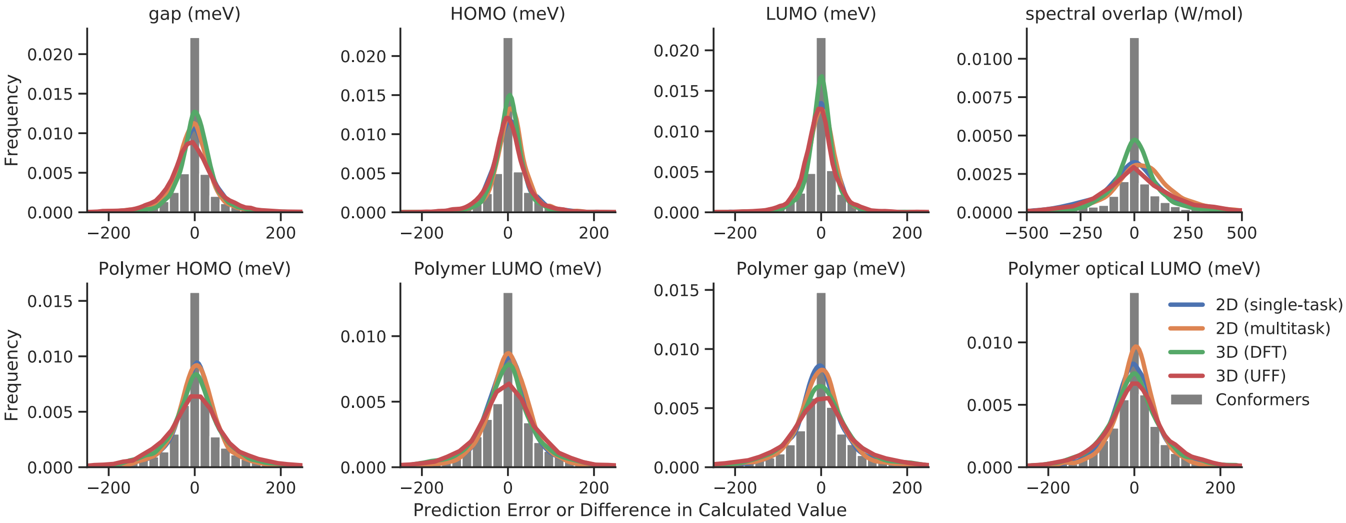

FIG. 2. Distributions of prediction error for held-out data. Distributions of prediction errors for test-set molecules from each model summarized in Table I. Histograms in differences in calculated values between pairs of conformational isomers are shown in gray. Lines represent kernel density estimates for prediction errors from each model.

in the dataset were reoptimized using the UFF force field, in order to determine approximate 3D coordinates at a much lower computational cost. Models were then retrained using these approximated geometries. The resulting prediction accuracies were worse than even the 2D models, indicating that using poor-quality molecular geometries gives worse results than omitting 3D features (Table I, “3D, UFF”).

We next explored the effect of training set size on prediction accuracy for models trained on 2D structures. Repeated optimizations of the multitask model were performed with subsampled training data with the validation set, test set, and model architecture held constant across all experiments. As expected, additional training data cause out-of-sample predictive performance to improve, shown in Fig. 3(a). The model's accuracy asymptotically approaches the optimal error rate at the largest training set sizes.

B. Transfer learning to an alternate DFT functional

Finally, we examined whether the molecular representations learned from the large-scale B3LYP/6-31g(d) dataset improved predictive performance on a related regression. End-to-end learning models perform two tasks: they extract salient features from the input data and recombine these features to generate a prediction. Inside the network, higher level representations of the data are produced by subsequent layers before ultimately leading to a predicted value. Transfer learning has previously shown to be effective in improving the predictive accuracy of models by combining large amounts of lower-level theory calculations with sparser, more accurate DFT calculations.[27,44] Transferring weights to a new model from a model trained on a closely correlated target can therefore preserve much of the logic and higher-level representations of the previous model. However, even transferring weights from a poorly correlated target can aid models by preserving low-level features useful for both targets.

To test the effectiveness of transfer learning with the proposed MPNN structure, a second, smaller dataset of polymer bandgap values calculated using the CAM-B3LYP/6-31g functional was used as a benchmark task. Two models trained on B3LYP/6-31g(d) data were used to initialize the weights for a new polymer bandgap prediction model: first, a model trained on the same parameter calculated via B3LYP/6-31g(d), and, as a more difficult example, a model trained on the B3LYP/6-31g(d) *monomer* bandgap. Correlation coefficients were used as a measure of the similarity between the old and new prediction targets. The correlation coefficients between the CAM-B3LYP/6-31g polymer bandgap and B3LYP polymer and monomer bandgaps were 0.93 and 0.48, respectively, for molecules present in both the CAM-B3LYP/6-31g and B3LYP/6-31g(d) datasets [Fig. 3(c)].

Test and validation sets of 2000 polymer species were reserved, and the remaining data were subdivided into training sets of increasing size. All transfer learning strategies were compared against a reference model with random weight initialization for all layers (i.e., no transfer learning). The results of all model predictions on the test set are shown in Fig. 3(b). For each model, performance is compared to an estimated upper-bound error. For the reference model, this error was equal to the data's standard deviation, assuming a worst-case model would always predict the mean value of the prediction target. For the models with transferred weights, upper-bound errors were found assuming new targets were calculated by linearly transforming the old prediction target to best match the new target. The root mean squared error (RMSE) for these two base-case models was calculated as 360 meV and 150 meV for the B3LYP/6-31g(d) monomer bandgap and polymer bandgap, respectively. For models with weight transfer, performance superior to these estimated upper error limits indicates that the model has retained the ability to extract and process salient features of the molecules related to the new prediction target—rather than simply recalling and rescaling the previously learned output.

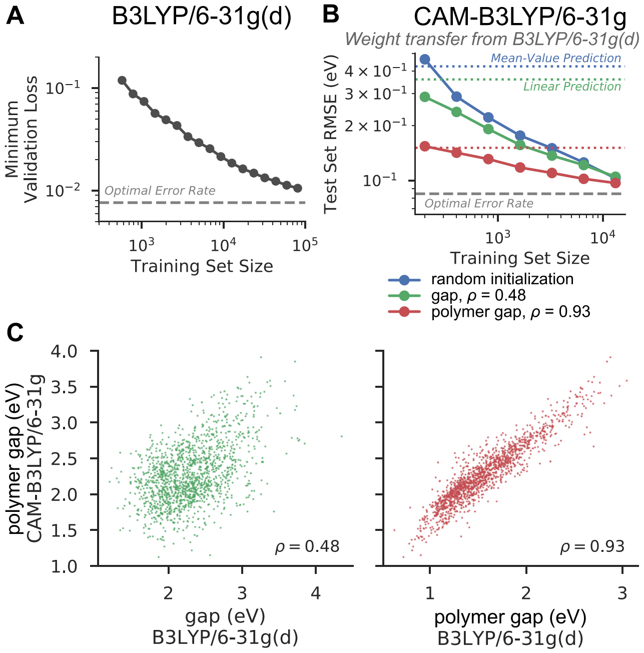

FIG. 3. Effect of training set size on predictive performance. (a) Training on B3LYP/6-31g(d). Models gradually approach the optimal error rate as training set size increases. (b) Transfer learning to predict the polymer bandgap calculated with CAM-B3LYP/6-31g. For each model, performance is compared to both the optimal error rate and an estimated upper error bound based on a simple linear model (dotted lines). (c) Illustration of the similarity between old and new prediction tasks considered during transfer learning. Plot of CAM-B3LYP/6-31g polymer bandgap (new) vs the single-target properties used for pretraining: monomer bandgap (left) and polymer bandgap (right). Points represent molecules with results calculated via both functionals.

For very small training set sizes (on the order of 200 molecules), models performed near the estimated upper error bound with the notable exception of the model with weakly correlated transferred weights, which had a substantially lower test set error than expected. This result demonstrates that pretraining models on even slightly related prediction targets could likely improve out-of-sample prediction accuracy when the available data are limited by allowing the MPNNs to learn useful molecular features. As the available training data are increased, both models with transferred weights demonstrate a concomitant decrease in their test set error below their estimated upper bound error. In particular, the model with weights transferred from the strongly correlated task shows superior performance at all training set sizes, requiring nearly an order of magnitude less data to reach RMSE values of 100 meV. At the largest training set sizes, all three models approach the optimal error rate (estimated through conformers with duplicated SMILES strings), indicating that knowledge encapsulated in transferred weights is eventually replaced with knowledge gained through the new training data.

IV. CONCLUSIONS

In this study, we have demonstrated near-equivalent prediction accuracies from both 2D and 3D structural features in MPNN architectures, both of which closely approach the estimated 2D lower-bound error from conformational optimization. While studies on the QM9 dataset have shown that 3D coordinates are required for accurate predictions, using these data as inputs mandates that full DFT calculations still be performed for each molecule. The necessity of 3D coordinates for the QM9 dataset might be explained by the substantially smaller molecules considered (≤29 atoms, including hydrogen atoms) when compared with our newly generated OPV database (≤201 atoms). Additionally, since they are exhaustively generated according to computational rules, molecules in QM9 frequently contain complex structural features that might only be captured through the explicit use of 3D coordinates. Our new public database might therefore serve as a more representative molecular learning benchmark for electronic structure calculations.

We have shown that a deep neural network pretrained on one DFT functional was able to improve predictive performance on a related DFT functional, especially in the case of limited data. This performance improvement is dependent on the correlation between tasks, but even weights transferred from a network trained on a weakly correlated task were able to improve accuracy. These results help confirm the immense value of machine learning approaches in scientific domains both to increase the fidelity of DFT simulations and to augment them, allowing for high throughput screening and guided search. Future work will therefore explore the ability of pretrained neural networks to improve prediction accuracy on experimental data and other important targets with limited available data.

SUPPLEMENTARY MATERIAL

See supplementary material for one table and one figure.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy under Contract No. DE-AC36-08GO28308 with the Alliance for Sustainable Energy, LLC, the Manager and Operator of the National Renewable Energy Laboratory. Funding was provided by the U.S. Department of Energy, Office of Bioenergy Technologies (DOE-BETO) under the Performance Advantaged Bioproducts working group. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.

REFERENCES

[1] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," APL Mater. **1**, 011002 (2013).

[2] S. A. Lopez, E. O. Pyzer-Knapp, G. N. Simm, T. Lutzow, K. Li, L. R. Seress, J. Hachmann, and A. Aspuru-Guzik, "The Harvard organic photovoltaic dataset," Sci. Data **3**, 160086 (2016).

[3] S.-D. Huang, C. Shang, X.-J. Zhang, and Z.-P. Liu, "Material discovery by combining stochastic surface walking global optimization with a neural network," Chem. Sci. **8**, 6327–6337 (2017).

[4] L. Ruddigkeit, R. van Deursen, L. C. Blum, and J.-L. Reymond, "Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17," J. Chem. Inf. Model. **52**, 2864–2875 (2012).

[5] F. Häse, C. Kreisbeck, and A. Aspuru-Guzik, "Machine learning for quantum dynamics: Deep learning of excitation energy transfer properties," Chem. Sci. **8**, 8419–8426 (2017).

[6] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, and O. A. von Lilienfeld, "Prediction errors of molecular machine learning models lower than hybrid DFT error," J. Chem. Theory Comput. **13**, 5255–5264 (2017).

[7] A. R. Katritzky, M. Kuanar, S. Slavov, C. D. Hall, M. Karelson, I. Kahn, and D. A. Dobchev, "Quantitative correlation of physical and chemical properties with chemical structure: Utility for prediction," Chem. Rev. **110**, 5714–5789 (2010).

[8] W. M. Brown, S. Martin, M. D. Rintoul, and J.-L. Faulon, "Designing novel polymers with targeted properties using the signature molecular descriptor," J. Chem. Inf. Model. **46**, 826–835 (2006).

[9] C. D. Maranas, "Optimal computer-aided molecular design: A polymer design case study," Ind. Eng. Chem. Res. **35**, 3403–3414 (1996).

[10] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, "Accelerating materials property predictions using machine learning," Sci. Rep. **3**, 2810 (2013).

[11] P. C. St. John, P. Kairys, D. D. Das, C. S. McEnally, L. D. Pfefferle, D. J. Robichaud, M. R. Nimlos, B. T. Zigler, R. L. McCormick, T. D. Foust, Y. J. Bomble, and S. Kim, "A quantitative model for the prediction of sooting tendency from molecular structure," Energy Fuels **31**, 9983–9990 (2017).

[12] D. D. Das, P. C. S. John, C. S. McEnally, S. Kim, and L. D. Pfefferle, "Measuring and predicting sooting tendencies of oxygenates, alkanes, alkenes, cycloalkanes, and aromatics on a unified scale," Combust. Flame **190**, 349–364 (2018).

[13] J. P. Janet and H. J. Kulik, "Predicting electronic structure properties of transition metal complexes with neural networks," Chem. Sci. **8**, 5137–5152 (2017).

[14] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gòmez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," Adv. Neural Inf. Process. Syst. **28**, 2224–2232 (2015).

[15] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in International Conference on Machine Learning, 2017.

[16] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. Dahl, A. Vaswani, K. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu, "Relational inductive biases, deep learning, and graph networks," e-print arXiv:1806.01261 (2018).

[17] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," Sci. Data **1**, 140022 (2014).

[18] K. Schütt, P.-J. Kindermans, H. E. Sauceda Felix, S. Chmiela, A. Tkatchenko, and K.-R. Müller, "SchNet: A continuous-filter convolutional neural network for modeling quantum interactions," Adv. Neural Inf. Process. Syst. **30**, 991–1001 (2017).

[19] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "Schnet—A deep learning architecture for molecules and materials," J. Chem. Phys. **148**, 241722 (2018).

[20] P. B. Jørgensen, K. W. Jacobsen, and M. N. Schmidt, "Neural message passing with edge updates for predicting properties of molecules and materials," e-print arXiv:1806.03146 (2018).

[21] T. S. Hy, S. Trivedi, H. Pan, B. M. Anderson, and R. Kondor, "Predicting molecular properties with covariant compositional networks," J. Chem. Phys. **148**, 241745 (2018).

[22] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436–444 (2015).

[23] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," Adv. Neural Inf. Process. Syst. **27**, 3320–3328 (2014).

[24] J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost," Chem. Sci. **8**, 3192–3203 (2017).

[25] K. Yao, J. E. Herr, D. Toth, R. Mckintyre, and J. Parkhill, "The TensorMol-0.1 model chemistry: A neural network augmented with long-range physics," Chem. Sci. **9**, 2261–2269 (2018).

[26] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," Nat. Commun. **8**, 13890 (2017).

[27] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, and A. Roitberg, "Outsmarting quantum chemistry through transfer learning," preprint chemRxiv:6744440 (2018).

[28] S. D. Oosterhout, N. Kopidakis, Z. R. Owczarczyk, W. A. Braunecker, R. E. Larsen, E. L. Ratcliff, and D. C. Olson, "Integrating theory, synthesis, spectroscopy and device efficiency to design and characterize donor materials for organic photovoltaics: A case study including 12 donors," J. Mater. Chem. A **3**, 9777–9788 (2015).

[29] M. Scharber, D. Mühlbacher, M. Koppe, P. Denk, C. Waldauf, A. Heeger, and C. Brabec, "Design rules for donors in bulk-heterojunction solar cells—Towards 10% energy-conversion efficiency," Adv. Mater. **18**, 789–794 (2006).

[30] N. Li, I. McCulloch, and C. J. Brabec, "Analyzing the efficiency, stability and cost potential for fullerene-free organic photovoltaics in one figure of merit," Energy Environ. Sci. **11**, 1355–1361 (2018).

[31] L. Wilbraham, R. S. Sprick, K. E. Jelfs, and M. A. Zwijnenburg, "Mapping binary copolymer property space with neural networks," Chem. Sci. **10**, 4973–4984 (2019).

[32] I. Y. Kanal, S. G. Owens, J. S. Bechtel, and G. R. Hutchison, "Efficient computational screening of organic polymer photovoltaics," J. Phys. Chem. Lett. **4**, 1613–1623 (2013).

[33] E. O. Pyzer-Knapp, K. Li, and A. Aspuru-Guzik, "Learning from the Harvard clean energy project: The use of neural networks to accelerate materials discovery," Adv. Funct. Mater. **25**, 6495–6502 (2015).

[34] P. B. Jørgensen, M. Mesta, S. Shil, J. M. García Lastra, K. W. Jacobsen, K. S. Thygesen, and M. N. Schmidt, "Machine learning-based screening of complex molecules for polymer solar cells," J. Chem. Phys. **148**, 241735 (2018).

[35] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci *et al.*, GAUSSIAN 09, Revision D.1, Gaussian, Inc., Wallingford, CT, 2009.

[36] R. Larsen, D. Olson, N. Kopidakis, Z. Owczarczyk, S. Hammond, P. Graf, T. Kemper, S. Sides, K. Munch, D. Evenson, and C. Swank, "Computational database for active layer materials for organic photovoltaic solar cells," https://organicelectronics.nrel.gov/.

[37] R. E. Larsen, "Simple extrapolation method to predict the electronic structure of conjugated polymers from calculations on oligomers," J. Phys. Chem. C **120**, 9650–9660 (2016).

[38] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," J. Chem. Inf. Model. **28**, 31–36 (1988).

[39] M. T. Lloyd, J. E. Anthony, and G. G. Malliaras, "Photovoltaics from soluble small molecules," Mater. Today **10**, 34–41 (2007).

[40] T. S. van der Poll, J. A. Love, T.-Q. Nguyen, and G. C. Bazan, "Non-basic high-performance molecules for solution-processed organic solar cells," Adv. Mater. **24**, 3646–3649 (2012).

[41] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," in International Conference on Learning, 2016.

[42] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," ICLR 2015, e-print arXiv:1412.6980 (2015).

[43] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff, "UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations," J. Am. Chem. Soc. **114**, 10024–10035 (1992).

[44] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, "Graph networks as a universal machine learning framework for molecules and crystals," e-print arXiv:1812.05055 (2018).