Deep learning total energies and orbital energies of large organic molecules using hybridization of molecular fingerprints

Obaidur Rahaman and Alessio Gagliardi*

Technische Universität München, Karlstr. 45, 80333 Munich, Germany

Machine learning, graph neural network, many body tensor representation, molecular descriptors

ABSTRACT: The ability to predict material properties without the need of resource consuming experimental efforts can immensely accelerate material and drug discovery. Although ab initio methods can be reliable and accurate in making such predictions, they are computationally too expensive at a large scale. The recent advancements in artificial intelligence and machine learning as well as availability of large quantum mechanics derived datasets enable us to train models on these datasets as benchmark and to make fast predictions on much larger datasets. The success of these machine learning models highly depends on the machine-readable fingerprints of the molecules that capture their chemical properties as well as topological information. In this work we propose a common deep learning based framework to combine different types of molecular fingerprints to enhance prediction accuracy. Graph Neural Network (GNN), Many Body Tensor Representation (MBTR) and a set of simple Molecular Descriptors (MD) were used to predict the total energies, Highest Occupied Molecular Orbital (HOMO) energies and Lowest Unoccupied Molecular Orbital (LUMO) energies of a dataset containing ~62k large organic molecules with complex aromatic rings and remarkably diverse functional groups. The results demonstrate that a combination of best performing molecular fingerprints can produce better results than the individual ones. The simple and flexible deep learning framework developed in this work can be easily adapted to incorporate other types of molecular fingerprints.

INTRODUCTION

Quantum mechanical calculations, especially Density Functional Theory (DFT) is a well-established tool to accurately calculate molecular energies and properties. However, these *ab initio* calculations are computationally rigorous and performing these calculations for large molecules become inhibitively expensive. Relatively inexpensive force field based methods can be used for this purpose and they are widely used for both inorganic systems as well as organic and biological molecules. However, these methods do not take electrons into account and fail to accurately reproduce the complex interactions in molecules.[1] Therefore, the search for a fast but accurate method for describing molecular properties is still ongoing.

Machine learning algorithms are shown to be capable of accurately predicting molecular properties in a test set by learning complex relationships between different molecular features in the training set. Once trained, they are orders of magnitude faster than DFT methods but can still produce DFT level accuracies. Recently some studies used machine learning methods to predict molecular properties at an accuracy comparable to the chemical accuracy.[2-4]

Deep learning, a neural network based subset of machine learning method, demonstrated remarkable accuracy in natural language processing,[5] computer vision,[6] speech recognition[7] and many other fields.[8] It is also finding applications in an increasing number of computational chemistry fields including drug discovery[9, 10] and materials property prediction.[11, 12] However, compared to its wide spread application in other areas, it is still in its infancy in the field of chemistry, materials science and biology. It is the best time to develop this field further and consequently discover new horizons in science and novel applications.

Graph neural network, a neural network that operates on graphs, is a natural choice for molecules. Recently some studies demonstrated that graph neural network consistently outperformed other machine learning methods in predicting molecular properties derived using DFT.[2-4, 13, 14]

The success of a machine learning method highly depends on the features of the training set, in this context molecular descriptors or fingerprints. In order to provide a molecule as an input to a machine learning model, it needs to be described by a number of features, typically represented by a fixed length vector. The descriptor should ideally capture as much information about the chemical environment of the atoms as possible. It should be sensitive to the local environment of the atoms as well as atomic positions. The descriptors should be rotationally and translationally invariant.[15] Many types of molecular descriptors are proposed, with their own strengths and limitations, producing different levels of accuracies. For example, Coulomb Matrix

(CM),[16] bag of bonds (BoB),[17] Many Body Tensor Representation (MBTR),[18] Bonds Angles Machine Learning (BAML),[19] Extended Connectivity Fingerprint (ECFP4)[20].

Although, there are numerous molecular fingerprinting strategies available in the literature, there is a lack of comparative studies for their performances in predicting properties. It is also known that the performances of different molecular fingerprints can vary depending on the target. There are no systematic guides available for selecting a particular type of fingerprinting strategy for a particular target. One way to overcome this problem is to combine different fingerprinting types. Such a stacking strategy has shown to improve on the performance for ligand-based virtual screening.[21]

Due to significant advancements in some key technologies, there has been an upsurge in the availability of high level and large-scale quantum mechanical data.[22-24] Most traditional approaches are unable to use these large datasets in an effective and transferable manner. Therefore, powerful and robust machine learning methods that can learn from these datasets and predict properties for a much larger dataset are highly sought after. This will eventually make the arduous search problem in drug discovery and novel materials design a reality.

In recent times, organic molecules are routinely used in many electronic devices like organic light emitting diodes,[25] organic solar cells[26] and organic field effect transistors.[27] The ability to accurately predict energies, especially molecular orbital energies of organic molecules can be highly beneficial for characterizing optoelectronic properties. This can in turn accelerate the rational design of novel materials for organic devices.

A few deep neural network based studies[2-4, 13] predicted the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies of ~134k organic molecules in the well known QM9 dataset.[22] To the best of our knowledge, the best results were produced by Chen et al. with a mean average error (MAE) of 0.038 eV for HOMO and 0.031 eV for LUMO. However, the QM9 dataset consists only of small organic molecules and it remains uncertain how a model trained on this dataset will perform on a new dataset, especially one that consists of larger organic molecules with complex and diverse functional groups and aromatic backbones.

Recently, Stuke et al. published a diverse benchmark spectroscopy dataset consisting of ~62k large organic molecules with diverse functional groups (OE62 dataset).[23] This is a significantly more challenging dataset than QM9 dataset for the prediction of HOMO and LUMO energies.

In this work, we used a set of different molecular descriptors, Many Body Tensor Representation (MBTR), graph neural network and a combination of them using a deep learning framework to predict total energies and HOMO/LUMO energies of the molecules in the OE62 dataset. We assess the individual performances of different types of descriptors for predicting different targets. We demonstrate that a selected combination of different descriptors can outperform a particular descriptor. This is an indication that more accuracy can be achieved by such combination strategy.

The models trained in this work are capable of making fast predictions of total and molecular orbital energies of large organic molecules that are not included in the dataset. This can be useful in screening novel molecules with desirable properties for optoelectronic applications.

The manuscript is organized in the following order: The METHODS section introduces the dataset followed by explanations of different strategies of descriptor generation followed by a summary of our deep learning architecture and training parameters. The RESULTS AND DISCUSSIONS section reports the major findings of the work in terms of the performances of different descriptors and their combinations in predicting total energies and HOMO/LUMO energies.

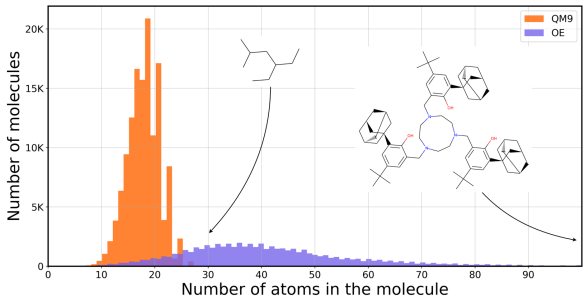
METHODS

Description of the dataset. With a goal to facilitate the design of organic semiconductors with high charge carrier mobility, Schober et al. constructed a diverse dataset of ~64k large organic molecules (referred as OE database).[24] They extracted these molecules from single-molecule crystal structures in the Cambridge Structural Database.[28] However this dataset is not yet publicly available. As a diverse benchmark spectroscopy dataset, Stuke et al. constructed a subset of this dataset (with ~62k molecules) which is publicly available (referred at OE62 dataset).[23] We used this OE62 database for fitting our models.

The OE62 dataset is significantly more challenging than the existing databases for predicting properties of organic molecules, for example QM9 dataset.[22] For example, the size distribution of the molecules in the OE62 dataset is considerably broader than that of QM9 dataset (Figure 1a). OE62 dataset is also chemically more diverse than the QM9 dataset. It contains 16 different element types, H, Li, B, C, N, O, F, Si, P, S, Cl, As, Se, Br, Te and I whereas QM9 dataset contains only five element types, H, C, N, O and F (Figure 1b).

The electronic structures of the molecules in the OE62 database are remarkably complex with large number of different scaffolds with large conjugated systems and unusual functional groups. Thus, predicting the properties of the molecules in OE62 dataset is quite challenging and therefore a good benchmark for developing predictive models.

a)



b)

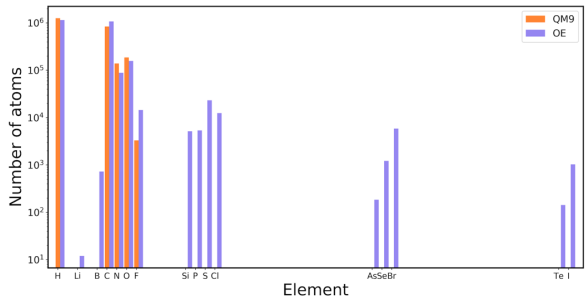


Figure 1. QM9 vs OE62, comparison of diversity of molecules. a) Size distribution of the molecules including hydrogen atoms. The largest molecules in both the datasets are shown. The largest molecule in OE62 has 174 atoms which falls outside the window depicted here for clarity. b) Element distribution of the molecules.

Feature generations. Fingerprinting molecules is an essential step for property prediction since machine learning typically requires some machine-readable features as input. Recently many innovative strategies are developed to engineer features from molecules. We selected three methods of fingerprinting the organic molecules based on their success and ease of calculation.

1. Graph neural network (GNN): During the last few years, graph neural network has emerged as a suitable, robust and powerful machine learning model for representing molecules and to predict material properties.[2-4, 14] Graph networks are natural choices for molecules since it can represent any molecule with any arbitrary connectivity.

   Gilmer et al. demonstrated that the variety of graph neural networks available in the literature can be unified with a common framework of message passing neural network.[4] The atoms in a molecule can be considered as the nodes of a graph G while the bonds between the atoms function like the edges. Neural networks can be used to model the data flow between the atoms. The inputs to these neural networks can be some atomic attributes like atomic number, type of hybridization, coordination number etc. as node features, x_v and some bond attributes like bond order, bond length etc. as edge features, e_{vw}. The messages m_v^{t+1} collected by each atom at a time step can be summed up by adding the messages of the neural nets corresponding to all the bonds that the atom is connected to:

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

   where $N(v)$ are the neighbors of v and h_v^t are the hidden states. The hidden state of a node stores the messages collected through all the edges by which it is connected to its immediate neighbors. Then the messages of each atom are updated using an update function U_t:

$$h_v^{t+1} = U_t(h_v^t, m_v^{t+1}) \quad (2)$$

   Finally, the readout function R computes a feature vector for the whole graph at T time steps:

$$\hat{y} = R(\{h_v^T \mid v \in G\}) \quad (3)$$

   The output of the whole graph/molecule can predict a molecular or material property. The parameters of the model can be optimized using a training set.

   The graph networks also satisfy the requirements of translational, rotational and permutational invariances. Faber et al. benchmarked many machine learning models on ~131k organic molecules from the QM9 data set and demonstrated that the graph based deep learning models consistently outperformed the classical machine learning models.[3] They also showed that the error in predicting the molecular properties by the best performing machine learning models were comparable to the DFT error at B3LYP level almost reaching chemical accuracy.

   As compared to the molecular systems, the crystal structures impose an additional problem since they have arbitrary sizes but most machine learning methods require fixed-length vectors as features. Xie and Grossman constructed convolutional neural networks on top of crystal

graphs representing crystal structures to accurately predict a broad range of material properties like absolute energy, band gap, Fermi energy and elastic properties.[14] Some machine learing models were also developed that can handle both molecular systems and crystals.[13] Very recently Chen et al. constructed a universal MatErials Graph Network (MEGNet) model by augmenting the attribute space with some state attributes like temperature, pressure and entropy in addition to the typical atomic and bond attributes.[2] They demonstrated that their model outperformed the previous graph networks in predicting various materials properties both for molecules and crystals.

In this work, we attributed several atomic and bond features to construct graph neural networks (GNN) for the organic molecules in the OE62 dataset. Table 1 provides the details of the features.

Table 1. Atomic and bond attributes uses to construct graph neural networks

type	attribute name	description
atom	atomic number	number of protons (integer)
	total valence	total valence (explicit + implicit) of the atom (integer)
	aromaticity	if the atom is a part of an aromatic ring (binary)
	number of Hydrogens	integer
	donor	donates electron (binary)
	acceptor	accepts electron (binary)
	hybridization	sp, sp2 or sp3 (one hot)
	atomic type	H, C, N, O etc (one hot)
bond	bond type	non bonded, single, aromatic, double or triple bond (one- hot)
	expanded distance	bond distance r expanded on Gaussian basis $\exp(-(r-r_0)^2/\sigma^2)$, where r_0 is assigned at 10 regular intervals between 0 and 4, and σ = 0.5
	Coulomb interaction	$(Z_i * Z_j)/R_{ij}$ where Z_i and Z_j are atomic numbers and R_{ij} is the bond distance

The attributes were calculated using Rdkit (www.rdkit.org) for each molecule. We note that the GNNs were constructed solely from the SMILES code provided in the dataset. The spatial information was generated by embedding the molecules using ETKDG method as implemented in Rdkit.

2. Many body tensor representation (MBTR): Based on Coulomb Matrix (CM)[16] and bag of bonds (BoB)[17], H. Huo and M. Rupp proposed a many body tensor representation of molecules and crystals.[18] It is an effective way to encode local chemical environment from three-dimensional geometric information to suitable features for machine learning.

   The one-body terms represent the atom types in the molecule. The two-body terms encode all the atom-pairs in the molecule, both bonded and non-bonded. The three-body terms capture the angular distributions of all possible atom triplets. The distributions are broadened into continuous Gaussian distributions. The broadening parameters: ρ_1, ρ_2 and ρ_3, that control the smearing of the atom type distributions are the hyperparameters of the representation and they can be fine-tuned to improve the performance of the machine learning model that uses MBTR as feature.

   Stuke et al. used MBTR to successfully reproduce the molecular orbital energies of OE dataset.[29] They observed that including the one-body term did not improve the performance but increased the computational time. We observed the same in our calculation. Therefore, we also included ony the two-body and three-body terms. We used *DScribe* package to generate the MBTR.[30]

3. Molecular descriptors (MD): Rdkit provides a convenient way to generate many molecular descriptors from the mol object, which can be generated from the SMILES code representing a molecule. Some of these descriptors can be good predictors of total energies or orbital energies. Therefore, as the third component of features, we constructed a set of 49 such descriptors, such as, molecular weight, number of valence electrons, number of hydrogen bond acceptors, number of hydrogen bond donors, number of hydrophobe, number of rotatable bonds etc. A full list of the descriptors can be found in Table S1 in supporting information.

Deep learning architecture. We used python[31] and PyTorch[32] programming languages to construct the deep learning framework. Figure 2 demonstrates the architecture we used to combine the different features (GNN, MBTR and MD) for deep learning the targets.

GNNs were built using the message passing neural network[4] (NNConv) as implemented in PyTorch. The number of convolution layers in GNN is hyperparameter of the model. However, we limited this to 3 in order to reduce the number of hyperparameters and thus complexity of the model. Gated Recurrent Unit (GRU)[33] was used as the update function for GNN.

Similarly, both the number of neurons and number of hidden layers processing MBTR and MD features, as well as the final features are hyper parameters of the model. In order to reduce the total number of hyperparameters, we fixed the number of hidden layers to 2. However, the number of neurons in the layer is kept as a hyperparameter. The number of neurons was gradually reduced in the more advanced layers, constructing a funnel like structure.

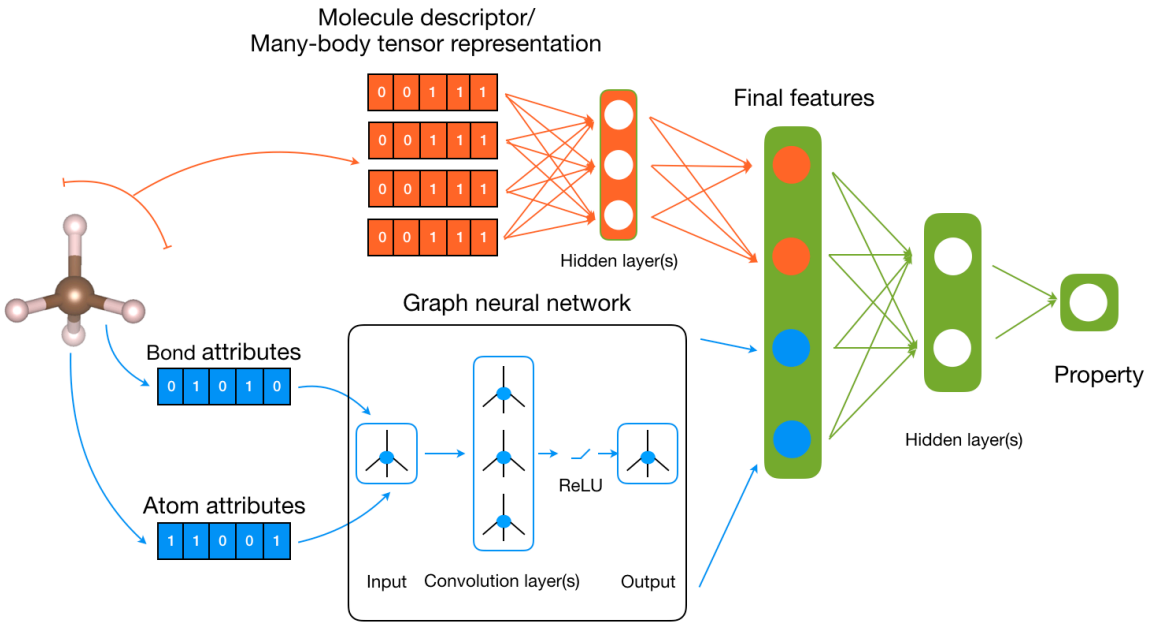


Figure 2. Deep learning architecture for hybridizing GNN, MBTR and MD.

Table 2. MAE of different targets in the OE62 dataset

Target	Units	GNN	MBTR	MD	Combination
Total energy	eV	1.028	1.035	1.016	1.005 (GNN + MD)
HOMO energy	eV	0.150	0.193	0.262	0.138 (GNN + MBTR)
LUMO energy	eV	0.158	0.215	0.318	0.136 (GNN + MBTR)

A batch size of 64 was used to balance between efficiency and accuracy. AdamW with amsgrad turned on was used as the optimizer. Applying a slight weight decay (=0.005) was useful in achieving regularization and preventing overfitting. The learning rate was kept as a hyperparameter of the model. A scheduler was used to modify the learning rate during the training. The training of the model was performed using GPU.

The OE62 dataset was randomly split into 70%, 15% and 15% for training, validation and test, respectively. The number of epochs required to achieve satisfactory accuracy was dependent on the type of the target.

RESULTS AND DISCUSSIONS

The OE62 dataset provides the total energies, HOMO energies and LUMO energies of all the molecules at both Perdew-Burker-Ernzerhof (PBE)[34] level including Tkatchenko-Scheffler van def Waals (TS-vdW) correction[35] and PBE hybrid (PBE0)[36] level of Density Functional Theory (DFT). In this work we selected the energies at the PBE0 level of theory for training our models.

Table 2 Summarizes the performances of different descriptors, when applied individually and when applied in combination with other descriptors for different targets in the dataset.

Prediction of total energy. Accurate prediction of total energy of a molecule by machine learning can be useful since it can bypass the need of expensive calculation at the DFT level.[37, 38] Ideally the three dimensional geometrical structures of the molecules should not be used as the input feature in the machine learning model because this information is something that is generated after the expensive DFT calculation. Is it possible to predict the DFT total energies of these molecules taking only the SMILES code into consideration?

We explored this question in this work using GNN and MD since these two strategies take only SMILES code as input. The OE62 dataset was used to train the deep learning architecture depicted in Figure 2.

The total energy was scaled by a logarithm to reduce its wide range. Figure 3 shows the performances of the fitted models using MD, GNN and GNN + MD on the out of bag test set. The performance of MD alone is decent with MAE of 1.016 eV. One the other hand, GNN performs slightly worse with MAE of 1.028 eV. It is interesting to note that a simple set of molecular descriptors performs better than much more complicated GNN. This could be due to the fact that the total energy is strongly correlated to the size of the molecule and the simple descriptor that counts the number of atoms in the molecule is capable of making a good prediction. This, in combination with other molecular descriptors are capable of at least predicting the relative energies of the molecules quite accurately.

When GNN is combined with MD, it produced a slightly better result than the individual ones with MAE of 1.005 eV.

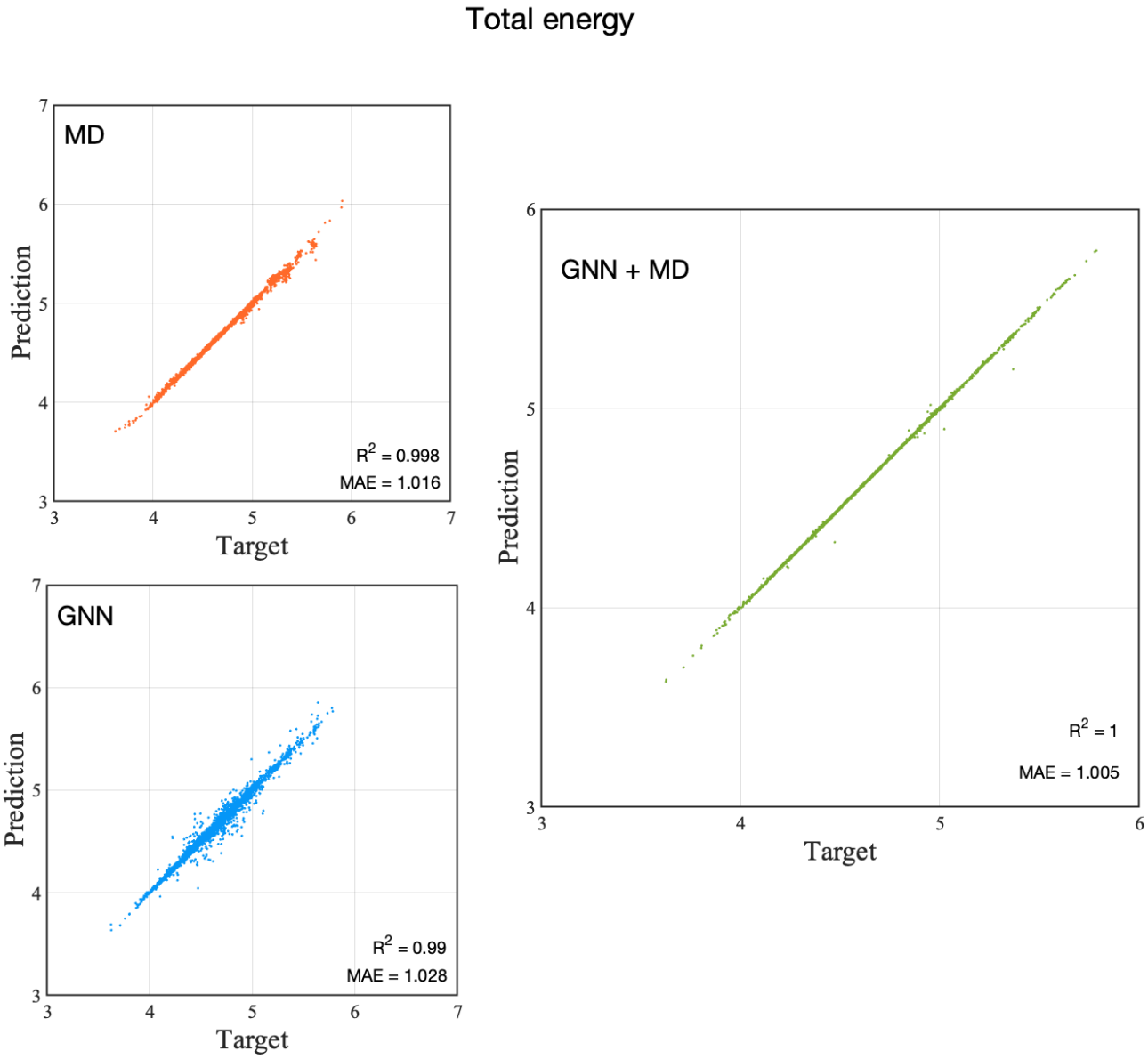


Figure 3. Prediction of total energy by MD, GNN and GNN + MD. The total energy was scaled by a logarithm before training the model and the results are also shown in the logarithm scale. The energy unit is eV.

There are a few studies that reported the performance of machine learning models for predicting total energies of organic molecules calculated by quantum mechanics.[39-41] However, the datasets considered by these studies consisted of only small organic molecules with a few different types of atomic elements. Therefore, the results of our work are not directly comparable with theirs.

Prediction of HOMO energy. GNN, MD and MBTR were individually applied to predict the HOMO energies. The performance of MD was significantly inferior compared to the other two. Therefore, we dropped MD from any further analysis. Figure 4 shows the performances of MBTR, GNN and GNN + MBTR for the prediction of HOMO energies.

MBTR alone can predict the HOMO energies with a MAE of 0.193 eV. GNN alone produced a MAE of 0.15 eV. When they were combined the MAE went down to 0.138 eV. Although, this reduction of error seems small, it is important to note that the reduction of error at a high accuracy level is always difficult. Ensemble strategies involving different machine learning models usually adds a marginal improvement. However, this marginal improvement can sometimes be significant and very difficult to achieve. The take home lesson here is that the reduction of MAE demonstrates that higher accuracy can be achieved by combining these two different types of strategies to describe the molecules, namely GNN and MBTR.

To the best of our knowledge, these results are not directly comparable to any other work in the literature. The closest is the work by Stuke at al.[29] They applied Kernel Ridge Regression

(KRR) method to predict HOMO energies of molecules in three datasets – QM9, OE62 and AA[42] using two typed of descriptors, namely Coulomb Matrix (CM) [16] and MBTR. According to their report the prediction of HOMO energies for OE62 dataset was significantly more difficult than the other two. This is due to the high diversity, structural complexity and larger sizes of the molecules in the OE62 dataset compared to the other two. They also observed that the performance of MBTR was significantly better than CM for the QM9 and OE62 datasets. Inspired by this work, we selected MBTR as one of our descriptors.

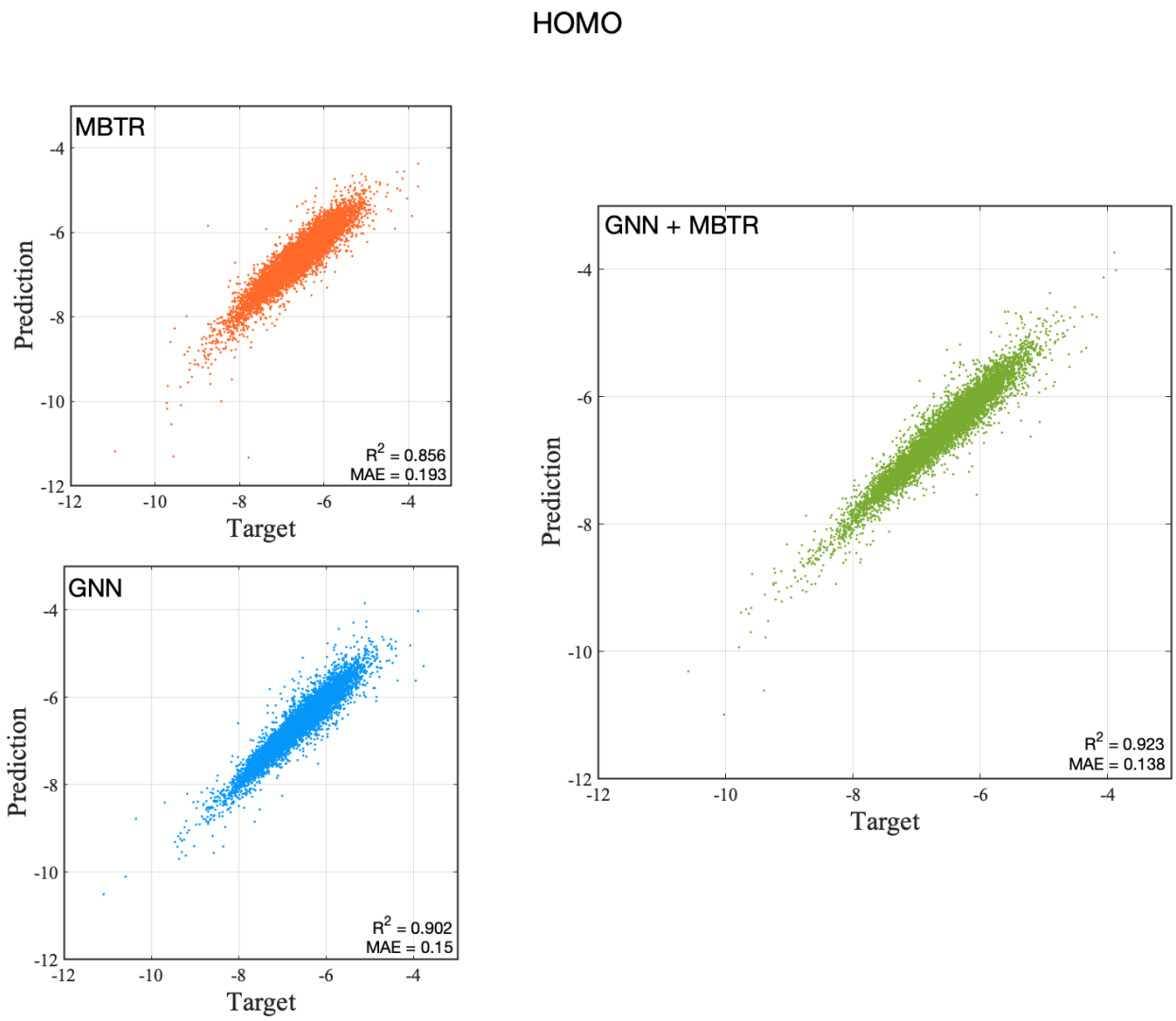


Figure 4. Prediction of HOMO energy by MBTR, GNN and GNN + MBTR. The energy unit is eV.

Stuke et al. reported a MAE of 0.173 eV for the prediction of HOMO energies for the molecules in OE62 dataset. However, for the sake of consistency, they reoptimized the structures of all the molecules in all three datasets using DFT and Perdew-Burke-Ernzerhof (PBE) functional including Tkatchenko-Scheffler van der Waals corrections and recalculated the HOMO energies. Whereas, we worked with the original structures and HOMO energies at PBE hybrid (PBE0) level of DFT as reported in the original OE62 dataset.

Prediction of LUMO energy. The results of LUMO energies followed similar trend as with HOMO energies. Figure 5 shows the results produced by MBTR, GNN and GNN + MBTR. MBTR alone produced a MAE of 0.215 eV whereas the MAE was 0.158 when GNN was applied alone. When both GNN and MBTR were applied, the MAE went down to 0.136 eV. The drop in the MAE was more pronounced for LUMO energies as compared to the HOMO energies.

Significance of errors. The descriptors and their combination produced a MAE in the range of 1 eV for total energy prediction. The errors reported in the literature are rather much smaller, in the range of kcal/mol.[39-41] However, these datasets consist of much smaller molecules. It would be interesting to compare the performance of our machine learning methods with similarly fast force field methods. Folmsbee et al. compared force field, semi empirical, DFT, ab initio and machine learing methods for their abilities to predict relative single point energies of small molecules. We note that no DFT derived structures were used to predict the total energies using GNN + MD scheme. They were predicted solely from the SMILES code of the molecules.

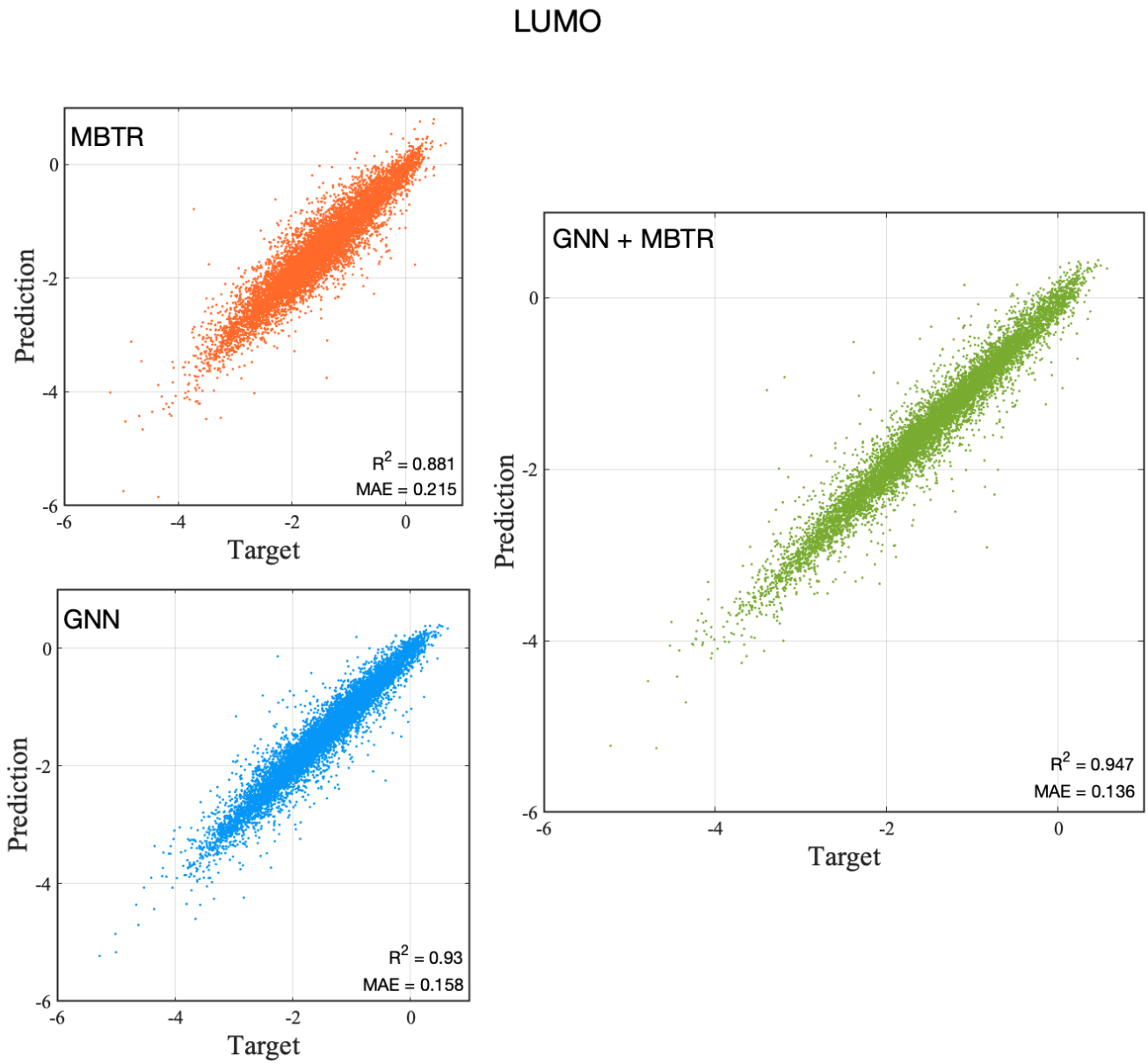


Figure 5. Prediction of LUMO energy by MBTR, GNN and GNN + MBTR. The energy unit is eV.

The error of theoretical prediction of HOMO/LUMO energies requires to be lower than 0.1 eV to be useful in spectroscopic applications. The MAEs for HOMO/LUMO energies obtained in this work, although not lower than this limit, are close. There are several opportunities to improve on these results. We hope that our work will inspire others to further develop methodologies following this direction.

Possibilities of further improvements. In this work, a reasonable accuracy is achieved for the prediction of total and molecular orbital energies of large organic molecules. However, the limit to how accurate we can get is not reached yet. Several opportunities of improvements remain to be explored in future work,

1. Hyperparameter tuning: Unlike other machine learning techniques, hyperparameter tuning for deep learning remains a bottleneck due to its long execution time and memory requirements. A simple grid search strategy for hyperparameter tuning quickly becomes prohibitively expensive with increasing number of hyperparameters. The inclusion of multiple types of descriptors in our deep learning framework resulted into quite a few hyperparameters, as described in the

method section. Therefore, we adopted a random search technique for optimizing the hyperparameters.

One of the drawbacks of both grid search and random search is that the trials are completely independent of each other. So, the search cannot be intelligently driven toward the desired direction. In some sense, they are both blind searches. A more sophisticated Bayesian optimization method learns from the previous trials using a surrogate model and makes new guesses with increasing confidence that it will drive the search toward better accuracy.[43] This method can be applied to our deep learning framework to improve results.

2. Modifying the architecture: The number of hidden layers for GNN, MBTR, MD as well as for the final features are all design parameters of the architecture. However, we kept these numbers fixed to some small integers in order to reduce the complexity and number of hyperparameters of the model. Higher accuracies can be achieved by modifying these numbers of hidden layers. To enable a deeper training of the model, residual netlike skip connections can be implemented between the nodes belonging to different descriptors.[2, 44]

   For the GNN, we applied the message passing neural network[4] (NNConv) with Gated Recurrent Unit (GRU).[33] However, it has been shown that augmenting graph neural network with adaptive attention weights that capture the local chemical environment can improve results.[45] Applying such attentional mechanism or implementing other types of GNN could possibly improve on the accuracy achieved in this work.

3. Inclusion of other molecular descriptors: We considered GNN and MBTR for their abilities in predicting material properties with high accuracy and MD for the ease they can be extracted without the need of expensive calculations. However, there are many other sophisticated molecular descriptors available that demonstrated high accuracy as well. Some of them could be promising candidates for including in our framework. The simplicity, flexibility and universality of the proposed deep learning framework allows for such experimentations and possibly further improve the results.

Novelty and originality. The motivation of this work is mainly to demonstrate that a combination of different types of features can produce better accuracy than the ones obtained by individual types. The works reporting top results in the prediction of molecular properties mostly used GNN. Although they combined different types of molecular features, they all fall within the framework of GNN. In this work, we went beyond that to hybridize GNN with completely different types of molecular fingerprints, namely MD and MBTR.

There are studies that adopted ensemble strategies based on stacking and blending of different machine learning models to outperform the individual models.[46, 47] However, we did not adopt a simple stacking or blending strategy to improve on results. Rather we developed a deep learning framework that seamlessly assembles the engineered features from individual descriptors into a set of final features, allows them to blend with each other and make the final prediction. This allows for different types of information embedded in different descriptors to mix and eventually create new information that can make better predictions. The flexibility in the deep neural network architecture enables an automated way of learning the best combinations of the final features instead of a manual way like stacking and blending.

The success of combining different types of features in a machine learning model largely depends on the complementarity of the features. The less they are correlated to each other the higher increment of accuracy can be expected by combining them. The complementarity between GNN and MBTR is quite clear. MBTR captures the fingerprint of a molecule purely based on topological information, specifically the distributions of atomic pairs and angles. No information about the interconnectivity between the atoms are captured. A model based on MBTR alone can possibly infer from the distance if there is a covalent bond between two particular atom types but it has no information about the whole network of atomic interactions. On the other hand, GNN intakes the information about the whole network of atoms and bonds with various attributes but it does not have a clear idea about the non-bonded interactions and angular distributions. To make this point clear, we can take an example of a small molecule A-B-C. MBTR will store the distances between A-B, B-C as well as A-C. In addition, it stores the angle ∠ABC. GNN will store the features of atoms A, B, C and bonds A-B and B-C. However, GNN does not have any information about the angle ∠ABC. Consequently, it cannot have any information about the A-C distance. Thus, MBTR and GNN capture different aspects of a molecule. Therefore, their combination can enrich the feature space and consequently improve the results, as evident from this work. We note that, the capacity of GNN has been enhanced to include non-bonded interactions by constructing virtual edges between any two atoms in the molecule that are within a cutoff of 4 Å. This significantly improves the performance of GNN. This enhancement enables GNN to capture long distance interactions. However, this extension is limited to a small cutoff. Increasing this cutoff to a higher value exorbitantly increases the number of edges in the network making the training time significantly longer, without gaining much accuracy. Although, some information about the long-distance pair interactions can be included in GNN using this strategy, the angular distributions still remain out of reach.

CONCLUSIONS

In this work we developed a deep learning framework to learn molecular properties by combining different types of descriptors. We selected Many Body Tensor Representation (MBTR) for its ability to capture topological information of a molecule that can be used to accurately predict its molecular property. We applied Graph Neural Network (GNN) as it is a natural choice for molecules and has shown impressive accuracy in predicting molecular properties, outperforming other machine learning and deep learning models. In addition, we assembled a set of easily derivable molecular descriptors (MD).

We applied the deep learning model to predict total energies, Highest Occupied Molecular Orbital (HOMO) energies and Lowest Unoccupied Molecular Orbital (LUMO) energies of about 62k large and complex organic molecules with diverse functional groups (OE62 dataset). A combination of GNN and MD produced the best results for the prediction of total energies whereas a combination of GNN and MBTR produced the best results for HOMO/LUMO energies. In all cases, the combinations of molecular descriptors produced better results than the individual ones.

As supported by the results, we propose our methodology as a universal framework for combining different types of molecular descriptors for the prediction of molecular properties.

ASSOCIATED CONTENT

Code availability

The code implementing the deep learning framework is available free of charge here: https://github.com/obaidur-raham an/GNN_MBTR_MD

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website.

A full list of all Molecular Descriptors (file type PDF)

AUTHOR INFORMATION

Corresponding Author

* alessio.gagliardi@tum.de

Funding Sources

ACKNOWLEDGMENT

ABBREVIATIONS

GNN Graph Neural Network; MBTR Many Body Tensor Representation; MD Molecular Descriptors

REFERENCES

1. Kanal, I. Y.; Keith, J. A.; Hutchison, G. R. A Sobering Assessment of Small-Molecule Force Field Methods for Low Energy Conformer Predictions. arXiv e-prints, arXiv:1705.04308, 2017; https://ui.adsabs.harvard.edu/abs/2017arXiv170504308K (accessed May 01, 2017).

2. Chen, C.; Ye, W. K.; Zuo, Y. X.; Zheng, C.; Ong, S. P., Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chemistry of Materials* **2019**, 31, 3564-3572.

3. Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A., Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *Journal of Chemical Theory and Computation* **2017**, 13, 5255-5264.

4. Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. arXiv e-prints, arXiv:1704.01212, 2017; https://ui.adsabs.harvard.edu/abs/2017arXiv170401212G (accessed April 01, 2017).

5. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; Klingner, J.; Shah, A.; Johnson, M.; Liu, X.; Kaiser, Ł.; Gouws, S.; Kato, Y.; Kudo, T.; Kazawa, H.; Stevens, K.; Kurian, G.; Patil, N.; Wang, W.; Young, C.; Smith, J.; Riesa, J.; Rudnick, A.; Vinyals, O.; Corrado, G.; Hughes, M.; Dean, J. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv e-prints, arXiv:1609.08144, 2016; https://ui.adsabs.harvard.edu/abs/2016arXiv160908144W (accessed September 01, 2016).

6. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E., Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience* **2018**, 2018, 7068349.

7. Nassif, A. B.; Shahin, I.; Attili, I.; Azzeh, M.; Shaalan, K., Speech Recognition Using Deep Neural Networks: A Systematic Review. *Ieee Access* **2019**, 7, 19143-19165.

8. LeCun, Y.; Bengio, Y.; Hinton, G., Deep learning. *Nature* **2015**, 521, 436-444.

9. Lavecchia, A., Deep learning in drug discovery: opportunities, challenges and future prospects. *Drug Discovery Today* **2019**, 24, 2017-2032.

10. Rifaioglu, A. S.; Atas, H.; Martin, M. J.; Cetin-Atalay, R.; Atalay, V.; Dogan, T., Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Briefings in Bioinformatics* **2019**, 20, 1878-1912.

11. Jha, D.; Ward, L.; Paul, A.; Liao, W. K.; Choudhary, A.; Wolverton, C.; Agrawal, A., ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition. *Scientific Reports* **2018**, 8.

12. Yang, Z. J.; Yabansu, Y. C.; Al-Bahrani, R.; Liao, W. K.; Choudhary, A. N.; Kalidindi, S. R.; Agrawal, A., Deep learning approaches for mining structure-property linkages in high contrast composites from simulation datasets. *Computational Materials Science* **2018**, 151, 278-287.

13. Schutt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Muller, K. R., SchNet - A deep learning architecture for molecules and materials. *Journal of Chemical Physics* **2018**, 148.

14. Xie, T.; Grossman, J. C., Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters* **2018**, 120.

15. von Lilienfeld, O. A.; Ramakrishnan, R.; Rupp, M.; Knoll, A., Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *International Journal of Quantum Chemistry* **2015**, 115, 1084-1093.

16. Rupp, M.; Tkatchenko, A.; Muller, K. R.; von Lilienfeld, O. A., Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical Review Letters* **2012**, 108.

17. Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Muller, K. R.; Tkatchenko, A., Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *Journal of Physical Chemistry Letters* **2015**, 6, 2326-2331.

18. Huo, H.; Rupp, M. Unified Representation of Molecules and Crystals for Machine Learning. arXiv e-prints, arXiv:1704.06439, 2017; https://ui.adsabs.harvard.edu/abs/2017arXiv170406439H (accessed April 01, 2017).

19. Huang, B.; von Lilienfeld, O. A., Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *Journal of Chemical Physics* **2016**, 145.

20. Rogers, D.; Hahn, M., Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, 50, 742-754.

21. Matsuyama, Y.; Ishida, T. Stacking Multiple Molecular Fingerprints for Improving Ligand-Based Virtual Screening. Cham, 2018; Springer International Publishing: Cham, 2018; pp 279-288.

22. Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A., Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **2014**, 1.

23. Stuke, A.; Kunkel, C.; Golze, D.; Todorovic, M.; Margraf, J. T.; Reuter, K.; Rinke, P.; Oberhofer, H., Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Scientific Data* **2020**, 7.

24. Schober, C.; Reuter, K.; Oberhofer, H., Virtual Screening for High Carrier Mobility in Organic Semiconductors. *Journal of Physical Chemistry Letters* **2016**, 7, 3973-3977.

25. Sekitani, T.; Nakajima, H.; Maeda, H.; Fukushima, T.; Aida, T.; Hata, K.; Someya, T., Stretchable active-matrix organic light-emitting diode display using printable elastic conductors. *Nature Materials* **2009**, 8, 494-499.

26. Zhao, J.; Li, Y.; Yang, G.; Jiang, K.; Lin, H.; Ade, H.; Ma, W.; Yan, H., Efficient organic solar cells processed from hydrocarbon solvents. *Nature Energy* **2016**, 1, 15027.

27. Muccini, M., A bright future for organic field-effect transistors. *Nature Materials* **2006**, 5, 605-613.

28. Allen, F. H., The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallographica Section B-Structural Science* **2002**, 58, 380-388.

29. Stuke, A.; Todorovic, M.; Rupp, M.; Kunkel, C.; Ghosh, K.; Himanen, L.; Rinke, P., Chemical diversity in molecular orbital energy

predictions with kernel ridge regression. *Journal of Chemical Physics* **2019**, 150.

30. Himanen, L.; Jager, M. O. J.; Morooka, E. V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S., DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications* **2020**, 247.

31. Van Rossum, G. a. D. J., Fred L, *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands: 1995.

32. Adam Paszke, S. G., Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, Adam Lerer, In *NIPS 2017 Workshop*; 2017.

33. Cho, K.; van Merrienboer, B.; Bahdanau, D.; Bengio, Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. arXiv e-prints, arXiv:1409.1259, 2014; https://ui.adsabs.harvard.edu/abs/2014arXiv1409.1259C (accessed September 01, 2014).

34. Perdew, J. P.; Burke, K.; Ernzerhof, M., Generalized gradient approximation made simple (vol 77, pg 3865, 1996). *Physical Review Letters* **1997**, 78, 1396-1396.

35. Tkatchenko, A.; Scheffler, M., Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Physical Review Letters* **2009**, 102.

36. Adamo, C.; Barone, V., Toward reliable density functional methods without adjustable parameters: The PBE0 model. *Journal of Chemical Physics* **1999**, 110, 6158-6170.

37. Sinitskiy, A. V.; Pande, V. S. Deep Neural Network Computes Electron Densities and Energies of a Large Set of Organic Molecules Faster than Density Functional Theory (DFT). arXiv e-prints, arXiv:1809.02723, 2018; https://ui.adsabs.harvard.edu/abs/2018arXiv180902723S (accessed September 01, 2018).

38. Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Muller, K. R., Bypassing the Kohn-Sham equations with machine learning. *Nature Communications* **2017**, 8.

39. Balabin, R. M.; Lomakina, E. I., Neural network approach to quantum-chemistry data: Accurate prediction of density functional theory energies. *Journal of Chemical Physics* **2009**, 131.

40. Dakota, F.; Geoffrey, H., *Assessing Conformer Energies using Electronic Structure and Machine Learning Methods*. 2020.

41. Laghuvarapu, S.; Pathak, Y.; Priyakumar, U. D., BAND NN: A Deep Learning Framework for Energy Prediction and Geometry Optimization of Organic Small Molecules. *J Comput Chem* **2020**, 41, 790-799.

42. Pereira, F.; Xiao, K.; Latino, D. A. R. S.; Wu, C.; Zhang, Q.; 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* **2017**, 57, 11-21.

43. Stuke, A.; Rinke, P.; Todorović, M. Efficient hyperparameter tuning for kernel ridge regression with Bayesian optimization. arXiv e-prints, arXiv:2004.00675, 2020; https://ui.adsabs.harvard.edu/abs/2020arXiv200400675S (accessed April 01, 2020).

44. He, K. M.; Zhang, X. Y.; Ren, S. Q.; Sun, J.; Ieee Deep Residual Learning for Image Recognition. In *2016 Ieee Conference on Computer Vision and Pattern Recognition*; 2016, pp 770-778.

45. Ryu, S.; Lim, J.; Hong, S. H.; Kim, W. Y. Deeply learning molecular structure-property relationships using attention- and gate-augmented graph convolutional network. arXiv e-prints, arXiv:1805.10988, 2018; https://ui.adsabs.harvard.edu/abs/2018arXiv180510988R (accessed May 01, 2018).

46. Zhang, R. An Ensemble Learning Approach for Improving Drug–Target Interactions Prediction. Cham, 2015; Springer International Publishing: Cham, 2015; pp 433-442.

47. Chen, C. H.; Tanaka, K.; Kotera, M.; Funatsu, K., Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications. *Journal of Cheminformatics* **2020**, 12.

Table of Contents