JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

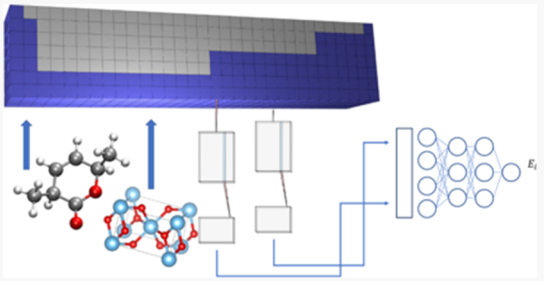
Application of Symmetry Functions to Large Chemical Spaces Using a Convolutional Neural Network

Balaranjan Selvaratnam, Ranjit T. Koodali, and Pere Miró*

Cite This: J. Chem. Inf. Model. 2020, 60, 1928−1935 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The use of machine learning in chemistry is on the rise for the prediction of chemical properties. The input feature representation or descriptor in these applications is an important factor that affects the accuracy as well as the extent of the explored chemical space. Here, we present the periodic table tensor descriptor that combines features from Behler−Parrinello's symmetry functions and a periodic table representation. Using our descriptor and a convolutional neural network model, we achieved 2.2 kcal/mol and 94 meV/atom mean absolute error for the prediction of the atomization energy of organic molecules in the QM9 data set and the formation energy of materials from Materials Project data set, respectively. We also show that structures optimized with a force field derived from this modelcan be used as input to predict the atomization energies of molecules at density functional theory level. Our approach extends the application of Behler−Parrinello's symmetry functions without a limitation on the number of elements, which is highly promising for universal property calculators in large chemical spaces.

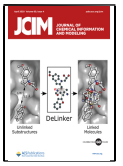


INTRODUCTION

Finding a new material with the desired properties for any industrial application is a difficult endeavor owing to the large materials search space.[1−5] Researchers have traditionally approached this problem based on the domain knowledge and chemical intuition.[2,6] However, the laborious nature of this approach limits the rate of discovery of new materials and the extent of the chemical space explored.[2,7] The introduction of high-throughput methods in both theoretical and experimental studies as a result of modern supercomputers, advanced algorithms, and robotics streamlines the identification of new promising materials for a wide variety of applications. However, the material search space remains too large to be explored using these tools within a reasonable time.[8] Fortunately, the use of high throughput methods has created large amounts of readily available chemical data such as Materials Project,[9] Aflowlib,[10] OQMD,[11] and so forth. For an extensive list, see Rinke and co-workers.[2] In addition, there are open-source software such as DScribe,[12] RDKit[13] that can calculate descriptors used for machine learning on chemical data. These data sets can be used to create predictive machine learning models to further accelerate the discovery of new materials. This approach has led to several works such as spectroscopic properties,[14] prediction of spectroscopic properties,[15,16] inverse design,[17−19] material discovery,[20] catalyst design,[21,22] synthesis,[23,24] structure elucidation,[25] and so forth. A recent example of this emerging approach was the use of machine learning to identify compositions of Co−V−Zr ternary systems that can form metallic glass.[7] The experiment-theory synergy lead to the identification of two new composition spaces with the potential to form metallic glasses. However, despite the rapid advances in the machine learning field, applying it to chemical spaces remains a challenge because of the need of descriptors that can encode the information about the chemical identities and local chemical environment to the machine learning model. Such a descriptor also needs to be invariant with respect to translation, rotation, and permutation in order to create a one-to-one mapping between any given structure and its fingerprint.[26,27]

For machine learning in the chemical spaces, many descriptors have proposed either based on the prior knowledge of the chemical space under study (e.g., ionization potential,[28] d-band center,[29] d-band filling factor,[29] etc.) or completely engineered from scratch such as Coulomb Matrix,[30,31] Bag of Bonds,[32] Partial Radial Distribution Function,[33] and so forth. These descriptors have been applied to data sets containing molecular and/or solid systems. For example, Rupp used a Coulomb matrix as a feature to predict atomization energies of organic molecules.[34] However, the length of the descriptor increases with the number of atoms. Smith et al. used modified Behler−Parrinello symmetry functions to predict the energy of organic molecules down to a root mean squared error (RMSE)

Received: September 26, 2019
Published: February 13, 2020



https://dx.doi.org/10.1021/acs.jcim.9b00835
J. Chem. Inf. Model. 2020, 60, 1928−1935

of 1.3 kcal/mol, which is near to chemical accuracy (<1 kcal/mol).[35] The descriptor used in this work represents the local chemical environment using two and three body Gaussian functions; hence the size will not change for a given number of elements. For systems with several elements, however, the descriptor size increases drastically, requiring the training of a large number of parameters. Janet and Kulik used size, electronegativity, and connectivity of atoms using revised auto correlation functions as the input to predict atomization energies of organic molecules with a mean absolute error (MAE) of 6 kcal/mol.[36] This approach was applied only to molecules and not to solid materials. For materials/solids, Zhou et al. used machine-learned atom vectors as the input to predict the formation energies with a RMSE of 150 meV/atom for elpasolites.[37] Jha et al. used the elemental composition as the input to predict the enthalpy of formation for over a quarter of a million compounds with a MAE of 50 meV/atom using a deep neural network model.[38] Atom vectors introduced by Zhou et al. and the elemental composition descriptor used by Jha et al. encode the compositional information and hence are not suitable for polymorphs. Jain and Bligaard used space group and Wyckoff-species matrix as representation to predict the formation energy with a MAE of 0.07 eV/atom using an universal—atomic position independent descriptor.[39] However, the approach was not tested for molecules. Faber et al. used similarity between query and training crystals as the descriptor to predict the formation energies of two million elpasolites with only ABC_2D_6 crystal structures with a MAE of 100 meV/atom.[8] Zheng et al. used periodic table representation (PTR) to predict the enthalpy of formation for full-Heusler X_2YZ type materials using a convolutional neural network (CNN) achieving a RMSE of 7 meV/atom.[40] These works focused on crystalline materials with specific lattices and were not tested for other systems.

Although these descriptors achieve low prediction errors, the applicability of these descriptors to a unified chemical space containing molecules and materials (periodic and amorphous) is limited. For example, in order to study the effect of molecular properties on the adsorption in porous solids, one needs a flexible, universal descriptor that can represent molecules and materials (periodic and amorphous).[41] Thus, the development of a universal chemical descriptor has the potential to remove the chemical space exploration constraints imposed by a specific descriptor. To this end, two unified descriptor-network approaches, have been already proposed to predict properties for both molecules and materials, SchNet and MEGNet. The former was developed by Schütt and co-workers and uses continuous-filter convolutional layers on learned atom embeddings,[42] and the latter uses atomic, bond, and state attributes as the input to a graph network composed of residual blocks to predict the desired properties.[43] Here, we propose a new periodic table tensor (PTT) descriptor, combining features from Behler−Parrinello's symmetry functions[44] and PTR of Zheng et al.[40]

METHODS

PTT Descriptor. In the Behler−Parrinello scheme, Gaussian functions with different parameters and cutoff radius are employed to describe the local atomic environment to a neural network, which predicts the total energy as a sum of atomic energies.[44] The two-body (G^2) and three body (G^4) symmetry functions (eqs 1 and 2, respectively) describe the radial and angular distribution of neighbor atoms within a cutoff radius

$$G_i^2 = \sum_{j\neq i}^{N} e^{-\eta(R_{ij}^2/R_c^2)} \times f_c(R_{ij}) \quad (1)$$

$$G_i^4 = 2^{1-\zeta}\sum_{j\neq i}^{N}\sum_{k\neq i,j}^{N} (1+\lambda\cos\theta_{ijk})^{\zeta} \times e^{-\eta(R_{ij}^2+R_{ik}^2+R_{jk}^2)/R_c^2} \times f_c(R_{ij}) \times f_c(R_{ik}) \times f_c(R_{jk}) \quad (2)$$

$$f_c(R_{ij}) = \begin{cases} 0.5 \times \left[\cos\left(\frac{\pi R_{ij}}{R_c}\right)+1\right] & \text{for } R_{ij} \le R_c \\ 0 & \text{for } R_{ij} > R_c \end{cases} \quad (3)$$

where R_{ij} is the distance between atom i and j, R_{ik} is the distance between atom i and k, R_{jk} is the distance between atom j and k, N is the total number of neighbor atoms, f_c is the cutoff function, η, ζ, and λ are Gaussian parameters, and $\theta_{ijk} = \cos^{-1}\left(\frac{R_{ij}\cdot R_{ik}}{R_{ij}R_{ik}}\right)$.

The symmetry functions for different element combinations are stored in a predefined order to create the atomic environment vector (AEV). The main disadvantage of this approach is that any change in the order of the AEV elements requires the model to be trained from scratch. Furthermore, in order to describe all the elements in the periodic table, the length of the AEV becomes a limiting factor as it is directly related with the number of parameters of the neural network to be trained. On the contrary, PTT descriptor uses features from both PTR and symmetry functions to create a descriptor that can work with any number of elements. Although the initial dimensions of this approach leads to a descriptor with a larger dimension, the number of neural network parameters to be trained can be reduced by using CNN, as demonstrated in this work.

Periodic Table Tensor. For machine learning, the order by which the information is presented in the descriptors should be maintained consistent for all samples. In Behler−Parrinello's implementation, this is maintained by placing the symmetry functions in a predefined order, for example, [G_1^2, G_2^2, G_3^2, ... G_n^2, G_1^4, G_2^4, G_3^4, ..., G_m^4], where each G_i^2 is uniquely specified by one element and one Gaussian parameter, η, and each G_j^4 is uniquely specified by two elements and three Gaussian parameters, (η, ξ, λ). Examples of G_i^2 and G_i^4 for a system with C and H are presented in Table 1.

The Behler−Parrinello's descriptor vector (AEV) constructed using the functions listed in Table 1 will look like [G_1^2, G_2^2, G_3^2, G_4^2, G_1^4, G_2^4, G_3^4, G_4^4], as shown in Figure 1a. In PTT, the radials (G_i^2) and angulars (G_i^4) are stored separately with a table of dimensions 7 × 32 as the basic

Table 1. Gaussian Functions and Parameters for a System Containing H and C

G_i^2	element, η	G_i^4	elements, η, ζ, λ
G_1^2	H, 0.01	G_1^4	H, C, 0.001, 2, 1
G_2^2	C, 0.01	G_2^4	C, C, 0.001, 2, 1
G_3^2	C, 0.1	G_3^4	H, H, 0.01, 4, −1
G_4^2	H, 0.1	G_4^4	H, C, 0.01, 4, −1

https://dx.doi.org/10.1021/acs.jcim.9b00835
J. Chem. Inf. Model. 2020, 60, 1928−1935

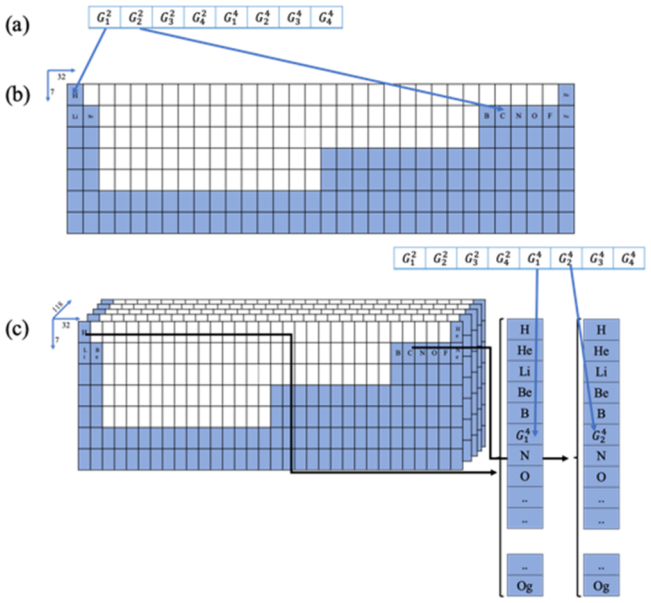


Figure 1. (a) AEV constructed for the example symmetry functions described in Table 1 using the Behler–Parrinello approach, (b) RT, and (c) AT proposed in this work for a water molecule. For clarity, only two values with same Gaussian parameters from radial and ATs are shown.

building block, based on the 7 rows and 32 columns (18 columns for the main group elements plus the 14 columns for the corresponding to the lanthanides and actinides) in the periodic table. The G_1^2 and G_2^2 values are stored in a 7 × 32 matrix, where G_1^2 is placed in row 1 and column 1 (cell address: 1, 1) corresponding to the hydrogen's position in the periodic table, and G_2^2 is placed in row 2 and column 28 (cell address: 2, 28) corresponding to the carbon's position (see Figure 1b). G_1^2 and G_2^2 are placed in a single 7 × 32 matrix because their Gaussian parameter is the same (η = 0.01). However, G_3^2 and G_4^2 are placed in separate the 7 × 32 matrix (because their η = 0.1) at positions corresponding to C (cell address: 2, 28) and H (cell address: 1, 1), respectively. Then, the two 7 × 32 matrices will be stacked to form a 7 × 32 × 2 tensor, which we refer to as the radial tensor (RT).

For the three body functions, G_i^4, the positions depend on two elements and Gaussian parameters. Hence, we use a three-dimensional tensor to store the values, where the first two dimensions (x, y) corresponds to the position of the first element (in the periodic table) and the third dimension (z) corresponds to the atomic number of the second element. Hence, we allocate a tensor of shape 7 × 32 × 118 for each unique set of Gaussian parameters (η, ζ, λ) (Figure 1c). For example, G_1^4 will be placed in a cell addressed by 1, 1, 6. The numbers, 1 and 1 correspond to the row and column numbers in the periodic table for the element H. The number 6 correspond to the atomic number of C. Likewise, G_2^4 will be placed in a cell with address (2, 28, 6). Similarly, G_3^4 and G_4^4 will be placed in another 7 × 32 × 118 tensor because the Gaussian parameters are different from those of G_1^4 and G_2^4. Finally, the two 7 × 32 × 118 matrices will be stacked to give a tensor of dimensions, 7 × 32 × 236, which we refer to as the angular tensor (AT). The advantage of this approach is that for a given set of Gaussian parameters, the dimensions of RT and AT will remain the same regardless of the number of elements. In this work, we used 16 radial functions (G^2) and 8 angular functions (G^4). The Gaussian parameters used for these functions are given in Table S1.

Convolutional Neural Network. It is apparent that training a feed forward neural network using this large descriptor will require optimizing a large number of weights. However, since the final dimension of RT and AT resembles that of an image with several color channels, we use a CNN to reduce the number of parameters. The CNN was constructed with two parallel convolutional layer blocks (one for RT and one for AT) as depicted in Figure 2. The hyperparameters, namely, convolution kernel size, number of convolution filters, dense layer structure, batch size, and activation function were optimized by training on 2048 organic molecules and their energies. The models were trained using the Adam optimizer with an initial learning rate of 0.001 until the loss on the validation set did not decrease for 100 epochs. From the results, the hyperparameters with the lowest loss for the evaluation set was selected for further studies. For hyperparameter optimization, a validation set containing 256 organic molecules was used for all other training, 10% of the respective data set was for validation.

RESULTS AND DISCUSSION

During the training, we noticed that our model was overfitting, and to avoid that, a dropout layer with a rate of 0.3 was added before the dense layers. The results of this hyperparameter optimization are given in the Supporting Information (Table S2) and the selected hyperparameters are tabulated in Table 2. The 1 × 1 kernel was found to be the optimum value for this network. Because the layers of the image-like input feature tensors (RT and AT) have information about a particular chemical environment (e.g., distribution of C–C bonds and distribution of H–C–H bonds), the 1 × 1 convolution

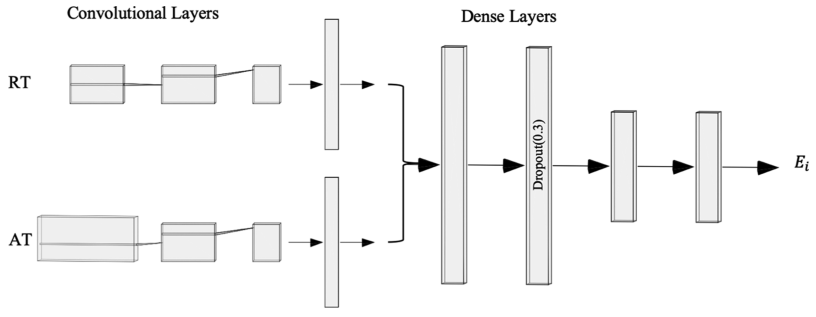


Figure 2. Network architecture of the CNN model showing the three parallel blocks each containing convolutional layers, concatenation of outputs from convolutional blocks, and the dense layers. This scheme is repeated for each element.

https://dx.doi.org/10.1021/acs.jcim.9b00835
J. Chem. Inf. Model. 2020, 60, 1928–1935

Table 2. Optimized Hyperparameters of PTT−CNN

hyperparameter	optimum value
convolution kernels	$1 \times 1^{a,b}$
number of convolution filters	$(16, 4)^{a,b}$
dense layers	(32, 32)
batch size	16
activation function	ReLu
optimizer	Adam
learning rate	0.001

[a]Radial block. [b]Angular block.

reduces the dimension and represents the information at a lower dimension. Using the optimized hyperparameters, we trained the final models with full training sets and the loss curve for the models trained using the training sets of QM9 (107,098), Materials Data set (64,264), QM7-FF (Force Field coordinates, 5680), and QM7-DFT density functional theory (DFT, coordinates, 5680) are given in Figure S1. From the optimized network structure, we calculated the number of neural network parameters for the Behler−Parrinello's model and our model for a data set containing 85 elements. For the PTT−CNN approach, including the parameters for convolution layers, the number of parameters per element is 73,993 whereas, for the reference Behler−Parrinello method as implemented in AMP by Peterson and Khorshidi, it is 980,321.[45] This shows the advantage of utilizing CNN to reduce the number of parameters. To evaluate the performance of the proposed universal descriptor, we applied it to a Materials Database collected from the Materials Project, QM9[46,47] molecular database, and a subset of GDB-13 data set containing 7,102 organic molecules optimized using a Force Field and their energies evaluated using DFT.[34]

Materials Database. For materials/solids, we collected the Materials Project data set used by Chen et al.[43] From this, structures containing elements that are poorly represented (less than 10 occurrence) in the dataset (Kr, He, Ar, and Ne) were removed. The formation energies of the materials in the data set were multiplied by the number of atoms present in the structure for training and then divided by the same number during inference. The resulting materials data set contains 85 chemical elements and 67,830 entries. The data set was randomly split into training, evaluation, and test sets with 80, 10, and 10 percent of the original data set. This split resulted in 54,264, 6783, and 6783 entries in training, evaluation, and test set, respectively. Our PTT−CNN model predicts the formation energy for structures in the test set with a MAE of 94 meV/atom which is less compared to the MAE between the DFT and experimental value[11] (Figure 3). The MAE for structures containing one to seven elements were calculated and the results are shown in Table S3. Among the materials in the test set, the MAE decreases with an increasing number of elements up to five elements (MAE is 37 meV/atom) and then increases to 54 and 61 meV/atom for structures with six and seven elements, respectively.

Molecular Database. For molecules, we used the QM9 data set which contains geometric, energetic, and thermodynamic properties of organic molecules. From this, we removed molecules containing only {C,N}, {C,F,N}, {C,F}, {H,N}, {H,O}, and {N,O} as these combinations occur less than 10 times in the data set (Table S5). The atomization energy of the molecules was used for training the models. The resulting organic data set contains 5 chemical elements and 133,873

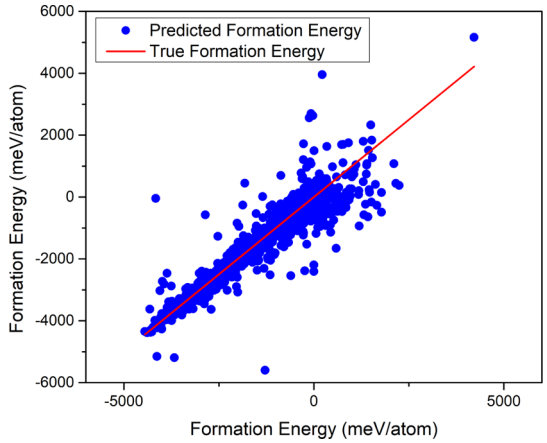


Figure 3. Parity plot showing the predicted and true formation energies for the structures in the test set of Materials Data set.

entries. This data set was randomly split into a training set with 107,098 entries, an evaluation set with 13,388 entries, and a test set with 13,387 entries corresponding to 80, 10, and 10 percent of the cleaned data set. In addition, to evaluate the influence of training data set size on the accuracy of prediction, we split the training set with 107,098 entries into five smaller data sets containing 2,048, 5,120, 10,240, 25,600, and 51,200 entries. Then, we trained different models using the aforementioned training sets with different sizes. The results are presented in Figure 4 and it shows that the MAE decreases with the increasing training set size until 25,600 and then the change is minimal.

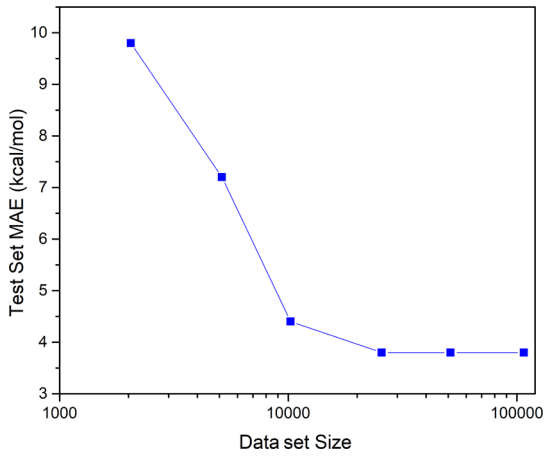


Figure 4. Change of MAE with training set size (from QM9).

Our PTT−CNN model predicts the total energy of the test set with a MAE of 2.2 kcal/mol when trained using the full training set with 107,098 entries. This error is greater than the chemical accuracy, 1 kcal/mol (Figure 5, left). However, the absolute error is within chemical accuracy (<1 kcal/mol) for 31% of the structures, below 3 kcal/mol for over 74% of the structures, and just only 0.9% of the structures have an absolute error above 10 kcal/mol (Figure 5, right).

To further analyze the performance of the proposed descriptor, a subset of 625 structural isomers of $C_7H_{10}O_2$ from the test set were extracted. The MAE for the isomer subset is 2.2 kcal/mol. The parity plots of structures containing 3-, 4-, 5-, and 6-membered rings are given in Figures S2−S9 in

https://dx.doi.org/10.1021/acs.jcim.9b00835
J. Chem. Inf. Model. 2020, 60, 1928−1935

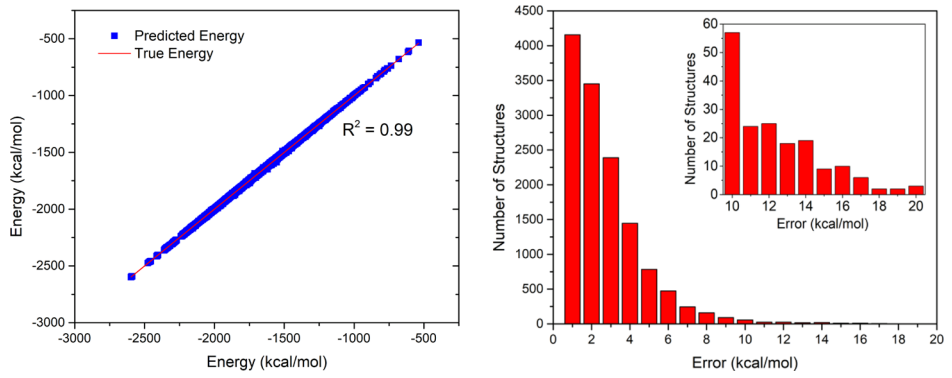


Figure 5. Parity plot showing the predicted and true atomization energies for the molecules in the test set of QM9 data set (left) and the distribution of the absolute error (right, inset shows the absolute error vs number of structures for structures with absolute error greater than 10 kcal/mol).

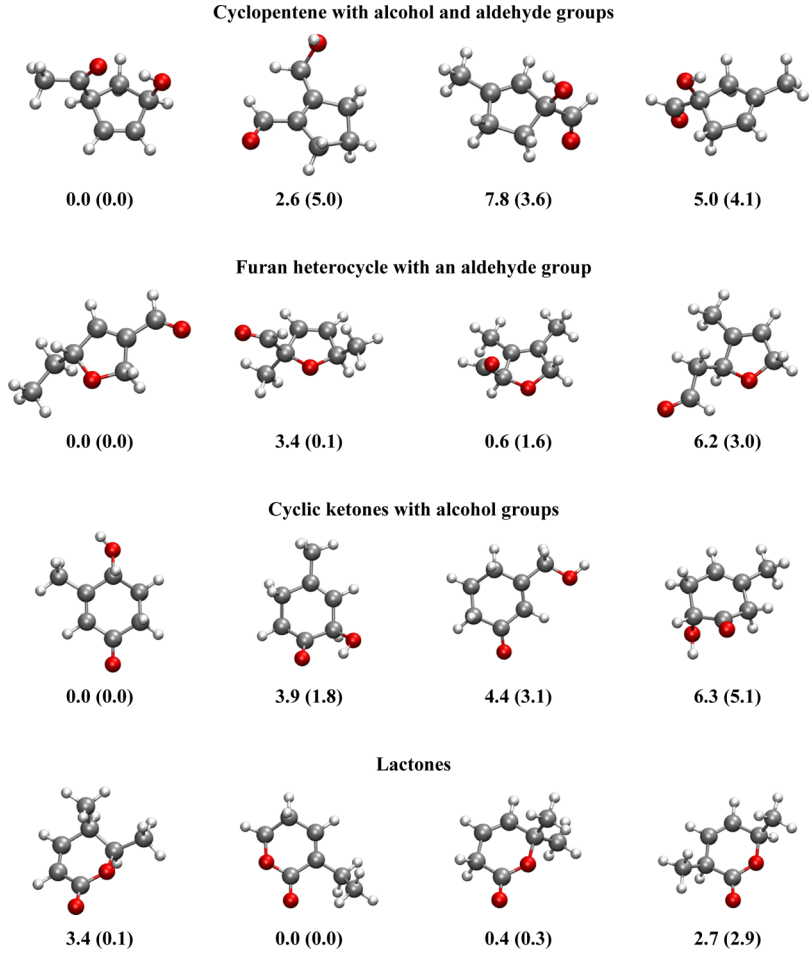


Figure 6. Structural isomers of $C_7H_{10}O_2$ of four different structural families: cyclopentene with an alcohol and aldehyde groups, furan heterocycle and an aldehyde group, cyclic ketone with an alcohol group, and lactone heterocycle. Relative energy to the most stable isomer using the PTT−CNN model are also shown, while the relative DFT energies are shown in parentheses. Only the four more stable isomers are shown. All energies are in kcal/mol.

the Supporting Information section. For isomers with similar functional groups such as cyclopentenes with an alcohol and aldehyde groups, furan heterocycles with aldehyde functional groups, cyclic ketones with an alcohol groups, and lactones, the PTT−CNN model predicts the correct minimum energy structure as DFT (Figure 6). Furthermore, the lowest energy structure predicted by the PTT−CNN model among the 625 isomers is the same as expected from the DFT energies (DFT: −1894.64 kcal/mol, PTT−CNN: −1893.96 kcal/mol).

Molecular Database with Coordinates Optimized by Force Field (QM7). The descriptor used in this work is calculated from optimized structures. However, obtaining the optimized structures is computationally time intensive. Hence, it would be interesting to see the accuracy of energies predicted using structures optimized by inexpensive methods

https://dx.doi.org/10.1021/acs.jcim.9b00835
J. Chem. Inf. Model. 2020, 60, 1928−1935

such as force fields. Hence, we used force field-optimized coordinates of 7102 organic molecules curated from GDB data set by Rupp.[34,48] From this, we removed one structure that only contained C and N atoms. Then, we split the aforementioned datas et into training, evaluation, and test set containing 5680, 710, and 711 entries (80, 10, and 10% of 7101), respectively. In order to assess the effect of structure optimization method, we trained two models: one with the force field-optimized coordinates (QM7-FF) and second model with DFT-optimized coordinates (QM7-DFT). In both cases, the energy obtained from DFT calculation was used as the target. After training, the MAE on the test set was found to be 3.6 kcal/mol for both models. This shows that the coordinates obtained using an inexpensive force field can be used to get energies corresponding to a higher-level theory.

A summary of the error of predictions on Materials Data sets, QM9 and QM7 data sets is indicated in Table 3. Although

Table 3. Comparison of Prediction Errors

	prediction error on the test set		
model	formation energy of materials (meV/atom)	energy of molecules (kcal/mol)	energy of molecules QM7FF (kcal/mol)
SchNet	35	0.44	not available
MEGNet	28	0.23	not available
PTT−CNN	94	2.2	3.6

our errors are higher than the results achieved by SchNet and MEGNet models on similar data sets, this work introduces a new descriptor-model which gives a single framework to access the unified chemical space. This enables the proposed approach to be used as an universal property calculator in large chemical spaces including molecular and solid systems. In addition, because PTT incorporates the coordinates of the atoms, hence it can be successfully used to study the energy differences in isomers. We also show that the coordinates obtained from force fields can be used as the input to predict the energies at a higher-level theory such as DFT without sacrificing the accuracy. Furthermore, the PTT approach is not restricted to chemical descriptors based on Behler−Parrinello's symmetry functions, and it can be easily adapted to other chemical descriptors that describe the local atomic environment.

CONCLUSIONS

In conclusion, this work demonstrates that our proposed PTT descriptor—CNN is able predict the energies of both molecular and solid systems. Using the proposed approach, we extend the use of Behler−Parrinello descriptors for systems with an arbitrary number of elements. Our approach takes advantage of a CNN to reduce the neural network parameters to be trained allowing the use of the PTT−CNN to large chemical spaces. In future studies, we will study performance of our model for databases containing several millions of calculations such as open quantum database, the Materials Project, the Novel Materials Discovery, or the Aflow.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.9b00835. The datasets and codes are available free of charge at figshare (10.6084/m9.figshare.11690787).

Parameters used for the angular symmetry functions; results of hyperparameter optimization; error for systems containing different number of elements; MAE error and R^2 value of structures containing different elements; element combinations and sample counts before and after the removal of low represented combinations; loss curves for QM9, materials data set, QM7-FF, and QM7-DFT; prediction of structures with 3, 4, 5, and 6 carbon member rings in the $C_7H_{10}O_2$ isomer subset; and prediction of structures with 3, 4, 5, and 6 member rings and one of them is oxygen in the $C_7H_{10}O_2$ isomer subset (PDF)

AUTHOR INFORMATION

Corresponding Author

Pere Miró − *Department of Chemistry, University of South Dakota 57069 Vermillion, South Dakota, United States;* orcid.org/0000-0002-3281-0708; Email: pere.miro@usd.edu

Authors

Balaranjan Selvaratnam − *Department of Chemistry, University of South Dakota 57069 Vermillion, South Dakota, United States*

Ranjit T. Koodali − *Department of Chemistry, University of South Dakota 57069 Vermillion, South Dakota, United States;* orcid.org/0000-0002-2790-3053

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.9b00835

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Computations supporting this project were performed on High-Performance Computing systems at the University of South Dakota funded by NSF (OAC-1626516). We are thankful to the Department of Chemistry of the University of South Dakota (USD) for Graduate Assistantship and to the USD Neuroscience, Nanotechnology, and Networks program (USD-N3) funded by the National Science Foundation (DGE-1633213). P.M. is thankful to the Department of Chemistry of the University of South Dakota (USD) for the start-up funds.

REFERENCES

(1) Le, T. C.; Winkler, D. A. Discovery and Optimization of Materials Using Evolutionary Approaches. *Chem. Rev.* **2016**, *116*, 6107−6132.

(2) Himanen, L.; Geurts, A.; Foster, A. S.; Rinke, P. Data-Driven Materials Science: Status, Challenges, and Perspectives. *Adv. Sci.* **2019**, *6*, 1900808.

(3) Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432*, 855−861.

(4) Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432*, 824−828.

(5) Reymond, J.-L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48*, 722−730.

(6) Medford, A. J.; Kunz, M. R.; Ewing, S. M.; Borders, T.; Fushimi, R. Extracting Knowledge from Data through Catalysis Informatics. *ACS Catal.* **2018**, *8*, 7403−7429.

(7) Ren, F.; Ward, L.; Williams, T.; Laws, K. J.; Wolverton, C.; Hattrick-Simpers, J.; Mehta, A. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci. Adv.* **2018**, *4*, No. eaaq1566.
(8) Faber, F. A.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Machine Learning Energies of 2 Million Elpasolite (ABC_2D_6) Crystals. *Phys. Rev. Lett.* **2016**, *117*, 135502.
(9) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002.
(10) Curtarolo, S.; Setyawan, W.; Wang, S.; Xue, J.; Yang, K.; Taylor, R. H.; Nelson, L. J.; Hart, G. L. W.; Sanvito, S.; Buongiorno-Nardelli, M.; Mingo, N.; Levy, O. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **2012**, *58*, 227−235.
(11) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **2015**, *1*, 15010.
(12) Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **2020**, *247*, 106949.
(13) RDKit. Open-source cheminformatics. http://www.rdkit.org (accessed Dec 2019).
(14) 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. *J. Chem. Inf. Model.* **2017**, *57*, 11−21.
(15) Stein, H. S.; Guevarra, D.; Newhouse, P. F.; Soedarmadji, E.; Gregoire, J. M. Machine learning of optical properties of materials − predicting spectra from images and images from spectra. *Chem. Sci.* **2019**, *10*, 47−55.
(16) Ghosh, K.; Stuke, A.; Todorović, M.; Jørgensen, P. B.; Schmidt, M. N.; Vehtari, A.; Rinke, P. Machine Learning: Deep Learning Spectroscopy: Neural Networks for Molecular Excitation Spectra (Adv. Sci. 9/2019). *Adv. Sci.* **2019**, *6*, 1970053.
(17) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360−365.
(18) Ma, W.; Cheng, F.; Liu, Y. Deep-Learning-Enabled On-Demand Design of Chiral Metamaterials. *ACS Nano* **2018**, *12*, 6326−6334.
(19) Noh, J.; Kim, J.; Stein, H. S.; Sanchez-Lengeling, B.; Gregoire, J. M.; Aspuru-Guzik, A.; Jung, Y. Inverse Design of Solid-State Materials via a Continuous Representation. *Matter* **2019**, *1*, 1370−1384.
(20) Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chem. Mater.* **2017**, *29*, 9436−9444.
(21) Huang, Y.; Chen, Y.; Cheng, T.; Wang, L.-W.; Goddard, W. A. Identification of the Selective Sites for Electrochemical Reduction of CO to C_{2+} Products on Copper Nanoparticles by Combining Reactive Force Fields, Density Functional Theory, and Machine Learning. *ACS Energy Lett.* **2018**, *3*, 2983−2988.
(22) Tran, K.; Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. *Nat. Catal.* **2018**, *1*, 696−703.
(23) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604.
(24) Kim, E.; Huang, K.; Jegelka, S.; Olivetti, E. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Comput. Mater.* **2017**, *3*, 53.
(25) Chan, L. L. a. M. L. a. M. K. Y., A Deep Learning Model for Atomic Structures Prediction Using X-ray Absorption Spectroscopic Data. **2019**, arXiv:1312.4400. arXiv preprint.
(26) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.
(27) Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**, *59*, 2545−2559.
(28) Takigawa, I.; Shimizu, K.-i.; Tsuda, K.; Takakusagi, S. Machine-learning prediction of the d-band center for metals and bimetals. *RSC Adv.* **2016**, *6*, 52587−52595.
(29) Li, Z.; Wang, S.; Chin, W. S.; Achenie, L. E.; Xin, H. High-throughput screening of bimetallic catalysts enabled by machine learning. *J. Mater. Chem. A* **2017**, *5*, 24131−24138.
(30) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(31) Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **2015**, *115*, 1094−1101.
(32) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.
(33) Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 205118.
(34) Rupp, M. Machine learning for quantum mechanics in a nutshell. *Int. J. Quantum Chem.* **2015**, *115*, 1058−1073.
(35) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192−3203.
(36) Janet, J. P.; Kulik, H. J. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure−Property Relationships. *J. Phys. Chem. A* **2017**, *121*, 8939−8954.
(37) Zhou, Q.; Tang, P.; Liu, S.; Pan, J.; Yan, Q.; Zhang, S.-C. Learning atoms for materials discovery. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, E6411−E6417.
(38) 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. *Sci. Rep.* **2018**, *8*, 17593.
(39) Jain, A.; Bligaard, T. Atomic-position independent descriptor for machine learning of material properties. *Phys. Rev. B* **2018**, *98*, 214112.
(40) Zheng, X.; Zheng, P.; Zhang, R.-Z. Machine learning material properties from the periodic table using convolutional neural networks. *Chem. Sci.* **2018**, *9*, 8426−8432.
(41) Rossi, K.; Cumby, J. Representations and descriptors unifying the study of molecular and bulk systems. *Int. J. Quantum Chem.* **2019**, No. e26151.
(42) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet − A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722.
(43) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31*, 3564−3572.
(44) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
(45) Khorshidi, A.; Peterson, A. A. Amp: A modular approach to machine learning in atomistic simulations. *Comput. Phys. Commun.* **2016**, *207*, 310−324.
(46) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.
(47) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.

https://dx.doi.org/10.1021/acs.jcim.9b00835
J. Chem. Inf. Model. 2020, 60, 1928−1935

(48) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732−8733.

https://dx.doi.org/10.1021/acs.jcim.9b00835
J. Chem. Inf. Model. 2020, 60, 1928−1935