THE JOURNAL OF PHYSICAL CHEMISTRY A

pubs.acs.org/JPCA

Article

# Solving Chemistry Problems via an End-to-End Approach: A Proof of Concept

*Published as part of The Journal of Physical Chemistry virtual special issue "Machine Learning in Physical Chemistry".*

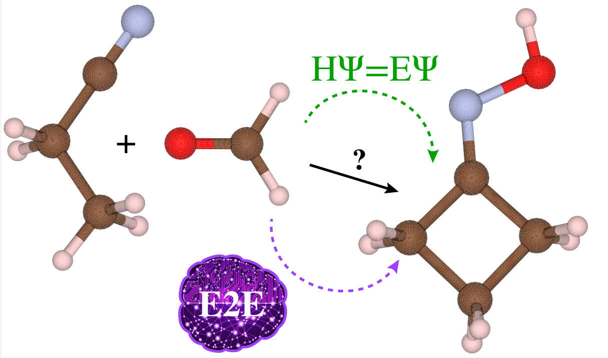
Xiaotong Liu, Tianfu Zhang, Tao Yang, Xiulei Liu, Xin Song, Yong Yang, Ning Li,
Gian-Marco Rignanese,* Yongwang Li, and Xiaodong Wen*

Cite This: *J. Phys. Chem. A* 2020, 124, 8866−8873

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

**ABSTRACT:** Traditionally, chemistry problems are solved by means of a deductive approach. The question to be addressed is typically related to the value of a property that is either measured experimentally, computed using quantum-chemistry software, or (more recently) predicted using a machine-learned model. In this paper, we demonstrate that an inductive approach can be adopted using End-to-End (E2E) machine learning. This approach is illustrated for tackling the following chemistry problems: (i) determine the fully coordinated (FC) and undercoordinated (UC) atoms in a molecule with one missing atom, (ii) identify the type of atom that is missing in such an incomplete molecule, and (iii) predict the direction of a reaction between two molecules according to an existing dataset. The E2E approach leads to accuracies higher than 99%, 98%, and 93% for these three problems, respectively. Finally, in order to achieve such accuracies, a descriptor for the molecules, called bag of clusters, is introduced and compared with a series previously proposed descriptors, highlighting a series of advantages.

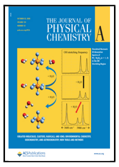


## 1. INTRODUCTION

In recent years, deep learning has been widely used in many fields, such as speech recognition, visual object recognition, object detection, drug discovery, and genomics.[1] In this framework, End-to-End (E2E) deep learning is one of the most exciting new developments. Traditional machine learning is a multistage procedure involving processing systems, which manually extract features from the raw data, and learning systems. In contrast, E2E deep learning ignores all of these stages and replaces them with a single neural network that outputs complex data types directly from the original raw features. For example, for speech recognition (something like Siri or Google Assistant), the traditional method that has been used for a long time and is still used today is to break the audio signal into phonemes (the fundamental building sound units), which can then used as features for the model generating the transcript. In contrast, the E2E approach goes straight from the audio waveform to the written transcript without going through the complex process of feature extraction.[2] Another example is self-driving cars. While the traditional approach is a very cumbersome process involving a lot of steps (getting images, identifying the types and positions of objects, self-positioning, traffic signal understanding, and using this information to calculate the trajectory, which can control the steering on the basis of a series of human made rules), the E2E approach goes straight from the images to steering and long-term planning.[3] It trains an artificial intelligence (AI) driver entirely from sensor input data and feedback from a human expert.

E2E deep learning is especially used in the field of perception and control[4] or object detection.[5] It usually performs better than traditional machine learning. In 2012, at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), the deep convolutional neural network alexNet[6] made the top-five classification error rate drop from 28.2% (best performance in 2010) to 16.4%.[7] Future refinements improved the results to a performance level of about 3%, which is better than that of humans, proving that E2E learning is a reliable way to solve a problem as long as enough labeled data are available. There are also lots of excellent works about E2E learning in the fields of music audio[8] and speech recognition.[9]

**Received:** July 10, 2020
**Revised:** September 30, 2020
**Published:** October 13, 2020



In chemistry, there have been numerous reports about the possible use of machine learning (see refs 10−12 for some recent reviews). Most of these machine learning studies consisted of the regression of properties either measured experimentally or calculated using quantum chemistry followed by the prediction of those properties for various new substances.[13−18] Often good results were achieved with error values lower than the widely accepted thresholds for chemical accuracy.[19] For example, the prediction error in the internal energy at 0 K can be as low as 9 meV/atom.[14] All of these very successful machine learning applications, however, followed an indirect route for solving chemistry problems, just like traditional empirical or quantum chemistry approaches. Indeed, the initial problem to be addressed is first related to some properties that can be measured, calculated with a quantum-chemistry simulation, or predicted by a machine learning model. Then the solution of the problem is determined from the values of those properties. However, such an indirect approach may show serious limitations. For instance, it was recently found by Bartel et al. that reasonably accurate predictions of formation energy do not imply accurate predictions of stability.[20]

In contrast, the E2E approach in this context would be a direct route from a given representation of the molecule (the raw data) to the results without resorting any intermediate property (measured or computed). This obviously raises the question of how the molecule can be introduced into a machine learning model, which can basically only process data in the form of vectors. While two-dimensional (2D) or three-dimensional (3D) images (consisting of pixels or voxels) are naturally represented by a vector, the structure of a molecule needs to be transformed before it can be fed to a neural network. Different from previous E2E attempts in the field of chemistry, which generally relied on the Simplified Molecular Input Line Entry Specification (SMILES) or other line notations of chemical structures[21−23] based on Natural Language Processing (NLP), we rather start from the 3D specification of the molecule. Many sophisticated descriptors have been developed to perform a regression from this representation. However, as we shall see, they are not totally suitable for an E2E approach. Therefore, we propose to use a descriptor called a bag of clusters (BoC), which reflects the chemistry information in a simple way. It fits very well in an E2E approach and can solve a series of chemistry problems with very high accuracy.

### 1.1. Chemistry Problems Addressed.

In this paper, three typical chemistry problems (illustrated in Figure 1) are addressed using an E2E approach:

1. determine the fully coordinated (FC) and under-coordinated (UC) atoms in a molecule with one missing atom;
2. identify the type of atom that is missing in such a molecule;
3. predict the direction of a reaction between two molecules according to an existing dataset.

The first two problems are related to one of the most fundamental concepts in chemistry, namely, bonding. Teaching students how to recognize typical bonding patterns is still the subject of various studies in chemical education.[24−26] It is thus interesting to understand what can be taught to a machine. The last problem is also among the major topics in chemistry.

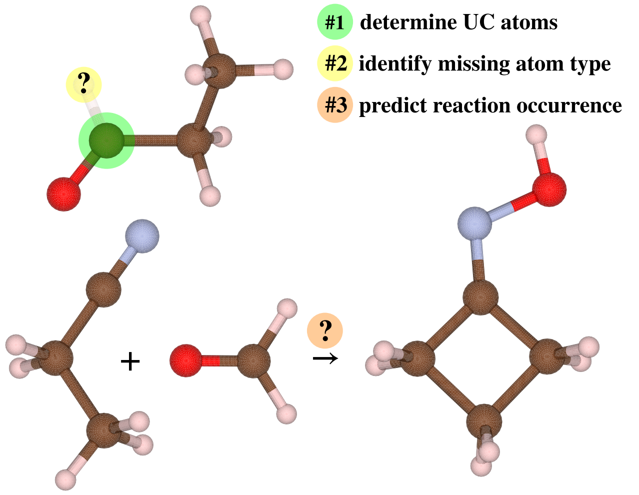


**Figure 1.** Graphical representation of the three chemistry problems addressed by an E2E approach.

## 2. METHODS

### 2.1. Traditional Approach.

These three problems are usually addressed by rule-based and quantum chemistry approaches and more recently by machine learning.[27,28] For problems 1 and 2, a reasonable quantum chemistry approach could be as follows. One would consider the different possibilities for the missing element (C, H, O, N, or F) one at a time. For each of them, the atom would be added close to the molecule, and then a global optimization of its position would be performed while the atoms of the incomplete molecule are kept fixed. The possibility with the highest binding energy would provide the type of the missing atom and even its position, and from the latter it would be possible to identify the UC and FC atoms. For problem 3, the reaction direction could be assessed by computing the energies of the reactants (A and B) and the product (C). In short, all three problems could be related to minimization of the energy, which could in turn be calculated using any kind of human designed approximation (from ab initio to empirical methods). The machine learning approaches used to date have simply aimed at replacing this latter calculation step by a model.

### 2.2. E2E Approach.

Our E2E workflow for solving these three problems starts by establishing a direct one-to-one relation between a given representation of the molecules and the results. A machine learning model is then trained to infer the answers from the descriptor of a given molecule. Specific training and test sets were generated starting from the QM9 dataset[29,30] (see section 4.1). For the representation of the molecules, we first considered commonly used descriptors (their complete definitions can be found in Supporting Information (SI) section A): the standard Coulomb matrix (CM),[31] the standard distance matrix (DM), the modified distance matrix with the proton number (DMP), the combined Coulomb/distance matrix (CDM), the combined Coulomb/inverse distance matrix (CIM), the coordinate data with the atomic number (XYZ), and graph neural network (GNN).[32] Because of the limitations of these descriptors, we also introduced a descriptor named bag of clusters (BoC), which is presented below.

### 2.3. Bag of Clusters.

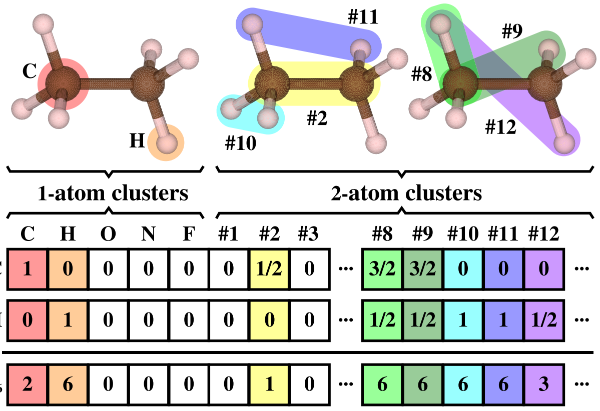
Besides the seven reference descriptors presented in SI section A, we used another one, called bag of clusters, to solve the three chemistry problems in a way that more resembles human intuition. In contrast with the previous ones, the BoC descriptor is not based specifically on the pair distances. It is actually inspired by the cluster expansion (CE) approach[33,34] and the bag of bonds (BoB) concept.[35]

The core concept of CE is to express the physical properties of materials as functions of the atomic configuration. The latter is described in terms of clusters of atoms of fixed shape to which an effective interaction is associated. It can be demonstrated that there is an exact infinite expansion for any physical property.[36,37] In practice, sufficient accuracy is often reached by limiting the expansion to clusters with a small number of atoms (e.g., one-, two-, and three-body clusters) that are relatively compact in size (e.g., 5−7 Å in diameter).

The BoB concept[35] is based on the bag of words (BoW) descriptor used in NLP. The latter encodes the frequencies of occurrence of words in text and is used for solving classification problems. Similarly, within BoB the molecules are described by a vector composed of bags, each representing a particular bond type (C−C, C−N, etc.). The entries in every bag are computed as $Z_iZ_j/|\mathbf{R}_i - \mathbf{R}_j|$ and sorted according to their magnitude. The vector is then obtained by concatenating all bags of bonds, padding each bag with zeros to give the bags equal sizes and to allow for dealing with other molecules with larger bags.

In BoC, the molecules are also described by a vector composed of bags, but now each bag represents a particular cluster of atoms of fixed shape, and each entry is simply the number of occurrences of each cluster. Here only clusters consisting of one or two atoms are considered, but this could easily be extended to larger clusters if needed. Since the molecules of the QM9 database are composed of five different types of atoms (C, H, O, N, and F), there are five different one-atom clusters. The different two-atom clusters are obtained by considering the 15 ($C_5^2$ + 5) different atom pairs generated by the five different atoms and distinguishing their different shapes, which depend only on the pair distance $d$. In fact, for a given pair, the different shapes are established by clustering (in the machine learning sense). The distribution of the distances is illustrated in Figure S1 for each of the atom pairs in the QM9 database. The counts are established using an interval of $10^{-3}$ Å, focusing on pair distances smaller than 5 Å. The counts that are less than 5% of the maximum count for that particular pair are considered as noise and thus are omitted. The standard $K$-means algorithm is used to process the data, requesting a number of groups $N_g$ between 2 and 20. The quality of the clustering is assessed through the goodness of variance fit (GVF), which ranges from 0 (bad) to 1 (perfect). The GVF is defined in terms of the sum of squared deviations from the class mean (SDCM) and the sum of squared deviations from the array mean (SDAM) as $1 - \frac{\text{SDCM}}{\text{SDAM}}$.[38] To prevent dimensional explosion, the smallest $N_g$ that provides a GVF above 99.5% is selected. Since the counts for the F−O pair are too small and there is only one cluster for the F−F pair, these two pairs are discarded from the descriptor. In the end, we obtain 92 different two-atom clusters, as shown in Table S1. Therefore, the molecules are described by a vector with 97 elements (five bags for the one-atom clusters and 92 bags for the two-atom clusters).

As an example, in Figure 2 we present the BoC descriptor of the ethane molecule. Interestingly, it can be easily split into the contributions of each of its atoms. For a particular atom, the latter is obtained by indicating 1 in the vector for the corresponding one-atom cluster and dividing by two the number of pairs in which this atom is involved. This possibility is used here to address problem 1.



| | C | H | O | N | F | #1 | #2 | #3 | ··· | #8 | #9 | #10 | #11 | #12 | ··· |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2×C | 1 | 0 | 0 | 0 | 0 | 0 | 1/2 | 0 | ··· | 3/2 | 3/2 | 0 | 0 | 0 | ··· |
| 6×H | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | ··· | 1/2 | 1/2 | 1 | 1 | 1/2 | ··· |
| $C_2H_6$ | 2 | 6 | 0 | 0 | 0 | 0 | 1 | 0 | ··· | 6 | 6 | 6 | 6 | 3 | ··· |

**Figure 2.** Determination of the BoC for the ethane molecule with the different atomic contributions. The molecule is composed of two single C atoms and six single H atoms. Hence, the five first components of the vector related to the one-atom clusters are [2, 6, 0, 0, 0]. The molecule shows one C−C pair with $d$ = 1.542 Å, which falls into cluster #2 in Table S1. It also displays six C−H pairs with $d$ = 1.092 Å (cluster #8), six C−H pairs with $d$ = 2.172 Å (cluster #9), six H−H pairs with $d$ = 1.765 Å (cluster #10), six H−H pairs with $d$ = 2.528 Å (cluster #11), and three H−H pairs with $d$ = 3.083 Å (cluster #12). Hence, the final BoC vector for the molecule is [2, 6, 0, 0, 0, 1, 0, 0, 0, 0, 0, 6, 6, 6, 6, 3, 0, ..., 0]. In terms of atomic contributions, each C atom is involved in one C−C pair with $d$ = 1.542 Å (cluster #2), three C−H pairs with $d$ = 1.092 Å (cluster #8), and three C−H pairs with $d$ = 2.172 Å (cluster #9). The contribution of each C atom to the BoC vector is [1, 0, 0, 0, 0, 0, 1/2, 0, 0, 0, 0, 0, 3/2, 3/2, 0, 0, 0, ..., 0]. Similarly, the contribution of each H atom to the BoC vector is [0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1/2, 1/2, 1, 1, 1/2, 0, ..., 0].

Even more excitingly, because there is no zero padding in BoC, the vector for any molecule can be subtracted from (or added to) that of another molecule. We refer to this possibility as ΔBoC. It will be used here to address problem 3.

It should be noted that upon careful analysis of Table S1 and Figure S1, we can see that the clustering result is not perfect. For instance, it does not distinguish C−C double and triple bonds. This is probably the case because the distance difference between the C−C double and triple bonds is too small and the count of double bonds in the QM9 distribution is quite low. We verified that if we preprocess the data by, e.g., converting $X$ versus $Y$ to $1 - \frac{1}{X^2}$ versus $\log(Y)$, the C−C double and triple bonds can be distinguished by the automatic clustering. Since our focus is on E2E, we prefer to try to feed raw data to the machine with minimal manual intervention, and hence, we do not apply this data preprocessing.

We also note that after developing BoC, we became aware of the existence of a similar descriptor called bag of fragments (BoF) that was also introduced recently[39,40] and is inspired by the BoW concept. The main difference is that in BoF the clusters consist only of bonded atoms. We verified that the reduction of the BoC basis is bad for the learning process. There is no need to introduce the concept of bond to the algorithm to achieve an E2E process. In fact, BoC can be seen as a generalization of BoF: for instance, it also includes two-atom clusters involving second-nearest neighbors (2NN), third-nearest neighbors (3NN), etc. However, in some extreme cases such as chiral molecules, additional rules are needed to further improve the distinguishing ability of BoC descriptors. BoC also presents some similarities to the work of Natarajan and Van der Ven,[37] who extended the CE approach by expressing the energy of a crystal as a sum of site energies while allowing for the site energies to be nonlinear functions of the local correlation functions. However, the E2E philosophy adopted here is quite different since we are not trying to machine-learn a specific property (e.g., the energy). Another interesting difference with all of those previous works lies in the method of clustering (in the machine learning sense) employed to automatically determine the clusters used in the descriptor. It would actually be interesting to determine which cluster type (two-atom, three-atom, etc.) and range (2NN, 3NN, etc.) are needed to improve the accuracy. One could also evaluate other clustering methods (e.g., mean shift) and test different clustering scores (e.g., silhouette score or Calinski−Harabasz score). However, that goes beyond the scope of the present paper and will be explored in the future.

## 3. RESULTS AND DISCUSSION

For problem 1, we need to split the descriptors into the contributions of each of the individual atoms. As already explained, this can be done easily for BoC. For DM, DMP, and CM, the different rows of the descriptor matrix (see SI section A) can be understood as atomic descriptors too. For XYZ, we can simply take the coordinates and atomic number of each atom as an atomic descriptor. For GNN, we have not yet figured out a way to obtain such an atomic descriptor, so that method could not be used to address this problem. For most of the descriptors apart from XYZ, the total accuracy is excellent (99.81% for BoC, 99.74% for DMP, and 99.66% for CM). For XYZ, the accuracy falls to 87.8%. In fact, as can be seen in SI section C (Tables S2−S4), this corresponds to guessing FC for every site: FC/(UC + FC) = 311441/354696 = 87.8%. The poor performance of XYZ can simply be understood by the fact that the atomic descriptor does not contain any information about the relations of that atom to the other ones. Indeed, the DM and DMP descriptors contain information about the distances of each atom to all of the other ones; the CM descriptor also includes information about the atomic numbers of the other atoms, and the BoC descriptor indicates which clusters the atom forms with the other ones.

In Figure 3, we show a partial enlarged view of the receiver operating characteristic (ROC) curves for BoC, DMP, and CM (detailed data are given in SI section C). In ROC evaluation, the curve that is closer to left-upper point of the figure represents a better-performing classifier for handling an unbalanced binary classification. As problem 1 is not a difficult challenge, most of the descriptors give very good accuracy results. Among them, BoC performs better than the others. The breakdown of species of the UC−FC classification are shown in Table S5. Most of errors originate from C and N atoms, around which the coordination environment is more complicated. H, F, and O atoms are easy to spot because of their small coordination number in most cases.

It is interesting to compare the performance of our E2E approach with that of the traditional electron counting (EC) method. The accuracy result tagged with the EC method is 90−96% and fluctuates with the determination of the bond distance (implementation details can be found in SI), which is far worse than E2E (99.8%). The EC method faces at least three problems. The first problem is the determination of the bond distance. We need to tell the machine what range of distances between two atoms can be viewed as a bond. This acts as a sensitive hyperparameter that influences the result a lot. Second, it is a multisolution problem to complete an incomplete formula of unsaturated molecular structure. For example, considering a propylene structure with one H atom removed from $CH_3$, we cannot determine the location of the double bond of the remaining $CH_2CHCH_2$ part if we rely only on the electron counting information. Third, the tagging error of the EC method doubles. One tagging error of an atom with more electrons inevitably results in a prediction error of another atom with fewer electrons.

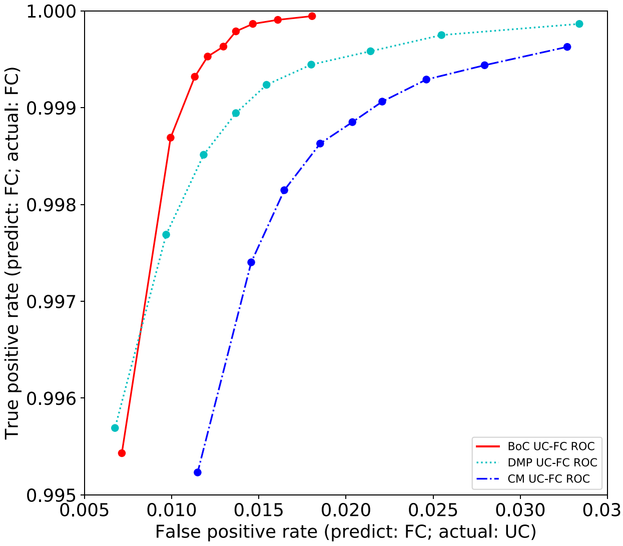


**Figure 3.** Partial enlarged view of the ROC curves for UC−FC classification.

This UC−FC tagging can be exploited to find the position of missing atom. In SI section E, we show that taking the average position of the UC atoms already provides a better guess than any direct regression based on the different descriptors. Furthermore, we use a conjugate gradient optimizer to refine the position taking into account the fact that the distance $d_i$ between atom $i$ and the missing one should be bigger (respectively smaller) than $D_1$ when atom $i$ is FC (respectively UC). However, when atom $i$ is UC, $d_i$ should be bigger than $D_0$ = 1.0 Å in order to avoid being too close to any atoms, as shown in Figure 4. To this end, we define the loss function of the optimizer as follows:

$$\begin{aligned}\text{Loss} &= \sum_i^{\text{FC}} [\min(d_i, D_1) - D_1]^2 \\ &+ \sum_i^{\text{UC}} \{[\max(d_i, D_1) - D_1]^2 \\ &+ [\min(d_i, D_0) - D_0]^2\}\end{aligned} \quad (1)$$

For problem 2, eight different descriptors are used to find the missing atom type. As shown in Figure 5, the molecular BoC descriptor shows the best performance, with an accuracy of about 98.2%. The descriptors of the CM series reach about 91−93% accuracy (91.9% for CM, 92.4% for CDM, and 93.0% for CIM), and GNN is a little better (94.6%). The atom coordinates and the descriptors of the DM series reach even

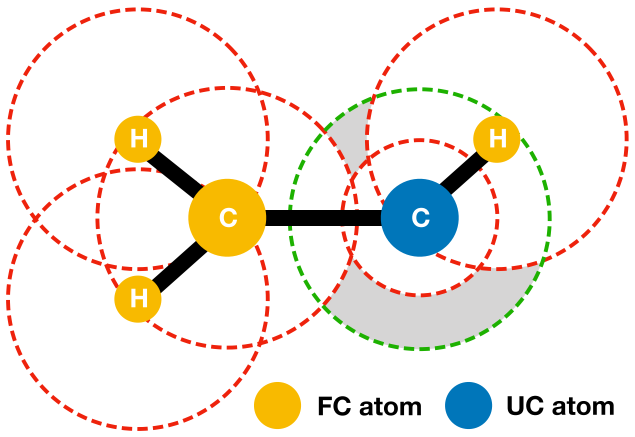


**Figure 4.** Schematic diagram of the loss function. The red spheres indicate exclusion zones: their radius is either $D_0$ for UC atoms or $D_1$ for FC atoms. The green sphere indicates where the missing atom should be; its radius is $D_1$. The missing atom position is constrained to be in the gray area.

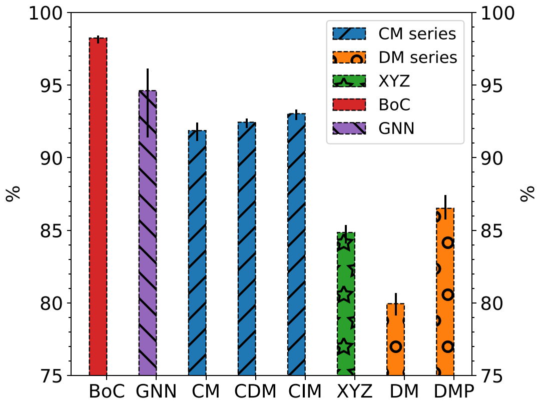


**Figure 5.** Prediction accuracies for the missing atom type.

lower accuracy (84.8% for XYZ, 79.9% for DM, and 86.5% for DMP). Molecular BoC seems to express the relations between the atoms in a better way. This enables its unique advantage in solving molecular-level problems, because finding the missing atom type can also be understood as classifying the remaining molecular group.

Small modifications of the Coulomb matrix (CDM and CIM) improve the performance. However, this improvement is not significant. This is understandable since the neural network can easily fit the simple multiplication and division operation of the input. DMP is significantly improved compared with DM. Obviously, the extra information about the element types (i.e., the atomic numbers) is very helpful in understanding the whole structure of the molecule. DMP is slightly better than coordinate data that also contains the atomic numbers. This indicates that simple extraction of specific distance information helps in this problem. Generally, we believe that the matrices from the DM and CM series are not able to express the essence of the problem when dealing with nonregression. GNN has recently attracted considerable attention in machine learning. Here it gives better accuracy than the other reference descriptors. The errors are comparable to those for BoC in certain cases, but the variance is very large. GNN may need further fine-tuning since it is really a different approach compared with the others. As can be seen in Table 1, the breakdown of atoms shows that the performance is good for C, H, O, and N but poor for F (the accuracy is only 12/(26 + 1 + 12) = 30.7%), which is disturbed by the large amount of H with similar electron property.

**Table 1. Breakdown of the BoC Prediction Results in Problem 2**

| prediction | | | | | |
|---|---|---|---|---|---|
| H | C | O | F | N | actual |
| 10412 | 7 | 14 | 11 | 5 | H |
| 0 | 7305 | 20 | 0 | 55 | C |
| 66 | 17 | 1515 | 1 | 45 | O |
| 26 | 0 | 1 | 12 | 0 | F |
| 24 | 61 | 46 | 0 | 1188 | N |

As we have seen, for problems 1 and 2 BoC shows the best performance. Although the performance of the other descriptors is slightly worse, they can also be used to address these two E2E problems (e.g., using DMP for problem 1 and GNN or CIM for problem 2). In contrast, for problem 3, BoC shows the unique advantage that the number of elements in the descriptor vector does not increase drastically when dealing with more atoms and more complex systems. This confers to BoC a wide application potential in the prediction of large molecular structures. Indeed, let us consider a system with $N$ atoms. For each atom added to the system, the CM and DM series matrix representations require $2N + 1$ additional elements, the XYZ representation requires four additional elements ($x$, $y$, $z$, and $Z$), and the BoC requires none (unless the additional atom type was not among those considered previously, in which case new clusters would need to be added to the list). Considering a large biological molecule such as hemoglobin (the oxygen carrier in red blood cells), which is made up of more than 10k atoms,[41] any CM and DM series matrix descriptor will contain more than 100 million elements. This is much larger than the number of pixels in a 720p (1280 × 720) image. Dealing with such a large number of elements would really be very expensive. Hence, the BoC approach is a much better choice. Furthermore, BoC does not require any zero padding for small molecules. This makes it possible within BoC to compute the difference between (or the sum of) the representations of two molecules, ΔBoC. ΔBoC is particularly useful to address problem 3. Indeed, as an input for the machine learning model, we use the difference between the BoC of the product (C) and the sum of the BoCs of the reactants (A and B): ΔBoC = Boc(C) − [BoC(A) + BoC(B)]. In Figure 6, the results obtained with this E2E approach are presented for four datasets generated using different tagging algorithms (see SI section F). The results obtained for the random assignment (Figure 6d) are completely wrong, indicating that the machine learning model cannot find any chemistry in that case. This is in clear contrast with the results obtained for the other three datasets.

All of these results demonstrate that E2E machine learning is able to solve different chemistry problems by adopting a direct route similar to human intuition without resorting to any intermediate property (measured or computed). A simple neural network with several dense layers is sufficient to solve these problems as long as there are enough training data.

## 4. COMPUTATIONAL DETAILS

**4.1. Datasets.** A dataset specific to each problem was generated starting from the QM9 dataset,[29,30] which contains

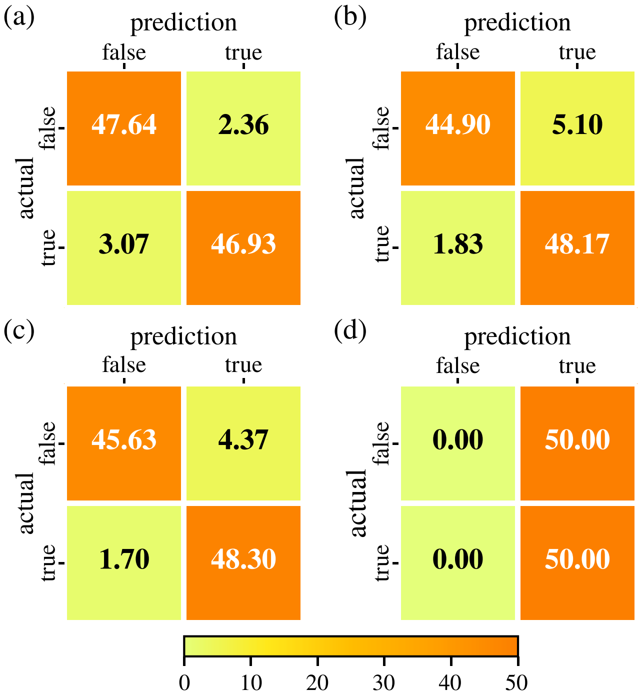


**Figure 6.** E2E classification of the reactions starting from different datasets: direction tagging (a) based on Gibbs free energy, (b) based on HOMO−LUMO, (c) based on both, or (d) at random.

134k small stable molecules made of five elements (C, H, O, N, and F).

For problems 1 and 2, a dataset was first generated by deleting randomly one atom from each molecule. This produced 134k high-quality data relating an incomplete molecule to the type (label) of the missing atom, which could be used right away for problem 2. For problem 1, however, it was still necessary to label the different atoms as being either FC or UC for each incomplete molecule. To this end, different algorithms were tested (as described in SI section G). Here we focus on the results obtained with the following procedure. First, we set a radius for the first-nearest-neighbor (1NN) shell for the different atom types according their typical maximum bond lengths as extracted from ref 42: $r^{\mathrm{C}}_{\mathrm{1NN}}$ = 1.86 Å, $r^{\mathrm{H}}_{\mathrm{1NN}}$ = 1.65 Å, $r^{\mathrm{O}}_{\mathrm{1NN}}$ = 1.86 Å, $r^{\mathrm{N}}_{\mathrm{1NN}}$ = 1.83 Å, $r^{\mathrm{F}}_{\mathrm{1NN}}$ = 1.85 Å. Then we tag as UC all of the atoms for which the missing atom (if present) would have been in their 1NN shell (i.e., within the sphere of radius $r_{\mathrm{1NN}}$ centered on them). At the end of the procedure, 87.8% (respectively 12.2%) of the atoms are FC (respectively UC) in the whole dataset.

For problem 3, the dataset was generated by selecting triplets of molecules from the QM9 database, two of them being considered as the reactants (A and B) and one as the product (C) of the chemical reaction A + B → C and by imposing that the total number of atoms of each type is the same on both sides of the reaction (i.e., for the reactants and the product). As a result, 136M (136 821 740) triplets of molecules were identified as potential candidates for a reaction. Then four different methods (as described in SI section F) were used to determine whether the reaction would occur. Alternatively, one might have used a database of known reactions, though it should be complemented by a list of reactions known not to occur. Finally, after all of the reactions were labeled, a limited dataset was built by randomly selecting 5M reactions tagged as occurring and 5M ones tagged as not occurring.

For problems 1 and 2, the 134k incomplete molecules obtained from QM9 were divided into three sets: 99k were used for training, 11k were employed for validation, and the rest (about 24k) were utilized for testing. For problem 1, the 134k incomplete molecules generate 2.27M atoms tagged as UC or FC. Basically, they were obtained from the previously divided datasets, so that the training set included 1.68M data, the validation set consisted of 0.19M data, and the test set comprised 0.4M data. For problem 3, the training set was taken to be 8 times as large as the validation and test sets, and four different datasets were considered depending on the labeling scheme adopted for assessing the direction of the reaction.

### 4.2. Machine Learning Model.

The molecules were handled using the Atomic Simulation Environment (ASE).[43] For problems 1 and 2, the neural networks were built using TensorFlow,[44] except for GNN, for which we utilized a modified version of the open source code[32] based on PyTorch.[45] In problem 3, we also employed PyTorch. Since the first five descriptors (CM, DM, DMP, CDM, and CIM) are 2D, they required an extra flattening layer to transform the format of the data from a 2D tensor to a 1D tensor. Afterward, the 1D tensor flowed through a sequence of four fully connected layers. Each dense layer was composed of 256 units with rectified linear unit (ReLU)[46] as the activation function.

For problems 1 and 3, the individual atomic contributions to the BoC vector and ΔBoC were used as the input with the same shape. The output of the neural network is a scalar with a sigmoid activation function. The loss function is the binary cross-entropy.

To find out the missing atom type (problem 2), an extra fully connected layer with five nodes and a softmax activation function was added at the end of the sequence. In our E2E method, this problem was understood as a molecular challenge, and the molecular BoCs were used as the input. The loss function of the model is the sparse categorical cross-entropy.

We used the Adam optimizer[47] for all neural network optimization.

## 5. CONCLUSIONS

We have demonstrated that it is possible to imitate human intuition for learning and solving chemistry problems in an E2E approach, which is inductive rather than deductive. This approach leads to accuracies higher than 99%, 98%, and 93% for tackling the following important chemistry problems: (i) determine the fully coordinated and undercoordinated atoms in a molecule with one missing atom, (ii) identify the type of the atom that is missing in such an incomplete molecule, and (iii) predict the direction of a reaction between two molecules according to an existing dataset. To achieve such accuracies, we have used a descriptor for the molecules called bag of clusters. We have compared it to a series of previously proposed descriptors, highlighting its advantages: it is a nonzero padding descriptor, and it can be used at different levels (atomic contributions, molecule, or sum and difference). We believe that our findings will generate new ideas and hence open a new way for machine learning in chemistry.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpca.0c06319.

> Common descriptors of molecular structures, definition of the clusters for BoC, UC−FC classification data for ROC testing, analysis of the performance of the classifier (BoC descriptor) on different elements, missing atom position inference, reaction tagging algorithms, UC−FC tagging algorithm, and explanation of technical terms (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Authors

**Xiaodong Wen** − *State Key Laboratory of Coal Conversion, Institute of Coal Chemistry, Chinese Academy of Sciences, Taiyuan 030001, P. R. China; National Energy Center for Coal to Liquids, Synfuels China Co., Ltd., Beijing 101400, P. R. China; Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Information Science and Technology University, Beijing 100101, P. R. China;* orcid.org/0000-0001-5626-8581; Email: wxd@sxicc.ac.cn

**Gian-Marco Rignanese** − *UCLouvain, Louvain-la-Neuve 1348, Belgium;* orcid.org/0000-0002-1422-1205; Email: gian-marco.rignanese@uclouvain.be

### Authors

**Xiaotong Liu** − *State Key Laboratory of Coal Conversion, Institute of Coal Chemistry, Chinese Academy of Sciences, Taiyuan 030001, P. R. China; National Energy Center for Coal to Liquids, Synfuels China Co., Ltd., Beijing 101400, P. R. China; University of Chinese Academy of Sciences, Beijing 100049, P. R. China; Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Information Science and Technology University, Beijing 100101, P. R. China*

**Tianfu Zhang** − *State Key Laboratory of Coal Conversion, Institute of Coal Chemistry, Chinese Academy of Sciences, Taiyuan 030001, P. R. China; National Energy Center for Coal to Liquids, Synfuels China Co., Ltd., Beijing 101400, P. R. China*

**Tao Yang** − *Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Information Science and Technology University, Beijing 100101, P. R. China*

**Xiulei Liu** − *Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Information Science and Technology University, Beijing 100101, P. R. China*

**Xin Song** − *National Energy Center for Coal to Liquids, Synfuels China Co., Ltd., Beijing 101400, P. R. China*

**Yong Yang** − *State Key Laboratory of Coal Conversion, Institute of Coal Chemistry, Chinese Academy of Sciences, Taiyuan 030001, P. R. China; National Energy Center for Coal to Liquids, Synfuels China Co., Ltd., Beijing 101400, P. R. China*

**Ning Li** − *Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Information Science and Technology University, Beijing 100101, P. R. China*

**Yongwang Li** − *State Key Laboratory of Coal Conversion, Institute of Coal Chemistry, Chinese Academy of Sciences, Taiyuan 030001, P. R. China; National Energy Center for Coal to Liquids, Synfuels China Co., Ltd., Beijing 101400, P. R. China; Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Information Science and Technology University, Beijing 100101, P. R. China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpca.0c06319

### Notes

The authors declare no competing financial interest.
Source code is available on https://github.com/liuxiaotong15/e2e_public for problems 1 and 2, https://github.com/liuxiaotong15/e2e_reaction_public for problems 3, and https://github.com/liuxiaotong15/qm9_electron_counting for the electron counting comparison.

## ■ ACKNOWLEDGMENTS

The authors are grateful for financial support from the National Natural Science Foundation of China (21972157 and 21872164) and the Key R&D Plan of Beijing Municipal Science and Technology Commission (Z181100005118014) as well as funding support from Synfuels China, Co. Ltd. We also thank Prof. Roald Hoffmann (Cornell University) for providing us with valuable comments on the work and acknowledge the Innovation Foundation of the Institute of Coal Chemistry, Chinese Academy of Sciences, the Hundred-Talent Program of the Chinese Academy of Sciences, the Shanxi Hundred-Talent Program, and the National Thousand Young Talents Program of China.

## ■ REFERENCES

(1) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436−444.

(2) Wang, D.; Wang, X.; Lv, S. An Overview of End-to-End Automatic Speech Recognition. *Symmetry* **2019**, *11*, 1018.

(3) Bojarski, M.; Del Testa, D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L. D.; Monfort, M.; Muller, U.; Zhang, J. End to End Learning for Self-Driving Cars. *arXiv (Computer Science.Computer Vision and Pattern Recognition)*, April 25, 2016, 1604.07316, ver. 1. https://arxiv.org/abs/1604.07316 (accessed 2020-03-21).

(4) Levine, S.; Finn, C.; Darrell, T.; Abbeel, P. End-to-End Training of Deep Visuomotor Policies. *J. Mach. Learn. Res.* **2016**, *17*, 1−40.

(5) Zhou, Y.; Tuzel, O. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. *arXiv (Computer Science.Computer Vision and Pattern Recognition)*, November 17, 2017, 1711.06396, ver. 1. http://arxiv.org/abs/1711.06396 (accessed 2020-03-21).

(6) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*; Curran Associates, 2012; pp 1097−1105.

(7) Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. et al. ImageNet Large Scale Visual Recognition Challenge. *arXiv (Computer Science.Computer Vision and Pattern Recognition)*, January 30, 2015, 1409.0575, ver. 3. https://arxiv.org/abs/1409.0575 (accessed 2020-03-21).

(8) Dieleman, S.; Schrauwen, B. End-to-End Learning for Music Audio. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* **2014**, 6964−6968.

(9) Amodei, D.; Anubhai, R.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Chen, J.; Chrzanowski, M.; Coates, A.; Diamos, G. et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. *arXiv (Computer Science.Computation and Language)*, December 8, 2015, 1512.02595, ver. 1. http://arxiv.org/abs/1512.02595 (accessed 2020-03-21).

(10) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **2019**, *5*, 83.

(11) Cova, T. F. G. G.; Pais, A. A. C. C. Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns. *Front. Chem.* **2019**, *7*, 809.

(12) de Almeida, A. F.; Moreira, R.; Rodrigues, T. Synthetic organic chemistry driven by artificial intelligence. *Nat. Rev. Chem.* **2019**, *3*, 589−604.

(13) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2016**, *2*, 16028.

(14) 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.

(15) Schütt, K. T.; Kessel, P.; Gastegger, M.; Nicoli, K. A.; Tkatchenko, A.; Müller, K.-R. SchNetPack: ADeep Learning Toolbox For Atomistic Systems. *J. Chem. Theory Comput.* **2019**, *15*, 448−455.

(16) Zhang, Z.; Schott, J. A.; Liu, M.; Chen, H.; Lu, X.; Sumpter, B. G.; Fu, J.; Dai, S. Prediction of Carbon Dioxide Adsorption via Deep Learning. *Angew. Chem., Int. Ed.* **2019**, *58*, 259−263.

(17) Bernstein, N.; Bhattarai, B.; Csányi, G.; Drabold, D. A.; Elliott, S. R.; Deringer, V. L. Quantifying Chemical Structure and Machine-Learned Atomic Energies in Amorphous and Liquid Silicon. *Angew. Chem., Int. Ed.* **2019**, *58*, 7057−7061.

(18) Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J. Constant size descriptors for accurate machine learning models of molecular properties. *J. Chem. Phys.* **2018**, *148*, 241718.

(19) 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. *J. Chem. Theory Comput.* **2017**, *13*, 5255−5264.

(20) Bartel, C. J.; Trewartha, A.; Wang, Q.; Dunn, A.; Jain, A.; Ceder, G. A critical examination of compound stability predictions from machine-learned formation energies. *arXiv (Condensed Matter.-Materials Science)*, January 28, 2020, 2001.10591, ver. 1. https://arxiv.org/abs/2001.10591 (accessed 2020-03-21).

(21) Kwon, S.; Yoon, S. End-to-End Representation Learning for Chemical-Chemical Interaction Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2019**, *16*, 1436−1447.

(22) Tsubaki, M.; Tomii, K.; Sese, J. Compound−protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* **2019**, *35*, 309−318.

(23) Verma, N.; Qu, X.; Trozzi, F.; Elsaied, M.; Tao, Y.; Larson, E.; Kraka, E. SSnet - Secondary Structure based End-to-End Learning model for Protein-Ligand Interaction Prediction. *bioRxiv* **2019**, DOI: 10.1101/2019.12.20.884841.

(24) Kiste, A. L.; Hooper, R. G.; Scott, G. E.; Bush, S. D. Atomic Tiles: Manipulative Resources for Exploring Bonding and Molecular Structure. *J. Chem. Educ.* **2016**, *93*, 1900−1903.

(25) Tsaparlis, G.; Pappa, E. T.; Byers, B. Teaching and learning chemical bonding: research-based evidence for misconceptions and conceptual difficulties experienced by students in upper secondary schools and the effect of an enriched text. *Chem. Educ. Res. Pract.* **2018**, *19*, 1253−1269.

(26) Tsaparlis, G.; Pappa, E. T.; Byers, B. Proposed pedagogies for teaching and learning chemical bonding in secondary education. *Chem. Teach. Int.* **2019**, *2*, 20190002.

(27) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725−732.

(28) Fooshee, D.; Mood, A.; Gutman, E.; Tavakoli, M.; Urban, G.; Liu, F.; Huynh, N.; Van Vranken, D.; Baldi, P. Deep learning for chemical reaction prediction. *Mol. Syst. Des. Eng.* **2018**, *3*, 442−452.

(29) 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.

(30) 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.

(31) 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.

(32) Tsubaki, M.; Mizoguchi, T. Fast and Accurate Molecular Property Prediction: Learning Atomic Interactions and Potentials with Neural Networks. *J. Phys. Chem. Lett.* **2018**, *9*, 5733−5741.

(33) Drautz, R.; Singer, R.; Fähnle, M. Cluster Expansion Technique: An Efficient Tool to Search for Ground-State Configurations of Adatoms on Plane Surfaces. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2003**, *67*, 035418.

(34) Huang, W.; Urban, A.; Rong, Z.; Ding, Z.; Luo, C.; Ceder, G. Construction of ground-state preserving sparse lattice models for predictive materials simulations. *npj Comput. Mater.* **2017**, *3*, 30.

(35) 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.

(36) Sanchez, J.; Ducastelle, F.; Gratias, D. Generalized cluster description of multicomponent systems. *Phys. A* **1984**, *128*, 334−350.

(37) Natarajan, A. R.; Van der Ven, A. Machine-learning the configurational energy of multicomponent crystalline solids. *npj Comput. Mater.* **2018**, *4*, 56.

(38) Coulson, M. R. C. In The Matter Of Class Intervals For Choropleth Maps: With Particular Reference To The Work Of George F Jenks. *Cartogr. Int. J. Geogr. Inf. Geovisualization* **1987**, *24*, 16−39.

(39) Hop, P.; Allgood, B.; Yu, J. Geometric Deep Learning Autonomously Learns Chemical Features That Outperform Those Engineered by Domain Experts. *Mol. Pharmaceutics* **2018**, *15*, 4371−4377. PMID: 29863875.

(40) Armitage, J.; Spalek, L. J.; Nguyen, M.; Nikolka, M.; Jacobs, I. E.; Marañón, L.; Nasrallah, I.; Schweicher, G.; Dimov, I.; Simatos, D.; et al. Fragment Graphical Variational AutoEncoding for Screening Molecules with Small Data. *arXiv (Physics.Data Analysis, Statistics and Probability)*, October 30, 2019, 1910.13325, ver. 2. https://arxiv.org/abs/1910.13325 (accessed 2020-03-21).

(41) Kotz, J. C.; Treichel, P. M.; Townsend, J. *Chemistry and Chemical Reactivity*; Cengage Learning, 2012.

(42) *Tables of Interatomic Distances and Configuration in Molecules and Ions*; Sutton, L., Ed.; The Chemical Society: London, 1965.

(43) Hjorth Larsen, A.; Jørgen Mortensen, J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; et al. The Atomic Simulation Environment—a Python Library for Working with Atoms. *J. Phys.: Condens. Matter* **2017**, *29*, 273002.

(44) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv (Computer Science.Distributed, Parallel, and Cluster Computing)*, March 16, 2016, 1603.04467, ver. 2. https://arxiv.org/abs/1603.04467 (accessed 2020-03-21).

(45) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L. et al. In *PyTorch: An Imperative Style, High-Performance Deep Learning Library*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, 2019; pp 8024−8035.

(46) Nair, V.; Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*; Omnipress: Madison, WI, 2010; pp 807−814.

(47) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv (Computer Science.Machine Learning)*, January 30, 2014, 1412.6980, ver. 9. https://arxiv.org/abs/1412.6980 (accessed 2020-03-21).