ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Screening toward the Development of Fingerprints of Atomic Environments Using Bond-Orientational Order Parameters

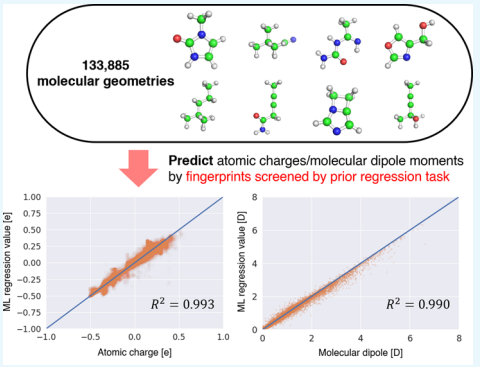
Hideo Doi,[‡] Kazuaki Z. Takahashi,[*,‡] and Takeshi Aoyagi

Cite This: *ACS Omega* 2022, 7, 4606−4613

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: A combination of atomic numbers and bond-orientational order parameters is considered a candidate for a simple representation that involves information on both the atomic species and their positional relation. The 504 candidates are applied as the fingerprint of the molecules stored in QM9, a data set of computed geometric, energetic, electronic, and thermodynamic properties for 133 885 stable small organic molecules made up of carbon, hydrogen, oxygen, nitrogen, and fluorine atoms. To screen the fingerprints, a regression analysis of the atomic charges given by Open Babel was performed by supervised machine learning. The regression results indicate that the 60 fingerprints successfully estimate Open Babel charges. The results of the dipole moments, an example of a property expressed by charge and position, also had a high accuracy in comparison with the values computed from Open Babel charges. Therefore, the screened 60 fingerprints have the potential to precisely describe the chemical and structural information on the atomic environment of molecules.



INTRODUCTION

A representation of atomic environments has become more needed than in the past decade because of the rapid development of material informatics and related technology.[1−13] Chemical representations without real-space information are mainstream because they are compact, lightweight, and usually adopted for molecular informatics. Indeed, some simple fingerprints, such as the simplified molecular-input line-entry system (SMILES)[14] and SYBYL line notation (SLN),[15] are useful, for instance, for similarity measurements between molecules as the simplest example. Structure representations have been developed separately from the chemical representations, e.g., Voronoi polyhedron of a central atom, angular Fourier series,[1] partial and generalized radial distribution functions, and the bond-orientational order parameter.[16] These were originally not intended for machine learning (ML) applications but are promising. However, separating the chemistry and structure of molecules is open to debate. For example, the Hohenberg−Kohn theorem,[17−19] which states that all physical quantities of a system can in principle be calculated from the electron density of the ground state, implies that there is an inseparable relationship between information about the electronic properties of atoms and the direct function of the geometric location of nuclei. Therefore, the species of atoms and their positional relations can hardly be separated to provide information on the atomic environment. This fact means that there is a need to consider a more proper representation of both chemistry and structure. One of the simplest ideas for such a representation is a combination of the existing chemical and structural representations. For example, a Coulomb matrix[20] is a generalization of an adjacency matrix representation and has been extended as the Ewald sum and the Sine matrix.[5] More advanced methods include histograms of distances (HDs), HD angles (HDAs), HDA dihedrals,[21] bag of bonds,[22] and the smooth overlap of atomic positions.[23]

The bond-orientational order parameter that uses spherical harmonics exhibits a highly generic performance for describing molecular local structures. It was first developed by Steinhardt and co-workers to investigate the structures of supercooled liquids and metallic glasses.[16] Lechner and co-workers improved the bond-orientational order parameter by locally averaging neighborhood molecules.[24] Their parameters successfully distinguished body-centered cubic (bcc), face-centered cubic (fcc), hexagonal close-packed (hcp), and liquid-like local structures and have been widely used to classify complex local structures.[25−32] Further modifications have been attempted to extend the entire accuracy or for a specific use.[33,34] The capability of their parameters has been reported for the identification of the crystal-like structures of a

Received: November 22, 2021
Accepted: December 29, 2021
Published: January 24, 2022

https://doi.org/10.1021/acsomega.1c06587
ACS Omega 2022, 7, 4606−4613

Lennard−Jones fluid,[35−37] water,[38−45] polyethylene,[46] and a liquid crystal and its polymer.[33,34] Such a high sensitivity for local structures is desirable for the precise expression of the structural contributions of the atomic environment.

In this work, we consider a combination of atomic numbers and bond-orientational order parameters as a candidate for the simple representation that involves information on both the atomic species and their positional relationship. This is applied as the fingerprint of the molecules stored in the QM9 data set,[47,48] which contains geometric, energetic, electronic, and thermodynamic properties for 133 885 stable small organic molecules made up of carbon, hydrogen, oxygen, nitrogen, and fluorine. A total of 504 fingerprints were systematically designed and then screened by the supervised ML for the regression analysis for the atomic charges of molecules. The results indicate that the 60 fingerprints successfully estimate the atomic charges. These fingerprints were used to compute the dipole moments of molecules as an example of a property expressed by both charge and position, and a high accuracy was obtained compared to the values computed from atomic charges. Therefore, the 60 screened fingerprints have the potential to precisely describe the chemical and structural information about the atomic environments of molecules.

METHODOLOGY

Real-Space Fingerprint Using a Generic Local-Order Parameter. Here we introduce the real-space fingerprint of the atomic environment considered in this work. First, the chemical information to be embedded is defined as the implant function $F(a_i, r_{ij}, r_c)$, where a_i is the atomic number of atom i, r_{ij} is the interatomic distance between atoms i and j, and r_c is the cutoff radius for judging whether particle j is a neighbor of particle i. Here we use four cutoff radii of 1.50, 1.75, 2.00, and 2.50 Å to accommodate various molecular geometries. We emphasize that $F(a_i, r_{ij}, r_c)$ is basically a simple combination of atomic numbers and distances among neighboring atoms. In this work, a total of 18 species of $F(a_i, r_{ij}, r_c)$ were calculated for each r_c, as shown in Table 1, where $\mathbf{N}_b(r_c, i)$ is an array that involves identifiers of neighboring atoms of atom i and N is the number of neighboring atoms contained in $\mathbf{N}_b(r_c, i)$. F was designed based on the following rules: (i) it should be a function of atomic number, (ii) the interaction distance to be emphasized should change depending on the weighting, and (iii) differences in the weighting with and without normalization should be considered. Rule (i) is essential as a fingerprint for the constituent elements of a molecule, but 0, 1, and 1.5 powers were considered in order to have variations in the effect of the atomic number. Rule (ii) is a concept similar to that of the radial basis function G^2 in Parrinello−Behler type descriptors,[1,49] where the weights are maximized for interaction distances of 0.0, 1.5, 2.0, and 2.5 Å. Rule (iii) was considered because the form of the function is clearly different depending on the presence or absence of normalization. From the above rules (i−iii), the number of implant functions became 18. Note that other expressions of F are possible for the chemical and physical properties as long as these are defined as a single scalar function. Second, the implant function was embedded in the bond-orientational order parameter developed by Steinhardt and co-workers[16] as

$$Q_l(r_c, F, i) = \sqrt{\frac{4\pi}{2l+1}\sum_{m=-l}^{l} |q_{lm}|^2} \quad (1)$$

Table 1. Description of the Implant Function $F(a_j, r_{ij}, r_c)$

name of function	formula
F_1	1
F_2	$1/N$
F_3	r_{ij}
F_4	$r_{ij}^{-1} / \sum_{k\in\mathbf{N}_b(r_c,i)} r_{ik}^{-1}$
F_5	a_j
F_6	$a_j / \sum_{k\in\mathbf{N}_b(r_c,i)} a_k$
F_7	$a_j r_{ij}^{-1}$
F_8	$a_j r_{ij}^{-1} / \sum_{k\in\mathbf{N}_b(r_c,i)} a_k r_{ik}^{-1}$
F_9	$a_j^{1.5} r_{ij}^{-1}$
F_{10}	$a_j^{1.5} r_{ij}^{-1} / \sum_{k\in\mathbf{N}_b(r_c,i)} a_j^{1.5} r_{ik}^{-1}$
F_{11}	$a_j^{1.5} r_{ij}$
F_{12}	$a_j^{1.5} r_{ij} / \sum_{k\in\mathbf{N}_b(r_c,i)} a_j^{1.5} r_{ik}$
F_{13}	$a_j \exp[-4(r_{ij} - 1.5)^2]$
F_{14}	$a_j \exp[-4(r_{ij} - 1.5)^2] / \sum_{k\in\mathbf{N}_b(r_c,i)} a_k \exp[-4(r_{ik} - 1.5)^2]$
F_{15}	$a_j \exp[-4(r_{ij} - 2.0)^2]$
F_{16}	$a_j \exp[-4(r_{ij} - 2.0)^2] / \sum_{k\in\mathbf{N}_b(r_c,i)} a_k \exp[-4(r_{ik} - 2.0)^2]$
F_{17}	$a_j \exp[-4(r_{ij} - 2.5)^2]$
F_{18}	$a_j \exp[-4(r_{ij} - 2.5)^2] / \sum_{k\in\mathbf{N}_b(r_c,i)} a_k \exp[-4(r_{ik} - 2.5)^2]$

$$q_{lm}(r_c, F, i) = \sum_{j\in\mathbf{N}_b(r_c,i)} F(a_j, r_{ij}, r_c) Y_{lm}(\mathbf{r}_{ij}) \quad (2)$$

where l is an arbitrary positive integer denoting the degree of the harmonic function, m is an integer that runs from $-l$ to $+l$, Y_{lm} is a spherical harmonics function, and $\mathbf{r}_{ij}$ is a vector from particle j to particle i. In this work, we set l values as 4, 6, 8, 12, 15, 18, and 20. Therefore, the total number of representations for the atomic environment for atom i becomes 504 (equal to the number of l × the number of r_c × the number of F). The atomic number and all the representations for atom i are stored in the atomic environment vector for atom i, $\boldsymbol{d}_i$, which has 505 (504 + 1) elements. Finally, the atomic environment vectors for all atoms are merged to the descriptor array D. In this work, the number of elements of D was 505 × 133 885. This was almost the same scale of data as that for the data set of a conventional representation for the atomic environment.[13] However, in order to attain an effective atomic environment it is desirable to have fewer elements. Therefore, we attempted to screen the fingerprints from 504 to a smaller number by means of supervised ML for a regression analysis of the atomic charges of molecules.

Machine Learning. To examine the capability of each fingerprint and to screen the fingerprints, a regression analysis of the atomic charges was performed in supervised ML. The molecular geometry and atomic charge for each molecule stored in QM9 were preliminarily optimized to the generalized amber force field (GAFF)[50] by Open Babel.[51] Figure 1 shows the actual ML flow used in this work. First, $\mathbf{N}_b(r_c, i)$ was

https://doi.org/10.1021/acsomega.1c06587
ACS Omega 2022, 7, 4606−4613

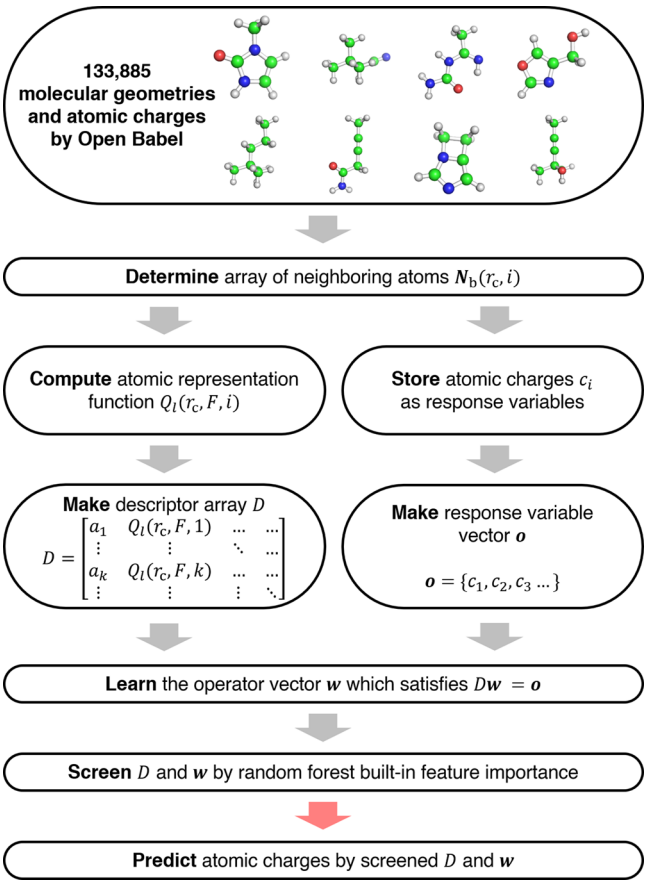


Figure 1. Machine learning flow for the regression analysis that predicts the atomic charges of molecules stored in QM9.

determined for each atom. Second, $Q_l(r_c, F, i)$ values were computed for each atom and were stored in $\boldsymbol{d}_i$. The atomic number a_i was also stored in $\boldsymbol{d}_i$. The $\boldsymbol{d}_i$ values for all atoms were merged with D. Third, the atomic charges for all atoms were merged with the response variable vector $\boldsymbol{o}$. Note that the response variable became the supervisor of the ML. Then, the operator vector $\boldsymbol{w}$, which satisfies the relation $D\boldsymbol{w} = \boldsymbol{o}$, was estimated through ML. The term $\boldsymbol{w}$ was estimated using the random forest method[52] implemented on Scikit-learn.[53] The random forest method has some useful characteristics as a learning algorithm of ML: (i) the learning routine is simple and thus has high performance of computing, (ii) the method prevents an overlearning, (iii) little or no data cleansing is needed, and (iv) the significance of data descriptors can be easily quantified. Note that overlearning is the situation where the learning results only fit to data used in the learning and do not fit to any new data. $\boldsymbol{w}$ was checked through a k-fold cross-validation implemented on Scikit-learn, where k denotes the number of cross-validations. Note that k-fold cross-validation is the method for checking overlearning. The fivefold cross-validation was done considering the quality and quantity of our data in this work. Namely, 1/5 of 2 407 756 local coordinates (~481 551) were used for each of a total of five cross-validations. Finally, the fingerprints were screened based on the function of the random forest method, which quantified the importance of the data descriptors.

RESULTS AND DISCUSSION

The time consumption to compute our fingerprint for all the molecules included in QM9 was approximately 72 h using an AMD Ryzen Threadripper 1950X CPU. Note that the fingerprints were eternally usable if these were prepared for ML once. Table 2 shows fingerprints in order of importance based on the random forest method. There is a bias in the importance of the fingerprints, which means that the top ranking fingerprints represent the typical chemical environments. Importantly, the development of highly sensitive fingerprints that contain such biases can be left to machine learning. Our proposed screening makes this possible, allowing us to select a relatively small set of candidates for the best fingerprint at any given time depending on the constantly changing (and typically increasing) amount of data in the database. The fingerprints ranked from first to 60th account for 97.3% of the importance. Figure 2 shows the importance of fingerprints from the first rank to the 60th rank. After the 55th rank, the importance level is generally saturated at about 0.0006. Therefore, in the following sections, only fingerprints from the first rank to the 60th rank were used.

For the most basic assessment of the capability of our fingerprint, the ML flow was performed using the GAFF molecular geometries and atomic charges derived from OpenBabel. Figure 3 shows the resulting regression curves for (a) hydrogen, (b) carbon, (c) nitrogen, (d) oxygen, (e) fluorine, and (f) total charges. The results for every atomic species demonstrated the high coefficient of determination (R^2) value, which indicated that the fingerprint captured well the charge variations derived from the difference of the atomic environment. The statistic scores for the regression results are shown in Table 3. The worst R^2 value was for the nitrogen charges. The mean squared error (MSE) was also the worst, and the mean absolute error (MAE) was the second worst. This implies that the variations in the charges of nitrogen atoms assigned by GAFF include systematic outliers. In contrast, the R^2 value and MSE for hydrogen charges were the best, and the MAE for hydrogen charges was second best. The regression results were also examined from the values of the molecular dipole moment. Figure 4 shows the resulting regression curve for the molecular dipole moments. The results demonstrated a high R^2 value of 0.990, which indicated that the fingerprint captured well the variations of dipole moment values derived from the difference of the atomic environment.

To evaluate the robustness of the fingerprints to variations in molecular geometries, the ML flow was performed using the DFT molecular geometries stored in QM9 before they were optimized in OpenBabel. The atomic charges, as the response variables, were the OpenBabel charges. We emphasize that response variables can be used for more than just OpenBabel charges. For example, not only the restrained electrostatic potential (RESP) charges[54] given by precise DFT calculations but also the charge and Lennard−Jones potential parameters given by CHARMM,[55] COMPASS,[56] and other force fields are possible candidates. However, as a simple and easy way to evaluate an example, we used the OpenBabel charge here. Figure 5 shows the resulting regression curves for (a) hydrogen, (b) carbon, (c) nitrogen, (d) oxygen, (e) fluorine, and (f) total charges. The results for every atomic species demonstrated high R^2 values, which indicated that the selected 60 fingerprints demonstrated a high robustness against the

https://doi.org/10.1021/acsomega.1c06587
ACS Omega 2022, 7, 4606−4613

Table 2. Fingerprints in Order of Importance Based on the Random Forest Method

rank	l	r_c (Å)	F	importance	rank	l	r_c (Å)	F	importance
1	12	1.75	F_{13}	0.688213177	31	12	2.50	F_{14}	0.001531455
2	6	1.50	F_8	0.092358318	32	8	2.50	F_9	0.001471098
3	12	2.50	F_{11}	0.028534423	33	6	2.50	F_{11}	0.001456022
4	6	1.75	F_{15}	0.023667266	34	6	2.50	F_{10}	0.001254301
5	12	1.50	F_{13}	0.016632251	35	8	2.00	F_8	0.001153517
6	12	2.50	F_9	0.014181471	36	4	1.50	F_8	0.001143127
7	12	1.50	F_8	0.012526058	37	8	2.50	F_{16}	0.001135169
8	6	1.75	F_8	0.011079143	38	4	2.50	F_{11}	0.001103377
9	4	2.50	F_8	0.006942812	39	12	1.75	F_{10}	0.001053001
10	6	1.50	F_{15}	0.004692287	40	4	2.50	F_9	0.000982597
11	12	2.50	F_{17}	0.004414363	41	8	2.50	F_{11}	0.000970519
12	12	1.75	F_{12}	0.003995392	42	4	1.75	F_{17}	0.000890206
13	12	1.75	F_8	0.003751858	43	8	2.50	F_{17}	0.000877437
14	4	1.75	F_{16}	0.003325598	44	12	2.00	F_8	0.000846427
15	12	2.50	F_{10}	0.002796547	45	4	2.50	F_{15}	0.000787806
16	6	2.50	F_{17}	0.002704569	46	4	2.50	F_{17}	0.000771686
17	4	2.50	F_{13}	0.002336904	47	6	2.00	F_{15}	0.000742759
18	4	2.50	F_{14}	0.002325804	48	4	2.50	F_{16}	0.000736547
19	12	2.50	F_{15}	0.002142383	49	6	2.50	F_{14}	0.000732413
20	8	2.50	F_{10}	0.002134398	50	4	1.75	F_{14}	0.000720596
21	6	2.50	F_{13}	0.002012958	51	12	2.50	F_{16}	0.000718018
22	12	2.50	F_8	0.001898805	52	12	2.00	F_{10}	0.000707675
23	4	2.50	F_{10}	0.001880449	53	6	2.00	F_8	0.000698370
24	6	2.50	F_8	0.001702691	54	8	2.50	F_{15}	0.000695468
25	6	2.00	F_{14}	0.001672374	55	4	2.00	F_8	0.000657664
26	8	2.50	F_8	0.001621629	56	8	1.50	F_{17}	0.000655981
27	6	2.50	F_9	0.001594413	57	4	2.00	F_{14}	0.000639511
28	4	1.75	F_8	0.001559824	58	6	2.50	F_{15}	0.000630099
29	12	1.75	F_{16}	0.001545018	59	12	1.75	F_{11}	0.000611518
30	8	1.75	F_8	0.001538792	60	4	1.50	F_{16}	0.000594030

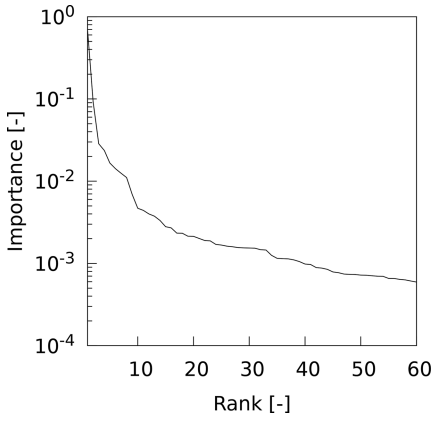


Figure 2. A logarithmic plot of the importance of the fingerprints from the first to the 60th rank.

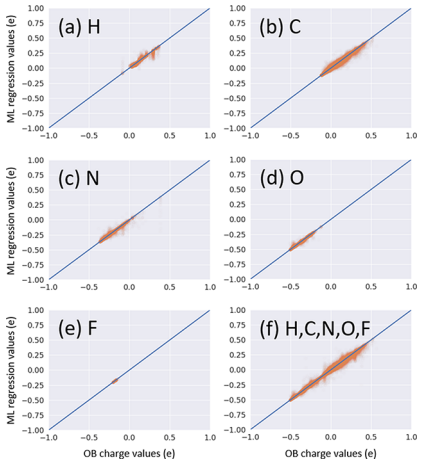


Figure 3. Regression curves for (a) hydrogen, (b) carbon, (c) nitrogen, (d) oxygen, (e) fluorine, and (f) total charges by the fingerprint of the GAFF structures.

deviation of the molecular geometries. The statistic scores for the regression results are shown in Table 4. The worst R^2 value was that for the fluorine charges. However, MAE and MSE values for fluorine charges were the smallest. This was because the charge variations of fluorine atoms were basically small. Thus, the regression accuracy for the fluorine charges was never low. The trends for the statistic scores for nitrogen charges were almost the same as those the regression results from the GAFF structures and OpenBabel charges. The regression results were also examined for the molecular dipole moment. Figure 6 shows the regression curve for molecular dipole moments. The results demonstrated an R^2 value of 0.990, which is comparable to that for the regression from GAFF structures and OpenBabel charges. Again, the accurate prediction of OpenBabel charges and dipole moments for DFT structures shows the high robustness of the 60 fingerprints for the molecular geometries.

Finally, to examine the prediction capability of the trained model from GAFF structures and OpenBabel charges, the charges were predicted from the trained operator vector $\boldsymbol{w}_{\mathrm{GAFF}}$,

https://doi.org/10.1021/acsomega.1c06587
ACS Omega 2022, 7, 4606−4613

Table 3. Statistic Scores for the Regression Results of the Fingerprint of the GAFF Structures

atom	R^2	MAE	MSE
H	0.984	0.002	0.00004
C	0.974	0.007	0.00020
N	0.951	0.006	0.00030
O	0.983	0.003	0.00006
F	0.967	0.001	0.00008
total	0.993	0.004	0.00012

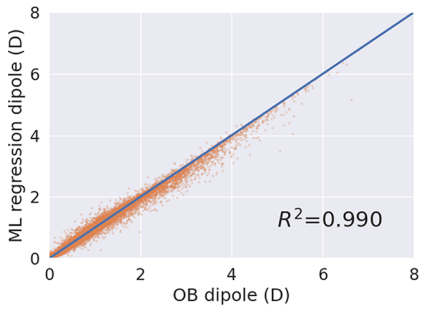


Figure 4. Regression curve for molecular dipole moments from the fingerprint of the GAFF structures.

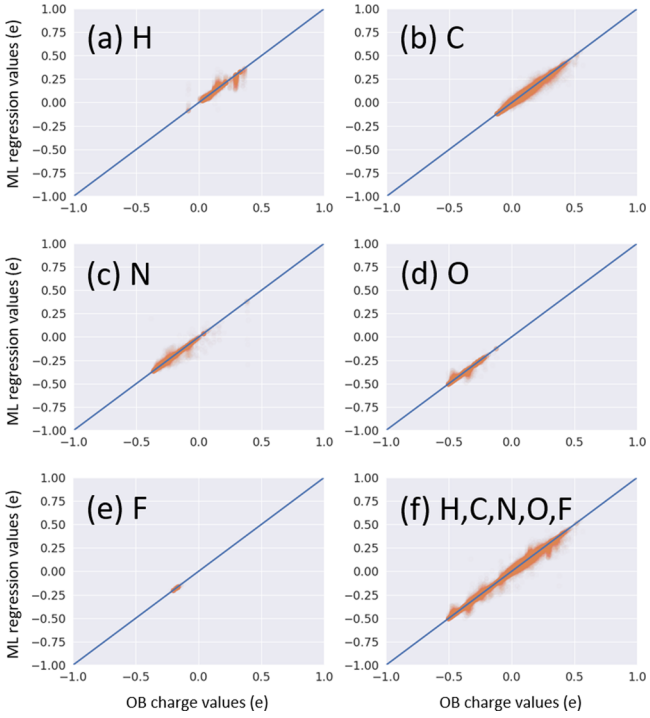


Figure 5. Regression curves for (a) hydrogen, (b) carbon, (c) nitrogen, (d) oxygen, (e) fluorine, and (f) total charges from the fingerprint of the DFT structures.

and the descriptor array was predicted from DFT structures D_{DFT}. Figure 7 shows a comparison of the predicted charges with OpenBabel charges for (a) hydrogen, (b) carbon, (c) nitrogen, (d) oxygen, (e) fluorine, and (f) total charges. The results for carbon and hydrogen had high R^2 values, and the other atomic species exhibited reasonable R^2 values. The statistical scores for the comparison results are shown in Table 5. The worst R^2, MAE, and MSE values were found for nitrogen charges, a feature consistent with the other results.

Table 4. Statistic Scores for the Regression Results from the Fingerprint of the DFT Structures

atom	R^2	MAE	MSE
H	0.995	0.003	0.00009
C	0.983	0.005	0.00010
N	0.947	0.006	0.00040
O	0.971	0.003	0.00010
F	0.944	0.001	0.00001
total	0.995	0.003	0.00009

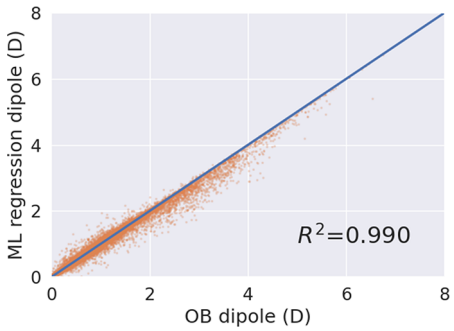


Figure 6. Regression curve for molecular dipole moments from the fingerprint of the DFT structures.

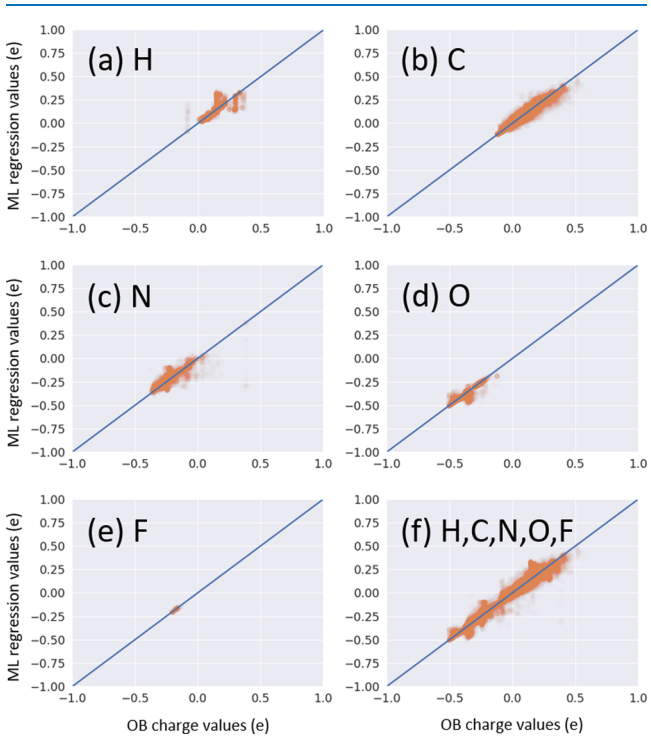


Figure 7. Comparison of predicted charges with OpenBabel charges for (a) hydrogen, (b) carbon, (c) nitrogen, (d) oxygen, (e) fluorine, and (f) total charges.

Overall, the model trained from GAFF structures and OpenBabel charges can adequately predict the OpenBabel charges for DFT molecular structures. However, the molecular dipole moments were not determined accurately when the predicted charges were used (data not shown). Despite the success of machine learning in predicting charges at a certain level, the low prediction accuracy of the dipole moment implies that the gap in molecular geometries between GAFF and DFT is not small.

https://doi.org/10.1021/acsomega.1c06587
ACS Omega 2022, 7, 4606−4613

Table 5. Statistical Scores for the Comparison of Predicted Charges with OpenBabel Charges

atom	R^2	MAE	MSE
H	0.913	0.004	0.00020
C	0.936	0.014	0.00050
N	0.764	0.022	0.00100
O	0.871	0.009	0.00050
F	0.860	0.004	0.00003
total	0.977	0.009	0.00040

CONCLUSIONS

We developed a candidate for simple representation that combines atomic species and the bond-orientational order parameter. The 504 candidates were applied as the fingerprint of molecules stored in QM9. To screen the fingerprints, a regression analysis of the atomic charges given by OpenBabel was performed by a supervised ML flow. The 60 fingerprints were selected based on importance value of the random forest method. The selected fingerprints was applied for the regression analysis of OpenBabel charges, which used GAFF molecular structures. The results exhibited high statistic scores for both the atomic charge and the dipole moment, which indicated that the fingerprint captured well the variations of charge derived from the difference of the atomic environment. Then, the regression analysis was performed using the DFT structures to assess the robustness of 60 fingerprints against the deviation of the molecular structures. The results exhibited high statistical scores, which indicated that the fingerprints demonstrated a high robustness against the deviation of the molecular structures. Finally, the atomic charges were predicted from $\boldsymbol{w}_{\mathrm{DFT}}$ and D_{GAFF} to examine the prediction capability of the trained model from GAFF structures and OpenBabel charges. The results showed that the trained model can adequately predict the OpenBabel charges for DFT molecular structures. Therefore, the screened 60 fingerprints have the potential to precisely describe the chemical and structural information on the atomic environment of molecules. Importantly, only 60 chemical representations give the exact atomic charges for 133 885 different molecular geometries. This fact means that our chemical representation successfully compresses the information content of fingerprints based on 3D molecular geometries. The robustness of the fingerprints to molecular geometries and the robustness of the learning model to charge prediction also indicate that our chemical representation can be used for iterations that simultaneously determine molecular geometries and charges. Furthermore, our representation can be used in combination with many other conventional atomic representations.[2−5,7,13] Thus, the prediction capability of the above-mentioned trained model may be improved for the deviation of molecular structures and molecules never included in training data. In addition to constructing precise big data for molecular structures, the development of atomic representations might assist in the high-throughput design of molecular models that include intra- and intermolecular interaction parameters without computationally expensive first-principles simulation techniques.

AUTHOR INFORMATION

Corresponding Author

Kazuaki Z. Takahashi − *Research Center for Computational Design of Advanced Functional Materials, National Institute of Advanced Industrial Science and Technology Tsukuba Central 2, Tsukuba, Ibaraki 305-8568, Japan*; orcid.org/0000-0001-6603-9862; Phone: +81-29-861-2972; Email: kazu.takahashi@aist.go.jp; Fax: +81-29-861-5375

Authors

Hideo Doi − *Research Center for Computational Design of Advanced Functional Materials, National Institute of Advanced Industrial Science and Technology Tsukuba Central 2, Tsukuba, Ibaraki 305-8568, Japan*

Takeshi Aoyagi − *Research Center for Computational Design of Advanced Functional Materials, National Institute of Advanced Industrial Science and Technology Tsukuba Central 2, Tsukuba, Ibaraki 305-8568, Japan*; orcid.org/0000-0001-9229-4226

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.1c06587

Author Contributions

‡H.D. and K.Z.T. contributed equally to this study. K.Z.T. and T.A. designed the study. K.Z.T. directed the study. H.D. performed the supervised machine learning and regression analyses. K.Z.T. wrote the first version of the paper. H.D., K.Z.T., and T.A. contributed to the final version of the paper.

Notes

The authors declare no competing financial interest.

The data that support the findings of this study are available from the corresponding author upon reasonable request based on the publication protocol of the research data as permitted by JPNP16010 commissioned by NEDO.

ACKNOWLEDGMENTS

This paper is based on results obtained from a project (JPNP16010) commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

REFERENCES

(1) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.

(2) Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K.-R.; Gross, E. K. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B* **2014**, *89*, 205118.

(3) Behler, J. Representing potential energy surfaces by high-dimensional neural network potentials. *J. Phys.: Condens. Matter* **2014**, *26*, 183001.

(4) 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. *Int. J. Quantum Chem.* **2015**, *115*, 1084−1093.

(5) 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.

(6) Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **2016**, *145*, 170901.

(7) Huang, B.; Von Lilienfeld, O. A. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *J. Chem. Phys.* **2016**, *145* (16), No. 161102.

(8) Landrum, G. *RDKit: Open-Source Cheminformatics Software*, 2016. https://www.rdkit.org/.

(9) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine learning in materials informatics: Recent applications and prospects. *npj Comput. Mater.* **2017**, *3*, 54.

(10) Botu, V.; Batra, R.; Chapman, J.; Ramprasad, R. Machine learning force fields: Construction, validation, and outlook. *J. Phys. Chem. C* **2017**, *121*, 511−522.
(11) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360−365.
(12) Carleo, G.; Cirac, I.; Cranmer, K.; Daudet, L.; Schuld, M.; Tishby, N.; Vogt-Maranto, L.; Zdeborová, L. Machine learning and the physical sciences. *Rev. Mod. Phys.* **2019**, *91*, 045002.
(13) Himanen, L.; Jäger, M. O.; 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. *Comput. Phys. Commun.* **2020**, *247*, 106949.
(14) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988**, *28*, 31−36.
(15) Homer, R. W.; Swanson, J.; Jilek, R. J.; Hurst, T.; Clark, R. D. SYBYL line notation (SLN): a single notation to represent chemical structures, queries, reactions, and virtual libraries. *J. Chem. Inf. Model.* **2008**, *48*, 2294−2307.
(16) Steinhardt, P. J.; Nelson, D. R.; Ronchetti, M. Bond-orientational order in liquids and glasses. *Phys. Rev. B* **1983**, *28*, 784.
(17) Hohenberg, P.; Kohn, W. Inhomogeneous electron gas. *Physical review* **1964**, *136*, B864.
(18) Levy, M. Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the v-representability problem. *Proc. Natl. Acad. Sci. U. S. A.* **1979**, *76*, 6062−6065.
(19) Vignale, G.; Rasolt, M. Density-functional theory in strong magnetic fields. *Physical review letters* **1987**, *59*, 2360.
(20) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters* **2012**, *108*, 058301.
(21) 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.
(22) 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. *journal of physical chemistry letters* **2015**, *6*, 2326−2331.
(23) Bartok, A. P.; Kondor, R.; Csanyi, G. Publisher's note: On representing chemical environments [Phys. Rev. B **87**, 184115 (2013)]. *Phys. Rev. B* **2013**, *87*, 219902.
(24) Lechner, W.; Dellago, C. Accurate determination of crystal structures based on averaged local bond order parameters. *J. Chem. Phys.* **2008**, *129*, 114707.
(25) Lü, Y.; Entel, P. Impact of medium-range order on the glass transition in liquid Ni-Si alloys. *Phys. Rev. B* **2011**, *84*, 104203.
(26) Bialké, J.; Speck, T.; Löwen, H. Crystallization in a dense suspension of self-propelled particles. *Physical review letters* **2012**, *108*, 168301.
(27) Kalikka, J.; Akola, J.; Larrucea, J.; Jones, R. O. Nucleus-driven crystallization of amorphous Ge 2 Sb 2 Te 5: A density functional study. *Phys. Rev. B* **2012**, *86*, 144113.
(28) Leocmach, M.; Tanaka, H. Roles of icosahedral and crystal-like order in the hard spheres glass transition. *Nat. Commun.* **2012**, *3*, 974.
(29) Russo, J.; Tanaka, H. The microscopic pathway to crystallization in supercooled liquids. *Sci. Rep.* **2012**, *2*, 505.
(30) Gasser, U.; Ziese, F.; Maret, G. Characterization of local structures with bond-order parameters and graphs of the nearest neighbors, a comparison. *European Physical Journal Special Topics* **2014**, *223*, 455−467.
(31) Yu, T.-Q.; Chen, P.-Y.; Chen, M.; Samanta, A.; Vanden-Eijnden, E.; Tuckerman, M. Order-parameter-aided temperature-accelerated sampling for the exploration of crystal polymorphism and solid-liquid phase transitions. *J. Chem. Phys.* **2014**, *140*, 214109.
(32) Eslami, H.; Khanjari, N.; Müller-Plathe, F. A local order parameter-based method for simulation of free energy barriers in crystal nucleation. *J. Chem. Theory Comput.* **2017**, *13*, 1307−1316.
(33) Doi, H.; Takahashi, K. Z.; Tagashira, K.; Fukuda, J.-i.; Aoyagi, T. Machine learning-aided analysis for complex local structure of liquid crystal polymers. *Sci. Rep.* **2019**, *9*, 16370.
(34) Takahashi, K. Z.; Aoyagi, T.; Fukuda, J.-i. Multistep nucleation of anisotropic molecules. *Nat. Commun.* **2021**, *12*, 5278.
(35) Kawasaki, T.; Tanaka, H. Structural evolution in the aging process of supercooled colloidal liquids. *Phys. Rev. E* **2014**, *89*, 062315.
(36) Coslovich, D.; Jack, R. L. Structure of inactive states of a binary Lennard-Jones mixture. *Journal of Statistical Mechanics: Theory and Experiment* **2016**, *2016*, 074012.
(37) Doi, H.; Takahashi, K. Z.; Aoyagi, T. Mining of effective local order parameters for classifying crystal structures: A machine learning study. *J. Chem. Phys.* **2020**, *152*, 214501.
(38) Quigley, D.; Rodger, P. Metadynamics simulations of ice nucleation and growth. *J. Chem. Phys.* **2008**, *128*, 154518.
(39) Reinhardt, A.; Doye, J. P.; Noya, E. G.; Vega, C. Local order parameters for use in driving homogeneous ice nucleation with all-atom models of water. *J. Chem. Phys.* **2012**, *137*, 194504.
(40) Sanz, E.; Vega, C.; Espinosa, J.; Caballero-Bernal, R.; Abascal, J.; Valeriani, C. Homogeneous ice nucleation at moderate supercooling from molecular simulation. *J. Am. Chem. Soc.* **2013**, *135*, 15008−15017.
(41) Hudait, A.; Moberg, D. R.; Qiu, Y.; Odendahl, N.; Paesani, F.; Molinero, V. Preordering of water is not needed for ice recognition by hyperactive antifreeze proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, 8266−8271.
(42) Montero de Hijes, P.; Espinosa, J.; Vega, C.; Sanz, E. Ice growth rate: Temperature dependence and effect of heat dissipation. *J. Chem. Phys.* **2019**, *151*, 044509.
(43) Doi, H.; Takahashi, K. Z.; Aoyagi, T. Searching local order parameters to classify water structures of ice Ih, Ic, and liquid. *J. Chem. Phys.* **2021**, *154*, 164505.
(44) Doi, H.; Takahashi, K. Z.; Aoyagi, T. Searching for local order parameters to classify water structures at triple points. *J. Comput. Chem.* **2021**, *42*, 1720−1727.
(45) Doi, H.; Takahashi, K. Z.; Aoyagi, T. Mining of Effective Local Order Parameters to Classify Ice Polymorphs. *J. Phys. Chem. A* **2021**, *125*, 9518−9526.
(46) Tang, X.; Yang, J.; Xu, T.; Tian, F.; Xie, C.; Li, L. Local structure order assisted two-step crystal nucleation in polyethylene. *Physical Review Materials* **2017**, *1*, 073401.
(47) 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.
(48) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data* **2014**, *1*, 140022.
(49) Artrith, N.; Behler, J. High-dimensional neural network potentials for metal surfaces: A prototype study for copper. *Phys. Rev. B* **2012**, *85*, 045439.
(50) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *Journal of computational chemistry* **2004**, *25*, 1157−1174.
(51) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminformatics* **2011**, *3*, 33.
(52) Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5−32.
(53) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(54) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **1993**, *97*, 10269−10280.

https://doi.org/10.1021/acsomega.1c06587
ACS Omega 2022, 7, 4606−4613

(55) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S. a.; Karplus, M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of computational chemistry* **1983**, *4*, 187−217.
(56) Sun, H. COMPASS: an ab initio force-field optimized for condensed-phase applications overview with details on alkane and benzene compounds. *J. Phys. Chem. B* **1998**, *102*, 7338−7364.

https://doi.org/10.1021/acsomega.1c06587
ACS Omega 2022, 7, 4606−4613