Feature selection in molecular graph neural networks based on quantum chemical approaches

Daisuke Yokogawa* and Kayo Suda

Graduate School of Arts and Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan

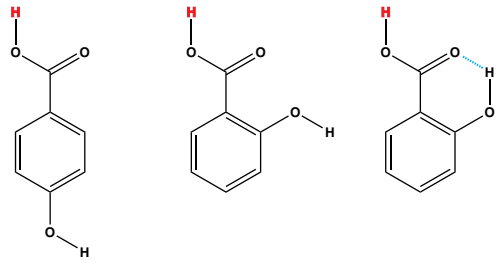
E-mail: c-d.yokogawa@g.ecc.u-tokyo.ac.jp

Abstract

Feature selection is an important topic that has been widely studied in data science. Recently, graph neural networks (GNNs) and graph convolutional networks (GCNs) have also been employed in chemistry. To enhance the performance characteristics of the GNN and GCN in the field of chemistry, the feature selection should also be discussed in detail from the chemistry viewpoint. Thus, this study proposes a new feature in molecular GNNs based on the quantum chemical approaches and discusses the accuracy, overcorrelation between features, and interpretability. From the overcorrelation and accuracy, the important graph convolution (IGC) with molecular-atomic properties (MAP) proposed herein showed good performance. Moreover, the integrated gradients analysis showed that the machine learning model with the IGC(MAP) explained the prediction outputs reasonably.

Introduction

What is required for good features in molecular graph neural networks (GNNs)? Several studies have been conducted concerning feature selection in data science, and it has been mentioned that good features should improve accuracy, overcorrelation, and interpretability.[1–3] Recently, GNNs and graph convolutional networks (GCNs) have been widely applied in chemistry.[4–7] The feature selection should also be discussed in detail from the chemistry viewpoint in order to enhance the performance of the GNN and GCN in the field of chemistry.



Scheme 1: Chemical structures of hydroxybenzoic acid (HB).

Accuracy is one of the most important points in feature selection. In chemistry, a small structural difference affects the molecular properties. For example, the acid dissociation constant is greatly affected by the positions of the functional groups. Scheme 1 shows three hydroxybenzoic acid (HB) structures. The difference is only the relative positions of the OH and COOH groups, and the orientation of the OH group. Despite the small difference, these conformations give different pKa values ($= -\log K_a$), where K_a is the acid dissociation constant. Good features should have the ability to distinguish the difference.

Overcorrelation is another critical point in GNN and GCN studies. The overcorrelation in features indicates that they have irrelevant or redundant information.[2,3] Jin et al. discussed the GNN performance based on the feature overcorrelation. Their model (DeCorr) reduced the feature correlation and performed better than the standard GNN approaches.[3] The feature correlation in the convolution step was also focused on the GCN. It was shown that the GCN shows the degradation of the performance when the correlation of the features

between the layers becomes large.[8,9] Thus, the overcorrelation in the features should be removed in molecular GNNs.

The accuracy and the overcorrelation are important points in the feature selection. However, in chemistry, interpretability is considered more seriously. Recently, due to the development of theoretical methods and computers, various molecular properties can be computed accurately. However, to understand the chemistry, the reasons behind such physical properties must be investigated. In the quantum chemical field, to discuss such reasons, population analysis is applied. If the atomic charges are assigned on each atomic site, the chemists can image the charge flow in a molecule, which leads to the design of new molecules. Therefore, when the molecular GNNs and GCNs are applied to chemistry, the obtained results should be explained with the employed features.

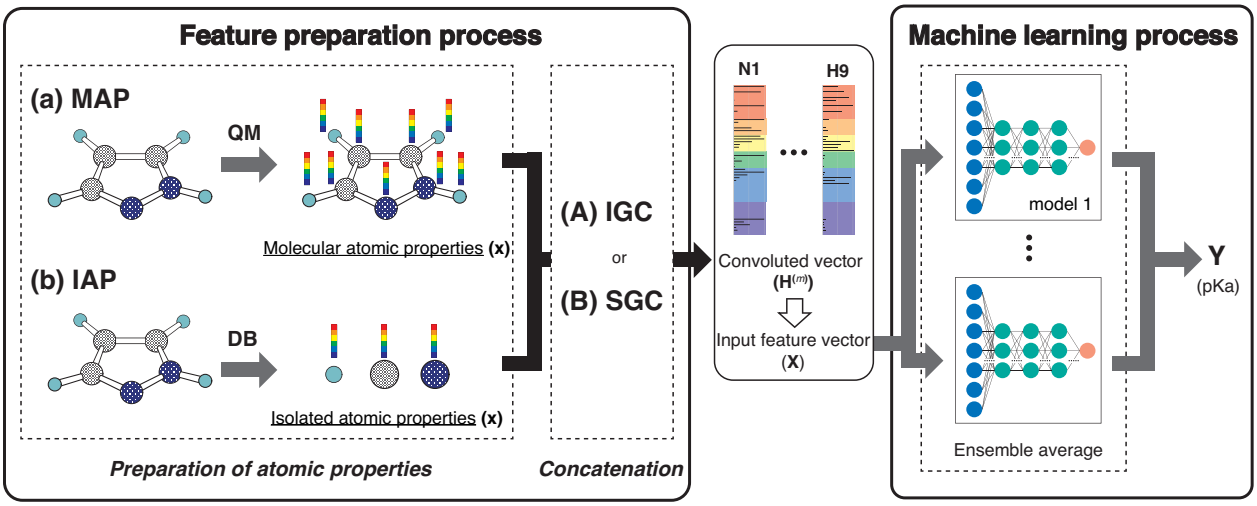
Feature selection has been widely studied in data science. However, to the best of our knowledge, detailed analysis in chemistry is limited. Hence, this study proposes a new feature in molecular GNNs based on quantum chemical approaches and discusses the accuracy, overcorrelation between features, and interpretability in detail.

Method

In this study, we proposed a new feature preparation process and constructed a machine learning (ML) process using the prepared features. Scheme 2 summarizes the flowchart of the present model. The feature preparation process comprises the preparation of atomic properties and concatenation. This section explains each step in detail.

Preparation of atom properties

Various atom features have been applied in GCN studies.[10,11] For example, Choudhary and DeCost employed the following nine atomic features in their GNN study: electronegativity, group number, covalent radius, valence electrons, first ionization energy, electron affinity,



Scheme 2: Schematic of the workflow in this study.

block, and atomic volume.[11] They are isolated atomic properties (IAP) and can be prepared without molecular information. In the quantum mechanical (QM) field, molecular-atomic properties (MAPs) are also employed for the analysis. After the QM calculations, various atomic properties are assigned on each atomic site using decomposition approaches.[12–15]

In this study, we used the IAPs and the MAPs in the feature preparation. In IAP, the following six atomic properties were applied: effective nuclear charge, atomic polarizability, atomic radius, ionization energy, electron affinity, and atomic mass. Concerning the MAPs, the following nine properties were used: the positive and negative values of the constrained spatial electron density distribution (cSED) charge ($Q+$ and $Q-$),[15] the positive and negative values of the isotropic magnetic shielding constant ($\sigma+$ and $\sigma-$), the positive and negative values of the molecular electrostatic potential (MEP)[16–18] change at the nucleus ($M+$ and $M-$), the positive value of the partial Fukui function ($F+$), volume (V), and atomic dispersion coefficient (C_6).[14] Although the partial Fukui function also takes positive and negative values, only the positive value is important. In MEP, the potential negatively increases as the atomic number increases. In order to remove the atomic number dependency in MEP, the $M+$ and $M-$ were computed by subtracting the MEP value computed in isolated atom from the MEP value computed in molecule.

Concatenation of atomic properties

To construct the ML features from the atom properties, the GNN was considered. The hidden feature of node v in the l-th layer is denoted by $h_v^{(l)}$ and $h_v^{(0)} = x_v$, where $\mathbf{x}$ are the node features (MAPs or IAPs). Moreover, $\mathbf{h}^{(l)}$ is formally given along the update step, as follows:

$$\mathbf{h}^{(l)} = U^{(l)}(\mathbf{h}^{(l-1)}), \tag{1}$$

where $U^{(l)}$ is the update function at the l-th layer.[19] By concatenating the obtained $\mathbf{h}^{(l)}$ $(l = 0, 1, \cdots, L)$, we prepared the following vector:

$$\mathbf{H}^{(L)} \equiv \mathbf{h}^{(0)} \oplus \mathbf{h}^{(1)} \oplus \cdots \oplus \mathbf{h}^{(L)} \tag{2}$$

where $\oplus$ is the concatenation of two vectors. This step is a simple version of Jumping Knowledge Networks.[20] The concatenated vector $\mathbf{H}^{(L)}$ is the feature vector for the ML process.

Many processes in the update step are given in eq 1. Concerning the simple graph convolution (SGC),[21] the update step is given as follows:

$$\mathbf{h}^{(l)} = \mathbf{S}\mathbf{h}^{(l-1)} = \mathbf{S}^l\mathbf{x} \tag{3}$$

where $\mathbf{S} = \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}$, $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, $\mathbf{A}$ is the adjacency matrix, and $\tilde{\mathbf{D}}$ is the degree matrix of $\tilde{\mathbf{A}}$. Although the update step with $\mathbf{S}$ is well employed, it is known that the elements of $\mathbf{S}^l$ converge to a fixed value when l is large.[22] To overcome the problem, the following update step was proposed as follows:

$$\mathbf{h}^{(l)} = \mathbf{a}^{(l)}\mathbf{x} \tag{4}$$

where $\mathbf{a}^{(l)}$ is the Hollow matrix, and the off-diagonal elements are defined as follows:

$$a_{ij}^{(l)} = \begin{cases} \bar{b}_{ij} & (l = 1) \\ \sqrt{\bar{a}_{ij}^{(l)} \bar{a}_{ji}^{(l)}} & (l > 1) \end{cases} \tag{5}$$

where

$$\bar{a}_{ij}^{(l)} = \delta^{\mathrm{H}} \left(\sum_k \bar{b}_{ik} a_{kj}^{(l-1)}; \max_{m=1,\cdots l-1} a_{ij}^{(m)} \right) \tag{6}$$

and $\delta^{\mathrm{H}}(x;\lambda)$ is the hard shrinkage function; $\bar{\mathbf{b}}$ is defined as follows:

$$\bar{\mathbf{b}} = \mathbf{D}^{-1/2} \mathbf{b} \mathbf{D}^{-1/2} \tag{7}$$

where $\{b_{ij}\}$ is the Wiberg bond index.[23] When $a_{ij}^{(l)}$ has a nonzero value, the path $i \leftrightarrow j$ at the l-th step is either the shortest, or not, but comprises strong bonds, such as double and triple bonds. Because the shortest path and the path through strong chemical bonds are important to transfer the information to each node, the new convolution process is termed an important graph convolution (IGC).

The employed atomic properties are defined with different units, and the maximum values in the properties differ. To remove the bias, we employed min-max normalization to $\mathbf{H}^{(L)}$. The element $h_i^{(l)}$ $(0 \leq l \leq L)$ in the $\mathbf{H}^{(L)}$ was normalized with

$$\tilde{h}_i^{(l)} = \frac{h_i^{(l)}}{h_i^{\max}}, \tag{8}$$

where $h_i^{\max}$ is the maximum value of the i-th property determined from the training and validation datasets.

ML process

To discuss the performance of the prepared features, the supervised learning algorithm for pKa recognition was employed. The feature vector $\mathbf{H}^{(L)}$ of the dissociated proton is chosen as an input vector $\mathbf{X}^{(0)}$ in the multilayer perceptron (Scheme 2). The output $\mathbf{Y}$ (pKa in this study) is obtained as follows:

$$\mathbf{X}^{(m)} = \sigma\left(\mathbf{X}^{(m-1)}\mathbf{\Theta}^{(m)} + \boldsymbol{\beta}^{(m)}\right), \tag{9}$$

$$\mathbf{Y} = \mathbf{X}^{(M-1)}\mathbf{\Theta}^{(M)} + \boldsymbol{\beta}^{(M)}, \tag{10}$$

where $\mathbf{\Theta}^{(m)}$ and $\boldsymbol{\beta}^{(m)}$ are the weight matrix and the bias vector of the layer m, respectively, and σ is a nonlinear activation function, e.g., a ReLU. The number of layer M is set to 4.

$\mathbf{Y}$ depends on the weight and the hyperparameters, such as the number of nodes in the hidden layers and the dropout ratio. If the weight is optimized with different hyperparameters, the different trained networks are produced. A linear combination of the corresponding outputs was taken as follows:[24,25]

$$\bar{\mathbf{Y}} = \sum_{j}^{p} \alpha_j \mathbf{Y}_j, \tag{11}$$

where $\mathbf{Y}_j$ is the output obtained with the j-th trained network, p is the number of trained networks, and α_j is the associated combination weight; $p = 5$, and an equal combination-weight was employed. The hyperparameters were chosen from the top five best in the hyperparameter fitting process.

Computational details

Datasets

The pKa values and molecular information were obtained from the training and test sets prepared in the previous study.[26] The pKa values were carefully cleaned and curated for this study, and uncertain values were removed (e.g., the pKa values are far from those of analog or the dissociated proton site is unclear). Moreover, the calculation was restricted to molecules with no iodide atom because of the current program limitation. Finally, 811, 203, and 316 pKa values were obtained for the training, validation, and test sets, respectively.

Hyperparameters

There are three layers, and the hidden size of the layers are n_0, n_1, and n_0, respectively, which are summarized in Figure S1 (Supporting information). A hyperparameter search for the optimal hidden size (n_0 and n_1) and the dropout rate was computed using Optuna,[27] where the Bayesian hyperparameter optimization was employed. The number of trial steps and epochs were 50 and 3000 epochs, respectively. The weight was further trained to 8000 epochs to improve its final accuracy.

Calculations of molecules

The molecular geometries were computed at the CAM-B3LYP/aug-cc-pVDZ level of theory.[28,29] The cSED charge, partial Fukui function, volume, atomic C_6 dispersion coefficient, and MEP were computed using the GAMESS program package,[30] and isotropic magnetic shielding constants on atom was computed using the Gaussian program package.[31]

The MAPs were also computed at the Hartree-Fock (HF)/6-31G** level of theory. Although a large difference in computational cost exists between HF/6-31G** and CAM-B3LYP/aug-cc-pVDZ, the difference in the predicted pKa was small, as shown in Figure S2 (Supporting information).

Results and discussion

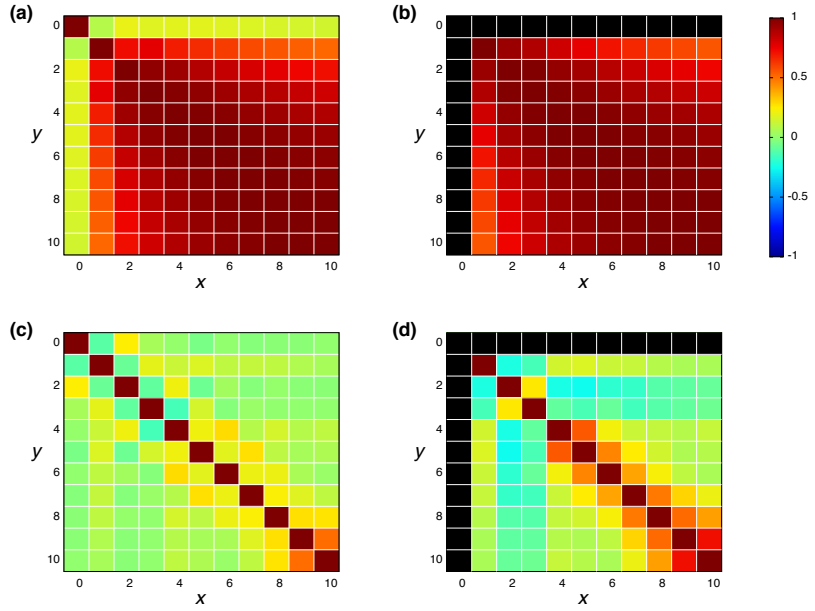


Figure 1: Heatmap of the Pearson's correlation coefficient between $\mathbf{h}(x)$ and $\mathbf{h}(y)$; $\mathbf{h}(x)$ is prepared with (a) and (b) SGC and (c) and (d) IGC. MAPs were employed for (a) and (c), and IAPs were employed for (b) and (d). The pairs with undefined correlation coefficients are shown in black.

The correlation between features was evaluated using Pearson's correlation coefficient,

$$\rho_{x,y} = \frac{\sum_i {}^t(\mathbf{h}_i^{(x)} - \bar{\mathbf{h}}^{(x)}) \cdot (\mathbf{h}_i^{(y)} - \bar{\mathbf{h}}^{(y)})}{\sqrt{\sum_i |\mathbf{h}_i^{(x)} - \bar{\mathbf{h}}^{(x)}|^2} \sqrt{\sum_i |\mathbf{h}_i^{(y)} - \bar{\mathbf{h}}^{(y)}|^2}}, \tag{12}$$

where $\mathbf{h}_i^{(x)}$ is the feature vector in the x-th layer of the molecular i, and $\bar{\mathbf{h}}^{(x)}$ is the mean value of $\mathbf{h}_i^{(x)}$. The input feature of molecule i was prepared by taking the concatenation of $\{\mathbf{h}_i^{(x)}\}$ (eq 2 and Scheme 2). Therefore, if $\rho_{x,y}$ is large, the information that $\mathbf{h}_i^{(x)}$ and $\mathbf{h}_i^{(y)}$ are similar, and the feature vector $\mathbf{X}$ contains the redundant data. In Figure 1, the heat maps of $\rho_{x,y}$ computed with IGC and SGC are shown. In the correlation calculations, IAPs and MAPs were employed as the atomic properties. Because $\mathbf{h}_i^{(0)}$ computed with IAPs have the same values among the molecules, the correlation coefficients ($\rho_{0,x}$ and $\rho_{x,0}$) cannot be defined, which was colored black in Figure 1. As shown in Figure 1(a), the correlation

between $\mathbf{h}^{(0)}$ and $\mathbf{h}^{(x)}$ ($x \geq 1$) was small in the case of SGC(MAP), whereas the correlations between $\mathbf{h}^{(x)}$ and $\mathbf{h}^{(y)}$ ($x, y \geq 1$) were large in both cases of MAP and IAP (Figures 1(a) and 1(b)). Therefore, the redundancy in the features of constructed $\mathbf{X}$ should be large when SGC(IAP) is employed. By contrast, the correlation between x-th and y-th layer vectors is small when IGC is chosen (Figures 1(c) and 1(d)). From the results, we concluded that the $\mathbf{X}$ constructed with IGC should have relevant features.

It is useful to consider the meaning of the small correlation between $\mathbf{h}^{(x)}$ and $\mathbf{h}^{(y)}$ ($x \neq y$) in IGC(MAP) based on spectral filtering. When $\mathbf{v}_i$ and λ_i are the i-th eigenvector and eigenvalue of Laplacian, respectively, the spectral filtering on graph signal $\mathbf{x}$ can be written as follows:

$$\mathbf{y} = \sum_k \mathbf{y}^{(k)}, \tag{13}$$

$$\mathbf{y}^{(k)} \equiv f(\lambda_k)\mathbf{v}_k\mathbf{v}_k^T\mathbf{x} \tag{14}$$

where $\mathbf{y}$ is the filtered signal and $f(\lambda_k)$ is the filter kernel. From the definition, $\mathbf{y}^{(k)}$ and $\mathbf{y}^{(l)}$ ($k \neq l$) are perpendicular to each other. From the similarity of eqs 4 and 14 and the no correlation between $\mathbf{y}^{(k)}$ and $\mathbf{y}^{(l)}$ ($k \neq l$), $\mathbf{h}^{(x)}$ can be considered to be a filtered vector, as in the case of graph spectral filtering.[32]

The redundancy in the feature vector $\mathbf{X}$ probably affects the accuracy of the predicted pKa values. To discuss the relationship between the accuracy and the redundancy in $\mathbf{X}$, the root mean square errors (RMSEs) of the pKa values were calculated. Because $\mathbf{X}$ is the concatenated vector $\mathbf{H}^{(L)}$ of the dissociated proton, the size of $\mathbf{X}$ is controlled by the convolution layers (L). In Figure 2, the RMSE computed with L=1, 2, 3, 5, 7, and 10 are shown. For comparison, the RMSEs were also computed with a freely available pKa prediction tool called OPERA[26] and MolGpKa.[33] In the case of MolGpKa, the model was optimized using the dataset employed in this study. When IAP was employed, there was a large difference in the accuracy between the convolution approaches, SGC and IGC. Although the error in

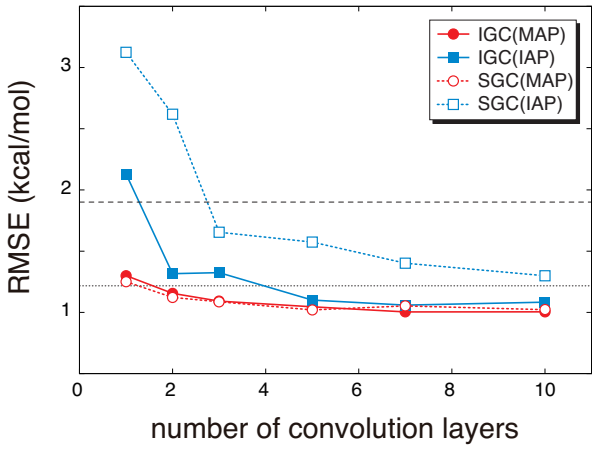


Figure 2: Root mean square errors (RMSEs) of pKa values obtained by two types of atomic features (IAP and MAP) and convolution processes (IGC and SGC) with different number of convolution layers (1, 2, 3, 5, 7, and 10). For comparison, the RMSEs computed with MolGpka and OPERA are also shown with black dashed and dotted lines, respectively.

IGC(IAP) and SGC(IAP) decreases as L increases, the error in IGC(IAP) is largely improved as L increases when compared with SGC(IAP). This is because the redundancy in $\mathbf{X}$ of IGC(IAP) is smaller than of SGC(IAP) (Figure 1). When MAP was employed, the RMSE is small in both cases of IGC and SGC because $\mathbf{X}$ has relevant information even with the small L value (Figures 1(a) and 1(c)). Figure 2 shows that the convolution vectors prepared with IGC or SGC(MAP) give a reasonable data.

Although the RMSE in Figure 2 is one of the good properties to discuss the accuracy, it is also important to check whether the prepared features can reproduce the pKa difference stemming from the structural difference (Scheme 1). In Table 1, the pKa values of hydroxybenzoic acids predicted with IGC(MAP), IGC(IAP), SGC(MAP), SGC(IAP), MolGpKa, and OPERA are shown. As a reference, the pKa values computed with QM approaches are also shown.[34] The obtained pKa values reproduced the pKa values computed with the QM, except for MolGpKa. Moreover, IGC(MAP), SGC(IAP), and OPERA can reproduce the QM result that the pKa of 4HB is larger than that of 2HB, suggesting that SGC and IGC can include the structural isomerism through convolution. However, the pKa difference between $2HB_a$ and $2HB_b$ was reproduced only with IGC(MAP), IGC(IAP), and SGC(MAP).

The results show that the IGC(MAP) gave a good feature to reproduce the pKa difference stemming from the structural difference.

Table 1: pKa of 4HB, $2HB_a$, and $2HB_b$ computed with IGC, SGC, MolGpKa, and OPERA. For comparison, the QM data computed in a previous study were also shown.[34] In IGC and SGC, IAP and MAP were employed as atomic properties and the number of layers is 10.

	IGC(MAP)	IGC(IAP)	SGC(MAP)	SGC(IAP)	MolGpKa	OPERA	QM
4HB	3.58	4.00	3.46	3.93	7.61	4.47	4.40
$2HB_a$	3.23	4.10	3.47	3.42	7.88	3.53	4.10
$2HB_b$	1.71	3.97	2.49	3.42	7.88	3.53	2.69

From the viewpoint of accuracy and the correlation between features, IGC with MAP is superior to others. However, with accuracy, it is difficult to say that the concatenated vector computed with IGC(MAP) is a good feature. To discuss the interpretability of the IGC(MAP), the integrated gradients (IGs) were computed.

$$IG_i(\mathbf{X}) = (X_i - \bar{X}_i) \int_{\alpha=0}^{1} \frac{\partial F(\bar{\mathbf{X}} + \alpha(\mathbf{X} - \bar{\mathbf{X}}))}{\partial X_i} d\alpha, \tag{15}$$

where $\mathbf{X}$ is the concatenated vector of a dissociated proton and $\bar{\mathbf{X}}$ is the baseline. Although it is well known that the baseline is important in calculating IGs, there is no universal rule to define the baseline. It is also difficult to determine the baseline of pKa. As shown in a previous study,[26] most of the DataWarrior acidic pKa values, which are freely available pKa dataset,[35] are whithin the range ($0 < \text{pKa} < 14$). Therefore, the middle of the pKa range ($\text{pKa} = 7$) is a candidate for the baseline. The **H**-value of 4-nitrophenol was chosen as the baseline because the pKa value is close to 7.

Figure 3 (a) summarizes the IGs of 4HB, $2HB_a$, and $2HB_b$ computed with the baseline. Because the pKa values of 4HB, $2HB_a$, and $2HB_b$ are < 7, the negative IGs are important. Figure 3 (a) shows that the difference among the molecules mainly comes from the properties, $M+(k = 0)$ and $M+(k = 2)$, where $M+$ is the positive MEP value. Previous studies[17,18] have shown that the MEP had a strong negative correlation with the sum of valence natural atomic orbital energies. Therefore, the IGs in Figure 3 (a) show that the pKa value decreases

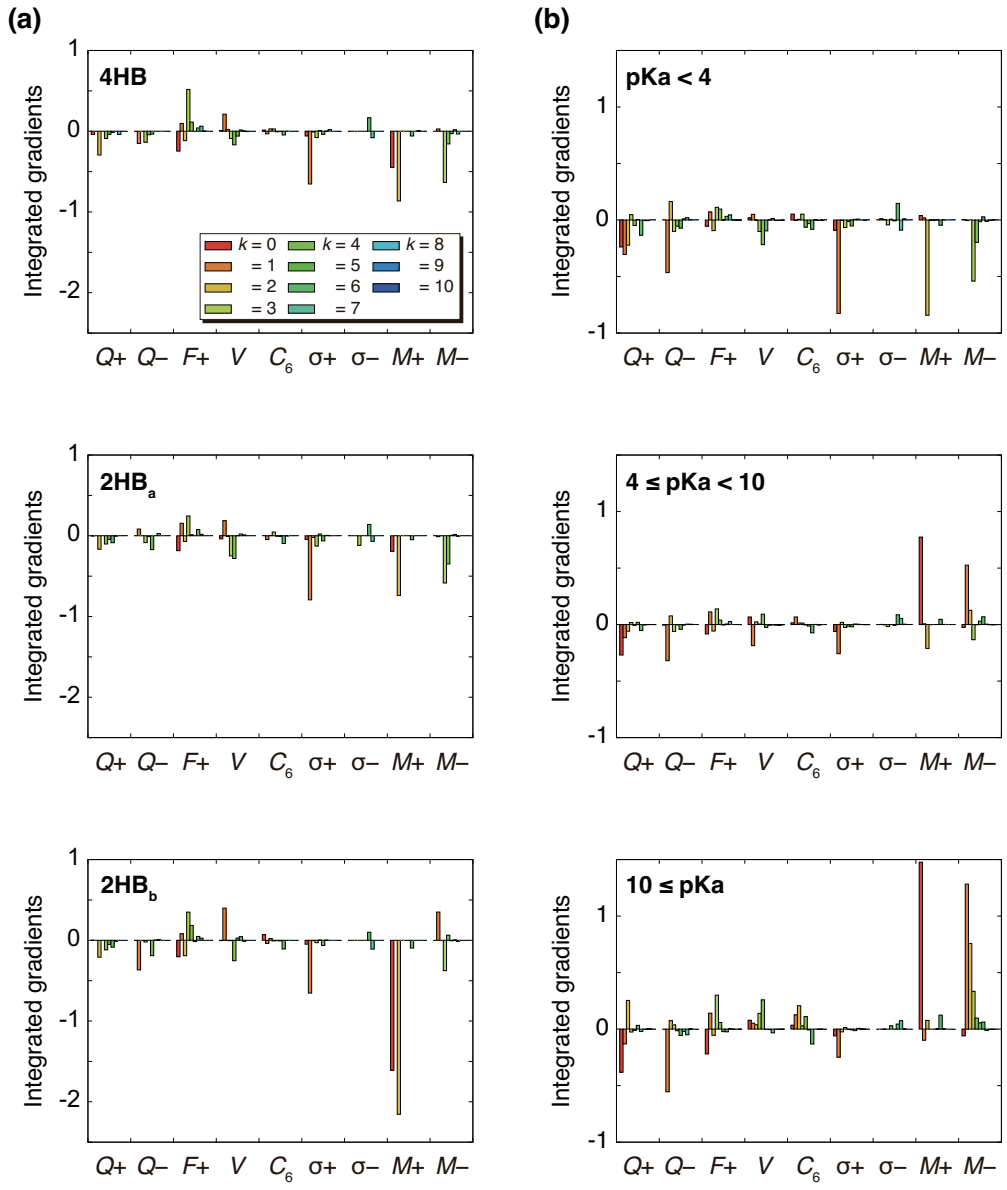
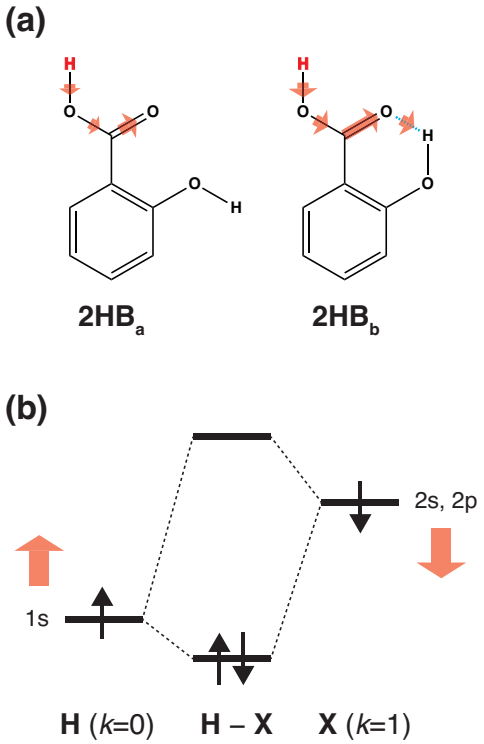


Figure 3: (a) Integrated gradients on carboxylic acid hydrogen site of hydroxybenzoic acids (4HB, 2HBa, and 2HBb) for $k = 10$ and (b) integrated gradients for $k = 10$ averaged in three pKa ranges (pKa < 4, $4 \leq$ pKa < 10, and $10 \leq$ pKa). Moreover, 4-nitrophenol was employed for the baseline molecules.

as the atomic orbital energy becomes negatively large. Because the electron-withdrawing atom makes the atomic orbital energy of the next atom more negative, the IGs in Figure 3 (a) also indicate that the pKa value decreases when the sites of $k = 0$ and 2 are surrounded by the more electron-withdrawing atoms. As shown in Scheme 3(a), in the case of $2HB_a$,



Scheme 3: (a) Interpretation of pKa difference between $2HB_a$, and $2HB_b$ derived from IGs. The deprotonated site is colored red, and the arrow size shows the electron-withdrawing strength schematically. (b) Interpretation of large pKa value in the range ($10 \leq$ pKa) based on the chemical bonding between the H and X atoms. The chemical bond (H-X) comprises 1s orbital on the H site and 2s, and 2p orbitals on the X site. The red arrows indicate orbital energy changes induced by increased $M+$ and $M-$ values.

and $2HB_b$, the sites of $k = 0$ and 2 are the proton H and carbonyl C sites, respectively. When the chemical structure is considered, the carbonyl O atom of $2HB_b$ can withdraw the electron on the carbonyl C atom more strongly than that of $2HB_a$. From the pKa difference between $2HB_a$ and $2HB_b$, and Scheme 3(a), the explanation by IGs is reasonable.

Although the interpretation in Scheme 3 (a) is reasonable for an acidic compound (2HB), checking the interpretation along the pKa value is also important. To discuss the interpretation change, the average of IGs in the three pKa ranges (pKa < 4, $4 \leq$ pKa < 10, and 10

$\leq$ pKa) were obtained. In Figure 3(b), the averaged IGs are shown. Although the averaged IGs in the range (pKa $<$ 4) are similar to Figure 3(a), the averaged IGs in the range (10 $\leq$ pKa) differ totally from Figure 3(a). In the weak acid condition (10 $\leq$ pKa), the IGs of $M+(k=0)$ and $M-(k=1)$ are positively large. The obtained IG is reasonable because of the following reasons. When the $M+$ and $M-$ values increase, the orbital energy difference decreases (Scheme 3(b)), and the polarity of the bond decreases. The large positive IG suggests that the low polarity in the chemical bond makes the pKa value positive (less acidic), which is reasonable from the chemical viewpoint if the size effect is omitted. Scheme 3 shows that the ML model obtained with IGC(MAP) gives a reasonable interpretation for chemists.

Conclusions

In this study, a new feature in molecular GNNs was proposed, and the accuracy, overcorrelation between features, and interpretability were discussed in detail. The overcorrelation and accuracy indicate that the IGC with MAP is superior to others. The prediction output with the IGC(MAP) was analyzed using the IGs method. From the analysis, positive values of MEP ($k = 0 and 2$) is important in acidic conditions, whereas the positive value of MEP ($k = 0$) and the negative value of MEP ($k = 1$) are important in basic conditions, which leads to reasonable interpretation from the chemistry viewpoint.

In this study, a part of the concatenated vectors $\{\mathbf{H}^{(L)}\}$ was employed in the ML model. In the future study, we will employ all $\{\mathbf{H}^{(L)}\}$ in a molecule to construct the ML model for predicting molecular properties, such as solvation free energy and octanol/water partition coefficient.

Acknowledgement

This study is supported by JST, PRESTO Grant Number JPMJPR21C9, and the Leading Initiative for Excellent Young Researchers. We also acknowledge Enago (www.enago.jp) for

the English language review.

Supporting Information Available

- Supporting_info.pdf: Details of the machine learning process and RMSEs of the pKa values obtained by IGC and two types of atom features (IAP and MAP) with a different number of convolution layers.

References

(1) Haury, A.-C.; Gestraud, P.; Vert, J.-P. The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures. *PLoS ONE* **2011**, *6*, e28210.

(2) Acharya, D. B.; Zhang, H. Feature Selection and Extraction for Graph Neural Networks. 2019.

(3) Jin, W.; Liu, X.; Ma, Y.; Aggarwal, C.; Tang, J. Feature Overcorrelation in Deep Graph Neural Networks. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022.

(4) Ishida, S.; Terayama, K.; Kojima, R.; Takasu, K.; Okuno, Y. Prediction and Interpretable Visualization of Retrosynthetic Reactions Using Graph Convolutional Networks. *J. Chem. Inf. Model.* **2019**, *59*, 5026–5033.

(5) Kojima, R.; Ishida, S.; Ohta, M.; Iwata, H.; Honma, T.; Okuno, Y. kGCN: a graph-based deep learning framework for chemical structures. *J. Cheminform.* **2020**, *12*, 32.

(6) Jiang, M.; Li, Z.; Zhang, S.; Wang, S.; Wang, X.; Yuan, Q.; Wei, Z. Drug-target affinity prediction using graph neural network and contact maps. *RSC Adv.* **2020**, *10*, 20701–20712.

(7) Kensert, A.; Bouwmeester, R.; Efthymiadis, K.; Van Broeck, P.; Desmet, G.; Cabooter, D. Graph Convolutional Networks for Improved Prediction and Interpretability of Chromatographic Retention Data. *Anal. Chem.* **2021**, *93*, 15633–15641.

(8) Li, Q.; Han, Z.; Wu, X.-M. Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning. 2018; `https://arxiv.org/abs/1801.07606`.

(9) Chen, M.; Wei, Z.; Huang, Z.; Ding, B.; Li, Y. Simple and Deep Graph Convolutional

Networks. Proceedings of the 37th International Conference on Machine Learning. 2020; pp 1725–1735.

(10) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V.; Leswing, K.; Wu, Z. *Deep Learning for the Life Sciences*; O'Reilly Media, 2019; `https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837`.

(11) Choudhary, K.; DeCost, B. Atomistic Line Graph Neural Network for improved materials property predictions. *npj Comput. Mater.* **2021**, *7*, 185.

(12) Reed, A. E.; Weinstock, R. B.; Weinhold, F. Natural population analysis. *J. Chem. Phys.* **1985**, *83*, 735–746.

(13) Bader, R. F. W. A quantum theory of molecular structure and its applications. *Chem. Rev.* **1991**, *91*, 893–928.

(14) Yokogawa, D. Isotropic Site-Site Dispersion Potential Constructed Using Quantum-Chemical Calculations and a Geminal Auxiliary Basis Set. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 748–753.

(15) Yokogawa, D.; Suda, K. Electrostatic Potential Fitting Method Using Constrained Spatial Electron Density Expanded with Preorthogonal Natural Atomic Orbitals. *J. Phys. Chem. A* **2020**, *124*, 9665–9673.

(16) Politzer, P.; Laurence, P. R.; Jayasuriya, K. Molecular electrostatic potentials: an effective tool for the elucidation of biochemical phenomena. *Environmental Health Perspectives* **1985**, *61*, 191–202.

(17) Liu, S.; Schauer, C. K.; Pedersen, L. G. Molecular acidity: A quantitative conceptual density functional theory description. *J. Chem. Phys.* **2009**, *131*, 164107.

(18) Liu, S.; Pedersen, L. G. Estimation of Molecular Acidity via Electrostatic Potential

at the Nucleus and Valence Natural Atomic Orbitals. *J. Phys. Chem. A* **2009**, *113*, 3648–3655.

(19) Lutzeyer, J. F.; Wu, C.; Vazirgiannis, M. Sparsifying the Update Step in Graph Neural Networks. 2021.

(20) Xu, K.; Li, C.; Tian, Y.; Sonobe, T.; Kawarabayashi, K.-i.; Jegelka, S. Representation Learning on Graphs with Jumping Knowledge Networks. 2018; `https://arxiv.org/abs/1806.03536`.

(21) Wu, F.; Zhang, T.; Souza, A. H. d.; Fifty, C.; Yu, T.; Weinberger, K. Q. Simplifying Graph Convolutional Networks. 2019.

(22) Liu, X.; Lei, F.; Xia, G.; Zhang, Y.; Wei, W. AdjMix: simplifying and attending graph convolutional networks. *Complex Intell. Syst.* **2022**, *8*, 1005–1014.

(23) Wiberg, K. B. Application of the Pople-Santry-Segal CNDO method to the cyclopropylcarbinyl and cyclobutyl cation and to bicyclobutane. *Tetrahedron* **1968**, *24*, 1083–1096.

(24) Hashem, S.; Schmeiser, B. Improving model accuracy using optimal linear combinations of trained neural networks. *IEEE Transactions on Neural Networks* **1995**, *6*, 792–794.

(25) Hashem, S. Optimal Linear Combinations of Neural Networks. *Neural Networks* **1997**, *10*, 599–614.

(26) Mansouri, K.; Cariello, N. F.; Korotcov, A.; Tkachenko, V.; Grulke, C. M.; Sprankle, C. S.; Allen, D.; Casey, W. M.; Kleinstreuer, N. C.; Williams, A. J. Open-source QSAR models for pKa prediction using multiple machine learning approaches. *J. Cheminform.* **2019**, *11*, 60.

(27) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019.

(28) Yanai, T.; Tew, D. P.; Handy, N. C. A new hybrid exchange-correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chem. Phys. Lett.* **2004**, *393*, 51–57.

(29) Kendall, R. A.; Dunning, Jr., T. H.; Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

(30) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. General atomic and molecular electronic structure system. *J. Comput. Chem.* **1993**, *14*, 1347–1363.

(31) Frisch, M. J. et al. Gaussian 16 Revision A.03. Gaussian Inc. Wallingford CT 2016.

(32) Opolka, F.; Zhi, Y.-C.; Lió, P.; Dong, X. Adaptive Gaussian Processes on Graphs via Spectral Graph Wavelets. Proceedings of The 25th International Conference on Artificial Intelligence and Statistics. 2022; pp 4818–4834.

(33) Pan, X.; Wang, H.; Li, C.; Zhang, J. Z. H.; Ji, C. MolGpka: A Web Server for Small Molecule pKa Prediction Using a Graph-Convolutional Neural Network. *J. Chem. Inf. Model.* **2021**, *61*, 3159–3165.

(34) Baba, T.; Matsui, T.; Kamiya, K.; Nakano, M.; Shigeta, Y. A density functional study on the pKa of small polyprotic molecules. *Int. J. Quantum Chem.* **2014**, *114*, 1128–1134.

(35) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *J. Chem. Inf. Model.* **2015**, *55*, 460–473.