Target-directed Atomic Importance Estimation via Reverse Self-attention

Anonymous authors
Paper under double-blind review

Abstract

Estimating the importance of each atom in a molecule is one of the most appealing and challenging problems in chemistry, physics, and material engineering. The most common way to estimate the atomic importance is to compute the electronic structure using density-functional theory (DFT), and then to interpret it using domain knowledge of human experts. However, this conventional approach is impractical to the large molecular database because DFT calculation requires huge computation, specifically, $O(n^4)$ time complexity w.r.t. the number of electrons in a molecule. Furthermore, the calculation results should be interpreted by the human experts to estimate the atomic importance in terms of the target molecular property. To tackle this problem, we first exploit machine learning-based approach for the atomic importance estimation. To this end, we propose reverse self-attention on graph neural networks and integrate it with graph-based molecular description. Our method provides an efficiently-automated and target-directed way to estimate the atomic importance without any domain knowledge on chemistry and physics.

1 Introduction

In molecules, each atom has the importance in manifesting the entire molecular properties, and estimating such atomic importance plays a key role in interpreting molecular systems. For these reasons, the atomic importance estimation has been consistently studied in the scientific communities (Yen & Winefordner, 1976; Tang et al., 2016; Pan et al., 2018). However, estimating the atomic importance is one of the most challenging tasks in chemistry and quantum mechanics because the importance of each atom is comprehensively determined based on atomic properties, neighbor atoms, bonding types, and target molecular property.

The most common approach for estimating the atomic importance is to interpret the electronic structure using density-function theory (DFT) (Sholl & Steckel, 2009). In this approach, the atomic importance is estimated through three steps: 1) A human expert selects appropriate functional and basis sets for a given molecule to apply DFT; 2) The electronic structure of the molecule is calculated based on DFT calculation; 3) The human expert estimates the atomic importance by interpreting the calculated electronic structure in terms of target molecular property. Although some methods are developed to estimate relative contributions of atoms in molecules, their generality is typically limited to the molecular properties (Marenich et al., 2012; Glendening et al., 2019). For this reason, DFT that can generate a general description of the molecule has been most widely used to interpret the molecular systems and to reveal important atoms for target molecular property (Crimme et al., 2010; Lee et al., 2018b; Chibani et al., 2018).

However, the conventional approach based on DFT has three fundamental limitations in efficiency, automation, and generality.

- **Efficiency:** As an example of the electronic structure computations, DFT calculation requires $O(n^4)$ time complexity to compute the electronic structure, where n is the number of basis functions that describe electrons in the molecule. Generally, molecules have more electrons than atoms.

- **Automation:** DFT cannot automatically generate all target-specified physical properties in principle, so human expert should manually select additional calculation method to com-

pute target molecular property from the electronic distributions. That is, domain knowledge of the human experts is necessarily required to estimate the atomic importance in terms of the target molecular property.

- **Generality:** For some molecular properties, the relationship between them and the electronic distributions is not clear. Moreover, sometimes the estimation is impossible because the relationships between molecular property and molecular structure are not interpretable.

For these limitations, estimating the atomic importance is remaining as one of the most challenging problems in both science and engineering such as physics, chemistry, pharmacy, and material engineering.

To overcome the limitations of the conventional approach in estimating the atomic importance, we first exploit machine learning-based approach. To this end, we propose a new concept of reverse self-attention and integrate it with the graph neural networks. The self-attention mechanism was originally designed to determine important elements within the input data to accurately predict its corresponding target or label in natural language processing (Vaswani et al., 2017). Similarly, in graph neural networks, self-attention is used to determine important neighbor nodes within the input graph to generate more accurate node or graph embeddings (Velickovic et al., 2018). Our reverse self-attention is defined as the inverse of the self-attention to calculate how important a selected node is considered in the graph. For a given molecule and target property, the proposed estimation method selects the atom that has the largest reverse self-attention score as the most important atom.

The proposed method estimates the target-directed atomic importance through two steps: 1) For the given molecular graphs and their corresponding target properties, self-attention-based graph neural network is trained; 2) After the training, the reverse self-attention scores are calculated, and then the atomic importance is estimated according to the reverse self-attention scores. As shown in this estimation process, neither huge computation nor human experts in chemistry and physics is required. Thus, the proposed method provides an efficient and fully-automated way to estimate the atomic importance in terms of the target molecular property via target-aware training of the graph self-attention.

To validate the effectiveness of the proposed method, we conducted comprehensive experiments and evaluated the estimation performance using both quantitative and qualitative analyses. The contributions of this paper are summarized as:

- This paper first proposes a machine learning-based approach to estimate the atomic importance in the molecule.
- The proposed method drastically reduced the computational cost for the atomic importance estimation from $O(n^4)$ time complexity to the practical time complexity of the graph-based deep learning.
- The proposed method provides a fully-automated and target-directed way to estimate the atomic importance.
- We comprehensively validated the effectiveness of the proposed method using both quantitative and qualitative evaluations. However, since none of a labeled dataset for the atomic importance estimation and a systematic way to quantitatively evaluate the estimation accuracy, we devised a systematic quantitative evaluation method and validated the effectiveness of the proposed method using it.

2 PRELIMINARIES

Before describing the atomic importance estimation based on the reverse self-attention, we briefly two essential concepts for understanding the proposed method in this section: 1) graph-based molecular analysis; 2) graph self-attention and graph attention network.

2.1 GRAPH-BASED MOLECULAR ANALYSIS

Graph-based molecular machine learning has attracted significant attention from both scientific and machine learning communities. It has shown state-of-the-art performance in various scientific applications such as molecular or crystal property prediction (Wu et al., 2018; Xie & Grossman, 2018; Lu
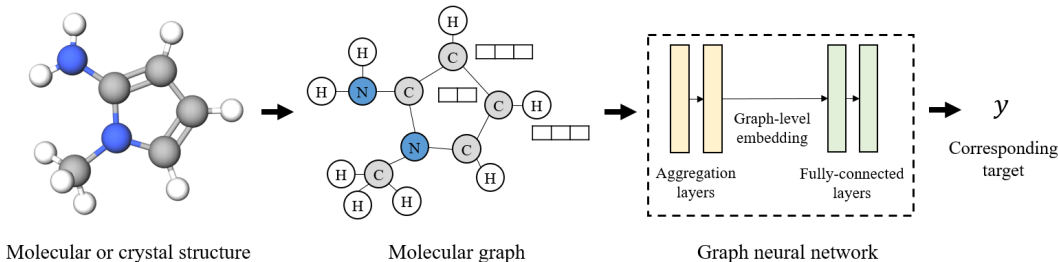
Figure 1: Overall process of the graph-based molecular machine learning. In the molecular graph, three and two-dimensional vectors mean atom-features and bond-features, respectively.

et al., 2019), molecular generation (You et al., 2018; Gao & Ji, 2019), and atomic reaction analysis (Coley et al., 2019). In the graph-based molecular machine learning, a molecule is represented as an undirected feature graph $\mathcal{G} = (\mathbb{V}, \boldsymbol{A}, \boldsymbol{X}, \boldsymbol{U})$, where $\mathbb{V}$ is a set of atoms (nodes); $\boldsymbol{A} \in \{0, 1\}^{|\mathbb{V}| \times |\mathbb{V}|}$ is an adjacency matrix indicating the existence of the atomic bonds (edges); $\boldsymbol{X} \in \mathbb{R}^{|\mathbb{V}| \times d}$ is a d-dimensional atom-feature matrix; $\boldsymbol{U} \in \mathbb{R}^{|\mathbb{B}| \times m}$ is a m-dimensional bond-feature matrix; and $\mathbb{B}$ is a set of atomic bonds. For a given labeled molecular dataset $D = \{(G_1, y_1), ..., (G_{|D|}, y_{|D|})\}$, the goal of the graph-based molecular machine learning is to build a model $f : G \rightarrow y$, where y can be target value, class label, or another molecular graph. In this paper, we will refer atom and bond as node and edge, respectively.

Fig. 1 shows the overall process of the graph-based molecular machine learning. First, molecular or crystal structures formatted by .xyz or .cif are converted into the molecular graph with the adjacency matrix $\boldsymbol{A}$, node-feature matrix $\boldsymbol{X}$, and edge-feature matrix $\boldsymbol{U}$. Then, graph neural network generates the graph-level embedding of the input molecular graph through the aggregation layers and predict the corresponding target y.

In the aggregation layer of the graph neural network, each node in the graph is converted into the node-embedding vector, and it is stacked to the node-embedding matrix in column-wise. The output node-embedding matrix of the kth aggregation layer, $\boldsymbol{H}^{(k)}$, is calculated by:

$$\boldsymbol{H}^{(k)} = \psi(\boldsymbol{A}, \boldsymbol{U}, \boldsymbol{H}^{(k-1)}, \boldsymbol{W}^{(k)}), \tag{1}$$

where ψ is an aggregation function, and $\boldsymbol{W}^{(k)}$ is a trainable weight matrix of the kth aggregation layer. Note that $\boldsymbol{H}^{(0)}$ is the input node-feature matrix $\boldsymbol{X}$. For example, in graph convolutional network (GCN) (Kipf & Welling, 2017), $\boldsymbol{H}^{(k)}$, is calculated by:

$$\boldsymbol{H}^{(k)} = \text{ReLU}(\tilde{\boldsymbol{A}} \boldsymbol{H}^{(k-1)} \boldsymbol{W}^{(k)}), \tag{2}$$

where $\tilde{A}$ is an adjacency matrix containing self-loop and can be calculated by $\tilde{A} = A + I$. After the node-embedding, the node-embedding matrix of the last aggregation layer is converted into the graph-level embedding vector by readout function (Zhou et al., 2018). For the implementation of the readout function, mean or max-based operations are commonly used (Lee et al., 2018a; Ying et al., 2018). Finally, the target corresponding the input graph is predicted by interpreting the output graph-level embedding vector through the fully-connected layers.

2.2 GRAPH ATTENTION AND GRAPH ATTENTION NETWORK

The attention and the self-attention mechanisms are originally introduced in natural language processing to refer other data and elements to improve prediction and classification accuracies, respectively (Luong et al., 2017; Vaswani et al., 2017). In graph neural networks based on neighbor node aggregation approach such as GCN, the self-attention mechanism is used to calculate the importance of neighbor nodes in generating the node-embedding vectors.

Graph attention network (GAT) (Velickovic et al., 2018) is the first neural network to apply the self-attention mechanism in graph-based deep learning. By exploiting the self-attention mechanism, GAT showed highly improved prediction and classification accuracies (Velickovic et al., 2018; Wang et al., 2019). In the kth aggregation layer of GAT, the attention coefficient between node i and its

neighbor node j is defined by:

$$e_{ij}^{(k)} = \text{LeakyReLU}\left(f\left(\boldsymbol{V}^{(k)}\boldsymbol{H}_{:,i}^{(k-1)} \oplus \boldsymbol{V}^{(k)}\boldsymbol{H}_{:,j}^{(k-1)}\right)\right), \tag{3}$$

where f is a feedforward neural network; $\boldsymbol{V}$ is a trainable weight matrix; and $\oplus$ is a vector concatenation. Based on the attention coefficient, the attention score, $\alpha_{ij}^{(k)}$, is calculated by:

$$\alpha_{ij}^{(k)} = \frac{\exp(e_{ij}^{(k)})}{\sum_{q\in\mathbb{N}(i)}\exp(e_{iq}^{(k)})}, \tag{4}$$

where $\mathbb{N}(i)$ is a set of neighbor nodes of node i in the graph. Finally, in the kth aggregation layer of GAT, the node-embedding vector of node i is calculated by:

$$\boldsymbol{H}_{:,i}^{(k)} = \psi\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(k)}\boldsymbol{V}^{(k)}\boldsymbol{H}_{:,j}^{(k-1)}\right). \tag{5}$$

In GAT, the aggregation layer that generates the node-embeddings based on the graph self-attention is called graph attention layer.

3 Machine Learning-based Atomic Importance Estimation

In this section, we explain our machine learning-based approach for estimating the atomic importance. To devise the machine learning-based importance estimator, we define a new concept of reverse self-attention and integrate it with GAT.

3.1 Reverse Self-attention

The self-attention mechanism in the molecular graph provides numerical importance of each neighbor atom in terms of target molecular property. To exploit the concept of the self-attention for estimating the atomic importance, reverse self-attention in the kth graph attention layer is defined as the inverse of the self-attention:

$$\rho_i^{(k)} = \sum_{j\in\mathbb{N}(i)} \alpha_{ji}^{(k)}. \tag{6}$$

That is, the reverse self-attention of node i means how much attention node i receives from its neighbor nodes. Fig. 2 shows an example of the reverse self-attention in the graph with five nodes. In this example, the reverse self-attention of n_1 is calculated as the sum of $\alpha_{21} + \alpha_{31} + \alpha_{41}$.

Self-attentions of n_1: $\mathbb{A}_1 = \{\alpha_{12}, \alpha_{13}, \alpha_{14}\}$ Reverse self-attention of n_1: $\rho_1 = \alpha_{21} + \alpha_{31} + \alpha_{41}$
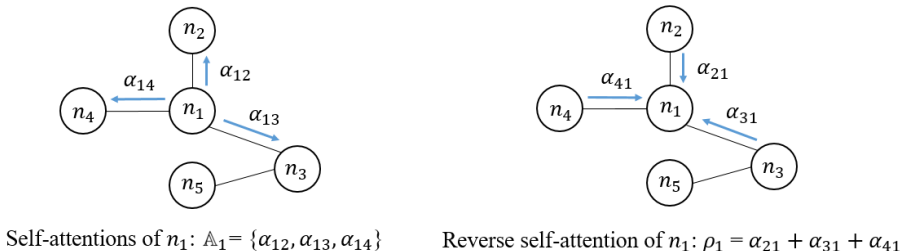
Figure 2: Example of reverse self-attention in the graphs containing five nodes.

As shown in the definition of the reverse self-attention, the additional time and space complexities of the reverse self-attention are negligible because calculating the reverse self-attention is just the sum of pre-computed self-attention scores in prediction or classification time.

In GAT, the self-attention scores of the kth graph attention layer are calculated based the node-embeddings of the $(k-1)^{\text{th}}$ graph attention layer. That is, the self-attention scores of the first layer are calculated by considering 1-hop neighbor nodes, and the second layer's self-attention scores are determined by considering 2-hop neighbor nodes. Thus, the reverse self-attention in the first
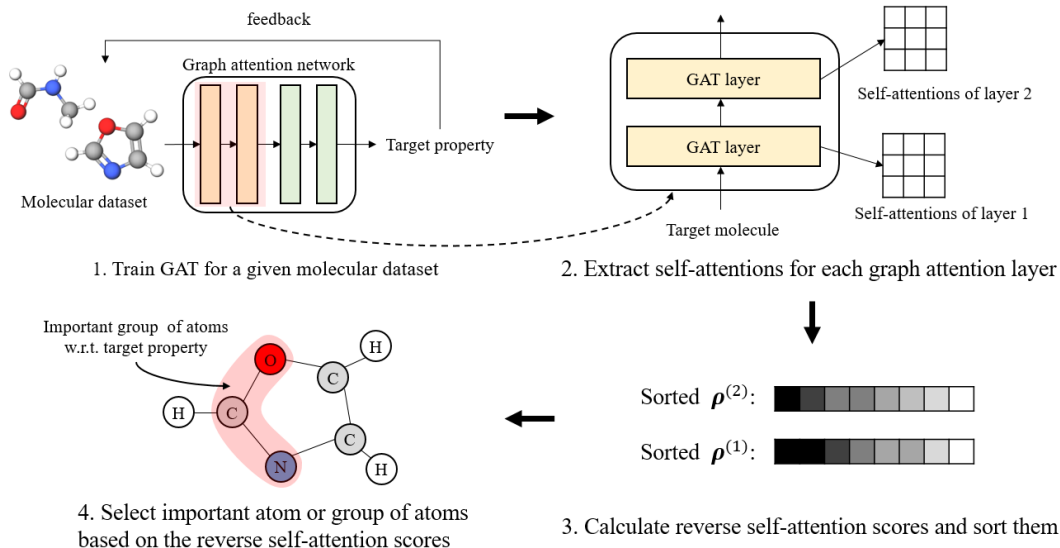
Figure 3: Overall process of building machine learning-based atomic importance estimator with the reverse self-attention mechanism and selecting important atom or group of atoms using the estimator.

graph attention layer indicates the importance of each atom itself, and the second layer's reverse self-attention means the importance of each atom group consists of 1-hop neighbor atoms, not an atom itself. These interpretations can be extended to the atomic importance for the group of atoms that consists of k-hop neighbor atoms by the reverse self-attention in the $(k+1)^{\text{th}}$ graph attention layer.

3.2 TARGET-DIRECTED ATOMIC IMPORTANCE ESTIMATOR

This section explains the way to build a fully-automated and target-directed atomic importance estimator based on the reverse self-attention and GAT. We call this estimator **M**achine **I**ntelligence-based **A**tomic **I**mportance **E**stimator (MIAIE). Fig. 3 shows overall process of building MIAIE and estimating the target-directed atomic importance via MIAIE. The atomic importance estimation using MIAIE consists of four steps:

- **Step 1:** Train GAT for regression or classification on the molecular dataset with the target molecular property. In this training process, GAT automatically learns the self-attentions scores in terms of predicting the target molecular property.
- **Step 2:** Predict the target property of the interesting molecule and extract self-attention scores for each graph attention layer.
- **Step 3:** Calculate the reverse self-attention scores using Eq. (6) and sort them for each graph attention layer.
- **Step 4:** Select atom or group of atoms that have the largest reverse self-attention score as the most important element in the molecule in terms of the target property.

As shown in the overall process of estimating the atomic importance using MIAIE, the estimation process is fully-automated and does not require any domain knowledge of the human experts in chemistry and physics. MIAIE is also incomparably efficient than the conventional approach using DFT calculations with $O(n^4)$ time complexity because it uses only the graph neural networks to describe the molecules, and the time complexity of the estimation process (step 3 and 4) is negligible.

4 EXPERIMENTS

To accurately validate the effectiveness of MIAIE, we conducted both quantitative and qualitative evaluations on two well-known molecular datasets. However, to the best of our knowledge, neither

a labeled dataset for the atomic importance estimation nor a systematic way to evaluate the performance of the atomic importance estimator exists. For this reason, we devised a validation method to quantitatively evaluate the performance of the atomic importance estimators. We will explain this validation method in Section 4.3. We used MolView[1] to visualize the estimation results of MIAIE.

4.1 BENCHMARK MOLECULAR DATASET

For the experiments, we used two well-known molecular datasets: Quantum Mechanics9 (QM9) (Ramakrishnan et al., 2014; Ruddigkeit et al., 2012) and Estimated SOLubility (ESOL) (Delaney, 2004). These datasets were randomly split into 90% train and 10% test subsets because the test dataset for them are not provided. In the experiments, we validated the effectiveness of MIAIE by splitting train and test datasets to evaluate the generalization capability as well as the atomic importance estimation accuracy. However, we can use MIAIE by fitting it to a given dataset only without considering the generalization capability—for example, training without L_2-regularization (Krogh & Hertz, 1992), dropout (Srivastava et al., 2014), or batch normalization (Ioffe & Szegedy, 2015)— if our analysis is only focused on specific molecular datasets containing the interesting molecules.

To quantitatively evaluate the effectiveness of MIAIE, QM9 dataset is used. It is a subset of GDB-17 database (Ramakrishnan et al., 2014; Ruddigkeit et al., 2012), in which the structural information of the molecules is given in Cartesian coordinates. QM9 dataset contains 133,886 organic molecules and several target molecular properties such as highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO), and their gap (HOMO-LUMO gap). In the experiments, we used HOMO-LUMO gap as the target molecular property because it is an essential property describing the molecular systems and one of the difficult properties to be interpreted from the molecular structure. Molecules in QM9 dataset have 0.0245∼0.6221 HOMO-LUMO gaps.

For the qualitative analysis, we used ESOL dataset that contains aqueous solubility as a target property because the important atoms in terms of aqueous solubility can be easily determined by a human expert. ESOL dataset was originally published with the aqueous solubility of 2,874 compounds but a smaller subset of 1,128 compounds is recently used in chemistry (Wu et al., 2018), and we also used the subset of ESOL for the experiment. Unlike QM9 dataset, the structural information of the molecules in ESOL is provided by SMILES (Weininger, 1988) representation that does not present hydrogen. For this reason, hydrogen in the molecule is ignored in estimating the atomic importance.

4.2 IMPLEMENTATION

We implemented MIAIE using PyTorch[2] and PyTorch Deep Graph Library (DGL)[3]. GAT in MIAIE is also implemented based on the neural network modules of DGL. In the experiments, we used GAT with two graph attention layers and two fully-connected layers. To generate graph-level embedding, mean-based readout function is applied. Mean squared error (MAE) was used as an objective function to train GAT, and L_2-regularization with 0.001 regularization coefficient was applied to improve the generalization capability of the model. As a training algorithm, Adam SGD (Kingma & Ba, 2015) was used with an initial learning rate of 0.001 to fit model parameters of GAT. To accelerate the training and improve the prediction performance of GAT, we concatenated additional molecular features about molecular weight and the number of rings to the graph-level embedding vector. The source code of MIAIE and the experiment scripts are available at `http://--------------------` (*open after the review*).

4.3 VALIDATION METHOD FOR QUANTITATIVE EVALUATION

To quantitatively evaluate the effectiveness of MIAIE, we assume that if the selected atom or group of atoms are truly important in terms of the target property, then the gap between the target properties of the original molecule and its selected sub-molecule (atom or group of atoms) will be small. Based

---

[1]http://molview.org/
[2]https://pytorch.org/
[3]https://www.dgl.ai/

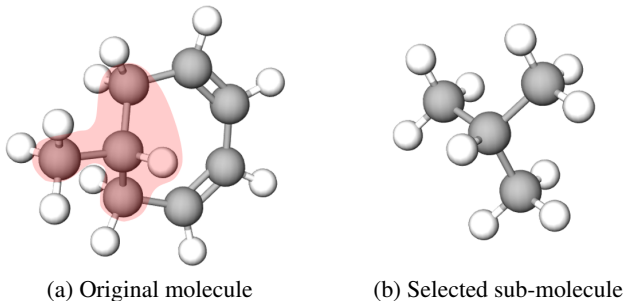(a) Original molecule   (b) Selected sub-molecule

Figure 4: An original molecule made of the carbon ring and its selected sub-molecule that consists of 1-hop neighbor atoms. Note that hydrogens are automatically attached to make a valid molecular structure. The group of atoms that is marked by a red mask means the most important atomic group estimated by MIAIE. White node: hydrogen (H); Gray node: carbon (C).

on this assumption, we define atomic importance estimation error (AIEE) as:

$$L_{AIE} = \frac{1}{|D|} \sum_{G_i \in D} \frac{|y_i - DFT(\widetilde{G}_i)|}{size(G_i)}, \tag{7}$$

where $\widetilde{G}_i$ is a sub-molecule of G_i selected by an atomic importance estimator; DFT means the value of the target property of $\widetilde{G}_i$ computed by DFT calculations; and $size(G_i)$ is the number of atoms in G_i. Since DFT can accurately calculate the value of the molecular properties in most cases even though it requires huge computation, we used DFT to calculate the value of the target molecular property of unknown $\widetilde{G}_i$. In AIEE, the error of each molecule is divided by the number of atoms in the molecule. It is an essential part of AIEE to accurately measure the target property gaps because the larger the size of the original molecule, the greater the probability of increasing the gap between the properties of the original molecule and its smaller sub-molecule with a fixed size.

4.4 QUANTITATIVE EVALUATION

In this experiment, we quantitatively evaluated the effectiveness of MIAIE using AIEE in Eq. (7) on QM9 dataset. We estimated the atomic importance using MIAIE with the reverse self-attention of the second graph attention layer for 13,3889 test molecules in terms of HOMO-LUMO gap and observed 0.00544 on AIEE. This error is relatively small when considering the fact that the molecular properties of the original molecule cannot be preserved completely in sub-molecules.

However, we observed a relatively large error of 0.01007 compared to the mean of error (0.00544) on the molecule made of the carbon ring as shown in Fig. 4. This large error is caused due to the chemical characteristics of HOMO-LUMO gap that it is determined by the overall electronic distributions in ring-shaped molecules with the same atoms such as Fig. 4-(a). Thus, this large error in HOMO-LUMO gap is inevitable in splitting a ring-shaped molecule with the same atoms into the sub-molecules because each atom has similar importance. It is a limitation of our validation method for the quantitative evaluation. One possible future work is to modify the estimation step of MIAIE to answer "each atom has similar importance" instead of selecting important atom for the molecules like Fig. 4-(a).

4.5 QUALITATIVE EVALUATION

For ESOL dataset, we conducted a qualitative evaluation because the important atom or group of atoms in terms of aqueous solubility can be easily determined by human experts. Furthermore, we denoted the normalized atomic importance of the most important group of atoms, which is selected by MIAIE. Due to the limitation of the space, we present only three natural and two interesting estimation results in this paper.

Fig. 5 shows the five atomic importance estimation results of MIAIE on ESOL dataset. Chemically, nitrogen and oxygen are known to affect aqueous solubility more than carbon and hydrogen in
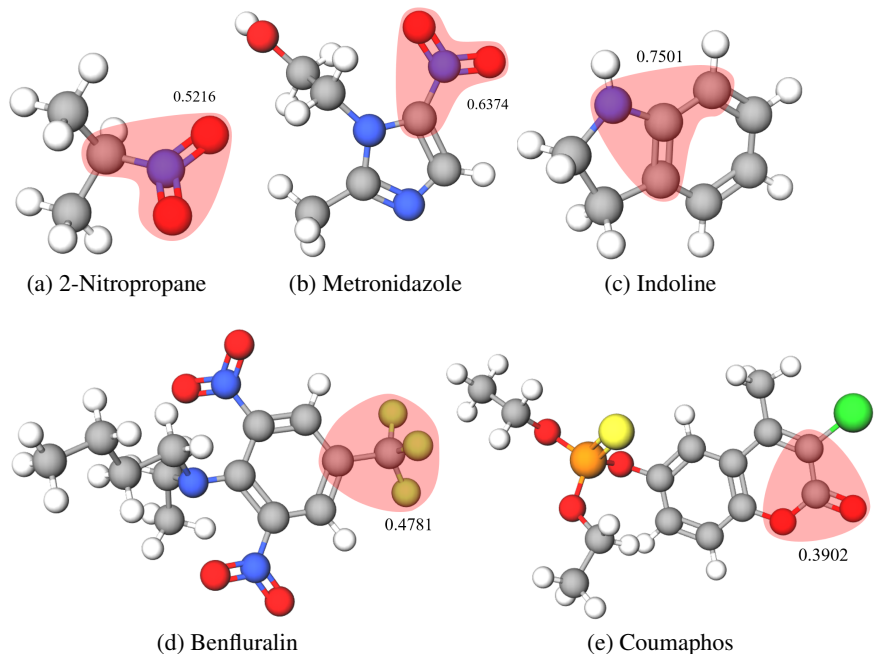
Figure 5: Original molecules and their selected sub-molecules in terms of aqueous solubility. Blue node: nitrogen (N); Red node: oxygen (O); Yellow-green node: Fluorine (F); Yellow node: Sulfur (S); Orange node: Phosphorus (P); Green node: Chlorine (Cl).

most cases. By these chemical characteristics, the estimation results of MIAIE in Fig. 5-(a)∼(c) can be easily justified: 1) Since nitrogen and oxygen exist only in the selected sub-molecule, the estimation result of 2-Nitropropane is natural; 2) Similar to the result of 2-Nitropropane, MIAIE accurately selected a group of atoms that contains the largest number of nitrogen and oxygen in Metronidazole; 3) In Indoline, only three sub-molecules that consists of 1-hop neighbor atoms can contain the nitrogen. However, among these sub-molecules, the selected group of atoms has the largest electronic charge that improve the reactivity to water, so the estimation result of MIAIE is also chemically resonable.

Some interesting results are Fig. 5-(d) and (e). Benfluralin in Fig. 5-(d) has a low aqueous solubility of -5.53 in log-scale because the C-CF$_3$ group is attached (Purser et al., 2008). Interestingly, although we did not tell this chemical observation about the relationship between the C-CF$_3$ group and aqueous solubility to MIAIE, it exactly selected the C-CF$_3$ group as the most important group of atoms in terms of aqueous solubility. Another interesting result is Coumaphos in Fig. 5-(e). In this molecule, the C-CO$_2$ group was selected as the most important group of atoms rather than PO$_3$S group that contains the largest number of oxygen. However, this estimation result is chemically reasonable because the carbons surrounding the PO$_3$S group prevent the reaction between the PO$_3$S group and water. In these estimation results, we observed that MIAIE can estimate the atomic importance by considering complex chemical rules related to the functional properties of the atomic groups and the overall environment of the molecules, even though it is trained without any domain knowledge of chemistry and quantum mechanics.

To check the correctness of the estimation result of MIAIE, we also measured the atomic importance on the molecules that have the symmetric structure because the structurally-equivalent groups of atoms must have the same atomic importance. Fig. 6 shows the estimation results of MIAIE for the molecules containing the symmetric structure. Benfluralin has a locally-symmetric structure in two C-NO$_2$ groups, and MIAIE correctly estimated the atomic importance of two C-NO$_2$ groups as the same value. Similarly, the C$_{24}$O$_4$ molecule has globally-symmetric structure throughout the overall molecular structure. For the C$_{24}$O$_4$ molecule, MIAIE correctly estimated two C-CO$_2$ groups as the most important group of atoms with the same score.
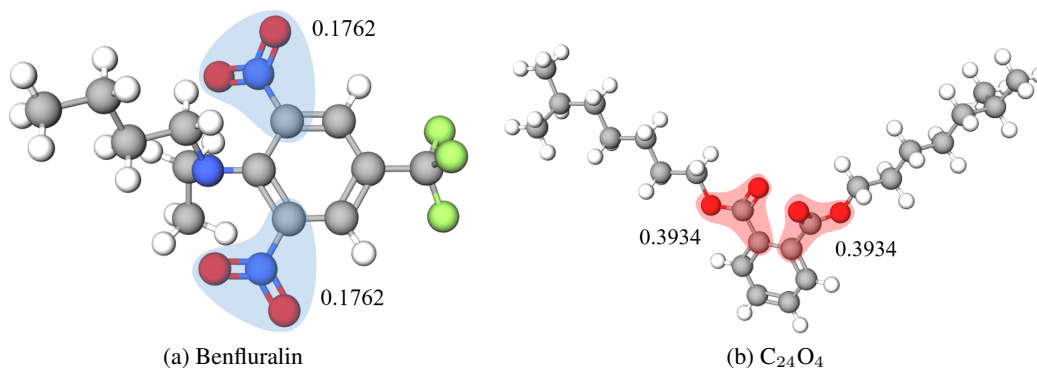
(a) Benfluralin

(b) $C_{24}O_4$

Figure 6: Two original molecules with symmetric structure and their selected sub-molecules in terms of aqueous solubility. The group of atoms that is marked by a blue mask means the second most important atomic group estimated by MIAIE.

5 CONCLUSION

This paper first exploited machine learning approach to estimate the atomic importance in molecules. To this end, the reverse self-attention was proposed and integrated with graph attention network. The proposed method is efficient and fully-automated. Furthermore, it can estimate the atomic importance in terms of the given target molecular property without human experts. However, the proposed method can estimate the importance of the group of atoms that consists of k-hop neighbor atoms only, even though some important group of atoms may have an arbitrary shape such as ring and bar. As the future work, it is necessary to modify the proposed method to estimate the importance of the group of atoms with arbitrary shape.

REFERENCES

Siwar Chibani, Michael Badawi, Thierry Loiseauu, Christophe Volkringer, Laurent Cantrel, and Jean-Francois Paul. A dft study of ruo4 interactions with porous materials: metalorganic frameworks (mofs) and zeolites. *Physical Chemistry Chemical Physics*, 20:16770–16776, 2018.

Connor W. Coley, Wengong Jin, Luke Rogers, Timothy F. Jamison, Tommi S. Jaakkola, William H. Green, Regina Barzilay, and Klavs F. Jensen. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical Science*, 10:370–377, 2019.

Stefan Crimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu. *Journal of Chemical Physics*, 132:154104, 2010.

John S. Delaney. Esol: Estimating aqueous solubility directly from molecular structure. *Chemical Informatio and Computer Sciences*, 44:1000–1005, 2004.

Hongyang Gao and Shuiwang Ji. Graph u-nets. *International Conference on Learning Representations (ICLR)*, 2019.

Eric D. Glendening, Clark R. Landis, and Frank Weinhold. Nbo 7.0: New vistas in localized and delocalized chemical bonding theory. *Journal of Computational Chemistry*, 40:2234–2241, 2019.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning (ICML)*, 2015.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR)*, 2017.

Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. *Conference on Neural Information Processing Systems (NIPS)*, 1992.

John Boaz Lee, Ryan Rossi, and Xiangnan Kong. Graph classification using structural attention. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2018a.

Yea-Lee Lee, Fangzhou Zhao, Ting Cao, Jisoon Ihm, and Steven G. Louie. Topological phases in cove-edged and chevron graphene nanoribbons: Geometric structures, z2 invariants, and junction states. *Nano Letters*, 18:7247–7253, 2018b.

Chengqiang Lu, Qi Liu, Chao Wang, Zhenya Huang, Peize Lin, and Lixin He. Molecular property prediction: A multilevel quantum interactions modeling perspective. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, 2017.

Aleksandr V. Marenich, Steven V. Jerome, Christopher J. Cramer, and Donald G. Truhlar. Charge model 5: An extension of hirshfeld population analysis for the accurate description of molecular interactions in gaseous and condensed phases. *Journal of Chemical Theory and Computation*, 8: 527–541, 2012.

Albert C. Pan, Daniel Jacobson, Konstantin Yatsenko, Duluxan Sritharan, Thomas M. Weinreich, and David E. Shaw. Atomic-level characterization of protein-protein association. *Proceedings of the National Academy of Sciences of the United Staes of America*, pp. 4244–4249, 2018.

Sophie Purser, Peter R. Moore, Steve Swallow, and Veronique Gouverneur. Fluorine in medical chemistry. *Chemical Society Reviews*, 2008.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structured and properties of 134 kilo molecules. *Nature Scientific Data*, 1, 2014.

Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Physical of Chemical Information and Modeling*, 52:2864–2875, 2012.

David S. Sholl and Janice A. Steckel. Density functional theory: A practical introduction. *Hoboken, J.J., Wiley*, 2009.

Nithish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

Yan Tang, Shu Zhao, Bo Long, Jin-Cheng Liu, and Jun Li. On the nature of support effects of metal dioxides mo2 (m = ti, zr, hf, ce, th) in single-atom gold catalysts: Importance of quantum primogenic effect. *The Journal of Physical Chemistry C*, 120:17514–17526, 2016.

Ashish Vaswani, Noam Shazeer, Niki Paramar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, and Lukasz Kaiser. Attention is all you need. *Conference on Neural Information Processing Systems (NIPS)*, 2017.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *International Conference on Learning Representations (ICLR)*, 2018.

Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S. Yu. Heterogeneous graph attention network. *International World Wide Web Conference*, 2019.

David Weininger. Smiles, a chemical language and information system. *Journal of Chemical Information and Computer Sciences*, 28:31–36, 1988.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9:513–530, 2018.

Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120:145301, 2018.

Tuan Vo Dinh Esther L. Yen and James D. Winefordner. Heavy-atom effect on room temperature phosphorimetry. *Analytical Chemistry*, 48:1186–1188, 1976.

Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. *Conference on Neural Information Processing Systems (NIPS)*, 2018.

Jiaxuan You, Bowen Liu, Rex Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. *Conference on Neural Information Processing Systems (NIPS)*, 2018.

Zie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *arXiv:1812.08434*, 2018.

A APPENDIX: EXTREME CASES ON QM9 DATASET

To provide some insight for the atomic importance estimation results of MIAIE, we present original molecules on QM9 dataset and their selected sub-molecules of two extreme cases: 1) the error is extremely small (< 0.0001); 2) the error is extremely large (> 0.01). These two extreme cases may give some chemical insight into the atomic importance in HOMO-LUMO gap and possible future work for the machine learning-based atomic importance estimation.
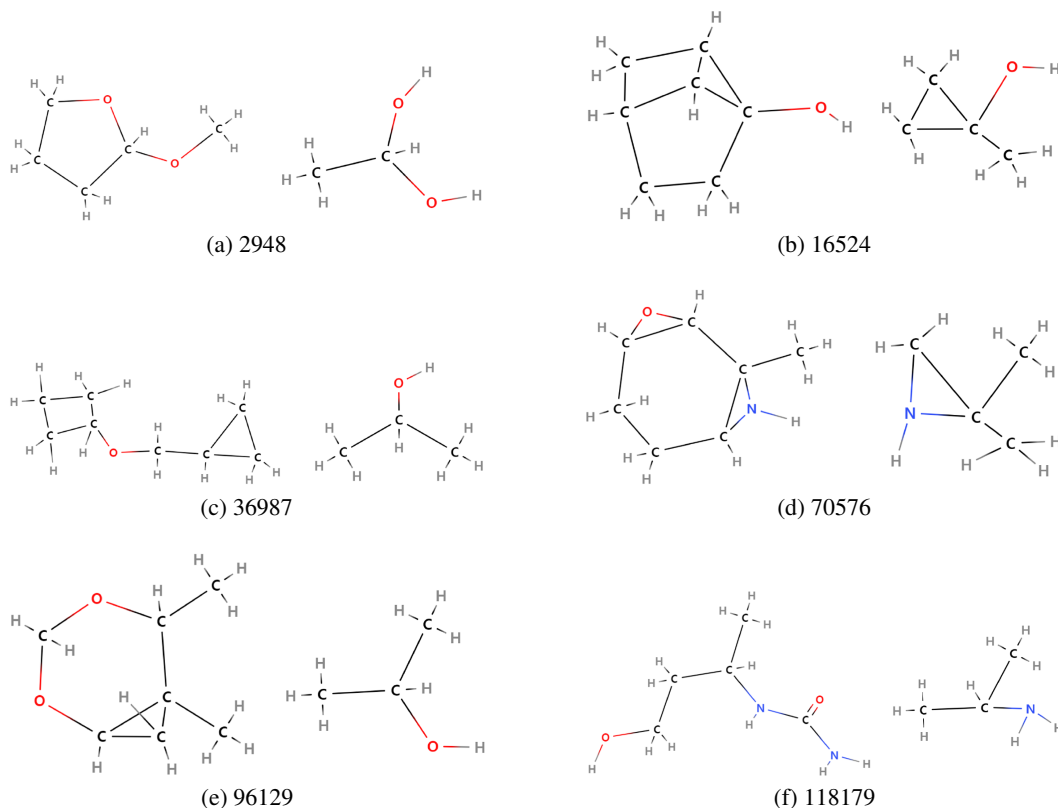

(a) 2948

(b) 16524

(c) 36987

(d) 70576

(e) 96129

(f) 118179

Figure 7: Original molecules (left) on QM9 dataset and their sub-molecules (right) showing extremely small error (< 0.0001) on AIEE. Sub-caption means the index of the original molecules on QM9 dataset.

Fig. 7 shows the original molecules and their selected sub-molecules with the extremely small error. As shown in the results, even though the molecules have carbon rings, the important group of atoms was correctly captured because the molecules are characterized by nitrogen or oxygen. On the other hand, Fig. 8 shows the original molecules and their sub-molecules with the extremely large error. Chemically, double bond plays an important role in determining HOMO-LUMO gap. However, as shown in the results of Fig. 8, sub-structures that do not contain the double bonds are selected as an important group of atoms. Thus, we need to develop a descriptor for the molecules that can emphasis the bond-features more strongly.

(a) 22299

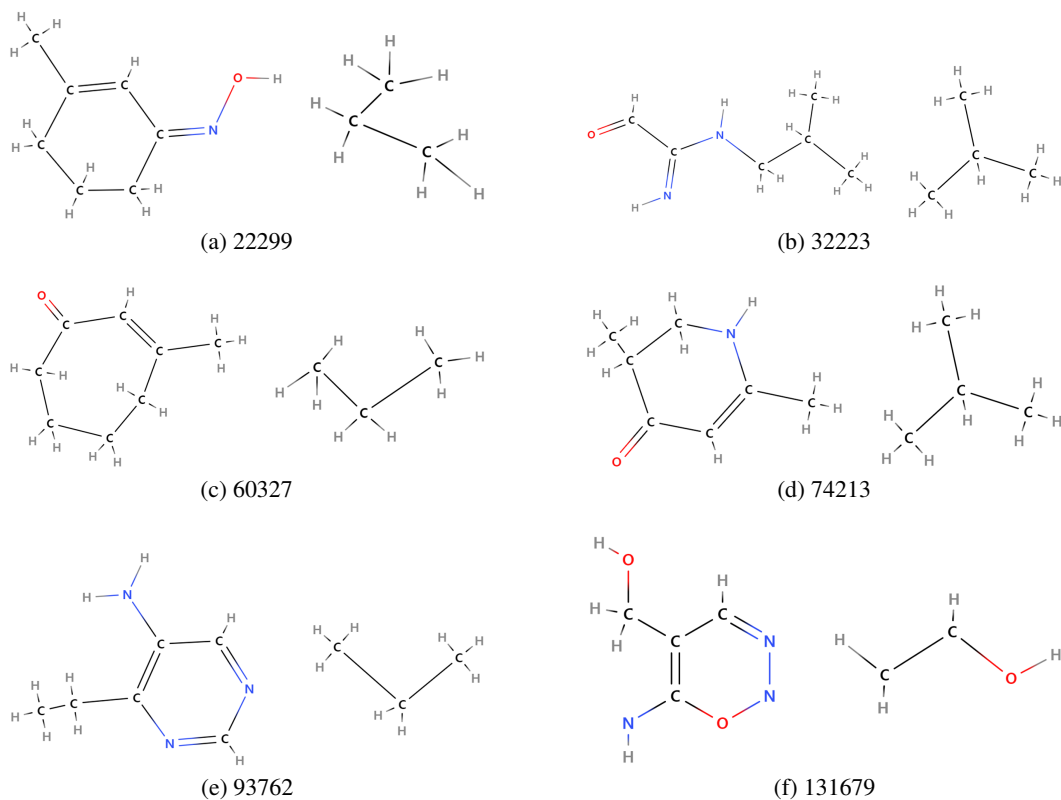(b) 32223

(c) 60327

(d) 74213

(e) 93762

(f) 131679

Figure 8: Original molecules (left) on QM9 dataset and their sub-molecules (right) showing extremely large error (> 0.01) on AIEE.