Supplementary Information of Message Passing Neural Network for Predicting Dipole Moment Dependent Core Electron Excitation Spectra

Kiyou Shibata[1,*] and Teruyasu Mizoguchi[1]

[1] *Institute of Industrial Sicence, the University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan.*

EVALUATION OF NOISE AND METRICS

In this study, we employed mean squared error (MSE) as the loss function during training and as a metric for evaluating the model's prediction accuracy. To validate the appropriateness of MSE and assess its performance under practical noise conditions in spectroscopy, we conducted an evaluation using simulated noise-added spectra. We assumed simple Poisson noise with a noise factor λ, introducing no correlation between each energy bin to replicate the noise observed in spectra:

$$S_{\text{noise}} = \text{Poisson}(S_{\text{mol}}(\mathcal{G}, \hat{\mathbf{n}}) \times \lambda)/\lambda, \tag{S1}$$

Figure S1a displays Poisson noise spectra with varying intensities, illustrating how spectral shapes change in response to different levels of Poisson noise. To evaluate MSE and the metric's performance, we randomly selected ten spectra from the dataset shown in Fig. S1b. The relationship between MSE and the Poisson noise factor λ is illustrated in Fig. S1c, depicting clear linearity between MSE and the intensity of Poisson noise for all spectra.

Additionally, we utilized spectral discriminatory entropy (SDE) to assess the similarity between two spectral metrics: MSE and spectral angle mapper (SAM). SAM between the i-th and j-th spectra, s_i and s_j, is defined as follows:

$$\text{SAM}(s_i, s_j) = \arccos\left(\frac{\langle s_i, s_j \rangle}{\|s_i\|\|s_j\|}\right). \tag{S2}$$

Note that spectral information divergence (SID) is also a metric for spectral similarity, but it was not utilized in this study due to the presence of zero values in the calculated C-K edge spectra, which causes SID to be undefined. Spectral discriminatory entropy (SDE) serves as the metric to evaluate the discriminatory power of metric m for the target spectrum t within the database Δ. It is defined as follows:

$$\begin{aligned} H^m(t;\Delta) &= -\sum_{j=1}^{J} p^m_{t,\Delta}(j) \log_2 p^m_{t,\Delta}(j), \\ p^m_{t,\Delta}(i) &= \frac{m(t, s_i)}{\sum_{j=1}^{J} m(t, s_j)}, \end{aligned} \tag{S3}$$

where p_i is the spectral descriminatroy probability of the i-th bin of the spectrum. Utilizing SDE enables us to evaluate the discriminatory power of both MSE and SAM for ten randomly selected target spectra within the entire C-K edge database, denoted as Δ.

Poisson noise factor dependence of SDE for the ten spectra is shown in Figs. S1d and e for MSE and SAM, respectively, within the same plot range. For both metrics, there is observable variation among the ten spectra, yet the general trend remains consistent: as the noise decreases, indicating an increase in spectral discernibility, the SDE of the spectra declines. There is a clear difference in the SDE between the two metrics, with MSE exhibiting a more significant decrease in SDE than SAM and a smaller distribution range. It should be noted that SAM is a metric that is irrespective of the scale of the spectra, whereas MSE is a metric that is sensitive to the scale of the spectra. From this point of view, SAM might be more practical than MSE for spectral shape comparisons involving spectra of unknown scale. However, for training spectral prediction models, MSE proves appropriate as the model's output scale is confined by the training data, preserving the relative scale between spectra.

ABLATION EXPERIMENTS

To assess the contribution of individual components within our model, we conducted ablation experiments by selectively modifying components of ISD-PaiNN and comparing the resulting prediction accuracies. Four distinct models were trained for the ablation experiments:

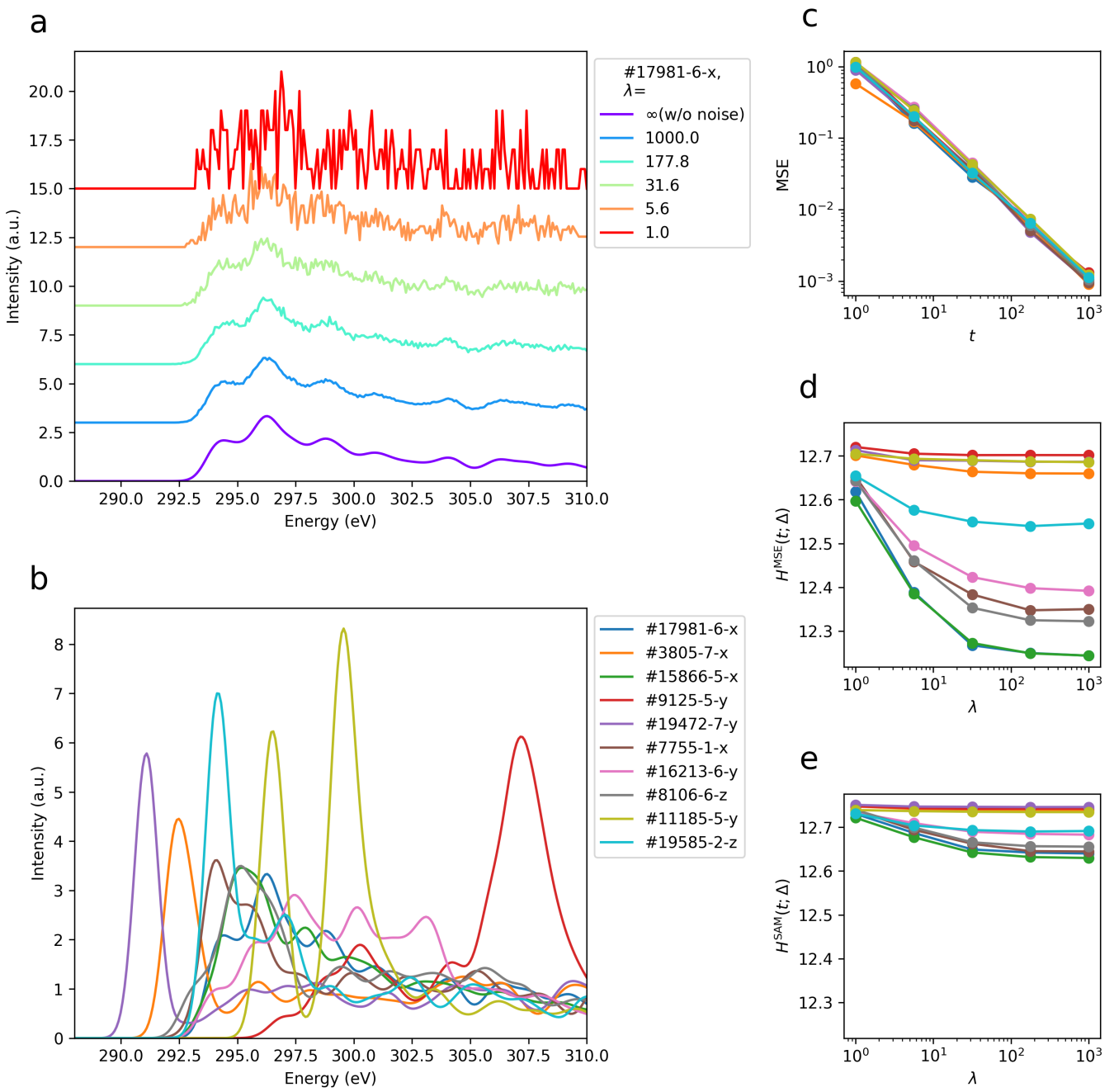

FIG. S1. (a) Spectra with no noise and Poisson noise with different Poisson noise factor λ. (b) The ten randomly selected spectra from the dataset. (c) Relationship between MSE and λ. (d) and (e) spectra discriminatory entropy of the predicted spectra λ with respect to λ for MSE and SAM, respectively. The colors of the lines in (c), (d), and (e) correspond to the line colors of the spectra in (b).

1. **ISD-PaiNN**
   The same model discussed and evaluated in the main text.

2. **Without Symmetric Message Layer (w/o SM)**
   This model omits the inversion symmetry and adopts the same message block as the original PaiNN.

3. **Without Directional Embedding (w/o DE)**
   In this model, node vector features are initialized with zero vectors, eliminating the directional embedding

component present in the original PaiNN.

4. **Without Symmetric Message Passing and Directional Embedding (w/o DE-SM)**
This model combines the modifications from both (2) and (3), removing both the symmetric message passing block and utilizing zero vector initialization for node vector features. This model is most similar to the original PaiNN model in terms of the message passing block and node vector feature initialization. The difference between this model and the original PaiNN model is the output block for predicting site-specific spectra.

The conversion from node-specific features to site-specific features is performed by the same output block as in ISD-PaiNN. Note that the original PaiNN model itself was not included in the ablation experiments, as its output is designed to predict energy or forces and is not tailored for predicting site-specific anisotropic spectra. These ablation experiments were performed under random splitting conditions, utilizing the same training and test data as presented in the main text.

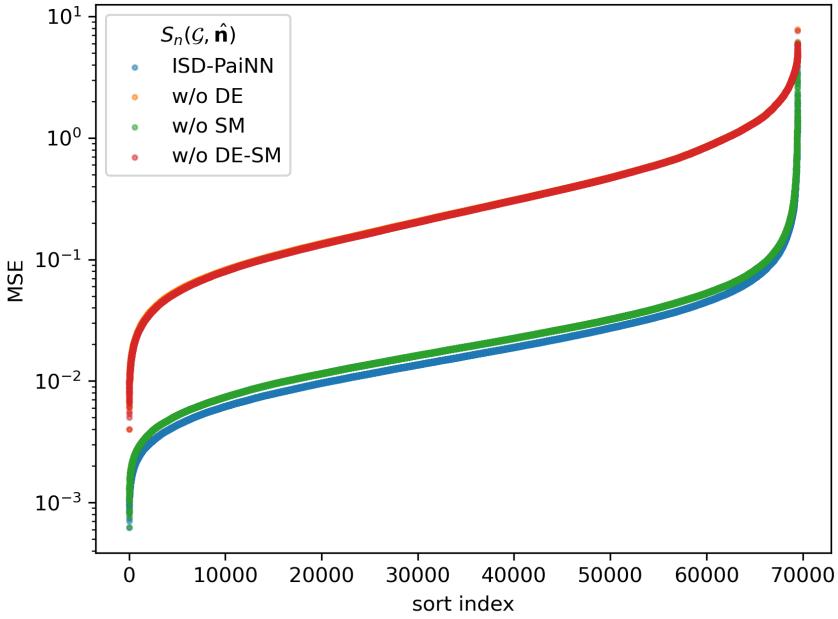

FIG. S2. Sorted MSE of the site anisotropic C K-edge spectra for the prediction by the four models on the test dataset for the random splitting. The data points for w/o DE and w/o DE-SM are almost overlapped.

The results, as compared by the sorted MSEs for the test spectra, are shown in Fig. S2. The results show that ISD-PaiNN exhibits the best prediction accuracy, followed by w/o SM. The other two models lacking directional embedding, w/o DE and w/o DE-SM, show comparatively lower prediction accuracy, indicating that directional embedding is essential for predicting anisotropic spectra. The results also reveal that symmetric message passing contributes to the prediction of anisotropic spectra, although the improvement in prediction accuracy is not as significant as with directional embedding.

EVALUATION OF PREDICTION ON MOLECULAR SPECTRA

The molecular spectra can be predicted by summing all site-specific sppectra in the molecule:

$$S_{\mathrm{mol}}(\mathcal{G}, \hat{\mathbf{n}}) = \sum_n S_n(\mathcal{G}, \hat{\mathbf{n}}). \tag{S4}$$

The prediction results of molecular spectra in the test data, using the model trained on the site-specific spectra under the random split, are illustrated in Fig. S3. Although the test data contains a total of 4,334 molecules, we display

results for 3,857 molecules where carbon sites are all symmetrically non-equivalent. This selection was made because the database only includes site spectra for one site among the symmetrically equivalent sites in the molecule, and directional information for the other equivalent sites is unavailable. The results for the typical percentiles of 0, 25, 50, 75, 100% of molecules id #273, #10228, #10578, #16771, and #23880 are shown in Figs. S4, S5, 2 in the main text, S6, and S7, respectively. Notably, molecule #23880, which exhibited the worst predicted MSE, corresponds to the molecule containing the trifluorocarbon site, for which the site-specific spectra also demonstrated the lowest prediction accuracy, as shown in Fig. 1 in the main text. It is plausible that the peak shift associated with the strong electronegativity of fluorine was not adequately learned due to the relatively small representation of molecules containing fluorine within the training dataset.

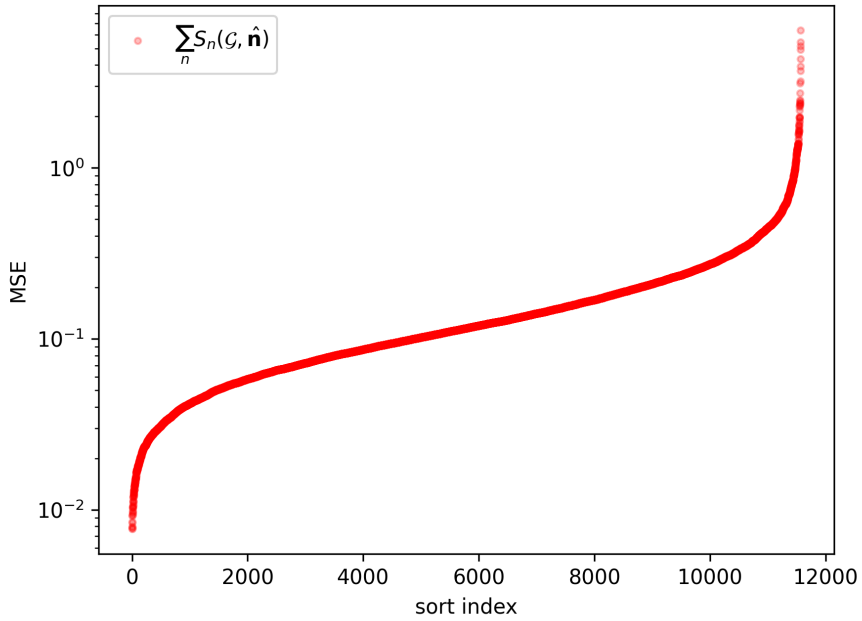

FIG. S3. Sorted MSE of the molecular anisotropic C K-edge spectra of the prediction on the test dataset.

SCAFFOLD SPLIT EVALUATION

To further validate the model's predictive performance beyond random splitting, we proceeded to evaluate its generalization using scaffold splitting[1], a technique that partitions datasets into training and testing sets while preserving structural diversity among molecules. Employing the ScaffoldSplitter class from the deepchem[2] module, the dataset was divided into train, validation, and test sets at a ratio of 6:2:2 based on molecular count. Consequently, the spectra were distributed across the train, validation, and test sets, resulting in 209,004, 69,669, and 67,731 spectra, respectively.

Figure S8 shows the prediction results for the test data using scaffold splitting. As shown in Fig. S8a, there is an overall degradation in predictive accuracy, as indicated by the increased MSE compared to that of random split shown in Fig. 1 in the main text. Figures S8b-e illustrate spectra at the 0, 50, 75, and 100 percentiles, representing typical examples. While an overall decrease in prediction accuracy is observed, there is a discernible capturing of general trends, up to the percentile at 75%. The decrease in prediction performance in scaffold split compared to the random split indicates that including a diverse range of structures is effective in training the predictive model. This underscores the importance of covering a wide range of molecular structures for effective model training, emphasizing the challenges posed by limited structural diversity in achieving robust predictive capabilities.

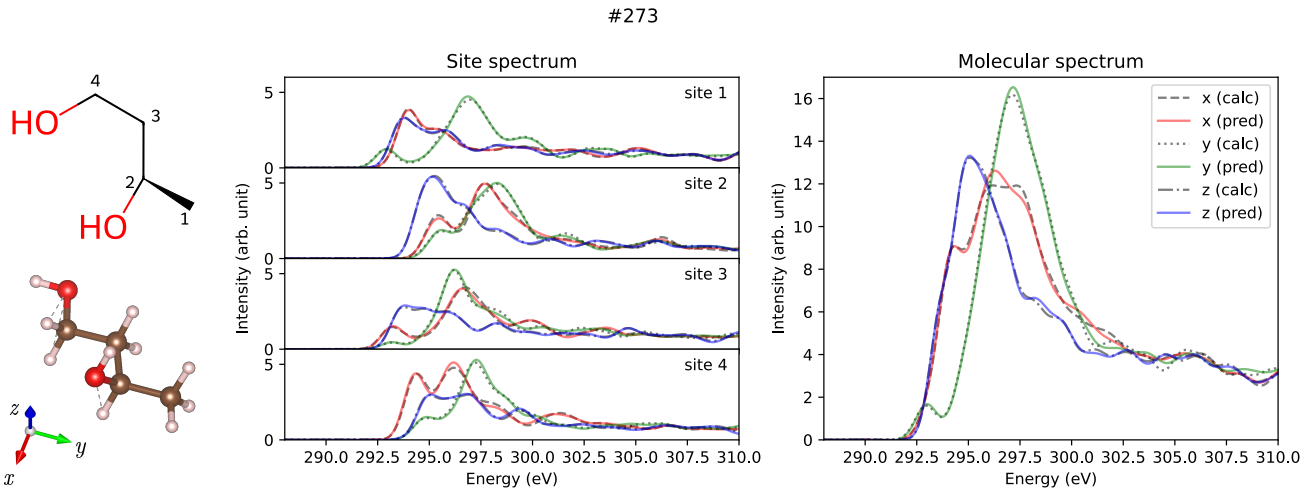

FIG. S4. Prediction results of site-specific and molecular anisotropic C K-edge spectra for molecude id #273 in QM9, which is located at 0% percentile (best accuracy). Left panel shows the molecular structure formula and the three-dimensional structure. Middle panel and right panel show the predicted (solid lines) and calculated (dotted lines) spectra for site-specific spectra for each C site and molecular spectra, respectively.

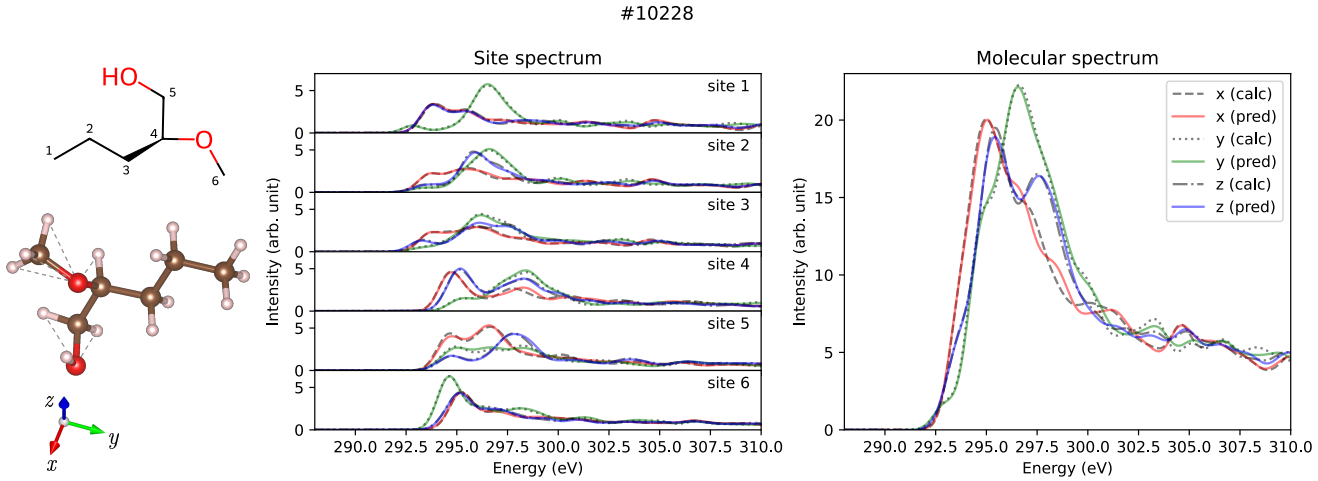

FIG. S5. Prediction results of site-specific and molecular anisotropic C K-edge spectra for molecude id #10228 in QM9, which is located at 25% percentile. Left panel shows the molecular structure formula and the three-dimensional structure. Middle panel and right panel show the predicted (solid lines) and calculated (dotted lines) spectra for site-specific spectra for each C site and molecular spectra, respectively.

TESTING ON LARGER MOLECULES

As mentioned in Sec. 2.3, the referenced C K-edge spectral dataset is composed of relatively small molecules, containing no more than eight non-hydrogen atoms. To assess our model's generalization capabilities beyond the training data domain concerning molecular size, we included four aromatic amino acids: Phenylalanine, Tyrosine, Tryptophan, and Histidine.

Figure S9 illustrates the predicted and calculated spectra for these four aromatic amino acids by the model trained on the random split described in Sec. 3. While the predicted spectra show minor differences in peak positions and intensities, they broadly capture the overall trends, including directional dependencies. This suggests the model's potential for generalizing to larger molecules beyond the training dataset.

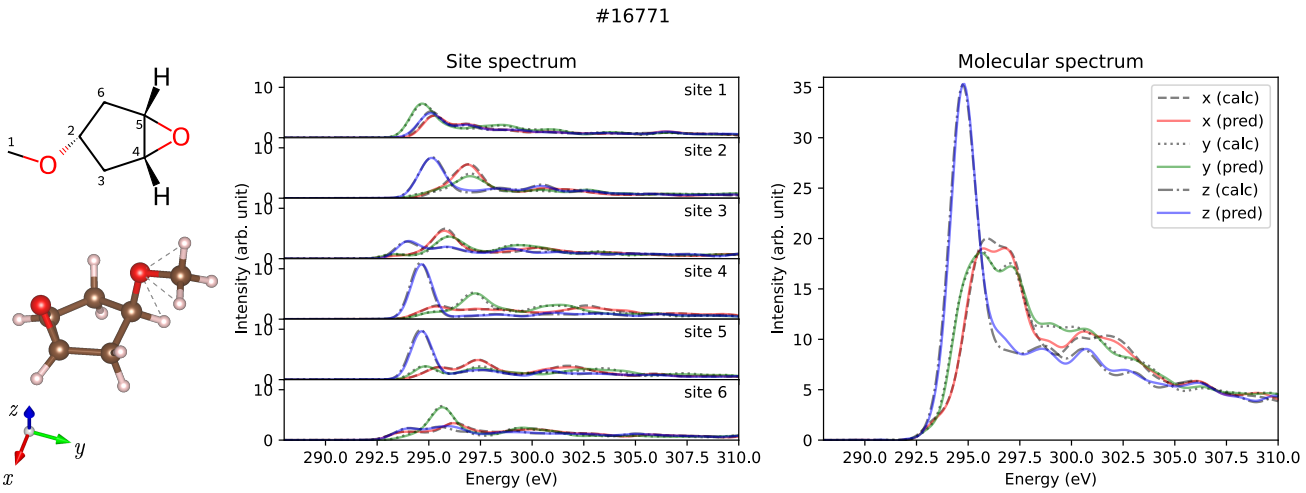

FIG. S6. Prediction results of site-specific and molecular anisotropic C K-edge spectra for molecude id #16771 in QM9, which is located at 75% percentile. Left panel shows the molecular structure formula and the three-dimensional structure. Middle panel and right panel show the predicted (solid lines) and calculated (dotted lines) spectra for site-specific spectra for each C site and molecular spectra, respectively.

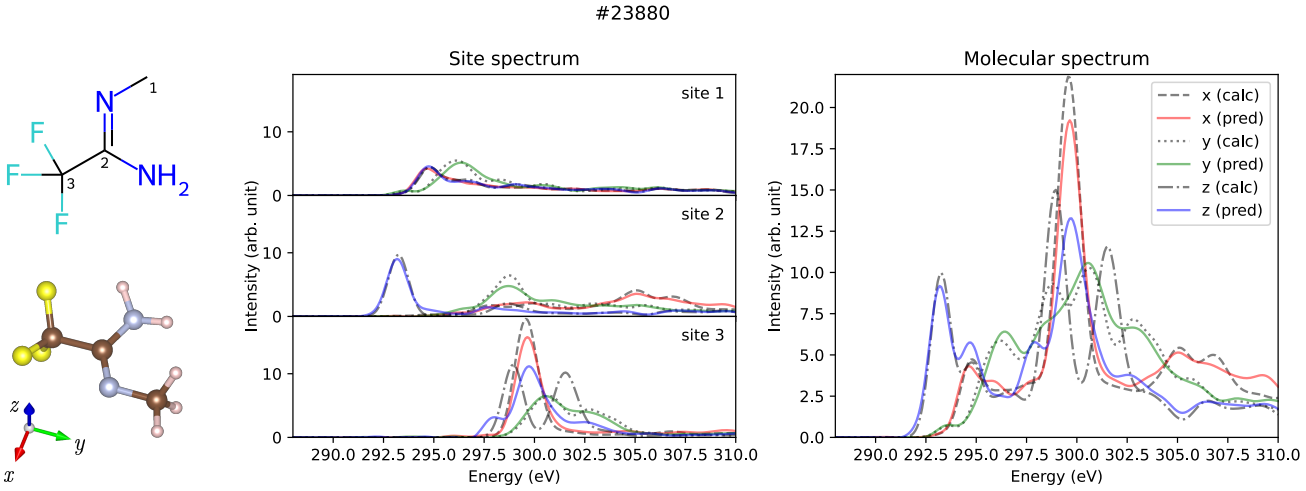

FIG. S7. Prediction results of site-specific and molecular anisotropic C K-edge spectra for molecude id #23880 in QM9, which is located at 100% percentile (worst accuracy). Left panel shows the molecular structure formula and the three-dimensional structure. Middle panel and right panel show the predicted (solid lines) and calculated (dotted lines) spectra for site-specific spectra for each C site and molecular spectra, respectively.

PREDICTION PERFORMANCE ON ORIENTATION DEPENDENCE

As previously examined in the validation of three different data splits, our focus was on assessing the predictive performance concerning various molecular structures, aiming to evaluate the model's robustness concerning structural graphs, denoted as $\mathcal{G}$. However, our model's dependence extends not only to the structural graph $\mathcal{G}$ but also to the directional vectors $\hat{\mathbf{n}}$ associated with dipole transition moments.

To investigate the robustness of our model concerning directional vectors $\hat{\mathbf{n}}$, we conducted an additional evaluation that explored the angle dependency within a specific molecule. Specifically, we performed computations using DFT, rotating a benzene molecule about the x-axis from 0 to 90 degrees. Figure S10 illustrates the calculated and predicted results for the x, y, and z-directional dipole transition moments, as well as the directional vector $\hat{\mathbf{n}}$, for the rotated benzene molecules. For the prediction, we used the same model trained on the C K-edge dataset by the random

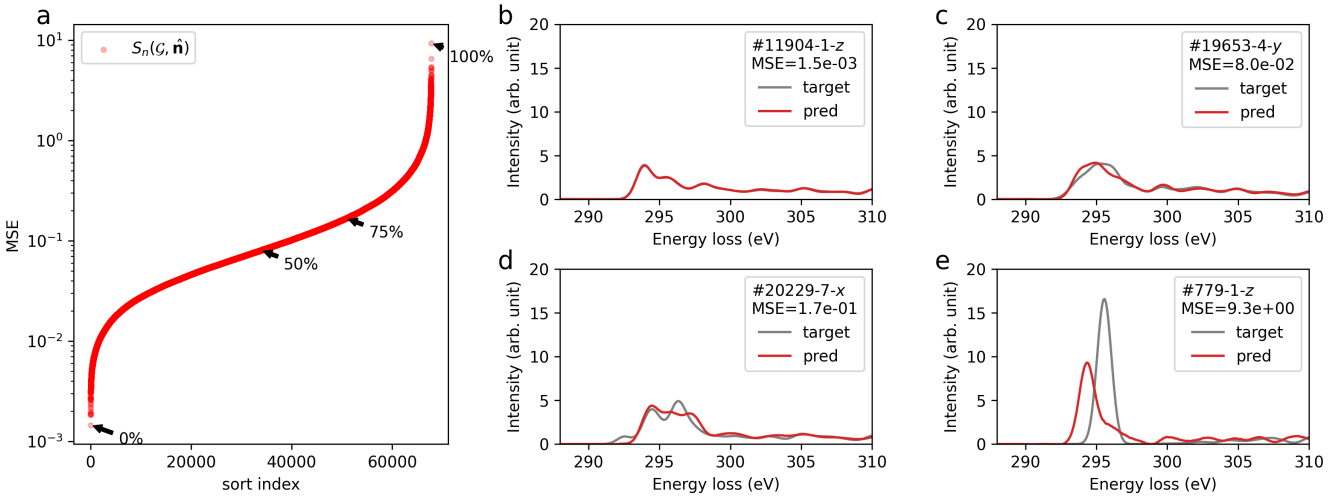

FIG. S8. Prediction results of site-specific anisotropic C K-edge spectra $S_n(\mathcal{G}, \hat{\mathbf{n}})$ for the test dataset split by scaffold splitting. (a) Sorted MSE of the prediction on the test dataset. (b-e) Predicted (red) and calculated (gray) spectra for typical percentiles in terms of MSE loss (0, 50, 75, and 100%) as denoted in (a). The inset in (b-e) shows the molecule id in QM9 corresponding $\mathcal{G}$, site index n, directional vector $\hat{\mathbf{n}}$, and the MSE value.

splitting evaluated in Sec. 3.

The spectra for the x-direction shown in Fig. S10 are theoretically expected to yield identical spectra as the rotation axis is parallel to the dipole transition moment. However, slight variations in the spectra concerning the rotation angles were observed in the first-principles calculations. This discrepancy likely stems from the finite cell size, periodic boundary conditions, and computational inaccuracies. However, our model accurately predicted identical spectra, adhering to the principle of rotational symmetry. This outcome highlights one of the advantages of our model, which effectively captures symmetry concerning both the graph and directional vector.

The results for the y- and z-directions in Fig. S10 reveal that the model excels in capturing the angle-dependent trends of both the approximate peak heights and energy positions. Furthermore, as physically predicted, results of $\hat{n} \parallel y$ for θ degree rotation are consistent with results of $\hat{n} \parallel z$ for $90 - \theta$ degrees. It should be emphasized that the training dataset only includes the three directional components (x, y, z) for each molecule, lacking densely sampled spectral data concerning orientation angular space, *i.e.* $\hat{\mathbf{n}}$-dependence for both benzene and other molecules. Despite this limitation, the ability to predict dense angular dependencies is a highly intriguing outcome that could be considered a successful extrapolation regarding directional predictions. The results indicate that the model has notable spectral feature angle-dependent prediction performance.

ADDITIONAL CALCULATION FOR EVALUATION OF PREDICTION PERFORMANCE

Additionally, we generated spectral data for evaluation purposes by computing C K-edge spectra for a rotation series of a benzene molecule and four aromatic amino acid molecules: phenylalanine ($C_9H_{11}NO_2$), tyrosine ($C_9H_{11}NO_3$), tryptophan ($C_{11}H_{12}N_2O_2$), and histidine ($C_6H_9N_3O_2$). The molecule structure of benzene for the rotation series was extracted from the C K-edge dataset[3] with molecule id #214, and the rotation series was generated by rotating the molecule around the z-axis with 18 degrees intervals. The molecular structures of aromatic amino acids were obtained directly from PubChem[4] with Compound Identifiers (CIDs) 6140, 6057, 6305, and 6274, respectively utilized for both prediction and first-principles calculations. The calculation was done based on DFT[5, 6] within the plane-wave basis pseudopotential method[7] implemented in the CASTEP code[8] under the same the calculation condition as in the C-K edge dataset[3] except that we used a $20 \times 20 \times 20 \text{Å}^3$ cubic cell with 2,000 extra bands for aromatic amino acids.

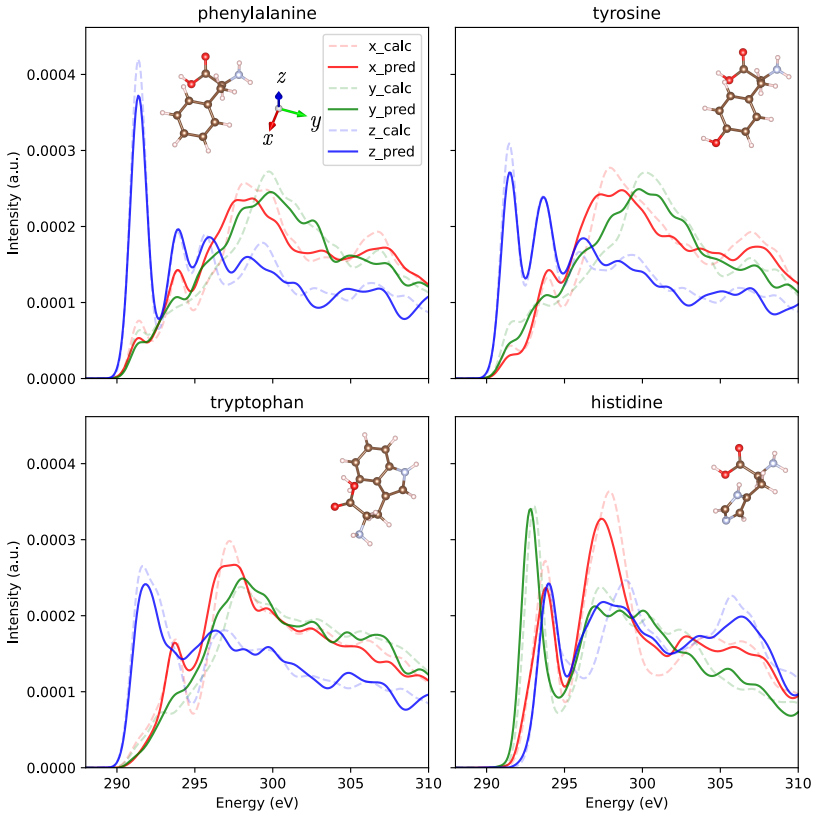

FIG. S9. Prediction and calculation results of the C K-edge molecular spectra for the four aromatic amino acids. The predicted spectra are shown in solid lines, and the calculated spectra are shown in dotted lines.

TIME COMPARISON WITH FIRST-PRINCIPLES CALCULATIONS

To assess the efficiency and speed of our model, we compared the time taken to acquire x, y, and z-directional spectra for each of the all 21,666 molecules included in the carbon K-edge dataset mentioned in Sec. 2.3.

For the prediction, we used the same model trained on the C K-edge dataset by the random splitting described in Sec. 3. We present a scatter plot in Fig. S11a illustrating the comparison between the time required for first-principles calculations (t_{DFT}) and the inference time for our model's predictions (t_{GNN}) for each molecule. The distribution range of t_{DFT} exhibits a relatively broad span, ranging from 1×10^3 to 1×10^4 seconds. In contrast, the range of t_{GNN} for predictions made after a one-time typical training with typical duration of 3 hours ($\sim 10^4$ seconds) is observed between 3.5×10^{-3} to 3.8×10^{-3} seconds, showcasing a speed enhancement of approximately 10^6 times. The marker color in Fig. S11 represents the number of non-equivalent excited sites (N_{ex}) for each molecule. A noticeable N_{ex}-dependent trend is evident along the t_{DFT} axis, while t_{GNN} displays negligible dependence. This difference arises from the need for individual spectral computations for each non-equivalent excited site within a molecule in first-principles calculations, resulting in time proportional to N_{ex}. Conversely, our model's prediction encompasses all excited site spectra with a single input, thereby eliminating Nex dependence. Furthermore, Figs. S11b and c displaying dependence of t_{DFT} and t_{GNN} dependence on number of atoms in each molecule (N) indicates an increase in t_{DFT} with a rise in the number of atoms, whereas t_{GNN} remains largely unaffected. Typically, first-principle calculations of the core electron excitation spectra of larger molecules in a given energy range require the inclusion of more unoccupied bands, which can significantly increase the computation time. These outcomes emphasize that our model's ability to swiftly predict spectra without dependency on molecular size or the number of excited sites, indicating its utility for rapid generation of reference spectra.

Regarding detailed rotation dependence, in some first-principles computation codes, only the three directional components (x, y, z) are outputted, and to acquire detailed angular dependencies, separate calculations for systems of the structures corresponding to each rotation angle are required. In contrast, as demonstrated in the preceding section, our model can provide dense angular dependencies in spectral outputs beyond the x, y, z directions. This capabil-

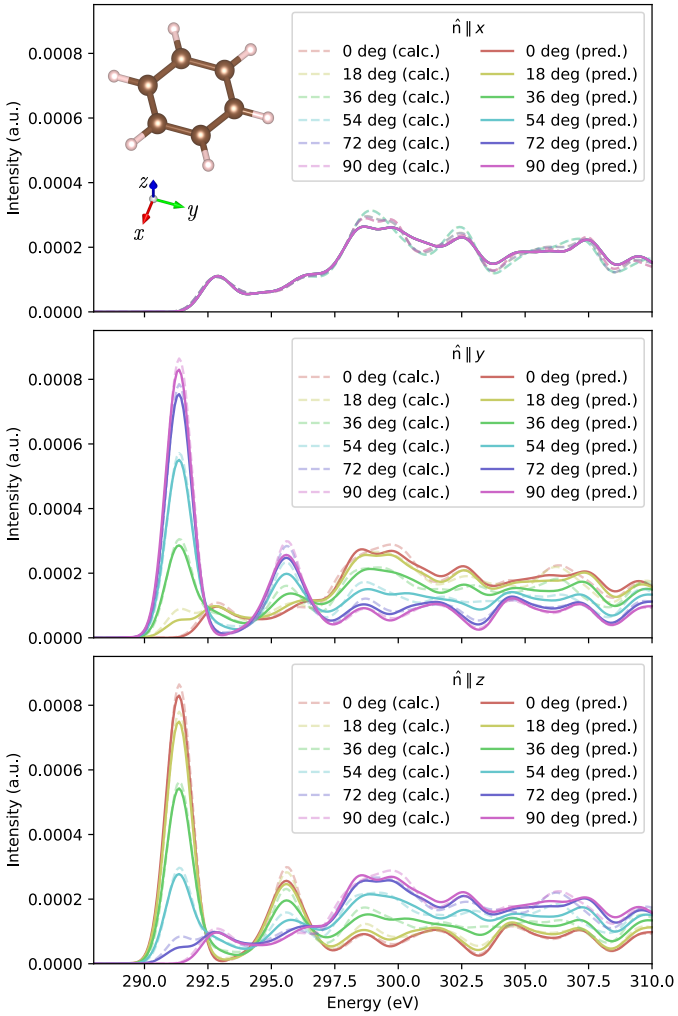

FIG. S10. Prediction results of C K-edge molecular spectra of a benzene molecule for rotation about x axis. The predicted spectra for each orientation of the dipole vector $\hat{\mathbf{n}}$ are shown in solid lines, and the calculated spectra are shown in dotted lines. The color of the lines corresponds to the rotation angle of the benzene molecule about x-axis, specifically, red to purple corresponds to 0 to 90 degrees.

ity enables the efficient exploration of detailed angular dependencies without the need for additional computations involving structural rotations.

COMPUTATIONAL CONDITIONS AND TIME MEASUREMENT

The calculation time for each molecule by first-principles calculation, t_{DFT} in Fig. 6 in the main text, was analyzed on the computations conducted to construct the C-K edge database[3]. The calculation procedure and conditions of the C-K edge spectra and excitation energies are described in the reference[3]. The calculations were performed using Intel(R) Xeon(R) Silver 4114 or Intel(R) Xeon(R) Gold 6130 processors. The time for each site in the molecules was calculated by summing up the computation time from the .castep files of the three calculation steps: ground state, excited state, and transition probability.

The calculation time for predicting the molecular spectra, t_{GNN} in Fig. 6 in the main text, was measured using a single NVIDIA GeForce RTX 4090 GPU.

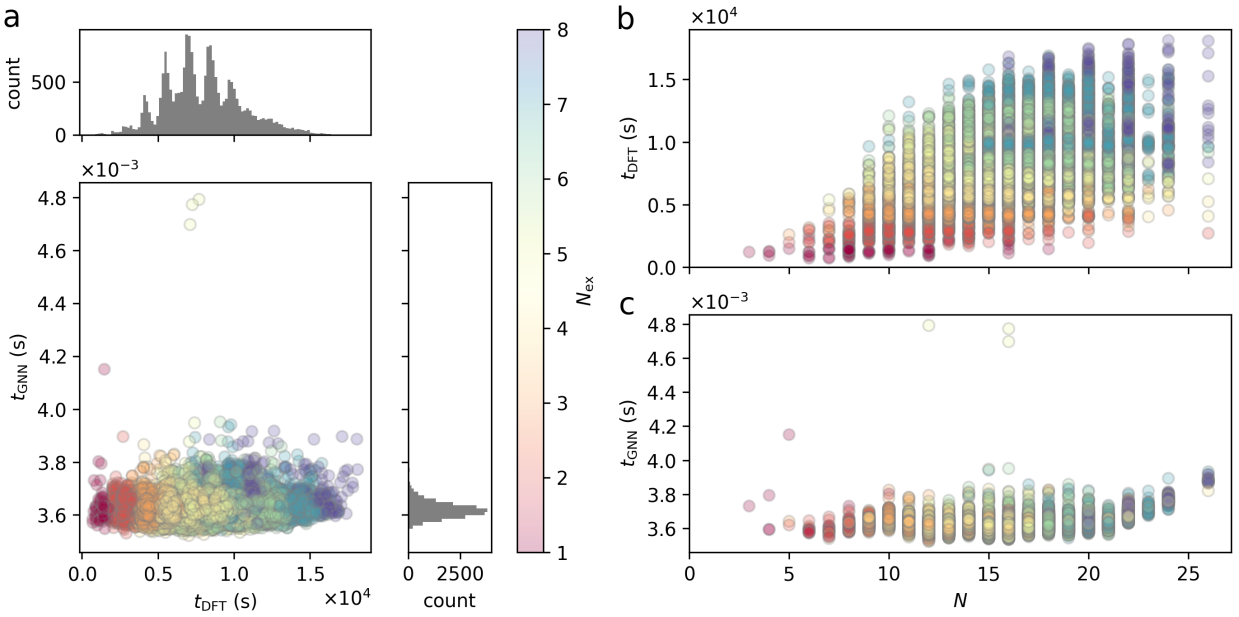

FIG. S11. Comparison between computational time of first-principles calculations (t_{DFT}) and that of our model (t_{GNN}). (a) Scatter plot between t_{DFT} and t_{GNN}, colored by number of unique excitation sites N_{ex}. (b, c) number of atom (N) dependence of t_{DFT} to t_{GNN}, colored by N_{ex} in the same scale as in (a).

* kiyou@iis.u-tokyo.ac.jp

[1] G. W. Bemis and M. A. Murcko, *Journal of Medicinal Chemistry*, 1996, **39**, 2887–2893.
[2] B. Ramsundar, P. Eastman, P. Walters, V. Pande, K. Leswing and Z. Wu, *Deep Learning for the Life Sciences*, O'Reilly Media, 2019.
[3] K. Shibata, K. Kikumasa, S. Kiyohara and T. Mizoguchi, *Scientific Data*, 2022, **9**, 214.
[4] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang and E. E. Bolton, *Nucleic Acids Research*, 2022, **51**, D1373–D1380.
[5] P. Hohenberg and W. Kohn, *Phys. Rev.*, 1964, **136**, B864–B871.
[6] W. Kohn and L. J. Sham, *Phys. Rev.*, 1965, **140**, A1133–A1138.
[7] M. C. Payne, M. P. Teter, D. C. Allan, T. Arias and J. D. Joannopoulos, *Rev. Mod. Phys.*, 1992, **64**, 1045–1097.
[8] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. J. Probert, K. Refson and M. Payne, *Z. Kristall.*, 2005, **220**, 567–570.