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

K. Shibata, T. Mizoguchi
Institute of Industrial Science
The University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan
{kiyou, teru}@iis.u-tokyo.ac.jp

Abstract

Absorption near edge structures in the core electron excitation spectra reflect the anisotropy of orbitals in the transition final state and can be used for analyzing local atomic environment including its orientation. So far, the analysis of fine structures is mainly based on a fingerprint-matching with high-cost experimental or simulated spectra. If core electron excitation spectra, including its anisotropy, can be predicted at low cost using machine learning, the application range of the core electron excitation spectra will be accelerated and extended for such as orientation and electronic structure analysis of liquid crystals and organic solar cells at high spatial resolution. In this study, we introduce a message-passing neural network for predicting core electron excitation spectra using a unit direction vector in addition to molecular graphs as input. Utilizing a database of calculated C K-edge spectra, we have confirmed that the network can predict core electron excitation spectra reflecting the anisotropy of molecules. Our model is expected to be expanded to other physical quantities in general that depend not only on molecular graphs but also on anisotropic vectors.

1 Introduction

Fine structures of core electron excitation spectra near the excitation edges, electron-energy loss near edge structure (ELNES) and X-ray absorption near edge structure (XANES or NEXAFS), have been widely used as one of the most effective fingerprints for determination of local atomic structures and electronic states. The fine structures reflect the excitation of an electron from an occupied core orbital to unoccupied orbitals. Since the fine structures differ depending on the symmetry of the orbital against the momentum transfer during the core electron excitation, orientation-dependence of the spectra can be used to analyze the local atomic environments including anisotropic nature of orbitals. This kind of analysis has been applied for analyzing orientation of molecules such as polymer films[1], liquid crystals[2] and organic solar cells[3, 4] through analysis on orientation dependence of characteristic peaks corresponding anisotropic orbitals. Meanwhile, quantitative analysis on fine structures has conventionally been done by fingerprint matching of reference spectra obtained from standard samples with known structures, but it is difficult to apply to unknown structures.

Recently, comparison with spectral shapes obtained by first-principles calculations has been used to analyze the experimental spectra[5, 6, 7]. As reference spectra can be obtained faster and on a larger scale with calculations than with experiments, databases of simulated core electron excitation spectra have been constructed and published[8, 9, 10, 11]. While experimental databases typically contain a few hundred spectra, simulated ones can include over 10,000 spectra. However, first-principles calculations are computationally expensive, and it is still difficult to obtain spectra exhaustively for a large number of candidate structures to identify unknown structure from experimental data without a

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

priori structural information. It would be useful to develop a low-computational-cost alternative to first-principles calculations that can obtain spectra exhaustively for a large number of structures.

In this context, attempts have been made to predict spectra from molecular structures using machine learning models, and there have been several reports: prediction from molecular graphs using a message passing graph neural network (GNN)[12], a deep neural network[13, 14, 15], and prediction using neural network ensemble through featurized local structural information[16]. However, as far as the authors know, the prediction of core electron excitation spectra considering anisotropy according to the momentum transfer has not been proposed.

From the perspective of predicting physical quantities in general, many machine learning models for material science, not limited to the spectral example above has been proposed. Previous research in GNN architecture includes notable models such as DimeNet[17] and GemNet[18] that leverage invariance, both emphasizing the utilization of invariance properties. Exploring equivariance, PaiNN[19] and NequIP[20] have been proposed, employing strict equivariance constraints. Moreover, attempts to enhance expressive capabilities by slightly relaxing equivariance constraints and incorporating nonlinear transformations have been done in SCN[21] and eSCN[22]. However, most models predict physical quantities that depend solely on structure, overlooking the explicit integration of material anisotropy and directional dependencies. It is crucial to consider such directional dependencies when dealing with anisotropic physical quantities, including strain tensors, resistivity tensors, dielectric polarization, and phenomena like dipole transitions. These physical quantities and phenomena reflect the anisotropic nature of materials, and neglecting them in model construction could limit accurate predictions of material properties.

In this study, we developed a message passing neural network[23] model that can predict spectra due to the dipole transition. Our model can predict site-specific spectra which is dependent on not only the molecular graph but also the dipole moment direction relative to the molecular graph. Notably, our model consists of transformations that satisfy the physically required invariance of symmetric operations on the input pair of molecular graph and dipole moment with respect to dipole transitions, ensuring an inductive bias with respect to the invariance.

2 Methods

2.1 Spatial symmetry constraints of dipole-transition spectra

In the limit of small angle electron scattering or long wavelength of incident X-ray in ELNES and XANES, respectively, the core electron excitation process can be described in the form of dipole transition. The transition matrix element for ELNES and XANES from an occupied initial state $|\psi_i\rangle$ to an unoccupied final state $|\psi_f\rangle$ is proportional to $\langle\psi_i|\mathbf{q}\cdot\mathbf{r}|\psi_f\rangle$ and $\langle\psi_i|\hat{\epsilon}\cdot\mathbf{r}|\psi_f\rangle$, respectively, where $\mathbf{r}$ is the relative position of the excited electron from the core, $\mathbf{q}$ is the momentum transfer, $\hat{\epsilon}$ is the polarization vector. In either case, the generalized oscillator strength is proportional to the square of the transition matrix element and the differential cross section of energy loss spectrum can be expressed as a function of energy loss E as:

$$\frac{\partial\sigma(E)}{\partial E} \propto \sum_{i,f} \|\hat{\mathbf{n}}\cdot\langle\psi_i|\mathbf{r}|\psi_f\rangle\|^2\delta(E-(E_f-E_i)), \tag{1}$$

where $\hat{\mathbf{n}}$ is the unit directional vector parallel to the momentum transfer $\mathbf{q}$ or polarization vector $\hat{\epsilon}$, and E_i and E_f are the energy eigenvalues of $|\psi_i\rangle$ and $|\psi_f\rangle$, respectively.

Based on Eq. 1, we consider the constraints imposed on spectra associated with dipole transitions. Let $\mathcal{G}$ the structural (molecular or crystal) graph defined by atomic numbers Z_i and positions $\vec{\mathbf{r}}_i$, $\hat{\mathbf{n}}$ the dipole vector, n site index as input, and the target site-specific spectrum of n-th site in $\mathcal{G}$ for transition regarding $\hat{\mathbf{n}}$, $S_n(\mathcal{G},\hat{\mathbf{n}})$. The invariance of $S_n(\mathcal{G},\hat{\mathbf{n}})$ against symmetry operations are as follows:

$$S_n(\mathcal{G},\hat{\mathbf{n}}) = S_n(\mathcal{G},i(\hat{\mathbf{n}})) = S_n(i(\mathcal{G}),\hat{\mathbf{n}}) = S_n(R(\mathcal{G}),R(\hat{\mathbf{n}})) = S_n(T(\mathcal{G}),\hat{\mathbf{n}}), \tag{2}$$

where i is the spatial inversion, $R\in\mathrm{SO}(3)$ a general 3D rotation, $T\in T(3)$ the displacement.

2.2 Model architecture

To satisfy the constraints in Eq. 2, we employ the message passing neural network (MPNN)[23] and referred to its variant, polarizable atom interaction neural network (PaiNN)[19].

2.2.1 PaiNN and its limitations about input

The PaiNN[19] model, classified among message passing neural networks[23], employs rotationally equivariant representations. It is designed by modeling equivariant interactions in Cartesian space, avoiding equivariant convolutions relying on spherical harmonics and Clebsch-Gordon transforms, which contributes to its conceptual simplicity and computational efficiency. The model takes atomic numbers $Z_i \in \mathbb{N}$ and atomic positions $\vec{\mathbf{r}}_i \in \mathbb{R}^3$ as inputs. Internally, it employs two node-wise features at each i-th site: an SO(3) equivariant 3D vector $\vec{\mathbf{v}}_i$ and an SO(3) invariant scalar $\mathbf{s}_i$. The features $\vec{\mathbf{v}}_i$ and $\mathbf{s}_i$ at i-th site are initialized with zero vector $\vec{\mathbf{v}}_i^0 = \vec{\mathbf{0}}$ and embedded by atomic number Z_i, respectively. Subsequently, they are updated by message constructed with relative positional vectors between site i and j, $\vec{\mathbf{r}}_{ij} = \vec{\mathbf{r}}_i - \vec{\mathbf{r}}_j$. The transformation of features $\vec{\mathbf{v}}_i$ and $\mathbf{s}_i$ while preserving their equivariance and invariance, respectively, is achieved by properly constructed message and update functions.

The design of PaiNN implies a dependency of its model output solely on the input structural graph $\mathcal{G}$. Consequently, PaiNN cannot incorporate additional information beyond $\mathcal{G}$, such as directional information respective to $\mathcal{G}$.

2.2.2 Element specific non-zero initialization of node-wise vector feature

Expanding on the above, we introduce directional details by setting node-wise vectors with non-zero values. These initial vectors need only adhere to the symmetry requirements concerning their orientation, allowing flexibility in their magnitude. Consequently, we initialized vector $\vec{\mathbf{v}}_i$ using a scaled vector constructed using the embedded representation specific to each elemental species:

$$\vec{\mathbf{v}}_i^0 = \mathbf{b}_{Z_i} \otimes \hat{\mathbf{n}} \in \mathbb{R}^{F\times 3}, \tag{3}$$

where $\mathbf{b}_{Z_i} \in \mathbb{R}^{F\times 1}$ is the embedded representation for the Z_i-th element, and F is the dimension of the embedded representation.

2.2.3 Message and update blocks

The invariance of the features under T on $\mathcal{G}$ is satisfied by using translational invariant $\vec{\mathbf{r}}_{ij}$ for the message block. The invariance and equivariance of features under R on both $\mathcal{G}$ and $\hat{\mathbf{n}}$ is also satisfied by the message function in PaiNN. However, the message function for $\vec{\mathbf{v}}_i$ in PaiNN is asymmetric under the inversion operation i on either $\mathcal{G}$ or $\hat{\mathbf{n}}$, which also results in $\mathbf{s}_i$ to be asymmetric under i due to the mixing of $\vec{\mathbf{v}}_i$ and $\mathbf{s}_i$ in the message and update functions. To satisfy the constraints concerning invariance against space inversion symmetry operation i, *i.e.* $S_n(i(\mathcal{G}), \hat{\mathbf{n}}) = S_n(\mathcal{G}, i(\hat{\mathbf{n}})) = S_n(\mathcal{G}, \hat{\mathbf{n}})$ intrinsically, the message function for $\vec{\mathbf{v}}_i$, $\Delta\vec{\mathbf{v}}_i^m$ is modified slightly from one in PaiNN as follows:

$$\Delta\vec{\mathbf{v}}_i^m = \sum_j \vec{\mathbf{v}}_j \circ \phi_{vv}(\mathbf{s}_j) \circ \mathcal{W}_{vv}(\|\vec{r}_{ij}\|) + \sum_j (\vec{\mathbf{v}}_j \cdot \vec{r}_{ij}) \circ \phi_{vs}(\mathbf{s}_j) \circ \mathcal{W}'_{vs}(\|\vec{r}_{ij}\|) \frac{\vec{r}_{ij}}{\|\vec{r_{ij}}\|}, \tag{4}$$

where ϕ_{vv} and ϕ_{vs} are the activation functions, W_{vv} and $\mathcal{W}'_{vs}$ are the weight functions as following Eq. (8) in the reference[19]. The modification is only on the second term and is just multiplying the inner product $\vec{\mathbf{v}}_j \cdot \vec{r}_{ij}$ which is odd against $\vec{\mathbf{v}}_i$ and $\vec{r}_{ij}$ so that the overall message function is odd and even against the inversion operation i on $\vec{\mathbf{v}}_i$ and $\vec{r}_{ij}$, respectively.

The other message function for $\mathbf{s}_i$ and the update functions are the same as PaiNN. This modification makes $\mathbf{s}_i$ to be updated with $\vec{\mathbf{v}}_i$, which reflects the input direction $\hat{\mathbf{n}}$. As a results, $\mathbf{s}_i$ posses even symmetry against i on $\mathcal{G}$ and $\hat{\mathbf{n}}$, which is the required invariance for $S_n(\mathcal{G}, \hat{\mathbf{n}})$ as described in Eq. 2.

2.2.4 Output block

The non-zero initialization of $\vec{\mathbf{v}}_i$ by Eq. 3 and the modified message function by Eq. 4 make $\mathbf{s}_i$ to satisfy the required symmetry in Eq. 2. The prediction of site-specific spectrum done by converting $\mathbf{s}_i$ at the site of interest by a multilayer perceptron consisting of two fully-connected layers.

2.3 Dataset and training for evaluation

To validate the model, a spectral dataset that includes dipole vector dependent and site-specific oscillation intensity is needed. In this study, we validate our model with a simulated C K-edge

spectral database of organic molecules[10], which is constructed by first-principles calculations based on density functional theory (DFT) and contains 117,340 site-specific spectra for three typical directions of dipole vectors of symmetrically unique sites in 22,155 molecules with no more than eight non-hydrogen atoms (C, O, N, and F) in the QM9 database[24, 25]. We preprocessed the spectra with Gaussian smearing of 0.5 eV in the energy range of 288-310 eV sampled equally spaced as 256 dimensional vectors for the objective variable. Spectral intensities were normalized by scaling the entire data set so that the averaged intensity of each spectrum in the energy range considered was 1, while preserving the relative magnitude relationship between the spectra per site.

The model was developed by modifying PaiNN implemented in the Open Catalyst Project's library opc[26] with PyTorch Geometric[27] and PyTorch[28], based on the description in Sec. 2.2.2-2.2.4. The mini-batch learning with batch size of 32 molecular graphs was performed. The loss function is the mean squared error (MSE) between the predicted and reference spectra on carbon sites with available spectra. We used the Adam[29] optimizer with a learning rate of 0.0001 for the learning process, implementing early stopping with a patience of 40 epochs.

3 Results and discussions

We evaluated the prediction performance using the processed C K-edge spectra dataset described in Section 2.3. The dataset was randomly divided: 80% for training and validation (in a ratio of 0.8:0.2) and 20% for testing based on molecules. The number of anisotropic and site-specific spectra for training, validation, and testing were 221,274, 55,653, and 69,477, respectively.

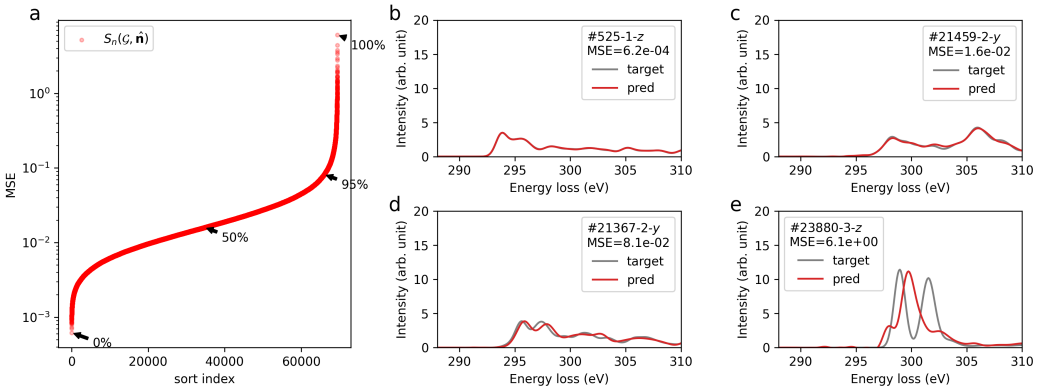

Figure 1: Prediction results of site-specific anisotropic C K-edge spectra $S_n(\mathcal{G}, \hat{\mathbf{n}})$. (a) Sorted MSE of the prediction on the test dataset. (b-e) Predicted (red) and calculated (gray) spectra for typical percentiles in terms of MSE loss (0, 50, 95, and 100%) as denoted in (a). 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.

Figure 1a shows the distribution of MSE, confirming that the over 95% data points has MSE below 0.1. Figures 1b-e show the predicted spectra (red line) and calculated results (gray line), sampled at typical percentiles at 0, 50, 95, and 100%, respectively as shown in Fig. 1a. For MSEs below 0.1 (Figs. 1b-d), the predicted spectra show good agreement with the calculated ones. Particularly, for the top 50% of MSEs, the predicted spectra closely resemble the calculated spectra. In the case of the worst-performing prediction (Fig. 1e), the general trend of peak distribution around 297eV rising and falling at 305eV is reproduced. However, accurate prediction of the major peaks is lacking. This specific spectrum corresponds to the trifluorocarbon site in molecule id #23880 (refer to Supplementary Fig. S6). The poor prediction results might stem from the limited trifluorocarbon site data, consisting of only 87 sites within the 117,340 sites in the database, and the distinct spectral features originating from chemical interactions with highly electronegative fluorine. Concerning the value of MSE and metrics for spectral similarity, we evaluated the MSE and its use by comparing it with another metric by simulating spectra with added noise (See Supplementary).

Predictions account for relative intensity enables the derivation of molecular spectra by aggregating site-specific ones. Figure 2 displays both the site-specific spectra and the molecular spectra of a representative molecule (id #10578) from QM9, which represents the median MSE for the molecular

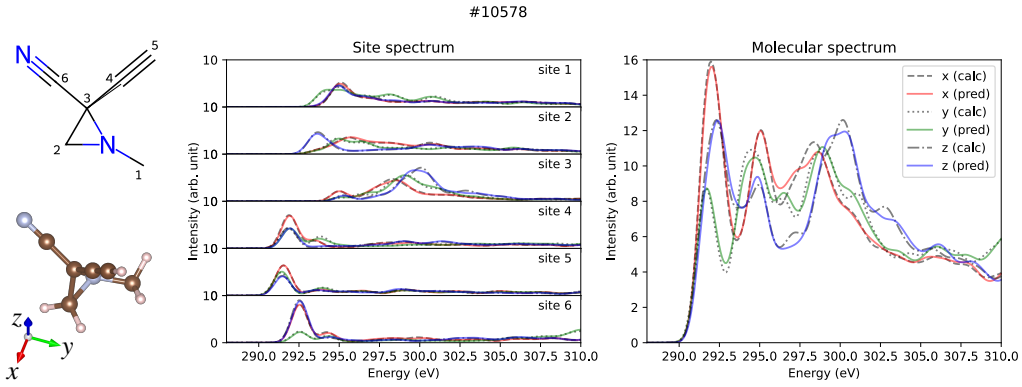

Figure 2: Prediction results of site-specific and molecular anisotropic C K-edge spectra for molecude id #10578 in QM9. Left panel shows the molecular structure formula and the three-dimensional structure drawn by RDKit[30] and VESTA[31], respectively. Middle and right panel show the predicted (solid lines) and calculated (dotted lines) spectra for site-specific spectra for each C site and molecular spectra, respectively.

spectra (refer to Supplementary). Not only the site-specific spectra but also the molecular spectra are predicted correctly including its dipole moment direction dependence, which confirms potential applications for molecular structure analysis taking advantage of anisotropic dipole transition.

Additionaly, we evaluated the generalization performance of the model by a scaffold splitting, a test on larger molecules, and prediction of orientation dependence of a specific molecule (see Supplementary). The results confirm the model's generalization ability to predict the spectra of molecules with different sizes and shapes. For the efficiency and speed of spectra prediction, we also compared the computational cost of our model with that of first-principles calculations, and comfirmed that our model is 10^6 times faster than first-principles calculations (see Supplementary).

4 Conclusion

We propose a message passing neural network based on PaiNN for predicting core electron excitation spectra considering its constraints on the physically required symmetry. We tested the model on C K-edge spectra of organic molecules, and it can predict the general shape of most of the site-specific spectra and molecular spectra including its dependence on the dipole moment direction.

So far, analysis considering anisotropy of core electron excitation spectra requires a highly accurate reference, whether by calculation or experiment. The cost of obtaining such reference may become a bottleneck in future, and our model, which is low-cost compared to first-principles calculations, will be useful for analyzing materials such as polymers, liquid cells, and organic solar cells.

Furthermore, the proposed model architecture is extended to be applied to other targets than ELNES/XANES by setting the required physical symmetry as an inductive bias to predict spectral data that physically depends on the orientation to molecules and crystals, which can advance GNN in materials science by explicitly incorporating directional dependencies in physical quantities alongside the representation of material graphs. This new approach complements the existing GNN paradigm by filling the gap in addressing directional aspects, fostering more accurate predictions in materials science. We anticipate that this novel strategy will contribute to an improved understanding of material anisotropy, offering innovative methodologies for detailed characterization.

Acknowledgments and Disclosure of Funding

This study was supported by the Grant-in-Aid for Scientific Research (Grant No. 19H05787) from the MEXT and CREST (Grant No. JPMJCR1993) from the JST. The authors declare no competing interests. We wish to thank Kento Nishio, Naoto Kawaguchi, and Izumi Takahara for their helpful discussions.

References

[1] M. G. Samant, J. Stöhr, H. R. Brown, T. P. Russell, J. M. Sands, and S. K. Kumar. Nexafs studies on the surface orientation of buffed polyimides. *Macromolecules*, 29(26):8334–8342, 1996.

[2] Nobuhiro Kawatsuki, Yusuke Taniguchi, Mizuho Kondo, Yuichi Haruyama, and Shinji Matsui. Comparison of the photoinduced orientation structure in the bulk and at the near-surface of a photoalignable liquid crystalline polymer film. *Macromolecules*, 48(7):2203–2210, 2015.

[3] Umut Aygül, David Batchelor, Ulf Dettinger, Seyfullah Yilmaz, Sybille Allard, Ullrich Scherf, Heiko Peisert, and Thomas Chassé. Molecular orientation in polymer films for organic solar cells studied by nexafs. *The Journal of Physical Chemistry C*, 116(7):4870–4874, 2012.

[4] Huifeng Yao, Long Ye, Hao Zhang, Sunsun Li, Shaoqing Zhang, and Jianhui Hou. Molecular design of benzodithiophene-based organic photovoltaic materials. *Chemical Reviews*, 116(12):7397–7457, 2016.

[5] J. J. Rehr and R. C. Albers. Theoretical approaches to x-ray absorption fine structure. *Rev. Mod. Phys.*, 72:621–654, Jul 2000.

[6] J.J. Rehr and A.L. Ankudinov. Progress in the theory and interpretation of xanes. *Coordination Chemistry Reviews*, 249(1):131–140, 2005. Synchrotron Radiation in Inorganic and Bioinorganic Chemistry.

[7] Hidekazu Ikeno and Teruyasu Mizoguchi. Basics and applications of ELNES calculations. *Journal of Electron Microscopy*, 66(5):305–327, oct 2017.

[8] Kiran Mathew, Chen Zheng, Donald Winston, Chi Chen, Alan Dozier, John J. Rehr, Shyue Ping Ong, and Kristin A. Persson. High-throughput computational x-ray absorption spectroscopy. *Scientific Data*, 5(1), July 2018.

[9] Yiming Chen, Chi Chen, Chen Zheng, Shyam Dwaraknath, Matthew K. Horton, Jordi Cabana, John Rehr, John Vinson, Alan Dozier, Joshua J. Kas, Kristin A. Persson, and Shyue Ping Ong. Database of ab initio L-edge X-ray absorption near edge structure. *Scientific Data*, 8(1):153, jun 2021.

[10] Kiyou Shibata, Kakeru Kikumasa, Shin Kiyohara, and Teruyasu Mizoguchi. Simulated carbon K edge spectral database of organic molecules. *Scientific Data*, 9(1):214, may 2022.

[11] Haoyue Guo, Matthew R. Carbone, Chuntian Cao, Jianzhou Qu, Yonghua Du, Seong-Min Bak, Conan Weiland, Feng Wang, Shinjae Yoo, Nongnuch Artrith, Alexander Urban, and Deyu Lu. Simulated sulfur K-edge X-ray absorption spectroscopy database of lithium thiophosphate solid electrolytes. *Scientific Data*, 10(1):349, jun 2023.

[12] Matthew R. Carbone, Mehmet Topsakal, Deyu Lu, and Shinjae Yoo. Machine-learning x-ray absorption spectra to quantitative accuracy. *Phys. Rev. Lett.*, 124:156401, Apr 2020.

[13] Marwah M. M. Madkhali, Conor D. Rankine, and Thomas J. Penfold. Enhancing the analysis of disorder in x-ray absorption spectra: application of deep neural networks to t-jump-x-ray probe experiments. *Phys. Chem. Chem. Phys.*, 23:9259–9269, 2021.

[14] C. D. Rankine and T. J. Penfold. Accurate, affordable, and generalizable machine learning simulations of transition metal x-ray absorption spectra using the XANESNET deep neural network. *The Journal of Chemical Physics*, 156(16):164102, 04 2022.

[15] Luke Watson, Conor D. Rankine, and Thomas J. Penfold. Beyond structural insight: a deep neural network for the prediction of pt l2/3-edge x-ray absorption spectra. *Phys. Chem. Chem. Phys.*, 24:9156–9167, 2022.

[16] Animesh Ghose, Mikhail Segal, Fanchen Meng, Zhu Liang, Mark S. Hybertsen, Xiaohui Qu, Eli Stavitski, Shinjae Yoo, Deyu Lu, and Matthew R. Carbone. Uncertainty-aware predictions of molecular x-ray absorption spectra using neural network ensembles. *Phys. Rev. Res.*, 5:013180, Mar 2023.

[17] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs, 2022.

[18] Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules, 2022.

[19] Kristof T. Schütt, Oliver T. Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra, 2021.

[20] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1), May 2022.

[21] C. Lawrence Zitnick, Abhishek Das, Adeesh Kolluru, Janice Lan, Muhammed Shuaibi, Anuroop Sriram, Zachary Ulissi, and Brandon Wood. Spherical channels for modeling atomic interactions, 2022.

[22] Saro Passaro and C. Lawrence Zitnick. Reducing so(3) convolutions to so(2) for efficient equivariant gnns, 2023.

[23] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR, 06–11 Aug 2017.

[24] 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. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, nov 2012.

[25] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, dec 2014.

[26] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 2021.

[27] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[28] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[30] RDKit: Open-source cheminformatics. `http://www.rdkit.org`.

[31] Koichi Momma and Fujio Izumi. *VESTA3* for three-dimensional visualization of crystal, volumetric and morphology data. *Journal of Applied Crystallography*, 44(6):1272–1276, Dec 2011.

Supplementary Material

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:

$$S_{\mathrm{noise}} = \mathrm{Poisson}(S_{\mathrm{mol}}(\mathcal{G}, \hat{\mathbf{n}}) \times \lambda)/\lambda, \tag{5}$$

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:

$$\mathrm{SAM}(s_i, s_j) = \arccos\left(\frac{\langle s_i, s_j \rangle}{\|s_i\|\|s_j\|}\right). \tag{6}$$

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:

$$\begin{aligned} H^m(t;\Delta) &= -\sum_{j=1}^{J} p^m_{t,\Delta}(j) \log_2 p^m_{t,\Delta}(j), \\ p^m_{t,\Delta}(i) &= \frac{m(t, s_i)}{\sum_{j=1}^{J} m(t, s_j)}, \end{aligned} \tag{7}$$

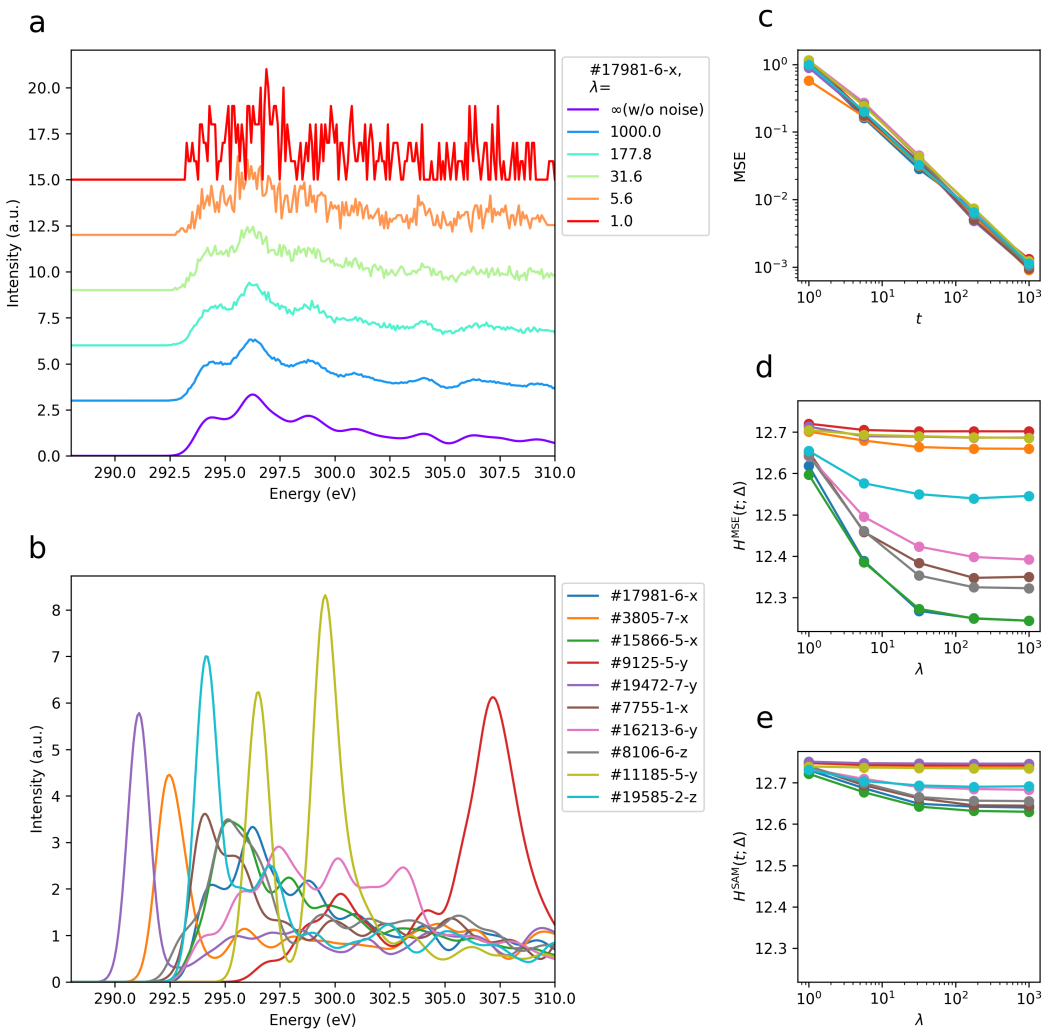

Figure S1: (a) Spectra with no noise and Poisson noise with different Poisson noise factor λ. (b) The ten randomly selected spectra from the dataset. (c) Relationship between MSE and λ. (d) and (e) spectra discriminatory entropy of the predicted spectra λ with respect to λ for MSE and SAM, respectively.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

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.

Evaluation of prediction on molecular spectra

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

$$S_{\mathrm{mol}}(\mathcal{G}, \hat{\mathbf{n}}) = \sum_n S_n(\mathcal{G}, \hat{\mathbf{n}}). \tag{8}$$

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. S2. 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. S3, S4, 2 in the main text, S5, and S6, respectively. All the molecular structure formula and the three-dimensional structures are drawn by RDKit[30] and VESTA[31], 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. 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.

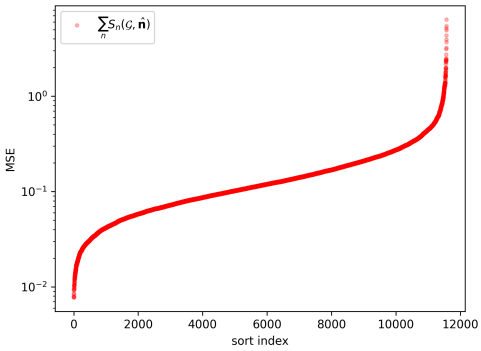

Figure S2: Sorted MSE of the molecular anisotropic C K-edge spectra of the prediction on the test dataset.

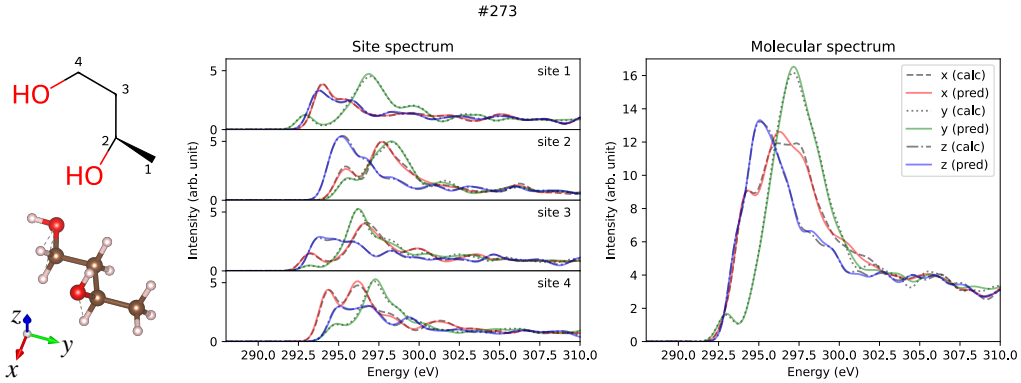

Figure S3: Prediction results of site-specific and molecular anisotropic C K-edge spectra for molecude id #273 in QM9, which is located at 0% percentile (best accuracy). Left panel shows the molecular structure formula and the three-dimensional structure. Middle panel and right panel show the predicted (solid lines) and calculated (dotted lines) spectra for site-specific spectra for each C site and molecular spectra, respectively.

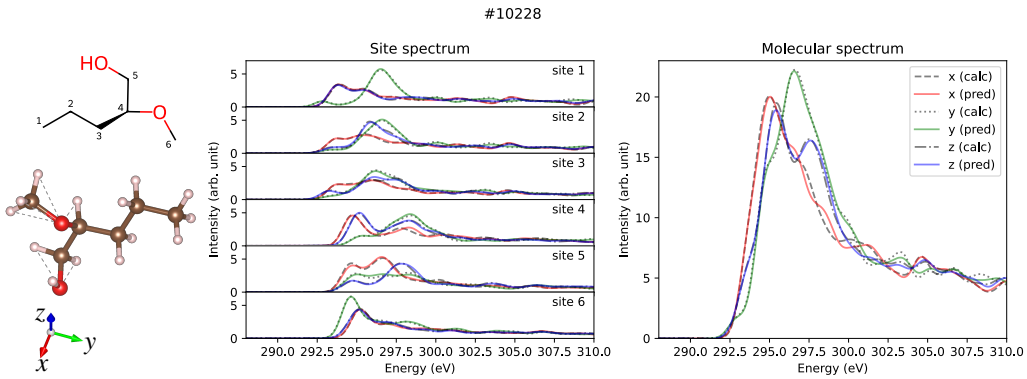

Figure S4: Prediction results of site-specific and molecular anisotropic C K-edge spectra for molecude id #10228 in QM9, which is located at 25% percentile. Left panel shows the molecular structure formula and the three-dimensional structure. Middle panel and right panel show the predicted (solid lines) and calculated (dotted lines) spectra for site-specific spectra for each C site and molecular spectra, respectively.

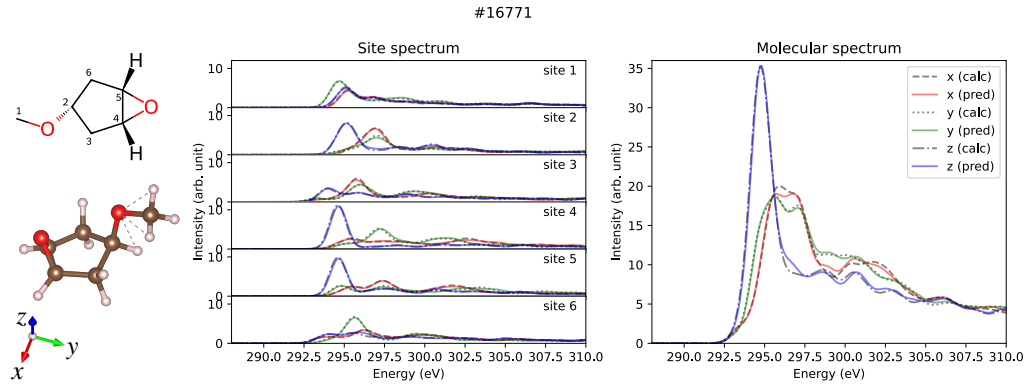

Figure S5: Prediction results of site-specific and molecular anisotropic C K-edge spectra for molecude id #16771 in QM9, which is located at 75% percentile. Left panel shows the molecular structure formula and the three-dimensional structure. Middle panel and right panel show the predicted (solid lines) and calculated (dotted lines) spectra for site-specific spectra for each C site and molecular spectra, respectively.

Scaffold split evaluation

To further validate the model's predictive performance beyond random splitting, we proceeded to evaluate its generalization using scaffold splitting[32], a technique that partitions datasets into training and testing sets while preserving structural diversity among molecules. Employing the ScaffoldSplitter class from the deepchem[33] 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 S7 shows the prediction results for the test data using scaffold splitting. As shown in Fig. S7a, there is an overall degradation in predictive accuracy, as indicated by the increased MSE compared to that of random split shown in Fig. 1. Figures S7b-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.

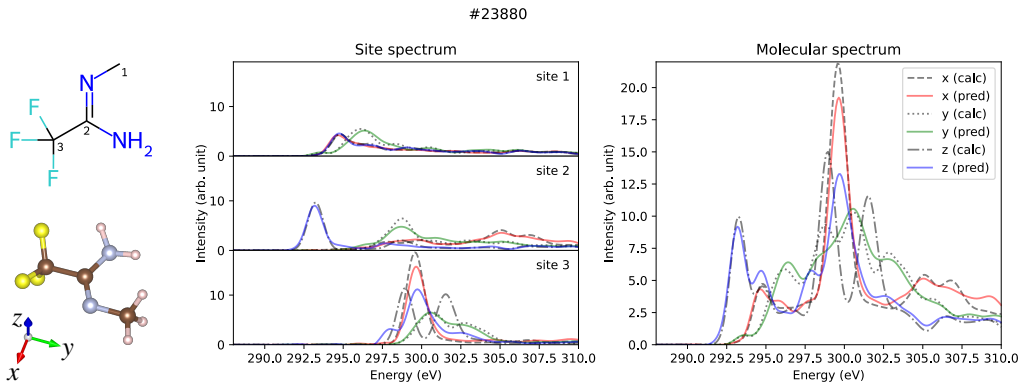

Figure S6: Prediction results of site-specific and molecular anisotropic C K-edge spectra for molecude id #23880 in QM9, which is located at 100% percentile (worst accuracy). Left panel shows the molecular structure formula and the three-dimensional structure. Middle panel and right panel show the predicted (solid lines) and calculated (dotted lines) spectra for site-specific spectra for each C site and molecular spectra, respectively.

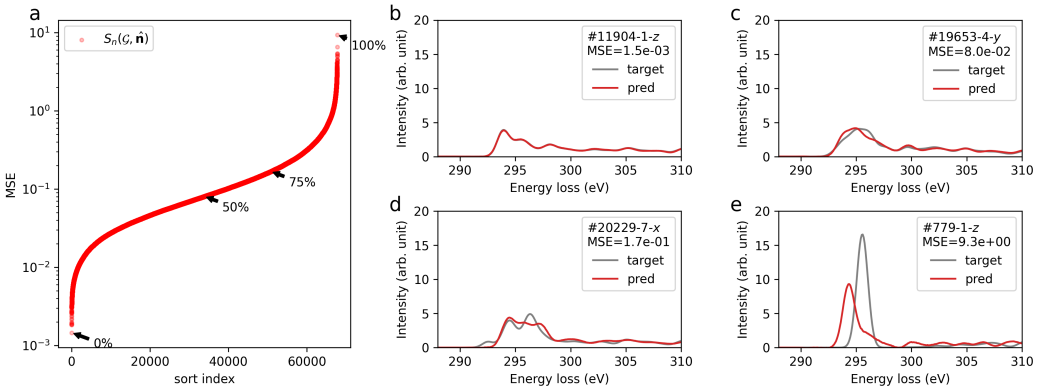

Figure S7: Prediction results of site-specific anisotropic C K-edge spectra $S_n(\mathcal{G}, \hat{\mathbf{n}})$ for the test dataset split by scaffold splitting. (a) Sorted MSE of the prediction on the test dataset. (b-e) Predicted (red) and calculated (gray) spectra for typical percentiles in terms 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.

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 S8 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.

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

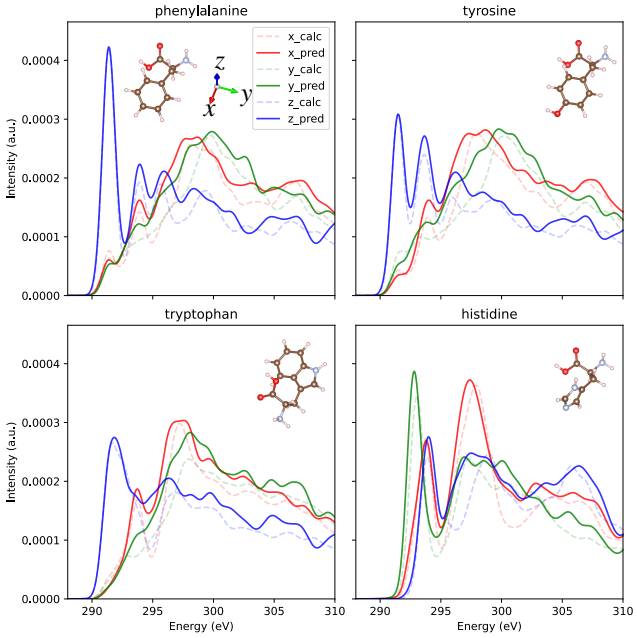

Figure S8: Prediction and calculation results of the C K-edge molecular spectra for the four aromatic amino acids. The predicted spectra are shown in solid lines, and the calculated spectra are shown in dotted lines.

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 S9 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 splitting evaluated in Sec. 3.

The spectra for the x-direction shown in Fig. S9 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. S9 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.

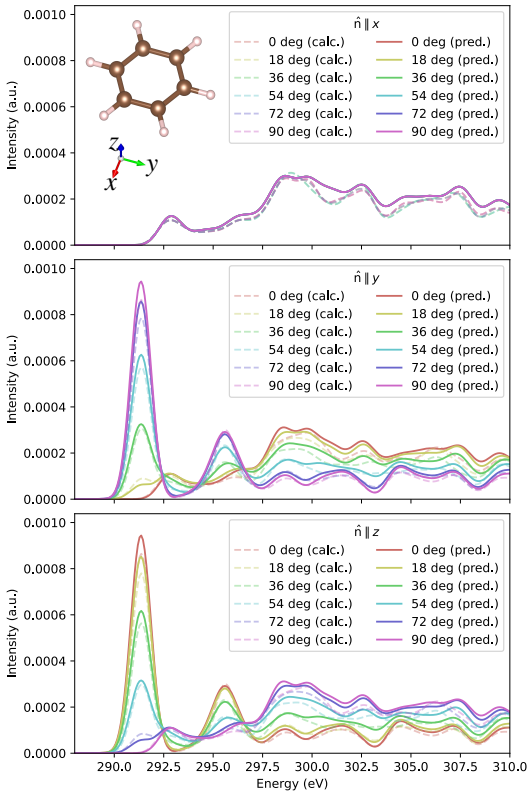

Figure S9: Prediction results of C K-edge molecular spectra of a benzene molecule for rotation about x axis. The predicted spectra for each orientation of the dipole vector $\hat{\mathbf{n}}$ are shown in solid lines, and the calculated spectra are shown in dotted lines. The color of the lines corresponds to the rotation angle of the benzene molecule about x-axis, specifically, red to purple corresponds to 0 to 90 degrees.

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[10] 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[34] 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[35, 36] within the plane-wave basis pseudopotential method[37] implemented in the CASTEP code[38] under the same the calculation condition as in the C-K edge dataset[10] except that we used a $20 \times 20 \times 20 \text{Å}^3$ cubic cell with 2,000 extra bands for aromatic amino acids.

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 Section 2.3 (See Supplementary for details).

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. S10a 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,

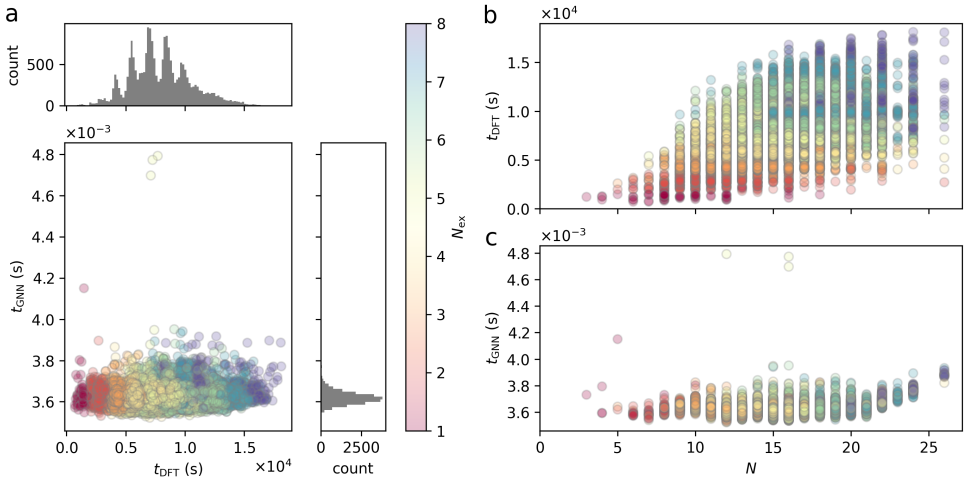

Figure S10: Comparison between computational time of first-principles calculations (t_{DFT}) and that of our model (t_{GNN}). (a) Scatter plot between t_{DFT} and t_{GNN}, colored by number of unique excitation sites N_{ex}. (b, c) number of atom (N) dependence of t_{DFT} to t_{GNN}, colored by N_{ex} in the same scale as in (a).

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. S10 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 $N\mathrm{ex}$ dependence. Furthermore, Figs. S10b 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 capability enables the efficient exploration of detailed angular dependencies without the need for additional computations involving structural rotations.

Computational conditions for time measurement

The calculation time for each molecule by first-principles calculation, t_{DFT} in Fig. S10, was analyzed on the computations conducted to construct the C-K edge database[10]. The calculation procedure and conditions of the C-K edge spectra and excitation energies are described in the reference[10]. 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. S10, was measured using a single NVIDIA GeForce RTX 4090 GPU.

References for supplementary materials

[32] Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. PMID: 8709122.

[33] Bharath Ramsundar, Peter Eastman, Patrick Walters, Vijay Pande, Karl Leswing, and Zhenqin Wu. *Deep Learning for the Life Sciences*. O'Reilly Media, 2019. `https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837`.

[34] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 10 2022.

[35] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, 1964.

[36] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, 1965.

[37] M. C. Payne, M. P. Teter, D. C. Allan, T.A. Arias, and J. D. Joannopoulos. Iterative minimization techniques for ab initio total-energy calculations - molecular-dynamics and conjugate gradients. *Rev. Mod. Phys.*, 64:1045–1097, 1992.

[38] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. J. Probert, K. Refson, and M.C. Payne. First principles methods using CASTEP. *Z. Kristall.*, 220:567–570, 2005.