LLM-based agent empowered with geometric deep learning, data-oriented approaches and quantum chemistry to unravel synchrotron data of *operando* catalysis

Bogdan Protsenko [1] **Sergey Guda** [1] **Kulaev Kirill** [1] **Mikhail Lifar** [1] **George Asaturov** [1] **Nazar Chubkov** [1] **Daniil Kuprianenko** [1] **George Kochiev** [1] **Alexander Guda** [1] **Alexander Soldatov** [1]

[1]*Southern Federal University, The Smart Materials Research Institute, Rostov-on-Don, 344090, Russian Federation.* Correspondence to: Alexander Guda aguda@sfedu.ru.

1. Introduction

Catalytical processes are at the heart of the chemical industry. In turn, X-ray Absorption Spectroscopy (XAS) is a powerful and unique tool for analyzing the electronic and structural properties of materials, particularly for elucidation the oxidation states and local atomic environments of catalysts' active sites under working conditions (*operando*). However, the interpretation of XAS spectra can be complicated ill-posed inverse problem. Therefore, usually the complex interplay between chemical reasoning, first-principles calculations and data-oriented approaches are needed to lift the conundrum of the catalyst action and poisoning mechanisms. The advent of autonomous LLM-based (Large Language Model) agents has already proven to be superior in tasks where planning, reasoning and tools' calling is needed. In this work we present the development of the LLM-based agent for automatic and robust XAS data analysis. We empower the agent developed with multiple tools, such as the use of the PyFitIt framework for data-oriented machine learning XAS analysis, multiple originally measured and public experimental XAS reference databases of well-defined species from the whole periodic table, theoretical database of single-site catalysts, E(3)-equivariant deep learning model for rapid prediction and quantum chemistry-informed structure refinement (Fig. 1). We evaluate each of the developed tool and test model in real-world tasks for *operando* catalysts' synchrotron experiments, namely industrially relevant Philips, Ziegler-Natta and hydroformulation Rh/NH3 systems. It is the first time when the problem of XAS data analysis is solved within one tool for all absorption edges of every element of the periodic table, drastically boosting both synchrotron and catalysis communities. Developed models are open-source and made available withing the *Telegram bot* services.

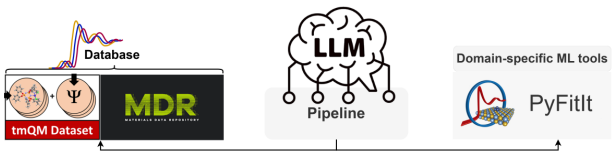

Fig. 1: LLM-agent for XAS spectra unraveling scheme

2. Methods

2.1 LLM agent and PyfitIt approach

Developed LLM-based agent for automatic XAS data analysis is implemented as LLM-model agnostic (Main models used are LLaMa3.2-40B-Instruct and DeepSeek-r1) agent with tool calling options and code-based actions, in which it integrates full range of the specific tools needed to analyze XAS data. The bottlenecks in the automatic processing of experimental data are the lack of chemically diverse XANES reference libraries and the systematic differences between theory and experiment. Therefore, compiling experimental reference libraries across the periodic table and rational application of ML methodology to small (in terms of data science) training data sets becomes increasingly important. This work revises the classical XANES fingerprint analysis by database augmentation, feature extraction, cross-validation, and uncertainty analysis automated within the PyFitIt framework to ensure the balance of ML methods and domain-specific knowledge.

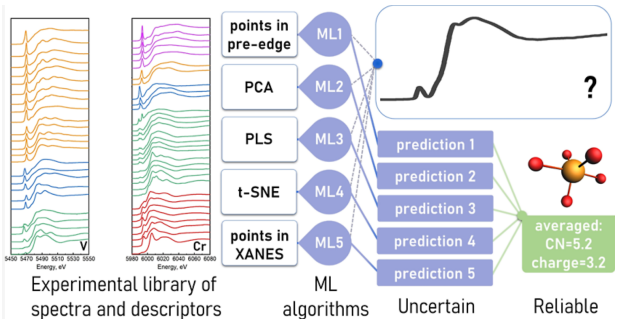

Fig. 2: Data-oriented cross-element and cross-edge approach to the XAS data analysis withing PyFitIt

2.2 Database compilation

The integration of data-driven methodologies in XAS has become indispensable for spectral interpretation. We use the Materials Data Repository (MDR) providing experimental XAS data of 2,500 spectra for almost all absorption edges and each element in the periodic table. Moreover we add to it our in-house collected 300 spectra of well-defined (meaning that structure of each was approved by NMR, IR, scXRD techniques) both molecular and bulk species of Cr, V, Rh, Pd, Pt and Ru. All experimental spectra were additionally labeled with through laborious manual work of several XAS specialists to en-

sure data quality. The theoretical database we calculated contains 60,000 spectra of 3d and 4d transition metal complexes based on tmQM database and full multiple scattering approach with self-consistent field calculations. Integration of this library into autonomous XAS analysis assistant aims to refine structural hypotheses iteratively using chemical reasoning of LLM.

2.3 DeepFit approach

To further advance XAS analysis capabilities, we developed a machine learning approach for rapid XAS spectra prediction and chemically- and physically-informed structure refinement. The model of E(3)-equivariant neural network was trained to predict theoretical XAS spectra directly from atomic structure. In conjunction with quantum chemical estimations of structure stability such approach enables to fit material's atomic structure to experimental spectra in differentiable manner combines both chemical and spectroscopic insights.

We used the E(3)-equivariant convolutional graph neural network from the e3nn library to approximate structure-spectrum mapping with forces from semi-empiric quantum chemistry xTB code.

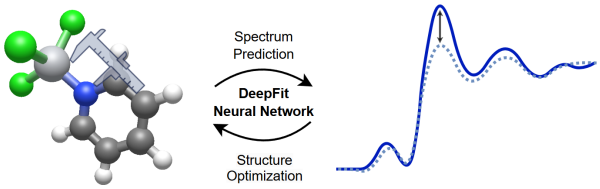

Fig. 3: DeepFit approach

Given feature representations of n atoms $X = (x_1, x_2, ..., x_n) \in \mathbb{R}^F$, which are at positions $\mathbf{R} = (\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_n) \in \mathbb{R}^3$. Thus, fitted atomic structure $\mathbf{R}^*$ can be obtained by minimizing the deviation between the spectrum predicted by the neural network and the ground truth. Due to the chemical relevance of the desired structure, the optimization result should be not only the minimum of the deviation between spectra (first term in the eq. 1), but also the minimum on the potential energy surface (second term in the eq. 1). Mentioned optimization task is given as:

$$\mathbf{R}^* = argmin(\int [F(\mathbf{R}, k) - \chi(k)]^2 dk + \lambda E(\mathbf{R})), \quad (1)$$

Where $\chi(k)$ is a ground truth (experimental) spectrum and $F(\mathbf{R}, k)$ is an output of the model, E is total electronic energy of the system and λ is some positive controllable parameter. The usage of a neural network allows us to estimate the gradient of the optimized function to the coordinates of atoms in analytical way. Nuclear gradients calculated by the quantum-chemical method are used. Thus overall gradient, expressed as:

$$\frac{\partial(\int [F(\mathbf{R}, k) - \chi(k)]^2 dk)}{\partial \mathbf{R}} + \lambda \frac{\partial E(\mathbf{R})}{\partial \mathbf{R}} \quad (2)$$

3. Conclusion

In this study, we have successfully developed an innovative LLM-based agent that integrates advanced geometric deep learning techniques, data-oriented approaches, and quantum chemistry principles to enhance the automatic and robust analysis of X-ray Absorption Spectroscopy (XAS) data in the context of *operando* catalysis. Our approach addresses the complexities associated with interpreting XAS spectra by combining robust machine learning frameworks, comprehensive reference databases, domain-specific tools, and cutting-edge predictive models inside the LLM-autonomous agent as a step forward to boost spectroscopy and ML techniques in catalysis.

Acknowledgments

The research was supported by the Strategic Academic Leadership Program of the Southern Federal University ("Priority 2030").

References

1. Protsenko, B.O.; Kakiuchi, Y.; Guda, S.A.; Trummer, D.; Zabilska, A.; Shapovalova, S.; Soldatov, A.V.; Safonova, O.V.; Copéret, C.; Guda, A.A. Fingerprint Analysis of X-Ray Absorption Spectra with the Machine-Learning Method Trained on the Multielement Experimental Library. J. Phys. Chem. C 2025, 129, 2525–2534.

2. Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks, 2022, arXiv:2207.09453.

3. Rehr, John J. and Kas, Joshua J. and Vila, Fernando D. and Prange, Micah P. and Jorissen, Kevin. Parameter-free calculations of X-ray spectra with FEFF9, 2010, Physical Chemistry Chemical Physics

4. Balcells, David and Skjelstad, Bastian Bjerkem. tmQM Dataset—Quantum Geometries and Properties of 86k Transition Metal Complexes, 2020, Journal of Chemical Information and Modeling

5. Ishii, et al. Integration of X-ray absorption fine structure databases for data-driven materials science, 2015, Science and Technology of Advanced Materials: Methods

6. Martini, A. and Guda, S.A. and Guda, A.A. et al., A.V. PyFitit: The software for quantitative analysis of XANES spectra using machine-learning algorithms, Computer Physics Communications, Volume 250, 107064, 0010-4655, 2020

7. A. Martini, S. A. Guda, A. A. Guda, E. Priola, E. Borfecchia, S. Smolders, K. Janssens, D. De Vos, A. V. Soldatov, Revisiting the Extended X-ray Absorption Fine Structure Fitting Procedure through a Machine Learning-Based Approach, J. Phys. Chem. A 2021, 125, 32, 7080–7091