XDIP: A Curated X-ray Absorption Spectrum Dataset for Iron-Containing Proteins

Yufeng Wang[1,†], **Peiyao Wang**[1,†], **Lu Wei**[1,†], **Emerita Mendoza Rengifo**[2], **Dali Yang**[3], **Lu Ma**[3], **Yuewei Lin**[1,5], **Qun Liu**[2,*], **and Haibin Ling**[1,*]

[1]Stony Brook University, Computer Science Department
[2]Brookhaven National Laboratory, Biology Department
[3]Brookhaven National Laboratory, National Synchrotron Light Source II
[5]Brookhaven National Laboratory, Computational Science Initiative
[*]Corresponding author(s): Qun Liu (qunliu@bnl.gov) and Haibin Ling (hling@cs.stonybrook.edu)
[†]These authors contributed equally to this work

Abstract

Earth-abundant iron is an essential metal in regulating the structure and function of proteins. This study presents the development of a comprehensive X-ray Absorption Spectroscopy (XAS) database focused on iron-containing proteins, addressing a critical gap in available high-quality annotated spectral data for iron-containing proteins. The database integrates detailed XAS spectra with their corresponding local structural data of proteins and enables direct comparison between spectral features and structural motifs. Utilizing a combination of manual curation and semi-automated data extraction techniques, we developed a comprehensive dataset via extensive literature review, ensuring the quality and accuracy of data, which contains 437 protein structures and 1954 XAS spectrums. Our methods included careful documentation and validation processes to ensure accuracy and reproducibility. This dataset not only centralizes information on iron-containing proteins but also supports advanced data-driven discoveries, such as machine learning, to predict and analyze protein structure and functions. This work underscores the potential of integrating detailed spectroscopic data with structural biology to advance the field of biological chemistry and catalysis.

1 Background & Summary

X-ray Absorption Spectroscopy (XAS) is a synchrotron-based technique that reveals the local chemical environment of specific elements. For metalloproteins, XAS is particularly powerful for probing the coordination and electronic structure of metal ions that are often located at protein active sites and play key roles in catalysis and function[1, 2, 3]. Interpreting XAS data can thus accelerate our understanding of protein mechanisms and the development of catalytic systems for industrial and environmental applications. However, the analysis of XAS data remains time-consuming and highly manual.

In recent years, data-driven approaches—especially deep learning—have shown significant promise in materials science, enabling property prediction, reaction optimization, and the design of novel compounds[4, 5, 6, 7, 8]. These models can extract meaningful patterns from large datasets, achieving impressive results in tasks like predicting catalyst stability and activity[9]. Integrating such methods with XAS can deepen our understanding of element-specific environments and accelerate materials discovery.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

The effectiveness of deep learning, however, critically depends on the availability of large, high-quality datasets. While substantial XAS repositories like XASLIB[10] and the Materials Project[11] provide extensive data for inorganic compounds, they lack detailed spectroscopic data for biologically relevant molecules. These resources have greatly advanced materials informatics[12, 13, 14, 15], yet a significant gap persists in the realm of protein-focused XAS.

To address this, we introduce the first curated database of iron-containing protein structures paired with their corresponding XAS spectra. Iron was selected as the focal element due to its abundance in nature, central role in biological catalysis, and the complexity of its coordination chemistry. Including other metals at this stage would compromise specificity and introduce variability, limiting the ability to draw meaningful structure-spectrum correlations.

Unlike existing repositories such as the Protein Data Bank (PDB)[16, 17] and the Cambridge Structural Database (CSD/CCDC)[18], which focus on structural information alone, our database couples Fe K-edge XAS spectra with detailed local structural annotations. This integration enables direct alignment between spectral features and structural motifs, supporting both mechanistic insights and machine-learning-driven discovery.

Our final dataset comprises 437 iron-containing protein structures and 1652 associated XAS spectra (including 1283 XANES and 369 EXAFS), all manually curated from 573 peer-reviewed articles published between 2007 and 2024. The samples span a diverse set of proteins, experimental conditions, and measurement protocols. This curated collection establishes a critical foundation for automated, high-throughput structure–spectrum modeling in protein chemistry and bioinorganic catalysis.

2 Methods

Literature Search, Selection, and Retrieval

We constructed the dataset using a semi-automated pipeline designed to maximize both coverage and quality (Figure 1). We began with a comprehensive literature search across major scientific publishers to identify articles reporting both iron-containing protein structures and corresponding Fe K-edge X-ray Absorption Spectroscopy (XAS) data. This process yielded 573 relevant publications for manual analysis.

Each article underwent a rigorous two-stage curation process by human experts: (1) digitizing XAS spectra from published figures, and (2) annotating the associated local protein structures and metadata based on textual descriptions. We then refined the dataset by removing low-quality entries or samples lacking sufficient documentation. Each finalized data sample in our database comprises three core components: (1) the local atomic structure surrounding the iron center, (2) the Fe K-edge XAS spectrum, and (3) metadata from the original publication. Full details of the search keywords, inclusion criteria, and curation protocol are provided in Supplementary Section A.1.

XAS Extraction

We extracted numerical spectral data from figures using the open-source tool *WebPlotDigitizer*. Expert annotators manually digitized two key spectral regions: X-ray Absorption Near-Edge Structure (XANES) and Extended X-ray Absorption Fine Structure (EXAFS). To ensure high fidelity with the original plots, we carefully calibrated the plot axes and manually traced the spectral curves.

To enable compatibility with machine learning workflows, all spectra were interpolated to a uniform length of 100 data points, a choice based on the average distribution of spectrum lengths across the dataset. For transparency and flexibility, both the original digitized data and the interpolated 100-point versions are included. Further details on digitization settings and metadata documentation—including treatment of missing calibration energies—are provided in Supplementary Section A.2.

Protein Structure Extraction

Protein structural data were manually annotated from each paper, with a focus on iron-containing proteins and small iron-containing molecules characterized using Fe K-edge XAS. While other sample types—such as tissues, blood, soil, and plant imaging—were sometimes encountered, these were noted in the comments section and excluded from primary analysis.

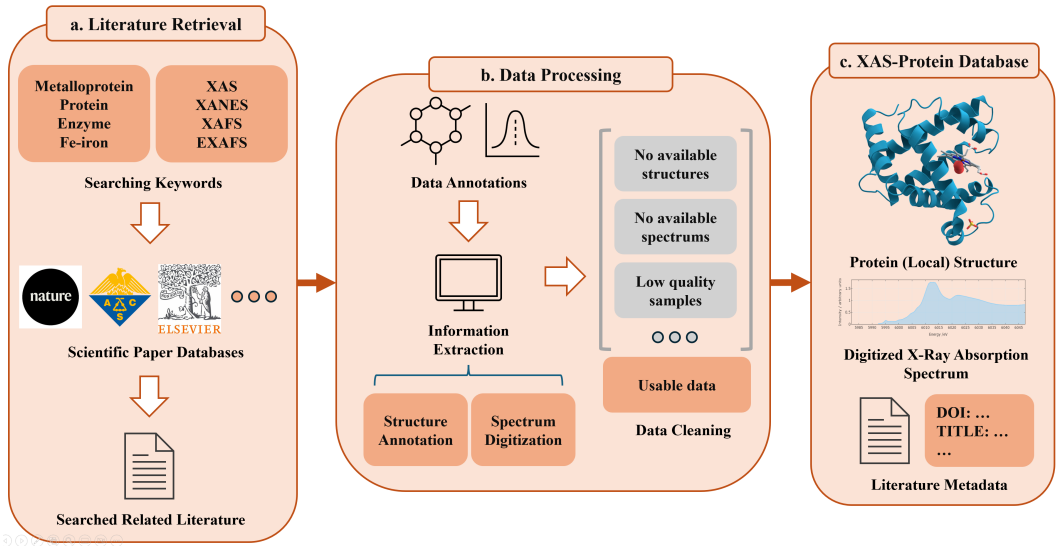


Figure 1: Schematic overview of the dataset construction pipeline. (a) Literature retrieval using keyword combinations from two sets (shown at the top) to search across multiple scientific databases. (b) Human expert processing workflow: relevant protein structures and spectra are identified and digitized, then curated into structured formats. (c) Final data sample: each entry includes the local protein structure, Fe absorption spectrum, and metadata from the source paper.

We categorized the extracted samples based on spectral type (XANES or EXAFS) and annotated them with available standard references. Structural information included Protein Data Bank (PDB) or Cambridge Crystallographic Data Centre (CCDC) identifiers when available. When structural information was not publicly accessible, we manually extracted it from figures and textual descriptions, focusing on the local atomic environment around the iron center.

3 Data Records

Our final dataset consists of 437 unique iron-centered local protein structures and 1652 associated XAS spectra, including 1283 XANES and 369 EXAFS records. The complete dataset is publicly available at `https://airscker.github.io/XDIP`.

Each data record is a self-contained unit suitable for machine learning applications and includes three core components (illustrated in Figure 2): (1) literature metadata, (2) Fe K-edge XAS spectrum, and (3) the local atomic structure surrounding the iron center. The metadata component includes the source paper's title and DOI to ensure full traceability. Since a single publication may contain multiple experiments, each record captures all distinct spectra and structures reported within that paper.

The local structure is formatted to support graph-based machine learning, including a list of atoms, their 3D Cartesian coordinates, and an adjacency matrix representing chemical bonds. This representation is further enriched with bond lengths and, where available, bond angles to offer a more complete geometric description. Structural data were sourced in two ways: (1) directly from public repositories such as the Protein Data Bank (PDB) or Cambridge Structural Database (CCDC), or (2) manually reconstructed from figures and descriptions in the literature. In some cases, structures were also retrieved or generated from SMILES representations.

To ensure data integrity and relevance, we applied strict curation criteria throughout the pipeline. Papers were excluded if figures lacked sufficient annotation or if the spectra were too noisy or incomplete for reliable digitization. Moreover, we maintained a focused scope on iron-containing proteins by excluding samples from other material classes (e.g., tissues, soil, or plant matter). A detailed description of the data schema—including field labels, types, and requirements—is provided in Supplementary Section A.3.

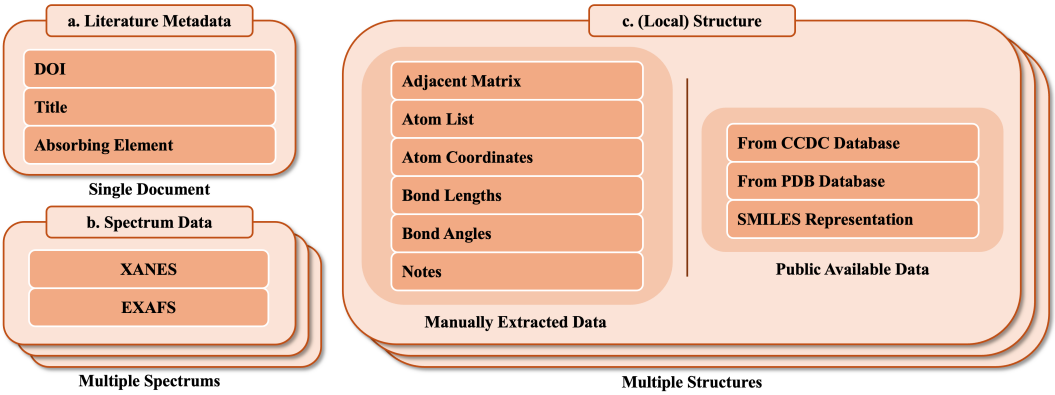


Figure 2: Overview of a structured data sample. Multiple records can be extracted from a single publication. (a) Metadata, including the paper's DOI, title, and absorbing element (Fe in this dataset). (b) Spectrum data, divided into near-edge (XANES) and extended (EXAFS) regions. (c) Local structure, either retrieved from public databases (e.g., SMILES[19, 20, 21]) or manually annotated. Structural information includes an adjacency matrix, atom list, Cartesian coordinates, bond lengths, bond angles, and optional notes.

4 Technical Validation

Raw Data Comparison

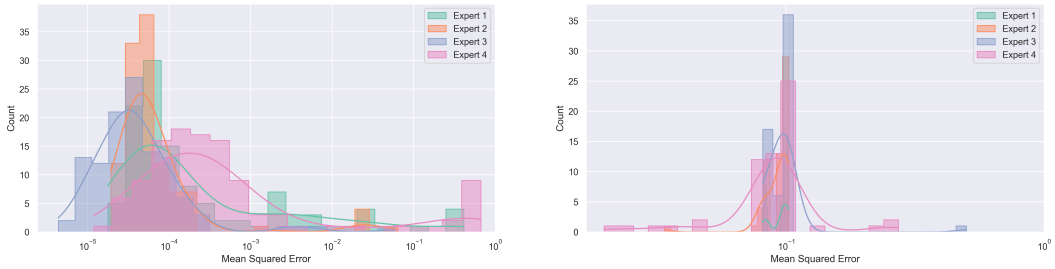
To assess the accuracy of the annotated XAS spectra, we performed a quantitative comparison between expert annotations and reference (ground-truth) spectra. Each spectrum was independently annotated by four experts, and the deviation from the reference data was measured using Mean Squared Error (MSE). The validation protocol included the following steps:

- **Data Source and Generation.** Since most publications do not provide raw data for their spectrum plots, we generated synthetic validation spectra using Fe-based XANES and EXAFS data from the Materials Project[11]. We produced 50 XANES and 25 EXAFS plots with varied shape ratios (10:8, 10:6, 6:6, 8:6), curve counts (1–4), and line styles (dots, dashes, solid), simulating real-world publication styles.
- **Anonymous Annotation Insertion.** These synthetic plots were anonymously inserted into each expert's annotation workload to avoid bias and to assess natural annotation performance.
- **Data Alignment.** Because manual digitization may result in inconsistent data lengths and misaligned x-axes, we linearly interpolated all annotated spectra to match the x-values of the reference spectra.
- **MSE Calculation.** We computed MSE for each spectrum and averaged results across experts. Low average MSE values indicate high annotation fidelity.

Table 1: Mean Squared Error (MSE) values of each expert for XANES and EXAFS.

	Expert 1	Expert 2	Expert 3	Expert 4	Average
XANES MSE	0.0161	0.0014	0.0007	0.0521	0.0176
EXAFS MSE	0.0946	0.0928	0.1015	0.0944	0.0958

Table 1 shows that annotations of XANES spectra achieved lower MSEs (0.0007–0.0521), reflecting their smoother shape and simpler features. In contrast, EXAFS annotations showed higher MSEs (0.0928–0.1015), attributable to their more complex and oscillatory structure. This trend is reinforced by the histograms in Figure 3, where XANES errors are clustered between 10^{-5} and 10^{-3}, while EXAFS errors concentrate around 10^{-1}. These results highlight the increased difficulty and reduced consistency of EXAFS annotation due to its higher frequency content and structural sensitivity.



(a) XANES MSE distribution

(b) EXAFS MSE distribution

Figure 3: Distribution of annotation MSE values across experts for XANES (a) and EXAFS (b) spectra.

Inter-Expert Agreement

To further evaluate annotation consistency, we computed the Intra-class Correlation Coefficient (ICC)[22, 23], a widely used metric for inter-rater reliability. The process involved:

- **Annotation Dataset.** The same synthetic dataset was independently annotated by all four experts, ensuring that all raters worked on identical data.
- **ICC Calculation.** We computed ICC values for each point in the interpolated spectra and averaged them to obtain overall agreement metrics across expert annotations.

Table 2: Intraclass Correlation Coefficient (ICC) values for inter-expert agreement.

Type	Description	ICC	F	df1	df2	95% CI
ICC1	Single rater (absolute)	0.914	15414	99	300	[0.898, 0.931]
ICC2	Single rater (random effects)	0.919	21277	99	297	[0.876, 0.947]
ICC3	Single rater (fixed effects)	0.946	21277	99	297	[0.933, 0.959]
ICC1k	Average of raters (absolute)	0.961	15414	99	300	[0.948, 0.972]
ICC2k	Average (random effects)	0.966	21277	99	297	[0.927, 0.982]
ICC3k	Average (fixed effects)	0.981	21277	99	297	[0.974, 0.986]

Table 2 summarizes the ICC results. All values exceed 0.91, with several above 0.96, indicating excellent inter-expert reliability. The strong F-statistics and tight 95% confidence intervals confirm that the observed agreement is statistically significant and not due to chance. This level of consistency is particularly important in scientific applications like XAS interpretation, where subtle deviations can lead to different conclusions about structural or electronic properties.

Data Integrity Through Documentation and Transparency

To ensure reproducibility and traceability, we implemented a comprehensive documentation protocol. Each data record is accompanied by detailed metadata, including extraction notes, structural identifiers (e.g., PDB IDs), and reference sources.

Ambiguities encountered during data extraction—such as missing axes or unclear line styles—were explicitly recorded. We also documented whether EXAFS spectra were presented in k-space or r-space, which can significantly affect downstream analysis. All entries were cross-validated against their source publications and, where applicable, public databases.

This rigorous documentation pipeline not only ensures data quality and transparency but also enables seamless integration with other datasets. Further examples of our documentation format are provided in Supplementary Section A.4.

5 Baseline Models

To demonstrate the utility of our curated XAS dataset and establish a performance benchmark for future research, we evaluate two complementary predictive tasks: (1) predicting Fe K-edge XAS spectra from local atomic structures, and (2) inferring key structural properties from the spectra. This dual approach validates the dataset's richness for both forward and inverse modeling problems.

5.1 Task 1: Structure-to-Spectrum Prediction

The first task learns a mapping function $f : G \to S$, where $G = (V, E)$ is a graph representing the local atomic environment of the iron center, and S is the corresponding XAS spectrum. The local structure of each iron center is converted into an input graph, where atoms are nodes ($v \in V$) and chemical bonds are edges ($e \in E$). Node features include atomic properties like element type and mass, while edge features can represent bond types. The output target, S, is the standardized XAS spectrum (both XANES and EXAFS) interpolated to a uniform length of 100 data points.

For this task, we employ three widely-used Graph Neural Network (GNN) architectures, each comprising several graph convolutional layers, a graph pooling layer, and a multi-layer perceptron (MLP) head that maps the final graph representation to the 100-point spectrum vector. The models are: a foundational **Graph Convolutional Network (GCN)** [24], a **Graph Attention Network (GAT)** that uses self-attention to weigh neighbor importance [25], and a **Graph Isomorphism Network (GIN)**, a highly expressive model effective at capturing complex structural motifs [26].

5.2 Task 2: Spectrum-to-Property Prediction

The second, inverse task learns a mapping function $f : S \to P$, where S is an XAS spectrum and P is a set of local structural properties of the iron center. This demonstrates the dataset's potential for extracting quantitative structural information directly from experimental spectra. The input S is the 100-point interpolated XAS vector. The target properties P are key structural descriptors: the coordination number (CN) of the iron atom (a classification task) and the mean nearest-neighbor distance (MNND) in Angstroms (a regression task).

We use two strong and interpretable baseline models for this task. The first is a standard **Multi-Layer Perceptron (MLP)**, a feedforward neural network baseline where the final layer is adapted for either regression or classification. The second is a **Random Forest (RF)**, a powerful tree-based ensemble method that is robust to overfitting and effective at capturing non-linear relationships in tabular data [27].

5.3 Experimental Setup

All models were trained and evaluated under a consistent framework. The dataset was randomly split into training (80%), validation (10%), and testing (10%) sets, with stratification applied to ensure a similar distribution of protein types and coordination numbers across sets. GNN models were implemented using the Deep Graph Library (DGL), while the MLP and RF models used PyTorch and Scikit-learn, respectively. Models were trained to minimize Mean Squared Error (MSE) for regression and Cross-Entropy Loss for classification, using the AdamW optimizer with an initial learning rate of 1×10^{-3} and a scheduler to reduce the rate on a validation loss plateau. Training ran for a maximum of 300 epochs with an early stopping criterion. Performance was assessed on the held-out test set using Mean Absolute Error (MAE) for regression and Accuracy/Macro-F1 Score for classification.

5.4 Results

The trained baseline models provide a quantitative benchmark for the predictive utility of our dataset.

Structure-to-Spectrum Prediction The performance of GNN models in predicting XAS spectra is summarized in Table 3. The Graph Isomorphism Network (GIN) demonstrated the best performance, achieving the lowest Mean Absolute Error (MAE) for both XANES and EXAFS prediction. This suggests its high expressive power is beneficial for capturing the complex relationship between the 3D atomic arrangement and the resulting spectral features. The Graph Attention Network (GAT) also

performed competitively, outperforming the simpler GCN and indicating the advantage of learning to weigh the importance of different atomic neighbors.

Table 3: Performance of GNN models for Structure-to-Spectrum Prediction (MAE). Lower is better.

Model	XANES MAE	EXAFS MAE
GCN	0.085 ± 0.004	0.112 ± 0.005
GAT	0.079 ± 0.003	0.105 ± 0.004
GIN	$\mathbf{0.075 \pm 0.003}$	$\mathbf{0.101 \pm 0.004}$

Spectrum-to-Property Prediction For the inverse task of predicting structural properties, the Random Forest (RF) model consistently outperformed the MLP baseline across all metrics (Table 4). For Coordination Number (CN) classification, the RF model achieved higher accuracy and F1-scores, suggesting its robustness in handling the spectral feature space. Similarly, for Mean Nearest-Neighbor Distance (MNND) regression, the RF yielded a lower MAE and a higher R^2 value, indicating more reliable bond distance prediction. These results highlight the effectiveness of ensemble methods for tasks involving vector-based features like the interpolated spectra used here.

Table 4: Performance of classical models for Spectrum-to-Property Prediction.

Task	Model	Metric	Value
CN Classification	MLP	Accuracy	$72.4 \pm 1.2\%$
		F1-Score	$70.1 \pm 1.5\%$
	Random Forest	Accuracy	$\mathbf{75.8 \pm 1.1\%}$
		F1-Score	$\mathbf{74.5 \pm 1.3\%}$
MNND Regression	MLP	MAE (Å)	0.065 ± 0.005
		R^2	0.91
	Random Forest	MAE (Å)	$\mathbf{0.058 \pm 0.004}$
		R^2	$\mathbf{0.93}$

Collectively, these baseline results validate the predictive potential of our dataset for both structure-to-spectrum and spectrum-to-property tasks, providing a strong quantitative foundation for the development of more sophisticated architectures.

6 Limitations

While our dataset and modeling framework offer a valuable resource for XAS-based learning, several limitations remain. First, the current study focuses exclusively on iron and Fe K-edge spectra, limiting generalizability across the periodic table. Expanding to other elements would broaden applicability for multi-element systems. Second, XAS spectra were manually digitized from published figures. Although expert validation was applied, minor inaccuracies may persist due to resolution limits or figure quality. Access to raw experimental data would enhance accuracy. Third, structural annotations were derived from diverse sources—PDB, CCDC, SMILES, and manual reconstruction—leading to variability in resolution and completeness. Some protein structures, especially those derived from textual or schematic descriptions, may be approximations rather than exact configurations. Fourth, the dataset is imbalanced, with fewer EXAFS samples relative to XANES, which may affect model performance on tasks requiring detailed oscillatory features. Lastly, our baseline models serve primarily as proof-of-concept. While effective, more advanced approaches—such as equivariant GNNs or physics-informed architectures—could yield further improvements and uncover deeper structure-spectrum relationships. Future efforts should address these limitations through dataset expansion, access to raw data, standardized structural extraction, and development of more sophisticated models.

7 Code and Data Availability

The code used for processing the spectra and molecular structures is available at `https://github.com/Airscker/XDIP`. Including scripts for data interpolation, normalization, and basic spectral analysis, providing a starting point for researchers who are interested in further processing the data. The extracted dataset is available at `https://airscker.github.io/XDIP`. The software used for extracting numerical points from spectrum plots, *i.e.*, WebPlotDigitizer V4, is open-source and available at `https://github.com/automeris-io/WebPlotDigitizer`.

8 Funding Disclosure

We express our sincere thanks to Rutwik Segireddy and Keying Jia, for their work in assisting our dataset annotation. The work was supported in part by the Physical Biosciences Program and the Photochemistry and Biochemistry group within the US Department of Energy (DOE), Office of Science, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences and Biosciences (KC030402).

9 Competing interests

The authors declare no competing interests.

References

[1] Thanit Tangcharoen, Wantana Klysubun, Chanapa Kongmark, and Wisanu Pecharapa. Synchrotron x-ray absorption spectroscopy and magnetic characteristics studies of metal ferrites (metal= ni, mn, cu) synthesized by sol–gel auto-combustion method. *physica status solidi (a)*, 211(8):1903–1911, 2014.

[2] Pankaj Singh Rawat, RC Srivastava, Gagan Dixit, and K Asokan. Structural, functional and magnetic ordering modifications in graphene oxide and graphite by 100 mev gold ion irradiation. *Vacuum*, 182:109700, 2020.

[3] Thanit Tangcharoen, Wantana Klysubun, and Chanapa Kongmark. Synchrotron x-ray absorption spectroscopy and cation distribution studies of nial2o4, cual2o4, and znal2o4 nanoparticles synthesized by sol-gel auto combustion method. *Journal of Molecular Structure*, 1182:219–229, 2019.

[4] Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.

[5] Jonathan Schmidt, Mário RG Marques, Silvana Botti, and Miguel AL Marques. Recent advances and applications of machine learning in solid-state materials science. *npj computational materials*, 5(1):83, 2019.

[6] Steven B Torrisi, Matthew R Carbone, Brian A Rohr, Joseph H Montoya, Yang Ha, Junko Yano, Santosh K Suram, and Linda Hung. Random forest machine learning models for interpretable x-ray absorption near-edge structure spectrum-property relationships. *npj Computational Materials*, 6(1):109, 2020.

[7] Alexander A Guda, Sergey A Guda, Andrea Martini, AN Kravtsova, Alexander Algasov, Aram Bugaev, Stanislav P Kubrin, LV Guda, Petr Šot, Jeroen A van Bokhoven, et al. Understanding x-ray absorption spectra by means of descriptors and machine learning algorithms. *npj Computational Materials*, 7(1):203, 2021.

[8] Oleg O Kartashov, Andrey V Chernov, Dmitry S Polyanichenko, and Maria A Butakova. Xas data preprocessing of nanocatalysts for machine learning applications. *Materials*, 14(24):7884, 2021.

[9] Boyuan Feng, Yuke Wang, Guoyang Chen, Weifeng Zhang, Yuan Xie, and Yufei Ding. Egemm-tc: accelerating scientific computing on tensor cores with extended precision. In *Proceedings of the 26th ACM SIGPLAN symposium on principles and practice of parallel programming*, pages 278–291, 2021.

[10] International X ray Absorption Society. X-ray absorption data library.

[11] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

[12] Fei Zhou, Matteo Cococcioni, Chris A Marianetti, Dane Morgan, and G Ceder. First-principles prediction of redox potentials in transition-metal compounds with lda+ u. *Physical Review B*, 70(23):235121, 2004.

[13] L Wang, T Maxisch, and G Ceder. A first-principles approach to studying the thermal stability of oxide cathode materials. *Chemistry of materials*, 19(3):543–552, 2007.

[14] David Waroquiers, Janine George, Matthew Horton, Stephan Schenk, Kristin A Persson, G-M Rignanese, Xavier Gonze, and Geoffroy Hautier. Chemenv: a fast and robust coordination environment identification tool. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 76(4):683–695, 2020.

[15] Matthew Kristofer Horton, Joseph Harold Montoya, Miao Liu, and Kristin Aslaug Persson. High-throughput prediction of the ground-state collinear magnetic order of inorganic materials using density functional theory. *npj Computational Materials*, 5(1):64, 2019.

[16] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[17] Christine Zardecki, Shuchismita Dutta, David S Goodsell, Robert Lowe, Maria Voigt, and Stephen K Burley. Pdb-101: Educational resources supporting molecular explorations through biology and medicine. *Protein Science*, 31(1):129–140, 2022.

[18] Colin R Groom, Ian J Bruno, Matthew P Lightfoot, and Suzanna C Ward. The cambridge structural database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 72(2):171–179, 2016.

[19] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[20] David Weininger, Arthur Weininger, and Joseph L Weininger. Smiles. 2. algorithm for generation of unique smiles notation. *Journal of chemical information and computer sciences*, 29(2):97–101, 1989.

[21] David Weininger. Smiles. 3. depict. graphical depiction of chemical structures. *Journal of chemical information and computer sciences*, 30(3):237–243, 1990.

[22] Gary G Koch. Intraclass correlation coefficient. *Encyclopedia of statistical sciences*, 2004.

[23] Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

[24] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[25] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[26] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[27] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[28] Norman Biggs. *Algebraic graph theory*. Number 67. Cambridge university press, 1993.

A Technical Appendices and Supplementary Material

A.1 Detailed Data Collection and Curation

Literature Search and Selection The data extraction pipeline is schematically overviewed in Figure 1 in the main paper. The first step involved constructing searching keywords to identify relevant literature. The search was conducted using combinations of keywords from two distinct sets: XAS, XANES, XAFS, EXAFS and Metalloprotein, Protein, Enzyme, Fe-iron. Each search term consisted of a pair of words, one from each set, such as "XAS Protein" or "EXAFS Metalloprotein".

These keywords were used to retrieve literature from various publishers, including Springer Nature, AAAS Science, American Chemical Society, Elsevier, Royal Society of Chemistry, Electrochemical Society, Wiley, MDPI, and others. To locate relevant literature, we utilized either the publishers' search APIs or conducted manual searches under copyright permission. After excluding duplicate results, we had a pool of 20,915 articles (Table 5).

Table 5: Number of papers retrieved using different keyword combinations.

Keywords	**XAS**	**XANES**	**XAFS**	**EXAFS**
Metalloprotein	389	401	130	877
Protein	3060	3282	1300	5139
Enzyme	2425	2824	976	4207
Fe-iron	6000	6000	3303	6000
Unioned Overall	20915			

From this pool, we manually selected papers that included both protein structure and the absorption spectra of iron elements. The selection criterion was the explicit mention of the protein and the inclusion of at least one Fe absorption spectrum. Eventually, we obtained 573 articles that met these criteria and were downloaded in PDF format for expert annotation.

Data Annotation and Refinement After gathering the literature, human experts annotated the text and XAS plot information, converting it into digitized data samples. The extracted dataset was then refined by removing low-quality samples, poorly presented XAS and structures, and outdated annotations. This combination of automatic searching and manual data extraction and cleaning ensures the final dataset's quality. Each data sample in our dataset consists of three main parts: (1) The entire or local protein structure (The first-coordination sphere of the element of interest), (2) The protein's corresponding Fe K-edge XAS spectra, and (3) the basic information of papers from which the protein structure and XAS spectrum were derived.

A.2 XAS Extraction and Processing Details

Digitization from Published Plots Extracting XAS data from the retrieved literature is a crucial step in constructing our database. XAS is divided into two regions: the near-edge spectra, also known as X-ray absorption near-edge structure (XANES), and the extended X-ray absorption fine structure (EXAFS), each providing specific information on the element studied. To extract numerical data points from the published XAS plots, we utilized the open-source data annotator WebPlotDigitizer V4. To guarantee the precision and reliability of data extraction, we followed the steps below:

- Image Preparation: Screenshots of the relevant plots were taken from the selected papers. High-resolution screenshots were used to improve the accuracy of digitization.
- Software Configuration: The WebPlotDigitizer tool was carefully configured to align with the axes and scales of the plots. This involved setting the axis points and calibrating the software to recognize the specific ranges and units used in the plots.
- Data Point Extraction: Using the calibrated software, data points were manually identified and extracted. The software allowed the refinement of point positions to ensure accuracy. The extracted numerical spectra were then well-organized and saved to the database.

In the process of extracting XAS data, we employ a meticulous digitization approach using the WebPlotDigitizer tool. Initially, we manually mask the spectral lines using the pencil tool to ensure

precise alignment. The color recognition distance is set variably at 2, 5, or 10 pixels based on the clarity and overlap of the plotted lines; clearer plots are annotated at a denser resolution of 1 point per 2 pixels, while more complex, fuzzy plots require a broader setting of 1 point per 10 pixels. This flexibility allows us to capture data with high fidelity, respecting the original plot's integrity without interpolation during the initial annotation phase.

Data Standardization and Metadata Annotation Post-annotation, to standardize the spectral data for consistent analysis and comparison, we interpolate all spectra to a uniform length of 100 data points. This is based on the mean distribution of spectra lengths observed across the dataset, where both XANES and EXAFS lengths peak around 100, despite ranging up to 600. In particular, we also preserve our dataset's original extracted data points. This dual provision enables researchers to select between the original or interpolated data depending on their specific analytical needs, thus offering flexibility while minimizing potential artifacts.

For the application of XAS in analysis, calibration energies are critical for accuracy. In our dataset, these energies are manually extracted from the text of each annotated paper. However, not all sources uniformly report this value; some mention using an iron foil standard without giving a specific energy, while others omit it entirely. In our efforts to maintain transparency, we have documented each instance of missing or incomplete information. Out of the dataset, 227 entries lack explicitly reported calibration energies. These cases are clearly labeled in our documentation files, enabling researchers to account for potential inconsistencies.

A.3 Detailed Data Record Format

The dataset developed in this study is hosted on `https://airscker.github.io/XDIP`. The format of the data records is detailed in Figure 2 in the main paper, which illustrates that each data sample consists of three parts: literature metadata, Fe K-edge XAS data, and the protein's local structure containing Fe.

The literature metadata in each data record includes the Digital Object Identifier (DOI), the title of the research paper, and the XAS absorbing element, which is Fe in this work. These metadata serve as a quick reference for identifying the original research papers and their primary findings. Since most papers cover multiple sets of iron K-edge XAS and corresponding structures, we have stacked multiple absorption spectra and protein structures to capture all relevant details in the literature.

The protein structure includes atom types, their spatial coordinates, and their relationships. We extracted the protein's structure or Fe-element neighborhood local structure through two methods: manual annotation of structure information and accessing data from public protein structure databases. The manually extracted protein (local) structures were derived through the following steps:

- Utilizing the adjacency matrix[28]: This matrix indicates the chemical bonds among different atoms within the extracted structure. The adjacency matrix is crucial for constructing the molecular graph when using computer algorithms to process the protein structure. It reflects the relationships among different atoms.
- Atom list and order: All atoms within the extracted local structure are included. The order of atoms in the atom list corresponds to the order of rows/columns in the adjacency matrix, where each element represents an atom.
- Cartesian coordinates: The atoms' Cartesian coordinates are presented in a list, with the order of coordinates corresponding to the atom list. If there are N atoms in the extracted structure, the shape of the atom list is $(N, \cdot)$, and the coordinate list's shape is $(N, 3)$, where 3 indicates the X, Y, Z dimensions of the atom coordinates.
- Bond lengths: All available bond lengths among different atoms within the local structure are included. The bond lengths are presented as a matrix with the same shape as the adjacency matrix. To construct the bond length matrix, we replace every non-zero element in the adjacency matrix with the corresponding bond length value.
- Bond angles: Some papers may present bond angles, which are crucial for accurately capturing the local structure, especially when atom coordinates are missing.
- Notes: This section includes expert-labeled information, such as identifying isomers of specific molecules.

Table 6: Format of each data record: description, key label, mandatory status, and data type.

Data Description	Data Key Label	Mandatory	Data Type
DOI of the original paper	DOI	Yes	string
Title of the original paper	Title	Yes	string
Absorbing element of XAS	Absorbing Element	Yes	string
XANES data points extracted from the paper	XANES	At least one of them desired	list of float numbers
EXAFS data points extracted from the paper	EXAFS		list of float numbers
Adjacent matrix of all atoms within the extracted (local) protein structure	Adjacent Matrix	Optional only if SMILES exists or publicly available. Otherwise must be provided.	matrix of 0/1 integers
The list of the atoms within the extracted structure	Atom List		list of strings
The list of Cartesian coordinates of atoms	Atom Coordinates		list of float numbers
The matrix of bond lengths between chemically bonded atoms	Bond Lengths		matrix of float numbers
The list of bond angles among neighboring bonds	Bond Angles		dictionary (key = angle vertex, value = float)
Notes about the extracted structure	Notes	Optional	string
The file path of the locally saved protein structure or its online link	PDB/CCDC Path	Optional only if the structure can be manually extracted. Otherwise, only one of them is desired.	string
The SMILES representation of the protein	SMILES		string

In contrast to manually extracted structures, extraction from open databases is more straightforward. In these cases, we simply include the URL for downloading the structure, the local path to the saved files, or the protein's SMILES representation. Table 4 provides a detailed description of the data format, including every key label, whether it is mandatory, and its data type.

A.4 Documentation and Transparency Protocol

The protein local structures and their corresponding XAS spectra have been carefully annotated and documented in two spreadsheets. One of them contains comprehensive details about each protein, including Protein Data Bank (PDB) IDs, chemical names, and structural information extracted from the scientific literature. These details are listed in Table 7. This thorough documentation supports the integrity and reproducibility of the dataset. Figure S1a presents an example table of protein structure annotation documentation, illustrating the documentation details and key points.

Another spreadsheet records the processes involved in extracting iron XAS spectra. It highlights various challenges encountered, such as missing axes on figures and ambiguities in figure interpretation. Such detailed records are crucial for enabling future researchers to understand and evaluate the decision-making processes affecting data inclusion or exclusion. Notably, we include and distinguish between EXAFS data represented in both k-space and r-space. This enhancement ensures that researchers can access and utilize the specific type of data they require for their analyses. We have meticulously labeled each dataset entry in our documentation files to reflect this categorization,

Table 7: Description of protein database documentation elements

Title		The title of the research article from which the protein structure or XAS data is derived.
URL		The direct URL to the source article for reference.
DOI		The Digital Object Identifier for the source article, ensuring precise location of the sources.
	Standards	Reference to any standard samples or controls used in the studies.
XANES	Ligation	Details of the ligands or binding groups associated with the protein structures.
	Sample	Describes the sample material as detailed in the source.
EXAFS		Describes the sample material as detailed in the source.
SMILES		SMILES notation providing a textual representation of the chemical structure.
PDB		Protein Data Bank ID, a unique identifier for the protein structure as deposited in the PDB.
CCDC		Cambridge Crystallographic Data Centre number, providing a reference to crystallographic data related to the protein.

Table 8: Description of XAS spectrum database documentation elements

Title	The title of the source article where the XAS data is described.
Images	Index of figures within the article that pertain to XAS data.
Comments	Notes or comments about the quality or issues identified in the spectral or structural data, such as missing axes or unclear lines.
EXAFS Type	Specifies the type of EXAFS data presented, whether it is raw signal data or Fourier transformed data.
DOI	Digital Object Identifier for each article, ensuring traceability back to the source.
URL	Direct URLs to the articles for quick access.

addressing a critical need for clarity. The details of this documentation are explained in Table 8, and Figure S1b shows an example of an XAS spectrum data documentation log.

Data entries within our datasets were cross-verified against original publications to ensure accuracy and reliability. Spectral and structural data were, wherever feasible, compared against established databases to ensure consistency. Discrepancies and significant findings are carefully documented. The methodologies employed for data extraction are thoroughly detailed, including the tools used, such as the WebPlotDigitizer for converting plot images to numerical data. Each dataset entry is accompanied by extensive metadata such as detailed URLs of original publications, covering all aspects from data extraction to final dataset compilation. This approach not only enhances reproducibility but also facilitates the integration of our data with other datasets. It ensures a clear understanding of the context and specifics of the data collected.

Title		Characteristics of the Isu1 C-terminus in relation to [2Fe-2S] cluster assembly and ISCU Myopathy
Url		https://doi.org/10.1007/s00775-022-01964-1
DOI		10.1007/s00775-022-01964-1
XANES	**Standards**	···
	Ligation	···
	Sample	2Fe-Isu1, 2Fe-ΔC10, 2Fe-ΔC17, and 2Fe-ΔC22
EXAFS		2Fe-Isu1, 2Fe-ΔC10, 2Fe-ΔC17, and 2Fe-ΔC22
SMILES		···
PDB		6NZU
CCDC		···

(a) Example of protein structure data documentation

Title	An in situ XAS study of ferric iron hydrolysis and precipitation in the presence of perchlorate, nitrate, chloride and sulfate
Images	4,7
Comments	Each figure has a large amount of colors
Calibration Energy	Energies were calibrated against the first inflection point of the iron foil at 7112 eV.
EXAFS Type	FT Applied
DOI	10.1016/j.gca.2016.01.021
Url	https://linkinghub.elsevier.com/retrieve/pii/S0016703716000478

(b) Example of XAS spectrum data documentation

Figure S1: Examples of protein structure and XAS spectrum data documentation. (a) This example presents a data entry from our database, including the paper's title, DOI, and URL. To accurately document the data, we included sections for both XANES and EXAFS to help identify the spectrum type corresponding to the protein structure. Due to missing descriptions in some papers, certain annotation elements, such as standards (which refer to the standard sample used) and ligation details (which provide information on ligands or binding groups associated with the protein structures), are omitted. Despite the missing information, we listed the samples mentioned in the paper that have both spectra and structures, with these key points shown in the sample section under the XANES and EXAFS parts. As outlined in the Data Records section, fields such as SMILES, PDB, and CCDC are selectively filled. (b) This table provides an example of an XAS spectrum data annotation log. We recorded the title of the literature, the location of the iron spectrum plot, its DOI, and URL. In addition to these key details, we included comments on the quality or issues identified in the spectral or structural data, such as missing axes or unclear lines. Furthermore, to ensure the accuracy and usability of our data, we highlighted the type of EXAFS spectrum extracted from the literature, noting whether the article presented Fourier-transformed EXAFS or not.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The dataset we have constructed was reflected in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of this work is discussed in Section 6

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All theoretical results are provided with full set of assumptions and complete proofs in Section 2, 3 and 4.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper fully discloses all the information needed to reproduce the main experimental results in Section 2 and Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All code and data links are available at Section 7.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All extraction method details are given in Section 2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The statistical analysis of data extraction process is available at Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All assets used in this paper are properly credited and their usage license and terms of use are properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The dataset constructed in this paper is well documented and documentation is available through Section 7.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.