Chemical Science

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: Y. Wang, S. Zhao, M. Luo, H. Zeng, Y. Feng, D. Liu, Y. Huang and J. Jiang, *Chem. Sci.*, 2026, DOI: 10.1039/D6SC00651E.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/chemical-science

View Article Online
DOI: 10.1039/D6SC00651E

Spectroscopy-Informed XANES–PXRD Framework for Multi-Property Prediction and Structure Inference

Yang Wang[1,#], Siyuan Zhao[1,#], Man Luo[1], Hantao Zeng[1], Yi Feng[1], Daobin Liu[1]*, Yan Huang[1]*, Jun Jiang[1]*

[1]State Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei, Anhui 230026, China

[#]Y. W. and S. Z. contributed equally to this work.

*Correspondence: ldbin@ustc.edu.cn; hyan@ustc.edu.cn; jiangj1@ustc.edu.cn

Abstract: Accelerating functional material characterization and rational design faces a fundamental circular bottleneck in which all mainstream computational screening methods rely entirely on a priori crystallographic knowledge, information inherently unavailable for novel, uncharacterized samples. To bypass this bottleneck, we utilize spectroscopy as a predictive driver. While X-ray absorption near-edge structure (XANES) captures element-specific local states, powder X-ray diffraction (PXRD) resolves global long-range order. Here, we integrate XANES and PXRD into a unified spectroscopic representation to address the challenge of structure inference from spectral data. Trained on 34,929 inorganic compounds with over 100,000 simulated spectra, our framework jointly predicts key physical properties (e.g., band gap, magnetism, density) while inferring oxidation states, coordination numbers, and crystal systems. Notably, a composition-aware partial-measurement strategy utilizing only transition-metal edges matches the accuracy of all-element models, significantly reducing experimental burden. Interpretability analyses reveal that transition-metal and non-transition-metal features cooperatively encode the correlations linking local electronic motifs to global symmetry. Crucially, the framework goes beyond property mapping to reconstruct charge-balanced formulas and retrieve structural templates from existing databases, enabling structure mining without prior structural knowledge. This approach achieves a top-1 accuracy exceeding 0.80 across binary to quinary systems, validated experimentally on eight representative samples including single-phase compounds and heterogeneous composites. These results establish spectroscopy as a quantitative, interpretable medium for decoding structure-property relationships, offering a practical pathway for spectroscopy-informed materials characterization and structure mining.

Introduction

Accelerating functional material characterization and rational design is fundamental to tracing the origin of material properties and addressing critical challenges in energy and environmental applications.[1-4] However, efficiently navigating the immense chemical space remains a formidable bottleneck. Traditional screening paradigms predominantly rely on structure-based approaches, utilizing first-principles calculations or machine learning models that require precise crystallographic information as input.[5-8] This reliance creates a circular dependency in the exploration of novel materials: accurate property predictions necessitate a

View Article Online
DOI: 10.1039/D6SC00651E

priori structural knowledge, yet such information is inherently unavailable for new, uncharacterized samples. Consequently, there is an urgent need for a high-throughput medium that can bypass this structural bottleneck. To break this deadlock, we propose to utilize spectroscopic signals not merely as diagnostic tools for post-synthesis characterization, but as predictive drivers for property prediction and structure inference. Spectroscopic data, such as X-ray absorption or diffraction patterns, serve as unique fingerprints that encode the intrinsic coupling between electronic states and atomic arrangements.[9-11] By decoding these signals directly, we can establish a spectroscopy-driven structure inference workflow, transforming spectra into high-throughput search queries that bridge the gap between raw experimental signals and the physical understanding required for functional design.

Among various spectroscopic techniques, XANES spectroscopy serves as a crucial window into the intrinsic properties of materials owing to its element selectivity and high sensitivity to local electronic states.[12, 13] The systematic variations in edge energy and white-line intensity directly reflect the density of unoccupied states and the distribution of electronic states near the Fermi level, providing quantitative insights into band gap evolution and carrier characteristics.[14-16] The pre-edge and near-edge features are highly sensitive to crystal-field splitting, exchange interactions, and spin-state transitions, thereby elucidating the formation mechanisms of magnetic ordering and spin polarization. Meanwhile, multiple scattering processes in the near-edge region respond to changes in bond lengths, bond angles, and metal-ligand interaction strength, mapping the stability of local configurations and the thermodynamic trends governed by the competition between enthalpy and entropy. In this way, XANES not only traces the evolution of valence states and coordination geometries but also encodes, within its spectral features, essential information about key physical properties such as band structure, magnetism, and stability.[17-19] However, its information content remains primarily confined to the local scale and cannot independently resolve global structural attributes such as long-range lattice periodicity, space-group symmetry, phase separation, or superstructure modulation, thus precluding the reconstruction of a complete crystallographic picture from XANES alone.

To overcome the limitations of XANES in capturing long-range order and global symmetry, PXRD provides a crucial channel for accessing comprehensive crystallographic information. PXRD directly resolves the space group, lattice parameters, and phase composition of a material, quantifies long-range ordering and microstrain, and thereby establishes its symmetry constraints and structural classification.[20-24] As such, XANES and PXRD are intrinsically complementary in their informational hierarchy.[25-28] The former probes local electronic states and coordination geometries, while the latter anchors global crystallographic systems, lattice constants, and phase relationships. Joint analysis of these two modalities enables bidirectional mapping between spectra and both structure and properties within a unified spectroscopic representation space. This integration not only supports multitask prediction of key properties such as band structure, magnetic ordering, and thermodynamic stability but also allows structural inference under the combined constraints of global symmetry and local coordination. Together they lay the foundation for constructing a unified framework that links structure, electronic states, and material functionality.

Despite the highly complementary nature of XANES and PXRD, few machine learning studies have fully integrated these two modalities for joint property prediction and structure inference. Most existing efforts focus on single-spectroscopy inputs. For instance, a recent

View Article Online
DOI: 10.1039/D6SC00651E

diffusion-model-based approach developed by Guo et al. enables ab initio structure solution directly from nanocrystalline powder diffraction patterns, yet this method relies exclusively on PXRD data and does not incorporate element-specific local electronic and coordination information from XANES.[29] Other multimodal attempts are either limited to narrow material systems, lack the ability to simultaneously predict functional properties and infer multiscale structural descriptors, or require full-element spectroscopic measurements that impose heavy experimental burdens.

In this study, we propose an integrated framework that deeply fuses XANES and PXRD to quantitatively link spectral features with material properties within a unified representation space (Figure 1). To resolve the circular dependency of structure-based screening, the framework synergistically combines the local electronic structure reflected by transition metal XANES with the global crystallographic symmetry and long-range order revealed by PXRD, thereby constructing a generalizable spectrum-crystal joint representation. Building on this, we develop a Spectral Fusion Network (SpecFusionNet) to predict key physical and structural information including band gap, magnetism, formation energy, Fermi level, density, coordination number, and oxidation state, and employ a CNN-Transformer architecture for high-accuracy crystal system identification. This multi-task architecture allows us to bypass a priori structural inputs, driving prediction directly from experimental fingerprints. These multi-scale structural and property predictions serve as the foundation for downstream structure inference.

Completing the workflow, the inferred structural priors enable constraint-based screening of existing databases to retrieve candidate structures, thereby facilitating structure mining in the absence of prior structural knowledge. Ablation and attribution analyses further delineate the critical information regions, quantitatively elucidating how local electronic motifs and global crystallographic symmetry act cooperatively to govern material functionality. Importantly, the framework exhibits excellent cross-domain generalization across large-scale theoretical datasets and diverse experimental systems, ranging from single-phase compounds to heterogeneous composites and a composition-aware partial-measurement strategy utilizing only transition-metal edges achieves comparable accuracy to all-element models, significantly reducing experimental burden. This demonstrates that even partial spectroscopic inputs can effectively disentangle dominant electronic features within complex mixed-phase environments, thereby encoding multiscale structural hierarchies. In doing so, this work establishes spectroscopy as a predictive, quantitative, and interpretable tool for data-driven materials characterization and structure mining, offering a general route for structural decoding and property inference without reliance on pre-solved structures.

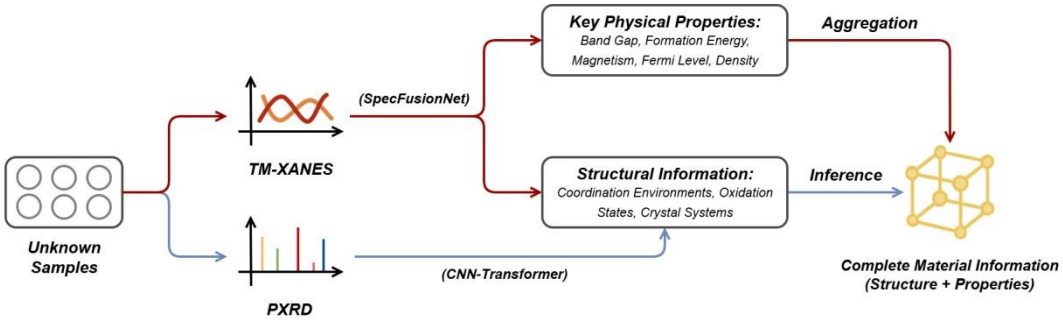

Figure 1. Integrated XANES–PXRD framework for multi property prediction and

View Article Online
DOI: 10.1039/D6SC00651E

structure inference. Unknown samples are described by TM-XANES and PXRD spectra, which can be either simulated or experimentally measured. TM-XANES spectra are encoded by SpecFusionNet to predict physical and local structural information, while PXRD patterns are encoded by a CNN-Transformer to classify crystal systems. The combined information enables structural-template retrieval and inference of complete material structures, providing a unified description of both structure and properties.

Results and discussion

XANES-based physical properties prediction

To establish a comprehensive foundation for data-driven spectroscopic learning, we first constructed a large-scale theoretical dataset that systematically connects chemical composition, crystal structure, and electronic signatures. Specifically, we curated 34,929 inorganic compounds containing two to five elements from the Materials Project[30, 31], and generated over 100,000 K-edge XANES spectra of transition-metal (TM) elements that span a wide chemical landscape encompassing oxides, sulfides, and mixed-anion systems (Figure 2a, Figures S1-S3). An additional dataset of all-element K-edge XANES spectra (including non-transition-metal constituents) was also prepared to construct a baseline model for performance comparison. Each XANES spectrum was preprocessed into 200 uniform intensity points within a 56 eV edge-centered window, and its fine structures were featurized via multi-scale convolution. For elemental featurization, we used fundamental atomic descriptors including atomic number, electronegativity, atomic radius, ionization energies, and element type. TM and non-TM elements are encoded separately.

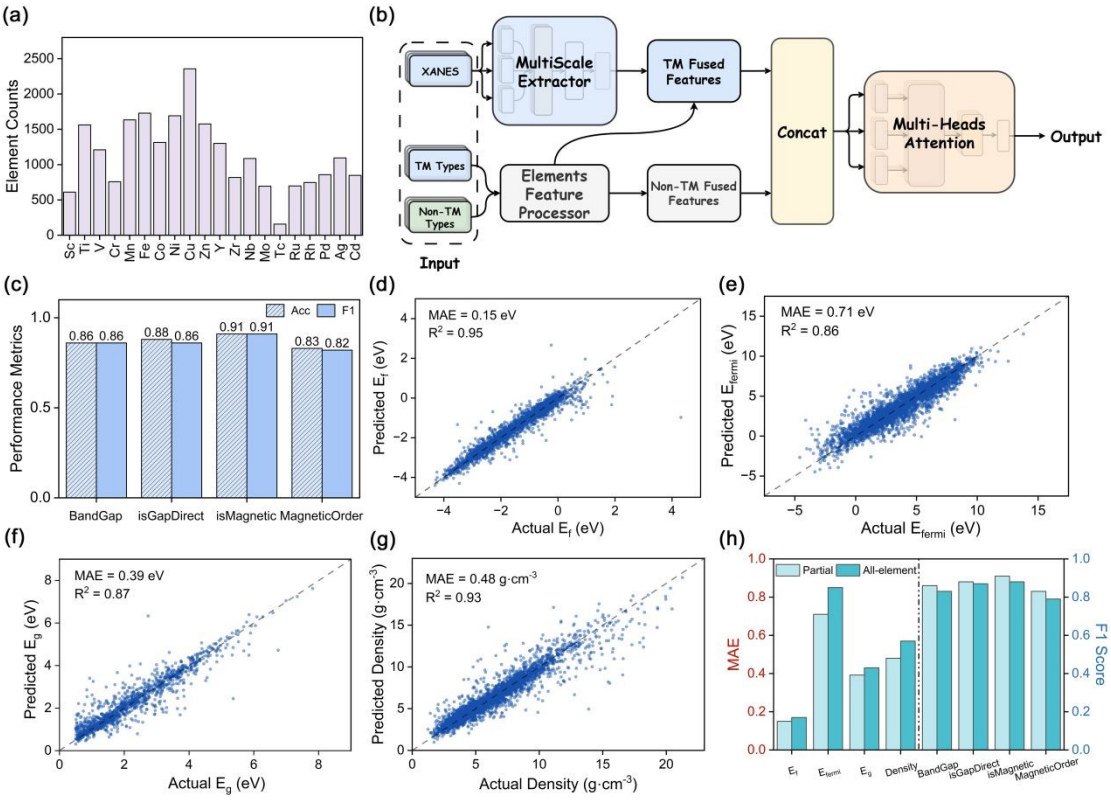

Figure 2. Prediction of physical properties using the XANES-based SpecFusionNet model. **(a)** Distribution of TM elements in the dataset, showing the frequency of occurrence for each

View Article Online
DOI: 10.1039/D6SC00651E

species. **(b)** Architecture of the SpecFusionNet model. A multi-scale extractor captures fine-grained features from TM XANES spectra, while an element-feature processor encodes both TM and non-TM information. The fused representations are concatenated and passed through a multi-head attention module to generate property predictions. **(c)** Classification performance across four representative tasks: band gap presence (BandGap), direct vs. indirect gap (isGapDirect), magnetic presence (isMagnetic), and magnetic ordering (MagneticOrder), evaluated by accuracy and F1 score. **(d-g)** Parity plots of predicted versus actual values for four regression targets (E_f, E_{fermi}, E_g, and density), with mean absolute error (MAE) and coefficients of determination (R^2) annotated. **(h)** Comparison between models trained with all-element spectra and those using a partial-spectra input strategy, evaluated by MAE (left axis) and F1 score (right axis) across different prediction tasks.

To interpret these inputs, we developed the Spectral Fusion Network model, which serves as the XANES encoder within the framework (Figure 2b). The model incorporates a multi-scale convolutional extractor that captures near-edge fine structures such as pre-edge and main-edge transitions at different resolutions (Figure S4). In parallel, an element-feature encoder processes both TM and non-TM elemental information (Figure S5). The spectral and compositional embeddings are then fused through a multi-head attention mechanism that adaptively weights each feature according to its relevance to the target property, enabling the model to learn physically interpretable, cross-scale correlations between local spectral fingerprints and physical properties.

To rigorously evaluate the predictive capacity of SpecFusionNet, we trained the model across a suite of classification and regression tasks encompassing diverse physical properties. The classification tasks encompassed the identification of band gap presence, direct versus indirect band gap, magnetism detection, and magnetic ordering, while the regression tasks targeted key quantitative descriptors including formation energy (E_f), Fermi level (E_{fermi}), band gap (E_g), and density. These four classification labels and four regression targets span chemically diverse and moderately imbalanced distributions in band-gap character, magnetism, formation energy, Fermi level, band-gap magnitude, and density, thereby providing a stringent and representative test bed for evaluating SpecFusionNet (Figures S6 and S7). Across all classification tasks, SpecFusionNet exhibits robust and transferable performance, with both accuracy and macro-F1 scores exceeding 0.82 (Figure 2c, Figure S8). For regression tasks, the model maintains low prediction errors, with mean absolute deviations below 0.71 eV for energetic quantities and 0.60 g cm^{-3} for density (Figures 2d-2g), underscoring its quantitative reliability. These results collectively confirm that transition-metal-centered XANES spectra encapsulate rich local electronic fingerprints that are sufficiently informative to reconstruct a broad spectrum of physical properties across chemically and structurally diverse materials, thereby establishing spectroscopy-driven learning as a powerful paradigm for universal materials representation and property inference.

We further benchmarked this configuration against an all-element baseline model that incorporated XANES spectra for every constituent element. The comparison revealed only marginal differences, with the mean absolute error (MAE) for band gap prediction increasing by merely 0.05 eV and classification F1 scores fluctuating by less than 6% (Figure 2h). These results indicate that omitting non-transition-metal edges does not compromise predictive accuracy and may even enhance robustness by suppressing noise arising from low-information

View Article Online
DOI: 10.1039/D6SC00651E

spectra. To further assess the generality of this strategy, we expanded the dataset to include K-edge XANES of alkaline-earth metals as well as $L_{2,3}$-edge XANES of 5d transition metals and lanthanides. SpecFusionNet maintained consistently high predictive performance across these extended chemical systems (Figures S9-S12), confirming that the composition-aware partial-measurement strategy provides a reliable and transferable foundation for quantitatively linking spectroscopic features with diverse physical properties.

Interpretability of physical properties prediction

To elucidate the role of localized spectroscopic information in physical properties prediction and to probe the internal attribution mechanisms of the SpecFusionNet, we conducted a series of interpretability analyses. Model performance was evaluated under two ablation settings in which either the TM-fusion or non-TM-fusion matrix was masked. In each case, the corresponding feature vectors were replaced with zeros while preserving tensor dimensions and parameter counts, thereby maintaining constant network complexity. Both ablations resulted in a marked decline in predictive accuracy (Figures 3a, Figure S13), revealing that transition-metal and non-transition-metal components act cooperatively to encode the key spectral-property correlations underpinning the predicted electronic and structural behaviors.

We further performed ablation analyses on the element-embedding features for both TM and non-TM components, focusing on key atomic descriptors including electronegativity (χ), atomic radius (r_{atom}), ionization energies (IE_1, IE_2, IE_3), atomic number (Z), and element type (Type). Sequential removal of these descriptors elucidated their distinct physical contributions. Of these descriptors, Z exerted the most pervasive influence, as its elimination caused a pronounced increase in MAE across all prediction tasks, underscoring the fundamental role of nuclear charge and electronic configuration in determining physical properties. The Type descriptor most strongly affected density prediction, consistent with its association with atomic packing motifs and bonding topology. Ionization energies were particularly important for band-structure-related quantities, while χ and r_{atom} contributed moderately yet significantly to properties governed by local coordination geometry and bond strength (Figures 3b-3c, Figure S14). Collectively, these results quantitatively confirm that the learned elemental embeddings encode intrinsic chemical periodicity and its coupling with local spectral information, thereby enhancing the physical transparency and interpretability of the model.

To visualize how the model extracts property-specific spectral signatures, Grad-weighted class activation mapping (Grad-CAM) was applied to the final convolutional layers. For each spectrum, attention intensities were integrated over pre-edge, edge, and post-edge regions to identify the most influential energy intervals. In Y_2TiO_5 (mp-17559), both Ti and Y K-edges contributed substantially to the prediction of physical properties, with the Ti edge exhibiting enhanced pre-edge and main-edge activations that reflect the sensitivity of Ti 3d-O 2p hybridization to band-structure variations (Figure 3d). The Y XANES spectrum shows a distinct activation pattern compared to Ti, and when predicting density, Grad-CAM highlights the Y K-edge almost exclusively, emphasizing the dominance of Y^{3+} coordination in determining lattice packing and overall cell volume (Figure 3e). This trend is consistent with the crystal structure of Y_2TiO_5 (Figure 3f). This task-dependent shift of spectral attention illustrates that the model adapts its learned representation to the physical origin of each property,

View Article Online
DOI: 10.1039/D6SC00651E

with local orbital transitions governing band-gap estimation and heavy-atom coordination dominating density inference.

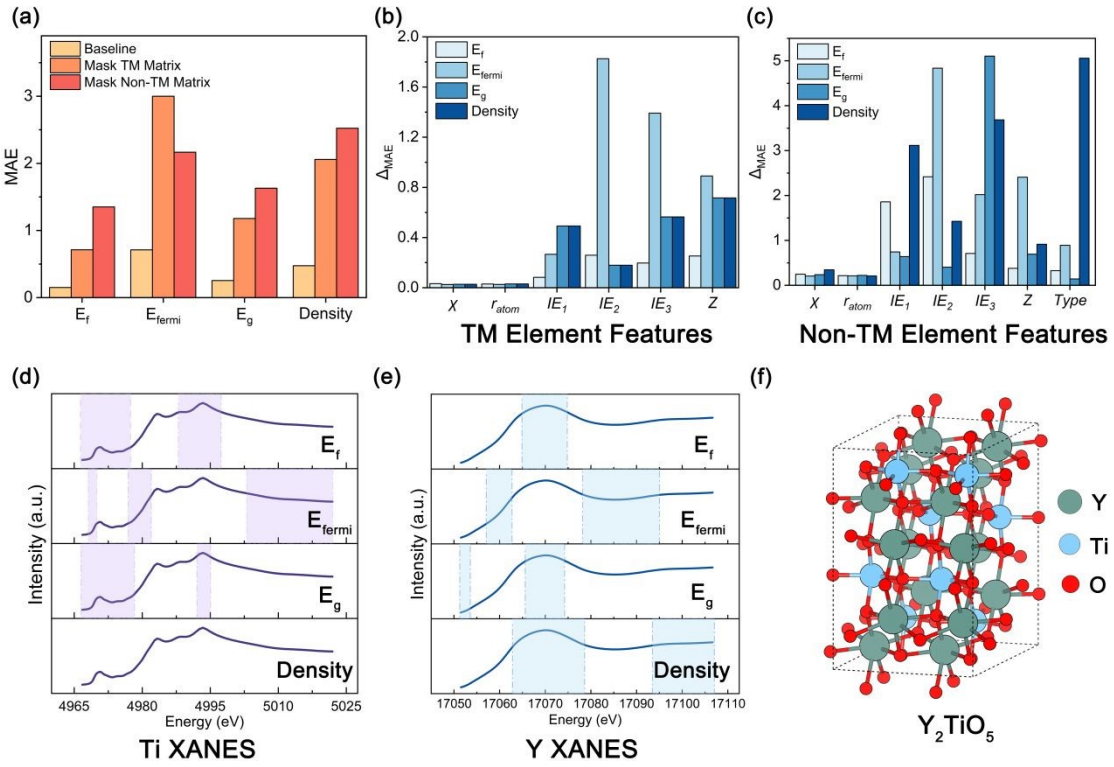

Figure 3. Interpretability and feature attribution in physical properties prediction. (a) Ablation analysis shows the effect of masking different element-fusion matrices on four regression targets. Bars compare the MAE of the baseline model with those obtained when the fusion matrices corresponding to TM or non-TM elements are masked. **(b, c)** Contribution of atomic descriptors for TM (**b**) and non-TM (**c**) elements to the four regression targets. Each bar represents the change in MAE (ΔMAE) relative to the baseline when the corresponding descriptor is removed. **(d, e)** Grad-CAM visualizations highlighting the importance of specific energy regions in the XANES spectra of Ti (**d**) and Y (**e**) in Y_2TiO_5 (mp-17559) for different regression tasks. The shaded areas denote spectral regions with normalized Grad-CAM values > 0.5, corresponding to the most influential features. (**f**) Ball and stick representation of the Y_2TiO_5 crystal structure, with Y, Ti, and O atoms shown in blue, green, and red, respectively.

Interestingly, this element-specific dominance is not universal across all Y-Ti systems. A similar pattern was observed in $Y_2Ti_2O_7$ and $YTiO_3$, where both Y and Ti edges contributed prominently to the predicted density. Grad-CAM maps revealed comparable activation across the Ti pre-edge and Y main-edge regions, implying cooperative spectral influence between the two cations (Figure S15 and S16). This behavior is consistent with the underlying crystal chemistry, as Ti primarily occupies octahedral sites whose connectivity modulates framework packing, while Y^{3+} coordination expands cell volume through its larger ionic radius.

Consequently, the model dynamically balances attention across multiple edges rather than privileging a single element, confirming that the learned XANES representation is chemically grounded and that element-resolved spectral features are quantitatively linked to macroscopic functionality.

View Article Online
DOI: 10.1039/D6SC00651E

Multiscale structural descriptors from XANES and PXRD

To extend our analysis from physical properties prediction to structural descriptors, we evaluated macro-F1 scores for oxidation-state and coordination-number classification using standardized XANES spectra for all transition-metal elements (Figure S17). Most transition metals exhibit macro-F1 scores above 0.90 for both tasks, reflecting the pronounced sensitivity of near-edge features to variations in d-orbital occupancy and ligand-field geometry. The most discriminative features for oxidation-state prediction occur near the absorption edge, where the energy shift correlates with effective nuclear charge and covalency, whereas coordination-number recognition depends primarily on pre-edge and near-edge oscillations associated with crystal-field splitting. For main-group cations which lack d states, small but systematic edge shifts still enable reliable identification of coordination environments, particularly in distorted or non-octahedral sites (Figures 4a-4b). These results demonstrate that the approach generalizes beyond transition metals, revealing that near-edge spectral signatures carry chemically transferable descriptors of local structure across diverse chemistries.

Whereas XANES offers element-specific sensitivity to local coordination environments, PXRD encodes the long-range crystallographic symmetry and lattice periodicity that cannot be resolved from near-edge spectra alone. To assess how reliably such global symmetry information can be extracted under realistic measurement conditions, we first convolved the simulated PXRD patterns with a Voigt function to emulate instrumental peak broadening, then added controlled Gaussian white noise to mimic counting statistics and background fluctuations (Figure S18). The resulting dataset spans seven representative crystal systems (Figure 4c) and was used to train a CNN-Transformer classifier based solely on PXRD inputs (Figure S19). Despite the presence of noise, the model attains an overall crystal-system classification accuracy of 0.81 (Figure 4d). High-symmetry systems such as cubic and hexagonal are predicted with accuracies above 0.90, reflecting their relatively simple and well-separated diffraction signatures. By contrast, orthorhombic, monoclinic and triclinic systems reach more modest accuracies of 0.70–0.78, as their peak-rich patterns exhibit substantial overlap and subtle angular splittings (Figure S20). The reduced performance for these low-symmetry systems arises from the intrinsic similarity of their diffraction profiles, in which weak symmetry-diagnostic reflections and minor intensity modulations are easily masked by the imposed noise, rendering symmetry-dependent distinctions challenging even in this idealized learning scenario.

To disentangle the roles of local and global spectroscopic information, we benchmarked models trained on XANES-only, PXRD-only, and fused XANES–PXRD inputs. PXRD delivers the highest accuracy (0.81), whereas fusion reaches 0.66 and XANES alone yields 0.48 (Figure S21). The inferior accuracy of the fused model stems from the exclusive dependence of crystal-system classification on long-range symmetry information, which is uniquely encoded in PXRD patterns; the incorporation of XANES input introduces extraneous local structural descriptors that distort the identification of symmetry-diagnostic diffraction features. This hierarchy highlights the intrinsic complementarity of the two spectroscopic modalities: PXRD encodes lattice periodicity and global long-range symmetry, whereas XANES probes element-specific local coordination environments and oxidation-state variations. This performance distinction justifies our framework's design, which assigns crystal-system classification explicitly to the PXRD encoder to maximize accuracy. By integrating these

View Article Online
DOI: 10.1039/D6SC00651E

complementary predictions in the downstream workflow, the framework establishes spectroscopy as a multiscale structural probe and provides a robust foundation for the subsequent structure-inference module.

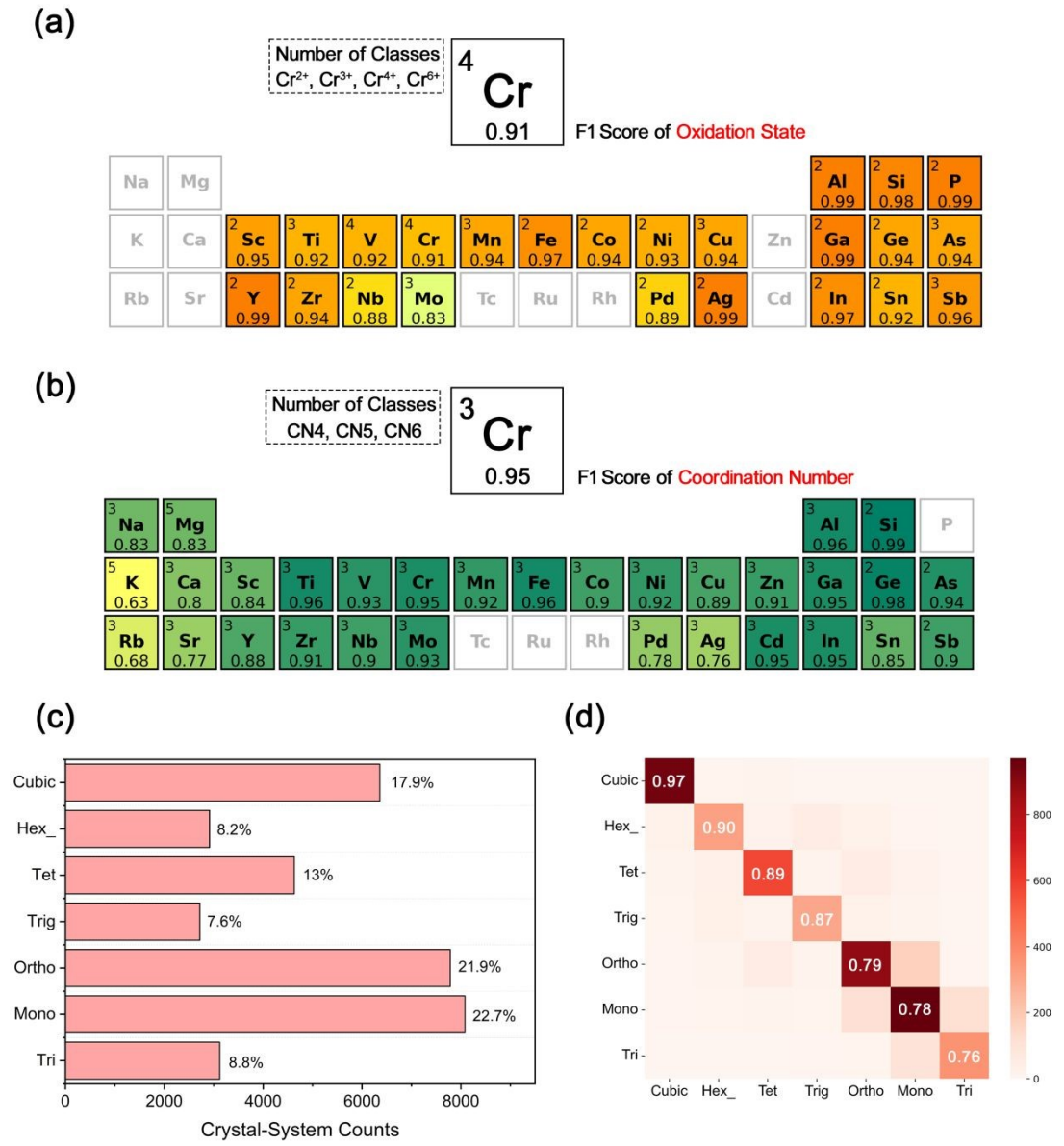

Figure 4. Prediction of local and global structural information from XANES and PXRD. **(a)** Element-wise oxidation-state classification based on XANES spectra. Each colored cell denotes an element with the number of oxidation states considered (upper left), the element symbol (center), and the corresponding macro-F1 score (bottom). Grey cells indicate elements with insufficient training data for reliable evaluation. **(b)** Element-wise coordination-number classification from XANES spectra. Each colored cell shows the number of coordination categories (e.g., CN4-CN6), the element symbol, and the corresponding macro-F1 score. Grey cells mark elements without sufficient samples for quantitative evaluation. **(c)** Crystal system count distribution across the dataset, expressed as absolute counts of materials per crystal system. **(d)** Confusion matrix for crystal-system classification using PXRD-only inputs, yielding an overall accuracy of 0.81 across seven crystal systems. Darker diagonal cells correspond to higher correct-classification counts.

View Article Online
DOI: 10.1039/D6SC00651E

Structure inference and framework validation

Building on the multi-scale structural descriptors extracted from XANES and PXRD, the structure-inference module of our framework is designed as a conditional data-mining and structural-template retrieval tool to match the most plausible crystal structures under spectroscopic constraints.

As shown in Figure 5a, the module takes three core predicted descriptors as input constraints: oxidation states and coordination numbers from the XANES encoder, and crystal system classification from the PXRD encoder, which together define the chemically feasible compositional and symmetry space for candidate screening. Within this constrained space, we first enumerate charge-balanced stoichiometries following integer-valence rules. We then retrieve representative crystallographic prototypes matching the predicted composition type and symmetry constraints from the full Materials Project inorganic crystal structure database (the module is also compatible with other mainstream crystallographic databases such as ICSD). Subsequently, elemental substitutions guided by coordination environment similarity and ionic radius compatibility are performed on the retrieved templates to generate chemically plausible candidate structures. Finally, all candidates are geometrically relaxed, and ranked by the consistency between their simulated spectra and the experimental input spectra, to identify the atomic configuration that best matches the experimentally inferred electronic and crystallographic signatures. Full details of the retrieval rules and workflow are provided in Sections S2.2 and S2.3 of the Supporting Information.

To evaluate the reliability of the inference module, we benchmarked its performance using theoretical spectra derived from first-principles calculations on the Materials Project dataset, with all evaluations conducted on its held-out theoretical subset. Across binary to quinary systems, the combined top-1 accuracy for simultaneous empirical-formula and crystal-system prediction exceeded 0.80 (Figure 5b), indicating that the spectroscopic constraints are sufficient to recover both composition and symmetry with high fidelity. A representative example, Ca_2MnAlO_5, illustrates the entire reconstruction workflow. Under charge-balance and coordination constraints, two chemically plausible formulas, $CaMnAlO_4$ and Ca_2MnAlO_5, were first enumerated (Figure S22). For each candidate, representative structural models were generated and their PXRD patterns and XANES spectra were simulated. Direct comparison with the reference PXRD and XANES spectra unambiguously selected Ca_2MnAlO_5 as the correct solution (Figure S23 and S24). Additional evaluations on three compounds excluded from the training set confirmed comparable reconstruction accuracy, demonstrating that the framework does not overfit to the training distribution and can be transferred to previously unseen chemical systems (Figure S25). These results show that the inference module can reliably translate spectral information into chemically and structurally consistent atomic configurations.

The structure-inference module was further validated using seven experimentally synthesized compounds. Their XANES spectra were processed using identical normalization and interpolation protocols as those applied to theoretical data to ensure consistent spectral treatment across modalities (Figures 5c-5d, Figures S26-S31). Across all samples, the experimental spectra closely reproduce the simulated XANES profiles in terms of edge positions, peak intensities, and fine-structure oscillations, demonstrating that the simulation

View Article Online
DOI: 10.1039/D6SC00651E

pipeline faithfully captures the underlying electronic and coordination environments. Complementary PXRD measurements for the same samples reveal similarly high agreement between predicted and experimental diffraction patterns, further corroborating the structural assignments (Figure S32). These results confirm that the integrated spectroscopic-inference framework generalizes from simulated to laboratory data with minimal loss of fidelity, enabling robust structure reconstruction under realistic experimental conditions.

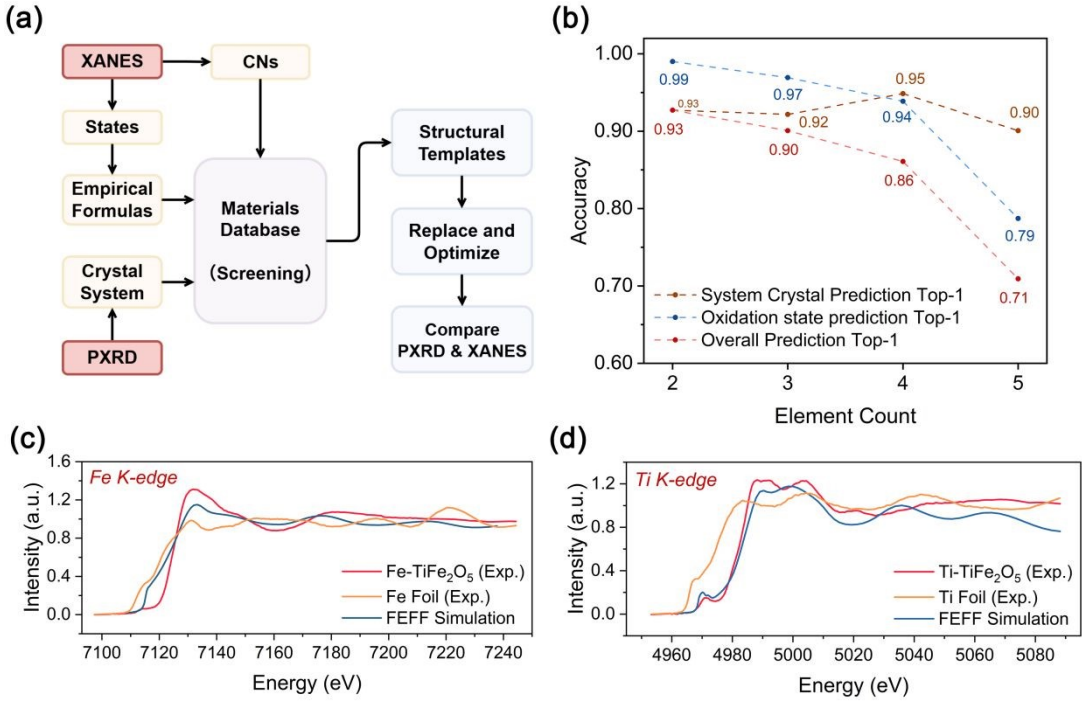

Figure 5. Structure inference and experimental validation of the integrated framework. **(a)** Schematic of the structure-inference stage, integrating predicted oxidation states, coordination numbers, and crystal systems to generate charge-balanced candidates. **(b)** Accuracy of empirical-formula prediction, crystal-system classification, and their combined top-1 performance as a function of element count (two to five). **(c)** and **(d)** Comparison of experimental, reference foil, and FEFF-simulated XANES spectra for $TiFe_2O_5$ at the Fe K-edge (c) and Ti K-edge (d), respectively. These spectra illustrate the cross-modal consistency between theoretical predictions and experimental results; all experimental spectra were background-subtracted and normalized for direct comparison.

To examine performance beyond structural concordance, we additionally evaluated property-level fidelity across the seven single-phase experimental validation samples. Band gaps and magnetic responses were measured by UV-Vis diffuse reflectance and magnetic hysteresis loops, respectively (Figures S33-S40), while other quantities were compared with theoretical references. The sample-wise prediction performance, including the consistency of structural classification tasks and the absolute error of quantitative property predictions, is summarized in Table 1. For structural classification tasks, the framework achieves 100% accuracy in crystal system prediction across all seven samples, with 6 out of 7 samples showing fully consistent oxidation state and coordination number predictions compared with experimental crystallographic results. For quantitative property prediction, the absolute errors are naturally larger than those on theoretical benchmarks due to domain shifts between simulated and experimental spectra. Specifically, band gap absolute errors range from 0.39 to

View Article Online
DOI: 10.1039/D6SC00651E

1.27 eV, density errors from 0.16 to 1.47 g cm^{-3}, and formation energy errors remain below 0.24 eV for all samples. Despite the increased numerical errors, the model correctly identifies the trend of band gaps across the seven compounds and consistently distinguishes insulating from metallic behavior. Moreover, the formation energy predictions maintain a remarkably low absolute error (< 0.24 eV) for all experimental samples, indicating that the learned spectral–energy mapping is robust against experimental noise. The full raw predicted and reference values for all properties are provided in Table S1 of the Supporting Information. A direct comparison between experimental and simulated XANES spectra shows that larger prediction errors correlate with greater spectral deviations, underscoring the sensitivity of the framework to experimental spectral fidelity.

The only sample with partial inconsistency in oxidation state and coordination number assignments is Co_3O_4, a spinel-structured compound with two non-equivalent Co sites: tetrahedrally coordinated Co^{2+} and octahedrally coordinated Co^{3+}. Its Co K-edge XANES spectrum is an inherent superposition of signals from these two mixed-valence, distinct coordination environments, which increases the difficulty of resolving two separate sets of local structural descriptors from a single, convoluted spectrum. This result further validates the high sensitivity of our model to element-specific local coordination and electronic states, and we have noted this specific challenge for multi-site mixed-valence systems as a key direction for future method optimization.

Table 1. Prediction performance on seven experimental compounds

Compound	E_f	E_{fermi}	E_g	Density	Valence & CN	Crystal System
$TiFe_2O_5$	0.08	1.06	0.72	0.53	✓	✓
MnS	0.03	1.30	1.27	0.16	✓	✓
$ZnFe_2O_4$	0.10	0.65	0.64	0.97	✓	✓
$CoFe_2O_4$	0.02	1.26	0.50	1.37	✓	✓
V_2O_5	0.23	1.06	0.52	0.45	✓	✓
Co_2O_3	0.21	1.04	0.39	1.47	✓	✓
Co_3O_4	0.24	0.72	0.42	1.21	✗	✓

Note: Values shown as the absolute error between predicted and experimental/reference values.

Units: formation energy (E_f), Fermi level (E_{fermi}), and band gap (E_g) in eV; density in g cm^{-3}. ✓ = predicted structural descriptors fully consistent with experimental crystallographic results; ✗ = partial inconsistency.

To further validate the generalization capability of the framework in heterogeneous systems, we investigated a Cu_2O/TiO_2 composite photocatalyst. While the material was fully characterized experimentally (PXRD and UV-Vis spectra are shown in Figures S41-S44), we utilized solely the experimental XANES spectra (Figure 6a) as input to probe the electronic structure. The model successfully identified the electronic dominance of the wide-bandgap TiO_2 component within the composite. As shown in Figure 6b, the predicted band gap (3.67 eV) exhibits excellent agreement with the experimental trend derived from UV-Vis measurements, proving that the model can extract critical electronic features even from the convoluted spectral signals of mixtures. This successful reconstruction underscores the framework's ability to bridge the gap between spectral diagnostics and physical inference, providing a robust tool for analyzing realistic, multi-component materials.

View Article Online
DOI: 10.1039/D6SC00651E

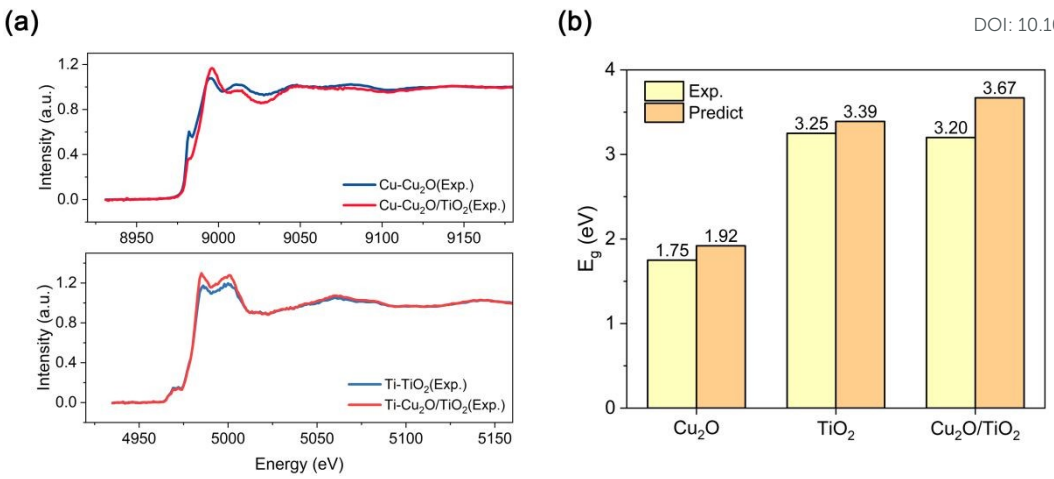

Figure 6. Experimental validation on a Cu_2O/TiO_2 heterogeneous composite. (a) Normalized experimental XANES spectra at the Cu K-edge (blue) and Ti K-edge (red) of the Cu_2O/TiO_2 composite. The Ti K-edge spectrum is dominated by features characteristic of TiO_2, while the Cu K-edge shows contributions from Cu_2O. **(b)** Comparison of the band gap predicted by the framework with the experimental value derived from UV-Vis diffuse reflectance spectroscopy (Tauc plot method).

Conclusions

This work establishes a unified XANES–PXRD framework that quantitatively links multi-scale spectroscopic features to both material properties and atomic structures. By integrating element-specific local electronic information from XANES with the global crystallographic symmetry captured by PXRD, the framework constructs a coherent representation that bridges the electronic and crystallographic domains. Within this representation, the model predicts diverse physical properties and structural descriptors, demonstrating the capability to disentangle dominant electronic signatures in heterogeneous systems, while a structure-inference module reconstructs charge-balanced formulas and retrieves structural templates under spectroscopic constraints. The framework achieves robust performance across theoretical and experimental domains, demonstrating accurate reconstruction of both electronic properties and crystal structures. While the current model exhibits high fidelity for broad material classes, we acknowledge challenges in differentiating complex low-symmetry systems or heavily disordered phases, highlighting areas for future algorithmic refinement.

Attribution and ablation analyses reveal that transition-metal and non-transition-metal features cooperatively encode the correlations that couple local electronic motifs with long-range symmetry, highlighting the physical interpretability of the model. These results suggest that spectroscopy can serve as a predictive, quantitative, and interpretable medium for decoding structure-property relationships, offering a practical framework for data-driven candidate screening and structure inference. Beyond the present results, the unified spectroscopic representation can be further extended to model task-specific properties across diverse materials domains. Crucially, by enabling the rapid identification of functional phases within complex or uncharacterized streams—such as industrial waste or low-value byproducts—this framework offers a tangible pathway for materials valorization, creating new opportunities for design from a circular-economy perspective.

View Article Online
DOI: 10.1039/D6SC00651E

Methods

Spectral preprocessing. To ensure consistent data quality, each XANES spectrum was interpolated within a 56 eV window centered at the absorption edge, yielding 200 uniformly spaced intensity points. Spectra were normalized to unit height and screened to remove artifacts or non-physical baselines. For PXRD, theoretical diffraction patterns were calculated over 10-90° (2θ) using relaxed crystal structures. The patterns were first broadened via convolution with a Voigt function (combining Gaussian and Lorentzian components) to simulate instrumental broadening, then resampled to 800 points. Subsequently, controlled Gaussian white noise was added to mimic experimental counting statistics and background fluctuations. This preprocessing procedure ensures that both XANES and PXRD datasets are standardized and noise-aware before model training.

Model architecture. The integrated framework comprises two complementary encoders. The XANES encoder (SpecFusionNet) employs multiscale convolutional branches (kernel sizes = 3, 5, 7) to capture pre-edge, main-edge, and post-edge features at different resolutions. Elemental information, including atomic number, electronegativity, atomic radius, ionization energies, and categorical element type, is transformed through embedding layers and fused with spectral features via a multi-head attention module. This design allows both transition-metal and non-transition-metal elements to contribute chemical context, even when only partial XANES measurements are available. A masking mechanism enables compounds containing variable numbers of elements to be processed without introducing artifacts. The PXRD encoder adopts a CNN-Transformer architecture to capture both local peak correlations and long-range symmetry information. The two encoders jointly generate hierarchical representations that connect local electronic environments with global crystallographic order. Detailed layer configurations and hyperparameters are provided in the Supporting Information.

Experimental details. PXRD patterns were recorded on a Rigaku SmartLab multifunctional rotating-anode diffractometer using Cu Kα radiation (λ = 1.54184 Å) operated at 45 kV and 200 mA. Data were collected in a continuous mode over a 2θ range of 10° to 90°, with a step size corresponding to a scanning speed of 10° min^{-1}. X-ray absorption near-edge structure (XANES) spectroscopy was performed on a laboratory-based spectrometer (RapidXAFS 2 M, Anhui Absorption Spectroscopy Analysis Instrument Co., Ltd.) equipped with a Mo target source operated at 20 kV and 20 mA. Energy selection was achieved using a spherically bent crystal analyzer (SBCA, 500 mm radius of curvature) chosen specifically for each absorption edge. Room-temperature magnetic hysteresis loops were measured using a Magnetic Property Measurement System (MPMS3, Quantum Design, USA). The magnetic field was swept from +50,000 Oe to -50,000 Oe and back to complete the loop. Solid-state UV-Vis-NIR spectra were collected with a Shimadzu SOLID3700 spectrophotometer equipped with an integrating sphere for diffuse reflectance measurements over the 200–800 nm wavelength range. The diffuse reflectance data (R) were converted to the Kubelka–Munk function $F(R) = (1-R)^2/(2R)$. The Tauc plot was then obtained by plotting $[F(R)h\nu]^2$ versus $h\nu$, from which the band gap was determined by extrapolating the linear portion of the curve to the energy axis.

View Article Online
DOI: 10.1039/D6SC00651E

Author contributions

J. J., Y. H., and D. L. conceived and supervised the project. Y. W. performed all data computation, model training, and framework design. S. Z. conducted the experimental measurements. Y. W., J. J., Y. H., and D. L. led the preparation of the manuscript with input from all other authors.

Conflicts of interest

The authors declare no conflict of interest.

Data availability

The code and data underlying the machine learning part of this paper are available in our public repository (https://github.com/WangyLab/XANES-XRD_WorkFlow).

Acknowledgements

The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of Chinese Academy of Science. Y.H. acknowledges the National Natural Science Foundation of China (22303091) and the Fundamental Research Funds for the Central Universities (WK2490250008) and the National Key Research and Development Program of China (2023YFA1508200). J.J. acknowledges the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB0450302) and the National Natural Science Foundation of China (22025304, 22033007) and the CAS Project for Young Scientists in Basic Research (YSBR-005) and the Innovation Program for Quantum Science and Technology (2021ZD0303303). D.L. gratefully acknowledges financial support by the University of Science and Technology of China (USTC) Startup Program (KY9990000209) and the National Key Research and Development Program of China (2024YFA1509500). The numerical calculations in this paper have been done on Hefei advanced computing center.

References

1. D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C. J. Brabec, B. Maruyama, K. A. Persson and A. Aspuru-Guzik, *Nature Reviews Materials*, 2018, **3**, 5-20.
2. M. Cheng, C.-L. Fu, R. Okabe, A. Chotrattanapituk, A. Boonkird, N. T. Hung and M. Li, *Nature Materials*, 2026, DOI: 10.1038/s41563-025-02403-7.
3. Y. Bai, K. Li, N. Han, J. Kim, R. Zhang, S. Mahesh, A. Shayesteh Zeraati, B. R. Sutherland, K. Chow, Y. Liang, S. Hoogland, J. E. Huang, D. Sinton, E. H. Sargent and J. Hattrick-Simpers, *Nature Catalysis*, 2026, DOI: 10.1038/s41929-025-01463-x.
4. J. Sun, D. Li, J. Zou, S. Zhu, C. Xu, Y. Zou, Z. Zhang and H. Lu, *npj Computational*

View Article Online
DOI: 10.1039/D6SC00651E

Materials, 2024, **10**, 181.

5. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547-555.
6. T. Xie and J. C. Grossman, *Physical Review Letters*, 2018, **120**, 145301.
7. C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, *Chemistry of Materials*, 2019, **31**, 3564-3572.
8. V. L. Deringer, M. A. Caro and G. Csányi, *Nature Communications*, 2020, **11**, 5461.
9. S. Wang and J. Jiang, *ACS Catalysis*, 2023, **13**, 7428-7436.
10. T. Yang, D. Zhou, S. Ye, X. Li, H. Li, Y. Feng, Z. Jiang, L. Yang, K. Ye, Y. Shen, S. Jiang, S. Feng, G. Zhang, Y. Huang, S. Wang and J. Jiang, *Journal of the American Chemical Society*, 2023, **145**, 26817-26823.
11. S. Feng, M. Huang, Y. Li, A. Cai, X. Yue, S. Wang, L. Chen, J. Jiang and Y. Luo, *Chemical Society Reviews*, 2025, **54**, 8243-8286.
12. A. Iglesias-Juez, G. L. Chiarello, G. S. Patience and M. O. Guerrero-Pérez, *The Canadian Journal of Chemical Engineering*, 2022, **100**, 3-22.
13. C. T. Chantler, G. Bunker, P. D'Angelo and S. Diaz-Moreno, *Nature Reviews Methods Primers*, 2024, **4**, 89.
14. M. S. Huzan, M. Fix, M. Aramini, P. Bencok, J. F. W. Mosselmans, S. Hayama, F. A. Breitner, L. B. Gee, C. J. Titus, M.-A. Arrio, A. Jesche and M. L. Baker, *Chemical Science*, 2020, **11**, 11801-11810.
15. C. Zheng, C. Chen, Y. Chen and S. P. Ong, *Patterns*, 2020, **1**.
16. A. A. Guda, S. A. Guda, A. Martini, A. N. Kravtsova, A. Algasov, A. Bugaev, S. P. Kubrin, L. V. Guda, P. Šot, J. A. van Bokhoven, C. Copéret and A. V. Soldatov, *npj Computational Materials*, 2021, **7**, 203.
17. Z. Chen, A. G. Walsh and P. Zhang, *Accounts of Chemical Research*, 2024, **57**, 521-532.
18. J. Timoshenko, F. T. Haase, S. Saddeler, M. Rüscher, H. S. Jeon, A. Herzog, U. Hejral, A. Bergmann, S. Schulz and B. Roldan Cuenya, *Journal of the American Chemical Society*, 2023, **145**, 4065-4080.
19. J. Timoshenko and A. I. Frenkel, *ACS Catalysis*, 2019, **9**, 10192-10211.
20. N. J. Szymanski, C. J. Bartel, Y. Zeng, M. Diallo, H. Kim and G. Ceder, *npj Computational Materials*, 2023, **9**, 31.
21. J. Venderley, K. Mallayya, M. Matty, M. Krogstad, J. Ruff, G. Pleiss, V. Kishore, D. Mandrus, D. Phelan, L. Poudel, A. G. Wilson, K. Weinberger, P. Upreti, M. Norman, S. Rosenkranz, R. Osborn and E.-A. Kim, *Proceedings of the National Academy of Sciences*, 2022, **119**, e2109665119.
22. J.-W. Lee, W. B. Park, J. H. Lee, S. P. Singh and K.-S. Sohn, *Nature Communications*, 2020, **11**, 86.
23. J. E. Salgado, S. Lerman, Z. Du, C. Xu and N. Abdolrahim, *npj Computational Materials*, 2023, **9**, 214.
24. N. Wang, X. Zhang, S. Tan, S. Lee and E. Hu, *Chemical Reviews*, 2025, **125**, 9834-9874.
25. T. A. Hamdalla, A. M. Aboraia, V. V. Shapovalov, A. A. Guda, N. V. Kosova, O. A. Podgornova, A. A. A. Darwish, S. A. Al-Ghamdi, S. Alfadhli, A. M. Alatawi and A.

View Article Online
DOI: 10.1039/D6SC00651E

Soldatov, *Scientific Reports*, 2023, **13**, 2169.

26. K. Mukai, T. Nonaka and T. Uyama, *Inorganic Chemistry*, 2020, **59**, 16882-16892.
27. A. Tsoukalou, P. M. Abdala, D. Stoian, X. Huang, M.-G. Willinger, A. Fedorov and C. R. Müller, *Journal of the American Chemical Society*, 2019, **141**, 13497-13505.
28. E. B. Deeva, A. Kurlov, P. M. Abdala, D. Lebedev, S. M. Kim, C. P. Gordon, A. Tsoukalou, A. Fedorov and C. R. Müller, *Chemistry of Materials*, 2019, **31**, 4505-4513.
29. G. Guo, T. L. Saidi, M. W. Terban, M. Valsecchi, S. J. L. Billinge and H. Lipson, *Nature Materials*, 2025, **24**, 1726-1734.
30. M. K. Horton, P. Huck, R. X. Yang, J. M. Munro, S. Dwaraknath, A. M. Ganose, R. S. Kingsbury, M. Wen, J. X. Shen, T. S. Mathis, A. D. Kaplan, K. Berket, J. Riebesell, J. George, A. S. Rosen, E. W. C. Spotte-Smith, M. J. McDermott, O. A. Cohen, A. Dunn, M. C. Kuner, G.-M. Rignanese, G. Petretto, D. Waroquiers, S. M. Griffin, J. B. Neaton, D. C. Chrzan, M. Asta, G. Hautier, S. Cholia, G. Ceder, S. P. Ong, A. Jain and K. A. Persson, *Nature Materials*, 2025, **24**, 1522-1532.
31. C. Zheng, K. Mathew, C. Chen, Y. Chen, H. Tang, A. Dozier, J. J. Kas, F. D. Vila, J. J. Rehr, L. F. J. Piper, K. A. Persson and S. P. Ong, *npj Computational Materials*, 2018, **4**, 12.

View Article Online
DOI: 10.1039/D6SC00651E

Spectroscopy-Informed XANES–PXRD Framework for Multi-Property Prediction and Structure Inference

Yang Wang[1,#], Siyuan Zhao[1,#], Man Luo[1], Hantao Zeng[1], Yi Feng[1], Daobin Liu[1]*, Yan Huang[1]*, Jun Jiang[1]*

[1]State Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei, Anhui 230026, China

[#]Y. W. and S. Z. contributed equally to this work.

*Correspondence: ldbin@ustc.edu.cn; hyan@ustc.edu.cn; jiangj1@ustc.edu.cn

Data Availability Statement

To maintain high standards of transparency and reproducibility, the data and code supporting the findings of this study are available as follows:

- **Source Code and Software:** The complete source code for the XANES-XRD Integrated Machine Learning Workflow is openly available on GitHub at https://github.com/WangyLab/XANES-XRD_WorkFlow. Detailed documentation and usage instructions are provided in the repository.
- **Training Dataset:** The large-scale theoretical dataset of 34,929 inorganic compounds used for model training was derived from the Materials Project. Access to this processed dataset is available via the GitHub repository mentioned above.
- **Experimental Data:** The experimental data supporting the validation results, including XANES spectra, PXRD patterns, and characterization measurements for the seven synthesized compounds, have been provided as part of the Supplementary Information accompanying this article.