Supplementary Information

Xiaoyue He[*1,2], Bowen Hou[*1,2], Hongyu Wu[*2], Jianbo Hu[4], Lu Zhang[4], Lipeng Chen[4], Ziping Wan[2], Siyu Chen[4], Mingjun Xiao[†1,2], Zhicheng Zhong[†2,3], Xin Chen[†2], and Bo Chen[†2]

[1]Department of Computer Science and Technology, University of Science and Technology of China, Hefei, China
[2]Suzhou National Laboratory, Suzhou, China
[3]School of Artificial Intelligence and Data Science, University of Science and Technology of China, Hefei, China.
[4]Zhejiang Laboratory, Hangzhou, China

[*]These authors contributed equally to this work.
[†]Corresponding author.

Contents

Supplementary Note 1
DPA Fine-tuning and Dataset Generation

We fine-tune a pretrained DPA3 model to predict ionic diffusion coefficients from crystal structures. To ensure data efficiency and preserve pretrained physical priors, the graph neural network backbone is frozen during fine-tuning, and only the task-specific multilayer perceptron (MLP) heads are updated.

The model is trained on the 1,203-sample training set. We optimize the model for 10,000 steps using an exponentially decaying learning rate schedule from 1×10^{-4} to 3×10^{-5}. The Smooth MAE loss is adopted with $\beta = 1.0$ to improve robustness against outliers.

Ionic conductivity is derived using the Nernst–Einstein relation:

$$\sigma = \frac{nq^2 D}{k_B T} H \tag{S1}$$

where n denotes the charge carrier concentration, q is the ionic charge, D is the diffusion coefficient, k_B is the Boltzmann constant, T is the absolute temperature, and H is the Haven ratio accounting for ion–ion correlations.

Using this pipeline, we generate a dataset of 1,603 crystal structures with corresponding diffusion coefficients. Among them, 1,203 samples are used for fine-tuning, while the remaining 400 samples are reserved for evaluation.

We apply a base-10 logarithmic transformation to the diffusion coefficients, which empirically improves regression stability. The data generation process is computationally intensive, requiring approximately 1.5–2.5 A800 GPU-hours per sample.

Evaluation on the held-out test set of 400 samples demonstrates strong predictive performance, achieving an R^2 score of 0.998 and a mean absolute error (MAE) of 6.95 on the log-transformed diffusion coefficients.

These results indicate that the fine-tuned DPA model accurately captures structure–transport relationships despite the limited amount of DFT-labeled data.

After fine-tuning, the DPA model is employed as a high-throughput surrogate to generate large-scale ionic conductivity data. Guided by domain experts, we expand the material space through compositional substitution and structural variation, resulting in 79,430 candidate crystal structures.

The fine-tuned model is used to predict logarithmic diffusion coefficients for all generated structures. The resulting predictions exhibit physically plausible distributions consistent with known trends in solid-state ionic conductors.

This large-scale predicted dataset alleviates the scarcity of ionic conductivity labels and serves as a foundation for subsequent multimodal training in LLAM. The proposed pipeline achieves a favorable balance between accuracy and scalability.

Supplementary Figures

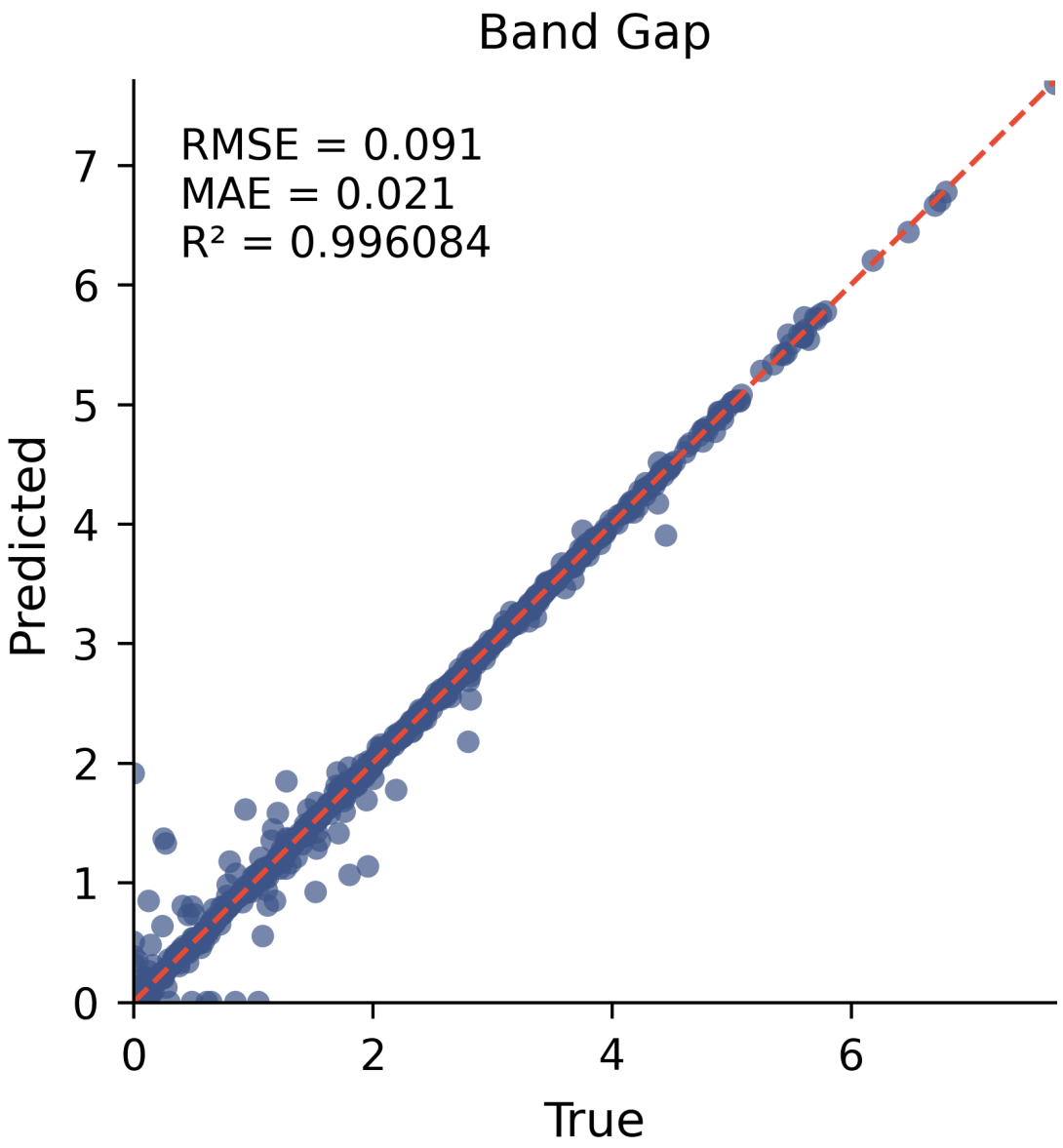

Figure S1: Correlation between true and predicted band gap. Each blue data point corresponds to a unique material structure, and the red dashed line represents the perfect prediction baseline.

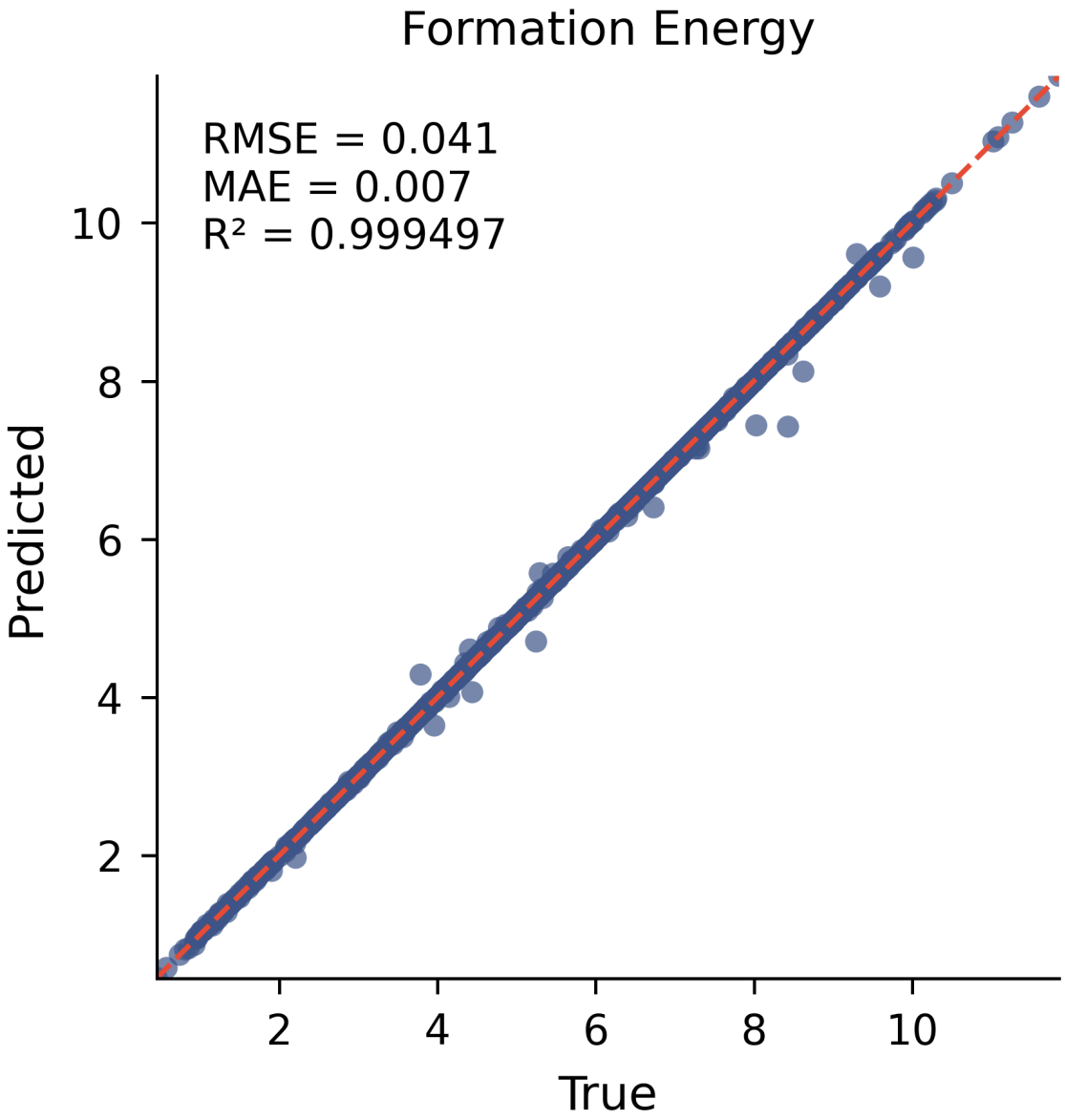

Figure S2: Correlation between true and predicted formation energy. Each blue data point corresponds to a unique material structure, and the red dashed line represents the perfect prediction baseline.

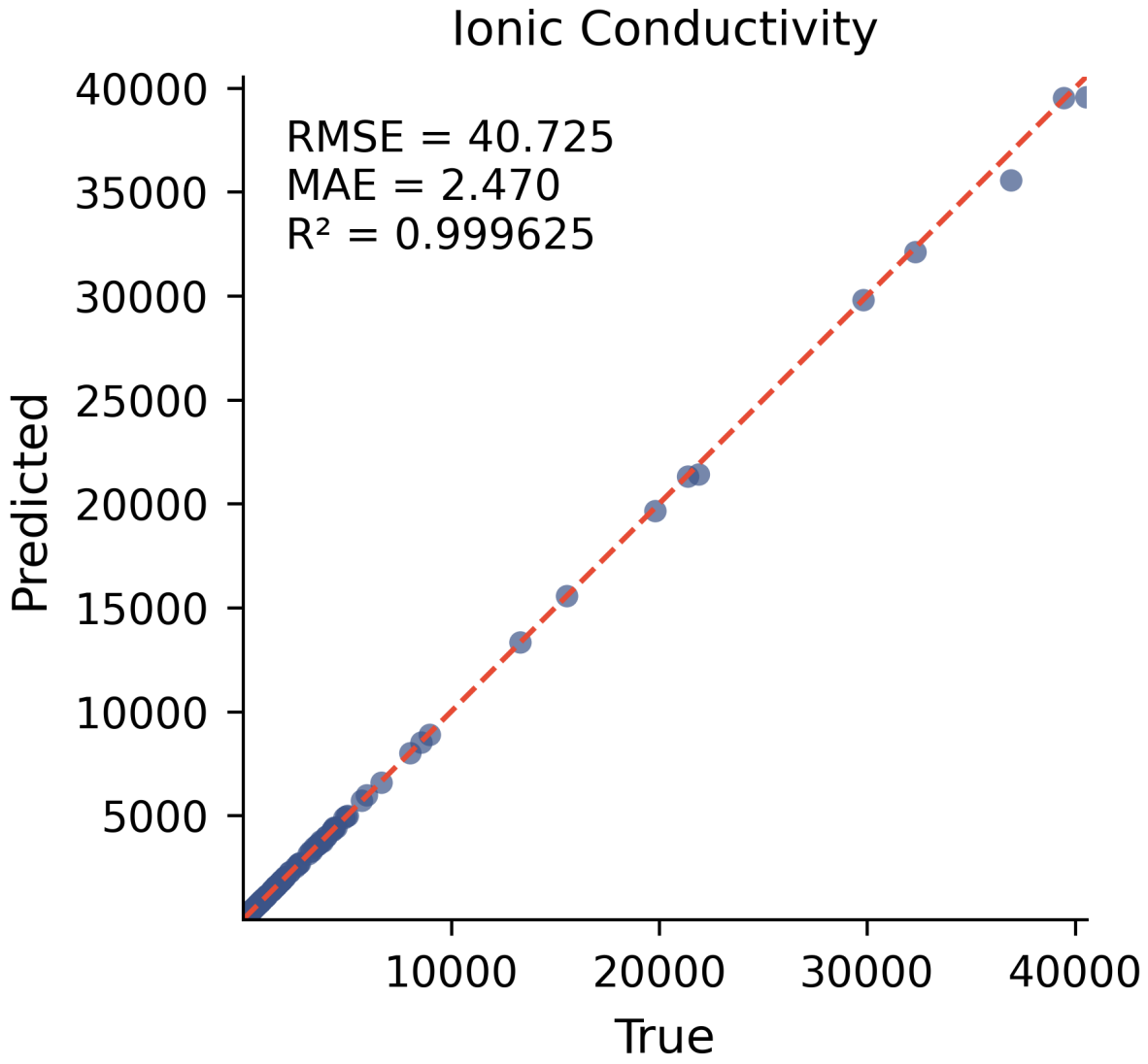

Figure S3: Correlation between true and predicted ionic conductivity. Each blue data point corresponds to a unique material structure, and the red dashed line represents the perfect prediction baseline.

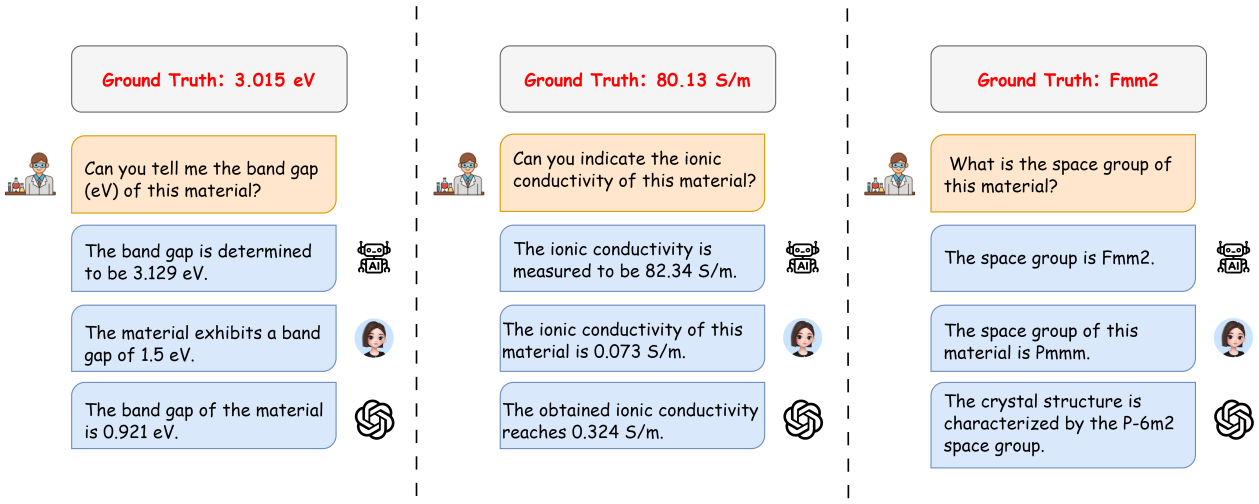

Figure S4: Performance comparison of the proposed LLAM model with baseline models including Doubao and GPT-5 for the prediction of band gap, ionic conductivity, and space group.

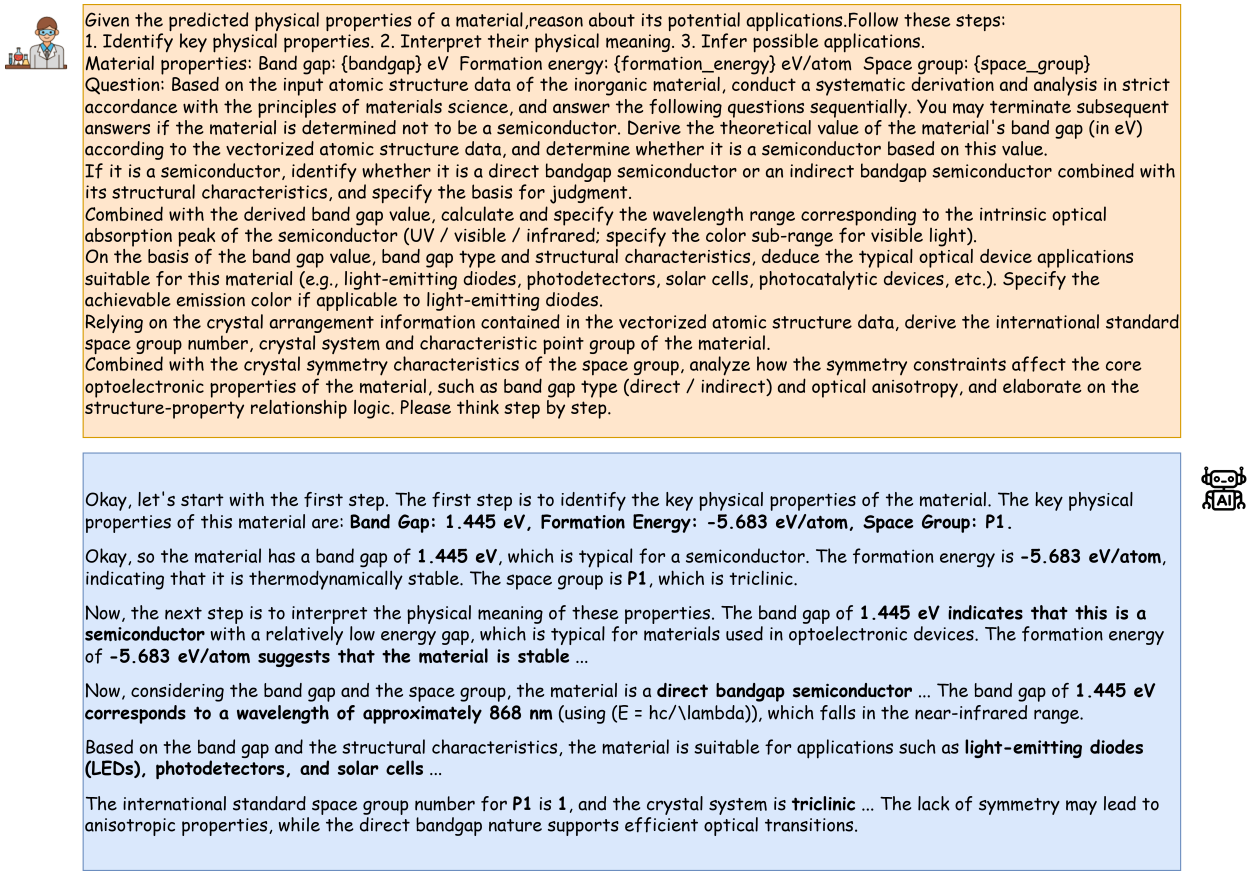

Figure S5: Quantitative reasoning from atomic structure to optoelectronic functionality. The framework processes key material properties (band gap, formation energy, space group, direct/indirect band gap) to classify semiconductor type, derive optical characteristics, and infer practical applications.

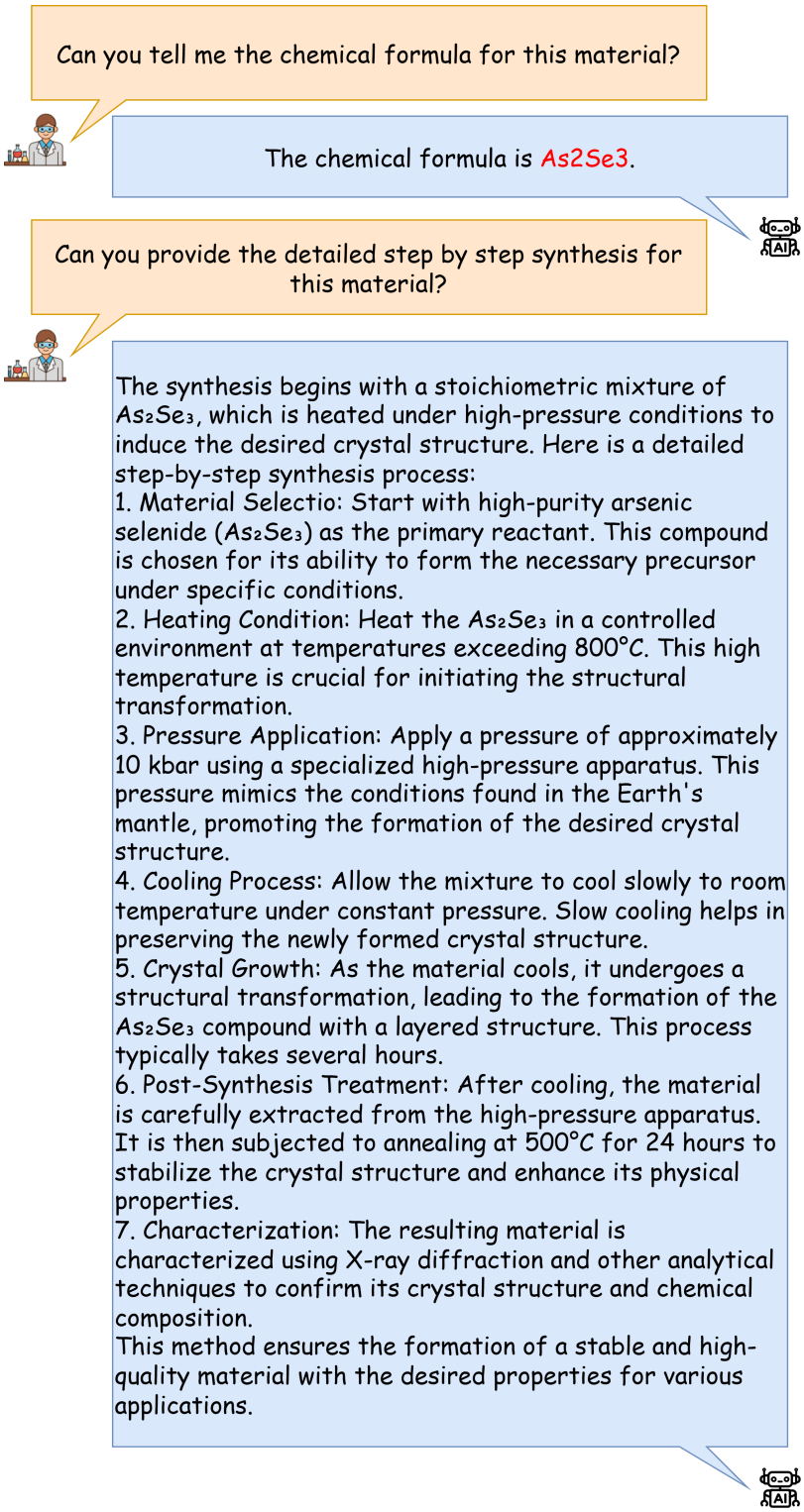

Figure S6: Step-by-step synthesis protocol of As_2Se_3 material. Schematic illustration detailing the chemical formula identification and systematic synthetic procedure of arsenic triselenide (As_2Se_3), covering precursor preparation, high-pressure reaction, crystal growth, and post-synthesis treatment.

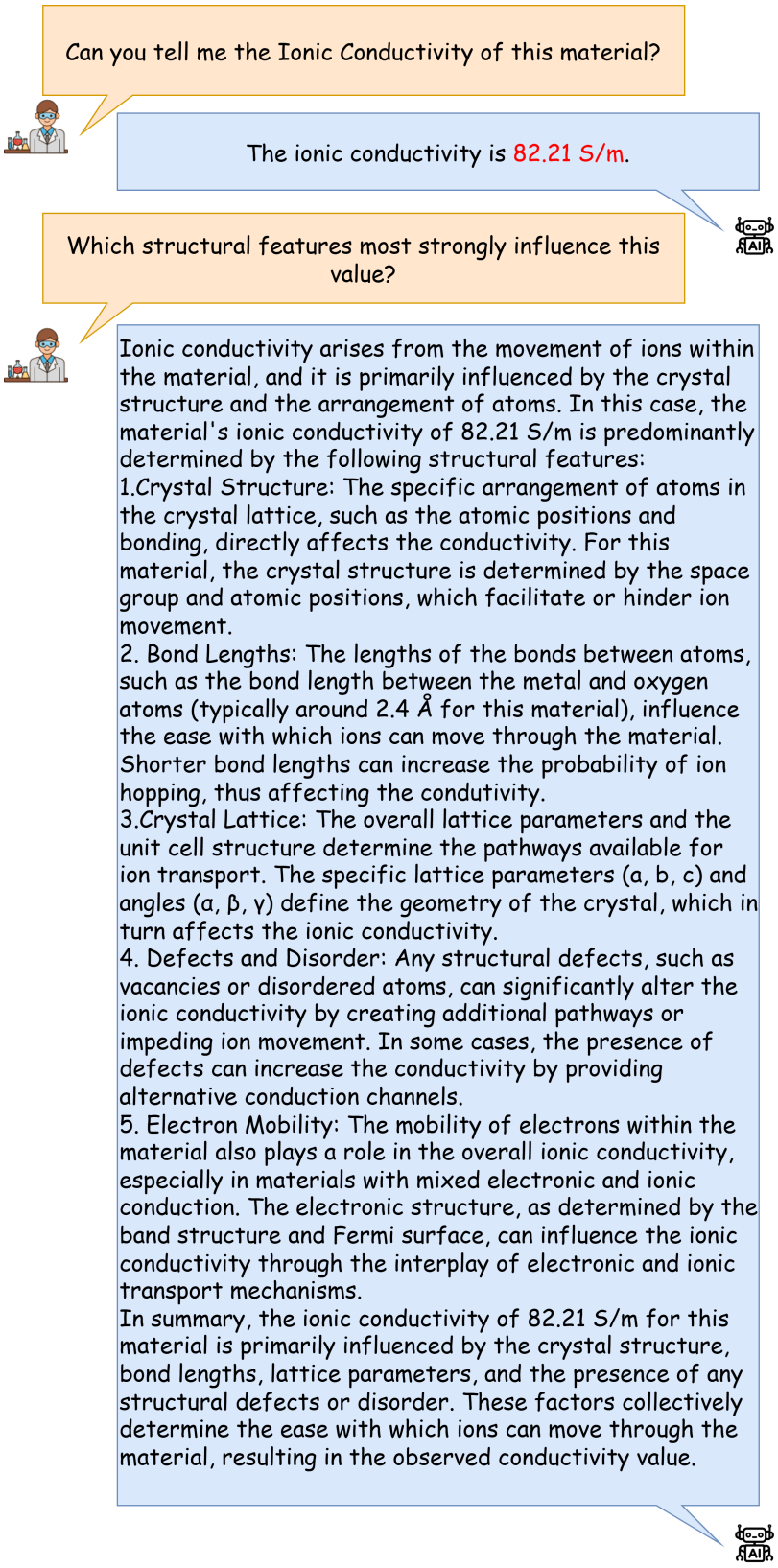

Figure S7: Mechanistic attribution of transport phenomena. Presentation of the measured ionic conductivity (82.21 S/m) and its mechanistic link to the material's crystal structure, bonding, and defect chemistry.

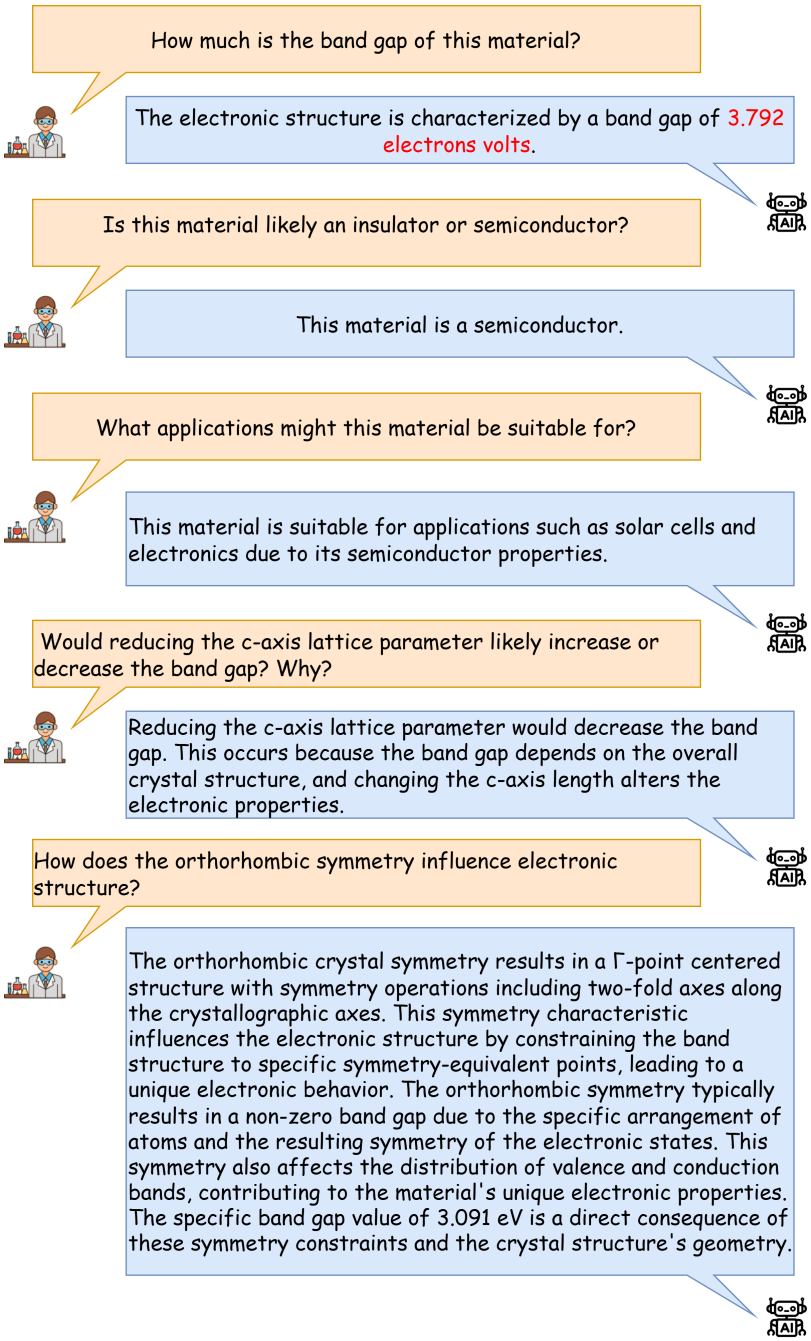

Figure S8: Electronic structure analysis of the semiconducting material. Dialogue-driven analysis of the material's electronic properties, including a band gap of 3.792 eV, classification as a semiconductor, potential applications in solar cells and electronics, the effect of c-axis lattice parameter modulation on band gap, and the influence of orthorhombic crystal symmetry on electronic structure.

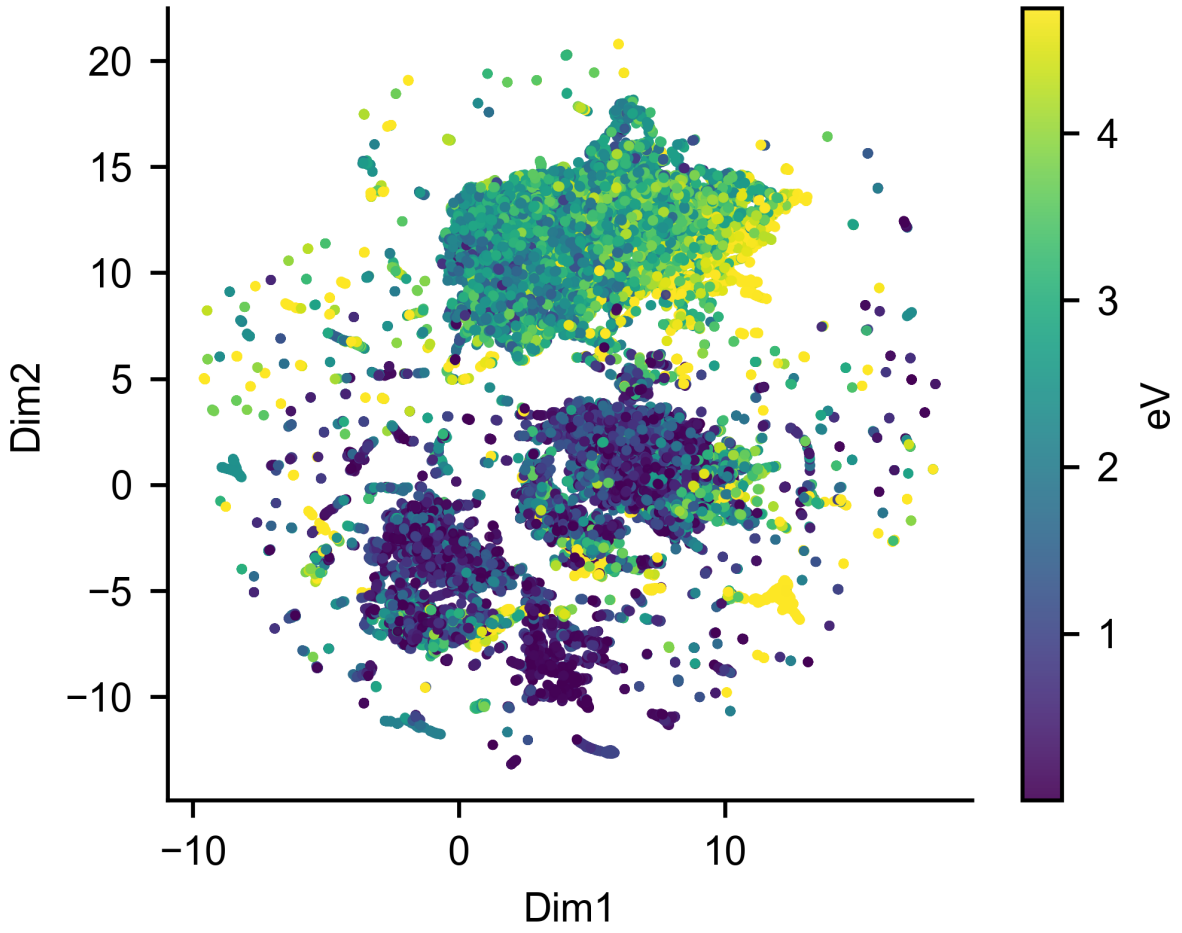

Figure S9: Two-dimensional UMAP projection of the learned atomic embeddings, color-coded by the DFT-calculated band gap (eV) of the corresponding materials.

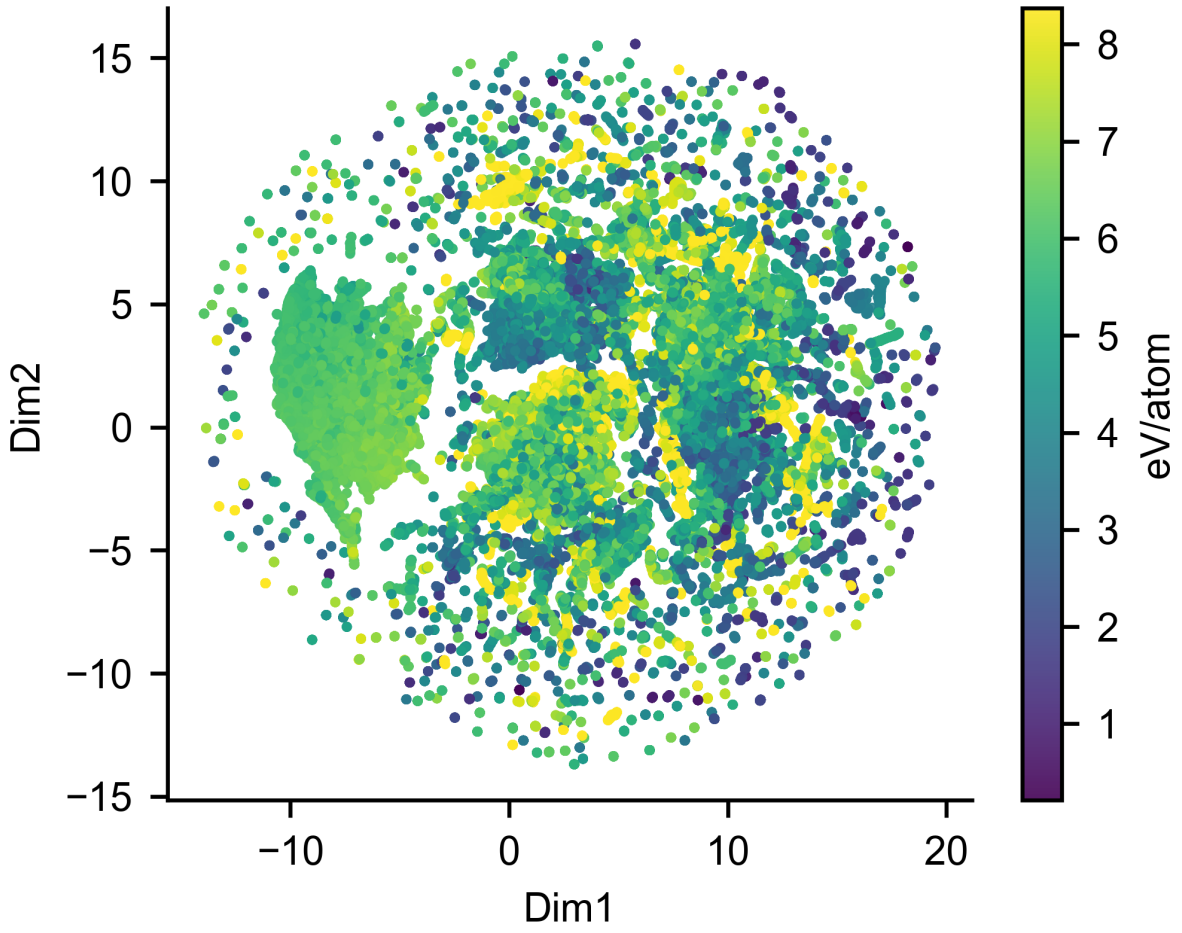

Figure S10: Two-dimensional UMAP projection of the learned atomic embeddings, color-coded by the DFT-calculated formation energy (eV/atom) of the corresponding materials.

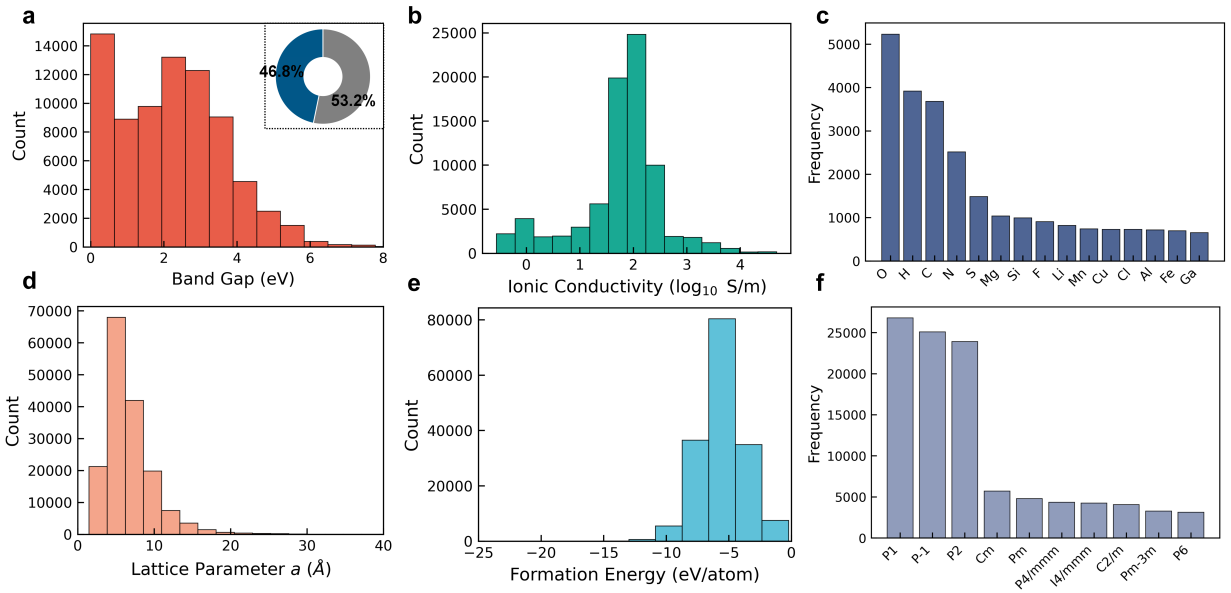

Figure S11: Distribution of the dataset. (a) Band gap, (b) Ionic conductivity, (c) Frequency distribution of elements, (d) Lattice parameter, (e) Formation energy (eV/atom), and (f) the top 10 most frequent space groups.