Seeing without Crystal Structure: Multimodal AI for Materials Characterization

Jithendaraa Subramanian1, Flora Chen1, Daniel Schweigert1, Santosh Suram1, Linda Hung1, **Weike Ye**1,
1Toyota Research Institute, Los Altos, CA, USA

weike.ye@tri.global

Abstract

Crystal-graph representations have been widely used in many large-scale materials models [1], enabling accurate property prediction on computational datasets and large in silico screening, yet in laboratory workflows atomic structures are often unknown or costly to resolve, making structure-dependent representations impractical for experimental discovery; experimentalists instead rely on readily available modalities such as synthesis records and characterization measurements (e.g., spectroscopy and microscopy). Multimodal learning [2] integrates these heterogeneous signals and enables scalable self-supervision via cross-modal alignment to improve accuracy, robustness, and transfer when labels are scarce; however, prior multimodal approaches still typically assume crystal structure as a required input [3,4], limiting experimental applicability. Here we present a structure-free multimodal framework that learns directly from elemental composition and XRD, two widely accessible laboratory modalities, without requiring crystal structure. Our architecture uses modality-specific encoders with cross-attention fusion for joint prediction: a CrabNet-based composition encoder [5] and a Transformer-style XRD encoder over 2θ capturing peak positions, shapes, and relative intensities; the resulting shared representation feeds downstream heads (Figure 1; core architecture and pretraining losses in [6]). We train at scale on the 5M-sample Alexandria dataset and develop self-supervised objectives: masked XRD modeling (MXM), masked composition modeling (MCM), and cross-modal contrastive alignment, where MXM masks 2θ regions to reconstruct missing spectra, MCM masks elements/stoichiometry to predict missing constituents (generalizing to sparse/uncertain compositions and disordered materials), and contrastive learning aligns paired composition–XRD embeddings by cosine similarity. To bridge sim-to-experiment gaps in XRD, we augment simulated patterns with experimental artifacts (grain-size broadening, preferred orientation, temperature effects, zero-shift, background noise), expanding the dataset from 5M to 40M, which improves sim-to-real transfer and robustness under noisy spectra without instrument calibration or structure refinement. We also evaluate robustness to missing/noisy composition, enabling composition-completion and use cases such as predicting dopants and doping ratios in compositionally disordered materials, and demonstrate transfer to low-data experimental targets (e.g., ion conductivity). Overall, we present a multimodal foundation model that learns from composition and XRD to build a more comprehensive materials representation and accelerate experimental materials discovery.

References

[1] Jacobs, Ryan, et al. "A practical guide to machine learning interatomic potentials–Status and future." *Current Opinion in Solid State and Materials Science* 35 (2025): 101214.

[2] Li, Junnan, et al. "Align before fuse: Vision and language representation learning with momentum distillation." *Advances in neural information processing systems* 34 (2021): 9694-9705.

[3] Belouadi, Jonas, Tamy Boubekeur, and Adrien Kaiser. "MultiMat: Multimodal Program Synthesis for Procedural Materials using Large Multimodal Models." *arXiv preprint arXiv:2509.22151* (2025).

[4] Mirza, Adrian, et al. "MatBind: Probing the multimodality of materials science with contrastive learning." *AI for Accelerated Materials Design-ICLR 2025*. 2025.

[5] Wang, Anthony Yu-Tung, et al. "Compositionally restricted attention-based network for materials property predictions." *Npj Computational Materials* 7.1 (2021): 77.

[6]J. Subramanian, L. Hung, D. Schweigert, S. Suram, W. Ye, arXiv:2507.01054 (2025).

Figures

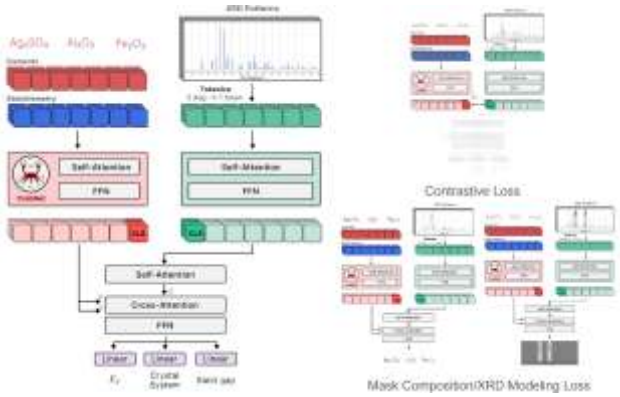

Figure 1. **Multimodal composition–XRD foundation model architecture (left) and pretraining objectives (right).**
Left: composition (elements and stoichiometry) and XRD patterns (tokenized peaks) are encoded by modality-specific Transformer encoders and fused via a cross-attention module, producing a joint representation for downstream property prediction.
Right: self-supervised pretraining losses used in this work including cross-modal contrastive alignment, masked modeling of composition and XRD (MCM/MXM), to learn transferable, experimentally grounded representations.