A Multimodal Conditional JEPA for Composite Materials

Abhiroop Bhattacharya [1,2] **Hangwei Qian** [2] **Ivor Tsang** [2]

[1]*École de technologie supérieure, Montreal, Canada* [2]*CFAR and IHPC, Agency for Science, Technology and Research (A*STAR), Singapore*. Correspondence to: Hangwei Qian qian_hangwei@a-star.edu.sg.

1. Introduction

Composite properties are governed by latent morphological and interfacial mechanisms that manifest across multiple data modalities [1], including **microstructure images** capturing spatial organization, **scientific text** encoding domain context, and **process and material parameters** such as cooling rate, fiber size, and filler volume fraction that strongly influence strength and stiffness [2]. Despite the inherently multimodal nature of composite materials, traditional machine learning approaches rely on a single data source and are further constrained by the high cost of experimental characterization, which limits the availability of annotated data [3, 4]. These challenges motivate a **multimodal** self-supervised foundation model that learns robust, invariant representations and can be efficiently adapted to diverse downstream tasks. While recent work has explored diffusion-based models for composite materials [4], in small data regimes, pixel-level reconstruction objectives often overfit acquisition artifacts [5], while contrastive methods depend on fragile negative sampling strategies that are ill-suited to small scientific datasets [6].

To address these challenges, we propose a conditional multimodal Joint Embedding Predictive Architecture (JEPA [7]) that learns invariant, material-relevant representations by predicting latent targets using context from complementary modalities, encouraging invariance to experimental measurement artifacts while retaining morphology and context-sensitive factors. Our contributions are threefold: (1) a **JEPA-based multimodal** conditional pretraining framework that leverages microstructure images, process parameters, and scientific text; (2) a **conditioning strategy** that incorporates textual and visual context into feature-wise tabular modeling; and (3) an experimental evaluation demonstrating improved prediction performance across multiple mechanical properties. **Our results indicate that the proposed model learns compact latent representations that generalize effectively across downstream material property prediction tasks.**

2. Proposed Model Architecture

Let $\mathbf{x} = [x_1, x_2, \ldots, x_F]$ denote the tabular input, where each x_i corresponds to a process or material feature. Each feature x_i is independently embedded into a d-dimensional token

$$\mathbf{e}_i = \phi_i(x_i) \in \mathbb{R}^d, \tag{1}$$

where $\phi_i(\cdot)$ denotes a feature-specific embedding function. The resulting sequence of feature tokens

$$\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_F] \in \mathbb{R}^{F\times d} \tag{2}$$

is processed by a **feature-wise Transformer encoder** $f_\theta(\cdot)$, which captures inter-feature dependencies via self-attention:

$$\mathbf{H} = f_\theta(\mathbf{E}) \in \mathbb{R}^{F\times d}. \tag{3}$$

Given a microstructure image I and a corresponding textual description T, modality-specific representations are obtained using pretrained encoders:

$$\mathbf{z}_I = g_{\text{ViT}}(I), \quad \mathbf{z}_T = g_{\text{BERT}}(T), \tag{4}$$

where $g_{\text{ViT}}(\cdot)$ denotes a Vision Transformer (ViT) [8] encoder and $g_{\text{BERT}}(\cdot)$ denotes a MatSciBERT [9] encoder. The image and text embeddings are concatenated and projected into a shared conditioning space.

$$\mathbf{z}_c = W_c[\mathbf{z}_I \,\|\, \mathbf{z}_T] \in \mathbb{R}^d, \tag{5}$$

where W_c is a learnable linear projection and $\|$ denotes concatenation. The conditioning vector $\mathbf{z}_c$ is incorporated by additive modulation of the feature token embeddings $\mathbf{e}_i$. The conditioned representation is given by:

$$\tilde{\mathbf{E}}_i = \mathbf{e}_i + \alpha\mathbf{z}_c, \tag{6}$$

where α is a scaling factor controlling the contribution of the multimodal conditioning. Two conditioned feature-wise Transformer encoders are employed, a context encoder $f_\theta^{(c)}$ and a target encoder $f_\theta^{(t)}$. Both share the same architecture but maintain separate parameters. Given the same multimodal conditioning $\mathbf{z}_c$, the encoders generate representations:

$$\mathbf{H}^{(c)} = f_\theta^{(c)}(\tilde{\mathbf{E}}), \quad \mathbf{H}^{(t)} = f_\theta^{(t)}(\tilde{\mathbf{E}}). \tag{7}$$

We randomly mask approximately 30% of the input features while keeping the rest visible. The visible context embeddings are passed as input to a Multilayer Perceptron (MLP) which predicts the masked features in latent space. The error between the predicted masked features and the encoded masked features is measured using L2 loss. Fig. 1 shows a schematic of the proposed model.

3. Experiments

Dataset We analyze a dataset of electrospun nanofiber-reinforced composites containing process parameters and fiber descriptors, including fiber mixture, flow rate, radius, thickness, and width. Nanofiber morphology is characterized using scanning electron microscopy (SEM). Mechanical properties are measured via uniaxial tensile tests, yielding

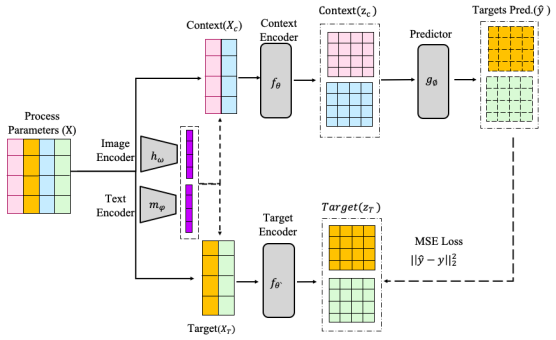

Fig. 1: An overview of the proposed approach. The tabular features are encoded using conditioned feature-wise Transformers. The fusion of text embeddings and image embeddings is used to condition the transformer. The error between the predicted masked features and the encoded masked features is measured using L2 loss.

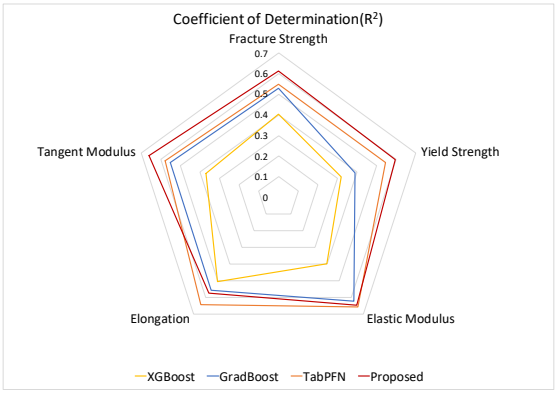

Fig. 2: The radar plot presents a comparison of the R^2 for the proposed algorithm with SoTA algorithms for tabular data.

fracture strength, yield strength, elastic modulus, tangent modulus, and fracture elongation. A binary indicator denotes the tensile loading direction. Details of the dataset construction are provided in [10].

4. Results

In this section, we compare the performance of the proposed model to other state-of-the-art (SoTA) models for tabular data. We only fine-tune the MLP prediction head while keeping the weights of the encoder frozen. Fig. 2 shows that the proposed method outperforms baseline models on most mechanical properties, with notable gains in strength and modulus related targets, while elongation remains comparable.

Changing the masking ratio We investigate the effect of the masking ratio by varying the number of masked columns while keeping all other model architecture and hyperparameters fixed. Performance improves as the masking ratio increased up to three columns, with similar results for masking two and three columns. However, further increasing the masking ratio led to a degradation in performance.

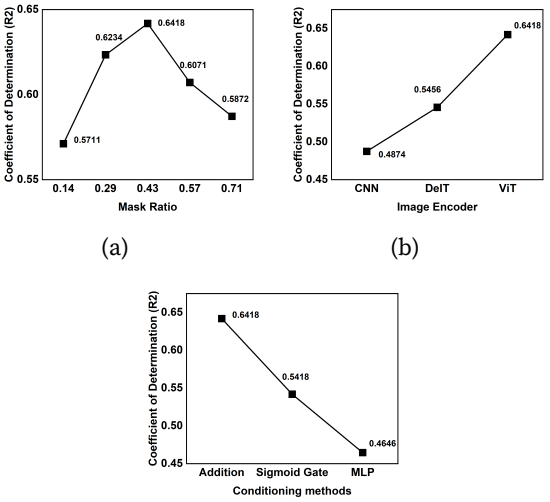

Fig. 3: The figures present the ablation studies for (a) target masking ratio, (b) microstructure image encoder and (c) multimodal conditioning strategy.

Fig. 3(a) shows this inverted U-shaped performance curve.

Vision Encoders We ablate the encoder used for encoding the microstructure images. We ran experiments using a CNN encoder, and a pretrained DeiT encoder [11], a pretrained Vision Transformer to encode the microstructure images. We observe that the ViT encoder outperforms the other encoders. Fig. 3(b) shows the trend.

Conditioning Mechanism We evaluate several strategies for conditioning the context and target encoders with text and image embeddings. The simplest approach adds a small projected bias to the feature representations, scaled by a factor. We also explore more expressive alternatives, including sigmoid-based gating with an MLP and a weighted averaging scheme learned via an MLP. Surprisingly, the simple bias-based method consistently achieves the best performance. We attribute this to the limited size of the tabular data. The difference in model performance is shown in Fig. 3(c).

5. Conclusion

In this work, we propose a multimodal conditional framework for predicting material properties of composite materials using tabular features, microstructure images, and literature-derived text. The proposed approach follows a JEPA-based formulation to predict masked material properties without reconstructing raw inputs or relying on contrastive memory banks. Our results show that the proposed model learns meaningful correlations across the available inputs and achieves performance in line with established SoTA tabular learning methods. Future work will explore improved robustness to incomplete data.

6. Acknowledgments

This research is supported by the National Research Foundation, Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No.: AISG2-GC-2023-010), "Design Beyond What You Know": Material-Informed Differential Generative AI (MIDGAI) for Light-Weight High-Entropy Alloys and Multi-functional Composites (Stage 1b). This research is supported by A*STAR Career Development Fund <Project No. C243512010>.

References

[1] Silu Huang, Qiuni Fu, Libo Yan, and Bohumil Kasal. Characterization of interfacial properties between fibre and polymer matrix in composite materials–a critical review. *Journal of Materials Research and Technology*, 13:1441–1484, 2021.

[2] Shutian Liu, Conglin Dong, Chengqing Yuan, and Xiuqin Bai. Study of the synergistic effects of fiber orientation, fiber phase and resin phase in a fiber-reinforced composite material on its tribological properties. *Wear*, 426:1047–1055, 2019.

[3] Yi Liang, Xinyue Wei, Yongyue Peng, Xiaohan Wang, and Xiaoting Niu. A review on recent applications of machine learning in mechanical properties of composites. *Polymer Composites*, 46(3):1939–1960, 2025.

[4] Hangwei Qian, Yang He, Bingjin Chen, Mohit Sharma, and Ivor Tsang. Physics-constrained diffusion for lightweight composite material design. In *AI for Accelerated Materials Design-NeurIPS 2025*, 2025.

[5] DJ Lekou, TT Assimakopoulou, and TP Philippidis. Estimation of the uncertainty in measurement of composite material mechanical properties during static testing. *Strain*, 47(5):430–438, 2011.

[6] Namkyeong Lee, Junseok Lee, and Chanyoung Park. Augmentation-free self-supervised learning on graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7372–7380, 2022.

[7] Yann LeCun. A path towards autonomous machine intelligence version 0.9.2, 2022-06-27. *OpenReview*, 62(1), 2022.

[8] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Tanishq Gupta, Mohd Zaki, NM Anoop Krishnan, and Mausam. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102, 2022.

[10] Yuhui Wu, Minmin Ding, Haonan He, Qijun Wu, Shaohua Jiang, Peng Zhang, and Jian Ji. A versatile multimodal learning framework bridging multiscale knowledge for material design. *npj Computational Materials*, 11(1):276, 2025.

[11] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.