Information fusion strategy based on language model and contrastive learning for knowledge retrieval of metallic materials

Zhouran Zhang[a], **Yongqian Peng**[a], **Yicong Ye**[a]

[a] *Department of Materials Science and Engineering, National University of Defense Technology, Changsha, China 410073* z_zhang@outlook.com

1. Introduction

The predominant AI4M (AI for Materials) approaches remain circumscribed by a reductionist perspective: simplifying materials properties into deterministic relationships with specific feature parameters and constructing such mappings through feature engineering and model training. This methodology has proven particularly efficacious in predicting intrinsic materials properties. Nevertheless, ductility is dependent on non-intrinsic factors (such as process-induced defects, porosity, shrinkage cavities and interface characteristics) which defy quantitative representation, and thus the prediction of ductility of is challenging. Materials science literature not only encompass experimental data but also encapsulate authors' profound experimental insights and distilled materials science knowledge.

2. Results and discussion

In this work, we present an innovative information fusion model architecture based on MatSciBERT and contrastive learning. Using refractory multi-principal alloy ductility prediction as a case study, we constructed a specialized corpus comprising 520 titanium alloy entries and 50 refractory multi-principal alloy entries, encompassing ductility data and influencing factors (processing methods and microstructural characteristics). Through comparative analysis with conventional machine learning models built on structured data, we validated the efficacy and feasibility of text-physical information fusion, demonstrating an improvement in prediction accuracy from 40% to 84.9%. Furthermore, employing contrastive learning strategies, we established relationships among physical parameters, processing methods, and microstructural characteristics, enabling ductility prediction even with incomplete microstructural information. Applying this contrastive learning model to the vast compositional space of the Ti-V-Zr-Nb-Hf-Ta hexinary system yielded predictions with minimal RMSE deviation (merely 3%) from reported values, proving the method's validity. Notably, the model autonomously identified a ductility peak near 15% Ti content, correlating remarkably with recent experimental findings. Analysis of the model's attention mechanisms revealed that this prediction largely stemmed from the model's comprehension of processing-structure relationships described in literature, suggesting that this methodology not only enables accurate property prediction but may also facilitate the discovery of novel materials science principles. This paradigm of integrating textual knowledge with physical features presents a novel approach to addressing complex property prediction challenges in materials science.

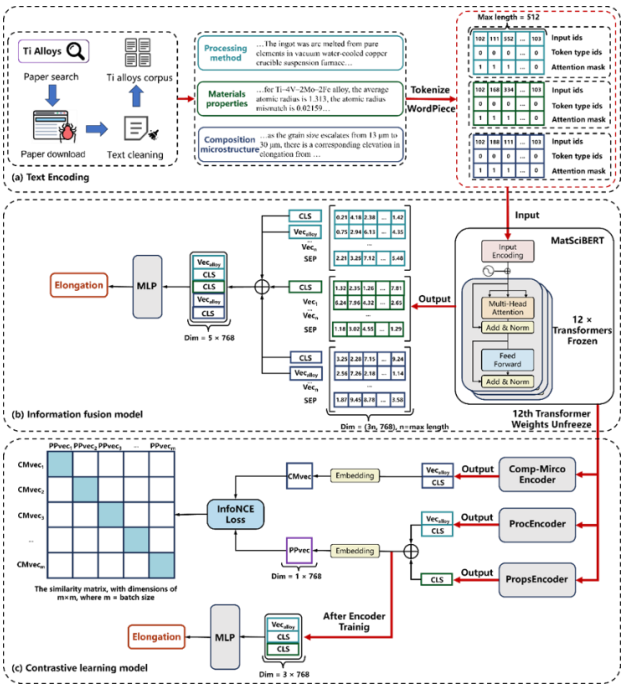

Fig. 1: Overview of our approach. (a) Text Encoding (b) Information fusion model (c) Contrastive learning model.

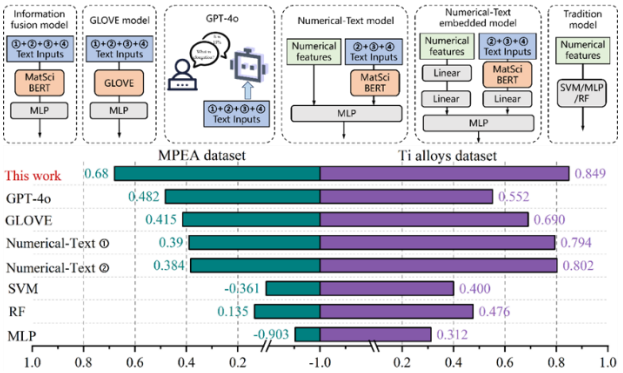

Fig. 2: The R^2 of different methods between MPEA dataset and titanium alloys dataset.

Acknowledgments

This study was funded by National Natural Science Foundation of China [11972372].

References

[1] Sasidhar, K. N. et al. Enhancing corrosion-resistant alloy design through natural language processing and deep learning. Science Advances 9, eadg7992, (2023).

[2] Tshitoyan, V., Dagdelen, J., Weston, L. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571, 95–98 (2019).

[3] Tian, S. et al. Steel design based on a large language model. Acta Materialia 285, 120663, (2025).