Discovery and recovery of crystalline materials with property-conditioned transformers

Cyprien Bone[1] and Keith T. Butler[1]
Matthew A.H. Walker[1]
Kuangdai Leng[2]
Luis M. Antunes[3]
Ricardo Grau-Crespo[4]
Amil Aligayev[5] and Javier Domiguez[5]

[1]Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, UK
[2]Earth Rover Program, 71-75 Shelton Street, Covent Garden, London WC2H 9JQ, UK
[3]Independent Researcher, Canada
[4]School of Engineering and Materials Science, Queen Mary University of London, London E1 4NS, UK
[5]NOMATEN Centre of Excellence, National Centre for Nuclear Research, ul. A. So ltana 7, Otwock, 05-400, Poland

cyprien.bone.24@ucl.ac.uk and k.t.butler@ucl.ac.uk

Generative models have recently shown great promise for accelerating the design and discovery of new functional materials. Conditional generation enhances this capacity by allowing inverse design, where specific desired properties can be requested during the generation process. However, conditioning of transformer-based approaches is constrained by discrete tokenisation schemes and the risk of catastrophic forgetting during fine-tuning. This work introduces CrystaLLM-π (property injection), a conditional autoregressive framework that integrates continuous property representations directly into the transformer's attention mechanism.[1]

Two architectures, Property-Key-Value (PKV) Prefix attention and PKV Residual attention, are presented in Figure 1. These methods bypass inefficient sequence-level tokenisation and preserve foundational knowledge from unsupervised pre-training on Crystallographic Information Files (CIFs) as textual input. We establish the efficacy of these mechanisms through systematic robustness studies and evaluate the framework's versatility across two distinct tasks. First, for structure recovery, the model processes high-dimensional, heterogeneous X-ray diffraction patterns, achieving structural accuracy competitive with specialised models. We also demonstrate the model's capability to resolve experimental structures and differentiate polymorphs from lab-measured patterns (Figure 2). Second, for materials discovery, the model is fine-tuned on a specialised photovoltaic dataset[2] to generate novel, stable candidates (Figure 3a) validated by Density Functional Theory (DFT). It implicitly learns to target optimal band gap regions for high photovoltaic efficiency (Figure 3b), demonstrating a capability to map complex structure–property relationships. CrystaLLM-π provides a unified, flexible, and computationally efficient framework for inverse materials design.

References

[1] Bone, C. Butler, K. Walker, M. Antunes, L. Leng, K. Grau-Crespo, R. Aligayev & A. Dominguez, J. (2025), arXiv:2511.21299 (pre-print)
[2] Matthew A.H. Walker and Keith. T. Butler, Mater. Horiz. (2026)
[3] Radford, A. Wu, J. Child, R. Luan, D. Amodei, D. & Sutskever, I. OpenAI Technical Report, 1 (2019)
[4] Gražulis, S. Chateigner, D. Downs, R. T. Yokochi, A. F. T. Quirós, M. Lutterotti, L. Manakova, E., Butkus, J. Moeck, P., & Le Bail, A., Journal of Applied Crystallography, 42(4) (2009) 726-729

Figures

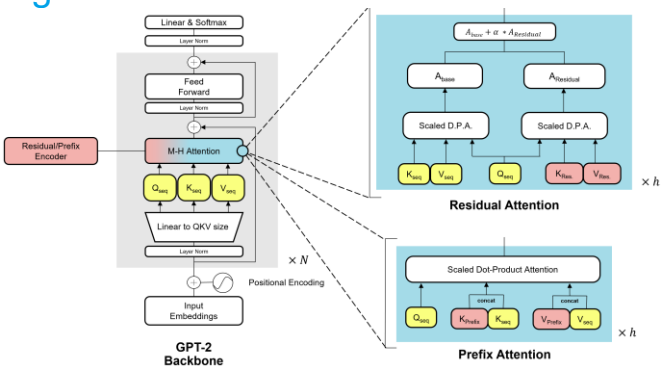

Figure 1. Architecture Diagrams for CrystaLLM-π, with Prefix vs Residual attention mechanisms. The base model is inspired by the GPT-2 architecture[3], with additional modules and modified attention to condition the textual outputs on continuous input properties.

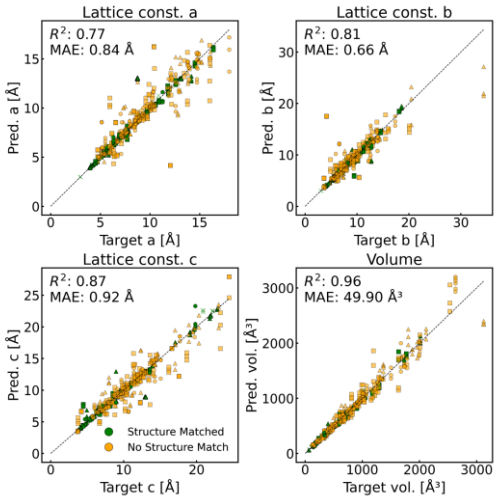

Figure 2. Predicted vs. True lattice parameters when model attempts to resolve experimental structures from the COD[4] using lab-measured XRD patterns.

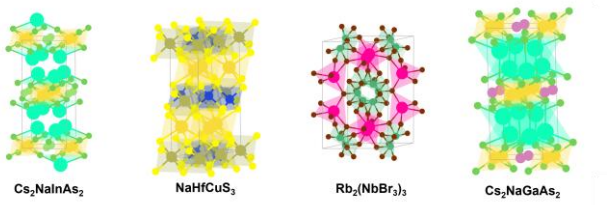

Figure 3a. PV Candidates generated by the model absent from the training set and validated with DFT. The predicted Spectroscopy Limited Maximum Efficiency (SLME) values (HSE06 functional) from left-to-right are: 26.4%, 23.3%, 13.3%, 24.4%.

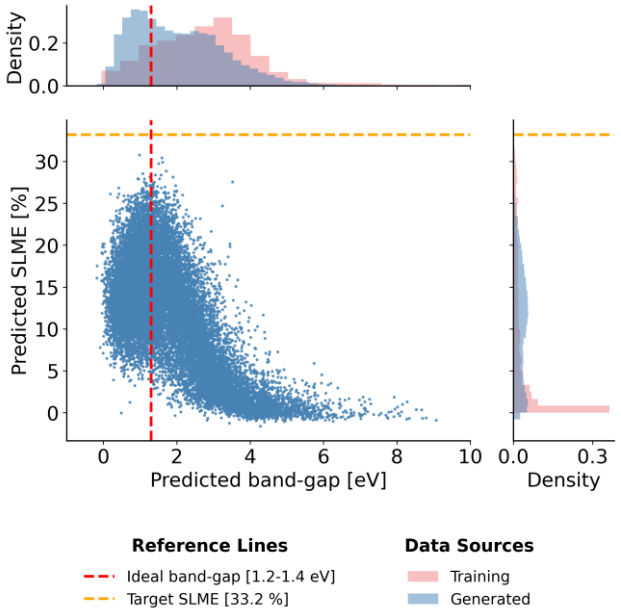

Figure 3b. Exploration of the photovoltaic candidate chemical space. Scatter plot of predicted SLMEs against predicted band gaps for generated structures. The distribution highlights the model's implicit focus on the Shockley-Queisser optimal band gap range (1.2-1.4 eV), despite the absence of explicit band gap supervision.