Learning the language of crystal chemistry: How methods from natural language can help discover new materials.

Keith T. Butler [1], Luis M. Antunes[2] , Ricardo Grau-Crespo[3], Cyprien Bone[1]
[1]UCL, London, UK
[2]Independent researcher, Toronto, Canada
[1]Queen Mary University of London, London, UK

k.t.butler@ucl.ac.uk

The discovery and design of new materials is critical for advancing carbon-emission-reducing technologies such as renewable energy and electric vehicles. While quantum-mechanical (QM) calculations have brought computational materials design within reach, their cost still limits screening to relatively small sets of candidate compounds, particularly when targeting the complex properties required for next-generation energy materials. This motivates the development of fast, data-driven models that can learn directly from the rapidly growing body of crystallographic data.

In this talk I will present examples of how we are adapting concepts from natural language processing to build efficient models for materials discovery. By treating crystal structures as a language, we can learn distributed representations of atomic species and local environments directly from large structural databases [1]. Extending this idea, we train transformer-based large language models on sequences derived from crystallographic information files, enabling next-token prediction to generate complete crystal structures [2]. These models produce syntactically valid and chemically plausible compounds, demonstrating that they capture important regularities in crystal chemistry and offering a new route to the long-standing challenge of predicting structure from composition.

However, a detailed analysis shows that the generated materials often remain close to the training distribution, raising important questions about how far such models can extrapolate beyond known chemistry. I will discuss recent work on conditioning strategies that incorporate experimental constraints or targeted properties into the generation process [3]. While this further highlights the difficulty of exploring truly novel regions of materials space, it also reveals a powerful and practical use case: the rapid proposal of structures consistent with experimental observations, providing new tools for accelerated materials characterisation and discovery.

The aim of this talk is to give a balanced perspective on the opportunities and limitations of language-model approaches in solid-state chemistry. These methods provide a compelling new framework for learning the “language of crystal chemistry” and for building fast surrogate models for materials design, but they also challenge us to think carefully about what it means for a model to generalise and to discover genuinely new materials.

References

[1] Luis M Antunes, Ricardo Grau-Crespo, Keith T Butler, npj Computational Materials, 8 (2022) 44
[2] Luis M Antunes, Keith T Butler, Ricardo Grau-Crespo, Nature Commun., 15 (2024) 10570
[3] Cyprien Bone, Matthew Walker, Kuangdai Leng, Luis M Antunes, Ricardo Grau-Crespo, Amil Aligayev, Javier Dominguez, Keith T Butler, arXiv, (2025) arXiv:2511.21299