SUSTENET

"Computational Approaches to Sustainable Thermoelectric Materials: Challenges and Opportunities"

WG1 Webinar

Tuesday, 28 April, 11:30 AM – 1:30 PM (CEST)

This inaugural webinar for the COST Action SUSTENET marks the first step in building a vibrant, interdisciplinary community dedicated to advancing sustainable thermoelectric technologies.

Organized by Working Group 1, which focuses on leveraging artificial Intelligence to accelerate material discovery and optimize system efficiency, this session brings together researchers exploring how computational tools can unlock the next generation of earth-abundant, high-performance thermoelectric materials.

Programme:

Keith Butler (UCL, UK): "***Learning the language of crystal chemistry: How methods from natural language can help discover new materials***"

Philippe Jund (Montpellier, France): "***Machine Learning Strategies for the Accelerated Discovery of Efficient Half-Heusler Thermoelectric Materials***"

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

Free event upon registration! Click below!

REGISTRATION PAGE

SUSTENET

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

Keith Butler
(UCL, UK)

Abstract

The discovery and design of new materials is central to advancing low-carbon energy technologies, including renewable energy systems, electrification, and energy harvesting. While quantum-mechanical (QM) calculations have made computational materials design increasingly accessible, their cost still restricts screening to relatively small sets of candidate compounds, particularly when targeting the complex, often competing properties required for next-generation energy materials such as thermoelectrics. This motivates the development of fast, data-driven models that can learn directly from the rapidly expanding 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, with potential relevance for identifying materials with tailored transport properties.

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, such as those relevant to electronic and thermal transport [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 the design of functional energy materials, while also prompting careful consideration of what it means for a model to generalise and to discover genuinely new materials.

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

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

“Computational Approaches to Sustainable Thermoelectric Materials: Challenges and Opportunities”
WG1 Webinar, Tuesday, 28 April, 11:30 AM – 1:30 PM (CEST)

SUSTENET

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

Philippe Jund
ICGM, Univ Montpellier, CNRS, ENSCM
Co-authors: Shoeb Athar, Adrien Mecibah

Abstract

Thermoelectric materials are pivotal to the global transition toward sustainable energy due to their ability to convert waste heat directly into electricity via the Seebeck effect. While machine learning (ML) can accelerate the discovery of high-performance thermoelectric (TE) materials, its efficacy is often compromised by the "garbage in, garbage out" phenomenon*. This work identifies critical inaccuracies in publicly available databases, stemming from Large Language Model (LLM)-assisted data curation, ambiguous nomenclature, and complex material structures. Using one of the most promising TE family of materials (half-Heuslers) as a case study, we first propose an innovative statistical filtering method based on the "round-robin" experimental error **. We can thus objectively and automatically identify and remove outliers from the dataset and then design a hybrid dataset creation workflow combining curated data from publicly available databases with the precision of manual literature extraction. The resulting ICGM dataset demonstrates significantly enhanced consistency through rigorous benchmarking.

Four ML models - SISSO, XGBoost, Random Forest, and Neural Networks - were then trained and validated using a PCA-based train/test split method to ensure generalizability. The SISSO (Sure Independent Screening and Sparsifying Operator) model demonstrated the highest predictive accuracy and provided physical interpretability, highlighting the heat of vaporization and electron affinity as dominant descriptors for the thermoelectric figure-of-merit (zT).

High-throughput screening of 3,780 pure hH compounds using ensemble-averaged ML predictions identified several promising candidates. Density Functional Theory (DFT) validation of the two top candidates predicted exceptional zT values of 2.1 at 650 K and 1.26 at 700 K, respectively. These findings demonstrate that integrating ML-driven screening with targeted DFT validation and experimental libraries can significantly reduce the time and cost associated with discovering next-generation energy materials.

* S. Athar and P. Jund, Artificial Intelligence Chemistry 4, 100113 (2026)
** S. Athar, A. Mecibah and P. Jund, Mater. Today Phys. 59, 101948 (2025)

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

"Computational Approaches to Sustainable Thermoelectric Materials: Challenges and Opportunities"
WG1 Webinar, Tuesday, 28 April, 11:30 AM – 1:30 PM (CEST)