# Getting better materials faster with ML – a question of representation and distributed platforms

**Tejs Vegge**[1,2], Raul Ortega-Ochoa[1], Alán Aspuru-Guzik[2,3], Tonio Buonassisi[2,4]

[1]DTU Energy and CAPeX Pioneer Center for Accelerating P2X Materials Discovery, Technical University of Denmark, DK 2800 Kgs. Lyngby, Denmark
[2]Acceleration Consortium, 700 University Ave, Toronto,
[3]Chemical Physics Theory Group, Department of Chemistry, University of Toronto, 80 St. George St, Toronto, Ontario M5S 3H6, Canada, Ontario M5G 1Z
[4]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

teve@dtu.dk

Here, we discuss two aspects of machine learning (ML) for materials centered around new and faster methods for predicting and discovering complex or "hard to obtain" properties of advanced materials.

Recent advancements in ML4Materials have demonstrated that apparently simple materials representations like the chemical formula without any structural information can sometimes achieve competitive property prediction performance in common tasks. Our physics-based intuition would suggest that such representations are "incomplete," which indicates a gap in our understanding. Using a *tomographic interpretation* of structure-property relations to bridge that gap by defining what is a material representation, material properties, the material, and the relationships between these [1]. We apply concepts from information theory to verify this framework by performing an exhaustive comparison of property-augmented representations on a range of materials property prediction objectives.

With this in mind, we, as scientists, might not know *a priori* which experiments or simulations will ultimately provide the most valuable information to capture the "ghost of the material," i.e., what is the fastest or least expensive path to obtain a complex material's property. Is it, e.g., more valuable to know the formation energy per atom than to know the total magnetization to predict the band gap (Figure 1)?

This then begs the question, how can we dynamically orchestrate the acquisition of such multimodal information and datasets of unknown dimensionality, which may not even be available in our own labs?

To begin to answer these questions, we will show two examples of how dynamic workflow orchestrators like PerQueue [2] are capable of orchestrating multimodal data acquisition from simulations and experiments, and the FINALES (Fast INtention-Agnostic LEarning Server) framework [3] for integration of data from geographically distributed Materials Acceleration Platforms (MAPs) [4] or self-driving laboratories (SDL) [5].

## References

[1] Raul Ortega, Alán Aspuru-Guzik, Tejs Vegge, Tonio Buonassisi, "A tomographic interpretation of structure-property relations for materials discovery", ArXiv, 10.48550/arXiv.2501.18163 (2025)
[2] Benjamin H. Sjølin et al., "PerQueue: managing complex and Dynamic workflows", Digital Discovery 5, 1832 (2024).
[3] Monika Vogler et al. "Autonomous battery optimisation by deploying distributed experiments and simulations", Adv. Energy Mater., 2403263 (2024).
[4] Simon P. Stier et al., "Materials Acceleration Platforms (MAPs): Accelerating Research and Development to Meet Urgent Societal Challenges", Adv. Mater. 2407791 (2024).
[5] Brenden Pelkie et al., "Democratizing self-driving labs through user-developed automation infrastructure", ChemRxiv, https://doi.org/10.26434/chemrxiv-2025-zhkrf (2025).

## Figures

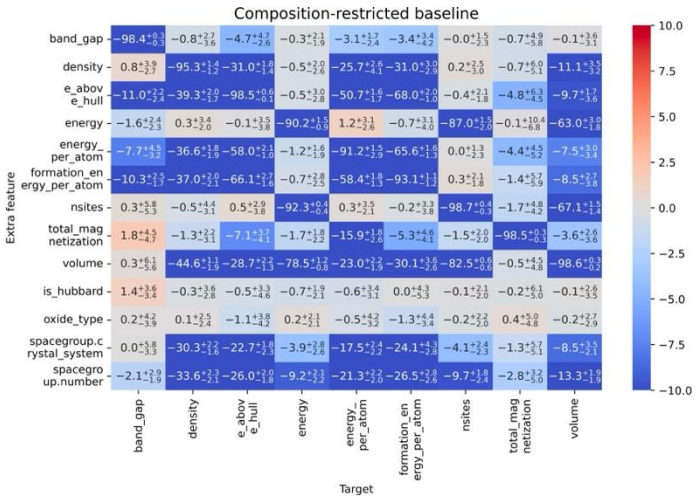


**Figure 1.** Percentage change in MAE of an augmented vs a non-augmented composition-restricted representation [1].