Wyckoff Set regression for Materials Discovery

Rhys E. A. Goodall[1], Abhijith S. Parackal[2], Felix A. Faber[1], Rickard Armiento[2], Alpha A. Lee[1]

[1]Cavendish Laboratory, University of Cambridge
[2] Department of Physics, Chemistry and Biology (IFM), Linköping University

Materials Discovery in a Vast Materials Space

In the modern era scientific efforts have led to the synthesis and structural characterisation of $O(10^5)$ different stable/meta-stable inorganic crystalline materials. The space of possible inorganic materials is believed to be of the order $O(10^{100})$. As such the discovery of novel stable materials epitomises a "needle in a haystack problem".

The Materials Discovery Pipeline

Historically the majority of progress in Materials Science has been made via Trial-Error-Improvement experimental workflows. Both in terms of synthesis products and procedures.

Recently the cost of computation has fallen making *ab-initio* quantum mechanical calculations of materials properties more accessible. These calculations can accelerate traditional workflows by allowing for more targeted experiments to be conducted. Critically, the majority of the these *ab-initio* methods make use of atomic positions as inputs which limits their application to high-throughput novel material discovery.

In recent years there has been increased interest in leveraging machine learning in conjunction with established high-throughput *ab-initio* materials databases to attempt to amortise and avoid bottlenecks in the current Materials Discovery Pipeline. The key challenge is selecting appropriate descriptors (i.e. model inputs) for the task.

Descriptors for Screening Novel Materials

Descriptors for high-throughput novel material discovery have two key requirements,

- computationally cheap to enumerate for the design space being considered.
- distinguish different material polymorphs (e.g. Graphite and Diamond).

In crystallography structures can be described fully by their spacegroup, lattice parameters and their set of occupied Wyckoff positions.

Wyckoff positions in a given crystal are described by Wyckoff letters, which reflect the site symmetries, and offsets, that localise the atoms to particular positions. Discarding the offsets gives an "anonymised" Wyckoff set. This descriptor can be combinatorially enumerated but also captures key information about the crystal symmetry.

	Cheap to Enumerate	Distinguishes Polymorphs
Material Composition	✓	✗
Crystal Structure	✗	✓
Anonymised Wyckoff Set	✓	✓

Table 1: Table showing the characteristics of different descriptor choices

Machine Learning Results

For the composition model we use the recently introduced *Roost* model [1], a set regression model that operates on the elements in a material. For the crystal structure based model we use the crystal graph convolutional neural network (*CGCNN*) [2]. For the "anonymised" Wyckoff set model we use an adaptation of the *Roost* model that we call *Wren* (**W**yckoff **Re**presentation **N**etwork). All models are trained to predict the formation enthalpy per atom of materials recorded in the Materials Project catalogue, 20% of the data is held-out as a test set.

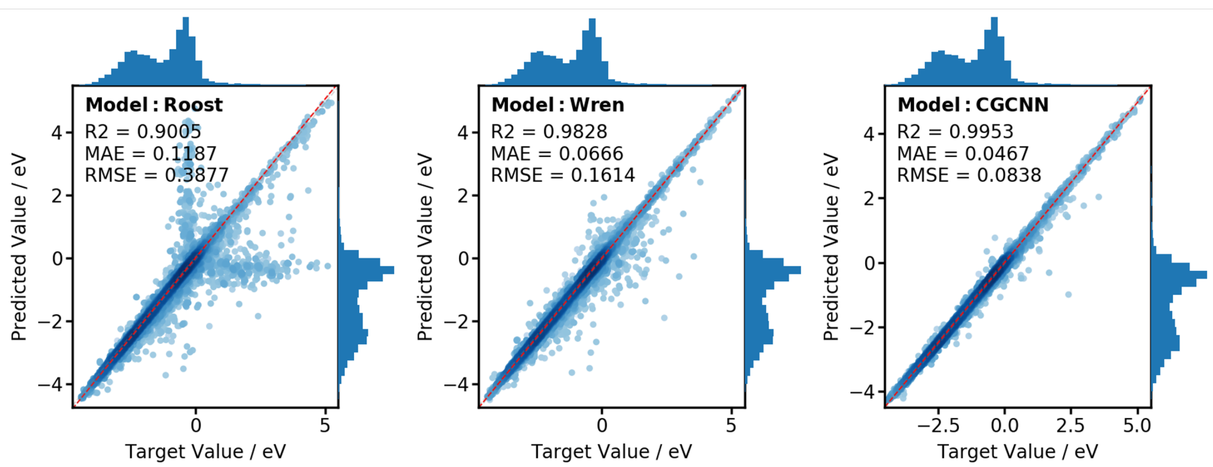

Figure 1: Scatter plots showing the predicted enthalpy against the target enthalpy. Above the scatter plots we show schematics indicating the information content of each descriptor.

These preliminary results show that "anonymised" Wyckoff set regression can distinguish polymorphs whilst still being cheap to enumerate. As such it offers a compelling alternative to prototyping to screen for novel stable materials.

References

[1] Rhys EA Goodall and Alpha A Lee.
Predicting materials properties without crystal structure: Deep representation learning from stoichiometry.
arXiv preprint arXiv:1910.00617, 2019.

[2] Tian Xie and Jeffrey C Grossman.
Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties.
Physical Review Letters, 120(14):145301, 2018.

reag2@cam.ac.uk