Wyckoff Set Regression for Materials Discovery

Rhys E. A. Goodall
Cavendish Laboratory
University of Cambridge
Cambridge, CB3 0HE, UK

Abhijith S. Parackal
IFM - Theoretical Physics,
Linköping University,
Linköping, SE-58183, Sweden

Felix A. Faber
Cavendish Laboratory
University of Cambridge
Cambridge, CB3 0HE, UK

Rickard Armiento
IFM - Theoretical Physics,
Linköping University,
Linköping, SE-58183, Sweden

Alpha A. Lee
Cavendish Laboratory
University of Cambridge
Cambridge, CB3 0HE, UK

Abstract

In recent years machine learning has been shown to be able to approximate the accuracy and amortise the computational cost of *ab-initio* quantum mechanics calculations. This has opened up many exciting use cases in the study of materials *in-silico*. However, the majority of the these works make use of atomic positions as inputs which limits their application to novel material discovery applications where crystal structures are *a-priori* unknown. For a model to see useful application in materials discovery we need to be able to enumerate its inputs over a possible design space of new materials. In this work, we build upon a recent machine learning framework for material science that operates on the stoichiometry of materials and extend it to look at Wyckoff sets. We show that operating on Wyckoff sets allows the model to handle compositions with multiple polymorphs, therefore, overcoming one of the major limitations of composition-based models whilst maintaining the key benefit of having a combinatorially enumerable input space.

1 Introduction

In recent years there has been a boom in academic work applying novel graph-based models to small organic molecules for applications in cheminformatics and drug discovery [1, 2]. Many works inspired by such pioneering activities have attempted to realise a similar paradigm in the study of inorganic crystalline materials [3, 4]. However, whilst it is possible to combinatorially generate huge numbers of potentially viable molecules using rules based chemistry [5] it is not possible to generate stable crystalline structures in a similar way. This limits the ways in which we can use structure-based machine learning models to discover novel materials. Although on-the-fly and hybrid DFT-ML force field workflows [6, 7, 8] have shown early promise for accelerating structure relaxations, even with these advances crystal structure searching remains computationally costly, limiting the number of candidates that can be considered.

In order to investigate larger search spaces several groups have argued for using composition-based models [9, 10, 11, 12] to triage discovery workflows as they do not require atomic positions as input. Such models can be used to de-risk the discovery process by restricting the search space to compositions more likely to give rise to stable polymorphs, therefore, reducing the number of computationally expensive crystal structure searches that need to be conducted. Alternatively, others have proposed restricting searches to given structure prototypes e.g. perovskites [13] or elpasolites [14] where the structural constraint allows for crystal structure searching to be avoided entirely.

Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.

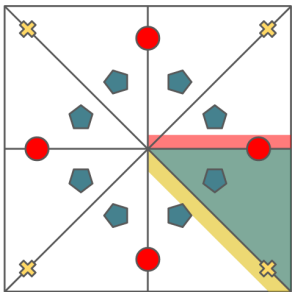

Figure 1: The figure shows a toy 2D crystal with 3 occupied Wyckoff positions. The shaded regions illustrate the regions the relevant atoms are constrained to lie in by specifying an anonymised Wyckoff position for that atom.

In this extended abstract we introduce preliminary work on a new model that combines the benefits of these two different approaches by imposing constraints on the symmetries of the structures through considering anonymised Wyckoff sets. In theory this approach maintains the flexibility of composition-based models to conduct combinatorial searches across materials space whilst simultaneously reducing the burden of the necessary crystal structure searching by providing symmetry constraints for the structure relaxations.

2 Discovery of Novel Stable Inorganic Materials

2.1 Premise

Only a small proportion of possible material compositions (believed to be of the order 10^{100} [15]) will have dynamically stable polymorphs with formation enthalpies per atom sufficiently low that they lie near to or on the convex hull of thermodynamically stable materials. Finding new compositions and the associated structures of the thermodynamically stable/meta-stable polymorphs *in silico* with minimal computational cost is a central problem in computational material science. Currently the only feasible high-throughput methods for structure searching are based on prototyping methods that substitute new atoms into known prototypes [16, 17]. Whilst these methods have allowed for rapid expansion of high-throughput databases large amounts of computation are still wasted on relaxing and determining formation enthalpies for prototypes structures that end up being unstable. A model that can be used to triage and select which relaxations to carry out would allow for cheaper and therefore more extensive expansion of high-throughput databases. However, building such a model hinges on identifying model input features that are sufficiently informative to allow for accurate predictions for the expected formation enthalpy, able to discriminate between polymorphs, and are cheap to enumerate for a novel design space. We propose that using model inputs derived from the concept of a Wyckoff set offers a promising compromise.

2.2 Wyckoff Set Regression

In crystallography we can completely specify a materials crystal structure with a combination of: 1) the spacegroup of the structure, 2) the dimensions of its unit cell, and 3) a set of Wyckoff positions and the elements that sit on them. Wyckoff positions describe sites that map onto equivalent sites under the symmetry transformations of the given spacegroup, as a consequence a single Wyckoff position can detail the positions of multiple atoms.

Here, we consider anonymised Wyckoff positions where we discard the information about the exact positions (see Figure 1) such that we can enumerate over a design space of potential Wyckoff sequences. Importantly it is possible, even common, for a material structure to contain multiple instances of a given anonymised Wyckoff position i.e. we need a model that can consider multi-sets of anonymised Wyckoff positions.

The key question is given an anomymised Wyckoff set for a material can we train a model that accurately predicts the formation enthalpy per atom? We tackle this as a multi-set regression problem using a message passing neural network architecture based on the *Roost* model [11], we call this

model *Wren* (**W**yckoff **Re**presentation **N**etwork). The principle idea behind the models is to consider all directed pairwise combinations in the set and use them to update the representations of elements in the set via message passing operations [18]. These message passing stages are repeated multiple times before a permutation invariant pooling operation is applied to the set to get a fixed-length representation. A full description of the architecture and update rules of the *Wren* model is given in Appendix A. We also note the existence of a previous model, *abcNN*, that attempted to tackle a similar problem using a sparse descriptor tensor as the input to a CNN-based architecture [19].

2.3 Data

Whilst screening via anonymised Wyckoff sets does not require explicit knowledge of the relaxed crystal structure we do require such knowledge in order to generate the labelled anonymised Wyckoff sets used to the train the model. Here we use structures drawn from the Materials Project catalogue [20] and train the model to predict the formation enthalpy per atom. We restrict the query to materials with less than 64 atoms and 16 Wyckoff positions in their unit cells. We clean the queried data to remove materials with a volume per atom of greater than 500 Å^3. We use the *spglib* python package [21] to assign spacegroups and Wyckoff positions to the structures in the data set, for ease we use the same tolerance values as used by Materials Project (positions to 0.1 Å and angles to 5°). After the cleaning and processing procedure we were left with 94,319 data points which were then randomly partitioned 80:20 into a training set and test set.

3 Results and Discussion

In order to get an understanding of the *Wren* model's performance we benchmark against the reference architectures for *Roost* [11] and *CGCNN* [22] released by their respective authors. The *CGCNN* model is a graph neural network that acts on the extended connectivity graph of the atom positions, the reference architecture considers up to 12 neighbours within 8 Å of each atom in the structure to build this graph. All the models were trained to minimise the L1 loss between the predicted and target formation enthalpies. The Adam optimiser was used with a fixed learning rate of 3×10^{-4}, weight decay parameter of 10^{-6} and mini-batch size of 128 for all the experiments. The models were trained for 200 epochs.

Figure 2 shows a comparison of the models and gives aggregate metrics of their performances. As expected *CGCNN* performs the best out of the three models. However, as it uses relaxed DFT structures to generate model inputs *CGCNN* it cannot be used to screen novel materials where the structures are *a priori* unknown. Whilst *Wren* doesn't quite match the accuracy of *CGCNN*, achieving a mean absolute error (MAE) of 0.07 eV per atom compared to 0.05 eV per atom, it completely avoids any costs associated with DFT when screening as its inputs can be enumerated combinatorially. The inputs to the *Roost* model can also be trivially enumerated, however, its inability to distinguish between polymorphs leads to it being far less accurate (MAE of 0.12 eV per atom). This comparison shows that *Wren* is indeed a compelling approach for accelerating materials discovery ripe for further investigation and development.

Given a structure, determining the Wyckoff sequence is actually a non-trivial task, selecting the tolerances is a careful balance between wanting to minimise the number of structures that approximately satisfy larger numbers of symmetries being assigned to relatively "uninformative" lower symmetry spacegroups (i.e. *P1* (No. 1) and $P\bar{1}$ (No. 2)) whilst equally not wanting to assign symmetries to structures that are not present due to the tolerances being too loose. However, perhaps unintuitively, we see that the model doesn't perform significantly worse on the "uninformative" space groups. While *a priori* we might consider *P1* to be the most "uninformative" spacegroup the model achieves a MAE of 0.10 eV per atom which is comparable to the average test set result of 0.07 eV per atom. In contrast, there are other highly symmetrical spacegroups, where we might expect better performance, that result in higher errors. One example is $Pm\bar{3}m$ (No. 221) where we get a MAE of 0.14 eV per atom despite the fact that it is relatively abundant in the training set with ∼2400 examples. At the other extreme, in *Pbam* (No. 55) and $I\bar{4}$ (No. 82) we achieve MAEs of 0.03 eV per atom with only ∼600 training set examples. The take away from this is that whilst the additional information available through the consideration of Wyckoff positions allows the *Wren* model to deal with polymorhpic freedoms how beneficial this information is for improving the fit of the model depends on the diversity within the training and testing data i.e. the $Pm\bar{3}m$ structures present in Materials Project might be far

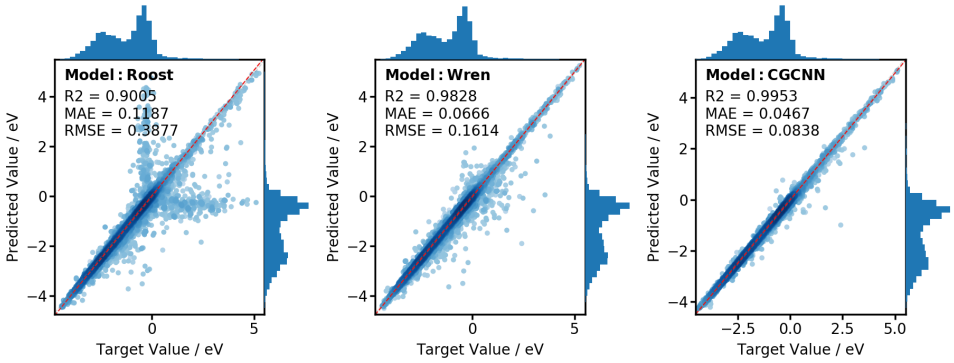

Figure 2: Scatter plots showing model performance on the test set for the *Roost*, *Wren*, and *CGCNN* models. The points are shaded by their log density and marginal histograms are shown to show how the results are distributed. The red dotted lines are robust (Huber) lines of best fit.

more structurally diverse than the *Pbam* structures and therefore making it harder for the model to fit that region of the materials space. Similarly the *P1* structures might have a low structural diversity in the data set such that the lack of symmetrical constraints is unimportant as only a restricted subset of such structures without symmetries have been observed in experiments[1]. This result is supported by the fact that the *Roost* model manages to achieve an error of 0.12 eV per atom on *P1* structures which is comparable to the *Wren* model whilst much larger jumps are seen for other spacegroups e.g. $Fm\bar{3}m$ (No. 225) jumps from 0.04 to 0.25 eV per atom when comparing *Wren* and *Roost* on the test set (This large jump can be attributed to the fact that for many binary compositions both α ($Pm\bar{3}m$) and β ($Fm\bar{3}m$) CsCl-type structures are present in Materials Project).

4 Conclusions and Future Work

The preliminary results of this work are promising with potential to provide a new efficacious alternative to structure prototyping [16, 17] for high-throughput materials discovery workflows. Whilst the model looks promising in retrospective testing prospective testing will be required in the future to confirm the utility of the approach. More broadly, we introduce materials discovery via anonymised Wyckoff sets as a challenging and important set regression task. We hope such a task can draw more attention to both inorganic materials discovery and set regression [18, 23] as a research areas within the machine learning community.

Before any prospective deployment we intend to extend the model to allow for uncertainty estimation. Suitable approaches include the creation of a *Deep Ensemble* [24] as well as extensions to the ensemble idea such as *Multi-SWAG* [25]. Uncertainty estimates would allow the model to be used as part of an active learning workflow allowing which allow for more efficient exploration of large search spaces [26].

In further work, we also hope to refine the pre-processing by augmenting the training set using relabelling operations allowed within the spacegroup (coset representatives of its affine normalizer) to obtain multiple equivalent Wyckoff sets for each structure. Doing this will improve the support provided by the training set without requiring additional data. In addition, we intend to investigate the influence that different symmetry finders have on the model. The symmetry finder is perhaps the most important hyperparameter of the model as the spacegroup and Wyckoff positions assigned vary non-trivially depending on the algorithms and tolerances used by the symmetry finder.

Supplementary work is also needed to investigate the implications of using anonymised Wyckoff sets as inputs. For example; what is the relationship between the number of possible Wyckoff positions in a spacegroup and the number of training examples needed to achieve a low model error? how frequently do distinct dynamically stable polymorphs occur with the same anonymised Wyckoff set and how do such collisions influence model performance?

[1]The majority of the entries in the Materials Project catalogue are derived from experimentally characterised structures recorded in the ICSD.

Broader Impact

Models that are able to alleviate current bottlenecks in *in-silico* materials discovery will play an important role in the development of new technologies. The discovery of new materials is often key to making technologies cheaper and more functional which is necessary if we are to be able to effectively tackle many current global issues e.g. decarbonisation of the economy relies on developing cheap alternatives to fossil fuel-based energy technologies for both energy generation and storage.

Acknowledgements

REAG and AAL acknowledge the support of the Winton Programme for the Physics of Sustainability. AAL acknowledges support from the Royal Society. FAF acknowledges funding from the Swiss National Science Foundation (Grant No. P2BSP2_191736). RA and ASP acknowledge funding from from the Swedish Research Council (VR) (Grant No. 2016-04810). The authors also extend thanks to the teams behind the Bilbao Crystallographic Server and the Materials Project for making their resources freely and easily accessible for academic use.

References

[1] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pages 2224–2232, 2015.

[2] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6):463–477, 2019.

[3] Jonathan Schmidt, Mário RG Marques, Silvana Botti, and Miguel AL Marques. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5(1):1–36, 2019.

[4] James E Saal, Anton O Oliynyk, and Bryce Meredig. Machine learning in materials discovery: Confirmed predictions and their underlying approaches. *Annual Review of Materials Research*, 50, 2020.

[5] Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

[6] Zhenwei Li, James R Kermode, and Alessandro De Vita. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Physical review letters*, 114(9):096405, 2015.

[7] Volker L Deringer, Davide M Proserpio, Gábor Csányi, and Chris J Pickard. Data-driven learning and prediction of inorganic crystal structures. *Faraday Discussions*, 211:45–59, 2018.

[8] Jonathan Vandermause, Steven B Torrisi, Simon Batzner, Yu Xie, Lixin Sun, Alexie M Kolpak, and Boris Kozinsky. On-the-fly active learning of interpretable bayesian force fields for atomistic rare events. *npj Computational Materials*, 6(1):1–11, 2020.

[9] Bryce Meredig, Ankit Agrawal, Scott Kirklin, James E Saal, JW Doak, Alan Thompson, Kunpeng Zhang, Alok Choudhary, and Christopher Wolverton. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Physical Review B*, 89(9):094104, 2014.

[10] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2:16028, 2016.

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

[12] Anthony Yu-Tung Wang, Steven K. Kauwe, Ryan J. Murdock, and Taylor D. Sparks. Compositionally-restricted attention-based network for materials property prediction. *ChemRxiv preprint chemrxiv:11869026*, 2020.

[13] Haiying Liu, Jiucheng Cheng, Hongzhou Dong, Jianguang Feng, Beili Pang, Ziya Tian, Shuai Ma, Fengjin Xia, Chunkai Zhang, and Lifeng Dong. Screening stable and metastable abo3 perovskites using machine learning and the materials project. *Computational Materials Science*, 177:109614, 2020.

[14] Felix Faber, Alexander Lindmaa, O Anatole von Lilienfeld, and Rickard Armiento. Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry*, 115(16):1094–1101, 2015.

[15] Daniel W Davies, Keith T Butler, Adam J Jackson, Andrew Morris, Jarvist M Frost, Jonathan M Skelton, and Aron Walsh. Computational screening of all stoichiometric inorganic materials. *Chem*, 1(4):617–627, 2016.

[16] Geoffroy Hautier, Chris Fischer, Virginie Ehrlacher, Anubhav Jain, and Gerbrand Ceder. Data mined ionic substitutions for the discovery of new compounds. *Inorganic chemistry*, 50(2):656–663, 2011.

[17] Scott Kirklin, James E Saal, Bryce Meredig, Alex Thompson, Jeff W Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *npj Computational Materials*, 1(1):1–15, 2015.

[18] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019.

[19] Ankit Jain and Thomas Bligaard. Atomic-position independent descriptor for machine learning of material properties. *Phys. Rev. B*, 98:214112, Dec 2018.

[20] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.

[21] Atsushi Togo and Isao Tanaka. *Spglib*: a software library for crystal symmetry search. *arXiv preprint arXiv:1808.01590*, 2018.

[22] 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.

[23] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in neural information processing systems*, pages 3391–3401, 2017.

[24] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[25] Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *arXiv preprint arXiv:2002.08791*, 2020.

[26] Turab Lookman, Prasanna V Balachandran, Dezhen Xue, and Ruihao Yuan. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials*, 5(1):1–17, 2019.

[27] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95, 2019.

[28] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1263–1272. JMLR. org, 2017.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

A - Full Model Description

In the *Wren* each Wyckoff position in a given material is represented by a vector. This initial vector is comprised of three parts: 1) a vector representation of the atom type from [27], 2) a OneHot vector representation of the Wyckoff position, and 3) the multiplicity of the Wyckoff position. These initial representations are then multiplied by a n by d learnable weight matrix where n is the size of the initial vector and d is the size of the internal representations used in the model which we set to 64. These internal representations are then updated based on the other Wyckoff species present in the material using a message passing neural network [28]. The mathematical form of the update process is

$$h_i^{t+1} = U_t(h_i^t, \boldsymbol{\nu}_i^t) \tag{1}$$

where h_i^t is the feature vector for the i^{th} Wyckoff position after t updates, $\boldsymbol{\nu}_i^t = \{h_\alpha^t, h_\beta^t, h_\gamma^t, ...\}$ is the set of other Wyckoff positions in the material, and U_t is the Wyckoff position update function for the $t + 1^{th}$ update. For this work, we use a weighted soft-attention mechanism for our Wyckoff position update functions. The first stage of the attention mechanism is to compute unnormalised scalar coefficients, e_{ij}, across pairs of Wyckoff positions in the material.

$$e_{ij}^t = f^t(h_i^t || h_j^t) \tag{2}$$

where $f^t(...)$ is a single-hidden-layer neural network for the $t + 1^{th}$ update, the j index runs over all the Wyckoff positions in $\boldsymbol{\nu}_i^t$, and $||$ is the concatenation operation. The coefficients e_{ij} are directional depending on the concatenation order of h_i and h_j. These coefficients are then normalised using a weighted softmax function where the weights, w_j, are the fractional multiplicities of the Wyckoff positions in the composition,

$$a_{ij}^t = \frac{w_j \exp(e_{ij}^t)}{\sum_k w_k \exp(e_{ik}^t)}. \tag{3}$$

where j is a given Wyckoff position from $\boldsymbol{\nu}_i^t$ and the k index runs over all the Wyckoff positions in $\boldsymbol{\nu}_i^t$. The internal representations are then updated in a residual manner [29] with learnt pair-dependent perturbations weighted by these soft-attention coefficients.

$$h_i^{t+1} = h_i^t + \sum_j a_{ij}^t g^t(h_i^t || h_j^t), \tag{4}$$

where $g^t(...)$ is a single-hidden-layer neural network for the $t + 1^{th}$ update and the j index again runs over all the Wyckoff positions in $\boldsymbol{\nu}_i^t$. The $f^t(...)$ and $g^t(...)$ neural networks use 256 hidden units and LeakyReLU activation functions.

A fixed-length representation for each material is determined via another weighted soft-attention-based pooling operation over all the Wyckoff positions. Finally, these material representations are taken as the input to a feed-forward output neural network to predict the formation enthalpy per atom for the material. The output network used has 5 hidden layers and ReLU activation functions. The number of hidden units in each layer is 1024, 512, 256, 126, and 64 respectively and linear skip connections are added between layers.