Rapid Discovery of Novel Materials by Coordinate-free Coarse Graining

Rhys E. A. Goodall,1 Abhijith S. Parackal,2 Felix A. Faber,1 Rickard Armiento,2,† and Alpha A. Lee1,†

1Department of Physics, University of Cambridge, Cambridge, UK
2Department of Physics, Chemistry and Biology, Linköping University, Sweden

A fundamental challenge in materials science pertains to elucidating the relationship between stoichiometry, stability, structure, and property. Recent advances have shown that machine learning can be used to learn such relationships, allowing the stability and functional properties of materials to be accurately predicted. However, most of these approaches use atomic coordinates as input and are thus bottlenecked by crystal structure identification when investigating novel materials. Our approach solves this bottleneck by coarse-graining the infinite search space of atomic coordinates into a combinatorially enumerable search space. The key idea is to use Wyckoff representations – coordinate-free sets of symmetry-related positions in a crystal – as the input to a machine learning model. Our model demonstrates exceptionally high precision in discovering new stable materials, identifying 1,558 materials that lie below the known convex hull of previously calculated materials from just 5,675 ab-initio calculations. Our approach opens up fundamental advances in computational materials discovery.

I. INTRODUCTION

Finding a needle in a haystack is often used as an analogy for materials discovery. Only a small proportion of viable material compositions (believed to be of the order $10^{10}$ – $10^{100}$ [1]) will have thermodynamically stable polymorphs that are experimentally accessible. Most approaches to tackle this challenge focus on predictive models for materials properties – metaphorical sieves that filter out the hay. Here we seek an alternative approach: Can we significantly cut down materials space by changing how we represent materials – making most of the hay disappear?

Our approach is motivated by a concept ubiquitous in science: coarse-graining. Taking molecular chemistry, for example, chemists typically build intuitions about chemical properties using molecular graphs. Molecular graphs are a coarse-grained representation of molecules with each graph corresponding to a unique ensemble of atomic coordinates. Searching in the enumerable space of molecular graphs, as opposed to the innumerable space of possible atomic coordinates, has enabled the development of powerful computational tools [2, 3] as well as efforts that exhaustively enumerate chemical space [4, 5].

In materials science, however, an analogous coarse-grained representation of crystal structures is missing. Thus, we are left confronting the innumerable search space problem. Composition-based approaches can somewhat overcome this challenge [6–9] but do so at the cost of discarding all information about the crystal structures of the materials being considered. As such, either extensive computational crystal structure searching or lab-based experiments are required to validate predictions.

One avenue to manoeuvre around this challenge has been to explore restricted classes of structure prototypes using novel descriptors e.g. Perovskites [10–12], quaternary Heuslers [13], or Elpasolites [14]. Specifying the prototype avoids the need for crystal structure searching empowering more extensive screening campaigns as the computational cost of validation is greatly reduced.

In this paper, we introduce an approach that generalises these prototype-restricted models by considering Wyckoff representations – coordinate-free sets of symmetry-related positions in a crystal. We use this to develop accurate machine learning models for the formation energy of inorganic materials that can distinguish and rank polymorphs from computably enumerable inputs. We first illustrate our approach by considering the problem of exploring Ti-Zn-N, Zr-Zn-N and Hf-Zr-N ternary systems, showing that our model finds low energy structures in the phase diagram with significantly lower computational effort. We then test the ability of our model to identify novel stable materials across a diverse range of chemistries, showing that it has a precision 170-290% higher than the state-of-the-art. Finally, we develop a materials exploration pipeline that, starting from an initial nucleus of known materials, screens nearby materials space and allows the efficient discovery of new stable materials. We identify 1,558 hitherto unknown materials below the known convex hull of previously calculated materials from just 5,675 ab-initio calculations.

II. RESULTS

A. Wyckoff Representation Regression

Building an accurate machine learning model hinges on identifying model inputs that are sufficiently informative to allow the target variable to be predicted. However, for a machine learning model to be useful in practice, these inputs need to be significantly cheaper to obtain than the cost of labelling data. In the context of material science, previous works have shown that highly accurate machine learning models can be built for the formation energies of inorganic crystals calculated via Kohn-Sham density functional theory (DFT) using the DFT-relaxed crystal structure as the model input [15–18]. Sadly, when it comes to materials discovery, this is a circular workflow because arriving at a DFT-relaxed structure necessitates the calculation of...
Figure 1. **Coarse-graining materials space using Wyckoff representation enables efficient structure search.** A) Schematic showing how a crystal structure can be encoded in a coordinate-free manner either using its Wyckoff representation or its composition. B) Original crystal structures can be recovered from the Wyckoff representation via symmetry constrained relaxations. In contrast, extensive crystal structure searching is needed to find the ground state structure when starting from the composition. C) A machine learning powered materials discovery workflow that takes advantage of the benefits of the proposed Wyckoff representation. From an enumerated library of Wyckoff representations, the workflow uses a machine learning model to predict formation energies for the candidate materials. These predicted formation energies are then compared against the known convex hull of stability. Structures satisfying the required symmetries are then generated and relaxed for materials predicted to be stable. The calculated energies of the relaxed structures can then be compared against the known convex hull to confirm whether the candidate is stable.

The energy using DFT multiple times.

To construct a model that is useful for materials discovery, the cost of DFT needs to be avoided entirely when generating model inputs. Several groups have therefore proposed to use composition-based inputs [6–9], which avoid the upfront need for structure identification. However, the composition is not expressive enough to differentiate polymorphs. This is a significant shortcoming as different polymorphs can have radically different properties, most famously the example of diamond and graphite. In this work, we propose model inputs that can distinguish polymorphs whilst also avoiding the cost of DFT.

In crystallography one way to completely specify the crystal structure of a material is via a combination of: 1) the spacegroup of the structure, 2) the dimensions...
of its unit cell, and 3) a set of Wyckoff positions with the elements that sit on them. The Wyckoff positions describe sites that map onto equivalent sites under the symmetry transformations of the given spacegroup [19]. As a consequence, a single Wyckoff position can encode the positions of multiple atoms. To construct model inputs from sets of Wyckoff positions we discard the information about the exact positions and lattice parameters. In the resulting coordinate-free representation, that we refer to as the Wyckoff representation, each Wyckoff position is simply labelled by a Wyckoff letter and the element at that position. Consequently, as the Wyckoff representation is discrete, it is possible to computationally enumerate Wyckoff representations that represent candidate materials for use in screening campaigns.

The procedure of obtaining the Wyckoff representation from a crystal structure can be viewed as a coarse-graining process that takes us from an unsymmetrised initial parameter space of size $4N + 6$, through the symmetrised Wyckoff position space of size $5M + 6$, to the much smaller coordinate-free space of Wyckoff representations with size $2M$, where $N$ is the number of sites in the unit cell and $M$ is the corresponding number of Wyckoff positions. The backmapping from the coarse-grained Wyckoff representation to the full structure can often be satisfactorily obtained via a single symmetry constrained DFT-relaxation of a prototype structure (see Figure 1).

To use the Wyckoff representation as the input for a machine learning model, we formulate the task of property prediction as a multi-set regression problem. A message-passing neural network architecture based on the Roost architecture [8] is used to do this – the Roost model performs materials property prediction via set regression on the set of elements in a material’s composition.

The principal idea behind the model architecture is to embed the coordinate-free Wyckoff positions of a given material into a vector space. The representations in this embedded space are then updated via message passing operations that consider all directed pairwise combinations of members in the multi-set. The messages propagate contextual information between Wyckoff positions leading to the emergence of material-specific representations. These message passing stages are repeated multiple times before a permutation invariant pooling operation is applied to the multi-set to get a fixed-length representation. These fixed-length representations are then fed into a feed-forward output neural network that returns the models predictions.

In this work, we focus primarily on models that predict the formation energy of inorganic crystalline materials, although the proposed framework and inputs are applicable to any material property. We call the proposed model Wren (Wyckoff Representation regression N). Throughout this work, we train Deep Ensembles consisting of 10 Wren models starting from different random initialisations [20], this allows us to estimate the model uncertainty as well as providing better point estimates. Details of the Wren architecture and the hyper-parameters used are given in the Supplementary Information.

### B. Datasets

#### 1. TAATA Dataset

The first dataset we consider is the TAATA dataset [21] which consists of 3 highly sampled ab-initio phase diagrams for the Hf-Zn-N, Ti-Zn-N and Zr-Zn-N ternary systems. The ternary systems studied in the TAATA dataset were investigated for their potential in piezoelectric devices and energy harvesting applications. The TAATA dataset contains a diverse range of stable and unstable structures for each composition.

After applying a canonicalisation and cleaning treatment (see Supplementary Information) we are left with 3,128 entries over 520 compositions in the Ti-Zn-N phase diagram, 2,722 entries over 449 compositions in the Zr-Zn-N phase diagram, and 3,406 entries over 595 compositions in the Hf-Zn-N phase diagram.

To obtain training and test sets for the TAATA dataset we perform a group split based on composition prototypes (e.g. {Ti, Zr, Hf}ZnN, {Ti, Zr, Hf}ZnN$_4$ etc.) and then in each instance randomly assign two of the chemical systems (if present in the data) to the training set and one to the test set.

#### 2. Materials Project Compatible Datasets

The second data source we consider is the Materials Project (MP) database [22] which is a highly curated database of high-throughput DFT calculations. At the time of access, the Materials Project database contains approximately 140k crystal structures. We apply the same canonicalization and cleaning procedure to this MP dataset as we applied to the TAATA dataset. This process leaves a final MP dataset containing approximately 105k distinct materials.

In addition to the official Materials Project database, we also consider an additional source of Materials Project compatible data from Ref. [23]. In this work, the authors calculated the energies and properties of a large number of crystal structures that were generated through the substitution of elements in known crystal structures with chemically similar elements [24]. We refer to this dataset as the WBM dataset. After deduplication and cleaning the WBM dataset contains approximately 220k materials.

#### C. Exploration of Structurally Diverse Chemical Systems

We start by examining the performance of Wren on the TAATA dataset. To put the Wren model into context we compare against the composition-based Roost model [8], which has been shown in a recent benchmark [25] to outperform other composition-based machine learning models for formation energy prediction. The architectural similarity between Roost and Wren also allows us to attribute any significant differences in model performance to the change in inputs. Figure 2 presents parity plots for both models showing the predicted energies against their DFT-calculated val-
D. Selecting Stable Materials from Diverse Chemical Spaces

To accelerate the screening of materials space for novel stable materials, a model must reduce the expected number of calculations needed to find a candidate below the known convex hull (here taken to be the convex hull of the MP dataset before cleaning). Using the MP and WBM datasets we can interrogate our model’s ability to screen for novel stable materials. To do this, we make predictions for the formation energies of the materials contained in the WBM dataset using a \textit{Wren} model trained on the MP dataset. We then assess how well the \textit{Wren} model selects potentially stable materials from the WBM dataset.

The relevant metrics are; the prevalence – the proportion of materials below the known convex hull (actual positives) in the entire screening space, the precision – how many of the predictions of potentially stable materials are correct (i.e. the ratio of true predicted positives to the total predicted positives), and the recall – how many of the actual materials below the known convex hull are found (i.e. the ratio of true predicted positives to actual positives). The ratio of the precision and the prevalence gives the enrichment factor or speed-up from using \textit{Wren}.

When calculating these metrics we exclude materials in the WBM dataset containing lanthanides and actinides. Approximately 51% of materials in the WBM data contain lanthanides or actinides compared to 27% of the MP dataset before cleaning. The preva-
Figure 3. *Wren* accelerates the recovery of low energy structures in unseen chemical systems. The figures show how the enrichment factor varies as we use *Wren* to direct exploration of the Ti-Zn-N, Zr-Zn-N and Hf-Zn-N chemical systems. The enrichment factor is the ratio of candidates found satisfying a given triage criterion to the number we would expect to have found via a random search. The enrichment factor is plotted for candidates within 10, 20 and 30 meV per atom from the convex hull of the full explored system. A light-grey guideline is included to show the performance expected from a random model – an enrichment factor of 1. The plots demonstrate that using *Wren* leads to a significant degree of early enrichment of low energy structures.

Figure 4. *Wren* efficiently identifies stable structures when screening diverse chemical spaces. Histogram of the energy to the convex hull for materials in the WBM dataset. The histogram is broken down into true positives, false negatives, false positives, and true negatives based on whether the *Wren* model predicts candidates to be below the known convex hull. *Wren* exhibits a high recall with the majority of materials below the convex hull being correctly identified by the model.

Figure 4 shows a stacked histogram of the energy to the convex hull of the full MP dataset for the materials from the WBM dataset. The histogram sections are coloured according to whether the model correctly predicts a candidate to be below the known convex hull given its Wyckoff representation. The precision using the *Wren* model to triage calculations is 33%. Consequently, using *Wren* leads to an enrichment factor of 2.7 given the 12% prevalence of the dataset under investigation. This translates into a significant improvement in efficiency allowing many more materials below the known hull to be identified with limited computational resources. We also observe a high recall of 76% meaning that *Wren* misses relatively few potentially computationally stable materials.

The screening performance of the model can be tuned by adjusting our triage criteria. For example, an alternative triage criterion would be to require that \( \Delta \hat{E}_{\text{Hull}} + \hat{\sigma} < 0 \) where \( \Delta \hat{E}_{\text{Hull}} \) is the predicted distance of a candidate material from the known convex hull and \( \hat{\sigma} \) is the predictive uncertainty of the model. This uncertainty adjusted criterion encourages the model to suggest candidates it is more certain about, leading to an increase in the precision from 33% to 47%, but a fall in the recall from 76% to 47%. Consequently, the choice of triage criteria should depend on the aims of a given workflow – the expected opportunity cost of false negatives vs false positives.

The strong performance of *Wren* can be explained by looking at how the mean absolute error changes as a function of the distance from the known convex hull. Figure 5 shows that near to the stability threshold, \( \Delta \hat{E}_{\text{Hull}} = 0 \), *Wren* makes highly accurate predictions of the formation energy. Larger errors are seen for materials far above and far below the hull. However, in these regions, the average error is less than the energy to the convex hull meaning that the model’s classifications are still reliable. The large errors far above the hull are due to the routine underestimation of the formation energy of unstable structures. This underestimation is a manifestation of a bias in the MP dataset towards structures with low formation energies. The bias arises from the fact that large numbers of the initial structures in the MP dataset are sourced from the ICSD [32]. This result highlights the importance of negative results for building generally applicable machine learning models [33–35].
Figure 5. Wren’s average error is below DFT error in the region around the stability threshold. Rolling mean absolute error on the WBM dataset as the energy to the convex hull is varied. The highlighted region shows the area in which the average absolute error is greater than the energy to the known convex hull – this is the region where the model is most at risk of miss-classifying structures. However, we see that in the majority of this region the model’s accuracy is well below the 100 meV per atom threshold considered to be the accuracy of semi-local DFT across diverse chemistries [28] and comparable to the ∼ 50 meV per atom threshold characteristic of fitted correction schemes [29–31]. This high level of accuracy around the stability threshold leads to the model’s impressive performance when identifying materials below the known convex hull.

E. Computational Prospecting for Novel Stable Materials

Having established the promise of Wyckoff representation regression in predicting the stability of unseen materials, we deploy Wren on the prospective challenge of discovering new stable materials. For this stage, we trained Wren on the union of the MP and WBM datasets. This combined dataset contains approximately 323k materials after canonicalisation and cleaning. We randomly sampled 5% of the dataset to use as a test set and trained on the remaining 95%. The resulting model has a mean absolute error of 31 meV per atom on this test set. The model’s accuracy as a function of training set size is shown in the Supplementary Information, revealing a power-law relationship. Reassuringly, the model does not appear to saturate in performance, suggesting that the representation is rich enough and further increases in model performance can be unlocked given more data.

Whilst the coarse-grained space of Wyckoff representations is computably enumerable and far smaller than the infinite space of atomic coordinates, attempting materials discovery by exhaustively screening all possible Wyckoff representations is computationally inefficient as the prevalence of stable materials remains vanishingly low even in the coarse-grained space. To construct a design space with a higher expected prevalence [13] we draw inspiration from previous work [23, 36] and generate candidates for screening by making elemental substitutions in crystal structures that are near to the known convex hull.

To obtain our substitution probabilities we extracted 39,164 ordered structures from the ICSD [32, 37] and binned them according to their Wyckoff representations. Within each prototype, all pairs of structures are compared and we count which element substitutions (including self substitutions) would be needed to change one structure into the other [24]. We only consider substitutions where all Wyckoff positions sharing one element-type are changed simultaneously and not per position substitutions. Once normalised, the rows of the count matrix can be interpreted as substitution probabilities for each element.

Using these data-mined probabilities, we generated a screening library of materials by substituting different elements into structures taken from the MP dataset. We only consider initial structures from the MP dataset with energies above the convex hull less than 100 meV.
per atom. This choice of this threshold means that we should be including most metastable structures within the MP dataset. We consider 10 different substitutions for each initial structure. Candidates that have the same composition as materials already present in the union of the MP and WBM datasets are removed from the library. Lanthanide and actinide-based materials and materials containing Noble elements were also excluded. This workflow produced a screening library of approximately 415k candidates.

Despite constraining our screening set to be close to known materials, it is likely that we are still asking models to make predictions in areas of materials space where it lacks support from the training data. As shown on the WBM dataset, uncertainty estimation allows us to reduce the risk in our materials screening process by factoring our model’s uncertainty into our triage criterion. For simplicity we use the same simple uncertainty adjusted criterion considered previously; $\Delta \hat{E}_{\text{Hull}} + \hat{\sigma} < 0$. In total 5,675 candidates satisfied this screening criterion.

We ran all selected candidates through a high-throughput DFT workflow which resulted in 4,721 completed calculations. Of these, 1,558 were confirmed to be below the convex hull of the union of the MP and WBM datasets. The precision for the completed calculations was 33%, confirming the workflow’s ability to significantly accelerate materials discovery.

III. CONCLUSION

In this work, we introduced the concept of using coarse-graining to accelerate materials discovery. We developed the framework of Wyckoff representation regression, $\text{Wren}$, and applied it to predict the formation energy of materials. $\text{Wren}$ collapses an infinite search space of atomic coordinates into a combinatorially enumerable search space enabling efficient data-driven exploration of materials space. On a set of challenging tasks curated from the literature, we find that our approach can accurately map the phase diagrams of unseen chemical systems, and is up to 4x more precise at predicting the stability of materials.

We developed a materials prospecting pipeline using $\text{Wren}$. As a proof-of-concept, we identified 1,558 new materials below the known convex hull from just 5,675 calculations. These results demonstrate that leveraging Wyckoff representation regression allows for more efficient and extensive expansion of computational material science databases. Such efforts are crucial to expedite the search for a wide variety of industrially desirable materials required for the transition to a low-carbon economy, e.g. thermoelectrics [38], piezoelectrics [21], fast-ion conductors [39], high voltage multi-valent cathode materials [40], and caloric materials [41].

DATA AVAILABILITY

The TAATA dataset is available on request from the original authors [21].
[1] 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.

[2] 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.

[3] Jessica Vanatthevan, 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.

[4] 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.

[5] Jean-Louis Reymond. The chemical space project. Accounts of Chemical Research, 48(3):722–730, 2015.

[6] 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.

[7] 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.

[8] Rhys EA Goodall and Alpha A Lee. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. Nature communications, 11(1):1–9, 2020.

[9] Anthony Yu-Tung Wang, Steven K Kauwe, Ryan J Murdock, and Taylor D Sparks. Compositionally restricted attention-based network for materials property predictions. npj Computational Materials, 7(1):1–10, 2021.

[10] Jonathan Schmidt, Jingming Shi, Pedro Borlido, Liming Chen, Silvana Botti, and Miguel AL Marques. Predicting the thermodynamic stability of solids combining density functional theory and machine learning. Chemistry of Materials, 29(12):5090–5103, 2017.

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

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

[13] Kyungwook Kim, Logan Ward, Jiangang He, Amar Krishna, Ankit Agrawal, and C Wolverton. Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary heusler compounds. Physical Review Materials, 2(12):123801, 2018.

[14] Felix Faber, Alexander Lindmaa, O Anatoile 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] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet—a deep learning architecture for molecules and materials. The Journal of Chemical Physics, 148(24):241722, 2018.

[16] 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.

[17] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. Chemistry of Materials, 31(9):3564–3572, 2019.

[18] Cheol Woo Park and Chris Wolverton. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. Physical Review Materials, 4(6):063801, 2020.

[19] Ralph Walter Graysom Wyckoff. The Analytical Expression of the Results of the Theory of Space-groups, volume 318. Carnegie institution of Washington, 1922.

[20] 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.

[21] Christopher Tholander, CBA Andersson, Rickard Armiento, Ferenc Tasnadi, and Björn Alling. Strong piezoelectric response in stable tizn2, zrzn2, and hfzn2 found by ab initio high-throughput approach. Journal of Applied Physics, 120(22):225102, 2016.

[22] Anubhav Jain, Shyue Ping Ong, Geoffrey 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.

[23] Hai-Chen Wang, Silvana Botti, and Miguel AL Marques. Predicting stable crystalline compounds using chemical similarity. npj Computational Materials, 7(1):1–9, 2021.

[24] Henning Glawe, Antonio Sanna, EKU Gross, and Miguel AL Marques. The optimal one dimensional periodic table: a modified tetragonal chemical scale from data mining. New Journal of Physics, 18(9):093011, 2016.

[25] Christopher J Bartel, Amalie Trewartha, Qi Wang, Alex Dunn, Anubhav Jain, and Gerbrand Ceder. A critical examination of compound stability predictions from machine-learned formation energies. npj Computational Materials, 6(1):97, 2020.

[26] Chris J Pickard and RJ Needs. Ab initio random structure searching. Journal of Physics: Condensed Matter, 23(5):053201, 2011.

[27] Yanchao Wang, Jian Lv, Li Zhu, and Yanming Ma. Calypso: A method for crystal structure prediction. Computer Physics Communications, 183(10):2063–2070, 2012.

[28] Scott Kirklin, James E Saal, Bryce Meredig, Alex Thompson, Jeff W Doak, Muratathan Akyol, 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.

[29] Vladan Stevanović, Stephan Lany, Xiuwen Zhang, and Alex Zunger. Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies. Physic
Rico Friedrich, Demet Usanmaz, Corey Oses, Andrew Supka, Marco Fornari, Marco Buongiorno Nardelli, Cormac Toher, and Stefano Curtarolo. Coordination corrected ab initio formation enthalpies. npj Computational Materials, 5(1):1–12, 2019.

Amanda Wang, Ryan Kingsbury, Matthew McDermott, Matthew Horton, Anubhav Jain, Shyue Ping Ong, Shyam Dwaraknath, and Kristin Persson. A framework for quantifying uncertainty in dft energy corrections, May 2021.

G Bergerhoff, ID Brown, F Allen, et al. Crystallographic databases. International Union of Crystallography, Chester, 360:77–95, 1987.

Paul Raccuglia, Katherine C Elbert, Philip DF Adler, Casey Falk, Malia B Wenny, Aurelio Mollo, Matthias Zeller, Sorelle A Friedler, Joshua Schrier, and Alexander J Norquist. Machine-learning-assisted materials discovery using failed experiments. Nature, 533(7601):73–76, 2016.

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.

MK Horton, S Dwaraknath, and KA Persson. Promises and perils of computational materials databases. Nature Computational Science, 1(1):3–5, 2021.

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.

Mariette Helenbrandt. The inorganic crystal structure database (icsd)—present and future. Crystallography Reviews, 10(1):17–22, 2004.

Shidong Wang, Zhao Wang, Wahyu Setyawan, Natalio Mingo, and Stefano Curtarolo. Assessing the thermoelectric properties of sintered compounds via high-throughput ab-initio calculations. Physical Review X, 1(2):021012, 2011.

Austin D Sendek, Qian Yang, Ekin D Cubuk, Karel-Alexander N Duerloo, Yi Cui, and Evan J Reed. Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials. Energy & Environmental Science, 10(1):306–320, 2017.

Pieremanuele Canepa, Gopalakrishnan Sai Gautam, Daniel C Hannah, Rahul Malik, Miao Liu, Kevin G Gallagher, Kristin A Persson, and Gerbrand Ceder. Odyssey of multivalent cathode materials: open questions and future challenges. Chemical reviews, 117(5):4287–4341, 2017.

Nikolai A Zarkevich, Duane D Johnson, and Vitalij K Pecharsky. High-throughput search for caloric materials: the calorcool approach. Journal of Physics D: Applied Physics, 51(2):024002, 2017.

Shyue Ping Ong, Shreyas Cholia, Anubhav Jain, Miriam Brafman, Dan Gunter, Gerbrand Ceder, and Kristin A Persson. The materials application programming interface (api): A simple, flexible and efficient api for materials data based on representational state transfer (rest) principles. Computational Materials Science, 97:209–215, 2015.
SI: Rapid Discovery of Novel Materials by Coordinate-free Coarse Graining

Rhys E. A. Goodall,1 Abhijith S. Parackal,2 Felix A. Faber,1 Rickard Armiento,2 and Alpha A. Lee1

1 Cavendish Laboratory, University of Cambridge, Cambridge, UK
2 Department of Physics, Chemistry and Biology, Linköping University, Sweden

I. METHODS

A. Wren Model Architecture

The bulk of the Wren architecture directly mimics that of Roost [1] and we refer the readers there for an in-depth description of how the message passing is formulated. The principle difference between the two architectures comes in that the nodes on the dense graph now represent different Wyckoff positions rather than just the different elemental species. The elemental information is encoded using the ‘matscholar’ embedding from [2] which has a dimensionality of \( d_{\text{el}} = 200 \). The remainder of the node embedding comprises the Wyckoff position embedding (described below) plus the fractional multiplicity of that Wyckoff position within the unit cell. The combined dimensionality of the Wyckoff proportion of the embedding is \( d_{\text{wyk}} = 445 \).

To reduce the total dimension of the node embeddings, we project both the elemental and Wyckoff embeddings into lower-dimensional spaces using learnt affine transformations. The low dimension embeddings are then concatenated to give the node embeddings. In this work we chose \( d_{\text{el}}^\text{ Wyck} = 32 \) and \( d_{\text{wyk}}^\text{ Wyck} = 32 \) giving a total dimensionality of \( d = 64 \) for the node embeddings.

We use 3 message passing layers each with a single attention head. We chose single-hidden-layer neural networks with 256 hidden units and LeakyReLU activation functions for both parts of the soft-attention mechanism. The output network consists of a feed-forward neural network with skip connections and ReLU activation functions. The output network used has 4 hidden layers containing 256, 256, 128, and 64 hidden units respectively.

B. Wyckoff Position Embedding

In total across the 230 crystallographic spacegroups in 3D there are 1731 different Wyckoff positions. The embedding we use is made up of 3 parts: a one-hot encoding of the crystal system (of which there are 6), a one-hot encoding of the Bravais lattice centring (of which there are 5), and an encoding constructed from the sum of multi-hot encodings of the equivalent sites within a given Wyckoff position. To construct the multi-hot encodings we first collate all the sites within all the allowed Wyckoff positions as recorded on the Bilbao crystallographic server [3]. Each site can be represented as a 3x4 matrix consisting of a 3x3 matrix encoding the algebraic terms (whether the position corresponds to a line, a plane, etc.) and a 3x1 matrix encoding the offset, e.g.

\[
(-x + y + 1/4, y, z + 3/4) = \begin{bmatrix} -1 & 1 & 0 & 1/4 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 3/4 \end{bmatrix},
\]

From this, we construct separate one-hot encodings for the unique algebraic and unique offset positions. We end up with 185 unique algebraic positions and 248 unique offset positions. A Wyckoff position is then represented by a sum of the embeddings of all the allowed sites. The resulting embedding has a dimensionality of 444 and 1038 out of 1731 of the Wyckoff positions are uniquely encoded.

C. Invariance to Equivalent Wyckoff Representations

The categorisation of Wyckoff positions depends on a choice of origin [4]. As such, there is not a unique mapping between the crystal structure and the Wyckoff representation. To ensure the model is invariant to the choice of origin we perform on-the-fly augmentation of Wyckoff positions with respect to this choice of origin. The augmented representations are then averaged at the end of the message passing stage to provide a single representation of equivalent Wyckoff representations to the output network. By pooling at this point we ensure that the model is invariant and that its training is not biased towards materials for which many equivalent Wyckoff representations exist.

D. Evaluation of Wyckoff Positions

For this work, we primarily make use of ‘spglib’ [5] to determine the spacegroup and Wyckoff positions for the structures in the datasets. We set the tolerance thresholds as 0.1 Å for positions and 5° for angles (Note, these are the same tolerances as used in Materials Project to calculate the spacegroup). We note that symmetry finders with adaptive tolerances [6] could potentially provide more reliable identification of the spacegroup and Wyckoff positions, in turn, reducing the number of structures erroneously discarded during our cleaning processes. However, preliminary investigations showed that the amounts of data discarded as a result of differences between symmetry finders did not significantly impact model accuracy. As a consequence ‘spglib’ was picked over other symmetry finders due to its speed.

E. Roost Model Architecture

The Roost models trained for this work used the architectural hyper-parameters described in Ref. [1].
F. Model Training

Throughout this work, we train Deep Ensembles of 10 models starting from different random initialisations for each data setup and architecture considered.

All the models examined in this work were trained using the AdamW optimiser [7] with a fixed learning rate of $3 \times 10^{-4}$. A mini-batch size of 128 and a weight decay parameter of $10^{-6}$ were used for all the experiments. The models were trained for 400 epochs (cycles through the training set).

Deep ensembles require the use of a proper scoring rule for training. Therefore, we train all models to minimise the following robust L1 loss function which is an example of a proper scoring rule for regression [8, 9],

$$
\mathcal{L} = \sum_i \sqrt{2} \frac{\|y_i - \hat{\mu}_a(x_i)\|_1 + \log(\hat{\sigma}_a(x_i))}{\hat{\sigma}_a(x_i)^2}
$$

(2)

where $\hat{\mu}_a(x_i)$ and $\hat{\sigma}_a(x_i)^2$ are a predictive mean and predictive aleatoric variance outputted by the model, and $y_i$ is the target label.

The expectations, $\hat{y}(x_i)$, and epistemic uncertainties, $\hat{\sigma}_e(x_i)^2$ from the ensemble are calculated as

$$
\hat{y}(x_i) = \frac{1}{W} \sum_w \hat{\mu}_{\theta_w}(x_i),
$$

(3)

$$
\hat{\sigma}_e^2(x_i) = \frac{1}{W} \sum_w \left( \hat{y}(x_i) - \hat{\mu}_{\theta_w}(x_i) \right)^2,
$$

(4)

where the index $w$ runs over the $W$ members of the ensemble. The total uncertainty of the ensemble expectation is simply the sum of the epistemic contribution and the average of the aleatoric contributions from each model in the ensemble.

$$
\hat{\sigma}^2(x_i) = \hat{\sigma}_e^2(x_i) + \frac{1}{W} \sum_w \hat{\sigma}_a^2(\theta_w)(x_i).
$$

(5)

G. Canonicalisation and Cleaning

All the data used to train models in this work went through a canonicalisation and cleaning process. Tables SI, SII and SIII show how much data is discarded at each stage.

The check for valid Wyckoff representations looks at whether the multiplicities of the Wyckoff positions returned correctly sum up to give the composition of the material. Invalid Wyckoff representations are believed to arise due to overlapping Wyckoff positions within the tolerance of the symmetry finder used. As the number of invalid Wyckoff representations is small these structures were simply discarded for this work. As such this is an area of data leakage that might be improved by using a symmetry finder with an adaptive tolerance scheme.

The canonicalisation stage removes higher energy structures that have equivalent Wyckoff representations to other structures in the dataset. The majority of removed structures by canonicalisation are triclinic. This is because the lack of symmetries in triclinic systems results in many distinct structures mapping to the same Wyckoff representation.

When we take the union of the MP and WBM datasets we remove duplicates from the WBM dataset that have since been included in the MP dataset. For duplicated elemental structures we kept the structures recorded in the MP dataset to ensure that our endpoints for the calculation of formation energies are consistent. During deduplication 2,446 additional materials were removed from the MP dataset leaving 322,915 materials in the union of the WBM and MP datasets.

H. Density Functional Theory Settings

The validation of predictions in our materials prospecting pipeline was carried out using Kohn-Sham DFT with the plane wave pseudopotential code VASP [10, 11]. Projector augmented wave (PAW) type pseudopotentials [12, 13] were used with the Perdew-Burke-Ernzerhof (PBE) generalised gradient approximation exchange-correlation functional [14]. All calculations were done using a 520 eV plane-wave energy cutoff. The pseudopotentials and Hubbard-\(U\) values were selected to ensure compatibility with data contained in the Materials Project. The Materials Project correction scheme was applied to allow the mixing of GGA and GGA+\(U\) calculations [15]. We used the High-
Throughput Toolkit (httk v1.0) introduced in Ref. [16] to manage the calculations.

II. ADDITIONAL DISCUSSION

A. Structure-based Models for Materials Discovery

Due to the inability to generate relaxed structures without incurring the costs of DFT, we cannot use relaxed structures as the inputs to structure-based models in materials discovery applications. Instead, using a structure-based model for materials discovery typically involves approximating the energies of relaxed structures using predictions made on unrelaxed prototype structures. This is a qualitatively different task to the task that structure-based models are trained to perform. In contrast, when using Wren the training and testing tasks are consistent by design. Despite this domain shift, Park et al. have used this approach alongside their “improved” CGCNN model to accelerate the discovery of novel stable compounds within the ThCr$_2$Si$_2$ structure-type [17] achieving a precision of 13%. These results show that structure-based models can still be useful even with the apparent issues in the problem setup.

Accordingly, several attempts have been made to improve the performance of structure-based models when making predictions on unrelaxed structures. The symmetry-labelled Voronoi graph convolutional neural network of [18] tries to do this by discarding explicit distance information. Instead, edges in the graph are labelled by the approximate symmetry of the Voronoi facets. This leads to a smaller deterioration in accuracy than seen for distance-based models. More recently, the BOWSR algorithm [19] has been introduced to directly minimise the domain shift between unrelaxed and relaxed structures using a surrogate model. The algorithm uses Bayesian optimisation to adjust the free parameters of unrelaxed symmetrised structures according to a trained structure-based energy model. This process yields pseudo-relaxed structures and results in smaller errors in materials discovery tasks.

B. Extrapolation in Material Space

The WBM dataset was generated using an iterative workflow where successful candidates from the first batch were included when generating candidates for the second batch. As a consequence, candidates considered in later batches are likely to be less similar to materials contained in the MP dataset. We use this stratification to probe how Wren’s performance changes as it is asked to make larger and larger extrapolations. Figure S1 shows split violin plots for the absolute errors and model uncertainties from the Deep Ensemble on each of the different batches. We observe a slight upwards trend in the quartiles of the absolute error distributions going to later batches. This is in keeping with our expectation that model performance should degrade the more it is required to extrapolate.

Figure S1. Violin plots of absolute error and uncertainty of the Wren model on the WBM dataset batches. Distributions of both the absolute errors and predictive uncertainty of the Wren model trained on the MP dataset when used to screen the different batches within the WBM dataset. The widths of the violins are scaled according to the number of points in each batch. We see a general trend that the average error of the model increases for the later batches. This agrees with the expectation that model performance should deteriorate when asked to make significant extrapolations.

Figure S2. Rolling MAE of the Wren model on the WBM dataset batches. Rolling mean absolute error on the batches of the WBM dataset as the energy to the convex hull is varied. A windowing period of 40 meV per atom is used when calculating the rolling average. The curves show that later batches, believed to be more likely to be chemically dissimilar to the training data, incur higher average errors.

Additionally, we can probe this same concept by looking at how the mean absolute error changes as a function of the distance from the convex hull for the different batches. In Figure S2 we see that, whilst the overall shape of the curves remains the same as the V-shape seen for the full dataset, there is an offset between the batches with later batches incurring higher errors on average. This suggests that iteratively retraining the model as more data is acquired should shrink the model’s extrapolation error leading to improved perfor-
Figure S3. Learning curve for the Wren model with increasing training data. Learning curves for both single and Deep Ensembles of Wren models. The curves are produced by plotting the error on a fixed test set (here 5% of the union of the MP and WBM datasets) as the amount of data used to train the model is increased. A grey guideline shows the size of the test set. The learning curve shows that a power-law relationship exists between the amount of data used to train the model and the MAE of the trained model.

A benefit of using a Deep Ensemble [20], beyond better point predictions, is that we can get an estimate of the predictive uncertainty of the model. Looking at the quartiles of the uncertainty distributions compared to the distributions of absolute errors in Figure S1 shows that the uncertainty is typically larger than the error when the error is small but smaller than the error when the error is large. This discrepancy indicates that the model uncertainty is miscalibrated as is typical of neural network-based models [9, 21]. It is important to note, as demonstrated in the main manuscript when using the uncertainty adjusted triage criterion, that having miscalibrated uncertainties does not preclude them from being useful in downstream applications.

C. Learning Efficiency

Applications in material science, particularly the investigation of functional properties of materials, often encounter issues with data scarcity. Given this the learning efficiencies of models used in material science applications are of critical importance. The best way to probe this is through the construction of learning curves [22, 23]. A learning curve depicts the error of a model on a fixed test set as the amount of data used to train the model is varied. We see in Figure S3 that both individual Wren models and Deep Ensembles of Wren models follow unbroken power law relationships. Consequently, we would expect as more data is obtained the accuracy of the proposed model should continue to improve.
Persson, and Gerbrand Ceder. Formation enthalpies by mixing gga and gga+u calculations. Physical Review B, 84(4):045115, 2011.

[16] Rickard Armiento. Database-Driven High-Throughput Calculations and Machine Learning Models for Materials Design, pages 377–395. Springer International Publishing, Cham, 2020.

[17] Cheol Woo Park and Chris Wolverton. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. Physical Review Materials, 4(6):063801, 2020.

[18] Peter Bjørn Jørgensen, Estefanía Garijo del Río, Mikkel N Schmidt, and Karsten Wedel Jacobsen. Materials property prediction using symmetry-labeled graphs as atomic position independent descriptors. Physical Review B, 100(10):104114, 2019.

[19] Yunxing Zuo, Mingde Qin, Chi Chen, Weike Ye, Xiangguo Li, Jian Luo, and Shyue Ping Ong. Accelerating materials discovery with bayesian optimization and graph deep learning. arXiv preprint arXiv:2104.10242, 2021.

[20] 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.

[21] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of Proceedings of Machine Learning Research, pages 1321–1330. PMLR, 2017.

[22] K-R Müller, Michael Finke, Noboru Murata, Klaus Schulten, and Shun-ichi Amari. A numerical study on learning curves in stochastic multilayer feedforward networks. Neural Computation, 8(5):1085–1106, 1996.

[23] Felix A Faber, Anders S Christensen, Bing Huang, and O Anatole von Lilienfeld. Alchemical and structural distribution based representation for universal quantum machine learning. The Journal of Chemical Physics, 148(24):241717, 2018.