Machine Learning Modeling of Materials with a Group-Subgroup Structure

Prakriti Kayastha and Raghunathan Ramakrishnan
Tata Institute of Fundamental Research, Centre for Interdisciplinary Sciences, Hyderabad 500107, India
(Dated: January 20, 2021)

A cornerstone of materials science is Landau’s theory of continuous phase transitions. Crystal structures connected by Landau-type transitions are mathematically related through group-subgroup relationships. We introduce “group-subgroup learning” and show training on small unit cell phases of materials to decrease out-of-sample errors for modeling larger phases. The proposed approach is generic and independent of the ML formalism, descriptors, or datasets; and extendable to other symmetry abstractions such as spin-, valency-, or charge order. Since available materials datasets are heterogeneous with too few examples for realizing the group-subgroup structure, we present the “FriezeRMQ1D” dataset of 8393 Q1D organometallic materials uniformly distributed across seven frieze groups and provide a proof-of-the-concept. For these materials, we report < 3% error with 25% training with the Faber–Christensen–Huang–Lilienfeld descriptor and compare its performance with a fingerprint representation that encodes materials composition as well as crystallographic Wyckoff positions.

I. INTRODUCTION

The upsurge in the use of machine learning (ML) modeling in computational chemistry or materials science is because ML models, once trained sufficiently well on computed properties, can deliver accurate new predictions at a cost lower than that of the reference method by orders of magnitude[1][2]. Developments of such novel, data-driven methods are fueled by pioneering efforts in designing and generating big data using high-throughput computation[3][9]. While for molecules such campaigns aim at complete coverage of synthetically feasible chemical compound space[10], for materials, the coverage is mostly inspired by experimentally known structures[11]. Symmetry-adaption in ML has its merits for potential energy surface modeling[12]. However, across stoichiometries, most of the information about the total electronic energy is encoded in atom-in-molecule-based fragments[13] that seldom show any symmetry relationships with the molecule.

In this study, we propose the group-subgroup learning approach for ML modeling on materials datasets containing multiple crystal structures related through a phase transition. Landau applied group theory for understanding continuous phase transition of material from a phase of high-symmetry space group, \(G_h\) to a phase of low-symmetry space group, \(G_l\), which is a subgroup, \(G_l \subset G_h\)[14]. Landau theory states that symmetry breaking occurs through a collective variable that transforms according to a single irreducible representation of \(G_l\)[15]. Systematic development of structural relationships between different crystal structures was done by Bärnighausen using crystallographic group–subgroup relations[16]. These relationships are represented in a graphical tree, (Bärnighausen-tree or \(B\)-tree), with the most symmetric space group placed at the top. A \(B\)-tree not only encodes group-subgroup relations but it also correlates the Wyckoff positions during symmetry reduction[16]. The so-called translationengleiche subgroups (t-subgroups) are related to their supergroup by a decrease in the point group symmetry of the lattice and conservation of the translation symmetry. A group can also have klassengleiche subgroups (k-subgroups), retaining the point group symmetry but the translational symmetry is compromised[16][17]. A material’s cohesive energy that largely arises from intra-unit cell atomic interactions in a group should be similar in its t-subgroups and t-supergroups. This similarity must be maximum when going from a group to its maximal subgroup or minimal supergroup.

For a target accuracy, ML modeling of materials belonging to a space group requires a smaller training set size compared to training on materials across space groups[13]. When modeling on a space group with a large unit cell, the proposed group-subgroup learning approach facilitates both a reduction in the training set size and instantaneous generation of the descriptor for a query. The key features of group-subgroup learning are: 1) For modeling on a class of materials belonging to one or more groups with large unit cells, the training set includes phases of the same materials with smaller unit cells. For conventional unit cells, moving from a group to its t-subgroup decreases the unit cell size by a factor, \(n\), which equals the ratio between orders of the group and its subgroup. 2) Query on large unit cell materials uses a prototype descriptor made of equilibrium geometries of smaller unit cells. For this, one applies the splitting pattern of the Wyckoff positions (listed in the \(B\)-tree) during symmetry lowering from \(G_h\) to its subgroup \(G_l\). Generating such commensurate prototype unit cells also ensures the descriptor size to be homogeneous to establish a faithful correlation with cohesive energy per atom, a size-intensive quantity.
II. RING-METAL Q1D MATERIALS IN SEVEN CRYSTALLOGRAPHIC FRIEZE GROUPS

Our dataset consists of Q1D materials comprising a ring-metal pair as the formula unit. In a previous study, we reported geometries, electronic and phonon properties of these materials with 11 monovalent metals (Na–Cs, Cu–Au, Al–Tl) and 109 heterocyclic rings generated by combinatorial substitution of C atoms in the cyclopentadienyl ring with B, N or S atoms of all possible valencies[19]. In this work, we extend the dataset by optimizing the lattice geometries according to 7 frieze group symmetries. In a frieze pattern, commonly encountered in architecture or textile design, a two-dimensional unit is repetitive in one direction. A frieze group is a set of all elements that define the symmetries of a frieze pattern. For every frieze group, there is an isomorphic crystallographic space group. Since the lateral packing of the FriezeRMQ1D materials are less relevant for their electronic properties, we restrict our attention to only 7 frieze groups rather than 230 space groups. The $B$-tree of the frieze groups showing group-subgroup interrelationships is presented in Figure 1a.

RMQ1D materials belonging of the smallest frieze group symmetry, $p1$, contain repeated ring-metal units. While in the earlier study, we performed a full relaxation of all 1199 FriezeRMQ1D materials in the $p1$ phase, here to conserve computational efforts, we have defined the most relevant internal coordinates involving translations, rotations, and reflections of the formula unit as shown...
TABLE I. Symmetry operations for generating the atomic coordinates for various formula units in a unit cell of polymeric multidecker sandwich complexes for seven frieze group symmetries: \( \hat{T}_j(\xi) \) denotes translation along the \( j \)-axis through a distance \( \xi \), \( \hat{R}_k(\tau) \) denotes two-fold rotation w.r.t. the \( k \)-axis, \( \sigma_{ij} \) denotes mirror planes and \( \{r^I_k\} \) denote the set of atomic coordinates belonging to the \( j \)-th unit. Each formula unit contains a five-membered cycle and a metal center. Also given are the number of degrees-of-freedom for the crystal structure relaxation

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Group} & N & \text{Wyckoff positions of formula units} & \text{Symmetry operations} \\
\hline
\alpha 1 & 6 & \{r^I_k\} = \{r_k\} & \\
\alpha 211 & 8 & \{r^I_k\} = T_y(\xi_1)T_z(\xi_2) \{r_k\} & \{r^I_k\} = \hat{R}_x(\pi) \{r^I_k\} \\
\alpha 1m1 & 7 & \{r^I_k\} = T_z(\xi) \{r_k\} & \{r^I_k\} = \hat{R}_x(\pi) \{r^I_k\} \\
\alpha 11m & 7 & \{r^I_k\} = T_y(\xi) \{r_k\} & \{r^I_k\} = \hat{R}_z(\pi) \{r^I_k\} \\
\alpha 11g & 7 & \{r^I_k\} = T_y(\xi) \{r_k\} & \{r^I_k\} = \hat{R}_z(\pi) \{r^I_k\} \\
\alpha 2mm & 8 & \{r^I_k\} = T_y(\xi_1)T_z(\xi_2) \{r_k\} & \{r^I_k\} = \hat{R}_z(\pi) \{r^I_k\} \\
\alpha 2mg & 8 & \{r^I_k\} = T_y(\xi_1)T_z(\xi_2) \{r_k\} & \{r^I_k\} = \hat{R}_z(\pi) \{r^I_k\} \\
\hline
\end{array}
\]

\(^a\) in all cases, the six internal degrees of freedom with in a unit is defined in Figure 1b.
\(^b\) for \( \alpha 11g \) and \( \alpha 2mg \) the Wyckoff position is \((x, y, z + 1/2)\).
\(^c\) for \( \alpha 2mg \) the Wyckoff position is \((x, y, z + 1/2)\).

in Figure 1b. In principle stable phases with a larger unit cell can be realized by beginning with more than one units in the unit cell. However, with a full relaxation, it is likely that the final crystal structure will be of lower symmetry. For finding minimum energy configurations corresponding to 6 larger frieze groups, constrained geometry relaxation is inevitable\(^{20}\). Such constrained optimizations can be performed using atomic coordinates and the Wyckoff positions of the frieze groups. Groups \( \alpha 211, \alpha 1m1, \alpha 11m, \) and \( \alpha 11g \) contain two Wyckoff positions, while the larger groups \( \alpha 2mm \) and \( \alpha 2mg \) contain four Wyckoff positions. The symmetry of these larger groups necessitate additional degrees of freedom to be relaxed as summarized in Table I. The structures of the seven frieze phases resulting from an application of the transformations presented in Figure 1b and Table I are shown in Figure 1c.

It must be noted that the number of atoms in a material’s unit cell is independent of the order of the crystallographic space group\(^{21,22}\). Consider structures in the widely studied perovskite family ABX\(_3\). The most symmetric structure with the smallest unit cell is cubic in the space group, \( Pm\overline{3}m \). A simple mechanism of tilting the rigid octahedral units results in structures belonging to 15 subgroups with larger unit cells\(^ {21}\). A counter example is the Peierls transition in a 1D chain of H-atoms. Here, the uniformly distributed chain belongs to the \( P1 \) group with one atom in the unit cell. Following Peierls distortion, the unit cell contains a H\(_2\) molecule and the system is in \( PT \) space group. Hence, the high-symmetry phase contains fewer atoms in the unit cell only when the phase is more compact (smaller lattice constants) than the low-symmetry phase.

Symmetry constrained geometry relaxations were performed with the all-electron, numerical atom-centered orbital code FHI-aims\(^{25}\) with the PBE\(^{26}\) functional. Since the goal of the present study is to explore the capabilities of the ML approach, rather than presenting

![FIG. 2. Structure-property diversity in the FriezeRMQ1D dataset: (a) Preference for hapticity (\( \eta \)) for intra formula unit metal-ring bonding in FriezeRMQ1D materials is shown through joint variation of the ring-metal distance, \( r \) (in A), and the ring-slip angle, \( \theta \) (in degrees). (b) Distribution of the thermodynamically most stable phase for each stoichiometry across 7 frieze groups shown for 1199 Cu-based materials.](image)
III. KERNEL-RIDGE REGRESSION

We use kernel-ridge regression\textsuperscript{[28, 29]} for modeling atomization energies, $E$, of the FriezeRMQ1D materials. The energy of a query material, $q$, is estimated as the linear combination of kernel functions, each centered on a training material, $t$. The kernel functions take as argument the similarity between the query and the training materials quantified through a descriptor, $d$

$$E^\text{est}(d_q) = \sum_{t=1}^{N} c_t k(d_q, d_t).$$

(1)

Here, we use a Laplacian kernel, $k(d_q, d_t) = \exp(-|d_q - d_t|/\sigma)$, where $\sigma$ defines the length scale of the kernel function. The fit-coefficients, $c_t$, are obtained by solving the linear system

$$[K + \lambda I] c = E^\text{DFT}$$

(2)

In all ML calculations, we used a fixed-value for the regularization strength, $\lambda = 10^{-6}$, as a preconditioning measure. For determining a suitable kernel-width, we followed the “single-kernel recipe” proposed in Ref.\textsuperscript{[30]}

$$\sigma_{opt}^\text{max} = d_{ij}^\text{max} / \log(2).$$

(3)

where $d_{ij}^\text{max}$ is the largest descriptor difference among training entries. Use of fixed hyperparameters enables $\lambda, \sigma$ enables rapid training facilitating adequate shuffling of the training set to prevent any bias. Since the goal of this study is to test the validity of modeling a superstructure phase by training on substructures, we limit our explorations to one of the best structure-based representation: FCHL\textsuperscript{[31, 32]}. This descriptor has a tensor structure and a direct evaluation of the kernel matrix elements without storing the descriptor is the preferred approach as implemented in the program QML\textsuperscript{[33]}. The scope of the problem presented here is not limited by the representation, hence, one can also apply other descriptors such as row-sorted Coulomb matrix\textsuperscript{[29]}, bag-of-bond (BoB),\textsuperscript{[34]} SLATM\textsuperscript{[35]}, MBTR\textsuperscript{[36]}, or SOAP\textsuperscript{[37]} to exploit group-subgroup structure in ML modeling of materials.

The FriezeRMQ1D dataset is dominated by compositional diversity than structural variations compared to larger datasets such as OQMD\textsuperscript{[6]}. Hence, we also explore a fingerprint representation that encodes the stoichiometry of the materials and the Wyckoff positions of the 7 frieze groups. For a similar problem of modeling the energetics of 2 million elpasolites, a fingerprint descriptor has delivered more accurate ML learning curves than one of best structure-based formalism, FCHL\textsuperscript{[38]}. Further, while modeling on ICSD\textsuperscript{[11]} subset of OQMD containing only few cases of multiple structures for the same chemical formula, augmenting structural information to a composition-based one did not improve the prediction accuracy\textsuperscript{[38]}. The problem can be rectified if a $\Delta$-ML strategy\textsuperscript{[39, 40]} can be adapted where an inexpensive theory can be used as a baseline for property estimation and to generate a structural descriptor. Due to the lack of rapid and reliable baseline models for the materials problem, studies have usually dependent on a fingerprint\textsuperscript{[41, 45]}, atom-level\textsuperscript{[37]}, or composition-based representations\textsuperscript{[38]} that do not require structural information for out-of-sample predictions.

![FIG. 3. Encoding of Wyckoff positions in the fingerprints of FriezeRMQ1D materials. Each blue rectangle is a 19 bit integer encoding the composition of the formula unit. White rectangles contain zeroes for padding. The direction of the arrows signify the bit sequence.](image-url)
IV. GROUP-SUBGROUP MACHINE LEARNING

To understand the advantages of group-subgroup learning over conventional ML, we begin by presenting the learning curve for modeling the atomization energy of 1199 FriezeRMQ1D materials in the $\mathbb{P}2$ phase (see Figure 4). We present results for ML models using a fingerprint representation (Figure 4a) and the FCHL formalism (Figure 4b). When training directly on $\mathbb{P}2$, with the fingerprint representation, the offset of the learning curve is over 1 eV/atom. The offsets drop for all cases of group-subgroup learning to 0.4-0.5 eV/atom indicating the materials in $\mathbb{P}2$ to have chemical similarity with those in the subgroups. The performance of $\mathbb{P}11$ which is not a $t$-subgroup of $\mathbb{P}2$ is only on par with the smallest subgroup $\mathbb{P}1$ implying the lattice arrangement in a $k$-subgroup to have less relevance to that in the supergroup. The only information that is common to the structures in both these subgroups are the intra-formula-unit features. Learning with $\mathbb{P}2$ and $\mathbb{P}1m$ show better rates than with $\mathbb{P}1$ for smaller training set sizes. It is important to note that with increasing training set size, the learning curves of subgroup learning must coincide with that of direct learning. The better the subgroup is, lower the out-of-sample errors will be hence approaching the direct learning error asymptotically. Learning with the best subgroup $\mathbb{P}1m$ shows consistently lower errors reaching an MAE of 0.07 eV/atom for 50% training.

Learning curves for modeling $\mathbb{P}2$ energies with FCHL are provided in Figure 4b. The offset of direct $\mathbb{P}2$ modeling for 0.1% training (number of training entity is 1) is over 3.1 eV/atom. While not evident from Figure 4, learning curves for structure-based representations such as FCHL typically exhibit continuity as approaching the limit of zero-training-set size, where the offset should be the mean of absolute target property values. This is in agreement with the actual offset, 4.1 eV, calculated over all 1199 $\mathbb{P}2$ energies. With FCHL, the offsets for all subgroup learning training curves are noticeably smaller compared to those with the fingerprint vector. The subgroup delivering the least error is consistent across the representations, in both cases the MAE with 50% training is 0.07 eV/atom.

We compare the performance of subgroup learning with a $\Delta$-ML [39], for which energies from the subgroups are used as baselines. While the atomization energies

| Training set size (%) | MAE (eV/atom) |
|-----------------------|--------------|
| 0.1                   | 0.1          |
| 1                     | 0.04         |
| 10                    | 0.07         |
| 50                    | 0.07         |

crystallographic Wyckoff positions (Fig. 3). Making the fingerprint vector commensurate across frieze groups requires a minimum of 8 formula units to be considered to apply the group operations. The size of the vector per formula unit is 19 bits, of which the first 15 are reserved for the 5 ring atom sites —each ring site requires 3 bits to store the 8 possible atomic combinations: C, CH, B, BH, N, NH, S and SH—while the last 4 to store 11 metal types.
of the subgroup structures better approximate the supergroup ones, their performance as a baseline for $\Delta$-ML was found to very poor (see Figure 5). For both the fingerprint and FCHL descriptors, the training curves poorly converging towards zero indicates weak learning. Rather than a good rank correlation between the baseline and target line energies, to account for training set bias, learning curves were generated for 100 shuffles. The mean of MAEs are shown in solid lines while the envelope denote the standard deviation of MAEs.

Percentage errors for various group-subgroup combinations are collected in Table II. For both high-symmetry phases $\mathcal{P}_{2mm}$ and $\mathcal{P}_{2mg}$, better performances are seen when using a maximal subgroup than $\mathcal{A}$. For $\mathcal{P}_{2mg}$, all three maximal subgroups have resulted in lower prediction errors compared to using $\mathcal{A}$ irrespective of the descriptor. Among all combinations of groups, lowest out-of-sample prediction error was noted for modeling $\mathcal{A}1m$ energies with the $\mathcal{A}1$ subgroup; 1.0 % error with 75% training using FCHL. Due to the arrangement of the formula units according to $\mathcal{A}1m$ symmetry one can expect the internal degrees of freedom in this phase to be similar to that of $\mathcal{A}1$. As stated in Section II, for materials such as perovskites, phases of low-symmetry contain larger unit cells. Hence, the target phase of interest corresponds to subgroups. The ML formalism presented here can be applied also for such cases when the roles of groups and subgroups are reversed compared to the situation in FriezeRMQ1D materials. To validate this point, we present results for all possible group-subgroup modeling in Fig. 7. For the FriezeRMQ1D materials, the computationally complicated phases are $\mathcal{A}2mm$ and $\mathcal{A}2mg$, including these in the training set to model materials in subgroups such as $\mathcal{A}1$ also yields favorable results compared to a direct modeling on $\mathcal{A}1$. Generalizing the results discussed above, the FCHL approach in combination with $\mathcal{A}1$ data, facilitates modeling of atomization energies in $\mathcal{A}2mm$ with a prediction error of $< 3\%$ using 10% reference data (see Figure 5). Although the FriezeRMQ1D materials

![FIG. 5. Out-of-sample prediction errors for modeling 1199 $\mathcal{P}_{2mm}$ energies with $\Delta$-ML using energies of $\mathcal{A}1m$ structures as the baseline. The inset features a scatterplot of the baseline and target line energies. To account for training set bias, learning curves were generated for 100 shuffles. The mean of MAEs are shown in solid lines while the envelope denote the standard deviation of MAEs.](image)

| Training set $\gamma_i$ | MPAE for predictions in $\gamma_h$ | $\mathcal{A}1$ | $\mathcal{A}1m$ | $\mathcal{A}1g$ | $\mathcal{P}_{2mm}$ | $\mathcal{P}_{2mg}$ |
|-------------------------|----------------------------------|----------------|----------------|----------------|----------------|----------------|
| Fingerprint             |                                  |                |                |                |                |                |
| $\gamma_1$              |                                  | 14.5 13.9 7.0 15.0 13.8 14.8 |                |                |                |                |
| 5                       |                                  | 6.1 6.0 5.9 5.9 7.1 8.1 |                |                |                |                |
| 10                      |                                  | 3.4 3.6 3.8 3.6 4.3 5.9 |                |                |                |                |
| 25                      |                                  | 2.4 2.8 2.6 2.8 3.1 4.2 |                |                |                |                |
| 50                      |                                  | 1.9 2.3 2.0 2.3 2.3 3.0 |                |                |                |                |
| 75                      |                                  | 1.4 1.9 1.6 1.8 1.9 2.3 |                |                |                |                |
| $\gamma_{211}$          |                                  | 1.9 1.4 |                |                |                |                |
| 75                      |                                  | 1.4 1.4 |                |                |                |                |
| $\gamma_{11m}$          |                                  | 1.9 |                |                |                |                |
| 75                      |                                  | 1.4 |                |                |                |                |
| $\gamma_{11g}$          |                                  | 1.4 |                |                |                |                |
| 75                      |                                  | 1.2 |                |                |                |                |
| FCHL                    |                                  |                |                |                |                |                |
| $\gamma_1$              |                                  | 4.0 1.6 4.1 1.9 4.2 4.2 |                |                |                |                |
| 5                       |                                  | 3.7 1.7 3.4 2.0 3.7 3.9 |                |                |                |                |
| 10                      |                                  | 3.0 1.5 2.2 1.7 2.7 3.1 |                |                |                |                |
| 25                      |                                  | 2.5 1.3 1.8 1.6 2.3 2.6 |                |                |                |                |
| 50                      |                                  | 2.1 1.1 1.5 1.4 1.8 2.0 |                |                |                |                |
| 75                      |                                  | 1.9 1.0 1.3 1.3 1.4 1.6 |                |                |                |                |
| $\gamma_{211}$          |                                  | 1.7 1.2 |                |                |                |                |
| 75                      |                                  | 1.4 1.2 |                |                |                |                |
| $\gamma_{11m}$          |                                  | 1.6 |                |                |                |                |
| 75                      |                                  | 1.2 |                |                |                |                |

TABLE II. Mean percentage absolute error (MPAE) for out-of-sample predictions on high-symmetry phases corresponding to groups $\gamma_h$. Training is done with materials in low-symmetry phases in subgroups $\gamma_i \subset \gamma_h$, along with fraction of high-symmetry examples. $\%\gamma_h=0$ implies training sets containing only the subgroup $\gamma_i$. |
are low-dimensional, the CPU costs for their DFT energy evaluation are a function of number of electrons in the unit cell. For the high-symmetry phases with the cyclopentadienyl ring and Cu metal the unit cell contains 256 electrons making their computational complexity comparable with complex phases of bulk ternary materials such as perovskites or transition metal based pnictides/chalcogenides. It must be noted that the cost for generating the data for several smaller unit cell materials is often negligible compared to that of large unit cell ones. For instance, the ratio between the time taken for single point energy calculations in $\pi_1$ and $\pi_{11m}$ phases for the cyclopentadienyl-Cu system is about 0.015. For modeling the energies of a few thousand ternary materials demands a data-generation cost for only a few hundred examples. Compared to this, the cost for performing DFT calculations of the entire set in a compact phase should be negligible.

V. CONCLUSIONS

We have presented the group-subgroup learning formalism for efficient modeling of materials properties in complex phases with multiple chemical formulas in a unit cell. The approach exploits transformation relationships between crystallographic space groups and is applicable for materials datasets where more than one crystal structures are available per stoichiometry. To provide a proof-of-the-concept, we have generated a new Q1D materials dataset “FriezeRMQ1D” with chemical compositions spanning 11 metals, 109 rings, and seven frieze group symmetries. To facilitate further symmetry-based explorations, all data generated for this study are made publicly available. For the resulting 8393 materials, minimum energy geometries of desired symmetries were calculated using constrained lattice relaxations. The target property of interest is the atomization energies calculated at the hybrid-DFT level. The dataset is one-hot encodable. Hence, we designed a fingerprint vector representation containing information about the cyclic structure of the ring, its stoichiometry, metal bonded to the ring, and the frieze group symmetry of the lattice. Besides, we have tested the performance of the FCHL formalism that captures structural and alchemical similarities. We depend on a % error as a reliable error metric than absolute errors because other studies have shown the prediction errors of ML models for materials to depend strongly on the dataset[31].

We have analyzed the performance of group-subgroup learning through all possible combinations of frieze groups. Irrespective of the materials representation, we find the maximal subgroup of a group to deliver the lowest prediction errors. Furthermore, we have shown the proposed formalism to be orthogonal to the $\Delta$-ML approach. Application of the $\Delta$-ML idea to this dataset by using the energies from the $\pi_{1m1}$ phase as a baseline to model $\pi_{2mm}$ energies exposes poor structure-property correlations between global descriptors and $\Delta$ energies. The most attractive feature of the group-subgroup learning is that it alleviates the explicit dependence of materials descriptors on DFT-level minimum energy geometries. While structural descriptors such as FCHL, MBTR or SOAP by design...
deliver best prediction accuracies when using minimum energy geometries to predict ground state total energies, such a dependency restricts the application domain of the models tremendously. Hence, constructing structural descriptors for materials phases with a complex unit cell arrangement using the geometries of compact phases with the knowledge of Wyckoff positions splitting widens the application domain of ML for materials modeling.

The general conclusions drawn from our results are independent of the ML approach, datasets or materials representation. It will be interesting to see if this strategy can be adapted to modeling economically important materials like perovskites [24] or Fe-based pnictides [16] that have received a lot of attention as superconductor candidates. Similarly, the chemistries common in non-magnetic and spin-collinear phases of materials can also be exploited using group-subgroup relations for the distribution of local magnetic moments in materials.

VI. DATA AVAILABILITY

Atomic coordinates of all 8393 FriezeRMQ1D materials, Wyckoff-encoded fingerprint vectors, and PBE0 atomization energies are available at http://moldis.tifrii.res.in/data/FriezeRMQ1D. Input and output files of corresponding calculations are deposited in the NOMAD repository (https://nomad-lab.eu/).

VII. ACKNOWLEDGMENTS

We acknowledge support of the Department of Atomic Energy, Government of India, under Project Identification No. RTI 4007. All calculations have been performed using the Helios computer cluster, which is an integral part of the MolDis Big Data facility, TIFR Hyderabad (http://moldis.tifrii.res.in).

[1] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenhölz, G. E. Dahl, O. Vinayals, S. Kearnes, P. F. Riley, and O. A. von Lilienfeld, J. Chem. Theory Comput. 13, 3525 (2017).
[2] J. Schmidt, M. R. Marques, S. Botti, and M. A. Marques, Npj Comput. Mater. 5, 1 (2019).
[3] J. Schönhölz, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amaral-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, and A. Aspuru-Guzik, J. Phys. Chem. Lett. 2, 2241 (2011).
[4] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, Sci. Data 1, 140022 (2014).
[5] S. Chakraborty, P. Kayastha, and R. Ramakrishnan, J. Chem. Phys. 150, 114106 (2019).
[6] S. Kirklin, J. E. Saal, B. Merezh, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, Npj Comput. Mater. 1, 1 (2015).
[7] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, et al., Comp. Mat. Sci. 58, 227 (2012).
[8] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Guenter, D. Skinner, G. Ceder, et al., APL Mater. 1, 011002 (2013).
[9] C. Kim, A. Chandrasekaran, T. D. Huan, D. Das, and R. Ramprasad, J. Phys. Chem. C 122, 17575 (2018).
[10] L. Ruddigkeit, R. Van Deursen, L. C. Blum, and J.-L. Reymond, J. Chem. Inf. Model. 52, 2864 (2012).
[11] M. Hellenbrandt, Crystallography Reviews 10, 17 (2004).
[12] J. Behler, S. Lorenz, and K. Reuter, J. Chem. Phys. 127, 07B603 (2007).
[13] B. Huang and O. A. von Lilienfeld, Nat. Chemistry 12, 945 (2020).
[14] H. T. Stokes and D. M. Hatch, Phys. Rev. B 30, 4962 (1984).
[15] L. D. Landau, Zh. Eksp. Teor. Fiz. 11, 19 (1937).
[16] U. Müller, Symmetry relationships between crystal structures: applications of crystallographic group theory in crystal chemistry, Vol. 18 (OUP Oxford, 2013).
[17] F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, Phys. Rev. Lett. 117, 135502 (2016).
[18] P. Kayastha and R. Ramakrishnan, arXiv preprint arXiv:2009.12519 (2020).
[19] M.-O. Lenz, T. A. Purcell, D. Hicks, S. Curtarolo, M. Scheffler, and C. Carbogno, Npj Comput. Mater. 5, 1 (2019).
[20] M. J. Mehl, D. Hicks, C. Toher, O. Levy, R. M. Hanson, G. Hart, and S. Curtarolo, Comp. Mat. Sci. 136, 51 (2017).
[21] D. Hicks, M. J. Mehl, E. Gossett, C. Toher, O. Levy, R. M. Hanson, G. Hart, and S. Curtarolo, Comp. Mat. Sci. 161, 51 (2019).
[22] D. Hicks, C. Toher, D. C. Ford, F. Rose, C. De Santo, O. Levy, M. J. Mehl, and S. Curtarolo, arXiv preprint arXiv:2010.04222 (2020).
[23] C. J. Howard and H. T. Stokes, Acta Crystallogr. B Struct. Crystal. Chem. 54, 782 (1998).
[24] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, Comput. Phys. Comm. 180, 2175 (2009).
[25] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. 77, 3865 (1996).
[26] C. Adamo and V. Barone, J. Chem. Phys. 110, 6158 (1999).
[27] B. Schölkopf, A. J. Smola, F. Bach, et al., Learning with kernels: support vector machines, regularization, optimization, and beyond (MIT press, 2002).
[28] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. 108, 058301 (2012).
[29] R. Ramakrishnan and O. A. von Lilienfeld, CHIMIA 69, 182 (2015).
[30] P. A. Faber, A. S. Christensen, B. Huang, and O. A. von Lilienfeld, J. Chem. Phys. 148, 2411717 (2018).
[32] A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. A. von Lilienfeld, J. Chem. Phys. 152, 044107 (2020).
[33] A. S. Christensen, F. A. Faber, B. Huang, L. A. Bratholm, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, “QML: A python toolkit for quantum machine learning,” (2017), https://github.com/qmlcode/qml.
[34] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, J. Phys. Chem. Lett. 6, 2326 (2015).
[35] B. Huang, N. O. Symonds, and O. A. von Lilienfeld, Handbook of Materials Modeling: Methods: Theory and Modeling, 1883 (2020).
[36] M. F. Langer, A. Goëßmann, and M. Rupp, arXiv preprint arXiv:2003.12081 (2020).
[37] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, Phys. Chem. Chem. Phys. 18, 13754 (2016).
[38] L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary, and C. Wolverton, Phys. Rev. B 96, 024104 (2017).
[39] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, J. Chem. Theory Comput. 11, 2087 (2015).
[40] A. Gupta, S. Chakraborty, and R. Ramakrishnan, arXiv preprint arXiv:2009.06814 (2020).
[41] S. A. Meldgaard, H. L. Mortensen, M. S. Jørgensen, and B. Hammer, J. Condens. Matter Phys. (2020).
[42] T. D. Huan, A. Mannodi-Kanakkithodi, and R. Ramprasad, Phys. Rev. B 92, 014106 (2015).
[43] R. Batra, H. D. Tran, C. Kim, J. Chapman, L. Chen, A. Chandrasekaran, and R. Ramprasad, J. Phys. Chem. C 123, 15859 (2019).
[44] O. A. von Lilienfeld, R. Ramakrishnan, M. Rupp, and A. Knoll, Int. J. Quantum Chem. 115, 1084 (2015).
[45] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, J. Chem. Phys. 148, 241730 (2018).
[46] M. Shatruk, J. Solid State Chem. 272, 198 (2019).