AFRL-AFOSR-UK-TR-2018-0040

Machine learning models of solid properties for high-throughput screening of condensed phase materials with chemical accuracy

Otto Anatole von Lilienfeld-Toal
UNIVERSITAT BASEL

08/20/2018
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ IOE
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
20-08-2018	Final	01 Dec 2014 to 30 Nov 2017

4. TITLE AND SUBTITLE
Machine learning models of solid properties for high-throughput screening of condensed phase materials with chemical accuracy

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-15-1-0026

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Otto Anatole von Lilienfeld-Toal

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITAT BASEL
PETERSPLATZ 1
Basel, 4003 CH

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
EOARD
Unit 4515
APO AE 09421-4515

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR IOE

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-UK-TR-2018-0040

12. DISTRIBUTION/AVAILABILITY STATEMENT
A DISTRIBUTION UNLIMITED: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
We have investigated several machine learning models which can be used to study solids. First findings were published in a peer-reviewed journal article. Within a second study, we developed a new representation, and applied it to the study of elpasolite crystals. Elpasolite was chosen since it is the predominant quaternary crystal structure (AlNaK2F6 prototype) reported in the Inorganic Crystal Structure Database. We generated machine learning model to calculate density functional theory quality formation energies of all the 2M pristine ABC2D6 elpasolite crystals which can be made up from main-group elements (up to bismuth). Our model's accuracy was improved systematically, reaching 0.1 eV/atom for a training set consisting of 10 k crystals. Important bonding trends are revealed and are reported. Subsequently, in 2016 and 2017, a more universal and improved representation was developed and tested on various data sets including molecules, solids, water clusters, and peptide side chain interactions. Systematic improvement of prediction error with training set size was demonstrated for all data sets, often reaching unprecedented predictive power. Impressive results were also obtained for various molecular electronic ground-state properties, dipole moments, HOMO-LUMO gaps, and polarizabilities. Interestingly, learning has even been observed when predicting systems containing chemical elements which were absent in training. The results of these studies were published in an article in 2018.

15. SUBJECT TERMS
EOARD, materials modeling, high-throughput screening, machine learning, computational materials science

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			FOLEY, JASON
Unclassified	Unclassified	Unclassified	SAR		19b. TELEPHONE NUMBER *(Include area code)* 011-44-1895-616036

Final Performance Report 30 Apr 2018
FA9550-15-1-0026
Machine Learning Models of Solid Properties for High-Throughout Screening of Condensed Phase Materials with Chemical Accuracy

PI: Dr. Anatole Von Lilienfeld-Toal
Period of Performance: 1 December 2014 to 30 November 2017
Period of Reporting: 1 December 2014 to 30 November 2017
(Dated: August 17, 2018)

CONTENTS

I. SUMMARY

We have investigated several machine learning models which can be used to study solids. First findings were published in Ref.[1]. Within a second study, we developed a new representation, and applied it to the study of elpasolite crystals. Elpasolite is the predominant quaternary crystal structure ($AlNaK_2F_6$ prototype) reported in the Inorganic Crystal Structure Database. We generated machine learning model to calculate density functional theory quality formation energies of all the 2 M pristine ABC_2D_6 elpasolite crystals which can be made up from main-group elements (up to bismuth). Our model's accuracy can be improved systematically, reaching 0.1 eV/atom for a training set consisting of 10 k crystals. Important bonding trends are revealed, fluoride is best suited to fit the coordination of the D site which lowers the formation energy whereas the opposite is found for carbon. The bonding contribution of elements A and B is very small on average. Low formation energies result from A and B being late elements from group (II), C being a late (I) element, and D being fluoride. Out of 2 M crystals, the three degenerate pairs $CaSrCs_2F_6/SrCaCs_2F_6$, $CaSrRb_2F_6/SrCaRb_2F_6$ and $CaBaCs_2F_6/BaCaCs_2F_6$ yield the lowest formation energies: −3.44, −3.41, and −3.39 eV/atom, respectively. In crystals with large negative formation energies unusual atomic oxidation states have been discovered for Sb and Te. This work was published in 2016[2]. Subsequently, in 2016 and 2017, a more universal and improved representation was developed and tested on various data sets including molecules, solids, water clusters, and peptide side chain interactions. Systematic improvement of prediction error with training set size was demonstrated for all data sets, often reaching unprecedented predictive power. Impressive results were also obtained for various molecular electronic ground-state properties, dipole moments, HOMO-LUMO gaps, and polarizabilities. Interestingly, learning has even been observed when predicting systems containing chemical elements which were absent in training. The resulting study was published in 2018 in the special J Chem Phys issue on "Data-enabled theoretical chemistry" by Faber (the PhD student predominantly funded by this project), Christensen, Huang, and von Lilienfeld (the PI)[3]. Correspondingly, the representation was dubbed as FCHL, i.e. the acronym resulting from the last names of the authors. Most recent work has dealt with the extension of FCHL to also account for response properties, such as forces on atoms or the electric dipole moment[4]. Consequently, throughout this report we frequently and freely quote from all the three Refs.[2–4].

II. INTRODUCTION

We have worked on the development and application of chemically accurate machine learning models for crystalline solids. This has become possible by hiring a PhD

student, Felix Faber, to work under the PI's direct guidance. The PhD program at the University of Basel, Switzerland, as well as the academic research environment in the group led by the PI at the Institute of Physical Chemistry, have provided the setting in which this ambitious work has started. Over the course of the first year, we have adapted the molecular representation introduced in Ref.[5] to also encode unit cells within periodic boundary conditions[1]. For consistent training and testing systematic materials data sets of electronic structure properties have been generated for many stoichiometrical mixtures. So far, our work has focussed on formation energies of quaternary crystals involving all main-group elements (I-VIII) up to Bi[2]. Other crystals, liquids, and interfaces as well as properties have been dealt with within our most recent work[3,4]. More specifics are given below. The PhD student is currently finalizing the writing of his thesis, and is expected to graduate in late 2018 or in 2019.

A. Proposed Milestones

In the narrative of the grant proposal the following milestones were specified:

1. Generate a consistent (representative and dense in materials space) database of well behaved main-group element based solids that we can use for the development of new representations (training and testing).
   (Year 1)

2. Explore various descriptor spaces to identify models which can be trained to predict properties of doped III-V semi-conductors, other maingroup elements, transition metal oxides, as well as alloys and defects, and even molecular crystals or liquids. More specifically, the model will be trained to estimate the deviation of an inexpensive base-line method (such as tight-binding density functional theory) from a desirable reference (such as hybrid density functional theory, GW, or even quantum Monte Carlo).
   (Year 1+2)

3. Screen larger sets with millions of materials candidates for interesting properties such as band-gap, desirable DOS, and convex hull in stoichiometrical mixtures.
   (Year 2)

4. Augment machine learning models with models of other properties of interest such as atomic forces, so that even crystal structure relaxation or molecular dynamics can be carried out.
   (Year 2+3)

5. Assess the performance for the simultaneous modeling of large numbers of different molecular crystals, liquids, and interfaces in parallel.
   (Year 3)

We have accomplished the first milestone (year 1) by calculating density functional theory (DFT) based formation energies for over 10,000 elpasolite crystals. For the second milestone (year 1+2) we investigated several descriptors for solids[1]. We tackled the third milestone (year 2) by using the elpasolite machine learning (ML) model to screen formation energies for 2 M crystals involving all main-group elements (I-VIII) up to Bi[2]. Milestones 4 and 5 have, at least partially, been resolved within the most recent work[3,4].

B. Elpasolites

Elpasolite ($AlNaK_2F_6$) is a glassy, transparent, luster, colorless, and soft quaternary crystal in the Fm3m space group which can be found in the Rocky Mountains, Virginia, or the Apennines. The elpasolite crystal structure (See Fig. 1) is not uncommon, it is the most abundant prototype in the Inorganic Crystal Structure Database[6,7]. Some elpasolites emit light when exposed to ionic radiation, which makes them interesting material candidates for scintillator devices[8,9]. One could use first-principle methods such as DFT[10,11] to computationally predict the existence and basic properties of every elpasolite. Unfortunately, even when considering crystals composed of only main group elements (columns I to VIII) the sheer number of all the 2 M possible combinations makes DFT based screening challenging—if not prohibitive. Recently, computationally efficient ML models were introduced for predicting molecular properties with the same accuracy as DFT[5,12]. Requiring only milliseconds per prediction, they represent an attractive alternative when it comes to the combinatorial screening of millions of crystals. While some ML model variants have already been proposed for solids[1,13,14], a generally applicable ML-scheme with DFT accuracy of formation energies is still amiss. We wrote a Letter where we introduced a newly developed ML model which we use to investigate the formation energies of *all* ∼2 M elpasolites made from all main-group elements up to Bi[2]. Resulting estimates are used to identify a new elemental order of descending elpasolite formation energy, crystals with peculiar atomic charges, as well as 250 elpasolites with lowest formation energies. The ML model achieves DFT accuracy or better, and can be generalized to any crystalline material. During the review process we were pressed to also investigate if we can identify the thermodynamically stable elpsasolites. This resulted in substantial additional work because all the competing phases for all elpasolites with negative formation energies had to be taken into account. We have accomplished this task by querying the Materials Project data base [`https://materialsproject.org`] in order to extract all the recorded competing formulations of ternary

and bindary compounds consistent for each of the quaternary elpasolite crystals for which the ML model predicted negative formation energies (∼200'000). To quote from the published paper[2]: *This resulted in many million queries from which we extracted 2133 elpasolites with energies indicating that they are on the convex hull of thermodynamic stability. We subsequently validated these structures using DFT, and 128 of them were confirmed to be on the convex hull. 38 of these structures were polymorphs (ABC_2D_6 vs. BAC_2D_6), resulting in 90 overall stoichiometries. Such a reduction (274,213 → 90) in number of crystal candidates is to be expected since sorting crystals by ML energies being lower than the convex hull systematically favors those with negative ML formation energy errors. We note that this does not amount to proof that the 90 crystals are stable: The MP database is not exhaustive. This implies that other new competing phases and materials, with even stronger stabilization, might still be discovered in the future. Also, the intrinsic error of the employed DFT method within the MP might still alter the outcome with respect to experiment. As such, the 90 new elpasolite DFT energies represent new upper bounds on the convex hull at the corresponding compositions. They have been submitted to the MP database, and most of them have been made available for further studies*

Among these elpasolites, metals, semiconductors and insulators are roughly distributed equally. All structures with an earth alkaline metal in crystal position 4 have a low or zero band-gap. We have noted an intriguing yet stable structure of a conductor, $NFAl_2Ca_6$ (MP ID: mp-989399) with Ca at position 4, instead of F or Cl. Bader charge analysisindicates an exotic negative oxidation state for Al (-II), previously only reported for Al in substantially larger Zintl phase unit cells ($Sr_{14}[Al_4]_2Ge_3$).Since Bader charges sometimes yield non intuitive results,calculated Hirshfeld and Voronoi deformation density charges confirm the negative oxidation state, albeit reduced by one unit (-I). The calculated phonon spectra of $NFAl_2Ca_6$ also indicate stability.

C. Representation

We have made significant progress regarding the development of an improved representation which was recently published in Ref.[3]. Inductive quantum machine learning (QML) models can infer properties of new materials directly from training data, and can even predict the electron density which in turn can be used to calculate properties[16]. As such, ML models can have an exceptional trade off between predictive accuracy and computational cost. For example, in 2017 we showed that QML models can estimate hybrid DFT atomization energies, as well as several other properties, of medium sized organic molecules with prediction errors lower than chemical accuracy (∼0.04 eV)—multiple orders of magnitude faster than hybrid DFT[17].

The system variables defining the ground-state properties of a given compound are its external potential, a simple function of interatomic distances and nuclear charges. However, using this information directly to measure similarity results in QML models with rather disappointing predictive power. This can be mitigated by transformation of system variables into "representations". Such transformations can either be designed by human intuition, or be included in the learning problem, e.g. when using neural networks (NN) which include representation learning in the supervised learning task. Letting a NN find the representation has proven to yield models with low out-of-sample prediction errors[18–20]. This approach, however, has the drawback that representation and model are intermingled within the NN, making it less amenable to human understanding, interpretation, adaptation, and further improvement. Furthermore, such machine designed representations do not necessarily lead to better QML performance than human design based representations (vide infra).

D. Response properties

We have also made significant progress regarding the development of ML models of response properties. For a motivation of why this is an important problem, we quote from the work posted recently[4]: *Time-independent electronic ground-state quantum properties can be expressed as expectation values of the electronic wavefunction and an operator, typically defined via the correspondence principle. The performance of supervised machine learning models of these quantum properties, a.k.a. quantum machine learning (QML),[2,12,21,22] can be conveniently assessed using learning curves which monitor the decay of the out-of-sample prediction error, i.e. the deviation of predicted properties from reference for query instances not included in training, as a function of training set size N. Due to the leading prediction error decaying as a/N^b, log-log plots have become the recommended practice in the field with* $\log(a)$ *and* b *denoting the off-set and learning rate (or efficiency), respectively[23–25]. While in principle, supervised ML models can be generated for any cause and effect relationship, it is the very philosophy of QML that representation (and kernel function when using kernel ridge regression) is property independent[26,27]. However, there is a select and highly relevant subset of quantum properties which can be understood as response properties, obtained through the use of response operators and perturbation theory. Common examples include derivatives of the energy with respect to e.g. the nuclear displacement, an external electric field, an external magnetic field, or nuclear magnetic moments, and can efficiently be accounted for within density functional theory[28,29]. We note in passing that energy response properties also form the basis for conceptual density functional theory[30,31], as well as computational alchemy[32,33]. It has previously been observed that prediction errors of*

quantum machine learning models of response properties can converge relatively slowly, even for machine models that are able to achieve remarkably high accuracy for energies.[3,12,17,26,34]*. In this paper we investigate if the use of response operators might be beneficial for deriving improved QML models which afford learning curves with lower off-sets and better learning rates.*

Maybe the most relevant quantum response property is the force exerted on each atom in the system, the first order energy derivative with respect to nuclear displacement[35]*. Quite recently, tremendous efforts have been made to predict atomic forcces accurately within QML models for the purpose of running* ab initio *quality molecular dynamics simulations at low computational cost.*[36–48]*. Treating the force as the first derivative of the energy is tantamount to using the gradient operator, as commonly implemented in quantum chemistry packages. Doing so leads directly to energy conservation, a crucial property for most statistical mechanics applications, which has already also been obtained by others*[45,49]*. Using the response operator, however, has not yet been applied generally to generate QML models for other response properties.*

Here, we extend the principle of using response operators to investigate (i) energy and its response to a change in the nuclear coordinates, and (ii) the energy response to a change in the external electric field, i.e. the dipole moments. Recently other QML models, capable of predicting dipole moments have also been published.[50–54] *The work by Schütt et al. presents a neural network that is able to predict the dipole moment of the QM9 dataset*[55,56] *with very high accuracy*[50]*, by simply training on the observable, the dipole moment vector itself. Other works use a charge model predicted from a neural network to estimate the intensities in an infrared spectrum when the vibrational frequencies are given from a molecular dynamics simulation.*[51,53] *Similarly to Schütt et al., we propose to learn the dipole moment by training on the quantum mechanical observable directly, but in contrast we train a model to describe the energy for which the dipole moment can be calculated as a response property simply by taking the derivative.*

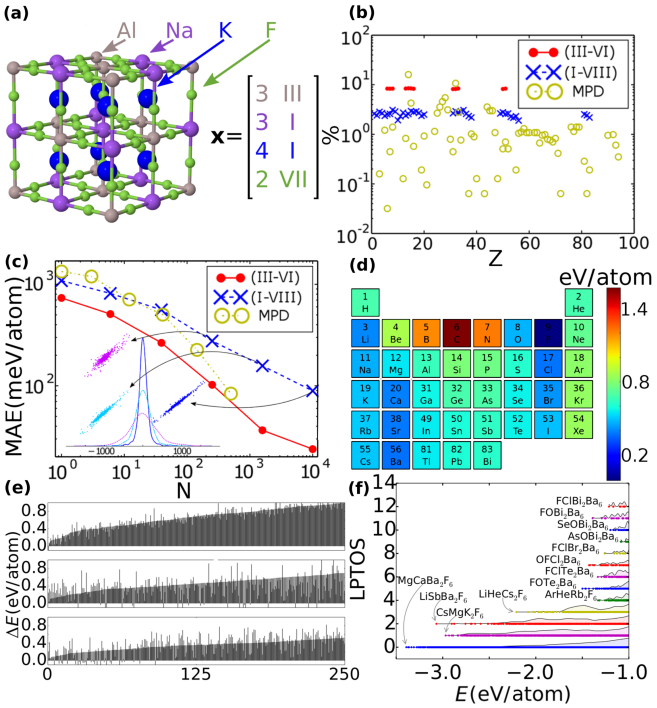

FIG. 1. (a) Illustration of elpasolite crystal ($AlNaK_2F_6$ structure). The four-tuple $x = (x_1, \ldots, x_4)$ representation of atomic sites is specified. (b) Frequency of elements (defined by nuclear charge Z) for the three data sets studied. (c) Mean absolute out-of-sample prediction error as a function of training set size for the three data sets studied. Inset: Error distributions and DFT vs. ML scatter plots for three training set sizes for the (I−VIII) data set. (d) Estimated mean energy contribution of each element to formation of any elpasolite crystal. The color code reflects the new elemental elpasolite order. (e) Lowest 250 ML model predicted formation energies of elpasolites in ascending order from (III−VI) (TOP) and (I−VIII) (MID and BOTTOM) data sets. Results in TOP and MID panel correspond to ML models trained on 2000 examples, BOTTOM panel results correspond to a ML model trained on 10k crystals. Validating DFT energies are shown aside. (f) Distributions of absolute lowest possible total oxidation states (LPTOS) in energies. Formulas indicate the lowest lying crystals.

III. METHODS, ASSUMPTIONS AND PROCEDURES

A. ML model for Elpasolite work

The ML-model is based on kernel ridge regression[57] which maps the non-linear energy difference between the actual DFT energy and an inexpensive approximate baseline model into a linear feature space[58]. More specifically, we construct a ML model of the energy difference to the sum of static, atom-type dependent, atomic energy contributions ϵ_{It}, obtained through fitting of each atom type t in all main group elements up to Bi. The energy-predicting function is a sum of weighted exponentials in similarity d between query and training crystal,

$$E(\mathbf{x}) = \sum_{I}^{N'} \epsilon_{It} + \sum_{i}^{N} \alpha_i e^{-d_i/\sigma}, \qquad (1)$$

where N' is the number of atoms/unit cell (10 in the case of elpasolites), and the second sum runs over all N training instances. α_i are the weights obtained through linear regression, and σ is the global exponential width, regulating the length scale of the problem. The similarity d_i is the Manhattan distance, i.e., $d_i = \|\mathbf{x} - \mathbf{x}_i\|_1$. While various crystal structure representations have previously been proposed[1,13,14,59], we have found the following representation to yield superior performance: $\mathbf{x}$ is a $n \times 2$ tuple that encodes any stoichiometry within a given crystal

prototype. For quaternary ($n = 4$) elpasolites, each x_{1-4} refers to the 4 representative sites, the atom type for each site is represented by its row (principal quantum number 2 to 6) and column (number of valence electrons) I to VIII in the periodic table, and sites are ordered according to the Wyckoff sequence of the crystal. As such, $\mathbf{x}$ implicitly represents the global energy minimum structure for a system restricted to this prototype—without explicitly encoding precise coordinates, lattice constants, or other (approximate) solutions to Schrödinger's equation. This representation is not restricted to the elpasolite structure, it can be used for any crystalline configuration: Below we also briefly discuss test results for small size ML models applied to ternary crystals.

B. DATA set for Elpasolite work

For training and evaluation, we have generated DFT data for two data sets of elpasolites, one small, (III−VI), made up from only 12 elements, C, N, O, Al, Si, P, S, Ga, Ge, As, Sn, and Sb; and one large, (I−VIII), containing all main-group elements up to Bi. Since (III−VI) only comprise ∼12 k possible permutations, we have used DFT to obtain a complete list of formation energies.(I−VIII) consists of 10 k structures, i.e. 0.5% of the total number of 2 M possible crystals. The (I−VIII) data set has been generated through random selection of elpasolites while ensuring an unbiased composition. To verify that the ML model is general and not only restricted to elpasolites, we have also included a materials project[60] dataset (MPD) consisting of ∼0.5k ternary crystals in $ThCr_2Si_2$ (I4/mmm) prototype and made up of 84 different atom types. The distribution of the chemical elements in the data sets are shown in Fig. 1(b).

C. Novel representation and response work

1. *Representation, distances and scalar products*

In this work, also kernel ridge regression models are used. In contrast to the Elpasolite work, however, we have not relied on a Δ-ML approach[58]. Rather, we investigated the direct performance of the models. Also, instead of Laplacian kernel functions, we have solely worked with Gaussian kernels.

We use a set of interatomic M-body expansions $\mathcal{A}_M(I) = \{A_1(I), A_2(I), A_3(I), \ldots, A_M(I)\}$ which contain up to M-body interactions to represent the structural and chemical environment of an atom I in compound $\mathbf{C}$. $A_m(I)$ is a weighted sum that runs over all m-body interactions. Each element in the sums consists of Gaussian basis functions, placed on structural and elemental degrees of freedom, and multiplied by a scaling function ξ_m. Structural values encode geometrical information about the system, such as interatomic distances or angles. As elemental parameters we use the period P and group G from the periodic table. The scaling functions ξ_m are used to weigh the importance of each Gaussian, based on internal system coordinates. We now consider only the first three distributions in $\mathcal{A}_M(I)$ for an atom I. We have also derived, implemented and tested the 4-body $A_4(I)$ distributions. However, the predictive accuracy improvements of resulting QML models were found to be negligible in comparison to the 3-body expansion. Also, the computational cost for generating large kernel matrices increases substantially when going from third to fourth order terms.

The first-order expansion $A_1(I)$ accounts for chemical composition (stoichiometry) and is modeled by a Gaussian function placed on period P_I and group G_I in the periodic table of element I:

$$A_1(I) = \mathcal{N}(\mathbf{x}_I^{(1)}) \;=\; e^{-\frac{(P_I-\chi_1)^2}{2\sigma_P^2}-\frac{(G_I-\chi_2)^2}{2\sigma_G^2}} \qquad (2)$$

where $\mathbf{x}_I^{(1)} = \{P_I, \sigma_P; G_I, \sigma_G\}$, with respective widths σ_P and σ_G. σ_P and σ_G can be seen as elemental smearing parameters, which control the near-sightedness of elements in the periodic table. χ_1 and χ_2 represent dummy variables for period and group, to be integrated out when evaluating the Euclidean distance (see Eq. (3)). For $A_1(I)$, the scaling function is set to unity, since stoichiometry is geometry independent. We are not aware of other representations in the literature which employ similar distribution functions in the periodic table.

$A_2(I)$ is a product of $A_1(I)$ and a sum that runs over all neighboring atoms i: $A_2(I) = \mathcal{N}(\mathbf{x}_I^{(1)})\sum_{i\neq I}\mathcal{N}(\mathbf{x}_{iI}^{(2)})\xi_2(d_{iI})$, $\mathbf{x}_{iI}^{(2)} = \{d_{iI}, \sigma_d; P_i, \sigma_P; G_i, \sigma_G\}$, where d_{iI} and σ_d correspond to the interatomic distance at which a Gaussian is placed, and its width, respectively. ξ_2 corresponds to the 2-body, interatomic distance dependent, scaling function which takes the form of the power laws discussed below. Note that letting σ_P and σ_G approach zero is equivalent to using a radial distribution function (RDF) for each element pair. This attribute of the representation holds for any of $A_m(I)$, i.e. $\sigma_P, \sigma_G \to 0$ is equivalent to creating a separate distribution for each chemical element $m-$tuple in $A_m(I)$. $A_3(I)$ is the logical extension from $A_2(I)$, it has a different scaling function with an additional summation, running over all neighboring atoms j: $A_3(I) = \mathcal{N}(\mathbf{x}^{(1)})\sum_{i\neq I}\mathcal{N}(\mathbf{x}_{iI}^{(2)})\sum_{j\neq i,I}\mathcal{N}(\mathbf{x}_{ijI}^{(3)})\xi_3(d_{iI}, d_{jI}, \theta_{ij}^I)$, $\mathbf{x}_{ijI}^{(3)} = \{\theta_{ij}^I, \sigma_\theta; P_j, \sigma_P; G_j, \sigma_G\}$. P_j and G_j, similarly to P_i and G_i, corresponds to the period and group of atom j. Again, $\xi_3(d_{iI}, d_{jI}, \theta_{ij}^I)$ is the (three-body) scaling function, and θ_{ij}^I the principal angle between the two distance vectors $\vec{r}_{Ii}$ and $\vec{r}_{Ij}$ which span from I to i and I to j, respectively. σ_θ is the width of the Gaussian placed at θ_{ij}^I. Letting σ_d go to infinity in A_3 is equivalent to using a type of angular distribution function (ADF), which in one form or another has already been used in several representations[3,61,62]. A_3 can therefore be seen as a generalized ADF containing more structural information. Fig. **??** illustrates how $A_3(I)$ looks for a

hydrogen, carbon, and the oxygen atom in ethanol.

The scaling functions ξ we have chosen for this work correspond to simple power laws. They have been modified from the leading order two- and three-body dispersion laws by London, $1/r^6$, and Axilrod-Teller-Muto[63,64], $1/r^9$. Such dispersion expressions were previously already used by some of us[61]. Our scaling functions, however, use different exponents for the radial decay, and set the C_6 and C_9 coefficients to unity, as early tests indicated better performance for this choice. For periodic systems, however, a very large cutoff radius would be needed in order to converge the distances between two atomic environments, when using the optimized exponents. We have therefore augmented the scaling functions by a previously used soft cutoff function[65], which goes to zero at 9 Å.

In order to train and evaluate the KRR model, proper distance measures must be specified. We have found good performance when using as a distance between two atomic environments $\mathcal{A}_M(I)$ and $\mathcal{A}_M(J)$ a weighted sum of the distances between each m-body expansion: $\Delta(\mathcal{A}_M(I), \mathcal{A}_M(J))^2 \equiv \sum_{m=0}^{M} \beta_m \Delta(A_m(I), A_m(J))^2$. Here, β_m is another hyperparameter, which weighs the importance of each expansion order.

The distances between each distribution term are evaluated as Euclidean (L_2) norms, as shown in Eq.3. ς_m are normalization factors, which ensures that all individual basis functions integrate to 1 in the L_2-norm. All integrals can be solved analytically since they consist of a sum of Gaussian products. The explicit form of the A_m integrals for $m = 1 \dots 3$ is shown in Eq. 4. Details on how to evaluate the A_3 and A_4 integrals in Fourier space can be found in the SI.

$$\Delta(A_m(I), A_m(J))^2 = \frac{1}{\varsigma_m^2} \int_{\mathbb{R}^{3m-1}} d\chi_1 \cdots d\chi_{3m-1} (A_m(I) - A_m(J))^2 \tag{3}$$

$$\begin{aligned}
\frac{1}{\varsigma_1^2} \int_{\mathbb{R}^2} d\chi_1 d\chi_2 A_1(I) A_1(J) &= \frac{1}{2} \exp\left(-\frac{(P_I - P_J)^2}{4\sigma_P^2} - \frac{(G_I - G_J)^2}{4\sigma_G^2}\right) \\
\frac{1}{\varsigma_2^2} \int_{\mathbb{R}^5} d\chi_1 \cdots d\chi_5 A_2(I) A_2(J) &= \frac{1}{2\sqrt{2}} \exp\left(-\frac{(P_I - P_J)^2}{4\sigma_P^2} - \frac{(G_I - G_J)^2}{4\sigma_G^2}\right) \\
&\quad \sum_{i \neq I}^{n_I} \xi_2(d_{iI}) \sum_{j \neq J}^{n_J} \exp\left(-\frac{(d_{jJ} - d_{iI})^2}{4\sigma_d^2} - \frac{(P_i - P_j)^2}{4\sigma_P^2} - \frac{(G_i - G_j)^2}{4\sigma_G^2}\right) \xi_2(d_{jJ}) \\
\frac{1}{\varsigma_3^2} \int_{\mathbb{R}^8} d\chi_1 \cdots d\chi_8 A_3(I) A_3(J) &= \frac{1}{16} \exp\left(-\frac{(P_I - P_J)^2}{4\sigma_P^2} - \frac{(G_I - G_J)^2}{4\sigma_G^2}\right) \\
&\quad \sum_{i \neq I}^{n_I} \sum_{j \neq J}^{n_J} \exp\left(-\frac{(d_{jJ} - d_{iI})^2}{4\sigma_d^2} - \frac{(P_i - P_j)^2}{4\sigma_P^2} - \frac{(G_i - G_j)^2}{4\sigma_G^2}\right) \\
&\quad \sum_{k \neq i, I}^{n_I} \xi_2(d_{iI}, d_{kI}, \theta_{ik}^I) \sum_{l \neq j, J}^{n_J} \exp\left(-\frac{(\theta_{ik}^I - \theta_{jl}^J)^2}{4\sigma_\theta^2} - \frac{(P_k - P_l)^2}{4\sigma_P^2} - \frac{(G_k - G_l)^2}{4\sigma_G^2}\right) \xi_3(d_{jJ}, d_{lJ}, \theta_{jk}^J)
\end{aligned} \tag{4}$$

Note that third and fourth order terms become prohibitively expensive to calculate directly. However, this can to a large extent be circumvented by slightly modifying the distributions, and solving the angular integrals in Fourier space.

2. *Response properties*

Within kernel-based regression[66–69], the total potential energy $\mathbf{U}$ of a query molecule C in its electronic ground-state, can be decomposed into a sum of local energies of its I atoms contributions, which are calculated using a basis of kernels:

$$U_C^* = \sum_{I \in C} U_{\text{local}}^*(q_I^*) = \sum_{I \in i} \sum_J \mathcal{k}(q_J, q_I^*)\, \alpha_J \tag{5}$$

where J is an atomic environment in the basis, α_I is its regression weight, and q_I is the representation of the I'th atom in the molecule.

Writing Eq. 5 in matrix form, we have:

$$\mathbf{U} = \mathbf{K}\boldsymbol{\alpha} \tag{6}$$

Note that in contrast to conventional KRR and Gaussian Process Regression (GPR) based QML models[27], this kernel matrix is no longer symmetric since it relies on atomic kernel functions as a basis set.

In this work, we approximate a response property ω, i.e. an observable which can be computed by applying a differential operator $\mathcal{O}$ acting on the energy U^*, defined in Eq. 5,

$$\omega = \mathcal{O}[\mathbf{U}] \approx \mathcal{O}[\mathbf{K}]\boldsymbol{\alpha} \tag{7}$$

The set of regression coefficients, $\boldsymbol{\alpha}$, can obtained e.g. by minimizing the Lagrangian

$$\begin{aligned} J(\boldsymbol{\alpha}) =& \sum_\gamma \beta_\gamma \|\mathcal{O}_\gamma[\mathbf{U}^{\text{ref}}] - \mathcal{O}_\gamma[\mathbf{K}\boldsymbol{\alpha}]\|^2_{L_2(\Omega_\gamma)} \equiv \\ \equiv& \sum_\gamma \beta_\gamma \int_{\Omega_\gamma} \Big[\mathcal{O}_\gamma[\mathbf{U}^{\text{ref}}] - \mathcal{O}_\gamma[\mathbf{K}\boldsymbol{\alpha}]\Big]^T \Big[\mathcal{O}_\gamma[\mathbf{U}^{\text{ref}}] - \mathcal{O}_\gamma[\mathbf{K}\boldsymbol{\alpha}]\Big] \end{aligned} \tag{8}$$

with respect to $\boldsymbol{\alpha}$ over some training set of known values of $\mathcal{O}[\mathbf{U}^{\text{ref}}]$. Ω_γ is the domain over which the corresponding operator should be minimized, e.g. all rotational degrees of freedom if the operator acts on a SO(3) group. For simplicity we pick Ω such that $\int_\Omega = 1$ for the remainder of this study. $\boldsymbol{\alpha}$ can be obtained e.g. by solving the associated normal equations or using an orthogonal factorization such as a QR or a singular-value decomposition (SVD). The corresponding normal equation (see supplementary materials for derivation) to this problem is given by

$$\boldsymbol{\alpha} = \Big[\sum_\gamma \beta_\gamma \int_{\Omega_\gamma} \mathcal{O}_\gamma[\mathbf{K}]^T \mathcal{O}_\gamma[\mathbf{K}]\Big]^{-1} \Big[\sum_\gamma \beta_\gamma \int_{\Omega_\gamma} \mathcal{O}_\gamma[\mathbf{U}^{\text{ref}}]^T \mathcal{O}_\gamma[\mathbf{K}]\Big] \tag{9}$$

However, solving the normal equations can be numerically unstable since it effectively squares the condition number, i.e. $\kappa(\mathbf{K}^T\mathbf{K}) = (\kappa(\mathbf{K}))^2$.

For the practical implementation and the results discussed in the following, an SVD factorization has been used to solve Eq. 8, as it is has several practical and efficient implementations. In contrast to the QR factorization, the SVD factorization is also numerically stable, even if $\mathbf{K}$ is rank-deficient, e.g. if $\mathbf{K}$ contains rows or columns that corresponding to atoms or molecules that are identical or only differ by symmetry operations to which the representation is invariant.

In the case of under-determined equations, the SVD factorization is performed ignoring singular values smaller than a threshold, which can be treated as a hyperparameter similarly to regularization and length-scale within ordinary KRR.

This section is dedicated to discussing some important response operators in quantum mechanics, defining the domain Ω over which the Lagrangian is to be minimized and to provide corresponding solutions to the integrals in Eq. 9.

We define the response operator for some external parameter $\vec{\eta} = \{\eta_x, \eta_y, \eta_z\}$ which can be written as $\mathcal{O}_{\delta\vec{\eta}} \equiv \dfrac{\partial}{\partial\vec{\eta}}$. Applying such an operator would map a the scalar field to a three dimensional vector field. All rotational degrees of freedom can then be integrated out with the following solutions. The solutions to the two integrals in Eq. 9, respectively, are thus

$$\int_{\Omega_{\delta\vec{\eta}}} \mathcal{O}_{\delta\vec{\eta}}[\mathbf{K}]^T \mathcal{O}_{\delta\vec{\eta}}[\mathbf{K}] = \frac{1}{3} \sum_{\nu\in x,y,z} \Big(\frac{\partial}{\partial\eta_k}\mathbf{K}\Big)^T \Big(\frac{\partial}{\partial\eta_\nu}\mathbf{K}\Big) \tag{10}$$

$$\int_{\Omega_{\delta\vec{\eta}}} \mathcal{O}_{\delta\vec{\eta}}[\mathbf{U}^{\text{ref}}]^T \mathcal{O}_{\delta\vec{\eta}}[\mathbf{K}] = \frac{1}{3} \sum_{\nu\in x,y,z} \Big(\frac{\partial}{\partial\eta_\nu}\mathbf{K}\Big)^T \Big(\frac{\partial}{\partial\eta_\nu}\mathbf{U}^{\text{ref}}\Big). \tag{11}$$

Similarly this procedure can be used to solve the equations for the second order response operator, with respect to two different perturbations $\vec{\eta}$ and $\vec{\eta'}$:

$$\begin{aligned} &\int_{\Omega_{\delta\vec{\eta}\delta\vec{\eta'}}} \mathcal{O}_{\delta\vec{\eta}\delta\vec{\eta'}}[\mathbf{K}]^T \mathcal{O}_{\delta\vec{\eta}\delta\vec{\eta'}}[\mathbf{K}] = \\ &\frac{1}{9} \sum_{\nu,\nu'\in x,y,z} \Big(\frac{\partial^2}{\partial\eta_\nu\partial\eta'_{\nu'}}\mathbf{K}\Big)^T \Big(\frac{\partial^2}{\partial\eta_\nu\partial\eta'_{\nu'}}\mathbf{K}\Big) \end{aligned} \tag{12}$$

$$\begin{aligned} &\int_{\Omega_{\delta\vec{\eta}\delta\vec{\eta'}}} \mathcal{O}_{\delta\vec{\eta}\delta\vec{\eta'}}[\mathbf{U}^{\text{ref}}]^T \mathcal{O}_{\delta\vec{\eta}\delta\vec{\eta'}}[\mathbf{K}] = \\ &\frac{1}{9} \sum_{\nu,\nu'\in x,y,z} \Big(\frac{\partial^2}{\partial\eta_\nu\partial\eta'_{\nu'}}\mathbf{U}^{\text{ref}}\Big)^T \Big(\frac{\partial^2}{\partial\eta_\nu\partial\eta'_{\nu'}}\mathbf{K}\Big) \end{aligned} \tag{13}$$

A step-by-step derivation of these equations is given in the supplementary materials.

Now we can explicitly write the matrix elements for the operators investigated within this study. In the following, the indices uppercase I, J, K correspond to atomic centers, and lower-case i, j and k correspond to molecules.

The unperturbed kernel corresponds to the energy or identity operator acting on the kernel. The elements of the unperturbed kernel $\mathbf{K}$ are given as:

$$(\mathbf{K})_{iJ} = \sum_{I\in i} k(q_J, q_I^*) \tag{14}$$

The kernel elements that correspond the force, i.e. minus the nuclear gradient operator acting on the kernel are given by:

$$-\frac{\partial}{\partial x_I^*}(\mathbf{K})_{IJ} = -\sum_{K\in i} \frac{\partial k(q_J, q_K^*)}{\partial x_I^*} \qquad \text{where} \qquad I \in i \tag{15}$$

The kernel elements that correspond to the response to the external electric field $\vec{E}$ are given by:

$$\frac{\partial}{\partial E_\nu^*}(\mathbf{K})_{i_\nu J} = \sum_{K\in i} \frac{\partial k(q_J, q_K^*)}{\partial E_\nu^*} \qquad \text{where} \qquad \nu \in \{x, y, z\} \tag{16}$$

Similarly, the nuclear Hessian kernel is given by:

$$\frac{\partial^2}{\partial x^*_{I'}\partial x^*_I}\left(\mathbf{K}\right)_{I'IJ} = \sum_{K\in i}\frac{\partial \mathcal{k}\left(q_J, q^*_K\right)}{\partial x^*_{I'}\partial x^*_I} \qquad \text{where} \qquad I', I \in i \tag{17}$$

Lastly, the kernel that yields the dipole derivatives necessary for the infrared intensities is written as the mixed second order derivative,

$$\frac{\partial^2}{\partial E^*_\nu \partial x^*_I}\left(\mathbf{K}\right)_{i_\nu IJ} = \sum_{K\in i}\frac{\partial \mathcal{k}\left(q_J, q^*_K\right)}{\partial E^*_\nu \partial x^*_I} \quad \text{where} \quad I \in i \quad \text{and} \quad \nu \in \{x, y, z\} \tag{18}$$

We are not aware of any other QML model which can account for these effects.

D. Data sets for novel representation and response work

The data sets newly developed and used for the studies published in Refs.[3,4] are so modest in scope that we refer to these papers for the description. Both studies predominantly relied on data sets which were previously published in the literature.

IV. RESULTS AND DISCUSSION

A. Results Elpasolites

Numerical results on display in Fig. 1(c) indicate systematic improvement of the predictive accuracy of the ML model with increasing training set size, for all three datasets. The inset details normally distributed errors and scatter plots which systematically improve with training set size for (III−VI) and (I−VIII) machine. The accuracy of our ML model can be compared to that of semi-local DFT as used in our data sets. Lany [70] reported prediction errors for heats of formation for general chemistries with filled *d*-shells which (assuming normal distributions) translates to a MAE of at least 0.19 eV/atom [71]. For transition metal oxides, results of Hautier et al. correspond to MAEs of at least 0.055 eV/atom (0.019 for DFT+U), but such errors are expected to increase when going beyond oxides, as in our datasets. For a training set of 10k, our ML model reaches a MAE of 0.1 eV/atom, which is roughly at the level of accuracy of semi-local DFT formation energies.

The converged performance for using all crystals of the (III−VI) data set as training set confirms that our representation captures all the information of a crystal necessary to determine its energy. While errors decay monotonically, the learning rate levels off for the (III−VI) data set when N approaches 10k. This is due to the employed relaxation threshold in the DFT calculations of ±10 meV/atom. Any inductive model will obviously fail to go below this level, and only numerically more precise reference numbers would mitigate this issue. In all validation tests dealing with energy predictions for random out-of-sample crystals, the ML model performance meets the expectations set in Fig. 1(c). For example, drawing 100 crystals at random from (III−VI) and (I−VIII), ML models perform as expected when compared to the result from validating DFT calculations.

Having established the performance of the ML model, we have subsequently used the 10 k training set model (I−VIII) for investigation of the elpasolite universe. Estimated formation energies for *all* 2 M elpasolites are featured in Fig. 2. The formation energies are clearly dominated by the chemical identity of position 4, followed by position 3 but according to a different pattern. Chemical identity at position 1 and 2 has the smallest influence and very similar impact. Due to the effective degeneracy of positions 1 and 2, all inner matrices in Fig. 2 appear largely symmetric. Figure 1(d) shows the average contribution of each element to the formation energies estimated by the 10 k ML model. These average contributions per element are used to order the elements in Fig. 2 to yield the smoothest elpasolite map. Arranging elements by their nuclear charge, or by their Pettifor order[72], results in a much more oscillatory map or stripe-like pattern due to underlying periodicities.

Figures 1(d) visualize the bonding emergent from the geometry and bond coordination of the elpasolite crystal structure.Fluorine and carbon are at the respective ends of the global scale of low and high formation energies. But also alkaline metals, alkaline earth metals, and oxygen contribute to lowering the formation energy. On average, the formation energies of elpasolites involving halogens, alkaline metals, noble gases increase as the periodic table is descended. The opposite holds for all other elements, except oxygen, boron, carbon and nitrogen, which all have a noticeably higher average formation energy than any other element. A saddle point can also be observed in the midst of the periodic table table as well as two valleys along the halogen and alkaline earth rows. Site-specific resolution indicates that fluorine fits best with the bond coordination of sites 1, 2, and 4, whereas the same does not apply to later halogens. In contrast, as the element on site 3 goes down column II in the periodic table, the formation energy is successively lowered, with Ca, Sr, and Ba contributing more than any halogen atom. On sites 1 and 2, the formation energy generally increases the most for heavy noble gases. On sites 3 and 4, it is carbon, followed by neighboring B and N that increase the formation energy the most. The accuracy of linear single atom energy models based on these scales, however, is not on par with the ML-model, and—maybe more importantly—cannot be improved systematically through increasing training set sizes but rather converges to a finite residual error.

In order to achieve satisfying accuracy of ±0.1 eV/atom for elpasolites, a relatively large training set of 10 k is needed. This is likely due to the sparsity

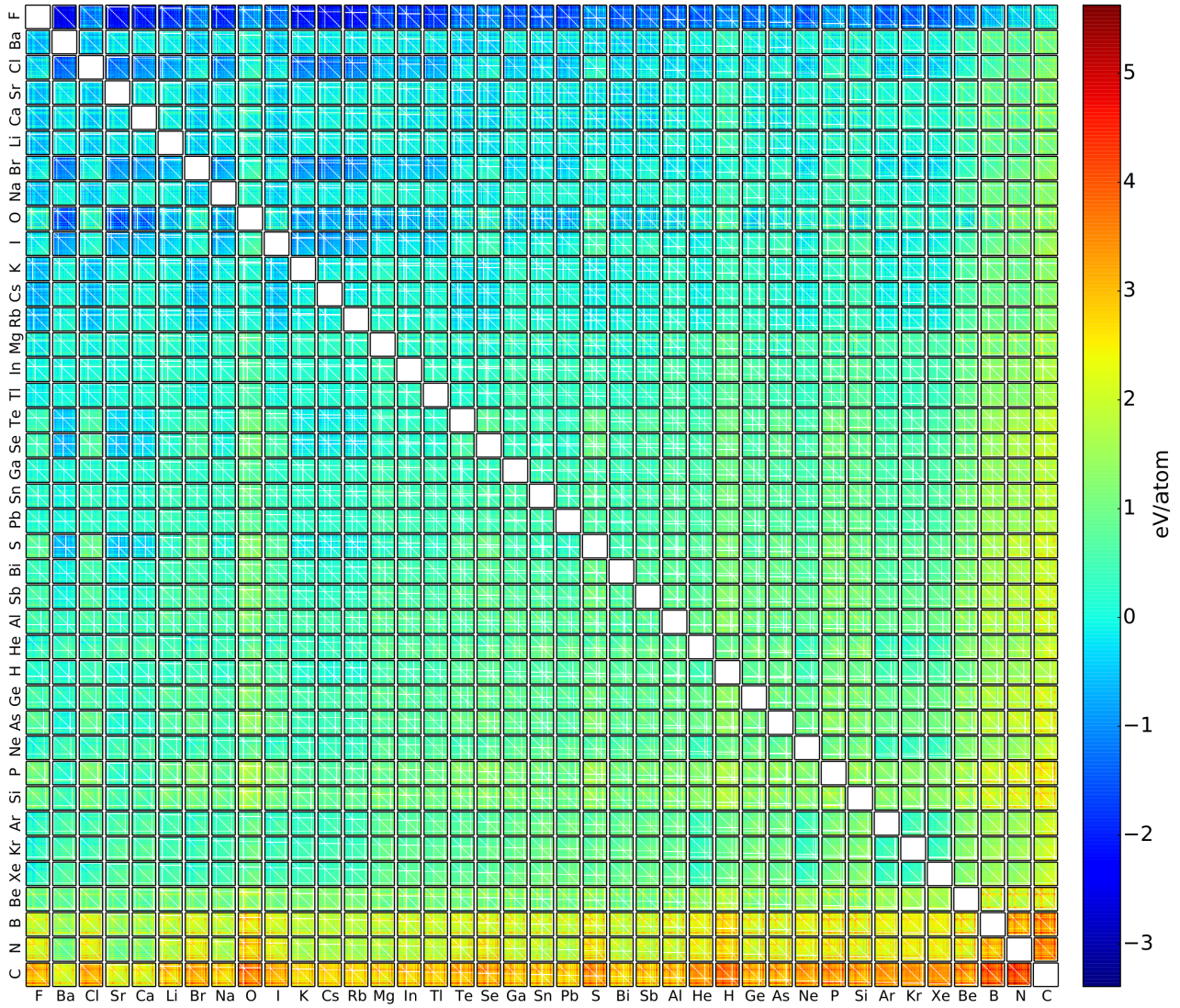

FIG. 2. Formation energies for all 2 M elpasolites made up of all main-group elements up to Bi predicted by the 10 k ML-model. The outer vertical and horizontal axis correspond to x_4 and x_3 symmetry position, respectively. Inner vertical and horizontal axis correspond to x_2 and x_1 symmetry position, respectively. Elemental sequence follows the elpasolite order of Fig. 1(d). White pixels correspond to subspaces of ternary, binary, or elementary non-elpasolite crystals.

of crystals at the opposite ends of the high and low formation energy spectrum; this results in a decreased predictive ML model accuracy for crystals in these regions. Nevertheless, the 10 k ML model readily identifies a larger set of lowest lying elpasolites for which the actual DFT minima can be obtained through subsequent DFT based screening. This is shown in Fig. 1(e) where the 250 crystals with the lowest ML predicted formation energies are shown in ascending order. Subsequent screening with DFT indicates the 26^{th} crystal $CaSrCs_2F_6$ (out of 2M) to be the global formation energy minimum at −3.44 eV/atom, closely followed a near-degenerate isomer $SrCaCs_2F_6$. The DFT energies of the next two degenerate pairs $CaSrRb_2F_6$/$SrCaRb_2F_6$ and $CaBaCs_2F_6$/$BaCaCs_2F_6$ correspond to −3.41, and −3.39 eV/atom, respectively. Overall, the elpasolites with the

most favorable formation energies, ABC_2D_6, correspond to A and B being late elements from group (II), and C and D being a late element from group (I) and fluoride, respectively. Populating the four sites with elements from groups (II),(II),(I), and (VIII), respectively, differs from the experimentally established stoichiometry $AlNaK_2F_6$. In fact, the lowest DFT energy crystal with a group-(III) element is $CsAlRb_2F_6$ (in 69^{th} position) with -3.09 eV/atom (ML energy: -2.96 eV/atom.

We have also used our predictions to analyse atomic oxidation states in elpasolites. In particular, we have found that roughly 6 % of the crystals with formation energies below -1 eV/atom exhibit unusual atomic charges: They are low in energy despite the fact that no combination of conventional atomic charges would result in a neutral system. In order to identify these crystals, we have used the absolute value of the lowest possible total oxidation state (LPTOS) that could possibly be realized using a list of typical atomic oxidation states on display in Table I. The lowest lying crystals have a LPTOS of 0 (-3 to -3.44 eV/atom formation energies). However, already at -3 eV/atom crystals with LPTOS of 2 or 1 start to occur. At formation energies of ~ -1.25 eV/atom and higher, the number of crystals with non-zero LPTOS increases rapidly, with LPTOS as high as 12. Corresponding crystal frequency distributions are shown in Fig. 1(e), along with formulas for the mutually lowest lying crystals. Interestingly, the number of crystals with zero LPTOS increases monotonically with formation energy, while for nonzero LPTOS crystals the distribution is oscillatory. In order to identify elements with unusual oxidation states we report atomic charges obtained according to Bader's scheme[73–75] in Table I for the 10 lowest lying crystals with non-zero LPTOS. We found the Bader analysis to indicate atomic charges consistent (after rounding to the next integer) with the conventional oxidation states in Table I for 95% of the 250 lowest lying crystals with zero LPTOS. Not surprisingly, due to its strong electronegativity, F always conserves its negative oxidation state of $\sim$-1 in x_4 position. For $CsMgRb_2F_6$, $CsMgK_2F_6$, or $BaNaCs_2F_6$, no unusual atomic charge is found, the non-zero LPTOS being rather due to the accumulation of relatively low charges on the six fluoride atoms. For the remaining seven crystals, however, unusual charges are found for atoms late in the periodic table and populating the energetically weakly contributing x_1 or x_2 sites. In particular, Bader's charge analysis indicate unusual oxidation states for elements Sb (0;1) and Te (0), suggesting that new chemistries could be explored for compounds involving these elements.

B. Results representation

Fig. 3 displays the performance overview for energy predictions on six different data sets (QM9, QM7b, SSI, water, elapsolites, OQMD). Mean absolute out-of-sample energy prediction errors are shown as a function of training set size. The results indicate remarkable performance for all data sets, indicating a well-working QML model yielding systematic improvement with increasing training set size. The learning curves also indicate out-of-sample MAEs which are consistently lower, or similar, than previously published models in the literature. For QM9, the MAE reaches the highly coveted chemical accuracy threshold (1 kcal/mol or ~ 0.043 eV for enthalpy of formation) with only 2k training points on the QM9 dataset. Previously published QML models had to include an order of magnitude more training molecules to reach such accuracy. This is similar to the amount of training molecules necessary when using the Coulomb matrix representation in conjunction with semi-empirical or DFT based baselines in order to estimate electron correlated energies, as demonstrated in 2015 with the Δ-ML model[58].

For QM9, atomic Spectral London Axilrod-Teller-Muto (aSLATM)[61] and SOAP multi kernel model[76,77] reach a performance nearly as good as our QML model. aSLATM, however, performs worse for the SSI and the Water cluster. The SOAP multi kernel QML model, however, performs an expansion in kernel function space acting on the distance for which all degrees of freedom have already been integrated out. As such it is, strictly speaking, not the same as as an improved representation, but rather an improved regressor. Note that single kernel based SOAP QML models perform significantly worse. The reader should take notice however that in the SOAP learning curve results presented in Fig 3, the $\sim$3k structures which had failed the SMILES consistency test, were included. As such, theses QML models are not exactly comparable, and the SOAP results are still likely to slightly improve if these faulty structures were to be removed. One should also note that the SOAP results shown for QM7b correspond to the multi-kernel SOAP kernel[41,77].

Other models presented correspond to Coulomb matrix (CM)[21], bags of bonds (BOB)[78], Bonds and Angles based Machine Learning (BAML)[79], Histogram of Distances, Angles, and Dihedrals (HDAD)[17], Spectral London Axilrod-Teller-Muto (SLATM), aSLATM[61], the crystal representation by Faber, Lindmaa, Lilienfeld, Armiento (FLLA)[2], the Sinematrix[1], and the many-body tensor representation (MBTR)[62]. We also compared to QML models which are not based on KRR, such as the message passing neural network model (enn-s2s)[19], and a Voronoi-tessellation based random forest model (Voronoi)[80].

The MAE of our new QML model is consistently the lowest for all data sets and large training sets. For the set of 4,000 non-equilibrium water clusters, there is a noticeable difference between the global (CM, BOB and SLATM) and the atomic representations (i.e., aSLATM and the new model we introduce in this work): The global models exhibit very little learning at first, only for larger N the learning curves begin to turn downward. The atomic models, however, our new representation based

TABLE I. Calculated atomic charges for the 10 lowest formation energy crystals with non-zero LPTOS. ML and DFT energies in eV/atom. Values for elements of unusual oxidation states are printed in bold.

Formula	LPTOS	E_{ML}	E_{DFT}	q_1	q_2	q_3	q_4
$MgSbBa_2F_6$	1	-2.88	-2.70	1.66	**0.42**	1.63	-0.89
$CaTeBa_2F_6$	1	-2.90	-2.68	1.58	**0.31**	1.67	-0.87
$TeCaBa_2F_6$	1	-2.83	-2.68	**0.31**	1.59	1.67	-0.87
$LiSbBa_2F_6$	2	-3.06	-2.62	0.89	**1.06**	1.62	-0.86
$CsMgRb_2F_6$	1	-2.93	-2.61	0.98	1.67	0.92	-0.75
$BeSbBa_2F_6$	2	-2.88	-2.60	1.68	**0.35**	1.62	-0.88
$CsMgK_2F_6$	1	-2.97	-2.58	1.01	1.68	0.92	-0.75
$SrSbBa_2F_6$	2	-2.90	-2.56	1.48	**0.60**	1.59	-0.88
$SrTeBa_2F_6$	2	-2.89	-2.55	1.70	**0.40**	1.66	-0.90
$BaNaCs_2F_6$	1	-2.84	-2.52	1.69	0.86	0.92	-0.73.

QML model as well as aSLATM, improve rapidly with increasing training data set size. We believe that sorting and crowding in the global representations makes it difficult to accurately account for the purely geometrical changes in structures that contribute to total energy variations.

Impressive predictive power is also observed for the OQMD dataset, a structurally and compositionally very diverse set of solids. Our new model has a lower out-of-sample MAE for all N when compared to the sine matrix representation on the OQMD dataset. The offset of the learning curve of our new model is larger compared to that of the Voronoi-based random-forest model[80]. However, the learning rate of our QML model is significantly steeper, surpassing the Voronoi model already at just $\sim$ 250 training samples. Results for a solid state variant of the CM, designed for use in periodic systems, has also been included (SineMatrix)[1]. It has a similar slope as the Voronoi model, but a substantially larger off-set.

For the elpasolite data set,[2], with large composition diversity but identical crystal structures, the learning-curve of the FLLA representation has a slightly higher off-set than our new QML model, yet exhibits a steeper learning curve. Our model converges towards the same slope for larger training set sizes. We can only speculate on the reasons for such behavior. The FLLA representation differs qualitatively from the other representations in this study: It does not include any explicit information about coordinates and only encodes periodic row and column of the elements which populate each crystal structure site. The QML model then learns to infer ground state energies without knowing the exact configuration. This leads to a very low dimensional model that is still unique for the system, which might be the cause of the lower slope. This however needs to be investigated more carefully before any conclusions can be drawn.

Further promising results are presented and discussed in Ref.[3], including predictions of multiple electronic ground state properties as well of energies for molecules containing elements which were absent in training.

C. Results response properties

Here, a selection of results published in Ref.[4] is discussed. In particular, we rely on evidence obtained for a molecular dynamics data set (MD17) consisting of eight organic molecules, as well as of 4000 data points for a set of isomers (ISO17).

Here we use the FCHL* representation within the presented machine learning algorithm to study two existing benchmark sets for learning forces and energies. The MD17 consists of molecular dynamics (MD) snapshots from MD trajectories of 8 different molecules, for which reference forces and energies are given[45]. Similarly, the ISO17 consists of MD snapshots of isomers with the chemical formula $C_7O_2H_{10}$. The ISO17 additionally comes with two different test sets[47,81]. One that consists only of isomers with a connectivity that is present in the training set ("known"), and one that only contains isomers with connectivity that is not present in the training set ("unknown"). Briefly the two datasets benchmark the conformational freedoms and constitutional freedoms of molecules, respectively. Since there is no electric field applied to the molecules in these data sets, note that the FCHL* representation reduces to the original FCHL representation[3].

Learning curves for the two datasets are displayed in Figures 4 and 5. For the MD17 dataset, the out-of-sample MAE errors of predicted energies are similar between FCHL*, GDML and SchNet, with SchNet being slightly less accurate in most cases (See Fig. 4). FCHL* and SchNet perform best for ethanol and malonhaldehyde, while GDML is the best for Salicyclic acid and naphtalene. The case of benzen and uracil is interesting. For benzene, all models show little to no progress for energies at rather low error values, and the force learning is very weak. Uracil is best modeled by GDML, with relatively poor SchNet forces, and FCHL being in between. At this point, we remind the reader that the GDML approach is only applicable to a given system, while FCHL* and SchNet are capable of learning across chemical space.

Performance across constitutional space is tested on the constitutional isomers in the ISO17 dataset (Fig. 5). For the two test sets of "known" and "unknown"

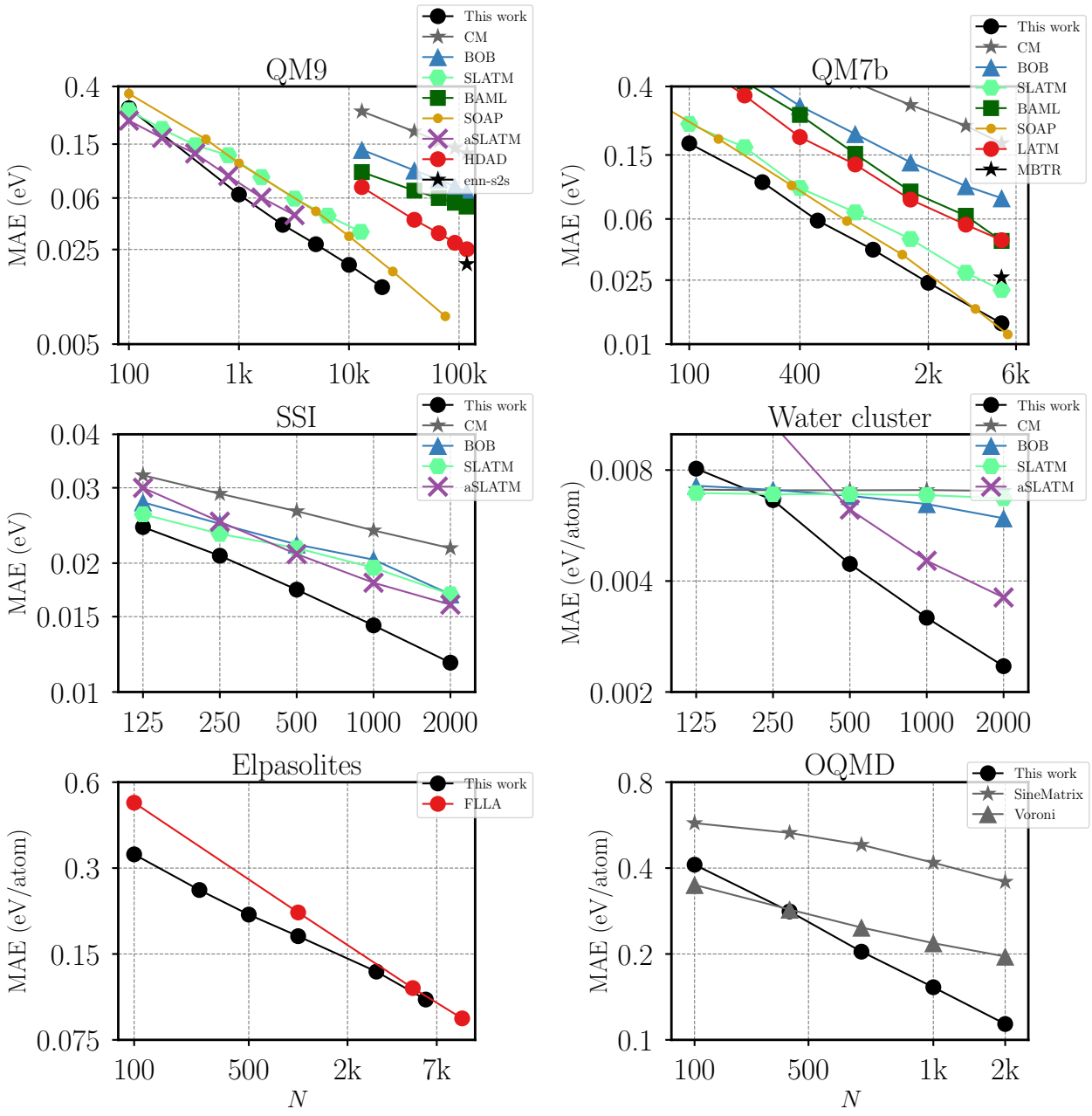

FIG. 3. Learning curves for atomization/formation energy predictions corresponding to various QML models. Out-of-sample MAE is shown as a function of training set size for molecules (QM9 and QM7b), protein side-chain dimers (SSI), liquid water ($(H_2O)_{40}$ snapshots (Water cluster) and crystalline (OQMD and Elpasolites) data-sets.

molecules in the ISO17, the FCHL* model displays a good learning rate, that is qualitatively comparable to the SchNet model. Note that here, the name "known" only implies that the isomers of the same constitution are known to the machine, but not the conformations in the test set. Unfortunately the learning curves between the FCHL* models and SchNet do not overlap, so the two models cannot be compared quantitatively here, but the out-of-sample accuracy seems comparable.

Overall, we find that our operator approach leads to forces with state-of-the-art accuracy, on par with two of the most accurate models already published in literature.

Prediction errors of machine learning models of dipole moments converge slowly for conventional QML models[3,17,26]. Here we demonstrate how including the underlying physics for the dipole moment into the representation improves the learning rate, compared to learning the dipole norm with conventional kernel-ridge regres-

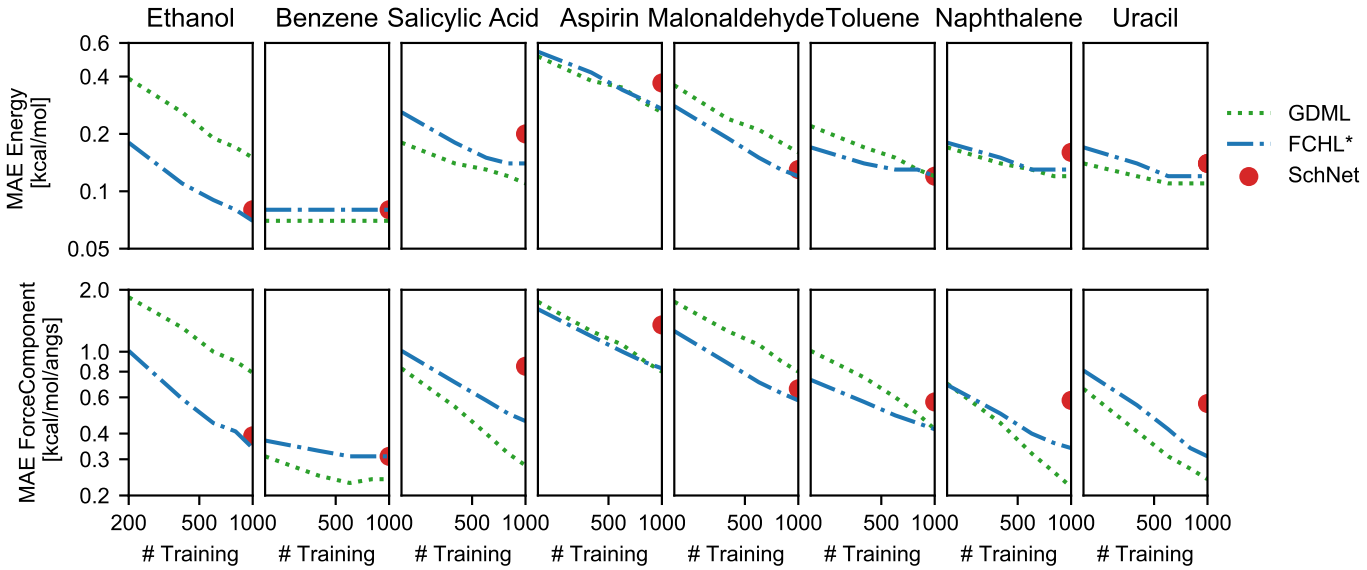

FIG. 4. The two figures show the learning curves of our model for the MD17 dataset, for each of the eight molecules. The out-of-sample MAE energy prediction (top row) and MAE force component prediction (bottom row) is shown for the presented FCHL* (blue) model as well as for the GDML[45] (green) and SchNet models (red).[47,81]

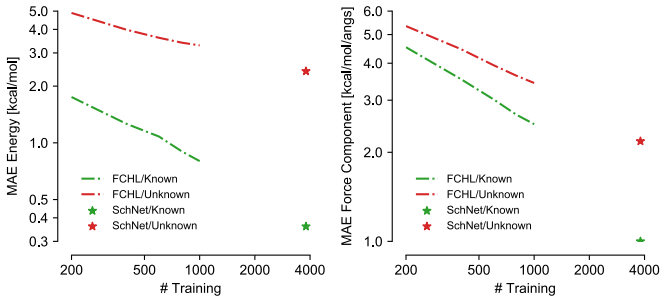

FIG. 5. The two figures show the learning curves of our model for the ISO17 dataset, in addition the accuracy for SchNet when using 4,000 training samples is shown. Left shows the out-of-sample MAE energy prediction for a set of isomers known to the trained machine ("known") and for a set of unknown to the machine ("unknown"). Right shows the out-of-sample MAE force prediction for the same two sets. Note that "known" in this context only concerns whether the isomers are included in the training set or not. In both cases only isomers with a conformation unknown to the machine are used to as test data.

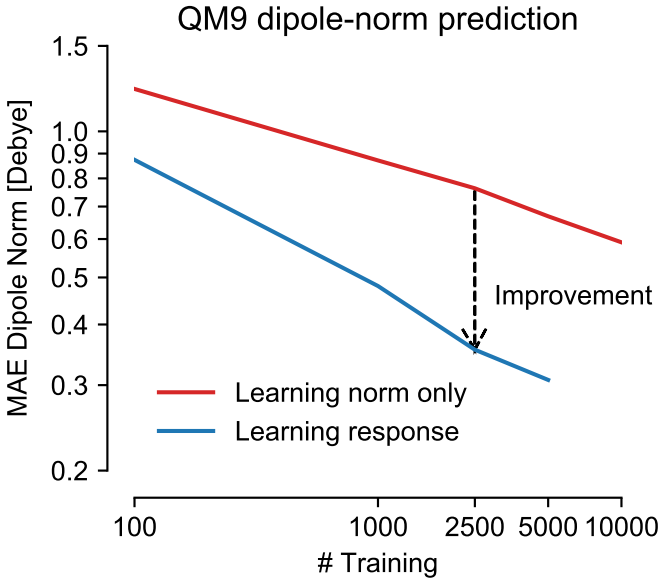

FIG. 6. The figure displays the out-of-sample prediction error of the dipole norm as a function of QM9 training data set size. The red curve corresponds to a conventional KRR model learning the scalar with the original FCHL representation (taken from Faber *et al.*[3]). The blue curve shows the predictions from a machine trained on the energy and dipole moments of QM9 molecules, which in turn predict the dipole vector, from which the norm is calculated.

sion. We compare two approaches to learn the dipole moment norm of the molecules in QM9; (1) using the FCHL* representation with the aforementioned machine learning approach to fit the dipole moments as derivatives of the energy, and (2) simply learning the dipole moment norm as a scalar using kernel-ridge regression with the FCHL representation as done in our earlier paper[3]. The learning curves of the two models are displayed in Fig. 6. The MAE out-of-sample predicted dipole moment norm is decreased substantially with our new approach. For instance, training on 5000 random molecules, the out-of-sample MAE error is reduced by 54% (From 0.67 Debye to 0.31 Debye). We also note that not only is the new learning curve offset lower than the conventional learning curve, but it is also substantially steeper, showing the strength of the approach due to using the correct response operators in the kernel to learn the corresponding

response properties.

Further promising results, for example for the prediction of normal modes and IR spectra, are discussed in Ref.[4].

V. CONCLUSIONS

In conclusion, in the beginning of this grant, we have developed and used ML-models of formation energies to investigate all possible elpasolites made up of main-group elements. We have presented numerical results for ∼2 M formation energies. The ML-model is only implicitly dependent on spatial coordinates, through reference data used for training. No spatial coordinates are needed for new queries, yet for a training set of 10 k crystals the model reaches ±0.1 eV/atom—comparable to DFT accuracy for solids. The results have been used to identify the most strongly bound elpasolites as well as to investigate energy and bonding trends at crystal structure sites, leading to a new "elpasolite order" of elements, consistent with the bonding physics in the elpasolite crystal structure. Crystals with lowest lying formation energies have been identified, and using Bader's charge analysis, Te and Sb have been found to exhibit unconventional oxidation states. During the second year we have performed and extensive thermodynamic analysis of our ML results. Making use of the materials project database we could use our ML predictions to identify 90 new and thermodynamically stable elpasolite crystals. While our analysis does not offer a 100% guarantee that all possible elpasolites made up of main group elements have been discovered in an exhaustive manner, it does seem extremely likely: Out of 200'000 crystals considered, 90 met the DFT criterion which is currently state of the art. Among these 90 crystals we identified one crystal with a very unusual negative oxidation state for Al. We believe that our results hold great promise for the computational screening of polymorphs, other crystal structure symmetries, solid mixtures, phase transitions, or defects at unprecedented rate and extent. Other crystal properties than energies could also be considered.

Subsequently, in year 3 and 4 we have developed a more universal representation which results in machine learning models with unprecedented performance, as assessed by learning curves, throughout materials compound space. It is applicable to any material, i.e. to molecules as well as clusters as well as periodic condensed matter[3]. Due to the quantum philosophy of the approach, other extensive properties are also accounted for with remarkable predictive power.

Most recent work from year 4, not yet published but already available at the `arxiv.org`, has dealt with the development of a ML methodology which can be used to train and predict forces and other response properties throughout chemical space[4]. State-of-the art predictive power for forces is obtained, as well as unprecedented accuracy for predicting other response properties, such as dipole moments.

ACKNOWLEDGEMENT

All material reported on is based upon work partially supported by the Air Force Office of Scientific Research, Air Force Material Command, USAF under Award No. FA9550-15-1-0026.

VI. REFERENCES

[1] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento. Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.*, 115:1094, 2015. `http://arxiv.org/abs/1503.07406`.

[2] Felix A. Faber, Alexander Lindmaa, O. Anatole von Lilienfeld, and Rickard Armiento. Machine learning energies of 2 million elpasolite (abC_2D_6) crystals. *Phys. Rev. Lett.*, 117:135502, Sep 2016.

[3] 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.

[4] Anders S Christensen, Felix A Faber, and O Anatole von Lilienfeld. Operators in machine learning: Response properties in chemical space. *arXiv preprint arXiv:1807.08811*, 2018.

[5] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.*, 108:058301, Jan 2012.

[6] Alec Belsky, Mariette Hellenbrandt, Vicky Lynn Karen, and Peter Luksch. New developments in the inorganic crystal structure database (ICSD): accessibility in support of materials research and design. *Acta Crystallographica Section B Structural Science*, 58(3):364–369, May 2002.

[7] G. Bergerhoff, R. Hundt, R. Sievers, and I. D. Brown. The inorganic crystal structure data base. *Journal of Chemical Information and Computer Sciences*, 23(2):66–69, May 1983.

[8] Pin Yang, F. Patrick Doty, Mark A. Rodriguez, Margaret R. Sanchez, Xiaowong Zhou, and Kanai S. Shah. The synthesis and structures of elpasolite halide scintillators. In *Symposium L Nuclear Radiation Detection Materials 2009*, volume 1164 of *MRS Online Proceedings Library*, 2009.

[9] Koushik Biswas and Mao-Hua Du. Energy transport and scintillation of cerium-doped elpasolite Cs_2LiYCl_6: Hybrid density functional calculations. *Phys. Rev. B*, 86:014102,

Jul 2012.

[10] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964.

[11] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965.

[12] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics*, 15(9):095003, 2013.

[13] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B*, 89:205118, May 2014.

[14] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B*, 89:094104, Mar 2014.

[15] Frank Jensen. *Introduction to Computational Chemistry*. John Wiley, West Sussex, England, 2007.

[16] Felix Brockherde, Li Li, Mark E. Tuckerman, Kieron Burke, and Klaus-Robert Müller. By-passing the kohn-sham equations with machine learning. *Nature Communications*, 8:872, 2017.

[17] Felix A Faber, Luke Hutchison, Bing Huang, Justin Gilmer, Samuel S Schoenholz, George E Dahl, Oriol Vinyals, Steven Kearnes, Patrick F Riley, and O Anatole von Lilienfeld. Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theory Comput.*, 13:5255–5264, 2017.

[18] Kristof T. Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R. Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.*, 8:13890, 2017.

[19] 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, ICML 2017*, 2017.

[20] 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 2215–2223, 2015.

[21] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.*, 108:058301, 2012.

[22] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad. Accelerating materials property predictions using machine learning. *Scientific reports*, 3:2810, 2013.

[23] Corinna Cortes, Lawrence D Jackel, Sara A Solla, Vladimir Vapnik, and John S Denker. Learning curves: Asymptotic values and rate of convergence. In *Advances in Neural Information Processing Systems*, pages 327–334, 1994.

[24] K. R. Müller, M. Finke, N. Murata, K. Schulten, and S. Amari. A numerical study on learning curves in stochastic multilayer feedforward networks. *Neural Comp.*, 8:1085, 1996.

[25] O. Anatole von Lilienfeld. Quantum machine learning in chemical compound space. *Angewandte Chemie International Edition*, 57:4164, 2018. http://dx.doi.org/10.1002/anie.201709686.

[26] R. Ramakrishnan and O. A. von Lilienfeld. Many Molecular Properties from One Kernel in Chemical Space. *CHIMIA*, 69:182, 2015. `http://arxiv.org/abs/1502.04563`.

[27] Raghunathan Ramakrishnan and O. Anatole von Lilienfeld. *Machine Learning, Quantum Chemistry, and Chemical Space*, volume 30, pages 225–256. John Wiley & Sons, Inc., 2017.

[28] Xavier Gonze. Adiabatic density-functional perturbation theory. *Physical Review A*, 52(2):1096, 1995.

[29] A. Putrino, D. Sebastiani, and M. Parrinello. Generalized variational density functional perturbation theory. *J. Chem. Phys.*, 113:7102–7109, 2000.

[30] R. G. Parr and W. Yang. *Density functional theory of atoms and molecules*. Oxford Science Publications, 1989.

[31] P. Geerlings, F. De Proft, and W. Langenaeker. Conceptual density functional theory. *Chem. Rev.*, 103:1793, 2003.

[32] O. A. von Lilienfeld. First principles view on chemical compound space: Gaining rigorous atomistic control of molecular properties. *International Journal of Quantum Chemistry*, 113(12):1676–1689, 2013.

[33] Stijn Fias, Farnaz Heidar-Zadeh, Paul Geerlings, and Paul W Ayers. Chemical transferability of functional groups follows from the nearsightedness of electronic matter. *Proceedings of the National Academy of Sciences*, 114(44):11633–11638, 2017.

[34] Wiktor Pronobis, Alexandre Tkatchenko, and Klaus-Robert Müller. Many-body descriptors for predicting molecular properties with machine learning: Analysis of pairwise and three-body interactions in molecules. *Journal of chemical theory and computation*, 2018.

[35] R. P. Feynman. Forces in molecules. *Phys. Rev.*, 56:340, 1939.

[36] B. G. Sumpter and D. W. Noid. Potential energy surfaces for macromolecules. a neural network technique. *Chemical Physics Letters*, 192(56):455 – 462, 1992.

[37] S. Lorenz, A. Gross, and M. Scheffler. Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks. *Chem. Phys. Lett.*, 395:210, 2004.

[38] S. Manzhos and T. Carrington, Jr. A random-sampling high dimensional model representation neural network for building potential energy surfaces. *J. Chem. Phys.*, 125:084109–084123, 2006.

[39] J. Behler and M. Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98:146401, 2007.

[40] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.*, 104(13):136403, 2010.

[41] Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87:184115, May 2013.

[42] Sergei Manzhos, Richard Dawes, and Tucker Carrington. Neural network-based approaches for building high dimensional and quantum dynamics-friendly potential energy surfaces. *Int. J. Quantum Chem.*, 115(16):1012–1020, 2015.

[43] Venkatesh Botu and Rampi Ramprasad. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *Int. J. Quantum Chem.*, 115(16):1074–1083, 2015.

[44] M. Rupp, R. Ramakrishnan, and O. A. von Lilienfeld. Machine learning for quantum mechanical properties of atoms in molecules. *J. Phys. Chem. Lett.*, 6:3309, 2015. `http://arxiv/abs/1505.00350`.

[45] Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5):e1603015, 2017.

[46] Justin Steven Smith, Olexandr Isayev, and Adrian E Roitberg. Ani-1: An extensible neural network potential with dft accuracy at force field computational cost. *Chem. Sci.*, 8(4):3192–3203, 2017.

[47] 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.

[48] Andrea Grisafi, David M Wilkins, Gábor Csányi, and Michele Ceriotti. Symmetry-adapted machine learning for tensorial properties of atomistic systems. *Physical review letters*, 120(3):036002, 2018.

[49] Aldo Glielmo, Peter Sollich, and Alessandro De Vita. Accurate interatomic force fields via machine learning with covariant kernels. *Physical Review B*, 95(21):214302, 2017.

[50] Kristof T. Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R. Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8:13890, 2017.

[51] Andrew E. Sifain, Nicholas Lubbers, Benjamin T. Nebgen, Justin S. Smith, Andrey Y. Lokhov, Olexandr Isayev, Adrian E. Roitberg, Kipton Barros, and Sergei Tretiak. Discovering a Transferable Charge Assignment Model Using Machine Learning. *ChemRxiv*, 6 2018.

[52] Benjamin Nebgen, Nicholas Lubbers, Justin S Smith, Andrew E Sifain, Andrey Lokhov, Olexandr Isayev, Adrian E Roitberg, Kipton Barros, and Sergei Tretiak. Transferable dynamic molecular charge assignment using deep neural networks. *Journal of chemical theory and computation*, 2018.

[53] Michael Gastegger, Jörg Behler, and Philipp Marquetand. Machine learning molecular dynamics for the simulation of infrared spectra. *Chemical science*, 8(10):6924–6935, 2017.

[54] Kristof T Schütt, Michael Gastegger, Alexandre Tkatchenko, and Klaus-Robert Müller. Quantum-chemical insights from interpretable atomistic neural networks. *arXiv preprint arXiv:1806.10349*, 2018.

[55] L. Ruddigkeit, R. van Deursen, L.C. Blum, and J.-L. Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *J. Chem. Inf. Model.*, 52:2684, 2012.

[56] R. Ramakrishnan, P. Dral, M. Rupp, and O. A. von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:140022, 2014.

[57] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer, New York, 2nd ed. 2009. corr. 7th printing 2013 edition edition, April 2011.

[58] R. Ramakrishnan, P. Dral, M. Rupp, and O. A. von Lilienfeld. Big Data meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.*, 11:2087, 2015.

[59] Luca M. Ghiringhelli, Jan Vybiral, Sergey V. Levchenko, Claudia Draxl, and Matthias Scheffler. Big data of materials science: Critical role of the descriptor. *Phys. Rev. Lett.*, 114:105503, Mar 2015.

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

[61] Bing Huang and O. Anatole von Lilienfeld. The "DNA" of chemistry: Scalable quantum machine learning with "amons". *arXiv preprint arXiv:1707.04146*, 2017. submitted to Nature.

[62] Haoyan Huo and Matthias Rupp. Unified representation for machine learning of molecules and crystals. *arXiv preprint arXiv:1704.06439*, 2017.

[63] B. M. Axilrod and E. Teller. Interaction of the van der Waals type between three atoms. *J. Chem. Phys.*, 11:299, 1943.

[64] Y. Muto. Force between nonpolar molecules. *J. Phys.-Math. Soc. Japan*, 17:629, 1943.

[65] Anders S Christensen, Marcus Elstner, and Qiang Cui. Improving intermolecular interactions in dftb3 using extended polarization from chemical-potential equalization. *The Journal of chemical physics*, 143(8):084123, 2015.

[66] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201, 2001.

[67] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[68] Vladimir Vovk. *Kernel Ridge Regression*, pages 105–116. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[69] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: data mining, inference and prediction*. Springer series in statistics. Springer, New York, N.Y., 2001.

[70] Stephan Lany. Semiconductor thermochemistry in density functional calculations. *Phys. Rev. B*, 78:245207, Dec 2008.

[71] R. C. Geary. The ratio of the mean deviation to the standard deviation as a test of normality. *Biometrika*, 27(3/4):310–332, 1935.

[72] D.G. Pettifor. *Bonding and Structure of Molecules and Solids*. Oxford university press, 2002.

[73] W Tang, E Sanville, and G Henkelman. A grid-based bader analysis algorithm without lattice bias. *Journal of Physics: Condensed Matter*, 21(8):084204, 2009.

[74] Edward Sanville, Steven D. Kenny, Roger Smith, and Graeme Henkelman. Improved grid-based algorithm for bader charge allocation. *Journal of Computational Chemistry*, 28(5):899–908, 2007.

[75] Graeme Henkelman, Andri Arnaldsson, and Hannes Jònsson. A fast and robust algorithm for bader decomposition of charge density. *Computational Materials Science*, 36(3):354 – 360, 2006.

[76] Albert P. Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R. Kermode, Gábor Csányi, and Michele Ceriotti. Machine learning unifies the modeling of materials and molecules. *Science Advances*, 3(12), 2017.

[77] Sandip De, Albert P. Bartok, Gabor Csanyi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.*, 18:13754–13769, 2016.

[78] K. Hansen, F. Biegler, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko. Interaction potentials in molecules and non-local information in chemical space. *J. Phys. Chem. Lett.*, 6:2326, 2015.

[79] Bing Huang and O. Anatole von Lilienfeld. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *J. Chem. Phys.*, 145(16), 2016.

[80] Logan Ward, Ruoqian Liu, Amar Krishna, Vinay I Hegde, Ankit Agrawal, Alok Choudhary, and Chris Wolverton. Including crystal structure attributes in machine learning models of formation energies via voronoi tessellations. *Physical Review B*, 96(2):024104, 2017.

[81] Kristof T Schütt, P-J Kindermans, Huziel E Sauceda, Alexandre Tkatchenko, and K-R Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *arXiv preprint arXiv:1706.08566*, 2018.

VII. LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

ABC_2D_6 General stoichiometry formula for elpasolite crystals. Atoms A, B, C, and D occupy the respective Wyckoff sites 1, 2, 3, and 4, as shown in Fig. 1. For example, the elpasolite prototype corresponds to $AlNaK_2F_6$.

DFT Density Functional Theory

MAE Mean Absolute Error

ML Machine Learning

LPTOS Lowest Possible Total Oxidation State