COMPUTING FORMATION ENTHALPIES THROUGH AN EXPLAINABLE MACHINE LEARNING METHOD: THE CASE OF LANTHANIDE ORTHOPHOSPHATES SOLID SOLUTIONS[*]

EDOARDO DI NAPOLI[†], XINZHE WU[‡], THOMAS BORNHAKE[§], AND PIOTR M. KOWALSKI[¶]

Abstract. In the last decade, the use of Machine and Deep Learning (MDL) methods in Condensed Matter physics has seen a steep increase in the number of problems tackled and methods employed. A number of distinct MDL approaches have been employed in many different topics; from prediction of materials properties to computation of Density Functional Theory potentials and inter-atomic force fields. In many cases the result is a surrogate model which returns promising predictions but is opaque on the inner mechanisms of its success. On the other hand, the typical practitioner looks for answers that are explainable and provide a clear insight on the mechanisms governing a physical phenomena. In this work, we describe a proposal to use a sophisticated combination of traditional Machine Learning methods to obtain an explainable model that outputs an explicit functional formulation for the material property of interest. We demonstrate the effectiveness of our methodology in deriving a new highly accurate expression for the enthalpy of formation of solid solutions of lanthanide orthophosphates.

Key words. Explainable learning, enthalpy, LASSO, ridge regression, sparsification, solid solutions.

MSC codes. 68T05, 62J07,

1. Introduction. In recent years, the role of existing Machine Learning (ML) methods have experienced a tremendous growth in many scientific computing domains including Materials Science and Quantum Chemistry [20, 28, 36]. Concurrently, new revolutionary methods and algorithms have appeared that expanded the range of applicability of existing state-of-the art techniques [17, 3]. This trend led to a high interest in general ML applications which, in turn, left the scientific community struggling in reconciling the need of developing refined tools with the assessment of their usefulness when applied to specific problems [40]. The assessment of their efficacy is particularly relevant when the target is an *explainable* learning method [29] and the domain knowledge is *integrated* in the final model [39].

In recent years several machine learning methods have been proposed to predict enthalpies of formation of several categories of materials [34, 6, 37]. Despite their progress, none of these efforts provides a fully explainable model which outputs a mathematical expression building on existing knowledge and improves it by adding to it additional terms that are statistically inferred.

In this article we propose a 3-step approach to use traditional machine learning tools to arrive at a scientifically consistent explainable method [29]. At first, we propose the use of *Kernel Ridge Regression* [11] methods to first assess which, out of a number of different kernels, provides the most reliable and transparent method. Second, we make a post-hoc interpretation of the model and we proceed to *reverse engineer* it so as to find which coefficients of the model are the most relevant to recover

a fully explainable mathematical expression for the target property. Finally, we integrate domain-specific knowledge by forcing scientific consistency through a constraint on how the input variables could be combined. The end result is a mathematical expression which relates the target property to the input variables in a functional dependence that replicates known results and add further terms substantially improving the accuracy of the final expression.

Our methodology is in part inspired by the work of Ghiringhelli et al. [9, 8] which uses the *Least Absolute Shrinkage and Selection Operator* (LASSO) [32] together with a sparsification process they term LASSO+ℓ_0 to learn the most relevant descriptors for a given target property. In this work, we go beyond their approach by constraining the functional form of the prior using knowledge coming from both the algorithmic model (the assessment of best kernel) and integration of domain-knowledge (to ensure scientific consistency). To demonstrate the feasibility of our approach we applied it to the specific problem of computing the excess enthalpy of formation of solid solution (enthalpy of mixing) of lanthanide orthophosphates (LnPO$_4$). We investigate the functional dependence of the mixing enthalpy for binary solid solutions of two distinct lanthanide cations (Ln), taking into account two distinct phases these materials form: monazite and xenotime [14].

1.1. Excess enthalpy of solid solution formation. Monazite is a phosphate mineral that contains rare-earth elements. Among these, lanthanide phosphates (LnPO$_4$) are the most widely distributed. These form monazite-type structure for $Ln = $ La, ..., Tb and xenotime-type (zircon) structure for heavier lanthanides [5, 24, 22, 33]. Among other potential applications, synthetic (monazite-type) ceramic solid matrices are suitable for the conditioning of long-lived radionuclides such as minor actinides (Np, Am, Cm) or Pu in a form of a synrock [7, 35, 19]. However, before these ceramics could be used in nuclear waste management, their physical and chemical properties and, most importantly, their thermodynamic parameters have to be well characterized and understood.

A solid solution is formed when two or more cations are incorporated into a solid host matrix on the same crystallographic site. When atoms of the solute solid phase are incorporated into the solvent solid host phase, the whole process can be interpreted as a sort of impurity inclusion into the host phase [10, 27]. Here, we consider a combination of two cations within a single phase, either monazite or xenotime. In reality, however, when lighter (monazite as stable structure) and heavier (xenotime as stable structure) lanthanide are mixed, such a solid solution has a wide miscibility gap, i.e. it is thermodynamically unstable in a wide solid solution range, with different stable phases of solid solution endmembers (solute and solvent). In these cases, the mixing enthalpy of single phases solid solutions is a key factor for describing the two-phase cases, such as monazite-xenotime system [21, 14].

Whether a single phase solution will stay stable or not, the result is driven by the excess enthalpy of the mixing [10, 18]. The latter is defined as the difference between the formation enthalpies of the mixed compounds and those of the solid solution endmembers, which could be measured [38] or accurately computed [2, 31, 1]. The single phase solid solutions such as monazite-type, resemble closely a symmetric solid solutions and are well described by a simple model, $H^E = m(1 - m)W$, with W being the Margules interaction parameter and m the solid solution ratio [26, 18]. With systematic DFT-based calculations, it was found [16] that for the monazite-type solid solutions, the Margules interaction parameter W is a function of the Young's modulus, Y, the unit-cell volume, V, and the unit-cell volume difference between the

solid solution endmembers (solid and solute), ΔV, namely [13]

$$(1.1) \qquad W \sim \frac{0.154Y}{V}\Delta V^2 \sim \frac{Y}{6V}\Delta V^2.$$

The relationship between the excess formation of mixing and the physical parameters has been a topic of discussion of various studies [21, 18, 16]. Among these, the ionic radius of the mixing cations is often used as the main discriminant parameter [21]. Such a choice, however, makes the thermodynamic parameters of the mixing only weakly material dependent. As such, the excess enthalpy of mixing of monazite- and xenotime-type solid solutions are described with very similar relationship as a function of $\Delta R/R$ (Fig. 6 of [21]).

In Figure 1 we illustrate how existing models describe the functional dependence on physical parameters of the excess of enthalpy for the data used in this work. Plot (a) shows the case of monazite for which the models of [16, 4] reproduce the data reasonably well. This is in part because both models use the difference in volumes of the endmembers as a parametric variable, while the model of [21] uses the difference in ionic radii.

The situation is diametrically different in the case of xenotime-type solid solutions (plot (b) of Fig. 1). Here, the models of [4] and [16] give predictions that are inconsistent with the *ab initio* data by a factor of ~ 2. This points towards the possibility of another, unaccounted term in the Margules parameter that could be quite relevant in the case of xenotime-type solid solutions but of minor importance for monazite-type solid solution.
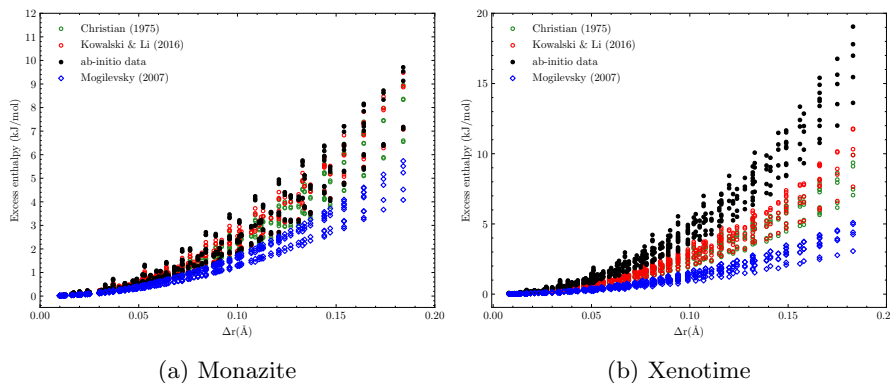

(a) Monazite   (b) Xenotime

Fig. 1: The excess enthalpy of mixing for monazite- and xenotime-type solid solution computed *ab initio* and from models of [4], [16] and [21]

A combination of *ab initio* and calorimetric studies, [23] has shown that the *ab initio* data themselves are not enough to constrain the values of the Margules parameter W, and that understanding of the dependence of W on the selected physical parameters is crucial for precise modeling of the stability of solid solutions. As such, the study of the excess of enthalpy for this type of solid solutions lends itself perfectly to test the explainable machine learning methodology we have devised.

2. Methodology and learning algorithms. In predicting materials' properties one needs a set of curated training data organized in *input* variables $\boldsymbol{x}$ and *target*

properties y. The set of input variables has to represented in terms of a number of *independent* descriptors that are invariant under known mathematical and physical symmetries and usually requires a considerable amount of domain expertise. In this work, the input variables are represented by properties of the elements constituting a solid solution (e.g. electron orbitals, nuclear charge, etc.) and the target property by the solution excess enthalpy of formation.

2.1. Elemental properties and descriptors. As mentioned in other works [9, 37, 34] and based on the physics of the solution process, properties associated to the electron orbitals and the nuclei are expected to carry most of the weight in determining the value of the enthalpy of formation. Moreover, all solid solutions that are part of our data set have in common the same phosphate group (PO_4). Consequently, properties of the atomic elements of such group are not taken into consideration. Based on the online database http://www.knowledgedoor.com/, we list in Table 2 a meaningful list of elemental properties (*elementals*) ϵ_k that are available for each and every lanthanide element.

Name	Symbol	Unit
Atomic number	Z	-
Atomic mass	m	[u]
Ionic radius coordination 8, 9	R^8, R^9	[Å]
Ionization potential +2, +3	IP^{2+}, IP^{3+}	[eV]
Pauli electronegativity	χ	-
Young's modulus of LnPO$_4$	Y	[GPa]
Effective nuclear charge	Z_{eff}	-
Macroscopic density	ρ	[g/cm^3]
Volume of LnPO$_4$	V	[Å^3]

Table 1: List of elemental properties and their physical units

These elementals can be arranged in an abstract vector $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \dots)^\mathsf{T} = (Z, m, R^8, R^9, IP^{2+}, IP^{3+}, \dots)^\mathsf{T}$. For each lanthanide $L_i \in (\text{La}, \text{Ce}, \text{Pr} \dots)$ there is one such vector. We build descriptors out of elementals. Since we are investigating solid solutions made of two lanthanides, our descriptors $x_k[\epsilon_k(L_i), \epsilon_k(L_j), m_i, m_j]$ are functions of elementals from two different lanthanides together with their mixing ratio m_i. The inclusion of m_is is necessary to distinguish between different solution ratios. However the descriptor is defined, it should be invariant to simultaneous permutation of lanthanide and mixture ratios. Noting that $m_i + m_j = 1$, we can actually use only one mixing ratio $m = m_i$ so that $m_j = 1 - m$.

Invariance under permutation can be expressed as

$$x_k[\epsilon_k(L_i), \epsilon_k(L_j), m] = x_k[\epsilon_k(L_j), \epsilon_k(L_i), 1 - m]$$

It is important that a descriptor x changes significantly with the mixing ratio. Additionally, we should include descriptors capturing certain processes were the enthalpy is strongly dependent on which lanthanide has the largest abundance. Last but not least, descriptors has to be homogeneous functions of elementals and cannot mix elementals with different physical units (unless conversion constants are involved). For the reasons above we selected three type of descriptors $x^{(1)}$, $x^{(2)}$ and $x^{(3)}$, listed in Table 2, for every elemental ϵ_k and every lanthanide pair (L_i, L_j).

Table 2: Descriptor types depending on lanthanide pairs (L_i, L_j), elemental ϵ_k, and mixing ratio m.

Name	Descriptor
Weighted mean	$x_k^{(1)} = m\epsilon_k(L_i) + (1-m)\epsilon_k(L_j)$
Quadratically weighted mean	$x_k^{(2)} = \frac{m^2\epsilon_k(L_i) + (1-m)^2\epsilon_k(L_j)}{m^2 + (1-m)^2}$
Absolute difference	$x_k^{(3)} = |\epsilon_k(L_i) - \epsilon_k(L_j)|$

Notice that the quadratically weighted mean is not quadratic in the actual values of the elementals ϵ_k but quadratic in the mixture ratio m. That means the descriptor $x^{(2)}$ will lean heavily to the value of the elementals of the more abundant lanthanide. For each combination of lanthanides pairs the ten elementals ϵ_k are organized in a descriptor vector $\boldsymbol{x}$ made of thirty descriptors in total.

$$
\boldsymbol{x}(L_i, L_j, m) = \begin{bmatrix} x_1^{(1)}[\epsilon_1(L_i), \epsilon_1(L_j), m] \\ x_1^{(2)}[\epsilon_1(L_i), \epsilon_1(L_j), m] \\ x_1^{(3)}[\epsilon_1(L_i), \epsilon_1(L_j), m] \\ x_2^{(1)}[\epsilon_2(L_i), \epsilon_2(L_j), m] \\ \vdots \end{bmatrix}
$$

Each vector $\boldsymbol{x} \in \mathcal{X}$ of size $d = 30$ makes up the input variables for the learning algorithm. The target value y is the excess enthalpy of formation H^E. For each choice of lanthanide pairs (L_i, L_j) and choice of mixing ratio m we have a data point $(\boldsymbol{x}, y)$. All data points together constitute a set holding N data points. We will see at the end of this section how this set of points is generated.

2.2. Learning algorithms. Since the data points in our set have both input values and target value, we use a common type of *supervised* learning algorithm: *kernel ridge regression* (KRR) [11]. Kernel ridge regression is a non-linear regression algorithm with a regularization term (from which the name "ridge") that is comparable to the well-known Support Vector Machine algorithm.

The simple linear regression algorithm aims at finding the unknown coefficients $\boldsymbol{\beta}$ of the function $f(\boldsymbol{z}) = \langle \boldsymbol{\beta}, \boldsymbol{z} \rangle$ minimizing the error $E[f(\boldsymbol{z}) - y]$ (also known as *loss function*) over the entire set of data. In order to alleviate over fitting of the data, a regularization term is usually added. In the *ridge regression*, the regularization amounts to adding a penalty term to the minimization problem. Choosing the squared error as loss function leads to the following minimization problem

$$
(2.1) \qquad \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^d} \sum_i (\langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle - y_i)^2 + \lambda \|\boldsymbol{\beta}\|_2^2
$$

By introducing a function $\boldsymbol{z} \rightarrow \phi(\boldsymbol{z})$ which maps the input space $\mathcal{X}$ to a feature space $\mathcal{H}$, the use of kernels generalizes the linear to non-linear regression (see [30] for a didactic introduction). In this context a kernel is an inner product in feature space $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$. The advantage of using kernels is that the function $f(\boldsymbol{z}) = \sum_i \alpha_i k(\boldsymbol{x}_i, \boldsymbol{z})$ is not expressed anymore as a sum over dimension of the input space d, but instead as a sum over the number of data N making up the training set.

With this set up, the minimization problem to be solved becomes

$$\arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^N} \sum_i^N (f(\boldsymbol{x}_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2.$$

In practice, the kernel function is expressed as a matrix of inner products between points of the training data set in feature space $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = K_{ij}$. Eventually, the solution of the minimization problem can be expressed by the linear equation

$$(2.2) \qquad\qquad \boldsymbol{\alpha} = (\boldsymbol{K} + \lambda \mathbf{1}_N)^{-1} \boldsymbol{y},$$

with $\boldsymbol{\alpha} \in \mathbb{R}^N$ being the vector that contains the information learned.

Almost all Machine Learning methods do not work directly out of the box but have a number of parameters that have to be fixed. In the case of the KRR, the level of regularization through the parameter λ is tuned for the dataset at hand and the selected kernel. Additionally, almost every kernel has some extra parameters that must also be tuned. The entire set of these adjustable parameters are called *hyperparameters.*

Given a kernel, its computation can still be performed in input space despite its value describes the inner product in feature space. In this work, we employed three different kernels with the same set of data: the polynomial, the Gaussian, and the Laplacian kernels, which are respectively based on the inner product, the ℓ_2-norm and the ℓ_1-norm

$$(2.3) \qquad\qquad k(\boldsymbol{x}, \boldsymbol{z}) = (\langle \boldsymbol{x}, \boldsymbol{z} \rangle + c)^p$$

$$(2.4) \qquad\qquad k(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{z}\|_2^2}{2\sigma^2}\right)$$

$$(2.5) \qquad\qquad k(\boldsymbol{x}, \boldsymbol{z}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{z}\|_1}{\sigma}\right).$$

The actual computation of $\boldsymbol{\alpha}$ amounts to solve a positive-definite numerical linear system $A\boldsymbol{\alpha} = \boldsymbol{y}$. Once $\boldsymbol{\alpha}$ is computed, it is used to compute predictions for any new data point with $\hat{y} = \sum_i \alpha_i k(\boldsymbol{x}_i, \hat{\boldsymbol{x}})$. For validation purpose, the results of such prediction are typically presented in a *scatter plot* where the predicted and computed target values are represented on the x and y, respectively (see plots of Fig. 2, for instance).

The Laplacian and Gaussian kernels are by far the most used in KRR because they provide the most effective map since they use a virtually limitless number of functions as the prior of our statistical model. The down side is that we cannot recover the explicit functional form which express the target value in terms of input variables. This is why we have also included polynomial kernels, since they could be virtually inverted and return a functional expression for the coefficients $\boldsymbol{\alpha}$ in terms of descriptors. Being able to statistically infer such functional expression would allow us to go beyond the prediction of target values for new solid solutions and understand which descriptors are more relevant and contribute the most to determine the target values.

In this sense, the polynomial kernels that will be closest in error to the Gaussian or Laplacian kernels will provide a hint on the order of polynomial functions that should be included in our prior. With this information we can manually construct

thousands of candidate functions of the original descriptors $\boldsymbol{x}_i^{(t)}$ that could faithfully represent the underlying surrogate space. In our application, we denote these candidate functions as $\boldsymbol{v} = [\boldsymbol{v_1}; \boldsymbol{v_2}; \cdots; \boldsymbol{v_M}]$, where M can range from $\mathcal{O}(10^3)$ to $\mathcal{O}(10^4)$. We then apply a *sparsification* technique, which amounts to find the most relevant among $\boldsymbol{v_k}$ by forcing as many coefficients of the statistical model to be zeros. In Section 3.1, justified by the KRR results, we show how each distinct $\boldsymbol{v_k}$ is constructed from a low degree p polynomial function of the original descriptor.

The objective of sparsification, is finding the most relevant term(s) among $\boldsymbol{v_k}$ which contributes the most to the target values. Moreover, the number of the relevant terms should be also controllable. A straightforward sparsification technique that one can employ is the LASSO [32] approach combined with an ℓ_0 regularization. This combination is able to sparsify the coefficients of LASSO regression into a small determined number. The minimization problem to be solved is given as follows

$$(2.6) \qquad \arg \min_{\boldsymbol{\gamma} \in \mathbb{R}^M} \sum_i^N (\langle \boldsymbol{\gamma}, \boldsymbol{v}_i \rangle - y_i)^2 + \hat{\lambda} \|\boldsymbol{\gamma}\|_0.$$

In this formula, the ℓ_0-norm of a coefficient vector $\boldsymbol{\gamma}$ is defined as

$$(2.7) \qquad \|\boldsymbol{\gamma}\|_0 = \#\{j : \gamma_j \neq 0\}, \quad j = 1, \ldots, M$$

indicating the number of non-zero elements of the $\boldsymbol{\gamma}$ vector. A vector $\boldsymbol{\gamma} \in \mathbb{R}^M$, it is called κ-sparse if $\|\boldsymbol{\gamma}\|_0 \leq \kappa$.

While the LASSO$+\ell_0$ is the exact problem we want to solve, it has a significant drawback: this minimization problem is not convex. This leads to a "NP-hard problem" [1] which is infeasible when M is large. Therefore, LASSO$+\ell_0$ cannot be directly applied to sparsify the candidate functions $\boldsymbol{v}$. In order to compromise between the convexity and sparsity of the coefficient vector, we first utilize a LASSO with a Manhattan ℓ_1-norm regularization $\lambda \|\boldsymbol{\gamma}\|_1$ (LASSO$+\ell_1$ in this paper) to carry out an initial *feature selection* out of which we can achieve the sparsification of $\boldsymbol{v}$ [8].

$$(2.8) \qquad \arg \min_{\boldsymbol{\gamma} \in \mathbb{R}^M} \sum_i^N (\langle \boldsymbol{\gamma}, \boldsymbol{v}_i \rangle - y_i)^2 + \hat{\lambda} \|\boldsymbol{\gamma}\|_1.$$

This latter optimization problem is convex, which promote also sparsity. In equation (2.8), $\hat{\lambda} > 0$ is the penalty parameter which is able to control the sparsity of the coefficient vector $\boldsymbol{v}$: the larger $\hat{\lambda}$ is, the smaller the ℓ_1-norm $\|\boldsymbol{\gamma}\|_1$ would be, hence higher sparsity is achieved and more candidate functions are eliminated.

In our application, we start from a very small $\hat{\lambda}$, then increase it with a fixed step. With the increase of $\hat{\lambda}$, we observed that a large number of constructed functions in $\boldsymbol{v}$ get quickly eliminated to rapidly flatten the curve which stagnates even for much larger λ values; in other words the minimization process with the ℓ_1-norm reduces the vector $\boldsymbol{\gamma}$ to be at most κ-sparse, κ being usually smaller than 20 in our application. In practice, we compressed a very large feature space spanned by $\boldsymbol{v}$ into a tiny one spanned by $\boldsymbol{v}'$, in which $\boldsymbol{v}'$ is a subset of $\boldsymbol{v}$ whose number of elements is smaller than κ. Thanks to the low dimension of feature space selected by LASSO$+\ell_1$, the Equation

---

[1] A NP-hard problem means that the solution can not be found in polynomial time with respect to M.

(2.6) of LASSO+ℓ_0 can be solved rather effectively with a brute force approach. More details of numerical results are given in Section 3.

This method, combining LASSO with a sparsification of the coefficients vector, has been developed in the context of *compressed sensing* [8], where one is interested in reproducing the gathered data with as few as possible degrees of freedom. Unlike in the [8], our work starts with a Kernel Ridge Regression and maps the descriptors $\boldsymbol{x}_i$ to a much larger but finite dimensional input space where the new features $\boldsymbol{v}_i$ are made out of a finite number of selected functions of the original descriptors $\boldsymbol{x}_i$. We additionally integrate domain-knowledge by modifying the initial selection of descriptors and features to ensure scientific consistency with the existing knowledge from the physics of solid solutions. The necessity of tuning of the descriptors and the features is also confirmed by the inefficiency of the LASSO+ℓ_1 method when the original descriptors $\boldsymbol{x}_i$ are used.

In the next section, we see that KRR polynomial kernel with degree 3 returns one of the lowest statistical error. We capitalize on such a result and make the plausible hypothesis that a polynomial map of degree at most 3 for LASSO should identify the most promising functions of the descriptors $x_k^{(t)}$. We then show that the results of *feature selection* by LASSO+ℓ_1 suggests the selection of a modified set of descriptors that guarantees scientific consistency. Finally we determine the desired functions which map the elemental descriptors to the excess of enthalpy for both monazite and xenotime.

2.3. Data set generation. The data set used by the selected learning methods was computed with the Quantum-ESPRESSO code using the approach of [18, 16]. The solid solutions were modeled with special quasi-random structures constructed using procedure of [41]. All the structures were modeled with supercells containing 192 atoms (32 formula units). We applied the PBEsol exchange-correlation functional [25] and the f electrons were included into pseudopotential core. It has been shown that this setup results in a very good predictions of structures and thermodynamic parameters of lanthanide phosphates, including formation enthalpies [2, 12, 15, 1]. The excess enthalpies of mixing and Margules interaction parameters were derived from differences in the total energies of the mixed cation structure and weighted (according to the solid solution composition) sum of the end members.

The dataset consists of excess enthalpies of formation between all 15 lanthanides, which leads to 105 possible combinations ($\sum_{i=1}^{14} i = 105$). Those 105 combinations were then modeled for five different mixture ratios $m = 0.25, 0.375, 0.50, 0.625, 0.75$ giving a total of 525 data points. Two distinct data sets were generated for the two lanthanide orthophosphate phases, monazite and xenotime, which correspond to the two possible coordination numbers of the lanthanides (see Table 1). In the following, we will test our models on three distinct configurations of these data: monazite only (525 data points), xenotime only (525 data points), and fused (1050 data points). Not all these points are used to train the learning model. A subset is reserved for testing and validation purposes.

3. Regression, sparsification and interpretable formulas. The first step before determining the coefficients $\boldsymbol{\alpha}$ of our KRR models is to determine the optimal value of the corresponding hyperparameters. To this aim, the entire given dataset is typically split in two parts. The *training* dataset $(\boldsymbol{x}_i, y_i) \in \mathcal{T}$ is used to compute the actual coefficients $\boldsymbol{\alpha}$ of the regression that appear in Eq. (2.2). The *testing* or *prediction* dataset $(\boldsymbol{x}_i', y_i') \in \mathcal{P}$ is kept separate and is used to evaluate the quality of a given set of hyperparameters values for predicting unseen data $\hat{\boldsymbol{y}} = \boldsymbol{\alpha}^\mathsf{T} \boldsymbol{K}'$, where

$K'_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}'_j)$. The optimal set of parameters is selected when the predictions (i.e. $\hat{y}_i$) are in the best possible agreement with the known values from the testing dataset (i.e. y'_i). Different error functions are commonly used to quantifying the quality of a prediction. The most common that we are also using in this work are Mean Squared Error (MSE), Mean Absolute Error (MAE), and Maximum Error (ME).

Because possible hyperparameters values may span over multiple dozens of orders of magnitude, brute force methods that quickly scan the entire space in a grid-like pattern are preferred to conventional minimization methods. Once an approximate local minimum is found, local optimizations are used to refine the values of the hyperparameters. In the following we do not report of the hyperparameter search results which is a standard procedure in the use of ML algorithms. In all the scatter plots and tables it is implicitly intended that all hyperparameters have been opportunely optimized following the procedure just described.
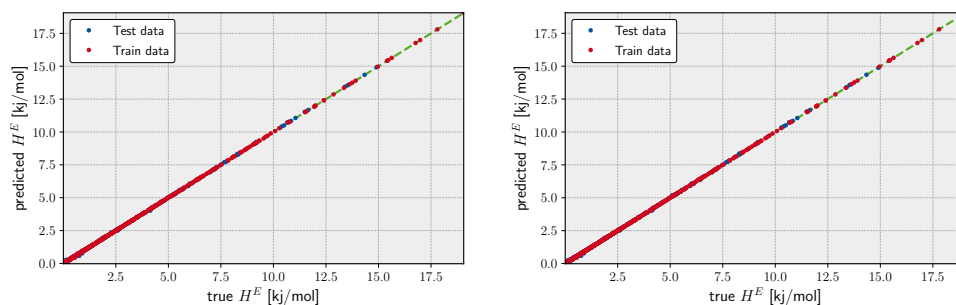

Fig. 2: Regression for the Gaussian Kernel (left) and the degree 3 polynomial Kernel (right). The scatter plots corresponds to a trained model whose hyperparameters have been already fitted. Gaussian best fit: $\lambda = 10^{-13.46}$, $\gamma = 10^{-5.90}$ with an MAE of 0.0107 kJ/mol. Degree 3 polynomial the best fit: $\lambda = 10^{4.74}$, $\gamma = 10^{5.79}$, $c = 10^{0.37}$ with an MAE of 0.0090 kJ/mol.

3.1. Predicting excess of enthalpies with KRR. In the search for optimal hyperparameters, we have split the total subset of data in two parts between $\mathcal{T}$ and $\mathcal{P}$. The ratio of the data size between $\mathcal{T}$ and $\mathcal{P}$ is $4:1$. Once the hyperparameters search is completed we fitted the data of the training set $\mathcal{T}$ using all three distinct kernels for the KRR model.

In Table 3 we report the results obtained for all kernels with all three type of errors (MSA, MAE, ME) for both sets $\mathcal{T}$ and $\mathcal{P}$ on the *fused* data set configuration (similar results are obtained for the other two configurations). Despite its remarkable low errors on the $\mathcal{T}$ set, the Laplacian kernel does not return a satisfactory result on the set of prediction data $\mathcal{P}$. In fact, no choice of hyperparameters returns a reliable regression for unseen data: the search space minimization return a value for λ numerically indistinguishable from zero. This is a typical sign that KRR with this kernel is overfitting the data and returning an in-sample error much smaller of the out-of-sample error. For this reason we discarded the learning model using the Laplacian kernel.

The low order polynomial and the Gauss kernels return a much nicer picture in terms of the errors. Already the degree 2 polynomial kernel is able to fit the data quite well. Its MAE differs only ≈ 0.01 kJ/mol from the error of the degree 3 kernel

Kernel	$\log \lambda$	$\log \gamma$	$\log c$		MAE	MSE	ME
Poly. 2D	-3.68	3.95	-0.447	$\mathcal{T}$	0.0098	0.0002	0.0917
				$\mathcal{P}$	0.0278	0.0025	0.4714
Poly. 3D	4.73	5.79	0.368	$\mathcal{T}$	0.0090	0.0001	0.1041
				$\mathcal{P}$	0.0228	0.0011	0.1865
Gaussian	-13.4	-5.90	-	$\mathcal{T}$	0.0107	0.0002	0.1140
				$\mathcal{P}$	0.0235	0.0011	0.1774
Laplacian	-20.0	-2.31	-	$\mathcal{T}$	7.6×10^{-14}	9.2×10^{-27}	3.1×10^{-13}
				$\mathcal{P}$	0.1102	0.0272	0.6428

Table 3: Kernel Ridge Regression results on the Excess Formation Enthalpy dataset for different Kernel types. The first two rows are polynomial kernels of degree 2 or 3 respectively. The errors displayed are **Mean Squared Error** (MSE), **Mean Absolute Error** (MAE) and **Maximum Error** (ME). $\mathcal{T}$=Training dataset, $\mathcal{P}$=Prediction dataset. Units are $[\text{kJ}^2/\text{mol}^2]$ (MSE) or $[\text{kJ/mol}]$ (MAE & ME). Note that the Laplacian kernel has by far the lowest regularization strength ($\lambda = 10^{-20}$) which leads to a perfect fit on the training data but by also to the worst performance on the testing dataset. This is a strong sign of overfitting.

and the Gaussian Kernel. Judging from the fact that the degree 3 polynomial kernel returns a MAE as low as $\approx 0.02\,\text{kJ/mol}$ for the $\mathcal{P}$ set indicates that the underlying function for the excess of Enthalpy could be represented by functions of the descriptors having up to cubic terms. Since the actual values for the excess enthalpy of formation for the lanthanide orthophosphates span a range going from of 0.5 to $10\,\text{kJ/mol}$, the relative errors of our model represent the same level of uncertainty returned by the DFT simulations. In other words, the prediction provided by the KRR model with either a degree 3 polynomial or the Gauss kernel are indistinguishable from the finite accuracy of the in silico simulation used to generate the data used in both sets $\mathcal{T}$ and $\mathcal{P}$ (see Fig.2).

What distinguishes the Gaussian and the degree 3 polynomial kernels are the value of the hyperparameters: the degree 3 kernel requires a quite large value of λ and γ respect to the Gaussian. This is not necessarily a negative result but it points out that our choice of descriptors may not be ideal when the kernel represents a finite set of prior functions like in the case of the degree 3 polynomial. When the set of prior functions becomes virtually infinite (the Gaussian kernel can be seen as an infinite series of polynomials), the descriptor choice becomes unimportant. We will see in the following subsection how the choice of descriptors becomes significant when one would like to sparsify the vector of coefficients $\boldsymbol{\alpha}$ starting from a finite set of prior functions of the descriptors.

To ensure that our KRR models provide a good fit for all data independently on how they are split between $\mathcal{T}$ and $\mathcal{P}$, we have used *cross validation*. In practice, we run the KRR models fitting several different choices of training and testing datasets always keeping the same choices of values for the hyperparameters that were previously selected. If the results with the original dataset were truly a product of chance, a fit with an entirely new combination of data points should show a different performance. Table 4 shows the results for repeating the KRR with polynomial kernel of degree 3 for five different subsets, some of which even showed slightly better performance than the original regression. The MAE over the five different sets ranges between

| Regressions ID	Set type	MAE	MSE	ME
Original	$\mathcal{T}$	0.0090	0.0001	0.1041
	$\mathcal{P}$	0.0228	0.0011	0.1865
1	$\mathcal{T}$	0.0126	0.0004	0.1705
	$\mathcal{P}$	0.0120	0.0004	0.1865
2	$\mathcal{T}$	0.0130	0.0004	0.1864
	$\mathcal{P}$	0.0108	0.0002	0.0832
3	$\mathcal{T}$	0.0123	0.0004	0.0171
	$\mathcal{P}$	0.0128	0.0005	0.1865
4	$\mathcal{T}$	0.0127	0.0004	0.1865
	$\mathcal{P}$	0.0126	0.0004	0.1108
5	$\mathcal{T}$	0.0120	0.0004	0.1865
	$\mathcal{P}$	0.0139	0.0005	0.1705

Table 4: Errors for cross-validation of polynomial kernel of degree 3 with five different combinations of data points evenly split between $\mathcal{T}$ and $\mathcal{P}$. For each new regression, we used the same hyperparameters λ, γ and c as in the original one.

$0.052 - 0.082$ kJ/mol, confirming the quality of the original regression.

3.2. LASSO + ℓ_0. In the previous section, we have concluded that KRR with Gaussian and degree 3 polynomial kernel performs very similarly with a slight advantage for the latter kernel. In this section we want to pursue the road of finding a surrogate model that is explainable: we aim at formulation of the excess enthalpy of mixing that can provide the domain scientist with a understandable function of a small number of descriptors. We are driven to recover this result by the observation that already polynomial functions of degree 3 provide enough prior to get an accurate KRR model. We achieve this result through a so-called *sparsification* process.

3.2.1. Sparsification with Descriptors in KRR. A large number of candidate functions have been built from the polynomial kernel of degree 3 based on the 27 elemental descriptors introduced in Section 2.1. Denoting the group of 27 elemental descriptors as $\mathcal{D}_1$, a group of candidate functions with polynomial of degree 2, denoted as $\mathcal{D}_2$, can be defined as a commutative element-wise vector product $d_i d_j$, with $d_i, d_j \in \mathcal{D}_1$. The number of descriptors in $\mathcal{D}_2$ is 378. The group of candidate functions with polynomial of degree 3, denoted as $\mathcal{D}_3$, can be defined in a similar way, as a commutative element-wise product of three vectors $d_i d_j d_k$ with d_i, d_j, $d_k \in \mathcal{D}_1$. The number of descriptors in $\mathcal{D}_3$ is 3,303. Therefore, all the candidate functions are a union of $\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{D}_3$, which is denoted as $\mathcal{D}$. $\mathcal{D}$ is a dense matrix of size $1{,}050 \times 3{,}708$ for the fused data set, and of size $525 \times 3{,}708$ for both monazite and xenotime data set configurations.

As described in Section 2.2, a *feature selection* step has been performed on $\mathcal{D}$ through the LASSO+ℓ_1 method. Increasing the penalty parameter λ, the size of feature space is reduced as more candidate functions are removed. In the numerical experiments, λ is increased from 0.001 to 0.096 in incremental steps of 0.005. We performed the same feature selection step not only for the fused data, but also separately

(a) fused

(b) monazite
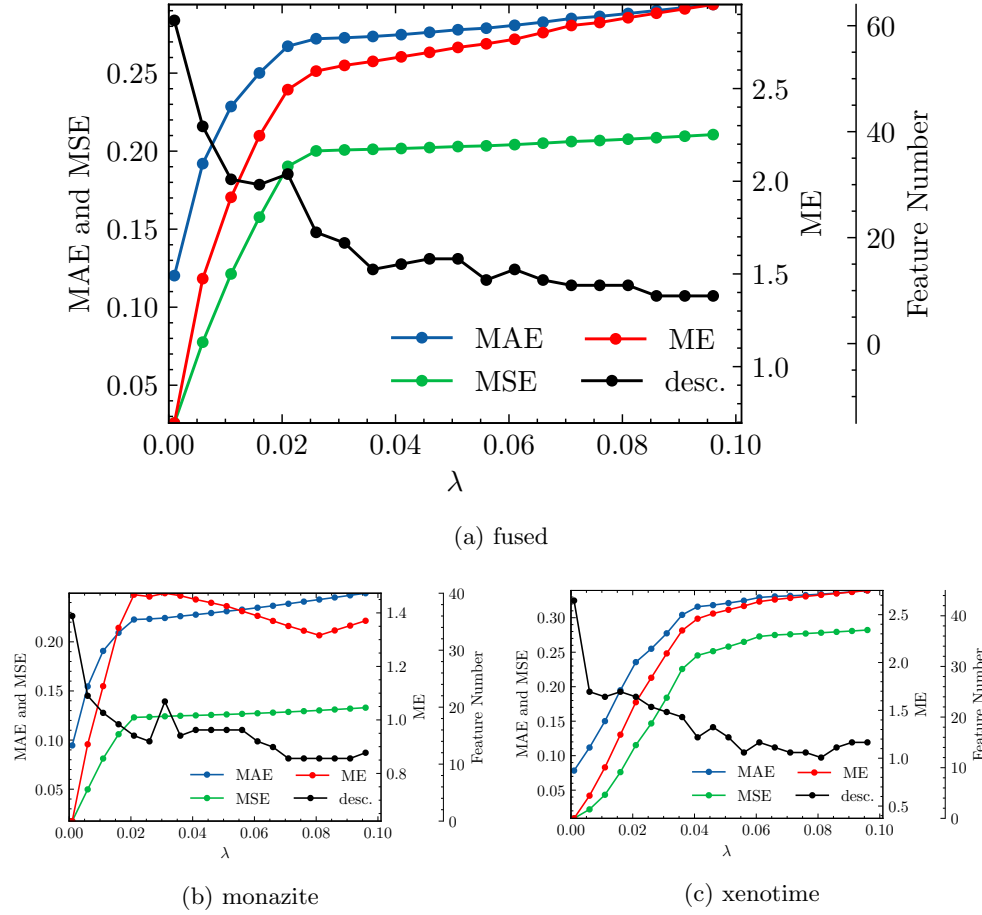
(c) xenotime

Fig. 3: LASSO+ℓ_1 results with descriptors used in KRR

for monazite and xenotime data configurations. The results are plotted in Fig. 3, which shows the changes of the errors (MAE, MSE, ME) and the size of the reduced feature space (marked as desc.).

From Figure Fig. 3 we infer that with the increase of λ, the feature space size can be quickly reduced to less than 30. This is good news, since sparsifying a feature space of size ~ 30 through LASSO+ℓ_0 into a determined size smaller than 5 is still feasible [2]. At the same time, we observe that the all the errors increase with the increase of λ. This behavior is to be expected and is the direct consequence of the reduction of the feature space size. Unfortunately, the errors increase too quickly and show an exponential growth that plateaus already for small values of λ. For example, for *fused* data, the MAE increases from ~ 0.03 to more than 0.25 when λ is only increased from 0.001 to 0.021. All the while the MSE and ME increase to 0.20 and 2.5, respectively. When increasing λ from 0.021 to 0.096, the feature space size continues to be reduced, however, the changes of error flatten. Similar trends can also be observed separately

---

[2]This is an NP-hard problem which can still be solved by brute-force for such small size
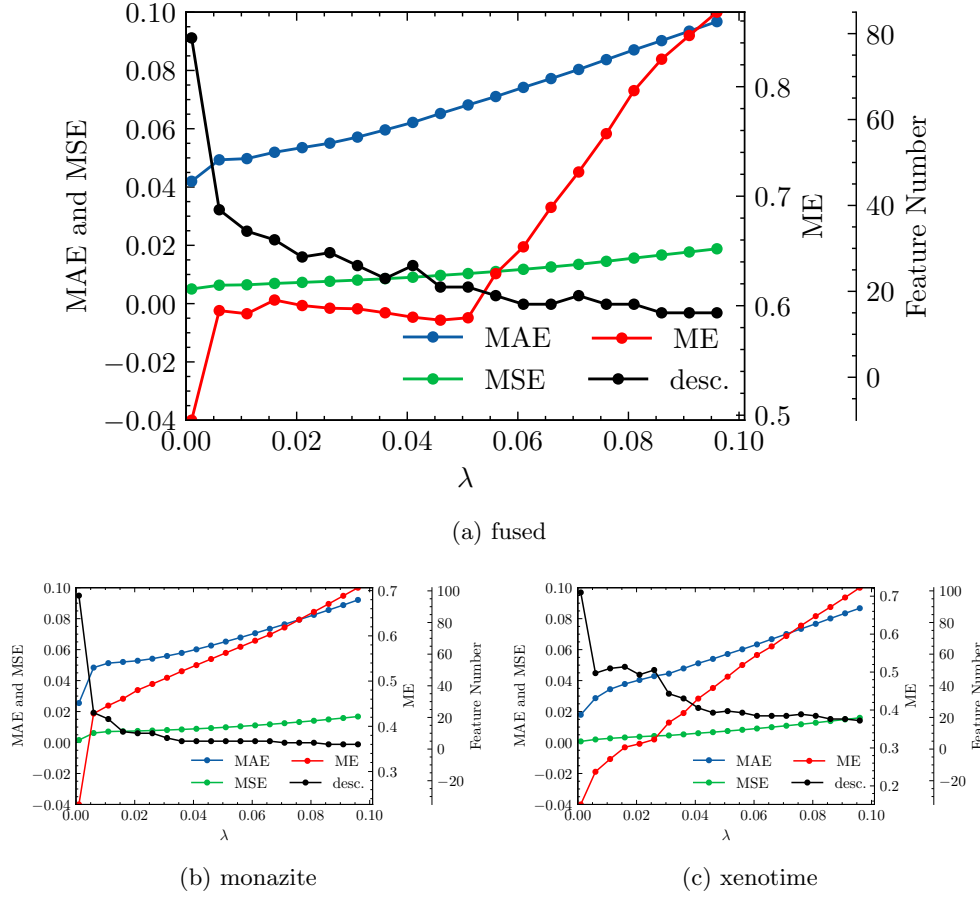
(a) fused

(b) monazite

(c) xenotime

Fig. 4: LASSO+ℓ_1 results with descriptors defined only by arithmetic means and absolute difference.

for the monazite and xenotime data. This result signals that the generated feature space doesn't capture well the functional dependence of the excess enthalpy on the descriptors. In other words, we cannot find a simple functional dependence of H^E and need many hundreds of functions to faithfully predict the enthalpy.

In order to confirm our concerns, we perform a LASSO+ℓ_0 step on the reduced feature space separately for the fused, monazite, and xenotime data set configurations. At most three leading terms have been selected, which are listed in Table 5, along with their corresponding MAEs, MSEs, and MEs. As expected, the errors of candidate functions selected by LASSO+ℓ_0 are more than one of order of magnitude larger than the ones derived by KRR. What is more remarkable is the fact that the errors don't seem to decrease much as we consider more terms. In fact, for the fused data configuration, the MAE and ME actually increase going from 2 to 3 leading terms. The errors for the other data configuration seems a bit better but are still far from what we would like to observe. From these tables and the plots in Figure 3, we conclude that the choice of prior candidate functions cannot be effectively exploited

nb.	Leading terms	MAE	MSE	ME
1	$(x_V^{(3)})^2 x_V^{(1)}$	0.3528	0.4400	3.6657
2	$(x_V^{(3)})^2,\ x_V^{(3)} x_Y^{(3)} x_Y^{(1)}$	0.2730	0.2053	2.4429
3	$(x_V^{(3)})^2,\ x_V^{(3)} x_V^{(2)} x_Y^{(2)},\ x_V^{(3)} x_Y^{(3)} x_Y^{(1)}$	0.2757	0.2017	2.5846

(a) fused

nb.	Leading terms	MAE	MSE	ME
1	$(x_V^{(3)})^2 x_R^{(1)}$	0.2274	0.1302	1.8870
2	$(x_V^{(3)})^2 x_V^{(1)},\ (x_V^{(3)})^2 x_R^{(1)}$	0.2201	0.1237	1.4692
3	$x_{Z_{eff}}^{(2)},\ x_R^{(2)} x_Z^{(2)},\ (x_V^{(3)})^2 x_\chi^{(1)}$	0.2076	0.0799	1.1821

(b) monazite

nb.	Leading terms	MAE	MSE	ME
1	$(x_V^{(3)})^2 x_Z^{(1)}$	0.3303	0.2865	2.3072
2	$(x_{Z_{eff}}^{(3)})^2 x_\chi^{(1)},\ (x_V^{(3)})^2 x_Z^{(1)}$	0.3247	0.2784	2.4090
3	$(x_Y^{(2)})^3,\ x_R^{(2)} x_Z^{(2)},\ (x_V^{(3)})^2 x_Z^{(1)}$	0.2078	0.0698	0.9729

(c) xenotime

Table 5: LASSO+ℓ_0 results for Fused, monazite and xenotime with descriptors used in KRR

by the sparsification process.

3.2.2. Sparsification with Descriptors built with Prior Knowledge. In order to overcome the shortcomings of the sparsification process seemingly caused by the choice of descriptors, we simplify the form of the descriptors and exploit the existing knowledge from the application domain. Based on the insight provided by the simple model with the Margules interaction parameter W and the expression in Eq. (1.1), we make three additional hypothesis: i) the polynomial degree of m and $1 - m$ may not be in accord with each other and the polynomial degree of the elemental property ϵ_k, ii) negative power of the descriptors may appear in the Margules parameter and, iii) the volume V, coordination number R , and mixing ratio m play a special role than the other elemental properties and may contribute with monomials of degree higher than 3.

The direct consequence of i) is that m and $1 - m$ have been decoupled from the elemental descriptors and included as descriptors on their own. This may seem a strange choice since they do not depend on the lanthanides but are the same for all. On the other hand, decoupling the mixing ratio allows more freedom in the way it appears in the degree 3 polynomial functions that are part of $\mathcal{D}$. The indirect consequence is that we do not have anymore three types of descriptors for each elemental property ϵ_k but only two: we drop the weighted quadratic mean, convert the weighted mean

nb.	Leading terms	MAE	MSE	ME
1	$\frac{(\Delta V \overline{V})^2}{\overline{R}^3}$	0.2798	0.2268	3.6657
2	$\frac{\Delta Y \Delta Z}{\overline{Z}_{eff}}$, $m^2(1-m)^2(\Delta V)^2$	0.0950	0.0254	1.5519
3	$\frac{\Delta Y \Delta Z}{\overline{Z}_{eff}}$, $m^2(1-m)^2(\Delta V)^2$, $(1-m)\Delta Y(\Delta V)^2$	0.0878	0.0164	0.7348

(a) fused

nb.	Leading terms	MAE	MSE	ME
1	$m(1-m)(\Delta V)^2$	0.0622	0.0095	0.4936
2	$m(1-m)(\Delta V)^2$, $m^2(1-m)^2(\Delta V)^2$	0.0575	0.0082	0.6309
3	$m(1-m)(\Delta V)^2$, $m(1-m)^2(\Delta V)^2$, $(\frac{\Delta V}{m})^3$	0.0516	0.0064	0.6195

(b) monazite

nb.	Leading terms	MAE	MSE	ME
1	$m(1-m)(\Delta V)^2$	0.1699	0.0688	1.1805
2	$m(1-m)(\Delta V)^2$, $\frac{(\Delta V)^2}{(\overline{V})^4}$	0.0793	0.0237	0.9359
3	$m(1-m)(\Delta V)^2$, $\overline{Z}_{eff}(\Delta V)^2$, $\frac{(\Delta V)^2}{(\overline{V})^4}$	0.0483	0.0051	0.4407

(c) xenotime

Table 6: LASSO+ℓ_0 results for Fused, monazite and xenotime with descriptors defined only by arithmetic means and absolute difference

to a simple arithmetic mean, $\overline{\epsilon_k}$ and keep the absolute difference $\Delta\epsilon_k$ (see Table 7).

The second hypothesis, implies that the inverse of each elemental property $\Delta\epsilon_k$ and $\overline{\epsilon_k}$ are also included explicitly in the descriptors. Additionally, due to the third hypothesis, we include as descriptors monomials of degree higher than one for m, V, and R. In particular, for m and $1-m$ we include up to quadratic terms (and their inverse), and for V and R we include up to cubic terms (and their inverse)

Table 7: Two basic descriptors of elemental property ϵ_k

Name	Descriptor
Absolute difference	$\Delta\epsilon_k = \frac{|\epsilon_K(L_i) - \epsilon_K(L_j)|}{2}$
Arithmetic mean	$\overline{\epsilon_k} = \frac{|\epsilon_k(L_i) + \epsilon_k(L_j)|}{2}$

All these descriptors build up a basic feature space $\mathcal{D}_1$ of size 58. Analogously as done for the original descriptors, the group $\mathcal{D}_2$ and $\mathcal{D}_3$ are built as the element-wise product of two or three features out of $\mathcal{D}_1$, respectively. Feature space $\mathcal{D}_2$ is of size 1647 after removing 6 features with standard deviation being 0, such as the element-wise product of $\frac{1}{m}$ and m. The size of feature space $\mathcal{D}_3$ is 30856. The sum

of all candidate functions are collected into $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3$. The size of data $\mathcal{D}$ is 1050×32561 and 525×32561 for fused and for monazite/xenotime data configurations.

We run the same LASSO+ℓ_1 as in Section 3.2.1, the results of which are shown in Fig. 4. Compared to the Fig. 3 the results we obtained are quite more promising. For fused data in Fig. 4, the size of reduced feature space drops down quickly below 40 already at very small value of λ. Afterwards, with the increase of λ, the decrease of the size of reduced feature space slows down, which implies that the reduced feature space contains always important features for the target problem. Meanwhile, with the increase of λ, the errors increase moderately and linearly, which is a second sign that the feature space is reduced into a good choice of representative functions.

After applying LASSO+ℓ_0 on the vector of descriptors spanning the reduced space, the first three leading terms and their corresponding errors for fused, monazite and xenotime data configurations are shown in Table 6. It is important to notice that while the errors may still be quite high for keeping only the leading term, they decrease rapidly when we include higher terms. Moreover the first leading term for the monazite and xenotime data configurations resemble very closely the expression of Equation (1.1). We will analyze in more details the results and interpret their physical meaning up to five leading terms in the next section.

4. Numerical results. As illustrated in Sec. 1.1, the discrepancies between existing models and the data makes the computation of the excess of enthalpy for solid solutions of both monazite and xenotime, a clear cut example to demonstrate the validity of our statistical approach in retrieving an explainable expression for H^E. We will (1) show which descriptor provides the most reliable leading term for the mixing enthalpy between ionic radii of the mixing cations and the volumes of the pure phases, (2) identify the first sub-leading term, which enhances the accuracy of the H^E prediction of xenotime-type solid solutions, (3) provide a exhausting statistical analysis for additional sub-leading terms.

Table 8: Fused: all elemental descriptors except Ionic radius coordination

#	Functions
1	$72.5497\frac{\Delta V \Delta Y}{IP^{2+}} + 0.3661$
2	$302.2691\frac{\Delta V \Delta Y}{\overline{Y}} + 14{,}068.2423 m^2(1-m)^2(\Delta V)^2 - 0.0271$
3	$267.3993\frac{\Delta V \Delta Y}{\overline{Y}} + 13{,}775.7831 m^2(1-m)^2(\Delta V)^2 + 7.1386(1-m)\Delta Y(\Delta V)^2$ $+0.0135$
4	$22.9437\frac{\Delta V \Delta Y}{IP^{2+}} + 13{,}090.1088 m^2(1-m)^2(\Delta V)^2 + 5.9124(1-m)\Delta Y(\Delta V)^2$ $+185.5363\frac{(\Delta V)^2\Delta Y}{\Delta M} + 0.0040$
5	$0.3198 m^2(1-m)^2\Delta Y + 0.0001\overline{IP^{3+}\Delta IP^{3+}\Delta Y} + 0.0335\overline{\Delta V IP^{3+}\Delta Y}$ $+12{,}977 m^2(1-m)^2(\Delta V)^2 + 7.3732(1-m)\Delta Y(\Delta V)^2 - 0.0603$

Error \ Desc. #	1	2	3	4	5
MAE	0.5581	0.1374	0.1249	0.1110	0.1075
MSE	0.6803	0.0526	0.0403	0.0329	0.0260
ME	4.0386	1.7908	1.5139	1.1327	1.2419

In order to address the point (1)-(3) we apply the sparsification integrated with prior knowledge to a variety of scenarios: (i) use all modified descriptors illustrated

Table 9: Fused: all elemental descriptors except volume

#	Functions
1	$72.5497\frac{\Delta R \Delta Y}{IP^{2+}} + 0.3661$
2	$302.2691\frac{\Delta R \Delta Y}{\overline{Y}} + 14{,}068.2423m^2(1-m)^2(\Delta R)^2 - 0.0271$
3	$267.3993\frac{\Delta R \Delta Y}{\overline{\overline{Y}}} + 13{,}775.7831m^2(1-m)^2(\Delta R)^2 + 7.1386(1-m)\Delta Y(\Delta R)^2$ $+0.0135$
4	$22.9437\frac{\Delta R \Delta Y}{IP^{2+}} + 13{,}090.1088m^2(1-m)^2(\Delta R)^2 + 5.9124(1-m)\Delta Y(\Delta R)^2$ $+185.5363\frac{(\Delta R)^2 \Delta Y}{\Delta M} + 0.0040$
5	$0.3198m^2(1-m)^2\Delta Y + 0.0001\overline{IP^{3+}}\Delta IP3\Delta Y + 0.0335\Delta R\overline{IP^{3+}}\Delta Y$ $12{,}977.1177m^2(1-m)^2(\Delta R)^2 + 7.3732(1-m)\Delta Y(\Delta R)^2 - 0.0603$

Error \ Desc. #	1	2	3	4	5
MAE	0.5581	0.1374	0.1249	0.1110	0.1075
MSE	0.6803	0.0526	0.0403	0.0329	0.0260
ME	4.0386	1.7908	1.5139	1.1327	1.2419

in Sec. 3.2.2, (ii) use all modified descriptors excluding the one based on ionic radii elementals R^8 and R^9, (iii) use all modified descriptors excluding the one based on the volume elemental V, and (iv) using only descriptors based on the volume V, mixing ratio m and Young modulus Y elementals. In addition, we used three different set of data—monazite, xenotime, fused—for each of the four cases (i)–(iv). In total, we produced 12 separate sparsification scenarios, each specifying functions up to five leading terms and the relative errors.

Table 10: Monazite: all elemental descriptors except Ionic radius coordination

#	Functions
1	$0.9246m(1-m)(\Delta V)^2 + 0.0173$
2	$0.8044m(1-m)(\Delta V)^2 + 0.5302m^2(1-m)^2(\Delta V)^2 + 0.0220$
3	$0.8668m(1-m)(\Delta V)^2 + 0.1626m(1-m)^2(\Delta V)^2 - 0.00005(\frac{\Delta V}{m})^3 + 0.0108$
4	$24.6538\frac{\Delta V}{IP^{3+}Z_{eff}} + 0.8378m(1-m)(\Delta V)^2 + 0.1708m(1-m)^2(\Delta V)^2$ $-0.00005(\frac{\Delta V}{m})^3 - 0.0337$
5	$-0.0003\frac{\Delta Y}{m^3} + 40.2702\frac{\Delta V}{IP^{3+}Z_{eff}} + 0.8019m(1-m)(\Delta V)^2$ $+0.2111m(1-m)^2(\Delta V)^2 - 0.00003(\frac{\Delta V}{m})^3 - 0.0395$

Error \ Desc. #	1	2	3	4	5
MAE	0.0622	0.0575	0.0516	0.0500	0.0460
MSE	0.0095	0.0082	0.0064	0.0058	0.0051
ME	0.4936	0.6309	0.6195	0.5149	0.4814

The (1) statement is inferred directly from the direct inspection of the tables for cases (ii) and (iii) applied to the fused data set (Tables 8 and 9). From these two

Table 11: Monazite: all elemental descriptors except volume

#	Functions
1	$19{,}964.3970 m^2(1-m)^2(\Delta R)^2 + 0.0382$
2	$12{,}358.3498 m^2(1-m)^2(\Delta R)^2 + 113.7001\frac{(\Delta R)^2\Delta Y}{\Delta Z_{eff}} - 0.0002$
3	$11{,}702.6552 m^2(1-m)^2(\Delta R)^2 + 116.2558\frac{(\Delta R)^2\Delta Y}{\Delta Z_{eff}} + 6{,}851.1011\frac{(\Delta R)^2\Delta\chi}{\Delta Y}$ -0.0278
4	$0.1684\frac{IP^{3+}\overline{\chi}}{IP^{2+}} + 11{,}865.2535 m^2(1-m)^2(\Delta R)^2 + 113.8385\frac{(\Delta R)^2\Delta Y}{\Delta Z_{eff}}$ $+5{,}585.8614\frac{(\Delta R)^2\Delta\chi}{\Delta Y} - 0.3904$
5	$0.6122 m^2(1-m)^2\Delta Y + 0.4085\Delta IP^{3+}\Delta\chi\Delta IP^{2+} + 34.8468\frac{\Delta R\Delta Y}{IP^{2+}}$ $+9{,}966.2662 m^2(1-m)^2(\Delta R)^2 + 175{,}670.7717\frac{(\Delta R)^3}{IP^{2+}\Delta Z_{eff}} - 0.0642$

Desc. # / Error	1	2	3	4	5
MAE	0.2344	0.0858	0.0837	0.0807	0.0810
MSE	0.1296	0.0176	0.0152	0.0145	0.0140
ME	1.4019	0.7153	0.6658	0.6053	0.5780

Table 12: Xenotime: all elemental descriptors except Ionic radius coordination

#	Functions
1	$1.2264 m(1-m)(\Delta V)^2 + 0.0630$
2	$62{,}658.0864\frac{(\Delta V)^2}{(\overline{V})^3} + 2.5431 m^2(1-m)^2(\Delta V)^2 + 0.0094$
3	$53{,}402.4559\frac{(\Delta V)^2}{(\overline{V})^3} + 2.5473 m^2(1-m)^2(\Delta V)^2 + 0.0003(1-m)\overline{Y}(\Delta V)^2$ $+0.0123$
4	$0.0432(1-m)(\Delta V)^2 + 1{,}524.3348(\frac{\Delta V}{\overline{V}})^2 - 1.3755 m(1-m)(\Delta V)^2$ $+5.7829 m^2(1-m)^2(\Delta V)^2 + 0.0047$
5	$0.0430(1-m)(\Delta V)^2 + 5{,}499.6413(\frac{\Delta V}{\overline{V}})^2 - 200{,}379.3047\frac{(\Delta V)^2}{(\overline{V})^3}$ $-3.5547 m(1-m)(\Delta V)^2 + 10.8335 m^2(1-m)^2(\Delta V)^2 + 0.0114$

| Desc. # / Error	1	2	3	4	5
MAE	0.1699	0.0817	0.0492	0.0396	0.0348
MSE	0.0688	0.0243	0.0058	0.0033	0.0024
ME	1.1805	0.9838	0.4178	0.3495	0.3412

tables, one observes that the functions of one up to five leading terms are exactly the same except that wherever in one table the ionic radius R appears in the other the volume V appears in exactly the same terms. Moreover, all of the coefficients of the leading terms are exactly the same across the two tables as well as the all the errors. This obviously points out to the fact that only either ionic radius or volume should be included in the list of elemental properties. In order to understand which among these two elemental properties should be eliminated, we look next to the cases (ii) and (iii)

Table 13: Xenotime: all elemental descriptors except volume

#	Functions
1	$9.8282\frac{\Delta Y \Delta Z_{eff}}{\overline{Y}} - 0.0317$
2	$18{,}347.9196 m^2(1-m)^2(\Delta R)^2 + 3{,}136.8386\frac{(\Delta R)^2\overline{IP^{3+}}}{\overline{Y}} - 0.0109$
3	$18{,}059.4867 m^2(1-m)^2(\Delta R)^2 + 2{,}625.5857\frac{(\Delta R)^2\overline{IP^{3+}}}{\overline{Y}}$ $+93{,}534.5878(1-m)\frac{(\Delta R)^3}{IP^{2+}} + 0.0577$
4	$0.3421\frac{\Delta Z_{eff}\Delta Y}{IP^{2+}} + 17{,}191.3100 m^2(1-m)^2(\Delta R)^2 + 12{,}680{,}655.2904\frac{(\Delta R)^3}{\overline{Y}\Delta M}$ $+89.9428(1-m)(\Delta R)^3\overline{IP^{3+}} + 0.0066$
5	$0.3441\frac{\Delta Z_{eff}\Delta Y}{IP^{2+}} + 6\times 10^{-5}\overline{IP^{3+}}\Delta IP^{3+}\Delta Y + 17{,}407.1 m^2(1-m)^2(\Delta R)^2$ $+10{,}464{,}580.7\frac{(\Delta R)^3}{\overline{Y}\Delta M} + 91.8(1-m)(\Delta R)^3\overline{IP^{3+}} + 0.0054$

Error \ Desc. #	1	2	3	4	5
MAE	0.3575	0.1547	0.1321	0.0871	0.0813
MSE	0.3691	0.0603	0.0335	0.0162	0.0139
ME	3.0150	1.5249	0.7334	0.6216	0.6191

applied separately to monazite and xenotime data sets (Tables 10, 11, 12, and 13).
Comparing the errors, it is immediately obvious that descriptors without the volume
V elemental returns always larger error than descriptors without the ionic radius R
elemental for both monazite and xenotime data sets. This is a strong indication that
V should be preferred over R as the elemental property of choice. This concludes
statement (1) and excludes from further analysis cases (i) and (iii).
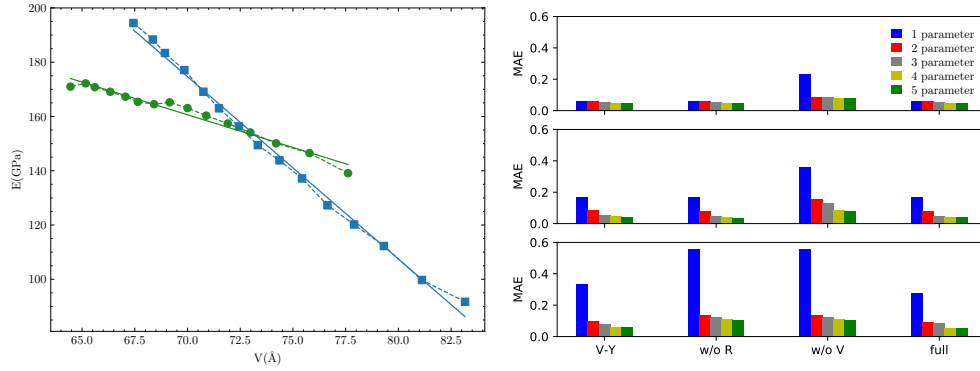
(a) The Young's modulus as a function of volume for monazite (green circles) and xenotime (blue squares) phases. The lines represent the linear regression fits.

(b) The mean absolute error in kJ/mol of the LASSO + ℓ_0 sparsification for the monazite, xenotime and fused cases obtained with the set of 1 to 5 leading terms.

Fig. 5: Young modulus vs Volume and MAE for the LASSO + ℓ_0 sparsification.

Next, we look at cases (iii) and (iv) applied to all three sets of data (Tables 14,

10, 15, 12, 16, and 8). The very first observation is that no matter what data set or case one looks at, all the one-term function are of the form $m(1-m)(\Delta V)^2$. Unequivocally this is the first leading term describing the excess of enthalpy H^E. We also notice that sub-leading terms for monazite and xenotime data sets are dominated by the volume elemental but this dominance manifests itself differently for each distinct data set and with errors that vary from case to case. In the case of xenotime, the prediction of H^E with only one term is not sufficient and additional terms are needed to get satisfactory accuracy. While the leading term is also $m(1-m)\Delta V^2$, the prefactor coefficient is substantially larger that the one obtained for monazite (1.23 vs. 0.94). Although this difference could be possibly explained with on average larger Young's moduli of xenotime phases than monazite (see Eq. 1.1), this demonstrates that additional terms are essential to obtain an equation that consistently described both structures. When we look at the solutions for xenotime alone, different terms appear, usually containing some power of the difference and average of volume and/or Young's modulus. Only simultaneous fit to monazite and xenotime data sets could shed a light on the composition of the second leading term.

Table 14: Monazite: only Volume and Young modulus

#	Functions
1	$0.9247m(1-m)(\Delta V)^2 + 0.0173$
2	$0.3573m(1-m)\Delta V + 0.8674m(1-m)(\Delta V)^2 - 0.066$
3	$0.00004\frac{(\Delta Y)^2}{m^3} + 0.8681m(1-m)(\Delta V)^2 + 0.1527m(1-m)^2(\Delta V)^2 + 0.0207$
4	$0.1028m(1-m)\Delta Y - 0.00004\frac{(\Delta Y)^2}{m^3} + 0.8243m(1-m)(\Delta V)^2$ $+0.1574m(1-m)^2(\Delta V)^2 - 0.0299$
5	$0.1204m(1-m)\Delta Y - 0.00007\frac{(\Delta Y)^2}{m^3} + 0.9109m(1-m)(\Delta V)^2$ $+0.2370m(1-m)^2(\Delta V)^2 - 0.5384m^2(1-m)^2(\Delta V)^2 - 0.0413$

Desc. # / Error	1	2	3	4	5
MAE	0.0622	0.0585	0.0546	0.0482	0.0495
MSE	0.0095	0.0082	0.0068	0.0057	0.0053
ME	0.4936	0.4239	0.6279	0.5357	0.4579

Looking at the fused data, it becomes clear that not only the volume but also the Young modulus Y elemental plays a central role in sub-leading terms. In addition, comparing Tables 16 and 8, it is immediately clear that using descriptors of just V and Y returns consistently smaller errors for all sub-leading terms. Incidentally, Young's moduli and volumes/ionic radii are the elemental properties used in models of [21] and [16]. From this analysis we infer that we should focus on the leading and sub-leading terms provided by table 16, where all elementals apart from V and Y are considered. The leading second term contains product of ΔV and ΔY. With combination of $m(1-m)\Delta V^2$ and $\Delta Y \Delta V/\bar{V}^2$ one can satisfactorily express H^E for both sets of data. This indicates that in addition to difference in volume, the difference in Young's modulus plays a significant role in determining the excess enthalpy of mixing. To better understand the role of this second term, we plot in Fig. 5a the variation of Young's modulus as a function of ΔV for the two phases. For the considered range of elastic moduli and volumes, there is a clear linear relationship between the two.

Table 15: Xenotime: only Volume and Young modulus

#	Functions
1	$1.2264m(1-m)(\Delta V)^2 + 0.0630$
2	$0.0010\overline{Y}(\Delta V)^2 + 2.5799m^2(1-m)^2(\Delta V)^2 + 0.0296$
3	$0.0008\overline{Y}(\Delta V)^2 + 2.5799m^2(1-m)^2(\Delta V)^2 + 0.0003(1-m)\overline{Y}(\Delta V)^2$ $+0.0296$
4	$-0.000008(\overline{Y})^2(\Delta V)^2 + 2.5772m^2(1-m)^2(\Delta V)^2 + 0.0003(1-m)\overline{Y}(\Delta V)^2$ $+11.8681\frac{\overline{Y}(\Delta V)^2}{(\overline{V})^2} + 0.0031$
5	$-0.00002(\overline{Y})^2(\Delta V)^2 - 0.5726m(1-m)(\Delta V)^2 + 3.9222m^2(1-m)^2(\Delta V)^2$ $+0.0003(1-m)\overline{Y}(\Delta V)^2 + 23.1551\frac{\overline{Y}(\Delta V)^2}{(\overline{V})^2} - 0.0154$

Desc. # / Error	1	2	3	4	5
MAE	0.1699	0.0850	0.0547	0.0449	0.0408
MSE	0.0688	0.0252	0.0067	0.0046	0.0037
ME	1.1805	1.0853	0.5429	0.3859	0.3759

However, the linear relationship is much steeper in case of xenotime than monazite, with the slope about three times higher in the former case. This makes the $\Delta Y \Delta V/\bar{V}^2$ term much larger (by a factor ~ 3) in case of xenotime, putting it on equal footing with the leading term. This explains why missing this term, there is a factor of ~ 2 difference between the *ab initio* data and the discussed, prior theoretical models (see Fig. 1).

In general, adding more terms improves the fit only marginally (see Fig. 5b). On the other hand, one can observe from table 16 that for any choice of number of terms, the factor $\Delta Y \Delta V/\bar{V}^2$ is always present, confirming its importance in contributing to the expression of H^E. This concludes the statements (2) and (3).

5. Summary and conclusions. In this work we report on a 3-step approach that combines distinct methods from classical Machine Learning to reach a *scientifically consistent explainable* methodology. First, we use a KRR approach on the generated data and evaluate which kernel returns the least amount of error. Then, using the results of KRR, we reverse engineer the model and proceed to sparsify the vector of coefficients using a so-called LASSO+ℓ_0 procedure. Finally we integrate domain-specific knowledge to force an a-posteriori scientific consistency of the reverse model.

In order to demonstrated the feasibility and potential of this methodology, we have applied it to the computation of the excess of enthalpy of formation H^E of solid solutions of lanthanide Orthophosphates. This particular class of materials is present in nature in two distinct crystal structure—monazite and xenotime—for which no single formula is capable of describing accurately H^E. Applying our machine learning based 3-step method to a set of in silico data, we were able to retrieve subleading corrections to known expressions for H^E, which represent an important step in resolving a conflicting description of the excess of enthalpy for both type of structures. We expect that the importance of accounting for the gradient of elastic moduli when estimating the excess enthalpy of mixing will trigger follow-up theoretical studies

Table 16: Monazite+Xenotime: only Volume and Young modulus

#	Functions
1	$1.1453m(1-m)(\Delta V)^2 - 0.0517$
2	$2.9453m^2(1-m)^2(\Delta V)^2 + 108.1079\frac{\Delta V \Delta Y}{(\overline{V})^2} + 0.0429$
3	$0.0066m(1-m)^2(\Delta Y)^2 + 0.7963m(1-m)(\Delta V)^2 + 54.3783\frac{\Delta Y \Delta V}{(\overline{V})^2}$ $+0.0064$
4	$0.0047m(1-m)^2(\Delta Y)^2 + 0.4672m(1-m)(\Delta V)^2 + 1.2743m^2(1-m)^2(\Delta V)^2$ $+70.8755\frac{\Delta V \Delta Y}{(\overline{V})^2} + 0.0195$
5	$0.5117m(1-m)\Delta V + 0.0001(1-m)\Delta V(\Delta Y)^2 + 0.2862m(1-m)(\Delta V)^2$ $+1.639m^2(1-m)^2(\Delta V)^2 + 78.1966\frac{\Delta V \Delta Y}{(\overline{V})^2} - 0.0764$

Error Desc. #	1	2	3	4	5
MAE	0.3316	0.0997	0.0801	0.0630	0.0594
MSE	0.2804	0.0262	0.0173	0.0091	0.0083
ME	2.8122	1.3163	0.7561	0.6165	0.5345

aiming at providing physical interpretation of the origin of this phenomenon. The successful application of our procedure shows the potential of its application to other areas of Quantum Chemistry and Materials Science where explainability of machine learning models is an essential feature.

REFERENCES

[1] G. BERIDZE, A. BIRNIE, S. KONISKI, Y. JI, AND P. M. KOWALSKI, *Dft + u as a reliable method for efficient ab initio calculations of nuclear materials*, Prog. Nucl. Energ., 92 (2016), pp. 142 – 146.

[2] A. BLANCA-ROMERO, P. M. KOWALSKI, G. BERIDZE, H. SCHLENZ, AND D. BOSBACH, *Performance of DFT plus U Method for Prediction of Structural and Thermodynamic Parameters of Monazite-Type Ceramics*, J. Comput. Chem., 35 (2014), pp. 1339–1346.

[3] D. S. BOLME, Y. M. LUI, B. A. DRAPER, J. R. BEVERIDGE, T. J. CHIN, U. JAMES, K. SCHINDLER, D. SUTER, A. ELGAMMAL, R. DURAISWAMI, D. HARWOOD, L. S. DAVIS, I. HARITAOGLU, R. CUTLER, D. HARWOOD, L. S. DAVIS, S. G. MIAOU, P. H. SUNG, C. Y. HUANG, Y. MING, G. WANG, C. FAN, D. DUQUE, H. SANTOS, P. CORTEZ, W. HU, T. TAN, L. WANG, S. MAYBANK, C. FERNÁNDEZ, P. BAIGET, F. X. ROCA, J. GONZLEZ, B. YAO, H. HAGRAS, D. ALGHAZZAWI, M. J. ALHADDAD, A. PSARROU, S. GONG, M. WALTER, A. BOUZIANE, Y. CHAHIR, M. MOLINA, F. JOUEN, A. MOKHBER, C. ACHARD, M. MILGRAM, Y. A. IVANOV, A. F. BOBICK, G. ACAMPORA, P. FOGGIA, A. SAGGESE, M. VENTO, T. DUONG, D. PHUNG, H. BUI, S. VENKATESH, S. K. MISHRA, K. S. BHAGAT, D. SCHREIBER, C. BELEZNAI, M. RAUTER, N. A. OGALE, Y. TAIGMAN, M. A. RANZATO, T. AVIV, M. PARK, R. M. ZUCKER, M. PAUL, S. M. E. HAQUE, S. CHAKRABORTY, C. RAGHAVACHARI, V. APARNA, S. CHITHIRA, V. BALASUBRAMANIAN, C. ROUGIER, J. MEUNIER, A. ST-ARNAUD, J. ROUSSEAU, Y. TAIGMAN, M. A. RANZATO, T. AVIV, AND M. PARK, *Deepface*, Department of Computer Science University Of Maryland College Park, (2015), https://doi.org/10.1109/CVPR.2014.220, https://arxiv.org/abs/NIHMS150003.

[4] J. W. CHRISTIAN, *The theory of transformations in metals and alloys: An advanced textbook*

in physical metallurgy, 2nd ed. ed., 1975.

[5] N. Clavier, R. Podor, and N. Dacheux, *Crystal chemistry of the monazite structure*, J. Eur. Ceram. Soc., 31 (2011), pp. 941–976.

[6] A. M. Deml, R. O'Hayre, C. Wolverton, and V. Stevanović, *Predicting density functional theory total energies and enthalpies of formation of metal-nonmetal compounds by linear regression*, Phys. Rev. B, 93 (2016), p. 085142, https://doi.org/10.1103/PhysRevB.93.085142, https://link.aps.org/doi/10.1103/PhysRevB.93.085142.

[7] R. Ewing and L. Wang, *Phosphates as nuclear waste forms*, Rev. Mineral. and Geochem., 48 (2002), pp. 673–699.

[8] L. M. Ghiringhelli, J. Vybiral, E. Ahmetcik, R. Ouyang, S. V. Levchenko, C. Draxl, and M. Scheffler, *Learning physical descriptors for materials science by compressed sensing*, New Journal of Physics, 19 (2017), p. 023017, https://doi.org/10.1088/1367-2630/aa57bf, https://dx.doi.org/10.1088/1367-2630/aa57bf.

[9] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, *Big data of materials science: Critical role of the descriptor*, Physical Review Letters, (2015), https://doi.org/10.1103/PhysRevLett.114.105503, https://arxiv.org/abs/1411.7437.

[10] P. Glynn, *Solid-solution solubilities and thermodynamics: Sulfates, carbonates and halides*, Reviews in Mineralogy and Geochemistry, 40 (2000), pp. 481–511, https://doi.org/10.2138/rmg.2000.40.10, https://doi.org/10.2138/rmg.2000.40.10.

[11] T. Hofmann, B. Schölkopf, and A. J. Smola, *Kernel methods in machine learning*, The Annals of Statistics, 36 (2008), pp. 1171 – 1220, https://doi.org/10.1214/009053607000000677, https://doi.org/10.1214/009053607000000677.

[12] Y. Ji, G. Beridze, D. Bosbach, and P. M. Kowalski, *Heat capacities of xenotime-type ceramics: An accurate ab initio prediction*, J. Nucl. Mater., 494 (2017), pp. 172 – 181.

[13] Y. Ji, G. Beridze, Y. Li, and P. M. Kowalski, *Large scale simulation of nuclear waste materials*, Energy Procedia, 127 (2017), pp. 416 – 424.

[14] Y. Ji, P. M. Kowalski, P. Kegler, N. Huittinen, N. A. Marks, V. L. Vinograd, Y. Arinicheva, S. Neumeier, and D. Bosbach, *Rare-earth orthophosphates from atomistic simulations*, Frontiers in Chemistry, 7 (2019), p. 197, https://doi.org/10.3389/fchem.2019.00197, https://www.frontiersin.org/article/10.3389/fchem.2019.00197.

[15] P. M. Kowalski, G. Beridze, V. L. Vinograd, and D. Bosbach, *Heat capacities of lanthanide and actinide monazite-type ceramics*, J. Nucl. Mater., 464 (2015), pp. 147–154.

[16] P. M. Kowalski and Y. Li, *Relationship between the thermodynamic excess properties of mixing and the elastic moduli in the monazite-type ceramics*, J. Eur. Ceram. Soc., 36 (2016), pp. 2093 – 2096.

[17] LeCun Yann, Cortes Corinna, and Burges Christopher, *THE MNIST DATABASE of handwritten digits*, The Courant Institute of Mathematical Sciences, (1998).

[18] Y. Li, P. M. Kowalski, A. Blanca-Romero, V. Vinograd, and D. Bosbach, *Ab initio calculation of excess properties of solid solutions*, J. Solid State Chem., 220 (2014), pp. 137 – 141.

[19] I. McGill, *Rare Earth Elements*, American Cancer Society, 2000, https://doi.org/10.1002/14356007.a22_607, https://onlinelibrary.wiley.com/doi/abs/10.1002/14356007.a22_607.

[20] E. Mjolsness and D. DeCoste, *Machine learning for science: State of the art and future prospects*, 2001, https://doi.org/10.1126/science.293.5537.2051.

[21] P. Mogilevsky, *On the miscibility gap in monazite–xenotime systems*, Phys. Chem. Miner., 34 (2007), pp. 201–214.

[22] S. Neumeier, Y. Arinicheva, Y. Ji, J. M. Heuser, P. M. Kowalski, P. Kegler, H. Schlenz, D. Bosbach, and G. Deissmann, *New insights into phosphate based materials for the immobilisation of actinides*, Radiochim. Acta, 105 (2017), pp. 961–984.

[23] S. Neumeier, P. Kegler, Y. Arinicheva, A. Shelyug, P. M. Kowalski, C. Schreinemachers, A. Navrotsky, and D. Bosbach, *Thermochemistry of la1-xlnxpo4-monazites (ln=gd, eu)*, J. Chem. Thermodyn., 105 (2017), pp. 396 – 403.

[24] Y. Ni, J. M. Hughes, and A. N. Mariano, *Crystal chemistry of the monazite and xenotime structures*, American Mineralogist, 80 (1995), p. 21, https://doi.org/10.2138/am-1995-1-203, http://dx.doi.org/10.2138/am-1995-1-203.

[25] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, *Restoring the density-gradient expansion for exchange in solids and surfaces*, Phys. Rev. Lett., 100 (2008), p. 136406.

[26] K. Popa, R. J. M. Konings, and T. Geisler, *High-temperature calorimetry of (la(1-x)ln(x))po4 solid solutions*, Journal of Chemical Thermodynamics, 39 (2007), pp. 236–239, https://doi.org/10.1016/j.jct.2006.07.010.

[27] M. Prieto, *Thermodynamics of solid solution-aqueous solution systems*, 70 (2009), pp. 47–85,

https://doi.org/10.2138/rmg.2009.70.2. 19th Annual V M Goldschmidt Conference, Davos, SWITZERLAND, JUN 19-21, 2009.

[28] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, *Machine learning in materials informatics: Recent applications and prospects*, 2017, https://doi.org/10.1038/s41524-017-0056-5, https://arxiv.org/abs/1707.07294.

[29] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, *Explainable machine learning for scientific insights and discoveries*, IEEE Access, 8 (2020), pp. 42200–42216, https://doi.org/10.1109/ACCESS.2020.2976199.

[30] M. Rupp, *Machine learning for quantum mechanics in a nutshell*, International Journal of Quantum Chemistry, 115 (2015), pp. 1058–1073, https://doi.org/https://doi.org/10.1002/qua.24954, https://onlinelibrary.wiley.com/doi/abs/10.1002/qua.24954, https://arxiv.org/abs/https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.24954.

[31] J. R. Rustad, *Density functional calculations of the enthalpies of formation of rare-earth orthophosphates*, Am. Mineral., 97 (2012), p. 791.

[32] F. Santosa and W. W. Symes, *Linear Inversion of Band-Limited Reflection Seismograms*, SIAM Journal on Scientific and Statistical Computing, (1986), https://doi.org/10.1137/0907087.

[33] H. Schlenz, J. Heuser, A. Neumann, S. Schmitz, and D. Bosbach, *Monazite as a suitable actinide waste form*, Z Kristallogr, 228 (2013), pp. 113–123.

[34] V. Stevanović, S. Lany, X. Zhang, and A. Zunger, *Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies*, Phys. Rev. B, 85 (2012), p. 115104, https://doi.org/10.1103/PhysRevB.85.115104, https://link.aps.org/doi/10.1103/PhysRevB.85.115104.

[35] W. Stoll, *Thorium and Thorium Compounds*, American Cancer Society, 2000, https://doi.org/10.1002/14356007.a27_001, https://onlinelibrary.wiley.com/doi/abs/10.1002/14356007.a27_001.

[36] C. Sutton, M. Boley, L. M. Ghiringhelli, M. Rupp, J. Vreeken, and M. Scheffler, *Identifying domains of applicability of machine learning models for materials science*, Nature Communications, 11 (2020), p. 4428, https://doi.org/10.1038/s41467-020-17112-9, https://doi.org/10.1038/s41467-020-17112-9.

[37] S. Ubaru, A. Miedlar, Y. Saad, and J. R. Chelikowsky, *Formation enthalpies for transition metal alloys using machine learning*, Phys. Rev. B, 95 (2017), p. 214102, https://doi.org/10.1103/PhysRevB.95.214102, https://link.aps.org/doi/10.1103/PhysRevB.95.214102.

[38] S. Ushakov, K. Helean, and A. Navrotsky, *Thermochemistry of rare-earth orthophosphates*, J. Mater. Res., 16 (2001), p. 2623.

[39] L. von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, M. Walczak, J. Pfrommer, A. Pick, R. Ramamurthy, J. Garcke, C. Bauckhage, and J. Schuecker, *Informed machine learning - a taxonomy and survey of integrating prior knowledge into learnig systems*, IEEE Transactions on Knowledge and Data Engineering, (2021), pp. 1–1, https://doi.org/10.1109/TKDE.2021.3079836.

[40] K. L. Wagstaff, *Machine learning that matters*, in Proceedings of the 29th International Conference on Machine Learning, ICML 2012, 2012, https://arxiv.org/abs/1206.4656.

[41] A. Zunger, S. Wei, L. Ferreira, and J. Bernard, *Special quasirandom structures*, PHYSICAL REVIEW LETTERS, 65 (1990), pp. 353–356, https://doi.org/10.1103/PhysRevLett.65.353.