Supervised deep learning prediction of the formation enthalpy of complex phases using a DFT database: the σ–phase as an example

Jean-Claude Crivello, Jean-Marc Joubert, Nataliya Sokolovska

Supervised deep learning prediction of the formation enthalpy of complex phases using a DFT database: the $\sigma-$phase as an example

Jean-Claude Crivello[1], Jean-Marc Joubert

Univ Paris Est Creteil, CNRS, ICMPE, UMR 7182, 2 rue Henri Dunant, 94320 Thiais, France

Nataliya Sokolovska

NutriOmics, INSERM, Sorbonne University Paris, France

Abstract

Machine learning (ML) methods are becoming the state-of-the-art in numerous domains, including material sciences. In this manuscript, we demonstrate how ML can be used to efficiently predict several properties in solid-state chemistry applications, in particular, to estimate the heat of formation of a given complex crystallographic phase (here, the $\sigma-$phase, $tP30$, $D8_b$). Based on an independent and unprecedented large first principles dataset containing about 10,000 $\sigma-$compounds with $n = 14$ different elements, we used a supervised learning approach to predict all the $\sim$500,000 possible configurations. From a random set of $\sim$1000 samples, predictions are given within a mean absolute error of $23\,\mathrm{meV\,at^{-1}}$ ($\sim 2\,\mathrm{kJ.mol^{-1}}$) on the heat of formation and $\sim 0.06\,\text{Å}$ on the tetragonal cell parameters. We show that deep neural network regression results in a significant improvement in the accuracy of the predicted output compared to traditional regression techniques. We also integrated descriptors having physical nature (atomic radius, number of valence electrons), and we observe that they improve the model precision. We conclude from our numerical experiments that the learning database composed of the binary-compositions only, plays a

*Corresponding author
Email address: crivello@icmpe.cnrs.fr (Jean-Marc Joubert)

major role in predicting the higher degree system configurations. Our results open a broad avenue to efficient high-throughput investigations of the combinatorial binary computations for multicomponent complex intermetallic phase prediction.

Keywords: intermetallic, $\sigma-$phase, heat of formation, machine learning, DFT

1. Introduction

Machine learning, being a domain of artificial intelligence, revolutionised research in many fields (image processing, natural language, speech processing, biology and medicine, *etc.* [1, 2, 3]), and the number of publications introducing new statistical methods has exploded within the last decades. Strange but true that the applications of the machine learning methods to the material sciences, although more and more visible [4], are still falling behind compared to other applications. The statistical machine learning (ML) is the art to construct statistical models from observational data. Among the successful applications of the ML to the materials science, is automatic extraction of predictive models from existing materials data [5, 6], and discovery of new classes of promising materials or composition, such as the high entropy alloys (HEA) [7, 8, 9, 10]. So far, materials scientists have used ML to build predictive models for a few applications such as to predict heat capacity [11], semiconducting band gap [12, 13] *etc.*, but also the heat of formation of intermetallic compounds [14, 15, 16, 17]. All these recent studies have emerged with the powerful use of high-throughput calculations, such as density functional theory (DFT) impulsed for large projects in the last decade (AFLOW [18], OQMD [18], NOMAD [19] *etc.*). In fact, the increasing availability of DFT data, combined with modern data mining and ML techniques, has enabled the construction of a predictive model to replace DFT calculations and accelerate data generation [20]. Prediction of crystal structure is still the holy Grail in inorganic chemistry, while the component prediction is one promising approach [21]. However, in a recent work of Kim *et al.* [22], the effect of the space quality has been investigated, and it was reported that ML

performance can be rather poor if there are several bad (noisy) training data quite close to good candidates [23].

The contribution of the current work is two-fold: (i) we introduce a general high performance ML-based framework for predicting the heat of formation, corresponding to the energy scale which measures the strength of chemical bonds in a compound, where an input to the ML methods are the combinations of the elements in a given crystal structure; and (ii) we present and explore new original data set that we constructed and manage.

Note that the heat of formation prediction was the aim of several studies and attempts, like the semi-empirical Miedema's model [24]. In fact, this fundamental value, called also enthalpy of formation ($\Delta_f H$ in meV) is the key parameter in thermodynamics modelling, such as in the calculation of phase diagrams (Calphad method) [25]. We have applied the $\Delta_f H$ determination to a large combinatorial challenge, yielded by the distribution of every atom from a given space (n-base) into every s non-equivalent crystallographic sites of a given phase. This kind of description is well known in thermodynamic modelling for addressing the energy of a multicomponent and non-stochiometric phases and is called the Compound Energy Formalism (CEF) [26]. In CEF, each crystal site is considered as a sublattice and the distribution of every atom generates n^s unique configurations, called end-members, the $\Delta_f H$ of which has to be given to use this model.

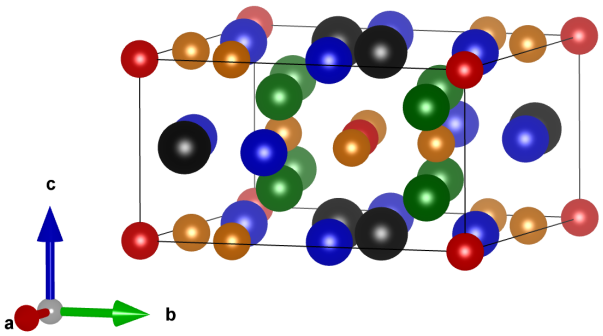

Figure 1: Representation of the primitive cell of the crystal phase $\sigma-$phase ($D8_b$), with its 5 non-equivalent sites: $2a$ (red balls), $4f$ (black), $8i_1$ (blue), $8i_1$ (orange) and $8j$ (green).

To illustrate the efficiency of our ML framework, we investigated an important intermetallic phase in the field of metallurgy: the $\sigma-$phase ($D8_b$). Its complex crystallographic structure is composed of 30 atoms in its tetragonal cell, occupying $s = 5$ distinct sites (i, j, k, l, m), as shown in Figure 1. The $\sigma-$phase belongs to the FrankKasper or topologically close packed phases, characterized by the unique presence of tetrahedral interstices, and a limited number of coordination polyhedra. Its features has been discussed in details [27]. Some details are summarized in Supplementary Materials A, SM-A.

This phase appears in many types of engineering alloys and its formation prediction requires reliable thermodynamical description. It is shown that it is important to keep a 5-sublattice model in CEF to properly describe the configuration of the $\sigma-$phase in multicomponent alloys [28, 29]. The difficulty lies in the large number of end members which must be considered in multicomponent systems. In fact, the description for a binary system (2 elements) leads to the generation of $2^5 = 32$ different ordered configurations to express, but this number rapidly increases with the degree of the system: ternary with $3^5 = 243$, quaternary with with 4^5=1024, ... up to a real alloys with ~ 14 different elements and its $14^5 = 537,824$ configurations. Since the corresponding huge number of $\Delta_f H$ cannot be calculated by classical DFT, their prediction using ML is computationally tractable and, therefore, looks attractive and is one of the major contributions of this paper.

2. Computational methods

2.1. General-purpose approach

The originality of current work is in construction of our learning database. Instead of a mishmash of massive data coming from several independent phases and from various high-throughput sources (calculated using different parameters), we built an original single σ-phase oriented database with our own consistent massive DFT calculations. In addition to some additional physical parameters, the main descriptors are the combination of $n = 14$ different elements

on the different $s = 5$ crystallographic sites that described the σ-phase. Among the combinatorial set (described in SM-B), a selection of data in the training set includes all the possible binary compositions (system degree $d = 2$), which represents only 0.5% of the all possible configurations ($\binom{14}{2} = 91$ purely binary systems with 30 unique configurations each). A graphical chart-flow of the present methodology is given on Figure 2.

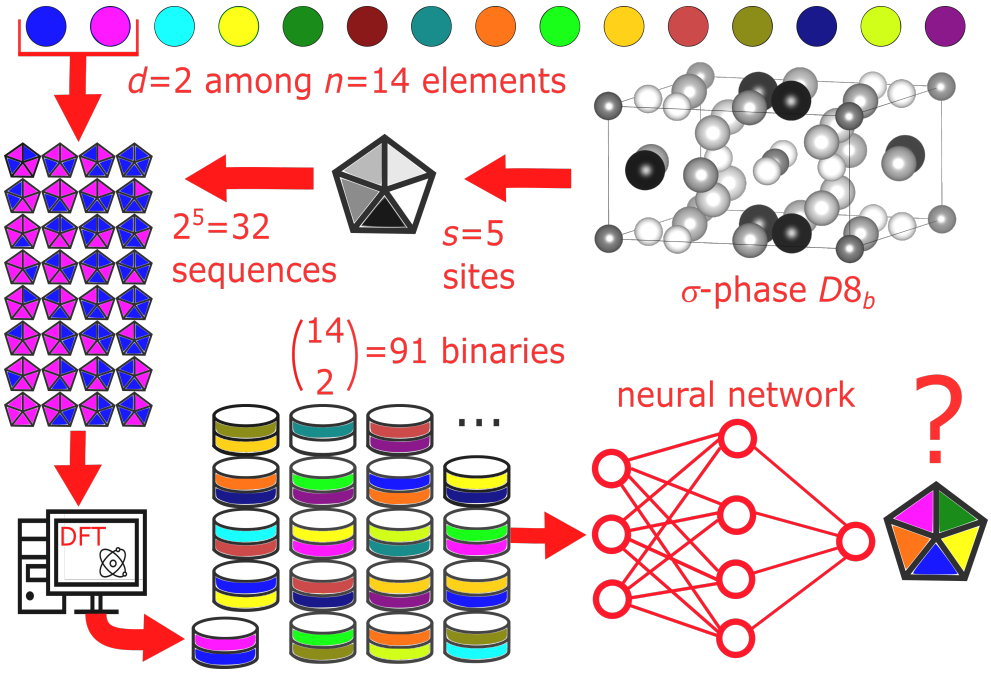

Figure 2: Chart-flow of the methodology presented in the paper. (*i*) The crystal structure is summarized as a s non-equivalent sites figure; (*ii*) from n available elements, a given system of system degree d is selected (*e.g.* $d = 2$ for binary); (*iii*) the permutation leads to d^5 unique configurations ; (*iv*) every configuration is calculated by DFT, forming a unit of data; (*v*) the stack of all $\binom{n}{d}$ units forms a learning database; then (*vi*) a supervised machine learning is used to predict multicomponent configurations.

2.2. Training database from DFT calculations

First, a database from DFT calculations has been compiled. Since 2008, many active groups have calculated σ-phase configurations in binary [33, 34, 32, 35, 36], ternary [37, 38, 39, 40] and quaternary systems [41]. Since all these sparse studies were calculated with different methods and parameters,

our present original database includes only new calculations obtained under the same conditions, and required millions of hours of CPU time to construct. The DFT methodology applied is a classical approach and its details are explained in SM-C. The heat of formation, $\Delta_f H$, of every configuration is given by the difference of its total DFT energy related to the element energies in their stable reference state.

The learning database includes $n = 14$ different elements: Al, Co, Cr, Fe, Mn, Mo, Nb, Ni, Pt, Re, Ru, V, W, Zr, and contains 9974 unique configurations embracing all the $\binom{14}{2} = 91$ binaries (degree $d = 2$), 33 on the $\binom{14}{3} = 364$ ternaries ($d = 3$), 9 on the 1001 quaternaries ($d = 4$) and only 1 on the 2002 possible quinaries ($d = 5$, see SM-D for the detailed list of included systems in our training database). The elemental distribution is not uniform because of some chemical reasons explaining that we wanted to have more data for pertinent systems (*e.g.* Zr-based σ-phase is not frequent). This analysis is illustrated on Figure 3 and could be seen in details from SM-D. In addition, an independent testing set for 1001 randomized configuration were calculated (detailed in SM-E).

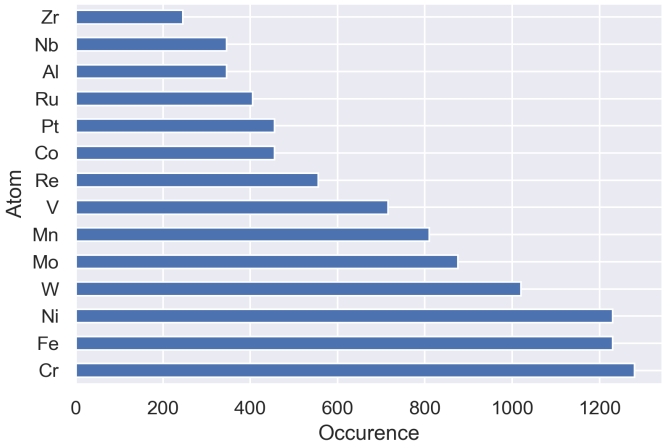

Figure 3: Occurrence of the $n = 14$ elements in the training database (9974 entries).

2.3. Database construction format

In a second step, data was arranged as a learning database, $X_{ijklm} \in \mathbb{R}^{(n+2)s\times N}$. The $n = 14$ elements are categorical variables but need to be treated with analytical methods that require numbers. Thus, each of the 5 crystal sites (i, j, k, l, m) has been considered as a 14-dim vector of dummies (spin variables: 0 or 1) by the one hot encoding method. In addition, because it is well known that the stability in this kind of compounds is driven by the two geometric and electronic constraints [27] (*e.g.* large electropositive atoms have a preference on high coordination sites), atom size and electron concentration have been used as additional descriptors. In total, we use a set of $p = (n + 2)s = 80$ attributes corresponding to the following features for each configuration X_{ijklm}, with (i, j, k, l, m) $\in$ { Al, Co, Cr, ..., V, W, Zr }:

- Ordering configuration of atoms in the crystal (14×5 vectors of dummies)
- Atomic radius (5 normalized values, related to the 5 atoms in $ijklm$ configuration)
- Number of valence electrons (5 normalized values)

leading to a 9974×80 matrix as the learning database, associated to the target y_{ijklm} vector, here the heat of formation, $\Delta_f H(ijklm)$, but could be any crystallographic properties such as cell parameters.

At last, based on the ML best results, the learning on the whole database (9974 configurations) was done and a final prediction of 1001 random configurations among the 537,824 was estimated (details in SM-D and SM-E respectively).

2.4. Estimation of the machine learning models

The machine learning models are estimated using Scikit-Learn 0.23 library in Python 3.5. Each approach offers different advantages, such as speed or interpretability, but our main goal was the high accuracy. Figure 4 illustrates the relationship between the model's simplicity (interpretability) and the generalising performance for the machine learning methods we considered. Note

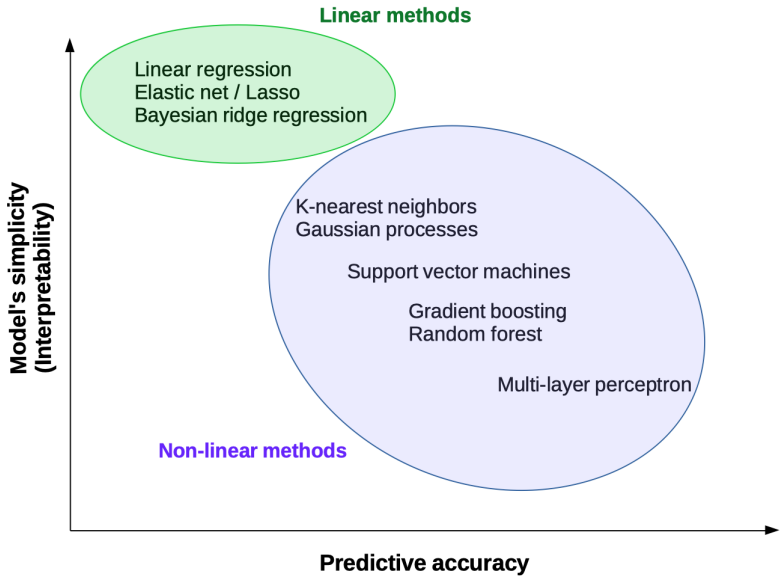

Figure 4: A schematic representation of the models' simplicity and the generalising accuracy for the tested Machine Learning approaches.

that this scheme (Figure 4) is approximate. What is true for our results is that the linear methods perform worse than the non-linear regressions. We also noticed that the grid search for an optimal Multi-Layer Perceptron architecture (number of hidden layers and units), as well as the number of trees and their maximal depth in the Random Forest and Gradient Boosting, is important, and the search for an optimal configuration can be computationally expensive.

For each method, the corresponding hyper-parameters were fixed using the grid search module and the cross-validation error rate. The generalising performance is the test accuracy using 10-fold cross validation (CV) procedure: the database is randomly split into 10 subsets (folds), and the model is trained on 9 parts, and tested on 1 part. The procedure is repeated 10 times. The average accuracy is the mean value over performances on the test data. All machine learning methods used are described in SM-F.

3. Results

The results of this study – our new and original data set, ML metrics used to test the statistical methods, and the corresponding results – are described below. We demonstrate its application to the heat of formation prediction, and, finally, we report the prediction performance.

3.1. Prediction of the heat of formation

The observations are the independent variables from configurations X_{ijklm} of the training database with $N \simeq 10{,}000$ data, and the aim of the regression analysis is to generate a statistical model that can predict a dependent variable, y_{ijklm} (the heat of formation in our case). Several regression algorithms have been investigated and for each of them, the (hyper)-parameters have been choosing by a cross-validation grid-search method to produce the most accurate generalising results. The evaluation of the prediction accuracy of a model is characterized using the coefficient of determination R^2, the mean absolute error MAE, and the root mean squared error RMSE, given as:

$$R^2 = 1 - \frac{\sum \left(y_{ijklm} - \hat{y}_{ijklm}\right)^2}{\sum \left(y_{ijklm} - \bar{y}\right)^2}, \tag{1}$$

$$\text{MAE} = \frac{1}{N} \sum_{\{i,j,k,l,m\}}^{N} \left|y_{ijklm} - \hat{y}\right|, \tag{2}$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{\{i,j,k,l,m\}}^{N} \left(y_{ijklm} - \hat{y}_{ijklm}\right)^2}, \tag{3}$$

where $\hat{y}$ is the predicted value based on the learned model, and $\bar{y}$ the mean of the observed data.

First, we have estimated a number of regression models from our learning database. Namely, we tested the Ridge Linear Regression, Elastic Net Linear Regression, Random Forest Regression (RFR), Multi-Layer Perceptron Regression (MPR), Gradient Boosting Machine (GBM), Support Vector Machine (SVM), K-Nearest Neighbours, Bayesian Ridge Regression, and Gaussian Process Regression (GPR). Our averaged results from 10-fold cross validation are

summarized in Table 1, and also fully shown in SM-G. Classical regression with various regularization such as LASSO (Least Absolute Shrinkage and Selection Operator), Ridge regression, or their combination, also known as Elastic Net, are not accurate enough, since the number of observations in the database is not very big. Moreover, the sparsity inducing penalties (LASSO and Elastic Net) are not very relevant to our case, since the number of parameters is quite small.

Table 1: Cross validation scores on the complete data set (average values from 10-fold) using various machine learning methods. MAE, MSE and RMSE in meV/at (1 meV∼0.0965 kJ/mol), illustrated in SM-G. An horizontal line separates the linear from the non-linear methods.

Algorithm	R^2	MAE	RMSE
Ridge Linear Regression	0.45	73	110
Bayesian Ridge Regression	0.45	73	110
Elastic Net Linear Regression	0.46	73	109
K-nearest neighbors	0.61	55	93
Gaussian Process Regression (GPR)	0.87	30	66
Random Forest Regressor (RFR)	0.89	31	56
Support Vector Machine Reg (SVM)	0.91	26	54
Gradient Boosting Machine (GBM)	0.95	19	37
Multi-layer Perceptron Regressor (MPR)	0.96	13	31

On the other hand, non-linear supervised learning methods achieve very reasonable performance. The R^2 closest to 1 are obtained with RFR, MPR, GBM and SVM regression algorithms. The associated best MAE (average from 10-fold CV) are obtained for the MPR method with 13 meV ($\sim$1 kJ mol^{-1}) using 3 hidden layers, each containing 500 units.

3.2. Comparison of testing performance on unseen dataset

Random Forest Regression, Multi-Layer Perceptron Regression, Gradient Boosting and Support Vector Machines have shown the best performance on the data set containing 9974 inputs (our training database, SM-D). We tested these regressors on a new, previously unobserved during the training procedure, set of configurations: 1001 randomized observations (our testing database, SM-E) among the 537,824 possible ones. Using MPR, the achieved accuracy of total energy, and therefore heat of formation, with a MAE of about 23 meV at^{-1}

(Figure 5) provides a very reasonable accuracy compared to other ML methods of the literature, where the standards are usually around MAE$\sim$50 meV at^{-1} for systems higher than binaries [14, 15, 20, 30, 31].

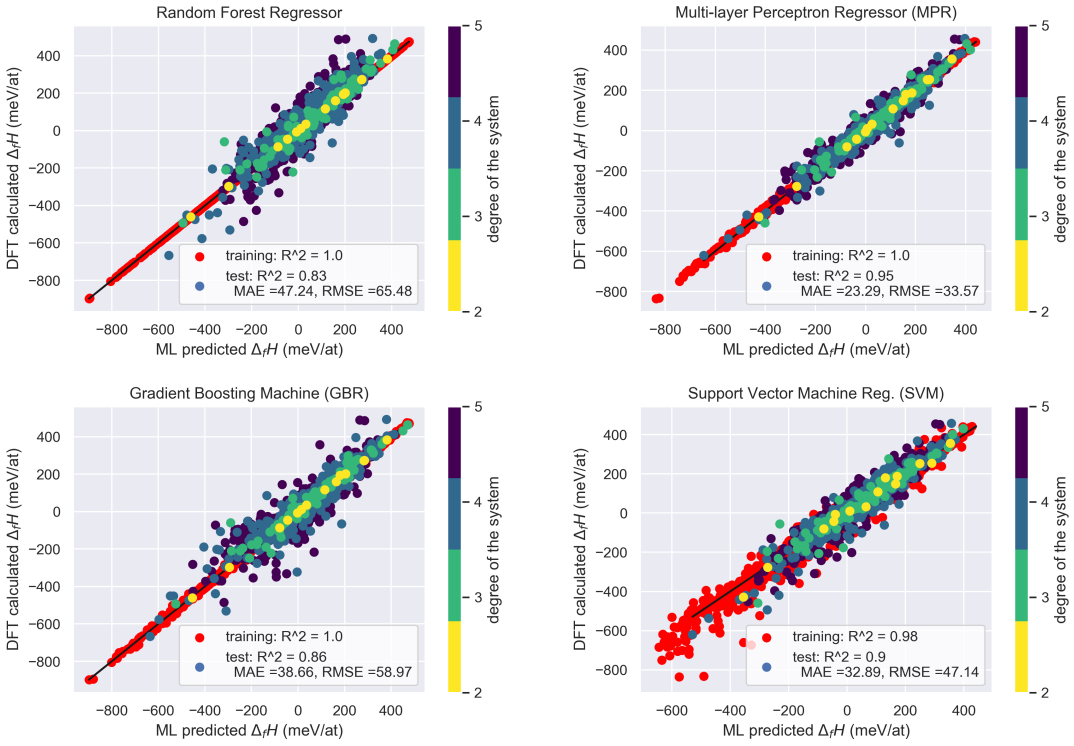

Figure 5: Prediction of randomized 1001 configurations among the 537,824 ones from the learning of the training database (9974 data in red). The tested 1001 configurations are reported in colors (blue to yellow) corresponding to the degree d of their system (right side legend). The diagonal line indicates the perfect agreement between DFT calculated and ML predicted values.

4. Discussion

4.1. Influence of the system degree

For the best method, here the regressor-type neural network MPR, the accuracy of prediction depends on the degree of the system of the tested configuration, illustrated by the color code of Figure 5. As an example, the MAE$\sim$23 meV for the whole testing set could be decomposed as contribution depending of the system degree d. It increases with d: 7 meV ($d = 2$), 22 ($d = 3$), 28 ($d = 4$) and 38 ($d = 5$). This result illustrates obviously that multi-component systems are more difficult to predict.

Another question could be addressed to the learning weight of binaries: are the whole $d = 1$ (14 elements) and $d = 2$ systems (here the 91 different subsystems $\times$30 configurations) are sufficient to predict higher degree systems? In other words, is it possible to predict accurately the whole possible 14^5 combinations only from all unary and binary configurations (2744 unique data), which is representative of only 0.5% of the total set? In order to answer, we merged our training and testing sets leading to 10941 unique configurations and split them in the 5 sub-systems: 14 "$d = 1$" , 2730 "$d = 2$", 5051 "$d = 3$", 2571 "$d = 4$" and 575 "$d = 5$" configurations. Then, from the all unique unary and binary configurations, we have tested the predictive behaviour for higher degree systems as shown in Figure 6. Whereas the ternary and quaternary systems are well predicted with MAE$\sim$18 and 22 meV respectively, the quinaries still present surprisingly reasonable results with MAE$\sim$34 meV respectively. However, from higher to lower system degrees, a learning from a portion of the ternary configurations (5051 among the 54,600 possible) gives larger dispersion on prediction on the binaries with higher MAE$\sim$40 meV and RMSE$\sim$73 meV. Other combinations of training/testing subsystems are illustrated in SM-H.

4.2. Contribution of additional descriptors

Introduction of additional physical descriptors improves slightly the learning scores. As shown in SM-I, the prediction with neither atomic radius nor the

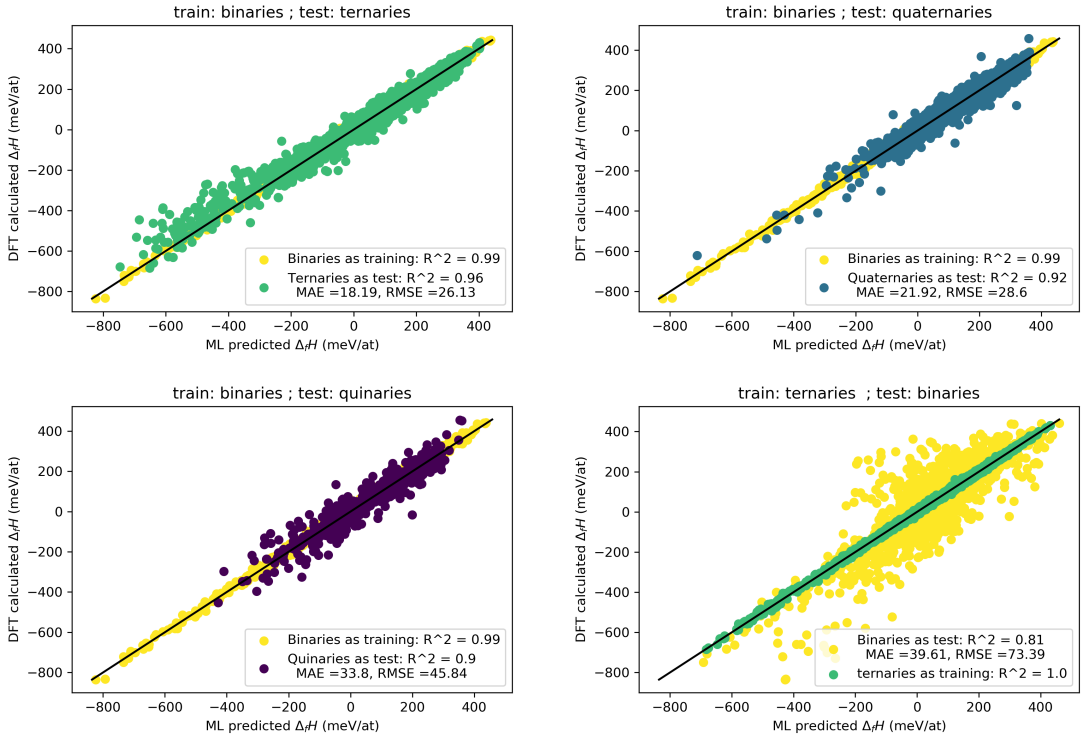

Figure 6: From training on only binaries, prediction of ternaries, quaternaries or quinaries. The diagonal line indicates the perfect agreement between predicted and real values.

number of valence electron is slightly worse (MAE∼24 with MPR). This result might seem unexpected at first glance. In fact, it is well known that topologically close packed (TCP) structures, as $\sigma-$phase, are driven by geometric arguments: since atoms are in the center of a coordination sphere, the atomic radius reflects the capability to occupy small or large coordination number (CN) sphere. The number of valence electron is also known to be important. In fact, for similar radius, a study has shown that the degeneracies of electronic levels play a role on the site preference [32]. To summarize this point, it seems that the total energy calculated by DFT contains these additional properties, and there is no need to provide it as separate descriptors, especially while using a versatile deep learning approach like the MPR.

4.3. Crystallographic prediction

The $\Delta_f H$ is only one predictive variable among many describing the crystal structure of the σ phase. Considering the only crystal properties, 9 variables are necessary to describe a configuration: 2 cell parameters (a, c) and 7 internal parameters (x^{4f}, x^{8i_1}, y^{8i_1}, x^{8i_2}, y^{8i_2}, x^{8j}, z^{8j}). In the present work, the supervised learning is optimized for predicting the $\Delta_f H$ but was also applied for every other variables. From our best model (optimized MPR) and from all available learning sets (9974 + 1001 = 10975 data), the predictive $\Delta_f H$ and the 9 other crystal variables are given for the every 537,824 configurations in SM-J. As an example, the prediction of both a & c tetragonal cell parameters presents a MAE $\sim$ 0.06 & 0.07 Å and a RMSE $\sim$ 0.08 & 0.10 respectively (Figures 7).

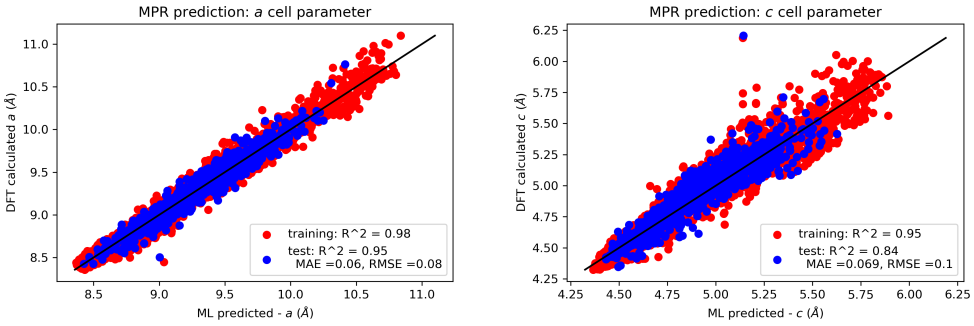

Figure 7: Prediction of both a and c tetragonal cell parameters of randomized 1001 testing configurations from the MPR learning of the training database (9974 data in red).

The prediction of crystal properties by ML is also very useful to help the initialization of new DFT input files and thus consequently reduces CPU times. In fact, the crystal properties of the 1001 random configurations, constituting the testing set, have been predicted from the learning database and were used as starting structures of the DFT relaxation steps. Since the initial state was close to the equilibrium structure for each configuration, we save about $\sim$10 times of the CPU consumption for the only DFT relaxation steps. This kind of ML approach can be thus conducted with a reinforcement learning to build a final database with a considerable reduction of the CPU time for DFT calculations.

4.4. Outlooks

Our ML approach, based on one-hot encoding from the only binary configurations as a learning set, is efficient in the present example of the $\sigma-$phase which is a complex TCP structure with 5 different Wyckoff positions. This methodology will be extended to other both simpler and more complex intermetallics, such as $C14$, χ or $\mu-$phases, to investigate the performance of the algorithms in additional future works (*e.g.* learning rate as a function of database size) in order to solve the issue of multicomponent $\Delta_f H$ estimation for thermodynamics database. As an outlook, the learning database size will be increased up to twenty elements, including Ta and Si. Moreover, the magnetism will be considered in future.

5. Conclusion

This work addresses the issue of the crystal phase stability from the machine learning viewpoint. Because the $\Delta_f H$ is the key descriptor to model the formation of compounds, we have investigated the prediction of this variable using a supervised approach, using a complex crystallographic structure as an example: the σ-phase. Based on an unprecedented large first principled dataset containing about 10,000 compounds with $n = 14$ different elements, we optimized several supervised learning approaches, where the Multi-layers Perceptron Regressor shows best results to predict all the ∼500,000 possible configurations within a mean absolute error of 23 meV (∼2 kJ.mol^{-1}) on the testing set. Additional descriptors with roots in the physical nature of the problem are minor contribution to the learning score in comparison with the only combinatorial DFT set. It is shown that the training database from the only binary-compositions (0.5% occurrence of whole set) are able to predict multicomponent configurations with a high accuracy. This result suggests that several complex phases including non-equivalent sites could be easily determined from the only binary contribution.

In addition to the heat of formation, the prediction of the lattice parameters and internal degrees of freedom seems to be very useful for reducing effort in the DFT calculations.

This work will be extended to other complex TCP phases with more than 2 sites to demonstrate the efficiency of our approach, such as $A12$, $C14$, *etc.* Indeed, it opens broad avenues in the study of complex structures with the only binary configurations as a learning set, this could be efficient even with low number of data.

Data and code availability

All data produced in this study (DFT results, ML code and properties prediction of every 14^5 configurations) are available on the code sharing platform at `https://github.com/crivello-jc/sigma-phase-prediction`.

Acknowledgements

Calculations were performed using HPC resources from GENCI-CINES (No. A0060906175) and supercomputer at IMR, Tohoku University (No. 16S0403). In addition, we acknowledge the financial support from the CNRS (programs MaLeFHYCe, PEPS, Cellule Energie CNRS and MALEpHYq, Emergence@INC).

Competing Interests

The authors declare no competing interests

Supplementary materials

The supplementary materials are available for this paper in several Appendix:

SM-A Crystal details of the σ-phase

SM-B Analysis of the combinatorial descriptions

SM-C DFT calculation methodology

SM-D The training database, 9974 configurations

SM-E The testing set, 1001 configurations

SM-F Machine learning methods

SM-G Cross validation results from the only training database

SM-H Prediction from several simulations of training and testing sets

SM-I Influence of additional physical featuring

SM-J Results of prediction of every 14^5 configurations

[1] S. H. Lee, "Natural language generation for electronic health records," *npj Digital Medicine*, vol. 1, p. 63, Nov. 2018.

[2] M. H. S. Segler, T. Kogej, C. Tyrchan, and M. P. Waller, "Generating focused molecule libraries for drug discovery with recurrent neural networks," *ACS Central Science*, vol. 4, no. 1, pp. 120–131, 2018. PMID: 29392184.

[3] P. Ruamviboonsuk, J. Krause, P. Chotcomwongse, R. Sayres, R. Raman, K. Widner, B. J. L. Campana, S. Phene, K. Hemarat, M. Tadarati, S. Silpa-Archa, J. Limwattanayingyong, C. Rao, O. Kuruvilla, J. Jung, J. Tan, S. Orprayoon, C. Kangwanwongpaisan, R. Sukumalpaiboon, C. Luengchaichawang, J. Fuangkaew, P. Kongsap, L. Chualinpha, S. Saree, S. Kawinpanitan, K. Mitvongsa, S. Lawanasakol, C. Thepchatri, L. Wongpichedchai, G. S. Corrado, L. Peng, and D. R. Webster, "Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program," *npj Digital Medicine*, vol. 2, no. 1, pp. 25–, 2019.

[4] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," *npj Computational Materials*, vol. 5, no. 1, pp. 83–, 2019.

[5] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain, "Unsupervised word embeddings capture latent knowledge from materials science literature," *Nature*, vol. 571, pp. 95–98, July 2019.

[6] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," *npj Computational Materials*, vol. 2, no. 1, pp. 16028–, 2016.

[7] Z. Zhou, Y. Zhou, Q. He, Z. Ding, F. Li, and Y. Yang, "Machine learning guided appraisal and exploration of phase design for high entropy alloys," *npj Computational Materials*, vol. 5, no. 1, pp. 128–, 2019.

[8] T. Kostiuchenko, F. Krmann, J. Neugebauer, and A. Shapeev, "Impact of lattice relaxations on phase transitions in a high-entropy alloy studied by machine-learning potentials," *npj Computational Materials*, vol. 5, no. 1, pp. 55–, 2019.

[9] Z. Pei, J. Yin, J. A. Hawk, D. E. Alman, and M. C. Gao, "Machine-learning informed prediction of high-entropy solid solution formation: Beyond the hume-rothery rules," *npj Computational Materials*, vol. 6, no. 1, pp. 50–, 2020.

[10] J. Peng, Y. Yamamoto, J. A. Hawk, E. Lara-Curzio, and D. Shin, "Coupling physics in machine learning to predict properties of high-temperatures alloys," *npj Computational Materials*, vol. 6, no. 1, pp. 141–, 2020.

[11] S. K. Kauwe, J. Graser, A. Vazquez, and T. D. Sparks, "Machine learning prediction of heat capacity for solid inorganics," *Integrating Materials and Manufacturing Innovation*, vol. 7, no. 2, pp. 43–51, 2018.

[12] G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, and T. Lookman, "Machine learning bandgaps of double perovskites," *Scientific Reports*, vol. 6, no. 1, pp. 19375–, 2016.

[13] V. Gladkikh, D. Y. Kim, A. Hajibabaei, A. Jana, C. W. Myung, and K. S. Kim, "Machine learning for predicting the band gaps of abx3 perovskites from elemental properties," *The Journal of Physical Chemistry C*, vol. 124, no. 16, pp. 8905–8918, 2020.

[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*, vol. 89, p. 094104, Mar 2014.

[15] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rhl, and C. Wolverton, "The open quantum materials database (oqmd):

assessing the accuracy of dft formation energies," *npj Computational Materials*, vol. 1, no. 1, p. 15010, 2015.

[16] S. Ubaru, A. Międlar, Y. Saad, and J. R. Chelikowsky, "Formation enthalpies for transition metal alloys using machine learning," *Phys. Rev. B*, vol. 95, p. 214102, Jun 2017.

[17] C. J. Bartel, A. Trewartha, Q. Wang, A. Dunn, A. Jain, and G. Ceder, "A critical examination of compound stability predictions from machine-learned formation energies," *npj Computational Materials*, vol. 6, no. 1, pp. 97–, 2020.

[18] C. Oses, C. Toher, and S. Curtarolo, "Data-driven design of inorganic materials with the automatic flow framework for materials discovery," *MRS Bulletin*, vol. 43, no. 9, p. 670675, 2018.

[19] C. Draxl and M. Scheffler, "Nomad: The fair concept for big data-driven materials science," *MRS Bulletin*, vol. 43, no. 9, p. 676682, 2018.

[20] 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*, vol. 93, p. 085142, Feb 2016.

[21] Y. Liu, T. Zhao, W. Ju, and S. Shi, "Materials discovery and design using machine learning," *Journal of Materiomics*, vol. 3, no. 3, pp. 159 – 177, 2017. High-throughput Experimental and Modeling Research toward Advanced Batteries.

[22] Y. Kim, E. Kim, E. Antono, B. Meredig, and J. Ling, "Machine-learned metrics for predicting the likelihood of success in materials discovery," *npj Computational Materials*, vol. 6, no. 1, pp. 131–, 2020.

[23] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary:

The materials project: A materials genome approach to accelerating materials innovation," *APL Materials*, vol. 1, no. 1, p. 011002, 2013.

[24] A. Miedema, R. Boom, and F. De Boer, "On the heat of formation of solid alloys," *J. Less-Common Met.*, vol. 41, pp. 283–298, July 1975.

[25] L. Kaufman and H. Bernstein, *Computer Calculation of Phase Diagrams with Special Reference to Refractory Metals*. Notes and Reports in Computer Science and Applied Mathematic, Academic Press, 1970.

[26] B. Sundman and J. Ågren, "A regular solution model for phases with several components and sublattices, suitable for computer applications," *J. Phys. Chem. Solids*, vol. 42, pp. 297–301, 1981.

[27] J.-M. Joubert, "Crystal chemistry and calphad modeling of the σ phase," *Prog. Mater. Sci.*, vol. 53, pp. 528–583, 2008.

[28] R. Mathieu, N. Dupin, J.-C. Crivello, K. Yaqoob, A. Breidi, J.-M. Fiorani, N. David, and J.-M. Joubert, "CALPHAD description of the Mo–Re system focused on the sigma phase modeling," *Calphad*, vol. 43, no. 0, pp. 18–31, 2013.

[29] N. Dupin, U. Kattner, B. Sundman, M. Palumbo, and S. Fries, "Implementation of an effective bond energy formalism in the multicomponent calphad approach," *J Res Natl Inst Stan*, vol. 123, p. 123020, 2018.

[30] D. Jha, L. Ward, A. Paul, W.-k. Liao, A. Choudhary, C. Wolverton, and A. Agrawal, "Elemnet: Deep learning the chemistry of materials from only elemental composition," *Scientific Reports*, vol. 8, no. 1, pp. 17593–, 2018.

[31] Z. Zhang, M. Li, K. Flores, and R. Mishra, "Machine learning formation enthalpies of intermetallics," *Journal of Applied Physics*, vol. 128, no. 10, p. 105103, 2020.

[32] M. H. F. Sluiter and A. Pasturel, "Site occupation in the Cr-Ru and Cr-Os σ phases," *Phys. Rev. B*, vol. 80, p. 134122, Oct 2009.

[33] C. Berne, M. H. F. Sluiter, Y. Kawazoe, T. Hansen, and A. Pasturel, "Site occupancy in the Re-W sigma phase," *Phys. Rev. B*, vol. 64, p. 144103, 2001.

[34] P. Korzhavyi, B. Sundman, M. Selleby, and B. Johansson, "Atomic, electronic, and magnetic structure of iron-based sigma-phases," *Mater. Res. Soc. Symp. Proc.*, vol. 842, p. S4.10.1, 2005.

[35] J. Cieslak, J. Tobola, and S. M. Dubiel, "Electronic structure of the σ phase of paramagnetic Fe-V alloys," *Phys. Rev. B*, vol. 81, p. 174203, May 2010.

[36] M. Palumbo, T. Abe, C. Kocer, H. Murakami, and H. Onodera, "Ab initio and thermodynamic study of the Cr-Re system," *Calphad*, vol. 34, pp. 495–503, Dec. 2010.

[37] K. Chvátalová, J. Houserová, M. Šob, and J. Vřeštál, "First-principles calculations of energetics of sigma phase formation and thermodynamic modelling in Fe-Ni-Cr system," *J. Alloys Comp.*, vol. 378, pp. 71–74, Sept. 2004.

[38] J.-C. Crivello, M. Palumbo, T. Abe, and J.-M. Joubert, "*Ab initio* ternary $\sigma-$phase diagram: the Cr-Mo-Re system," *Calphad*, vol. 34, pp. 487–494, 2010.

[39] M. Palumbo, T. Abe, S. G. Fries, and A. Pasturel, "First-principles approach to phase stability for a ternary sigma phase: Application to Cr-Ni-Re," *Phys. Rev. B*, vol. 83, p. 144109, Apr. 2011.

[40] K. Yaqoob, J.-C. Crivello, and J.-M. Joubert, "Study of site occupancies in Mo-Ni-Re sigma-phase by both experimental and ab initio methods," *Inorg. Chem.*, vol. 51, pp. 3071–3078, 2012.

[41] J.-C. Crivello, R. Souques, A. Breidi, N. Bourgeois, and J.-M. Joubert, "Zengen, a tool to generate ordered configurations for systematic first-principles calculations: The Cr–Mo–Ni–Re system as a case study," *Calphad*, vol. 51, pp. 233 – 240, 2015.