Accelerating the discovery of crystalline materials with desired intrinsic properties by machine learning

Zur Erlangung des Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.)
Genehmigte Dissertation von Teng Long aus Shandong, China
Tag der Einreichung: 27.07.2021, Tag der Prüfung: 11.11.2021

1. Gutachten: Prof. Dr. Hongbin Zhang
2. Gutachten: Prof. Dr. Baixiang Xu
Darmstadt

Materials and Earth Sciences Department
Institute of Materials Science
Theorie magnetischer Materialien

Accelerating the discovery of crystalline materials with desired intrinsic properties by machine learning

Accepted doctoral thesis by Teng Long

1. Review: Prof. Dr. Hongbin Zhang
2. Review: Prof. Dr. Baixiang Xu

Date of submission: 27.07.2021
Date of thesis defense: 11.11.2021

Darmstadt

Dedicate to my family and beloved friends

Erklärungen laut Promotionsordnung

§8 Abs. 1 lit. c PromO

Ich versichere hiermit, dass die elektronische Version meiner Dissertation mit der schriftlichen Version übereinstimmt.

§8 Abs. 1 lit. d PromO

Ich versichere hiermit, dass zu einem vorherigen Zeitpunkt noch keine Promotion versucht wurde. In diesem Fall sind nähere Angaben über Zeitpunkt, Hochschule, Dissertationsthema und Ergebnis dieses Versuchs mitzuteilen.

§9 Abs. 1 PromO

Ich versichere hiermit, dass die vorliegende Dissertation selbstständig und nur unter Verwendung der angegebenen Quellen verfasst wurde.

§9 Abs. 2 PromO

Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient.

Darmstadt, 27.07.2021

T. Long

Contents

Acknowledgement

I would first like to express my deep thanks to Prof. Dr. Hongbin Zhang, who gave me the opportunity to work in his group. And I am grateful for his supervision, productive discussions and encouragement, which improved my work and scientific attitude.

I would also like to thank Ms. Maria Walker, the administrator of our group, for her kind help in everyday life.

Many thanks to the meaningful discussions and constant support from my dear colleagues: Nuno Miguel dos Santos Fortunato, Yixuan Zhang, Ilias Samathrakis, Niloofar Hadaeghi, Harish Kumar Singh, Chen Shen, Yi Xiao, Mian Dai, Dr. Ruiwen Xie, Dr. Fei Liang, Dr. Ingo Opahle, Dr. Qiang Gao, Dr. Xinru Li, Dr. Tingting Lin, Dr. Lei Xu, Dr. Ye Han, Yaqian Guo, Ling Fan and so on.

Special thanks to Prof. Baixiang Xu for her willingness to serve as Co-referee. I would also like to thank Prof. Oliver Gutfleisch, Prof. Karsten Albe and Prof. Dierk Raabe for their kind suggestions to my work.

I would like to acknowledge the financial support from Chinese Scholarship Council (CSC) and the support by the Deutsche Forschungsgemeinschaft (DFG – German Research Foundation) and the Open Access Publishing Fund of Technical University of Darmstadt. I gratefully acknowledge computational time on the Lichtenberg High Performance Supercomputer.

Last but not least, I would like to thank my family and friends for their love and support.

Abstract

As an emergent research paradigm, data-driven methods (*e.g.*, machine learning) have recently been applied extensively to materials science, which provide valuable solutions to map out the process-structure-property relationships, thus enabling autonomous materials designs. In this thesis, focusing on the mapping between crystal structures and intrinsic physical properties, both forward modelling (to predict physical properties with crystal structures as input) and inverse design (to predict novel crystal structures with desired properties) have been performed, accelerating the design of crystalline materials with desired properties.

For the forward modelling, Curie temperature of 1749 ferromagnetic materials was collected to carry out machine learning modelling based on the two-step random forest method. The resulting accuracy is about 91% for evaluating the Curie temperature, which has been further validated by 85 experimental results. In this regard, it provides a practical solution to accelerate designing functional ferromagnetic materials, as the Curie temperature is one of the three key intrinsic magnetic properties (in addition to magnetization and magnetic anisotropy energy). Furthermore, in collaboration with Yixuan Zhang, we demonstrated that both the total energies and forces on atoms could be modelled accurately, leading to a reliable construction of machine-learning interatomic potentials for further atomistic simulations like molecular dynamics. Therefore, the forward modelling can be applied to predict the intrinsic physical properties and to bridge quantitative simulations across the electronic and atomistic length scales.

In terms of inverse design, constrained crystal deep convolutional generative adversarial networks (CCDCGAN) have been developed, directly predicting crystal structures distinct from the known cases based on the image-based continuous representation (of the crystal structures) forming a latent space. Moreover, the intrinsic properties of generated structures can be optimized in the latent space through a back propagator (applied on the pre-trained forward model) and an appropriately defined loss function (of CCDCGAN). In this thesis, formation energy has been integrated into CCDCGAN as a forward model, and correspondingly CCDCGAN can design stable crystal structures. It has been successfully applied on a binary system (Bi-Se) and multicomponent systems (binary, ternary, and quaternary systems). It is observed that unreported crystal structures below the con-

vex hull defined by the known experimental cases can be obtained. Interestingly, it is suspected that the CCDCGAN model can be generalized to multi-objective optimization when forward models on different properties are applied, leading to a systematic way of designing novel crystalline materials.

In the future, inverse design of crystalline materials with multi-objective optimization of various physical properties can be realized based on our current implementation of CCDCGAN, *e.g.*, by incorporating mechanical and other magnetic properties to design magnetic materials with optimal performance. Additionally, such image-based deep learning algorithms can be straightforwardly generalized to model the microstructures, which will provide a systematic method to inverse design the microstructures and hence a decent solution to map out the structure-property relationship.

Zusammenfassung

Als aufstrebendes Forschungsparadigma wurden datengesteuerte Methoden (z. B. maschinelles Lernen) in letzter Zeit ausgiebig in der Materialwissenschaft angewandt. Sie bieten wertvolle Lösungen, um die Prozess-Struktur-Eigenschafts-Beziehungen abzubilden, und ermöglichen daher autonomes Materialdesign. In dieser Arbeit, die sich auf die Abbildung zwischen Kristallstrukturen und intrinsischen physikalischen Eigenschaften konzentriert, wurden sowohl Vorwärtsmodellierung (zur Vorhersage physikalischer Eigenschaften mit Kristallstrukturen als Input) als auch inverses Design (zur Vorhersage neuartiger Kristallstrukturen mit gewünschten Eigenschaften) durchgeführt, was das Design kristalliner Materialien mit gewünschten Eigenschaften beschleunigt.

Zur Vorwärtsmodellierung wurde die Messdaten der Curie-Temperaturen von 1749 ferromagnetischen Materialien gesammelt, um eine maschinelle Lernmodellierung, die auf der zweistufigen Random-Forest-Methode basiert, durchzuführen. Die sich daraus ergebende Genauigkeit beträgt ca. 91% zur Evaluierung der Curie-Temperaturen, die durch 85 experimentelle Ergebnisse weiter validiert wurde. In dieser Hinsicht wird eine praktische Lösung geboten, die die Entwicklung funktionaler ferromagnetischer Materialien beschleunigt, weil die Curie-Temperatur eine der drei wichtigsten intrinsischen magnetischen Eigenschaften ist (neben der Magnetisierung und der magnetischen Anisotropieenergie). Darüber hinaus haben wir in Zusammenarbeit mit Yixuan Zhang gezeigt, dass sowohl die Gesamtenergien als auch die Kräfte auf die Atome genau modelliert werden können, was zu einer zuverlässigen Konstruktion von maschinenlernenden interatomaren Potentialen für weitere atomistische Simulationen wie Molekulardynamik führt. Daher kann die Vorwärtsmodellierung zur Vorhersage der intrinsischen physikalischen Eigenschaften und zur Überbrückung quantitativer Simulationen über die elektronischen und atomistischen Längenskalen eingesetzt werden.

Zum Zwecke des inversen Designs wurden Constrained Crystal Deep Convolutional Generative Adversarial Networks (CCDCGAN) entwickelt, die direkt Kristallstrukturen prognostizieren, die sich von den auf die bildbasierte kontinuierlichen Darstellung (der Kristallstrukturen) basierenden bekannten Fällen unterscheiden. Diese Darstellung bildet einen latenten Raum. Darüber hinaus können die intrinsischen Eigenschaften der generierten Strukturen im latenten Raum durch einen Rückpropagator (im vortrainier-

ten Vorwärtsmodell angewandt) und eine entsprechend definierte Verlustfunktion (des CCDCGAN) optimiert werden. In dieser Arbeit fand die Formationsenergie in CCDCGAN als Vorwärtsmodell die Anwendung, und dementsprechend kann CCDCGAN stabile Kristallstrukturen entwerfen. Es wurde erfolgreich auf ein binäres System (Bi-Se) und Mehrkomponentensysteme (binäre, ternäre und quaternäre Systeme) angewendet. Es wird beobachtet, dass bisher unbekannte Kristallstrukturen unterhalb der durch die bekannten experimentellen Fälle definierten konvexen Hülle gewonnen werden können. Interessanterweise wird vermutet, dass das CCDCGAN-Modell für die Mehrzieloptimierung verallgemeinert werden kann, wenn Vorwärtsmodelle in Bezug auf verschiedene Eigenschaften angewendet werden, was zu einem systematischen Design neuartiger kristalliner Materialien führt.

In der Zukunft kann das inverse Design von kristallinen Materialien mit multikriterieller Optimierung verschiedener physikalischer Eigenschaften auf der Grundlage unserer aktuellen Implementierung von CCDCGAN realisiert werden, indem mechanische und andere magnetische Eigenschaften einbezogen werden, um magnetische Materialien mit optimaler Leistung zu entwerfen. Zusätzlich können solche bildbasierten Deep-Learning-Algorithmen geradeheraus verallgemeinert werden, um die Mikrostrukturen zu modellieren, was eine systematische Methode zum inversen Design der Mikrostrukturen und somit eine angemessene Lösung zur Abbildung der Struktur-Eigenschafts-Beziehung zur Verfügung stellt.

Abbreviations

CCDCGAN	Constrained crystal deep convolutional generative adversarial networks
PSPP	Processing–structure–property–performance
SEM	Scanning electron microscope
TEM	Transmission electron microscopy
APT	Atom probe tomography
MGI	Materials Genome Initiative
ML	Machine learning
T_C	Curie temperature
MD	Molecular dynamics
DFT	Density functional theory
GA	Genetic algorithm
SVM	Support vector machine
RF	Random forest
BO	Bayesian optimization
ANN	Artificial neural network
CNN	Convolutional neural network
RNN	Recurrent neural network
AP	Affinity propagation
PCA	Principle component analysis
t-SNE	t-distributed stochastic neighbor embedding
GAN	Generative adversarial network
VAE	Variational autoencoder
ICSD	Inorganic Crystal Structure Database
UHCSDB	UltraHigh Carbon Steel Micrograph DataBase
ASM	ASM Alloy Phase Diagram Database
ICDD	International Centre for Diffraction Data
MP	Materials Projects
OQMD	Open Quantum Materials Database
AFLOW	Automatic FLOW for Materials Discovery databases

NOMAD	Novel Materials Discovery
2D database	2D Materials Encyclopedia
COD	Crystallography Open Database
QM9	Quantum machine 9
HTEM	High Throughput Experimental Materials Database
FAIR principles	Findability, accessibility, interoperability, and reusability
NFDI	National Research Data Infrastructure
BCC	Body-centered cubic
FCC	Face-centered cubic
SOAP	Smooth overlap of atomic position
ACSF	Atom centered symmetry functions
MAGPIE	Materials-agnostic platform for informatics and exploration
CGCNN	Crystal graph convolutional neural network
NLP	Natural language processing
MAE	Mean absolute error
MSE	Mean square error
GD	Gradient descent
SGD	Stochastic gradient descent
ADAM	Adaptive moment estimation
R^2	Coefficient of determination
GAP	Gaussian approximation potential
h-BN	Hexagonal boron nitride
3D	Three-dimensional
NN	Neural network
ReLU	Rectified linear unit
Leaky ReLU	Leaky rectified linear unit
DMFT	Dynamical mean-field theory

List of Figures

List of Tables

1. Introduction

1.1. Motivation

Materials play an essential role in our society. The stages of our history, such as stone ages, bronze ages, iron ages, and silicon ages [1], have been labeled by the prototypical materials. The discovery and exploitation of such materials have an apparent spillover effect on both technology and industry, thus benefiting the welfare of humankind [2]. Currently, two of apparent challenges for our society are the shortage of energy and the explosion of information. The former refers to the surge in energy demand, while the latter indicates the rapid increase in the amount of data and the consequent difficulties in data storage and management [3]. Such challenges entail the development of more efficient materials in harvesting energy from natural resources [4] or materials storing more data with less mass [5], respectively. Of course, many other types of materials are also needed, such as materials in pharmaceuticals [6], sensors [7], communication of data [8], and visual displays [9].

Therefore, there is a strong impetus to develop sustainable materials with optimized properties. The exponential growth of the global population and the transformation of lifestyles in developing countries have triggered a scarcity of resources [10]. However, the currently used materials are mainly produced by depleting non-recyclable resources, which cannot be sustainable for the future [11]. The above two reasons advocate for the development of new materials with properly implemented circularity [12], *e.g.*, the replacement of polyethylene by polylactic acid [13]. Such demands can be achieved by but not limited to the following three strategies: 1) reduce the cost of developing new materials; 2) restrict the resources extracted from the earth system; 3) limit the emissions back to earth [14, 15]. The latter two are challenges faced by both industry and materials science, while the eco-design strategy is predominantly dependent on materials science.

From the scientific point of view, the fundamental problem for materials science is mapping out the processing–structure–property–performance (PSPP) relationships [16] as demonstrated in Figure 1.1. Such a PSPP mapping involves the following three aspects:

- To explore the connection between processes and structures, *e.g.*, to stabilize the de-

sired phases (*i.e.*, compositions and crystal structures) or to tailor the microstructures by manipulating the processing [17], where microstructures can be characterized experimentally by scanning electron microscope (SEM) [18], transmission electron microscopy (TEM) [19], and atom probe tomography (APT) [20], etc.

- To establish the relationship between the properties and structures of materials. Note that structures include both the crystal structures and microstructures, correspondingly the properties contain both the intrinsic and extrinsic properties, respectively.
- To connect the performance of materials with their properties.

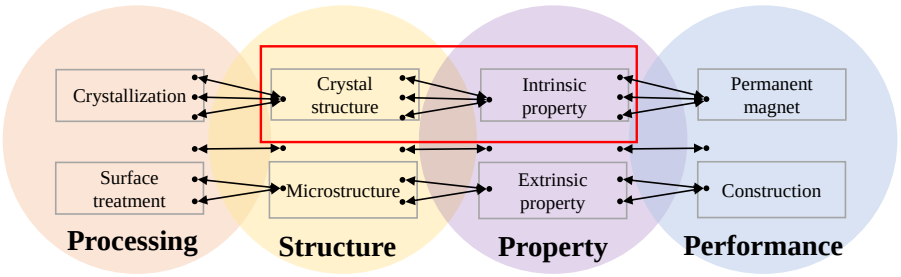


Figure 1.1.: The PSPP relationships of materials science. On the one hand, multiple processings, structures, and properties can determine the same structure, property, and performance. On the other hand, the same structure, property, and performance can be reached by multiple processings, structures, and properties. It is worth noting that properties can be divided into intrinsic and extrinsic properties, and the structures are categorized as crystal structures and microstructures [21].

However, there are still no established systematic solutions to explore the entire PSPP mapping quantitatively. One of the main reasons is that the PSPP mapping dictates detailed investigations across several length scales, *i.e.*, ranging from the electronic scale to the device scale. Although the underlying theoretical principles for an individual length scale can be defined, the main challenge is how to connect them so that comprehensively quantitative predictions can be delivered [22]. Another reason is that the PSPP mapping is a many-to-one mapping [21], *e.g.*, various processing routes, multiple structures and different properties can potentially lead to the same structure, property and performance, respectively.

Notwithstanding such challenges, one of the most famous attempts in the exploration of PSPP mapping is the Materials Genome Initiative (MGI) proposed by the US government in

2011 [23], aiming at amalgamating theories, computations, and experiments to accelerate the discovery, design, and deployment of new materials [24]. MGI relies heavily on collecting digitized data by experiments and high-throughput computations, which has three crucial features [25]: a vast number of escalating data, efficient utilization of computational resources, and interdisciplinary integration. For example, self-assembly polymers [26], polar metals [27], and organic light-emitting diodes [28] are all developed with the support of MGI. However, the increasing amount of data is tremendous pressure in data analysis. [24]. For instance, over 1.6 million molecules are checked in search of new organic light-emitting diodes [28], which is beyond the capability of analysis by human beings.

Thus, machine learning (ML) techniques, which are subordinate to statistics, are utilized to extract actionable insights from big data and conduct materials design at various levels [29]. For instance, the expeditious estimation of population parameters (predictivity of the multiscale model) and uncertainty quantification to access the reliability are inherent in ML, leading to pivotal impacts on identifying the PSPP relationships [30]. As a result, objective-based experimental design using the Bayesian optimization (BO) [31], Curie temperature ($\mathrm{T_C}$) prediction for ferromagnetic materials through only chemical composition [32], as well as the automated generation of porous materials [33] can be achieved.

Notably, the ML approach 1) can help decipher the PSPP relationships via forward models, which in turn 2) can discover novel materials through inverse design. On the one hand, forward models predict the structure of a material based on the processing used [34], properties contingent on the structure [35], and performance determined by properties [36], which benefit the materials discovery approaches [37, 29]. On the other hand, inverse design aims to determine the parameters of processing and structures that would yield materials with desired structures and properties. For example, expediting the determination of process paths to realize specific (micro)structures [34], and design of crystal structures with certain properties [38]. Thus, it is of great importance to apply ML techniques in materials science.

In this thesis, how to accelerate the discovery of new crystalline materials by ML from the two perspectives mentioned above will be discussed, *i.e.*, the forward prediction of intrinsic properties by crystal structures and the inverse design of crystal structures with desired intrinsic properties. For the following sections of this chapter, firstly, a brief review of four paradigms in materials science will be given in Section 1.2 to elucidate the potential solutions to the current challenges. Then, the choice of ML techniques in materials discover will be justified in Section 1.3 by rummaging through the details of these methods. In the end, a short summary will be presented in Section 1.4.

1.2. Four paradigms in materials design

Materials science research can be classified into four paradigms, as sketched in Figure 1.2. To be more specific, the paradigms are the first paradigm based on experiments, the second supported by classical theories, the third contingent on computational quantum theories, and the fourth driven by big data [21].

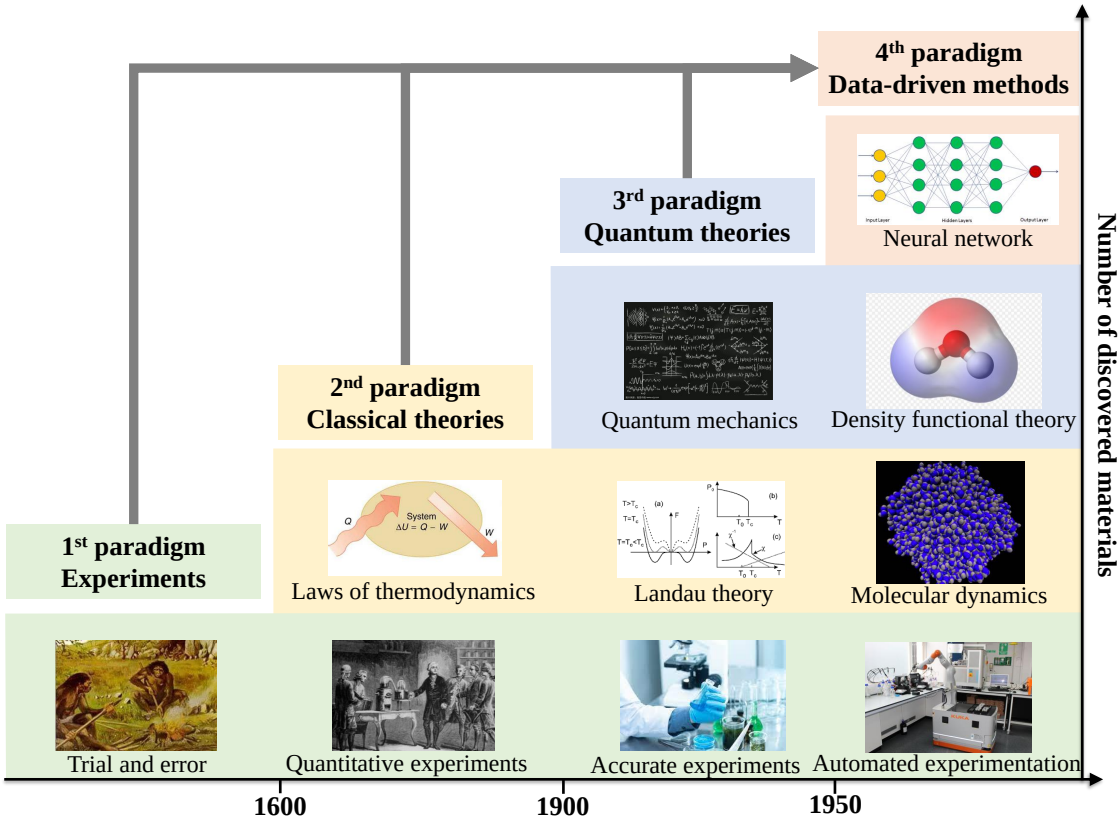


Figure 1.2.: Four paradigms in materials design. The first paradigm is the approach to synthesize new materials by experiments, while the second paradigm is the calculations and simulations supported by the parameterized mathematical formulas (classical theories). Similarly, the third paradigm enables the identification of new materials by simulations without prior determined parameters (quantum theories). Finally, the fourth paradigm is driven by big data, *e.g.*, materials discovery by machine learning.

Conducting experiments in a trial and error way functions as the first paradigm of materials design [21]. It means the repeatable tests of processing and logical analysis

of the results, aiming to synthesize novel materials [39]. For example, steel was first discovered as a byproduct in the smelting of bronze [40], and its production was gradually mastered by human beings over 800 years. As a result, the application of steels improves the bronze farming tools in hardness and endurance, which brought advancements in the agricultural production [41]. Hitherto, such a paradigm has been developed into mature techniques requiring domain expertise of experimentalists. Thus, it becomes the most fundamental way to study materials science due to its direct relationship with applications. Nevertheless, the discovery of materials by trial and error is sluggish and costly, attributed to its requirement of a considerable amount of time and a large number of raw materials.

The modelling based on classical theories became the second paradigm of materials science marked by the publication of *Mathematical Principles of Natural Philosophy* [42] in 1687. One of the essential features of these theories is to parameterize the experimental phenomena into mathematical formulas, which can be used to predict the properties of materials under specific conditions. For example, the laws of thermodynamics explain the relationship between free energy and temperature or pressure, allowing the prediction of phase transitions of alloys and the establishment of phase diagrams [43]. Later the Landau theory expresses the free energy as an expansion of the order parameter, leading to the explanation of phase transitions such as the magnetic phase transition [44]. With the help of computers, molecular dynamics (MD) simulations were also developed. They treat the atoms as classical particles and determine interactions between particles by empirical potentials, *e.g.*, the Lennard-Jones potential [45]. As a result, the PSPP mapping can be calculated quantitatively, which improves the understanding of the behaviour of materials. In addition, the experiments were also promoted into the quantitatively controlled level, which then enabled massive manufacture of discovered materials, *e.g.*, the flourishing of the metallurgical industry before the 1950s.

The materials discovery is further enriched by calculations based on the quantum theory. It was introduced by Max Planck in 1900 [46], which describes the behaviour of particles (electrons, protons, and neutrons) by wave functions. Although wave functions for more than two particles cannot be analytically solved, accurate numerical approximations can be acquired, *e.g.*, the application of density functional theory (DFT) [47, 48]. Particularly, with the development of supercomputers, calculations of millions of materials have been conducted [49, 50], where the newly discovered materials can be further validated by experiments [51]. Furthermore, the quantum theories improve the first two paradigms. For instance, the ab initio MD methods calculate the interactions between atoms by DFT, which avoids the inaccuracy of empirical interatomic potentials [52]. The quantum theories can also conduct more accurate predictions of materials, which guide the experiments.

The data-driven methods have been developed as the fourth paradigm for materials design, based on materials databases established by the previous three paradigms. They

summarise the statistical features of known materials and design materials based on the learned features [53]. Data-driven approaches make the best usage of all existing data and therefore enable the automatic PSPP mapping in a more holistic manner [54]. For example, genetic algorithms (GA) and ML approaches [55] can directly predict crystal structures of new materials by "learning" the known structures. However, the method is data demanding, *e.g.*, a deep learning model requires samples of the order of magnitude of 10^3 to start training [56]. So it was not until the collection of enough materials that the data-driven methods have become practical [21]. Moreover, the data-driven paradigm enhances the performance of previous paradigms [16]. For example, the ML techniques facilitate the efficiency of DFT calculation [57], as the underlying Kohn-Sham equations are partial differential equations. Machine-learning interatomic potentials have been developed to balance the accuracy and efficiency of MD, *i.e.*, achieving higher accuracy than the classical potentials in MD simulations but costing less time than ab initio MD. In the meantime, the automated experimental robots contingent on artificial intelligence and ML can perform more experiments efficiently, which expedite the pace of discovering new materials by experimental methods [58].

1.3. Machine learning

ML is an algorithm to optimize its performance (P) over a specific task (T) spontaneously by capitalizing on experience (E) without explicit programming [59, 60], which can also be expressed as the following equation:

$$\max_{X \in E} P_T(X) \tag{1.1}$$

where X is one sample of the experience. Assuming the ML algorithm is a player of a video game, T is the objective of the game, *i.e.*, to win the game. P is the possibility for the player to win the game by analyzing the existing experience X from previous games [61]. The player then establishes the relationship between X and P and tries to maximize P by selecting proper X. With the exponential increase in the computational capability and agglomeration of information [62], ML techniques are successfully utilized in self-driving vehicles [63], image and speech recognition [64, 65], spam filtering [66], industrial robotic manipulator [67], etc. Undeterred by the low interpretability of the decision, ML has even demonstrated superiority over human beings in terms of performance, learning efficiency, and knowledge portability such as playing Go and Starcraft [61, 68].

The superiority is achieved at the expense of interpretability. Occam's razor theorem favors the simple solutions over the complex ones [69], indicating that complicated and

hierarchical relationships involving massive combinatorial spaces and nonlinear processes can be addressed with a relatively easy solution. This applies to the training of ML models: without the pressure to justify their action, the ML models automatically reduce the redundant parameters unrelated to the task and grasp the essential relationship between the performance and actions [70]. For example, the element symbols and atomic numbers are both meaningful to determine the electronic structures of an atom, but the ML algorithms can ignore the element symbols and only focus on the atomic numbers, which improves their efficiency.

However, the application of ML algorithms should keep a balance between causality and statistics. In other words, the sacrifice of interpretation for efficiency is not always worthwhile, as research in the natural sciences still needs to provide a plausible explanation for a phenomenon. Nevertheless, the ML models assist the understanding, which is indispensable for researchers. For example, the protein folding problem is a grand challenge for biology, which is recently partially solved by an ML algorithm called AlphaFold by learning 170,000 known protein structures [71]. Despite the ability to predict protein structures supposed to reveal the relationship between the shape of proteins and their functions, such valuable knowledge has not been transferred to human beings. This requires further efforts from researchers to gain such understanding. Similarly, the crop yield prediction and nitrogen status estimation in agriculture [72] and the forest management in geology [73] could not give an explicit explanation on the underlying mechanism. However, such relationships inspired insights for researchers to gain a better understanding of these subjects.

The ML techniques can be applied to materials science as well. As discussed above, the goal of materials science is the determination of PSPP mapping, *i.e.*, the deductive relationship of processing, structure, property, and performance, as illustrated in Figure 1.1. Notably, each connection in PSPP mapping is reversible. Thus, ML for materials science has two aspects: 1) the scientific prediction of structures, properties, and performance, based on the processing, structures, and properties of materials in a forward way; 2) designing materials with optimal performance on desired properties in the reverse direction, *i.e.*, inverse design [74]. Particularly, the ML-based inverse design methods actively search for materials with desired properties in latent space instead of selecting materials with desired properties from multiple materials, as demonstrated in Figure 1.3. For instance, Iyer et al. predicted the microstructure of silica nanoparticles dispersed in polymethylmethacrylate polymers with the processing parameters as inputs by ML models [34]. In the meantime, they used the acquisition function to select the processing parameters with desired dispersion. Xie et al. developed ML models to predict the formation energies, band gaps, and shear/bulk modulus based on crystal structures [35]. In addition, Noh et al. used continuous image-based representation generated by an autoencoder to directly

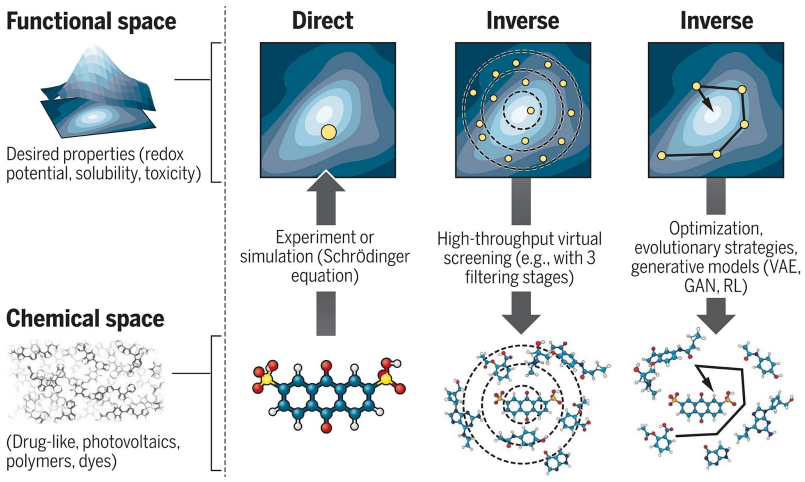


Figure 1.3.: Illustration of ML-based inverse design [2]. The upper part of the figure reflects the design target in the functional space, while the lower part of the figure demonstrates the corresponding structures in the chemical space. The ML-based inverse design method is more effective than traditional inverse design in new structure exploration. It does not require the extensive calculation of multiple structures and selects the structures with desired properties. On the contrary, it automatically searches for structures with desired properties.

design new stable crystal structures of V_xO_y [75].

In this work, **using ML to explore the relationship between crystal structures and intrinsic properties** will be extensively discussed. Crystal structures describe the arrangement of atoms. Furthermore, the intrinsic properties are the properties dependent on the chemical composition and crystal structures [76], like the formation energies [77], band gaps [78], magnetic moments [79], etc. Thus, such a relationship reflects the connection between crystal structures and resulting properties. The applications of ML on the crystal structures and intrinsic properties mapping can be conducted in the forward inference. For example, formation energies have been evaluated by ML models with crystal structures as inputs [80, 81], which indicated the stability of materials. There are three key intrinsic magnetic properties for magnetic materials [82]: magnetization [83], magnetic anisotropy energy [84], and the critical ordering temperature [32]. All of them can be predicted by ML models, accelerating the design of new functional magnetic materials. The ML models can also predict band gaps [85], which assist the discovery of insulators [86]. Another interesting application is to do inverse design, *i.e.*, the prediction of structures with desired properties. For example, the search for anti-cancer medicines can be achieved with "qualitative estimate of drug-likeness" as the target [87]. Using low formation energy

as the target, ML models can design the crystal structures of thermodynamically stable perovskites [88]. Similarly, the objectives to reduce the absorption energy of the crystal surfaces enable the design of catalysts [89]. There is also the model to seek out structures of new semiconductors by aiming at specified band gap ranges [90].

One of the challenges faced by ML for materials science is overfitting [91]. It means that the ML models have good performance over the known data but have lousy performance for the unknown data. This is because the ML models are not optimized, which is expected to happen when the data of ML models have large dimensions. Since the input data to represent materials are usually multi-dimensional, the potential of overfitting is high in ML for materials science. Consequently, despite that ML has revolutionized numerous fields, the application of ML in materials science did not receive intensive attention one decade ago [29]. However, as the more delicate ML methods and techniques are developed to overcome overfitting, an extensive collection of ML models to examine the relationship between intrinsic properties and crystal structures has been developed, as demonstrated in the previous paragraph.

Thus, it is of great significance to elaborate ML for materials science in the following four aspects:

- Objectives, which are the target of the ML models.
- Data, which are (numeric) information collected through observations [92].
- Representations (or descriptors), which mean a set of parameters to describe the data, aiming at transforming the collected data into a form that can be identified by ML algorithms [93].
- Algorithms, which refer to mathematical processes that take the representation as the input to achieve the objectives, where parameters in ML models can be automatically updated during the processes [94].

1.3.1. Objectives

Objectives indicate the purpose of ML models. From this perspective, ML, in general, can be classified into three main categories, *i.e.*, supervised learning, unsupervised learning, and reinforcement learning, as demonstrated in Figure 1.4. It is worth noting that semi-supervised learning is a hybridization of supervised and unsupervised learning [95]. However, semi-supervised learning and unsupervised learning are similar, so semi-supervised learning is combined with unsupervised learning in this thesis.

Firstly, supervised learning is an ML model to map the known inputs (X) to the unknown outputs by a mathematical function (f), which is established from the labeled training output (y) [96]. It can be expressed as:

$$\min_f ||y - f(X)|| \tag{1.2}$$

which is also known as the objective function or loss function. Depending on the format of the outputs (a real value or a Boolean value), the supervised learning can be further divided into regression or classification, respectively. Taking the handwritten number recognition as an example, only when the written number X is predicted correctly by f, the resulting loss function can be 0. In contrast, the wrong predictions will significantly increase the loss, which motivates the modification of f to reduce the loss. Thus, it is easy to tell that supervised learning can only do forward modelling in materials science, to directly predict the intrinsic properties [35] or to train the machine-learning interatomic potentials to calculate the intrinsic properties [97]. In addition, the most commonly used supervised learning models include but not limited to: support vector machine (SVM) [98], random forest (RF) [99], BO [100], artificial neural network (ANN) [101], convolutional neural network (CNN) [102], recurrent neural network (RNN) [103] and transformer [104].

Secondly, unsupervised learning is a model that summarizes the commonalities (f) of the inputs (X), which can be used in clustering, dimension reduction, and data generation [105]. It can be interpreted as the mapping to itself:

$$\min_f ||X - f(X)|| \tag{1.3}$$

Compared with supervised learning, unsupervised learning does not have a specific target. For clustering models like k-means [106] and affinity propagation (AP) [107], f is equivalent to searching center of corresponding clusters. Thus, their objective functions aim at reducing the distance between inputs with the centers. For dimension reduction such as principal component analysis (PCA) [108] and t-distributed stochastic neighbor embedding (t-SNE) [109], f refers to the mapping between real space and latent space. Therefore, their objective functions focus on increasing the similarity between original inputs and dimension-reduced counterparts. As for the generation models, *e.g.*, generative adversarial network (GAN) [110], and variational autoencoder (VAE) [111], f is the generative function to design new data. Thus, generative models can be utilized to do ML-based inverse design.

Last but not least, reinforcement learning is a model to take optimized actions (A) by maximizing the rewards (R) in a determined environment (S) [112]. It is best represented

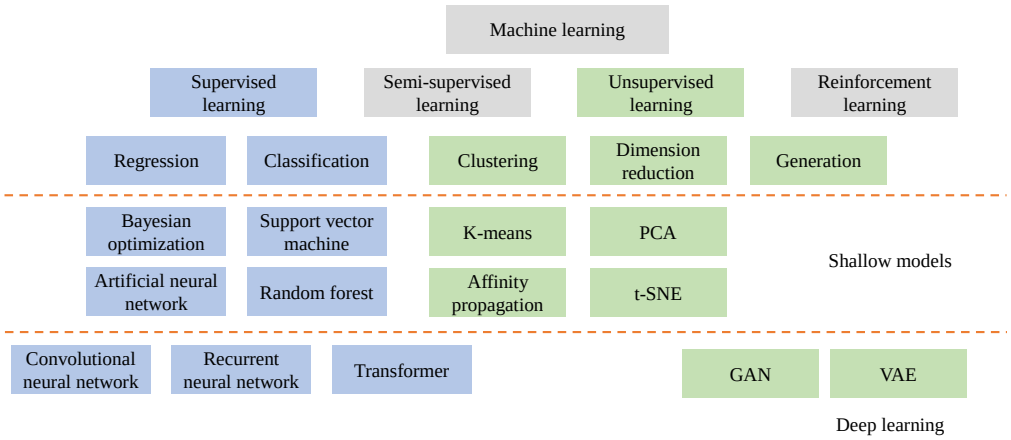


Figure 1.4.: Machine learning model categories. The machine learning model can be divided into supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. The supervise learning has two further categories namely the regression and classification, including: SVM, RF, BO, ANN, CNN, RNN, transformer, where the previous four belong to shallow models while the latter three are deep learning models. Similarly, the unsupervised learning can do clustering, dimension reduction and generation, including: k-means, AP, PCA, t-SNE, GAN and VAE, where the latter two are deep learning models.

as:

$$\max_{A} R_S(A) \tag{1.4}$$

Among all three categories, reinforcement learning is the only model learning from interactions rather than the existed data. Thus it has the potentials to do more creative work at a higher training cost. For example, the AlphaGo is a reinforcement learning model conducted under the rules of the Go games. It uses the winning of a Go game as the final reward of the model, where the actions are the movement of players [61]. The ability of the model can be polished through intensive training by self-playing or playing with human players and finally achieve a level beyond world champions. However, the training time is equivalent to the restless playing of a human for 200 years [68].

1.3.2. Data

Data serve as the basis of ML models. Data are a set of variables about one object [92]. In this thesis, data include both crystal structures and intrinsic properties. Moreover, a database is a collection of data, where there is increasing availability of materials databases such as the experimental databases and calculated databases in Table 1.1. The existing

experimental databases compile mostly the chemical compositions and crystal structures, of which only 33% have experimentally measured intrinsic properties. While the calculated databases usually comprise the optimized crystal structures and the calculated intrinsic properties, which are suitable for the purpose of examining the relationship between crystal structures and intrinsic properties.

Nevertheless, both of the calculated databases and the experimental databases only examine a limited chemical space. For example, the AFLOW database has covered 30,426 ternary systems (the combination of three elements [118], but the possible number of binary combinations is 266,916, *i.e.*, only 11.3% of ternary systems are considered for the past few decades. The ratio is even worse for quaternary systems, only 1.9% (150,622 out of 7,673,835). Thus, it is of great significance 1) to collect more data by more experiments or calculations, and 2) to better utilize the existed databases. The former relies on the improvement of calculation capability and accumulation over time, which cannot be easily overcome in the coming decades. However, the latter could be ameliorated by database curation and management.

There are three commonly used methods in database curation and management: 1) imposing FAIR principles [124], 2) demonstration of metadata [125], and 3) adoption of ontology [126].

- The FAIR principles aim at improving the findability, accessibility, interoperability, and reusability of the data. They could enforce the disclosure of computational details for each material and allow the direct comparison of the same material between different databases. Following this spirit, the National Research Data Infrastructure (NFDI) is launched by the German government to provide services and advice on research data management [127]. Thus, it will provide one united platform to search for the information of materials (findability), to allow the download of the found information (accessibility), to enable the direct comparison of the same material between different databases (interoperability), and to acquire enough details to reproduce and improve the current results (reusability).

- The metadata are “data that provides information about other data” [125], illustrating the general feature of data in databases. For example, the total number of available crystal structures and the distribution of one property in the database are typical metadata. Metadata helps the selection of databases for ML methods when the objectives have been determined because metadata provides general features of the database. For instance, a database with only Boolean values of magnetic or not could be used for a classification model, while a database with continuous formation energies is suitable for regression.

Table 1.1.: 13 commonly used databases containing structures and intrinsic properties in materials science, where 6 are experimental databases and the other 7 are computational. They includes: Inorganic Crystal Structure Database (ICSD) [113], AtomWork adv. [114], UltraHigh Carbon Steel Micrograph DataBase (UHCSDB) [115], ASM Alloy Phase Diagram Database (ASM) [116], and International Centre for Diffraction Data (ICDD) [117], Materials Projects (MP) [51], Open Quantum Materials Database (OQMD) [49, 50], Automatic FLOW for Materials Discovery (AFLOW) databases [118], Novel Materials Discovery (NOMAD) [119], 2D Materials Encyclopedia (2D database) [120], Crystallography Open Database (COD) [121], Quantum machine 9 (QM9) [122], and High Throughput Experimental Materials Database (HTEM) [123]

Database	Number	Type	Structures	Properties
ICSD	210,000	Experiment	Crystal structures	None
AtomWork adv.	334,450	Experiment	Crystal structures & X-ray diffraction & phase diagrams	483 properties covering electronic and electrical, magnetic, mechanical, optical, phase transition, superconductivity, thermal and thermodynamical
UHCSDB	1,700	Experiment	SEM images	None
ASM	40,300	Experiment	Phase diagrams	None
ICDD	1,047,661	Experiment	X-ray diffraction	None
MP	144,595	Calculation	Crystal structures & X-ray diffraction & phase diagrams	Properties covering thermodynamic properties (*e.g.*, formation energy), mechanical properties (*e.g.*, elasticity), electronic properties (*e.g.*, band structure, dielectric constant, piezoelectricity etc.), and magnetic properties (*e.g.*, magnetization)
OQMD	815,654	Calculation	Crystal structures & phase diagrams	thermodynamic properties (*e.g.*, formation energy and energy above convex hull)
AFLOW	3,513,989	Calculation	Crystal structures & phase diagrams	Properties covering thermodynamic properties, thermal properties (*e.g.*, heat capacity), mechanical properties, electronic properties, and magnetic properties
NOMAD	11,159,336	Calculation	Crystal structures & molecules	Properties covering thermodynamic properties, thermal properties, mechanical properties, electronic properties, and magnetic properties
2D database	6,351	Calculation	2D materials	Includes formation energy, band structures and exfoliation energy
COD	475,225	Calculation	Crystal structures of organic, inorganic, metal-organics	None
QM9	133,885	Calculation	Molecules	Including dipole moment, isotropic polarizability, band gap, free energy and heat capacity
HTEM	47,213	Experiment	X-ray diffraction of thin films	Absorption coefficient and band gap

- The ontology established a widely accepted formal representation to describe data. It is easier for researchers to find an internal link and understand the underlying mechanism between syntheses, chemical compositions, crystal structures, microstructures, and physical properties because the ontology visualizes the above connections from both experimental measurements and theoretical modellings. It further helps the selection of a representation of the materials to improve the performance of ML models due to the underlying mechanism between different descriptors are visualized as well [126]. However, the application of ontology in the materials informatics community is only in its infancy, so it is still necessary for researchers to develop proper representations of known materials.

1.3.3. Representations

Representations (or descriptors) are a set of efficiently descriptive parameters of the materials [128, 129], *i.e.*, use the least information (but adequate to perform the tasks) to represent materials. For example, chemical compositions, space group numbers, lattice constants, and Wyckoff positions, etc., compose a proper set of representations to understand the crystal structures of materials as demonstrated in Figure 1.5.

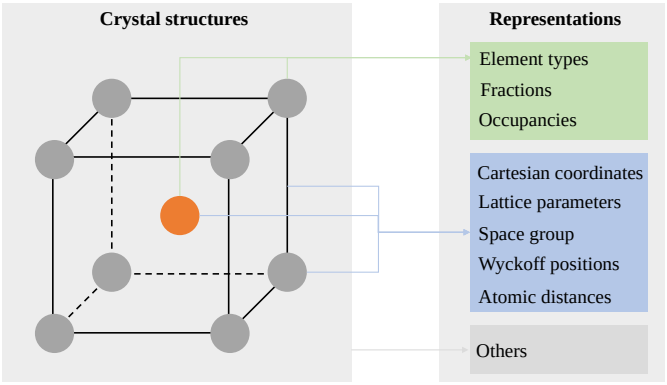


Figure 1.5.: Schematics of representations for crystal structures. The left panel indicates a crystal structure, while the right panel denotes a representation vector, in which the green depicts the chemical information, while the blue reflects the structural information.

In general, a set of representations has four common features: reproducible, meaningful, distinguishable, and straightforward.

- "Reproducible" means that the identical set of representations can be acquired for

the same material by following the same procedure. For example, the position of one electron in a material cannot be used as a representation because it is dependent on the measurement, but its wave function could be a good representation.

- "Meaningful" indicates that the generated representations reflect intrinsic features of the materials, which allow further analysis of the materials. Otherwise, the representations ignore the important information of the materials and therefore lead to unreasonable conclusions. For example, it is hard to identify the chemical composition of a material by analyzing its space group, so space group is hardly a meaningful representation in this task. However, "meaningful" is also related to the purpose of representation. Following the same example, the space group is one important representation to measure the hardness of materials.
- "Distinguishable" suggests that different materials should possess distinct representations and therefore enables the identification of materials with only a set of representations. Take Fe as an example, the chemical composition of α-Fe and γ-Fe are the same, but they are with different structures, *i.e.*, body-centered cubic (BCC) and face-centered cubic (FCC), respectively. Thus, a set of representations containing both chemical and structural information can still distinguish them.
- Last but not least, "straightforward" means the representations could not be more complicated than the materials themselves. Otherwise, the usage of representation is inefficient. For example, the chemical compositions and crystal space groups are proper representations for one material, but the Cartesian coordinates of all atoms in this material would be too complicated to be used as representations.

Apart from the features mentioned above, the representations used for ML should be homogeneous and numerical as well. "Homogeneous" means that representations of different materials should always have the same dimension. For example, element symbol vectors of compounds are not homogeneous because binary compounds have two elements while ternary compounds have three. So such vectors cannot be directly used as representations in most ML models. In practice, when dealing with heterogeneous representations, it is common to pad zeros into the vectors to make them homogeneous. "Numerical" indicates that the representations should consist of real numbers, not strings. Taking magnetic or not as an example, true or false is not a proper numerical representation, but 1 or 0 can serve as a numerical representation. This originates from ML models' training process, where only numerical data can be used to calculate the derivatives of loss functions.

For the generative model, the representation should be reversible as well. "Reversible" suggests that all representations should contain enough information that allows

them to transform back to their original data. Taking the magnetic anisotropy energy as an example, it can serve as a representation of a material, which can be used to predict another magnetic properties of the material. However, the crystal structure of this material cannot be directly reverted from the magnetic anisotropy energy. This means that magnetic anisotropy energy is not a reversible representation of a crystal structure.

However, as the ugly duckling theorem indicates [130], there is no problem-independent representation suitable for all objectives. Thus, there are no commonly used descriptors for all materials until now, but researchers did develop descriptors that can be used for an aimed purpose. There are Coulomb matrix [131], Ewald sum matrix [132], sine matrix [132], smooth overlap of atomic position (SOAP) [97], and atom centered symmetry functions (ACSF) [133] designed to predict interatomic potentials (mostly from structural perspectives). The materials-agnostic platform for informatics and exploration (MAGPIE) [134] can be used to predict other intrinsic properties. Crystal graphs also play an important role in the regression of intrinsic properties [35]. Apart from these, the image-based continuous representation could be developed in a reversible manner for generative purposes [135, 136]. All these representations will be discussed in detail in Chapter 2.

1.3.4. Algorithms

The algorithms are the essence of ML techniques [137]. In this thesis, the algorithms specifically indicate the processes to establish, train, and evaluate ML models. Before discussing the details of the algorithms, it is necessary to clarify: ML models are the mathematical equations to calculate outputs from inputs, while the ML algorithms include the process to establish, train, and evaluate the ML models.

That is to say, an algorithm covers three different aspects: 1) a model itself, 2) a training process, and 3) a test process.

- Model: the selection of a model is always the most important step in ML. As the no free lunch theorem indicates, there is no ML model suitable for all problems [138]. Thus, usually, one can tell which model is better suited for one specific task by comparing their performance on the same task.
- Training: the processes to train different models follow two steps [139]: 1) to quantitatively evaluate the difference between the predicted values and the target values (loss function); 2) to make the modification of the weights in the ML models (optimizer) based on the difference. By repeating these two steps, the performance of all ML models will be improved, *i.e.*, the loss function will be minimized through the training process.

- Evaluation: mathematically speaking, "all models (including ML models) are wrong", *i.e.*, unable to be entirely sure of the results [140]. This leads to uncertainty quantification to determine the reliability of the results [141]. Although the loss function could evaluate the performance of models, the loss functions focus on comparing the performance of the same model at different training steps. When it comes to comparing the relative performance of the different models, more general metrics like relative error will be utilized.

Model

As previously mentioned, an ML model is a mathematical formula that takes the representation as inputs and predicts the desired values as outputs. So it is easy to conclude that the selection of an ML model is determined by the output (objective) of the model as well as the inputs (representation), as demonstrated in Figure 1.6.

On the one hand, the selection of ML models is dependent on the tasks. For example, in order to perform natural language processing (NLP) like the translation from English to German, RNN [103] and transformer [104] are better choices because they are designed to capture the correlation between words in one sentence. Whereas, CNN is better suited for identifying a pattern on images due to the architecture of the convolutional layer [102]. Regarding dimension reduction, PCA [142] and t-SNE [109] will map the data points into a low dimensional space (linearly or non-linearly) since they are designed to find the commonality of data points. Classification problems are better treated by SVM [98] or RF [99] since they can separate data with different features. In contrast, BO [100] and ANN [101] would be better to do regression because these methods are good at making predictions. Note that BO and ANN can also be used to classify by using logistic loss (which will be discussed later in this section), but they are better suited in the regression task. Similarly, SVM and RF can also do regression. It is also possible to combine the models mentioned above to conduct more complicated jobs, like the autoencoder to compress data, VAE [111] and GAN [110] to conduct generation.

On the other hand, the choice of models is related to the representation as well. For example, the word-based representations are better processed by RNN and transformer. As for the vector-based representations mentioned in the previous section, SVM, BO, RF, and ANN can be a better choice, but these four algorithms focus on different aspects. For example, the SVM and BO methods usually take kernels [143] as inputs, which describe the similarity of different representation points. In contrast, ANN, RF, PCA, and t-SNE can work on the representations directly. To be more specific, ANN must work on continuous and normalized vectors. PCA and t-SNE only require continuous vectors, while RF can work on discrete vectors. When the number of data is limited, kernels effectively increase

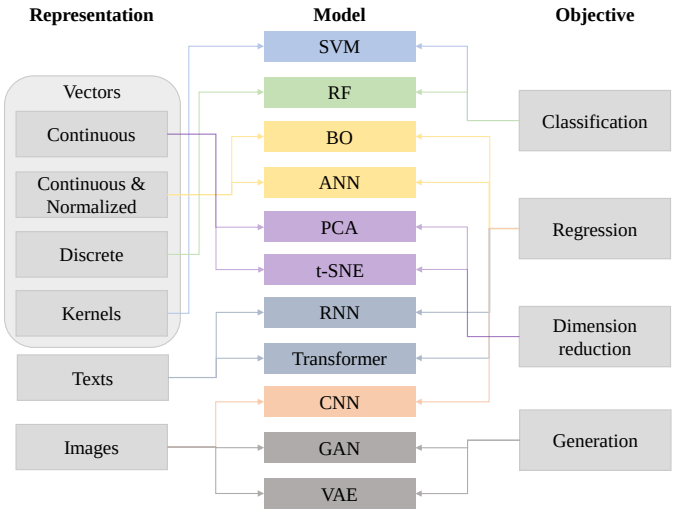


Figure 1.6.: Schematics of selecting ML models. The middle part denotes ML models, including SVM, RF, BO, ANN, PCA, t-SNE, RNN, Transformer, CNN, GAN, VAE. This part is connected with their best matching representation type (in the left part) and objectives (in the right part). Note that the connections reflect only the best selection, not the possible selections.

the number of inputs, but the latter two models could have better performance with more data. When it comes to images, CNN is good at extracting the necessary information from images. In general, the selection of an ML model is affected by many different aspects, and commonly used models will be discussed in detail in Chapter 3.

Training

After determining a model to use, the next step is the training of the model. The model itself is only a formula, where the parameters in the equation are not optimized yet at the beginning. The training process aims to determine the parameters in the model to improve the model's performance. As mentioned previously, this will be divided into two different aspects: the loss function and the optimizer, as demonstrated in the ML workflow in Figure 1.7.

The main purpose of a loss function is to quantitatively evaluate the difference between predicted values and the target values, reflecting the accuracy of the predictions. Taking the forward modelling as an example, the mean absolute error (MAE) and mean square

error (MSE) are good loss functions for the regression model [144]. They are defined as:

$$\mathrm{MAE}(\theta) = \frac{\sum_{i=1}^{n} |y_i - f_\theta(X_i)|}{n} \tag{1.5}$$

$$\mathrm{MSE}(\theta) = \frac{\sum_{i=1}^{n} |y_i - f_\theta(X_i)|^2}{n} \tag{1.6}$$

where y_i represents the target values of a certain property, X_i denotes the inputs, f indicates the ML model, θ means the parameters in the ML model, i is the index of the sample, n represents the total number of samples. It is evident that MAE is more sensitive to the slight deviation, while MSE is determined by the significant deviation, where the smaller value indicates better prediction accuracy. Similarly, the absolute loss function of a binary classification problem (logistic loss) can be defined as [145]:

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n}[-y_i log(f_\theta(X_i)) - (1-y_i)log(1-f_\theta(X_i))] \tag{1.7}$$

where J is the logistic loss, and others have the same meanings as the regression loss.

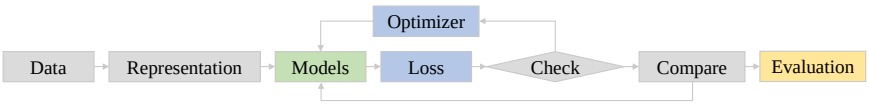


Figure 1.7.: Workflow of a machine learning algorithm, where the model, training and evaluation are in green, blue and yellow, respectively. The training process are the same for almost all machine leaning models, that is: collecting data, generating representations, putting them into the models, calculating the loss function and optimizing the model until the loss is acceptable, finally evaluating the performance of the model.

The optimizers could then guide the modification of the parameters (weights) in the ML model to minimize that loss function, *i.e.*, make a more accurate prediction. Taking the most straightforward gradient descent (GD) optimizer as an example [146], it first calculates the partial derivative of the loss function over the individual weights. The gradients could then indicate the direction to modify the weights towards a lower loss function, but the magnitude of modification cannot be calculated in this process. This magnitude is then determined by the learning rate, which is decided before the training process. If the learning rate is too large, then the modification of the weights will be so significant that the weights go beyond the optimized point. On the contrary, if the learning

rate is too low, the training process would be too slow, making the training procedure very time-consuming. Nevertheless, with a reasonable learning rate, the best performance can always be achieved for any convex function. However, practical problems cannot always be formulated as a minimization problem with guaranteed convexity of the loss function in the phase space. More robust optimizers like the stochastic gradient descent (SGD) optimizer should be used in this case [147]. Instead of calculating the gradients for all training examples at each training step, the SGD optimizer only randomly selects a subset of training examples to calculate gradients. This gives some randomness of the descending direction and thus helps the weight to jump out of the saddle point, which finally makes the SGD optimizer able to optimize non-convex problems. Although SGD optimizer has the possibility to go to the wrong directions under extreme conditions, this method is helpful in practice. Recently, the ADAM optimizer [148], which stands for adaptive moment estimation, has become more popular. It utilizes the concept of momentum by adding fractions of previous gradients to the current one, *i.e.*, calculating the weighted average of the gradients, which is more robust compared with the SGD optimizer. Similarly, other optimizers like Adagrad [149], Adadelta [150], RM-Sprop [151], etc. are also proposed. However, until now, no optimizer guarantees to find the global minimum of all non-convex problems [152, 153].

Evaluation

After training a model, the general performance of it should be evaluated. Although the loss function mentioned above can be used as a metric, both MAE and MSE represent the absolute deviations for the target values, making it hard to make a direct comparison of the results. For example, the MAE of formation energy prediction in the Bi-Se system could be 0.019 eV/atom while this value is 0.1 eV/atom for all binary compounds, but the first model is not necessarily better than the second one because the formation energy ranges of these two models are not the same [135, 136].

Under such circumstances, relative error evaluations for the regression and classification are required. A relative deviation of a regression model can be measured by the coefficient of determination (R^2) [154]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|^2}{\sum_{i=1}^{n} |y_i - \bar{y}|^2} \tag{1.8}$$

where $\bar{y}$ is the average value of all target values, and $\hat{y}_i = f_\theta(X_i)$ means the predicted values. The value of R^2 is a direct evaluation of accuracy. For instance, when R^2 (accuracy) is 0.85 (85%), it means that the prediction model could give the accurate predictions for

85% of data. Similarly, the logistic regression is not explicit to make comparison between models, the accuracy of classification can be defined as [155]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{1.9}$$

where TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative, respectively. Taking binary classification (1 or 0) as an example, true positive means that the target value is 1 and the prediction is also 1, while true negative is that the target value is 0 and the prediction value is also 0. False negative indicates that the target value is 1, but the prediction result is 0; and false positive denotes that the target value is 0, but the prediction value is 1.

To better evaluate the performance of ML models towards unknown samples, in forward modelling, data are usually split into three datasets [156]: training, validation, and test sets, with typical splitting ratios being 70%/20%/10% or 80%/10%/10%. The training set is used to fit the parameters in an ML model, *i.e.*, the tenable parameters in a forward model are adjusted to match the outputs with target outputs. Thus, ML models can usually achieve high accuracy over the training sets (low accuracy in the training sets is called underfitting), but there is no way to tell whether the trained models only work for the training data (overfitting) or can be generalized to other data. Thus, a small portion of randomly select data (test set) is saved to examine the performance of the trained model. Since the test sets are randomly selected and unknown to the model, the corresponding accuracy could reflect the performance of the trained model. However, there is no guarantee that the trained model is the best model we can get. For example, another model with better performance could be achieved by tuning hyper-parameters (like the learning rate). Thus, a validation set (which is another set of the test) is split to test the performance of the trained model with different hyper-parameters, where only the best-performed model will be adopted, and only the performance of the best model will be evaluated by the test set. In practice, when the number of available samples is limited, the validation set and test set can be combined as one to increase the training set ratio. In summary, the training set is used to train the model, the validation set is to select the best performance model, and the test set is to demonstrate the ability of the best model. To ensure the performance of a model is not caused by a particular data splitting, several splittings are usually applied on the same model, which is also dubbed as cross-validation [157]. When the model's performance is stable in different splittings, the model can be robust when applied to unknown samples.

1.4. Summary

To summarize, ML can be applied to tackle the PSPP mapping in materials science and facilitate more efficient materials design. As a result, ML techniques revolutionize the research on materials science as the fourth paradigm and improve the other three paradigms by conducting experiments in an automatic way, increasing the accuracy of simulations based on classical theories, and improving the efficiency of quantum theories based calculations. As summarized in Figure 1.8, the fundamental aspects of applying ML to materials science are:

- Objectives, which divide the ML methods into forward modelling and inverse design.
- Databases, which collect the crystal structures and the intrinsic properties.
- Representations, which efficiently describe the crystal structures to predict intrinsic properties.
- Algorithms, which are used to perform the tasks based on the objectives and representations.

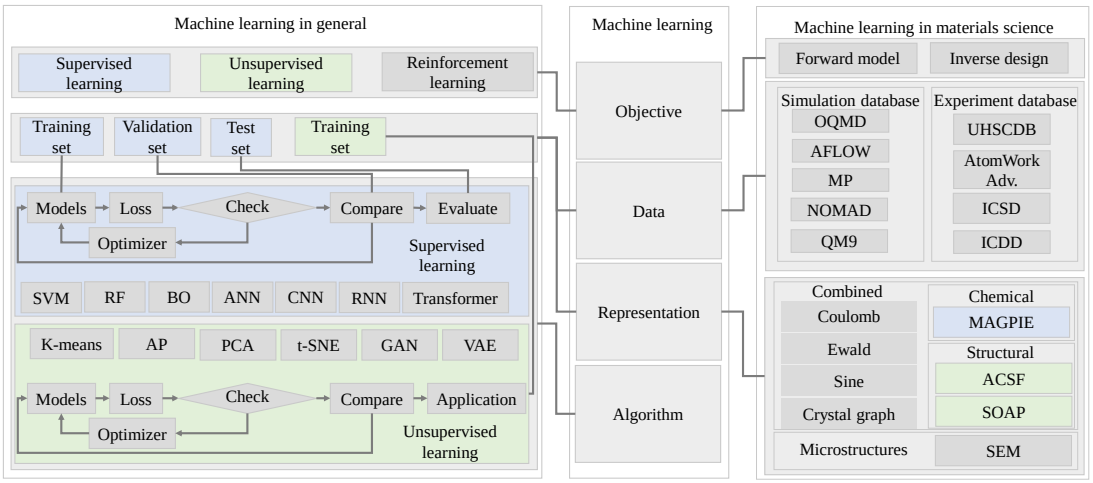


Figure 1.8.: Schematics of machine learning, *i.e.*, delineating machine learning from objective, data, representation and algorithm perspectives. The left block depicts the machine learning in general, while the right block describes the machine learning in materials science.

In the following part of the thesis, I will focus on how to establish the relationship between crystal structures and intrinsic properties. Chapter 2 and 3 will focus on the

details of two essential procedures in ML, *i.e.*, the representations and models, respectively. Chapter 4 will illustrate how ML can be conducted using my publications as examples, explaining how to 1) predict the $\mathrm{T_C}$ of the ferromagnets, 2) develop an inverse design model to predict stable crystal structures, as well as 3) establish machine-learning interatomic potentials for hexagonal boron nitride (h-BN). Chapter 5 will present the conclusions and provide an outlook of ML for materials science.

2. Representations

As mentioned in Section 1.3.3, representations are a set of descriptive parameters of the chemical and structural information of the materials [128, 129]. In this chapter, the commonly used representations in materials science covering chemical descriptors and structural descriptors are discussed, corresponding to the structural and chemical information of materials. The structural representations describe the arrangement of atoms, which includes two different aspects: 1) the translational and rotational symmetry of crystal structures, 2) the interaction between atoms (local environment) [128]. In contrast, the chemical representations numerate the chemical compositions of materials using statistics such as the average atomic numbers of chemical compositions [134]. However, both descriptors mentioned above focus on partial features of the materials. The combined descriptors, including both chemical and structural information, can describe a material completely.

2.1. Chemical descriptors

The chemical descriptors reflect the most fundamental information of the materials. Ward et al. developed MAGPIE, which is a set of descriptors that only consist of chemical information [134]. MAGPIE collects the fundamental properties of the corresponding elemental compounds and calculates their statistical metrics to form information vectors. In general, such vectors have 145 dimensions, compromising four categories:

- Stoichiometric features (in total 6), which ignore the exact element compositions. Such features represent the fractions of elements, expressed by the L^p norms of the fractions:

$$|x|_p = \sum_{i=1}^{n} |x_i^p|^{1/p} \tag{2.1}$$

  where x is the fraction vector, and x_i is the fraction term. In order to catch a large difference in stoichiometry, p with 0, 2, 3, 5, 7, and 10 are considered in MAGPIE, where $|x|_0$ is the number of elements in the compounds.

- Elemental property statistics (in total 132), which are defined as the mean, absolute deviation, range (*i.e.*, difference between maximum and minimum), minimum (*i.e.*, the property of the element with the smallest value), maximum (*i.e.*, the property of the element with the largest value), and mode (*i.e.*, the property of the most prevalent elements) of 22 different elemental properties:
    1. Atomic number
    2. Mendeleev number
    3. Atomic weight
    4. Melting temperature
    5. Row on the periodic table
    6. Column on the periodic table
    7. Covalent radius
    8. Electronegativity
    9. Number of s valence electrons
    10. Number of p valence electrons
    11. Number of d valence electrons
    12. Number of f valence electrons
    13. Total number of valance electrons
    14. Number of unfilled s states (*i.e.*, the number of electrons in a partially-occupied s orbital subtracted from the total number of electrons allowed in that orbital)
    15. Number of unfilled p states
    16. Number of unfilled d states
    17. Number of unfilled f states
    18. Total number of unfilled states
    19. Specific volume of 0 K ground state
    20. Band gap energy of 0 K ground state
    21. Magnetic moment (per atom) of 0 K ground state
    22. Space group number of 0 K ground state

The means and absolute deviations are calculated by the following two equations:

$$\bar{f} = \sum_{i=1}^{n} x_i f_i \tag{2.2}$$

$$\sigma_f = \sum_{i=1}^{n} x_i |f_i - \bar{f}| \tag{2.3}$$

where f represents the corresponding properties.

- Electronic structure features, which are the average fraction of electrons from the s, p, d, and f valence shells for all elements in the composition (in total 4). Their mathematical forms are defined as:

$$\bar{F}_\alpha = \frac{\sum_{i=1}^{n} x_i N_i^\alpha}{\sum_{i=1}^{n} x_i N_i^{total}} \tag{2.4}$$

  where F is the fraction, N is the valence electron number, and α can be s, p, d, or f.

- Ionic features, which determine the strength of ionic bonds in a material (in total 3). The first one is a Boolean value to indicate whether it is possible to form an ionic bond; while the other two are the largest and mean value of "ionic character", which can be calculated by the following equation:

$$I(X_i, X_j) = 1 - e^{-0.25(X_i - X_j)^2} \tag{2.5}$$

  where X means the electronegativity of corresponding elements, i and j represent the elements [158].

In general, the MAGPIE descriptors are effective. On the one hand, they are easy to construct when the chemical composition of one material is available, because the construction does not require further information. For example, Stanev et al. used only the MAGPIE descriptors to predict the critical temperatures of superconductors [54]. On the other hand, the MAGPIE descriptors do not require further processing, because they are homogeneous (145 dimensions). However, MAGPIE does not take structural information into consideration, which prevents it from doing more predictions. For instance, the structural information improves the prediction accuracy of $\mathrm{T_C}$ for ferromagenets [159, 32].

2.2. Structural descriptors

2.2.1. Atom centered symmetry functions

Structures can be described from local environments. ACSF use a set of many-body fingerprints to represent the local environment in the neighbours of one certain atom [133]. One important feature of ACSF is that they include all the two-body and three-body interactions, referred as radical and angular functions, respectively. Mathematically speaking, three radical functions and two angular function are considered, as demonstrated in Figure 2.1. The first radical function can be expressed as the following equation:

$$G_i^{1,Z_n} = \sum_j^{Z_n} f_c(R_{ij}) \tag{2.6}$$

where $R_{ij} = |\mathbf{R}_i - \mathbf{R}_j|$ indicates the distance between two atoms, Z_n is the atomic number of the corresponding element. f_c is the cutoff function, which is defined as:

$$f_c(R_{ij}) = \begin{cases} \frac{1}{2}[\cos(\pi \frac{R_{ij}}{R_{cut}}) + 1] & R_{ij} \leq R_{cut} \\ 0 & R_{ij} > R_{cut} \end{cases} \tag{2.7}$$

where R_{cut} is the cutoff radius defined by the users. The first radical function is simply the summation of all the cutoff functions with the neighbor atoms. While the second radical function includes the Gaussian distribution as well:

$$G_i^{2,Z_n} = \sum_j^{Z_n} e^{-\eta_1 (R_{ij} - R_S)^2} \cdot f_c(R_{ij}) \tag{2.8}$$

where η_1 defines the width of the Gaussian distribution function, R_S describes the center of the Gaussian distribution. Following the same spirit, the third radical function takes the Fourier transform into consideration by including the cosine functions:

$$G_i^{3,Z_n} = \sum_j^{Z_n} \cos(\kappa R_{ij}) \cdot f_c(R_{ij}) \tag{2.9}$$

where κ affects the periodicity of the function. In contrast, two angular functions depict the three-body interactions, which are defined in the following equations:

$$G_i^{4,Z_n,Z_m} = 2^{1-\zeta} \sum_j^{Z_n} \sum_k^{Z_m} (1 + \lambda \cos\theta)^\zeta \cdot e^{-\eta_2 (R_{ij}^2 + R_{ik}^2 + R_{jk}^2)} \cdot f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}) \tag{2.10}$$

$$G_i^{5,Z_n,Z_m} = 2^{1-\zeta} \sum_j^{Z_n} \sum_k^{Z_m} (1 + \lambda \cos\theta)^\zeta \cdot e^{-\eta_2(R_{ij}^2 + R_{ik}^2)} \cdot f_c(R_{ij}) f_c(R_{ik})) \tag{2.11}$$

where ζ is called the angular resolution (higher ζ leads to narrower angular function), η_2 has the same function as η_1, θ is the angle between three atoms (atom i, atom j, and atom k) with atom i in the center. The G_i^4 function ensures the condition that all three atoms are within the cutoff range to each other, while the G_i^5 function ignores the affect on atom i by the interaction between atom j and k.

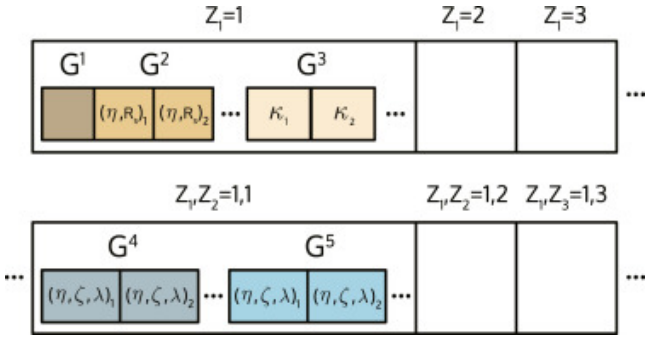


Figure 2.1.: Structure of the ACSF output vector [160], where the values of the two-body symmetry functions and the values of the three-body symmetry functions are given in two columns, respectively. The two-body symmetry functions are also dubbed as the radical functions, reflecting the interactions between two atoms. Similarly, the three-body symmetry functions are known as angular functions, indicating the interactions between three atoms.

In general, ACSF is suitable for describing small molecules. As demonstrated in the five equations mentioned above, ACSF describes the local environments of an atom within a pre-determined cutoff range. The cutoff range in ACSF is usually small in order to reduce the computational cost. Thus, only the short-range interactions between atoms are captured [133]. However, the long-range interactions between atoms are missing, which prevents the integration of rotation and translation symmetry into the descriptors. As a result, ACSF cannot be extensively used to describe crystalline materials but can represent small molecules.

ACSF is commonly used to train machine-learning interatomic potentials. As mentioned above, ACSF uses interactions between atoms to represent the local environment of an atom, which are calculated by the coordinates of the atoms. Such interactions are similar to the empirical potentials in MD. Therefore it is possible to apply ACSF to train machine-learning interatomic potentials. In addition, ACSF describes the local environment by

individual atoms, but none of them alone can represent the molecules. Instead, all of them should be included in the representation of molecules. For instance, water molecules are represented by the local environments for H atoms and O atoms in water molecules. However, different materials usually have a different number of atoms. Thus, their representations by ACSF contain a different number of atoms, *i.e.*, heterogeneous. Such heterogeneity limits their usage in predicting the properties of different materials and restricts their applications.

2.2.2. Smooth overlap of atomic positions

SOAP describes the structures from both local and global aspects. SOAP achieves this by a spherical expansion of the interactions between atoms, as illustrated in Figure 2.2, where the local environment of one atom are expressed as the Gaussian density function:

$$\rho^Z(\mathbf{r}) = \sum_i^Z e^{-\frac{1}{2\sigma^2}|\mathbf{r}-\mathbf{R}_i|^2} \tag{2.12}$$

where Z represents elements. $\mathbf{r}$ indicates the coordinates of the target atom, $\mathbf{R}_i$ is the coordinates of neighboring atoms. σ denotes the standard deviation of the Gaussian distribution function. Such a function can be expanded in the following format:

$$\rho^Z(\mathbf{r}) = \sum_{nlm} c^Z_{nlm} g_n(r) Y_l^m(\theta, \phi) \tag{2.13}$$

where c^Z_{nlm} is the coefficient, which can be calculated by the following equation:

$$c^Z_{nlm} = \iiint_V dV g_n(r) Y_l^m(\theta, \phi) \rho^Z(\mathbf{r}) \tag{2.14}$$

where V means the whole real space. $g_n(r)$ is called radial basis functions, which also forms an orthogonal complete set. It can be mathematical expressed as:

$$g_n(r) = \sum_\alpha^N W_{n\alpha} \psi_\alpha(r) \tag{2.15}$$

where N is the number of atoms within the cutoff range. $\mathbf{W}$ is coefficient matrix, defined as:

$$\mathbf{W} = \mathbf{S}^{-1/2} \tag{2.16}$$

and $\mathbf{S}$ can be determined by the following equation:

$$S_{\alpha\beta} = \int_0^{r_{cut}} \psi_\alpha(r)\psi_\beta(r)dr \tag{2.17}$$

where r_{cut} is the cutoff range, and $\psi_\alpha(r)$ can be expressed as following:

$$\psi_\alpha = \sqrt{\frac{2\alpha+5}{r_{cut}^{2\alpha+5}}}(r_{cut}-r)^{\alpha+2} \tag{2.18}$$

where α can be integers from 1 to N. $Y_l^m(\theta,\phi)$ is the spherical harmonics, which can be expressed as:

$$Y_l^m(\theta,\phi) = (-1)^m\sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_m^l(\cos\theta)e^{im\phi} \tag{2.19}$$

where P_m^l is the associated Legendre Polynomial.

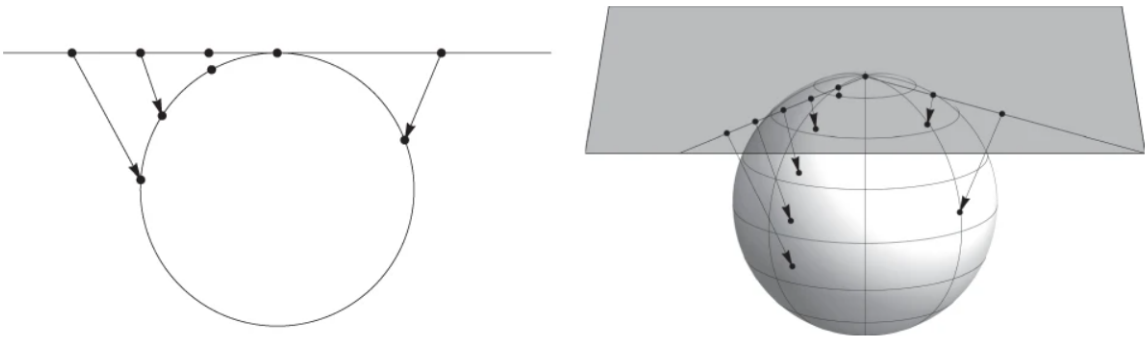


Figure 2.2.: Mapping of a flat space in one and two dimensions onto the surface of a sphere in one higher dimension [97, 29]. This illustrates the process of generating SOAP representation, where the interactions between atoms are expanded as the spherical functions.

In general, SOAP has wider application than ACSF. Individual interactions between atoms in SOAP are examined through a spherical expansion, which is equivalent to the many-body fingerprints in ACSF. Thus, SOAP can also describe the local environment of individual atoms. However, the expansion terms in SOAP are periodic functions, which consider periodicity. When the cutoff radius in SOAP is larger than the size of a unit cell, SOAP allows the description of symmetries, and therefore the representations of crystal structures. As a result, SOAP can represent both molecules and crystal structures, which has wider application than ACSF.

SOAP is also commonly used in machine-learning interatomic potentials. The essence of SOAP is also to represent the local environment by interactions between atoms, so it can be used to train machine-learning interatomic potentials as ACSF. When taking SOAP as inputs in Gaussian approximation potential (GAP) [161, 97], it takes 10 times longer than MD with 10 times higher accuracy [162, 163]. Nevertheless, SOAP is also heterogeneous in representing materials, and SOAP dimensions are dependent on the number of atoms in materials. However, SOAP can be applied to represent materials. By calculating the average between SOAP descriptors for atoms in materials, they improve the prediction accuracy of T_C for ferromagents [159]. Although such application in ML is possible for SOAP, ambiguous physical meanings of the average calculation limit its usage.

2.3. Combined descriptors

2.3.1. Matrix descriptors

The combined descriptors consider both atomic species and interatomic distances of a compound and construct a correlation matrix of every two atoms in different formats. The matrix descriptors are the most straightforward way to describe the local environment of materials. Usually, they are not homogeneous in describing different materials, but they are undoubtedly homogeneous for a system with a fixed number of atoms in the unit cell. These are precisely the cases to train machine-learning interatomic potentials. Thus matrix descriptors can be used to train interatomic potentials. However, the combination of chemical and structural information endows the matrix descriptors with a wide range of applications in predicting intrinsic properties. To be more specific, there are the Coulomb matrix, Ewald summation matrix, and Sine matrix, as demonstrated in Figure 2.3.

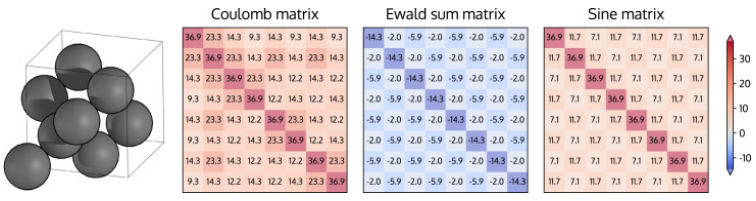


Figure 2.3.: Illustration of the Coulomb matrix, Ewald sum matrix and sine matrix for a periodic diamond structure [160]. From left to right are the crystal structure of diamond, Coulomb matrix, Ewald sum matrix and sine matrix, respectively. All of them capture the interactions between all the atoms in the crystal structures, where the only difference is the way in calculating interactions.

Coulomb matrix calculates the correlation matrix of every two atoms in the format of

Coulomb interactions. Its mathematical form is as following:

$$M_{ij}^{Coulomb} = \begin{cases} 0.5Z_i^{2.4} & i = j \\ \frac{Z_i Z_j}{|\mathbf{R_i} - \mathbf{R_j}|} & i \neq j \end{cases} \tag{2.20}$$

where Z_i denotes the atomic number of atom i and $|\mathbf{R}_i - \mathbf{R}_j|$ represents the distance (Euclidean) between atom i and j. Naturally, the Coulomb matrix represents a material with n atoms by a $n \times n$ matrix. For materials with finite atoms like molecules, the Coulomb matrix is able to capture all the interactions, which describes their local environment perfectly. However, the Coulomb matrix would become a matrix with an astronomical number of dimensions for a real crystal because a crystal usually has an enormous number of atoms. Thus, the Coulomb matrix is not a proper descriptor for crystalline materials.

The first attempt to take periodicity into consideration in matrix descriptors is the Ewald sum matrix [132], where the idea of the Ewald summation is considered [164]. Such descriptors separate the contribution of interactions into three terms, *i.e.*, the short-range interaction term M_{ij}^{real}, the long-range interaction term $M_{ij}^{reciprocal}$, and the self-interaction term M_{ij}^{self}, corresponding to the real space sum, reciprocal space sum and constant term in the Ewald summation [160]. Mathematically, the Ewald sum matrix can be expressed as:

$$M_{ij}^{Ewald} = M_{ij}^{real} + M_{ij}^{reciprocal} + M_{ij}^{self} \tag{2.21}$$

where the short range interaction term is given by the following equation:

$$M_{ij}^{real} = \begin{cases} \frac{Z_i Z_j}{2} \sum\limits_{\mathbf{n} \neq \mathbf{0}} \frac{\mathrm{erfc}(\alpha |\mathbf{R_i} - \mathbf{R_j} + \mathbf{n}|)}{|\mathbf{R_i} - \mathbf{R_j} + \mathbf{n}|} & i = j \\ Z_i Z_j \sum\limits_{\mathbf{n}} \frac{\mathrm{erfc}(\alpha |\mathbf{R_i} - \mathbf{R_j} + \mathbf{n}|)}{|\mathbf{R_i} - \mathbf{R_j} + \mathbf{n}|} & i \neq j \end{cases} \tag{2.22}$$

where $\mathbf{n}$ is the lattice vector ($\mathbf{n} = h\mathbf{a} + k\mathbf{b} + l\mathbf{c}$). $\mathrm{erfc}(x) = 1 - \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2} du$ is called the complimentary error function , which decreases monotonically as x increases, indicating that the short range interaction decreases with an increasing distance between atoms. α is a constant number (dubbed as screening parameter) related to the density of atoms in the unit cell, which can be defined as $\alpha = \sqrt{\pi}(\frac{0.01N}{V})^{\frac{1}{6}}$, where N and V are the number of atoms and volume of a unit cell, respectively [165]. This suggests that higher atomic density results in less interactions from atoms in large distance. The long range interaction term is given by the following equation:

$$M_{ij}^{reciprocal} = \begin{cases} \frac{Z_i Z_j}{2\pi V} \sum\limits_{\mathbf{G} \neq \mathbf{0}} \frac{e^{-|\mathbf{G}|^{\mathbf{2}}/(\mathbf{2}\alpha)^{\mathbf{2}}}}{|\mathbf{G}|^{\mathbf{2}}} \cos\left(\mathbf{G} \cdot (\mathbf{R_i} - \mathbf{R_j})\right) & i = j \\ \frac{Z_i Z_j}{\pi V} \sum\limits_{\mathbf{G} \neq \mathbf{0}} \frac{e^{-|\mathbf{G}|^{\mathbf{2}}/(\mathbf{2}\alpha)^{\mathbf{2}}}}{|\mathbf{G}|^{\mathbf{2}}} \cos\left(\mathbf{G} \cdot (\mathbf{R_i} - \mathbf{R_j})\right) & i \neq j \end{cases} \tag{2.23}$$

where $\mathbf{G}$ represents the reciprocal space lattice vector. This term implies that a higher symmetry contributes more to the long range interactions between atoms. The self interaction term added correction beyond the Ewald summation, which enables it to describe charged unit cell, which is demonstrated in the following equation:

$$M_{ij}^{self} = \begin{cases} -(Z_i^2)\frac{\alpha}{\sqrt{\pi}} - (Z_i^2)\frac{\pi}{2V\alpha^2} & i = j \\ -(Z_i^2 + Z_j^2)\frac{\alpha}{\sqrt{\pi}} - (Z_i^2 + Z_j^2)\frac{\pi}{2V\alpha^2} & i \neq j \end{cases} \tag{2.24}$$

Although the Ewald sum matrix describes the local environment, the computational cost is relatively high compared with other matrix descriptors. Besides, the heterogeneity of the Ewald sum matrix of different materials cannot be avoided, which also limits its application. Nevertheless, it enlarged the scope of matrix descriptors and inspired the design of other vector-based descriptors.

A sine matrix considers periodic condition without requiring heavy computational cost to balance the computational cost and encoded information. However, this is at the cost of physics meaning of descriptors, *i.e.*, previous two descriptors are also meaningful for humans, but sine matrix only makes sense for computer. Its form can be expressed as the following equation:

$$M_{ij}^{sine} = \begin{cases} 0.5Z_i^{2.4} & i = j \\ Z_i Z_j |\mathbf{B} \cdot \sum\limits_{k=x,y,z} \mathbf{e}_k \sin^2 |[\pi \mathbf{B}^{-1} \cdot (\mathbf{R}_i - \mathbf{R}_j)]|^{-1} & i \neq j \end{cases} \tag{2.25}$$

where $\mathbf{B}$ is the lattice matrix, and $\mathbf{e}_k$ is a unit lattice vector in the Cartesian coordination.

In general, matrix descriptors are easy to construct, and they combine both chemical and structural information. Thus, they still play an essential role in representations. However, the computational cost of matrix descriptors is proportional to the square of the number of considered atoms, which requires a sizeable computational resource for an extensive system. As a result, their application is gradually replaced by more object-orientated representations, *e.g.*, SOAP, ACSF, crystal graphs, etc.

2.3.2. Crystal graphs

Crystal graphs use mathematical graphs to describe materials, as demonstrated in Figure 2.4. It was first proposed by Xie et al. in their work called CGCNN [35], where both sites and bond of crystalline materials is stored in the graphs as nodes and edges, respectively. Such graphs reflect the hierarchical nature of crystalline materials, which can be divided into 1) the properties of atoms and 2) the connection between the atoms:

- The nodes of graphs are represented as the feature vectors v_i of atomic properties:

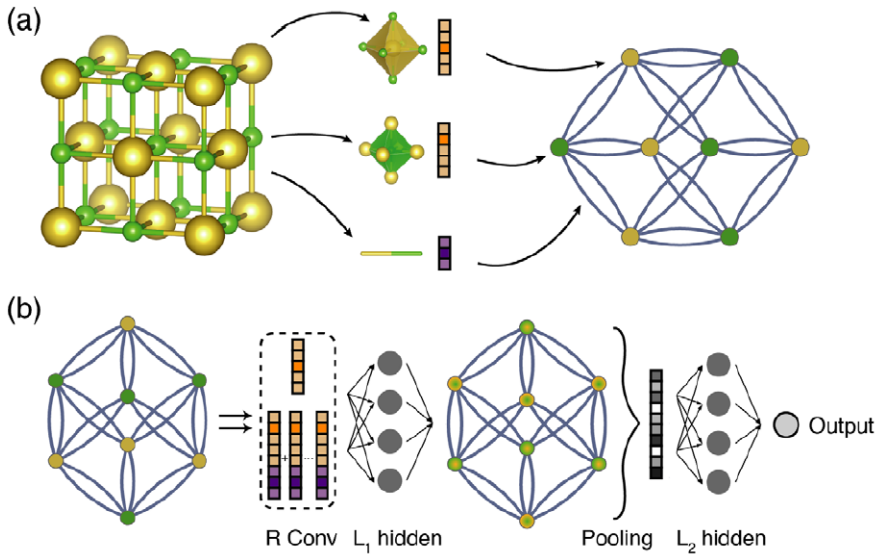


Figure 2.4.: Illustration of the crystal graph convolutional neural networks [35]. (a) Construction of the crystal graph, where both atoms and bonds are represented as one-hot features, and the one-hot features are stored in different channels. (b) Structure of the convolutional neural network on top of the crystal graph, where different channels of both atoms and bonds are considered in making the final prediction.

1. Group number
2. Period number
3. Electronegativity
4. Covalent radius
5. Valence electrons
6. First ionization energy
7. Electron affinity
8. Block
9. Atomic volume

To better fit in the channels of CGCNN, continuous properties are transformed into categorized one-hot features, while the discrete properties are able to metamorphose easily. A one-hot is a vector of bits (1 or 0) with only one bit being 1, where the location of 1 indicates the value [166]. For example, in a three-dimensional one-hot,

there are only three available vectors, *i.e.*, 110, 010, 001 where 100 is the largest and 001 is the smallest.

- The edges reflect the bonding interactions, which are constructed analogously to the atomic properties as one-hot feature vectors $u(i,j)_k$, representing the k^{th} bond between atom i and j. In order to demonstrate the periodicity of the crystal structures, the developed crystal graphs allow multiple edges between two nodes, enabling the connection between atoms with their counterparts in adjacent unit cells. Note that only the atomic distance is considered for the edges.

Crystal graphs overcome the heterogeneous problem by using one-shot features. This avoids further processing of representations and increases the efficiency of forward modelling, which can predict intrinsic properties accurately [35, 88]. However, the noncontinuous nature of crystal graphs prevents their usage in inverse design.

2.3.3. Image-based continuous representations

Image-based continuous representation can be used in both forward modelling and inverse design. As mentioned previously, Noh et al. developed the image-based continuous representation to achieve the reversibility of representation (to crystal structures) [75], as illustrated in Figure 2.5. Specifically, three-dimensional (3D) voxel grids are applied to crystal structures. Taking a binary compound as an example, two grids are saved to record the atomic positions of two elements separately, and the third one to store the lattice constants, *i.e.*, the lengths and angles of/between them. Hereafter, such grids are referred as site voxel and lattice voxel, respectively. Using the probability density based on Gaussian functions, the lattice voxel is defined as:

$$p_{x,y,z} = e^{-\frac{\mathbf{r}_{x,y,z}^2}{2\sigma^2}} \tag{2.26}$$

where x, y, z denote the grid index, p is the probability density of the denoted grid, $\mathrm{r_c}$ is the distance between this grid to the center of unit cell, and σ denotes the standard deviation of a Gaussian function. Similarly, sites voxel is defined as:

$$p_{x,y,z} = e^{-\frac{\mathbf{r}_{n,x,y,z}^2}{2\sigma^2}} \tag{2.27}$$

where n is number of atoms and r is the distance between this grid point and the atom. In this way, the inverse transformation is trivial for the lattice voxel while that for the sites voxel relies on the image filter technique.

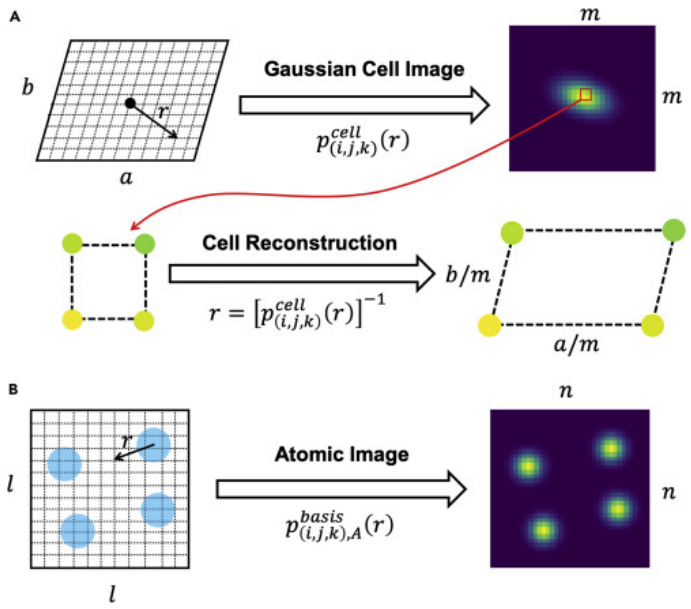


Figure 2.5.: Image representation of crystal structure and inverse transform [75]. Gaussian functions are used to transform the lattice constants and atomic positions into the crystal images, where the inverse process can transform them back to their original formats.

Autoencoder can encode inputs into low-dimensional vectors, as well as decode the vectors back. In this work, the autoencoder is applied to encode 3D voxel data into one-dimensional vectors in the latent space, which is realized based on 3D CNN autoencoder. The loss function of autoencoder consists of two parts: the first is the loss of information, *i.e.*, the difference between inputs and outputs, and the second is the regulation term to prevent over fitting, which yields

$$loss = |\mathbf{x} - \mathbf{y}|^2 + \frac{\lambda}{2}\sum \boldsymbol{\omega}^2 \tag{2.28}$$

where $\mathbf{x}$ and $\mathbf{y}$ are the input and output of the autoencoder, λ is a regulation coefficient, and $\boldsymbol{\omega}$ is weight in the autoencoder.

Continuous representation can be used in both inverse design and forward modelling [135]. Since the continuous representations can reconstruct the original structures, they should contain all the materials' essential chemical and structural information. Thus, the continuous representations are one-to-one mappings of real materials in latent space, *i.e.*, a newly generated continuous representation can be reverted to a crystal structure. This enables the inverse design in the latent space. Due to the same reason, the continuous

representations can be directly used to predict the materials' intrinsic properties, enabling their application in forward modelling.

2.4. Summary

In this chapter, three categories of representations have been discussed, *i.e.*, chemical representation, structural representation, and combined representation. In practice, representations should be selected based on the data and the objectives. MAGPIE is a chemical representation generated by only chemical composition, which can be used as inputs to predict intrinsic properties. SOAP and ACSF are the structural representation calculated by structures, and they can be used as inputs for the direct prediction of intrinsic properties for similar structures. The Matrix descriptors (including Coulomb, Ewald summation, and sine matrix) are combined representations, enabling a prediction of properties for structures with a significant difference at high computation cost. In addition, crystal graphs use nodes and edges in graphs to reflect the atoms and bonds in materials, respectively, enabling accurate prediction for intrinsic properties. Last but also most important, the image-based continuous representation can represent both chemical and structural information of crystal structures as images. Meanwhile, they serve as a one-to-one mapping for materials in a latent space, enabling the reconstruction of crystal structures from the representations.

3. Machine learning models

ML models are mathematical equations depicting the relationship between inputs and outputs, where the parameters of the equations can be determined by the existing samples. However, as the no free lunch theorem indicates: there is no common solution that is suitable for all problems [138]. Thus, the selection of ML models depends on the available representations as well as the objectives. As mentioned in Chapter 1, two objectives are 1) establishing forward models to predict intrinsic properties with crystal structures as inputs and 2) building inverse design to directly generate crystal structures with target properties. So in this chapter, commonly used methods in forward modelling and their advantages and disadvantages are firstly discussed. Then how inverse models work and how inverse models could combine with forward models to design new materials with desired properties are illustrated.

3.1. Forward models

3.1.1. Support vector machines

SVM [98] is one of the most robust supervised learning models. It constructs a hyperplane (or sets of hyperplanes) to separate two (many) clusters of data in a high-dimensional space, as demonstrated in Fugure 3.1, which can be utilized for classification, regression, or outliers detection [155]. Traditionally, an SVM classifier assigns one category to the inputs for binary classification, making it a non-probabilistic linear binary classifier. Although separations can be achieved in numerous approaches, a hyperplane that has the largest functional margin, *i.e.*, distance to the nearest training point, is more desirable. Taking the binary classification as an example, the objective is to maximize the distance between the hyperplane and the point x_i from either group. Such a hyperplane (dubbed as linear classifier) can be represented as a set of points x subject to $w^T x - b = 0$, where w is the weights in this model, and the loss function can be expressed as:

$$l = \frac{1}{n}\sum_{i=1}^{n}[1 - y_i(w^T x_i - b)] + \lambda|w|^2 \tag{3.1}$$

where λ is the hyperparameter to avoid overfitting, l is the loss, and y_i are either 1 or -1 indicating the belonging class.

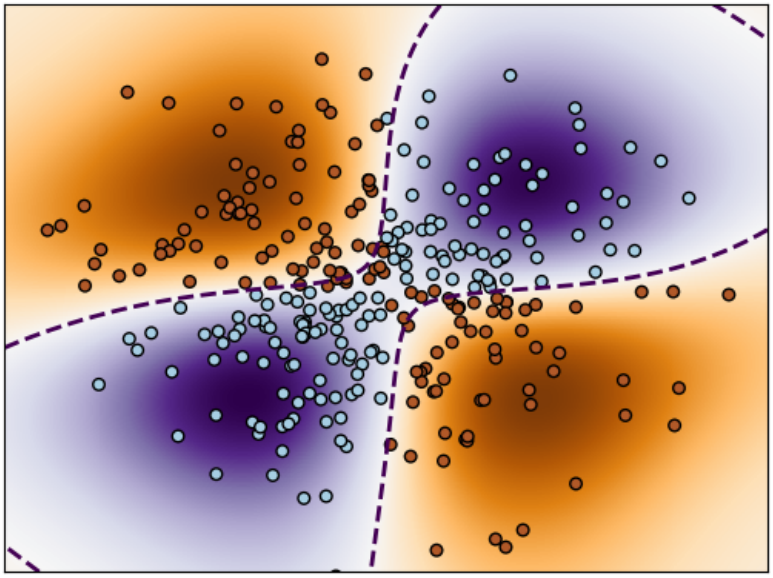


Figure 3.1.: Illustration of SVM [155], where the cyan and brown points represent samples with different labels, and the dash line denotes the hyperplane discovered by SVM to separate them. Depending on the applied kernels, the hyperplane can be either linear or non-linear.

In addition to linear classifiers, a non-linear classification is more common in practice, where SVM can use kernels to map their inputs into high-dimensional feature spaces [167]. It can be expressed mathematically as the following equations:

$$l = \frac{1}{n}\sum_{i=1}^{n}[1 - y_i(\sum_{j=1}^{n} w^T k(x_i, x_j) - b)] + \lambda|w|^2 \tag{3.2}$$

where $k(x_i, x_j)$ is the kernel function. Apart from this, regression by SVM is also possible, but objective term, $1 - y_i(\sum_{j=1}^{n} w^T k(x_i, x_j) - b)$, becomes the distance between points,

as defined in the following equations:

$$l = \frac{1}{n}\sum_{i=1}^{n}[\sum_{j=1}^{n}|w^T k(x_i, x_j) - b - y_i|] + \lambda|w|^2 \tag{3.3}$$

In general, SVM models can do classification and regression, which can serve as a promising approach for forward modelling. Specifically, both non-linear classification and regression have endowed SVM models with the ability to examine the complicated relationship between structures and properties. However, it is noteworthy that the explosive computational cost of kernels has prevented the usage of a large amount of data. Therefore, the SVM models are more efficient at exploiting intrinsic properties and crystal structures relationship with a limited number of data.

3.1.2. Bayesian method

The Bayesian method refers to the BO with the Gaussian processes [168]. Technically, the BO is one optimizer to search for the global optimum of a black-box function without assuming a functional form [100], as demonstraded in Figure 3.2. To be specific, it places a prior distribution (which will be updated to form a posterior distribution) over an objective function, whose mathematical expression is $\max_{x\in A} f(x)$, where A is a set of data points. Though $f(x)$ is sophisticated and hard to be evaluated, the prior/posterior distribution is usually determined by the Gaussian processes [168]. Thus, the Bayesian method transfers the fitting of objective functions into the uncertainty quantification of distribution functions [169].

A Gaussian process is a stochastic process to determine multivariate normal distribution of random variables, *i.e.*, the estimation of the value of a certain data point and its uncertainty based on the measurement of similarity between data points (the kernel function). It can be mathematically expressed as $f(x) \sim N(0, K(\theta, x, x'))$, where x are the data points, $f(x)$ is the objective function, $K(\theta, x, x')$ is the covariance matrix between all possible pairs (x, x') for a given set of hyperparameters θ. The log possibility of the evaluation is defined as:

$$\log p(f(x) \mid \theta, x) = -\frac{1}{2}f(x)^T K(\theta, x, x')^{-1} f(x') - \frac{1}{2}\log\det(K(\theta, x, x')) - \frac{n}{2}\log 2\pi \tag{3.4}$$

and the training process is to maximize the possibility by tuning the unobserved values $f(x^*)$ at coordinates x^* from the predictive distribution $p(y^* \mid x^*, f(x), x) = N(y^* \mid A, B)$ where the posterior mean estimate A is defined as:

$$A = K(\theta, x^*, x)K(\theta, x, x')^{-1} f(x) \tag{3.5}$$

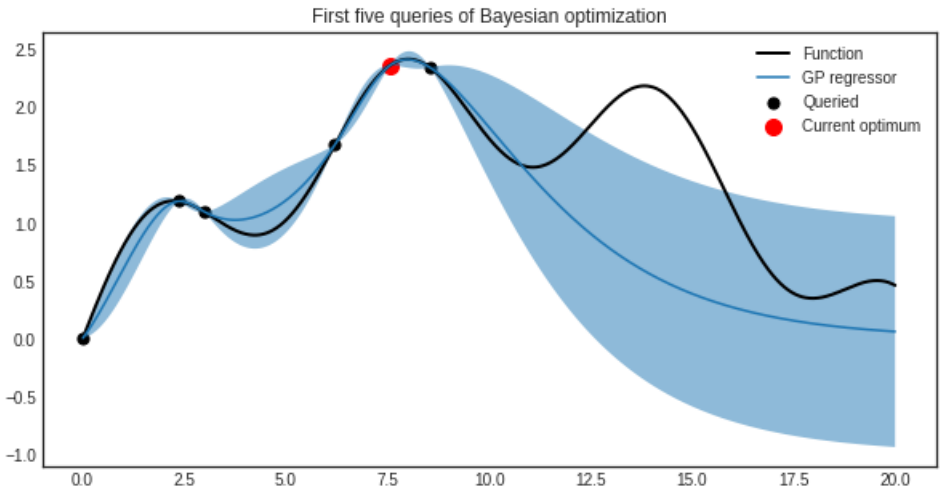


Figure 3.2.: Illustration of Bayesian methods [170], where the black line represents the target function, the blue line denotes the prediction of Bayesian methods, the black points are the known samples. After getting a few known samples, one can already assume that the distribution of known samples complies with Gaussian distributions, but the exact distribution format is hard to get. As a result, by combining the possible distributions, it can be used to predict the target as well as evaluate the uncertainty.

and the posterior variance estimate B is defined as:

$$B = K(\theta, x^*, x^*) - K(\theta, x^*, x)K(\theta, x, x')^{-1}K(\theta, x^*, x)^T \tag{3.6}$$

where $K(\theta, x^*, x)$ is the covariance between the new coordinate of estimation x^* and all other observed coordinates x, $K(\theta, x^*, x^*)$ is the variance at point x^*.

The Bayesian method can make predictions as well as evaluate the reliability of the results. Although there is no additional term in the Bayesian method to prevent overfitting like SVM, the Bayesian method can evaluate the likelihood of the results. This avoids the deception of high accuracy. It has been extensively applied on determining experimental parameters to produce better materials [34]. However, the usage of kernels also limits the number of training data since the dimension of a kernel is the square of the samples.

3.1.3. Random forests

RF has been extensively used in ML for materials science. RF is the ensemble ML method by firstly constructing a large number of decision trees and then the evaluation of final results through voting, *i.e.*, the bootstrap aggregating of individual trees [99], as demonstrated in

Figure 3.3. Intuitively, the RF model can only be used to be applied for classification. Yet, the ensemble method accommodates it with the ability to do regression as well, which can be mathematical represented as:

$$\min \sum_{i} |y_i - \hat{f}_i| \tag{3.7}$$

where the $\hat{f}_i$ can be defined as:

$$\hat{f}_i = \frac{1}{B} \sum_{b=1}^{B} f_b(x_i) \tag{3.8}$$

where (x_i, y_i) are the training samples, b denotes the label of decision trees, B is the total number of estimators, and each estimator randomly takes a subset of the total training set.

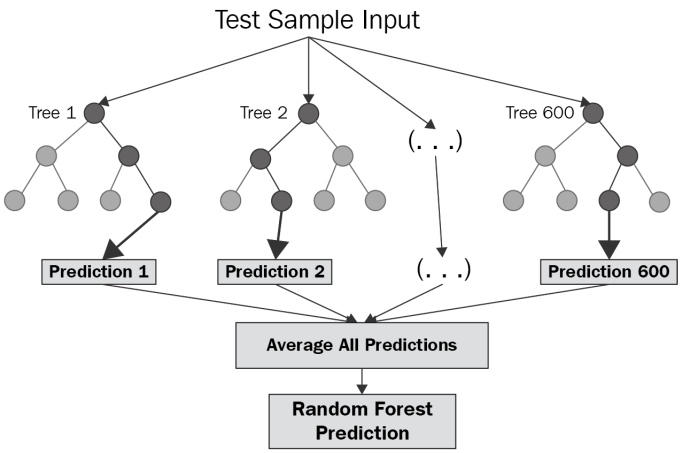


Figure 3.3.: Illustration of random forest [171]. The random forest model comprises of a large number of decision trees, each tree will use limit number of features to predict the target results. But the final output of the model is evaluated by doing weighted average of the predicted values of all trees.

However, as shown in Eq.3.8 and 3.9, the factors that dominate the prediction results are divided into two: the accuracy of the decision tree and their weights. The former is at the mercy of hyperparameters like the depths of decision trees or the number of leaves of each branch, while the latter is called feature bagging. Traditionally, the weights are determined by the number of decision trees using the same features over the total

number of decision trees in the forest, which can be further assisted by the standard deviations [172], as the following equation demonstrated:

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x) - \hat{f})^2}{B-1}} \tag{3.9}$$

where the number of decision trees is a hyperparameter that should be determined by validation [173]. Further, the extremely randomized trees (ExtraTrees) are similar to ordinary RF with two main differences: 1) each decision tree is trained by the whole learning samples, and 2) the top-down splitting in the tree learner is randomized [174].

In general, the RF model is the most adaptive method, which can be used for classification or regression without requiring a large number of training data. For example, an accuracy of more than 90% can both be achieved by RF models with the size of their databases varying from 100 [175] to 2000 [32]. Besides, it can work on the string type data if necessary, making it one of the most common methods. For instance, true or false is also a valid input for RF models. However, RF is best suited in interpolation tasks because the decision trees cannot generate an unknown category in the training samples.

3.1.4. Artificial neural networks

ANN is currently the most popular ML method, inspired by the biological neural networks with connected neurons [101]. Similarly, ANN is based on a collection of connected artificial neurons, which receive, process, and transmit signals from/to the connected neurons. A signal is a real number, and an edge (processing) is conducted by non-linear functions, *i.e.*, take inputs and predict outputs. ANN models are also simply referred as neural networks (NN) in practice. The training process of ANN is to optimize the parameters in the processing procedure. A critical feature for the connections of neurons in ANN is exhaustive and one-direction, so the most straightforward architecture of ANN can be described by only two hyperparameters, *i.e.*, number of layers and the number of neurons on layers, as demonstrated in Figure 3.4. To be more specific, the neurons in the previous layer are connected to all neurons in the layer by edges, and all neurons in this layer will also connect to all neurons in the next layer (the first layer in an ANN model is referred as the input layer, the last layer is called output layer, and others are denoted as hidden layers).

Mathematically speaking, for a given set of artificial neurons in one layer denoting k, providing $m+1$ input signals from x_0 to x_m and edge weights from w_{k0} to w_{km}, where

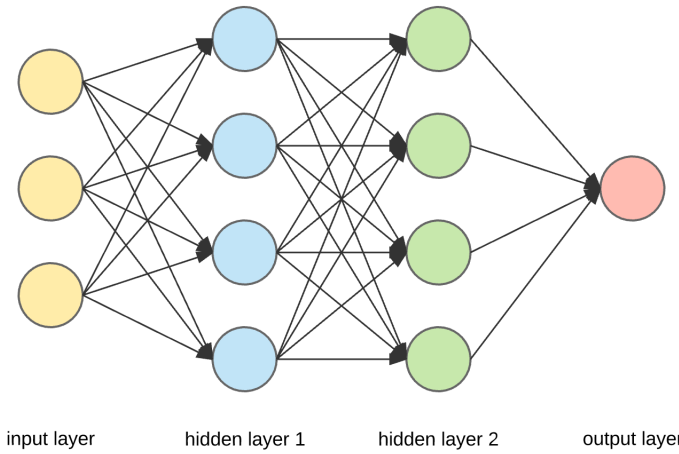


Figure 3.4.: Illustration of an ANN model with 2 hidden layers and 4 neurons on each hidden layer [176]. The input layer is responsible for collecting the input data, hidden layers are used to process the data and pass the information into the next layer, then the output layer will generate the predicted results.

bias is defined as $w_{k0} = b_k$ and x_0 is assigned to be 1, so the output of a neuron is:

$$y_k = \varphi\left(\sum_{j=0}^{m} w_{kj} x_j\right) \tag{3.10}$$

where φ is the activation function to add non-linearity into the ANN model. Otherwise, no matter how much the number of layers and neurons are, the ANN model could only catch the linear relationships. The commonly used activation functions include but are not limited to: sigmoid, tanh, rectified linear unit (ReLU), and leaky rectified linear unit (Leaky ReLU). The sigmoid function can be defined as the following equations [177]:

$$\varphi(x) = \frac{1}{1 + e^{-x}} \tag{3.11}$$

where φ denotes the activation function, x is the input of the neuron. No matter how large or small the x would be, the given value can only be in the range of 0 to 1, which enables the usage of sigmoid activation in the classification model. Similarly, the tanh activation function has precisely the same function, with the only difference being the output range (from -1 to 1), which can be expressed as the following equation [178]:

$$\varphi(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{3.12}$$

For the activation function in the hidden layers, the ReLU [179] and leaky ReLU [180] are the most commonly used functions, which can be defined as the following equations:

$$\varphi(x) = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases} \tag{3.13}$$

$$\sigma(x) = \begin{cases} \alpha x & x < 0 \\ x & x \geq 0 \end{cases} \tag{3.14}$$

Despite their simple form, they both mimic the activation threshold of biological neurons, which perform extremely well in NN.

However, the edge weights in artificial neurons should be modified during the training process by an optimizer. As demonstrated in Section 1.3.4, the training process of ML models has two aspects: calculation of loss functions and modification of weights by optimizers. The essence of optimizers is evaluating the contributions of each weight to the loss function and making modifications, *i.e.*, calculate partial derivatives with respect to the weights, and modify the weights correspondingly. When the ADAM optimizer is used, the training process can be mathematically described as:

$$w^{(t+1)} = w^{(t)} - \eta \frac{\hat{m}_w}{\sqrt{\hat{v}_w} + \epsilon} \tag{3.15}$$

where *t* indicates the number of iteration, *w* is the weights to be optimized, η is the learning rate. $\hat{m}_w$ and $\hat{v}_w$ are called the momentum which can be defined as:

$$\hat{m}_w = \frac{m_w^{(t+1)}}{1 - \beta_1} \tag{3.16}$$

$$\hat{v}_w = \frac{v_w^{(t+1)}}{1 - \beta_2} \tag{3.17}$$

where β_1 and β_2 are called forgot parameter to determine how many step in the past will be considered, $m_w^{(t+1)}$ and $v_w^{(t+1)}$ can be further determined as:

$$m_w^{(t+1)} = \beta_1 m_w^{(t)} + (1 - \beta_1) \nabla_w L^{(t)} \tag{3.18}$$

$$v_w^{(t+1)} = \beta_2 v_w^{(t)} + (1 - \beta_2) \nabla_w L^{(t)} \tag{3.19}$$

where *L* is the loss.

ANN has become the most attractive ML model. Given enough data, ANN is the most accurate model to reveal the nonlinear relationships due to the delicate design of artificial neurons. When the number of layers and neurons is increased, the accuracy of prediction also becomes higher. However, increasing in ANN size also results in signal degradation in optimizers, which causes inefficient improvements during the training processes. Thus, ANN models advocate for new techniques like normalization or dropout layers [181] to overcome the challenges.

Such techniques also lead to the birth of deep learning. To be more specific, NN models with more than three layers areare usually referred to deep learning [182] due to their superiority in prediction compared with shallow ML models. Thus, the previously discussed models are all shallow models, where the following models are deep learning models. The deep learning models not only increase the size of NN models but also revolutionize the architecture of the models. For instance, CNN, RNN, transformer and etc., have different ways to connect neurons, and therefore can be used in image recognition and NLP.

3.1.5. Convolutional neural networks

CNN is one deep learning NN specially designed to process images. It uses convolution techniques to replace the matrix multiplication in the fully connected layers of ANN in at least one layer [183], as illustrated in Figure 3.5. The emergence of the convolution techniques is originated from the challenges faced by ANN: once the position of the target patterns is shifted to other positions, the trained models are no longer able to identify it [184], *i.e.*, the ANN can only recognize images with fixed positions of target patterns, which are not the cases in reality. Where a convolutional layer is designed to identify the trained pattern (known as feature maps) from every local subsection of images in the training and latter search for this pattern from every subsections of new images [185], which in general has three attributes:

- Convolutional filters/kernels with certain width and height, *i.e.*, identification of the pattern with the size of the kernels.
- The number of filters, which is reflected by the number of channels, where the number of input channels of a layer should equal to the number of channels of the previous layer.
- The selection of subsection by padding and stride, which dictates the process of the boundaries of images and the distance between different subsection.

However, the width and height of kernels, or the number of channels, as well as the paddings and strides, are all hyperparameters, which can only be determined in practice

according to the no free lunch theorem [138]. Mathematically speaking, the convolutional layer performs a dot product of the kernel with the subsection of the input matrix, and the results are further calculated by activation functions as demonstrated in the following equation:

$$X_j^l = f(\sum_{i \in M_j} X_i^{l-1} \dot{W}_{ij}^l + b_j^l) \tag{3.20}$$

where X_j^l is the output of the l^{th} layer and the j^{th} feature map, W and b indicate the weights and bias, respectively, f is the activation function.

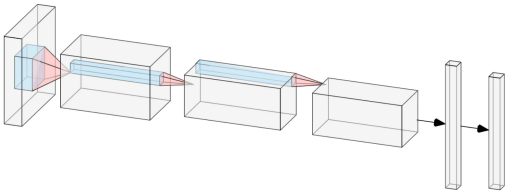


Figure 3.5.: Topology of a convolutional neural network starting with convolutional layers with multiple filters followed by pooling and two fully connected layers [29]. The convolutional layers could identify the target pattern from images, but patterns with larger size cannot be recognized. However, the pooling layer will shrink the size of the images to make the pattern recognizable for the filters.

Nevertheless, other techniques are also required. Firstly, the pooling layers are used to normalize the image pattern. Providing a pattern in one image is larger than the kernel size, the convolutional layer alone cannot recognize the pattern [185]. The pooling layer condenses pixel blocks on the image into smaller ones, where the patterns in the new images have the potential to be fitted in the feature maps. Secondly, to prevent overfitting, the dropout layers and batch normalization layers are added to randomly disconnect edges between some neurons and normalized the outputs of one layer to the range of 0 to 1, respectively [181]. The mechanism of dropout is to avoid the extensive dependency on a specific neuron, while batch normalization is to reduce the dependency between layers. The final few layers in CNN are usually the fully connected layer as ANN to achieve the final prediction of outputs.

CNN has strong advantages in visual feature extraction. CNN achieves high accuracy, efficiency, and adaptability in image learning problems when compared with all previously mentioned ML methods. Given the fact that crystalline materials can be expressed as images, the application of CNN has provided valuable insights for forward modelling [35]. However, the nature of deep learning endows CNN with millions of parameters to be determined, which also requires the highest computational cost. Based on the same reason, the risk of overfitting is also high, which calls for tremendous efforts in hyperparameter

tuning.

3.1.6. Transformer

A transformer is a deep learning model utilizing the attention mechanism [186]. Essentially, transformers use attention to extract information from each subsection of the inputs. However, unlike RNN and CNN models, transformers neither require the sequential processing of sentences nor the segmentation of images. Instead, the attention operation combines every two contexts/pixels in the input sequences/images. Therefore transformers can extract the implicit information from the inputs, which increases the prediction accuracy compared with RNN and CNN. However, this also makes the transformers computationally expensive, which calls for parallelization training to reduce training time.

The fundamental building block of the transformer is the attention layer, where the scaled dot-product attention units serve as the counterpart of convolutional kernels in CNN, as illustrated in Figure 3.6. Taking NLP as an example, attention weights are calculated by tokens in input sentences. To be more specific, each attention units consist of three weight matrices: the query weights W_Q, the key weights W_K, and the value weights W_V, where the query vector $q_i = x_i W_Q$, key vector $k_i = x_i W_K$, and value vector $v_i = x_i W_V$ can be calculated from the input word embedding x_i, respectively. Attention weights can then be calculated by the query and key vectors, *i.e.*, the attention weight a_{ij} for token i from token j is the dot product of q_i and k_j, where the results are divided by the square root of the dimension of the key vectors, $\sqrt{d_k}$, and normalized by a softmax to sum to 1 [187]. Since the commutative law is not satisfied by the dot product of matrices in general, the attention weights of a_{ij} and a_{ji} are not the same, enabling the non-symmetric affection between token i and token j. The attention calculation can be expressed as the following equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{3.21}$$

The matrices Q, K and V are defined as the query, key, and value matrices where the i^{th} rows are vectors q_i, k_i, and v_i, respectively.

Transformers have become highly efficient for NLP problems [188], with the parallelization training implemented. This enables the application of NLP on large databases. The techniques to evaluate connections between tokens can also be applied on the cloud data like atomic positions in molecules, with the only drawback being the high computational cost. However, the development of SE(3)-Transformers has overcome such drawbacks by only considering chemical bonds [189].

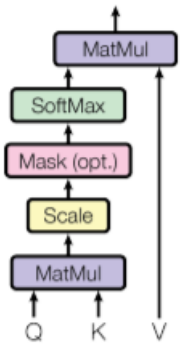


Figure 3.6.: Illustration of attention [104], where K, Q, and V represent the query, key, and value matrices, respectively. Query and key will be used to calculate the connection between the interested subsection with all the subsections, while value store the information of all the subsections. By doing the attention process, the information of other related subsections will be incorporated into the interested subsection.

3.2. Inverse design models

3.2.1. Variaitonal autoencoder

VAE is one generative ML model composed of both encoder and decoder, with a prior assumption of the distribution of the encoded data [190]. Since the VAE is a variation of autoencoder, an ordinary autoencoder should be scrutinized first. An autoencoder could be used as a dimension reduction method to compress information efficiently, *i.e.*, the autoencoder endeavors to minimize the reconstruction error by searching for the best pair of encoder and decoder. Mathematically, the objective of the autoencoders can be represented as:

$$\underset{(e,d)\in E\times D}{min}\ \epsilon(x, d(e(x))) \tag{3.22}$$

where e and d represent the encoder and decoder, respectively, $\epsilon(x, d(e(x)))$ defines the reconstruction error between the input x and the output $d(e(x))$. However, when both encoders and decoders adopt the neural networks, the autoencoder backpropagates the error through the architecture (the encoder followed by the decoder) to update the weights of the networks. The loss function is simplified to:

$$\text{loss} = |x - d(e(x))|^2 \tag{3.23}$$

where $e(x)$ represents data points on a latent space. When manipulating the vectors in the latent space, they could be decoded to potentially new crystal structures, no matter how the latent space is organized.

However, a randomly distributed latent space cannot guarantee the generation of valid data. In this case, a regularized latent space is supposed to be established. Thus, VAE can be an autoencoder whose training is regularized by a prior defined distribution function to ensure the continuity of the latent space, as demonstrated in Figure 3.7. This can be mathematically expressed as

$$\text{loss} = |x - d(e(x))|^2 + D_{KL}(N(\mu_x, \sigma_x), N(0, 1)) \tag{3.24}$$

where D_{KL} is called as Kullback–Leibler divergence [191]. It can evaluate the difference between two distributions, *i.e.*, the smaller the value is, more likely these two distributions are the same. Kullback–Leibler divergence can be defined as:

$$D_{KL}(P, Q) = \int_{-\infty}^{\infty} p(x) log(\frac{p(x)}{g(x)}) dx \tag{3.25}$$

where P and Q are distribution functions and $p(x)$ and $g(x)$ are their density functions.

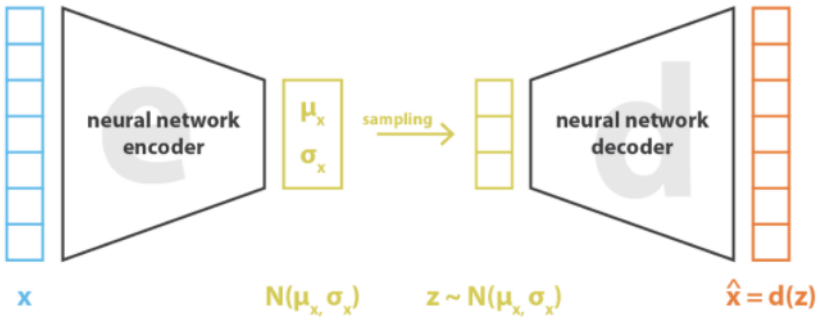


Figure 3.7.: Illustration of a VAE [192], where the input x can be encoded into z that comply with the Gaussian distribution in a latent space. The VAE model does not only encode the information into the latent space, it also enforces the encoded information to comply with a certain distribution. This enables the generation with pure interpolation in the latent space.

Unfortunately, the distribution function is not prior known to us, whose form demands profound knowledge in statistics and a comprehensive understanding of the input data. On the one hand, such assumptions of the distribution function integrate the physics and chemistry insights into the VAE model, which also attracts intensive attention from material science [75, 88]. On the other hand, an assumption makes the training of VAE relatively cheap compared with models not specifying a distribution. Although the deep learning nature of VAE brings high training cost, which usually takes longer time than the forward modelling, the prior determined distribution reduces the number of parameters in generative models.

3.2.2. Generative adversarial networks

GAN is another type of generative model. The main feature of GAN is to replace the direct comparison of distribution function into an indirect competition between two neuron networks called generator and discriminator, whose objective is to make the generated distribution closer to the true distribution [110]. A discriminator takes samples of existing and generated data as inputs in a GAN architecture, trying to identify their origins. Moreover, a generator fools the discriminator by generating samples statistically similar to the existing data. Both networks endeavor to achieve their goals and perfect their performance during the training process [193]. From the perspective of game theory, the generator and discriminator are trying to reach a Nash equilibrium in a two-players game [194]. Mathematically speaking, the generator minimizes the difference between the existing and generated data, while the discriminator maximizes such difference, which can be expressed as the mathematical formula:

$$\max_{D}(\min_{G}(\frac{1}{2}E_{x\sim p_t}(1-D(x))+\frac{1}{2}E_{x\sim p_g}D(x))) \tag{3.26}$$

where D and G represents the discriminator and generator, respectively, p_t and p_g are the possibility density function of the existing data and the generated data.

In general, the main advantage of GAN over VAE is no requirement of the prior defined distribution. Thus, the GAN model has been extensively used in inverse design [33, 195]. However, the GAN models are relatively fragile because the generators in GAN are NN models instead of a fixed distribution function. Distribution functions have only two or three parameters to be determined, and their outputs remain reasonable even if parameters are not optimized. In contrast, NN models have to determine a large number of parameters, and the corresponding distribution can potentially lead to unreasonable generations without optimized parameters.

3.3. Summary

In this chapter, the ML models for both forward modelling and inverse design are introduced. For forward modelling, SVM and RF are best suited for classification tasks, while ANN and BO are better used for regression tasks with vector-based inputs. CNN and transformer can take images and sentences as inputs, respectively, and conduct regression tasks. For the inverse design models, both VAE and GAN can use the trained forward models as part of the objective functions and design new materials.

ML models can be separated into shallow ML models and deep learning models as well. The NN models with larger than three layers are called deep learning models, and all

the other models are referred to as shallow ML models. The shallow ML models are less accurate but require fewer training samples. In contrast, the deep learning models are more powerful but require more training data.

4. Publications

4.1. An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature

For magnetic materials, there are three intrinsic magnetic properties: magnetization, magnetic anisotropy energy, and critical ordering temperature [82]. The $\mathrm{T_C}$ is the critical ordering temperature of the magnets transforming from paramagnetism to ferromagnetism, *i.e.*, the local moments are randomly distributed above it, while all aligned below it. [196]. To design magnetic materials, DFT calculations can be applied to evaluate the magnetization and magnetic anisotropy energy in a high-throughput manner [197], but the evaluation of $\mathrm{T_C}$ is complex. There are, in principle, two ways to obtain the critical magnetic ordering temperatures. One is to solve the Hubbard model, *e.g.*, based on DFT+dynamical mean-field theory (DMFT), to look for the temperature of vanishing ordered moments [198]. The other is to project the electronic structure onto an effective spin model (*e.g.*, Heiserberg model) followed by Monte Carlo modelling [82]. However, both methods are numerically expensive and with limited predictive power, so it is of great significance to evaluate $\mathrm{T_C}$ efficiently, which can further accelerate the discovery of new ferromagnets with high $\mathrm{T_C}$.

In this work, 1749 ferromagnetic and 1056 antiferromagnetic materials are collected, and the $\mathrm{T_C}$ of ferromagnetic materials are trained using ordinary RF models to distinguish the ferro/antiferro-magnetic materials as well as to predict $\mathrm{T_C}$ for ferromagnetic materials based on a two-step RF method. The resulting accuracy is 87% for the classification and 91% for the regression, where the $\mathrm{T_C}$ of 85 intermetallic materials have further been applied to validate the regression with a resulting accuracy of 91%. This work provides a practical solution to accelerate the evaluation of $\mathrm{T_C}$ for ferromagnetic materials, which paves the way to expedite the discovery of ferromagnetic materials.

MATER. RES. LETT.
2021, VOL. 9, NO. 4, 169–174
https://doi.org/10.1080/21663831.2020.1863876

MATERIALS RESEARCH LETTERS
Taylor & Francis
Taylor & Francis Group

REPORT

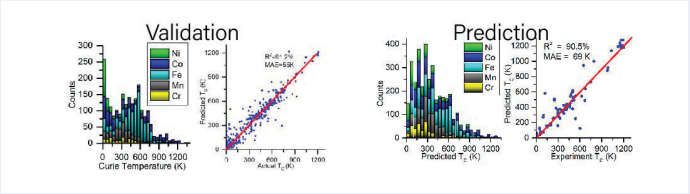
An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature

Teng Long, Nuno M. Fortunato, Yixuan Zhang, Oliver Gutfleisch and Hongbin Zhang

Institute of Materials Science, Technische Universität of Darmstadt, Darmstadt, Germany

ABSTRACT

Magnetic materials have a plethora of applications from information technologies to energy harvesting. However, their functionalities are often limited by the magnetic ordering temperature. In this work, we performed random forest on the magnetic ground state and the Curie temperature (T_C) to classify ferromagnetic and antiferromagnetic compounds and to predict the T_C of the ferromagnets. The resulting accuracy is about 87% for classification and 91% for regression. When the trained model is applied to magnetic intermetallic materials in Materials Project, the accuracy is comparable. Our work paves the way to accelerate the discovery of new magnetic compounds for technological applications.



ARTICLE HISTORY
Received 22 June 2020

KEYWORDS
Magnetic ground state; Curie temperature; machine learning; random forest

IMPACT STATEMENT

A machine learning model to classify ferromagnetic and antiferromagnetic compounds and to predict the Curie temperature of the ferromagnets is constructed, which could accelerate the discovery of new magnetic compounds.

Introduction

Magnetic materials have a wide spectrum of applications, particularly in efficient energy utilization [1]. Specifically, permanent magnets (PMs) are the key components for energy related technologies, such as conventional generators, e-mobility, automatization and refrigeration [2]. Moreover, ferromagnetic (FM) materials have been widely applied in spintronics, such as sensing, memory and logic, whereas the emerging antiferromagnetic (AFM) spintronics have recently drawn intense attention [3]. Three fundamental intrinsic properties desired for promising candidate magnetic materials are a FM ground state, a high Curie temperature (T_C) and the magnetocrystalline anisotropy (MCA). These properties are also important for magnetic refrigeration which promises enhanced energy efficiency over the conventional cooling technologies [4]. In this work, we focus on the machine learning modeling of the magnetic ground state and the T_C, as both the magnetization and MCA can be obtained via density functional theory (DFT) calculations for most compounds even in a high-throughput manner [5] and using machine learning [6], though evaluation of MCA is numerically expensive.

Although T_C is experimentally measurable, synthesis, characterization and optimization of a wide range of real materials are time-consuming and costly. Thus, a methodology to accelerate the development of magnetic materials with a theoretical pre-screening is of obvious interest. Typical theoretical approaches to evaluate T_C rely on the parameterization of DFT electronic structure to construct a Heisenberg Hamiltonian. This approach has good accuracy in predicting the Curie temperature of magnetic compounds in the localized limit (e.g. bcc Fe), but may fail for systems with tricky origin

CONTACT Hongbin Zhang hzhang@tmm.tu-darmstadt.de Institute of Materials Science, Technische Universität of Darmstadt, Darmstadt 64287, Germany

This article has been republished with minor changes. These changes do not impact the academic content of the article.

Supplemental data for this article can be accessed here. https://doi.org/10.1080/21663831.2020.1863876

of magnetism (e.g. fcc Ni) [7–9]. Moreover, DFT is insufficient in describing the strong-correlated 4f electrons in rare-earths [10], while the orbital dependent functional (e.g. DFT + U) treatment is often chosen to fit to experiments. The state-of-the-art DFT plus dynamical mean field theory (DMFT) method can be applied to tackle the electronic correlation problem, which is one of the most sophisticated methods to evaluate the magnetic properties. It provided many successful predictions, [11] but the T_C evaluated for bcc Fe based on DFT + DMFT ranges from 840 K to 1800 K, due to different ways to treat the Coulomb interactions, and the selection the U and J parameters [12,13]. Besides, DFT + DMFT is numerically expensive, thus it cannot be used for extensive predictions. Therefore, there is a great impetus for a predictive approach to obtain T_C, which is applicable to compounds with arbitrary compositions and crystal structures.

Machine learning is an emerging tool in materials science, being applied successfully to model the thermodynamic stability [14], band gap [15], inter-atomic potentials [16] and in predicting potential high temperature superconductors [17]. However, regression models to predict ordering temperature of magnetic materials have only been reported in a limited scope while classification models to distinguish AFM and FM are absent in literature to the best of our knowledge. Sanvito et al. trained a linear regression model using 40 intermetallic Heusler alloys (with experimental T_C), and made predictions for another 20 compounds. Through experimental validation, they discovered Co_2MnTi with a remarkably high T_C of 900 K [18]. Dam et al. focused on selecting the best features for fitting T_C of binary 3d-4f intermetallic compounds by applying Gaussian kernel regression on 108 compounds. The add-one-in test accuracy can reach above 95% when only eight descriptors are used, with the rare-earth concentration being the most relevant [19]. Nelson et al. demonstrated that only chemistry is required to model the T_C, where the accuracy is about 88% with a MAE of 50 K [20].

In this work, we develop a FM/AFM classification model along with a regression model to predict the T_C for intermetallic FM compounds, using the random forest method. These models are then used to identify the magnetic ground state of 5193 magnetic intermetallic compounds from the Materials Project database and to predict the T_C of those classified as FM. It is demonstrated that our machine learning framework is efficient and predictive, and can be used to accelerate the screening for FM compounds which are promising for spintronics and permanent magnets applications. Please find the workflow of our machine learning modelling in Figure S1.

Materials and methods

Data

1749 FM and 1056 AFM inter-metallic compounds are collected from the AtomWork database [21], which contains *experimental* structures and properties of magnetic materials. By excluding oxides, along with compounds without either Cr, Mn, Fe, Co, and Ni atoms, which are the typical magnetic atoms in transition metal based intermetallic magnetic materials. The distribution of experimental FM ordering temperatures is shown in Figure 1(a). The periodic table in Figure 1(b) highlights the distribution of the chemical elements for the FM compounds in our database, covering a large compositional phase space that includes not only the typical magnetic elements (*3d* and *4f*) but also the *s* and *p* blocks.

Descriptors

The Materials Agnostic Platform for Informatics and Exploration (MAGPIE) [22] proposed by Ward et al. is used to obtain the chemical descriptors, labeled as CHEM. As for structural descriptors (labeled as STR), Smooth Overlap of Atomic Positions (SOAP) is used to describe the local crystalline environment such as coordination and distance between atoms [23]. Space group number and volume of the unit cell are considered as a structural descriptor as well. In total, 139 (26) CHEM (STR) descriptors for each compound are considered.

Random forest

The random forest method is used both for identifying magnetic ground state of intermetallic compounds as well as predicting T_C for FM. The AtomWork database is spitted into training/validation set by the ratio of 80%/20%, the MP database is used to demonstrate the prediction ability. It can capture complicated relationship, analyze feature importance and is insensitive to extreme data compared with other machine learning methods (like neural network) [24].

Results

Classification

To enable the prediction of the magnetic ordering in a computationally inexpensive fashion, we perform a classification of the AFM or FM ground state. The training set consisted of 80% of the database and 20% are used for validation. The confusion matrix (CM) is a table that represents the instances in a predicted class versus the ones in the actual class, as shown in Figure 2(a), together with

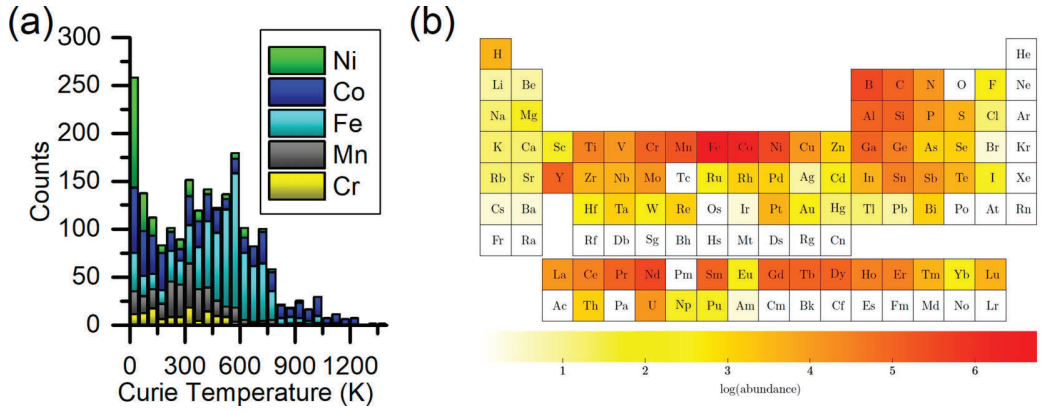


Figure 1. Distribution of T_C. (a) Histogram of T_C for 1749 FM materials in the database. (b) Distribution of the elements for compounds in the FM database.

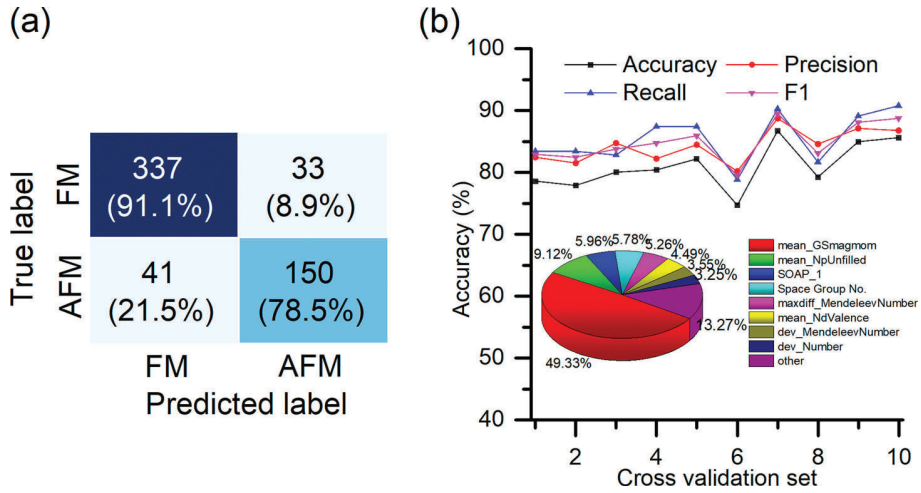


Figure 2. Performance of the classification model. (a): Confusion matrix of FM and AFM classification test set (b): Line graph represents the accuracy, precision, recall and F_1 score of the 10-folder cross validation of classification data, and the pie chart denotes the most important features in classification model (cf. Supplementary for detailed explanation).

the resulting metrics for 10 cross-validation sets plotted in Figure 2(b). The best accuracy for a model trained with all CHEM and STR features is found to be 87.3% as demonstrated in Figure S6. When the number of features used is reduced to 15, by considering the feature correlation and SHARP analysis, the accuracy drops slightly to 86.8%, with 91.1% FM and 78.5% AFM compounds being correctly classified (Figure 2(a)). This combined with an F1 score of 90.1% indicates good predictability, with a slight bias towards predicting compounds as FM, which might be due to the unbalanced number of FM/AFM compounds in the database. By performing 10-fold cross validation (Figure 2(b)), the average accuracy is about 81.1%, meaning the random forest model has neither overfitting nor biased sampling.

Interestingly, the descriptor 'mean GSmagmom' (the average of magnetic moment of the elemental solids for atoms in a given compound) is selected as the most important feature, contributing 49% of feature importance. In this regard, it suggests that 'mean GSmagmom' is critical, which accounts for the number of *3d/4f* electrons and the elemental properties of corresponding atoms. It is noted that the STR features contribute almost 15%, which proves its significance in determining the magnetic ground state. This is unsurprising given that it is known that interatomic distance is a determining factor for the nature of the AFM/FM exchange interaction following the Bethe-Slater curve [25]. The STR descriptors (especially the volume of the unit cell) could then be expected to be more important when the FM and AFM states are competing. However, it is noted that the relative importance of the STR feature depends also on the contents of our database. By replacing 682 experiment crystal structures in the training set with the DFT relaxed structures based on GGA from MP, the features importance of the model has change slightly, indicating our

model is robust against the subtle changes in the crystal structures.

Regression

Turning now to modeling the T_C, we construct and compare three random forest models with different feature combinations, using a 80%/20% partition for training/validation of the 1749 FM compounds. For the first two models we proceed in the same way as for the classification. The first model is trained with all CHEM and STR features, giving a R^2 of 91.24% and mean absolute error (MAE) of 54.99 K (Figure S7). The second model is trained using the reduced number of features as obtained by the dimensional reduction, with a comparable R^2 and MAE of 90.4% and 58 K, respectively (Figure S8). The last one is a two-step random forest model, intended to better describe the compounds with isomers (i.e. compounds with the same composition but different crystal structures). In the first step, we train a model using a reduced set of 17 CHEM features to predict the T_C, and in a second step, we train a new model using only the T_C predicted in the first step and 6 STR features as descriptors. The best R^2 obtained from the validation is as high as 91.2%, indicating very good agreement between the experimental and predicted values. As shown in Figure 3(a), the corresponding MAE is 55 K. It is noted that, compared to the accuracy of 90% for the T_C of Heusler compounds obtained based on DFT calculations [26], the accuracy (about 91.2%) of our machine learning model is improved, which can be applied on intermetallic compounds with various crystal structures.

The two-step random forest model generates comparable results to the one-step models but uses fewer features, so we choose to adopt the two-step random forest model for the regression of T_C and discuss the most important features selected in each step. In the first step, 'mean GSmagmom' is assigned with the highest importance of 52%, which highlights the relationship between electronic structure and magnetic properties of the elements and those of the compounds, as in the case of classification. For the second step, the T_C predicted by only CHEM contributes the most around 95%, which leaves the contribution of STR to be 5%, where 'Space Group No.' is the most important feature among them. Though the absolute contribution of the STR descriptors is not high, they are crucial in distinguishing the T_C of isomers. Even though, the CHEM descriptors alone can already provide good accuracy (90%, 60 K), T_C of isomers are not distinguishable without considering structure descriptors. For example, the experimental T_C for $DyMn_2$ with space group number 227 and 194 are 40 K and 37 K, respectively. By excluding all isomers in the training set to train the model, when using only CHEM descriptors both structures have the same (as expected) T_C of 62.37 K, however the values after considering the STR descriptors are 40.43 K and 36.37 K (Table S6). In general, considering STR descriptors increase the prediction accuracy of isomers from 90% to 91% and decrease the MAE from 74 K to 67 K (Figure S9), the average T_C difference between structures is only 18.7 K making it a reasonable improvement.

Prediction

Having established the predictive power of our models we now turn to its application on establishing the ground state and T_C of compounds in an existing database. We

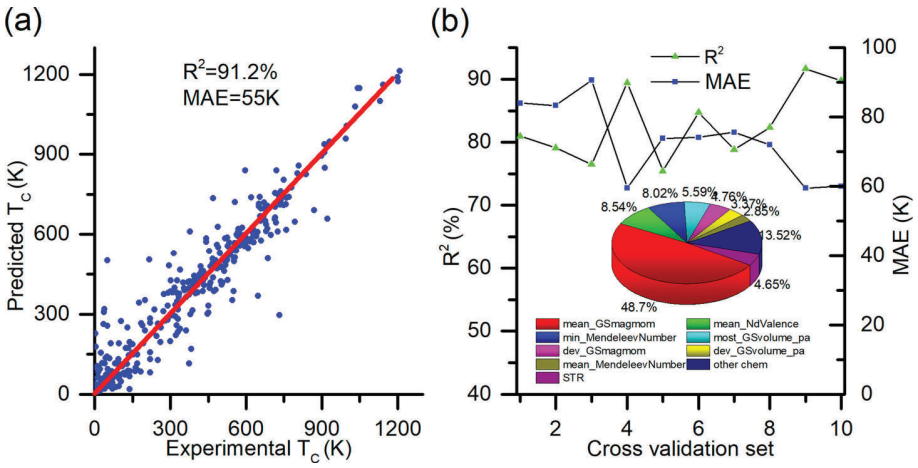


Figure 3. Regression model performance and feature importance. (a) Predicted vs. experimental T_C for test set. (b) Line graph represents the MAE and R^2 score of the 10-folder cross validation of our database, and the pie chart denotes the most important features in two-step regression model (cf. Supplementary for detailed explanation).

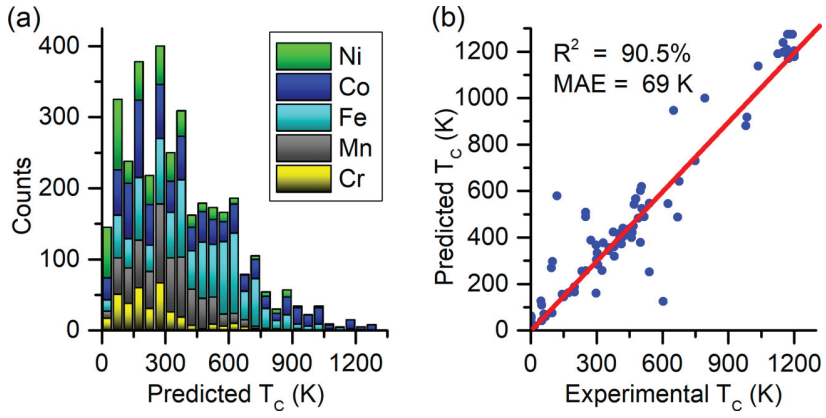


Figure 4. T_C prediction. (a) Histogram of predicted T_C for 3002 FM materials in the database. (b) Scatter plot of predicted vs. experimental T_C for 85 compounds with experimental T_C.

perform AFM/FM classification of 5193 magnetic intermetallic compounds from Materials Project, leading to 3002 (2191) FM (AFM) (see Data file S1 in the Supplementary). Figure 4(a) shows the predicted T_C for the classified FM compounds, ranging from 5 K to 1280 K. Obviously, the compounds with T_C higher than 600 K are again mostly Co and Fe based, and the element-wise distribution is comparable to that of our training set as shown in Figure S11. Furthermore, we randomly select 548 compounds from 3002 FM materials, and find that 85 among them have experimentally reported T_C. As shown in Figure 4(b), comparing the predicted T_C with the experimental values reveals that the accuracy of our prediction reaches as high as 90.5% with a corresponding MAE of 69 K. For example, despite that the calculation of compounds containing rare earth element has been a well-known trouble, $GdCo_4B$ (space group P6/mmm) and $Gd_3Co_{11}B_4$ (space group P6/mmm) have been predicted to be FM with a T_C of 490 K and 333 K, which are quite close to the experimental values of 517 K and 303 K, respectively. This is in line with the results obtained via the construction of the machine learning models, validating the workflow that we propose. Detailed values of all predicted and the experimental values are presented in Data file S2 and Table S1, respectively.

Discussion

It is demonstrated that machine learning using the random forest algorithm is able to distinguish materials with FM and AFM ordering, and further to predict the T_C of FM compounds, solving two critical problems in designing magnetic materials. For classification, the accuracy reaches 87%, outperforming the DFT calculations [27], which are applied on a selected set of compounds. For the resulting FM compounds, the magnetization can be straightforwardly evaluated using DFT. Furthermore, the T_C can be accurately modeled with R^2 about 91% and MAE about 55 K. This enables us to reduce the number of candidates for further computational and experimental characterizations. For instance, the MCA can be evaluated in a high throughput way [5], which sets an upper limit for the coercivity. Thus, the machine learning model developed in this work in conjunction with DFT calculations enables us to get all three essential intrinsic magnetic properties evaluated.

This paves the way to develop FM materials with systematic characterization of the intrinsic magnetic properties, with the help of further high throughput DFT calculations.

Acknowledgements

We acknowledge support by the Deutsche Forschungsgemeinschaft (DFG - German Research Foundation) and the Open Access Publishing Fund of Technical University of Darmstadt.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Teng Long thanks the financial support from the China Scholarship Council. Part of this work was supported by the European Research Council under the European Union's Horizon 2020 research and innovation program [grant number 743116 – project Cool Innov] and the Deutsche Forschungsgemeinschaft [grant number 405553726 – TRR 270].

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Teng Long http://orcid.org/0000-0003-0262-299X
Oliver Gutfleisch http://orcid.org/0000-0001-8021-3839

References

[1] Skokov KP, Gutfleisch O. Heavy rare earth free, free rare earth and rare earth free magnets – vision and reality. Scripta Mater [Internet]. 2018 [cited 2019 Jul 3]; 154:289–294. Available from: http://www.sciencedirect.com/science/article/pii/S1359646218300599

[2] Sander D, Valenzuela SO, Makarov D, et al. The 2017 magnetism roadmap. J Phys D Appl Phys [Internet]. 2017 [cited 2019 Jul 2]; 50:363001. Available from: https://doi.org/10.1088%2F1361-6463%2Faa81a1

[3] Jungwirth T, Sinova J, Manchon A, et al. The multiple directions of antiferromagnetic spintronics. Nature Phys [Internet]. 2018 [cited 2019 Jul 2]; 14:200. Available from: https://www.nature.com/articles/s41567-018-0063-6

[4] Gutfleisch O, Gottschall T, Fries M, et al. Mastering hysteresis in magnetocaloric materials. Philos Trans A Math Phys Eng Sci. 2016;374:20150308.

[5] Drebov N, Martinez-Limia A, Kunz L, et al. *Ab initio* screening methodology applied to the search for new permanent magnetic materials. New J Phys. 2013;15:125023.

[6] Möller JJ, Körner W, Krugel G, et al. Compositional optimization of hard-magnetic phases with machine-learning models. Acta Mater. 2018;153:53–61.

[7] Kübler J. Theory of itinerant electron magnetism. New York (NY): Oxford University Press; 2017.

[8] Sponza L, Pisanti P, Vishina A, et al. Self-energies in itinerant magnets: a focus on Fe and Ni. Phys Rev B [Internet]. 2017 [cited 2020 Sep 8]; 95:041112. Available from: https://link.aps.org/doi/10.1103/PhysRevB.95.041112

[9] Zhang H. High-throughput design of magnetic materials. arXiv:200812907 [cond-mat] [Internet]. 2020 [cited 2020 Sep 8]; Available from: http://arxiv.org/abs/2008.12907

[10] Matsumoto M, Akai H. Calculated Curie temperatures for rare-earth permanent magnets: ab initio inspection on localized magnetic moments in d-electron ferromagnetism. arXiv:181204842 [cond-mat] [Internet]. 2018 [cited 2019 Jul 13]; Available from: http://arxiv.org/abs/1812.04842

[11] Katanin AA, Belozerov AS, Anisimov VI. Nonlocal correlations in the vicinity of the phase transition in iron within a DMFT plus spin-fermion model approach. Phys Rev B. 2016;94:161117.

[12] Han Q, Birol T, Haule K. Phonon softening due to melting of the ferromagnetic order in elemental iron. Phys Rev Lett [Internet]. 2018 [cited 2018 Dec 14]; 120. Available from: https://link.aps.org/doi/10.1103/PhysRevLett.120.187203

[13] Kvashnin YO, Grånäs O, Di Marco I, et al. Exchange parameters of strongly correlated materials: extraction from spin-polarized density functional theory plus dynamical mean-field theory. Phys Rev B [Internet]. 2015 [cited 2020 Aug 26]; 91:125133. Available from: https://link.aps.org/doi/10.1103/PhysRevB.91.125133

[14] Ward L, Liu R, Krishna A, et al. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. Phys Rev B [Internet]. 2017 [cited 2019 Jul 4]; 96:024104. Available from: https://link.aps.org/doi/10.1103/PhysRevB.96.024104

[15] Zhuo Y, Mansouri Tehrani A, Brgoch J. Predicting the band gaps of inorganic solids by machine learning. J Phys Chem Lett [Internet]. 2018 [cited 2019 Jul 4]; 9:1668–1673. Available from: https://doi.org/10.1021/acs.jpclett.8b00124

[16] Behler J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. J Chem Phys [Internet]. 2011 [cited 2019 May 16]; 134:074106. Available from: http://aip.scitation.org/doi/10.1063/1.3553717

[17] Stanev V, Oses C, Kusne AG, et al. Machine learning modeling of superconducting critical temperature. Comp Mater [Internet]. 2018 [cited 2018 Oct 10]; 4:29. Available from: https://www.nature.com/articles/s41524-018-0085-8

[18] Sanvito S, Oses C, Xue J, et al. Accelerated discovery of new magnets in the Heusler alloy family. Sci Adv [Internet]. 2017 [cited 2019 May 27]; 3:e1602241. Available from: http://advances.sciencemag.org/lookup/doi/10.1126/sciadv.1602241

[19] Dam HC, Nguyen VC, Pham TL, et al. A regression-based feature selection study of the Curie temperature of transition-metal rare-earth compounds: prediction and understanding. arXiv:170500978 [cond-mat] [Internet]. 2017 [cited 2019 Mar 12]; Available from: http://arxiv.org/abs/1705.00978

[20] Nelson J, Sanvito S. Predicting the Curie temperature of ferromagnets using machine learning. arXiv:190608534 [cond-mat, physics:physics] [Internet]. 2019 [cited 2019 Jun 23]; Available from: http://arxiv.org/abs/1906.08534

[21] Xu Y, Yamazaki M, Villars P. Inorganic materials database for exploring the nature of material. Jpn J Appl Phys [Internet]. 2011 [cited 2019 Jun 19]; 50:11RH02. Available from: https://iopscience.iop.org/article/10.1143/JJAP.50.11RH02/meta

[22] Ward L, Agrawal A, Choudhary A, et al. A general-purpose machine learning framework for predicting properties of inorganic materials. NPJ Comput Mater [Internet]. 2016 [cited 2019 Apr 12]; 2:16028. Available from: https://www.nature.com/articles/npjcompumats201628

[23] Bartók AP, Kondor R, Csányi G. On representing chemical environments. Phys Rev B [Internet]. 2013 [cited 2019 Apr 13]; 87:184115. Available from: https://link.aps.org/doi/10.1103/PhysRevB.87.184115

[24] Géron A. Hands-on machine learning with Scikit-learn and tensor flow [Internet]. [cited 2019 Jun 19]. Available from: http://shop.oreilly.com/product/0636920052289.do

[25] Sommerfeld A, Bethe H. Elektronentheorie der Metalle. In: Smekal A, editor. Aufbau Der Zusammenhängenden Materie [Internet]. Berlin: Springer; 1933. [cited 2020 Jan 20]. p. 333–622. DOI:10.1007/978-3-642-91116-3_3

[26] Kovacik R, Mavroupolos P, Blügel S. Critical temperature and effective magnetic moment of experimentally known magnetic Heusler alloys from first principles. Regensburg: DPG-Frühjahrstagung; 2019. Report No.: MA 55.8: Vortrag.

[27] Horton MK, Montoya JH, Liu M, et al. High-throughput prediction of the ground-state collinear magnetic order of inorganic materials using density functional theory. Npj Comput Mater. 2019;5:2.

Supplementary Materials

Section 1:

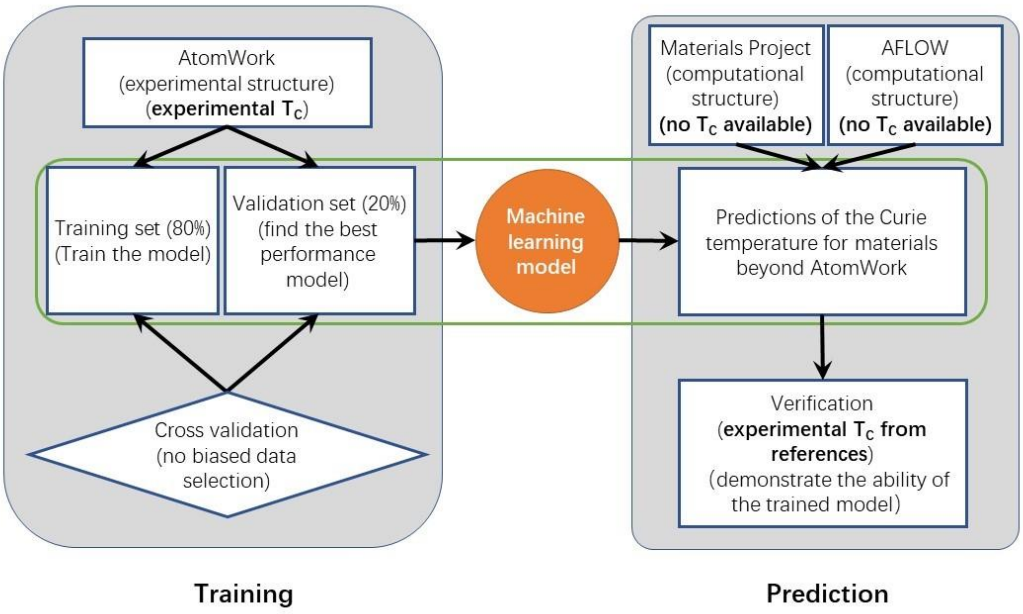


Fig. S1. The workflow of our machine learning modelling, together with the databases used. The "Training" part consists of training and validation using the data from AtomWork, which leads to a machine learning model. In the "Prediction" part, our machine learning model is applied to make predictions based on crystal structures from Materials Project and AFLOW, which are partially verified by explicit literature survey on 85 compounds. The green box denotes the training, validation and prediction steps, in analogue to the training, validation and test in traditional machine learning models.

The workflow of our machine learning modelling is showed in Fig. S1. It contains two essential parts. In the first part, our machine learning model is trained using the AtomWork database, which contains the experimental crystal structures and Curie/Neel temperatures, which have been collected from published papers. In the second part, after obtaining the machine learning model, we applied it to predict the Curie temperatures for compounds in the Materials Project and AFLOW databases. It is noted that Materials Project and AFLOW provide only the crystal structures, but neither experimental nor computational values of the Curie temperature.

The corresponding crystal structures are collected from the AtomWork and Inorganic Crystal Structure Database (ICSD).[1] For compounds with multiple magnetic phase transitions, the critical temperature is defined as the magnetic transition temperature from a disordered paramagnetic state to an ordered FM state. In this way, materials with first-order magneto-volume, magneto-structural, and temperature dependent spin-reorientation transitions are excluded for the current work.

The distribution of experimental FM ordering temperatures is shown in Fig. 1(a). It is dominated by compounds with low T_C, with a maximum value of around 1400 K, *e.g.*, 1410 K and 1388 K for Co in face-centered-cubic (space group 225) and hexagonal-closed-pack (space group 194)

structures, respectively. The element resolved T_C distributions are highlighted in Fig. S2(B-D) for Fe-, Co-, and Mn-based compounds, whereas the T_C for Ni-based compounds are mostly in the low-temperature range (Fig. S2(A) and Fig. S2(E)) and there are limited number of Cr-based ferromagnetic materials (Fig. S2(A)). It is clear that all compounds with T_C higher than 1200K are Co-based (Fig. 1(d)), the Fe-based compounds (Fig. S2(C)) consist of a third of the database with T_C normally distributed around 600K, while the T_C of Mn-based compounds (Fig. S2(B)) are mostly found at relative low temperature (with a peak at 300 K) range. Thus, the Fe- and Co-based compounds are optimal for high temperature applications, and in the room temperature range all three classes are interesting.

Moreover, the distribution of experimental AFM ordering Neel temperatures (T_N) is shown in Fig. S2(a), where T_N of most compounds are less than 100K. The element resolved T_N distributions in Fig. S2(b-f) indicate that Fe- and Mn- based compounds are more suitable for high temperature application. It is noted that there are many more AFM compounds such as oxides, thus the collection of AFM intermetallic compounds in this work serves only for classification and we save the regression of T_N for future study. We have also collected 5193 magnetic intermetallic compounds from Materials Project (with no T_C but only crystal structure provided),[2] in order to make predictions by applying the machine learning models for magnetic ground state classification and T_C regression, as discussed in detail below.

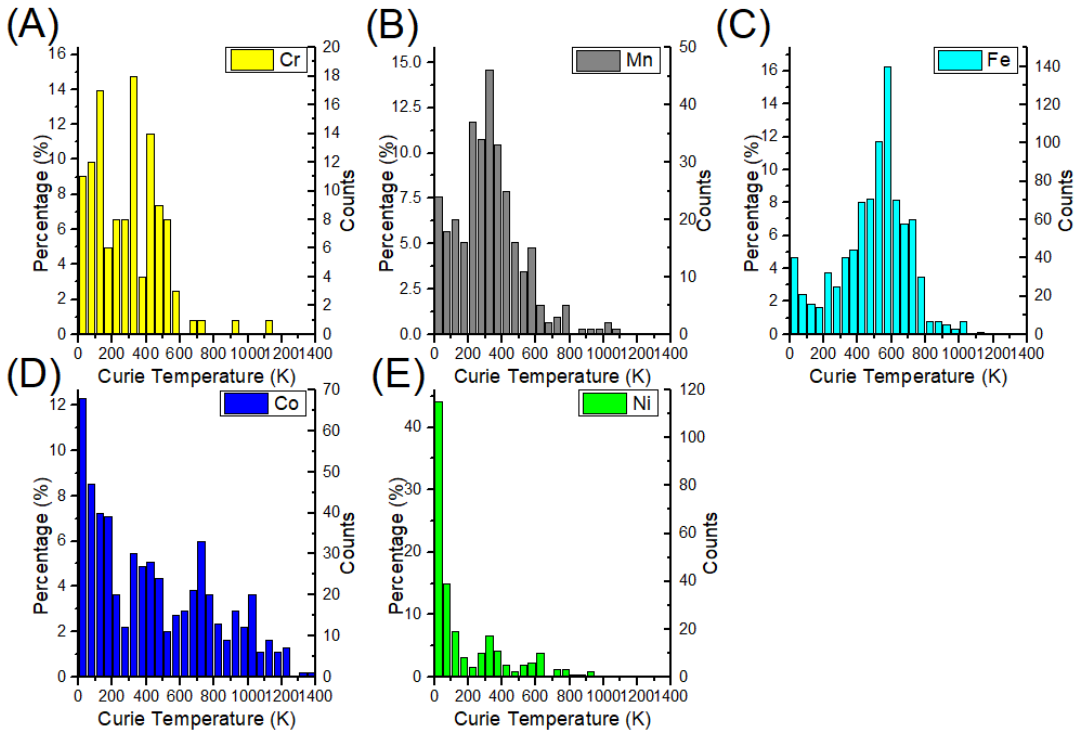


Fig. S1. Histogram of FM compound categorized by T_C. (A) Histogram of Cr based FM compound categorized by T_C. **(B)** Histogram of Fe based FM compound categorized by T_C. **(C)** Histogram of Co based FM compound categorized by T_C. **(D)** Histogram of Mn based FM compound categorized by T_C. **(E)** Histogram of Ni based FM compound categorized by T_C. The bin size is fixed to be 50K.

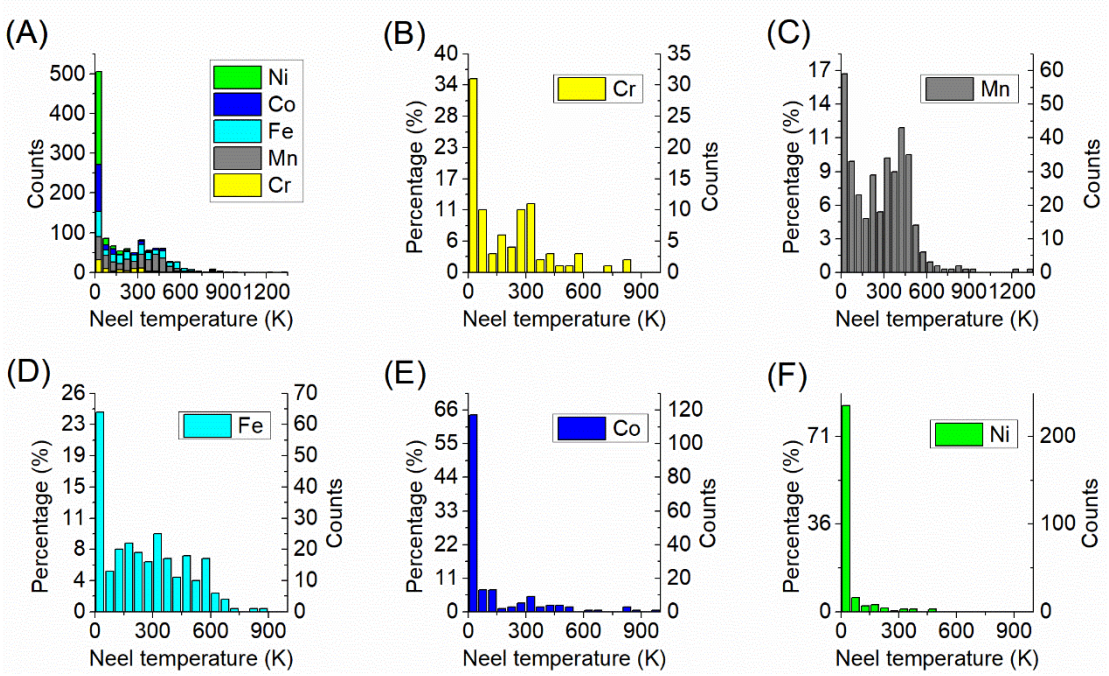


Fig. S3. Distribution of T_N for the AFM database. (A) Histogram of TC for 1056 AFM materials in the database. Green, blue, cyan, gray and yellow represents Ni, Co, Fe, Mn and Cr based compounds, respectively. **(B)** Histogram of Cr based AFM compound categorized by T_N. **(C)** Histogram of Mn based AFM compound categorized by T_N. **(D)** Histogram of Fe based AFM compound categorized by T_N. **(E)** Histogram of Co based AFM compound categorized by T_N. **(F)** Histogram of Ni based AFM compound categorized by T_N. The bin size is fixed to be 50K.

The AtomWork database[3] contains 5487 compounds labeled as FM and 5896 as AFM. As we aim at screening for permanent magnets and materials for metallic spintronics, non-intermetallic (*i.e.*, oxides, sulfites, chlorides and fluorides) compounds are excluded. Moreover, we chose compounds with at least one of Cr, Mn, Fe, Co, or Ni elements, which are the typical magnetic atoms for magnetic intermetallic compounds. Their structural information is collected from AtomWork and Inorganic Crystal Structure Database (ICSD).[1] For intermetallic compounds with multiple magnetic phase transitions, the critical temperature is defined as the magnetic transition temperature from an ordered FM state to a paramagnetic state. As such first-order magneto-volume or martensitic transitions, as well as spin-reorientation are excluded. Additionally, we consider the ferrimagnetic materials as FM, as they have a non-vanishing total magnetic moment, as in the case of FM. Based on above criteria, 1749 FM and 1056 AFM inter-metallic compounds are selected.

As for the prediction, we collected all magnetic compounds from Materials Project, following the same criteria specified above and with total magnetic moment larger than $0.1\mu_B$.[2] After removing those which are already included in our database, 5193 compounds are obtained. Prediction over some compounds in AFLOW can be found in Table S7 in comparison with the prediction over Materials Projects. They have almost the same predicted values, so readers can choose either of these two databases.

Section 2:

In this work, we regroup them in the following 5 classes, namely, norm (L^p norms of the fractions), magnetic moment of constituting elements, atomic number, valence electrons and other chemical descriptors. MAGPIE aims at describing a compound through the properties of their constituent elements by calculating the average, standard deviation and other statistic metrics of the experimental atomic properties.[4] Originally Magpie including 4 categories: stoichiometric attributes (L^p norms of the fractions), element properties statistics (mean magnetic moment, range of atomic number and deviation of columns of periodic table), electron structure attributes (the average fraction of electrons from the s, p, d and f valence shells between all present elements) and ionic compound attributes. For more details, please refer to work by *Ward et al.* (npj Computational Materials volume 2, Article number: 16028 (2016)). In this work, the rest are divided into five categories: L^p norms of the fractions, magnetic moment of constituting elements, atomic number (properties strong linear correlated with atomic number like Mendeleev number, atomic weight etc. are included as well), valence electrons (unoccupied orbitals) (a descriptors related with valence electrons and unoccupied orbitals, like electron structure attributes, melting temperature and other element properties related with electrons) and other CHEM descriptors (volume of element solids, electronegativity and band gap).

SOAP provides a representation of the local environment (like coordination and distance between atoms), which is expected to be important in studying magnetic properties. It uses Laplace spherical harmonics to expand Gaussian function and then transfers the coefficients to rotation, translation and perturbation invariant descriptors as following equations describe.[5]

$$\rho(r) = \sum_i e^{-(r-r_i)^2}$$

$$c_{nlm} = \rho|g_n(r)Y_{lm} = \int_V g_n(r)Y_{lm}(\theta,\phi)\rho(r,\theta,\phi)dV$$

$$P_{nn'l} = \sum_m c_{nlm}c^*_{n'lm}$$

However, such descriptors can only generate homogeneous data for atoms.[6] DScribe library in python developed by Aalto University generates SOAP descriptors for each atom and averages them for compounds automatically.[7] But their dimensions are still related to the number of elements in one compound. By doing average over different local environments, we can get homogeneous STR descriptors and their dimensions are solely determined by parameters in spherical harmonics.[8] Space group number of compounds is used as other STR descriptor as well.

Firstly, we remove those descriptors with a correlation of 70% or more, reducing the number of features to 42 for both models. Further, based on the Shapley additive explanations (SHAP) analysis,[9]

the feature importance can be defined, and those with a contribution smaller than 0.25% are then dropped, leading to 15 features for the classification and 23 for the regression.

SHAP (SHapley Additive exPlanations) is a feature importance analysis method proposed by Lundberg and Lee.[10] The essential idea is to calculate the SHAP value, which marks the contribution of one feature by comparison between predicted values with and without this feature. TreeSHAP was proposed in 2018 as a generalization of SHAP for tree-based machine learning models.[9] All the analysis has been conducted using the SHAP package.[9] In order to analyze the feature importance more explicitly, in this work, we evaluate the absolute of each SHAP value and calculate the percentage of each of them to the feature importance.

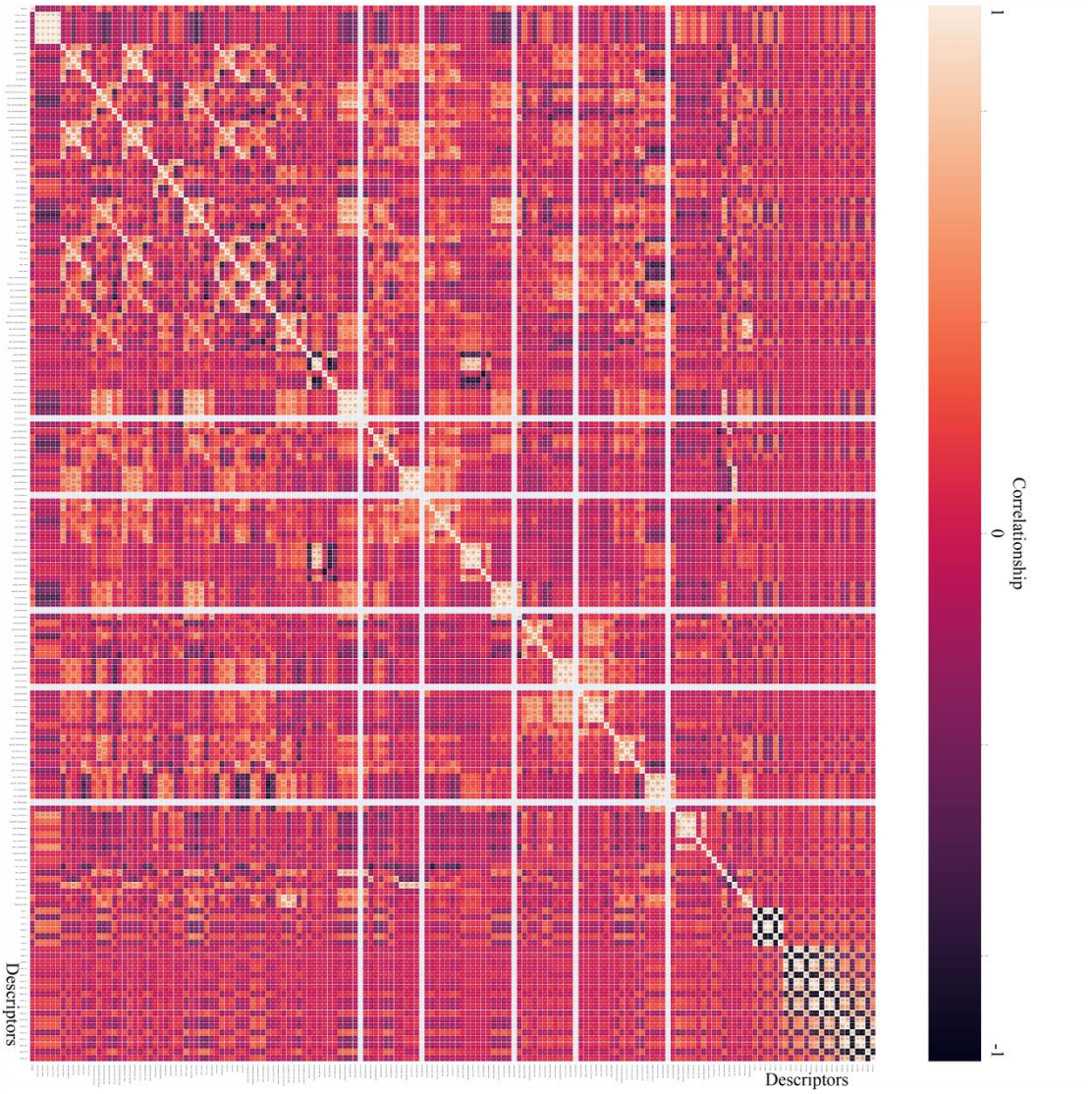


Fig. S4. Correlation heatmap of descriptors in classification model, both axes are descriptors used in this model; color represents the correlation, *i.e.* 1 means positive related and -1 means negative related.

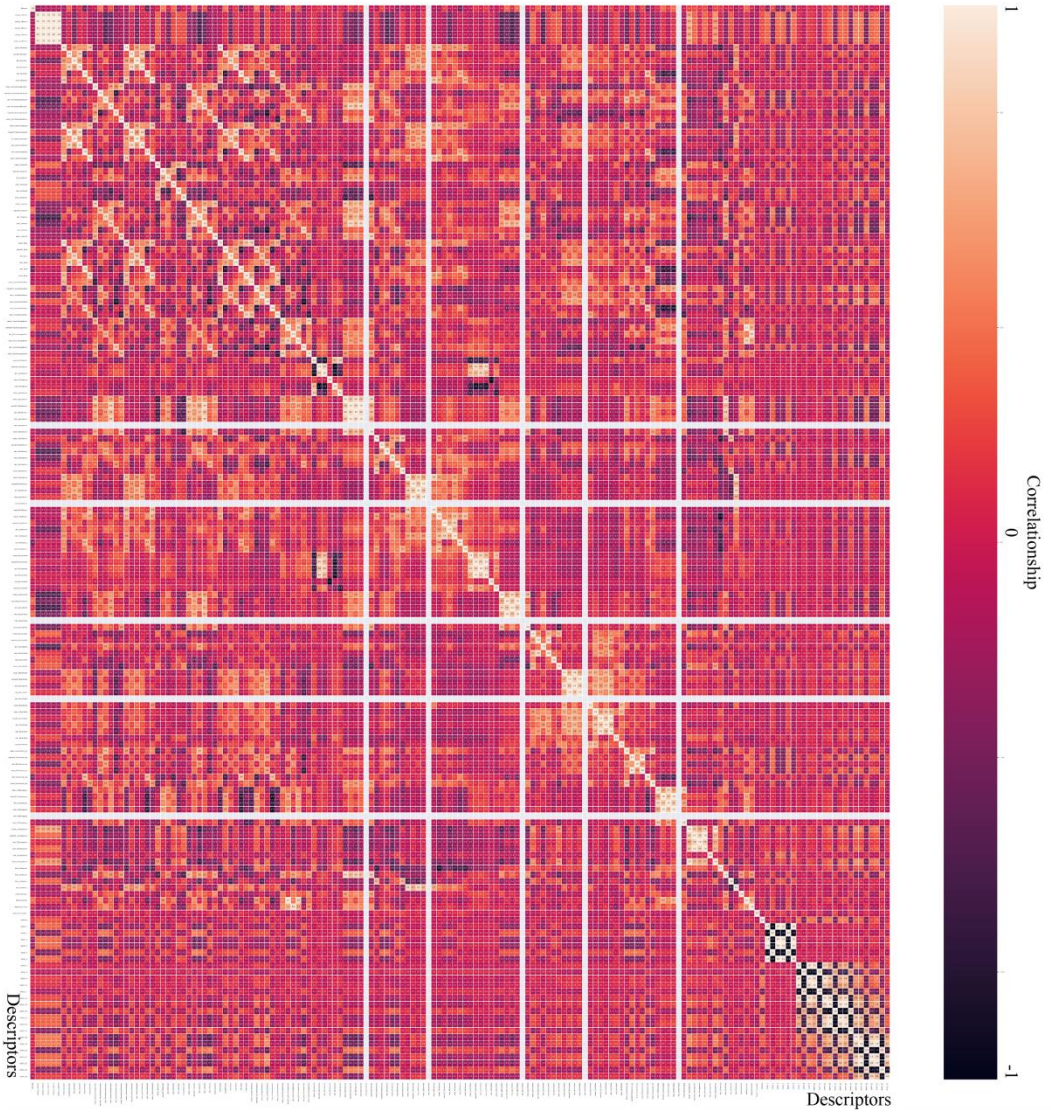


Fig. S5. Correlation heatmap of features in regression model, both axes are descriptors used in this model; color represents the correlation, *i.e.* 1 means positive related and -1 means negative related.

Section 3:

The training and test splitting is conducted by "train_test_split" function of "sklearn" package, with the best model using random seed 2. Splitting with random seed ranging from 0 to 1000 gives the accuracy ranging from 88% to 91%, which proves the high accuracy is not due to bias data selection.

Neural network and Gaussian kernel method are also trained to solve this problem, but accuracy can barely reach 70%. As a result, we use the random forest model despite some apparent disadvantages.

For the classification model, the number of estimators is 23 and the number of features used is 15, while the minimum value of the number of leafs in each node is 3. For the regression, the model based only on the CHEM descriptors has 23 estimators, 17 features and at least 3 leafs in each node. The corresponding number for the second stage of regression are 23, 7 and 3 respectively. All these procedures along with the metrics to evaluate models (accuracy, precision, recall, F1,

MAE and R^2) are all conducted in scikit-learn package.[11]

The quality of the classification model is judged in terms of the following statistic metrics. By looking at ture positives (tp), false positives (fp), false negatives (fn) and true negatives (tn), the accuracy, precision, recall and F1 can be evaluated as following:

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

$$precision = \frac{tp}{tp + fp}$$

$$recall = \frac{tp}{tp + fn}$$

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

with true being FM and false AFM. The accuracy represents the overall quality of the prediction, the precision is the proportion of those correctly classified as FM within all classified as FM, the recall is the proportion of those correctly identified as FM with all known to be FM, and the F1 bridges the recall and precision metrics, denoting if there is a bias towards classifying one particular label. The confusion matrix (CM) is a table that represents the instances in a predicted class versus the ones in the actual class, as shown in Fig. 2(a), together with the resulting metrics for 10 cross-validation sets plotted in Fig. 2(b).

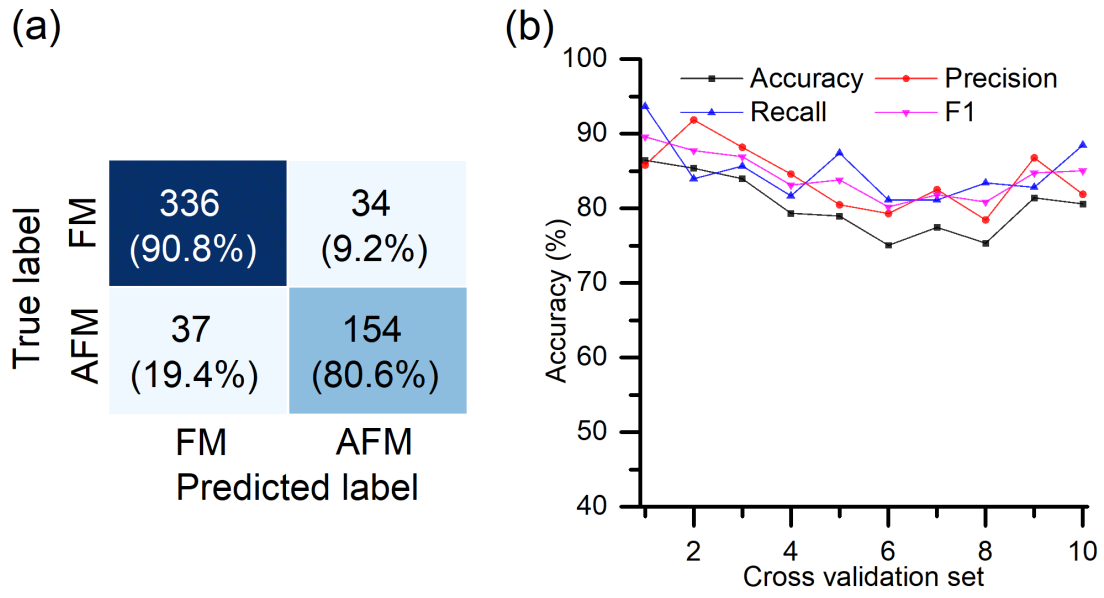


Fig. S6. Performance of the classification model with all features. (a): Confusion matrix of FM and AFM classification test set of 561 compounds. **(b)**: Statistical metrics for 10-folder cross validation.

Section 4:

R^2 score is called the coefficient of determination, which describes the accuracy of regression; closer to 1 means the prediction accuracy is closer to 100%. It is defined as the following equation:

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

y_i and f_i represent experimental value and predicted value, respectively.

The second most important descriptor of regression, "mean NpUnfilled" (average number of unfilled *p* electrons across the elements in the stoichiometry), only constitutes 9%, followed by the structure descriptors, SOAP, volume of unit cell and space group numbers. As for "mean NpUnfilled", it can be understood by considering the relative abundance of *p*-block elements in our dataset, but nonetheless relates to electronic properties. It is clear that both of these features relate to the electronic structure of the elements involved and thus indirectly to the electronic structure and magnetism of the compounds.

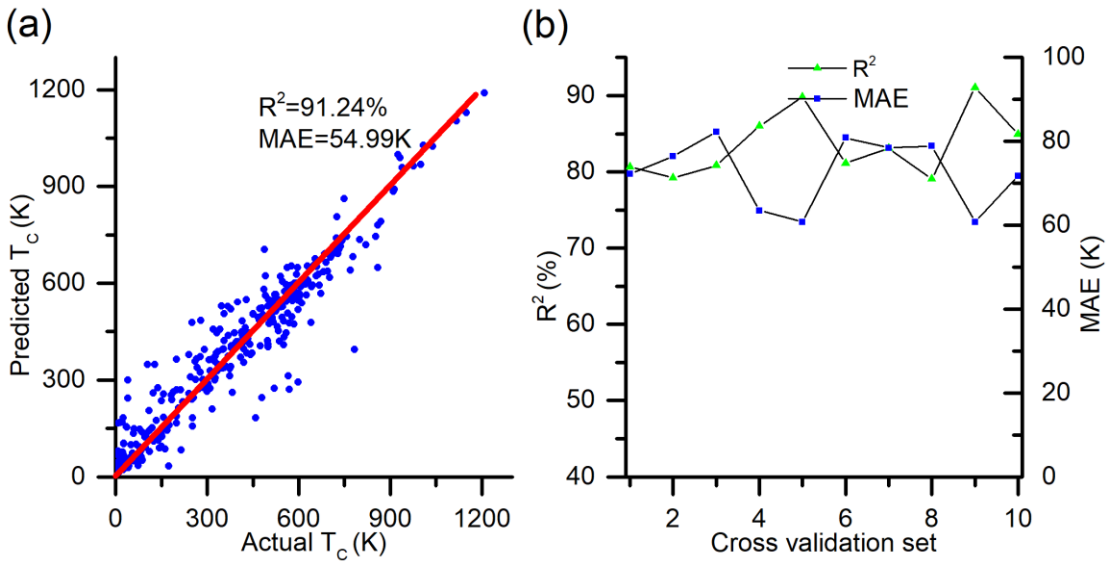


Fig. S7. One-step regression model performance with all features. (a) Predicted vs. experiment T_C for test set in general regression model. **(b)** 10-folder cross validation.

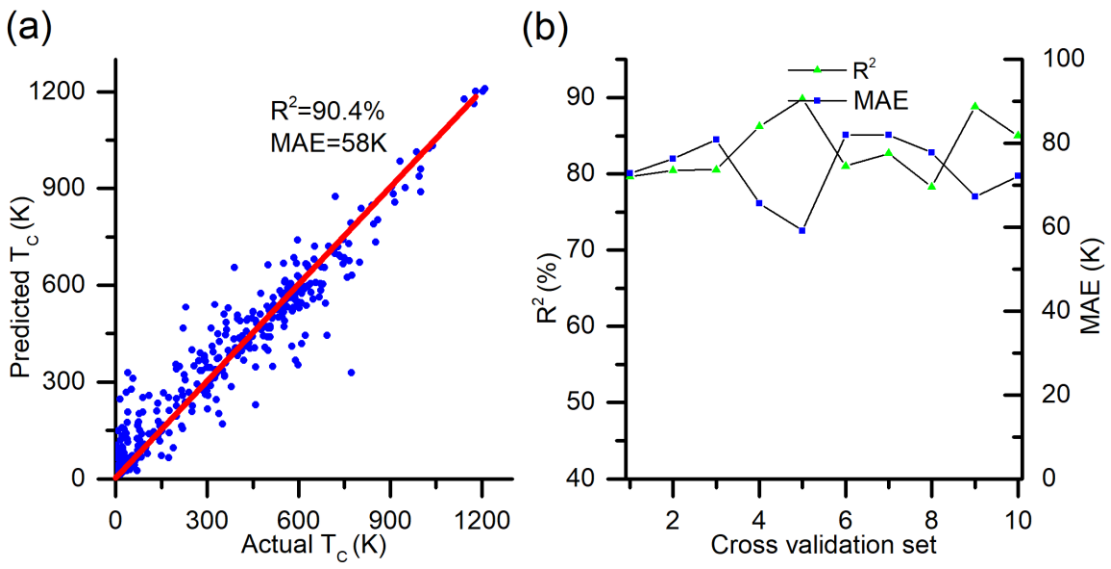


Fig. S8. One-step regression model performance with reduced features. (a) Predicted vs. experiment T_C for test set in general regression model. **(b)** 10-folder cross validation.

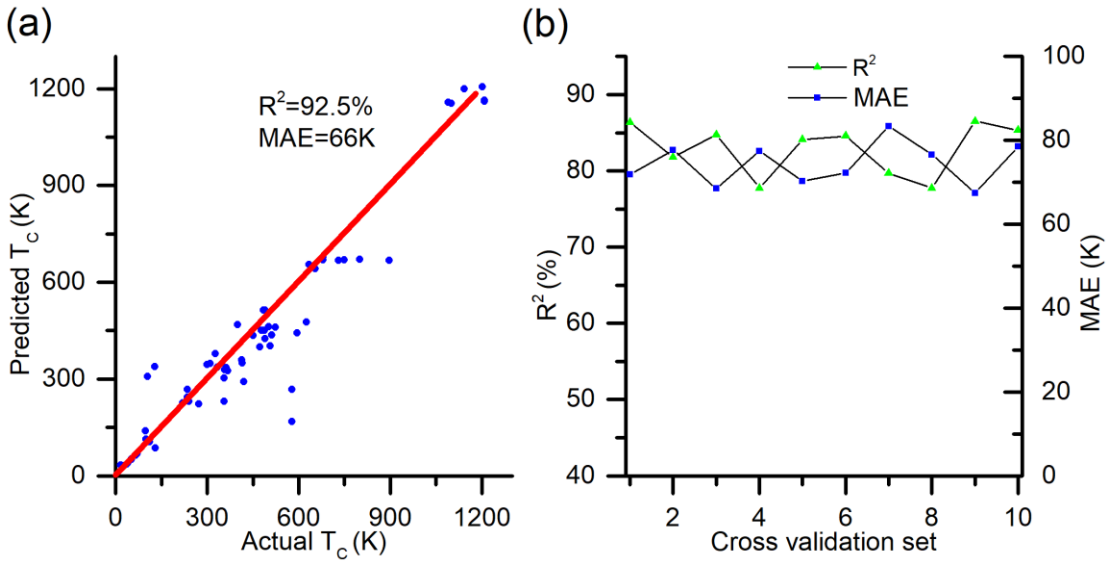


Fig. S9. Two-step regression model performance for isomers only in the test set. (a) Predicted vs. experiment T_C for test set in general regression model. **(b)** 10-folder cross validation.

Section 5:

Furthermore, in order to accelerate the training process, we performed active learning (AL), starting with 10% of the training data (not the whole database) and take automatically 10% more samples data which are the most outliers (where the predicted value has the most absolute difference from the experiment value) in the rest of the training set. In this way, the feature space that has not been covered by the previous training set will be included, enabling us to effectively select most representative data in the training set. As shown in Fig. S6, using only 50% of the data, the resulting accuracy based on AL is already comparable with the standard model. Such training accelerating process can be used to collect more experimental data more efficient and thus make the model easy to improve.

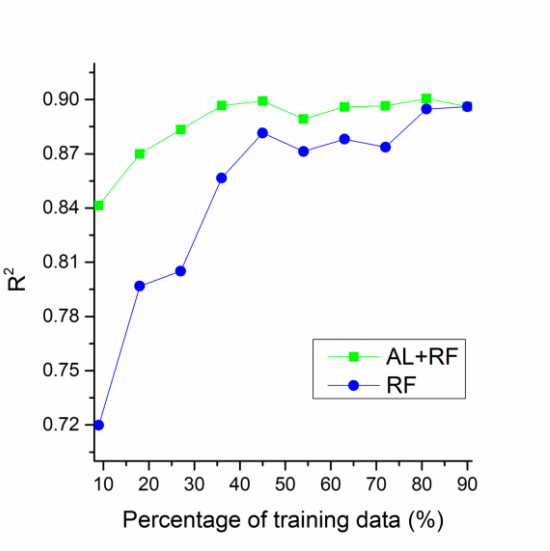


Fig. S10. Learning curve of RF vs. AL+RF in regression.

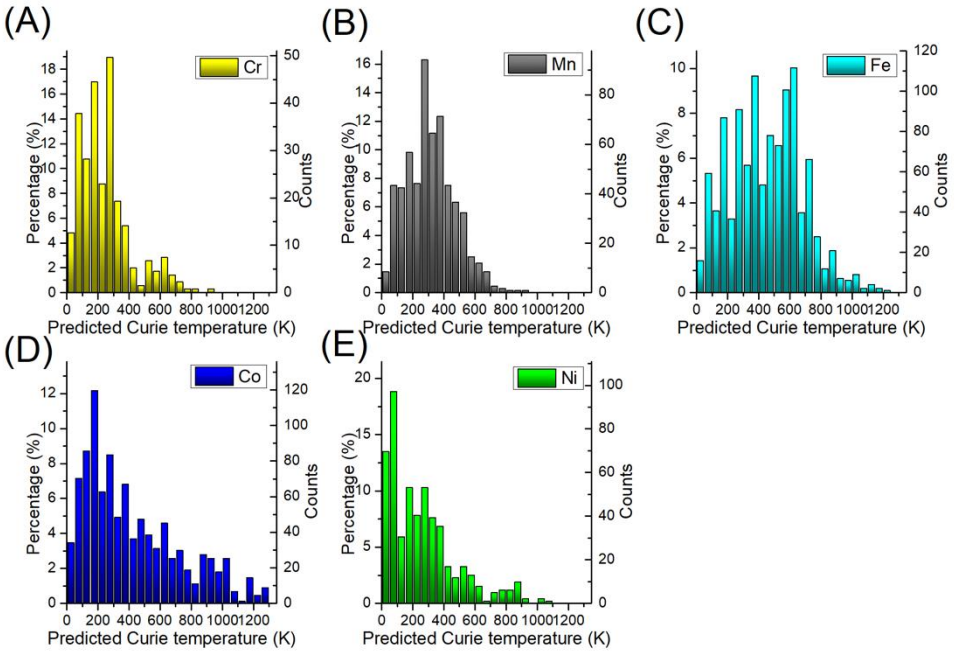


Fig. S11. Histogram of predicted result analysis. (A) Histogram of Cr based FM compound categorized by T_C. **(B)** Histogram of Mn based FM compound categorized by T_C. **(C)** Histogram of Fe based FM compound categorized by T_C. **(D)** Histogram of Co based FM compound categorized by T_C. **(E)** Histogram of Ni based FM compound categorized by T_C. (bin size is 50K)

Table S1. Predicted vs. experimental T_C of predicted data.

MP-id	Formula	Space group No.	Experiment T_C (K)	Predicted T_C (K)
mp-16932	$Co_{17}Th_2$	166	1035	1137
mp-1190790	$Fe_{17}N_3Sm_2$	166	749	730
mp-1190611	$Fe_{17}H_3Nd_2$	166	478	566
mp-1095461	Co_3Nd	166	381	319
mp-13268	BCo_4Gd	191	517	490
mp-604537	$GaMnNi_2$	12	379[12]	359
mp-1192498	$Mn_{23}Tm_6$	225	404	416
mp-1190704	$B_3Co_7Er_2$	191	303	333
mp-22673	Cu_2InMn	225	506.15	525
mp-5491	$AlAu_2Mn$	225	233	255
mp-8793	$B_6Fe_2Nd_5$	166	67	58
mp-1189379	Ni_7Tb_2	166	98	74
mp-1188905	$Co_3Sb_4U_3$	220	10	25
mp-1201816	$Co_{17}Gd_2$	194	1200	1180
mp-1204082	$Co_{17}Lu_2$	194	1192	1274
mp-1199370	$Co_{17}Tb_2$	194	1150	1238
mp-1196360	$Co_{17}Tm_2$	194	1170	1274
mp-1219785	$Co_{17}PrSm$	160	1200	1192
mp-1219295	$Co_{17}GdSm$	160	1200	1177
mp-1220026	$Co_{17}ErPr$	160	1167	1210
mp-1200096	$Co_{17}Sm_2$	194	1150	1197
mp-1215870	$Co_{17}PrYb$	194	1178	1195
mp-1199900	$Co_{17}Yb_2$	194	1176	1274
mp-1216133	$Co_{33}Yb_4Zr$	156	1175	1174
mp-1215883	$Co_{34}Pr_3Yb$	8	1173	1170
mp-1219047	$Co_{17}SmY$	160	1200	1203
mp-356	$Co_{17}Nd_2$	166	1125	1191
mp-1198821	$CuFe_{13}Nd_6$	140	463	421
mp-20780	Mn_6Sn_6Tb	191	376	424
mp-605854	$B_4Co_{11}Gd_3$	191	454	439
mp-22790	Co_2GaV	225	300	305
mp-9277	Ce_2CrN_3	71	50	42
mp-22760	Ge_2LaMn_2	139	306	282
mp-1195842	$Fe_{17}Lu_2$	194	298	367
mp-22625	Cr_2CuTe_4	227	326	257
mp-30450	$Bi_4Cu_4Mn_3$	225	460	400
mp-1193393	Er_6Mn_{23}	225	430	416
mp-1193270	$Mn_{23}Yb_6$	225	406	401
mp-1189397	$As_{12}Fe_4Pr$	204	18	20
mp-13464	$CaFe_4Sb_{12}$	204	60	71
mp-22300	Co_2FeGe	225	980	880

MP-id	Formula	Space group No.	Experiment T_C (K)	Predicted T_C (K)
mp-2794	C_2Fe_5	15	500	601
mp-1729	Fe_2Sm	227	676	642
mp-1190681	Fe_2Zr	194	625	545
mp-1095566	BCo_4La	191	420	403
mp-1198314	Mn_6SmSn_6	71	405	393
mp-601285	$GaMnNi_2$	139	355	343
mp-1188589	$B_6Co_{12}Er$	166	144	155
mp-581681	$Ca_{21}Mn_4Sb_{18}$	12	2.4	58
mp-1200648	$Fe_{14}Ga_3Y_2$	194	540	546
mp-1191681	$Co_{12}Eu_2P_7$	174	151	144
mp-1193555	$Fe_{23}Sm_6$	225	442	434
mp-1205374	$Fe_{15}Nd_2Si_2$	166	490	482
mp-1191470	$Fe_{17}H_3Th_2$	166	471	541
mp-1200417	$Fe_{15}Si_2Tm_2$	194	467	448
mp-20228	$GaMnNi_2$	225	414	371
mp-643614	Mn_6NdSn_6	71	357	355
mp-3706	$Co_{12}Ho_2P_7$	174	150	148
mp-1105798	$B_6Co_{12}Sm$	166	172	162
mp-1194031	$MnPr_4Sb_9$	12	50	108
mp-1103869	Cr_5KTe_8	12	250	257
mp-601400	Co_9GdSi_4	140	47	127
mp-1193687	$Ce_6Ni_6P_{17}$	217	1.4	61
mp-1080653	FeNiP	189	330	377
mp-4492	Co_2MnSi	225	985	917
mp-1191073	$Fe_{17}H_3Pr_2$	166	480	566
mp-13494	$Fe_{29}Nd_3$	12	421	439
mp-16085	$B_{10}Co_{29}Nd_3Si_4$	129	200	165
mp-23047	Bi_2Ho_6Mn	189	200	188
mp-30217	$Ca_{21}Mn_4Sb_{18}$	15	2.4	55
mp-1196805	$Fe_{17}Gd_2$	194	500	379
mp-1194380	CCo_3Mo_3	227	100	296
mp-1200877	Fe_5HoP_3	62	298	160
mp-1189764	$FeSb_3Zr_5$	193	540	251
mp-5529	Fe_2MnSi	225	250	508
mp-19905	Co_2Si	62	120	579
mp-1204045	$Fe_{15}Ge_2U_2$	194	505	619
mp-27505	Fe_3Sn_2	166	670	488
mp-1197772	$Co_{15}Ge_2U_2$	194	650	946
mp-3805	AlB_2Fe_2	65	274	388
mp-1197003	$BCo_{14}Pr_2$	136	793	999
mp-1193316	$Fe_{23}Tb_6$	225	250	490
mp-21469	Co_2SnV	225	95	269
mp-2317	$LaNi_5$	191	603	124

Taking $GdCo_4B$ (space group P6/mmm) as an example, significant magnetic momentum of Co elemental solids provides high possibility for the compound to be FM, which is further strengthened by the unfilled p-shell of all atoms and the crystal structure. The regression model uses the average of the magnetic momentum and d-valence electrons etc. to make a prediction on T_C to be 489 K. This value is then updated by its crystal structure to be 490 K, which is quite close to its experimental value, 517 K.

Table S3. Feature importance of classification model.

Descriptors	Contribution	Importance
mean_GSmagmom	8.773955	0.493264
mean_NpUnfilled	1.623009	0.091244
SOAP_1	1.060342	0.059611
Space Group No.	1.027325	0.057755
maxdiff_MendeleevNumber	0.934828	0.052555
mean_NdValence	-0.79883	0.04491
dev_MendeleevNumber	0.631341	0.035493
dev_Number	-0.57801	0.032495
mean_MendeleevNumber	-0.42977	0.024161
SOAP_18	0.40581	0.022814
min_Number	-0.39057	0.021957
maxdiff_MeltingT	0.383366	0.021553
dev_NdUnfilled	-0.33552	0.018863
mean_MeltingT	-0.31288	0.01759
Volume	-0.10199	0.005734

Table S4. Feature importance of regression model step 1.

Descriptors	Contribution	Importance
mean_GSmagmom	3.912247	0.526456
mean_NdValence	0.686272	0.092349
min_MendeleevNumber	-0.64408	0.086671
most_GSvolume_pa	0.449208	0.060448
dev_GSmagmom	0.382552	0.051479
dev_GSvolume_pa	-0.27059	0.036412
mean_MendeleevNumber	0.228628	0.030766
dev_NdValence	0.185824	0.025006
dev_MeltingT	0.161416	0.021721
max_NdValence	0.113833	0.015318
maxdiff_NdUnfilled	0.10454	0.014067
dev_CovalentRadius	0.095342	0.01283
min_Number	-0.06556	0.008822
dev_Number	-0.06162	0.008292
Comp_L10Norm	0.054779	0.007371
dev_NdUnfilled	0.012707	0.00171
dev_MendeleevNumber	-0.0021	0.000283

Table S5. Feature importance of regression model step 2.

Descriptors	Contribution	Importance
TC_chem	1.98537	0.952096
Space Group No.	0.036274	0.017396
SOAP_1	0.022814	0.010941
Volume	-0.02242	0.010752
SOAP_15	0.01669	0.008004
SOAP_7	0.000868	0.000416
SOAP_21	0.000825	0.000396

Table S6. Predicted vs. experiment T_C of isomers.

Compound	Space group number	Experiment TC	Predicted TC STR	Predicted TC +STR
GdNiAl	189	70	70.76639	67.96717
GdNiAl	194	66	70.76639	63.58434
$Y_2Fe_{15}Al_2C$	166	400	480.681	469.0419
$Y_2Fe_{15}Al_2C$	194	501	480.681	461.9752
$MnAs_{0.8}P_{0.2}$	62	420	291.2587	291.3955
$MnAs_{0.8}P_{0.2}$	194	220	291.2587	226.6723
Fe_3B	82	800	624.5149	671.1125
Fe_3B	62	897	624.5149	667.2541
$Y_2Fe_{15}Si_2C$	166	524	466.3141	460.6831
$Y_2Fe_{15}Si_2C$	194	511	466.3141	436.3283
$Er_2Fe_{17}C_3$	166	680	626.0689	669.6111
$Er_2Fe_{17}C_3$	194	680	626.0689	677.9212
$Cd_{0.75}Cr_2Sb_{0.25}Se_4$	70	130	106.456	86.77514
$Cd_{0.75}Cr_2Sb_{0.25}Se_4$	227	111.8	106.456	105.1845
Ce_2Co_{17}	194	1100	1150.005	1154.594
Ce_2Co_{17}	166	1090	1150.005	1157.409
Rb_2CrCl_4	139	52.8	78.47344	50.64846
Rb_2CrCl_4	64	52.3	78.47344	52.95379
Co	194	1388	1146.911	1157.409
Co	225	1410	1146.911	1157.409
Gd_2Co_{17}	194	1142	1184.031	1198.979
Gd_2Co_{17}	166	1202	1184.031	1205.983
Y_2Co_{17}	194	1208	1157.682	1160.373
Y_2Co_{17}	166	1208	1157.682	1164.562
MnCoGe	62	355	294.0392	303.1521
MnCoGe	194	273	294.0392	222.0136
MnCoSb	216	478	425.0712	450.4354
MnCoSb	225	490	425.0712	424.6246
$NbCo_2Sn$	51	128	328.4883	338.1481
$NbCo_2Sn$	225	105	328.4883	308.2084
$NdCo_2$	227	98	159.812	139.5273

Compound	Space group number	Experiment TC	Predicted TC STR	Predicted TC +STR
$NdCo_2$	141	100	159.812	114.2584
$TbCo_2$	227	235	232.533	243.389
$TbCo_2$	166	240	232.533	230.6885
$DyMn_2$	227	40	62.36883	40.42719
$DyMn_2$	194	37	62.36883	36.37044
Dy_2Fe_{17}	194	367	346.8356	325.3343
Dy_2Fe_{17}	166	360	346.8356	331.2172
$NdErFe_{17}$	194	327	371.1559	379.0717
$NdErFe_{17}$	166	334	371.1559	337.3882
$LiVMnF_6$	102	17	43.14117	33.6176
$LiVMnF_6$	150	26	43.14117	31.79423
$Y_2Fe_{14}Ga_3$	166	485	525.4222	513.662
$Y_2Fe_{14}Ga_3$	194	490	525.4222	513.2205
$Y_2Fe_{15}Si_2$	194	450	462.5268	434.699
$Y_2Fe_{15}Si_2$	166	488	462.5268	451.0083
Gd_2Fe_{17}	166	472	379.8516	398.9523
Gd_2Fe_{17}	194	506	379.8516	402.2726
Tb_2Fe_{17}	166	415	350.9004	349.7142
Tb_2Fe_{17}	194	413	350.9004	359.8412
Y_2Fe_{17}	194	310	374.4155	347.8273
Y_2Fe_{17}	166	300	374.4155	344.3138
FeRhP	189	355	245.8468	230.5933
FeRhP	59	235	245.8468	267.7877
$HfFe_2$	227	625	453.9123	475.6601
$HfFe_2$	194	595	453.9123	442.6048
Fe_3Ga	225	750	660.7569	669.7664
Fe_3Ga	221	730	660.7569	668.171
Fe_3Ge	194	634	578.8974	654.4591
Fe_3Ge	221	655	578.8974	642.7742
Fe_7S_8	152	578	241.6827	167.9677
Fe_7S_8	15	578	241.6827	267.67
$MnNi_2Ga$	225	363	373.6219	335.8301
$MnNi_2Ga$	139	358	373.6219	328.1502

Reference

(1) Hellenbrandt, M. The Inorganic Crystal Structure Database (ICSD)—Present and Future. *Crystallography Reviews* **2004**, *10* (1), 17–22. https://doi.org/10.1080/08893110410001664882.

(2) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Materials* **2013**, *1* (1), 011002. https://doi.org/10.1063/1.4812323.

(3) Xu, Y.; Yamazaki, M.; Villars, P. Inorganic Materials Database for Exploring the Nature of Material. *Jpn. J. Appl. Phys.* **2011**, *50* (11S), 11RH02. https://doi.org/10.1143/JJAP.50.11RH02.

(4) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials. *npj Computational Materials* **2016**, *2*, 16028. https://doi.org/10.1038/npjcompumats.2016.28.

(5) Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B* **2013**, *87* (18), 184115. https://doi.org/10.1103/PhysRevB.87.184115.

(6) Deringer, V. L.; Csányi, G. Machine Learning Based Interatomic Potential for Amorphous Carbon. *Phys. Rev. B* **2017**, *95* (9), 094203. https://doi.org/10.1103/PhysRevB.95.094203.

(7) Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of Descriptors for Machine Learning in Materials Science. *arXiv:1904.08875 [cond-mat]* **2019**.

(8) De, S.; P. Bartók, A.; Csányi, G.; Ceriotti, M. Comparing Molecules and Solids across Structural and Alchemical Space. *Physical Chemistry Chemical Physics* **2016**, *18* (20), 13754–13769. https://doi.org/10.1039/C6CP00415F.

(9) Lundberg, S. M.; Erion, G. G.; Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv:1802.03888 [cs, stat]* **2019**.

(10) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; pp 4765–4774.

(11) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D. Scikit-Learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON* 6.

(12) Kanomata, T.; Endo, K.; Kudo, N.; Umetsu, R. Y.; Nishihara, H.; Kataoka, M.; Nagasako, M.; Kainuma, R.; Ziebeck, K. R. A. Magnetic Moment of Cu-Modified Ni2MnGa Magnetic Shape Memory Alloys. *Metals* **2013**, *3* (1), 114–122. https://doi.org/10.3390/met3010114.

4.2. Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures

In terms of inverse design, *i.e.*, prediction of novel materials with desired properties, constrained crystal deep convolutional generative adversarial networks (CCDCGAN) has been developed, which can directly predict crystal structures distinct from the known cases based on a proper construction of the latent space. Moreover, intrinsic properties can be optimized in the latent space, which offers the opportunity to achieve multi-objective optimization. Currently, formation energy has been integrated into the generative model as a constraint. Correspondingly, CCDCGAN can design stable crystal structures. It has been successfully applied on a randomly chosen binary system (Bi-Se). It is observed that unreported crystal structures that are more stable than the structures in the training set (such as Bi_2Se_4) can be obtained.

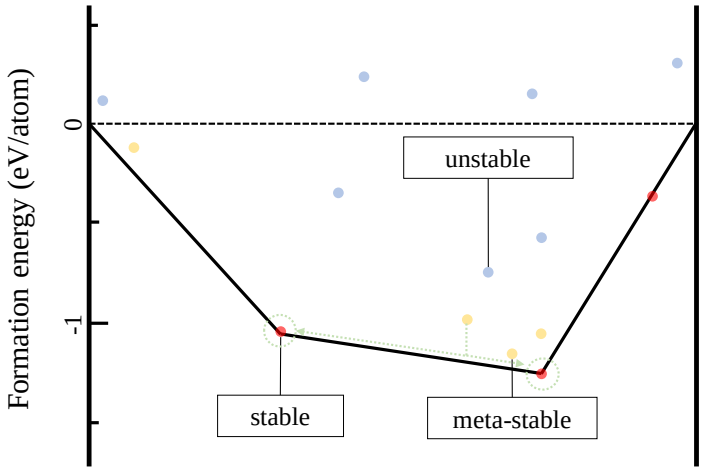


Figure 4.1.: Schematic of formation energy and convex hull for a binary system, where the x axis represents the composition, the y axis denotes the formation energy, red, yellow and blue points indicates the stable, meta-stable and unstable compounds, respectively.

However, the formation energy only measures the relative stability with respect to the elements. In order to check whether a designed material can be synthesized, the distance to the convex hull should be checked, which is defined with respect to the available stable

phases (mostly not the elements). For instance, the metastable phase marked in Figure 4.1 has a negative formation energy. However, it can decompose into the neighboring stable phases at the thermodynamic equilibrium, as denoted by the arrow in Figure 4.1. It is noted that whether the metastable phases can be synthesized experimentally does not depend on only the thermodynamics but also the kinetics. Thus, those phases with distances to the convex hull smaller than a tolerance value (100 meV/atom) can possibly be synthesized experimentally.

The successful application of CCDCGAN on the Bi-Se system is based on a training set of 10,000 hypothetical structures, which can, in principle, be done for all the other binary cases, entailing tremendous computational time. Thus, an interesting question is when the chemical and structural space can be sufficiently sampled so that inverse design can be performed for elements in the whole periodic table. To address this problem, we recently generalized CCDCGAN to the multicomponent systems, *i.e.*, designing crystal structures of random binary, ternary, or quaternary systems. The workflow of the generalized CCDCGAN is illustrated in Figure 4.2, where the crystal images are defined in a larger latent space when compared to the latent space of Bi-Se. After training the extended model with about 50,000 crystalline materials from Materials Project, it is observed that many unreported crystal structures below the convex hull can be obtained, leading to a systematic way of designing novel materials, in particular intermetallic compounds. More details are demonstrated in Appendix B.

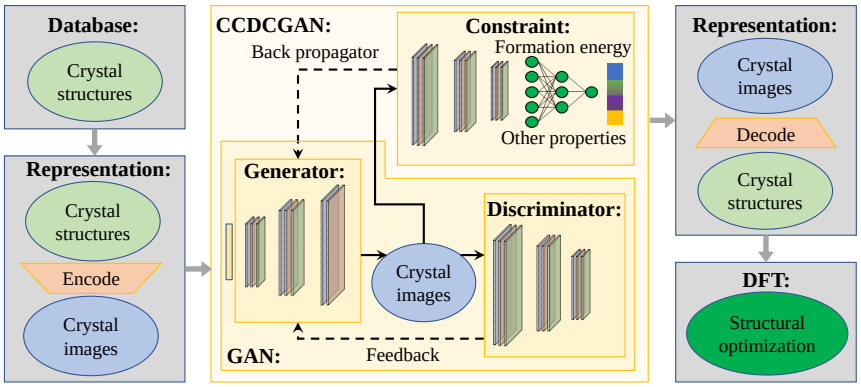


Figure 4.2.: Workflow of CCDCGAN. Three key elements of CCDCGAN, *i.e.*, generator, discriminator, and constraint, are highlighted by shaded blocks in yellow. Such models act on the crystal images, which are one-to-one mapping to the crystal structures, with the transformations managed by the autoencoder.

npj Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures

Teng Long[1]✉, Nuno M. Fortunato[1], Ingo Opahle[1], Yixuan Zhang[1], Ilias Samathrakis[1], Chen Shen[1], Oliver Gutfleisch[1] and Hongbin Zhang[1]✉

Autonomous materials discovery with desired properties is one of the ultimate goals for materials science, and the current studies have been focusing mostly on high-throughput screening based on density functional theory calculations and forward modeling of physical properties using machine learning. Applying the deep learning techniques, we have developed a generative model, which can predict distinct stable crystal structures by optimizing the formation energy in the latent space. It is demonstrated that the optimization of physical properties can be integrated into the generative model as on-top screening or backward propagator, both with their own advantages. Applying the generative models on the binary Bi-Se system reveals that distinct crystal structures can be obtained covering the whole composition range, and the phases on the convex hull can be reproduced after the generated structures are fully relaxed to the equilibrium. The method can be extended to multicomponent systems for multi-objective optimization, which paves the way to achieve the inverse design of materials with optimal properties.

npj Computational Materials (2021)7:66 ; https://doi.org/10.1038/s41524-021-00526-4

INTRODUCTION

The discovery and exploitation of materials have enormous benefits for the welfare of society and technological revolutions[1], which motivates the launch of the Materials Genome Initiative in 2011[2,3]. Till now, high-throughput (HTP) workflows based on density functional theory (DFT) enable massive calculations on existing and hypothetical compounds, accelerating materials discovery dramatically[4]. For instance, crystal structure predictions can be performed based on brute force substitution of the known prototypes or the evolutionary algorithms as implemented in CALYPSO[5] and USPEX[6]. Nevertheless, the soaring computation cost prevents exhaustive screening over immense phase space, limiting the application of such methods and calling for more efficient solutions. Global optimization methods added a target function to guide the searching and therefore reduce the computational cost[7], e.g., crystal structure prediction with Bayesian optimization can efficiently select the most stable structure from a large number of candidate structures with a lower number of searching trials and has been successfully applied to predict structures of NaCl and Y_2Co_{17}[8]. The random search method has also demonstrated its potential to find the global minimum with highly parallel implementation, as manifested by the applications on elemental boron, silicon cluster, and lithium hydrides[9,10]. Last but not least, the evolutionary algorithm is promising in predicting thermodynamically stable crystal structures, such as demonstrated for tungsten borides, carbon, and TiO_2[11,12].

The emergent machine learning techniques have become the fourth paradigm for materials science[13], as exemplified by the rapid developments and applications in the last decade[14]. Particularly, the generative machine learning methods have been utilized to discover distinct materials, e.g., applying neural networks to assist molecule design[15,16]. Two particularly powerful methods therein are the variational autoencoder (VAE)[17] and generative adversarial network (GAN)[18], resulting in significant progresses on molecule design[19]. Combined with the successful forward modeling of physical properties[20], generative machine learning on the existed structures enables inverse design, i.e., prediction of distinct structures with the desired properties[1,21]. Unfortunately, unlike molecules that can be represented using simplified molecular input line entry specification (SMILES)[22], the inverse design of three-dimensional (3D) crystal structures has been rare due to the challenges in obtaining a continuous representation in the latent space[20].

Recently, Nouira et al. developed a CrystalGAN model to design ternary A-B-H phases starting from binary A-H and B-H structures using a vector-based representation[23], and successfully applied it on the Ni-Pd-H system. Another promising representation is the two-dimensional (2D) crystal graphs constructed based on the voxel and autoencoder methods[24–26]. For instance, Noh et al. developed the iMatGen model to represent the image-based crystals continuously, and successfully generated distinct V_xO_y using VAE[27]. The ZeoGAN scheme developed by B. Kim et al. is based on the same representation applied on zeolites, where Wasserstein GAN (WGAN) is used to generate porous materials[28]. S. Kim et al. also developed the Crystal-WGAN to design Mn-Mg-O system using such a continuous representation[29]. These models exhibit excellent abilities in generating distinct structures, but not all predicted phases host the desired functionality. For example, only ZeoGAN integrates the optimization of the heat of adsorption, whereas the other schemes exert mostly constraints to filter the generated structures. It is noted that by nature, inverse design entails the optimization of properties in the latent space, and thus internalizes the desired property as a joint objective of the generator[30].

In this work, we developed a GAN-based inverse design framework for crystal structure prediction with target properties and applied it to the binary Bi-Se system. It is first demonstrated that our deep convolutional generative adversarial network (DCGAN) can be applied to generate distinct crystal structures[31]. Taking formation energy as the target property, its optimization is integrated into the DCGAN model in two schemes: DCGAN +

[1]Institute of Materials Science, Technische Universität Darmstadt, Darmstadt, Germany. ✉email: tenglong@tmm.tu-darmstadt.de; hongbin.zhang@tu-darmstadt.de

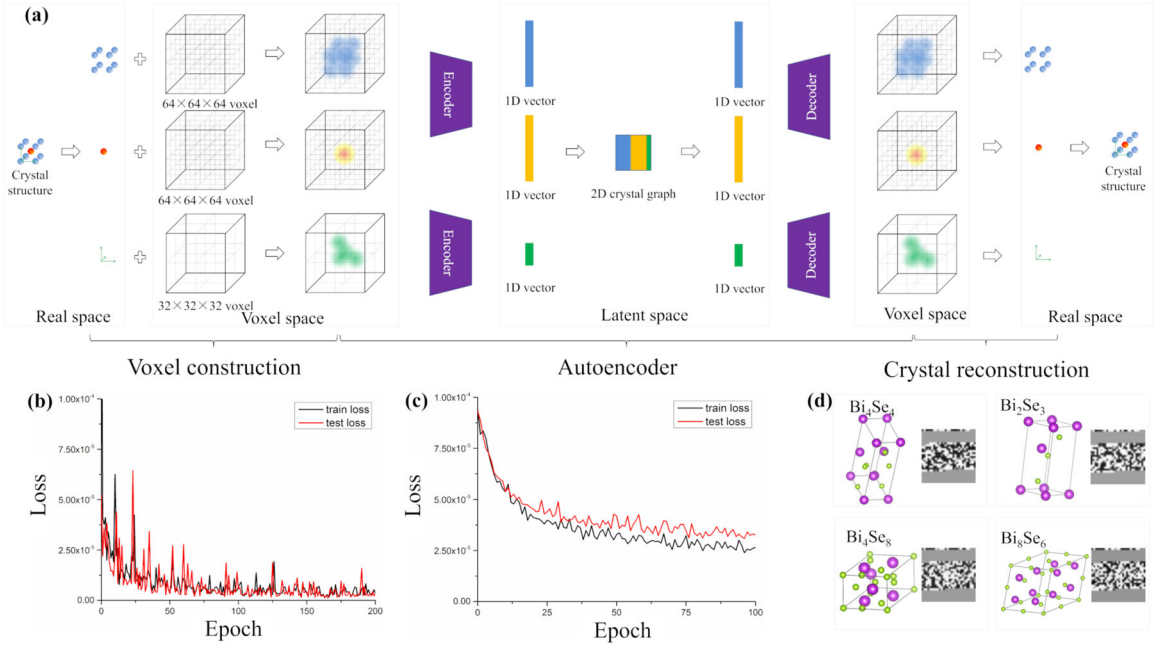


Fig. 1 Construction of 2D crystal graphs. (a) Schematic diagram of generating 2D crystal graphs; (**b**) Learning curve of lattice autoencoder, where the red line represents the test loss and the black line represents the training loss; (**c**) Learning curve of sites autoencoder, where the red line represents the test loss and the black line represents the training loss; (**d**) Four typical crystal structures (Bi_4Se_4, Bi_2Se_3, Bi_4Se_8, Bi_8Se_6) and their corresponding 2D crystal graphs.

constraint to select structures following the conventional screening method, and constrained crystals deep convolutional generative adversarial network (CCDCGAN) with an extra feedback loop for automatic optimization. The performance of the DCGAN + constraint and CCDCGAN models are comparatively evaluated, and it is demonstrated CCDCGAN is more efficient in generating stable structures, while the DCGAN + constraint model has a higher generation rate of meta-stable structures. Both schemes can be generalized to perform multi-objective inverse design for multicomponent systems[32], hence our work paves the way to achieve autonomous inverse design of distinct crystal phases with desired properties.

RESULTS

Data for generative adversarial network

Typically, DCGAN requires 10^3 known crystal structures as positive examples during training. However, there are mostly tens of known crystalline phases for a given binary system, e.g., Materials Projects (MP)[33] provides only 17 already known Bi_xSe_y materials, which are not sufficient for deep learning. In order to get enough training data, DFT calculations are performed on the prototype binary structures used in Ref. [27], where are selected based on two criteria: (1) there are no >20 atoms in the unit cell and (2) the maximum lattice constant is smaller than 10 Angstrom. Starting from 10981 prototype structures, we managed to converge 9810 cases after substituting Bi and Se atoms[34], which are taken as training databases. Hereafter, such a database is referred to as the Bi-Se database. Out of the 15 experimentally achievable phases[35], 8 structures (i.e., $BiSe_3$, Bi_2Se_3, Bi_4Se_5, Bi_3Se_2, Bi_5Se_3, Bi_2Se, Bi_7Se_3, and Bi_3Se) are included in the database, whereas the other 7 compounds (i.e., $BiSe_2$, Bi_3Se_4, Bi_8Se_9, $BiSe$, Bi_8Se_7, Bi_6Se_5, and Bi_4Se_3) are excluded by the selection criteria, as demonstrated in Supplementary Fig. 1. Furthermore, out of the Bi-Se database, 155 (1.56%) structures of the database are stable (i.e., formation energy lower than 0 eV/atom and distance to the convex hull smaller than 100 meV/atom) and 707 (7.2%) cases are meta-stable (i.e., with negative formation energy and distance to the convex hull smaller than 150 meV/atom).

Continuous representation for crystal structures

Although the crystallographic and chemical information is stored in the crystallographic information file (CIF), the discontinuous and heterogeneous formats are not a suitable choice of representation in a generative model, thus a continuous and homogeneous representation including both the chemical and structural information is required. Following Ref. [27], the lattice constants and atomic positions are translated into the voxel space, followed by encoding into a 2D crystal graph through autoencoder, as demonstrated in Fig. 1(a). In this way, a continuous latent space is constructed. The whole process is reversible, i.e., a random 2D crystal graph can be reconstructed into a crystal structure in real space, which is essential for a generative model. When applied to the Bi-Se database, 9420 of 9810 (i.e., 96%) crystal structures are successfully reconstructed. Such a high ratio is not caused by overfitting, illustrated by the learning curve of both autoencoders as shown in Fig. 1(b, c). The differences of training set loss and test set loss are both negligible, suggesting that the 2D crystal graphs in the latent space contain adequate information to reconstruct crystal structures. This is confirmed by the diversity of twelve typical 2D crystal graphs in the latent space (cf. Fig. 1(d) for four cases). Please refer to "Methods" as well as Supplementary Information for more details.

Construction and prediction of DCGAN

Turning now to the realization of generative models, VAE[17] and GAN[18] are two most popular algorithms. VAE is a mutation of the autoencoder discussed above, which assumes a specific (such as Gaussian) distribution of data (in our case 2D crystal graphs) in the latent space. To be generative, such a distribution function should be defined properly, i.e., to be consistent with the distribution of 2D crystal graphs of the known crystal structures in the latent space. In addition, the distribution function has to be specified prior to the training process and its form determines the

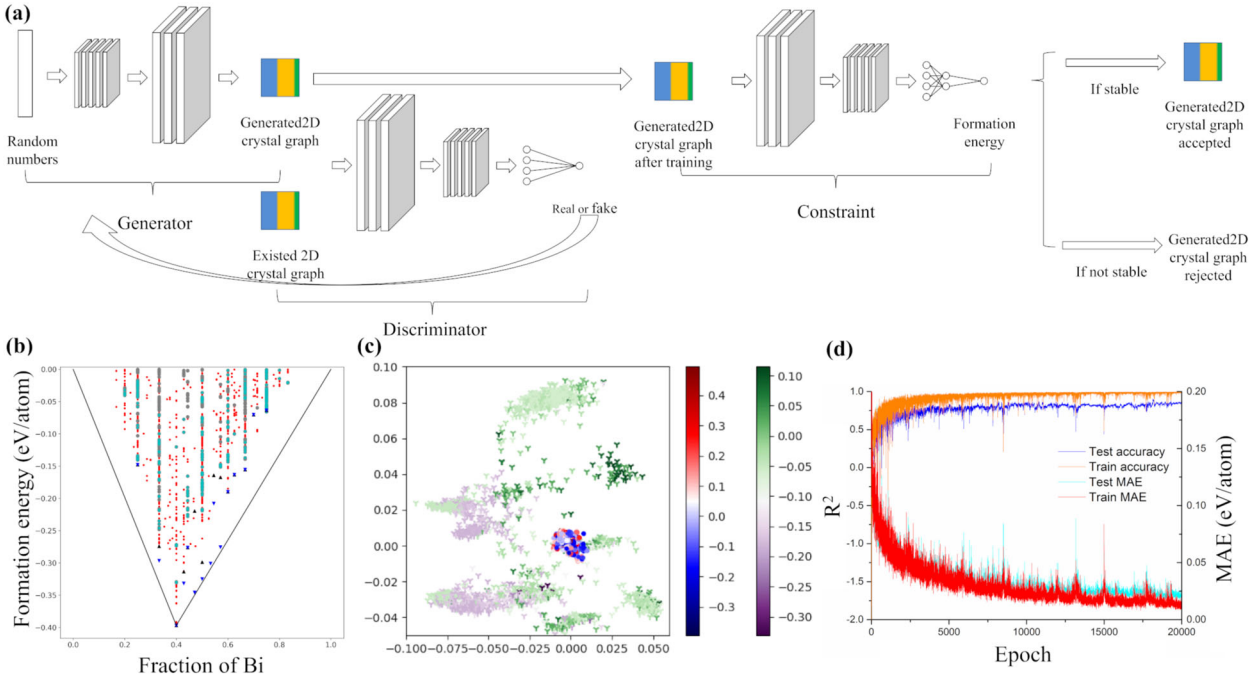


Fig. 2 Illustration on the DCGAN + constraint model. (**a**) Schematic diagram of the DCGAN + constraint model; (**b**) Scatter plot of formation energy vs. composition for structures generated by DCGAN and DCGAN + constraint models, where the black line represents the convex hull, red points denote the crystal structures considered in our machine learning database, gray circles indicate the generated structures, cyan stars denote the DCGAN + constraint structures, blue triangles indicate the experimentally achievable phases (cf. Supplementary Table 7), black triangles mark the crystal structures of the available Bi-Se phases in the Bi-Se database; (**c**) Distribution of the crystal structures in the latent space, the circles denote the structures in the Bi-Se database, while the stars denote the generated structures. The color maps denote the corresponding formation energy; (**d**) Learning curves of the training process of the DCGAN model, red and orange represent the MAE and accuracy of the training set, while blue and cyan represent the accuracy and MAE of the test set.

performance of the generative model, which demands domain expertise in statistics and profound understanding of the input data. For instance, a recent work by Ren et al. used sigmoid function in VAE to design inorganic crystals and find structures that do not exist in training set[36]. Court et al. implemented VAE to generate crystal structures and successfully applied it to a broad class of materials including binary and ternary alloys, perovskite oxides, and Heusler compounds[37]. In contrast, GAN trains two mutual-competitive neural networks (i.e., generator and discriminator) to generate data statistically the same as the training data without assuming the distribution function[18]. That is, the discriminator tries to distinguish the generated data from the training data, while the generator attempts to fool the discriminator by generating data similar to the training data. Mathematically the objective of GAN is defined by the following equation:

$$\max_D \left(\min_G \left(\frac{1}{2} \cdot E_{\mathbf{x}\sim p_x}[1 - D(\mathbf{x})] + \frac{1}{2} \cdot E_{\mathbf{x}\sim p_g}[D(\mathbf{x})] \right) \right) \quad (1)$$

where D is the discriminator, G is the generator, E means expectation value, $\mathbf{x}$ is the original data, $D(\mathbf{x})$ is the output of discriminator with $\mathbf{x}$ as input, p_x is the possibility density function of the original data, while p_g is the possibility density function of the generated data. Through the competition between the generator and discriminator, the distribution of generated data becomes hardly distinguishable with the distribution of training data. Compared with VAE, GAN does not require specifying a distribution function in the latent space (p_g), which makes the generation process more robust, and thus is adopted in this work[19].

Such an implementation of DCGAN can generate crystal structures with a high success rate (defined as the ratio of the number of generated crystals over the number of generated 2D crystal graphs), e.g., 2832 crystal structures are transformed from 13,000 generated 2D crystal graphs. The generated structures cover a large composition range as shown in Fig. 2(b), where the red points denote the original data in the Bi-Se database and the gray circles mark the generated structures by DCGAN. One interesting observation is that the generated structures are mostly with negative formation energies, on average lower than that of the original database (Supplementary Fig. 2(a)). For instance, 2548 structures out of 2832 have negative formation energies after follow-up DFT relaxations, which consists of 89.9% of the generated structures, whereas only 46.8% (4588 out of 9810) of the Bi-Se database have negative formation energies. Out of 2832 generated structures, it is found that 1233 of them are meta-stable, and 58 are stable.

Detailed analysis reveals that 476 out of the 2832 structures are considered to be distinct structures, i.e., different from those in the Bi-Se database and different from each other, leading to 73 (15) meta-stable (stable) structures, respectively. For instance, Bi_2Se_4 and $BiSe_3$ shown in Supplementary Fig. 2(b, c) are the automatic generated structures, whose formation energies (distances to the convex hull) are −0.192 (0.136) eV/atom and −0.135 (0.111) eV/atom, respectively. Furthermore, DCGAN explores a much larger phase space when generating structures. As illustrated in Fig. 2(c), the structures in the Bi-Se database are concentrated in a small region in the latent space, whereas much larger phase space is covered by DCGAN. Thus, DCGAN can generate distinct crystal structures beyond the phase space of the known crystal structures.

Training of constraint model

As mentioned above, the objective of inverse design is to design compounds with desired properties, including thermodynamical, mechanical, and functional properties. In this work, we take the formation energy (i.e., thermodynamic stability) as the target

property. In order to be able to optimize the formation energy in the latent space, another convolutional neural network (CNN) model is trained taking 2D crystal graphs in the latent space as descriptors and formation energies as the output physical property. The training is carried out by a 90%/10% training / test-set ratio over the Bi-Se database, with the resulting R^2 score and mean absolute error (MAE) being about 85% and 0.019 eV/atom (Supplementary Fig. 3(a)), respectively. Such a MAE is even smaller than that (0.021 eV/atom) obtained using the state-of-the-art CGCNN model[38]. That is, the 2D crystal graphs in the latent space can be considered as effective descriptors to perform forward prediction of physical properties. We note that overfitting is a marginal issue for such high accuracy, as seen in the learning curve of the training process in Fig. 2(d). Details on the CNN model are described in "Methods" and Supplementary Information.

Construction and prediction of DCGAN + constraint

A straightforward way to implement the optimization of physical properties, e.g., formation energy in this work, is to apply an add-on constraint on the generated structures using DCGAN, as sketched in Fig. 2(a). The advantage of doing so is that there is no need to train another model, which saves training time. In addition, all the existing machine learning models can be transplanted no matter whether such forward predictions can take 2D crystal graphs in the latent space or vector-based chemical and structural information as descriptors. This allows the optimization of a wide spectrum of physical properties[20]. Nevertheless, this method is essentially a selection of the structures generated by DCGAN, thus it cannot search automatically for a specific region in the latent space to reach the local optimal values. More details are described in Supplementary Information.

The DCGAN + constraint model demonstrates its advantage in generating metastable structures. Applying the constraint on the formation energy (with a tolerance of 0.3 eV/atom) on 2832 generated structures by DCGAN, 2148 crystal structures are selected and optimized by further DFT calculations. Such a screening procedure does not affect the diversity of the composition as shown in Supplementary Fig. 2(d). The application of such a constraint has evident effect in the latent space, as demonstrated in Supplementary Movie 1 where the constraint screens the unwanted points out leading to a shrunk region in the latent space. With such a constraint applied, the ratio of structures with negative formation energy is higher, increasing from 89.9% for DCGAN to 94.8% (2037/2148). The ratio of meta-stable structures also becomes higher compared to the DCGAN case, i.e., 1145 meta-stable structures are generated from 2037 structures, reaching the ratio of 56.2% which is better than 43.5% in the DCGAN model. However, the number of generated stable structures reduces to 36 due to the selection process.

An interesting observation is that applying the constraint jeopardizes the generation of distinct structures. After performing DFT calculations, only 247 distinct structures remain, which are significantly reduced compared with the previous 476 cases obtained by DCGAN. The reason is two-fold. On the one hand, the generated crystal structures by DCGAN are not guaranteed to be at the mechanical and dynamical equilibrium, i.e., the lattice constants and atomic positions will change during DFT relaxation. On the other hand, the CNN model applied as constraint is not good enough in predicting formation energy, though the MAE for the formation energy is only about 0.019 eV/atom. Thus, 629 candidates have been screened out by considering a tolerance of 0.3 eV/atom. After comparing with the known crystal structures in the Bi-Se database, 67 distinct meta-stable structures and 12 distinct stable structures are obtained, e.g., two of them Bi_3Se and Bi_2Se_2 are demonstrated in Supplementary Fig. 2(e, f), with the formation energies (distances to the convex hull) being −0.065 (0.099) eV/atom and −0.202 (0.126) eV/atom, respectively. Overall, generated structures via DCGAN + constraint have a higher ratio of meta-stable structures.

Construction and prediction of CCDCGAN

The constraint can also be integrated into DCGAN as a back propagator, as illustrated as CCDCGAN in Fig. 3(a), to realize automated optimization in the latent space so that inverse design can be realized. In this way, the constraint gives rise to an additional optimization objective of the generator, which can be mathematically described as:

$$\min_G \left(\frac{1}{2} \cdot E_{\mathbf{z} \sim p_x}[D(\mathbf{z})] + \omega \cdot E_f(\mathbf{z}) \right), \mathbf{z} \sim p_g, \tag{2}$$

where E_f is the formation energy predicted by the constraint model, $\mathbf{z}$ is the generated 2D crystal graph, and ω is defined as the weight of formation energy loss. Note that such an additional optimization objective cannot outweigh the primary objective, leading to lower weight for the formation energy loss (0.1 in this work) than the discriminator loss. Unlike the DCGAN + constraint model, CCDCGAN can accomplish automated searching for the local minima in the latent space and thus improve the efficiency of discovering distinct stable structures. It is noted that CCDCGAN requires training from scratch, which takes 18 h on a Quadro P2000 GPU machine, which is 2 h longer than the time needed for training the DCGAN + constraint model. Please refer to "Methods" and Supplementary Information for more details.

In comparison with the other two models (i.e., DCGAN and DCGAN + constraint models), CCDCGAN has a higher success rate of generating crystal structures, i.e., 3743 crystal structures are transformed successfully from 13,000 generated 2D crystal graphs. Correspondingly, the number of structures that are with negative formation energy and meta-stable is 3556 and 1910, respectively. Furthermore, CCDCGAN has a higher generation efficiency of stable structure (307 out of 3743) when compared with the other two models, which suggests that the back propagation indeed causes optimization in the latent space. This is also proved by the average formation energy of generated structures, where the average formation energy (−0.11 eV/atom) for the CCDCGAN generated structures is lower than that from the DCGAN (−0.08 eV/atom) and DCGAN + constraint (−0.09 eV/atom).

The most intriguing result is that the number of distinct structures is bigger than that from DCGAN or DCGAN + constraint, where 511 distinct structures have been identified after performing DFT relaxations on 3743 structures. Correspondingly, the number of distinct meta-stable and stable structures is 187 and 42, respectively. Such numbers are similar to those obtained using iMatGen and Crystal-WGAN[27,29], indicating comparable performance of our DCGAN-based models with the existing ones. Three particularly promising structures with formation energy lower than their counterparts in Bi-Se database are Bi_2Se_4, Bi_4Se_2, and Bi_6Se_2, whose formation energies (distance to the convex hull) are −0.287 (0.041) eV/atom, −0.139 (0.069) eV/atom, and −0.063 (0.101) eV/atom, respectively. In this regard, CCDCGAN performs better than both the DCGAN + constraint and DCGAN models in generating distinct structures.

To explore the full competence of the CCDCGAN model, we generated 100,000 crystal structures and performed DFT calculation on the generated structures. The CCDCGAN model is able to regenerate most (11 out of 15) experimentally achievable phases. As seen from Fig. 3(c) and Supplementary Table 7, all the 8 experimentally achievable phases in the Bi-Se database can be regenerated using our CCDCGAN model, i.e., exactly the same crystal structures as the experimentally achievable phases of $BiSe_3$, Bi_2Se_3, Bi_4Se_5, Bi_3Se_2, Bi_5Se_3, Bi_2Se, Bi_7Se_3, and Bi_3Se can be obtained. Furthermore, our CCDCGAN model can also generate three out of seven non-included experimentally achievable phases, namely, $BiSe_2$, Bi_3Se_4, and Bi_6Se_5, with exactly the same crystal structures as

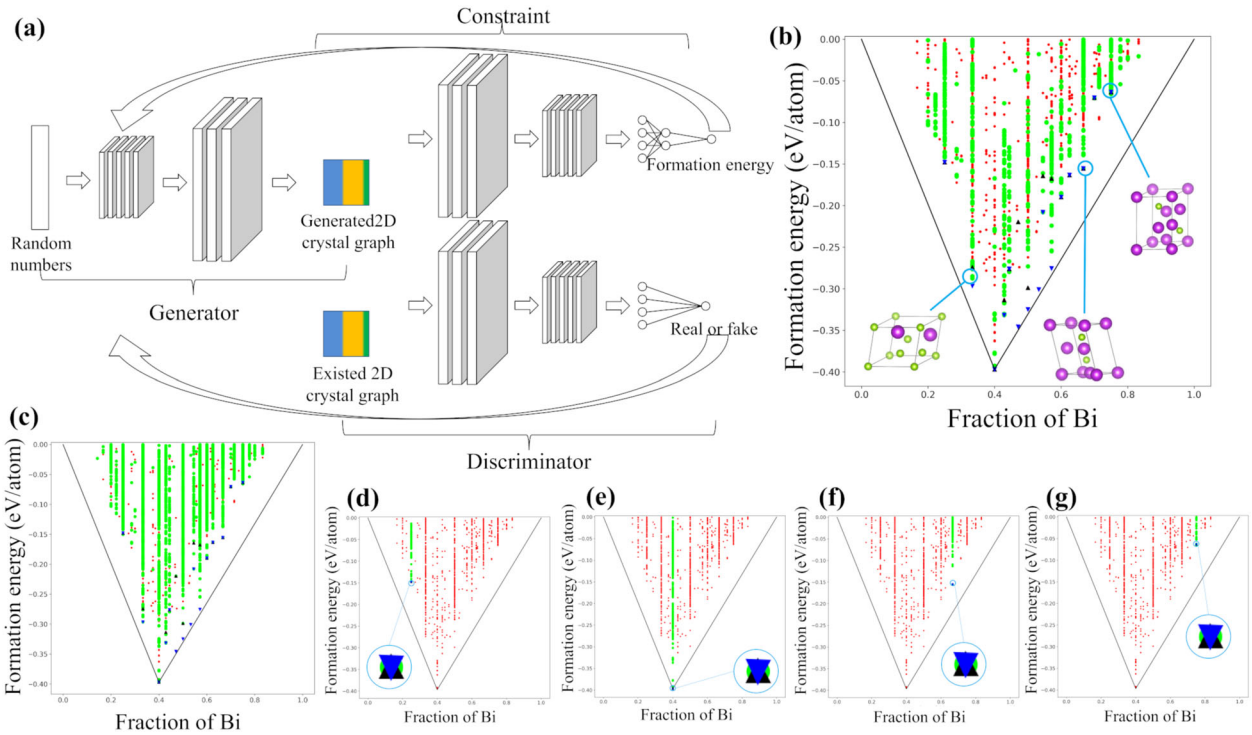


Fig. 3 Illustration on the CCDCGAN model. (**a**) Schematic diagram of the CCDCGAN model; (**b**) Scatter plot of formation energy vs. composition of CCDCGAN generated structures, where the black line represents the convex hull, red points denote the crystal structures considered in our machine learning database, green circles represent the generated structures, blue triangles indicate the experimentally achievable phases, black triangles are the crystal structures of the available Bi-Se phases in the Bi-Se database; and crystal structure of generated Bi_2Se_4, Bi_4Se_2, and Bi_6Se_2; (**c**) Scatter plot of formation energy vs. composition of 100,000 structures generated by CCDCGAN; (**d**) The $BiSe_3$ phases regenerated by CCDCGAN, which are trained with the $BiSe_3$ composition removed; (**e**) The Bi_2Se_3 phases regenerated by CCDCGAN, which are trained with the Bi_2Se_3 composition removed; (**f**) The Bi_2Se phases regenerated by CCDCGAN, which are trained with the Bi_2Se composition removed; (**g**) The Bi_3Se phases regenerated by CCDCGAN, which are trained with the Bi_3Se composition removed.

reported. Unfortunately, the other four phases, namely, Bi_8Se_9, BiSe, Bi_8Se_7, and Bi_4Se_3, out of seven non-included phases cannot be reproduced, where the generated structures of BiSe and Bi_4Se_3 have comparable formation energy as the experimentally achievable structures, while the formation energies of generated Bi_8Se_9 and Bi_8Se_7 structures are higher than those reported phases. It is suspected that this can be attributed to the constraints (i.e., unit cell dimension and number of atoms) we applied to select the training database, which will be extended in the future.

To further test the predictive capability of the CCDCGAN model, we deliberately removed the crystal structures of 4 specific compositions (i.e., $BiSe_3$, Bi_2Se_3, Bi_2Se, and Bi_3Se) from the training database, and observed that the experimental crystal structures of such four phases can still be successfully regenerated, as shown in Fig. 3(d, g). This proves that the CCDCGAN model can generate crystal structures for unknown compositions, and thus it could accelerate the discovery of distinct crystal phases.

DISCUSSION

We have developed an inverse design framework CCDCGAN consisting of a generator, discriminator, and constraint and successfully applied it to design unreported crystal structures with low formation energy for the binary Bi-Se system. It is demonstrated that 2D crystal graphs can be used to construct a latent space with the continuous representation of the known crystal structures, which serve as effective descriptors for modeling the physical properties (e.g., formation energy) and can be decoded into real space crystal structures enabling the generation of distinct crystal structures. Such an inverse design model can be easily generalized to the multicomponent cases, as demonstrated by Ref. [29]. Importantly, we elucidate that the optimization of physical properties (e.g., formation energy) can be integrated into the generative deep learning model as explicit constraints or back propagators. This allows further development of a multi-objective inverse design framework to optimize other physical properties by modifying the objective function, as demonstrated by Ref. [28]. One apparent challenge is how to get the generated structures into their mechanical and dynamical equilibria, entailing further exploration such as relaxation using the machine learning interatomic potential[39].

METHODS

Data for generative adversarial network

The Bi-Se database is based on the substitution of all binaries of MP, in the spirit of high-throughput calculation, previous research has selected 10981 structures of them by eliminating the large unit cell, i.e. only select structures with the maximum number of atoms in unit cell < 20 and the maximum length of unit cell smaller than 10 Angstrom, and we do the substitution based on their result[27,29]. Following this criteria, several stable phases like BiSe, Bi_4Se_3, Bi_8Se_7 close to the convex hull are screened out. We use the high-throughput environment to optimize the structure and determine the thermodynamical stability of Bi-Se database and the generated structures[40], where thermodynamical stability is evaluated by calculation of the formation energy. All data in the Bi-Se database are considered to calculate the convex hull. For VASP setting[41], Perdew–Burke–Ernzerhof (PBE) is adopted, the energy cutoff is 500 eV, and K space density is 40 in each direction for the Brillouin zone. The Bi-Se database is in Supplementary Data 1.

Continuous representation for crystal structures

Specifically, 3D voxel grids are used for a typical binary compound: two grids to record the atomic positions of two elements separately and the

third one to store the lattice constants, i.e., the lengths and angles of/between them. Hereafter such grids are referred to as site voxel and lattice voxel, respectively. Using the probability density based on Gaussian functions, the lattice voxel is defined as

$$p_{x,y,z} = e^{-\frac{r^2_{x,y,z}}{2\sigma^2}}, \quad (3)$$

where *x, y, z* denotes the grid index, *p* is the probability density of the denoted grid, **r** is the real space distance between this grid to the center of the unit cell, and σ denotes the standard deviation of the Gaussian function. Similarly, sites voxel is defined as

$$p_{x,y,z} = \sum_{i=1}^{n} e^{-\frac{r^2_{n,x,y,z}}{2\sigma^2}}, \quad (4)$$

where *n* is number of atoms and r is the real space distance between this grid point and the atom. In this way, the inverse transformation is trivial for the lattice voxel while that for the sites voxel relies on the image filter technique[27].

Autoencoder is a type of artificial neural network, which is able to encode inputs into low dimension vectors, as well as decode the vectors back[42]. In this work, the autoencoder is applied to encode 3D voxel data into one-dimensional (1D) vectors in the latent space, which is realized based on 3D convolutional neural network (CNN) autoencoder. The loss function of autoencoder consists of two parts: the first is the loss of information, i.e., the difference between inputs and outputs, and the second is the regulation term to prevent overfitting, which yields

$$\text{loss} = \|\mathbf{x} - \mathbf{y}\|^2 + \frac{\lambda}{2} \cdot \sum \boldsymbol{\omega}^2, \quad (5)$$

where **x** and **y** are the input and output of the autoencoder, λ is regulation coefficient, and ω is weight in the autoencoder[17]. The construction of autoencoder for both site and lattice voxel are the same, where both are trained using a 90%/10% training/test set ratio. Both site and lattice voxel autoencoders have high reconstruction ratio. All of them are considered to be identical by the crystal structure comparison subroutine available in pymatgen[43], with fractional length tolerance being 0.2, sites tolerance 0.3, and angle tolerance 5 degrees.

For data transformation between real space and voxel space, the voxel used for atomic positions is $64 \times 64 \times 64$, while the voxel for lattice parameters is $32 \times 32 \times 32$ and the σ used in Gaussian function is always 0.15, function "gaussian_filter" from scipy package is used[27]. The autoencoders constructed by tensorflow in python. The sites autoencoder consists of 10 3D convolutional layers, i.e. 5 for encoder and 5 for decoder, strides except the first and last on are $2 \times 2 \times 2$ and activation function is leakyRELU, Adam optimizer is used, learning rate is 0.003, λ is 0.000001, β_1 is 0.9 and β_2 is 0.99[27]. Similar, the lattice autoencoder consists of only 8 3D convolutional layers (4 for encoder and 4 for decoder), other parameters are the same as the previous one[27]. Detailed design of these two models can be found in Supplementary Tables 1 and 2.

Training of DCGAN

The left half of Fig. 2(a) illustrates our implementation of DCGAN to generate crystal structures. First, both the generated 2D crystal graphs by the generator and original 2D crystal graphs are fed into the discriminator, which is trained to distinguish such graphs. Afterward the generator is further trained by the feedback from the discriminator through back propagation to generate structures more similar to the original structures. By repeating the processes, the generator will evolve to be able to eventually generate structures that are indistinguishable (by the discriminator) to the original structures. Details to construct DCGAN can be found in "Methods" and Supplementary Information.

The DCGAN model consists of generator and discriminator, they both use Adam optimizer with 0.002 learning rate, β_1 is 0.5 and β_2 is 0.999. 2D convolutional layer, dropout and batch normalization are used in both model, while RELU, leakyRELU, tanh, and sigmoid are used as activation functions. 1,000,000 steps are trained for this model for 16 h on Quadro P2000 GPU. The design of the generator and discriminator are listed in Supplementary Tables 3 and 4, respectively. The latent space of the GAN model has 200 dimensions.

Training of constraint model

The constraint model consists of 4 convolutional 2D layers and 6 fully connected layers, conducted by keras[44]. It also uses Adam optimizer, learning rate 0.002, β_1 is 0.5 and β_2 is 0.999, optimization loss is mean square error, activation function is leakyRELU. To prevent overfitting, dropout, and batch normalization are used after each convolutional layer[45]. The detailed design of this model is described in Supplementary Table 5.

Training of CCDCGAN

The generator, discriminator, and constraint have exactly the same design as the previous model, all parameters are same as parameters in DCGAN. The only difference is the optimization objective becomes Eq. (2), where the weight of the formation energy loss is 0.1. The training time is about 18 h under the same condition as DCGAN. Typical generated structures are listed in Supplementary Table 6, more data is available in Supplementary Data 2.

DATA AVAILABILITY

All data needed to produce the work are available from the corresponding author.

CODE AVAILABILITY

CCDCGAN code is available upon reasonable request from the corresponding author.

Received: 18 August 2020; Accepted: 29 March 2021;
Published online: 10 May 2021

REFERENCES

1. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **361**, 360–365 (2018).
2. de Pablo, J. J. et al. New frontiers for the materials genome initiative. *npj Comput. Mater.* **5**, 1–23 (2019).
3. de Pablo, J. J. et al. The materials genome initiative, the interplay of experiment, theory and computation. *Curr. Opin. Solid State Mater. Sci.* **18**, 99–117 (2014).
4. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM* **65**, 1501–1509 (2013).
5. Wang, Y., Lv, J., Zhu, L. & Ma, Y. CALYPSO: a method for crystal structure prediction. *Comput. Phys. Commun.* **183**, 2063–2070 (2012).
6. Glass, C. W., Oganov, A. R. & Hansen, N. USPEX—Evolutionary crystal structure prediction. *Comput. Phys. Commun.* **175**, 713–720 (2006).
7. Noh, J., Ho, Gu,G., Kim, S. & Jung, Y. Machine-enabled inverse design of inorganic solid materials: promises and challenges. *Chem. Sci.* **11**, 4871–4881 (2020).
8. Yamashita, T. et al. Crystal structure prediction accelerated by Bayesian optimization. *Phys. Rev. Mater.* **2**, 013803 (2018).
9. Deringer, V. L., Pickard, C. J. & Csányi, G. Data-driven learning of total and local energies in elemental boron. *Phys. Rev. Lett.* **120**, 156001 (2018).
10. Pickard, C. J. & Needs, R. J. Ab initio random structure searching. *J. Condens. Matter Phys.* **23**, 053201 (2011).
11. Li, Q., Zhou, D., Zheng, W., Ma, Y. & Chen, C. Global structural optimization of tungsten borides. *Phys. Rev. Lett.* **110**, 136403 (2013).
12. Lyakhov, A. O. & Oganov, A. R. Evolutionary search for superhard materials: methodology and applications to forms of carbon and TiO_2. *Phys. Rev. B* **84**, 092103 (2011).
13. Tolle, K. M., Tansley, D. S. W. & Hey, A. J. The fourth paradigm: data-intensive scientific discovery [point of view]. *Proc. IEEE* **99**, 1334–1337 (2011).
14. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
15. Pyzer-Knapp, E. O., Li, K. & Aspuru-Guzik, A. Learning from the Harvard Clean Energy Project: the use of neural networks to accelerate materials discovery. *Adv. Funct. Mater.* **25**, 6495–6502 (2015).
16. Ryan, K., Lengyel, J. & Shatruk, M. Crystal structure prediction via deep learning. *J. Am. Chem. Soc.* **140**, 10158–10168 (2018).
17. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. Preprint at https://arxiv.org/abs/1312.6114 (2013).
18. Goodfellow, I. et al. Generative adversarial networks. Preprint at https://arxiv.org/abs/1406.2661 (2014).
19. Schwalbe-Koda, D. & Gómez-Bombarelli, R. Generative models for automatic chemical design. *Mach. Learn. Meets Quantum. Phys*. 445–467 (2020).
20. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **5**, 1–36 (2019).

21. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
22. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
23. Nouira, A., Sokolovska, N. & Crivello, J.-C. CrystalGAN: Learning to discover crystallographic structures with generative adversarial networks. Preprint at https://arxiv.org/abs/1810.11203 (2018).
24. Hoffmann, J. et al. Data-driven approach to encoding and decoding 3-D crystal structures. Preprint at https://arxiv.org/abs/1909.00949 (2019).
25. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).
26. Kaufmann, K. et al. Paradigm shift in electron-based crystallography via machine learning. Preprint at https://arxiv.org/abs/1902.03682 (2019).
27. Noh, J. et al. Inverse design of solid-state materials via a continuous representation. *Matter* **1**, 1370–1384 (2019).
28. Kim, B., Lee, S. & Kim, J. Inverse design of porous materials using artificial neural networks. *Sci. Adv.* **6**, eaax9324 (2020).
29. Kim, S., Noh, J., Gu, G. H., Aspuru-Guzik, A. & Jung, Y. Generative adversarial networks for crystal structure prediction. *ACS Cent. Sci.* **6**, 1412–1420 (2020).
30. Bojanowski, P., Joulin, A., Lopez-Paz, D. & Szlam, A. Optimizing the latent space of generative networks. Preprint at https://arxiv.org/abs/1707.05776 (2019).
31. Zhang, H. et al. Topological insulators in Bi_2Se_3 Bi_2Te_3 and Sb_2Te_3 with a single Dirac cone on the surface. *Nat. Phys.* **5**, 438–442 (2009).
32. Heim, E. Constrained generative adversarial networks for interactive image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10753–10761 (2019).
33. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
34. Singh, H. K. et al. High-throughput screening of magnetic antiperovskites. *Chem. Mater.* **30**, 6983–6991 (2018).
35. Binary Alloy Phase Diagram—an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/engineering/binary-alloy-phase-diagram.
36. Ren, Z. et al. Inverse design of crystals using generalized invertible crystallographic representation. Preprint at https://arxiv.org/abs/2005.07609 (2020).
37. Court, C. J., Yildirim, B., Jain, A. & Cole, J. M. 3-D inorganic crystal structure generation and property prediction via representation learning. *J. Chem. Inf. Model.* **60**, 4518–4535 (2020).
38. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
39. Deringer, V. L., Caro, M. A. & Csányi, G. Machine learning interatomic potentials as emerging tools for materials science. *Adv. Mater.* **31**, 1902765 (2019).
40. Opahle, I., Madsen, G. K. H. & Drautz, R. High throughput density functional investigations of the stability, electronic structure and thermoelectric properties of binary silicides. *Phys. Chem. Chem. Phys.* **14**, 16197–16202 (2012).
41. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
42. Kramer, M. A. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.* **37**, 233–243 (1991).
43. Ong, S. P. et al. Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
44. Chollet, F. et al. Keras. https://github.com/fchollet/keras (2015).
45. Gron, A. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (O'Reilly Media, 2017).

ACKNOWLEDGEMENTS

The authors gratefully acknowledge computational time on the Lichtenberg High-Performance Supercomputer. Teng Long thanks the financial support from the China Scholarship Council (CSC). Part of this work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant No. 743116-project Cool Innov). This work was also supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 405553726 – TRR 270. We also acknowledge support by the Deutsche Forschungsgemeinschaft (DFG – German Research Foundation) and the Open Access Publishing Fund of Technical University of Darmstadt.

AUTHOR CONTRIBUTIONS

This work originated from the discussion of H.Z., T.L., and Y.Z. H.Z. and O.G. supervised the research. T.L. worked on the machine learning model. T.L., N.F., I.O., and C.S. worked on the DFT calculations. T.L., I.S., and Y.Z. worked on data analysis. All authors contributed in the writing.

FUNDING

Open Access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

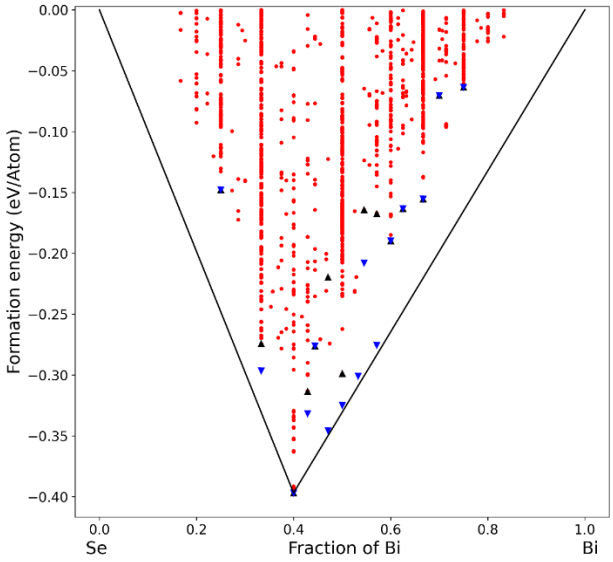
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-021-00526-4.

Correspondence and requests for materials should be addressed to T.L. or H.Z.

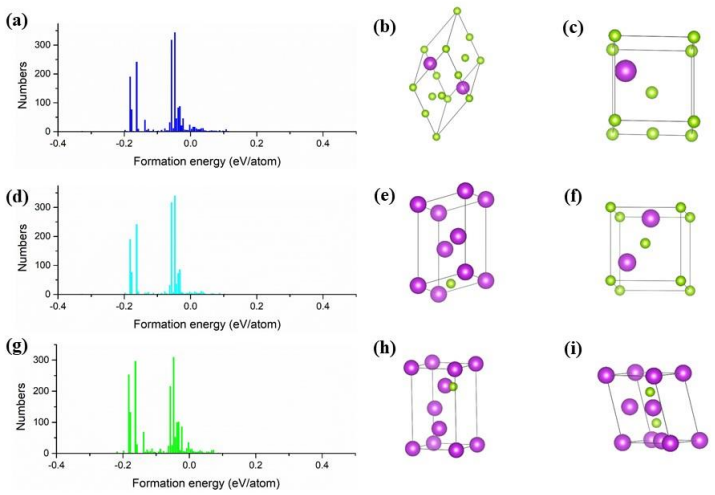
Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

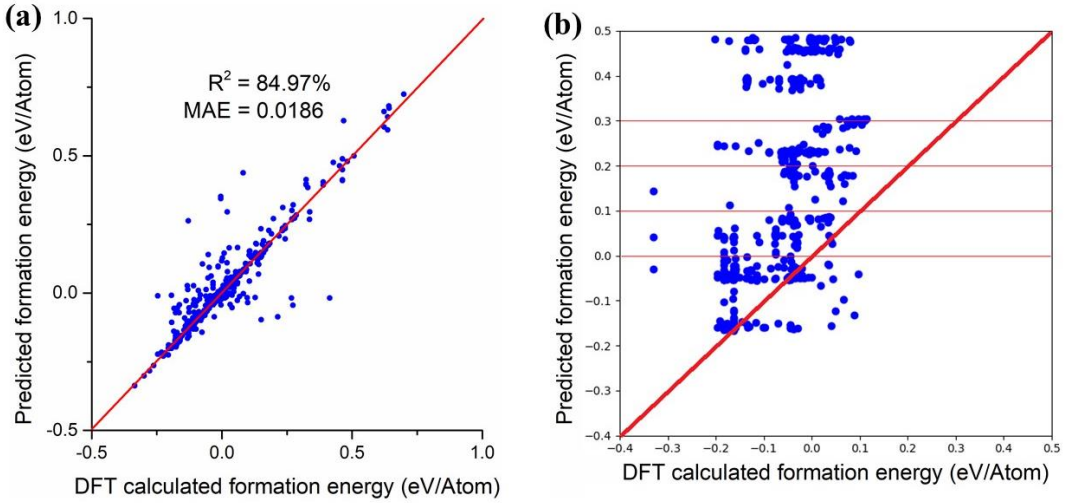
Supplementary



Supplementary Figure 1. Convex hull of Bi-Se system, where the black line represents the convex hull, red points are the structures, blue triangles are the experimentally achievable phases, black triangles are the corresponding structures in the MP database.



Supplementary Figure 2. Distribution and structures of generated data. (a) Distribution of formation energy of DCGAN generated data, where the size of bin is 5 meV; (b) crystal structure of generated Bi_2Se_4; (c) crystal structure of generated Bi_1Se_3 ; (d) Distribution of formation energy of DCGAN + constraint generated data, where the size of bin is 5 meV; (e) crystal structure of generated Bi_3Se_1; (f) crystal structure of generated Bi_2Se_2; (g) Distribution of formation energy of CCDCGAN generated data, where the size of bin is 5 meV; (h) crystal structure of generated Bi_4Se_1; (i) crystal structure of generated Bi_4Se_2.



Supplementary Figure 3. Comparison of formation energy. (a) DFT calculated formation energy vs predicted formation energy for test set of original database; (b) Predicted formation energy of generated structures by DCGAN before structure relaxation vs their corresponding formation energy after relaxation.

Supplementary Table 1. Design of sites autoencoder.

Model type	Layer type	Stride	Activation	Padding	Input size	Output size
Encoder	3D Convolution	2,2,2	LeakyRELU(0.2)	SAME	64,64,64	32,32,32
	3D Convolution	2,2,2	LeakyRELU(0.2)	SAME	32,32,32	16,16,16
	3D Convolution	2,2,2	LeakyRELU(0.2)	SAME	16,16,16	8,8,8
	3D Convolution	2,2,2	LeakyRELU(0.2)	SAME	8,8,8	4,4,4
	3D Convolution	1,1,1	tanh	VALID	4,4,4	200
Decoder	3D Convolution		LeakyRELU(0.2)	VALID	200	4,4,4
	3D Convolution		LeakyRELU(0.2)	SAME	4,4,4	8,8,8
	3D Convolution		LeakyRELU(0.2)	SAME	8,8,8	16,16,16
	3D Convolution		LeakyRELU(0.2)	SAME	16,16,16	32,32,32
	3D Convolution		sigmoid	SAME	32,32,32	64,64,64

Supplementary Table 2. Design of lattice autoencoder.

Model type	Layer type	Stride	Activation	Padding	Input size	Output size
Encoder	3D Convolution	2,2,2	LeakyRELU(0.2)	SAME	32,32,32	16,16,16
	3D Convolution	2,2,2	LeakyRELU(0.2)	SAME	16,16,16	8,8,8
	3D Convolution	2,2,2	LeakyRELU(0.2)	SAME	8,8,8	4,4,4
	3D Convolution	1,1,1	tanh	VALID	4,4,4	200
Decoder	3D Convolution		LeakyRELU(0.2)	VALID	200	4,4,4
	3D Convolution		LeakyRELU(0.2)	SAME	4,4,4	8,8,8
	3D Convolution		LeakyRELU(0.2)	SAME	8,8,8	16,16,16
	3D Convolution		sigmoid	SAME	16,16,16	32,32,32

Supplementary Table 4. Design of discriminator.

Layer type	Stride	Activation	Padding	Input size	Output size
2D Convolution	2,2	LeakyRELU(0.2)	SAME	28,28,1	14,14,32
Dropout(0.25)				14,14,32	14,14,32
2D Convolution	2,2	LeakyRELU(0.2)	PADDING	14,14,32	8,8,64
BatchNormallization(0.8)				8,8,64	8,8,64
Dropout(0.25)				8,8,64	8,8,64
2D Convolution	2,2	LeakyRELU(0.2)	SAME	8,8,64	4,4,128
BatchNormallization(0.8)				4,4,128	4,4,128
Dropout(0.25)				4,4,128	4,4,128
2D Convolution	1,1	LeakyRELU(0.2)	SAME	4,4,128	4,4,256
BatchNormallization(0.8)				4,4,256	4,4,256
Dropout(0.25)				4,4,256	4,4,256
Flatten				4,4,256	4096
Fully connected		sgimoid		4096	1

Supplementary Table 5. Design of constraint model.

Layer type	Stride	Activation	Padding	Input size	Output size
2D Convolution	2,2	LeakyRELU(0.2)	SAME	28,28,1	14,14,32
Dropout(0.25)				14,14,32	14,14,32
2D Convolution	2,2	LeakyRELU(0.2)	PADDING	14,14,32	8,8,64
BatchNormallization(0.8)				8,8,64	8,8,64
Dropout(0.25)				8,8,64	8,8,64
2D Convolution	2,2	LeakyRELU(0.2)	SAME	8,8,64	4,4,128
BatchNormallization(0.8)				4,4,128	4,4,128
Dropout(0.25)				4,4,128	4,4,128
2D Convolution	1,1	LeakyRELU(0.2)	SAME	4,4,128	4,4,256
BatchNormallization(0.8)				4,4,256	4,4,256
Dropout(0.25)				4,4,256	4,4,256
Flatten				4,4,256	4096
Fully connected		LeakyRELU(0.2)		4096	1024
Fully connected		LeakyRELU(0.2)		1024	256
Fully connected		LeakyRELU(0.2)		256	256
Fully connected		LeakyRELU(0.2)		256	256
Fully connected		LeakyRELU(0.2)		256	64
Fully connected		LeakyRELU(0.2)		64	1

Table Supplementary 6. Generated distinct structures.

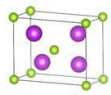
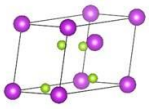
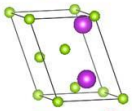
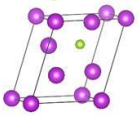
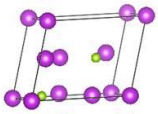
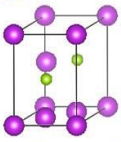
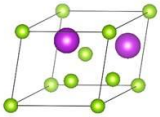
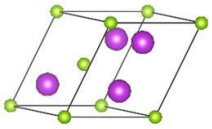
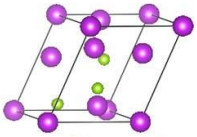
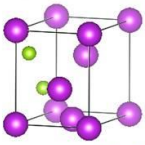
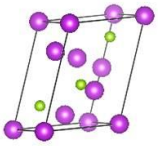
ID	Formula	Formation energy (eV/atom)	Convex hull (eV/atom)	Structure	Space group #
DCGAN Epoch_11731	Bi_4Se_2	-0.1391	0.0798		141
DCGAN Epoch_2368	Bi_2Se_4	-0.2481	0.0803		12
DCGAN Epoch_1651	Bi_2Se_4	-0.2447	0.0837		2
DCGAN Epoch_1343	Bi_5Se_1	-0.021	0.0893		8
DCGAN Epoch_876	Bi_6Se_2	-0.066	0.0981		12
DCGAN Epoch_7598	Bi_4Se_2	-0.1201	0.0988		8
CCDCGAN Epoch_315	Bi_2Se_4	-0.2871	0.0413		2
CCDCGAN Epoch_8688	Bi_4Se_2	-0.1394	0.0795		141
CCDCGAN Epoch_12310	Bi_6Se_3	-0.1314	0.0876		1
CCDCGAN Epoch_4803	Bi_4Se_2	-0.1258	0.0931		10
CCDCGAN Epoch_9982	Bi_6Se_3	-0.1214	0.0975		1

Table Supplementary 7. Compilation of the experimentally achievable (EXP) phases and the corresponding generated phases (CCDCGAN) of the binary Bi-Se system. The fifth column shows the formation energy of the structures with corresponding compositions included in the training. [1]

Phase	Composition % of Bi	Formation energy (EXP) (eV/atom)	Distance to the convex hull (EXP) (eV/atom)	Formation energy (training) (eV/atom)	Formation energy (CCDCGAN) (eV/atom)
$BiSe_3$	0.25	-0.15	0.10	-0.15	-0.15
$BiSe_2$	0.333	-0.30	0.03	-0.27	-0.30
Bi_2Se_3	0.4	-0.39	0	-0.39	-0.39
Bi_3Se_4	0.429	-0.33	0.06	-0.31	-0.33
Bi_4Se_5	0.444	-0.28	0.09	-0.28	-0.28
Bi_8Se_9	0.471	-0.35	0.01	-0.22	-0.22
BiSe	0.5	-0.32	0.01	-0.30	-0.30
Bi_8Se_7	0.533	-0.30	0.01	None	-0.02
Bi_6Se_5	0.545	-0.21	0.13	-0.16	-0.21
Bi_4Se_3	0.571	-0.28	0.01	-0.17	-0.26
Bi_3Se_2	0.6	-0.19	0.07	-0.19	-0.19
Bi_5Se_3	0.625	-0.16	0.08	-0.16	-0.16
Bi_2Se	0.667	-0.16	0.08	-0.16	-0.16
Bi_7Se_3	0.7	-0.07	0.13	-0.07	-0.07
Bi_3Se	0.75	-0.07	0.10	-0.07	-0.07

Supplementary Table 3. Design of generator.

Layer type	Stride	Activation	Padding	Input size	Output size
Fully connected		RELU		200	7,7,128
2D Convolution		RELU	SAME	7,7,128	14,14,128
BatchNormallization(0.8)				14,14,128	14,14,128
2D Convolution		RELU	SAME	14,14,128	28,28,128
2D Convolution		RELU	SAME	28,28,128	28,28,64
BatchNormallization(0.8)				28,28,64	28,28,64
2D Convolution		tanh	SAME	28,28,64	28,28,1

Batch training is conducted for both sites autoencoder and lattice autoencoder, the batch size is 822 and 311 respectively, and the epoch used in training are 100 and 200 respectively. Exact model designs are listed in Supplementary Table 1 and 2. The generated vectors are resized into a 28×28 2d graphs where the rest are padding zeros.

Batch training is conducted for DCGAN model, batch size is 128 and total epoch is 500000. Detailed model design is listed in Supplementary Table 3 and 4. The loss function used in the GAN model is “binary_crossentropy”, generator loss is to minimize the entropy between 1 and generated structures, while the discriminator loss is to minimize the entropy between 1 and real

structures.

Batch training is conducted for constraint model, batch size is 128 and total epoch is 2000. Detailed model design is listed in Supplementary Table 5.

The performance of constraint model is in Supplementary Figure 3 (a). In the DCGAN + constraint model, the selection criterion is 0.3 eV/atom. The selection is calculated for two reasons: first, the MAE of the model is not exactly 0, so cutting at 0 eV/atom is highly likely to drop some reasonable structures; second, DFT is highly likely to reduce the formation energy of generated structures through relaxation as Supplementary Figure 3 (b) demonstrates, so we select this number to balance these two factors. And the effects of constraint in the latent space are demonstrated in movie file Supplementary Movie 1.

All parameters and designs are the same as DCGAN, except the loss of generator, it is the combination of "binary_crossentropy" and the average of generated structures predicted by the constraint.

Reference

1. *ASM handbook*. (ASM International, 1990).

4.3. Thermal conductivity of h-BN monolayers using machine learning interatomic potential

As mentioned above, this thesis focuses on examining the relationship between crystal structures and intrinsic properties by ML techniques, where most training data are evaluated by the third paradigm. However, the calculation of intrinsic properties by DFT is slow even in a high-throughput manner, so machine-learning interatomic potentials are utilized to increase the efficiency of the evaluation of intrinsic properties. For example, the thermal conductivity is evaluated based on the Boltzmann transport theory, with the relaxation time approximation calculated by three-phonon interactions [199]. Force constants are calculated by configurations with different displacement patterns to evaluate the phonon–phonon interactions. The atomic forces for each atom and the total energies of these displaced configurations are calculated by DFT. However, the calculation by DFT is slow even in a high-throughput manner, so I adopted the machine-learning interatomic potential to increase the efficiency of the evaluation of atomic forces. So in this work, we use the GAP model and SOAP descriptors developed by Bartok et al. to describe the interatomic potential for h-BN [161, 97]. This enables the fast evaluation of total energy for h-BN with MAE less than 10^{-5} eV/atom. With the accurate and fast evaluation of the atomic forces, the thermal conductivity of h-BN can be calculated. The result only 1.12% deviates from the actual value, which demonstrates the capability of the GAP.

IOP Publishing Journal of Physics: Condensed Matter
J. Phys.: Condens. Matter **33** (2021) 105903 (6pp) https://doi.org/10.1088/1361-648X/abcf61

Thermal conductivity of h-BN monolayers using machine learning interatomic potential

Yixuan Zhang, Chen Shen*, Teng Long and Hongbin Zhang*

Institute of Materials Science, Technical University of Darmstadt, Darmstadt 64287, Germany

E-mail: yixuan.zhang@tmm.tu-darmstadt.de, chenshen@tmm.tu-darmstadt.de and hongbin.zhang@tu-darmstadt.de

Received 30 September 2020, revised 23 November 2020
Accepted for publication 1 December 2020
Published 23 December 2020

Abstract

Thermal management materials are of critical importance for engineering miniaturized electronic devices, where theoretical design of such materials demands the evaluation of thermal conductivities which are numerically expensive. In this work, we applied the recently developed machine learning interatomic potential (MLIP) to evaluate the thermal conductivity of hexagonal boron nitride monolayers. The MLIP is obtained using the Gaussian approximation potential method, and the resulting lattice dynamical properties and thermal conductivity are compared with those obtained from explicit frozen phonon calculations. It is observed that accurate thermal conductivity can be obtained based on MLIP constructed with about 30% representative configurations, and the high-order force constants provide a more reliable benchmark on the quality of MLIP than the harmonic approximation.

Keywords: machine learning interatomic potential, lattice thermal conductivity, anharmonic lattice dynamics

Supplementary material for this article is available online

(Some figures may appear in colour only in the online journal)

1. Introduction

Hexagonal boron nitride (h-BN), as one of the most interesting III–V compounds with layered structures, has drawn intensive attentions in the last few decades [1–6], as h-BN exhibits many intriguing properties such as excellent mechanical properties [7], high in-plane thermal conductivity [4], and behaves like ionic half-metal [8, 9] due to the different electronegativity of the boron and nitrogen atoms. Such properties make h-BN a promising thermal management material which is essential for engineering miniaturized electronic devices. On the other hand, the thermal conductivities play also a crucial role in optimizing the thermoelectric performance [10]. Previous theoretical studies predicted an in-plane thermal conductivity as high as 550 W m^{-1} K^{-1} [6] at room temperature, consistent with the reported experimental values ranging from 220 to 420 W m^{-1} K^{-1} [2, 6, 11]. The evaluation of thermal conductivities requires accurate calculations on phonon-phonon interaction, where at least third-order force constants have to be calculated. For instance, such calculations of thermal conductivities are usually done using the frozen phonon method based on density functional theory (DFT) [12, 13], which are numerically demanding.

Recently, machine learning interatomic potential (MLIP) has become a valuable solution to construct atomistic models with dominant short-range interactions. For instance, Gaussian approximation potential (GAP) introduced by Bartók *et al* [14, 15] has been applied to describe the potential energy surface with sufficient accuracy, where the interatomic potential (energy, forces, and strain) are fitted using the Gaussian kernel method [16, 17] on a multidimensional configurational space of atomic positions. Szlachta *et al* established a

* Authors to whom any correspondence should be addressed.

comprehensive GAP model based on a database of bcc tungsten and applied this model to calculate the properties such as screw dislocation, the Peierls barrier and vacancy-dislocation interaction energies [15]. John and Csányi introduced a framework describing general many-body coarse-grained interactions of molecules. By combining molecular dynamics (MD) and GAP potential, a decent model which can precisely describing the free energy surface of molecular liquids were established [18]. Qian and Yang developed two GAP models for both hcp and bcc phase of zirconium crystals, and used them in MD calculations to predict the phonon dispersion at finite temperature [19]. Recently, Deringer *et al* gave a brief introduction of MLIP and highlighted some representative applications of that in material science, e.g., the prediction of crack behavior of Si, the simulation of phase-change materials, the calculation of nanoparticles for catalysis and the investigation of carbon-based nanomaterials [20].

In this work, we demonstrated a MLIP obtained using GAP which can be applied to accurately evaluate the thermal conductivity of h-BN monolayer. Detailed comparisons on the phonon dispersion, density of states (DOSs), phonon lifetime, and thermal conductivities are done between the explicit frozen phonon and MLIP calculations. Particularly, we studied how the numerical efforts can be reduced while keeping the accuracy via testing the minimum size of dataset for the GAP model. This establishes a workflow combining DFT and ML to evaluate the thermal conductivities efficiently, which paves the way for future design of materials for thermal management and thermoelectric applications.

2. Numerical methods

To evaluate the phonon and phonon–phonon interactions, second- and third-order force constants are calculated using the Alamode package [21]. The configurations with different displacement patterns are generated using the frozen phonon method. The atomic forces for each atoms and total energies of these displaced configurations are calculated by DFT, performed using the projector augmented-wave method as implemented in Vienna *ab initio* simulation package [22, 23]. The energy cutoff is 550 eV and the energy convergence tolerance is 10^{-8} eV. The exchange correlation functional is chosen to be the generalized gradient approximation parametrized by Perdew–Burke–Ernzerhof [24]. A supercell of $6 \times 6 \times 1$ is constructed, with the interlayer distance being 20 Å. The Brillouin zone is sampled using the gamma centered Mokhorst–Pack scheme with a $4 \times 4 \times 1$ k-mesh with Bloch corrections [25].

The thermal conductivity is evaluated based on the Boltzmann transport theory with the relaxation time approximation [26], which yields

$$\kappa_{\mathrm{ph}}^{\mu\upsilon}(T) = \frac{1}{\Omega N_q}\sum_{q,j} C_{q,j}(T) v_{qj}^{\mu} v_{qj}^{\upsilon} \tau_{qj}(T), \qquad (1)$$

where Ω is the volume of the unit cell and N_q is the number of q-point, $C_{q,j}(T) = \hbar\omega_{qj}\partial n_{qj}/\partial T$ is the constant-volume specific heat with $n_{qj} = 1/(\mathrm{e}^{\hbar\omega_{qj}/kT} - 1)$ being the Bose–Einstein distribution function, $v_{qj}^{\mu/\upsilon}$ denotes the group velocities along the Cartesian μ/υ-direction, $\tau_{qj}(T)$ is the phonon lifetime which can be estimated according to the Matthiessen's rule as $\tau_{qj}(T) = \frac{1}{2}(\Gamma_{qj}(T))^{-1}$ with $\Gamma_{qj}(T)$ is the imaginary part of the phonon self-energy considering the lowest-order three-phonon interaction.

Table 1. The number of the training sets of manual and different incremental method and their relative proportion out of the database.

	Training set	Proportion (%)
Manual	325	37.5
Incremental-2	82	9.5
Incremental-3	122	14.1
Incremental-4	162	18.7
Incremental-5	202	23.3
Incremental-6	242	27.9

The use of $6 \times 6 \times 1$ supercell results in 867 third order frozen phonon configurations, where the atomic displacement amplitude is set to 0.04 Å. The total energies and forces of such supercells are calculated and fed into the GAP [14] to construct MLIP. It is noted that due to the 2D nature of BN monolayers, the lattice dynamics along the z-direction is different from the in-plane directions, thus we performed a classification of frozen-phonon configurations based on the direction of atomic displacement. That is, 867 configurations are separated into two groups: one group (dubbed database1 with 554 configurations) with only in-plane displacements and the other group (dubbed database2 with 313 configurations) with z-direction displacement. To ensure the diversity and continuity of the training sets, and to develop a thorough understanding on how to pick up representative configurations as well, we propose a general but robust scheme (denoted as 'manual') for selecting training set. We consider the configurations with displacements on the same atom but of opposite directions as similar configurations, such configurations are grouped as subsets for database1 and database2. The training sets are constructed by choosing one configuration from each subsets, leading to 160 out of 554 configurations from database1 and 165 out of 313 configurations from database2.

Moreover, to estimate the minimum number requirement of configurations for accurately reproducing the thermal conductivity, and also the feasibility of employing active learning to make this workflow automatically, another selecting method (denoted as 'incremental') was tested, where the most representative configurations are selected out as the training set with the help of MLIP. Specifically, starting from a random configuration for each database, at each iteration, 25 (15) configurations for database1 (database2) were screened out and added into the training set for studying the model for the next iteration. Then the results of different iteration steps (2–6) were compared. The selected training set with the corresponding percentage are shown in table 1 for different selecting methods. Note that the crystalline symmetry is not considered here, in order to test different ways to obtain the training sets.

The construction of GAP is performed using QUIP package [27]. In GAP, the total energy of MLIP consists of the local

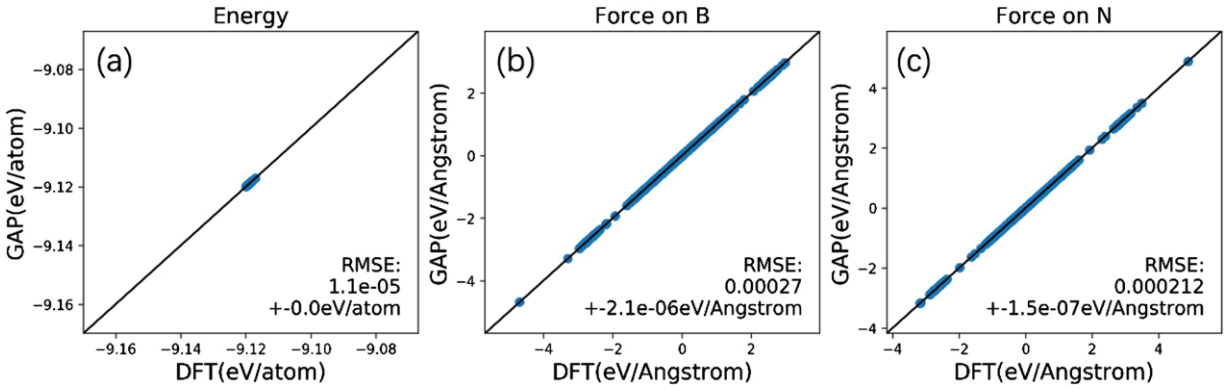


Figure 1. The comparison of energies and interatomic forces between DFT and MLIP trained using the Manual dataset. (a) The energy; (b) the forces on B atoms; (c) the forces on N atoms.

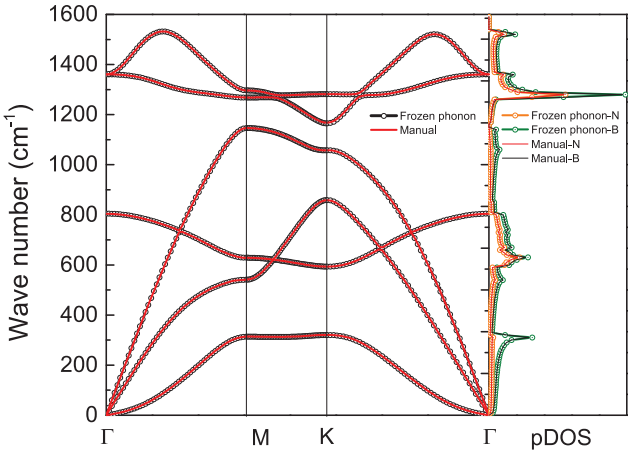


Figure 2. The comparison for phonon dispersion (left panel) and partial pDOSs (right panel) for frozen-phonon and MLIP based on the manual training set.

energies of each individual atoms, and the covariance of two local energies is built in the forms of kernel function:

$$E = \sum_i \varepsilon\left(\boldsymbol{q}_i\right) \tag{2}$$

$$\langle \varepsilon(\boldsymbol{q}_i)\varepsilon(\boldsymbol{q}_j)\rangle = \alpha K(\boldsymbol{q}_i, \boldsymbol{q}_j), \tag{3}$$

where E is the total energy of the system, $\varepsilon(\boldsymbol{q}_i)$ and $\varepsilon(\boldsymbol{q}_j)$ are the energy contribution from atom i and j, $\boldsymbol{q}$ is a descriptor whose function is to characterize the atomic environment, e.g., positions of atoms, in the neighborhood of the target atom. $\boldsymbol{q}_i$ here can be considered as the descriptor of atom i of the configurations to be predicted and $\boldsymbol{q}_j$ of atom j of the configurations in the training set. $K(q_i, q_j)$ is the kernel function which describes the similarity of atomic environment between the atom i and atom j. And α can be understood as the weight of this kernel, which can be determined by the fit. Then the interatomic forces can be calculated as the partial derivative of total energy versus displacement:

$$f_k = -\frac{\partial E}{\partial r_k}, \tag{4}$$

where r_k is the x, y or z component of the Cartesian coordinates of atom k. The weights are fitted by minimizing the lost function in kernel ridge regression, where the unknown function is expanded as a linear combination of kernel functions:

$$F(\boldsymbol{q}) = \sum_i \alpha_i K(\boldsymbol{q}, \boldsymbol{q}_i) \tag{5}$$

$$L = \sum_i (t_i - F(\boldsymbol{q_i}))^2 + \lambda \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha}, \tag{6}$$

where t_i and $F(\boldsymbol{q_i})$ denote the observed and predicted total energy of system or interatomic force on atom, λ is a hyper-parameter and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_j, \ldots)$ is the vector of the weights to be fitted. $\boldsymbol{K}$ is a covariance matrix. The form of this matrix for energies and forces are shown respectively in equations (7) and (8):

$$\boldsymbol{K}_{\text{energy}} = \langle E_N E_M \rangle = \alpha \sum_{i\in N} \sum_{j\in M} K(\boldsymbol{q_i}, \boldsymbol{q}_j) \tag{7}$$

$$\begin{aligned} \boldsymbol{K}_{\text{force}} &= \left\langle \frac{\partial E_N}{\partial r_k} E_M \right\rangle = \frac{\partial \langle E_N E_M \rangle}{\partial r_k} \\ &= \alpha \sum_{i\in N} \sum_{j\in M} \nabla_{\boldsymbol{q}_i} K(\boldsymbol{q_i}, \boldsymbol{q}_j) \frac{\partial \boldsymbol{q}_i}{\partial r_k}. \end{aligned} \tag{8}$$

In GAP, the spatial cutoff function is employed to ensure the computational efficiency. The implemented cutoff function here is:

$$f(x) = \begin{cases} 1 & \text{for } r \leqslant r_{\text{cut}} - d \\ \left[\cos\left(\pi \dfrac{r - r_{\text{cut}} + d}{d}\right) + 1\right]/2 & \text{for } r_{\text{cut}} - d < r \leqslant r_{\text{cut}} \\ 0 & \text{for } r > r_{\text{cut}} \end{cases} \tag{9}$$

with d determines the width of the cutoff region.

The descriptors used in this manuscript are pair and triplets descriptors. For pair descriptor, it is simply defined as the distances between the pair $|\boldsymbol{r}_i - \boldsymbol{r}_j|$ with using cutoff function of $f_{\text{cut}}(\boldsymbol{r}_{ij})$. For triplets, the descriptor is in form of a vector $[\boldsymbol{r}_{ik} + \boldsymbol{r}_{ij}, (\boldsymbol{r}_{ik} - \boldsymbol{r}_{ij})^2, \boldsymbol{r}_{jk}]$ with center atom i, the used cutoff function is $f_{\text{cut}}(\boldsymbol{r}_{ij}) f_{\text{cut}}(\boldsymbol{r}_{ik})$ [27].

The parameters used for the pair descriptor distance_2b are cutoff = 8.0, theta_uniform = 1.0, delta = 0.5, and for the triplets descriptor angle_3b are cutoff = 6.0, theta_fac = 0.5, delta = 0.5. The nature of these two-body and three-body

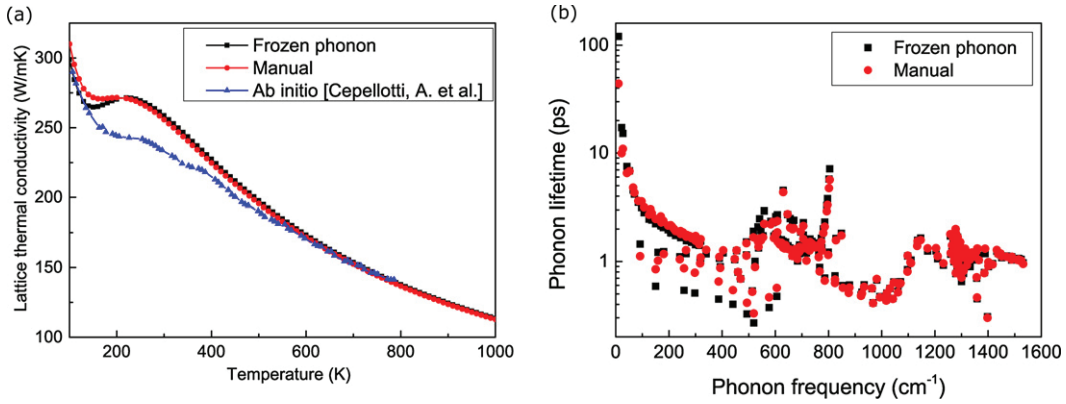


Figure 3. The comparison of (a) lattice thermal conductivity and (b) phonon lifetime of h-BN monolayers for frozen phonon and MLIP based on the manual training set. Blue triangles in panel (a) denote the results of Cepllotti *et al* [28].

Table 2. RMSE comparison of energies and interatomic forces between DFT and MLIP data trained with (a) incremental-2, (b) incremental-4, and (c) incremental-6 training-sets (cf table 1).

	Energy (eV/atom)	Force on B (eV Å^{-1})	Force on N (eV Å^{-1})
Incremental-2	5×10^{-6}	$2.82 \times 10^{-4} \pm 3.1 \times 10^{-7}$	$2.32 \times 10^{-4} \pm 2.1 \times 10^{-7}$
Incremental-4	6×10^{-6}	$2.37 \times 10^{-4} \pm 1.5 \times 10^{-7}$	$2.01 \times 10^{-4} \pm 1.0 \times 10^{-7}$
Incremental-6	1×10^{-5}	$2.32 \times 10^{-4} \pm 1.4 \times 10^{-7}$	$1.96 \times 10^{-4} \pm 9.0 \times 10^{-8}$

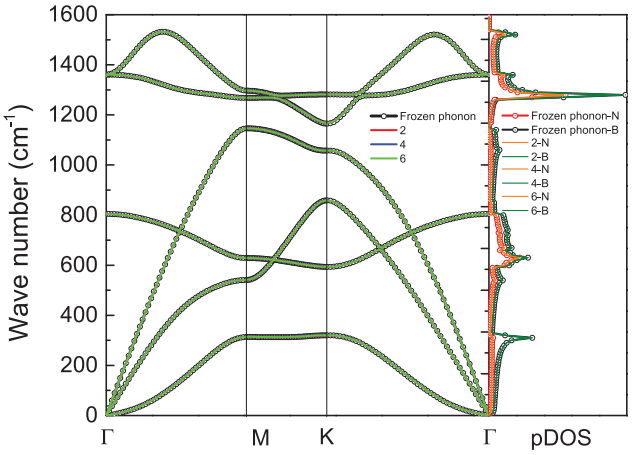


Figure 4. The comparison for phonon dispersion and partial pDOSs obtained using frozen phonon and MLIP based on incremental-2, 4 and 6 training sets.

descriptors makes them perfectly appropriate for the second and third order phonon–phonon interactions.

3. Results and discussions

To verify the construction of MLIP, figure 1 shows the comparison of total energies and interatomic forces between DFT and GAP results for the test sets, where the training is done using the 325 manual configurations. The root-mean-squared errors (RMSE) of dataset for energy is 1.1×10^{-5} eV/atom compared with the mean value of energy (−9.12 eV). Whereas the RMSE for forces on the boron (nitrogen) atoms are $0.000\,27 \pm 2.1 \times 10^{-6}$ eV Å^{-1} ($0.000\,212 \pm 1.55 \times 10^{-7}$ eV Å^{-1}), respectively, with the interatomic forces ranging from −4 eV Å^{-1} to 4 eV Å^{-1}. The high accuracy and excellent consistency of the so obtained MLIP guarantee the accuracy of the predicted properties, as discussed later.

Figure 2 shows the phonon dispersion and phonon density of states (pDOS) obtained based on the explicit frozen-phonon calculations and MLIP (manual). It is observed that both the phonon spectra and DOS obtained using MLIP are in excellent consistency with the frozen phonon results. That is, MLIP can be applied to get reliable harmonic phonon spectra. Such excellent agreement indicates that both the phonon velocities and distribution function (equation (1)) can be well reproduced using MLIP.

Now turn to the thermal conductivity which is originated from the phonon–phonon scattering. According to equation (1), the phonon lifetime, which is an essential quantity for the thermal conductivity, can be obtained from the phonon linewidths at irreducible *k*-points. In order to get the phonon linewidths, the anharmonic terms beyond the quadratic order should be considered, and in this work we considered only the third-order terms. Figure 3(a) shows the thermal conductivities in comparison with previous DFT calculations by Cepelloti *et al* [28], and figure 3(b) displays the phonon lifetime obtained using frozen phonon and MLIP.

Obviously, the phonon lifetime obtained using frozen phonon and MLIP is very comparable, especially at high phonon frequency range. This indicates that the MLIP based on the GAP model can account well for the anharmonicity effects. However, at the middle and low frequency range, the phonon lifetime is slightly overestimated in the MLIP results. This is due to the long-range phonon–phonon interactions with

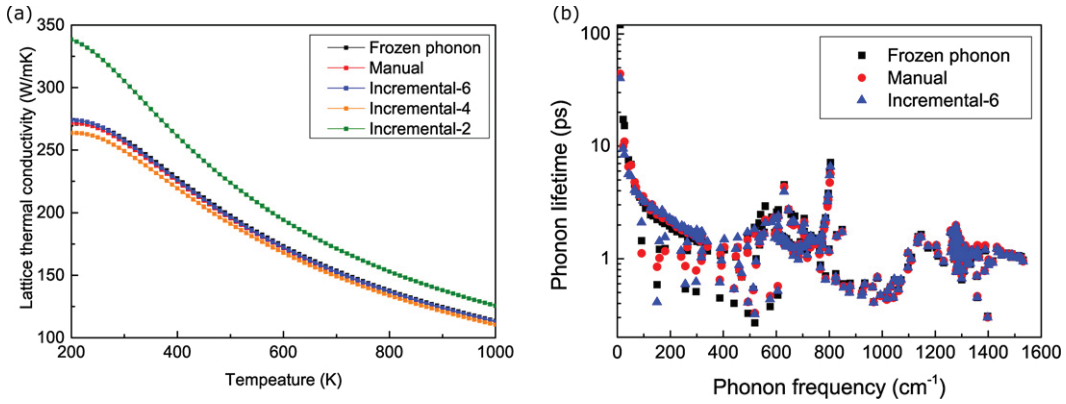


Figure 5. The comparison of: (a) lattice thermal conductivity of frozen phonon and MLIP based on the manual and incremental-2/4/6 training-sets; (b) phonon life time of frozen phonon and MLIP based on manual and incremental-6 training-sets.

low vibrational frequencies, which are not well considered in the training set. Nevertheless, it is fair to say that the MLIP results based on manual training set can be quantitatively compared with the original frozen phonon results.

Importantly, the thermal conductivities obtained based on frozen phonon and MLIP are in good agreement, as shown in figure 3(a). For instance, the deviation of the thermal conductivity obtained using MLIP is as small as -0.73 W m^{-1} K^{-1} at 1000 K, with a relative error only about 0.71%. Such a good accuracy persists to the medium temperature range between 200 K and 500 K, e.g., the deviation at 400 K is 2.50 W m^{-1} K^{-1}, with the relative error of 1.12%. However, significant deviation is observed below 200 K. This can be attributed to the missing long-range phonon–phonon interactions in the MLIP model at the low frequency range as observed in the phonon lifetime (figure 3(b)), which plays an important role for thermal conductivity at low temperatures. Last but not least, our frozen phonon results are in good agreement in the medium to high temperature range with those obtained by Cepelloti *et al* [28], but not for the low temperature below 500 K. The deviation may be originated from the differences in isotopic concentrations and cutoff distance when evaluating the force constants. It is noted that in the low temperature range, the quantum effect should be considered, which is beyond the scope of this work.

Furthermore, in order to verify whether such representative configurations can be selected automatically during the training processes and find out the minimal set, we tested the Incremental method. Five different MLIP models trained with increasing number of configurations were obtained, corresponding to 9.5% (incremental-2), 14.1% (incremental-3), 18.7% (incremental-4), 23.3% (incremental-5) and 27.9% (incremental-6) of the total configurations, respectively.

Table 2 shows the RMSE comparison of total energies and atomic forces where incremental-2, 4, and 6 training-sets are used to construct MLIP. All the models shows good agreement with the DFT results. Note that the incremental-2 dataset corresponds to only 10% of the data, suggesting already a good selection of the training set. As the iteration times increase, i.e., more configurations are selected as the training set, the RMSE of the resulting interatomic forces decreases. Nevertheless, for the energy prediction, both incremental-2 and 4 has RMSE energies smaller than that from the manual set, whereas that of incremental-6 abnormally increases. It is suspected that this is caused by the fact that specific critical structures are not involved in the training set, so these models are not explicitly enough in predicting the forces. On the other hands, these models also avoid giving huge offsets when facing critical data, so that the RMSE of energy predictions are smaller.

Although the RMSEs of the incremental models are slightly deviated from each other, all the models give rise to accurate phonon spectra as shown in figure 4. In this regard, a small number of well-selected configurations are enough to reproduce the harmonic properties.

Figure 5(a) shows the resulting thermal conductivities obtained using different methods as specific above, the plot of all methods is presented in figure S3 (https://stacks.iop.org/JPCM/33/105903/mmedia). Apparently, the more configurations used in the training set, the more accurate the resulting thermal conductivities, in comparison with the frozen phonon results. Although all incremental models give rise to excellent harmonicitic lattice dynamical properties (figure 4), the thermal conductivities based on most incremental models are overestimated, except for the incremental-4 case with 18.7% of the configurations where the thermal conductivity is slightly underestimated (table 2 incremental-4). This suggests that the anharmonic terms, i.e., the phonon–phonon interactions, requires higher accuracy of MLIP. Furthermore, a clear trend of convergence is observed (figure 5(a)), and the result from the incremental-6 case (with 27.9% configurations) are in good agreement with the frozen phonon result (figure 5(a)), e.g., the deviation of the incremental-6 result is about -0.648 W m^{-1} K^{-1} at 1000 K, with a relative error about 0.63%. This is even slightly better than that based on the MLIP obtained using the manual training set. This can be understood based on the phonon lifetime shown in figure 5(b), where the MLIP obtained with incremental-6 training-set can reproduce well the major features of the phonon lifetime. In this regard, the computational time for lattice thermal conductivities can be

significantly reduced by selecting the most representative configurations as the training set. It is admitted that currently we have evaluated the forces for all the configurations based on DFT calculations, which are used to benchmark the MLIP trained using part of the configurations. In order to be able to implement such a scheme without doing calculations on the whole set, we suspect that active learning with uncertainty quantification can be applied so that only the most representative configurations will be selected and calculated explicitly based on DFT [29].

4. Summary

In conclusion, GAP-type MLIPs constructed using different training sets are applied to evaluate the lattice dynamics and thermal conductivities of h-BN monolayers. It is illustrated that the phonon spectra can be accurately obtained using a small set (about 30%) of configurations, leading to an efficient way of performing calculations on the lattice thermal conductivities. It is further observed that for three-dimensional materials like beta-Sn (not shown), even less than 30% configurations can be selected automatically and used to construct a MLIP to evaluate the thermal conductivities accurately. Moreover, the thermal conductivity calculations require more accurate construction of MLIPs than the harmonic lattice properties, which can be applied to benchmark the quality of MLIPs.

It is suspected that active learning with proper uncertainty quantification can be applied to select the most representative configurations as the training set, so that the numerical efforts to evaluate the thermal conductivity can be significantly reduced, which will be saved for future studies. And it is also suspected that such a workflow is capable in constructing MLIPs for MD calculations as well, while proper configurations related to desired properties should be generated and considered to construct the more suitable MLIPs.

Acknowledgments

Yixuan Zhang thanks the financial support from Fullbright-Cottrell Foundation. Hongbin Zhang acknowledges the financial support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project-ID 405553726–TRR 270. Teng Long thanks the financial support from China Scholarship Council. Calculations of this research were conducted on the Lichtenberg high performance computer of TU Darmstadt.

ORCID iDs

Yixuan Zhang https://orcid.org/0000-0001-6165-0646
Hongbin Zhang https://orcid.org/0000-0002-1035-8861

References

[1] Geick R, Perry C H and Rupprecht G 1966 *Phys. Rev.* **146** 543
[2] Sichel E K, Miller R E, Abrahams M S and Buiocchi C J 1976 *Phys. Rev.* B **13** 4607
[3] Kubota Y, Watanabe K, Tsuda O and Taniguchi T 2007 *Science* **317** 932
[4] Lin Z, Mcnamara A, Liu Y, Moon K-S and Wong C-P 2014 *Compos. Sci. Technol.* **90** 123
[5] Tran T T, Bray K, Ford M J, Toth M and Aharonovich I 2016 *Nat. Nanotechnol.* **11** 37
[6] Jiang P, Qian X, Yang R and Lindsay L 2018 *Phys. Rev. Mater.* **2** 064005
[7] Kiran M S R N, Raidongia K, Ramamurty U and Rao C N R 2011 *Scr. Mater.* **64** 592
[8] Kim K and Kim J 2014 *Ceram. Int.* **40** 5181
[9] Zheng F, Zhou G, Liu Z, Wu J, Duan W, Gu B-L and Zhang S B 2008 *Phys. Rev.* B **78** 205415
[10] Chiu W-T, Chen C-L and Chen Y-Y 2016 *Sci. Rep.* **6** 23143
[11] Simpson A and Stuckes A D 1971 *J. Phys. C: Solid State Phys.* **4** 1710
[12] Fu C L, Freeman A J, Wimmer E and Weinert M 1985 *Phys. Rev. Lett.* **54** 2261
[13] Frank W, Elsässer C and Fähnle M 1995 *Phys. Rev. Lett.* **74** 1791
[14] Bartok A P, Payne M C, Kondor R and Csanyi G 2010 *Phys. Rev. Lett.* **104** 136403
[15] Szlachta W J, Bartok A P and Csanyi G 2014 *Phys. Rev.* B **90** 104108
[16] Rasmussen C E 2004 *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2–14, 2003, Tübingen, Germany, August 4–16, 2003, Revised Lectures* (Berlin: Springer) p 63
[17] Seeger M 2004 *Int. J. Neural Syst.* **14** 69
[18] John S T and Csányi G 2017 *J. Phys. Chem.* B **121** 10934
[19] Qian X and Yang R 2018 *Phys. Rev.* B **98** 224108
[20] Deringer V L, Caro M A and Csányi G 2019 *Adv. Mater.* **31** 1902765
[21] Tadano T, Gohda Y and Tsuneyuki S 2014 *J. Phys.: Condens. Matter.* **26** 225402
[22] Kresse G and Furthmüller J 1996 *Phys. Rev.* B **54** 11169
[23] Kresse G and Furthmüller J 1996 *Comput. Mater. Sci.* **6** 15
[24] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865
[25] Blöchl P E, Jepsen O and Andersen O K 1994 *Phys. Rev.* B **49** 16223
[26] McGaughey A J H and Kaviany M 2004 *Phys. Rev.* B **69** 094303
[27] Bartók A P and Csányi G 2015 *Int. J. Quantum Chem.* **115** 1051
[28] Cepellotti A, Fugallo G, Paulatto L, Lazzeri M, Mauri F and Marzari N 2015 *Nat. Commun.* **6** 6400
[29] Frederiksen S L, Jacobsen K W, Brown K S and Sethna J P 2004 *Phys. Rev. Lett.* **93** 165501

Supplementary

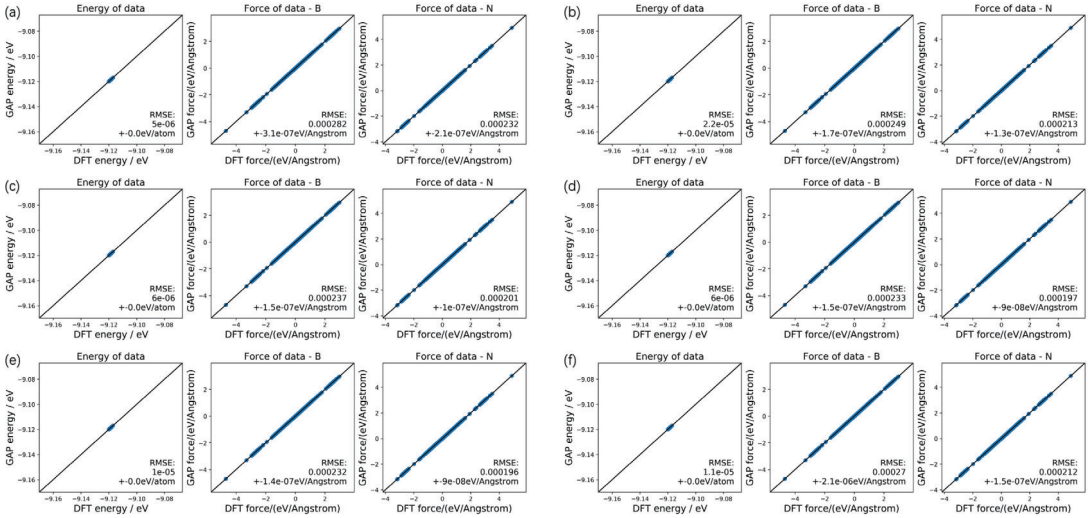


Figure S1: The comparison of energies and interatomic forces between DFT and MLIP trained using different datasets: (a) Incremental-2; (b) Incremental-3; (c) Incremental-4; (d) Incremental-5; (e) Incremental-6; (f) Manual.

5. Conclusion and outlook

To summarize, we have successfully applied machine learning techniques to perform forward modelling and inverse design of crystalline materials, focusing particularly on mapping out the relationship between crystal structures and the intrinsic properties, such as Curie temperature for ferromagnets, formation energy, and atomic forces.

Firstly, forward modelling can provide a fast and accurate evaluation of the intrinsic properties directly. Given that it is challenging to evaluate $\mathrm{T_C}$ using DFT-based methods, an RF model is established to make the prediction. RF models can distinguish the ferro/antiferro-magnetic materials and predict $\mathrm{T_C}$ for ferromagnetic materials. The resulting accuracy is 87% for the classification and 91% for the regression. 85 experimental values have been further validated with a resulting accuracy of 91%. Thus, it provides a practical solution to accelerate the evaluation of $\mathrm{T_C}$ for ferromagnetic materials, which paves the way to accelerate the discovery of ferromagnetic materials.

Secondly, the forward modelling can predict the total energies and atomic forces with respect to the local crystalline environments, *i.e.*, a fast and accurate prediction of interaction between atoms, and then machine-learning interatomic potentials. For instance, the GAP method developed by Bartok et al. was used, allowing the accurate evaluation of total energies with an MAE less than 10^{-5} eV/atom. Additionally, the atomic forces with respect to local displacement can be well interpolated using SOAP as descriptors, and the resulting thermal conductivity of h-BN predicted by the GAP at room temperature is only 1.12% deviated from the value obtained by explicit DFT calculations.

To further explore the capability of ML techniques, I implemented and demonstrated a CCDCGAN model, which can predict stable crystal structures distinct from the known cases based on constructing the latent space. The most salient feature of CCDCGAN is that the intrinsic properties can be optimized in the latent space, which offers the opportunity to achieve multi-objective optimization. Currently, formation energy has been implemented into the generative model as a constraint, leading to the prediction of stable crystal structures. Such a model has been successfully applied on the binary Bi-Se system and multicomponent systems, such as Li-Cd, Co-Hf, Al-Cr-Mn, Co-Mn-Si, etc. It is observed that many unreported crystal structures below the convex hull defined by the known experimental cases can be obtained, leading to a systematic way of designing novel

materials, in particular intermetallic compounds.

In the future, it is believed that ML can be applied to further accelerate the discovery of materials with optimal properties. For instance, the inverse design based on CCDCGAN can be easily generalized to optimize the other physical properties such as hardness and magnetization. This will unlock multi-objective optimization to achieve the goal of designing functional materials with optimal performance.

Currently, I focus on exploring the relationship between crystal structures and intrinsic properties, where sufficient training data can be obtained via high-throughput DFT calculations. A relatively common deduction is to extend the current framework to establish a relationship between microstructures and extrinsic properties. For example, the hardness of a material depends not only on its crystal structure but also on its microstructure. Thus, it would be of great significance to take both crystal structures and microstructures into consideration when predicting the hardness of materials.

Besides, quantum machine learning is an emerging topic in this area as well. On the one hand, it refers to take advantage of the development of quantum computers and then uses quantum computers to implement ML learning algorithm [200]. This enables the processing of immense quantities of data to improve computational efficiency. Although the challenges of hardware and software are still considerable, these have been applied to experimental design in chemistry [201]. On the other hand, quantum machine learning is associated with classical machine learning algorithms applied to quantum problems, such as instructing the experiments to create complex quantum states [202]. Thus, it is interesting to adapt the quantum machine learning techniques in materials design.

In conclusion, with the improvement of computational facilities and the development of new ML techniques. It is hoped that by combining with the other research paradigms, the complete PSPP mapping can be efficiently conducted so that sustainable materials can be developed. On the other hand, as an interdisciplinary field, studies in materials science will enrich the ML techniques from the mathematical complexity, mechanistic understanding, and algorithm development points of view. Thus, ML techniques can revolutionize scientific research in the future.

Reference

[1] Yuval Noah Harari. *Sapiens: A brief history of humankind*. Random House, 2014.

[2] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. “Inverse molecular design using machine learning: Generative models for matter engineering”. In: *Science* 361.6400 (2018), pp. 360–365.

[3] Christine Burton and Carol Scott. “Challenges for the 21st century”. In: *Museum management and marketing* (2007), pp. 49–65.

[4] Xueqiang Zhang, Xinbing Cheng, and Qiang Zhang. “Nanostructured energy materials for electrochemical energy conversion and storage: a review”. In: *Journal of energy chemistry* 25.6 (2016), pp. 967–984.

[5] Charbel Tannous and R Lawrence Comstock. “Magnetic information-storage materials”. In: *Springer Handbook of Electronic and Photonic Materials*. Springer, 2017, pp. 1–1.

[6] John C Mauro et al. “Accelerating the design of functional glasses through modeling”. In: *Chemistry of Materials* 28.12 (2016), pp. 4267–4277.

[7] S Adarsh et al. “Performance comparison of Infrared and Ultrasonic sensors for obstacles of different materials in vehicle/robot navigation applications”. In: *IOP Conference Series: Materials Science and Engineering*. Vol. 149. 1. IOP publishing. 2016, p. 012141.

[8] M Bulut Coskun et al. “Detecting subtle vibrations using graphene-based cellular elastomers”. In: *ACS applied materials & interfaces* 9.13 (2017), pp. 11345–11349.

[9] Jun Gao et al. “Bipolar electrode array embedded in a polymer light-emitting electrochemical cell”. In: *ACS applied materials & interfaces* 9.37 (2017), pp. 32405–32410.

[10] John R Weeks. *Population: An introduction to concepts and issues*. Cengage Learning, 2020.

[11] Paul James. *Urban sustainability in theory and practice: circles of sustainability*. Routledge, 2014.

[12] Steve Harris, Michael Martin, and Derek Diener. “Circularity for circularity’s sake? Scoping review of assessment methods for environmental performance in the circular economy.” In: *Sustainable Production and Consumption* (2020).

[13] Donald Garlotta. “A literature review of poly (lactic acid)”. In: *Journal of Polymers and the Environment* 9.2 (2001), pp. 63–84.

[14] Blanca Corona et al. “Towards sustainable development through the circular economy—A review and critical assessment on current circularity metrics”. In: *Resources, Conservation and Recycling* 151 (2019), p. 104498.

[15] Thomas E Graedel. “Industrial ecology”. In: (1995).

[16] Keith T Butler et al. “Machine learning for molecular and materials science”. In: *Nature* 559.7715 (2018), pp. 547–555.

[17] Zhifeng Liu et al. “Design and manufacturing model of customized hydrostatic bearing system based on cloud and big data technology”. In: *The International Journal of Advanced Manufacturing Technology* 84.1-4 (2016), pp. 261–273.

[18] Debbie Stokes. *Principles and practice of variable pressure/environmental scanning electron microscopy (VP-ESEM)*. John Wiley & Sons, 2008.

[19] Ernst Ruska. “The early development of electron lenses and electron microscopy”. In: *Microscopica acta. Supplement* Suppl 5 (1980), pp. 1–140.

[20] Thomas F Kelly and Michael K Miller. “Atom probe tomography”. In: *Review of Scientific Instruments* 78.3 (2007), p. 031101.

[21] Ankit Agrawal and Alok Choudhary. “Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science”. In: *Apl Materials* 4.5 (2016), p. 053208.

[22] Karel Matouš et al. “A review of predictive nonlinear theories for multiscale modeling of heterogeneous materials”. In: *Journal of Computational Physics* 330 (2017), pp. 192–220.

[23] National Science and Technology Council (US). *Materials genome initiative for global competitiveness*. Executive Office of the President, National Science and Technology Council, 2011.

[24] Juan J de Pablo et al. “The materials genome initiative, the interplay of experiment, theory and computation”. In: *Current Opinion in Solid State and Materials Science* 18.2 (2014), pp. 99–117.

[25] Juan J de Pablo et al. “New frontiers for the materials genome initiative”. In: *npj Computational Materials* 5.1 (2019), pp. 1–23.

[26] Gurdaman Khaira et al. “Derivation of multiple covarying material and process parameters using physics-based modeling of X-ray data”. In: *Macromolecules* 50.19 (2017), pp. 7783–7793.

[27] TH Kim et al. “Polar metals by geometric design”. In: *Nature* 533.7601 (2016), pp. 68–72.

[28] Rafael Gómez-Bombarelli et al. “Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach”. In: *Nature materials* 15.10 (2016), pp. 1120–1127.

[29] Jonathan Schmidt et al. “Recent advances and applications of machine learning in solid-state materials science”. In: *npj Computational Materials* 5.1 (2019), pp. 1–36.

[30] Xinbo Qi et al. “Applying neural-network-based machine learning to additive manufacturing: current applications, challenges, and future perspectives”. In: *Engineering* 5.4 (2019), pp. 721–729.

[31] Mahdi Imani and Seyede Fatemeh Ghoreishi. “Bayesian optimization objective-based experimental design”. In: *2020 American Control Conference (ACC)*. IEEE. 2020, pp. 3405–3411.

[32] James Nelson and Stefano Sanvito. “Predicting the Curie temperature of ferromagnets using machine learning”. In: *Physical Review Materials* 3.10 (2019), p. 104405.

[33] Baekjun Kim, Sangwon Lee, and Jihan Kim. “Inverse design of porous materials using artificial neural networks”. In: *Science advances* 6.1 (2020), eaax9324.

[34] Akshay Iyer et al. “Data-centric mixed-variable bayesian optimization for materials design”. In: *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. Vol. 59186. American Society of Mechanical Engineers. 2019, V02AT03A066.

[35] Tian Xie and Jeffrey C Grossman. “Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties”. In: *Physical review letters* 120.14 (2018), p. 145301.

[36] Turab Lookman et al. “Statistical inference and adaptive design for materials discovery”. In: *Current Opinion in Solid State and Materials Science* 21.3 (2017), pp. 121–128.

[37] Jeffrey M Rickman, Turab Lookman, and Sergei V Kalinin. “Materials informatics: From the atomic-level to the continuum”. In: *Acta Materialia* 168 (2019), pp. 473–510.

[38] Juhwan Noh et al. “Machine-enabled inverse design of inorganic solid materials: promises and challenges”. In: *Chemical Science* 11.19 (2020), pp. 4871–4881.

[39] Patricia M Stohr-Hunt. “An analysis of frequency of hands-on experience and science achievement”. In: *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching* 33.1 (1996), pp. 101–109.

[40] Hideo Akanuma. “The Significance of Early Bronze Age Iron Objects from Kaman-Kalehöyük, Turkey”. In: *Anatol Archaeol Stud* (2008), pp. 313–320.

[41] John Collis. *The European iron age*. Routledge, 2003.

[42] Isaac Newton. *Mathematical principles of natural philosophy*. A. Strahan, 1802.

[43] Alan A Luo. “Material design and development: From classical thermodynamics to CALPHAD and ICME approaches”. In: *Calphad* 50 (2015), pp. 6–22.

[44] Lev Landau. “The theory of phase transitions”. In: *Nature* 138.3498 (1936), pp. 840–841.

[45] John Edward Jones. “On the determination of molecular fields.—I. From the variation of the viscosity of a gas with temperature”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 106.738 (1924), pp. 441–462.

[46] Jagdish Mehra, Helmut Rechenberg, and Silvan S Schweber. “Volume 1: The Quantum Theory of Planck, Einstein, Bohr and Sommerfeld: Its Foundation and the Rise of Its Difficulties, 1900 1925 (Parts 1 and 2) and Volume 2: The Discovery of Quantum Mechanics, 1925 and Volume 3: The Formulation of Matrix Mechanics and Its Modifications, 1925 1926 and Volume 4: The Fundamental Equations of Quantum Mechanics, 1925 1926. The Reception of the New Quantum Mechanics, 1925 1926 and Volume 5: Erwin Schrödinger and the Rise of Wave Mechanics. Part 1: Schrödinger in Vienna and Zurich, 1887 1925. Part 2: The Creation of Wave Mechanics; Early Response and Application, 1925 1926”. In: *Physics Today* 40.9 (1987), p. 83.

[47] Pierre Hohenberg and Walter Kohn. “Inhomogeneous electron gas”. In: *Physical review* 136.3B (1964), B864.

[48] Walter Kohn and Lu Jeu Sham. “Self-consistent equations including exchange and correlation effects”. In: *Physical review* 140.4A (1965), A1133.

[49] James E Saal et al. “Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD)”. In: *Jom* 65.11 (2013), pp. 1501–1509.

[50] Scott Kirklin et al. “The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies”. In: *npj Computational Materials* 1.1 (2015), pp. 1–15.

[51] Anubhav Jain et al. “Commentary: The Materials Project: A materials genome approach to accelerating materials innovation”. In: *APL materials* 1.1 (2013), p. 011002.

[52] Dominik Marx and Jurg Hutter. “Ab initio molecular dynamics: Theory and implementation”. In: *Modern methods and algorithms of quantum chemistry* 1.301-449 (2000), p. 141.

[53] Anthony JG Hey, Stewart Tansley, Kristin Michele Tolle, et al. *The fourth paradigm: data-intensive scientific discovery*. Vol. 1. Microsoft research Redmond, WA, 2009.

[54] Valentin Stanev et al. “Machine learning modeling of superconducting critical temperature”. In: *npj Computational Materials* 4.1 (2018), pp. 1–14.

[55] Artem R Oganov, Andriy O Lyakhov, and Mario Valle. “How Evolutionary Crystal Structure Prediction Works and Why”. In: *Accounts of chemical research* 44.3 (2011), pp. 227–237.

[56] Neil C Thompson et al. “The computational limits of deep learning”. In: *arXiv preprint arXiv:2007.05558* (2020).

[57] Li Li et al. “Kohn-Sham equations as regularizer: Building prior knowledge into machine-learned physics”. In: *Physical Review Letters* 126.3 (2021), p. 036401.

[58] Benjamin Burger et al. “A mobile robotic chemist”. In: *Nature* 583.7815 (2020), pp. 237–241.

[59] Tom M Mitchell. “Machine learning and data mining”. In: *Communications of the ACM* 42.11 (1999), pp. 30–36.

[60] John R Koza et al. “Automated design of both the topology and sizing of analog electrical circuits using genetic programming”. In: *Artificial Intelligence in Design’96*. Springer, 1996, pp. 151–170.

[61] Satinder Singh, Andy Okun, and Andrew Jackson. “Learning to play Go from scratch”. In: *Nature* 550.7676 (2017), pp. 336–337.

[62] Robert R Schaller. “Moore’s law: past, present and future”. In: *IEEE spectrum* 34.6 (1997), pp. 52–59.

[63] Jinkyu Kim and John Canny. “Interpretable learning for self-driving cars by visualizing causal attention”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2942–2950.

[64] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[65] Sercan O Arik et al. “Neural voice cloning with a few samples”. In: *arXiv preprint arXiv:1802.06006* (2018).

[66] Thiago S Guzella and Walmir M Caminhas. “A review of machine learning approaches to spam filtering”. In: *Expert Systems with Applications* 36.7 (2009), pp. 10206–10222.

[67] Yudha P Pane et al. “Reinforcement learning based compensation methods for robot manipulators”. In: *Engineering Applications of Artificial Intelligence* 78 (2019), pp. 236–247.

[68] Oriol Vinyals et al. “Grandmaster level in StarCraft II using multi-agent reinforcement learning”. In: *Nature* 575.7782 (2019), pp. 350–354.

[69] Jürgen Schmidhuber. “Deep learning in neural networks: An overview”. In: *Neural networks* 61 (2015), pp. 85–117.

[70] Frank Emmert-Streib et al. “An introductory review of deep learning for prediction models with big data”. In: *Frontiers in Artificial Intelligence* 3 (2020), p. 4.

[71] Andrew W Senior et al. “Improved protein structure prediction using potentials from deep learning”. In: *Nature* 577.7792 (2020), pp. 706–710.

[72] Anna Chlingaryan, Salah Sukkarieh, and Brett Whelan. “Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review”. In: *Computers and electronics in agriculture* 151 (2018), pp. 61–69.

[73] Abigail Lavallin and Joni A Downs. “Machine learning in geography–Past, present, and future”. In: *Geography Compass* (2021), e12563.

[74] Gregory B Olson. “Computational design of hierarchically structured materials”. In: *Science* 277.5330 (1997), pp. 1237–1242.

[75] Juhwan Noh et al. “Inverse design of solid-state materials via a continuous representation”. In: *Matter* 1.5 (2019), pp. 1370–1384.

[76] Umesh Mishra and Jasprit Singh. *Semiconductor device physics and design*. Springer Science & Business Media, 2007.

[77] Alan D McNaught, Andrew Wilkinson, et al. *Compendium of chemical terminology*. Vol. 1669. Blackwell Science Oxford, 1997.

[78] Weihua Zhu and Heming Xiao. “First-principles band gap criterion for impact sensitivity of energetic crystals: a review”. In: *Structural Chemistry* 21.3 (2010), pp. 657–665.

[79] EH Kennard. “Moment of Momentum of Magnetic Electrons”. In: *Physical Review* 19.4 (1922), p. 420.

[80] Logan Ward et al. “Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations”. In: *Physical Review B* 96.2 (2017), p. 024104.

[81] Wojciech J Szlachta, Albert P Bartók, and Gábor Csányi. “Accuracy and transferability of Gaussian approximation potential models for tungsten”. In: *Physical Review B* 90.10 (2014), p. 104108.

[82] Hongbin Zhang. “High-throughput Design of Magnetic Materials”. In: *Electronic Structure* (2020).

[83] Sebastian Schaffer et al. “Machine learning methods for the prediction of micromagnetic magnetization dynamics”. In: *arXiv preprint arXiv:2103.09079* (2021).

[84] Alessandro Lunghi and Stefano Sanvito. “Surfing multiple conformation-property landscapes via machine learning: Designing single-ion magnetic anisotropy”. In: *The Journal of Physical Chemistry C* 124.10 (2020), pp. 5802–5806.

[85] Yue Liu et al. “The method of insulator recognition based on deep learning”. In: *2016 4th International Conference on Applied Robotics for the Power Industry (CARPI)*. IEEE. 2016, pp. 1–5.

[86] Siqi Shi et al. “Multi-scale computation methods: Their applications in lithium-ion battery research and development”. In: *Chinese Physics B* 25.1 (2015), p. 018212.

[87] Rafael Gómez-Bombarelli et al. “Automatic chemical design using a data-driven continuous representation of molecules”. In: *ACS central science* 4.2 (2018), pp. 268–276.

[88] Callum J Court et al. “3-D Inorganic Crystal Structure Generation and Property Prediction via Representation Learning”. In: *Journal of chemical information and modeling* 60.10 (2020), pp. 4518–4535.

[89] Jessica G Freeze, H Ray Kelly, and Victor S Batista. “Search for catalysts by inverse design: artificial intelligence, mountain climbers, and alchemists”. In: *Chemical reviews* 119.11 (2019), pp. 6595–6612.

[90] Rui Xin et al. “Active learning based generative design for the discovery of wide bandgap materials”. In: *arXiv preprint arXiv:2103.00608* (2021).

[91] Xue Ying. “An overview of overfitting and its solutions”. In: *Journal of Physics: Conference Series*. Vol. 1168. 2. IOP Publishing. 2019, p. 022022.

[92] OECD. Publishing. *OECD glossary of statistical terms*. Organisation for Economic Co-operation and Development, 2008.

[93] Richard C Holt. “Data descriptors: A compile-time model of data and addressing”. In: *ACM Transactions on Programming Languages and Systems (TOPLAS)* 9.3 (1987), pp. 367–389.

[94] Merriam Webster. “Merriam-Webster online dictionary”. In: *Retrieved June* 20 (2006), p. 2013.

[95] Hai Nguyen, Shin-ichi Maeda, and Kenta Oono. “Semi-supervised learning of hierarchical representations of molecules using neural message passing”. In: *arXiv preprint arXiv:1711.10168* (2017).

[96] Stuart J Russell and Peter Norvig. “Artificial Intelligence: A Modern”. In: *Approach, 2nded., PrenticeHall, December2002* (2010).

[97] Albert P Bartók, Risi Kondor, and Gábor Csányi. “On representing chemical environments”. In: *Physical Review B* 87.18 (2013), p. 184115.

[98] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine learning* 20.3 (1995), pp. 273–297.

[99] Tin Kam Ho. “Random decision forests”. In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE. 1995, pp. 278–282.

[100] Jonas Mockus. *Bayesian approach to global optimization: theory and applications*. Vol. 37. Springer Science & Business Media, 2012.

[101] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

[102] Maria V Valueva et al. “Application of the residue number system to reduce hardware costs of the convolutional neural network implementation”. In: *Mathematics and Computers in Simulation* 177 (2020), pp. 232–243.

[103] Samuel Dupond. “A thorough review on the current advance of neural network structures”. In: *Annual Reviews in Control* 14 (2019), pp. 200–230.

[104] Ashish Vaswani et al. “Attention is all you need”. In: *arXiv preprint arXiv:1706.03762* (2017).

[105] Geoffrey E Hinton, Terrence Joseph Sejnowski, et al. *Unsupervised learning: foundations of neural computation*. MIT press, 1999.

[106] Hans-Peter Kriegel, Erich Schubert, and Arthur Zimek. “The (black) art of runtime evaluation: Are we comparing algorithms or implementations?” In: *Knowledge and Information Systems* 52.2 (2017), pp. 341–378.

[107] Brendan J Frey and Delbert Dueck. “Clustering by passing messages between data points”. In: *science* 315.5814 (2007), pp. 972–976.

[108] IT Jolliffe. “Principal component analysis”. In: *Technometrics* 45.3 (2003), p. 276.

[109] Geoffrey Hinton and Sam T Roweis. “Stochastic neighbor embedding”. In: *NIPS*. Vol. 15. Citeseer. 2002, pp. 833–840.

[110] Ian J Goodfellow et al. “Generative adversarial networks”. In: *arXiv preprint arXiv:1406.2661* (2014).

[111] Jinwon An and Sungzoon Cho. “Variational autoencoder based anomaly detection using reconstruction probability”. In: *Special Lecture on IE* 2.1 (2015), pp. 1–18.

[112] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. “Reinforcement learning: A survey”. In: *Journal of artificial intelligence research* 4 (1996), pp. 237–285.

[113] Guenter Bergerhoff et al. “The inorganic crystal structure data base”. In: *Journal of chemical information and computer sciences* 23.2 (1983), pp. 66–69.

[114] Yibin Xu, Masayoshi Yamazaki, and Pierre Villars. “Inorganic materials database for exploring the nature of material”. In: *Japanese Journal of Applied Physics* 50.11S (2011), 11RH02.

[115] Brian L DeCost et al. “Uhcsdb: ultrahigh carbon steel micrograph database”. In: *Integrating Materials and Manufacturing Innovation* 6.2 (2017), pp. 197–205.

[116] Metals Handbook. *Desk edition*. 1998.

[117] Ron Jenkins, Mark Holomany, and Winnie Wong-Ng. “On the need for users of the Powder Diffraction File to update regularly”. In: *Powder Diffraction* 2.2 (1987), pp. 84–87.

[118] Stefano Curtarolo et al. “AFLOW: An automatic framework for high-throughput materials discovery”. In: *Computational Materials Science* 58 (2012), pp. 218–226.

[119] Claudia Draxl and Matthias Scheffler. “NOMAD: The FAIR concept for big data-driven materials science”. In: *Mrs Bulletin* 43.9 (2018), pp. 676–682.

[120] Jun Zhou et al. “2DMatPedia, an open computational database of two-dimensional materials from top-down and bottom-up approaches”. In: *Scientific data* 6.1 (2019), pp. 1–10.

[121] S Grazulis. “Daskevic A. Merkys A. Chateigner D. Lutterotti L. Quiros M. Serebryanaya NR Moeck P. Downs RT Le Bail A.(2012) Crystallography Open Database (COD): an open-access collection of crystal structures and platforms for world-wide collaboration”. In: *Nucleic acids research* 40 (), pp. D420–D427.

[122] Rongxin Xia and Sabre Kais. “Quantum machine learning for electronic structure calculations”. In: *Nature communications* 9.1 (2018), pp. 1–6.

[123] Andriy Zakutayev et al. “An open experimental database for exploring inorganic materials”. In: *Scientific data* 5.1 (2018), pp. 1–12.

[124] Mark D Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific data* 3.1 (2016), pp. 1–9.

[125] NISO Press. “Understanding metadata”. In: *National Information Standards* 20 (2004).

[126] Lauren Takahashi and Keisuke Takahashi. “Visualizing Scientists’ Cognitive Representation of Materials Data through the Application of Ontology”. In: *The journal of physical chemistry letters* 10.23 (2019), pp. 7482–7491.

[127] advice f "u r information infrastructures. *Achievement from diversity: Recommendations on structures, processes and financing of research data management in Germany*. Rat f "u r information infrastructures (RfII), 2016.

[128] Luca M Ghiringhelli et al. “Big data of materials science: critical role of the descriptor”. In: *Physical review letters* 114.10 (2015), p. 105503.

[129] Atsuto Seko, Atsushi Togo, and Isao Tanaka. “Descriptors for machine learning of materials data”. In: *Nanoinformatics*. Springer, Singapore, 2018, pp. 3–23.

[130] Satosi Watanabe. “Knowing and Guessing a Quantitative Study of Inference and Information”. In: (1969).

[131] Matthias Rupp et al. “Fast and accurate modeling of molecular atomization energies with machine learning”. In: *Physical review letters* 108.5 (2012), p. 058301.

[132] Felix Faber et al. “Crystal structure representations for machine learning models of formation energies”. In: *International Journal of Quantum Chemistry* 115.16 (2015), pp. 1094–1101.

[133] Jörg Behler. “Atom-centered symmetry functions for constructing high-dimensional neural network potentials”. In: *The Journal of chemical physics* 134.7 (2011), p. 074106.

[134] Logan Ward et al. “A general-purpose machine learning framework for predicting properties of inorganic materials”. In: *npj Computational Materials* 2.1 (2016), pp. 1–7.

[135] Teng Long et al. “CCDCGAN: Inverse design of crystal structures”. In: *arXiv preprint arXiv:2007.11228* (2020).

[136] Teng Long et al. “Inverse design of crystal structures for multicomponent systems”. In: *arXiv preprint arXiv:2104.08040* (2021).

[137] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. “A systematic review of algorithm aversion in augmented decision making”. In: *Journal of Behavioral Decision Making* 33.2 (2020), pp. 220–239.

[138] David H Wolpert. “The lack of a priori distinctions between learning algorithms”. In: *Neural computation* 8.7 (1996), pp. 1341–1390.

[139] Sebastian Raschka and Vahid Mirjalili. “Python Machine Learning: Machine Learning and Deep Learning with Python”. In: *Scikit-Learn, and TensorFlow. Second edition ed* (2017).

[140] George EP Box. “Science and statistics”. In: *Journal of the American Statistical Association* 71.356 (1976), pp. 791–799.

[141] Shing Chan and Ahmed H Elsheikh. “A machine learning approach for efficient uncertainty quantification using multiscale methods”. In: *Journal of Computational Physics* 354 (2018), pp. 493–511.

[142] TP Barnett and R Preisendorfer. “Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis”. In: *Monthly Weather Review* 115.9 (1987), pp. 1825–1850.

[143] Mauricio A Alvarez, Lorenzo Rosasco, and Neil D Lawrence. “Kernels for vector-valued functions: A review”. In: *arXiv preprint arXiv:1106.6251* (2011).

[144] Cort J Willmott and Kenji Matsuura. “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance”. In: *Climate research* 30.1 (2005), pp. 79–82.

[145] Lorenzo Rosasco et al. “Are loss functions all the same?” In: *Neural computation* 16.5 (2004), pp. 1063–1076.

[146] Claude Lemaréchal. “Cauchy and the gradient method”. In: *Doc Math Extra* 251.254 (2012), p. 10.

[147] Léon Bottou and Olivier Bousquet. “13 the tradeoffs of large-scale learning”. In: *Optimization for machine learning* (2011), p. 351.

[148] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[149] John Duchi, Elad Hazan, and Yoram Singer. “Adaptive subgradient methods for online learning and stochastic optimization.” In: *Journal of machine learning research* 12.7 (2011).

[150] Matthew D Zeiler. “Adadelta: an adaptive learning rate method”. In: *arXiv preprint arXiv:1212.5701* (2012).

[151] Tijmen Tieleman, Geoffrey Hinton, et al. “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude”. In: *COURSERA: Neural networks for machine learning* 4.2 (2012), pp. 26–31.

[152] Zijun Zhang. “Improved adam optimizer for deep neural networks”. In: *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*. IEEE. 2018, pp. 1–2.

[153] Sebastian Bock and Martin Weiß. “A proof of local convergence for the Adam optimizer”. In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2019, pp. 1–8.

[154] Nico JD Nagelkerke et al. “A note on a general definition of the coefficient of determination”. In: *Biometrika* 78.3 (1991), pp. 691–692.

[155] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[156] Cullen Schaffer. “Selecting a classification method by cross-validation”. In: *Machine Learning* 13.1 (1993), pp. 135–143.

[157] S Madeh Piryonesi and Tamer E El-Diraby. “Data analytics in asset management: Cost-effective prediction of the pavement condition index”. In: *Journal of Infrastructure Systems* 26.1 (2020), p. 04019036.

[158] William D Callister Jr and David G Rethwisch. *Fundamentals of materials science and engineering: an integrated approach*. John Wiley & Sons, 2020.

[159] Teng Long et al. “An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature”. In: *Materials Research Letters* 9.4 (2021), pp. 169–174.

[160] Lauri Himanen et al. “DScribe: Library of descriptors for machine learning in materials science”. In: *Computer Physics Communications* 247 (2020), p. 106949.

[161] Albert P Bartók et al. “Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons”. In: *Physical review letters* 104.13 (2010), p. 136403.

[162] Miguel A Caro. “Optimizing many-body atomic descriptors for enhanced computational performance of machine learning based interatomic potentials”. In: *Physical Review B* 100.2 (2019), p. 024112.

[163] Lixin Sun et al. “Two-tier machine learning acceleration of molecular dynamics with enhanced sampling: surface reactions and restructuring on metal catalysts”. In: *Bulletin of the American Physical Society* (2021).

[164] Abdulnour Y Toukmaji and John A Board Jr. “Ewald summation techniques in perspective: a survey”. In: *Computer physics communications* 95.2-3 (1996), pp. 73–92.

[165] RA Jackson and CRA Catlow. “Computer simulation studies of zeolite structure”. In: *Molecular Simulation* 1.4 (1988), pp. 207–224.

[166] David Harris and Sarah Harris. *Digital design and computer architecture*. Morgan Kaufmann, 2010.

[167] Derek A Pisner and David M Schnyer. “Support vector machine”. In: *Machine Learning*. Elsevier, 2020, pp. 101–121.

[168] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[169] Warren E Walker et al. “Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support”. In: *Integrated assessment* 4.1 (2003), pp. 5–17.

[170] Tivadar Danka and Peter Horvath. “modAL: A modular active learning framework for Python”. In: *arXiv preprint arXiv:1805.00979* (2018).

[171] Afroz Chakure. *Random Forest Regression*. `https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f`. 2019.

[172] Tin Kam Ho. “The random subspace method for constructing decision forests”. In: *IEEE transactions on pattern analysis and machine intelligence* 20.8 (1998), pp. 832–844.

[173] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[174] Pierre Geurts, Damien Ernst, and Louis Wehenkel. “Extremely randomized trees”. In: *Machine learning* 63.1 (2006), pp. 3–42.

[175] Hieu Chi Dam et al. “A regression-based feature selection study of the Curie temperature of transition-metal rare-earth compounds: prediction and understanding”. In: *arXiv preprint arXiv:1705.00978* (2017).

[176] Steven Gong. *How does a Neural Network work intuitively in code?* `https://gongster.medium.com/how-does-a-neural-network-work-intuitively-in-code-f51f7b2c1e3f`. 2019.

[177] Eric W Weisstein. “Sigmoid function”. In: *https: // mathworld. tungsten. com /* (2002).

[178] Engui Fan. “Extended tanh-function method and its applications to nonlinear equations”. In: *Physics Letters A* 277.4-5 (2000), pp. 212–218.

[179] Jason Brownlee. “A gentle introduction to the rectified linear unit (ReLU)”. In: *Machine learning mastery* 6 (2019).

[180] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. “Rectifier nonlinearities improve neural network acoustic models”. In: *Proc. icml*. Vol. 30. 1. Citeseer. 2013, p. 3.

[181] François Chollet. *keras*. `https://github.com/fchollet/keras`. 2015.

[182] Yoshua Bengio, Aaron Courville, and Pascal Vincent. “Representation learning: A review and new perspectives”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[183] Ian Goodfellow et al. *Deep learning*. Vol. 1. 2. MIT press Cambridge, 2016.

[184] Anamika Dhillon and Gyanendra K Verma. “Convolutional neural network: a review of models, methodologies and applications to object detection”. In: *Progress in Artificial Intelligence* 9.2 (2020), pp. 85–112.

[185] Wang Zhiqiang and Liu Jun. “A review of object detection based on convolutional neural network”. In: *2017 36th Chinese Control Conference (CCC)*. IEEE. 2017, pp. 11104–11109.

[186] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”. In: *arXiv preprint arXiv:1409.0473* (2014).

[187] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[188] Igor V Tetko et al. “State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis”. In: *Nature communications* 11.1 (2020), pp. 1–11.

[189] Fabian B Fuchs et al. “Iterative SE (3)-Transformers”. In: *arXiv preprint arXiv:2102.13419* (2021).

[190] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[191] Tim Van Erven and Peter Harremos. “Rényi divergence and Kullback-Leibler divergence”. In: *IEEE Transactions on Information Theory* 60.7 (2014), pp. 3797–3820.

[192] Joseph Rocca. *Understanding Variational Autoencoders (VAEs)*. https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73. 2019.

[193] Antonia Creswell et al. “Generative adversarial networks: An overview”. In: *IEEE Signal Processing Magazine* 35.1 (2018), pp. 53–65.

[194] Roger B Myerson. “Nash equilibrium and the history of economic theory”. In: *Journal of Economic Literature* 37.3 (1999), pp. 1067–1082.

[195] Sungwon Kim et al. “Generative adversarial networks for crystal structure prediction”. In: *ACS central science* 6.8 (2020), pp. 1412–1420.

[196] Soshin Chikazumi and Chad D Graham. *Physics of Ferromagnetism 2e*. 94. Oxford University Press on Demand, 2009.

[197] Nedko Drebov et al. “Ab initio screening methodology applied to the search for new permanent magnetic materials”. In: *New Journal of Physics* 15.12 (2013), p. 125023.

[198] AA Katanin, AS Belozerov, and VI Anisimov. “Nonlocal correlations in the vicinity of the α- γ phase transition in iron within a dmft plus spin-fermion model approach”. In: *Physical Review B* 94.16 (2016), p. 161117.

[199] Alan JH McGaughey and M Kaviany. “Quantitative validation of the Boltzmann transport equation phonon thermal conductivity model under the single-mode relaxation time approximation”. In: *Physical Review B* 69.9 (2004), p. 094303.

[200] Jacob Biamonte et al. “Quantum machine learning”. In: *Nature* 549.7671 (2017), pp. 195–202.

[201] O Anatole Von Lilienfeld. “Quantum machine learning in chemical compound space”. In: *Angewandte Chemie International Edition* 57.16 (2018), pp. 4164–4169.

[202] Mario Krenn et al. “Automated search for new quantum experiments”. In: *Physical review letters* 116.9 (2016), p. 090405.

Appendices

A. Insightful discussion of the publications

A.1. An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature

Question 1:

> *"i) On page three a statement is made concerning the electron states of Co and Ni, and their strong itinerant nature being responsible for why DFT based theory fails in describing their Curie temperature. This is completely erroneous. Most 3d intermetallics and compounds, such as the ones analyzed in this paper, have the same degree of itineracy, when judged by the conventional measures; Hubbard U compared to bandwidth W. It is true that Ni has signals of correlations in the electronic structure, as judged by the so called "6 eV satellite" in the measured valence band spectrum, but Co lacks this feature. Furthermore, the statement that DFT fails for these systems is simply wrong. In the textbook by Kubler (Theory of itinerant electron magnetism, Oxford publications) the calculated Curie temperature of hcp Co is 1205 K, which can be compared to the experimental value of ~1400 K. Furthermore, bcc Fe is listed with a theoretical Curie temperature of 1095 while the experimental value is 1044 K. It is true that calculations of fcc Ni are less accurate, but this has nothing to do with the degree of itinerancy of the electron states, but rather that the assumptions behind mapping DFT results to a spin-Hamiltonian (as described in Kublers book, and many other works in the field) and treating this Hamiltonian classically using Monte Carlo simulations, is less accurate for Ni. The complications arise due to the small moment of Ni, which prevents an accurate description using classical spins. The description in the paper by Long et al, is completely misleading on this point and must be improved. This involves the statement on page 8 that these types of methods lack accuracy and generality."*

We thank the Reviewer for his/her insightful comments. We agree with the Reviewer that the magnetism in most transition-metal-based intermetallic compounds is of significant itinerant nature, in contrast to that in the insulating materials such as transition metal oxides and rare-earth based systems. In this regard, the U vs. W is a good descriptor classifying such two categories of magnetic materials, but it cannot be applied to assess whether an intermetallic compound is more localized or itinerant. Instead, the Rhodes-Wohlfarth plot (as shown in Figure 1, Santiago, J. M., C. L. Huang, and E. Morosan. "Itinerant magnetic metals." *Journal of Physics: Condensed Matter* 29.37 (2017): 373002.) should be used, where the ratio of Curie-Weiss constant q_c and the saturation magnetization q_s is plotted with respect to the Curie temperature. Obviously, Fe shows a typical localized moment behavior as $q_c/q_s \sim 1$ like EuO, whereas Ni is deviated from the localized limit due to enhanced itinerancy.

Furthermore, we agree with the Reviewer that the calculated Curie temperature of Co seems okay, but we wanted to emphasis that such a success on Co is different from Fe, *e.g.*, DFT fails to get the correct orbital moments where orbital polarization correction is needed (Eschrig, H., Sargolzaei, M., Koepernik, K., & Richter, M. "Orbital polarization in the Kohn-Sham-Dirac theory." *EPL (Europhysics Letters)* 72.4 (2005): 611.). Additionally, it is well known that the local spin density approximation (LSDA) gives wrong ground state of Fe, *i.e.,* LSDA predicts the nonmagnetic fcc phase being more stable than the FM bcc phase (Peltzer y Blanca, E. P., Rodríguez, C. O., Shitu, J., & Novikov, D. L. "Degree of localization of the exchange-correlation hole and its influence on the ground-state (structural and magnetic) properties of d metals." *Journal of Physics: Condensed Matter* 13.42 (2001): 9463.). Therefore, careful DFT calculations should be done in order to obtain the proper electronic structure even for such elemental magnets. Such subtle complications are well represented by fcc Ni, where the ferromagnetism and associated electronic structure are driven by the nonlocal charge fluctuation and local spin fluctuations (both missing in standard LSDA), and such effects cancel thus render the magnetic moments of Ni fortuitously good in the LSDA,

following Ref. 8 of the updated manuscript (L. Sponza, P. Pisanti, A. Vishina, D. Pashov, C. Weber, M. van Schilfgaarde, S. Acharya, J. Vidal, and G. Kotliar "Self-energies in itinerant magnets: a focus on Fe and Ni." *Physical Review B* 95.4 (2017): 041112.). Therefore, the current DFT methods should be applied carefully in order to deliver a reliable description of magnetism in Fe, Co, and Ni, not to mention the other intermetallic magnetic materials.

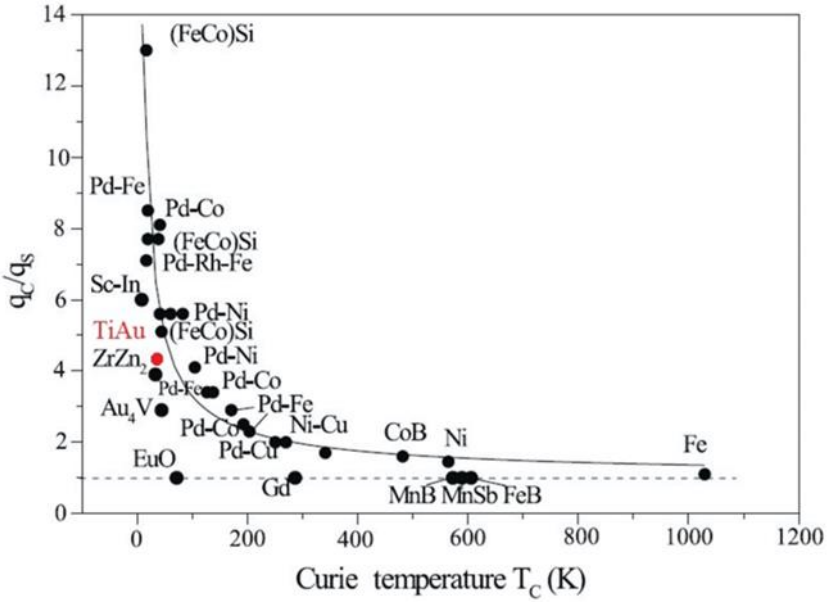


Figure 1. The Rhodes-Wohlfarth ratio q_c/q_s for various magnetic materials.

In this regard, it is believed that the multi-orbital Hubbard model with hybridizations accounted for at the DFT level is a good choice to obtain the accurate electronic structure of such magnetic intermetallic compounds. Thus, there are in principle two ways to obtain the critical magnetic ordering temperatures: the one is to solve the Hubbard model, *e.g.*, based on DFT+DMFT, to look for the temperature of vanishing ordered moments; the other is to project the electronic structure onto an effective spin model (*e.g.*, Heiserberg model) followed by Monte Carlo modelling. It is noted that the projected Heisenberg model is only valid in the strong coupling limit (*i.e.*, localized limit) of the Hubbard model (*cf.* Fazekas, Patrik. *Lecture notes on electron correlation and magnetism*. Vol. 5. World scientific, 1999.). Therefore, as pointed out by the Reviewer, it is questionable whether mapping DFT electronic structure to a spin Hamiltonian is valid for all the magnetic intermetallic compounds, given the unknown magnitude of itinerancy. We note also that usually the classical Monte Carlo model is performed instead of the quantum Monte Carlo to evaluate the Curie temperature.

Therefore, there are at least two origins of inaccuracy for evaluating the Curie temperature by parameterizing the DFT electronic structure into a spin Hamiltonian: (1) whether the electronic structure is reliably described and (2) whether it is justified to use the Heisenberg model, as discussed thoroughly in our recent review (Zhang, Hongbin. "High-throughput Design of Magnetic Materials." *arXiv preprint arXiv:2008.12907* (2020).). The former is tricky because all the intermetallic compounds lie at *a priori unknown* crossovers between the localized and itinerant limits. This further dictates the latter.

We agree with the Reviewer that our original statements were not detailed thus they may appear misleading. Based on the arguments above, our intention was to emphasize that even for elemental magnets like Fe, Co,

and Ni, how to obtain accurate electronic structure and how to construct the spin Hamiltonian (and whether it is justified to do so) are challenging tasks. Not to mention the accuracy of theoretically evaluated Curie temperatures for *all* the known magnetic intermetallic compounds, which to the best of our knowledge has not yet been done.

Question 2:

> *"ii) A bit further down on the same page, it is listed that DMFT yields a value for bcc Fe that is 50 % off compared to experiment. Why do the authors not quote the calculated values from other works? In PRB 91, 125133 (2015) the DMFT value for bcc Fe is 840 K. Also, there are several other papers that report DMFT values of Fe, Co and Ni. To simply select one particularly bad value is poor practice when describing the research field, and the paper needs improvement on this point. The reason for why different practitioners in DMFT get different results should also be discussed. "*

We appreciate this comment from the Reviewer. We did not deliberately select the bad value, which is obtained based on the state-of-the-art implementation of DFT+DMFT. Instead we were simply trying to give an example, as we did not aim to reconcile the discrepancies therein. On the other hand, regarding fluctuating values of Curie temperature obtained using DMFT for Fe, Co, and Ni, the reasons are possibly due to (a) the implementation DFT+DMFT, *i.e.*, how to define the local orbitals (*e.g.*, Wannier function vs. embedding), (b) whether there is charge self-consistency, (c) how the Coulomb interaction is considered (*e.g.*, density-density correlations vs. rotational invariant), and (d) the values of U and J.

Question 3:

> *"Related to this is the implementation of the machine learning approach. 80% of the data is used for training while the other 20% are used for validation. How was this partitioning made? If completely random, a second try of partioning the data randomly would generate a new set of training (80%) and validation (20%) data. Results from such a re-run should be compared to the present results and would justify that the results are not biased due to data selection. With machine learning it's customary to split your data into three datasets: train, validation, and test. The split is recommended to be performed in a reproducible way, e.g. by assigning a random seed and shuffling the dataset. In this work I interpret that the initial dataset (AtomWork) is split into two and the tested on a completely different dataset (Materials Project)."*

We appreciate this insightful comment from the Reviewer. First of all, as discussed in the reply to the previous comment, we used only the AtomWork database to construct our machine learning model, where the data are split into training and validation sets following the 80%-20% rule. Furthermore, the splitting into the training and validation sets is done by using random seeds, which is conducted using the standard "*train_test_split*" function of "*sklearn*" package. In order to reproduce our result in the manuscript, one can simply set the random seed (random state of the function) to 2. Lastly, to avoid possible biased splitting, we performed cross validations for both classification (*i.e.,* classifying the FM and AFM materials) and regression (*i.e.,* fitting the Curie temperature of the FM materials), as shown in Fig. 2(b) and Fig. 3(b) in the main text. Cross validation is a fundamental technique in machine learning to switch data between the training and validation sets. In addition, training of other 100 splitting is also conducted, *i.e.*, with random seeds from 0 to 100. The accuracy for the validation on the FM/AFM classification ranges from 85% to 87%

as shown in Figure 2, with an average accuracy of 85.5%. The accuracy for the regression of Curie temperature for the FM materials ranges 89% to 91% as shown in Figure 3, with an average accuracy of 90.1%. These confirm that our machine learning model is not biased due to training data selection.

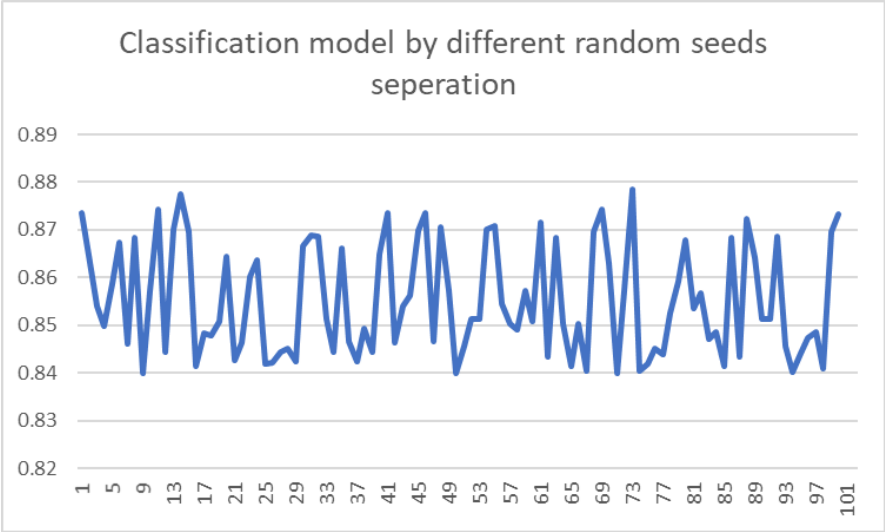


Figure 2. Validation accuracy of the FM/ AFM classification model using different random seeds. The x-axis represents the random seeds, and y-axis is the resulting accuracy.

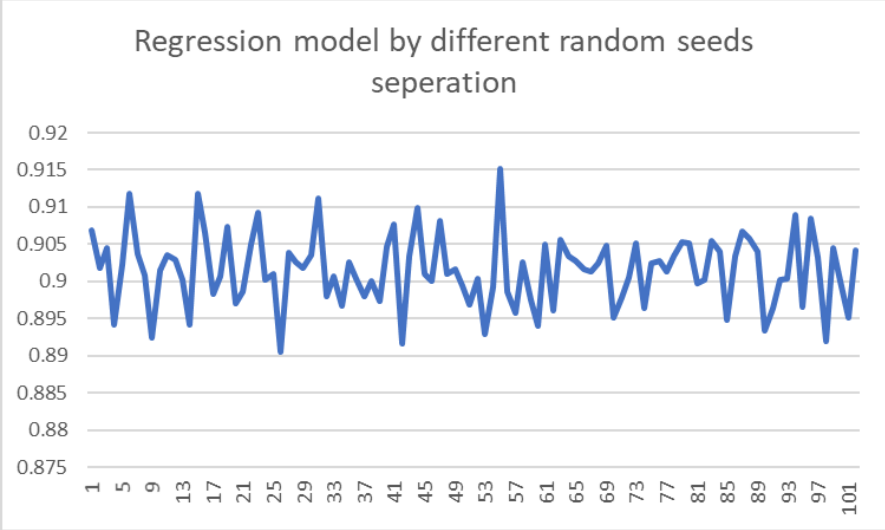


Figure 3. Validation accuracy of the two-step regression model using different random seeds. The x-axis represents the random seeds, and y axis is the resulting accuracy.

As the Reviewer mentioned, for a conventional machine learning modelling, data should be split into three datasets: training, validation and test. The training set is used to train the model, the validation set is to select the best performance model, and the test set is to demonstrate the ability of the best model. In our work, as shown in Figure 1, after getting the model trained and validated using the data from AtomWork, we performed predictions for compounds in another database (Materials Project). By further validating our predictions using the experimental Curie temperatures collected for 85 compounds (account to about 5% of the data from AtomWork) *by hand*, the resulting high accuracy of 90.5%. This confirms that our model exhibits the predictive power, and it is also expected to be a more robust and stricter test of our machine learning model than the traditional way.

Additionally, it is noted that in machine learning for materials science, splitting of the database into only training and test is also a common practice. This is because that the hyperparameter determination is not highly demanded, and possible biased data splitting can be checked by cross validation. For example, Stanev, Valentin, et al. "Machine learning modeling of superconducting critical temperature." *npj Computational Materials* 4.1 (2018): 1-14. used the training/test (validation) splitting to model the superconductor critical temperature. In this regard, we believe the training/validation splitting of the AtomWork data with on-top predictions done in our work is also justified.

Question 4:

> *"In the introduction it is stated that state-of-the-art DFT+DMFT methods are numerically expensive and could, due to lack of consistency in the numerical implementation, give quite spread in Tc as exemplified for bcc Fe. Here the impression is that DFT+DMFT is far from useful, however, there are many great examples where it has been successful. Why would machine learning on a broad selection of materials with rather "simple" input perform better then DFT+DMFT. "*

We cannot agree more with the Reviewer that DFT+DMFT provides the state-of-the-art solution to evaluate the critical temperature of magnetic materials, as one of the coauthors (Hongbin Zhang) has applied the methods on a series of compounds. However, exactly due to the fact that the DFT+DMFT methods are still under developments, the resulting Curie temperature of ferromagnetic materials are quite distributed, as shown in Table 1 and Table 2 for fcc Ni and bcc Fe. There are two main reasons for such inconsistencies, *i.e.,* (1) arbitrariness in choosing the U and J parameters and (2) the implementation of the Coulomb interaction, *e.g.*, rotational invariant or only density-density interaction. It is also noted that the DFT+DMFT calculations are expensive, particularly for rare-earth based magnetic compounds with correlated *4f*-shells. Thus, we believe that there is still a long way to go until the DFT+DMFT methods can be applied to evaluate extensively the Curie temperatures for the known materials in the database and to make further predictions.

Table 1. Curie temperature of fcc Ni obtained based on the DFT+DMFT methods

Publication	Code	U (eV)	J (eV)	Coulomb interaction	Calculated T_C (K)
Hausoel, et al. *Nature communications* 8.1 (2017): 1-9.	VASP + w2dynamics	2.71	0.8	Non-local exchange	530
			0.9		650
			1.08		860
Belozerov, et al. *Physical Review B* 87.12 (2013): 125138.	Quantum-ESPRESSO	2.3	1.0	dens.-dens./ rot. inv.	840 / 400
Lichtenstein, et al. *Physical Review Letters* 87.6 (2001): 067205.		3.0	0.9	density-density	700

Table 2. Curie temperature of bcc Fe obtained based on DFT+DMFT

Publication	Code	U (eV)	J (eV)	Coulomb interaction	Calculated T_C (K)
Han, et al. *Physical review letters* 120.18 (2018): 187203.	Wien2k + EDMFT	5.5	0.84	Ising/Full Coulomb interaction	2500 / 1550

Belozerov, et al. *Journal of Physics: Condensed Matter* 26.37 (2014): 375601.	Quantum-ESPRESSO	4	0.8	dens.-dens./ rot. inv.	1750 / 970
		1.5	0.9		2030 / 1230
		2.3	0.9		2050 / 1260
		4.0	0.9		2060 / 1270
		6.0	0.9		1980 / 1200
		4.0	1.0		2390 / 1560
Hausoel, et al. *Nature communications* 8.1 (2017): 1-9.	VASP + w2dynamics	2.71	0.8	Non-local exchange	1500
			0.9		1900
			1		2300
Belozerov, et al. *Physical Review B* 87.12 (2013): 125138.	Quantum-ESPRESSO	2.3	0.9	dens.-dens./ rot. inv.	2050 / 1260
Kvashnin, et al. *Physical Review B* 91.12 (2015): 125133.	RSPt	2.3	0.9	SPTF	840
Katanin, et al. Anisimov. *Physical Review B* 94.16 (2016): 161117.	exciting-plus + Quantum-ESPRESSO	4.0	0.9	spin-rotationally-invariant	1185
Belozerov, et al. *Physical Review B* 96.7 (2017): 075108.	exciting-plus + Quantum-ESPRESSO	4.0	0.9	density-density	2150
Lichtenstein, et al. *Physical Review Letters* 87.6 (2001): 067205.		2.3	0.9	density-density	1900
Leonov, et al. *Physical review letters* 106.10 (2011): 106405.		1.8	0.9	density-density	1600

However, by no means, we want to negate the DFT+DMFT methods. Nevertheless, given the non-converging results for fcc Ni (Table 1) and bcc Fe (Table 2) and the tremendous numerical efforts needed to perform DFT+DMFT calculations on real materials, it is fair to say that DFT+DMFT cannot provide an immediate support to make prediction. On the other hand, we believe that the machine learning model presented in our work is a valuable pragmatic solution to guide the design and engineering of magnetic materials. Additionally, the machine learning model is trained by learning on thousands of Curie temperature from experiments, which provides a valid statistical perspective, rather than focusing on the physical causality.

Question 5:

> *"It is found that the average of magnetic moment of the elemental solids for atoms in a given compound (mean GSmagmom) is most important feature. Moreover, it is mentioned that the second most important feature is the structural descriptors could be expected to be more important when the FM and AFM states are competing, especially in terms the unit-cell volume. How sensitive then would such descriptors be depending on the xc-potential used, say comparing LDA vs GGA vs higher-ranked xc-potentials? Some bias are to be expected depending on the xc-potential used, with its inherent cons and flaws, and its impact on the electronic structure and internal interactions. Showing such comparison followed by a thorough discussion would benefit the present work significantly."*

We appreciate this comment from the Reviewer. Firstly, it is noted that our original statements mean that the "mean GSmagmon" is the most important feature, but the structural descriptors are not the second important feature. What we wrote was *"It is noted that the STR features contribute almost 15%, which proves its significance in determining the magnetic ground state."* As we specified in the *Method, Descriptors*, STR features are a series of descriptors that includes cell volume, space group and 24 SOAP descriptors. After feature reduction, only four STR descriptors are left in the classification model, and the importance of each STR descriptor is listed in Table S3. These STR descriptors contribute 15% in determining the magnetic ground state, which are important, but meanwhile the CHEM descriptors contribute 85%, which are dominant.

Secondly, during the training process, we used only the experimental structures from AtomWork, thus the structures we used are not affected by the xc-potentials. Nevertheless, following the Reviewer's suggestion, we collected relaxed crystal structures using GGA for 682 compounds (the best we can do using the latest pymatgen API) from Materials Project. After substituting such crystal structures into the database and retraining the model, it is observed that the feature importance has exactly the same order as our previous mode, as shown in Table 3. For instance, the contribution of "mean_GSmagmon" reached 50%, while the contribution of the volume has dropped to 0.3%. Generally, all the feature importance of STR descriptors have slightly dropped, but the total contribution of STR still stands at 14% (Table 3). Therefore, considering the experimental crystal structures as obtained from a *hypothetical* xc-functional, we draw the conclusion that the feature importance of our machine learning model is not sensitive to the crystal structures from different xc-potentials.

Table 3. Feature importance of classification model.

Descriptors	Class	Importance (trained by experimental structures)	Importance (trained by GGA relaxed structures)
mean_GSmagmom	CHEM	0.493264	0.501322
mean_NpUnfilled	CHEM	0.091244	0.093346
SOAP_1	STR	0.059611	0.054591
Space Group No.	STR	0.057755	0.054466
maxdiff_MendeleevNumber	CHEM	0.052555	0.053214
mean_NdValence	CHEM	0.04491	0.046792
dev_MendeleevNumber	CHEM	0.035493	0.03621
dev_Number	CHEM	0.032495	0.031145
mean_MendeleevNumber	CHEM	0.024161	0.023289
SOAP_18	STR	0.022814	0.022032
min_Number	CHEM	0.021957	0.021287
maxdiff_MeltingT	CHEM	0.021553	0.020734
dev_NdUnfilled	CHEM	0.018863	0.019682
mean_MeltingT	CHEM	0.01759	0.018553
Volume	STR	0.005734	0.003337

Question 6:

> *"How are the AFM configurations defined in the present work? Is the spin configuration for a given materials defined in the AtomWork database? As of now the AFM is undefined or given the impression of being undefined. Is the AFM configuration defined within the unicell alone or in multiple unitcells as needed when describing, e.g. G-AFM of BiFeO3 (and many other perovskites). Or is the input entirely based on TN irrespective of how the AFM looks like?"*

We thank the Reviewer for the insightful comment. The AFM/FM label used in this work is provided by the AtomWork database directly. However, if the compounds exhibit multiple magnetic transitions, we labeled them as FM or AFM based on the first transition from the paramagnetic state, *i.e.*, we aim at modeling the high-temperature magnetic ordering temperature. Additionally, the ferrimagnetic compounds are also considered as FM, covering the most rare-earth based magnets.

Furthermore, the original AFM compounds in AtomWork are collected from experiments, and they are all labelled by T_N. Unfortunately, there is no spin configuration for such compounds in the AtomWork database. Therefore, our current machine learning model cannot predict the spin configurations, and the input is entirely based on T_N irrespective of the underlying spin configuration. This is also the reason why we focus on performing regression on the Curie temperature, though we can classify the FM and AFM materials. Nevertheless, machine learning modeling of spin configurations is an interesting subject for the future, and we are trying to collected such spin configurations from the literature and to generate a database based on magnetic cif files.

A.2. Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures

Reply to the Reviewer 1:

> *"The paper "Constrained crystals deep convolutional generative adversarial network: Inverse design of crystal structures" is an original work based on generative algorithms applied to chemical crystal structures design (# NPJCOMPUMATS-02049-T)."*
>
> *This is an interesting article that combines conventional computational chemistry techniques with the state-of-the-art machine learning methods to generate new structures of inorganic materials, here the Bi-Se compounds as an example. Recently, there have been numerous suggestions in the literature that the use of generated adversarial networks (GAN) or variation of autoencoders (VAE) would be away to "fill the gaps" in a sparse configuration space of materials with a larger number of unobserved compounds. The present work is trying to answer this problem and delivers an original method that successfully couples a reversible crystal description model including the stability (heat of formation) as the main descriptor in the machine learning approach. I particularly appreciate the use of the inverse crystal graph through autoencoder where crystal structures are reconstructed with 96%."*

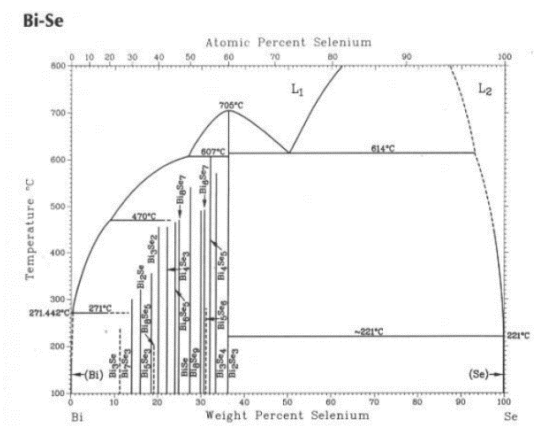
We thank the Reviewer for his/her positive impression on our work, and we are grateful for the constructive comments.

Question 1:

> *"(1) The Bi-Se system is already known to present numerous ordered phases as Bi3Se2, BiSe, Bi2Se3, and even at low temperature: Bi3Se, Bi7Se3, Bi2Se, ... reported on the Okamoto's handbook.*
> *Among all the generative approaches introduced in this paper (DCGAN , CCDCGAN), the only generated compound clearly stable is the Bi3Se2. It is fine to generate thousands of structures but my first question is simple: how to evaluate an algorithm if it is not able to reproduce the expected ground state?"*

We thank the Reviewer for her/his very insightful comments.

In our previous DFT calculations and machine learning modelling, we considered Bi_2Se_3 as the only stable phase on the convex hull. As the Reviewer pointed out, there are several metastable phases which are experimentally accessible, as shown in the phase diagram (Figure 1). We have collected the corresponding experimental crystal structures of such Bi-Se phases and reconstructed the phase diagram based on further DFT calculations, as shown in Figure 2. Obviously, the formation energies of the phases reported in Okamoto's handbook are on average about 0.06 eV/atom higher than the convex hull, *i.e.,* they are metastable phases. Such phases are incorporated into our machine learning modelling, as our goal is to identify not only the phases on the convex hull but also possible metastable phases.



H. Okamoto, unpublished

Phase	Composition, wt% Se	Pearson symbol	Space group
(αBi)	0	*hR*2	$R\bar{3}m$
Bi_7Se_3	14	*hR*20	$R\bar{3}m$
Bi_2Se	15.9	*hP*9	$P\bar{3}m1$
Bi_3Se_3(a)	18.5	*hP*48	$P\bar{3}m1$
Bi_3Se_2	20	*hP*30	$P\bar{3}m1$
Bi_4Se_3	22.1	*hR*7	$R\bar{3}m$
Bi_6Se_5	24.0	*hP*33	$P\bar{3}m1$
Bi_8Se_7	24.8	*hP*45	$P\bar{3}m1$
BiSe	27.4	*hP*12	$P\bar{3}m1$
Bi_8Se_9	29.8	*hP*17	$R\bar{3}m$
Bi_6Se_7	30.6	*hP*39	$P\bar{3}m1$
Bi_4Se_5	32.1	*hP*27	$P\bar{3}m1$
Bi_3Se_4	33.5	*hP*42	$P\bar{3}m1$
Bi_2Se_3	36	*hR*5	$R\bar{3}m$
(Se)	100	*hP*3	$P3_121$
Metastable phases			
BiSe(b)	27.4	*cF*8	$Fm\bar{3}m$
Bi_2Se_3IIIa	36	*c***	...
High-pressure phases			
Bi_2Se_3II(c)	36	*oP*20	*Pnma*
Bi_2Se_3III	36	*tP*40	$P4_2/nmc$
$BiSe_2$	43.1	...	...

(a) Laitakarite. (b) Thin film. (c) Bismuthite

Figure 1. The phase diagram of Bi-Se system. [1]

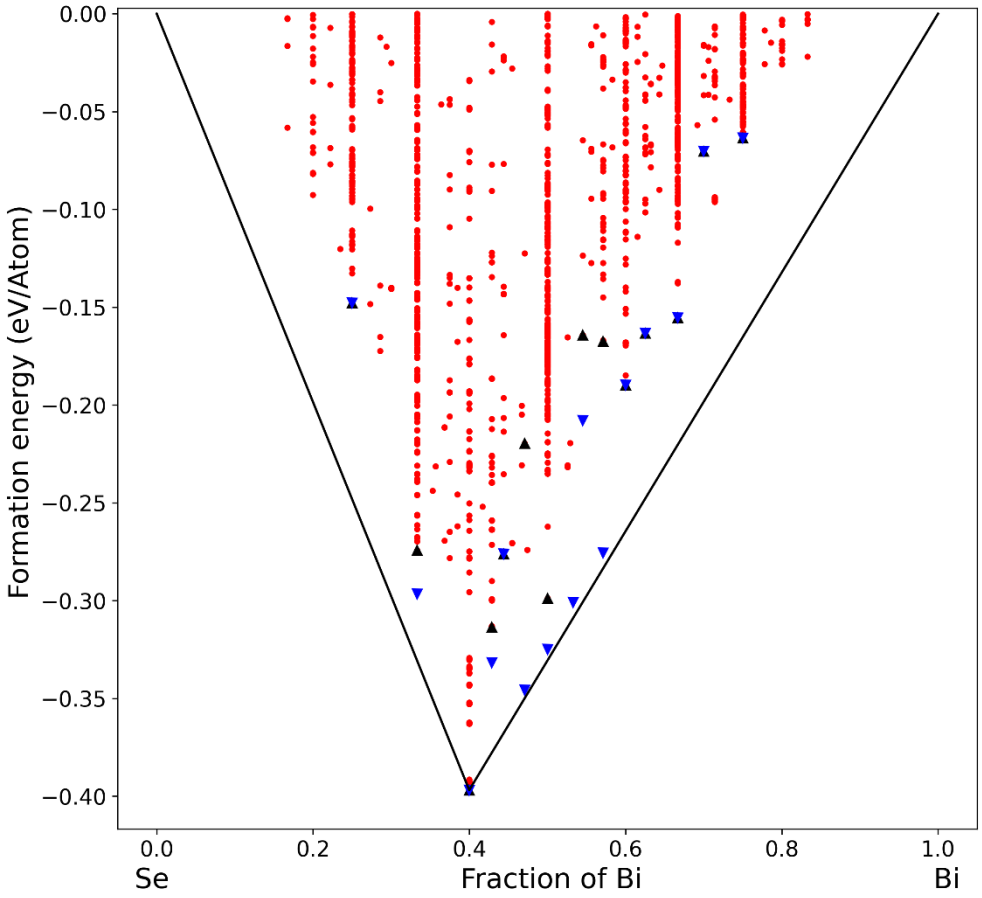


Figure 2. Convex hull of the Bi-Se system, where the black line represents the convex hull, red points denote the crystal structures considered in our machine learning database, blue triangles indicate the experimentally achievable phases, black triangles are the crystal structures of the available Bi-Se phases in the Bi-Se database (as we defined in the manuscript, the Bi-Se database is the database consisting of 9810 Bi-Se structures).

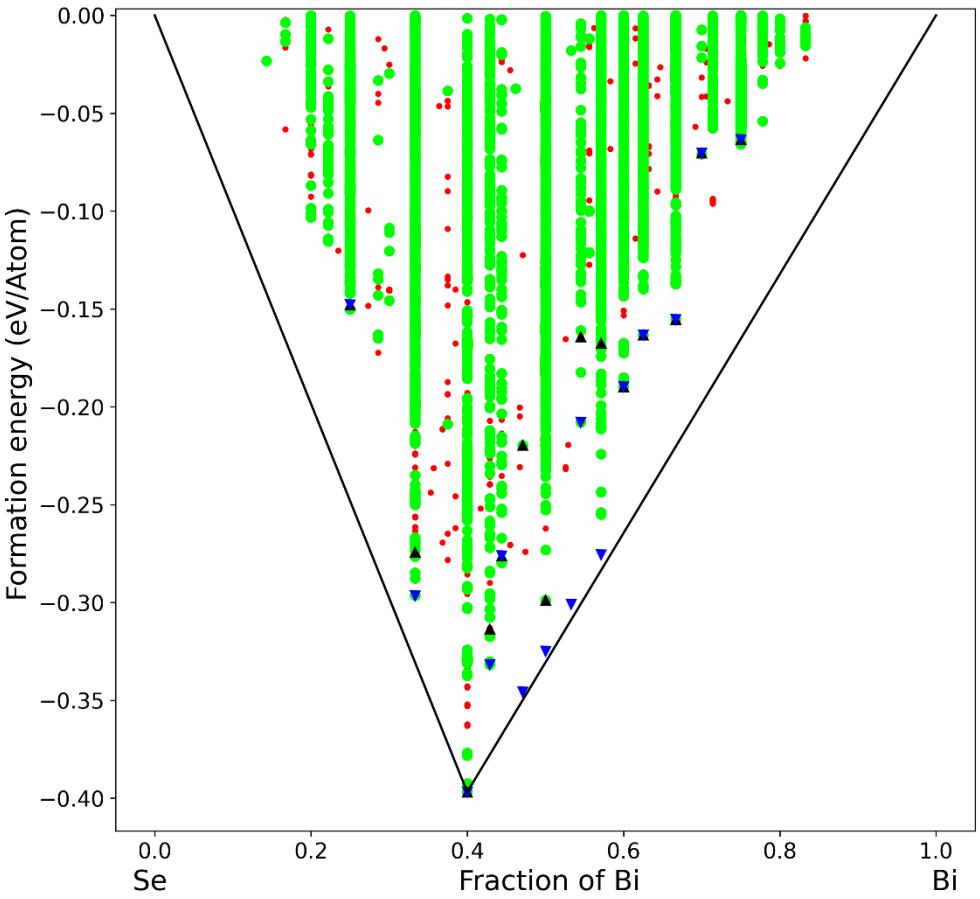


Figure 3. Convex hull of the Bi-Se system including the CCDCGAN generated structures, where the black line represents the convex hull, red points denote the crystal structures considered as training dataset, green circles represent the generated structures, blue triangles indicate the experimentally achievable phases, black triangles are the crystal structures of the available Bi-Se phases in the Bi-Se database.

Figure 3 summarizes the updated results based on our CCDCGAN model, and it is observed that:

- Out of 15 experimentally achievable phases, 8 structures (*i.e.*, $BiSe_3$, Bi_2Se_3, Bi_4Se_5, Bi_3Se_2, Bi_5Se_3, Bi_2Se, Bi_7Se_3, and Bi_3Se) are included in the training database, whereas the other 7 compounds (*i.e.,* $BiSe_2$, Bi_3Se_4, Bi_8Se_9, BiSe, Bi_8Se_7, Bi_6Se_5, and Bi_4Se_3) are not included because of their large unit cell dimensions (larger than 10 Angstrom for at least one lattice constant) or big number of atoms (more than 20 atoms);
- All the 8 experimentally achievable phases can be regenerated using our CCDCGAN model, *i.e.,* exactly the same crystal structures can be reproduced;
- Our CCDCGAN model can also generate three out of seven non-included phases, namely $BiSe_2$, Bi_3Se_4, and Bi_6Se_5, with exactly the same crystal structures as reported. In this regard, the generation step of our CCDCGAN model can go beyond the 10-Angstrom and 20-atom constraints on the unit cell dimension and number of atoms, though only structures fulfilling such constraints are used in training;
- The other four phases, namely, Bi_8Se_9, BiSe, Bi_8Se_7, and Bi_4Se_3, out of seven non-included phases cannot be reproduced. However, for both the Bi_4Se_3 and BiSe compositions, crystal structures with comparable formation energies as the experimentally achievable structures can be obtained. For the

Bi_8Se_9 and Bi_8Se_7 compositions, the generated structures based on the current CCDCGAN model have higher formation energies than those of the experimentally achievable structures. We suspect this can be attributed to the constraints (*i.e.,* unit cell dimension and number of atoms) we applied to select the training database;

- As summarized in Table 1, comparing the structures in the training dataset and those generated using CCDCGAN, it is obvious that crystal structures with lower or equal formation energies can be obtained for all the compositions;
- To further test the predictive power of the CCDCGAN model, we deliberately removed crystal structures corresponding to specific compositions from the training database and could still successfully regenerate the exact crystal structures of $BiSe_3$, Bi_2Se_3, Bi_2Se, and Bi_3Se, as shown in Figure 4. It is noted extended DFT calculations should be done in order to optimize the crystal structures generated by CCDCGAN, thus we have tested only on such four compositions.

Based on the observations above, after incorporating the experimentally metastable phases, our CCDCGAN model is able to reproduce 11 out of 15 experimentally achievable phases. It is believed that after generalizing the model for cases with larger unit cells, its predictive power will be further enhanced.

Table 1. Compilation of the experimentally achievable (EXP) phases and the corresponding generated phases (CCDCGAN) of the binary Bi-Se system. The fifth column shows the formation energy of the structures with corresponding compositions included in the training.

Phase	Composition % of Bi	Formation energy (EXP) (eV/atom)	Distance to the convex hull (EXP) (eV/atom)	Formation energy (training) (eV/atom)	Formation energy (CCDCGAN) (eV/atom)
$BiSe_3$	0.25	-0.15	0.10	-0.15	-0.15
$BiSe_2$	0.333	-0.30	0.03	-0.27	-0.30
Bi_2Se_3	0.4	-0.39	0	-0.39	-0.39
Bi_3Se_4	0.429	-0.33	0.06	-0.31	-0.33
Bi_4Se_5	0.444	-0.28	0.09	-0.28	-0.28
Bi_8Se_9	0.471	-0.35	0.01	-0.22	-0.22
BiSe	0.5	-0.32	0.01	-0.30	-0.30
Bi_8Se_7	0.533	-0.30	0.01	None	-0.02
Bi_6Se_5	0.545	-0.21	0.13	-0.16	-0.21
Bi_4Se_3	0.571	-0.28	0.01	-0.17	-0.26
Bi_3Se_2	0.6	-0.19	0.07	-0.19	-0.19
Bi_5Se_3	0.625	-0.16	0.08	-0.16	-0.16
Bi_2Se	0.667	-0.16	0.08	-0.16	-0.16
Bi_7Se_3	0.7	-0.07	0.13	-0.07	-0.07
Bi_3Se	0.75	-0.07	0.10	-0.07	-0.07

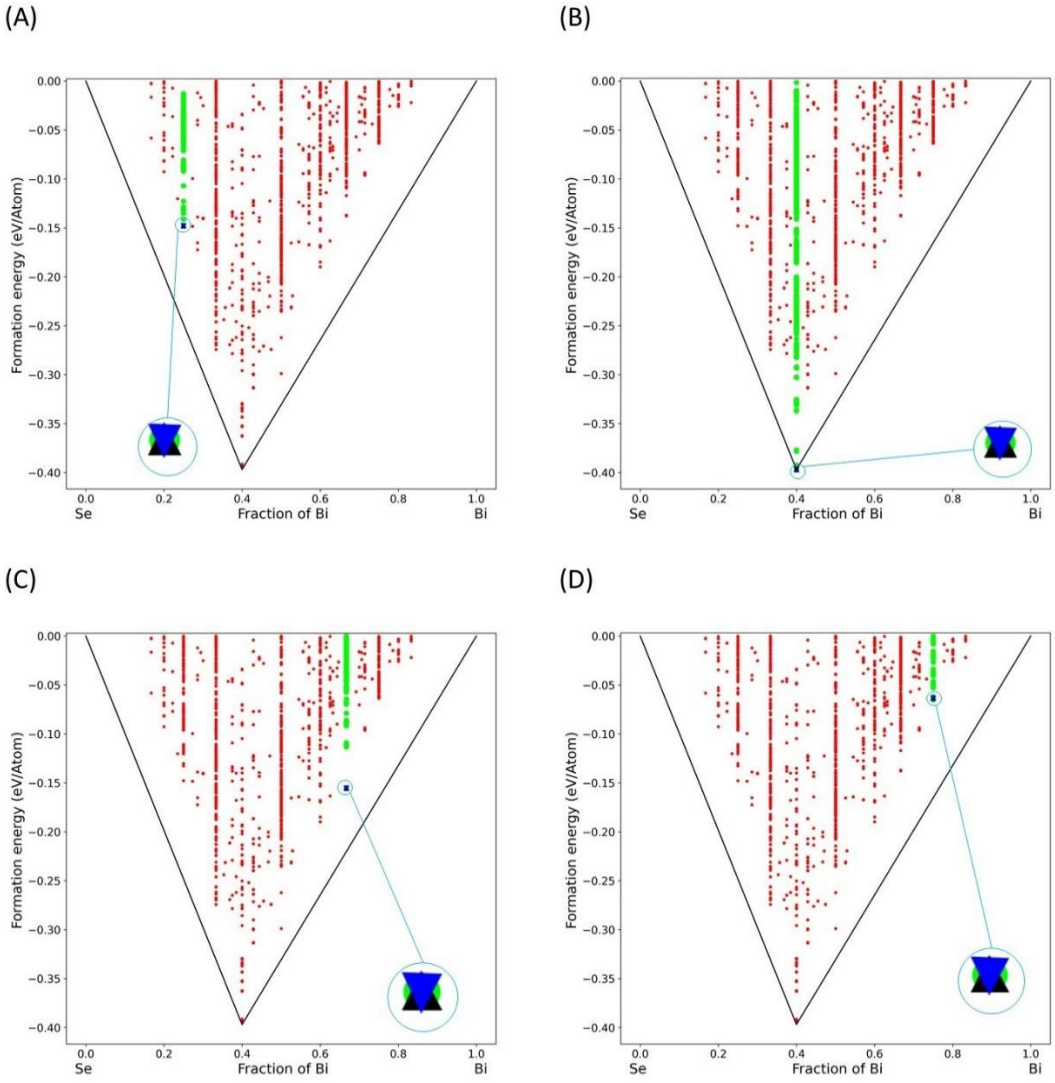


Figure 4. The formation energy of regenerated phases (A) The $BiSe_3$ phases regenerated by CCDCGAN which are trained with the $BiSe_3$ composition removed; (B) The Bi_2Se_3 phases regenerated by CCDCGAN which are trained with the Bi_2Se_3 composition removed; (C) The Bi_2Se phases regenerated by CCDCGAN which are trained with the Bi_2Se composition removed; (D) The Bi_3Se phases regenerated by CCDCGAN which are trained with the Bi_3Se composition removed. The insets show the zoom-in plots about the structures with the lowest energies for such crystal structures.

Question 2:

"(2) In the first paragraph of results, data for generative adversial network, it is claimed that among the ~10,000 compounds artificial structures generated and calculated, only 155 are stable. However the author's definition of "stable" is strange to me: "negative formation energy and distance of 100meV/at to the convex hull".

If the energy is lower than the convex hull, then the convex hull is modified and includes the new compound. This is the case of all ground-state figures presented in this paper with a V-shape indicating the only Bi3Se2 stable.

If the energy is higher, then the compounds are not stable. The definition of stable compounds has to be clarified.
Moreover, the formation energy plot of Fig3 indicates some degeneracies in composition: at a given Bi fraction, there are several compounds. But again, only the lowest is a candidate to be stable in regards to the whole ground state.
Does this definition change has a consequence on their objective function?"

As sketched in Figure 5, in our calculations, the formation energy measures the relative stability of a crystalline phase with respect to decomposition into the elements, *i.e.,* into the Bi and Se elements for binary Bi-Se phases in this work. Additionally, the distance to the convex hull is defined with respect to the decomposition into the available stable phases (mostly not the elements). For instance, the metastable phase marked in Figure 5 has negative formation energy, *i.e.,* it is stable with respect to decomposing into the elements. However, it can decompose into the neighboring stable phases at the thermodynamic equilibrium, as denoted by the arrow for one metastable phase in Figure 5. It is noted that whether the metastable phases can be achieved experimentally does not depend only on the thermodynamics but also the reaction kinetics, *i.e.,* depending also on the experimental set-up. Thus, those phases with distances to the convex hull smaller than a tolerance value can possibly be synthesized experimentally.

This is best demonstrated using the experimentally achievable phases of Bi-Se. According to Figure 2 and Table 1, both Bi_6Se_5 and Bi_7Se_3 are with a distance to the convex hull as large as 130 meV/atom and such phases can be synthesized. Of course, our evaluation of the convex hull is done at 0 K based on DFT, which does not correspond exactly to the experimental conditions at finite temperature. However, it is believed that such results are still valuable in making reasonable predictions. Therefore, in our manuscript, we define the criterion for a phase being stable (*i.e.,* synthesizable) to be "with negative formation energy and distance of 100 meV/atom to the convex hull", and being metastable (i.e., less likely to be synthesizable) to be "with negative formation energy and distance of 150 meV/atom to the convex hull". It is noted that such a criterion is widely used in high-throughput screening for "stable" compounds.[2,3]

We agree with Reviewer that the degeneracies in composition occur and the structure with the lowest formation energy is stable. Nevertheless, we included also such phases with higher formation energies in Fig. 3, intending to show that the CCDCGAN model can optimize the formation energy in the latent space, *i.e.,* generating structures not only at the global minima but also at the local minima. Furthermore, the definition of "stable" does not affect the objective function. The DCGAN model itself does not have an objective function related to the formation energy, and is therefore not affected the definition of stable. In the CCDGAN model, the constraint is to find the lowest possible formation energy, so changing the definition has no effect on the objective function as well.

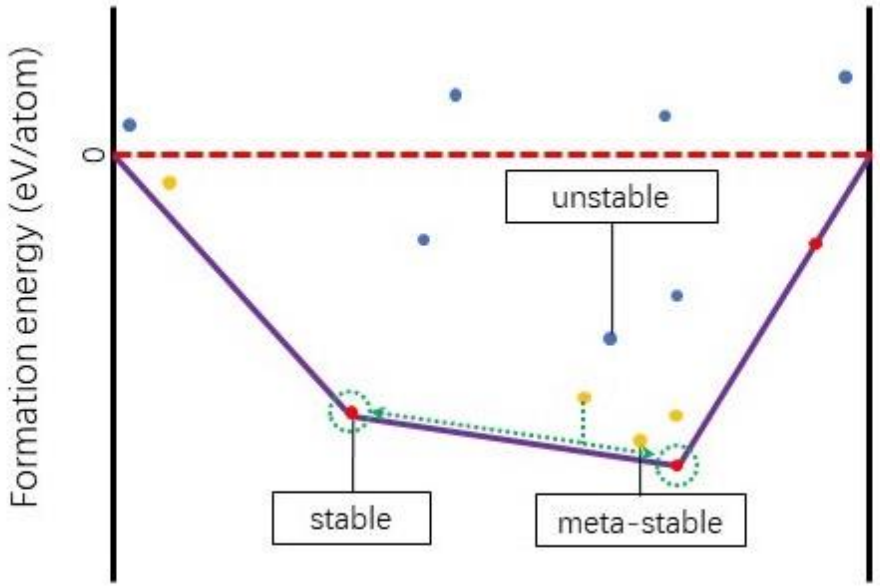


Figure 5. The sketch of formation energy and convex hull, where the y-axis represents the formation energy, and the purple line represnts the convex hull; red circle are stable structures, orange circle are meta-stable structures, and the blue circle are unstable structures.

Question 3:

"(3) My other interrogation is about the DFT calculation. It is not clear what is the set of calculated configurations. In the paper, the authors claim that they use all binary structures from Material Project (10981), but later in the paper, authors chose only structures with interatomic distance criteria, and later, in the methodologies part, they exclude structures with more than 20 atoms…? As an example, they reject some known stable structures in Bi-Se. This is completely inconsistent to me. A clear policy should have been chosen to build the learning database. Instead of a brutal force size criterion, I believe that it is better to include known existing phases and exclude structures with a different chemical interaction nature.

Moreover, It is not clear if the DFT relaxation was done before the training on all structures. If yes, does the relaxation scheme has been done carefully in order to keep the original symmetry?

I believe that authors are not familiar so much with DFT: for example, they choose a k-meshing of 40 points in all directions whatever the structure, this is not a pertinent optimization.

"

We are grateful to the Reviewer for the comments.

For the training database, we have applied consistent selection criteria, but the text may be confusing. Our DFT calculations are performed on the prototype binary structures used in the work of *Noh et al.,*[4] where they selected such structures based on two criteria: (1) structures with no more than 20 atoms in the unit cell and (2) the maximum length of the lattice constants is shorter than 10 Angstrom. Starting from 10981 prototype structures, we managed to converge 9810 cases for Bi-Se systems, which are then used as the training set of our DCGAN models. Therefore, we are following the same selection criteria as established in the work of *Noh et al.,*[4] as described in the *Methods* section. Such selection criteria can be justified based on

the statistics of all the binary compounds in the Materials Project database as shown in Figure 6 (A) & (B), *e.g.,* more than 80% of the binary compounds are with less than 20 atoms per unit cell and about 80% of them are with the maximal lattice constant longer than 10 Angstroms.

We agree with the Reviewer that such selection criteria reject some experimentally achievable Bi-Se phases in the training procedure, as discussed above in detail in the reply to the first question. We also tested to construct machine learning models with such structures considered, and it is observed that the reproduction ratio of the autoencoder dropped significantly to 70% (from 96%) and the generation ratio of CCDCGAN is decreased to less than 0.1%. Such a straightforward test jeopardizes the overall performance of the model. Therefore, we decided to exclude these structures after weighting the advantages and disadvantages, which will be saved for future investigation.

On the other hand, for the generated crystal structures, we have tried to implement several constraints in order to facilitate the generation of reasonable crystal structures. As shown in Figure 6 (B), the minimal interatomic distance of all the selected 9810 structures is found to be larger than 2 Angstrom. Thus, it was believed that the generated structures with interatomic distance smaller than 2 Angstrom are not reasonable, which will not be considered in further DFT calculations in order to save the computational time. We realized recently that this constraint is not realistic, *i.e.*, it removes a large number of stable structures (Please refer to the reply to Question 4 for more detailed discussion). Therefore, in the revised version of the manuscript, we have removed the interatomic distance constraint for all the models and updated all the results.

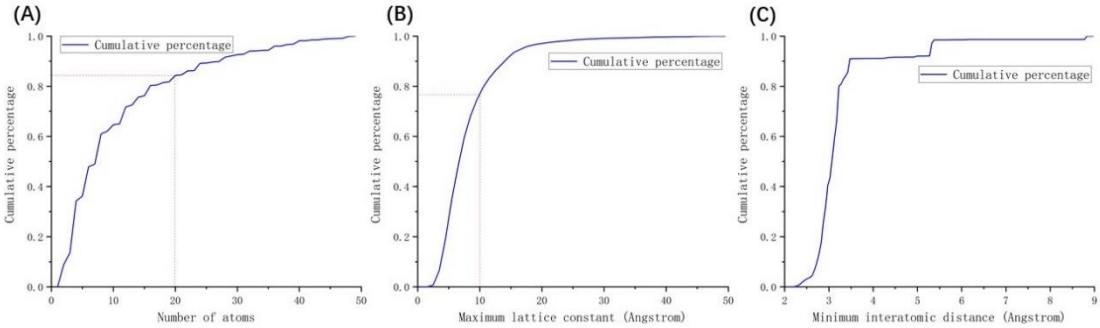


Figure 6. (A) The cumulative percentage of all binaries in Materials Project with respect to the number of atoms in the unit cell, where the x-axis represents the number of atoms in the unit cell, and the y-axis represents the cumulative percentage. **(B)** The cumulative percentage of all binaries in Materials Project with respect to the maximum lattice constant, where the x-axis represents the maximum lattice constant, and the y-axis represents the cumulative percentage. **(C)** The cumulative percentage of Bi-Se system with respect to the minimum interatomic distance, where the x-axis represents the minimum interatomic distance, and the y-axis represents the cumulative percentage.

Lastly, the DFT relaxations are carried out with the symmetries switched on, thus it is guaranteed that the original symmetry is kept. Nevertheless, it is possible that the relaxed structures may end up with a higher symmetry. In addition, the k-mesh density of 40 means that the product of the number of k-points and the corresponding lattice constant in Angstrom is chosen to be 40, following the standard VASP documentation https://www.vasp.at/wiki/index.php/KPOINTS. Based on our experience, such a k-mesh density is good enough to guarantee the convergence of total energies for most compounds.

Question 4:

"Minor points:
(a) The global policy is confusing: it is claimed that the DCGAN developed in this paper generates crystal structures without applying any constraint. However the two scheme comparison, including the CCDCGAN are using structures from the training database with the constrain of a minimum interatomic distance larger than 2 Angstom (before or after the DFT relaxation?). Once again, the authors should clarify the policy of their constrained training set. "

We thank the Reviewer for the comment.

In the previous version, the interatomic distance was applied as a constraint during the generation. This is based on the observation that the interatomic distance of all the 9810 structures in the training database is found to be larger than 2 Angstrom, as shown in Figure 6 (B). Therefore, in order to reduce the number of DFT calculation, we applied the constraint so that the generated structures with minimal interatomic distance smaller than 2 Angstroms will not be passed over to further relaxation for all three models.

In the revised version, inspired by this question and also our observation that the generated crystal structures are not at the equilibria, we removed the constraint on the interatomic distance for all the models in our database, and rechecked the generated structures. Interestingly, the number of stable and metastable structures generated by all three models has significantly increased, indicating that it is not necessary to apply such a constraint. Nevertheless, it is noted that the volume of such directly generated structures with interatomic distance smaller than 2 Angstrom will change significantly during the relaxation, as we observed in the DFT calculations. Correspondingly, we have updated the results for all three models in the revised manuscript, where the main conclusions of our manuscript remain unchanged.

On the other hand, as an essential feature of inverse design (*i.e*., deep learning) where physical properties are supposed to be optimized, formation energy is considered in the current work so that our machine learning models can generate automatically the stable crystal structures. In this regard, we have tried out three different models:

- For the DCGAN model, the generated crystal structures correspond to random vectors in the latent space, *i.e.*, there is no optimization/constraint at all;
- For the DCGAN+constraint model, the generated crystal structures are screened by their formation energies via an additional neural network which models the formation energies;
- For the CCDCGAN model, the optimization of the formation energy is integrated into DCGAN as a back propagator, so that stable crystal structures can be automatically searched in the latent space by optimizing the formation energy.

In this regard, the constraint on formation energy as applied in the DCGAN+constrain and CCDCGAN models should be considered as an optimization procedure.

Question 5:

"(d) The abstract gives the outlooks that the method can be extended to a multicomponent system for multi objective optimization, which is a great promise. However this idea is not developed later in the manuscript."

We thank the Reviewer for this insightful comment. After collecting our results on the Bi-Se binary system as summarized in the current manuscript, we realized that it is straightforward to generalize this method to multicomponent systems for multi-objective optimization.

In terms of multicomponent systems, in a recent work by S. Kim *et al.,* they have successfully applied the crystal graph representation on the ternary Mg-Mn-O system.[5] As a continuation of the current work, we recently have applied the methods on all the compounds in the Materials Project database and have achieved a reproduction ration of 87% as shown in Figure 7. We are now on the way to test the generation of novel crystal structures. Thus, we are convinced that the method can be extended to multicomponent systems.

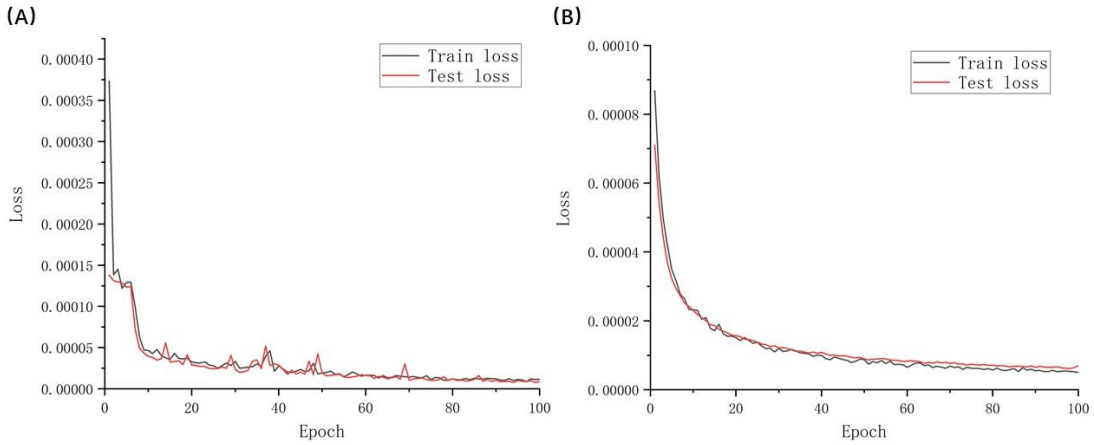


Figure 7. The learning curve of MP database autoencoder, the black line represents training loss, while the red line represents test loss. (A) the learning curve of the lattice autoencoder. (B) the learning curve of site autoencoder.

Regarding the multi-objective optimization, B. Kim *et al.* demonstrated that a ZeoGAN model based on 2D crystal graph representation to optimize the heat of adsorption of methane for the generated structures.[6] We have tried to train neural networks using 2D crystal graphs as descriptors to model the other physical properties such as band gaps and Poisson ratio, with resulting MAE of 0.29 eV and 0.06, respectively (cf. Figure. 8). Such accuracies are comparable with the results of one of the state-of-the-art algorithms CGCNN with MAE for band gaps and Poisson ratio being 0.30eV and 0.05, respectively.[7] Therefore, it is believed that the inverse design models based on 2D crystal graphs can be extended to perform multi-objective optimization as well.

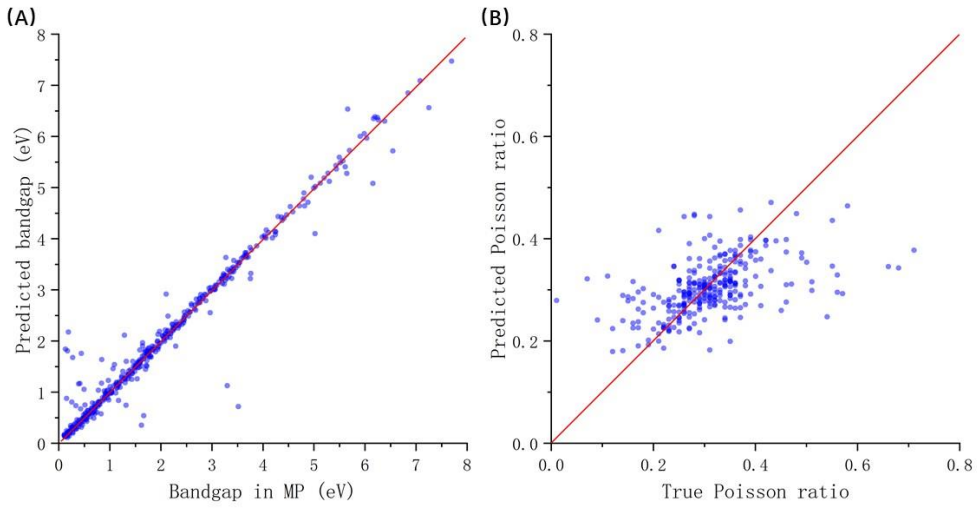


Figure 8. The prediction of band gap and Poisson ratio of compounds in the Materials Project database using 2D crystal graph as descriptors.

Reply to the Reviewer 2:

> *"This work by Long et al reports a GAN model to generate new crystal structures for the BiSe system. To further guide the structure prediction, the authors added a formation energy term in the Loss. Overall, first, I find that the work is a combination of previous works (crystal representation by Bengio et al, standard GAN, and the formation energy regression by Ren et al), and does not offer any distinct novelty within the model. Second, the work does not contain more meaningful applications towards inverse design beyond the formation energy of BiSe. Third, the performance of the model is in question, or at least not well established, with no comparison with baseline models. I think the paper is premature and it would be more suited to more specialized journal such as J Chem Inf Model after addressing some specific comments I can suggest:"*

We thank the Reviewer for his/her comments.

Firstly, we agree with the Reviewer that what we managed to accomplish in the current work is an integrated framework combining the crystal representation [4] and standard GAN, which had not yet been done upon the submission of our manuscript by the others. It is noted that what we used is not standard GAN, it is DCGAN, which solves the gradient degradation of traditional GAN model. Importantly, we demonstrated the optimization of the formation energies in the latent space, either as an explicit constraint or as an integrated back-propagator. To the best of our knowledge, the optimization of formation energy in the latent space of GAN model has not been reported. Such an implementation has been successfully applied on the randomly chosen Bi-Se binary system, which can be straightforwardly generalized to the multicomponent systems and other physical properties (please refer to our reply to Q7 of the first Reviewer). Additionally, our work is parallel to the work of Ren, where variational autoencoder is used. Note also that the preprint of Ren appeared on arXiv firstly on May 15th, 2020, whereas our draft on July 22nd, 2020. Therefore, we believe our work marks a step forward in designing crystal structures based on deep learning.

Secondly, we agree with the Reviewer that the ultimate goal of inverse design is to design functional materials. To this goal, we focus on how to evaluate the stability of the materials first, which is a fundamental property before evaluating the other properties, where the Bi-Se system is picked up to as an illustrative case. We are now very confident that our formalism works for the other binary and multicomponent systems. However, it is noted that in order to prepare the training set for the Bi-Se system, about 10,000 DFT calculations have been done to get the training dataset, which took us one month on a supercomputer. Not to mention the follow-up calculations on the crystal structures generated using our implementation, which are several times more expensive than the calculations on the training dataset.

Lastly, the main working hypothesis for our work is "whether it is possible to generate unreported distinct stable crystal structures based on deep learning", and we believe such a question is well addressed in our manuscript. The work done in our manuscript is based on the crystal graph representation of crystal structures proposed by Noh et al.[4] with significant improvements (*i.e.*, integration of DCGAN and optimization of formation energies). Therefore, the performance of our model is at least comparable with that of imatgen.[4] For instance, they discovered 40 new stable V-O structures out of 3600 generated structures, while we obtained 42 out of 3743. On the other hand, we cannot get access to the source codes of the other inverse

design frameworks, so direct comparison cannot be performed. Last but not least, we agree with the Reviewer that the comparison with crystal structure predictions done using the generative/evolutionary algorithms is an interesting question. However, this is beyond the scope of the current manuscript.

Therefore, we believe that our work represents a step forward towards the goal of inverse design of functional materials, though there are many pending challenges for the whole community. We thank the Reviewer for his/her insightful suggestions/comments, which are addressed point-to-point as detailed below.

Question 1:

> *"- The effectiveness of using the additional loss term (formation energy) as defined in Eq. (2) is not clear. In Fig. S2, there is no significant difference in distribution of the formation energy of the generated data with and without the additional loss term. In addition, the difference of the average formation energy between CCDCGAN and the other two models is only extremely marginal, at best. The average formation energy of CCDCGAN, DCGAN, and DCGAN+constraint are reported to be -0.0895, -0.0788, and -0.0868 eV/atom, respectively, which in my opinion are essentially the same for most practical purposes. Considering numerical variations of periodic DFT calculations, I would even report these numbers as -0.09, -0.08, and -0.09 eV/atom. It is odd that the authors state that one number is "noticeably lower" than the other two in line 200."*

We agree that the average formation energy seems comparable for structures generated using three models, this is caused by the constraint on the interatomic distance as pointed out by Reviewer 1. By removing this unnecessary constraint (which dictates more follow-up DFT calculations though), the average formation energy for CCDCGAN becomes -0.11eV/atom, while this value remains the almost unchanged for DCGAN (-0.08eV/atom) and DCGAN+constraint (-0.09eV/atom), as demonstrated in Figure 9 and Figure 10.

The difference of 0.02 eV/atom is noticeable though it might be insignificant for non-specialists. Nevertheless, based on our observations, it is unrealistic to expect CCDCGAN outperform the other models significantly, due to the following reasons: (1) the accuracy of the forward modeling of formation energy applied as constraint. Despite the accuracy in modeling the formation energy using 2D crystal graphs as descriptors is comparable with the state-of-the-art method *(e.g.*, CGCNN) at Bi-Se database, it is still not adequate, thus both the generation of the CCDCGAN model and the selection of DCGAN+constraint are affected. This is a general challenge which will be saved for future investigation; (2) The generated structures need further relaxation, which will certainly reduce the formation energy of the generated crystal structures.

Therefore, we believe the incorporation of the formation energy loss term is important, at least as a proof-of-concept procedure. As a matter of fact, the generation efficiency of the CCDCGAN model can be significantly enhanced in comparison to the bare DCGAN model. For instance, starting from 13000 directly generated 2D crystal graphs, the CCDCGAN model leads to 3743 reasonable structures, whereas the DCGAN and DCGAN+constraint model leads to only 2832 and 2148 structures after relaxation, respectively.

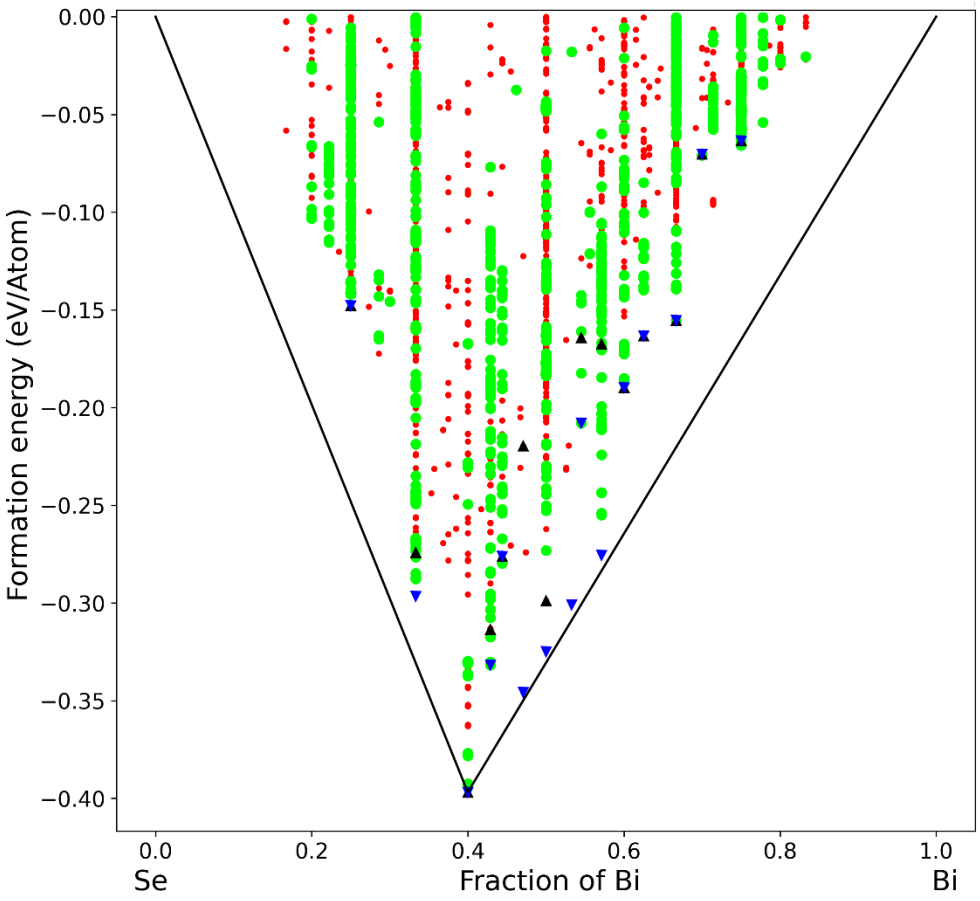


Figure 9. Performance of CCDCGAN, where the black line represents the convex hull, red points denote the crystal structures considered in our machine learning database, green circles represent the generated structures blue triangles indicate the experimentally achievable phases, black triangles are the crystal structures of the available Bi-Se phases in the Bi-Se database.

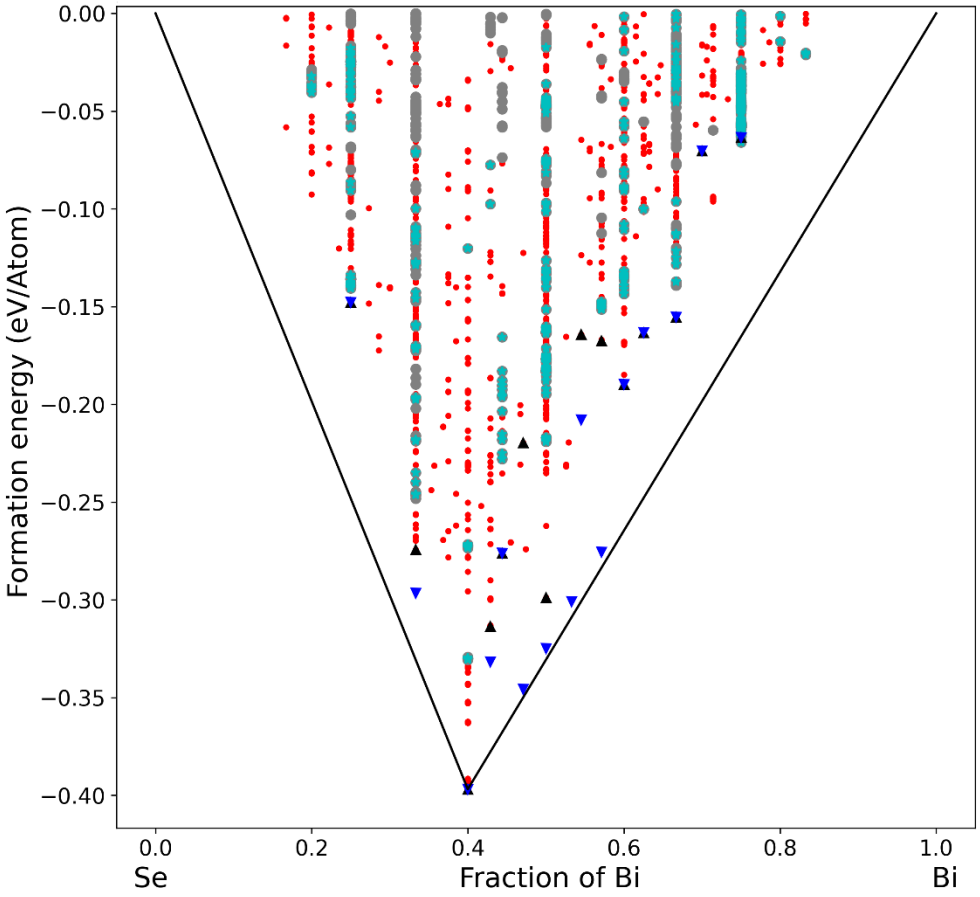


Figure 10. Performance of DCGAN and DCGAN+constraint, where the black line represents the convex hull, red points denote the crystal structures considered in our machine learning database, gray circles represent the generated structures, cyan stars denote the DCGAN+constraint structures, blue triangles indicate the experimentally achievable phases, black triangles are the crystal structures of the available Bi-Se phases in the Bi-Se database.

For the relaxed structures, CCDCGAN demonstrates its advantage in the generation efficiency of stable structures, while the DCGAN+constraint is better at the generation efficiency of metastable structures. For example, only 36/2148 structures generated by DCGAN+constraint is stable, while CCDCGAN has generated 307/3743 stable structures. This again proves that the formation energy has been successfully optimized in the latent space, and therefore improve the efficiency of generation. But the DCGAN+constraint model demonstrates its advantage in generating metastable structures (1145/2148, 53.3%) compared to CCDCGAN (1910/3743, 51%).

In conclusion, the effectiveness of the formation energy loss term is significant. The modest improved performance of CCDCGAN in comparison to bare DCGAN is due to two general challenges as discussed above, requiring further considerations.

Question 2:

"- The author analyzed the generated structures in terms of structural uniqueness and stability compared to the materials existing in the database. However, given the key advantage of generative

models to generate entirely new structures beyond the database, analyses and discussions on the generated structures in comparison with the training set structures would be needed."

We note that all the models in our manuscript (DCGAN, DCGAN+constraint, and CCDCGAN) are trained using the entire database including 9810 crystal structures, thus there is no separation of training and test sets of crystal structures, and consequently we compare the generated structures to all the 9810 input crystal structures in order to identify novel crystal structures. This is consistent with how the machine learning is usually done. For instance, in supervised learning, the performance of the machine learning model is usually validated using the test set. Whereas for unsupervised/generative models, their performance cannot be measure by a specific target property, thus the database is usually not split into training and test sets. Instead, one feeds the generative model as many structures as possible and then analyzes the uniqueness of the generated structures, as we did in the manuscript.

The models based on DCGAN in our manuscript can generate distinct crystal structures in comparison to all the input crystal structures, as measured using the "structure.match" module from pymatgen.

Question 3:

"- There is no comparison with baseline models in literature. Many other theoretical approaches are widely used for crystal structure prediction to design novel functional materials to date. Considering the required DFT calculations to construct generative model and calculate the energy above convex hull, the effictiveness of the model should be established with proper baseline models. Potential baseline models can include those that have source codes available publicly, such as evolutionary algorithm-based crystal structure predictions and existing GAN and VAE-based crystal generative models."

We agree with the Reviewer that there are many theoretical approaches widely used for crystal structure prediction, which can be summarized into three categories, *i.e.*, high-throughput screening, global optimization (including evolutionary algorithm), and generative ML models, as defined in the review paper of *Noh et al.* [8]. High-throughput screening is a conventional method, which consists of elemental substitution on the existing crystal structural prototypes and follow-up DFT calculations. This method has been studied intensively in our group[3,9]. The Evolutionary algorithm is a popular method of global optimization.[10] Both methods have marginal connection with machine learning. In contrast, the generative ML models build a continuous latent space to encode the crystal structures and to generate novel cases,[4,11] which is the focus of our work. *"This approach is a potential solution to the long-sought goal of the community of inverse design,"*[5] *i.e*., automatic design of new structures with desired properties[11]. Therefore, the generative ML models enable inverse design, whereas the methods based on high-throughput calculations and evolutionary algorithm are more like forward modelling.

The performance of generative ML models and evolutionary algorithm has been conducted by *Noh et al*.[4] and *Kim et al.*[5], and their conclusions are that in order to generate the same amount of stable new structures, the generative ML models require 14,000 DFT calculations (include the 11,000 calculation of training data), while the required DFT calculation for evolutionary algorithm is 93,600 (did not include the calculation of training data). [4] Thus, the generative ML models have demonstrated their advantage over the evolutionary algorithm. Although we agree with the Reviewer that a direct comparison of our CCDCGAN with the

evolutionary algorithms is an interesting question, this is beyond the scope of the current manuscript.

The performance of our CCDCGAN model is comparable with that of *Noh et al.* and *Kim et al.*. For instance, Noh *et al*. have discovered 40 new stable structures out of 3600 generated structures, in comparison to 42 out of 3743 based on our CCDCGAN model. Kim *et al*. reported 753 out of 6000 Mg-Mn-O materials as theoretically meta-stable, similar to what we obtained using CCDCGAN (*e.g.,* 1910 out of 3743).

Lastly, we cannot get access to the source codes of the other inverse design frameworks, so direct comparison cannot be performed.

Question 5:

> *"- The introduction should include more discussions on the crystal structure prediction (since this is what the authors are doing essentially). For example, currently, random crystal structure search, Bayesian approach and etc are missing."*

We appreciate this suggestion. As mentioned in the reply to Question 3, there are three crystal structure prediction methods, *i.e.,* high-throughput screening, global optimization, and generative ML models. Global optimization includes but is not limited to random search, Bayesian approach, or the genetic algorithm. For instance, crystal structure prediction with Bayesian optimization can efficiently select the most stable structure from a large number of candidate structures with a lower number of searching trials, and has been successfully applied to predict structures of NaCl and Y_2Co_{17}.[12] The random search method has demonstrated their effectiveness to find global minimum with highly parallel implementation in elemental boron, silicon cluster and lithium hydrides.[13,14] Evolutionary algorithms are also promising in prediction of thermodynamically stable tungsten borides, carbon and TiO_2. [15,16] In contrast, the generative ML methods have been implemented mostly based on variational autoencoder (VAE) and generative adversarial network (GAN), which make use of the latent space to generate and optimize the crystal structures.

Reference:

(1) *ASM Handbook*, 10th editon.; ASM International, Ed.; ASM International: Materials Park, Ohio, 1990.

(2) Sun, W.; Dacek, S. T.; Ong, S. P.; Hautier, G.; Jain, A.; Richards, W. D.; Gamst, A. C.; Persson, K. A.; Ceder, G. The Thermodynamic Scale of Inorganic Crystalline Metastability. *Science Advances* **2016**, *2* (11), e1600225. https://doi.org/10.1126/sciadv.1600225.

(3) Singh, H. K.; Zhang, Z.; Opahle, I.; Ohmer, D.; Yao, Y.; Zhang, H. High-Throughput Screening of Magnetic Antiperovskites. *Chem. Mater.* **2018**, *30* (20), 6983–6991. https://doi.org/10.1021/acs.chemmater.8b01618.

(4) Noh, J.; Kim, J.; Stein, H. S.; Sanchez-Lengeling, B.; Gregoire, J. M.; Aspuru-Guzik, A.; Jung, Y. Inverse Design of Solid-State Materials via a Continuous Representation. *Matter* **2019**, *1* (5), 1370–1384. https://doi.org/10.1016/j.matt.2019.08.017.

(5) Kim, S.; Noh, J.; Gu, G. H.; Aspuru-Guzik, A.; Jung, Y. Generative Adversarial Networks for Crystal Structure Prediction. *ACS Cent. Sci.* **2020**, *6* (8), 1412–1420. https://doi.org/10.1021/acscentsci.0c00426.

(6) Kim, B.; Lee, S.; Kim, J. Inverse Design of Porous Materials Using Artificial Neural Networks. *Science Advances* **2020**, *6* (1), eaax9324. https://doi.org/10.1126/sciadv.aax9324.

(7) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120* (14), 145301. https://doi.org/10.1103/PhysRevLett.120.145301.

(8) Noh, J.; Ho Gu, G.; Kim, S.; Jung, Y. Machine-Enabled Inverse Design of Inorganic Solid Materials: Promises and Challenges. *Chemical Science* **2020**, *11* (19), 4871–4881. https://doi.org/10.1039/D0SC00594K.

(9) Opahle, I.; Madsen, G. K. H.; Drautz, R. High Throughput Density Functional Investigations of the Stability, Electronic Structure and Thermoelectric Properties of Binary Silicides. *Phys. Chem. Chem. Phys.* **2012**, *14* (47), 16197–16202. https://doi.org/10.1039/C2CP41826F.

(10) *Computational Materials Discovery*; 2018. https://doi.org/10.1039/9781788010122.

(11) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361* (6400), 360–365. https://doi.org/10.1126/science.aat2663.

(12) Yamashita, T.; Sato, N.; Kino, H.; Miyake, T.; Tsuda, K.; Oguchi, T. Crystal Structure Prediction Accelerated by Bayesian Optimization. *Phys. Rev. Materials* **2018**, *2* (1), 013803. https://doi.org/10.1103/PhysRevMaterials.2.013803.

(13) Deringer, V. L.; Pickard, C. J.; Csányi, G. Data-Driven Learning of Total and Local Energies in Elemental Boron. *Phys. Rev. Lett.* **2018**, *120* (15), 156001. https://doi.org/10.1103/PhysRevLett.120.156001.

(14) Pickard, C. J.; Needs, R. J. Ab Initio Random Structure Searching. *J Phys Condens Matter* **2011**, *23* (5), 053201. https://doi.org/10.1088/0953-8984/23/5/053201.

(15) Li, Q.; Zhou, D.; Zheng, W.; Ma, Y.; Chen, C. Global Structural Optimization of Tungsten Borides. *Phys. Rev. Lett.* **2013**, *110* (13), 136403. https://doi.org/10.1103/PhysRevLett.110.136403.

(16) Lyakhov, A. O.; Oganov, A. R. Evolutionary Search for Superhard Materials: Methodology and Applications to Forms of Carbon and TiO${}_{2}$. *Phys. Rev. B* **2011**, *84* (9), 092103. https://doi.org/10.1103/PhysRevB.84.092103.

B. Unpublished work on the inverse design of multicomponent systems

Inverse design of crystal structures for multicomponent systems

Teng Long*, Yixuan Zhang, Nuno M. Fortunato, Chen Shen, Mian Dai, Hongbin Zhang*

**Correspondence to Teng Long and Hongbin Zhang*

E-mail: tenglong@tmm.tu-darmstadt.de/ hongbin.zhang@tu-darmstadt.de

Institute of Materials Science, Technical University of Darmstadt, Otto-Berndt-Straße 3, 64287 Darmstadt, Germany

We developed an inverse design framework enabling automated generation of stable multi-component crystal structures by optimizing the formation energies in the latent space based on reversible crystal graphs with continuous representation. It is demonstrated that 9,160 crystal structures can be generated out of 50,000 crystal graphs, leading to 8,310 distinct cases using a training set of 52,615 crystal structures from Materials Project. Detailed analysis on 15 selected systems reveals that unreported crystal structures below the convex hull can be discovered in 6 material systems. Moreover, the generation efficiency can be further improved by considering extra hypothetical structures in the training. This paves the way to perform inverse design of multicomponent materials with possible multi-objective optimization.

Keywords: inverse design; machine learning; generative adversarial network; crystal structures;

Introduction

Data-driven materials design has developed rapidly in the last decade, commencing with launching of the Materials Genome Initiative [1,2], with the goal to engineer materials with desired properties (dubbed as inverse design hereafter) [3]. There are three widely applied approaches to achieve inverse design of crystalline materials, *i.e.,* high-throughput calculations, global optimizations, and machine learning generative models [4,5]. The high-throughput approach relies on massive density functional theory (DFT) calculations to evaluate the physical properties from a large number of hypothetical structures [6,7]. This method has been successfully applied to identify Li-ion battery anode materials [8], magnetic materials [9], and topological insulators [10,11]. Meanwhile, the global optimization method generalizes the characteristics of materials as an objective function and then predicts the crystal structures (rather than physical properties) by finding the optimal solutions [12]. This approach is best exemplified by the genetic algorithm as implemented in USPEX and CALYPSO [13,14], which has been used to design superconductors [15], superhard materials [16], and magnetic materials [17]. To better utilize the existing data available from databases such as Materials Project (MP) [18], OQMD [6,19], NOMAD [20], and AFLOWlib [21], the machine learning generative models [4] can be divided into two categories: the probabilistic machine learning model and deep learning generative model. The former one predicts physical properties based on the crystal structures with uncertainty quantification, and uses an

acquisition function to select potential structures with the desired properties, which has found structures of NaCl and Y_2Co_{17} with limited training data [22,23]. While the latter one relies on a properly defined continuous latent space (explained in details in the following paragraph) which enables optimization of the physical properties, which has been successfully applied on V-O, Mg-Mg-O and Bi-Se systems [24–26].

Focusing on the deep learning generative model, it is essential to construct a continuous latent space. As first demonstrated by Gómez-Bombarelli *et al.* in the inverse design of molecules [27], properties of the designed molecules can be efficiently optimized in the latent space. Similarly, in order to design crystal structures, a latent space consisting of crystal graphs is first established by Xie *et al.* [28] and further improved by Park *et al.* [29], where the crystal graphs can be used as descriptors to optimize the physical properties such as formation energies, band gaps, and so on. However, without the ability to reconstruct crystal structures, such crystal graphs can only conduct forward prediction but not inverse design. Thereafter, with the refinement of continuous representation (voxel and autoencoder) [24,30–32], crystal graphs are constructed to be reversible, which serve as one-to-one mappings of the crystal structures (via encoding and decoding). Therefore, the resulting continuous latent space enables the inverse design of crystalline structures.

The generation model is another key component of a deep learning generative model, which can be best exemplified by variational autoencoder (VAE) [33] and generative adversarial network (GAN) [34]. VAE is a variant of autoencoder, which assumes a distribution function (*e.g.*, Gaussian) of the crystal graphs in the latent space [35]. For instance, Noh *et al.* developed the iMatGen model, and used it to generate novel V_xO_y crystals [24]. Ren *et al.* also used VAE to design inorganic crystals and find novel structures that do not exist in the training set [36]. Court *et al.* generated binary alloys, ternary perovskites, and Heusler compounds by a VAE model [37]. In contrast, the GAN model trains two competing neural networks (*i.e.*, a generator and a discriminator), where the generator is responsible for generating crystal graphs (aiming to minimize their difference with existing crystal graphs), while the discriminator is responsible for distinguishing them (maximize the difference) [26]. This model does not require prior assumptions of a distribution function to start training, which makes it a more attractive method [34]. For example, Nouira *et al.* developed a CrystalGAN model to design ternary A-B-H phases starting from binary A-H and B-H structures, and applied it to the Ni-Pd-H system [38]. The ZeoGAN scheme developed by Kim *et al.* was applied on zeolites, where Wasserstein GAN (WGAN) was used to generate porous materials [39]. Kim *et al.* also developed the Crystal-WGAN to design new structures in the Mn-Mg-O system [25]. Furthermore, our constrained crystals deep convolutional generative adversarial network (CCDCGAN) model has been applied successfully to design novel thermodynamically stable structures in the Bi-Se system [26].

In this work, we generalized CCDCGAN to the multicomponent systems covering all

compositions from the whole periodic table. As shown in Fig. 1, one salient feature of our CCDCGAN is to integrate a formation energy prediction model in latent space as a back propagator for the generator [26]. After training on 52,615 structures in the Materials Project database (please refer to the methods for details), the CCDCGAN model has produced 9,160 crystal structures (of which 8310 are distinct, meaning unreported in the training set) out of 50,000 crystal graphs for 667 systems (meaning element combinations, *e.g.*, Cd-Li, Co-Li-O, etc.) covering 71 elements in the periodic table. Detailed analysis on 15 selected systems indicates that 72% of the newly designed crystal structures have negative formation energies after DFT relaxation, and unreported crystal structures below the convex hulls can be identified for 6 systems. Taking the Co-Hf system as an example, we further investigated how to improve the generation efficiency by including additional hypothetical structures in the training. We believe such an implementation can be further generalized to optimize other physical properties beyond the formation energy [28], so that multi-objective optimization can be achieved.

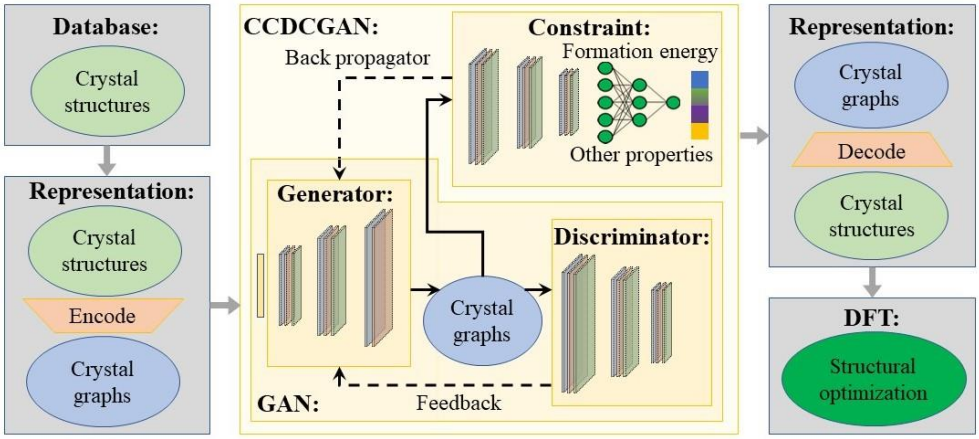


Fig. 1. Workflow of CCDCGAN. Three key elements of CCDCGAN, i.e., generator, discriminator, and constraint, are highlighted by shaded blocks in yellow. Such models act on the crystal graphs, which are one-to-one mapping to the crystal structures, with the transformations managed by the autoencoder, *i.e.*, encoder and decoder.

Methods

Data

Crystal structures and corresponding formation energies from the MP database are used for training in this work, which amount to 125,617 cases (accessed on September 8th 2020). Following Ref. [17] and [20], compounds with less than 20 atoms in primitive cell and with none lattice constant exceeding 10 Å are selected as the training set, leading to 52,615 compounds comprising 84 elements in the periodic table and 25,246 systems. It is straightforward to construct generalized crystal graphs to include compounds with more than 20 atoms or larger dimensions, but the crystal reconstruction ratio will drop significantly because considering larger cells will make it impossible to distinguish subtle crystal structure differences. It is also noted that all the 52,615 compounds are included in the training set for our generative model, which are validated by comparing with the

generated structures. Such training set is referred as MP database from here on out.

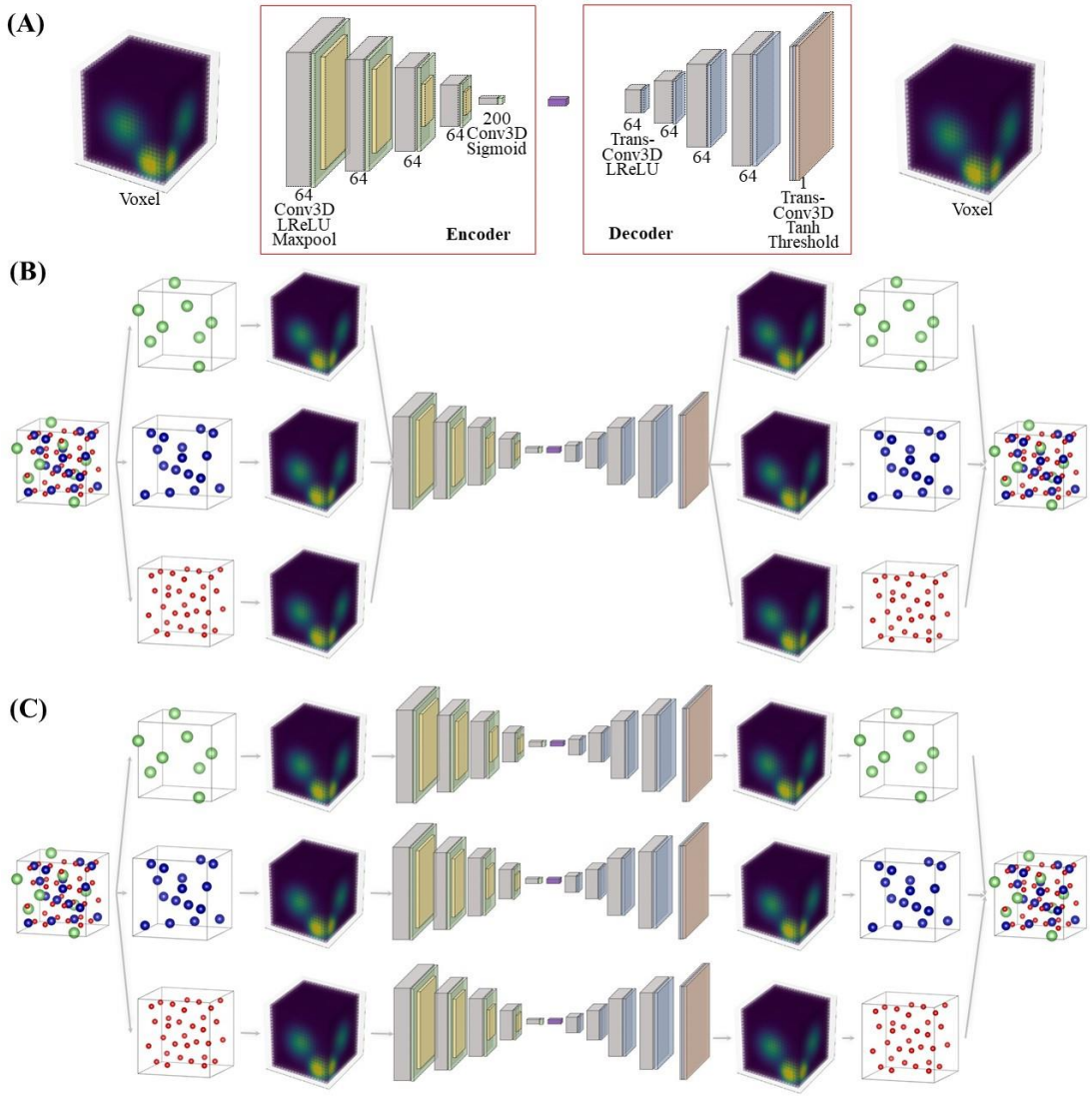


Fig. 2. (A) Voxel autoencoder, the autoencoder takes the voxel of crystal structures as input and encodes it into vector in latent space, then the decoder restores the voxel from the encoded vector. (B) Workflow of combined model, different elements share the same autoencoder. (C) workflow of separated model, each element has its own autoencoder.

Representation

The crystal graphs are 160×160 matrices, which serve as one-to-one mappings of the crystal structures in the latent space. As shown in Fig. 2(A), such crystal graphs are obtained by encoding 64×64×64 (32×32×32) voxel images for atomic positions (lattice constants) through a 3D convolutional neural network (CNN) typed autoencoder. Importantly, the encoding transformation process can be reversed, *i.e.*, 3D crystal

structures can be obtained by decoding the crystal graphs. Correspondingly, the reproduction ratio is an essential indicator on the quality of the crystal graphs. Consistent with Ref. [24] and [26], it is observed that the reproduction ratio can be higher than 90% if only one specific system is considered with sufficient (around 10,000) training structures. However, the encoding-decoding procedure corresponds to Fig. 2(B) can only reach a reproduction ratio of 70% when taking the MP database as input [30]. This can be attributed to the overlapping of different atomic species in the generated crystal graphs. We found that separating the voxels and autoencoders for different elements (as shown in Fig. 2(C)) can reduce the reconstruction loss (Fig. 3(E)) and increase the reproduction ratio to 87%. This leads to reasonable crystal graphs for a large number of multicomponent systems.

GAN

To generate new crystal graphs and hence unreported crystal structures, the GAN model is used to sample the latent space, where both the generator and discriminator adopt a CNN model, as shown in Fig. 3 (A) and (B). The generator takes a randomly generated 800-dimensional array obeying the Gaussian distribution as input and generate a 160×160 crystal graph, while the discriminator takes the crystal graph as input and identifies its source. From the mathematical point of view, the goal of the generator (discriminator) is to minimize (maximize) the statistical difference between generated crystal graphs and original crystal structures, which can be expressed for the GAN model as:

$$\max_{D}\left(\min_{G}\left(\frac{1}{2}\cdot E_{\mathbf{x}\sim p_t}\left[1-D(\mathbf{x})\right]+\frac{1}{2}\cdot E_{\mathbf{x}\sim p_g}\left[D(\mathbf{x})\right]\right)\right) \quad (1)$$

where D is the discriminator, G is the generator, E means the expectation value, $\mathbf{x}$ represents the crystal graphs, $D(\mathbf{x})$ is the output of discriminator, *i.e.*, close to 1 (0) means higher probability to be a generated (original) crystal graph, p_t is the possibility density function of the crystal graphs in MP database, while p_g is the possibility density function of generated crystal graphs. In this way, driven by the competition between the generator and discriminator, the GAN model will be competent to design reasonable crystal graphs and thus crystal structures.

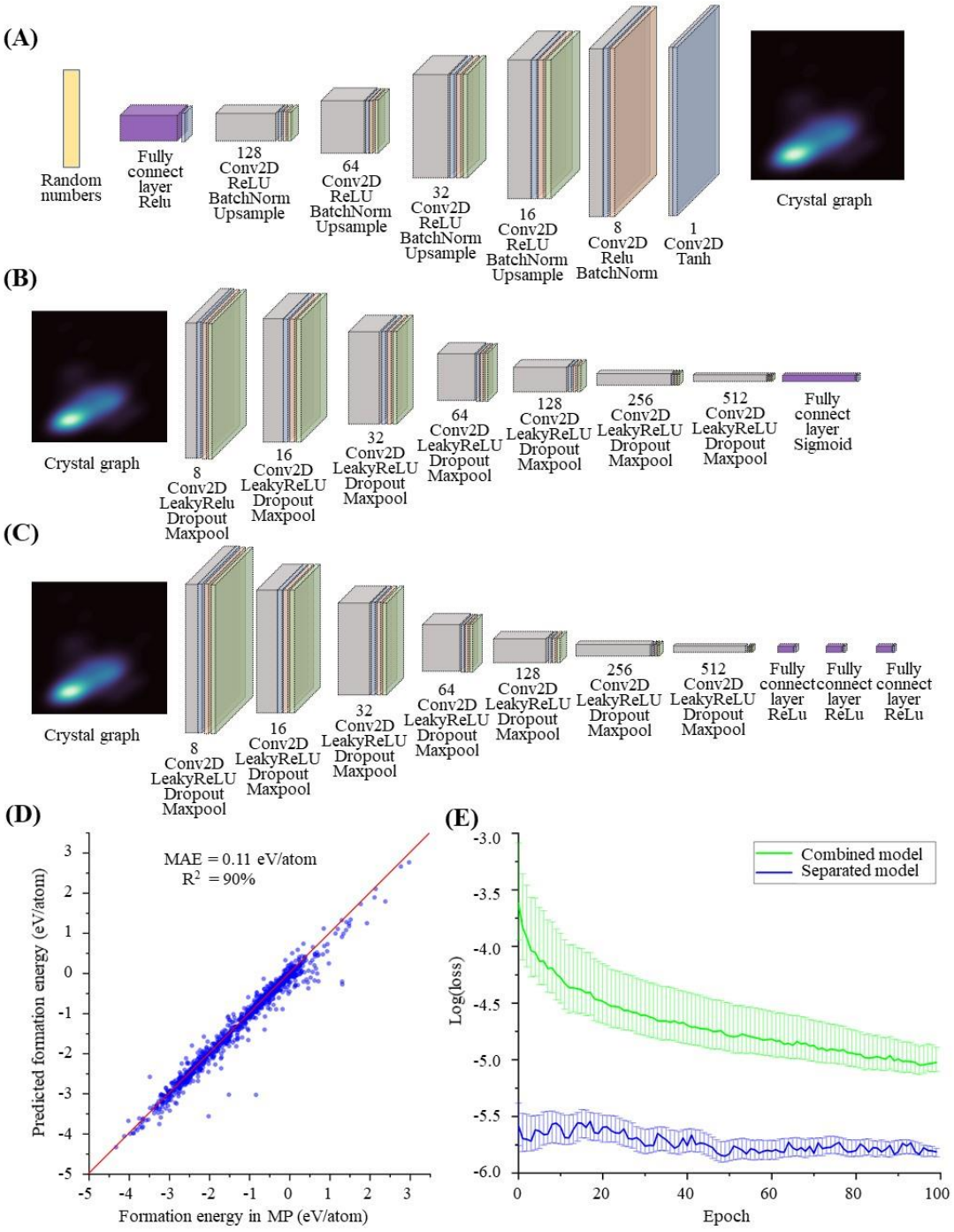


Fig. 3. (A) Generator, the generator takes random number as input and generate crystal graph. (B) Discriminator, the discriminator takes the crystal graph as input and judge whether it is generated. (C) Constraint model, the constraint model takes the crystal graph as the input and predict the corresponding properties, *e.g.*, formation energy. (D) Formation energy prediction evaluation by our constraint model. The x-axis is the formation energy in MP database, the y-axis is the formation energy predicted by constraint model. (E) The learning curve of MP database autoencoder, the green line represents test loss of combined model, while the blue line denotes loss of separated model, error bar is the Q_1 and Q_3 point.

Constraint

In order to design crystal structures with desired properties, constraints can be applied to perform optimization in the latent space. In this work, we consider formation energy as the target property so that stable crystal structures can be obtained. Correspondingly, such a constraint can be formulated as a CNN similar to the discriminator, *i.e.*, it takes the crystal graphs as input and predicts their formation energies, as Fig.3(C) demonstrates. For instance, based on the MP database which is divided into a training/validation/test sets following the 80%/10%/10% splitting, the accuracy of predicting formation energies using the crystal graphs as descriptors reaches 90%, as measured by the coefficient of determination (R^2), with a corresponding mean absolute error (MAE) of 0.11eV/atom (Fig. 3(D)), which is comparable with the results of CGCNN (Fig. S1)[28].

CCDCGAN

Such a constraint on the formation energies has been implemented in our CCDCGAN model as a back propagator coupled to the GAN model, as shown in Fig. 1. That is, the constraint model works also as a feedback mechanism for the generator like the discriminator, but remains unchanged during the CCDCGAN training process, which ensures only the adjustment of generator. It can be mathematically expressed as:

$$\min_{G}\left(\frac{1}{2}\cdot E_{\mathbf{z}\sim p_t}\left[D(\mathbf{z})\right]+\omega\cdot e^{E_f(\mathbf{z})}\right),\mathbf{z}\sim p_g \tag{2}$$

where E_f is the formation energy predicted by the constraint model, $\mathbf{z}$ is the generated crystal graph, and ω is defined as the weight of formation energy loss, which is 0.1 in this model. It is noted that more constraints on the other properties can be integrated so that our CCDCGAN can perform multi-objective optimization, which will be saved for future studies.

Results and discussion

Based on our CCDCGAN model trained using the MP database with the formation energies as the constraint, 9,160 crystal structures can be obtained out of 50,000 randomly generated crystal graphs. Such structures comprise 71 elements (Fig. 4(A)), giving rise to 667 systems. And 8,310 structures among them are distinct, indicating variety in structure type. Defining the elemental density as the number of generated structures containing the element divided by the total number of generated structures, the elemental densities of the generated structures are shown in Fig. 4(B), where the elemental densities of the generated structures fluctuated around the density of existed structures, suggesting a similar distribution of elements. Thus, the CCDCGAN model is able to generate a wide range of crystal structures.

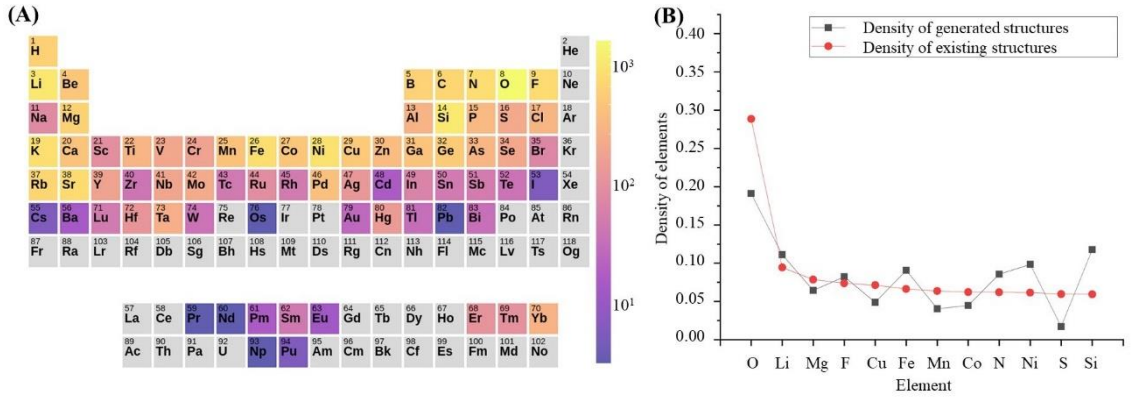


Fig. 4. (A) Element distribution of 9,810 generated structures. The Color represents the number of compounds containing the elements, and the color bar is demonstrated in log format. Gray means that this element is not covered by the database. (B) Density of elements of generated structures and existed structures from MP database. The black square represents the density of the generated structures, while the red circle represents the density of the existed structures.

To verify the quality of generated structures, detailed analysis is done on 15 systems focusing on intermetallic compounds (with reasons specified below), including Cr-Ta, Mn-Ta, Fe-Ta, Co-Ta, Ni-Ta, Co-Hf, Al-Cr-Mn, Co-Mn-Si, Cd-Li, Cr-Se, Mn-Se, Fe-Se, Co-Se, Ni-Se, and Co-Li-O, as listed in Table 1. Using the change of formation energies (ΔE_f) of the generated structures before and after DFT optimization as an indicator, it is observed that the average ΔE_f for 493 generated structures is 0.10 eV/atom, with the largest average ΔE_f being 0.47 eV/atom for the Cd-Li system. In contrast, ΔE_f ranges between 0.57 eV/atom and 3.34 eV/atom for structures designed based on the VAE model [37]. In this regard, CCDCGAN is promising in generating nearly optimized crystal structures. Moreover, it is observed that 355 out of 493 (72%) the generated compounds exhibit negative formation energies. That is, compounds in 9 systems have an average formation energy less than zero while those of the other 6 systems are smaller than 70 meV/atom (Table 1). This leads to an average energy above convex hull (referring to the convex hull defined by existing phases) for 6 systems less than 0.1 eV/atom, indicating such generated crystal structures can be synthesized experimentally.

Furthermore, it is observed that distinct structures below the convex hulls can be obtained for 6 out of 15 systems. For instance, the Cd-Li system is an interesting case, where not only all three experimentally known phases ($CdLi_3$, CdLi and Cd_3Li) [40] can be reproduced but also two structures (*i.e.*, Cd_3Li_4 with a C/2m space group and Cd_2Li with Immm) are predicted, as shown in Fig. 5(A). For other 5 systems, *i.e.*, Mn-Se, Mn-Ta, Ni-Ta, Co-Li-O and Al-Cr-Mn, CCDCGAN generates structures which redefines the convex hull, despite it cannot reproduced all the experimentally known phases. As shown in Fig. 5(B-F), $MnSe_2$ (formation energy -0.47eV/atom, C/2m) and Ni_2Ta (formation energy -0.35eV/atom, I4/mmm) are stable phases that exist experimentally, but are not yet

included in the MP database, thus demonstrating the predictive power of CCDCGAN. Additionally, Mn_4Se_5 (formation energy -0.41eV/atom, P1), $MnTa_2$ (formation energy -0.17eV/atom, I4/mmm), $AlCr_3Mn$ (formation energy -0.12eV/atom, R3m), Al_2Cr_3Mn (formation energy -0.18eV/atom, Amm2), AlCrMn (formation energy -0.22eV/atom, I4mm), and $AlCr_2Mn_3$ (formation energy -0.11eV/atom, P-3m1) are possible novel stable phases which have not been experimentally reported. Last but not least, two new Co-Li-O phases, *i.e.*, Co_7LiO_{12} (formation energy -1.40eV/atom, P1) and $CoLi_3O_7$ (formation energy -1.32eV/atom, P1), are predicted.

Table 1. Statistical analysis on the generated crystal structures for the 15 selected systems. The second column is number of generated structures (# Gen), while the third column is their average formation energy difference before and after DFT optimization (Mean ΔE_f). The fourth column is the number of structures with $\Delta E_f < 0.2$ eV/atom (# $\Delta E_f < 0.2$) with parenthesis denoting the ratio between # Gen and # $\Delta E_f < 0.2$. The last two column are the average formation energy (Mean E_f) and average energy above convex hull (Mean E_{ab}), respectively.

Systems	# Gen	Mena ΔE_f (eV/atom)	# $\Delta E_f < 0.2$	Mean E_f (eV/atom)	Mean E_{ab} (eV/atom)
Cd-Li	69	0.47	44 (63.8%)	-0.21	0.002
Cr-Se	11	0.02	11 (100%)	0.06	0.31
Mn-Se	11	0.03	11 (100%)	-0.11	0.09
Fe-Se	36	0.25	34 (94.4%)	-0.06	0.18
Co-Se	101	0.14	91 (90.1%)	-0.05	0.15
Ni-Se	36	0.12	31 (86.1%)	-0.07	0.14
Cr-Ta	4	0.13	4 (100%)	0.07	0.12
Mn-Ta	8	0.09	7 (87.5%)	0.02	0.15
Fe-Ta	15	0.05	15 (100%)	0.01	0.11
Co-Ta	32	0.06	23 (71.9%)	0.001	0.10
Ni-Ta	11	0.04	11 (100%)	-0.07	0.12
Al-Cr-Mn	9	0.09	8 (88.9%)	-0.04	0.10
Co-Li-O	88	0.09	82 (93.2%)	-1.18	0.07
Co-Mn-Si	11	0.15	10 (90.9%)	-0.07	0.19
Co-Hf	51	0.18	49 (96.1%)	-0.05	0.07

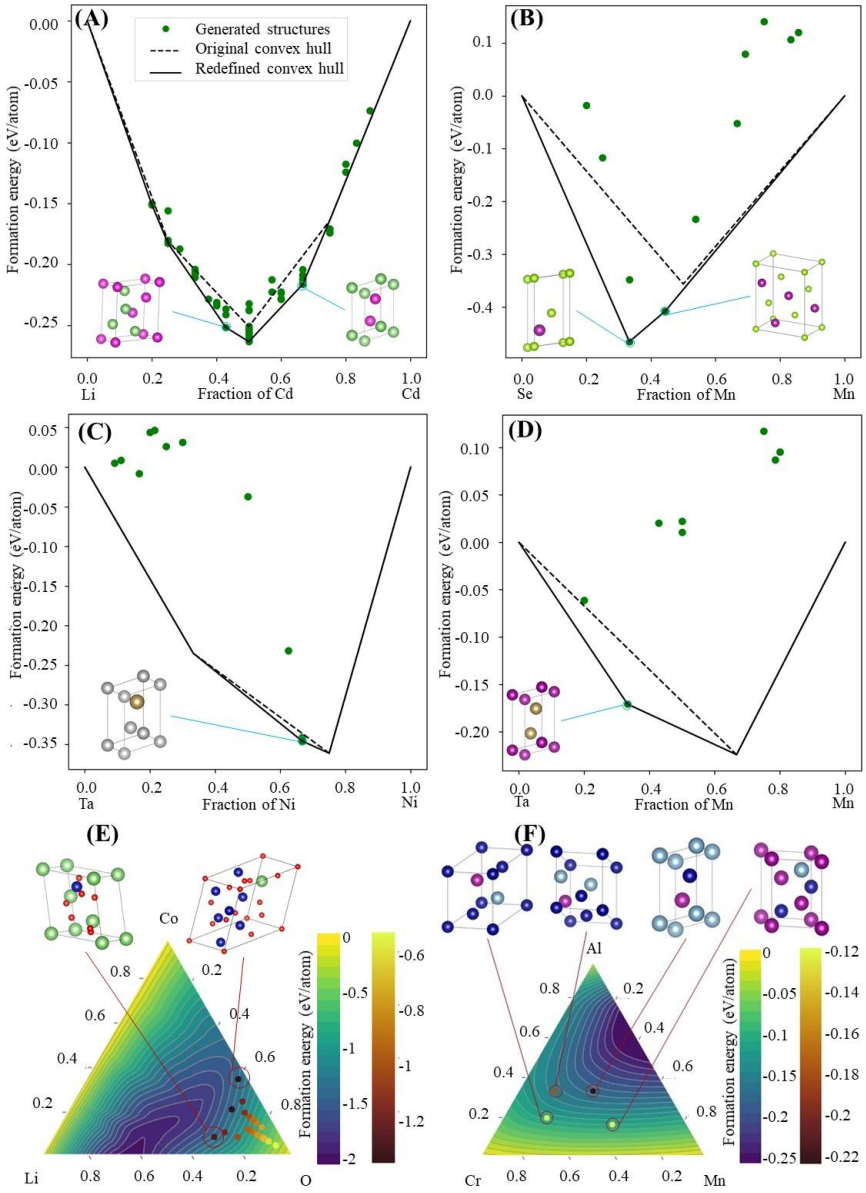


Fig. 5. (A) Convex hull of the Cd-Li system, where the red circles denote the structures generated by DCGAN model, green circles indicate the structures generated by CCDCGAN model, dash line and solid line represent the convex hull of MP database and redefined convex hull respectively; crystal structures of Cd_3Li_4 and Cd_2Li are demonstrated as well, where green indicates Li atom and magenta denotes Cd atom. (B) Convex hull of the Mn-Se system and crystal structure of $MnSe_2$ and Mn_4Se_5, where green represents Se atom and purple means Mn atom. (C) Convex hull of the Ni-Ta system and structure of Ni_2Ta, where grey indicates Ni atom and yellow denotes Ta atom. (D) Convex hull of Mn-Ta system and structure of $MnTa_2$. (E) Convex hull of Co-Li-O system, the left one is for the original convex hull and the right one is the redefined convex hull, the color bar represents formation energy; and structures of Co_7LiO_{12}, $CoLi_3O_7$, red indicates O atom, blue denotes Co atom. (F) Convex hull of Al-Cr-Mn system, where the cold color triangle represents the convex hull and the warm color circles denote the generated structures, values of the color bar are the corresponding formation energy; the crystal structures of $AlCr_3Mn$, Al_2Cr_3Mn, $AlCrMn$, and $AlCr_2Mn_3$ is demonstrated as well, where grey denotes Al atom, blue means Cr atom, magenta indicates Mn atom.

It is noted that for the transition metal oxides, proper calculations should be done based on the DFT+U method to obtain reasonable electronic structure and hence the total energies, due to the strongly correlated nature of the partially filled d-shells. However, the transferability of the U values between compounds with different compositions is limited, which causes possible error in the high-throughput DFT calculations. Besides, additional correction terms are required to evaluate the formation energies and convex hulls of such correlated transition metal oxides with respect to the competing phases [41]. Such a challenge resides in DFT calculations, which is also applicable for rare-earth based systems. Nevertheless, it is beyond the scope of this work, and we suspect our implementation of CCDCGAN is best applicable for predicting novel intermetallic compounds and oxides where mixing DFT+U and DFT calculations is required.

A pending problem is whether CCDCGAN is able to reproduce the experimentally known structures and predict novel structures that redefine the convex hull for the rest 9 systems. Taking the Co-Hf system as an example, all the 51 crystal structures obtained from 50,000 generated cases are above the convex hull, except the CoHf phase is reproduced, as shown in Fig. 6(A). Nevertheless, when the number of generated Co-Hf structures is increased to 2,743, all three known structures (*i.e.*, $CoHf_2$, CoHf, Co_2Hf) [40] are reproduced, including the $CoHf_2$ phase (formation energy -0.31 eV/atom, Fd-3m) which is not included in the training set. In addition, a Co_2Hf_5 (formation energy -0.27 eV/atom, P-1) phase is obtained which lies below the convex hull (Fig. 6(B)). This suggests that CCDCGAN is capable of reproducing the known phases and predicting new crystal structures. Moreover, in order to improve the generation efficiency, we performed explicit DFT calculations on 10,912 hypothetical Co-Hf crystal structures as we did for the Bi-Se system [26], and used these structures as the training set. Such a model can reproduce all three experimental structures and predict Co_2Hf_5 phase again after generating only 714 structures, as shown in Fig. 6(C), *i.e.*, 4 times more efficient. Therefore, we believe that increasing the number of compounds in the training set (by taking other database into consideration) will improve the generation efficiency of CCDCGAN significantly.

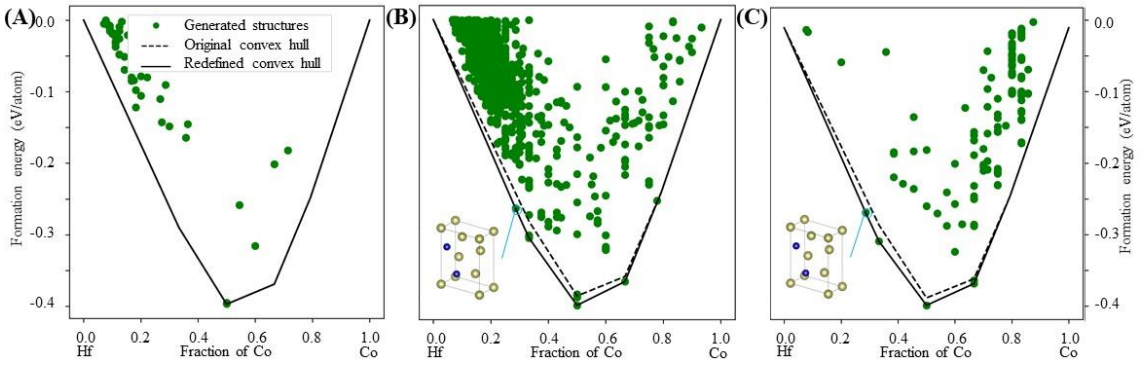


Fig. 6. (A)Convex hull of Co-Hf system and the 51 generated structures. (B) Convex hull of Co-Hf system and the 2743 generated structures; the crystal structure of Co_2Hf_5, blue means Co atom, yellow is the Hf atom. (C) Convex hull of Co-Hf system and the generated 714 structures after including 10,912 Co-Hf structures in the training set.

Conclusions

To summarize, we have developed an inverse design framework CCDCGAN to design novel stable crystalline materials by optimizing the formation energies in the latent space defined by crystal graphs, and applied it successfully on multicomponent systems coving most elements in the periodic table. Trained by 52,615 structures in the MP database, a reproduction ratio of as high as 87% is achieved by extending the separated continuous representation model. Correspondingly, CCDCGAN can be used to design distinct crystal structures over a substantial phase space with the generated crystal structures close to their thermodynamic equilibria. Additionally, it is demonstrated that its generation efficiency can be further improved by enlarging the training set by including hypothetical structures. Thus, combined with massive DFT calculations, the CCDCGAN model enables designing novel crystal structures efficiently and hence accelerating the discovery of new materials. In particular, it is speculated that the other physical properties can be optimized in the latent space within the same framework, which will unlock multi-objective optimization to achieve the goal of designing functional materials with optimal performance.

Acknowledgements

The authors gratefully acknowledge computational time on the Lichtenberg High Performance Supercomputer. Teng Long thanks the financial support from the China Scholarship Council (CSC). Yixuan Zhang thanks the financial support from Fulbright-Cottrell Foundation. Part of this work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant No. 743116-project Cool Innov). This work was also supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 405553726 – TRR 270.

Reference

[1] J.J. de Pablo, N.E. Jackson, M.A. Webb, L.-Q. Chen, J.E. Moore, D. Morgan, R. Jacobs, T. Pollock, D.G. Schlom, E.S. Toberer, J. Analytis, I. Dabo, D.M. DeLongchamp, G.A. Fiete, G.M. Grason, G. Hautier, Y. Mo, K. Rajan, E.J. Reed, E. Rodriguez, V. Stevanovic, J. Suntivich, K. Thornton, J.-C. Zhao, New frontiers for the materials genome initiative, Npj Computational Materials. 5 (2019) 1–23. https://doi.org/10.1038/s41524-019-0173-4.

[2] Cabinet-level National Science and Technology Council, Materials Genome Initiative for Global Competitiveness, (2011).

[3] B. Sanchez-Lengeling, A. Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, Science. 361 (2018) 360–365. https://doi.org/10.1126/science.aat2663.

[4] J. Noh, G. Ho Gu, S. Kim, Y. Jung, Machine-enabled inverse design of inorganic

solid materials: promises and challenges, Chemical Science. 11 (2020) 4871–4881. https://doi.org/10.1039/D0SC00594K.
[5] J.M. Rickman, T. Lookman, S.V. Kalinin, Materials informatics: From the atomic-level to the continuum, Acta Materialia. 168 (2019) 473–510. https://doi.org/10.1016/j.actamat.2019.01.051.
[6] J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD), JOM. 65 (2013) 1501–1509. https://doi.org/10.1007/s11837-013-0755-4.
[7] A. Jain, G. Hautier, C.J. Moore, S. Ping Ong, C.C. Fischer, T. Mueller, K.A. Persson, G. Ceder, A high-throughput infrastructure for density functional theory calculations, Computational Materials Science. 50 (2011) 2295–2310. https://doi.org/10.1016/j.commatsci.2011.02.023.
[8] S. Kirklin, B. Meredig, C. Wolverton, High-Throughput Computational Screening of New Li-Ion Battery Anode Materials, Advanced Energy Materials. 3 (2013) 252–262. https://doi.org/10.1002/aenm.201200593.
[9] H. Zhang, High-throughput Design of Magnetic Materials, Electron. Struct. (2020). https://doi.org/10.1088/2516-1075/abbb25.
[10] M.G. Vergniory, L. Elcoro, C. Felser, N. Regnault, B.A. Bernevig, Z. Wang, A complete catalogue of high-quality topological materials, Nature. 566 (2019) 480–485. https://doi.org/10.1038/s41586-019-0954-4.
[11] Z. Zhang, R.-W. Zhang, X. Li, K. Koepernik, Y. Yao, H. Zhang, High-Throughput Screening and Automated Processing toward Novel Topological Insulators, J. Phys. Chem. Lett. 9 (2018) 6224–6231. https://doi.org/10.1021/acs.jpclett.8b02800.
[12] A.R. Oganov, C.J. Pickard, Q. Zhu, R.J. Needs, Structure prediction drives materials discovery, Nature Reviews Materials. 4 (2019) 331–348. https://doi.org/10.1038/s41578-019-0101-8.
[13] Y. Wang, J. Lv, L. Zhu, Y. Ma, CALYPSO: A method for crystal structure prediction, Computer Physics Communications. 183 (2012) 2063–2070. https://doi.org/10.1016/j.cpc.2012.05.008.
[14] C.W. Glass, A.R. Oganov, N. Hansen, USPEX—Evolutionary crystal structure prediction, Computer Physics Communications. 175 (2006) 713–720. https://doi.org/10.1016/j.cpc.2006.07.020.
[15] Uranium polyhydrides at moderate pressures: Prediction, synthesis, and expected superconductivity | Science Advances, (n.d.). https://advances.sciencemag.org/content/4/10/eaat9776?utm_source=TrendMD&utm_medium=cpc&utm_campaign=TrendMD_1 (accessed March 31, 2021).
[16] H. Wang, Y. Wang, J. Lv, Q. Li, L. Zhang, Y. Ma, CALYPSO structure prediction method and its wide application, Computational Materials Science. 112 (2016) 406–415. https://doi.org/10.1016/j.commatsci.2015.09.037.
[17] Z. Allahyari, A.R. Oganov, Coevolutionary search for optimal materials in the space of all possible compounds, Npj Computational Materials. 6 (2020) 1–10. https://doi.org/10.1038/s41524-020-0322-9.

[18] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, APL Materials. 1 (2013) 011002. https://doi.org/10.1063/1.4812323.

[19] S. Kirklin, J.E. Saal, B. Meredig, A. Thompson, J.W. Doak, M. Aykol, S. Rühl, C. Wolverton, The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies, Npj Computational Materials. 1 (2015) 1–15. https://doi.org/10.1038/npjcompumats.2015.10.

[20] C. Draxl, M. Scheffler, NOMAD: The FAIR concept for big data-driven materials science, MRS Bulletin. 43 (2018) 676–682. https://doi.org/10.1557/mrs.2018.208.

[21] S. Curtarolo, W. Setyawan, G.L.W. Hart, M. Jahnatek, R.V. Chepulskii, R.H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M.J. Mehl, H.T. Stokes, D.O. Demchenko, D. Morgan, AFLOW: An automatic framework for high-throughput materials discovery, Computational Materials Science. 58 (2012) 218–226. https://doi.org/10.1016/j.commatsci.2012.02.005.

[22] T. Yamashita, N. Sato, H. Kino, T. Miyake, K. Tsuda, T. Oguchi, Crystal structure prediction accelerated by Bayesian optimization, Phys. Rev. Materials. 2 (2018) 013803. https://doi.org/10.1103/PhysRevMaterials.2.013803.

[23] Y. Zhang, D.W. Apley, W. Chen, Bayesian Optimization for Materials Design with Mixed Quantitative and Qualitative Variables, Scientific Reports. 10 (2020) 4924. https://doi.org/10.1038/s41598-020-60652-9.

[24] J. Noh, J. Kim, H.S. Stein, B. Sanchez-Lengeling, J.M. Gregoire, A. Aspuru-Guzik, Y. Jung, Inverse Design of Solid-State Materials via a Continuous Representation, Matter. 1 (2019) 1370–1384. https://doi.org/10.1016/j.matt.2019.08.017.

[25] S. Kim, J. Noh, G.H. Gu, A. Aspuru-Guzik, Y. Jung, Generative Adversarial Networks for Crystal Structure Prediction, ACS Cent. Sci. 6 (2020) 1412–1420. https://doi.org/10.1021/acscentsci.0c00426.

[26] T. Long, N.M. Fortunato, I. Opahle, Y. Zhang, I. Samathrakis, C. Shen, O. Gutfleisch, H. Zhang, CCDCGAN: Inverse design of crystal structures, ArXiv:2007.11228 [Cond-Mat, Physics:Physics]. (2020). http://arxiv.org/abs/2007.11228 (accessed October 10, 2020).

[27] R. Gómez-Bombarelli, J.N. Wei, D. Duvenaud, J.M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T.D. Hirzel, R.P. Adams, A. Aspuru-Guzik, Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules, ACS Cent. Sci. 4 (2018) 268–276. https://doi.org/10.1021/acscentsci.7b00572.

[28] T. Xie, J.C. Grossman, Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties, Phys. Rev. Lett. 120 (2018) 145301. https://doi.org/10.1103/PhysRevLett.120.145301.

[29] J. Schmidt, M.R.G. Marques, S. Botti, M.A.L. Marques, Recent advances and applications of machine learning in solid-state materials science, Npj Computational Materials. 5 (2019) 1–36. https://doi.org/10.1038/s41524-019-0221-0.

[30] J. Hoffmann, L. Maestrati, Y. Sawada, J. Tang, J.M. Sellier, Y. Bengio, Data-Driven

Approach to Encoding and Decoding 3-D Crystal Structures, ArXiv:1909.00949 [Cond-Mat, Physics:Physics, Stat]. (2019). http://arxiv.org/abs/1909.00949 (accessed January 18, 2021).

[31] S. De, A. P. Bartók, G. Csányi, M. Ceriotti, Comparing molecules and solids across structural and alchemical space, Physical Chemistry Chemical Physics. 18 (2016) 13754–13769. https://doi.org/10.1039/C6CP00415F.

[32] K. Kaufmann, C. Zhu, A.S. Rosengarten, D. Maryanovsky, T.J. Harrington, E. Marin, K.S. Vecchio, Paradigm shift in electron-based crystallography via machine learning, ArXiv:1902.03682 [Cond-Mat]. (2019). http://arxiv.org/abs/1902.03682 (accessed June 4, 2019).

[33] D.P. Kingma, M. Welling, Auto-Encoding Variational Bayes, ArXiv:1312.6114 [Cs, Stat]. (2014). http://arxiv.org/abs/1312.6114 (accessed March 31, 2021).

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Nets, in: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 27, Curran Associates, Inc., 2014: pp. 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf (accessed April 6, 2020).

[35] D. Schwalbe-Koda, R. Gómez-Bombarelli, Generative Models for Automatic Chemical Design, ArXiv:1907.01632 [Physics, Stat]. (2019). http://arxiv.org/abs/1907.01632 (accessed September 5, 2019).

[36] Z. Ren, J. Noh, S. Tian, F. Oviedo, G. Xing, Q. Liang, A. Aberle, Y. Liu, Q. Li, S. Jayavelu, K. Hippalgaonkar, Y. Jung, T. Buonassisi, Inverse design of crystals using generalized invertible crystallographic representation, ArXiv:2005.07609 [Cond-Mat, Physics:Physics]. (2020). http://arxiv.org/abs/2005.07609 (accessed June 25, 2020).

[37] C.J. Court, B. Yildirim, A. Jain, J.M. Cole, 3-D Inorganic Crystal Structure Generation and Property Prediction via Representation Learning, J. Chem. Inf. Model. 60 (2020) 4518–4535. https://doi.org/10.1021/acs.jcim.0c00464.

[38] A. Nouira, N. Sokolovska, J.-C. Crivello, CrystalGAN: Learning to Discover Crystallographic Structures with Generative Adversarial Networks, ArXiv:1810.11203 [Cs, Stat]. (2018). http://arxiv.org/abs/1810.11203 (accessed September 9, 2019).

[39] B. Kim, S. Lee, J. Kim, Inverse design of porous materials using artificial neural networks, Science Advances. 6 (2020) eaax9324. https://doi.org/10.1126/sciadv.aax9324.

[40] ASM International, ed., ASM handbook, 10th editon, ASM International, Materials Park, Ohio, 1990.

[41] A. Jain, G. Hautier, S.P. Ong, C.J. Moore, C.C. Fischer, K.A. Persson, G. Ceder, Formation enthalpies by mixing GGA and GGA $+$ U calculations, Phys. Rev. B. 84 (2011) 045115. https://doi.org/10.1103/PhysRevB.84.045115.

C. Statement of personal contribution in publications

C.1. An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature

Statement of personal contribution

An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature.
Long, T., Fortunato, N. M., Zhang, Y., Gutfleisch, O., & Zhang, H.
Materials Research Letters, 2021, 9(4), 169-174.

The idea was originated in the discussion of Prof. Gutfleisch and Prof. Zhang. The research was designed by Teng Long under the guidance of Prof. Zhang. The machine learning models are trained by Teng Long. Teng Long, N.M. Fortunato, Y. Zhang and Prof. Zhang collected the crystal structures for machine learning. Teng Long, N. M. Fortunato and Prof. Zhang analyzed the results. All authors contributed the writing of the thesis.

The first author Teng Long recognized the contribution of Teng Long in "Long, T., Fortunato, N. M., Zhang, Y., Gutfleisch, O., & Zhang, H. (2021). An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature. *Materials Research Letters*, *9*(4), 169-174."

Signature: Teng Long

30.06.2021

Statement of personal contribution

An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature.
Long, T., Fortunato, N. M., Zhang, Y., Gutfleisch, O., & Zhang, H.
Materials Research Letters, 2021, 9(4), 169-174.

The idea was originated in the discussion of Prof. Gutfleisch and Prof. Zhang. The research was designed by Teng Long under the guidance of Prof. Zhang. The machine learning models are trained by Teng Long. Teng Long, N.M. Fortunato, Y. Zhang and Prof. Zhang collected the crystal structures for machine learning. Teng Long, N. M. Fortunato and Prof. Zhang analyzed the results. All authors contributed the writing of the thesis.

Coauthor Nuno M. Fortunato recognized the contribution of Teng Long in "Long, T., Fortunato, N. M., Zhang, Y., Gutfleisch, O., & Zhang, H. (2021). An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature. *Materials Research Letters*, *9*(4), 169-174."

Signature: Nuno Fortunato

07.07.2021

Statement of personal contribution

An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature.
Long, T., Fortunato, N. M., Zhang, Y., Gutfleisch, O., & Zhang, H.
Materials Research Letters, 2021, 9(4), 169-174.

The idea was originated in the discussion of Prof. Gutfleisch and Prof. Zhang. The research was designed by Teng Long under the guidance of Prof. Zhang. The machine learning models are trained by Teng Long. Teng Long, N.M. Fortunato, Y. Zhang and Prof. Zhang collected the crystal structures for machine learning. Teng Long, N. M. Fortunato and Prof. Zhang analyzed the results. All authors contributed the writing of the thesis.

Coauthor Yixuan Zhang recognized the contribution of Teng Long in "Long, T., Fortunato, N. M., Zhang, Y., Gutfleisch, O., & Zhang, H. (2021). An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature. *Materials Research Letters*, *9*(4), 169-174."

Signature: 

30.06.2021

Statement of personal contribution

An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature.
Long, T., Fortunato, N. M., Zhang, Y., Gutfleisch, O., & Zhang, H.
Materials Research Letters, 2021, 9(4), 169-174.

The idea was originated in the discussion of Prof. Gutfleisch and Prof. Zhang. The research was designed by Teng Long under the guidance of Prof. Zhang. The machine learning models are trained by Teng Long. Teng Long, N.M. Fortunato, Y. Zhang and Prof. Zhang collected the crystal structures for machine learning. Teng Long, N. M. Fortunato and Prof. Zhang analyzed the results. All authors contributed the writing of the ~~thesis~~ paper.

Coauthor Prof. Oliver Gutfleisch recognized the contribution of Teng Long in "Long, T., Fortunato, N. M., Zhang, Y., Gutfleisch, O., & Zhang, H. (2021). An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature. *Materials Research Letters*, *9*(4), 169-174."

Signature:

Prof. Oliver Gutfleisch
Technische Universität Darmstadt
Materialwissenschaft, FB 11
FG Funktionale Materialien
Alarich-Weiss-Str. 16
D - 64287 Darmstadt

Statement of personal contribution

An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature.
Long, T., Fortunato, N. M., Zhang, Y., Gutfleisch, O., & Zhang, H.
Materials Research Letters, 2021, 9(4), 169-174.

The idea was originated in the discussion of Prof. Gutfleisch and Prof. Zhang. The research was designed by Teng Long under the guidance of Prof. Zhang. The machine learning models are trained by Teng Long. Teng Long, N.M. Fortunato, Y. Zhang and Prof. Zhang collected the crystal structures for machine learning. Teng Long, N. M. Fortunato and Prof. Zhang analyzed the results. All authors contributed the writing of the thesis.

Coauthor Prof. Hongbin Zhang recognized the contribution of Teng Long in "Long, T., Fortunato, N. M., Zhang, Y., Gutfleisch, O., & Zhang, H. (2021). An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature. *Materials Research Letters*, *9*(4), 169-174."

Signature: 

05.07.2021

C.2. Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures

Statement of personal contribution

Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures
Long, T., Fortunato, N. M., Opahle, I., Zhang, Y., Samathrakis, I., Shen, C., Gutfleisch, O., & Zhang, H.
2021, *npj Computational Materials*, *7*(1), 1-7.

This work originated from the discussion of Teng Long, Y. Zhang and Prof. H. Zhang. Prof. H. Zhang and Prof. O. Gutfleisch supervised the research. Teng Long worked on the machine learning model. Teng Long, N. M. Fortunato, I. Opahle, and C. Shen worked on the DFT calculations. Teng Long, I. Samathrakis, and Y. Zhang worked on data analysis. All authors contributed in the writing.

The first author Teng Long recognized the contribution of Teng Long in "Long, T., Fortunato, N. M., Opahle, I., Zhang, Y., Samathrakis, I., Shen, C., Gutfleisch, O., & Zhang, H. (2021). Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures. *npj Computational Materials*, *7*(1), 1-7."

Signature: 

30.06.2021

Statement of personal contribution

Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures
Long, T., Fortunato, N. M., Opahle, I., Zhang, Y., Samathrakis, I., Shen, C., Gutfleisch, O., & Zhang, H.
2021, *npj Computational Materials*, *7*(1), 1-7.

This work originated from the discussion of Teng Long, Y. Zhang and Prof. H. Zhang. Prof. H. Zhang and Prof. O. Gutfleisch supervised the research. Teng Long worked on the machine learning model. Teng Long, N. M. Fortunato, I. Opahle, and C. Shen worked on the DFT calculations. Teng Long, I. Samathrakis, and Y. Zhang worked on data analysis. All authors contributed in the writing.

Coauthor Nuno M. Fortunato recognized the contribution of Teng Long in "Long, T., Fortunato, N. M., Opahle, I., Zhang, Y., Samathrakis, I., Shen, C., Gutfleisch, O., & Zhang, H. (2021). Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures. *npj Computational Materials*, *7*(1), 1-7."

Signature: Nuno Fortunato

07.07.2021

Statement of personal contribution

Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures
Long, T., Fortunato, N. M., Opahle, I., Zhang, Y., Samathrakis, I., Shen, C., Gutfleisch, O., & Zhang, H.
2021, *npj Computational Materials*, *7*(1), 1-7.

This work originated from the discussion of Teng Long, Y. Zhang and Prof. H. Zhang. Prof. H. Zhang and Prof. O. Gutfleisch supervised the research. Teng Long worked on the machine learning model. Teng Long, N. M. Fortunato, I. Opahle, and C. Shen worked on the DFT calculations. Teng Long, I. Samathrakis, and Y. Zhang worked on data analysis. All authors contributed in the writing.

Coauthor Ingo Opahle recognized the contribution of Teng Long in "Long, T., Fortunato, N. M., Opahle, I., Zhang, Y., Samathrakis, I., Shen, C., Gutfleisch, O., & Zhang, H. (2021). Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures. *npj Computational Materials*, *7*(1), 1-7."

Signature:

02.07.2021

Statement of personal contribution

Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures
Long, T., Fortunato, N. M., Opahle, I., Zhang, Y., Samathrakis, I., Shen, C., Gutfleisch, O., & Zhang, H.
2021, *npj Computational Materials*, *7*(1), 1-7.

This work originated from the discussion of Teng Long, Y. Zhang and Prof. H. Zhang. Prof. H. Zhang and Prof. O. Gutfleisch supervised the research. Teng Long worked on the machine learning model. Teng Long, N. M. Fortunato, I. Opahle, and C. Shen worked on the DFT calculations. Teng Long, I. Samathrakis, and Y. Zhang worked on data analysis. All authors contributed in the writing.

Coauthor Yixuan Zhang recognized the contribution of Teng Long in "Long, T., Fortunato, N. M., Opahle, I., Zhang, Y., Samathrakis, I., Shen, C., Gutfleisch, O., & Zhang, H. (2021). Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures. *npj Computational Materials*, *7*(1), 1-7."

Signature: 

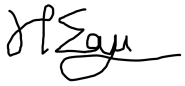
30.06.2021

Statement of personal contribution

Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures
Long, T., Fortunato, N. M., Opahle, I., Zhang, Y., Samathrakis, I., Shen, C., Gutfleisch, O., & Zhang, H.
2021, *npj Computational Materials*, *7*(1), 1-7.

This work originated from the discussion of Teng Long, Y. Zhang and Prof. H. Zhang. Prof. H. Zhang and Prof. O. Gutfleisch supervised the research. Teng Long worked on the machine learning model. Teng Long, N. M. Fortunato, I. Opahle, and C. Shen worked on the DFT calculations. Teng Long, I. Samathrakis, and Y. Zhang worked on data analysis. All authors contributed in the writing.

Coauthor Ilias Samathrakis recognized the contribution of Teng Long in "Long, T., Fortunato, N. M., Opahle, I., Zhang, Y., Samathrakis, I., Shen, C., Gutfleisch, O., & Zhang, H. (2021). Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures. *npj Computational Materials*, *7*(1), 1-7."

Signature: 

30.06.2021

Statement of personal contribution

Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures
Long, T., Fortunato, N. M., Opahle, I., Zhang, Y., Samathrakis, I., Shen, C., Gutfleisch, O., & Zhang, H.
2021, *npj Computational Materials*, *7*(1), 1-7.

This work originated from the discussion of Teng Long, Y. Zhang and Prof. H. Zhang. Prof. H. Zhang and Prof. O. Gutfleisch supervised the research. Teng Long worked on the machine learning model. Teng Long, N. M. Fortunato, I. Opahle, and C. Shen worked on the DFT calculations. Teng Long, I. Samathrakis, and Y. Zhang worked on data analysis. All authors contributed in the writing.

Coauthor Chen Shen recognized the contribution of Teng Long in "Long, T., Fortunato, N. M., Opahle, I., Zhang, Y., Samathrakis, I., Shen, C., Gutfleisch, O., & Zhang, H. (2021). Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures. *npj Computational Materials*, *7*(1), 1-7."

Signature:

05.07.2021

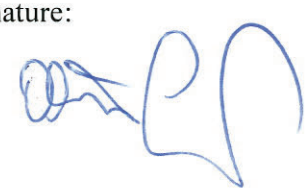
Statement of personal contribution

Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures
Long, T., Fortunato, N. M., Opahle, I., Zhang, Y., Samathrakis, I., Shen, C., Gutfleisch, O., & Zhang, H.
2021, *npj Computational Materials*, *7*(1), 1-7.

This work originated from the discussion of Teng Long, Y. Zhang and Prof. H. Zhang. Prof. H. Zhang and Prof. O. Gutfleisch supervised the research. Teng Long worked on the machine learning model. Teng Long, N. M. Fortunato, I. Opahle, and C. Shen worked on the DFT calculations. Teng Long, I. Samathrakis, and Y. Zhang worked on data analysis. All authors contributed in the writing.

Coauthor Prof. Oliver Gutfleisch recognized the contribution of Teng Long in "Long, T., Fortunato, N. M., Opahle, I., Zhang, Y., Samathrakis, I., Shen, C., Gutfleisch, O., & Zhang, H. (2021). Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures. *npj Computational Materials*, *7*(1), 1-7."

Signature:



08.07.21

Prof. Oliver Gutfleisch
Technische Universität Darmstadt
Materialwissenschaft, FB 11
FG Funktionale Materialien
Alarich-Weiss-Str. 16
D - 64287 Darmstadt

Statement of personal contribution

Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures
Long, T., Fortunato, N. M., Opahle, I., Zhang, Y., Samathrakis, I., Shen, C., Gutfleisch, O., & Zhang, H.
2021, *npj Computational Materials*, *7*(1), 1-7.

This work originated from the discussion of Teng Long, Y. Zhang and Prof. H. Zhang. Prof. H. Zhang and Prof. O. Gutfleisch supervised the research. Teng Long worked on the machine learning model. Teng Long, N. M. Fortunato, I. Opahle, and C. Shen worked on the DFT calculations. Teng Long, I. Samathrakis, and Y. Zhang worked on data analysis. All authors contributed in the writing.

Coauthor Prof. Hongbin Zhang recognized the contribution of Teng Long in "Long, T., Fortunato, N. M., Opahle, I., Zhang, Y., Samathrakis, I., Shen, C., Gutfleisch, O., & Zhang, H. (2021). Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures. *npj Computational Materials*, *7*(1), 1-7."

Signature: 

05.07.2021

C.3. Thermal conductivity of h-BN monolayers using machine learning interatomic potential

Statement of personal contribution

Thermal conductivity of h-BN monolayers using machine learning interatomic potential
Zhang, Y., Shen, C., Long, T., & Zhang, H.
2020, *Journal of Physics: Condensed Matter*, *33*(10), 105903.

This work originated from Prof. H. Zhang. Y. Zhang conducted the research under the supervision of Teng Long and Prof. H. Zhang. Y. Zhang and Teng Long worked on the machine learning interatomic potential. Y. Zhang and C. Shen worked on the thermal conductivity calculation. All authors contributed in the writing.

The first author Yixuan Zhang recognized the contribution of Teng Long in "Zhang, Y., Shen, C., Long, T., & Zhang, H. (2020). Thermal conductivity of h-BN monolayers using machine learning interatomic potential. *Journal of Physics: Condensed Matter, 33*(10), 105903."

Signature: 

30.06.2021

Statement of personal contribution

Thermal conductivity of h-BN monolayers using machine learning interatomic potential
Zhang, Y., Shen, C., Long, T., & Zhang, H.
2020, *Journal of Physics: Condensed Matter*, *33*(10), 105903.

This work originated from Prof. H. Zhang. Y. Zhang conducted the research under the supervision of Teng Long and Prof. H. Zhang. Y. Zhang and Teng Long worked on the machine learning interatomic potential. Y. Zhang and C. Shen worked on the thermal conductivity calculation. All authors contributed in the writing.

Coauthor Chen Shen recognized the contribution of Teng Long in "Zhang, Y., Shen, C., Long, T., & Zhang, H. (2020). Thermal conductivity of h-BN monolayers using machine learning interatomic potential. *Journal of Physics: Condensed Matter*, *33*(10), 105903."

Signature:

05.07.2021

Statement of personal contribution

Thermal conductivity of h-BN monolayers using machine learning interatomic potential
Zhang, Y., Shen, C., Long, T., & Zhang, H.
2020, *Journal of Physics: Condensed Matter*, *33*(10), 105903.

This work originated from Prof. H. Zhang. Y. Zhang conducted the research under the supervision of Teng Long and Prof. H. Zhang. Y. Zhang and Teng Long worked on the machine learning interatomic potential. Y. Zhang and C. Shen worked on the thermal conductivity calculation. All authors contributed in the writing.

Coauthor Teng Long recognized the contribution of Teng Long in "Zhang, Y., Shen, C., Long, T., & Zhang, H. (2020). Thermal conductivity of h-BN monolayers using machine learning interatomic potential. *Journal of Physics: Condensed Matter*, *33*(10), 105903."

Signature: 

30.06.2021

Statement of personal contribution

Thermal conductivity of h-BN monolayers using machine learning interatomic potential
Zhang, Y., Shen, C., Long, T., & Zhang, H.
2020, *Journal of Physics: Condensed Matter*, *33*(10), 105903.

This work originated from Prof. H. Zhang. Y. Zhang conducted the research under the supervision of Teng Long and Prof. H. Zhang. Y. Zhang and Teng Long worked on the machine learning interatomic potential. Y. Zhang and C. Shen worked on the thermal conductivity calculation. All authors contributed in the writing.

Coauthor Prof. Hongbin Zhang recognized the contribution of Teng Long in "Zhang, Y., Shen, C., Long, T., & Zhang, H. (2020). Thermal conductivity of h-BN monolayers using machine learning interatomic potential. *Journal of Physics: Condensed Matter*, *33*(10), 105903."

Signature: 

05.07.2021

Teng Long

Ph.D. candidate
TU Darmstadt

+ 49 6151 16-20827
Otto-Berndt-Straße 3, 64287
Darmstadt, Germany
tenglong@tmm.tu-darmstadt.de

PERSONAL INFORMATION

Nationality: P. R. China

Date of Birth: 12.03.1993

Gender: Male

EDUCATION

Technische Universität Darmstadt — **Darmstadt, Germany**
Ph.D. candidate in Materials Science — Oct. 2018 – Nov. 2021

Shandong University — **Jinan, China**
M. Sc. of Engineering in Materials Science and Engineering — Sept. 2015 - July 2018

Shandong University — **Jinan, China**
Bachelor of Economics in Financial Mathematics and Financial Engineering — Sept. 2011 - July 2015

RESEARCH INTEREST

- Materials design by machine learning
  -- multi-objective inverse design of crystal structures via deep learning
  -- reinforcement learning
  -- forward modelling of physical properties
  -- machine learning interatomic potential
  -- transformers
- Density functional theory and high-throughput calculations
  -- thermal conductivities
  -- dielectric function and band gap of insulators

PUBLICATIONS

1. **Long, T.**, Fortunato, N. M., Zhang, Y., Gutfleisch, O., and Zhang, H. (2021). An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature. *Materials Research Letters*, *9*(4), 169-174.
2. **Long, T.,** Fortunato, N. M., Opahle, I., Zhang, Y., Samathrakis, I., Shen, C., Gutfleisch, O., and Zhang, H. (2021). Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures." *npj Computational Materials*, 7(1), 1-7.
3. Zhang, Y., Shen, C., **Long, T.**, and Zhang, H. (2020). Thermal conductivity of h-BN monolayers using machine learning interatomic potential. *Journal of Physics: Condensed Matter*, *33*(10), 105903.
4. Samathrakis, I., **Long, T.**, Zhang, Z., Singh, H. and Zhang, H. (2021). "Enhanced anomalous Nernst effects in ferromagnetic materials driven by Weyl nodes." *Journal of Physics D: Applied Physics, 55(7)*, 074003.