CrystalGAN: Learning to Discover Crystallographic Structures with Generative Adversarial Networks

Asma Nouira[1], Nataliya Sokolovska[2], Jean-Claude Crivello[1]

[1]University Paris Est, ICMPE (UMR 7182)
CNRS, UPEC, F-94320 Thiais, France
[2]Sorbonne University, INSERM, NutriOmics team, Paris France

Abstract

Our main motivation is to propose an efficient approach to generate novel multi-element stable chemical compounds that can be used in real world applications. This task can be formulated as a combinatorial problem, and it takes many hours of human experts to construct, and to evaluate new data. Unsupervised learning methods such as Generative Adversarial Networks (GANs) can be efficiently used to produce new data. Cross-domain Generative Adversarial Networks were reported to achieve exciting results in image processing applications. However, in the domain of materials science, there is a need to synthesize data with higher order complexity compared to observed samples, and the state-of-the-art cross-domain GANs can not be adapted directly. In this contribution, we propose a novel GAN called CrystalGAN which generates new chemically stable crystallographic structures with increased domain complexity. We introduce an original architecture, we provide the corresponding loss functions, and we show that the CrystalGAN generates very reasonable data. We illustrate the efficiency of the proposed method on a real original problem of novel hydrides discovery that can be further used in development of hydrogen storage materials.

Keywords: Generative Adversarial Nets, Cross-Domain Learning, Materials Science, Higher-order Complexity.

Introduction

In modern society, a big variety of inorganic compositions are used for hydrogen storage owing to its favorable cost (Crivello et al., 2016). A vast number of organic molecules are applied in solar cells, such as organic light-emitting diodes, conductors, and sensors (Yang et al., 2017). Synthesis of new organic and inorganic compounds is a challenge in physics, chemistry and in materials science. Design of new structures aims to find the best solution in a big chemical space, and it is in fact a combinatorial optimization problem.

The number of applications of data mining methods in chemistry and materials science increases steadily (Seko, Togo, and Tanaka, 2017). There is a hope that recent developments in machine learning and data mining will accelerate the progress in materials science. Machine learning methods, namely, generative models, are reported to be efficient in new data generation (Friedman, Tibshirani, and Hastie, 2009), and nowadays we have access to both, techniques to generate a huge amount of new chemical compounds, and to test the properties of all these candidates.

In this work, we focus on applications of hydrogen storage, and in particular, we challenge the problem to investigate novel chemical compositions with stable crystals. Traditionally, density functional theory (DFT) plays a central role in prediction of chemically relevant compositions with stable crystals (Seko et al., 2018). However, the DFT calculations are computationally expensive, and it is not acceptable to apply it to test all possible randomly generated structures.

A number of machine learning approaches were proposed to facilitate the search for novel stable compositions (Butler et al., 2018). There was an attempt to find new compositions using an inorganic crystal structure database, and to estimate the probabilities of new candidates based on compositional similarities. These methods to generate relevant chemical compositions are based on recommender systems (Hu, Koren, and Volinsky, 2008). The output of the recommender systems applied in the crystallographic field is a rating or preference for a structure. A recent approach based on a combination of machine learning methods and the high-throughput DFT calculations allowed to explore ternary chemical compounds (Schmidt et al., 2018), and it was shown that statistical methods can be of a big help to identify stable structures, and that they do it much faster than standard methods. Recently, support vector machines were tested to predict crystal structures (Oliynyk et al., 2017) showing that the method can reliably predict the crystal structure given its composition. It is worth mentioning that data representation of observations to be passed to a learner, is critical, and data representations which are the most suitable for learning algorithms, are not necessarily scientifically intuitive (Swann et al., 2018).

Deep learning methods were reported to learn rich hierarchical models over all kind of data, and the GANs (Goodfellow et al., 2014) is a state-of-the-art model to synthesize data. Moreover, deep networks were reported to learn transferable representations (Ren and Lee, 2017). The GANs were already exploited with success in cross-domain learn-

ing applications for image processing (Zhu et al., 2017; Kim et al., 2017; Janz et al., 2017).

Our goal is to develop a competitive approach to identify stable ternary chemical compounds, i.e., compounds containing three different elements, from observations of binary compounds.

Nowadays, there does not exist any approach that can be applied directly to such an important task of materials science. The state-of-the-art GANs are limited in the sense that they do not generate samples in domains with increased complexity, e.g., the application where we aim to construct crystals with three elements from observations containing two chemical elements only. An attempt to learn many-to-many mappings was recently introduced by Almahairi et al. (2018), however, this promising approach does not allow to generate data of a higher-order dimension.

Our contribution is multi-fold:

- To our knowledge, we are the first to introduce a GAN to solve the scientific problem of discovery of novel crystal structures, and we introduce an original methodology to generate new stable chemical compositions;

- The proposed method is called *CrystalGAN*, and it consists of two cross-domain GAN blocks with constraints integrating prior knowledge including a feature transfer step;

- The proposed model generates data with increased complexity with respect to observed samples;

- We demonstrate by numerical experiments on a real challenge of chemistry and materials science that our approach is competitive compared to existing methods;

- The proposed algorithm is efficiently implemented in Python, and it will be publicly available shortly.

This paper is organized as follows. First, we discuss the related work. Second, we provide the formalization of the problem, and introduce the CrystalGAN. The results of our numerical experiments are shown in the experimental section. Concluding remarks and perspectives close the paper.

Related Work

Our contribution is closely related to the problems of unsupervised learning and cross-domain learning, since our aim is to synthesize novel data, and the new samples are supposed to belong to an unobserved domain with an augmented complexity.

In the adversarial nets framework, the deep generative models compete with an adversary which is a discriminative model learning to identify whether an observation comes from the model distribution or from the data distribution (Goodfellow, 2016). A classical GAN consists of two models, a generator $\mathbf{G}$ whose objective is to synthesize data and a discriminator $\mathbf{D}$ whose aim is to distinguish between real and generated data. The generator and the discriminator are trained simultaneously, and the training problem is formulated as a two-player minimax game. A number of techniques to improve training of GANs were proposed by Arjovsky, Chintala, and Bottou (2017); Gulrajani et al. (2017); Salimans et al. (2016).

Learning cross domain relations is an active research direction in image processing. Several recent papers (Zhu et al., 2017; Kim et al., 2017; Almahairi et al., 2018) discuss an idea to capture some particular characteristics of one image and to translate them into another image. This problem is formalized as image-to-image translation, and there exist multiple applications, e.g., converting a grayscale image to a color image, or converting an image from one representation of a given scene to another. The state-of-the-art methods of Zhu et al. (2017); Kim et al. (2017) are based on the property that the translation has to be cycle consistent. If a translator $\mathbf{G} : A \to B$ is used, then there exist another translator $\mathbf{F} : B \to A$ so that $\mathbf{G}$ and $\mathbf{F}$ are inverse of each other, and the mappings are bijective. The mappings $\mathbf{G}$ and $\mathbf{F}$ are trained simultaneously under the cycle consistency assumption what encourages $\mathbf{F}(\mathbf{G}(x)) \approx x$, and $\mathbf{G}(\mathbf{F}(x')) \approx x'$. The objective function includes the adversarial losses on domains A and B, and the cycle consistency loss.

A conditional GAN for image-to-image translation is considered by Isola et al. (2017). An advantage of the conditional model is that it allows to integrate underlying structure into the model. The conditional GANs were also used for multi-model tasks (Mirza and Osindero, 2014). An idea to combine observed data to produce new data was proposed in (Yazdani, 2017), e.g., an artist can mix existing pieces of music to create a new one.

An approach to learn high-level semantic features, and to train a model for more than a single task, was introduced by Ren and Lee (2017). In particular, it was proposed to train a model to jointly learn several complementary tasks. This method is expected to overcome the problem of overfitting to a single task. An idea to introduce multiple discriminators whose role varies from formidable adversary to forgiving teacher was discussed by Durugkar, Gemp, and Mahadevan (2017).

Several GANs were adapted to some materials science and chemical applications. So, Objective-Reinforced GANs that perform molecular generation of carbon-chain sequence taking into consideration some desired properties, were introduced in (Sanchez-Lengeling et al., 2017), and the method was shown to be efficient for drug discovery. Another avenue is to integrate rule-based knowledge, e.g., molecular descriptors with the deep learning. Chem-Net (Goh et al., 2017) is a deep neural network pre-trained with chemistry-relevant representations obtained from prior knowledge. The model can be used to predict new chemical properties. However, as we have already mentioned before, none of these methods generates crystal data of augmented complexity.

CrystalGAN: an Approach to Generate Stable Ternary Chemical Compounds

In this section, we introduce our approach. The CrystalGAN consists of three procedures:

1. *First step GAN* which is closely related to the cross-domain GANs, and that generates pseudo-binary samples where the domains are mixed.

AH:	First domain, H is hydrogen, and A is a metal
BH:	Second domain, H is hydrogen, and B is another metal
$\mathbf{G}_{AHB_1}$:	Generator function that translates input features x_{AH} from (domain) AH to BH
$\mathbf{G}_{BHA_1}$:	Generator function that translates input features x_{BH} from (domain) BH to AH
$\mathbf{D}_{AH}$ and $\mathbf{D}_{BH}$:	Discriminator functions of AH domain and BH domain, respectively
AHB_1:	x_{AHB_1} is a sample generated by generator function $\mathbf{G}_{AHB_1}$
BHA_1:	y_{BHA_1} is a sample produced by generator function $\mathbf{G}_{BHA_1}$
$AHBA_1$ and $BHAB_1$:	Data reconstructed after two generator translations
AHB_g and BHA_g:	Data obtained after feature transfer step from domain AH to domain BH,
	and from domain BH to domain AH, respectively
	Input data for the second step of CrystalGAN
$\mathbf{G}_{AHB_2}$:	Generator function that translates x_{AHB_g}
	Features generated in the first step from AHB_g to AHB_2
$\mathbf{G}_{BHA_2}$:	Generator function that translates y_{BHA_g}
	Data generated in first step from BHA_g to BHA_2
$\mathbf{D}_{AHB}$ and $\mathbf{D}_{BHA}$:	The discriminator functions of domain AHB_g and domain BHA_g, respectively
AHB_2:	x_{AHB_2} is a sample generated by the generator function $\mathbf{G}_{AHB_2}$
BHA_2:	y_{BHA_2} is a sample produced by the generator function $\mathbf{G}_{BHA_2}$
$AHBA_2$ and $BHAB_2$:	Data reconstructed as a result of two generators translations
AHB_2 and BHA_2:	Final new data (to be explored by human experts)

Table 1: Notations used in CrystalGAN.

2. *Feature transfer procedure* constructs higher order complexity data from the samples generated at the previous step, and where components from all domains are well-separated.

3. *Second step GAN* synthesizes, under geometric constraints, novel ternary stable chemical structures.

First, we describe a cross-domain GAN, and then, we provide all the details on the proposed CrystalGAN. We provide all notations used by the CrystalGAN in Table 1. The GANs architectures for the first and the second steps are shown on Figure 1.

A Cross-Domain GAN: Problem Formulation

DiscoGAN (Kim et al., 2017) and CycleGAN (Zhu et al., 2017) propose a promising modification compared to the classic GAN: the model does not take the noise but samples from another domain, resulting in cross-domain learning.

We consider a function $\mathbf{G}_{ABZ}$ that maps elements from domains A and B to domain Z which includes the co-domains A and B. In an unsupervised learning scenario, $\mathbf{G}_{ABZ}$ can be arbitrarily defined, however, to apply it to real-world applications, some conditions on the relation of interest have to be well-defined.

In an idealistic setting, the equality

$$\mathbf{G}_{ABZ} \circ \mathbf{G}_{ZAB}(x_A, x_B) = (x_A, x_B) \qquad (1)$$

is satisfied. However, this constraint is a hard constraint, it is not straightforward to optimize it, and a relaxed soft constraint is preferred. As a soft constraint, we can consider the distance

$$d\left(\mathbf{G}_{ABZ} \circ \mathbf{G}_{ZAB}(x_A, x_B), (x_A, x_B)\right), \qquad (2)$$

and minimize it using a metric function such as L_1 or L_2.

$$-\mathbb{E}_{x_A, x_B \sim P_{A,B}} \left[\log \mathbf{D}_Z(\mathbf{G}_{ABZ})(x_A, x_B)\right]. \qquad (3)$$

The cross-domain GANs were shown to be efficient to discover relations between two different domains from unpaired samples, without any explicit labels, and to find a mapping from one domain to another. However, neither DiscoGAN, or CycleGAN are not able to generate data with increased complexity.

Problem Formulation for Applications with Augmented Complexity

We now propose a novel architecture based on the cross-domain GAN algorithms with constraint learning to discover higher order complexity crystallographic systems. We introduce a GAN model to find relations between different crystallographic domains, and to generate new materials.

To make the paper easier to follow, without loss of generality, we will present our method providing a specific example of generating ternary hydride compounds of the form "A (a metal) - H (hydrogen) - B (a metal)".

The training algorithm observes stable binary compounds containing chemical elements A+H which is a composition of some metal A and the hydrogen H, and B+H which is a mixture of another metal B with the hydrogen. So, a machine learning algorithm has access to observations $\{(x_{AH_i})\}_{i=1}^{N_{AH}}$ and $\{(y_{BH_i})\}_{i=1}^{N_{BH}}$. *Our goal is to generate novel ternary, i.e. more complex, stable data x_{AHB} (or y_{BHA}) based on the properties learned from the observed binary structures.*

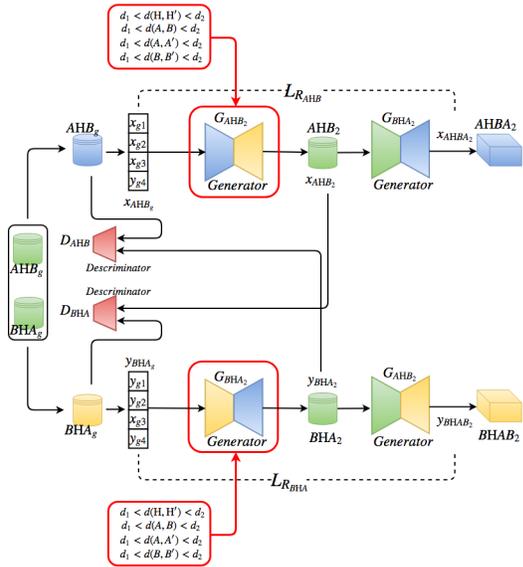We describe the architecture of the CrystalGAN on Figure 1.

Steps of CrystalGAN

Our approach consists of two consecutive steps with a feature transfer procedure inbetween.

First Step The first step of CrystalGAN generates new data with increased complexity. The adversarial network
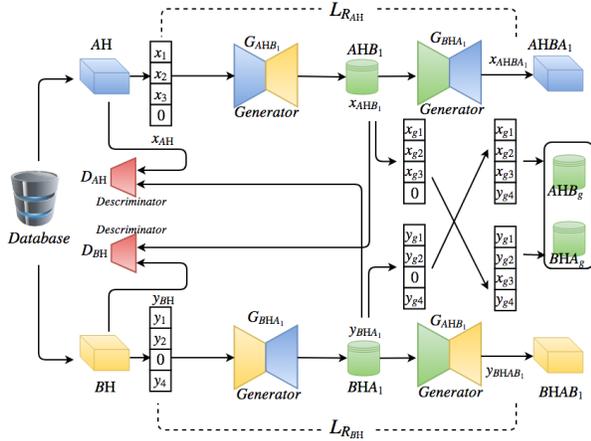
(a) First step of CrystalGAN.

(b) Second step of CrystalGAN
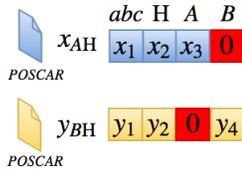
Figure 1: The CrystalGAN architecture.

Figure 2: Encoding of x_{AH} and y_{BH} with placeholders.

takes $\{(x_{AH_i})\}_{i=1}^{N_{AH}}$ and $\{(y_{BH_i})\}_{i=1}^{N_{BH}}$, and synthesizes

$$x_{AHB_1} = \mathbf{G}_{AHB_1}(x_{AH}), \qquad (4)$$

$$x_{AHBA_1} = \mathbf{G}_{BHA_1}(x_{AHB_1}) = \mathbf{G}_{BHA_1} \circ \mathbf{G}_{AHB_1}(x_{AH}). \qquad (5)$$

and

$$y_{BHA_1} = \mathbf{G}_{BHA_1}(y_{BH}), \qquad (6)$$

$$y_{BHAB_1} = \mathbf{G}_{AHB_1}(y_{BHA_1}) = \mathbf{G}_{AHB_1} \circ \mathbf{G}_{BHA_1}(y_{BH}). \qquad (7)$$

Figure 1a summarizes the first step of CrystalGAN.

The reconstruction loss functions take the following form:

$$\mathcal{L}_{R_{AH}} = d(x_{AHBA_1}, x_{AH}) = d(\mathbf{G}_{BHA_1} \circ \mathbf{G}_{AHB_1}(x_{AH}), x_{AH}), \qquad (8)$$

$$\mathcal{L}_{R_{BH}} = d(y_{BHAB_1}, y_{BH}) = d(\mathbf{G}_{AHB_1} \circ \mathbf{G}_{BHA_1}(y_{BH}), y_{BH}). \qquad (9)$$

Ideally, $\mathcal{L}_{R_{AH}} = 0$, $\mathcal{L}_{R_{BH}} = 0$, and $x_{AHBA_1} = x_{AH}$, $y_{BHAB_1} = y_{BH}$, and we minimize the distances $d(x_{AHBA_1}, x_{AH})$ and $d(y_{BHAB_1}, y_{BH})$.

The generative adversarial loss functions of the first step of CrystalGAN aim to control that the original observations are reconstructed as accurate as possible:

$$\mathcal{L}_{\mathrm{GAN}_{BH}} = -\mathbb{E}_{x_{AH} \sim P_{AH}}[\log(\mathbf{D}_{BH}(\mathbf{G}_{AHB_1}(x_{AH})))], \qquad (10)$$

and

$$\mathcal{L}_{\mathrm{GAN}_{AH}} = -\mathbb{E}_{y_{BH} \sim P_{BH}}[\log(\mathbf{D}_{AH}(\mathbf{G}_{BHA_1}(y_{BH})))]. \qquad (11)$$

The generative loss functions contain the two terms defined above:

$$\mathcal{L}_{\mathbf{G}_{AHB_1}} = \mathcal{L}_{\mathrm{GAN}_{BH}} + \mathcal{L}_{R_{AH}}, \qquad (12)$$

$$\mathcal{L}_{\mathbf{G}_{BHA_1}} = \mathcal{L}_{\mathrm{GAN}_{AH}} + \mathcal{L}_{R_{BH}}. \qquad (13)$$

The discriminative loss functions aim to discriminate the samples coming from AH and BH:

$$\mathcal{L}_{\mathbf{D}_{BH}} = -\mathbb{E}_{y_{BH} \sim P_{BH}}[\log(\mathbf{D}_{BH}(y_{BH}))] \qquad (14)$$
$$-\mathbb{E}_{x_{AH} \sim P_{AH}}[\log(1 - \mathbf{D}_{BH}(\mathbf{G}_{AHB_1}(x_{AH})))],$$

$$\mathcal{L}_{\mathbf{D}_{AH}} = -\mathbb{E}_{x_{AH} \sim P_{AH}}[\log(\mathbf{D}_{AH}(x_{AH}))] \qquad (15)$$
$$-\mathbb{E}_{y_{BH} \sim P_{BH}}[\log(1 - \mathbf{D}_{AH}(\mathbf{G}_{BHA_1}(y_{BH})))].$$

Now, we have all elements to define the full generative loss function of the first step:

$$\mathcal{L}_{\mathbf{G}_1} = \mathcal{L}_{\mathbf{G}_{AHB_1}} + \mathcal{L}_{\mathbf{G}_{BHA_1}} \qquad (16)$$
$$= \lambda_1 \mathcal{L}_{\mathrm{GAN}_{BH}} + \lambda_2 \mathcal{L}_{R_{AH}} + \lambda_3 \mathcal{L}_{\mathrm{GAN}_{AH}} + \lambda_4 \mathcal{L}_{R_{BH}},$$

where λ_1, λ_2, λ_3, and λ_4 are real-valued hyper-parameters that control the ratio between the corresponding terms, and the hyper-parameters are to be fixed by cross-validation.

The full discriminator loss function of this step $\mathcal{L}_{\mathbf{D}_1}$ is defined as follows:

$$\mathcal{L}_{\mathbf{D}_1} = \mathcal{L}_{\mathbf{D}_{AH}} + \mathcal{L}_{\mathbf{D}_{BH}}. \qquad (17)$$

Feature Transfer The first step generates pseudo-binary samples MH, where M is a new discovered domain merging A and B properties. Although these results can be interesting for human experts, the samples generated by the first step are not easy to interpret, since the domains A and B are completely mixed in these samples, and there is no way to deduce characteristics of two separate elements coming from these domains.

So, we need a second step which will generate data of a higher order complexity from two given domains. We transfer the attributes of A and B elements, this procedure is also shown on Figure 1a, in order to construct a new dataset that will be used as a training set in the second step of the CrystalGAN.

In order to prepare the datasets to generate higher order complexity samples, we add a placeholder. (E.g., for domain AH, the fourth matrix is empty, and for domain BH, the third matrix is empty.) This implementation detail is sketched on Figure 2.

Second Step of the CrystalGAN The second step GAN takes as input the data generated by the first step GAN and modified by the feature transfer procedure. The results of the second step are samples which describe ternary chemical compounds that are supposed to be stable from chemical viewpoint. *The geometric constraints control the quality of generated data.*

A crystallographic structure is fully described by a local distribution. This distribution is determined by distances to all nearest neighbors of each atom in a given crystallographic structure. We enforce the second step GAN with the following geometric constraints which satisfy the geometric conditions of our scientific domain application. The implemented constraints are also shown on Figure 1b.

Let $S = \{s_i\}_{i=1}^m$ be the set of distances of the first neighbors of all atoms in a crystallographic structure. There are two geometric constraints to be considered while generating new data.

The first geometric (geo) constraint is defined as follows:

$$\mathcal{L}_{\text{geo}_1} = f(d_1, s_1, ..., s_m) = \min_{s \in S} \| d_1 - s \|_2^2, \quad (18)$$

where d_1 is the minimal distance between two first nearest neighbors in a given crystallographic structure.

The second geometric constraint takes the following form:

$$\mathcal{L}_{\text{geo}_2} = f(d_2, s_1, ..., s_m) = -\min_{s \in S} \| d_2 - s \|_2^2, \quad (19)$$

where d_2 is the maximal distance between two first nearest neighbors.

The loss function of the second step GAN is augmented by the following geometric constraints:

$$\mathcal{L}_{\text{geo}} = \mathcal{L}_{\text{geo}_1} + \mathcal{L}_{\text{geo}_2}. \quad (20)$$

Given x_{AHB_g} and y_{BHA_g} from the previous step, we generate:

$$x_{AHB_2} = \mathbf{G}_{AHB_2}(x_{AHB_g}), \quad (21)$$

$$x_{AHBA_2} = \mathbf{G}_{BHA_2}(x_{AHB_2}) = \mathbf{G}_{BHA_2} \circ \mathbf{G}_{AHB_2}(x_{AHB_g}). \quad (22)$$

and

$$y_{BHA_2} = \mathbf{G}_{BHA_2}(y_{BHA_g}), \quad (23)$$

$$y_{BHAB_2} = \mathbf{G}_{AHB_2}(y_{BHA_2}) = \mathbf{G}_{AHB_2} \circ \mathbf{G}_{BHA_2}(y_{BHA_g}). \quad (24)$$

The reconstruction loss functions are given:

$$\begin{aligned} \mathcal{L}_{R_{AHB}} &= d(x_{AHBA_2}, x_{AHB_g}) \quad (25) \\ &= d(\mathbf{G}_{BHA_2} \circ \mathbf{G}_{AHB_2}(x_{AHB_g}), x_{AHB_g}), \end{aligned}$$

$$\begin{aligned} \mathcal{L}_{R_{BHA}} &= d(y_{BHAB_2}, y_{BHA_g}) \quad (26) \\ &= d(\mathbf{G}_{AHB_2} \circ \mathbf{G}_{BHA_2}(y_{BHA_g}), y_{BHA_g}). \end{aligned}$$

The generative adversarial loss functions are given by:

$$\mathcal{L}_{\text{GAN}_{BHA_g}} = -\mathbb{E}_{x_{AHB_g} \sim P_{AHB_g}}[\log(\mathbf{D}_{BHA}(\mathbf{G}_{AHB_2}(x_{AHB_g})))], \quad (27)$$

$$\mathcal{L}_{\text{GAN}_{AHB_g}} = -\mathbb{E}_{y_{BHA_g} \sim P_{BHA_g}}[\log(\mathbf{D}_{AHB}(\mathbf{G}_{BHA_2}(y_{BHA_g})))]. \quad (28)$$

The generative loss functions of the this step are defined as follows:

$$\mathcal{L}_{\mathbf{G}_{AHB_2}} = \mathcal{L}_{\text{GAN}_{BHA_g}} + \mathcal{L}_{R_{AHB}}, \quad (29)$$

$$\mathcal{L}_{\mathbf{G}_{BHA_2}} = \mathcal{L}_{\text{GAN}_{AHB_g}} + \mathcal{L}_{R_{BHA}}. \quad (30)$$

The losses of the discriminator of the second step can be defined:

$$\begin{aligned} \mathcal{L}_{\mathbf{D}_{BHA}} = &- \mathbb{E}_{y_{BHA_g} \sim P_{BHA_g}}[\log(\mathbf{D}_{BHA}(y_{BHA_g}))] \quad (31) \\ &- \mathbb{E}_{x_{AHB_g} \sim P_{AHB_g}}[\log(1 - \mathbf{D}_{BHA}(\mathbf{G}_{AHB_2}(x_{AHB_g})))], \end{aligned}$$

$$\begin{aligned} \mathcal{L}_{\mathbf{D}_{AHB}} = &- \mathbb{E}_{x_{AHB_g} \sim P_{AHB_g}}[\log(\mathbf{D}_{AHB}(x_{AHB_g}))] \quad (32) \\ &- \mathbb{E}_{y_{BHA_g} \sim P_{BHA_g}}[\log(1 - \mathbf{D}_{AHB}(\mathbf{G}_{BHA_2}(y_{BHA_g})))]. \end{aligned}$$

Now, we have all elements to define the full generative loss function:

$$\mathcal{L}_{\mathbf{G_2}} = \mathcal{L}_{\mathbf{G}_{AHB_2}} + \mathcal{L}_{\mathbf{G}_{BHA_2}} + \mathcal{L}_{\text{geo}} \quad (33)$$

$$\begin{aligned} = \lambda_1 \mathcal{L}_{\text{GAN}_{BHA_g}} + \lambda_2 \mathcal{L}_{R_{AHB}} + \lambda_3 \mathcal{L}_{\text{GAN}_{AHB_g}} + \lambda_4 \mathcal{L}_{R_{BHA}} \\ + \lambda_5 \mathcal{L}_{\text{geo}_1} + \lambda_6 \mathcal{L}_{\text{geo}_2}, \end{aligned}$$

where λ_1, λ_2, λ_3, λ_4, λ_5, and λ_6 are real-valued hyperparameters that control the influence of the terms.

The full discriminative loss function of the second step $\mathcal{L}_{\mathbf{D_2}}$ takes the form:

$$\mathcal{L}_{\mathbf{D_2}} = \mathcal{L}_{\mathbf{D}_{AHB}} + \mathcal{L}_{\mathbf{D}_{BHA}}. \quad (34)$$

To summarize, in the second step, we use the dataset issued from the feature transfer as an input containing two domains x_{AHB_g} and y_{BHA_g}. We train the cross-domain GAN taking into consideration constraints of the crystallographic environment. We integrated geometric constraints proposed by crystallographic and materials science experts to satisfy environmental constraints, and to increase the rate of synthesized stable ternary compounds. The second step of CrystalGAN is drafted on Figure 1b.

GAN Architecture

A generator network is defined as $\mathbf{G}_{AHB_1} : \mathbb{R}_{AH}^{l\times m}, \mathbb{R}_{BH}^{l\times m} \rightarrow \mathbb{R}_{AHB_1}^{k\times m}$, where AH, BH are the input domains, AHB_1 is the output domain, and l and m are the dimensions of the input, k and m dimensions of output samples.

The discriminator network is denoted as $\mathbf{D}_{AH}$: $\mathbb{R}_{AHB_1}^{k\times m} \rightarrow [0,1]$, and it discriminates samples in domain AHB_1. Each generator takes an observation of the size $l\times m$, and passes it to the encoder-decoder pair. Note that $\mathbf{G}_{BHA_1}$, $\mathbf{G}_{AHB_2}$, $\mathbf{G}_{BHA_2}$, $\mathbf{D}_{BH}$, $\mathbf{D}_{AHB}$ and $\mathbf{D}_{BHA}$ are similarly defined. The encoder and the decoder are composed of fully-connected layers. The number of layers ranges from 5 to 10 depending on a domain. The discriminator has an additional layer, a sigmoid function to output a predicted label.

Experiments

Task Description: Exploring Novel Hydrides

Hydrides, compounds which associate hydrogen atoms with other chemical elements, are actively used in storage battery technologies such as nickel-metal hydride battery. A number of hydrides have been explored as a means of hydrogen storage for fuel-cell powered electric cars.

Crystallographic structures can be represented using the POSCAR files which are input files for the DFT calculations under the VASP code (Kresse and Joubert, 1999). These are coordinate files, they contain the lattice geometry and the atomic positions, as well as the number (or the composition) and the nature of atoms in the crystal unit cell.

We use a dataset constructed from (Bourgeois et al., 2017; Villars and Cenzual, 2017) by experts in materials science. Our training data set contains the POSCAR files, and the proposed CrystalGAN generates also POSCAR files. Such a file contains three matrices: the first one is abc matrix, corresponding to the three lattice vectors defining the unit cell of the system, the second matrix contains atomic positions of H atom, and the third matrix contains coordinates of metallic atom A (or B). The information from the files is fed into 4-dimensional tensors. An example of a POSCAR file, and its corresponding representation for the GANs is shown on Figure 3. On Figure 4 we show the corresponding structure in 3D. Note that we increase the data complexity by the feature transfer procedure by adding placeholders.

Our training dataset includes 1,416 POSCAR files of binary hydrides divided into 63 classes where each class is represented as a 4-dimensional tensor. Each class of binary MH hydride contains two elements: the hydrogen H and another element M from the periodic table. This later is selected from the 63 highlighted M elements (in yellow) in the Figure 7. In our experiments, after discussions with materials science researchers, we focused on exploration of ternary compositions "Palladium - Hydrogen - Nickel" from the binary systems observations of "Palladium - Hydrogen" and "Nickel - Hydrogen". So, AH = PdH, and BH = NiH. We also considered another task to generate ternary compounds "Magnesium - Hydrogen - Titanium".

From each system (domain), we have selected 35 crystal structures (stable and metastable) which include experimen-
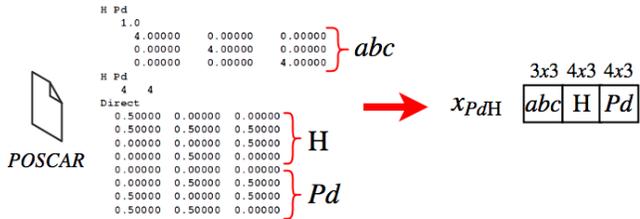

Figure 3: An example of a POSCAR file describing the composition of Palladium and Hydrogen, and the data representation in the CrystalGAN.

tally observed prototypes. Here is a brief data description for this task:

Input dataset	Dimension
PdH	$[35, 4, 18, 3]$
NiH	$[35, 4, 18, 3]$

where 18 and 3 are the maximal numbers of lines and columns in each matrix respectively.

In the CrystalGAN, we need to compute all the distances of the nearest neighbors for each generated POSCAR file. The distances between hydrogen atoms H in a given crystallographic structure should respect some geometric rules, as well as the distances between the atoms $A - B$, $A - A$', and $B - B'$. We applied the geometric constraints on the distances between the neighbors (for each atom in a crystallographic structure) introduced in the previous section. Note that the distances $A - $H and $B - $H are not penalized by the constraints.

Implementation Details

In order to compute the distances between all nearest neighbors in the generated data, we used the pythonic library Pymatgen (Ong et al., 2012) specifically developed for material analysis.

For all experiments in this paper, the distances are fixed by our colleagues in crystallographic and materials science to $d_1 = 1.8$ Å (angstrom, 10^{-10} meter) and $d_2 = 3$ Å. We set all the hyper-parameters by cross validation, however, we found that a reasonable performance is reached when all λ_i have similar values, and are quite close to 1. We use the standard AdamOptimizer with learning rate $\alpha = 0.0001$, and $\beta_1 = 0.5$. The number of epochs is set to 1000 (we verified that the functions converge). The mini-batch size equals 35.

Each block of the CrystalGAN architecture (the generators and the discriminators) is a multi-layer neural network with 5 hidden layers. Each layer contains 100 units. We use the rectified linear unit (ReLU) as an activation function of the neural network. All these parameters were fixed by cross-validation (for both chosen domains "Palladium - Hydrogen" and "Nickel - Hydrogen").

Our code is implemented in Python (TensorFlow). We run the experiments using GPU with graphics card NVIDIA Quadro M5000.

Composition	GAN (standard)	DiscoGAN	CrystalGAN without constraints	CrystalGAN with geometric constraints
Pd - Ni - H	0	0	4	9
Mg - Ti - H	0	0	2	8

Table 2: Number of ternary compositions of good quality generated by the tested methods.
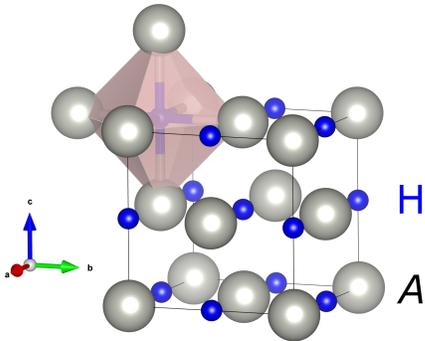

Figure 4: A visualization of a stable structure.

Results

In our numerical experiments, we compare the proposed CrystalGAN with a classical GAN, the DiscoGAN Kim et al. (2017), and the CrystalGAN but without the geometric constraints. All these GANs generate POSCAR files, and we *evaluate the performance of the models by the number of generated ternary structures which satisfy the geometric crystallographic environment*. Table 2 shows the number of successes for the considered methods. The classical GAN which takes Gaussian noise as an input, does not generate acceptable chemical structures. The DiscoGAN approach performs quite well if we use it to generate novel pseudo-binary structures, however, it is not adapted to synthesize ternary compositions. We observed that the CrystalGAN (with the geometric constraints) outperforms all tested methods.

Figure 5 illustrates characteristics of a newly generated ternary (H-Pd-Ni) stable structure: on the left we show the distances between the nearest neighbours in the crystallographic structure, and on the right we visualise the generated POSCAR file. We would like to underline that the generated structure respects the geometric constraints.

Discussion

Here we provide some important remarks on the task considered in this contribution. Discovery of stable chemical structures in general, and of new materials for hydrogen storage in particular, is a challenging task.

From multiple discussions with experts in materials science and chemistry, first, we know that the number of novel stable compounds can not be very high, and it is already considered as a success if we synthesize several stable structures which satisfy the constraints. Hence, we can not really reason in terms of accuracy or error rate which are widely used metrics in machine learning and data mining.

Second, evaluation of a stable structure is not straightforward. Given a new composition, only the result of density functional theory (DFT) calculations can provide a conclusion whether this composition is stable enough, and whether it can be used in practice. However, the DFT calculations are computationally too expensive, and it is out of question to run them on all data we generated using the CrystalGAN. In our work, to avoid the DFT computations, we imply the geometric constraints proposed by the human experience to control the properties of the generated compounds, such as the Switendick criterion (Switendick, 1979). It is planned to run the DFT calculations on some pre-selected generated ternary compositions to take a final decision on practical utility of the chemical compounds.

The evaluation of generated crystallographic structures can also be done by laboratory experiments, exploring geometric properties of the compositions based on the distances between atoms. For example, Figure 4 illustrates a stable structure in cubic NaCl prototype. Another representation of a synthesized data is a histogram of the number of nearest neighbors at a given distance which forms a pair distribution function (PDF). Figure 6 shows a PDF profile for a stable structure where the minimal distance between atoms is $d_{\min}(A, H) = 2 \, \text{Å}$ (angstrom) for 6 first nearest neighbours (cubic cell parameter is 4Å in this example).

Conclusions

Our goal was to develop a principled approach to generate new ternary stable crystallographic structures from observed binary, i.e. containing two chemical elements only. We propose a learning method called CrystalGAN to discover cross-domain relations in real data, and to generate novel structures. The proposed approach can efficiently integrate, in form of constraints, prior knowledge provided by human experts.

CrystalGAN is the first GAN developed to generate scientific data in the field of materials science. To our knowledge, it is also the first approach which generates data of a higher-order complexity, i.e., ternary structures where the domains are well-separated from observed binary compounds. The CrystalGAN was, in particular, successfully tested to tackle the challenge to discover new materials for hydrogen storage.

Currently, we investigate different GANs architectures, also including elements of reinforcement learning, to produce data even of a higher complexity, e.g., compounds containing four or five chemical elements. Note that although the CrystalGAN was developed and tested for applications in materials science, it is a general method where the constraints can be easily adapted to any scientific problem.

Atom 1	Atom 2	Distance
H 1		
	Pd	0.8103
	Ni	2.4366
	Pd	2.4782
	H	2.6765
	H	2.9976
H 2		
	Pd	1.9035
	Ni	2.2619
	Pd	2.2755
	H	2.6765
	H	2.9976
Pd 3		
	H	0.8103
	H	1.9035
	Ni	2.9513
Pd 4		
	Ni	1.9463
	H	2.2755
	H	2.4782
	Ni	2.9662
Ni 5		
	H	2.2619
	Ni	2.3474
	H	2.4366
	Pd	2.9513
	Pd	2.9662
Ni 6		
	Pd	1.9463
	Ni	2.3474

```
H Pd Ni

1.0

   0.31783467    0.38348212    6.43965053

  -0.14229795    4.61861133   -1.02656614

   5.30012273    2.00472569   -0.25500044

H Pd Ni

2  2  2

Cartesian

   1.58695459    0.61697345    0.02452001

   4.09158802    1.54180848   -0.16272403

   2.38770532    0.72835826    0.07921831

   0.40653645    0.16036039    4.67786407

   0.04466104    2.44065856   -0.45760279

  -0.29513394    1.82811617    3.96068835
```

Figure 5: The list of the nearest neighbours (on the left); the corresponding generated POSCAR file (on the right).

Figure 6: Number of nearest neighbors at a given distance for each atom in a structure.

Figure 7: The elements included in our data set are highlighted.

References

Almahairi, A.; Rajeshwar, S.; Sordoni, A.; Bachman, P.; and Courville, A. 2018. Augmented cycleGAN: Learning many-to-many mappings from unpaired data. In *ICML*.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein GAN. *arXiv:1701.07875*.

Bourgeois, N.; Crivello, J.-C.; Cenedese, P.; and Joubert, J.-

M. 2017. Systematic first-principles study of binary metal hydrides. *ACS Combinatorial Science* 19(8):513–523.

Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; and Walsh, A. 2018. Machine learning for molecular and materials science. *Nature* 559.

Crivello, J.-C.; Dam, B.; Denys, R. V.; Dornheim, M.; Grant, D. M.; Huot, J.; Jensen, T. R.; de Jongh, P.; Latroche, M.; Milanese, C.; Milcius, D.; Walker, G. S.; Webb, C. J.; Zlotea, C.; and Yartys, V. A. 2016. Review of magnesium hydride-based materials: development and optimisation. *Applied Physics A*.

Durugkar, I.; Gemp, I.; and Mahadevan, S. 2017. Generative Multi-Adversarial Networks. In *International Conference on Learning Representations (ICLR)*.

Friedman, J. H.; Tibshirani, R.; and Hastie, T. 2009. *The elements of statistical learning*. Springer.

Goh, G. B.; Siegel, C.; Vishnu, A.; and Hodas, N. 2017. ChemNet: A Transferable and Generalizable Deep Neural Network for Small-molecule Property Prediction. In *NIPS Workshop on Machine Learning for Molecules and Materials*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27 (NIPS)*.

Goodfellow, I. 2016. NIPS 2016 Tutorial: Generative Adversarial Networks. *arxiv:1701.00160*.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. 2017. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems 30 (NIPS)*.

Hu, Y.; Koren, Y.; and Volinsky, C. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *ICDM*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*.

Janz, D.; van der Westhuizen, J.; Paige, B.; Kusner, M. J.; and Hernández-Lobato, J. M. 2017. Learning a Generative Model for Validity in Complex Discrete Structures. In *NIPS Workshop on Machine Learning for Molecules and Materials*.

Kim, T.; Cha, M.; Kim, H.; Lee, J. K.; and Kim, J. 2017. Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*.

Kresse, G., and Joubert, D. 1999. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical Review B*.

Mirza, M., and Osindero, S. 2014. Conditional Generative Adversarial Nets. *arXiv:1411.1784*.

Oliynyk, A. O.; Adutwum, L. A.; Rudyk, B. W.; Pisavadia, H.; Lotfi, S.; Hlukhyy, V.; Harynuk, J. J.; Mar, A.;

and Brgoch, J. 2017. Disentangling Structural Confusion through Machine Learning: Structure Prediction and Polymorphism of Equiatomic Ternary Phases ABC. *Journal of the American Chemical Society*.

Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; and Ceder, G. 2012. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science 68 (2013) 314–319*.

Ren, Z., and Lee, Y. J. 2017. Cross-domain self-supervised multi-task feature learning using synthetic imagery. *ArXiv preprint arXiv:1711.09082*.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems 29 (NIPS)*.

Sanchez-Lengeling, B.; Outeiral, C.; Guimaraes, G. L.; and Aspuru-Guzik, A. 2017. Optimizing distributions over molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC). *Preprint: chemrxiv:5309668*.

Schmidt, J.; Chen, L.; Botti, S.; and Marques, M. A. L. 2018. Predicting the stability of ternary intermetallics with density functional theory and machine learning. *Journal of Chemical Physics 148, 241728 (2018)*.

Seko, A.; Hayashi, H.; Kashima, H.; and Tanaka, I. 2018. Matrix- and tensor-based recommender systems for the discovery of currently unknown inorganic compounds. *Phys. Rev. Materials 2, 013805*.

Seko, A.; Togo, A.; and Tanaka, I. 2017. Descriptors for Machine Learning of Materials Data. *arXiv:1709.01666*.

Swann, E.; Sun, B.; Cleland, D. M.; and Barnard, A. 2018. Representing molecular and materials data for unsupervised machine learning. *Molecular simulation*.

Switendick, A. 1979. Band structure calculation for metal hydrogen systems. *Z. Phys. Chem NF* 117:89.

Villars, P., and Cenzual, K., eds. 2017. *Pearson's Crystal Data Crystal Structure Database for Inorganic Compounds*. ASM International.

Yang, X.; Zhang, J.; Yoshizoe, K.; Terayama, K.; and Tsuda, K. 2017. ChemTS: an efficient Python library for de novo molecular generation. *Communications in materials informatics*.

Yazdani, M. 2017. RemixNet: Generative Adversarial Networks for Mixing Multiple Inputs. In *Semantic Computing (ICSC)*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.