









The use of generative models to speed up the discovery of materials

Andrea Gregores Coto¹ , Christian Eike Precker^{1*} , Tom Andersson² ,
Anssi Laukkanen² , Tomi Suhonen², Pilar Rey Rodriguez¹ ,
Santiago Muñnos-Landín^{1*} 

¹ AIMEN Technology Centre, Smart Systems and Smart Manufacturing, Artificial Intelligence and Data Analytics Laboratory, Pl. Cataboi, 36418 Pontevedra, Spain.

² VTT Technical Research Centre of Finland Ltd., 02044 Espoo, Finland.

Abstract

Material Science is a key factor in the evolution of many industrial sectors. Fields such as the aeronautics, automotive, construction, and biotechnology industries have experienced tremendous development with the introduction of advanced, high-performance materials. Such materials not only provide new functionalities to products, but also significant consequences in terms of economic and environmental sustainability of the products and processes triggered by the more efficient use of energy that they provide. Under this scenario, materials that provide such high performance, such as high entropy alloys (HEAs) or polymer derived ceramics (PDCs), have captured the attention of both industry and researchers in recent years. However, the remarkable number of resources required to develop such materials, from its design phase to its synthesis and characterization, means that the discovery of new high-performance materials is moving at a relatively low pace. This fact places emergent strategies based on artificial intelligence (AI) for the design of materials in a good position to be used to accelerate the whole process, providing an impulse in the initial phases of materials design. The enormous number of combinations of elements and the complexity of synthesizability conditions of HEAs and PDCs respectively, paves the way to the deployment of AI techniques such as Generative Models addressed in this work to create synthetic HEAs and PDCs for highly intensive industrial processes. A specific conditional tabular generative adversarial network (CTGAN) was developed to be used on tabular data to generate novel synthetic compounds for each kind of material. The generated synthetic data was based on the conventional parametric design parameters used for HEAs and PDCs, with specific datasets created for them. The real and generated data are compared, calculation of phase diagrams (CALPHAD) simulations are provided to evaluate the performance of the generated samples and a verification of the novel generated compositions is done in open materials databases available in the literature.

Keywords: artificial intelligence, materials science, high-performance materials, generative models

1. Introduction

1.1. Artificial intelligence in material science

Computational methods have attracted significant interest in the materials science community due to their

ability to design new functional materials and compounds through rapid and comprehensive prediction of the stability and properties of such materials (Butler et al., 2016; Shevlin et al., 2021). In recent years, it has been shown how data mining, machine learning, and mathematical optimization makes it possible to

* Corresponding authors: christian.precker@aimen.es, santiago.muinos@aimen.es

ORCID ID's: 0000-0002-9459-9047 (A. Gregores Coto), 0000-0003-0828-6835 (Ch. Eike Precker), 0000-0002-3416-6055 (T. Andersson), 0000-0002-2308-744X (A. Laukkanen), 0000-0003-4661-4983 (P. Rey Rodriguez), 0000-0002-2493-8891 (S. Muñnos-Landín)

© 2023 Authors. This is an open access publication, which can be used, distributed and reproduced in any medium according to the Creative Commons CC-BY 4.0 License requiring that the original work has been properly cited.

systematically reveal the material processing-structure-property-performance relations in order to discover novel compounds and designs (Lee et al., 2021; Zhou et al., 2019). The generic materials informatics workflow begins with the identification and selection of the key components of the dataset after extracting and pre-processing data from the materials database and experiments. The reduced dataset is further examined to discover relationships between the components of interest. Thus, these relationships are utilized to generate the so-called inverse and forward models, the former of which can be used to design materials with the desired properties, whereas the latter is used for predictive analytics. In the final step, new experiments and computer simulations based on theoretical models are used to generate new data for the materials databases (Curtarolo et al., 2012; Kirklin et al., 2015; Saal et al., 2013), thus closing the loop. But even with current computing power being substantially enhanced through parallel computation and the use of Graphic Processing Units (GPUs), is still difficult to perform calculations (such as in density functional theory) in high numbers of samples and large volumes of atoms within a timeframe that allows the exploitation of such feedback loops.

To face the challenge of exploring such a huge compositional space, artificial intelligence (AI) techniques have been employed in chemical and material science in recent years (Hart et al., 2020, 2021; Mazheika et al., 2022). The emergence of large datasets (Curtarolo et al., 2012; Saal et al., 2013), driven by those previous computational studies and experiments compilations, opened a new landscape for the application of machine learning methods in the prediction of material properties, phase transitions and stability in solid state domains (Lee et al., 2021; Zhou et al., 2019). Methods that may be supervised, semi-supervised, or unsupervised, depending on the type and amount of available data (Mitchell et al., 1990). So far, supervised learning (Cunningham et al., 2008) is the most mature and powerful of these approaches and is used in the majority of machine-learning studies in the physical sciences, such as in the mapping of chemical composition to a property of interest. On the other hand, non-supervised learning (Barlow, 1989) can be used for more general analysis and classification of data or to identify previously unrecognized patterns in large datasets. Non-supervised learning techniques based on generative models have driven highly valuable methods to design new chemical compounds (Tong et al., 2021). Nevertheless, the application of machine learning procedures in crystalline solids has a significant delay compared with that in molecular

chemistry. So far, the description of crystalline structures represents a remarkable constraint for feature engineering which is essential for the employment of AI. As a consequence, supervised learning approaches focused on specific structures have mostly been used so far to predict the properties of crystalline solids (Lee et al., 2021; Zhou et al., 2019). Yet the role of AI in compound design remains a territory to be explored in the case of solids.

In this work, we present the use of generative models for the automatic generation of two different materials using the same network architecture. Specifically, the use of a conditional tabular generative adversarial network (CTGAN) (Xu et al., 2019) is introduced to digitally synthesize high entropy alloys (Cantor et al., 2004; Wang et al., 2014) and polymer derived ceramics (Colombo et al., 2013; Sujith et al., 2021). These two materials were selected for two reasons. The first is the mechanical and thermal properties of both, which have captured the attention of processing industries such as steel or aluminium. The second reason is related to the intrinsic complexity of both materials, in terms of configurability for high entropy alloys and their dependency on experimental conditions for polymer derived ceramics. The results of this approach are validated in terms of the novelty and performance of the materials considering industrial demands that have been defined in the context of the European Project ACHIEF (<https://www.achief.eu>), which addresses the discovery of novel high-performance materials.

1.2. High entropy alloys

In 2004, two independent research groups developed so-called high entropy alloys (HEAs) (Cantor et al., 2004; Yeh et al., 2004), a class of materials containing multiple principal chemical elements in near-equimolar proportions. These kinds of materials are of interest in many fields due to their remarkable physical properties, such as superior hardness, strength, and great wear resistance (Wang et al., 2021). Before the introduction of the HEAs concept, the conventional alloying approach was based on a primary element, e.g., iron, followed by the addition of small amounts of secondary elements, e.g., chromium, to increase corrosion resistance and carbon to increase the strength, see Figure 1. This primary element method makes the combination space of elements limited, whereas, in the case of HEAs, we have a perfect scenario to apply ML and AI methods. Furthermore, many exploitable combinations are still open for discovery with improved mechanical and thermodynamic performance.

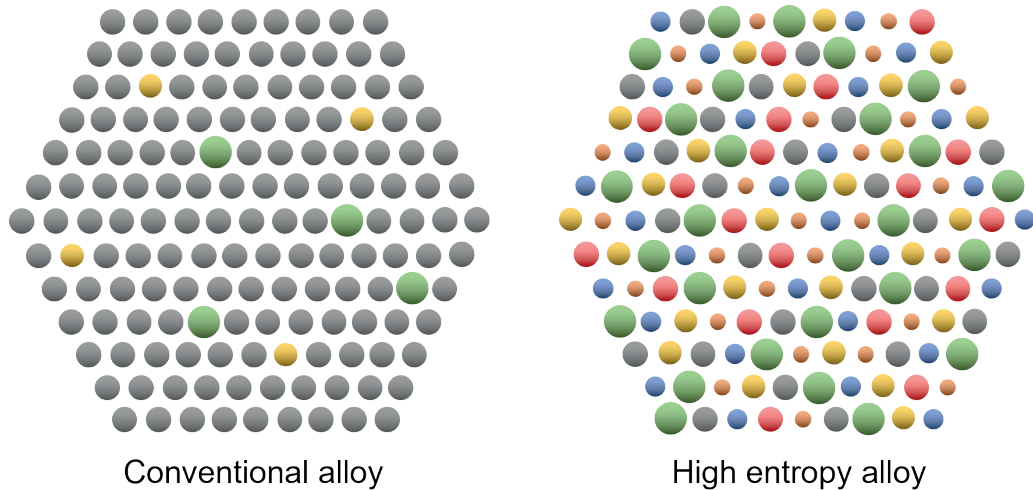


Fig. 1. Sketch of a conventional alloy and a HEA. A conventional alloy is composed of a primary element and small amounts of secondary elements. A HEA has no dominant element, being near-equiatomic, and the variability in atoms' positions contributes to the high entropy effect

In HEAs, the presence of multiple chemical elements in near-equiatomic proportions (composed of five or more principal elements, possessing between 5 at. % and 35 at. %) increases sufficiently the entropy of mixing, overcoming the enthalpy formation of the compounds, giving rise to stable solid solution formations, rather than intermetallic compounds (Yeh et al., 2004). HEAs can also be defined in terms of mixing entropy by

$$\Delta S_{\text{mix}} = -R \sum_{i=1}^n c_i \ln c_i \quad (1)$$

where c_i is the stoichiometric ratio of the i -th component in the alloy, and $R = 8.314 \text{ J} \cdot \text{mol}^{-1} \cdot \text{K}$ is the gas constant (Yeh, 2015). The mixing entropy can be written in terms of the gas constant R , so the HEAs are defined when a composition has $\Delta S_{\text{mix}} > 1.5R$. For $1R < \Delta S \leq 1.5R$ compounds are defined as medium entropy alloys (MEAs) and for $\Delta S_{\text{mix}} < 1R$ low entropy alloys (LEAs) (Miracle et al., 2014; Yeh, 2013). From equation (1) it is possible to see that with the increase of the number of elements, the entropy also does, e.g., an alloy containing five equiatomic elements has $\Delta S_{\text{mix}} = 1.61R$ and another one containing six equiatomic elements has $\Delta S_{\text{mix}} = 1.79R$.

The high entropy effect in HEAs is important because it can enhance the formation of phases. Among the phases in which HEAs can be found, the alloys can be classified as amorphous (AM), intermetallic (IM), solid solution (SS), or a mixture of them. The SS phase means a significant or complete mixing of all constituent elements in the structures of body-centred cubic (BCC), face-centred cubic (FCC), or hexagonal close-packed (HCP). IM phases mean stoichiometric com-

pounds with specific *Strukturbericht* designation, such as B2 (NiAl) and L1₂ (Ni₃Al) (Wang et al., 2014; Yeh, 2013). The phase is an important parameter for HEAs since it determines the physical properties. For example, to achieve high hardness, the SS is indicated, for better elasticity, the AM, and for great wear resistance, IM (Tsai et al., 2019; Wang et al., 2014).

1.3. Polymer derived ceramics

Polymer derived ceramics (PDCs) have been synthesized since the 1960s and have attracted considerable interest due to their excellent behaviour at high temperatures (Sujith et al., 2021), exhibiting good oxidation and creep resistance, as well as being additive-free ceramic materials (Colombo et al., 2013).

It began with the synthesis of organosilicon polymers to get PDCs, but progress in this area (Ainger & Herbert, 1960; Chantrell & Popper, 1964; Verbeek, 1974; Verbeek & Winter, 1974; Winter et al., 1974; Yajima et al., 1975, 1978) showed that precursors can be inorganic or organometallic systems (Colombo et al., 2013; Sujith et al., 2021). The precursor to PDC route is a complex process that consists of shaping and cross-linking the precursor, then the pyrolysis step, and finally, the crystallization to get the final ceramic (Colombo et al., 2013; Fu et al., 2019; see Figure 2). Apart from its complexity, this whole process has a huge impact on the behaviour of the final PDC. The final properties depend on external factors to the PDCs composition, e.g., pyrolysis temperature, the atmosphere where pyrolysis is done, pyrolysis time, etc. (Colombo et al., 2013; Cross et al., 2006; Sujith et al., 2021).

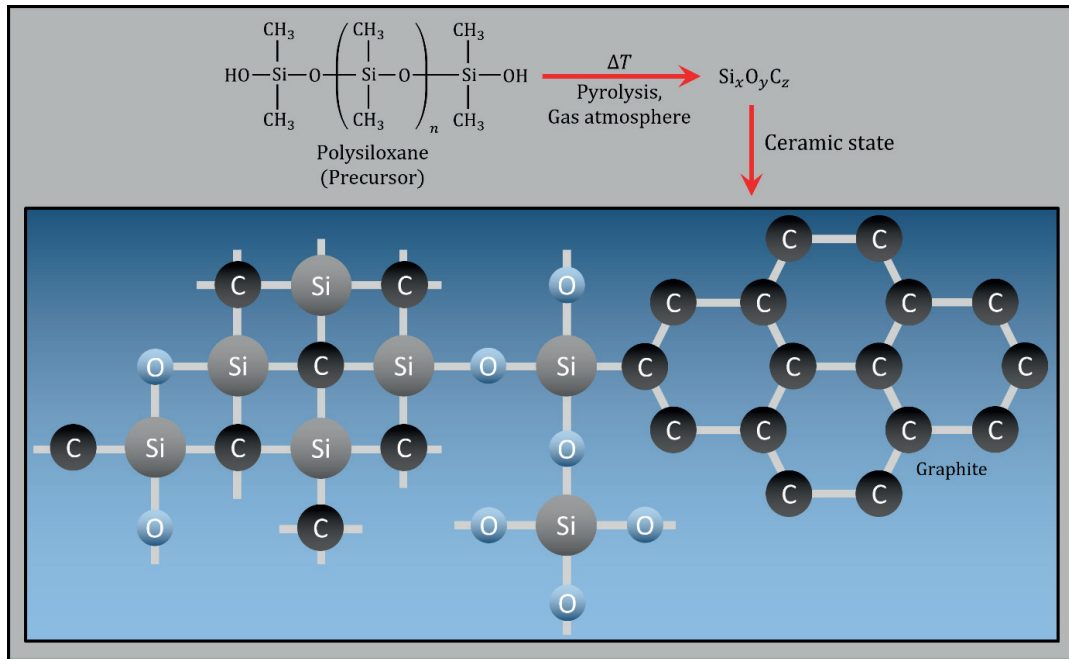


Fig. 2. Polymer derived ceramics synthesis by pyrolysis of polysiloxanes

Although there can be different kinds of precursors, organosilicons are the most studied ones. Due to their mechanical, thermal, biological, and electrical properties (Colombo et al., 2013; Fu et al., 2019; Riedel et al., 2006), are of great interest for industrial applications such as information technology, energy, nano-systems, and biomedicine (Colombo et al., 2013). For the synthesis of PDCs, the most frequently used precursors are polysilane, polycarbosilane, polysilazane, and polysiloxane (Fu et al., 2019).

2. Methodology

2.1. HEAs dataset

For the purpose of feeding a generative model, a dataset was developed. Based on previous work (Lee et al., 2021; Zhou et al., 2019), it was concluded that some design parameters are needed in order to make it possible for the model to learn the relations between features and chemical composition of HEAs.

HEAs were collected from works available in the literature and merged (Gorsse et al., 2018; Miracle & Senkov, 2017; Tsai et al., 2019; Vaidya et al., 2019; Zhou et al., 2019). After prior filtering and the removal of duplicated compounds, the given dataset ended with 1117 entries (Precker et al., 2021a). The phases were used as conditional training parameters, and because this information for some compounds was unknown, these were removed, so the dataset used to train the

CTGAN model had at the end 1103 entries, composed of 195 AM, 362 IM, 350 SS, and 196 SS+IM.

Previous studies on predicting HEAs phases have used parametric approaches based on the Hume–Rothery rules, which concern the mutual solubility at high temperatures (Gorsse et al., 2018; Huang et al., 2019; Li et al., 2020; Wang et al., 2021; Zhang et al., 2012, 2019). Based on these works, 15 design parameters were chosen (see Table 1), calculated, and included in the preliminary dataset, i.e., the mean atomic radius a , where c_i is the stoichiometric ratio and r_i is the atomic radius of the i -th component in the alloy, δ is the atomic size difference, K is the mean bulk modulus, where K_i is the bulk modulus of the i -th component in the alloy, σ_K is the bulk modulus's standard deviation, χ is the Pauling electronegativity, where χ_i is the Pauling electronegativity of the i -th component in the alloy, $\Delta\chi$ is the Pauling electronegativity standard deviation, T_m is the average melting temperature, where T_{mi} is the melting temperature of the i -th component in the alloy, σ_{T_m} is the melting temperature standard deviation, ΔH_{mix} is the mixing enthalpy, where H_{ij} is the binary mixing enthalpy in the liquid phase, $\sigma_{\Delta H}$ is the binary mixing enthalpy standard deviation, ΔS_{mix} is the mixing entropy, where R is the gas constant, G is the mean shear modulus, where G_i shear modulus of the i -th component in the alloy, VEC is the valence electron concentration, where VEC_i is the valence electron concentration of the i -th component in the alloy, σ_{VEC} is the valence electron concentration standard deviation, and E is the Young's modulus, where E_i is the Young's modulus of the i -th component in the alloy.

Table 1. HEAs design parameters

| Parameter | Equation |
|---|---|
| Mixing entropy | $\Delta S_{\text{mix}} = -R \sum_{i=1}^n c_i \ln c_i$ |
| Mixing enthalpy | $\Delta H_{\text{mix}} = 4 \sum_{i,j}^n c_i c_j H_{ij}$ |
| Standard deviation of mixing enthalpy | $\sigma_{\Delta H} = \sqrt{\sum_{i \neq j}^n (H_{ij} - \Delta H_{\text{mix}})^2}$ |
| Mean atomic radius | $a = \sum_{i=1}^n c_i r_i$ |
| Atomic size difference | $\delta = \sum_{i,j}^n c_i c_j \left(1 - \frac{r_i}{a}\right)^2$ |
| Electronegativity | $\chi = \sum_{i=1}^n c_i \chi_i$ |
| Electronegativity standard deviation | $\Delta \chi = \sqrt{\sum_{i=1}^n c_i (\chi_i - \chi)^2}$ |
| VEC | $VEC = \sum_{i=1}^n c_i VEC_i$ |
| Standard deviation of VEC | $\sigma_{VEC} = \sqrt{\sum_{i=1}^n c_i (VEC_i - VEC)^2}$ |
| Mean bulk modulus | $K = \sum_{i=1}^n c_i K_i$ |
| Standard deviation of bulk modulus | $\sigma_K = \sqrt{\sum_{i=1}^n c_i (K_i - K)^2}$ |
| Average melting temperature | $T_m = \sum_{i=1}^n c_i T_{mi}$ |
| Standard deviation of melting temperature | $\sigma_{T_m} = \sqrt{\sum_{i=1}^n c_i \left(1 - \frac{T_{mi}}{T_m}\right)^2}$ |
| Average Young's modulus | $E = \sum_{i=1}^n c_i E_i$ |
| Average Shear's modulus | $G = \sum_{i=1}^n c_i G_i$ |

Finally, 78 chemical elements and their corresponding fraction in the alloy were included, as well as the name and the number of chemical elements of each alloy, which led to a dataset structure of 96 columns. This dataset is already available online via open-access

at Zenodo (Precker et al., 2021a). However, after data pre-processing, some columns were not taken into account such as the name of the compounds, the number of chemical elements, and some of the 78 chemical elements to be part of the HEAs since they were not present in any of the alloys in the dataset. This meant that dimensionality was reduced for training to 73 vector components (see Table 3).

2.2. PDCs dataset

As for the case of the HEAs, it was necessary to define some characteristic parameters of PDCs related to their final composition. As mentioned before, the PDCs structure depends on external factors, with the initial precursor being an important one. It is normally an organosilicon polymer, but sometimes can also be an inorganic system. This initial polymer is shaped, crosslinked, and then the pyrolysis process takes place. Finally, as the last step, it takes the crystallization route to obtain the final ceramic. Apart from the precursor, pyrolysis conditions modify the final PDCs behaviour (Colombo et al., 2013). As a first approach, pyrolysis temperature, pyrolysis time, heating rate, dwelling time and gas atmosphere were selected as design parameters. For this purpose, PDCs found in the literature were collected, and pyrolysis conditions were studied. Despite this, after having collected all the possible information, there was a lack of data for heating rate and dwelling time, so finally, compounds with at least three of the features were taken into consideration, as can be seen in Table 2.

Table 2. PDCs design parameters

| Parameter | Description |
|-----------------------|---|
| Precursor | Initial polymer which after pyrolysis process will become a PDC |
| Pyrolysis temperature | Temperature at which pyrolysis process takes place |
| Pyrolysis time | Time that lasts the pyrolysis process |
| Gas atmosphere | Atmosphere where pyrolysis takes place |

In the training dataset, repeated final PDCs compositions are to be found, since different experimental conditions can lead to the same final polymer derived ceramic. For this reason, filtering was conducted to avoid duplicated compounds in the dataset. So, after all the above-mentioned processes, 181 PDCs were collected to form the dataset (Precker et al., 2021b), which is not a large number in comparison to HEAs, but due to the lack of data in the literature, it was not possible

to create a larger one. Apart from the design parameters shown in Table 2, the number of chemical elements, the name of the compounds, and the concentration of 78 chemical elements which could be a possible part of the compounds were also included in the dataset, together with those 4 mentioned features, which led to a dataset constituted of 85 columns which is open access in Zenodo (Precker et al., 2021b). As for HEAs, a pre-processing of data was done to reduce dimensionality for the training of the algorithm and at the end, only 17 dimensions remained (see Table 4), since many columns were removed, such as name of the compounds, the number of elements, and some chemical elements which were to be part of the final PDCs.

2.3. Neural network

A neural network (NN) can be fed with a dataset, so that it will learn the relationship between the features present in that dataset, i.e., what characterizes an output, and classify it accordingly. NNs are also able to create synthetic data that is very close to the real data, and that is exactly what the generative adversarial networks (GANs) do due to their ability to generate realistic fake content. A GAN is a generative model first used to create images (Goodfellow et al., 2014), but now the scope is extended to create other contents, e.g., furniture designs for 3D printing (González-Val & Muñíos-Landín, 2020).

The GANs work with two NN models, one competing against the other. One of the models is called the generator (G), responsible for generating synthetic data from a noisy entry z (a bunch of random values, e.g., randomized values from a normal distribution between 0 and 1),

which tries to generate a synthetic sample as close as possible to a real one. The other model is called discriminator (D), which is trained with both real and fake data, learning the difference between them, and classifying the data from the generator as real or fake (see Figure 3). The results from the discriminator's classification are used as input for the generator, which learns from these results and calibrates its weight to generate samples that appear closer to the real samples. After the generator improves its generated data, the discriminator is also improved, being updated by the new samples coming from the generator, calibrating its weights, and working as a loop, where the discriminator tries to become better at differentiating the real data from the generated.

Generative adversarial networks are a large family, but the most suitable GAN for this work is the conditional tabular generative adversarial network (CTGAN) (Xu et al., 2019). This specific kind of GAN provides solutions to data problems such as mixed data types, non-Gaussian distributions, multi-modal distributions, learning from sparse one-hot-encoded vectors, and highly imbalanced categorical columns, aspects which normal GANs do not address. Since the datasets are comprised of mixed types of data, containing discrete and continuous values, the CTGAN addresses the needs imposed by the data, and it can be used to generate new synthetic tabular data. The CTGAN can be conditioned using some extra information y , which can be any kind of auxiliary information, feeding the network with an additional input layer, e.g., class labels or data from other modalities. In a conditional GAN, the networks G and D are trained and optimized in an adversarial learning framework, called the objective function, as in Equation (2) (Mirza & Osindero, 2014).

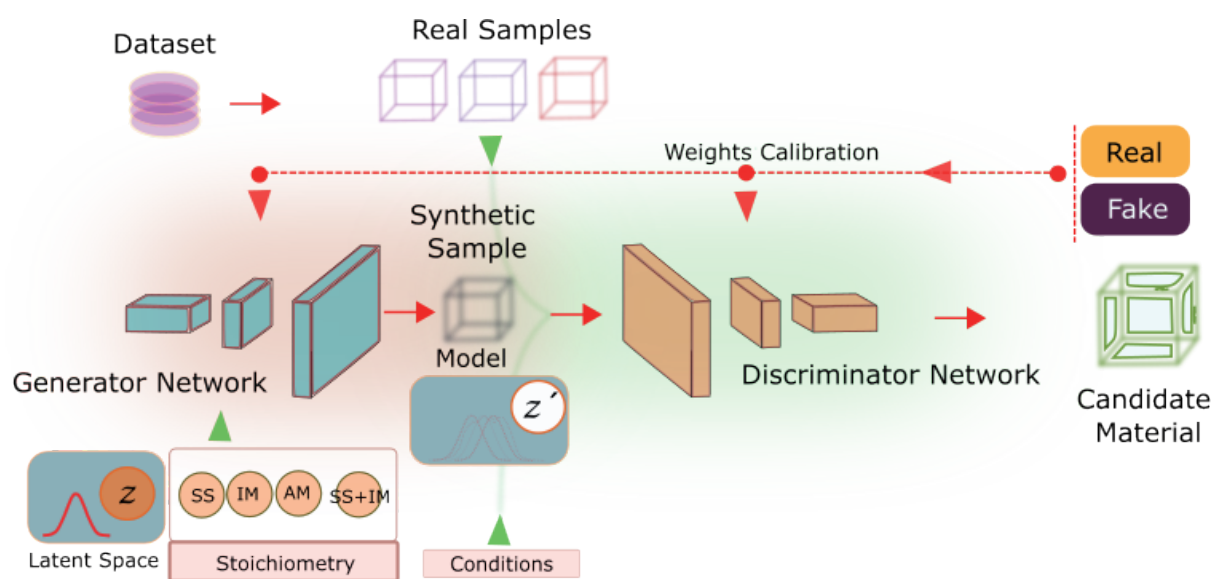


Fig. 3. Generative adversarial network (GAN) architecture composed of a generator G , and a discriminator D . The G takes a random entry z and creates new samples. D tries to differentiate real samples from generated ones

$$\min_G \max_D V(D, G) = E_{x \sim P_{data}(x)} [\log D(x | y)] + E_{z \sim P_z(z)} [\log(1 - D(G(z | y)))] \quad (2)$$

where x represents the real data, and y the conditioned information. In the conditional training, the CTGAN encodes the conditioned tabular data columns and categorical variables in condition vectors, using these vectors as generator inputs. This architecture uses recent GAN approaches where the quality and stability of the generated data are improved, e.g., it uses the discriminator of the PacGAN (Lin et al., 2018) and the loss function of the WGAN-GP (Gulrajani et al., 2017), defined as:

$$L = E_{G(z) \sim P_g} [D(G(z))] - E_{x \sim P_r} [D(x)] + \lambda E_{\hat{x} \sim P_{\hat{x}}} [\left(\|\nabla_{\hat{x}} D(\hat{x})\| - 1 \right)^2] \quad (3)$$

where: the two first terms are the original loss of the WGAN (Arjovsky et al., 2017) and the last term the gradient penalty loss, implemented to control the discriminator's gradient for random samples, $\hat{x} \sim P_{\hat{x}} \cdot \hat{x}$

represents samples that are interpolated by the real data, λ is the gradient coefficient penalty, and the distribution of the real and generated data are represented by P_r and P_g .

Figure 3 shows CTGAN's architecture sketch, comprised of the generator and discriminator models with the conditional entries used in this work, i.e., the phases, stoichiometry, and PDCs' design parameters to obtain the desired modes from the trained model. Table 3 and Table 4 summarize the used architecture. The following parameters were used in both G and D neural networks models: Adam optimizer with a learning rate of 2×10^{-4} and weight decay of 1×10^{-6} . For G , the ReLU activation was used in the input and hidden layers, and a softmax activation function in the case of the conditioned data, ensuring only positive values and a tanh activation function for the rest of the outputs. For D , the LeakyReLU activation was used in the input and hidden layers, and the sigmoid activation function in the output. Finally, the batch size was bigger for the HEAs' case than for the PDCs', due to the difference in the quantity of data used for training.

Table 3. CTGAN architecture for HEAs training

| Layer | Generator | | Discriminator | |
|----------|----------------|-----------|------------------|-----------|
| | type | dimension | type | dimension |
| Input | latent + cond. | 90 | features + cond. | 73 |
| Hidden 1 | dense layer | 256 | dense layer | 256 |
| Hidden 2 | dense layer | 128 | dense layer | 128 |
| Output | dense layer | 73 | dense layer | 1 |

Table 4. CTGAN architecture for PDCs training

| Layer | Generator | | Discriminator | |
|----------|----------------|-----------|------------------|-----------|
| | type | dimension | type | dimension |
| Input | latent + cond. | 50 | features + cond. | 17 |
| Hidden 1 | dense layer | 256 | dense layer | 256 |
| Hidden 2 | dense layer | 128 | dense layer | 128 |
| Output | dense layer | 17 | dense layer | 1 |

3. Results and validation

The CTGAN was fed for each material case with data from real compounds as input, containing the stoichiometry and the design parameters. The loss function for G and D versus the training epochs for each training episode is shown in Figure 4. The convergence of the loss function in both G and D for the HEAs' case occurs at approximately epoch 50, which means that from this point, the model reached a limit where G and D stopped

evolving. In the case of the PDCs, the convergence begins at approximately epoch 100.

The evaluation metrics used to get the score of the model were CSTest, KSTest, KSTestExtended, and ContinuousKLDivergence, which are statistical metrics found in the ecosystem of libraries of the synthetic data vault (SDV) (Patki et al., 2016). The average score value obtained from all these metrics together reached a value of 93% for HEAs and 81% for PDCs.

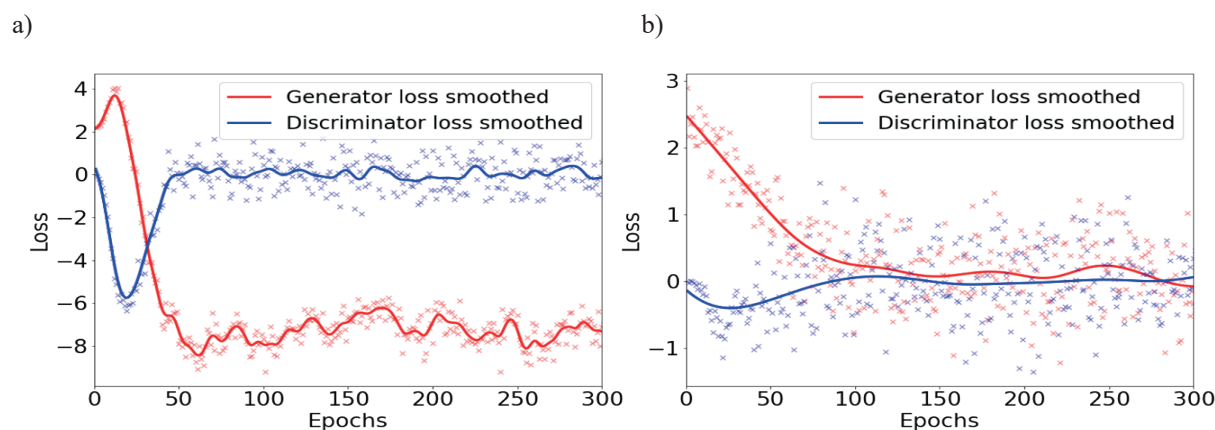


Fig. 4. Evolution of the loss functions for the: a) HEAs' case, where the convergence of discriminator and generator after epoch 50 means that one of the models stopped evolving, and consequently, the second model also stops to evolve since it's a competition between generating fake samples that look real and discriminating these samples as real and fake; b) PDCs' case where the convergence begins at approximately epoch 100

Once the model was trained, synthetic data based on the knowledge acquired during the training process was generated. For HEAs, the outputs provided by the CTGAN were the same 73 parameters used as inputs, i.e., the 15 design features, the 4 phases, and the columns containing the chemical elements fraction. For PDCs, the output was composed of 17 parameters, where there were the 4 design parameters and the columns referred to the stoichiometry.

3.1. Validation of high entropy alloys

Given the trained model with the created dataset, the neural network generated synthetic candidates, which must be evaluated in order to check if they accomplish the needed properties to be HEAs.

Within the generation, new compositions were also delivered by the CTGAN. Between the generated compounds, an experimentally known HEA that was not included in the initial dataset was generated, the TiZrCuNiBe, and the correct phase AM was attributed (Ding & Yao, 2013), which means that this method really opens the possibility of generating real HEAs. Some other examples of possible HEAs candidates (experimentally not proved) were generated, such as B_2CoGa_2VZr , $Al_{0.5}BCoCr_3FeMn$, $AlCoNdNiTi$, and $CoCuFeSn_3TiZn_{0.5}$.

Some of the present compounds in the initial dataset and some generated compounds were taken for evaluation in DFT-based open databases for materials, the Open Quantum Materials Database (OQMD) (Kirklin et al., 2015; Saal et al., 2013) and Automatic-FLOW for Materials Discovery (AFLOW) (Curtarolo et al., 2012). Figure 5 shows in the upper part four aleatory

selected compounds from the HEAs dataset (real compounds), and at the bottom, four compounds selected from the CTGAN generation (synthetic compounds). They are classified according to their phase, i.e., the real phase for the dataset compounds and the expected phase attributed by the CTGAN for the generated compounds. The bars inside the phase areas compare the mixing enthalpy ΔH for the real compounds (in the case of the dataset, calculated from Table 1) and synthetic with the mixing enthalpy calculated from the OQMD. When compared, the values are in good agreement in both cases for the real and synthetic data. Note that for comparison purposes, the mixing enthalpy modulus $|\Delta H|$ was used in Figure 5. Other generated compounds were found in the database of AFLOW, e.g., $AlCuTi$, $Al_{0.5}CuV$, $AlFeNi$, and $AlCrNiTi$, which once more validates the CTGAN as a generative candidate model for the discovery of novel HEAs.

An objective reinforced architecture was planned for the main GAN approach, including function performance parameters from the simulations as a reward. Under such an approach (ORGAN), a numeric function is defined to provide a reward or a penalty to the GAN architecture in order to promote some types of generated candidates over others. However, due to the interaction time between AI and simulation results, a Reinforcement Learning approach could not cope with such a low number of learning episodes. To overcome this limitation, a direct screening was done iteratively instead of a numeric function following different criteriums in terms of toxicity, material costs, and stability. This was to ensure that the developed tool only generates compounds that fulfil no toxicity, cost minimization, and stability assessment. Based on stability analysis, compounds that include Ni were always promoted,

ensuring that none of the chemical elements of the composition goes beyond 35% so that the HEA definition remains unviolated. This last conclusion was taken from diagram phases, so as it is convenient, the presence of gamma prime phase (FCC_L1₂) for applications at high temperatures, its presence was analysed. As it shows the phase diagram for Al_{0.15}Co_{0.19}Fe_{0.19}Ni_{0.28}Mn_{0.19} Figure 6a it was concluded that in comparison with other compounds such as Al_{0.36}Co_{0.17}Fe_{0.11}Ni_{0.2}Ti_{0.17} Figure 6b components with more Ni, had an adequate quantity of this mentioned phase at higher temperatures. Regard-

ing costs and toxicity Li, N, Na, P, Sc, V, Cr, Zn, Ga, Sr, Y, Zr, Nb, Mo, Ru, Rh, Cd, In, Sn, La, Ce, Pr, Sm, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Hf, Ta, W, Re, Pt, Pb and Bi were identified as elements to be avoided by the generator in the developed approach. Based on the high entropy alloy definition, compounds that have between 5 and 10 elements were taken into account. Finally, *VEC* values given in the output were used to check their accordance with those provided by stoichiometry. Screening removed those compounds that did not accomplish this requirement.

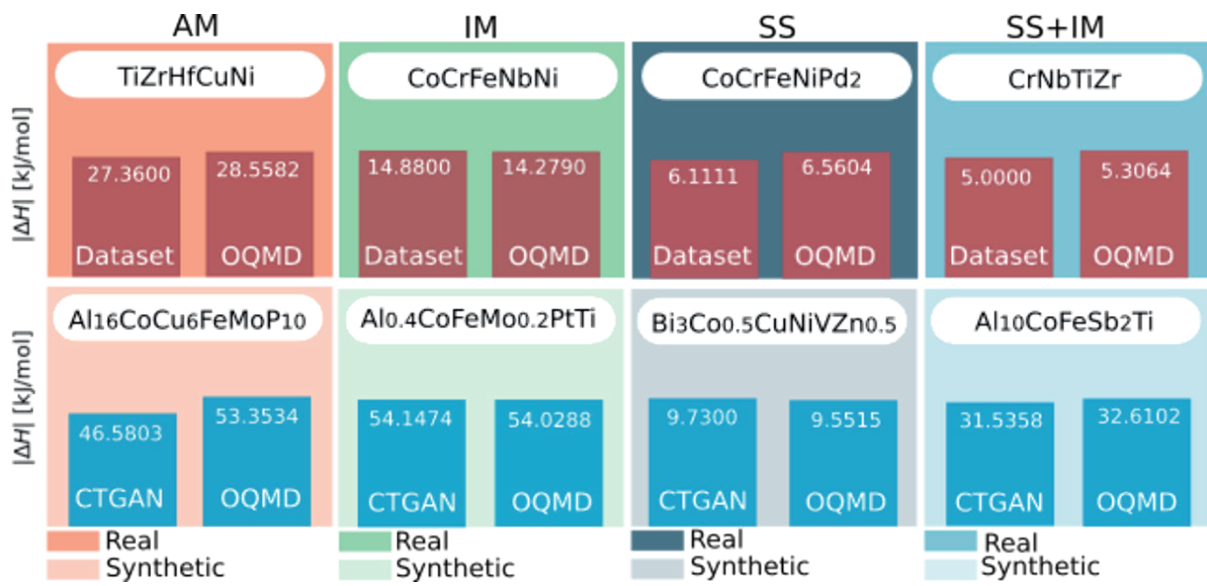


Fig. 5. Main results from the generative approach. Comparison of real values of ΔH for HEAs and those generated by our CTGAN. The figure shows results for the four different phases, AM, IM, SS, and SS+IM, separated in columns. Note that absolute values are shown (all the obtained values for ΔH are negative). The dark boxes contain compounds and ΔH values from real compounds (dark red columns), while those contained in the faded color boxes (blue columns) have been obtained with the CTGAN. The values of ΔH from the dataset and also those generated by the CTGAN are compared with the calculations performed using the OQMD

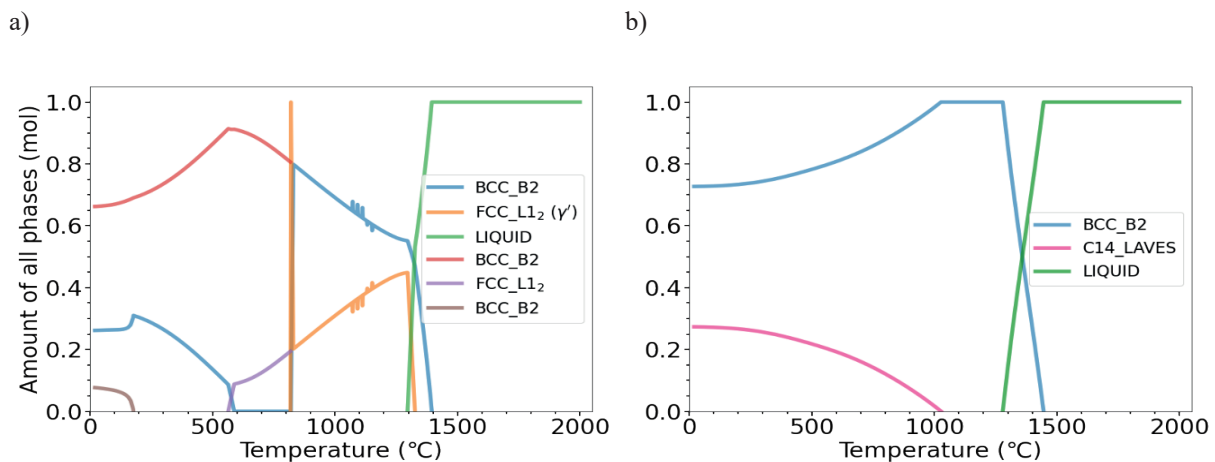


Fig. 6. Phase diagram of: a) Al_{0.15}Co_{0.19}Fe_{0.19}Ni_{0.28}Mn_{0.19} which presents the gamma prime phase (FCC_L1₂ (γ')) at higher temperatures than other compounds studied; b) Al_{0.36}Co_{0.17}Fe_{0.11}Ni_{0.2}Ti_{0.17} which does not present the gamma prime phase

3.2. Validation of polymer derived ceramics

Once the CTGAN generates samples as possible PDC candidates, some validation method was needed in order to assure that those synthetic compounds achieve the properties required of the desired material.

It is widely known that SiOC and SiCN systems are the most studied factors in this respect. SiOC systems present more negative enthalpy of a formation relative to their crystalline constituents. It ranges from -20 to -70 kJ/mol and some of them reach -128 kJ/mol, while SiCN system's enthalpy of formation is less negative, and some of them are near zero or slightly, turning them less stable than the SiOC systems (Wen et al., 2020). As a way of validation, some generated PDCs of both families were introduced in the DFT-based platform OQMD (Kirklin et al., 2015; Saal et al., 2013), so that their enthalpy of formation was computed, and with this, it was checked as expected that the ones from the SiOC family presented a more negative enthalpy of formation than those of the SiCN family, which is consistent with those synthetic candidates as possible PDC candidates.

Since the generation for PDCs was based on experimental features, external feedback was needed. Based on previous experimental conditions (Vry et al., 2020) a previous screening prior to validation was done. The pyrolysis temperature was set between 1000°C and 1400°C , the atmosphere used was argon, the precursor polysiloxane, and the pyrolysis time was no longer than 4 h due to production costs. Finally, only SiOC PDCs were considered since they show improved mechanical properties such as Young's modulus, hardness, and creep resistance (Wen et al., 2020). A way to compute the free carbon quantity was also found with a stoichiometric formula for this family of compounds,

so since a big amount of this can reduce some mechanical properties (Soraru et al., 2018; Sujith et al., 2021; Wen et al., 2020), candidates with no more of 50% of free carbon were taken into consideration.

After this screening process, a CALPHAD analysis with ThermoCalc was conducted on the generated PDCs to gain indications of which phases could be present at pyrolysis temperatures. Since ThermoCalc is not calibrated for this kind of material, the phase diagrams need to be considered indicative at best. The phase diagrams were generated depending on temperature and ternary ones at fixed temperatures.

Regarding the temperature-depending ones, SiC, SiO₂, and the graphite phases, the results found are in agreement with the literature. Then, taking into consideration the graphite phase (free carbon), its amount at room temperature was compared with the one computed using the stoichiometric formula mentioned in Table 5 (Martínez-Crespiera et al., 2011). It can be said that both values were in the same range for most of the cases, so it can be concluded that the stoichiometric formula is in accordance with the simulations.

For the ternary ones, as for the use case, it required a temperature above 750°C , fixed at 1400°C . The most stable compounds are expected to be localized in the middle of the diagram in Figure 8, so taking this into account alongside the amount of free carbon computed with the equations from Table 5, and having compared it with the quantity given by the phase diagram of Figure 7a at room temperature, it can be concluded that the compound SiO_{0.04}C (free carbon from Table 5: 1.96%) could be a good candidate. However, in the case of Figure 7b, there is no free carbon at pyrolysis temperature and it is not placed in the middle of Figure 8, so it can be concluded that it would not be a good candidate.

Table 5. Equations to compute mol fractions of SiO₂, SiC, and free carbon of SiOC systems

| Si _x O _y C _z | Amount [mol] | Mol fraction [%] |
|---|-------------------|---|
| SiO ₂ | $y/2$ | $\frac{y}{2z + y} \times 100$ |
| SiC | $x - (y/2)$ | $2 \left(\frac{\left[x - \left(\frac{y}{2} \right) \right]}{2z + y} \right) \times 100$ |
| "Free" C | $z - [x - (y/2)]$ | $2 \left(\frac{\left[z - \left[x - \left(\frac{y}{2} \right) \right] \right]}{2z + y} \right) \times 100$ |
| Total | $z + (y/2)$ | 100 |

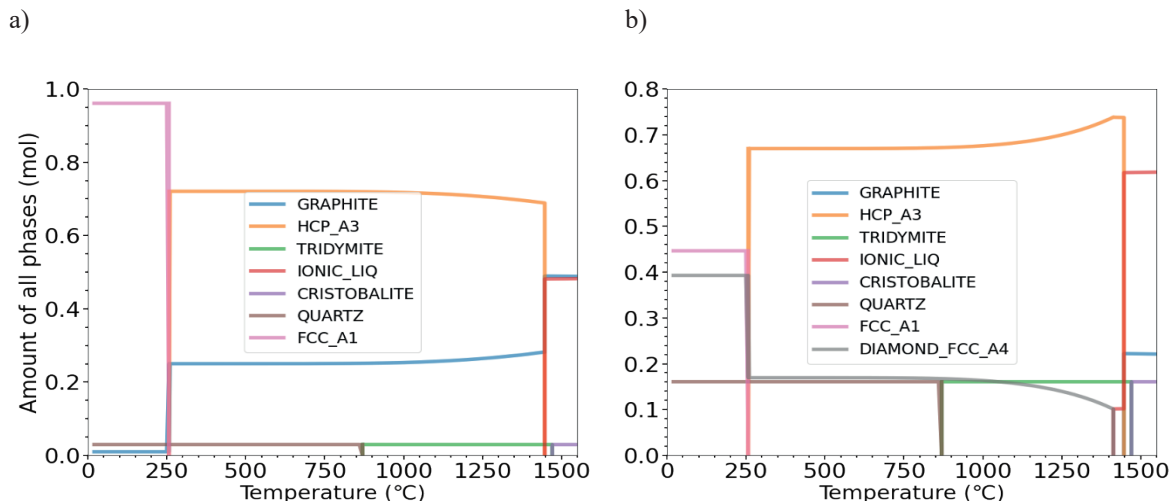


Fig. 7. Temperature dependence of the phases for the generated: a) PDC $\text{SiO}_{0.04}\text{C}$; b) PDC $\text{Si}_3\text{OC}_{0.48}$

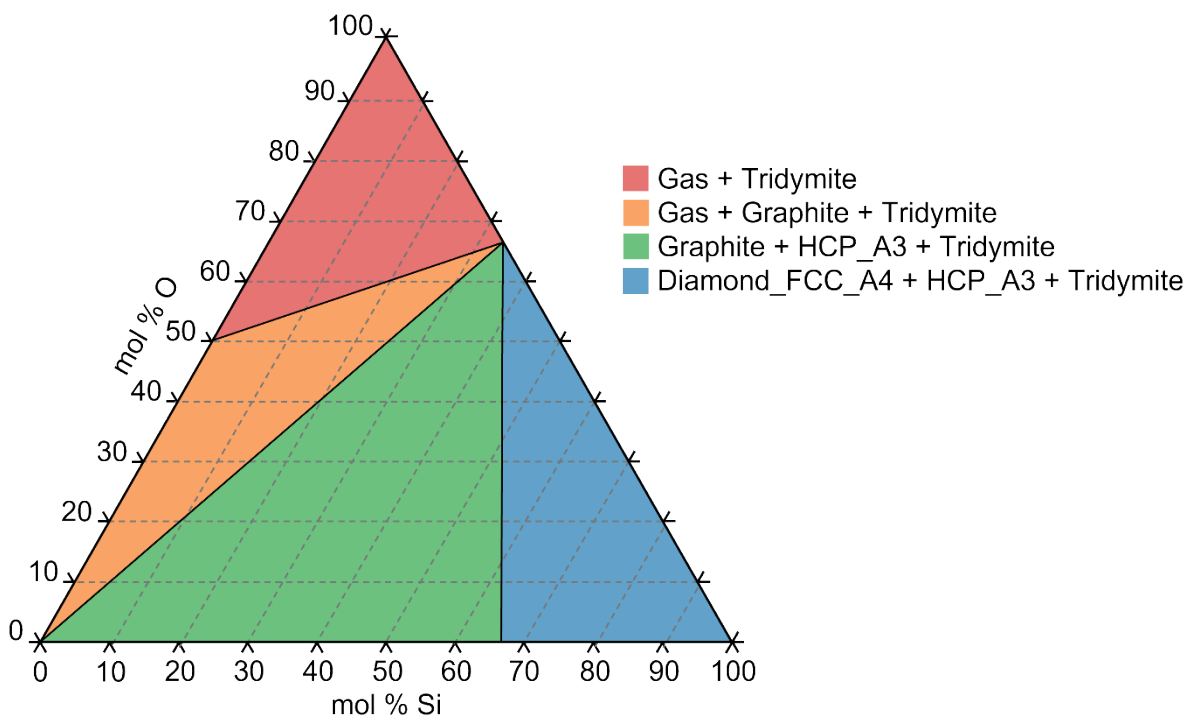


Fig. 8. Ternary diagram at a fixed temperature of 1400°C to predict the most stable PDC candidates

4. Conclusion

This work has presented how generative models can be used to design two different materials: high-entropy alloys and polymer derived ceramics. For that purpose, two specific datasets were designed and presented with a specific selection of features for each case based on the existing data. The outcome of the models has also been validated through computational methods, demonstrating the utility of the AI-based tool. The presented work might be an important contribution to the Gen-

erative-based methodologies (Menon & Ranganathan, 2022) for materials design and a first step towards the AI-driven design of high-performance materials based on industrial requirements or resource availability constraints. Once developed, the contrast between the time required for a compound discovery is profound. While months are typically required for experimentation and physical testing to drive the discovery of a material, now only in a few seconds, the method presented in this paper provides hundreds of possible alloys or PDCs. Future work focused on other types of materials will

reinforce the utility of such a tool, but also its integration within an easy-to-use and explainable framework will enhance the impact of developments like the one presented in this work.

Declaration of competing interest

The authors declare no conflict of interest.

Data availability

Datasets related to this article can be found at (Precker et al., 2021a, 2021b).

Acknowledgements



This research has received funding from the European Union's Horizon 2020 research and innovation programme under the project ACHIEF for the discovery of novel materials to be used in industrial processes with Grant Agreement 958374. The authors want to thank the comments and fruitful discussions with all the members of the Artificial Intelligence and Data Analytics Lab (AIDA-Lab) of the Smart Systems and Smart Manufacturing (S3M) and the Advanced Manufacturing Processes departments of the AIMEN Technology Centre.

References

- Ainger, F., & Herbert, J. (1960). The preparation of phosphorus-nitrogen compounds as non-porous solids. *Special Ceramics*, 168, 168–182.
- Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. *Proceedings of Machine Learning Research*, 70, 214–223.
- Barlow, H.B. (1989). Unsupervised learning. *Neural Computation*, 1(3), 295–311. <https://doi.org/10.1162/neco.1989.1.3.295>.
- Butler, K.T., Frost, J.M., Skelton, J.M., Svane, K.L., & Walsh, A. (2016). Computational materials design of crystalline solids. *Chemical Society Reviews*, 45(22), 6138–6146. <https://doi.org/10.1039/C5CS00841G>.
- Cantor, B., Chang, I.T.H., Knight, P., & Vincent, A.J.B. (2004). Microstructural development in equiatomic multicomponent alloys. *Materials Science and Engineering: A*, 375–377, 213–218. <https://doi.org/10.1016/j.msea.2003.10.257>.
- Chantrell, P., & Popper, P. (1964). Inorganic polymers for ceramics. In *Special Ceramics 1964* (pp. 87–103). Academic Press.
- Colombo, P., Mera, G., Riedel, R., & Sorarù, G.D. (2013). Polymer-derived ceramics: 40 years of research and innovation in advanced ceramics. In R. Riedel, I-Wei Chen (Eds.), *Ceramics Science and Technology* (vol. 4: *Applications*, pp. 245–320). Wiley-VCH Verlag GmbH & Co. KGaA. <https://doi.org/10.1002/9783527631971.ch07>.
- Cross, T.J., Raj, R., Cross, T.J., Prasad, S.V., & Tallant, D. R. (2006). Synthesis and tribological behavior of silicon oxycarbonyl thin films derived from poly(urea)methyl vinyl silazane. *International Journal of Applied Ceramic Technology*, 3(2), 113–126. <https://doi.org/10.1111/j.1744-7402.2006.02070.x>.
- Cunningham, P., Cord, M., & Delany, S.J. (2008). Supervised learning. In M. Cord, P. Cunningham (Eds.), *Machine Learning Techniques for Multimedia. Case Studies on Organization and Retrieval* (pp. 21–49). Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-75171-7_2.
- Curtarolo, S., Setyawan, W., Hart, G.L.W., Jahnatek, M., Chepulskii, R.V., Taylor, R.H., Wang, S., Xue, J., Yang, K., Levy, O., Mehl, M.J., Stokes, H.T., Demchenko, D.O., & Morgan, D. (2012). AFLOW: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58, 218–226. <https://doi.org/10.1016/j.commatsci.2012.02.005>.
- Ding, H.Y., & Yao, K.F. (2013). High entropy Ti₂₀Zr₂₀Cu₂₀Ni₂₀Be₂₀ bulk metallic glass. *Journal of Non-Crystalline Solids*, 364, 9–12. <https://doi.org/10.1016/j.jnoncrysol.2013.01.022>.
- Fu, S., Zhu, M., & Zhu, Y. (2019). Organosilicon polymer-derived ceramics: An overview. *Journal of Advanced Ceramics*, 8(4), 457–478. <https://doi.org/10.1007/s40145-019-0335-3>.
- González-Val, C., & Muñños-Landín, S. (2020). *Generative design for Social Manufacturing*. Zenodo. <https://zenodo.org/record/4597558>.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative adversarial nets*. Arxiv. <https://doi.org/10.48550/arXiv.1406.2661>.
- Gorsse, S., Nguyen, M.H., Senkov, O.N., & Miracle, D.B. (2018). Database on the mechanical properties of high entropy alloys and complex concentrated alloys. *Data in Brief*, 21, 2664–2678. <https://doi.org/10.1016/j.dib.2018.11.111>.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumolin, V., & Courville, A.C. (2017). Improved training of Wasserstein GANs. Arxiv. <https://doi.org/10.48550/arXiv.1704.00028>.
- Hart, G.L.W., Mueller, T., Toher, C., & Curtarolo, S. (2020). A roadmap for machine learning in alloy modelling. *Bulletin of the American Physical Society*, 65(1), X43.00012.
- Hart, G.L.W., Mueller, T., Toher, C., & Curtarolo, S. (2021). Machine learning for alloys. *Nature Reviews Materials*, 6(8), 730–755. <https://doi.org/10.1038/s41578-021-00340-w>.
- Huang, W., Martin, P., & Zhuang, H.L. (2019). Machine-learning phase prediction of high-entropy alloys. *Acta Materialia*, 169, 225–236. <https://doi.org/10.1016/j.actamat.2019.03.012>.

- Kirklin, S., Saal, J.E., Meredig, B., Thompson, A., Doak, J.W., Aykol, M., Rühl, S., & Wolverton, C. (2015). The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *Npj Computational Materials*, 1(1), 15010. <https://doi.org/10.1038/npjcompumats.2015.10>.
- Lee, S.Y., Byeon, S., Kim, H.S., Jin, H., & Lee, S. (2021). Deep learning-based phase prediction of high-entropy alloys: Optimization, generation, and explanation. *Materials & Design*, 197, 109260. <https://doi.org/10.1016/j.matdes.2020.109260>.
- Li, L., Fang, Q., Li, J., Liu, B., Liu, Y., & Liaw, P. K. (2020). Lattice-distortion dependent yield strength in high entropy alloys. *Materials Science and Engineering: A*, 784, 139323. <https://doi.org/10.1016/j.msea.2020.139323>.
- Lin, Z., Khetan, A., Fanti, G., & Oh, S. (2018). *PacGAN: The power of two samples in generative adversarial networks*. Arxiv. <https://doi.org/10.48550/arXiv.1712.04086>
- Martínez-Crespiera, S., Ionescu, E., Kleebe, H.-J., & Riedel, R. (2011). Pressureless synthesis of fully dense and crack-free SiOC bulk ceramics via photo-crosslinking and pyrolysis of a polysiloxane. *Journal of the European Ceramic Society*, 31(5), 913–919. <https://doi.org/10.1016/j.jeurceramsoc.2010.11.019>.
- Mazheika, A., Wang, Y.-G., Valero, R., Viñes, F., Illas, F., Ghiringhelli, L.M., Levchenko, S. V., & Scheffler, M. (2022). Artificial-intelligence-driven discovery of catalyst genes with application to CO₂ activation on semiconductor oxides. *Nature Communications*, 13(1), 419. <https://doi.org/10.1038/s41467-022-28042-z>.
- Menon, D.B., & Ranganathan, R. (2022). A generative approach to materials discovery, design, and optimization. *ACS Omega*, 7(30), 25958–25973. <https://doi.org/10.1021/acsomega.2c03264>.
- Miracle, D.B., & Senkov, O.N. (2017). A critical review of high entropy alloys and related concepts. *Acta Materialia*, 122, 448–511. <https://doi.org/10.1016/j.actamat.2016.08.081>.
- Miracle, D.B., Miller, J.D., Senkov, O.N., Woodward, C., Uchic, M.D., & Tiley, J. (2014). Exploration and development of high entropy alloys for structural applications. *Entropy*, 16(1), 494–525. <https://doi.org/10.3390/e16010494>.
- Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *Arxiv*. <https://doi.org/10.48550/arXiv.1411.1784>.
- Mitchell, T., Buchanan, B., DeJong, G., Dietterich, T., Rosenbloom, P., & Waibel, A. (1990). Machine learning. *Annual Review of Computer Science*, 4, 417–433. <https://doi.org/10.1146/annurev.cs.04.060190.002221>.
- Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 399–410. <https://doi.org/10.1109/DSAA.2016.49>.
- Precker, C.E., Gregores Coto, A., & Muñios-Landín, S. (2021a). *Materials for design open repository. High entropy alloys*. Zenodo. <https://zenodo.org/record/6403257>.
- Precker, C.E., Gregores Coto, A., & Muñios-Landín, S. (2021b). *Materials for design open repository. Polymer derived ceramics*. Zenodo. <https://zenodo.org/record/5801992>.
- Riedel, R., Mera, G., Hauser, R., & Kloneczynski, A. (2006). Silicon-based polymer-derived ceramics: Synthesis properties and applications – a review. *Journal of the Ceramic Society of Japan*, 114(1330), 425–444. <https://doi.org/10.2109/jcersj.114.425>.
- Saal, J.E., Kirklin, S., Aykol, M., Meredig, B., & Wolverton, C. (2013). Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM*, 65(11), 1501–1509. <https://doi.org/10.1007/s11837-013-0755-4>.
- Shevlin, S., Castro, B., & Li, X. (2021). Computational materials design. *Nature Materials*, 20(6), 727–727. <https://doi.org/10.1038/s41563-021-01038-8>.
- Soraru, G.D., Kundanati, L., Santhosh, B., & Pugno, N. (2018). Influence of free carbon on the Young's modulus and hardness of polymer-derived silicon oxycarbide glasses. *Journal of the American Ceramic Society*, 102(3), 907–913. <https://doi.org/10.1111/jace.16131>.
- Sujith, R., Jothi, S., Zimmermann, A., Aldinger, F., & Kumar, R. (2021). Mechanical behaviour of polymer derived ceramics – a review. *International Materials Reviews*, 66(6), 426–449. <https://doi.org/10.1080/09506608.2020.1784616>.
- Tong, X., Liu, X., Tan, X., Li, X., Jiang, J., Xiong, Z., Xu, T., Jiang, H., Qiao, N., & Zheng, M. (2021). Generative models for de novo drug design. *Journal of Medicinal Chemistry*, 64(19), 14011–14027. <https://doi.org/10.1021/acs.jmedchem.1c00927>.
- Tsai, M.-H., Tsai, R.-C., Chang, T., & Huang, W.-F. (2019). Intermetallic phases in high-entropy alloys: Statistical analysis of their prevalence and structural inheritance. *Metals*, 9(2), 247. <https://doi.org/10.3390/met9020247>.
- Vaidya, M., Muralikrishna, G.M., & Murty, B.S. (2019). High-entropy alloys by mechanical alloying: A review. *Journal of Materials Research*, 34(5), 664–686. <https://doi.org/10.1557/jmr.2019.37>.
- Verbeek, W. (1974). Production of shaped articles of homogeneous mixtures of silicon carbide and nitride. *US Patent US3853567A*.
- Verbeek, W., & Winter, G. (1974). Formkoerper aus siliciumcarbid und verfahren zu ihrer herstellung. *DE Patent, 2236078A1*.
- Vry, S., Roumanic, M., Laucournet, R., & Bernard-Granger, G. (2020). Transmission electron microscopy investigations on a polysiloxane preceramic polymer pyrolyzed at high temperature in argon. *Ceramics*, 3(4), 421–427. <https://doi.org/10.3390/ceramics3040035>.
- Wang, Z., Guo, S., & Liu, C. T. (2014). Phase selection in high-entropy alloys: from nonequilibrium to equilibrium. *JOM*, 66(10), 1966–1972. <https://doi.org/10.1007/s11837-014-0953-8>.
- Wang, X., Guo, W., & Fu, Y. (2021). High-entropy alloys: emerging materials for advanced functional applications. *Journal of Materials Chemistry A*, 9(2), 663–701. <https://doi.org/10.1039/D0TA09601F>.
- Wen, Q., Yu, Z., & Riedel, R. (2020). The fate and role of in situ formed carbon in polymer-derived ceramics. *Progress in Materials Science*, 109, 100623. <https://doi.org/10.1016/j.pmatsci.2019.100623>.
- Winter, G.D., Verbeek, W.D., & Mansmann, M.D. (1974). Formkoerper aus homogenen mischungen von siliciumcarbid und siliciumnitrid und verfahren zu ihrer herstellung. *DE Patent, DE2243527A1*.
- Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). *Modeling tabular data using conditional gan*. Arxiv. <https://doi.org/10.48550/arXiv.1907.00503>.

- Yajima, S., Okamura, K., & Hayashi, J. (1975). Structural analysis in continuous silicon carbide fiber of high tensile strength. *Chemistry Letters*, 4(12), 1209–1212. <https://doi.org/10.1246/cl.1975.1209>.
- Yajima, S., Hasegawa, Y., Okamura, K., & Matsuzawa, T. (1978). Development of high tensile strength silicon carbide fibre using an organosilicon polymer precursor. *Nature*, 273(5663), 525–527. <https://doi.org/10.1038/273525a0>.
- Yeh, J.-W. (2013). Alloy design strategies and future trends in high-entropy alloys. *JOM*, 65(12), 1759–1771. <https://doi.org/10.1007/s11837-013-0761-6>.
- Yeh, J.-W. (2015). Physical metallurgy of high-entropy alloys. *JOM*, 67(10), 2254–2261. <https://doi.org/10.1007/s11837-015-1583-5>.
- Yeh, J.-W., Chen, S.-K., Lin, S.-J., Gan, J.-Y., Chin, T.-S., Shun, T.-T., Tsau, C.-H., & Chang, S.-Y. (2004). Nanostructured high-entropy alloys with multiple principal elements: novel alloy design concepts and outcomes. *Advanced Engineering Materials*, 6(5), 299–303. <https://doi.org/10.1002/adem.200300567>.
- Zhang, Y., Yang, X., & Liaw, P. K. (2012). Alloy design and properties optimization of high-entropy alloys. *JOM*, 64(7), 830–838. <https://doi.org/10.1007/s11837-012-0366-5>.
- Zhou, Z., Zhou, Y., He, Q., Ding, Z., Li, F., & Yang, Y. (2019). Machine learning guided appraisal and exploration of phase design for high entropy alloys. *Npj Computational Materials*, 5(1), 128. <https://doi.org/10.1038/s41524-019-0265-1>.