ISBN: 978-628-95207-4-3. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2023.1.1.651

Evolutive screening of candidates for new materials using genetic algorithms and deep learning

David Tatis Posada, B.E.[1], María Ramos Álamo, B.S. [2], Heidy Sierra, Ph.D.[3], and Emmanuel Arzuaga, Ph.D.[4]
[1]Department of Electrical and Computer Engineering, University of Puerto Rico, Mayagüez, Puerto Rico, david.tatis@upr.edu
[2]Graduate Program of Biongineering, University of Puerto Rico, Mayagüez, Puerto Rico, maria.ramos21@upr.edu
[3,4]Department of Computer Science and Engineering, University of Puerto Rico, Mayagüez, Puerto Rico, heidy.sierra1@upr.edu, emmanuel.arzuaga@upr.edu

***Abstract*—Different mechanisms are used for the discovery of materials. These include creating a material by trial-and-error process without knowing its properties. Other methods are based on computational simulations or mathematical and statistical approaches, such as Density Functional Theory (DFT). A well-known strategy combines elements to predict their properties and selects a set of those with the properties of interest. Carrying out exhaustive calculations to predict the properties of these found compounds may require a high computational cost. Therefore, there is a need to create methods for identifying materials with a desired set of properties while reducing the search space and, consequently, the computational cost. In this work, we present a genetic algorithm that can find a higher percentage of compounds with specific properties than state-of-the-art methods, such as those based on combinatorial screening. Both methods are compared in the search for ternary compounds in an unconstrained space, using a Deep Neural Network (DNN) to predict properties such as formation enthalpy, band gap, and stability; we will focus on formation enthalpy. As a result, we provide a genetic algorithm capable of finding up to 60% more compounds with atypical values of properties, using DNNs for their prediction.**

***Keywords*— genetic algorithm, deep learning, materials discovery, materials properties, combinatorial screening**

I. Introduction

The discovery of materials is an essential task in the world of science since it allows for creating technological elements that are more efficient, resistant, economical, and environmentally friendly. Different mechanisms have been developed to discover materials based on various methods, including mathematical, statistical, and, more recently, Deep Learning (DL). Among the mathematical and statistical methods is the Density Functional Theory (DFT), which is mainly based on the two theorems by Hohenberg and Kohn [1] [2]. The first theorem states that the ground-state electron density determines the electronic wave function and, consequently, all ground-state properties of an electronic system. The second theorem states that the energy of an electron distribution can be described as a functional of the electron density, which is a minimum for the ground-state density [3].

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

Calculations through DFT have great utility in multiple areas of science [4]. DFT calculations for multiple compounds, are used to create databases such as The Open Quantum Materials OQMD is a collection of consistently calculated DFT total energies and relaxed crystal structures [5] [6]. In addition, it provides the calculation for the DFT thermodynamic and structural properties of 1,022,603 materials, and this total keeps increasing.

A. DNNs methods

Machine Learning (ML) and Deep Neural Networks (DNN) models have been used for different tasks [7] by using available information to predict a behavior or make decisions. DNNs perform better than traditional ML methods in some applications, such as predicting properties in chemical compounds. In [8], the authors compare methods such as random forest and a DNN showing that DNN methods lower Mean Absolute Error (MAE). Also, the authors compare different amounts of data, finding that for a small number of data, the ML methods obtain a better performance than the DNN.

In our preliminary work, a deep neural network based on a ElemNet [8] model was implemented to predict material properties using only the elemental composition. The model was trained with the OQMD dataset to predict formation enthalpy. Its performance was compared with conventional ML approaches. The results showed that a better performance in terms of speed and accuracy was obtained by using the DNN model.

The deep regression network with individual residual learning (IRNet) [9] is another model that has been used for material discovery. This model has been trained and evaluated with the OQMD and Materials Project (MP) data. The model was trained to predict formation enthalpy, bandgap, energy, and volume properties. The results show an improvement of the performance when compared to traditional ML techniques such as Random Forest, Kernel Ridge Regression, Lasso, and Support Vector Machine. A comparison between a plain network and a stacked residual network (SRNet) with shortcut connections was also performed after a stack of multiple layers [9].

DNN models with different architectures have recently been published [10, 11]. Further work [12], includes variations of the aforementioned models to improve the accuracy on predicting the formation enthalpy. The main characteristic of these

This work was supported by the National Science Foundation, Award No. OIA-1849243 and Grant No. OAC-1750970s.

networks is that they have residual jumps between the layers, allowing the increase of parameters without having the well-known gradient problem of the DNNs [13]. Nevertheless, DNN models are expected to be less accurate than the experimental calculations made, or by using the DFT [14]. Thus, the method to achieve such task can be selected on the required confidence level.

B. Material discovery

Currently, there are several challenges to achieve a systematic method for the discovery of materials. Such as computational limitations and the specific knowledge that some methods require [15]. For example, in the case of ternary or higher-order compounds, making combinations of elements to find some that have specific properties is an intractable task [14]. Therefore, using mathematical methods such as DFT to perform a combinatorial screening remains [15]. Machine Learning (ML) and DL methods provide an alternative to scan millions of compounds and ranking them in term of the predicted property [15] [16]. In this context DNNs methods are capable of predicting properties with a small error compared to ML methods [8]. More recent work consists of implementing methods that vary the number of elements and restricting the number of atoms. Subsequently, they predict the properties for stable compounds with a Convex Hull [8]. Performing a combinatorial screening using 86 elements and some restrictions leads to the prediction of around 450M compounds (binary, ternary and quaternary).

Recently, another screening method has been implemented, changing how elements are combined [16]. For example, authors in [16] propose a greedy screening, which initially selects a material, selects the constituent elements of an identified material, and performs all possible combinations of these elements. This method uses databases to select initial materials and materials with combinations of the elements of interest. Despite its innovation in the greedy algorithm, searching databases can cause limitations.

C. Genetic Algorithm (GA)

Material discovery can be described as a combinatorial optimization problem. A combination of elements results in a compound that may or may not have a target property [17]. The number of elements defines the space dimensionality, for example in [16], the combinatorial problem is seen as an n-dimensional knapsack problem. To solve this type of problem evolutionary algorithms are a good alternative [18].

Genetic algorithms (GA) are a type of evolutionary algorithms that were introduced several decades ago primarily as a stochastic method for solving combinatorial and optimization problems. GAs have been used for polymer design [21], vehicle routing problems [22], designing mixed refrigerant cryogenic processes [23], and more applications in the field of materials science and engineering.

GAs start from an initial population, in which crossover and mutation operations are subsequently performed. For the discovery of materials, these operations generate a new evolved population with the potential of containing the target material properties [20] [19].

II. METHODS

The materials studied in [8] are binary (A_wB_x), ternary ($A_wB_xC_y$), and quaternary ($A_wB_xC_yD_z$) compounds. All the possible combinations of compounds were considered under the following restrictions: *A*, *B*, *C*, and *D* represent one out of the 86 possible elements in OQMD, where the order of elements varies based on the electronegativity (Equation 1); *w*, *x*, *y*, and *z* are positive integers representing the amount of the corresponding element in the composition, and satisfy Equation 2. Here, we only consider the ternary compounds; Equation 3 shows the total of possible combinations.

$$ABC \implies \binom{86}{3} = 1{,}022{,}34 \tag{1}$$

$$|\{w + x + y \leq 10 \mid w, x, y \in \mathrm{N}\}| = 109 \tag{2}$$

$$A_wB_xC_y \Rightarrow 1{,}022{,}340 \times 109 = 11{,}155{,}060 \tag{3}$$

A. DNNs to predict formation enthalpy

The IRNET-CV architecture has been reported as a model to predict material properties such as stability and bandgap [12]. In this work a IRNET-CV of 17 was used to predict the formation enthalpy property (Figure 1). The model receives a list of 86 elements as an input, where each position of the list represents an element of the periodic table. Each element is described by the element's percentage of atoms in the compound.

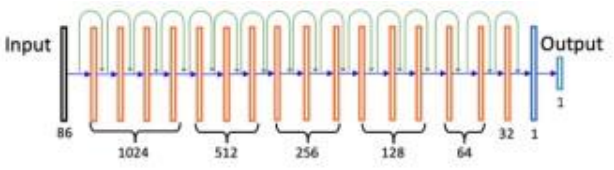

Fig. 1. IRNET Architecture: The orange rectangles represent fully connected layers (FC) with a batch normalization and a Relu activation function. The blue rectangle represents a FC layer with a linear activation function. The numbers in the bottom are the number of units of each layer.

The model was trained with data from the OQMD. For the training and testing phases 307,305 and 34,145 compounds were used respectively. This resulted in a MAE of 0.035 similarly to the FCUnet, and FCMnR architectures reported in the literature [12]. One of the advantages of DNN models is that a trained architecture allows to load different weights (compound elements distribution) to predict different properties.

B. Genetic Algorithm

To address the problem of finding materials with specific properties, we build a GA using five functions as follow: Initialize population, evaluate fitness, Selection, Crossover, and New population (reproduction) as described in Figure 2.

1) Initialize population: This function randomly generates N compounds each of three elements without repetitions, with

a random proportion of atoms (to satify equation 2). The compount are generated with the following elements (the same set of elementos used in DNNs training) :*H, Li, Be, B, C, N, O, F, Na, Mg, Al, Si, P, S, Cl, K, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Ge, As, Se, Br, Kr, Rb, Sr, Y, Zr, Nb, Mo, Tc, Ru, Rh, Pd, Ag, Cd, In, Sn, Sb, Te, I, Xe, Cs, Ba, La, Ce, Pr, Nd, Pm, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Hf, Ta, W, Re, Os, Ir, Pt, Au, Hg, Tl, Pb, Bi, Ac, Th, Pa, U, Np, Pu. The N compounds are returned in a N x 86 matrix.*

2) Evaluate fitness: This function will be executed in each generation. The model and the compound properties are evaluated. We make use of the IRNET trained model to predict the formation for the N compounds. The output of this function is an array of dimensions (N,n_properties); in our case (N,1).

3) Selection: This function receives the predictions of the compounds and a range of values to evaluate the property. For our case, the minimum and maximum value of formation enthalpy was used. This function can be modified to predict only stable compounds or those with other specific properties. The returned elements are stored as possible desired compounds and are part of the final output of the algorithm.

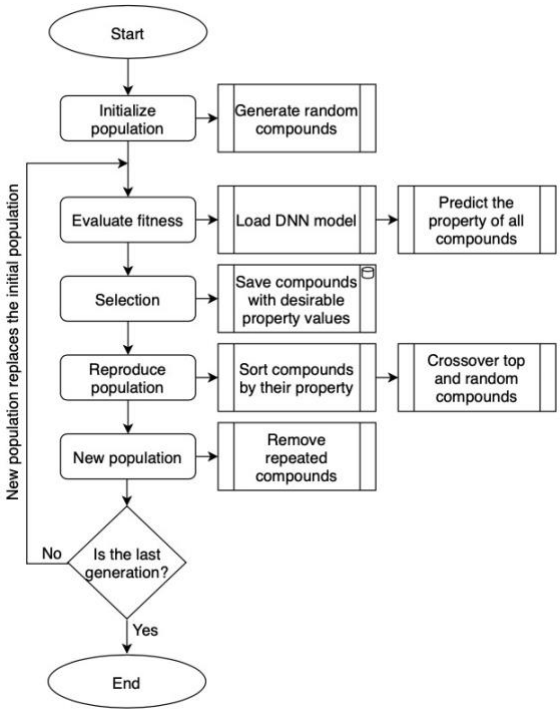

Fig. 2. Diagram of the Genetic Algorithm proposed for the discovery of candidate materials.

4) Crossover: Up to this point, the algorithm remains a combinatorial screening. After executing this and the following functions, the GA evolves the search after executing multiple generations.

In the crossover function, two compounds are received, and their elements are combined to create new child compounds, while maintaining the same proportion of atoms. In the case of ternary compounds, we replace the first parent's element with the second parent's first element, resulting in a new compound. Then the second element of the first compound is exchanged for the second element of the second compound. Finally, the same process is repeated with the third element. Subsequently, the same procedure uses the second parent as a base compound. Figure 3 shows a particular case of two randomly generated compounds. Their vector representation is shown for each one, having the proportion of atoms in the position of each element. Figure 4 shows the resulting compounds after the crossover operations described above, with a total of 6 Child compounds.

5) Reproduce population: This function ranks the compounds with properties equal to or close to the target. Those compounds with properties outside the range are also randomly selected. Next, the top and random compounds are selected as parents to enhance a crossover of all these. After obtaining all the children of the compounds, we eliminate repetitions and finally evalute the new generation. Once this new generation of compounds is created, the steps of evaluating the population's fitness, selection, and reproduction are repeated. The number of times this process is performed corresponds to the number of generations.

The GA is created parametrically, thus that performing experiments by varying the compounds properties becomes less complex. The main variables of the algorithm are min property, max property, model, random parents, top parents, generations, and the number of the initial population. In the next section, we will show different experiments that include the variation of these parameters.

Fig. 3. Example of parents to be used in the Crossover function.

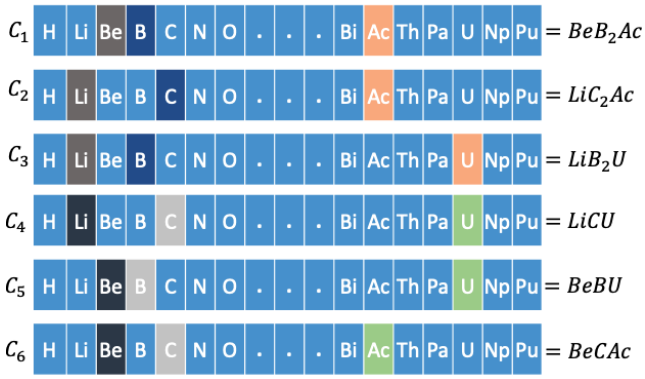

Fig. 4. Resulting children generated by the Crossover function

III. RESULTS

Initially, we perform an exhaustive combinatorial screening, testing all possible combinations of ternary compounds with the abovementioned restrictions. The prediction of the formation enthalpy was performed for 11,155,060 ternary generated compounds. The predicted average value of this property was −0.043 with a standard deviation of 0.717 (see Figure 5).

A first experiment, consisted of searching for compounds with different values of formation enthalpy, ranging from medium values such as −0.05 and 0.05 to less common values such as −1.25 and −1.20. As we see in Figure 6, the selection percentage per generation increases after the first generation. Although the selection for the last generations is between 15% and 20%, some compounds may be present in more than one generation. Therefore, in the following experiments, we focus on the percentage of finds instead of selections. Figure 5 shows that the values from −1.25 to −1.20 are more challenging to find.

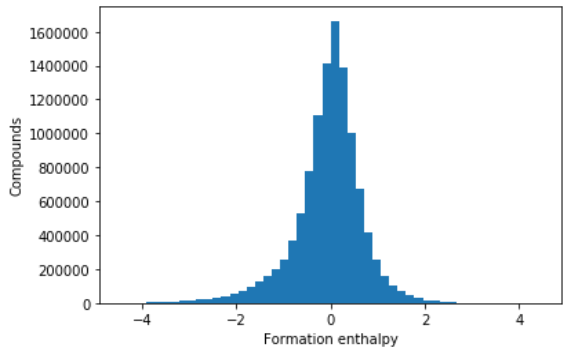

Fig. 5. Formation enthalpy of ternary compounds predicted by IRNet.

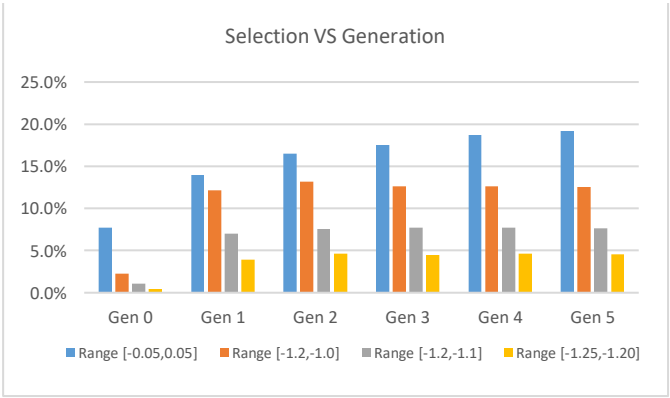

Fig. 6. Selection percentage for each generation and each range of desired formation enthalpy. GAs use 1M compounds as the initial population, 500 top parents, and five generations.

Another experiment consisted of generations (GA)using unique compounds. The percentage of finds for each range was compared to the results obtained with the combinatorial screening method (generation 0). As shown in Figure 7 the GA can find 60% more unique compounds than the combinatorial screening for those property values far from the mean value.

Another experiment consisted of varying the number of generations. For this, we used an initial population of 1.5M. Figure 8 shows a comparison of the percentage of selections and the finds per generation. It is observed that around the 3rd generation, the percentage of finds starts to decrease.

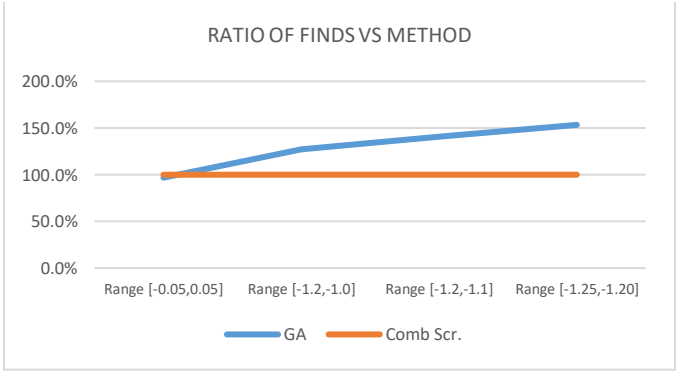

Fig. 7. Percentage of the difference between GA and combinatorial screening for each range of formation enthalpy. GA uses 1M compounds as the initial population, 500 top parents, and five generations.

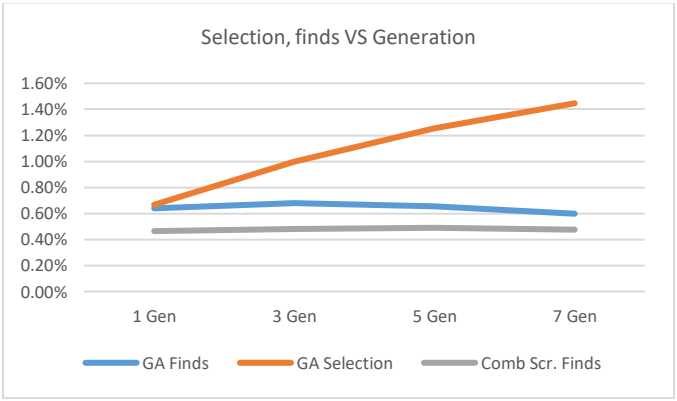

Fig. 8. Selection percentage and finding percentage using different numbers of generations, GA using 1.5M compounds as the initial population and 500 top parents.

These experiments were performed with 500 top parents. In other words, the new generations were created after a crossover of these 500 compounds for the formation enthalpy.

Figure 9 shows the percentage of finds of the GA after varying the number of top parents for an initial population of 1M compounds. It should be noted that if the number of parents is greater than the number of selected compounds (in the property range), then the crossover is performed for the last number of compounds. If the initial population decreases, fewer possible parents are found, and the performance is therefore changed. Finally, we find that for several cases, increasing the number of

random parents (without satisfying the property in the range) decreases the percentage of finds; this can be seen on Figures 10 and 11.

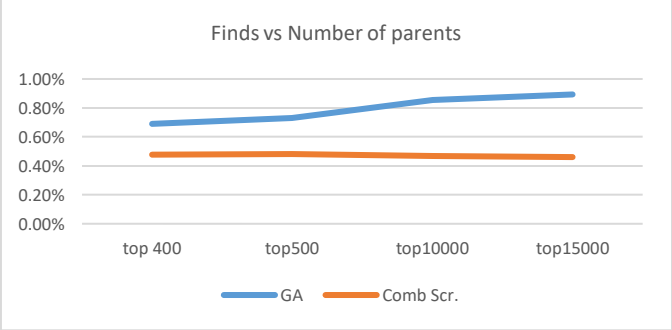

Fig. 9. Finds percentage using different amounts of top parents, GA using 1M compounds as the initial population and five generations.

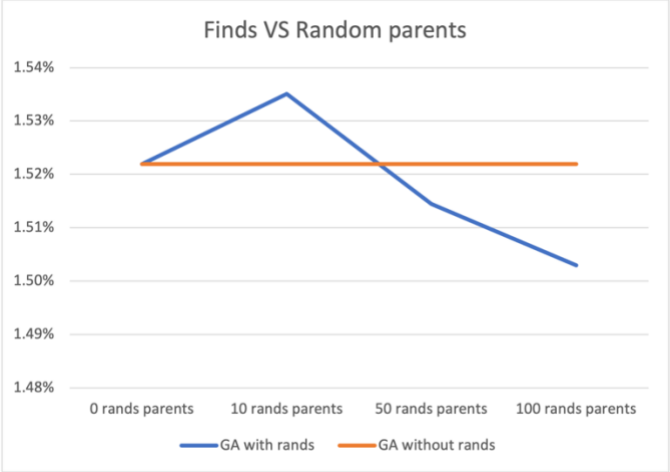

Fig. 10. Finds percentage using different amount of top and random parents. GA using 100K compounds as initial population and 5 generations.

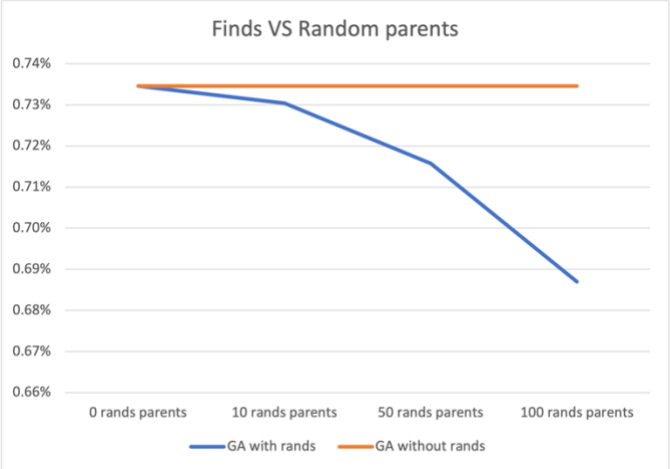

Fig. 11. Finds percentage using different amount of top and random parents. GA uses 1M compounds for the initial population and 5 generations.

Figure 10 shows an apparent improvement in the percentage of finds when using up to ten random parents and 490 top parents; however, after ten random parents the find percentage starts to decrease. Similarly, as the initial population is increased to 1M, there is a decrease in the number of finds, as shown in Figure 11.

IV. CONCLUSIONS AND FUTURE WORK

After the different experiments with the GA, we found a higher percentage of finds in the combinatorial screening method, for compounds that are randomly generated. After experimenting with different search ranges of the formation enthalpy property, we determined that the GA is more effective in finding compounds. However, we note that the crossover process between compounds with the same properties will likely generate more compounds with similar properties values. Even though there were repetitions between generations, the single population is still greater than that found by the combinatorial screening. Additionally, including random compounds with properties outside the desired range does not significantly impact the number of finds.

Although the GA requires a smaller number of compounds to be analyzed there is more complexity in performing the crossover and reproduction when compared to the combinatorial screening method. However, if a complex method such as the DFT is used to predict the properties, it is expected that the GA will perform faster.

The focus of these experiments was on the number of compounds that can be found for specific properties. In future work, an analysis of time and computational complexity will be performed.

REFERENCES

[1] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Physical Review*, vol. 136, no. 3B, p. B864, 1964.
[2] N. Argaman and G. Makov, "Density functional theory: An introduction," *American Journal of Physics*, vol. 68, no. 1, pp. 69–79, 2000.
[3] M. Orio, D. A. Pantazis, and F. Neese, "Density functional theory," *Photosynthesis Research*, vol. 102, no. 2, pp. 443–453, 2009.
[4] J. Hafner, C. Wolverton, and G. Ceder, "Toward computational materials design: the impact of density functional theory on materials research," MRS Bulletin, vol. 31, no. 9, pp. 659–668, 2006.
[5] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd)," *Jom*, vol. 65, no. 11, pp. 1501–1509, 2013.
[6] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Ru ̈hl, and C. Wolverton, "The open quantum materials database (oqmd): assessing the accuracy of DFT formation energies," *NPJ Computational Materials*, vol. 1, no. 1, pp. 1–15, 2015.
[7] I. El Naqa and M. J. Murphy, "What is machine learning?" in *machine learning in radiation oncology*. Springer, 2015, pp. 3–11.
[8] D. Jha, L. Ward, A. Paul, W.-K. Liao, A. Choudhary, C. Wolverton, and A. Agrawal, "Elemnet: Deep learning the chemistry of materials from only elemental composition," *Scientific Reports*, vol. 8, no. 1, 2018.
[9] D. Jha, L. Ward, Z. Yang, C. Wolverton, I. Foster, W.-K. Liao, A. Choudhary, and A. Agrawal, "Irnet: A general purpose deep residual regression framework for materials discovery," *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, 2019.

[10] D. Jha, V. Gupta, L. Ward, Z. Yang, C. Wolverton, I. Foster, W. Keng Liao, A. Choudhary, and A. Agrawal, "Enabling deeper learning on big data for materials informatics applications," *Scientific Reports*, vol. 11, no. 1, pp. 1–12, 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-83193-1.
[11] A. Y. T. Wang, S. K. Kauwe, R. J. Murdock, and T. D. Sparks, "Compositionally restricted attention-based network for materials property predictions," *NPJ Computational Materials*, vol. 7, no. 1, pp. 1–10, 2021. [Online]. Available: http://dx.doi.org/10.1038/s41524-021- 00545-1.
[12] D. Tatis, H. Sierra, and E. Arzuaga, "Residual neural network architectures to improve prediction accuracy of properties of materials," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 2915–2918.
[13] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in European conference on computer vision. Springer, 2016, pp. 630–645.
[14] K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C. W. Park, A. Choudhary, A. Agrawal, S. J. Billinge, *et al.*, "Recent advances and applications of deep learning methods in materials science," *NPJ Computational Materials*, vol. 8, no. 1, pp. 1–26, 2022.
[15] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, "Combinatorial screening for new materials in unconstrained composition space with machine learning," *Physical Review B*, vol. 89, no. 9, p. 094104, 2014.
[16] N. M. Twyman, A. Walsh, and T. Buonassisi, "Environmental stability of crystals: A greedy screening," *Chemistry of Materials*, vol. 34, no. 6, pp. 2545–2552, 2022.
[17] A. Rezoug, M. Bader-El-Den, and D. Boughaci, "Guided genetic algorithm for the multidimensional knapsack problem," *Memetic Computing*, vol. 10, no. 1, pp. 29–42, 2018.
[18] A. Radhakrishnan and G. Jeyakumar, "Evolutionary algorithm for solving combinatorial optimization—a review," *Innovations in Computer Science and Engineering*, pp. 539–545, 2021.
[19] H. Mühlenbein, M. Gorges-Schleuter, and O. Kra¨mer, "Evolution algorithms in combinatorial optimization," *Parallel computing*, vol. 7, no. 1, pp. 65–85, 1988.
[20] E. Manduchi, P. R. Orzechowski, M. D. Ritchie, and J. H. Moore, "Exploration of a diversity of computational and statistical measures of association for genome-wide genetic studies," *BioData mining*, vol. 12, no. 1, pp. 1–16, 2019.
[21] C.Kim, R.Batra, L.Chen, H.Tran, and R.Ramprasad, "Polymerdesign using genetic algorithm and machine learning," *Computational Materials Science*, vol. 186, p. 110067, 2021.
[22] M. Ibrahim, F. Nurhakiki, D. Utama, and A. Rizaki, "Optimised genetic algorithm crossover and mutation stage for vehicle routing problem pick-up and delivery with time windows," in *IOP Conference Series: Materials Science and Engineering*, vol. 1071, no. 1. IOP Publishing, 2021, p. 012025.
[23] A. Ebrahimi, J. Tamnanloo, S. H. Mousavi, E. Soroodan Miandoab, E. Hosseini, H. Ghasemi, and S. Mozaffari, "Discrete-continuous genetic algorithm for designing a mixed refrigerant cryogenic process," *Industrial & Engineering Chemistry Research*, vol. 60, no. 20, pp. 7700–7713, 2021.