

Article

# Crystal-Site-Based Artificial Neural Networks for Material Classification

Juan I. Gómez-Peralta <sup>1,\*</sup> , Nidia G. García-Peña <sup>2</sup> and Xim Bokhimi <sup>1,\*</sup><sup>1</sup> Instituto de Física, Universidad Nacional Autónoma de México, A.P. 20-364, México City 01000, Mexico<sup>2</sup> Department of Applied Physics, CINVESTAV-IPN, Antigua Carretera a Progreso km 6, A.P. 37, Mérida 97310, Mexico; gapenig@gmail.com

\* Correspondence: gomezperalta.ai@gmail.com (J.I.G.-P.); bokhimi@fisica.unam.mx (X.B.)

**Abstract:** In materials science, crystal structures are the cornerstone in the structure–property paradigm. The description of crystal compounds may be ascribed to the number of different atomic chemical environments, which are related to the Wyckoff sites. Hence, a set of features related to the different atomic environments in a crystal compound can be constructed as input data for artificial neural networks (ANNs). In this article, we show the performance of a series of ANNs developed using crystal-site-based features. These ANNs were developed to classify compounds into halite, garnet, fluorite, hexagonal perovskite, ilmenite, layered perovskite, -o-tp- perovskite, perovskite, and spinel structures. Using crystal-site-based features, the ANNs were able to classify the crystal compounds with a 93.72% average precision. Furthermore, the ANNs were able to retrieve missing compounds with one of these archetypical structure types from a database. Finally, we showed that the developed ANNs were also suitable for a multitask learning paradigm, since the extracted information in the hidden layers linearly correlated with lattice parameters of the crystal structures.

**Keywords:** crystal structures; material discovery; deep learning; feature engineering; multitask learning



**Citation:** Gómez-Peralta, J.I.; García-Peña, N.G.; Bokhimi, X. Crystal-Site-Based Artificial Neural Networks for Material Classification. *Crystals* **2021**, *11*, 1039. <https://doi.org/10.3390/cryst11091039>

Academic Editor: Hartmut Schlenz

Received: 6 August 2021

Accepted: 26 August 2021

Published: 29 August 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In recent years, machine learning algorithms have irrupted as an alternative tool to model the properties and structure of materials [1–11]. These algorithms have allowed scientists to work with large particle systems at shorter times and lower computational costs with respect to the recurred quantum methods [12–15]. Agrawal and Choudhary [16] have suggested that machine learning constitutes nowadays a fourth modeling paradigm in science, which relies on the information stocked in large databases. In addition, strategic initiatives [17–20] that seek to accelerate the material discovery–commercialization process have come to the public scene. Among all machine learning algorithms, artificial neural networks (ANNs) are perhaps the most extended algorithms, mainly due to their success in the automatization of tasks that have been regarded as exclusive to humans [21–25]. Deriving from this success, deep learning [26,27] has emerged from machine learning as a discipline that gathers all the activity related to the current ANNs—the zoo.

An important factor influencing the performance of ANNs is the input data [28,29]. In fact, conceiving the components of the input data, which are called descriptors or features, falls in a major area called feature engineering. In this sense, it turns out that it is necessary to revisit how the input data was conceived in reported crystal-chemical works.

Fedorov and Shamanaev [30] have described inorganic crystal compounds in terms of topological centers [31] to estimate thermodynamical variables with feed-forward ANNs. These topological centers contained chemical features such as electronegativity, oxidation state, molar mass, atomic covalent radius, and distance between a central atom and its neighbors. Recently, Jiang et al. [32] also characterized the crystal structure using algebraic topology to assess the DFT-formation energy, obtaining presumably better estimations than

previously informed approaches based on the Coloumb matrix [33], which models electrostatic interaction between atom pairs. Saidi and coworkers [34] concatenated properties of the atoms and molecules forming  $ABX_3$  perovskites to estimate the lattice constant, octahedral tilt angle, and even band gap energies using convolutional neural networks (CNNs); with these CNNs, hidden information was extracted from the 1D-input data. Pathak and coworkers [35] developed a strategy using conditional variational autoencoders to generate candidate materials that suit formation energy, energy per atom, and volume per atom constraints. Since the characterization of the materials was based on concatenated one-hot vectors, information concerning the structure was not necessary to generate the candidate materials. Each one-hot vector represents an element of the periodic table. The elements of this one-hot vector were zeros and ones, depending on the number of atoms in the material formula. Similar non-structural approaches have been reported by Jha et al. [36] and Goodall et al. [37]

Embeddings of elemental properties have been perhaps the most concurred material characterization approach [3,38–41], which allows the combination of properties of different magnitudes of scale and nature. Isayev and coworkers [40] characterized the materials for the AFLOW repository with the so-called property-labeled materials fragments. The starting point of this characterization is Voronoi tessellation, which establishes the connection between atoms. The connected atoms allow the feature calculation using elemental properties such as valence electrons, electron affinity, heat capacity, and chemical hardness, for instance. Recently, Choubisa et al. [41] embedded atomic-like properties to each specie of the  $ABX_3$  perovskite compound within the unit cell. These atomic-like properties were related to frontier orbital energies, electron population, and polarization, which were derived from DFT calculations.

Despite all the mentioned efforts, a more accurate approach can be tackled: Crystal compounds are periodical systems. Due to periodicity, the description of a  $10^{23}$  particle system can be formulated in terms of the atoms within the unit cell. In this manner, property embeddings could be constructed for each atom within the unit cell. Additionally, all atoms of a crystal occupy positions in space that are characterized by point-symmetry groups. These point-symmetry groups correspond to Wyckoff sites [42]. The atoms in the same orbit of a Wyckoff site have the same local order and, as a consequence, the description of the crystals can even be simplified in terms of the number of Wyckoff sites, i.e., number of different chemical environments in a crystal. This natural consequence of the periodicity of crystal systems can be exploited to characterize crystal materials to develop ANNs. However, there barely are reports that use material characterization based on symmetry sites. Ryan and coauthors [43] trained ANNs to distinguish chemical elements based on the topology of the crystallographic environments. The local topology around a crystallographic unique atom is represented by crystal structure fingerprints [44]. Ye et al. [45] characterized garnet and perovskite compounds with the ionic radii and electronegativity of the cations to estimate their DFT-formation energy. Interestingly, this work is one of the first to associate the cation geometries with their crystal sites.

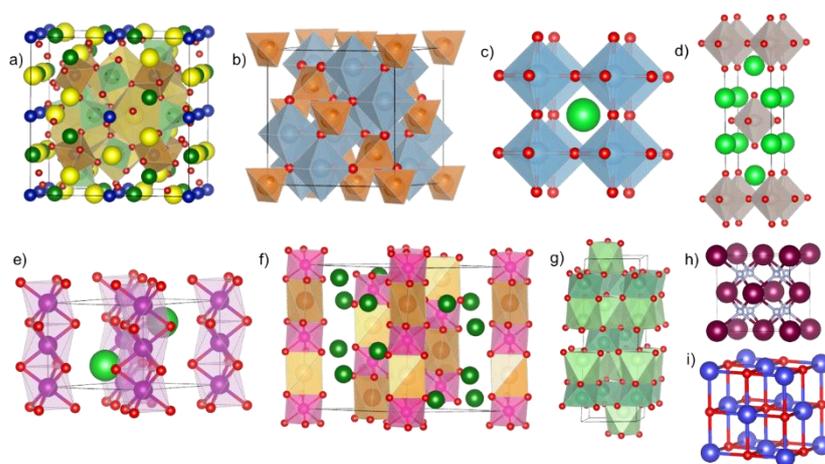
In this article, ANNs were developed to classify crystal compounds in terms of their structure types. The compounds were characterized with a reported methodology by ourselves [46,47], which uses the number of different Wyckoff sites to construct the features. The reported methodology has only been proven in a binary classification model in terms of perovskite or non-perovskite structures. In this sense, this work can be considered as an extension to multiclass classification. Contrary to the elemental-property embeddings, which combine variables of different scale and nature, all constructed features describe the compounds in terms of structural factors, as well as in terms of the atomic local order of the different chemical environments of the crystal. The performance of the developed ANNs validated the ability of the feature construction here implemented. In addition, the used methodology has two highlighted criteria: (a) quantum calculations are not required to obtain a feature, and (b) different compounds with the same structure type can be managed regardless of their crystal system. This last fact is a consequence of the crystal definition.

The ANNs were developed using the code `patolli.py`, which is available at GitHub [48]. Patolli was the name of a game in ancient Mexico, which also had divinatory properties. With the code `patolli.py`, the aim was to create ANNs, called `patollis` focused on classification and even prediction of the structure type of crystal materials.

## 2. Materials and Methods

### 2.1. Nomenclature

The compounds with the perovskite, spinel, garnet, hexagonal perovskite, layered perovskite, -o-tp- perovskite, ilmenite, halite, and fluorite structures (Figure 1) were used to create different collections to develop the ANNs. A brief review of the mentioned structure types is provided in the Supplementary Materials. In this article, we refer to those compounds characterized only by a vertex-shared octahedral framework as perovskite structures. Additionally, we refer to Ruddlesden–Popper and Dion–Jacobson structures as layered perovskites [49].



**Figure 1.** Crystal structure types used in this work to classify the compounds: (a) garnet structure, (b) spinel structure, (c) perovskite structure. An example of the layered perovskite structure is shown in (d), which corresponds to the Ruddlesden–Popper phase. Similarly, an example of hexagonal perovskite is shown in (e). The -o-tp- perovskite structure is depicted in (f), whereas the ilmenite structure is shown in (g). The fluorite and halite structures are labeled (h,i). The characteristic polyhedral framework in the fluorite and halite structures are not shown to facilitate their visualization.

In addition to the structure type of Figure 1, compounds without any of the previously mentioned structure types were used as examples of not-identified structure type. We refer to these not-identified compounds as the “others” structure type.

All the ANNs were feed-forward, full-connected type and had two hidden layers. The architecture of these ANNs is described in Table 1. The ANNs were named 4S4O-NEF, 4S4O-WEF, 6S4O-NEF, 6S4O-WEF, 6S8O-NEF, 6S8O-WEF, and 6S10O-WEF. The nomenclature refers to the number of crystal sites, *S*, characterizing the compounds of the used collection to develop the ANNs (either 4 or 6). The number of outputs in the ANNs was pointed out after the letter *O*, which could be 4, 8, or 10. The number of outputs corresponded to the different structure types. Additionally, if a set of extra features was included in the characterization of the compounds, it was indicated as WEF (with extra features) or NEF (no extra features). This set of extra features corresponded to the average atomic radius and electronegativity of the crystal sites, as well as to the density of the crystal compound.

The number of crystal sites influenced the number of features (input data) to characterize the crystal compounds. When the set of extra features was used, the number of features was 42 and 163 in the 4*S*- and 6*S*-ANN, respectively. Otherwise, the number of features was 33 and 150.

**Table 1.** Summary of the input data and the architectures of the developed ANNs. The first column enlists each of the ANNs. The second column refers to the dimension of the input vector of the ANNs, whereas the last column indicates the number of nodes in each layer.

ANN	Features in Input Data	Architecture (2 Hidden Layers + Output Layer)	Structure Type Outputs
4S4O-NEF	33	231, 132, 4	Garnet, perovskite, spinel, and others
6S4O-NEF		900, 750, 4	Garnet, perovskite, spinel, and others
6S8O-NEF	150	900, 750, 8	Garnet, hexagonal perovskite, ilmenite, layered perovskite, -o-tp- perovskite, perovskite, spinel, and others
4S4O-WEF	42	210, 126, 4	Garnet, perovskite, spinel, and others
6S4O-WEF		1141, 652, 4	Garnet, perovskite, spinel, and others
6S8O-WEF	163	815, 163, 8	Garnet, hexagonal perovskite, ilmenite, layered perovskite, -o-tp- perovskite, perovskite, spinel, and others
6S10O-WEF		815, 652, 10	Fluorite, garnet, halite, hexagonal perovskite, ilmenite, layered perovskite, -o-tp- perovskite, perovskite, spinel, and others

Except for the 4S4O-ANNs, all ANNs were developed with a collection of compounds having up to six Wyckoff sites. For the 4S4O-ANNs, a collection of compounds with up to four Wyckoff sites was used. The collections used to develop the 4S4O- and 6S4O-ANNs had compounds with the garnet, perovskite, spinel, and “others” structure types. Similarly, the collection used to develop the 4S8O-ANNs had compounds with either the garnet, hexagonal perovskite, ilmenite, layered perovskite, -o-tp- perovskite, perovskite, spinel, and “others” structure types. The collection used to develop the ANN 4S10O-WEF also included the fluorite and halite structure types.

## 2.2. Features

The methodology published by Gómez-Peralta and Bokhimi [46,47] uses the number of symmetry sites to characterize a crystal compound. The features are related to structural factors, such as geometric and packing ones, as well as local functions related to the chemical environment of the atoms in the crystal sites. The local functions [46,47] model the interaction of all neighbor atoms in a  $j$ -Wyckoff site over a central atom in the  $i$ -Wyckoff site, within cutoff radius  $R_c = 25 \text{ \AA}$  (Equation (1)).

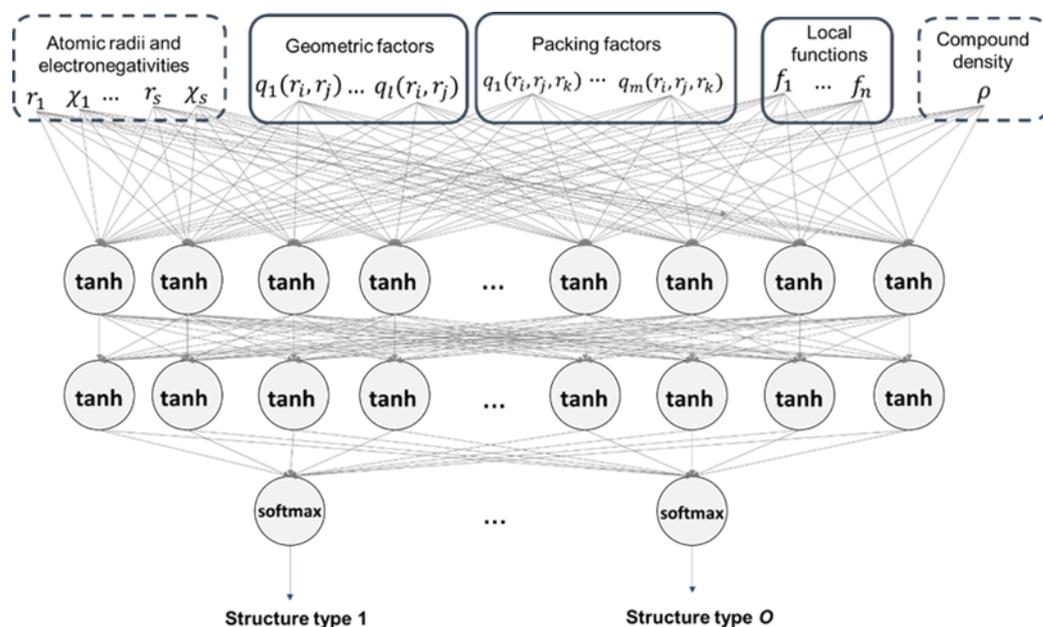
$$f_{ij} = (\chi_i - \chi_j) \sum_{n=1}^{d_{ij[n]} \leq R_c} \left[ \frac{1}{2} \left( \cos \frac{\pi d_{ij[n]}}{R_c} + 1 \right) \right] \exp \left[ - \left( \frac{d_{ij[n]}}{r_i^{norm} + r_j^{norm}} \right)^2 \right] \quad (1)$$

The function used to model this interaction has a Gaussian profile: the neighbor atoms closer to the central atom have a larger contribution to the local function. Additionally, the magnitude of the interaction is modulated depending on the nature of the involved atom pair.

The detailed list of the used features may be consulted in the Supplementary Materials. Since the referred methodology uses the number of combinations to compute the features, there were 6 geometric factors, 15 packing factors, and 12 local chemical environment functions to characterize the collection of compounds with up to four sites. Similarly, there were 15 geometric factors, 105 packing factors, and 30 chemical environment functions to characterize the collection of compounds with up to six sites. This makes 33 and 150 features for the collections with four and six sites. However, we also added the average atomic radius and electronegativity of each site, which provided 8 and 12 features more,

as well as the density of the compound. With these extra features, there were 42 and 163 features for the collections of compounds with up to four and six sites, respectively.

The features were arranged in the method described in Figure 2. This arrangement did not influence the training of the ANNs, but it is important to look out once the ANNs are developed.



**Figure 2.** General architecture of the ANNs developed in this work. All ANNs had three layers. The activation function in the hidden layers was a hyperbolic tangent, whereas in the output layer the SoftMax function was used. The input data consisted of a series of features that could be grouped in the manner described above in the figure. The features enclosed in dashed lines were not used in the NEF ANNs, but they were in the WEF type.

### 2.3. Software and Computational Infrastructure

The ANNs were developed with the Python code `patolli.py`, which is available via GitHub. `Patolli` uses a collection of crystal compounds taken from the Crystallography Open Database [50–53]. This collection of compounds contains information about the CIF, formula, number of Wyckoff sites, number of different elements in the formula, space group, number of atoms in the unit cell, and the occupation of each Wyckoff site. The atomic occupation of each Wyckoff site was assessed with the Python library `pymatgen` [54]. In fact, this description of occupied Wyckoff sites is crucial for the definition of the features in the input data.

The code `patolli.py` is executed via the terminal and calls two text files that need to be specified by the user: one defining all structure types in terms of their space groups and occupied Wyckoff sites, and another that defines the characteristics of the ANNs to be trained, as well as the hyperparameters of these. The crystal definition of the structure types used in the research is given in the Supplementary Materials. The crystallographic definition of the occupied sites in the perovskite and spinel structure was consulted in references [55–58]. Similarly, compounds with the hexagonal perovskite, layered perovskite, and -o-tp- perovskite structure were consulted according to the information provided by Tilley [49]. The other structure types were crystallographically defined after visual inspection of some of their compounds with the VESTA software package [59].

When `patolli.py` is executed, the code will ask the user whether further constraints will be taken into account to create the collection of compounds to develop the ANNs. These constraints are related to the number of atoms within the unit cell, the number of different elements in the formula, and the highest number of Wyckoff sites to consider of a

structure type. The code `patolli.py` also asks the user whether the extra features related to the atomic radii, electronegativities, and density of the compound are included in the input data. Finally, the code `patolli.py` asks whether the ANNs will be trained to differentiate not identified compounds. This subset of not-identified compounds corresponds to the label “others” and is created by randomly choosing compounds that do not match the crystal definitions provided.

During `patolli.py` execution, the collection of compounds was created according to the crystal structure definitions provided in the structure text files. The code automatically computed the features and splits the whole collection into the TRAINING–VALidation (traval) set and the test set. After splitting the collection, csv files containing the compounds as well as NumPy files, which contains the features, were created. The ANNs were trained with the compounds of the traval set, which were previously standardized. After the training of all ANNs was completed, these were tested with the compounds reserved in the test set. If the user declared to perform a second test, `patolli.py` tested the ANNs with all the remaining compounds of the integrated collection, which were not used in the traval or test sets.

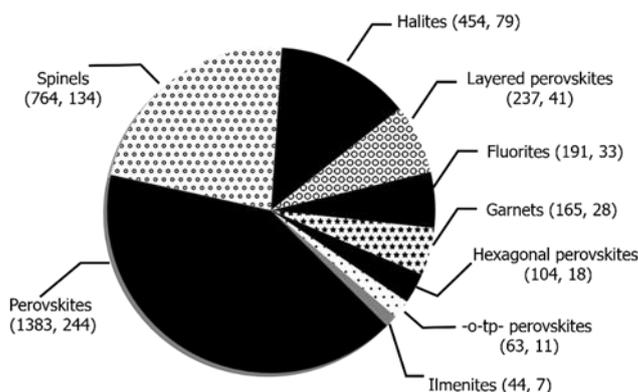
The library Keras 2.2.4 is required to satisfactorily execute `patolli.py`. The Keras backends that have been proven to work are Theano 1.0.3 and TensorFlow 1.14.0. In addition, the libraries `pydot` and `graphviz` were required to display some items related to the library Keras. The 6S-ANNs were developed in the computer Mixkhua, which used the NVIDIA GPU Tesla K40c. The training of the 6S-ANNs in this GPU lasted 3–6 min, depending on the extension of the ANN. Mixkhua is located in the Artificial Intelligence Laboratory of the Institute of Physics, UNAM. On the contrary, the 4S-ANNs were trained on a personal computer with a processor AMD 9-9420 Radeon R5 and 8 GB RAM. On average, the training of the 4S-ANNs lasted 5 min in the personal computer. Training the 6S-ANNs on this personal computer required longer times, even up to 1 h. In all cases, the NumPy random seed used to develop the ANNs was 10. This seed is preconfigured in `patolli.py`.

The linear regressions between the extracted features of the 6S10O-WEF ANNs and lattice parameters were established via the Python library `scikit-learn`. For this purpose, a Jupyter Notebook was prepared. This Jupyter Notebook can be consulted electronically (see Data Availability section at the end of the article).

### 3. Results and Discussion

#### 3.1. Classification of Crystal Compounds

Figure 3 shows the distribution of the used compounds to develop the 6S10O-WEF in terms of the structure types of Figure 1. Based on Figure 3, the collection of compounds used to develop the 6S8O-ANNs did not include those with the fluorite and halite structure. Similarly, the collection of compounds to develop the 6S4O- and 4S4O-ANNs only contained the perovskite, garnet, and spinel compounds. It is important to remember that the collection to develop the 4S4O-ANNs did not have compounds with more than four Wyckoff sites. Additionally, it did not include the spinel compounds with the hausmannite structure [57] in the collection to develop the 4S4O-ANNs, which are described with three Wyckoff sites. The reason behind this will be explained later in the text. In all the collections, the number of compounds labeled as “others” structure type was equal to the sum of the compounds of all structure types.



**Figure 3.** Distribution of the compounds of Figure 1 in the collection to develop 6S10O-WEF. The number of compounds in the traval and test sets is shown in parenthesis. The collection to develop 6S10O-WEF was the most complete of all. The collection used to develop 6S8O-ANNs did not include fluorite and halite compounds. Similarly, the collection to develop 6S4O- and 4S4O-ANNs only contained the garnet, perovskite, and spinel structures. Since the collection used for 4S4O only contained compounds with up to four Wyckoff sites, there were 193 and 57 compounds less with the perovskite and spinel structure. In all the collections, the number of compounds belonging to the “others” structure was equal to the sum of the compounds of all structure types.

Table 2 shows the precision in the classification of the test set compounds into each structure type by the developed ANNs. Since the compounds of the traval set were used during the learning process (optimization) of the ANNs, the precisions obtained with that compounds set are provided to the reader in the Supplementary Materials. The precisions obtained with the traval set were 2–3% higher than those of the test set, which is normally observed in the development of this kind of models. The metric known as precision is defined as the next quotient:

$$precision = \frac{TP}{TP + FP}$$

where *TP* stands for true positive cases, whereas *FP* is for false positive ones (misclassified sample after comparing with the actual label). This metric is important for predictive purposes since it tracks the rate of correct predictions. The mean precision of all developed ANNs was 93.72% after the test set compounds and 95.92% after the compounds of the traval test. These values are similar to those reported by Gómez-Peralta and Bokhimi [46] for a binary classification model, where the ANN outputs the probability of adopting the perovskite structure.

The highest precisions in the classification were obtained for the compounds with the garnet (97.60%, on average), perovskite (94.58%), spinel (95.25%), layered perovskite (96.37%), and -o-tp- perovskite structure (91.67%). In addition to these, the compounds labeled as “others” structure type were classified by the ANNs with a mean precision of 94.65%. In contrast, the precision obtained for the compounds fluorite, halite, ilmenite, and hexagonal perovskite were mild. The mean precisions of the mentioned structure types were 87.10%, 87.01%, 80.95%, and 77.78%, respectively. Since the number of compounds with the hexagonal perovskite and ilmenite structures was low, we can suggest that the precisions may improve after enrichment of the dataset with compounds of these structure types. Furthermore, the diversity within the hexagonal perovskite compounds could also be a factor of this diminished performance since different polytypes of the hexagonal perovskite structure (such as 2H-, 6H-, 8H-, and 10H- structures) were included with the same label compounds. The low number of available compounds in the data set was the reason for gathering the mentioned structures under the label hexagonal perovskite.

**Table 2.** Precisions in the classifications of the structure types (first column) achieved by each ANNs (from second column to the last) after the compounds of the test set.

	4S4O-NEF	4S4O-WEF	6S4O-NEF	6S4O-WEF	6S8O-NEF	6S8O-WEF	6S10O-WEF
Garnet	100.00	96.55	100.00	100.00	93.33	93.33	100.00
Perovskite	94.93	95.43	93.13	95.98	92.62	94.14	95.82
Spinel	98.44	96.92	94.24	96.32	95.52	94.93	90.41
Hexagonal perovskite	NA	NA	NA	NA	75.00	75.00	83.33
Ilmenite	NA	NA	NA	NA	71.43	85.71	85.71
Layered perovskite	NA	NA	NA	NA	97.22	94.59	97.30
-o-tp- perovskite	NA	NA	NA	NA	90.91	90.91	91.67
Fluorite	NA	NA	NA	NA	NA	NA	87.10
Halite	NA	NA	NA	NA	NA	NA	87.01
Others	97.54	98.61	92.74	92.47	93.63	94.25	93.32
Average	97.73	96.88	95.03	96.19	88.71	90.36	91.17

It is important to mention that all the compounds, except those adopting the fluorite or halite structure, were described with at least three Wyckoff sites. Fluorite and halite structures require two Wyckoff sites to be described. Since the methodology developed by Gómez-Peralta and Bokhimi depends on the number of occupied sites, there were not enough features to characterize compounds with halite and fluorite structures. In fact, the referred methodology expected three non-zero features in the input data. To alleviate this issue, the average atomic radius and electronegativities by site, as well as the density of the compound, were included in the feature set. The inclusion of compound density is justified since it could potentially model the presence of vacancies or solid solutions. After including the mentioned features, there were eight non-zero features for the fluorite and halite compounds. Nevertheless, the obtained precisions for fluorite and halite suggest that more features are needed to improve the performance for these structure types.

Regarding each ANN developed, 4S4O-ANNs had the highest precisions, with 97.73% for 4S4O-NEF ANN and 96.88% for 4S4O-WEF ANN (last row of Table 2). The result can be explained in terms of a sufficiently large number of compounds with garnet, spinel, and perovskite structure, as well as a small number of outputs in those ANNs. Similarly, 6S4O-NEF and 6S4O-WEF ANNs had the same structure type outputs as 4S-ANNs and classified the compounds with a mean precision of 95.03% and 96.19%, respectively. 6S8O-NEF and 6S8O-WEF included in their outputs the hexagonal perovskite, ilmenite, layered perovskite, and -o-tp- perovskite and had mean precisions of 88.71% and 90.36%. 6S8O-NEF had the lowest average score of all developed ANNs. It is interesting to compare the performance of the 6S8O-NEF (88.71%) and 6S10O-WEF (91.17%). The differences between these ANNs were the use of i) more features (average atomic radii and electronegativity per site, as well as the compound density) and ii) the inclusion of the halite and fluorite compounds. In addition, the use of the halite and fluorite structure type seem to improve the metrics for the perovskite, hexagonal perovskite, and ilmenite compounds but did not do so with the spinel compounds. The lower precision for the spinel compounds in the 6S10O-WEF with respect to 6S8O-NEF may be related to the number of different elements in their formula, which sometimes can be similar to fluorite and halite.

We performed a second and larger test on the developed ANNs. For this second test, the remaining compounds of the database were used, i.e., these compounds were not used either in the training or test sets. In principle, it can be considered that the remaining compounds belonged to the structure type labeled as “others”, since these compounds did not match the provided definition of the crystal structure types. For this second test, 4S4O, 6S4O, 6S8O, and 6S10O ANNs were tested with 12,264, 19,229, 18,179, and 16,667 compounds, respectively. The difference in the number of compounds used in the second test depended on the number of compounds in the training and test sets. The results in the classification of the remaining “others” structure compounds available with the ANNs are in Table 3. More details regarding the number of compounds with respect to their Wyckoff

sites are provided in Supplementary Materials. The results in Table 3 corresponded to the metric known as recall, which is defined as follows:

$$recall = \frac{TP}{TP + TN}$$

*TN* stands for true negative cases, whereas *TP* stands for true positive cases.

**Table 3.** Recall with the remaining compounds having structure types different to those depicted in Figure 1 (second test).

ANN	Recall (%) by Number of Wyckoff Sites					
	1	2	3	4	5	6
4S4O-WEF	99.83	94.22	96.97	96.85	NA	NA
4S4O-NEF	100.00	98.42	96.41	96.55	NA	NA
6S4O-WEF	99.92	90.78	93.72	96.22	83.89	93.02
6S4O-NEF	100.00	97.33	92.72	96.15	82.71	92.53
6S8O-WEF	99.83	99.07	93.02	96.57	85.35	91.28
6S8O-NEF	100.00	97.61	94.12	96.01	85.75	91.06
6S10O-WEF	97.22	85.14	94.12	94.53	84.56	93.43

Table 3 shows the recall of “others” compounds in terms of their number of Wyckoff sites. Recall of the “others” compounds was consistent with the precision shown in Table 2. The NEF ANNs slightly outperformed the WEF ANNs in classifying the compounds with less than three Wyckoff sites. The lower value in the classification of the two sited compounds with the 6S10O-WEF can be attributed to confusion with the halite and fluorite structures. Therefore, we may suggest that the compounds with the halite and fluorite structure might have worked to improve the performance in the classification of the compounds with some of the structure types of Figure 1. In contrast, the recall of the five-site compounds was the lowest, which was a consequence of the small number of existing compounds with that number of Wyckoff sites in the collection.

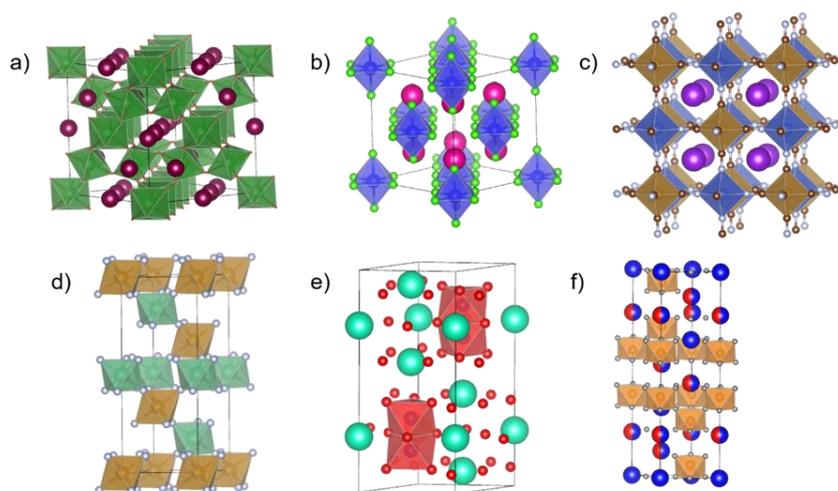
As previously established, the developed ANNs should classify all the remaining compounds as “others” structure type. Except for the ANN 6S10O-WEF, the developed ANNs were able to discriminate almost perfectly the compounds with less than three Wyckoff sites as “others”. This result is noteworthy since none of the compounds of Figure 1 are described with less than three Wyckoff sites, except those with the fluorite and halite structures, as is well known by an experienced crystal chemist. This capability to discriminate the compounds with less than three Wyckoff sites can be ascribed to the quality of the used features.

### 3.2. Retrieval of Compounds with an Archetypal Structure Type

In general, it was observed that the misclassifications in the test set occurred between the compounds belonging to the structure types of Figure 1 and the “others” structure type. In fact, the purpose of the “others” structure output was to prevent ANNs from systematically mixing up two structure types of Figure 1. The compounds confused by the ANN between the phases depicted in Figure 1 were BaMnO<sub>3</sub>, Ba<sub>2</sub>Co<sub>2</sub>O<sub>5.56</sub>, and CsNiF<sub>3</sub>, which were hexagonal perovskites misclassified as perovskites; MgSiO<sub>3</sub>, which was an ilmenite misclassified as perovskite; Cs<sub>2</sub>NaYCl<sub>6</sub> and Cs<sub>2</sub>AgAuCl<sub>6</sub>, which were perovskites misclassified as -o-tp- perovskites; FeBO<sub>3</sub>, GaBO<sub>3</sub>, and Na<sub>0.22</sub>FeF<sub>3</sub>, which were perovskites misclassified as spinel; and Fe<sub>1.2</sub>Mn<sub>1.6</sub>O<sub>4</sub> and Ca(InS<sub>2</sub>)<sub>2</sub>, which were spinel misclassified as ilmenite and perovskite, respectively. The structures of the mentioned compounds were visually verified.

So far, we have not identified a pattern for the misclassified compounds of Figure 1 with the label “others”. The comprehension of these misclassified compounds can help us to design new features that can avoid these errors. In contrast, we found that some compounds of the test set initially labeled as “others” were systematically classified as one

of the structure types of Figure 1. This was the case for the spinel compounds  $\text{Ag}_2\text{SO}_4$  (space group  $Fddd$ ) and  $\text{Li}_2\text{SO}_4$  ( $Cmcm$ ); the ilmenite compounds  $\text{Fe}_{0.33}\text{Sc}_{0.33}\text{O}$ ,  $\text{Nb}_2\text{Mn}_4\text{O}_9$  and  $\text{BiTi}_{0.375}\text{Fe}_{0.25}\text{Mg}_{0.375}\text{O}_3$  ( $R3c$ ); the hexagonal perovskite compound  $\text{Ba}_3\text{CrS}_5$  ( $P6_3cm$ ); and several perovskite compounds with the space groups  $Fd\bar{3}m$  ( $\text{Pb}_2\text{ReMnO}_6$ ),  $Fm\bar{3}m$  ( $\text{Rb}_2\text{ZrCl}_6$ ),  $Pa\bar{3}$  ( $\text{Ba}_8\text{U}_{2.668}\text{Fe}_{4.8}\text{In}_{0.532}\text{O}_{24}$ ),  $R\bar{3}m$  ( $\text{Ba}_2\text{Ca}_{0.9}\text{Nb}_{1.05}\text{O}_{5.65}$ ),  $R3m$  ( $\text{BaFe}_{0.5}\text{Ta}_{0.5}\text{O}_3$ ),  $P4mm$  ( $\text{Pb}_{0.998}\text{Ti}_{0.964}\text{O}_{2.9}$ ), and  $Amm2$  ( $\text{K}_{0.73}\text{Na}_{0.27}\text{NbO}_3$ ). After visual inspection, we verified that the ANNs correctly classified the enlisted compounds. For example, the  $\text{Pb}_2\text{ReMnO}_6$  had the distorted vertex-shared octahedral framework;  $\text{Rb}_2\text{ZrCl}_6$  can be considered a double cubic perovskite structure where half of the octahedral sites were occupied, to mention some (Figure 4).



**Figure 4.** Some of the compounds initially labeled as “others” were retrieved with one of the structure types of Figure 1: (a)  $\text{Pb}_2\text{ReMnO}_6$  ( $Fd\bar{3}m$ ) had the characteristic framework of the perovskite structure; (b)  $\text{Rb}_2\text{ZrCl}_6$  ( $Fm\bar{3}m$ ) had the perovskite structure where half of the octahedra were occupied by Zr; (c)  $\text{K}_2\text{FeCu}(\text{CN})_6$  ( $Fm\bar{3}m$ ) had the characteristic octahedral framework spaced by cyanide units; (d)  $\text{NbFeF}_6$  ( $R\bar{3}$ ) had the perovskite structure without occupied cuboctahedral sites; (e)  $\text{Cs}_3\text{W}_2\text{Cl}_9$  ( $P6_3/mmc$ ) had the hexagonal perovskite structure, without occupied octahedra connecting face-shared columns; and (f)  $\text{RbBa}_2\text{Fe}_2\text{F}_9$  ( $R\bar{3}m$ ), had the perovskite structure with vacant octahedra within the face-shared columns.

In addition, a similar trend was observed with some compounds of the second test. It is important to mention that the proportion of these compounds is a small fraction of the not recovered compounds (errors) in Table 3. The list of these compounds is provided to the reader in the Supplementary Materials. Interestingly, the ANNs were able to recognize the hexagonal perovskite structure in compounds with vacancies such as  $\text{Cs}_3\text{W}_2\text{Cl}_9$  ( $P6_3/mmc$ ) and  $\text{RbBa}_2\text{Fe}_2\text{F}_9$  ( $R\bar{3}m$ ), as well as cyanide compounds with the perovskite structure such as  $\text{K}_2\text{FeCu}(\text{CN})_6$  ( $Fm\bar{3}m$ ), and  $\text{NbFeF}_6$  ( $R\bar{3}$ ) had the distorted vertex-shared octahedral framework with unoccupied cuboctahedral sites, for instance. It is also noteworthy to mention that all hausmannite [57] compounds ( $I4_1/amd$ ), which were not used in the collections to develop the 4S4O-ANNs, were retrieved as spinel structures by the trained ANNs.

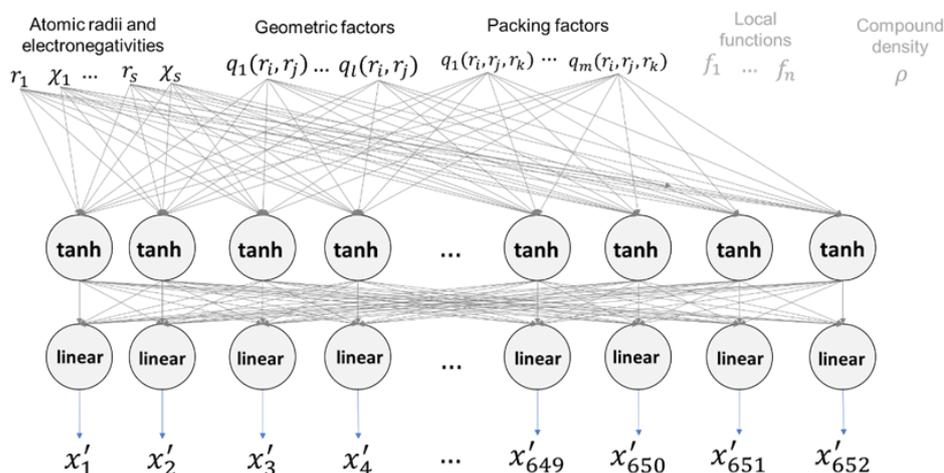
The systematic confusions of compounds initially labeled as the “others” structure type have their origin in the provided definition of the structure types: neither the space groups nor the Wyckoff occupation of these mislabeled compounds were considered. Nevertheless, the developed ANNs were able to recognize the characteristic polyhedral pattern of the structure types in Figure 1.

### 3.3. Lattice Parameter Assessment with the Extracted Features by the ANN

So far, we have focused the obtained results by the ANNs toward applications such as automatized classification and retrieving of the compounds from a crystal database. The

success of these ANNs is ascribed to the quality of the features used in the characterization of the crystal compounds. Beyond these applications, it could be of interest to establish correlations with other crystal variables using the processed information within the ANNs' hidden layers. In fact, the information processed in the hidden layer constitute new finer features that may be more accurate to consider for certain tasks.

We used the extracted features in the second layer of the ANN 6S10O-WEF to establish a correlation with the lattice parameters of the simple cubic perovskite ( $Pm\bar{3}m$ ), double cubic perovskite ( $Fm\bar{3}m$ ), orthorhombic perovskite ( $Pnma$ ), trigonal perovskite ( $R\bar{3}c$ ), garnets ( $Ia\bar{3}d$ ), spinels ( $Fd\bar{3}m$ ), and tetragonal Ruddlesden–Popper structures ( $I4/mmm$ ). According to the architecture of ANN 6S10O-WEF (Table 1), there were 652 extracted features in the second layer for each compound. The extracted features were computed after feeding the ANN 6S10O-WEF with the input data except the local functions and the density of the compound, which were set to zero (Figure 5). It was necessary to hide the local functions and the density of the compound since their calculation requires knowledge of the lattice parameter a priori.



**Figure 5.** Feature extraction from the second hidden layer of the ANN 6S10O-WEF. Since the local functions and the compound density were not fed to the ANN, the connections of these features with the first layer are hidden. It is important to mention that the extracted features ( $x'_1, x'_2, \dots, x'_{652}$ ) were the values computed by the second hidden layer prior to their activation. The extracted features were afterward used to fit a linear regression model with the lattice parameter of certain compounds.

Table 4 resumes the results of these linear regressions. In all, 344, 205, 165, 714, 388, 134, and 196 compounds of the training set with the simple cubic perovskite, double cubic perovskite, garnet, spinel, orthorhombic perovskite, trigonal perovskite, and tetragonal Ruddlesden–Popper structures, respectively, were used to fit the data. Additionally, 54, 33, 28, 127, 87, 22, and 31 compounds of the test set of the mentioned crystal structures were used to test the linear regression fit. We found compounds of the test set that did not follow the linear regression fit. For these outlier compounds, we found that the lattice parameters were exaggerated, and therefore they were not considered in the calculation of the mean square error. The number of non-outlying compounds appears also in Table 4. In all cases, a correlation coefficient above 0.90 was established between the extracted features and the lattice parameters of the studied structures types.

**Table 4.** Results of the linear regression of the lattice parameters with the extracted features in the second layer of the 6S10O-WEF ANN.

Structure Type	R <sup>2</sup>	MSE Fitting	MSE Test	Mean Lattice Parameters (Å)
Simple cubic perovskite ( <i>Pm</i> $\bar{3}$ <i>m</i> )	0.9996	0.0001 (344)	0.1057 (39 of 54)	4.1698 ± 0.4315
Double cubic perovskite ( <i>Fm</i> $\bar{3}$ <i>m</i> )	0.9992	0.0002 (205)	0.0984 (23 of 33)	8.4807 ± 0.5310
Garnet ( <i>Ia</i> $\bar{3}$ <i>d</i> )	0.9998	3 × 10 <sup>−5</sup> (165)	0.0931 (23 of 28)	12.0808 ± 0.3654
Spinel ( <i>Fd</i> $\bar{3}$ <i>m</i> )	0.9682	0.0256 (714)	0.0676 (107 of 127)	8.7000 ± 0.8976
Orthorhombic perovskite ( <i>Pnma</i> )	0.9836	0.0035 (388)	0.0683 (70 of 87)	5.5148 ± 0.3834 5.6433 ± 0.4045 7.8291 ± 0.5659
Trigonal perovskite ( <i>R</i> $\bar{3}$ <i>c</i> )	0.9994	0.0002 (134)	0.0297 (15 of 22)	5.3548 ± 0.3351 13.9118 ± 1.0555
Tetragonal Ruddlesden–Popper ( <i>I4/mmm</i> )	0.9422	0.0085 (196)	0.0673 (18 of 31)	3.9434 ± 0.3340 15.9746 ± 3.9851

It is expected that having a larger amount of compounds of each crystal structure will lower the number of outlier cases. The outlier compounds can be explained as cases where it was not possible to establish a good interpolation. A larger collection of compounds to fit a linear regression could also improve the metrics of the linear regression performance. In addition, it is important to mention that these correlations between the extracted features and lattice parameters were not possible to establish with the nodes of the first (which provided 815 extracted features) or third layer (10 extracted features).

Other than the well-known relationship between the ionic radii in the aristotype perovskite structure with the lattice parameter, an analog of this has been barely sketched for spinels, garnets, and Ruddlesden–Popper structures. Concerning the perovskite compounds, it is worthy to mention that Javed et al. [60] and Majid et al. [61,62] used similar features to assess the lattice parameters of cubic, monoclinic, and orthorhombic perovskite compounds with supported vector machines. Recently, Zhang et al. [63] established a relationship via Gaussian process regression between the ionic radii and the lattice parameter of monoclinic double perovskites. Interestingly, Song et al. [64] deduced a tolerance factor for garnet structures after reversion of the distorted dodecahedral cations to a regular cube.

### 3.4. Features' Influence on the Performance of the ANNs

Finally, we used the compounds of the test set to analyze the influence of the features on the performance in the classification of crystal structures. Table 5 contains the results of this analysis. This analysis was performed after hiding the features of a given block of Figure 2, i.e., the hidden features were set to zero. With this approach, the ANNs do not receive complete information, thus some connections between the nodes will be not triggered.

**Table 5.** Precision of each developed ANN after hiding features according to Figure 2. The precision obtained in this table were computed after the compounds in the first test set.

ANN	Precision (%) in the Test Set with					
	All Features	No Atomic Radii and Electronegativities	No Geometric Factors	No Packing Factors	No Local Functions	No Density
4S4O-NEF	97.73	NA	94.93	88.59	22.35	NA
6S4O-NEF	95.03	NA	95.00	76.60	50.84	NA
6S8O-NEF	88.71	NA	88.06	49.65	16.23	NA
4S4O-WEF	96.88	91.41	97.09	94.57	28.79	97.46
4S6O-WEF	96.19	86.28	96.16	81.40	41.27	95.33
4S8O-WEF	90.36	74.15	89.52	60.52	42.37	88.96
4S10O-WEF	91.17	80.30	89.67	69.09	37.04	89.35

Results in Table 5 points out that the most crucial set of features in the performance of the ANNs were the local functions. After remotion of the local functions, the precision dropped at least to 50.84% in the 6S4O-NEF, and to 16.23% in the 6S8O-NEF. The local functions sum up the interactions of all neighbor atoms over a central atom in a crystal site. The magnitude of the interaction depended on the nature of the involved atom pair. The second most important block of features was the packing factors. The packing factors measured the efficiency of the space-filling by the atoms of the crystal site. One of the most known packing factors is the Goldschmidt tolerance factor, which is used in the perovskite compounds. The most notorious reductions in the precision were obtained for 6S8O-NEF (49.65%), 6S8O-WEF (60.52%), and 6S10O-WEF (69.09%), which are the ANNs with more structure type outputs. In contrast, hiding the geometric factors did not affect the performance of the ANNs. The geometric factors were quotients of atomic radii, which are related to the geometry defined by the first neighbor atoms.

The average atomic radii and electronegativities by site and the density of the compound were features used in the WEF ANNs. In these types of ANNs, the omission of the average atomic radii and electronegativities block affected the performance in the classification. The precisions dropped to 91.41%, 86.28%, 74.15%, and 80.30% with the ANNs 4S4O-WEF, 4S6O-WEF, 4S8O-WEF, and 4S10-WEF, respectively. In contrast, the omission of the density of the compound did not affect the performance of the WEF ANNs.

#### 4. Conclusions

We have shown that crystal-site-based features enabled the ANNs to classify the crystal compounds with an average precision of 93.72%. The ANNs classified compounds with garnet, perovskite, spinel, layered perovskite, -o-tp- perovskite, fluorite, halite, ilmenite, and hexagonal perovskite structures with mean precisions of 97.60%, 94.58%, 95.25%, 96.37%, 91.67%, 87.10%, 87.01%, 80.95%, and 77.78%, respectively. The low scores obtained by the ANNs were ascribed to the availability in the database of compounds with a structure type. In addition, the compounds not belonging to any of the mentioned structures were classified by the ANNs with a mean precision of 94.65%. Hence, the ANNs developed with the used feature construction may find application in automatized systems for classifying and retrieving compounds from crystal databases.

In addition to the mentioned application, we were able to establish linear correlations between the lattice parameters with the extracted features in the ANN's hidden layers. More specifically, correlation coefficients above  $R^2 = 0.9422$  were established with the lattice parameters of garnet, spinel, Ruddlesden–Popper, and perovskite compounds. We suggest that the information derived from the ANN's hidden layers may serve to establish other correlations with optical or electronic properties, for instance. Thus, the developed ANNs are suitable for multitask learning applications.

It is important to mention that the ANNs were developed using the crystallographic information of already-synthesized compounds. Each developed ANN constitutes a function to map the chemical composition and atomic spatial arrangement into a structure type. Therefore, we expect that the crystal-site-based ANNs may also serve as a more accurate tool to probe the space of possible new crystal compounds.

**Supplementary Materials:** The review of crystal structures, the list of used features, the crystallographic definitions of the structure types, the results in the travel and second test, and the retrieved compounds in the second test are available online at <https://www.mdpi.com/article/10.3390/cryst11091039/s1>.

**Author Contributions:** Conceptualization, J.I.G.-P. and X.B.; methodology, J.I.G.-P.; software, J.I.G.-P.; validation, J.I.G.-P. and N.G.G.-P.; formal analysis, J.I.G.-P. and X.B.; investigation, J.I.G.-P.; resources, X.B.; data curation, N.G.G.-P.; writing—original draft preparation, J.I.G.-P.; writing—review and editing, N.G.G.-P. and X.B.; visualization, N.G.G.-P.; supervision, J.I.G.-P.; project administration, J.I.G.-P. and X.B.; funding acquisition, X.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are openly available in [https://github.com/gomezperalta/support\\_data\\_cs-anns](https://github.com/gomezperalta/support_data_cs-anns).

**Acknowledgments:** J.I.G.-P. thanks the Artificial Intelligence Lab of the Institute of Physics, UNAM, for the given support.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Tang, B.; Lu, Y.; Zhou, J.; Chouhan, T.; Wang, H.; Golani, P.; Xu, M.; Xu, Q.; Guan, C.; Liu, Z. Machine learning-guided synthesis of advanced inorganic materials. *Mater. Today* **2020**, *41*, 72–80. [[CrossRef](#)]
2. Raccuglia, P.; Elbert, K.C.; Adler, P.D.F.; Falk, C.; Wenny, M.B.; Mollo, A.; Zeller, M.; Friedler, S.A.; Schrier, J.; Norquist, A.J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533*, 73–77. [[CrossRef](#)] [[PubMed](#)]
3. Sorkun, M.C.; Astruc, S.; Koelman, J.M.V.A.; Er, S. An artificial intelligence-aided virtual screening recipe for two-dimensional materials discovery. *NPJ Comput. Mater.* **2020**, *6*, 106. [[CrossRef](#)]
4. Kaufmann, K.; Vecchio, K.S. Searching for high entropy alloys: A machine learning approach. *Acta Mater.* **2020**, *198*, 178–222. [[CrossRef](#)]
5. Dai, D.; Liu, Q.; Hu, R.; Wei, X.; Ding, G.; Xu, B.; Xu, T.; Zhang, J.; Xu, Y.; Zhang, H. Method construction of structure-property relationship from data by machine learning assisted mining for materials design applications. *Mater. Des.* **2020**, *196*, 109194. [[CrossRef](#)]
6. Lee, J.W.; Park, W.B.; Lee, J.H.; Singh, S.P.; Sohn, K.S. A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic XRD powder patterns. *Nat. Commun.* **2020**, *11*, 86. [[CrossRef](#)] [[PubMed](#)]
7. Park, W.B.; Chung, J.; Jung, J.; Sohn, K.; Singh, S.P.; Pyo, M.; Shim, N.; Sohn, K.S. Classification of crystal structure using a convolutional neural network. *IUCr* **2017**, *4*, 486–494. [[CrossRef](#)] [[PubMed](#)]
8. Schütt, K.T.; Sauceda, H.E.; Kindermans, P.J.; Tkatchenko, A.; Müller, K.R. SchNet—A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722. [[CrossRef](#)] [[PubMed](#)]
9. Pilania, G.; Gubernatis, J.E.; Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **2017**, *129*, 156–163. [[CrossRef](#)]
10. Pilania, G.; Balachandran, P.V.; Kim, C.; Lookman, T. Finding new perovskite halides via machine learning. *Front. Mater.* **2016**, *3*, 19. [[CrossRef](#)]
11. Gómez-Peralta, J.I.; Bokhimi, X. Ternary halide perovskites for possible optoelectronic applications revealed by Artificial Intelligence and DFT calculations. *Mater. Chem. Phys.* **2021**, *267*, 124710. [[CrossRef](#)]
12. Hong, Y.; Hou, B.; Jian, H.; Zhang, J. Machine learning and artificial neural network accelerated computational discoveries in materials science. *WIREs Comput. Mol. Sci.* **2020**, *10*, e1450. [[CrossRef](#)]
13. Schütt, O.; VandeVondele, J. Machine learning adaptive basis sets for efficient large scale density functional theory simulation. *J. Chem. Theory Comput.* **2018**, *14*, 4168–4175. [[CrossRef](#)] [[PubMed](#)]
14. Smith, J.S.; Isayev, O.; Roitberg, A.E. ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192–3203. [[CrossRef](#)]

15. Goh, G.B.; Hodas, N.O.; Vishnu, A. Deep learning for computational chemistry. *J. Comput. Chem.* **2017**, *38*, 1291–1307. [CrossRef]
16. Agrawal, A.; Choudhary, A. Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. *APL Mater.* **2016**, *4*, 053208. [CrossRef]
17. Materials Genome Initiative, Strategic Plan, National Science Technology Council. Available online: <https://obamawhitehouse.archives.gov/mgi> (accessed on 4 August 2021).
18. Report of the Clean Energy Materials Innovation Challenge Expert Workshop, Mission Innovation. Available online: <http://mission-innovation.net/wp-content/uploads/2018/01/Mission-Innovation-IC6-Report-Materials-Acceleration-Platform-Jan-2018.pdf> (accessed on 4 August 2021).
19. Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R.S.; Gold-Parker, A.; Voigt, L.; Brockway, A.M.; Aspuru-Guzik, A. The Harvard Clean Energy Project: Large-scale computational screening and design of organic photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241–2251. [CrossRef]
20. Vom Material zur Innovation. Rahmenprogramm zur Förderung der Materialforschung “Bundesministerium für Bildung und Forschung”. Available online: <https://www.ptj.de/vom-material-zur-innovation> (accessed on 4 August 2021).
21. Voulodimos, A.; Doulami, N.; Doulamis, A.; Protopapadakis, E. Recent developments in Deep Learning for engineering applications. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [CrossRef]
22. Lee, J.G.; Jun, S.; Cho, Y.W.; Lee, H.; Kim, G.B.; Seo, J.B.; Kim, N. Deep learning in medical imaging: General overview. *Korean J. Radiol.* **2017**, *18*, 570–584. [CrossRef] [PubMed]
23. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
24. Tkac, M.; Verner, R. Artificial neural networks in business: Two decades of research. *Appl. Soft Comput.* **2016**, *38*, 788–804. [CrossRef]
25. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef]
26. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
27. Sze, V.; Chen, Y.H.; Yang, T.J.; Emer, J.S. Efficient processing of deep neural networks: A tutorial and survey. *Proc. IEEE* **2017**, *105*, 2295–2329. [CrossRef]
28. Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **2016**, *145*, 170901. [CrossRef]
29. Ghiringelli, L.M.; Vybiral, J.; Levchenko, S.V.; Draxl, C.; Scheffler, M. Big data of materials science: Critical role of the descriptor. *Phys. Rev. Lett.* **2015**, *114*, 105503. [CrossRef]
30. Fedorov, A.; Shamanaev, I.V. Crystal structure representation for neural networks using topological approach. *Mol. Inf.* **2017**, *36*, 1600162. [CrossRef]
31. Thimm, G. Crystal topologies—The achievable and inevitable symmetries. *Acta Crystallogr. A* **2009**, *65*, 213–226. [CrossRef]
32. Jiang, Y.; Chen, D.; Chen, X.; Li, T.; Wei, G.W.; Pan, F. Topological representations of crystalline compounds for the machine-learning prediction of materials properties. *NPJ Comput. Mater.* **2021**, *7*, 28. [CrossRef]
33. Faber, F.; Lindmaa, A.; von Lillienfeld, O.A.; Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **2015**, *115*, 1094–1101. [CrossRef]
34. Saidi, W.A.; Shadid, W.; Castelli, I.E. Machine-learning structural and electronic properties of metal halide perovskites using a hierarchical convolutional neural network. *NPJ Comput. Mater.* **2020**, *6*, 36. [CrossRef]
35. Pathak, Y.; Juneja, K.S.; Varma, G.; Ehara, M.; Priyakumar, U.D. Deep learning enabled inorganic material generator. *Phys. Chem. Chem. Phys.* **2020**, *22*, 26935. [CrossRef]
36. Jha, D.; Ward, L.; Paul, A.; Liao, W.K.; Choudhary, A.; Wolverton, C.; Agrawal, A. ElmNet: Deep learning the chemistry of materials from only elemental composition. *Sci. Rep.* **2018**, *8*, 17593. [CrossRef]
37. Goodall, R.E.A.; Lee, A.A. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nat. Commun.* **2020**, *11*, 6280. [CrossRef]
38. Xie, T.; Grossman, J.C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **2018**, *120*, 145301. [CrossRef] [PubMed]
39. Schmidt, J.; Shi, H.; Borlido, P.; Chen, L.; Botti, S.; Marques, M.A.L. Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chem. Mater.* **2017**, *29*, 5090–5103. [CrossRef]
40. Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **2017**, *8*, 15679. [CrossRef] [PubMed]
41. Choubisa, H.; Askerka, M.; Ryczko, K.; Voznyy, O.; Mills, K.; Tamblyn, I.; Sargent, E.H. Crystal site feature embedding enables exploration of large chemical spaces. *Matter* **2020**, *3*, 433–448. [CrossRef]
42. Brown, I.D. *The Chemical Bond in Inorganic Chemistry. The Bond. Valence Model*, 1st ed.; Oxford University Press: Oxford, UK, 2002; pp. 121–133. [CrossRef]
43. Ryan, K.; Lengyel, J.; Shatruk, M. Crystal structure prediction via Deep Learning. *J. Am. Chem. Soc.* **2018**, *140*, 10158–10168. [CrossRef]
44. Valle, M.; Oganov, A.R. Crystal fingerprint space—A novel paradigm for studying crystal-structure sets. *Acta Crystallogr. A* **2010**, *66*, 507–517. [CrossRef] [PubMed]
45. Ye, W.; Chen, C.; Wang, Z.; Chu, I.H.; Ong, S.P. Deep neural networks for accurate predictions of crystal stability. *Nat. Commun.* **2018**, *9*, 3800. [CrossRef]

46. Gómez-Peralta, J.I.; Bokhimi, X. Discovering new perovskites with artificial intelligence. *J. Solid State Chem.* **2020**, *285*, 121253. [[CrossRef](#)]
47. Gómez-Peralta, J.I. Descubrimiento de Compuestos con Estructura Tipo Perovskita a Través de Inteligencia Artificial y Cálculos Químico Cuánticos. Ph.D. Thesis, National Autonomous University of Mexico, Mexico City, Mexico, 19 February 2021. Available online: <http://132.248.9.195/ptd2020/diciembre/0805987/Index.html> (accessed on 5 August 2021).
48. Patolli 2021. Available online: [https://github.com/gomezperalta/patolli\\_2021](https://github.com/gomezperalta/patolli_2021) (accessed on 4 August 2021).
49. Tilley, R.J.D. *Perovskites: Structure–Property Relationships*, 1st ed.; Wiley: Chichester, UK, 2016; pp. 123–154.
50. Gražulis, S.; Chateigner, D.; Downs, R.T.; Yokochi, A.T.; Quiros, M.; Lutterotti, L.; Manakova, E.; Butkus, J.; Moeck, P.; Le Bail, A. Crystallography Open Database—An open-access collection of crystal structures. *J. Appl. Crystallogr.* **2009**, *42*, 726–729. [[CrossRef](#)]
51. Gražulis, S.; Daškevič, A.; Merkys, A.; Chateigner, D.; Lutterotti, L.; Quirós, M.; Serebryanaya, N.R.; Moeck, P.; Downs, R.T.; LeBail, A. Crystallography Open Database (COD): An open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Res.* **2012**, *40*, D420–D427. [[CrossRef](#)]
52. Gražulis, S.; Merkys, A.; Vaitkus, A.; Okulič-Kazarinas, M. Computing stoichiometric molecular composition from crystal structures. *J. Appl. Crystallogr.* **2015**, *48*, 85–91. [[CrossRef](#)]
53. Vaitkus, A.; Merkys, A.; Gražulis, S. Validation of the Crystallography Open Database using the crystallographic information framework. *J. Appl. Crystallogr.* **2021**, *54*, 661–672. [[CrossRef](#)]
54. Ong, S.P.; Richards, W.D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V.L.; Persson, K.A.; Ceder, G. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013**, *68*, 314–319. [[CrossRef](#)]
55. Woodward, P.M. Octahedral tilting in perovskites. I. Geometrical considerations. *Acta Crystallogr. B* **1997**, *53*, 32–43. [[CrossRef](#)]
56. Woodward, P.M. Octahedral tilting in perovskites. II. Structure stabilizing forces. *Acta Crystallogr. B* **1997**, *53*, 44–66. [[CrossRef](#)]
57. Haas, C. Phase transitions in crystals with the spinel structure. *J. Phys. Chem. Solids* **1965**, *26*, 1225–1232. [[CrossRef](#)]
58. Sickafus, K.E.; Wills, J.M. Structure of spinel. *J. Am. Ceram. Soc.* **1999**, *82*, 3279–3292. [[CrossRef](#)]
59. Momma, K.; Izumi, F. VESTA: A three dimensional visualization system for electronic and structural analysis. *J. Appl. Crystallogr.* **2008**, *41*, 653–658. [[CrossRef](#)]
60. Javed, S.G.; Khan, A.; Majid, A.; Mirza, A.M.; Bashir, J. Lattice constant prediction of orthorhombic ABO<sub>3</sub> perovskites using support vector machines. *Comput. Mater. Sci.* **2007**, *39*, 627–634. [[CrossRef](#)]
61. Majid, A.; Khan, A.; Choi, T.S. Predicting lattice constant of complex cubic perovskites using computational intelligence. *Comput. Mater. Sci.* **2011**, *50*, 1879–1888. [[CrossRef](#)]
62. Majid, A.; Khan, A.; Javed, G.; Mirza, A.M. Lattice constant prediction of cubic and monoclinic perovskites using neural networks and support vector regression. *Comput. Mater. Sci.* **2010**, *50*, 363–372. [[CrossRef](#)]
63. Zhang, Y.; Xu, Z. Machine learning lattice parameters of monoclinic double perovskites. *Int. J. Quantum Chem.* **2021**, *121*, e26480. [[CrossRef](#)]
64. Song, Z.; Zhou, D.; Liu, Q. Tolerance factor and phase stability of the garnet structure. *Acta Crystallogr. C* **2019**, *75*, 1353–1358. [[CrossRef](#)]