batteries

MDPI

Article

Crystal Group Prediction for Lithiated Manganese Oxides Using Machine Learning

Pier Paolo Prosini

ENEA, Energy Department, C.R. Casaccia, Santa Maria di Galeria 301, 00123 Roma, Italy;
pierpaolo.prosini@enea.it

Abstract: This work aimed to predict the crystal structure of a compound starting only from the knowledge of its chemical composition. The method was developed to select new materials in the field of lithium-ion batteries and tested on Li-Fe-O compounds. For each testing compound, the correspondence with respect to the training compounds was evaluated simply by calculating the Euclidean distance existing between the stoichiometric coefficients of the elements constituting the two compounds. At the compound under test was assigned the crystal structure of the training compound for which the distance value was minimum. The results showed that the model can predict the crystalline group of the test compound with an accuracy higher than 80% and a precision higher than 90%, for a cut-off distance higher than four. The method was then used to predict the crystalline group of manganese-based compounds (Li-Mn-O). The analysis conducted on twenty randomly selected compounds showed an accuracy of 70%. Out of ten valid predictions, nine were true positives, with a precision of 90%.

Keywords: crystal structure prediction; machine learning; K-nearest neighbours; lithium-ion battery; cathodes; iron; manganese

1. Introduction

Lithium-ion batteries have, without any doubt, contributed significantly to the development of portable electronics that has occurred over the past thirty years. They are now used for the construction of electric vehicles and for stationary applications serving the electricity grid. These applications have been made possible by the fact that lithium-ion batteries have a higher energy density than other rechargeable battery systems, also thanks to the development and implementation of high energy density active materials [1]. At the basis of their operation there is the process of intercalation. Intercalation is the chemical reaction wherein lithium is inserted into a host matrix with essential retention of the crystal structure. For the intercalation to take place, the material must have a layered or stratified structure and only some materials possess such a structure. The first cathode used in a lithium-ion battery was stratified $LiCoO_2$ in which lithium and cobalt ions are arranged on alternating planes of the rock salt structure with a compact cubic matrix of oxide ions [2]. Spinel lithium manganese oxide ($LiMn_2O_4$), in which lithium occupies the tetrahedral sites and manganese the octahedral ones, can also intercalate lithium ions [3]. As an alternative, transition metal polyanions have been used. Among them, olivine lithium iron phosphate ($LiFePO_4$) has been studied as a suitable cathode material for Li-ion batteries [4]. In theory, materials that crystallize in one of these crystalline forms can be used as a cathode in lithium batteries. For this reason, the prediction of the material structure is determining if it will work or not as the cathode of a Li-ion battery.

The discovery of new materials and the understanding of the composition–structure–property relationships could lead to a rapid improvement in the performance of lithium batteries. Machine learning (ML) is emerging as one of the most promising tools that can accelerate workflows and material property discovery [5]. ML uses data available in the

check for updates

Citation: Prosini, P.P. Crystal Group Prediction for Lithiated Manganese Oxides Using Machine Learning. *Batteries* **2023**, *9*, 112. https://doi.org/10.3390/batteries9020112

Received: 24 December 2022
Revised: 13 January 2023
Accepted: 30 January 2023
Published: 5 February 2023

literature, both experimental and ab initio, to build accurate and computationally light statistical models. Such models can be used to predict the properties of materials that have yet to be synthesized. Alternatively, the models can be used to direct the research activities. Two basic approaches are usually used in ML: supervised learning and unsupervised learning. The main difference is that the former uses labelled data to predict outcomes, while the latter does not. Unsupervised learning is becoming an essential tool to analyse the increasingly large amounts of data produced by atomistic and molecular simulations. Unsupervised learning finds wide application in different fields such as material science, solid state physics, biophysics, and biochemistry [6]. The unsupervised learning can be carried out with different clustering algorithms. These include the K-means clustering algorithm, the hierarchy-based clustering, anomaly detection, the dimensionality reduction techniques such as Principal Component analysis (PCA), the a priori algorithm, and the K-nearest neighbours (K-NN). The K-means algorithm is generally the most known and used clustering method [7]. Distinct patterns are evaluated, and similar datasets are grouped together. The variable k represents the number of groups in the data. It is easy to implement and allows the identification of unknown groups of data from complex datasets. On the other hand, K-means does not allow the development of an optimal set of clusters and, for effective results, the clusters should be selected beforehand. Furthermore, it tends to produce clusters with uniform sizes even when the input data have different sizes. The hierarchical clustering technique builds clusters based on the similarity between different objects in the set. It examines the various characteristics of the data points and looks for the similarity between them [8]. The advantage of hierarchical clustering is the possibility to not pre-specify the clusters. However, it does not work very well on large amounts of data or huge datasets and only gives the best results in some cases. Anomaly detection is the process of finding the patterns in a dataset whose behaviour is not normal or expected [9]. The most relevant areas of application of such methods are seen in medicine and payment systems. The PCA is a statistical method that simplifies the complexity of the data [10]. PCA improves the capabilities of the ML algorithm as it deletes correlated variables that do not contribute to any decision making. Using the PCA can lead to some loss of information if the right number of principal components needed to describe the dataset and its variance are not selected. The a priori algorithm is generally used to evaluate association rules between objects [11]. This process is time-consuming and requires an iterative scan of the entire database for each set of candidates. The k-NN algorithm is a nonparametric method, usually used for classification and regression problems [12]. It is a kind of lazy learning algorithm that does not need offline training. During the classification stage for a given testing example, the k-NN algorithm directly searches through all the training examples by calculating the distances between the testing example and the training data to detect its nearest neighbours and produce the classification output. Since this is a lazy learning algorithm, there is no need to train the model; instead, all data points are used at the time of the forecast.

Machine learning's ability to generalize from a set of training data to discover unknown spaces makes it attractive for many challenges in materials science [13]. However, a key obstacle to widespread application remains the definition of model input data, the so-called descriptors. So far, most applications of ML in materials science have used descriptors based on knowledge of crystal structure. There are many ways to manually encode the crystal structure in three-dimensional space and the chemical composition of a compound into a suitable representation. For example, a compound can be represented as a list of interatomic potentials [14]. Another approach is to create a fingerprint of the compound, transforming the internal coordinates into a fixed set of numbers. For example, this can be performed by projecting coordinates onto a set of basis features or using structure topology [15]. Internal coordinate distribution represents another systematic approach, which has been shown to produce well-performing models applicable throughout the chemical space [16]. However, the use of structure-based descriptions in the search for new compounds limits the resulting models, as they are subject to the same limitations inherent in the ab initio approaches used to determine the descriptors [17]. A possible solution to this problem is to develop descriptors based on stoichiometry alone. In the literature, there are numerous works that have used ML to predict the crystalline structure of compounds. For example, deep neural networks have been proposed as a machine-learning tool for analysing crystallographic data using input data in the form of multi-perspective atomic fingerprints [18]. In about 30% of the cases, the known compositions were exactly identified among the top-10 most likely candidates proposed by the model. To accelerate the search for the crystal structure with the lowest enthalpy of formation, a machine-learning approach to crystal structure prediction was used in which a network of graphs is employed to establish a correlation model between crystal structure and enthalpies of training in the given database [19]. The neural network model can be trained to effectively recognize chemical elements based on the topology of their crystallographic environment. A unique methodology for crystal structure forecasting that relies on a machine-learning algorithm was able to determine the isomorphism of crystal structures composed by two given chemical compositions with an accuracy of around 96.4% [20]. An interesting work has been carried out for the prediction of the lattice parameters of perovskite-type structures [21]. The prediction model accuracy for the crystal structure classification averaged nearly 88% using the genetic algorithm-supported neural network, while for the lattice constant regression model the genetic algorithm-supported support vector regression provided about 95% on average. A ML approach that uses only stoichiometry as input and automatically learns appropriate and systematically amendable descriptors from the data has also been developed [17]. The advantage of this method is that the descriptor becomes automatically improvable as more data become available. This approach is inspired by revolutionary methods in chemistry that directly use a molecular graph as input and learn the best molecule-to-descriptor map from the data. In the same direction, ML algorithms have been proposed and evaluated to determine the type of structure of materials given only their composition [22]. Coupling of different models with different characteristics was used to predict the crystal system and space group of materials. Four types of models for the prediction of crystalline systems and space groups were proposed, trained, and evaluated. In the light of these works, an attempt has been made to predict the structure of a compound starting only from the knowledge of the chemical compositions. In general, this approach is doomed to fail, as two substances can be completely different from a chemical point of view and have the same crystalline structure (isomorphism). Isomorphism can also occur when replacing one atom with another; for this substitution to take place, the atoms must have the same chemical properties (such as, for example, the same atomic radius). This phenomenon is called vicariance. A typical example of an isomorphic mineral is olivine. Olivine forms a complete isomorphic series between forsterite (Mg_2SiO_4) and fayalite (Fe_2SiO_4). The two substitute atoms are therefore iron (Fe) and magnesium (Mg).

The concept behind this work is to restrict our input data to a well-defined class of chemical compounds. It will be presumable that a strong proximity in the chemical composition is reflected in the same spatial organization of the atoms in the space. In this way, it will be sufficient to simply use the stoichiometry of the compound as a descriptor to predict its crystalline structure. Furthermore, this concept can be extended to exploit the phenomenon of vicariance to explore compounds that are very similar to each other. In the mid-1920s, Victor Goldschmidt developed a set of rules of thumb to explain the ion substitution that can occur in a crystalline structure. As regards the dimensions, for the ions of one element to be replaced by another, Goldschmidt stated that the difference between their ionic radii must be less than 15%. The Goldschmidt empirical rule opens new perspectives within the prediction of space groups. In fact, it is possible to imagine, as was performed for iron, that other lithiated oxides containing substitute ions instead of iron and whose chemical formula is identical or similar to each other could have the same crystalline structure. The ionic radius of an element depends primarily on its electric charge and the state of coordination in the crystal. Manganese, with the same electric charge and state of coordination, has an ionic radius that does not exceed the 15% rule indicated by

Goldschmidt. For this reason, it could be possible to predict the crystal structure of lithiated manganese oxides using as the training dataset the structures of lithiated iron oxides.

The aim of this study was to develop a simple and lean ML technique capable of predicting the crystal structure of materials suitable to work as cathodes in Li-ion batteries. For this reason, a ML model that learns the crystal structure of iron lithium oxides to be analysed directly from the crystal structure of the training compounds will be described. The same ML model was applied to predict the crystal group of manganese lithium oxides by replacing iron with manganese, two neighbouring elements in the periodic table that have very strong similarities to each other. Despite its simplicity, this model has proven useful for obtaining highly predictive results.

2. Materials and Methods

2.1. Dataset

The dataset to train and evaluate the prediction model was extracted from the Crystallographic Open Database [23] which is an open-access collection of crystalline structures of inorganic, organic, metal–organic, and mineral compounds. The database currently hosts 495,690 entries. Attention has focused on potentially active iron-containing materials such as cathodes for lithium batteries. Therefore, a search was carried out which returned all the materials in the database, which simultaneously contained lithium, iron, and oxygen. The search returned a dataset containing 720 entries. Compounds containing carbon or with atomic number (AN) exceeding 56, except tungsten (AN = 74), mercury (AN = 80), and lead (AN = 82), were eliminated. This effectively excluded metal–organic compounds, lanthanides, precious metals, uranium, transuranic, and other uncommon elements from the dataset. Thus, the dataset was reduced to 419 entries. Further reduction of the dataset was performed by eliminating those compounds that were repeated multiple times. After these refinements, the final dataset contained 276 entries. Each entry contained the chemical formula of the unit cell and the relative crystallographic group it belonged to.

2.2. Descriptors

The molecular structures were transformed into descriptors before being used to train the machine-learning model. The descriptors used in this work were strings formed by the following characters: the identification number, the chemical formula of the elementary cell, the crystalline group it belongs to, the number of different chemical elements present in the compound, their atomic number, and the stoichiometric coefficient with which the elements appeared in the chemical formula. Some descriptors are shown in Table 1. For example, for the first compound reported in Table 1 are listed: the ID (7221070 in the COD database), the chemical formula of the unit cell, the crystalline phase (14 in the international system which corresponds to the P121/c1 phase of the monoclinic system), the number of the different elements that compose the elementary cell (4), the atomic number of arsenic (33), iron (26), lithium (3), and oxygen (8), followed by the stoichiometric coefficient with which they appear in the chemical formula of the elementary cell (12 for arsenic, 8 for iron, 12 for lithium, and 48 for oxygen). The strings were collected to form the data matrix as shown in Table 1.

Table 1. Graphical representation of the data matrix.

	ID	Unit Cell Formula	Group	N° Elements	Atomic Number/Stoichiometric Coefficients									
1	7221070	$As_{12}Fe_8Li_{12}O_{48}$	14	4	33	26	3	8	12	8	12	48		
2	7222177	$Fe_{12}Li_{28.8}O_{72}P18$	148	4	26	3	8	15	12	28,8	72	18		
3	7222395	$Fe_4Li_4O_{16}P_4$	63	4	26	3	8	15	4	4	16	4		
. . . .														
. . . .														
275	7710154	$Fe_{32}Li_{192}Nb_{192}O_{864}$	56	4	26	3	41	8	32	192	192	864		
276	7221015	$Fe_{0.08}Li_4Nb_{7.92}O_{26}Sr_{4.14}$	139	5	26	3	41	8	38	0.08	4	7.92	26	4.14

2.3. ML Method

The ML method is extremely simple. The matrix that contains the dataset is made up of 276 rows and 100 columns. The number of columns can accommodate compounds containing up to 48 different elements ([100 − 4]/2), a number certainly higher than the maximum number of elements contained in a single formula in the dataset. A graphical representation of the data matrix is shown in Table 1.

The (n − 1) rows represent the training matrix. The stoichiometric coefficients of the compound to be analysed (contained in the nth row) were then subtracted from the (n − 1) rows which form the training matrix. The result in the internal representation is a number representing the Euclidean distance of the stoichiometry of the compound under testing from that of the training element. Mathematically we have:

$$\begin{gathered} d = \sqrt[2]{\sum_{i=1}^{n}(qi-pi)^2} = \\ = \sqrt[2]{(q1-p1)^2+(q2-p2)^2+(q3-p3)^2+\ldots+(qn-pn)^2} \end{gathered} \tag{1}$$

where d is the distance existing between the two compounds, while qi and pi are the stoichiometric coefficients of the i^{th} element which constitutes the formula of the compound to be analysed and that of the training one, respectively.

2.4. Distance Calculation

The single entries have been transformed into a matrix composed of 59 columns by 3 rows. The 58 columns correspond to the maximum number of elements that could be found in our dataset (the first 56 elements, excluding C, plus W, Hg, and Pb). The position within the matrix corresponds to the AN of the elements contained in the formula, while the rows contain the stoichiometric coefficients of the compound under test and that of the training one, respectively. In the third row was placed the intermediate distance value calculated as the square of the difference of the stoichiometric coefficients.

Table 2 shows an example of the matrix when the compound to be analysed is $Fe_2Li_2O_{14}P$ and the training compound is $Fe_2H_2Li_4O_{16}P_4$.

Table 2. Example of the distance matrix.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	…
$Fe_2Li_2O_{14}P_4$			2					14							4											2	
$Fe_2H_2Li_4O_{16}P_4$	2		4					16							4											2	
$(q-p)^2$	4		4					4							0											0	

The values contained in the third row are algebraically added, and the square root of the sum is calculated. This value represents the distance between the two compounds. In the case shown, the distance is equal to the square root of 12, i.e., just over 3.46. The distance value is entered into a distance vector comprising (n − 1) values corresponding to the (n − 1) training compounds.

2.5. Classification Model

The K-nearest neighbours (KNN) method, a non-parametric supervised learning classifier which employs proximity to make classifications or predictions about the clustering of a single data point, was used. While it can be used for regression or classification problems, it is typically used as a classification algorithm, based on the assumption that similar points can be found close to each other. Once the distance vector was completed, the minimum value was searched for in it. The crystallographic group of the training compound with the minimum distance value is then assigned to the compound to be analysed. After this operation, in turn, each of the compounds that were part of the training set become the compound to be analysed, and the entire procedure is repeated. This method is known as one-off cross-validation. This method is generally preferred to the one that leaves out multiple elements because it does not suffer from being computationally intensive and because the number of possible combinations is equal to the number of data points in the original sample. At the end, for each of the n compounds forming the dataset, it is possible to have a prediction of the crystalline group to which it belongs.

2.6. Space Group Prediction

The same technique adopted to classify the compounds based on lithium, oxygen, and iron was used to predict the crystallographic group of the compounds containing lithium, manganese, and oxygen. Manganese was therefore treated as a vicarious atom with respect to iron. Only those compounds whose space group was contained in the previously processed dataset have been selected since, as is logical, it would not be possible for the program to predict space groups not contained within it. The prediction technique was the same as previously reported: the compound was compared with all those present in the database, and the proximity between the cell formulas was calculated. The manganese has been treated as if it was iron in the sense that, in the calculation, the stoichiometric coefficient of the manganese has been subtracted from that of the iron. For this reason, manganese compounds that also contained iron were not taken into consideration. The first six compounds closest to the compound to be tested were then extrapolated. The normalized exponential function was used to normalize the outputs, changing them from weighted-sum values into probabilities that add up to one.

For example, for the compound of formula $Li_8Mn_4O_{12}$, the program found six close neighbours with different space group values, as shown in Table 3.

Table 3. Calculation of the logarithmic cross-entropy loss.

ID	Cell Formula	Group	D	Exp(D)	1/Softmax	Norm	y	yi
7022166	$Fe_2Li_8O_{12}Sb_2$	12	2.828	16,919	201,467	0.873	1	0.873
1000196	$Fe_4Li_{4.66}O_{16}Sb_{2.0122}Sn_{1.3278}$	186	5.736	309,723	11,005	0.048	0	0.926
1001230	$Fe_6Li_4O_{16}Sb_2$	186	6.325	558,110	6107	0.026	0	
4001982	$F_2Fe_2Li_4O_8P_2$	1	6.633	759,948	4485	0.019	0	0.981
7244316	$Fe_4Li_{3.76}N_{0.64}O_{15.36}P_4$	62	6.758	861,316	3957	0.017	0	0.983
4003017	$Fe_{4.1416}Li_{3.86}0_{12}O_{16}P_{3.624}$	26	6.805	902,567	3777	0.016	0	0.984

The first three columns contain, respectively, the ID, the cell formula, and the space group of the six compounds whose distance is closest to the compound under test. The D column contains the distance measured between the compound under test and the reference compounds, as calculated by the model. The Exp(D) column contains the exponential value of the distance (e^D). The 1/Softmax column contains the inverse of the Softmax equation (see supplementary material for explanation). The Norm column contains the probability that the structure of the reference compound is the same as that under test, obtained by normalization of the previous column. Column y assigns a value of "1" to the component with the highest probability value and a "0" to all the others. The yi column contains the corrected probability that corresponds to the probability listed in the Norm column if y = 1 or (1-Norm) if y = 0. The logarithmic cross-entropy loss (LCEL) can be calculated by adding the negative log (with base two) of the corrected probabilities column (yi) divided by the number of columns:

$$\mathrm{LCEL} = {}^{1}\!/_{n} \sum_{1}^{n} -\mathrm{Log}_2\, \mathrm{yi} \tag{2}$$

If two or more entries have the same spatial group, their probability is summed together. For example, in the case shown, the second and third compounds belong to the same crystalline group so that the probability corresponding to this group is obtained by adding the two probabilities (0.048 + 0.026 = 0.074). The corrected probability (yi) is therefore 1 — 0.074 = 0.926.

3. Results

Traditional evaluation metrics were used to determine the effectiveness of the method. To identify the number of true or false positives and negatives, the confusion table shown in Figure S1 was used. To predict a true/false label, a cut-off threshold must be applied, and its value affects the distributions of true/false labels. The distance value was compared to the cut-off value. If the expected space group coincides with the real one (condition expressed as Correct) and the distance value is lower than the cut-off value, it is a true positive. If the distance is greater than the cut-off value, we have a false positive. If the expected space group does not coincide with the real one (condition expressed as Wrong) and the distance value is less than the cut-off value, it is a false negative. If the distance is greater than the cut-off value, we have a true negative. Table 4 reports the values of the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) as a function of the cut-off distance.

Table 4. The values of the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) as a function of the cut-off distance.

Cut-Off Value	TP	TN	FP	FN
1	101	57	105	12
2	143	49	61	19
3	178	39	28	30
4	191	33	15	35
5	192	28	14	41
6	193	24	13	45
7	195	23	11	46
8	195	19	11	49
9	198	16	8	53
10	198	15	8	54

By using the equations from one to six in the supplementary material, it is possible to calculate sensitivity, precision, F1_score, selectivity, accuracy, and false positive ratio for the proposed method. Figure 1 shows the sensitivity (also called recall or true positive ratio) and the precision as the cut-off distance increases.

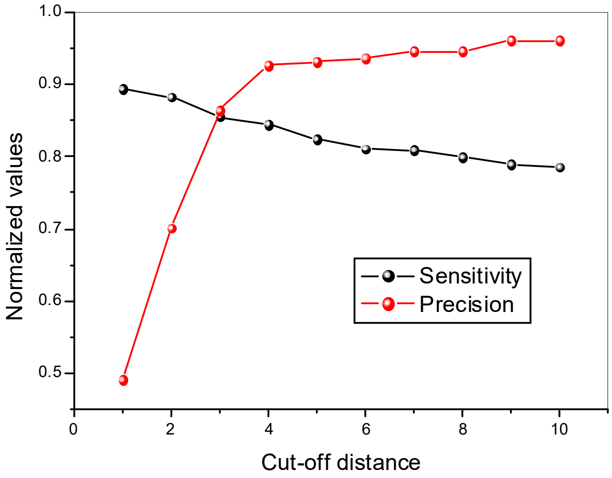

Figure 1. Calculated sensitivity and precision as a function of the cut-off distance for the developed model.

As it is logical, by increasing the cut-off distance an increase in precision and a decrease in sensitivity are observed. The initial value of the sensitivity approaches 0.90. The increase in the cut-off distance leads to a continuous decline in sensitivity. Precision increases from 0.49 to over 0.9 by increasing the cut-off distance by four points. Further increasing the cut-off distance does not significantly change the precision. In general, a model is reliable when it has high precision and high sensitivity. These two values add up in the F1_score, which is higher as the model is reliable. The value of the F1_score is shown in Figure 2.

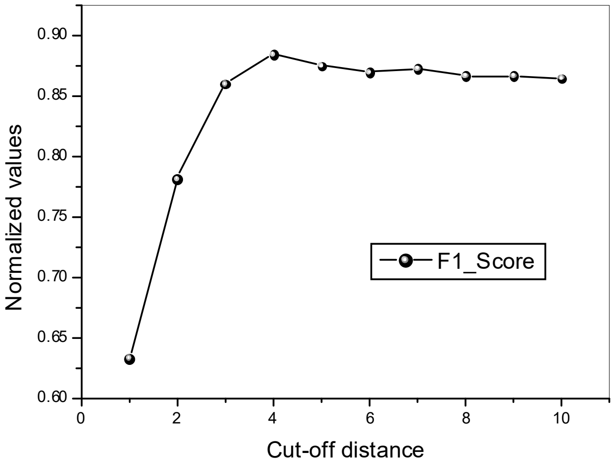

Figure 2. F1_score calculated as harmonic mean between sensitivity and precision as a function of the cut-off distance.

The F1_Score value is relatively low (0.63) for the cut-off distance equal to one but rapidly rises by increasing the cut-off distance and has a maximum for the cut-off distance equal to four. For further increases in the cut-off distance, the F1_score value tends to remain stable above 0.84. Therefore, by limiting the cut-off distance to four, the method is not only effective in correctly classifying the positive results among all the positive ones but also in returning a high ratio between how many cases have been correctly identified as positive and how many have been evaluated as positive.

The selectivity and accuracy of the method are shown in Figure 3. Both tend to increase as the cut-off distance increases, reaching a maximum when the cut-off distance is equal to four. The accuracy of the test shows how many positive results were predicted out of the total number of items evaluated. Its value starts from 0.58 for a cut-off distance equal to one and increases at values around 0.80 for cut-off distances higher than three. The selectivity of a test shows its ability to correctly determine true negatives. The selectivity is low for distances up to three, and it fluctuates between 0.63 and 0.69 for higher distance values so that the model is unable to specify with a high probability whether a negative value is a true negative or rather a false positive.

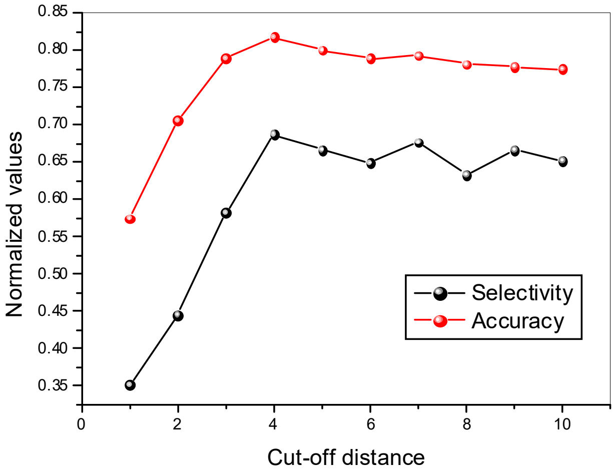

Figure 3. Selectivity and accuracy as a function of the cut-off distance calculated for the developed model.

For this type of application, it is important, in addition to having a high number of true positives, to reduce the false positives, i.e., to avoid identifying compounds which apparently appear to belong to a determinate crystalline group and instead belong to another one. Figure 4 show the false positive ratio; its value is relatively high for a cut-off distance of one and tends to decrease with increasing the distance cut-off. Its value passes from 0.65 to 0.31 by increasing the cut-off value from one to four. For further increments in the cut-off distance the false positive rate oscillates between values lower than 0.35.

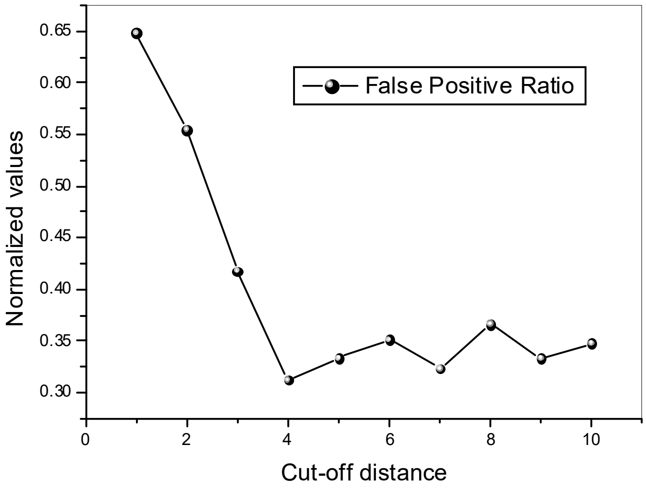

Figure 4. False positive rate (FPR) as a function of the cut-off distance calculated for the developed model.

It should be noted that false negatives do not affect the result, except in quantitative terms since, being erroneously treated as a negative result, these will not be taken into consideration and eliminated together with the true negatives.

The method was then used to predict the crystalline group of compounds containing a different transition metal and in particular manganese. The logarithmic cross-entropy loss was used in evaluating the forecast results obtained on lithiated manganese oxides. If the value of the logarithmic cross-entropy loss was less than 0.25 and the prediction was correct, the result was treated as a true positive. If the result had been wrong, it would have been a false positive. If the value of the logarithmic cross-entropy loss was greater than 0.25 and the prediction was correct, the result was treated as a false negative. If the result had been wrong, it would have been a true negative. The crystallographic group of the training compound with the maximum probability value is then assigned to the compound under testing. Table 5 reports the results obtained with twenty randomly selected lithiated manganese oxides contained in the COD database. The complete calculation of the logarithmic binary cross-entropy loss function can be found from Tables S1–S20.

The first three columns of the table show the ID, the stoichiometric formula of the unitary cell, and the space group of the compound under investigation. This is followed by the minimum value of the distance (as calculated by the program), the logarithmic cross-entropy loss, and the space group of the compound (as predicted by the program). The penultimate column shows if the prediction of the spatial group was successful. The last column reports the true/false condition as expressed by the confusion table shown in Figure S2.

Despite the low number of items, we can try to make a statistic: out of twenty randomly selected items, the program has identified fourteen that are correct with an accuracy of 14/20 = 0.70. Having selected the value of 0.25 of the logarithmic cross-entropy loss as the upper limit to decide if the prediction is true, five positive results are over this threshold (i.e., four false negatives) (FN = 5). The true positives are therefore TP = 14 − 5 = 9 (TP = 9). The sensitivity is 9/(9 + 5) = 0.75. Of the six results not correctly identified, five have a value greater than 0.25 (TN = 5) while one has a value lower than 0.25, which makes them seem correct: it is a false positive (FP = 1). The selectivity is 5/(5 + 1) = 0.83. If we do not take into consideration the ten doubtful entries, i.e., those that have a logarithmic cross-entropy loss value greater than 0.25 and highlighted in red in Table 5, ten entries remain, one of which is incorrect. The precision is equal to 9/9 + 1 = 0.90. The F1_score is equal to 2/(1/0.90 + 1/0.75) = 0.82. Among the incorrectly identified results, the number of false positives (1) is lower than the number of true negatives (5). The false positive rate is FPR = 1/(1 + 5) = 0.17. These values agree with those found previously.

Table 5. Result of the analysis conducted on twenty randomly selected manganese lithium oxides. Entries with a logarithmic cross-entropy loss value greater than the threshold value (0.025) are highlighted in red.

ID Test	Cell Formula	Group (Real)	Minimum Distance	LCEL	Group (Pred.)	Pred. Succ.	T/F P/N
1008373	$Li_8Mn_4O_{12}$	12	2.828	0.077	12	✓	TP
2311016	$Al_{0.68}Li_{0.32}MnO_4$	12	1.861	0.427	2	–	TN
1009066	$Li_{10.304}Mn_{13.728}O_{29.856}$	227	5.375	0.053	227	✓	TP
1532000	$Li_8Mg_{0.328}Mn_{12.348}Ni_{3.16}O_{32}$	212	4.015	0.031	227	✓	FP
1011089	$Li_4Mn_4O_{16}P_4$	62	0.000	0.197	62	✓	TP
1513972	$Li_4Mn_4O_{16}$	63	2.509	0.597	186	–	TN
1513958	$Li_{16}Mn_8O_{24}$	15	11.314	0.658	15	✓	FN
1541720	$Al_{23.0463}B_9Li_{1.1997}Mn_{2.7}Na_{2.67}O_{93}S_{18}$	160	4.005	0.000	160	✓	TP
1513974	$Li_{0.4}Mn_{16}O_{32}$	227	4.400	0.689	227	✓	FN
7022079	$F_8Li_6Mn_2O_{12}P_4$	14	2.828	0.539	12	–	TN
1513960	$Li_{6.64}Mn_{14.08}O_{32}$	227	3.846	0.150	227	✓	TP
1531701	$Cr_{2.4}Cu_{1.52}Li_{7.6}Mn_{12.48}O_{32}$	227	4.049	0.022	227	✓	TP
1514037	$Li_{1.998}Mn_{2.002}O_{4.008}$	225	0.144	0.196	225	✓	TP
4002444	$Co_{0.99999}Li_{2.49}Mn_{0.99999}$ $Ni_{1.002}O_6$	166	1.258	0.152	166	✓	TP
7221081	$Cu_4Li_8Mn_{12}O_{32}$	227	4.000	0.013	227	✓	TP
4335951	$Li_4Mn_2O_{16}S_4$	14	2.000	0.296	14	✓	FN
4343135	$Li_8Mn_{8.0208}O_{32}Ti_{7.9936}$	212	4.214	0.433	212	✓	FN
4110715	$Li_{3.66668}Mn_8O_{16}$	62	2.848	0.400	186	–	TN
1514077	$Li_{1.6}Mn_{15.2}O_{32}$	227	4.389	0.265	227	✓	FN
7045768	$Li_8Mn_4O_{28}P_8$	14	9.402	0.593	15	–	TN

4. Conclusions

In this paper, a very simple machine-learning algorithm for predicting the space groups of transition metal lithiated oxides (Fe and Mg) starting from their compositions has been proposed and evaluated. The calculation routines are extremely light and give results with very low execution times. Several control metrics were used to validate the results. The F1_score scores show that machine-learning algorithms and descriptors, while reliable, still have a fair margin of error, necessitating the development of further adjustments. One possible reason is that the refinement of the data has not been very thorough. There are numerous possibilities for improvement, as the learning model was developed considering only the lithiated iron oxides. Learning models based on more transition metals could be developed in the future to expand the number of crystalline groups that can be detected, thus improving the system's performance. This space-group prediction model enables rapid large-scale structural screening of materials based on their composition alone. This possibility paves the way for the design of new materials, not yet synthesized, predicting their structural feature. In this way, it will be possible to direct the synthesis towards those materials with a crystalline structure such as to be able to intercalate lithium ions.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/batteries9020112/s1. Figure S1: Confusion matrix for determining the reliability of the results obtained on lithiated iron oxides. Figure S2: Confusion table for determining crystallographic group of lithiated manganese oxides. Tables S1–S20: Calculation of the logarithmic binary cross entropy loss function for 20 selected lithiated manganese compounds.

Funding: Part of this research was funded by the Ministry of Enterprise and Made in Italy within the second important project of common European interest in the battery sector (IPCEI Batterie 2).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Manthiram, A. A reflection on lithium-ion battery cathode chemistry. *Nat. Commun.* **2020**, *11*, 1550. [CrossRef]
2. Mizushima, K.; Jones, P.C.; Wiseman, P.J.; Goodenough, J.B. Li_xCoO_2 ($0 < x < 1$): A new cathode material for batteries of high energy density. *Mater. Res. Bull.* **1980**, *15*, 783–798. [CrossRef]
3. Thackeray, M.M.; David, W.I.F.; Bruce, P.G.; Goodenough, J.B. Lithium insertion into manganese spinels. *Mater. Res. Bull.* **1983**, *18*, 461–472. [CrossRef]
4. Padhi, A.K.; Nanjundaswamy, K.S.; Goodenough, J.B. Phospho-Olivines as positive electrode materials for rechargeable lithium batteries. *J. Electrochem. Soc.* **1997**, *144*, 1188–1194. [CrossRef]
5. Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Deng, Z.; Ong, S.P. A Critical Review of Machine Learning of Energy Materials. *Adv. Energy Mater.* **2020**, *10*, 1903242. [CrossRef]
6. Glielmo, A.; Husic, B.E.; Rodriguez, A.; Clementi, C.; Noé, F.; Laio, A. Unsupervised Learning Methods for Molecular Simulation Data. *Chem. Rev.* **2021**, *121*, 9722–9758. [CrossRef]
7. Sinaga, K.P.; Yang, M.-S. Unsupervised K-Means Clustering Algorithm. *IEEE Access* **2020**, *8*, 80716–80727. [CrossRef]
8. Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: An overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2017**, *7*, e1219. [CrossRef]
9. Omar, S.; Ngadi, A.; Jebur, H.H. Machine learning techniques for anomaly detection: An overview. *Int. J. Comput. Appl.* **2013**, *79*, 33–41. [CrossRef]
10. Karamizadeh, S.; Abdullah, S.M.; Manaf, A.A.; Zamani, M.; Hooman, A. An Overview of Principal Component Analysis. *J. Signal Process. Syst.* **2013**, *4*, 173–175. [CrossRef]
11. Al-Maolegi, M.; Arkok, B. An Improved Apriori Algorithm for Association Rules. *Int. J. Nat. Lang. Comp.* **2014**, *3*, 21–29. [CrossRef]
12. Hu, L.Y.; Huang, M.W.; Ke, S.W.; Tsai, C.F. The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus* **2016**, *5*, 1304. [CrossRef]
13. Ho, G.G.; Juhwan, N.; Inkyung, K.; Yousung, J. Machine learning for renewable energy materials. *J. Mater. Chem. A* **2019**, *7*, 17096–17117. [CrossRef]
14. Mishin, Y. Machine-learning interatomic potentials for materials science. *Acta Mater.* **2021**, *241*, 116980. [CrossRef]
15. Chen, X.; Chen, D.; Weng, M.; Jiang, Y.; Wei, G.-W.; Pan, F. Topology-Based Machine Learning Strategy for Cluster Structure Prediction. *J. Phys. Chem. Lett.* **2020**, *11*, 4392–4401. [CrossRef] [PubMed]
16. Faber, F.A.; Christensen, A.S.; Huang, B.; von Lilienfeld, O.A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **2018**, *148*, 241717. [CrossRef]
17. Goodall, R.E.A.; Lee, A.A. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nat. Commun.* **2020**, *11*, 6280. [CrossRef]
18. Ryan, K.; Lengyel, J.; Shatruk, M. Crystal Structure Prediction via Deep Learning. *J. Am. Chem. Soc.* **2018**, *140*, 10158–10168. [CrossRef] [PubMed]
19. Cheng, G.; Gong, X.G.; Yin, W.J. Crystal structure prediction by combining graph network and optimization algorithm. *Nat. Commun.* **2022**, *13*, 1492. [CrossRef]
20. Kusaba, M.; Liu, C.; Yoshida, R. Crystal structure prediction with machine learning-based element substitution, Computational. *Mater. Sci.* **2022**, *211*, 111496. [CrossRef]
21. Jarin, S.; Yuan, Y.; Zhang, M.; Hu, M.; Rana, M.; Wang, S.; Knibbe, R. Predicting the Crystal Structure and Lattice Parameters of the Perovskite Materials via Different Machine Learning Models Based on Basic Atom Properties. *Crystals* **2022**, *12*, 1570. [CrossRef]
22. Yong, Z.; Cui, Y.; Xiong, Z.; Jin, J.; Liu, Z.; Dong, R.; Hu, J. Machine Learning-Based Prediction of Crystal Systems and Space Groups from Inorganic Materials Compositions. *ACS Omega* **2020**, *5*, 3596–3606. [CrossRef]
23. Crystallographic Open Database. Available online: http://www.crystallography.net/cod/ (accessed on 1 February 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.