Supplementary Materials

1. Dataset

The dataset to train and evaluate the CSP model was obtained from the COD. The COD predominantly collects data published in the peer-reviewed scientific papers. Each structure deposited into the COD receives a unique seven-digit number, called COD number. A COD number identifies a particular instance of a structure determination. The data in the COD are stored in the Crystallographic Interchange File/Framework (CIF) format, one structure per file. Each file contains all data necessary to describe the structure, interpret experimental data and find the corresponding publication. The COD website allows for the searching of the database using queries such as unit cell parameters, chemical composition, and bibliographic data (authors, journal names, paper titles). The selection is powered by the Structured Query Language (SQL) database, which is accessible by several protocols. Each structure in the COD crystallographic database is described as an entry in the SQL database. Such entries are generated automatically from the COD CIFs and consist of bibliography and parameters that describe the size and contents of a unit cell, space group, the diffraction experiment, and the quality of the data (such as the R factor and goodness-of-fit parameter). Parameters of the unit cells are stored together with their respective measurement precisions. Both Hermann-Mauguin and Hall symmetry space group symbols are included in the entries. To avoid errors, the Hall and Hermann-Mauguin space group symbols are regenerated from the symmetry lines, replacing the original entries if necessary. The correspondence of cell lengths and angles to the space group symmetry constraints are checked as well. Three datasets were extracted from the COD containing lithiated oxide of iron manganese and cobalt. The search returned a dataset containing 720 entries for iron, 618 for manganese and 220 for cobalt. 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 avoided taking into consideration organometallic compounds, lanthanides, precious metals, uranium, transuranic, and other less common elements. In such a way, the dataset was decreased to 419 entries for iron, 439 for manganese and 109 for cobalt. A further reduction was carried out by eliminating compounds repeated several times. The final dataset comprised 276 entries for iron, 220 for manganese and 93 for cobalt.

2. Descriptors

The descriptors used in this work were contained in strings each of which represents a specific crystallographic structure. Each string was made up of the following elements: the ID number (as reported in the COD), the elementary cell chemical formula, the crystalline group, the number of chemical elements present in the compound, the atomic number of the elements, and the stoichiometric coefficient with which the elements appeared in the cell formula. Three of these strings are reported in Tab. 1. For example, for $Fe_{1.9}Li_{2.1}O_4$ are reported: the ID (1541312), the crystalline phase (225 in the International System which corresponds to the Fm-3m phase of the cubic system ), the number of elements that make up the unit cell (3), the atomic number of iron (26), lithium (3), and oxygen (8) followed by the stoichiometric coefficient with which they appear in the chemical formula of the elemental cell (1.9 for iron, 2.1 for lithium and 4 for oxygen).

| **ID** | **Cell formula** | **Crystal Group** | **N° of elements** | **Atomic number 1th elem.** | **Atomic number 2nd elem.** | **Atomic number 3th elem.** | **Stech. Coeff. 1th elem.** | **Stech. Coeff. 2nd elem.** | **Stech. Coeff. 3th elem.** |
|---|---|---|---|---|---|---|---|---|---|
| 1541312 | $Fe_{1.9}Li_{2.1}O_4$ | 225 | 3 | 26 | 3 | 8 | 1.9 | 2.1 | 4 |
| 1541958 | $Fe_{20}Li_4O_{32}$ | 212 | 3 | 26 | 3 | 8 | 20 | 4 | 32 |
| 1542046 | $Fe_8Li_{40}O_{32}$ | 61 | 3 | 26 | 3 | 8 | 8 | 40 | 32 |

Table 1. Graphical representation of part of the data matrix.

3. ML method

The iron dataset was used as training matrix while the manganese and cobalt matrix were use as test matrices. The stoichiometric coefficients of the compound to be analysed was subtracted from the 276 rows which formed the training matrix. The result, in the internal feature, is a number representing the Euclidean

distance (d) that separates the stoichiometry of the compound under test from that of the training element. Mathematically we have:

$$d = \sqrt[2]{\sum_{i=1}^{n}(\boldsymbol{qi} - \boldsymbol{pi})^2} = \sqrt[2]{(\boldsymbol{q1} - \boldsymbol{p1})^2 + (\boldsymbol{q2} - \boldsymbol{p2})^2 + \cdots + (\boldsymbol{qn} - \boldsymbol{pn})^2} \quad (1)$$

Where qi and pi are the stoichiometric coefficients of the i$^{th}$ element which constitutes the formula of the compound to be analysed and of the training one, respectively. Iron, manganese, and cobalt were treated as a generic transition metal and the stoichiometric coefficient of manganese or cobalt was subtracted from the stoichiometric coefficient of iron. For this reason, compounds containing the three elements simultaneously were not taken into consideration. The individual entries have been transformed into a distance matrix composed of 59 columns and 3 rows. The 59 columns represent the maximum number of elements in the data set, consisting of the first 56 elements (excluding C) to which W, Hg and Pb must be added. The first two rows of the distance matrix contained the stoichiometric coefficients of the compound under test and that of the training one, respectively. The third row contained 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 $Co_{3.54}Li_{0.46}O_4$ and the training compound is $Fe_{1.9}Li_{2.1}O_4$.

| **Atomic Number** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **….** | **….** | **22** | **23** | **24** | **TM (25, 26, 27)** | **28** | **29** | **….** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Co_{3.54}Li_{0.46}O_4$ | | | 0,46 | | | | | 4 | | | | | | | 3,54 | | | |
| $Fe_{1.9}Li_{2.1}O_4$ | | | 2,10 | | | | | 4 | | | | | | | 1,90 | | | |
| $(q-p)^2$ | | | 2,6896 | | | | | 0 | | | | | | | 2,6896 | | | |

Table 2. The distance matrix used to calculate the Euclidean distance between $Co_{3.54}Li_{0.46}O_4$ and $Fe_{1.9}Li_{2.1}O_4$.

To evaluate the distance that separates the two compounds, the values contained in the third column were algebraically added and the square root of the sum was calculated. In the case shown, the distance is equal to the square root of 5,3792, i.e., just over 2,319. The distance value is entered into a distance vector comprising (n) values corresponding to the (n) testing compounds.

4. 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. 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.

5. Space group prediction

Manganese and cobalt were treated as vicarious atoms with respect to iron. The compound under test was compared with all those present in the training dataset and the proximity between the cell formulas was calculated. The first six compounds closest to the compound to be tested were then extrapolated. The normalized exponential function (Softmax, equation 2) was used to normalize the outputs, converting them from weighted-sum values into probabilities that add up to one. The Softmax function has the following expression:

$$\mathbf{Softmax}_{(i)} = \frac{e^{d(A,B)}}{\sum_{i=1}^{n} e^{d(A,B)}} = \frac{\mathbf{1}}{\boldsymbol{pi}} \quad (2)$$

The Softmax function has been reversed to give greater weight to smaller distances. Each value in the Softmax function output was interpreted as the membership probability for each space group. At the compound under test was assigned the space group of the training compound that compared with the highest percentage.

6. Evaluation metrics

Traditional evaluation metrics such as accuracy, precision, sensitivity, selectivity, and F1_score were used to evaluate the performance of the method. In addition, false positive ratio (FPR) was evaluated as an additional performance measure.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (5)$$

$$Selectivity = \frac{TN}{TN+FP} \quad (6)$$

$$F1\_Score = \frac{2}{\frac{1}{Precision}+\frac{1}{Sensitivity}} \quad (7)$$

$$FPR = \frac{FP}{FP+TN} \quad (8)$$

The logarithmic cross-entropy loss (LCEL) was used in evaluating the predicted results.

$$LCEL = -\frac{1}{n}\sum_{i=1}^{n}[yi * \log(pi) + (1 - yi) * \log(1 - pi)] \quad (9)$$

For example, for the compound of formula the $Co_{3.54}Li_{0.46}O_4$, the program found six close neighbours with different space group as reported in table 3:

| ID | Cell formula | Group | d | Exp(d) | 1/Softmax | Norm | y | yi | LCEL |
|---|---|---|---|---|---|---|---|---|---|
| 1541312 | $Fe_{1.9}Li_{2.1}O_4$ | 225 | 2,319 | 10,169 | 15,882 | 0,314 | 1 | 0,314 | 1,672 |
| 1008561 | $Fe_{4.0008}Li_{1.7392}O_3$ | 227 | 2,420 | 11,246 | 14,360 | 0,284 | 0 | 0,716 | 0,481 |
| 9012897 | $Fe_{0.99999}HLiO_5P$ | 2 | 3,121 | 22,678 | 7,121 | 0,141 | 0 | 0,859 | 0,219 |
| 1525051 | $Cl_2Fe_2LiO_2$ | 59 | 3,265 | 26,192 | 6,166 | 0,122 | 0 | 0,878 | 0,187 |
| 4342290 | $Fe_{2.376}H_2Li_{1.58}O_2Se_2$ | 129 | 3,820 | 45,590 | 3,542 | 0,140 | 0 | 0,860 | 0,217 |
| 4342295 | $Fe_{2.378}H_2Li_{1.586}O_2Se_2$ | 129 | 3,820 | 45,619 | 3,540 | | | | |

Table 3. Table for the calculation of the logarithmic cross-entropy loss.

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. Column (d) 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. The column (Norm) 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 column Norm if y= 1 or (1-Norm) if y = 0. The logarithmic cross-entropy loss can be calculated by adding the negative log (with base two) of the corrected probabilities column (yi) divided by the number of entries:

$$\mathbf{LCEL} = {}^{1}/_{5}\sum_{1}^{5} -\log_2 yi = 0,555$$

If two or more entries have the same spatial group their probability is summed together. For example, in the case shown, the fifth and sixth compounds belong to the same crystalline group (129) so that the probability corresponding to this group is obtained by adding the two probabilities (0.07+0.07 = 0.14). The corrected probability (yi) is therefore (1-0.14) = 0.86. The crystallographic group of the training compound with the maximum probability value is then assigned to the compound under testing.