LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-577712

Analysis of the Band Gap Type Dataset

C. Kamath

August 29, 2012

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Analysis of the Band Gap Type Dataset

Chandrika Kamath
kamath2@llnl.gov
Lawrence Livermore National Laboratory
August 24, 2012

1 Introduction

This report summarizes the work done as part of the *MINDES: Data Mining for Inverse Design* project [3] to mine the datasets generated by the Center for Inverse Design [1], an Energy Frontier Research Center (EFRC) of the Office of Science, US Department of Energy. In the course of the MINDES project, two datasets were analyzed, one on the formation enthalpy of spinels, and the other on the band gap type of the class of ternary compounds generated at NREL; the latter is the focus of this report.

The dataset considered in this analysis is from computer simulations of ternary compounds. Given the composition of a compound, that is, the three elements and their percentages in the compound, the simulations calculate various quantities, and associate with each compound a band gap type, which can take values 1, 2, 3, or 4. The primary focus of the analysis was to determine which properties of the elements are associated with band gap type 1 compounds, which are potentially more useful as photovoltaic material. Any insights obtained on the other band gap types in the course of the analysis were also deemed to be of interest. Further, it is expected that the insights gained in the course of this study will also apply when attempting to correlate other physical properties of compounds to their individual constituents.

This analysis falls in the broad area of design of computer experiments [2], where an ensemble of simulations is used to guide physical experiments and gain insights into the design space which maps the inputs of the simulations to the output(s). As the simulations are often computationally expensive, the ensemble must be carefully designed to gain the greatest insights into the physical phenomenon of interest using as few simulations as possible. A possible solution is to consider an incremental approach where we analyze the input/output data from the simulations that have already been run to identify the next set of simulations such that these new simulations would add the greatest insights, by either refining the original data set in a region of interest, or exploring new regions in the design space.

Therefore, in the context of our problem, we can analyze the simulations that have been run thus far to understand what properties of the elements comprising a compound are relevant to the band gap of type 1. By combining this information with physics insights, we can suitably create other compounds and expect that their simulation will likely indicate the compounds to be of band gap type 1.

In this report, we start by describing the dataset in Section 2, followed by the initial exploratory analysis of the data in Section 3. The detailed analysis of the data to gain insights into band gap type 1 compounds is described in Section 4, followed by Section 5 which provides other insights into the data, including an analysis of band gap types 2, 3, and 4. We conclude with a brief summary and some thoughts for future work.

Disclaimer - This report is written from the point of view of a data miner, not a materials scientist, a physicist, or a chemist. The analysis is purely data driven and is not influenced by any domain-specific biases. Also, any conclusions drawn must be interpreted with care as the analysis reflects the characteristics and quality of the data provided; the availability of additional data may change the results.

2 Description of the data

The data were provided in the form of a table, consisting of 487 compounds (also referred to as instances), each described by 83 features (see Table 3 in Appendix A). Of these 83 features, 3 are the atomic species (i.e., elements) that make up the compound, referred to as "A", "B", and "C"; 3 are the compositions of each species (the values are all less than 1.0 and sum to 1.0), referred to as "p", "q", and "r"; and 3 are other variables ("E1", "E4", and "sg"). "sg" is the space group, reflecting the crystal symmetry and "E1" and "E4" are properties of the compound. Since our goal is to use the properties of the atomic species to predict the gap type of the compound, we ignore features E1 and E4. Of the remaining 74 features, 25 describe species A, 23 describe species B, and 26 describe species C.

Associated with each compound is the gaptype, which takes on values 1 through 4. Of the 487 instances, 76 are of gaptype 1 (15.6%), 100 of gaptype 2 (20.5%), 133 of gaptype 3 (27.3%), and 178 of gaptype 4 (36.6%).

Observation: In the context of our analysis, this is an unbalanced data set as the percentage of gap type 1 compounds is quite small. This can make it difficult to ascertain if an observation is a physics insight or simply the result of too few samples.

We also observe that the dataset does not include the same set of properties for each of the three species as some properties are unavailable for some elements. Table 4 in Appendex A lists each of the three species A, B, and C, and the properties that are available for them.

An analysis of the features indicated that, for a compound, each feature was generated by taking the value of a property of the element and weighting it by its composition. This meant that if a compound had a certain element, say Ax, occurring at species "A", at a composition of 0.25, then the features corresponding to "A" for that compound would be the same as the "A" features for any other compound which also had Ax occurring at species "A" at a composition of 0.25.

Observation: This approach to generating features from the properties results in several repeating values, which can cause problems with some algorithms. It is also unclear if these features appropriately represent the characteristics that determine the band gap type of a compound. In other words, does the current representation of a compound capture what distinguishes one band type from another, or should we also consider a different representation?

3 Exploratory analysis

The first step in the analysis was to explore the data to check for inaccuracies or errors, and to evaluate the quality of the data for the task at hand. This resulted in the following observations:

- We found one compound, of gap type 2, whose space group was not listed. This was removed, resulting in 486 instances in the dataset.
- We found two pairs of instances which were exact duplicates. Since this was a small number, these instances were not removed from the dataset.
- We found five pairs of compounds which had the same values of "A", "B", "C", "p", "q", "r", and sg, but different band gap types (as well as different E1 and E4 values). These pairs indicate that the dataset contains multiple instances of compounds with the same composition and space group, but different crystal structure. These pairs, their line numbers in the original data file, and their key features are:

```
Values of  A, B, C, p, q, r, E1, E4, sg, and band gap type:

lines 81 and 133:
Na Ga S 0.428571 0.142857 0.428571 4.630000 0.010000 14 2
Na Ga S 0.428571 0.142857 0.428571 4.900000 0.010000 14 3

lines 100 and 162
Cu Al O 0.250000 0.250000 0.500000 4.860000 0.950000 166 3
Cu Al O 0.250000 0.250000 0.500000 3.910000 0.150000 166 4

lines 244 and 297
Ag P S 0.307692 0.153846 0.538462 2.940000 0.000000 15 1
Ag P S 0.307692 0.153846 0.538462 0.160000 0.060000 15 2

lines 245 and 439
Na P O 0.277778 0.166667 0.555556 5.990000 0.000000 15 1
Na P O 0.277778 0.166667 0.555556 7.100000 0.620000 15 4

lines 255 and 396
Na P O 0.200000 0.200000 0.600000 8.760000 0.000000 14 1
Na P O 0.200000 0.200000 0.600000 9.120000 0.070000 14 3
```

  Since "A", "B", "C", "p", "q", and "r" determine the features that are derived from the properties of the elements, this inconsistency indicates that some important features related to the crystal structure are missing from the dataset. This is also supported by the fact that the values of E1 and E4 are different for the five pairs listed above.

- We observed that there are compounds with same values of "A", "B", "C", "p", "q", and "r", but different sg and band gap types. This indicates that sg is likely an important feature in determining the gap type. See Sections 4 and 5 for more details.

Species A		Species B		Species C	
Element	Count	Element	Count	Element	Count
Ag	75	Al	42	S	130
Cu	86	Ga	42	Se	108
K	76	In	50	Te	46
Li	59	B	36	O	202
Na	87	Y	24		
Rb	50	Tl	23		
Cs	53	Sc	14		
		Ta	24		
		Bi	30		
		P	38		
		As	36		
		Sb	72		
		Nb	23		
		V	32		

Table 1: Counts of the different elements that are used for species "A", "B", and "C". Note that the elements do not overlap, that is, an element appears only in one of the three columns.

- Table 1 lists the elements used for "A", "B", and "C" and the number of times each element is used in the dataset. Note that the three groups do not overlap, that is, an element is never used for species "A" in one compound and in species "B" or "C" in another compound. Note also that elements for species "A" are from columns 1 and 11 of the periodic table; elements for species "B" are from columns 3, 5, 13, and 15 of the periodic table; and elements for species "C" are from column 16 of the periodic table

  Observation: When we consider the 76 instances of band gap type 1, we found that certain species occur rarely in certain locations. For example, for the element in species "B", there is one instance with V, two with Nb, and three with Y and Ta. Similarly, for the element in species "A", there are only three instances with Rb. This may be a reflection of the dataset, or an indication that certain elements in certain locations are unlikely to result in a compound of band gap type 1.

4 Detailed analysis for band gap type 1

Based on the exploratory analysis, there are several factors which make the analysis of band gap type 1 difficult, including:

- the relatively small percentage (15.6%) of instances with band gap type 1,
- the inconsistency in the dataset, where compounds with the same composition and space group can have different band gap types, and

Property (to keep)	for species	while removing the following correlated properties
pauling	A, C	mulliken_jaffe, sanderson
pauling	A	bulk_modulus
single_bond_radius	B, C	triple_bond_radius
pauling	C	allen
atomic_number	A, B, and C	atomic_weight
single_bond_radius	A, B, and C	orbital_radii_p, orbital_radii_s, covalent_radius, double_bond_radius, atomic_radius
pauling	A, B, and C	ionization_energies_1, pettifor, allred_chow

Table 2: Identifying the properties to remove as they are highly correlated to other properties which are included in the analysis. See also Table 4.

- the approach used to generate the features from the properties of the elements forming the compound. Since there are few elements used in species "A", "B" and "C", and a limited number of values for "p", "q" and "r", there are compounds with different band gap types but the same values for certain features. This repetition in feature values causes problems with some classification algorithms, such as decision trees.

Based on these observations, we expect that a straightforward application of analysis tools is unlikely to yield insights into band gap type 1 compounds. This was confirmed when we analyzed the data using parallel plots - we did not find any of the features discriminating and there appeared to be little difference between the compounds belonging to the four gap types. Further, as expected, a decision tree classifier was not be able to automatically classify compounds of band gap type 1 with a low error rate.

Given this, we next considered three avenues of exploration:

- First, we reduced the number of features by removing those that were correlated (see Table 4 in Appendix A). We take a conservative approach, removing a property Y only if it is correlated to another property Z for each of the atomic species for which values are available. Table 2 lists the properties that are retained for each species and the ones that are removed as their correlation coefficient with a property that is retained is greater than 0.9.

- Second, we considered alternate ways of representing the data derived from the original set of features. This included (i) using properties of the elements directly, without weighting them by the percentages (referred to by the prefix AU, BU, and CU to indicate unscaled values); (ii) taking ratios of the properties which are available for all three species (referred to by the prefix ABR, BCR, and ACR); (iii) taking ratios of the unscaled properties which are available for all three species (referred to by the prefix ABUR, BCUR, and ACUR); (iv) taking differences of the properties available for all three species (referred to by the prefix ABD, BCD, and ACD); and (v) taking differences of the unscaled properties available for all three species (referred to by the prefix ABUD, BCUD, and ACUD). Thus, the feature ABURmelting_point represents the ratio of the

unscaled melting points of the elements for species "A" and "B", where the original values, which had been scaled by the relative percentages of the two species in the original dataset, are now unscaled.

Note that these alternate representations use properties which are available for all the three species, so properties such as density and van der Waal's radius are not included (see Table 4 in Appendix A). The data without the correlated features and with only features available for all three species is referred to as the "cleaned" feature set.

- Third, we tried to determine if, using the new features above, we could identify parts of feature space where a large percentage of points were likely to be of band gap type 1.

  To ensure that the "rules" which identify parts of feature space with a large percentage of type 1 compounds are meaningful, we impose two constraints: first, we require that at least 30% of the compounds satisfying the rule be of type 1, which is twice the percentage of type 1 compounds in the data and second, we require that the number of type 1 compounds satisfying the rule be at least 25 (one third the number of type 1 compounds in the original data).

Our results, summarized in Table 5, Appendix B, indicate that, for the original dataset, we could not find a region which had a high percentage of type 1 compounds. However, if we remove the correlated features and features that are available for only one or two species, we find one region where 30 of the 67 compounds in the region (=45%) are of type 1. Using ratios or differences of the original scaled or the unscaled properties provides additional insights into regions with a greater than normal percentage of type 1 compounds. These rules also indicate which features are relevant to defining these regions. For example, the occurrence of sg in many of the rules indicates that the space group is an important features for band gap type 1, as predicted in Section 3.

The rules in Table 5, Appendix B, do not form an exhaustive list; they are the rules we could identify easily in the data. There may be other regions in the data which also have a high percentage of type 1 compounds but are not as easily identifiable. Once we have identified the rules, it may be possible to gain further insights by extracting the compounds which satisfy the rules from the dataset and examining their properties to determine if we can detect any similarities.

5 Other insights on the data

We next repeated the analysis with compounds of band gap type 2, 3, and 4 to determine if we could discover rules similar to the ones we found for band gap type 1 compounds. Our results are summarized in Tables 6 - 8, Appendix B. We also make the following observations on these compounds:

- Band gap type 2 compounds: We found only two cases where species "B" has value Y, two with value Sc, and three with Nb.

  Of the 486 compounds, 99 (= 20.37%) were of band gap type 2. Thus, to identify useful rules, we required that the number of compounds satisfying the rule should be at least 33 (= one-third of the number of type 2 compounds)), of which, at least 40% (twice 20.37%) were of band gap type 2. The results are presented in Table 6, Appendix B.

- Band gap type 3 compounds: Of the 486 compounds, 133 (= 27.36%) were of band gap type 3. Thus, to identify useful rules, we required that the number of compounds satisfying the rule should be at least 44 (= one-third of the number of type 3 compounds)), of which, at least 55% (twice 27.36%) were of band gap type 3. The results are presented in Table 7, Appendix B.

- Band gap type 4 compounds: We found only one case where species "B" has value Sc.

  Of the 486 compounds, 178 (= 36.6%) were of band gap type 4. Thus, to identify useful rules, we required that the number of compounds satisfying the rule should be at least 60 (= one-third of the number of type 4 compounds)), of which, at least 72% (twice 36.6%) were of band gap type 4. However, we found that these constraints were too stringent and it was difficult to identify rules that satisfied both constraints. Since there is a higher percentage of type 4 compounds in the data, it may be unreasonable to simply double this percentage to obtain one of the constraints, as we do for other band-gap types. So, we relaxed the constraints, and in Table 8, Appendix B include rules that do not quite meet the original constraints.

These analysis results confirm the relationship between the space group and the band gap type. To explore the issue further, Table 9, Appendix B is a simple count of the number of times a space group results in a particular gap type. The table only lists the space groups which have more than 3 instances. We observe that space group 122 is associated with type 1 or 2 compounds, and space group 198 is linked to type 3 compounds, while space group 2 favors type 4 compounds.

6 Conclusions and future work

In this brief report, we discussed the analysis of a data set of ternary compounds in an attempt to determine the properties of the elements that lead to a specific band gap type compound. The original focus was on band-gap type 1 compounds, though the scope of the study was expanded to include band-gap types 2, 3, and 4 as well. The analysis was made difficult by the small size of the dataset, the relatively small number of compounds of certain types, inconsistencies in the dataset, and appropriateness of the representation of the compounds. As a result, we found that traditional data mining techniques, such as classification algorithms, could not be used to atuomatically assign a gap type to a compound based on its composition.

Despite these issues with the quality of the dataset, we wanted to determine if it was possible to gain some insights into what properties of the elements resulted in a specific gap type of compound. We considered alternative representations and determined regions of feature space which had a higher percentage of a specific type of compounds than found in the full dataset. The resulting rules also provide some indication of which features are important.

Based on the analysis, an obvious next step would be to address the inconsistency in the dataset which indicates that additional information about the crystal structure (beyond its space group) is necessary to uniquely determine the band-gap type. Additional data points, prompted perhaps by the rules identified through the analysis, would also be helpful. And finally, the question of how to represent each compound remains an open one; a representation that helps to separate the different band gap types would be ideal.

7 Acknowledgment

This work was performed as part of the ARRA-funded DOE SciDAC-e project, *MINDES - Data Mining for Inverse Design*, in support of the Center for Inverse Design EFRC. The analysis was done in collaboration with Alberto Franceschetti from NREL, who provided the data and the domain expertise. The analysis of band-gap type 1 was funded by the SciDAC-e program; the analysis of band-gap types 2-4 was unfunded work. I appreciate the support and interest of Alberto Franceschetti, as well as the EFRC Directors, William Tumas and Alex Zunger.

LLNL-TR-577712: This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

[1] Center for Inverse Design web page, 2012. `http://www.centerforinversedesign.org/`.

[2] Fang, K.-T., Li, R., and Sudjianto, A. *Design and Modeling for Computer Experiments*. Chapman and Hall/CRC Press, Boca Raton, FL, 2005.

[3] MINDES: Data Mining for Inverse Design Project web page, 2012. `https://computation.llnl.gov/casc/StarSapphire/MINDES.html`.

A List of variables in the dataset

A	B	C	p	q
r	E1	E4	sg	Amulliken_jaffe
Asingle_bond_radius	Apauling	Amolar_volume	Abulk_modulus	Asanderson
Aatomic_weight	Amelting_point	Aorbital_radii_s	Aorbital_radii_p	Athermal_conductivity
Aionization_energies_1	Aionization_energies_2	Avaporization	Aatomic_number	Acovalent_radius
Afusion	Apettifor	Aatomization	Aelectron_affinity	Aboiling_point
Adensity	Adouble_bond_radius	Aallred_rochow	Aatomic_radius	Bsingle_bond_radius
Bpauling	Bmolar_volume	Batomization	Batomic_weight	Btriple_bond_radius
Bmelting_point	Borbital_radii_s	Borbital_radii_p	Bthermal_conductivity	Bionization_energies_1
Bionization_energies_2	Bvaporization	Batomic_number	Bcovalent_radius	Bfusion
Bpettifor	Belectron_affinity	Bboiling_point	Bdensity	Bdouble_bond_radius
Ballred_rochow	Batomic_radius	Cmulliken_jaffe	Csingle_bond_radius	Cpauling
Cmolar_volume	Catomization	Csanderson	Catomic_weight	Ctriple_bond_radius
Cmelting_point	Callen	Corbital_radii_s	Corbital_radii_p	Cthermal_conductivity
Cionization_energies_1	Cionization_energies_2	Cvaporization	Catomic_number	Ccovalent_radius
Cfusion	Cpettifor	Cvan_der_waals_radius	Celectron_affinity	Cboiling_point
Cdouble_bond_radius	Callred_rochow	Catomic_radius		

Table 3: Features for the data set. A, B, and C are the three atomic species. p, q, and r, are their respective percentages (they add to 1.0). E1 and E4 are properties of the compounds and are not used in the analysis. sg is the space group. The remaining features represent the properties of the different atomic species as indicated by the first character in the feature.

Property for species A	Property for species B	Property for species C
Amulliken_jaffe (X)		Cmulliken_jaffe (X)
Asingle_bond_radius	Bsingle_bond_radius	Csingle_bond_radius
Apauling	Bpauling	Cpauling
Amolar_volume	Bmolar_volume	Cmolar_volume
Abulk_modulus (X)		
Asanderson (X)		Csanderson (X)
Aatomic_number	Batomic_number	Catomic_number
Aatomic_weight (X)	Batomic_weight (X)	Catomic_weight (X)
Amelting_point	Bmelting_point	Cmelting_point
Aorbital_radii_s (X)	Borbital_radii_s (X)	Corbital_radii_s (X)
Aorbital_radii_p (X)	Borbital_radii_p (X)	Corbital_radii_p (X)
Athermal_conductivity	Bthermal_conductivity	Cthermal_conductivity
Aionization_energies_1 (X)	Bionization_energies_1 (X)	Cionization_energies_1 (X)
Aionization_energies_2	Bionization_energies_2	Cionization_energies_2
Avaporization	Bvaporization	Cvaporization
Acovalent_radius (X)	Bcovalent_radius (X)	Ccovalent_radius (X)
Afusion	Bfusion	Cfusion
Apettifor (X)	Bpettifor (X)	Cpettifor (X)
Aatomization	Batomization	Catomization
Aelectron_affinity	Belectron_affinity	Celectron_affinity
Aboiling_point	Bboiling_point	Cboiling_point
Adensity	Bdensity	
Adouble_bond_radius (X)	Bdouble_bond_radius (X)	Cdouble_bond_radius (X)
Aallred_rochow (X)	Ballred_rochow (X)	Callred_rochow (X)
Aatomic_radius (X)	Batomic_radius (X)	Catomic_radius (X)
	Btriple_bond_radius (X)	Ctriple_bond_radius (X)
		Callen (X)
		Cvan_der_waals_radius

Table 4: Properties for the atomic species A, B, and C organized by species to indicate missing properties: A has 25, B has 23, and C has 26. Several of the properties are correlated to others (correlation coefficient greater than 0.90) and have been marked with an X to indicate they are not being considered further in the analysis. We take a conservative approach, removing a property Y only if it is correlated to another property Z for each of the atomic species for which values are available. See also Table 2.

B Analysis results

Dataset	Rules	# results returned	#results of type 1	percent
Original scaled data with correlated variables	-	-	-	-
Original scaled data, with cleaned feature set	Bmelting_point $<$ 107.438 and Cionization_energies_2 $<$ 1.126e+06	67	30	44.8%
Unscaled data, with cleaned feature set	sg $<$ 141 and CUpauling $<$ 3.44 and sg $\geq$ 15 and AUsingle_bond_radius $<$ 128	43	20	46.5 %
	sg $<$ 141 and CUpauling $<$ 3.44 and sg $\geq$ 15 and BUmolar_volume $<$ 18.19	120	37	30.8%
Ratios of original scaled data, with cleaned feature set	BCRmelting_point $<$ 0.581793	124	46	37.1%
	BCRmelting_point $<$ 0.581793 and sg $<$ 126	79	38	48.1%
	BCRmelting_point $<$ 0.581793 and ACRmolar_volume $<$ 0.396468	50	28	56%
	BCRmelting_point $<$ 0.581793 and ACRmolar_volume $<$ 0.396468 and sg $<$ 126	28	23	82%
Differences of original scaled data, with cleaned feature set	sg $<$ 141 and sg $\geq$ 20 and ACDmolar_volume $<$ -0.731583	110	44	40%
Ratios of unscaled data, with cleaned feature set	sg $<$ 141 and ABURmelting_point $\geq$ 1.01359	108	44	40.7%
	sg $<$ 141 and ABURmelting_point $\geq$ 1.01359 and ABURmolar_volume $<$ 1.027	71	31	43.7%
Differences of unscaled data, with cleaned feature set	sg $<$ 141 and ABUDmelting_point $\geq$ 4.11516	108	44	40.7%
	sg $<$ 141 and ABUDmelting_point $\geq$ 4.11516 and ABUDvaporization $<$ 85300	52	31	59.6%
	sg $<$ 141 and ABUDmelting_point $\geq$ 4.11516 and ABUDvaporization $<$ 85300 and BCUDelectron_affinity $<$ -150100	34	26	76.5%

Table 5: Rules for gap type 1 indicating some of the regions where there is a higher percentage of type 1 compounds than in the full dataset. The cleaned data refers to the original data minus the correlated features and features unavailable for all three species. For rules involving the feature sg, a rule of the form "sg $<$ 141", should be interpreted to mean the 15 values of sg in Table 9 which have values lower than 141.

Dataset	Rules	# results returned	#results of type 2	percent
Original scaled data with correlated variables	Bpettifor $<$ 0.308571 and sg $<$ 143.5 and Bthermal_conductivity $<$ 10.8 and Amolar_volume $\geq$ 17.2275	40	25	62.5 %
Original scaled data, with cleaned feature set	Bmelting_point $\geq$ 107.438 and Bionization_energies_2 $<$ 227843	64	24	37.5 %
Unscaled data, with cleaned feature set	sg $<$ 141 and CUpauling $<$ 3.44 and sg $\geq$ 15 and AUsingle_bond_radius $\geq$ 128	103	40	38.8 %
	sg $<$ 141 and sg $\geq$ 15 and AUsingle_bond_radius $\geq$ 128 and BUfusion $<$ 36000	132	51	38.6 %
Ratios of original scaled data, with cleaned feature set	BCRmelting_point $\geq$ 0.581793 and sg $<$ 141 and ABRboiling_point $\geq$ 0.351574 and sg $\geq$ 15	107	45	42.1%
Differences of original scaled data, with cleaned feature set	sg $<$ 141 and sg $\geq$ 20 and ACDmolar_volume $\geq$ -0.731583	44	26	59.1 %
Ratios of unscaled data, with cleaned feature set	sg $<$ 141 and sg $\geq$ 15 and ABURelectron_affinity $\geq$ 0.49031 and ABURfusion $<$ 0.645414	101	41	40.6%
	sg $<$ 141 and sg $\geq$ 15 and ABURelectron_affinity $\geq$ 0.49031 and ABURfusion $<$ 0.338889	77	37	48.0%
Differences of unscaled data, with cleaned feature set	sg $<$ 141 and sg $\geq$ 15 and BCUDpauling $<$ -0.37 and ABUDatomization $<$ -190500	79	36	45.6%
	sg $<$ 141 and sg $\geq$ 15 and BCUDpauling $\geq$ -1.345 and BCUDpauling $<$ -0.37 and ABUDatomization $<$ -190500	49	31	63.3%

Table 6: Rules for gap type 2 indicating some of the regions where there is a higher percentage of type 2 compounds than in the full dataset. The cleaned data refers to the original data minus the correlated features and features unavailable for all three species. For rules involving the feature sg, a rule of the form "sg $<$ 141", should be interpreted to mean the 15 values of sg in Table 9 which have values lower than 141.

Dataset	Rules	# results returned	#results of type 3	percent
Original scaled data with correlated variables	Bpettifor < 0.308571 and sg ≥ 143.5	83	50	60.2 %
Original scaled data, with cleaned feature set	Bmelting_point ≥ 107.438 and Bionization_energies_2 ≥ 227843 and Apauling < 0.4825 and sg ¿ 141	89	44	49.4 %
Unscaled data, with cleaned feature set	sg ≥ 141 and BUionization_energies_2 $< 1.8207e+06$	126	67	53.2 %
	sg ≥ 141 and BUionization_energies_2 $< 1.8207e+06$ and BUsingle_bond_radius < 163	110	60	54.5 %
Ratios of original scaled data, with cleaned feature set	BCRmelting_point ≥ 0.581793 and sg ≥ 141	132	65	49.2%
	BCRmelting_point ≥ 0.581793 and sg ≥ 141 and BCRpauling < 0.198643	40	29	72.5%
Differences of original scaled data, with cleaned feature set	sg ≥ 141 and BCDthermal_conductivity < 3.99	54	32	59.2 %
Ratios of unscaled data, with cleaned feature set	sg ≥ 141 and ABURmelting_point < 0.848183	111	59	53.1%
	sg ≥ 141 and ABURmelting_point < 0.848183 and sg < 225	91	53	58.2 %
Differences of unscaled data, with cleaned feature set	sg ≥ 141 and ABUDmelting_point $<$ -207.87	99	56	56.5%

Table 7: Rules for gap type 3 indicating some of the regions where there is a higher percentage of type 3 compounds than in the full dataset. The cleaned data refers to the original data minus the correlated features and features unavailable for all three species. For rules involving the feature sg, a rule of the form "sg < 141", should be interpreted to mean the 15 values of sg in Table 9 which have values lower than 141.

Dataset	Rules	# results returned	#results of type 4	percent
Original scaled data with correlated variables	Bpettifor $\geq$ 0.308571 and Aorbital_radii_p $\geq$ 0.175951	135	73	54.1 %
Original scaled data, with cleaned feature set	Bmelting_point $\geq$ 107.438 and sg $<$ 14	49	31	63.3 %
Unscaled data, with cleaned feature set	sg $\geq$ 141 and BUionization_energies_2 $\geq$ 1.8207e+06	46	27	58.7 %
	sg $<$ 15 and CUpauling $<$ 3.44	88	40	45.5%
	sg $<$ 15 and CUpauling $<$ 3.44 and AUsingle_bond_radius $\geq$ 128	81	37	45.7%
Ratios of original scaled data, with cleaned feature set	BCRmelting_point $\geq$ 0.58179 and sg $<$ 141 and ABRboiling_point $<$ 0.351574	74	44	59.5%
	BCRmelting_point $\geq$ 0.58179 and sg $\geq$ 141 and BCRpauling $\geq$ 0.198643	94	43	45.7%
Differences of original scaled data, with cleaned feature set	sg $<$ 20	160	76	47.5 %
	sg $<$ 20 and ABDionization_energies_2 $<$ 124988	52	36	69.2 %
	sg $\geq$ 166 and BCDthermal_conductivity $\geq$ 3.99	96	48	50.0%
Ratios of unscaled data, with cleaned feature set	sg $<$ 15	117	53	45.3%
	sg $<$ 15 and ACURelectron_affinity $<$ 0.422695	96	45	46.9 %
	sg $\geq$ 141 and ABURmelting_point $\geq$ 0.848183	61	35	57.4 %
Differences of unscaled data, with cleaned feature set	sg $\geq$ 141 and ABUDmelting_point $\geq$ -207.87	73	39	53.4%
	sg $<$ 15 and BCUDatomization $<$ 559000	116	53	45.7 %

Table 8: Rules for gap type 4 indicating some of the regions where there is a higher percentage of type 4 compounds than in the full dataset. The cleaned data refers to the original data minus the correlated features and features unavailable for all three species. For rules involving the feature sg, a rule of the form "sg $<$ 141", should be interpreted to mean the 15 values of sg in Table 9 which have values lower than 141.

SG value	Type 1	Type 2	Type 3	Type 4	Total
2	2	1	3	16	22
11	1	0	5	1	7
12	0	7	4	16	27
14	8	6	16	18	48
15	3	11	2	19	35
19	0	2	2	4	8
31	4	2	1	0	7
33	10	2	3	1	16
55	1	0	0	3	4
61	0	3	0	1	4
62	3	16	9	9	37
63	0	3	0	2	5
121	2	1	1	1	5
122	18	6	0	1	25
140	0	3	0	3	6
148	0	2	1	4	7
161	0	1	1	5	7
166	1	5	11	18	35
194	1	3	9	11	24
198	3	1	12	0	16
215	0	0	5	3	8
216	5	0	3	0	8
217	0	5	2	0	7
225	2	1	5	16	24
227	0	0	3	3	6

Table 9: Count (if total > 3) of different band types for each space group.