**DISCOVERY OF DIRECT BAND GAP PEROVSKITES FOR LIGHT HARVESTING BY USING MACHINE LEARNING**

Smarak Ratha, Sudha Priyanga G.b, Nagappan N.a and Tiju Thomasa$

aDepartment of Metallurgical and Materials Engineering, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India – 600036

bDepartment of Physics, Mepco Schlenk Engineering College, Sivakasi, Tamil Nadu, India – 626005

$ Corresponding author. Tel: +91-44-2257-5781 (Office), Fax: +91-44-2257-4752, Email: tt332@cornell.edu

**DATA AND CODE**

To make our work reproducible, we have uploaded the following files in our GitHub repository located at https://github.com/smarakrath/MI-2021- ,

1. ABX3\_NBG.csv – This is the dataset containing the chemical formula, space group, Materials Project ID and the nature of band gap of 1528 ABX3 compounds.
2. ABX3.json – This file contains the same dataset along with pymatgen structure objects for all the materials.
3. ABX3\_with\_features.csv – This file contains the entire dataset along with the 118 features which were used for training the ML algorithms.
4. ABX3 (creation of pymatgen structure objects).ipynb – This Jupyter notebook file contains the code for creating pymatgen structure objects for all the materials in the dataset.
5. ABX3 (final).ipynb – This Jupyter notebook file contains the code for executing all the steps of our work from the beginning to the end.
6. Observations.csv – This file contains the entire dataset along with the actual and predicted (by XGBOOST) values of the nature band gap of all the materials. This file can be used to find out which materials were misclassified by the XGBOOST classifier.

**LIST OF FEATURES**

The following features were created using the ‘Structural Heterogeneity’ module of Matminer:

1. mean absolute deviation in relative bond length
2. max relative bond length
3. min relative bond length
4. minimum neighbor distance variation
5. maximum neighbor distance variation
6. range neighbor distance variation
7. mean neighbor distance variation
8. avg\_dev neighbor distance variation
9. mean absolute deviation in relative cell size

The following features were created using the ‘Meredig’ module of Matminer:

1. H fraction
2. He fraction
3. Li fraction
4. Be fraction
5. B fraction
6. C fraction
7. N fraction
8. O fraction
9. F fraction
10. Ne fraction
11. Na fraction
12. Mg fraction
13. Al fraction
14. Si fraction
15. P fraction
16. S fraction
17. Cl fraction
18. Ar fraction
19. K fraction
20. Ca fraction
21. Sc fraction
22. Ti fraction
23. V fraction
24. Cr fraction
25. Mn fraction
26. Fe fraction
27. Co fraction
28. Ni fraction
29. Cu fraction
30. Zn fraction
31. Ga fraction
32. Ge fraction
33. As fraction
34. Se fraction
35. Br fraction
36. Kr fraction
37. Rb fraction
38. Sr fraction
39. Y fraction
40. Zr fraction
41. Nb fraction
42. Mo fraction
43. Tc fraction
44. Ru fraction
45. Rh fraction
46. Pd fraction
47. Ag fraction
48. Cd fraction
49. In fraction
50. Sn fraction
51. Sb fraction
52. Te fraction
53. I fraction
54. Xe fraction
55. Cs fraction
56. Ba fraction
57. La fraction
58. Ce fraction
59. Pr fraction
60. Nd fraction
61. Pm fraction
62. Sm fraction
63. Eu fraction
64. Gd fraction
65. Tb fraction
66. Dy fraction
67. Ho fraction
68. Er fraction
69. Tm fraction
70. Yb fraction
71. Lu fraction
72. Hf fraction
73. Ta fraction
74. W fraction
75. Re fraction
76. Os fraction
77. Ir fraction
78. Pt fraction
79. Au fraction
80. Hg fraction
81. Tl fraction
82. Pb fraction
83. Bi fraction
84. Po fraction
85. At fraction
86. Rn fraction
87. Fr fraction
88. Ra fraction
89. Ac fraction
90. Th fraction
91. Pa fraction
92. U fraction
93. Np fraction
94. Pu fraction
95. Am fraction
96. Cm fraction
97. Bk fraction
98. Cf fraction
99. Es fraction

100) Fm fraction

101) Md fraction

102) No fraction

103) Lr fraction

104) mean AtomicWeight

105) mean Column

106) mean Row

107) range Number

108) mean Number

109) range AtomicRadius

110) mean AtomicRadius

111) range Electronegativity

112) mean Electronegativity

113) avg s valence electrons

114) avg p valence electrons

115) avg d valence electrons

116) avg f valence electrons

117) frac s valence electrons

118) frac p valence electrons

119) frac d valence electrons

120) frac f valence electrons

Thus, 9 features were created using the ‘Structural Heterogeneity’ module and 120 features were created using the ‘Meredig’ module. After the generation of the correlation matrix, 11 features were removed which are listed as follows:

1) mean neighbor distance variation

2) range neighbor distance variation

3) min relative bond length

4) frac f valence electrons

5) mean Electronegativity

6) mean Number

7) frac p valence electrons

8) mean absolute deviation in relative cell size

9) frac d valence electrons

10) mean Row

11) mean AtomicRadius

Out of the 129 features initially generated, 11 were removed because they were highly correlated. Thus, the remaining 118 features were used for training the ML algorithms.