Supplementary Information (SI) for Journal of Materials Chemistry C.

Supplementary:
Predicting Bandgap of ABX_2 Materials: Supervised Machine Learning on Small Datasets

Upendra Kumar,[1,*] Muhammad Wasi Ullah,[1] Vipin Kumar,[2] Joshua M R Muir,[1,†] and Feiwu Zhang[1,‡]

[1]*State Key Laboratory of Critical Mineral Research and Exploration, Institute of Geochemistry, Chinese Academy of Sciences, Guiyang 550081, China.*
[2]*Department of Electronics and Communication Engineering, School of Engineering, SR University, Warangal-506371, Telangana, India.*

[*] upendra@mail.gyig.ac.cn
[†] j.m.r.muir@mail.gyig.ac.cn
[‡] zhangfeiwu@vip.gyig.ac.cn

CONTENTS

I. DATASET

To focus exclusively on semiconductors and insulators i.e., materials with non-zero bandgaps, we excluded all entries with zero bandgap values (five in total) from the regression analysis, resulting in a final dataset of 92 samples (97 -5 = 92).

	compound	BandGap (eV)	Volume (Å^3)	vdW Rad$_{avg}$(pm)	vdW Rad$_{min}$(pm)	vdW Rad$_{max}$(pm)	CR$_{avg}$(pm)	CR$_{min}$(pm)	CR$_{max}$(pm)	AR$_{avg}$(pm)	AR$_{min}$(pm)	AR$_{max}$(pm)	AD$_{avg}$(pm)	AD$_{min}$(pm)	AD$_{max}$(pm)	Phase
0	$ZnGeN_2$	2.44	181.39	168.33	139.00	211.00	104.33	71.00	122.00	108.33	65.00	135.00	4.42	0.81	7.13	ST
1	$ZnGeN_2$	3.02	181.83	168.33	139.00	211.00	104.33	71.00	122.00	108.33	65.00	135.00	4.42	0.81	7.13	KT
2	$ZnGeN_2$	2.73	181.56	168.33	139.00	211.00	104.33	71.00	122.00	108.33	65.00	135.00	4.42	0.81	7.13	CP
3	$AlGaN_2$	4.59	178.46	175.33	155.00	187.00	104.67	71.00	122.00	106.67	65.00	130.00	3.14	0.81	5.91	ST
4	$AlGaN_2$	4.63	178.36	175.33	155.00	187.00	104.67	71.00	122.00	106.67	65.00	130.00	3.14	0.81	5.91	KT
5	$AlGaN_2$	4.46	178.51	175.33	155.00	187.00	104.67	71.00	122.00	106.67	65.00	130.00	3.14	0.81	5.91	CP
6	$ZnSiN_2$	4.52	164.13	168.00	139.00	210.00	101.33	71.00	122.00	103.33	65.00	135.00	3.42	0.81	7.13	ST
7	$AlGaAs_2$	1.76	367.95	185.33	184.00	187.00	120.67	119.00	122.00	123.33	115.00	130.00	4.78	2.70	5.91	ST
8	$AlGaAs_2$	2.06	367.95	185.33	184.00	187.00	120.67	119.00	122.00	123.33	115.00	130.00	4.78	2.70	5.91	KT
9	$AlGaAs_2$	2.13	368.07	185.33	184.00	187.00	120.67	119.00	122.00	123.33	115.00	130.00	4.78	2.70	5.91	CP
10	$AlGaP_2$	2.47	329.24	183.67	180.00	187.00	116.67	107.00	122.00	118.33	100.00	130.00	3.48	1.82	5.91	ST
11	$AlGaP_2$	2.71	329.30	183.67	180.00	187.00	116.67	107.00	122.00	118.33	100.00	130.00	3.48	1.82	5.91	KT
12	$AlGaP_2$	2.76	329.30	183.67	180.00	187.00	116.67	107.00	122.00	118.33	100.00	130.00	3.48	1.82	5.91	CP
13	$GaInN_2$	1.32	215.88	178.33	155.00	193.00	111.67	71.00	142.00	116.67	65.00	155.00	4.68	0.81	7.31	ST
14	$GaInN_2$	1.50	215.60	178.33	155.00	193.00	111.67	71.00	142.00	116.67	65.00	155.00	4.68	0.81	7.31	KT
15	$GaInN_2$	1.22	215.74	178.33	155.00	193.00	111.67	71.00	142.00	116.67	65.00	155.00	4.68	0.81	7.31	CP
16	$AlInN_2$	2.64	203.87	177.33	155.00	193.00	111.33	71.00	142.00	115.00	65.00	155.00	3.61	0.81	7.31	ST
17	$AlInN_2$	3.00	203.20	177.33	155.00	193.00	111.33	71.00	142.00	115.00	65.00	155.00	3.61	0.81	7.31	KT
18	$AlInN_2$	2.70	203.16	177.33	155.00	193.00	111.33	71.00	142.00	115.00	65.00	155.00	3.61	0.81	7.31	CP
19	$CdSiN_2$	2.45	188.76	174.33	155.00	210.00	108.67	71.00	144.00	110.00	65.00	155.00	3.93	0.81	8.65	ST
20	$CdSiN_2$	3.05	188.94	174.33	155.00	210.00	108.67	71.00	144.00	110.00	65.00	155.00	3.93	0.81	8.65	KT
21	$CdSiN_2$	3.16	187.61	174.33	155.00	210.00	108.67	71.00	144.00	110.00	65.00	155.00	3.93	0.81	8.65	CP
22	$ZnGeP_2$	0.88	316.45	176.67	139.00	211.00	116.33	107.00	122.00	120.00	100.00	135.00	4.76	1.82	7.13	ST
23	$ZnGeP_2$	1.73	316.78	176.67	139.00	211.00	116.33	107.00	122.00	120.00	100.00	135.00	4.76	1.82	7.13	KT
24	$ZnGeP_2$	1.94	316.85	176.67	139.00	211.00	116.33	107.00	122.00	120.00	100.00	135.00	4.76	1.82	7.13	CP
25	$CdGeN_2$	1.15	208.75	174.67	155.00	211.00	111.67	71.00	144.00	115.00	65.00	155.00	4.93	0.81	8.65	ST
26	$CdGeN_2$	2.04	209.26	174.67	155.00	211.00	111.67	71.00	144.00	115.00	65.00	155.00	4.93	0.81	8.65	KT
27	$CdGeN_2$	1.73	208.27	174.67	155.00	211.00	111.67	71.00	144.00	115.00	65.00	155.00	4.93	0.81	8.65	CP
28	$GaInP_2$	1.87	365.76	186.67	180.00	193.00	123.67	107.00	142.00	128.33	100.00	155.00	5.01	1.82	7.31	ST
29	$GaInP_2$	2.16	365.24	186.67	180.00	193.00	123.67	107.00	142.00	128.33	100.00	155.00	5.01	1.82	7.31	KT
30	$GaInP_2$	2.06	365.42	186.67	180.00	193.00	123.67	107.00	142.00	128.33	100.00	155.00	5.01	1.82	7.31	CP
31	$AlInP_2$	2.62	364.04	185.67	180.00	193.00	123.33	107.00	142.00	126.67	100.00	155.00	3.94	1.82	7.31	ST
32	$AlInP_2$	2.98	363.66	185.67	180.00	193.00	123.33	107.00	142.00	126.67	100.00	155.00	3.94	1.82	7.31	KT
33	$AlInP_2$	2.89	363.64	185.67	180.00	193.00	123.33	107.00	142.00	126.67	100.00	155.00	3.94	1.82	7.31	CP
34	$AlInAs_2$	1.55	405.59	187.33	184.00	193.00	127.33	119.00	142.00	131.67	115.00	155.00	5.25	2.70	7.31	ST
35	$AlInAs_2$	1.70	405.08	187.33	184.00	193.00	127.33	119.00	142.00	131.67	115.00	155.00	5.25	2.70	7.31	KT
36	$AlInAs_2$	1.67	405.22	187.33	184.00	193.00	127.33	119.00	142.00	131.67	115.00	155.00	5.25	2.70	7.31	CP
37	$GaInAs_2$	0.27	407.41	188.33	185.00	193.00	127.67	119.00	142.00	133.33	115.00	155.00	6.32	5.73	7.31	ST
38	$GaInAs_2$	0.43	406.71	188.33	185.00	193.00	127.67	119.00	142.00	133.33	115.00	155.00	6.32	5.73	7.31	KT
39	$GaInAs_2$	0.38	407.10	188.33	185.00	193.00	127.67	119.00	142.00	133.33	115.00	155.00	6.32	5.73	7.31	CP
40	$CdSiAs_2$	0.15	380.71	184.33	158.00	210.00	124.67	111.00	144.00	126.67	110.00	155.00	5.57	2.33	8.65	ST
41	$CdSiAs_2$	1.47	379.89	184.33	158.00	210.00	124.67	111.00	144.00	126.67	110.00	155.00	5.57	2.33	8.65	KT
42	$CdSiAs_2$	1.34	379.43	184.33	158.00	210.00	124.67	111.00	144.00	126.67	110.00	155.00	5.57	2.33	8.65	CP
43	$CdGeAs_2$	0.11	400.11	184.67	158.00	211.00	127.67	119.00	144.00	131.67	115.00	155.00	6.57	5.32	8.65	KT
44	$CdGeAs_2$	0.07	399.98	184.67	158.00	211.00	127.67	119.00	144.00	131.67	115.00	155.00	6.57	5.32	8.65	CP
45	$ZnGeAs_2$	0.07	358.94	178.33	139.00	211.00	120.33	119.00	122.00	125.00	115.00	135.00	6.06	5.32	7.13	ST
46	$ZnGeAs_2$	0.58	359.20	178.33	139.00	211.00	120.33	119.00	122.00	125.00	115.00	135.00	6.06	5.32	7.13	KT
47	$ZnGeAs_2$	0.50	359.26	178.33	139.00	211.00	120.33	119.00	122.00	125.00	115.00	135.00	6.06	5.32	7.13	CP
48	$ZnSnN_2$	0.94	212.68	170.33	139.00	217.00	110.67	71.00	139.00	115.00	65.00	145.00	5.08	0.81	7.31	ST
49	$ZnSnN_2$	0.97	213.04	170.33	139.00	217.00	110.67	71.00	139.00	115.00	65.00	145.00	5.08	0.81	7.31	KT
50	$ZnSnN_2$	0.79	213.22	170.33	139.00	217.00	110.67	71.00	139.00	115.00	65.00	145.00	5.08	0.81	7.31	CP
51	$ZnSiAs_2$	0.78	342.27	178.00	139.00	210.00	117.33	111.00	122.00	120.00	110.00	135.00	5.06	2.33	7.13	ST
52	$ZnSiAs_2$	1.81	342.30	178.00	139.00	210.00	117.33	111.00	122.00	120.00	110.00	135.00	5.06	2.33	7.13	KT
53	$ZnSiAs_2$	1.95	342.28	178.00	139.00	210.00	117.33	111.00	122.00	120.00	110.00	135.00	5.06	2.33	7.13	CP
54	$CdSiP_2$	1.21	335.64	182.67	158.00	210.00	120.67	107.00	144.00	121.67	100.00	155.00	4.27	1.82	8.65	ST
55	$CdSiP_2$	2.70	335.73	182.67	158.00	210.00	120.67	107.00	144.00	121.67	100.00	155.00	4.27	1.82	8.65	KT
56	$CdSiP_2$	2.33	334.95	182.67	158.00	210.00	120.67	107.00	144.00	121.67	100.00	155.00	4.27	1.82	8.65	CP
57	$CdGeP_2$	0.36	355.55	183.00	158.00	211.00	123.67	107.00	144.00	126.67	100.00	155.00	5.26	1.82	8.65	ST
58	$CdGeP_2$	1.78	355.29	183.00	158.00	211.00	123.67	107.00	144.00	126.67	100.00	155.00	5.26	1.82	8.65	KT
59	$CdGeP_2$	1.62	354.99	183.00	158.00	211.00	123.67	107.00	144.00	126.67	100.00	155.00	5.26	1.82	8.65	CP
60	$CdSnN_2$	0.10	244.97	176.67	155.00	217.00	118.00	71.00	144.00	121.67	65.00	155.00	5.59	0.81	8.65	ST
61	$CdSnN_2$	0.45	245.51	176.67	155.00	217.00	118.00	71.00	144.00	121.67	65.00	155.00	5.59	0.81	8.65	KT
62	$CdSnN_2$	0.15	245.50	176.67	155.00	217.00	118.00	71.00	144.00	121.67	65.00	155.00	5.59	0.81	8.65	CP
63	$ZnSnAs_2$	0.21	400.03	180.33	139.00	217.00	126.67	119.00	139.00	131.67	115.00	145.00	6.72	5.73	7.31	ST
64	$ZnSnAs_2$	0.27	400.34	180.33	139.00	217.00	126.67	119.00	139.00	131.67	115.00	145.00	6.72	5.73	7.31	KT
65	$ZnSnAs_2$	0.24	400.63	180.33	139.00	217.00	126.67	119.00	139.00	131.67	115.00	145.00	6.72	5.73	7.31	CP
66	$ZnSnP_2$	1.39	356.34	178.67	139.00	217.00	122.67	107.00	139.00	126.67	100.00	145.00	5.42	1.82	7.31	ST
67	$ZnSnP_2$	1.67	356.57	178.67	139.00	217.00	122.67	107.00	139.00	126.67	100.00	145.00	5.42	1.82	7.31	KT
68	$ZnSnP_2$	1.69	356.84	178.67	139.00	217.00	122.67	107.00	139.00	126.67	100.00	145.00	5.42	1.82	7.31	CP
69	$CdSnP_2$	0.75	399.77	185.00	158.00	217.00	130.00	107.00	144.00	133.33	100.00	155.00	5.93	1.82	8.65	ST
70	$CdSnP_2$	1.29	400.10	185.00	158.00	217.00	130.00	107.00	144.00	133.33	100.00	155.00	5.93	1.82	8.65	KT
71	$CdSnP_2$	1.14	400.28	185.00	158.00	217.00	130.00	107.00	144.00	133.33	100.00	155.00	5.93	1.82	8.65	CP
72	$ZnSiP_2$	1.53	300.31	176.33	139.00	210.00	113.33	107.00	122.00	115.00	100.00	135.00	3.76	1.82	7.13	ST
73	$ZnSiP_2$	2.35	300.56	176.33	139.00	210.00	113.33	107.00	122.00	115.00	100.00	135.00	3.76	1.82	7.13	KT
74	$ZnSiP_2$	2.20	300.44	176.33	139.00	210.00	113.33	107.00	122.00	115.00	100.00	135.00	3.76	1.82	7.13	CP
75	$CdSnAs_2$	0.07	446.11	186.67	158.00	217.00	134.00	119.00	144.00	138.33	115.00	155.00	7.23	5.73	8.65	ST
76	$CdSnAs_2$	0.05	446.14	186.67	158.00	217.00	134.00	119.00	144.00	138.33	115.00	155.00	7.23	5.73	8.65	KT
77	$CdSnAs_2$	0.13	446.31	186.67	158.00	217.00	134.00	119.00	144.00	138.33	115.00	155.00	7.23	5.73	8.65	CP
78	$ZnGeSb_2$	0.05	447.38	185.33	139.00	211.00	127.00	120.00	139.00	135.00	125.00	145.00	6.38	5.32	7.13	KT
79	$ZnGeSb_2$	0.02	447.96	185.33	139.00	211.00	127.00	120.00	139.00	135.00	125.00	145.00	6.38	5.32	7.13	CP
80	$CdSiSb_2$	0.88	474.04	191.33	158.00	210.00	131.33	111.00	144.00	136.67	110.00	155.00	5.89	2.33	8.65	KT
81	$CdSiSb_2$	0.75	474.07	191.33	158.00	210.00	131.33	111.00	144.00	136.67	110.00	155.00	5.89	2.33	8.65	CP
82	$ZnSiSb_2$	0.78	431.46	185.00	139.00	210.00	124.00	111.00	139.00	130.00	110.00	145.00	5.38	2.33	7.13	KT
83	$ZnSiSb_2$	1.15	432.01	185.00	139.00	210.00	124.00	111.00	139.00	130.00	110.00	145.00	5.38	2.33	7.13	CP
84	$CdGeSb_2$	0.09	492.76	191.67	158.00	211.00	134.33	120.00	144.00	141.67	125.00	155.00	6.89	5.32	8.65	KT
85	$CdGeSb_2$	0.19	493.09	191.67	158.00	211.00	134.33	120.00	144.00	141.67	125.00	155.00	6.89	5.32	8.65	CP
86	$ZnSnSb_2$	0.02	493.42	187.33	139.00	217.00	133.33	122.00	139.00	141.67	135.00	145.00	7.04	6.69	7.31	ST
87	$ZnSnSb_2$	0.05	493.83	187.33	139.00	217.00	133.33	122.00	139.00	141.67	135.00	145.00	7.04	6.69	7.31	KT
88	$ZnSnSb_2$	0.04	494.12	187.33	139.00	217.00	133.33	122.00	139.00	141.67	135.00	145.00	7.04	6.69	7.31	CP
89	$CdSnSb_2$	0.06	544.10	193.67	158.00	217.00	140.67	139.00	144.00	148.33	145.00	155.00	7.55	6.69	8.65	ST
90	$CdSnSb_2$	0.06	544.27	193.67	158.00	217.00	140.67	139.00	144.00	148.33	145.00	155.00	7.55	6.69	8.65	KT
91	$CdSnSb_2$	0.09	544.54	193.67	158.00	217.00	140.67	139.00	144.00	148.33	145.00	155.00	7.55	6.69	8.65	CP

TABLE I. Descriptor table used in ML Model.

II. STATISTICAL ANALYSIS

```
from scipy import stats

bandgap_1 = df[df['BandGap_class'] == 1]
bandgap_0 = df[df['BandGap_class'] == 0]

# Mean and standard deviation
print("Mean and Std for Bandgap = 1")
print(bandgap_1[['Volume', 'density_Avg', 'atom_rad_Avg']].mean())
print(bandgap_1[['Volume', 'density_Avg', 'atom_rad_Avg']].std())

print("\nMean and Std for Bandgap = 0")
print(bandgap_0[['Volume', 'density_Avg', 'atom_rad_Avg']].mean())
print(bandgap_0[['Volume', 'density_Avg', 'atom_rad_Avg']].std())

# t-test for Volume * density_Avg
t_stat, p_val = stats.ttest_ind(bandgap_1['Volume'] * bandgap_1['density_Avg'],
                                bandgap_0['Volume'] * bandgap_0['
                                density_Avg'])
print(f'\nT-test for Volume * density_Avg: t_stat = {t_stat}, p_val = {p_val}')

# t-test for Volume * atom_rad_Avg
t_stat, p_val = stats.ttest_ind(bandgap_1['Volume'] * bandgap_1['atom_rad_Avg'],
                                bandgap_0['Volume'] * bandgap_0['
                                atom_rad_Avg'])
print(f'T-test for Volume * atom_rad_Avg: t_stat = {t_stat}, p_val = {p_val}')
#------------------
Mean and Std for Bandgap = 1
Volume          179.367448
density_Avg       3.609099
atom_rad_Avg    107.309824
dtype: float64
Volume          6.081955
density_Avg     0.429419
atom_rad_Avg    1.774551
dtype: float64

Mean and Std for Bandgap = 0
Volume          332.748750
density_Avg       5.088611
atom_rad_Avg    123.912639
dtype: float64
Volume          80.278173
density_Avg      0.911569
atom_rad_Avg     7.538284
dtype: float64

T-test for Volume * density_Avg: t_stat = -14.118456605134813, p_val = 7.
                                883531279971937e-29
T-test for Volume * atom_rad_Avg: t_stat = -15.629622987637985, p_val = 1.
                                1609327950005447e-32
```

III. CLASSIFICATION REPORT DETAIL

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall (also called Sensitivity) is the ratio of correctly predicted positive observations to all actual positives:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-Score is the harmonic mean of Precision and Recall:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Support is the number of actual occurrences of each class in the dataset.
Accuracy is the ratio of correctly predicted observations to the total observations:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Macro Average computes the metric independently for each class and then takes the average:

$$\text{Macro Average} = \frac{1}{N} \sum_{i=1}^{N} \text{Metric}_i$$

Weighted Average computes the metric for each class and weights it by the number of true instances (support) for that class:

$$\text{Weighted Average} = \frac{1}{\sum_{i=1}^{N} \text{Support}_i} \sum_{i=1}^{N} (\text{Support}_i \cdot \text{Metric}_i)$$

IV. ML OUTPUT

A. Classification

TABLE II:

Index	Material	Bandgap	Bandgap Actual Class	Bandgap Predicted Class
0	$ZnGeN_2$	2.44	0	1
1	$ZnGeN_2$	3.02	1	1
2	$ZnGeN_2$	2.73	0	1
3	$AlGaN_2$	4.59	1	1
4	$AlGaN_2$	4.63	1	1
5	$AlGaN_2$	4.46	1	1
6	$ZnSiN_2$	4.52	1	1
7	$AlGaAs_2$	1.76	0	0
8	$AlGaAs_2$	2.06	0	0
9	$AlGaAs_2$	2.13	0	0
10	$AlGaP_2$	2.47	0	0
11	$AlGaP_2$	2.71	0	0
12	$AlGaP_2$	2.76	0	0
13	$GaInN_2$	1.32	0	0
14	$GaInN_2$	1.50	0	0
15	$GaInN_2$	1.22	0	0
16	$AlInN_2$	2.64	0	0
17	$AlInN_2$	3.00	0	0
18	$AlInN_2$	2.70	0	0
19	$CdSiN_2$	2.45	0	1
20	$CdSiN_2$	3.05	1	1
21	$CdSiN_2$	3.16	1	1
22	$ZnGeP_2$	0.88	0	0
23	$ZnGeP_2$	1.73	0	0
24	$ZnGeP_2$	1.94	0	0
25	$CdGeN_2$	1.15	0	0
26	$CdGeN_2$	2.04	0	0
27	$CdGeN_2$	1.73	0	0
28	$GaInP_2$	1.87	0	0
29	$GaInP_2$	2.16	0	0
30	$GaInP_2$	2.06	0	0
31	$AlInP_2$	2.62	0	0
32	$AlInP_2$	2.98	0	0
33	$AlInP_2$	2.89	0	0
34	$AlInAs_2$	1.55	0	0
35	$AlInAs_2$	1.70	0	0
36	$AlInAs_2$	1.67	0	0
37	$GaInAs_2$	0.27	0	0
38	$GaInAs_2$	0.43	0	0
39	$GaInAs_2$	0.38	0	0
40	$CdSiAs_2$	0.15	0	0
41	$CdSiAs_2$	1.47	0	0
42	$CdSiAs_2$	1.34	0	0
43	$CdGeAs_2$	0.11	0	0
44	$ZnGeAs_2$	0.58	0	0
45	$ZnGeAs_2$	0.50	0	0

Continued on next page

Index	Material	Bandgap	Bandgap Actual Class	Bandgap Predicted Class
46	$ZnSnN_2$	0.94	0	0
47	$ZnSnN_2$	0.97	0	0
48	$ZnSnN_2$	0.79	0	0
49	$ZnSiAs_2$	0.78	0	0
50	$ZnSiAs_2$	1.81	0	0
51	$ZnSiAs_2$	1.95	0	0
52	$CdSiP_2$	1.21	0	0
53	$CdSiP_2$	2.70	0	0
54	$CdSiP_2$	2.33	0	0
55	$CdGeP_2$	0.36	0	0
56	$CdGeP_2$	1.78	0	0
57	$CdGeP_2$	1.62	0	0
58	$CdSnN_2$	0.10	0	0
59	$CdSnN_2$	0.45	0	0
60	$CdSnN_2$	0.15	0	0
61	$ZnSnAs_2$	0.21	0	0
62	$ZnSnAs_2$	0.27	0	0
63	$ZnSnAs_2$	0.24	0	0
64	$ZnSnP_2$	1.39	0	0
65	$ZnSnP_2$	1.67	0	0
66	$ZnSnP_2$	1.69	0	0
67	$CdSnP_2$	0.75	0	0
68	$CdSnP_2$	1.29	0	0
69	$CdSnP_2$	1.14	0	0
70	$ZnSiP_2$	1.53	0	0
71	$ZnSiP_2$	2.35	0	0
72	$ZnSiP_2$	2.20	0	0
73	$CdSnAs_2$	0.13	0	0
74	$CdSiSb_2$	0.88	0	0
75	$CdSiSb_2$	0.75	0	0
76	$ZnSiSb_2$	0.78	0	0
77	$ZnSiSb_2$	1.15	0	0
78	$CdGeSb_2$	0.19	0	0

B. Regression

TABLE III: Bandgap Results

Index	Material	Bandgap Actual (eV)	Bandgap LASSO (eV)	Bandgap LASSO (eV) with Phase
0	$ZnGeN_2$	2.44	2.61	2.34
1	$ZnGeN_2$	3.02	2.61	2.40
2	$ZnGeN_2$	2.73	2.61	2.40
3	$AlGaN_2$	4.59	4.52	4.46
4	$AlGaN_2$	4.63	4.52	4.53
5	$AlGaN_2$	4.46	4.52	4.52
6	$ZnSiN_2$	4.52	4.04	3.89

Continued on next page

Index	Material	Bandgap Actual (eV)	Bandgap LASSO (eV)	Bandgap LASSO (eV) with Phase
7	$AlGaAs_2$	1.76	1.74	1.79
8	$AlGaAs_2$	2.06	1.74	1.84
9	$AlGaAs_2$	2.13	1.74	1.84
10	$AlGaP_2$	2.47	2.67	2.96
11	$AlGaP_2$	2.71	2.67	3.08
12	$AlGaP_2$	2.76	2.67	3.08
13	$GaInN_2$	1.32	1.41	1.41
14	$GaInN_2$	1.50	1.41	1.51
15	$GaInN_2$	1.22	1.41	1.51
16	$AlInN_2$	2.64	2.68	2.69
17	$AlInN_2$	3.00	2.69	2.77
18	$AlInN_2$	2.70	2.69	2.77
19	$CdSiN_2$	2.45	2.87	2.96
20	$CdSiN_2$	3.05	2.87	3.02
21	$CdSiN_2$	3.16	2.90	3.03
22	$ZnGeP_2$	0.88	1.50	0.80
23	$ZnGeP_2$	1.73	1.50	1.83
24	$ZnGeP_2$	1.94	1.50	1.83
25	$CdGeN_2$	1.15	1.36	1.41
26	$CdGeN_2$	2.04	1.36	1.48
27	$CdGeN_2$	1.73	1.37	1.49
28	$GaInP_2$	1.87	1.96	1.76
29	$GaInP_2$	2.16	1.96	2.08
30	$GaInP_2$	2.06	1.96	2.08
31	$AlInP_2$	2.62	2.61	2.58
32	$AlInP_2$	2.98	2.61	2.79
33	$AlInP_2$	2.89	2.61	2.79
34	$AlInAs_2$	1.55	1.42	1.61
35	$AlInAs_2$	1.70	1.42	1.70
36	$AlInAs_2$	1.67	1.42	1.70
37	$GaInAs_2$	0.27	0.34	0.41
38	$GaInAs_2$	0.43	0.34	0.43
39	$GaInAs_2$	0.38	0.34	0.43
40	$CdSiAs_2$	0.15	1.17	0.88
41	$CdSiAs_2$	1.47	1.17	1.29
42	$CdSiAs_2$	1.34	1.17	1.29
43	$CdGeAs_2$	0.11	0.27	0.31
44	$CdGeAs_2$	0.07	0.27	0.31
45	$ZnGeAs_2$	0.07	0.28	0.25
46	$ZnGeAs_2$	0.58	0.28	0.29
47	$ZnGeAs_2$	0.50	0.28	0.29
48	$ZnSnN_2$	0.94	0.87	0.99
49	$ZnSnN_2$	0.97	0.87	1.06
50	$ZnSnN_2$	0.79	0.86	1.06
51	$ZnSiAs_2$	0.78	1.27	1.01
52	$ZnSiAs_2$	1.81	1.27	1.36
53	$ZnSiAs_2$	1.95	1.27	1.36
54	$CdSiP_2$	1.21	1.97	1.07
55	$CdSiP_2$	2.70	1.97	2.22
56	$CdSiP_2$	2.33	1.98	2.23
57	$CdGeP_2$	0.36	1.40	0.54

Continued on next page

Index	Material	Bandgap Actual (eV)	Bandgap LASSO (eV)	Bandgap LASSO (eV) with Phase
58	$CdGeP_2$	1.78	1.40	1.75
59	$CdGeP_2$	1.62	1.40	1.75
60	$CdSnN_2$	0.10	0.20	0.14
61	$CdSnN_2$	0.45	0.20	0.22
62	$CdSnN_2$	0.15	0.20	0.22
63	$ZnSnAs_2$	0.21	0.08	0.07
64	$ZnSnAs_2$	0.27	0.08	0.15
65	$ZnSnAs_2$	0.24	0.08	0.15
66	$ZnSnP_2$	1.39	1.18	1.03
67	$ZnSnP_2$	1.67	1.18	1.53
68	$ZnSnP_2$	1.69	1.18	1.53
69	$CdSnP_2$	0.75	0.86	0.95
70	$CdSnP_2$	1.29	0.86	1.52
71	$CdSnP_2$	1.14	0.86	1.52
72	$ZnSiP_2$	1.53	2.09	1.52
73	$ZnSiP_2$	2.35	2.09	2.43
74	$ZnSiP_2$	2.20	2.09	2.43
75	$CdSnAs_2$	0.07	0.02	0.09
76	$CdSnAs_2$	0.05	0.02	0.21
77	$CdSnAs_2$	0.13	0.02	0.21
78	$ZnGeSb_2$	0.05	0.21	0.19
79	$ZnGeSb_2$	0.02	0.21	0.18
80	$CdSiSb_2$	0.88	0.88	1.10
81	$CdSiSb_2$	0.75	0.88	1.10
82	$ZnSiSb_2$	0.78	1.03	1.08
83	$ZnSiSb_2$	1.15	1.03	1.08
84	$CdGeSb_2$	0.09	0.19	0.23
85	$CdGeSb_2$	0.19	0.19	0.23
86	$ZnSnSb_2$	0.02	0.00	0.08
87	$ZnSnSb_2$	0.05	0.00	0.06
88	$ZnSnSb_2$	0.04	0.00	0.06
89	$CdSnSb_2$	0.06	0.07	0.11
90	$CdSnSb_2$	0.06	0.07	0.12
91	$CdSnSb_2$	0.09	0.06	0.12

C. Model Performance

(a) Hypothesis Testing

(i) Pearson Correlation Analysis

We use the Pearson correlation coefficient, denoted r, to find the linear relationship between each descriptor and the bandgap. It measures the strength and direction of the linear association between two continuous variables. A value of $r = +1$ indicates a perfect positive linear relationship. A value of $r = -1$ indicates a perfect negative linear relationship. A value of $r = 0$ indicates no linear association. The results are presented in Table (IV).

Descriptor	**Pearson** r	**p-value**	**Significant** ($\alpha = 0.05$)**?**
d_1	0.9366	9.04×10^{-43}	Yes
d_2	0.0432	6.82×10^{-1}	No
d_3	-0.4086	5.27×10^{-5}	Yes

TABLE IV. Pearson correlation (r) between each descriptor and bandgap.

Descriptor d_1 yielded $r = 0.9366$ with $p = 9.04 \times 10^{-43}$, indicating a very strong and highly significant positive linear relationship with bandgap. Descriptor d_2 yielded $r = 0.0432$ with $p = 0.682$, indicating a negligible and statistically insignificant correlation. Descriptor d_3 yielded $r = -0.4086$ with $p = 5.27 \times 10^{-5}$, indicating a moderate and statistically significant negative linear relationship.

(ii) Multiple Linear Regression

We fit a Multiple Linear Regression model using Ordinary Least Squares (OLS). The general form of the model is:

$$\hat{y} = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3$$

where $\hat{y}$ is the predicted bandgap, β_0 is the intercept (the predicted value of bandgap when all descriptors equal zero), and $\beta_1, \beta_2, \beta_3$ are the regression coefficients. Each coefficient shows the change in bandgap for a unit change in the corresponding standardized descriptor (d_1, d_2 and d_3), while keeping all other descriptors constant.

a. Coefficient of Determination (R^2): The R^2 measures the proportion of variance in bandgap explained by the three descriptors. It is defined as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where SS_{res} is the sum of squared residuals and SS_{tot} is the total sum of squares. An R^2 of 0.927 indicates that 92.7% of the variance in the bandgap is explained by the model, signifying an excellent fit.

b. Adjusted($\bar{R}^2$): The $\bar{R}^2$ penalises R^2 for additional predictors. It is defined as:

$$\bar{R}^2 = 1 - \frac{(1 - R^2)(n-1)}{n-k-1}$$

where n is the number of observations and k is the number of predictors. It is preferred over R^2 in multiple regression. This is because it does not inflate with the addition of irrelevant variables. The $\bar{R}^2$ = 0.925 confirms that the model explains 92.5% of the variance in bandgap. Even after penalizing for the number of predictors, this indicates that the added descriptors are genuinely useful. They are not inflating the fit artificially.

c. F-statistic and Overall Model Significance: The F-statistic tests the overall significance of the regression model. The null hypothesis is:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

The obtained value is F = 375.1, with an associated probability $\Pr(F) = 5.29 \times 10^{-50}$, which is far below $\alpha = 0.05$. This confirms the model as a whole is highly statistically significant.

d. Individual Coefficients: The individual t-statistics and their p-values test whether each coefficient is significantly different from zero:

$$H_0 : \hat{\beta}_j = 0$$

All three descriptors have $p \approx 0.000$, confirming each contributes significantly to the model after accounting for the others. The standardised coefficients are:

$$\hat{\beta}_0 = 1.4436, \quad \hat{\beta}_1 = 1.0818, \quad \hat{\beta}_2 = 0.2005, \quad \hat{\beta}_3 = -0.1343$$

Descriptor d_{001} has by far the largest coefficient, confirming it is the dominant predictor.

e. Confidence Intervals: The 95% confidence intervals indicate the range within which the true coefficient lies with 95% probability. None of these intervals include zero, further confirming significance. All the above details are mentioned in Table(V).

Term	$\hat{\beta}$	**Std. Err.**	t	$P > \|t\|$	**[0.025**	**0.975]**
const	1.4436	0.033	43.31	0.000	1.377	1.510
d_1	1.0818	0.036	30.36	0.000	1.011	1.153
d_2	0.2005	0.035	5.77	0.000	0.131	0.270
d_3	−0.1343	0.036	−3.76	0.000	−0.205	−0.063

TABLE V. OLS regression coefficients and diagnostics (standardised predictors).

f. Durbin–Watson Statistic: The Durbin–Watson statistic (DW = 2.425) tests for autocorrelation in the residuals. Values close to 2 indicate no autocorrelation. A value of 2.425 is acceptable and suggests residuals are largely independent of one another.

g. Residual Normality Tests: The Omnibus test and Jarque–Bera (JB) test assess the normality of residuals. Both return low p-values ($p_{\text{Omnibus}} = 0$ and $p_{\text{JB}} = 3.39 \times 10^{-8}$), indicating the residuals are not perfectly normally distributed.

h. Condition Number: The Condition Number ($\kappa = 1.48$) measures the numerical stability of the regression. Values below 30 indicate no harmful multicollinearity. A value of 1.48 confirms very low collinearity among the predictors.

i. Variance Inflation Factor (VIF): The VIF is computed to formally test for multicollinearity among the predictors, discussed in the Table(VI). For each predictor x_j, the VIF is defined as:

$$\text{VIF}_j = \frac{1}{1 - \text{R}_j^2}$$

where R_j^2 is the coefficient of determination obtained from regressing x_j on all other predictors. A value of VIF = 1 implies no correlation with other predictors. Values between 1 and 5 indicate low multicollinearity.

Feature	**VIF**	**Interpretation**
const	1.000	—
d_1	1.143	No multicollinearity
d_2	1.086	No multicollinearity
d_3	1.149	No multicollinearity

TABLE VI. Variance Inflation Factors for each predictor.

Values above 10 are generally considered problematic. All VIF values are close to 1. This confirms that the three descriptors are nearly orthogonal to each other. There is no multicollinearity concern in this model. The regression coefficient estimates are therefore stable and reliable. All the details of the code used are provided in the GitHub link.

(b) ML Model without Feature Engineering

If no feature engineering is applied and raw features are used directly in ML model. The ML model performance on the test data shows an R^2 of 0.75 and RMSE of 0.56 eV for LASSO, and an R^2 of 0.84 and RMSE of 0.43 eV for Random Forest. Even without feature engineering, the Random Forest model predicts the bandgap with reasonably good accuracy. A key limitation of black-box models, like Random Forest, is that they do not provide an interpretable analytical formula. This is the main reason for applying feature engineering with SISSO. It allows us to derive compact and physically meaningful descriptor-based expressions. The close agreement between training and test accuracy in the Random Forest model suggests good generalization. There are no signs of overfitting. This indicates that the selected raw features contain meaningful predictive information. The complete code is available on GitHub link.

(c) RepeatedKFold Cross-validation

We perform RepeatedK-fold[1] cross-validation, which combines Monte Carlo and K-fold techniques, on the LASSO model. The features used in the model were selected through SISSO feature engineering. During the training phase, we achieved an R^2 value of 0.92 and an RMSE of 0.31. For the test data, the model shows an R^2 of 0.91 and an RMSE of 0.31. These results demonstrate that the model generalizes well to unseen data. The close agreement between the training and test scores indicates that there is no sign of overfitting. The code is available at the GitHub link.

(d) LASSO Regression with Bootstrap Validation

Bootstrapping is a statistical method. It generates multiple new datasets. These datasets are created by randomly sampling with replacement from the original data. Each new dataset is called a bootstrap sample. It is the same size as the original dataset. Some data points are repeated, and some are missing. We train a model on each bootstrap sample. We evaluate the model on the data points not selected in the sampling. These points are called out-of-bag (OOB) data. Bootstrapping helps detect overfitting. If a model performs well on the OOB data, it suggests no overfitting.

We used Least Absolute Shrinkage and Selection Operator (LASSO) regression combined with bootstrap resampling to assess overfitting. To ensure unbiased model evaluation, a bootstrap validation scheme was adopted over $B = 1000$ iterations. In each iteration, a bootstrap sample of $n = 92$ observations was drawn with replacement. This resulted in approximately 59 unique samples per iteration. The remaining 33 observations, about 36.5% of the data, were OOB observations. The OOB samples, ranging from 25 to 43 per iteration, are used as an independent test set. The optimal regularization parameter was determined via 5-fold cross-validation using `LassoCV`, yielding $\alpha^* = 0.0011$. The results from the bootstrap analysis are presented in Fig.(1).

The LASSO bootstrap analysis produced a mean coefficient of determination of $R^2 = 0.92$. The 95% confidence interval (CI) for $R^2 = 0.92$ is [0.84, 0.96]. The mean root mean square error (RMSE) is 0.32 eV, with a 95% CI of [0.23, 0.40] eV. These results indicate strong and consistent predictive performance

[1] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RepeatedKFold.html

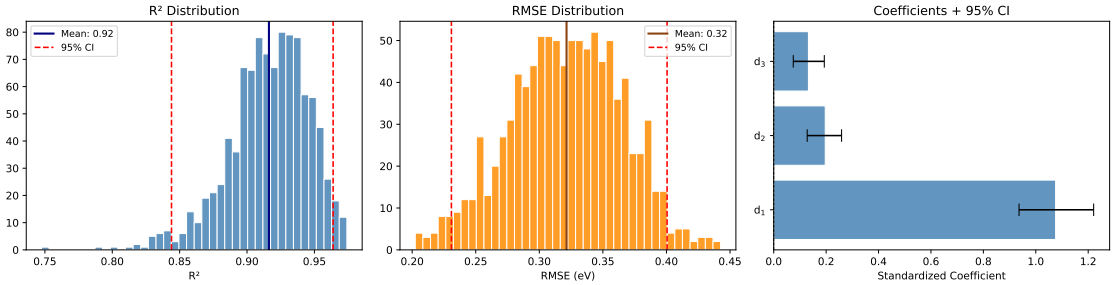

FIG. 1. LASSO bootstrap analysis results for bandgap prediction.

across all bootstrap resamples. The tightness of the confidence intervals reflects the model's stability. It confirms that the reported performance is not due to a favorable data partition. Analysis of the standardized regression coefficients revealed that d_1 exerted the dominant positive influence on band gap ($\beta_{d_1} = +1.07$, 95% CI: $[+0.94, +1.22]$), followed by d_2 ($\beta_{d_2} = +0.20$, 95% CI: $[+0.13, +0.26]$) and d_3 ($\beta_{d_3} = 0.13$, 95% CI: $[0.075, 0.19]$). Notably, none of the three features were zeroed out in any of the 1000 bootstrap iterations. This confirms that all three d-spacing descriptors contribute statistically meaningfully and consistently to band gap prediction. The full code is available at the GitHub link.

V. SISSO MODEL ACROSS MULTIPLE FOLDS

Fold	Features and Constants	RMSE	R^2
1	$d_1 = \left(\frac{AR_{min}}{MD_{avg}+MD_{min}}\right) \times \left(\frac{1}{\sqrt{\frac{Volume}{vdW\ Rad_{min}}}}\right)$ $d_2 = \frac{\left(\frac{AR_{avg}}{vdW\ Rad_{avg}} - \frac{AR_{min}}{CR_{avg}}\right)}{\log(MD_{min})}$ $d_3 = \left\lvert \frac{(AR_{min} - vdW\ Rad_{avg})}{(AR_{max} - vdW\ Rad_{max})} - \frac{vdW\ Rad_{avg}}{CR_{min}} \right\rvert$ Bandgap SISSO = $0.42d_1+2.15d_2+0.66d_3$ - 2.32	Train: 0.25 Test: 0.11	Train: 0.95 Test: 0.98
2	$d_1 = \frac{\left(\frac{AR_{min}}{Volume}\right) \cdot (vdW\ Rad_{min} + AR_{min})}{MD_{avg}+MD_{max}}$ $d_2 = \left\lvert \exp\left(\frac{vdW\ Rad_{avg}}{Volume}\right) - \frac{(CR_{min}+AR_{min})}{AR_{min}} \right\rvert$ $d_3 = \frac{(AR_{max} - vdW\ Rad_{min})}{\left(\frac{AR_{min}}{AR_{max}} - \frac{CR_{avg}}{vdW\ Rad_{avg}}\right)}$ Bandgap SISSO $= 0.29d_1 + 1.91d_2 + 0.32 \times 10^{-3}d_3 - 2.45$	Train: 0.32 Test: 0.17	Train: 0.91 Test: 0.98
3	$d_1 = \frac{\frac{AR_{min}}{MD_{avg}+MD_{min}}}{\sqrt{\frac{Volume}{vdW\ Rad_{min}}}}$ $d_2 = \frac{\left(\frac{AR_{avg}}{CR_{min}} - \frac{vdW\ Rad_{avg}}{AR_{avg}}\right)}{\log(MD_{min})}$ $d_3 = \left\lvert \left(\frac{AR_{min}}{CR_{avg}}\right) CR_{max} - (\lvert CR_{min} - AR_{min} \rvert + CR_{min}) \right\rvert$ Bandgap SISSO $= 0.39d_1 + 1.82d_2 - 0.031d_3 - 1.76$	Train: 0.31 Test: 0.25	Train: 0.92 Test: 0.91
4	$d_1 = \left\lvert \frac{CR_{min}}{AR_{max}} - \frac{MD_{avg}}{MD_{max}} \right\rvert + \log(MD_{avg})$ $d_2 = \left\lvert \frac{vdW\ Rad_{max}}{AR_{max}} - \frac{AR_{min}}{AR_{max}} \right\rvert - \frac{AR_{min}}{AR_{max}}$ $d_3 = \left\lvert \frac{(CR_{avg})^2}{vdW\ Rad_{avg} - CR_{avg}} - (vdW\ Rad_{min} + CR_{min}) \right\rvert$ Bandgap SISSO $= -4.58d_1 - 1.17d_2 + 0.68 \times 10^{-2}d_3 + 9.70$	Train: 0.33 Test: 0.39	Train: 0.91 Test: 0.92
5	$d_1 = \frac{\left(\frac{AR_{min}}{Volume}\right) \cdot (vdW\ Rad_{min} + AR_{min})}{MD_{avg}+MD_{min}}$ $d_2 = \frac{\left\lvert \frac{vdW\ Rad_{avg}}{AR_{avg}} - \frac{AR_{avg}}{AR_{min}} \right\rvert}{AR_{min} \cdot MD_{avg}}$ $d_3 = \frac{\lvert vdW\ Rad_{avg} - AR_{max} \rvert}{(AR_{max} - CR_{min}) - (CR_{max} - vdW\ Rad_{avg})}$ Bandgap SISSO $= 0.32d_1 - 0.93 \times 10^{3}d_2 + 0.39 \times 10^{-1}d_3 - 1.51$	Train: 0.30 Test: 0.39	Train: 0.93 Test: 0.76
6	$d_{001} = \frac{(AR_{avg} - AR_{min} - CR_{max})}{\frac{Volume \cdot MD_{min}}{vdW\ Rad_{min}}}$ $d_{002} = \left\lvert \frac{CR_{min}}{CR_{max}} - \frac{MD_{avg}}{MD_{max}} \right\rvert + \log(MD_{avg})$ $d_{003} = \left\lvert (AR_{avg} - vdW\ Rad_{min}) \left(\frac{MD_{avg}}{MD_{max}}\right) - (AR_{max} - vdW\ Rad_{min}) \right\rvert$ Bandgap SISSO $= -0.07d_1 - 2.90d_2 - 0.034d_3 + 5.76$	Train: 0.33 Test: 0.25	Train: 0.90 Test: 0.97
7	$d_1 = \frac{\left(\frac{AR_{min}}{Volume}\right) \cdot (vdW\ Rad_{min} + AR_{min})}{MD_{avg}+MD_{min}}$ $d_2 = \left\lvert \frac{vdW\ Rad_{avg}+CR_{min}}{AR_{min}} - \left(\frac{AR_{avg}}{AR_{min}}\right)^3 \right\rvert$ $d_3 = \frac{(AR_{avg} - vdW\ Rad_{max})}{\left(\frac{CR_{min}}{MD_{avg}}\right) - \left(\frac{vdW\ Rad_{max}}{MD_{max}}\right)}$ Bandgap SISSO $= 0.29d_1 - 0.59d_2 - 0.48 \times 10^{-2}d_3 - 0.90$	Train: 0.30 Test: 0.54	Train: 0.93 Test: 0.73
8	$d_1 = \frac{\left(\frac{AR_{min}}{Volume}\right) \cdot (vdW\ Rad_{min} + AR_{min})}{MD_{avg}+MD_{min}}$ $d_2 = \left\lvert \frac{vdW\ Rad_{avg}+CR_{max}}{CR_{min}+AR_{min}} - \frac{AR_{avg}}{AR_{min}} \right\rvert$ $d_3 = \frac{\lvert CR_{max} - AR_{min} \rvert}{\lvert (AR_{avg} - vdW\ Rad_{min}) - (AR_{min} - CR_{max}) \rvert}$ Bandgap SISSO $= 0.30d_1 - 4.13d_2 - 0.037d_3 - 1.02$	Train: 0.29 Test: 0.33	Train: 0.93 Test: 0.82
9	$d_1 = \exp\left(\frac{vdW\ Rad_{min}}{Volume}\right) \frac{CR_{min}}{MD_{avg}+MD_{min}}$ $d_2 = \left\lvert \left(\frac{AR_{avg}}{CR_{min}}\right)^3 - \frac{MD_{avg}+MD_{max}}{MD_{avg}} \right\rvert$ $d_3 = \frac{(AR_{avg} - CR_{max})}{\left(\frac{CR_{min}}{vdW\ Rad_{avg}} - \frac{MD_{avg}}{MD_{min}}\right)}$ Bandgap SISSO $= 0.15d_1 - 0.56d_2 - 0.67 \times 10^{-3}d_3 - 1.45$	Train: 0.30 Test: 0.39	Train: 0.92 Test: 0.83
10	$d_1 = \exp\left(\frac{VDW\ Rad_{Min}}{Volume}\right) \times \frac{AR_{Max}}{MD_{Avg}+MD_{Min}}$ $d_2 = \frac{AR_{Avg}^2 - AR_{Min} \times VDW\ Rad_{Avg}}{\frac{\log(MD_{Min})}{\exp(-MD_{Avg})}}$ $d_3 = -\frac{}{(AR_{Avg} - CR_{Max}) + (AR_{Min} - CR_{Min})}$ Bandgap SISSO $= 0.16d_1 + 6.0 \times 10^{-5}d_2 + 26.82d_3 - 1.92$	Train: 0.31 Test: 0.30	Train: 0.92 Test: 0.85
		Avg. Train: 0.30 **Avg. Test: 0.31**	**Avg. Train: 0.92** **Avg. Test: 0.88**

TABLE VII. Descriptors and final predictions for the SISSO model across multiple folds, along with their statistical measures.