Supplementary Material: "Towards machine learning for microscopic mechanisms: a formula search for crystal structure stability based on atomic properties"

Udaykumar Gajera[1,2], **Loriano Storchi**[3], **Danila Amoroso**[1,4], **Francesco Delodovici**[1], **Silvia Picozzi**[1]

1. Consiglio Nazionale delle Ricerche, CNR-SPIN c/o Università "G. D'Annunzio", 66100 Chieti, Italy

2. Chemistry Department and NIS, University of Turin, via Pietro Giuria, 7, 10125, Torino, Italy

3. Dipartimento di Farmacia, Universitá degli Studi "G. D'Annunzio", 66100 Chieti, Italy

4. NanoMat/Q-mat/CESAM,Universite de Liege, B-4000 Liege, Belgium

I. VALIDATION OF THE LINEAR REGRESSION

We employed different verification parameters to check the model's efficiency, namely: RMSE, Pearson correlation coefficient, R^2 values, and classification accuracy. RMSE represents the root mean square error of the test dataset. The Pearson correlation coefficient, defined in equation 1, represents a measurement of input and output property dependence[1]. If two properties are highly dependent one on the other, one gets values closer to 1 or -1, whereas values closer to zero show a much lower dependence. R^2, *i.e.* the coefficient of determination defined in equation 2, describes how properly the regression line interpolates the data. Finally, the classification accuracy shows the ability of the linear model to qualitatively distinguish different classes of the dataset, in our case RS and ZB as stable phases.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{1}$$

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} \tag{2}$$

In equation 2: $SS_{total} = \sum_i (y_i - \bar{y})^2$, $SS_{residual} = \sum_i (y_i - f_i)^2$ where f_i are the predicted values; x_i and y_i are values of input and output properties; $\bar{x}$ and $\bar{y}$ are mean of the input and output values.

II. ANALYSIS OF THE BEST 1D, 2D AND 3D DESCRIPTORS

Table-S.1 reports other verification parameters calculated for the best descriptors of each generator, including those presented in the paper. The relevance of calculating avg(RMSE train) and avg(RMSE test) lies in analysing the bias-variance tradeoff[2] in LR. Max_E and Min_E indicate the maximum and minimum error in the prediction for the specific descriptor.

Figure S.1 reports in panel **a**(**b**) the *ab-initio* energies against the energies predicted through the 1D descriptor proposed in Ref.[3] (obtained within GEN3). In addition, panels **(c)** and **(d)** of figure S.1 report the absolute errors obtained employing the 1D descriptors mentioned above, for certain compounds. From a comparison of the two scatter plots and of the two bar-graphs, one can infer the improved accuracy obtained with the GEN3 descriptor.

A. Dependence of energy difference on the atomic features

Figure S.2 reports the dependence of the total energy difference between RS and ZB phases as a function of the atomic features, except for r_p (since that is reported in the main text). It appears clearly that only r_s is strongly correlated with the energy difference, at variance with other atomic features where the correlation is small or absent.

B. Formula optimization using automated optimization methods

To find the relative contribution of individual atomic features in the descriptor, we employed a grid search method. We further bench-marked our grid-search optimization for the best 1D descriptor constructed using GEN3, defined in equation 3, by means of other automated optimization methods: Nelder-Mead[4], Conjugate Gradient(CG)[5], Broyden–Fletcher–Goldfarb–Shanno(BFGS)[6] and truncated Newton(TNC)[7]:

$$\frac{r_p(B) + \sqrt{|r_d(A)|}}{r_p(A)^3 + r_p(B)^3} \tag{3}$$

if we introduce different coefficients (a,b,c,d) multiplying each atomic feature, then the expression can be written as:

$$\frac{a \cdot r_p(B) + b \cdot \sqrt{|r_d(A)|}}{c \cdot r_p(A)^3 + d \cdot r_p(B)^3} \tag{4}$$

Through the automated optimization methods mentioned above, one can obtain the set of coefficients that give the lowest RMSE. Each step of the optimization is followed by LR. Thus, the optimization modifies the slope (m) and intercept moving towards the coefficients combination with the lowest possible RMSE. From 80 to 100 iterations are needed for each method to reach the minimum RMSE, as reported in panel **a** of figure S.3. Panels from **b** to **e** report the trend of the ratios $a/b, c/d, m \times a/c$ and $m \times b/d$. All these quantities appear to converge to constant values in the final steps of the optimization.

	Details	avg(RSME train)	avg(RMSE test)	RMSE	R^2	Pearson_coeff	success rate	Max_E	Min_E	Std_daviation
0	Ref_1D[3]	0.1420	0.1455	0.1422	0.89	0.947	89%	0.0523	0.0041	0.0081
1	1D_GEN1	0.1186	0.1296	0.1192	0.92	0.963	90%	0.0743	0.0014	0.0083
2	1D_GEN2	0.1305	0.1367	0.1309	0.91	0.956	91%	0.0676	0.0027	0.0105
3	1D_GEN3	0.0961	0.0995	0.0963	0.95	0.976	94%	0.0234	0.0016	0.0032
4	1D_GEN4	0.1055	0.1103	0.1058	0.94	0.971	96%	0.0330	0.0012	0.0044
5	Ref_2d[3]	0.0983	0.1041	0.0987	0.95	0.975	96%	0.0323	0.0019	0.0044
6	2D_GEN1	0.0941	0.0988	0.0943	0.95	0.977	89%	0.0419	0.0020	0.0040
7	2D_GEN2	0.1095	0.1163	0.1099	0.93	0.969	87%	0.0489	0.0011	0.0083
8	2D_GEN3	0.0875	0.0911	0.0878	0.96	0.980	88%	0.0178	0.0014	0.0026
9	2D_GEN4	0.0951	0.0995	0.0954	0.95	0.977	93%	0.0221	0.0016	0.0033
10	Ref_3d[3]	0.0751	0.0814	0.0755	0.97	0.985	93%	0.0185	0.0009	0.0031
11	3D_GEN1	0.0929	0.1003	0.0933	0.95	0.978	90%	0.0282	0.0024	0.0038
12	3D_GEN2	0.1200	0.1300	0.1205	0.92	0.963	91%	0.1119	0.0021	0.0103
13	3D_GEN3	0.0832	0.0874	0.0834	0.96	0.982	98%	0.0199	0.0015	0.0026
14	3D_GEN4	0.0915	0.0989	0.0919	0.96	0.978	93%	0.0227	0.0013	0.0035

TABLE S.1. Different verification parameters for 1D, 2D and 3D descriptors calculated in the present work and descriptors presented in Ref.[3]. Here, avg(RMSE train), avg(RMSE test) and RMSE indicate the root mean squared error for training data, test data and full dataset, respectively. R^2 and Pearson coeff are goodness parameters. Max_E and Min_E show the maximum and minimum absolute error in prediction.

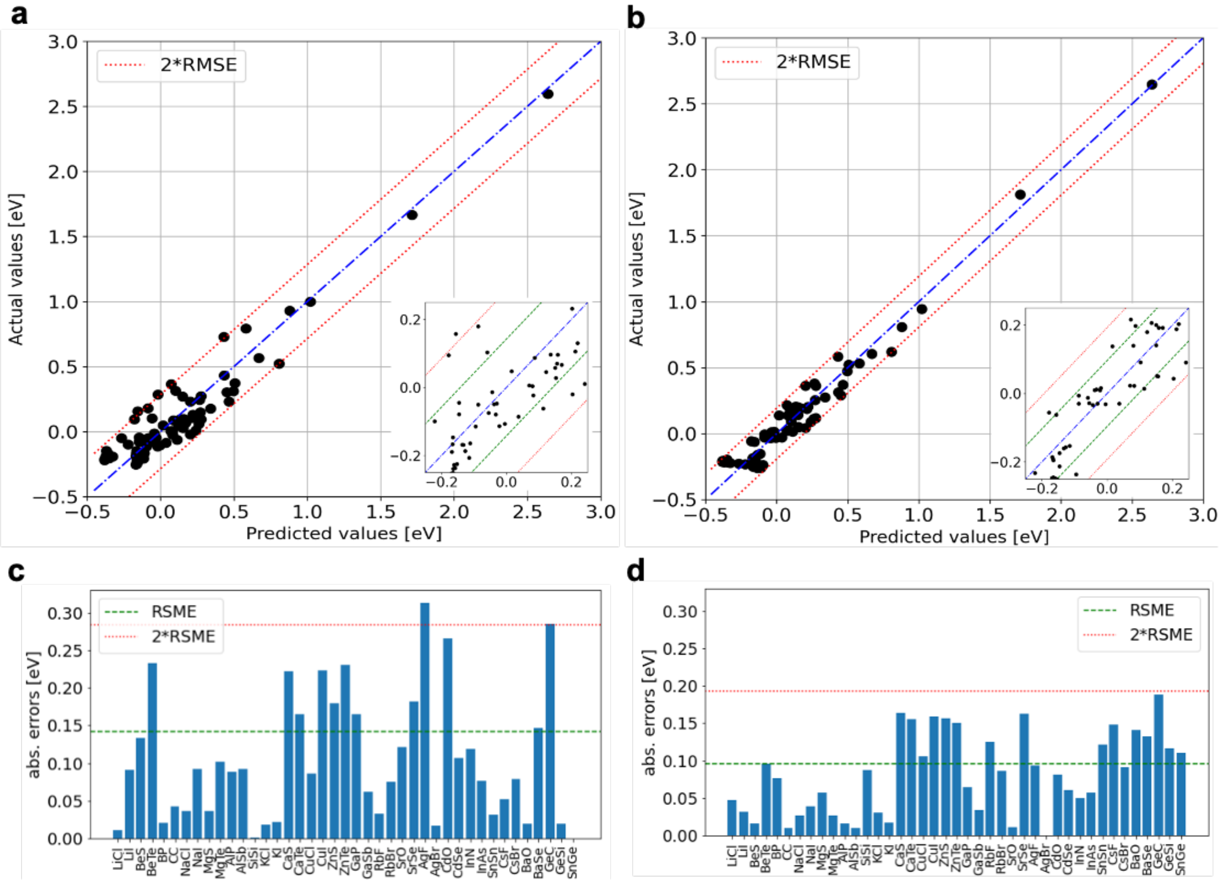

FIG. S.1. Panel **a** reports the predicted against actual (*i.e* DFT) ΔE values for 1D descriptor presented in Ref.[3]; panel **b** reports those obtained by *GEN*3 as a comparison. Absolute error for the same formula for different compounds in the bar graph (panels **c** and **d**). The related descriptors used to calculate the values can be inferred from the main text.

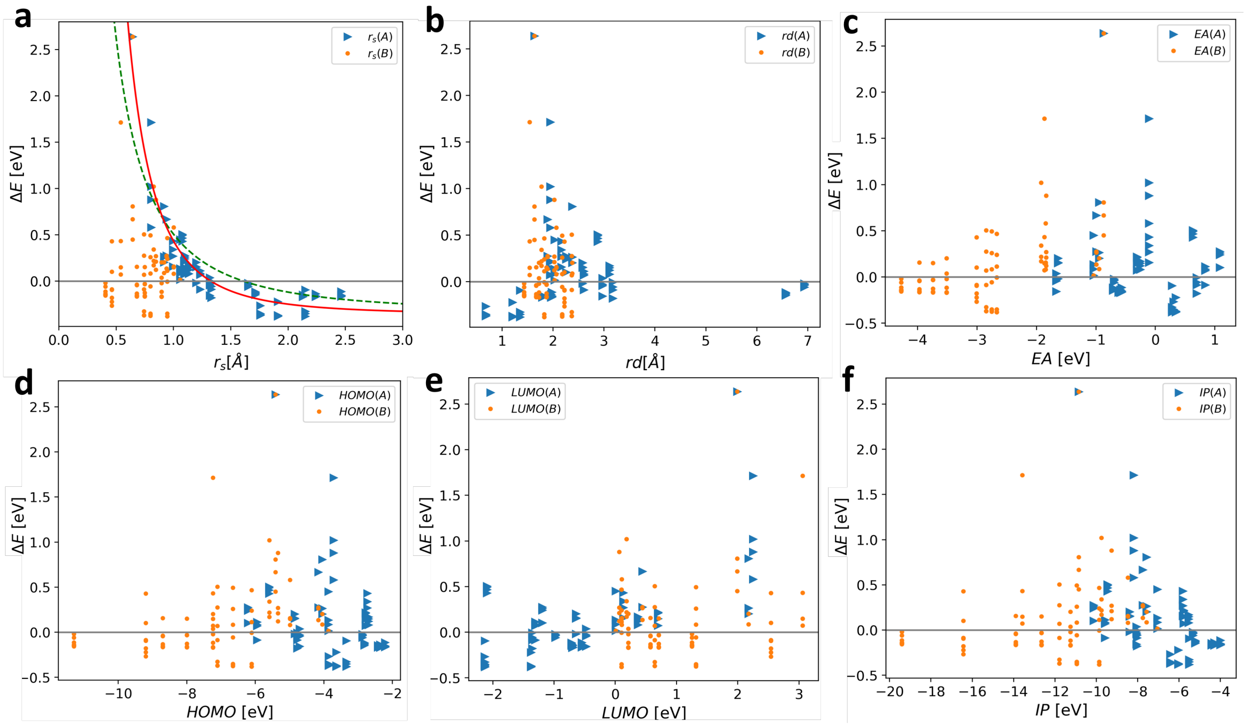

FIG. S.2. Dependence of ΔE on atomic features: a) r_s, b) r_d, c) EA, d) $LUMO$, e) $HOMO$ and f) IP. Orange dots (blue triangles) indicate values relative to the A (B) atoms. In panel-**a**, we perform a fit using a function $f(x)$ proportional to x^{-2} (dotted green line) and to x^{-3} (straight red line).

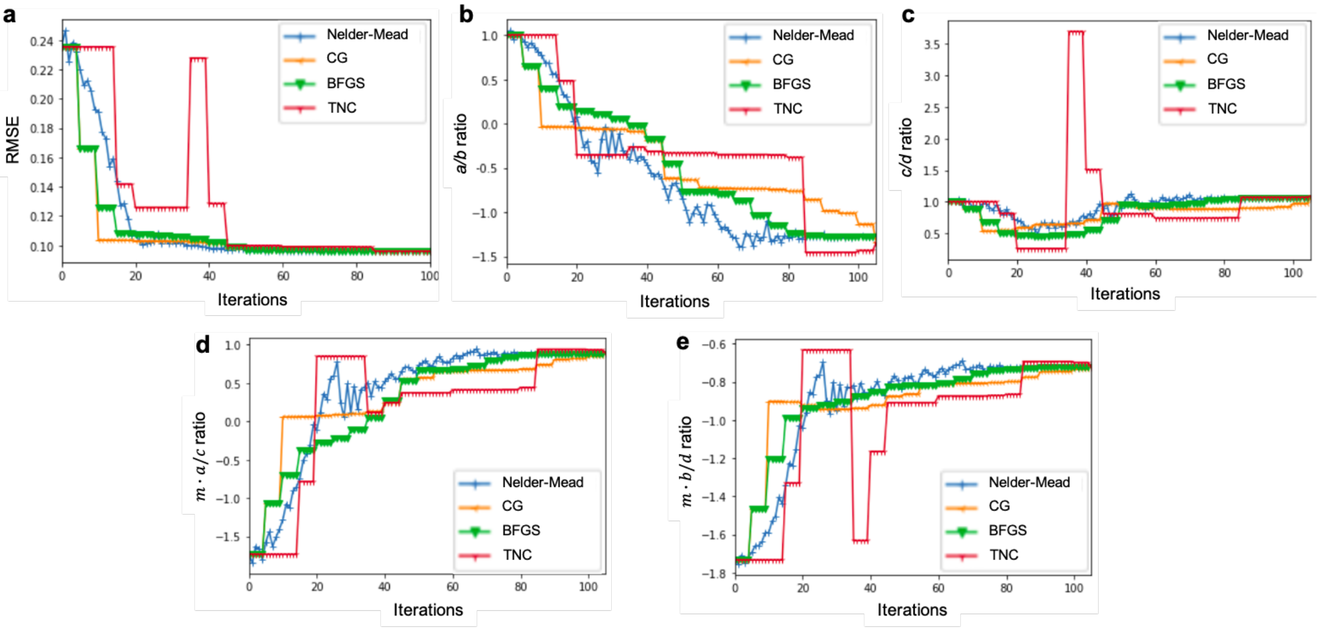

FIG. S.3. Evolution of different parameters at each iteration using different automated optimizing algorithms: Nelder-Mead (Blue lines), CG (orange lines), BFGS (green lines) and TNC (red lines). Here, we show the evolution of RSME, $a/b, c/d, m \times a/c$ and $m \times b/d$ in panels **a, b, c, d** and **e** respectively.

A	B	DFT Classification	ΔE	IP(A)	EA(A)	HOMO(A)	LUMO(A)	r_s(A)	r_p(A)	r_d(A)	IP(B)	EB(B)	HOMO(B)	LUMO(B)	r_s(B)	r_p(B)	r_d(B)
Li	F	RS	-0.059	-5.329	-0.698	-2.874	-0.978	1.652	1.995	6.930	-19.404	-4.273	-11.294	1.251	0.406	0.371	1.428
Li	Cl	RS	-0.038	-5.329	-0.698	-2.874	-0.978	1.652	1.995	6.930	-13.902	-3.971	-8.700	0.574	0.679	0.756	1.666
Li	Br	RS	-0.033	-5.329	-0.698	-2.874	-0.978	1.652	1.995	6.930	-12.650	-3.739	-8.001	0.708	0.749	0.882	1.869
Li	I	RS	-0.022	-5.329	-0.698	-2.874	-0.978	1.652	1.995	6.930	-11.257	-3.513	-7.236	0.213	0.896	1.071	1.722
Be	O	ZB	0.430	-9.459	0.631	-5.600	-2.098	1.078	1.211	2.877	-16.433	-3.006	-9.197	2.541	0.462	0.427	2.219
Be	S	ZB	0.506	-9.459	0.631	-5.600	-2.098	1.078	1.211	2.877	-11.795	-2.845	-7.106	0.642	0.742	0.847	2.366
Be	Se	ZB	0.495	-9.459	0.631	-5.600	-2.098	1.078	1.211	2.877	-10.946	-2.751	-6.654	1.316	0.798	0.952	2.177
Be	Te	ZB	0.466	-9.459	0.631	-5.600	-2.098	1.078	1.211	2.877	-9.867	-2.666	-6.109	0.099	0.945	1.141	1.827
B	N	ZB	1.713	-8.190	-0.107	-3.715	2.248	0.805	0.826	1.946	-13.585	-1.867	-7.239	3.057	0.539	0.511	1.540
B	P	ZB	1.020	-8.190	-0.107	-3.715	2.248	0.805	0.826	1.946	-9.751	-1.920	-5.596	0.183	0.826	0.966	1.771
B	As	ZB	0.879	-8.190	-0.107	-3.715	2.248	0.805	0.826	1.946	-9.262	-1.839	-5.341	0.064	0.847	1.043	2.023
C	C	ZB	2.638	-10.852	-0.872	-5.416	1.992	0.644	0.630	1.631	-10.852	-0.872	-5.416	1.992	0.644	0.630	1.631
Na	F	RS	-0.146	-5.223	-0.716	-2.819	-0.718	1.715	2.597	6.566	-19.404	-4.273	-11.294	1.251	0.406	0.371	1.428
Na	Cl	RS	-0.133	-5.223	-0.716	-2.819	-0.718	1.715	2.597	6.566	-13.902	-3.971	-8.700	0.574	0.679	0.756	1.666
Na	Br	RS	-0.127	-5.223	-0.716	-2.819	-0.718	1.715	2.597	6.566	-12.650	-3.739	-8.001	0.708	0.749	0.882	1.869
Na	I	RS	-0.115	-5.223	-0.716	-2.819	-0.718	1.715	2.597	6.566	-11.257	-3.513	-7.236	0.213	0.896	1.071	1.722
Mg	O	RS	-0.178	-8.037	0.693	-4.782	-1.358	1.330	1.897	3.171	-16.433	-3.006	-9.197	2.541	0.462	0.427	2.219
Mg	S	RS	-0.087	-8.037	0.693	-4.782	-1.358	1.330	1.897	3.171	-11.795	-2.845	-7.106	0.642	0.742	0.847	2.366
Mg	Se	RS	-0.055	-8.037	0.693	-4.782	-1.358	1.330	1.897	3.171	-10.946	-2.751	-6.654	1.316	0.798	0.952	2.177
Mg	Te	RS	-0.005	-8.037	0.693	-4.782	-1.358	1.330	1.897	3.171	-9.867	-2.666	-6.109	0.099	0.945	1.141	1.827
Al	N	ZB	0.072	-5.780	-0.313	-2.784	0.695	1.092	1.393	1.939	-13.585	-1.867	-7.239	3.057	0.539	0.511	1.540
Al	P	ZB	0.219	-5.780	-0.313	-2.784	0.695	1.092	1.393	1.939	-9.751	-1.920	-5.596	0.183	0.826	0.966	1.771
Al	As	ZB	0.212	-5.780	-0.313	-2.784	0.695	1.092	1.393	1.939	-9.262	-1.839	-5.341	0.064	0.847	1.043	2.023
Al	Sb	ZB	0.150	-5.780	-0.313	-2.784	0.695	1.092	1.393	1.939	-8.468	-1.847	-4.991	0.105	1.001	1.232	2.065
Si	C	ZB	0.668	-7.758	-0.993	-4.163	0.440	0.938	1.134	1.890	-10.852	-0.872	-5.416	1.992	0.644	0.630	1.631
Si	Si	ZB	0.275	-7.758	-0.993	-4.163	0.440	0.938	1.134	1.890	-7.758	-0.993	-4.163	0.440	0.938	1.134	1.890
K	F	RS	-0.146	-4.433	-0.621	-2.426	-0.697	2.128	2.443	1.785	-19.404	-4.273	-11.294	1.251	0.406	0.371	1.428
K	Cl	RS	-0.165	-4.433	-0.621	-2.426	-0.697	2.128	2.443	1.785	-13.902	-3.971	-8.700	0.574	0.679	0.756	1.666
K	Br	RS	-0.166	-4.433	-0.621	-2.426	-0.697	2.128	2.443	1.785	-12.650	-3.739	-8.001	0.708	0.749	0.882	1.869
K	I	RS	-0.168	-4.433	-0.621	-2.426	-0.697	2.128	2.443	1.785	-11.257	-3.513	-7.236	0.213	0.896	1.071	1.722
Ca	O	RS	-0.266	-6.428	0.304	-3.864	-2.133	1.757	2.324	0.679	-16.433	-3.006	-9.197	2.541	0.462	0.427	2.219
Ca	S	RS	-0.369	-6.428	0.304	-3.864	-2.133	1.757	2.324	0.679	-11.795	-2.845	-7.106	0.642	0.742	0.847	2.366
Ca	Se	RS	-0.361	-6.428	0.304	-3.864	-2.133	1.757	2.324	0.679	-10.946	-2.751	-6.654	1.316	0.798	0.952	2.177
Ca	Te	RS	-0.350	-6.428	0.304	-3.864	-2.133	1.757	2.324	0.679	-9.867	-2.666	-6.109	0.099	0.945	1.141	1.827
Cu	F	RS	-0.019	-8.389	-1.638	-4.856	-0.641	1.197	1.680	2.576	-19.404	-4.273	-11.294	1.251	0.406	0.371	1.428
Cu	Cl	ZB	0.156	-8.389	-1.638	-4.856	-0.641	1.197	1.680	2.576	-13.902	-3.971	-8.700	0.574	0.679	0.756	1.666
Cu	Br	ZB	0.152	-8.389	-1.638	-4.856	-0.641	1.197	1.680	2.576	-12.650	-3.739	-8.001	0.708	0.749	0.882	1.869
Cu	I	ZB	0.203	-8.389	-1.638	-4.856	-0.641	1.197	1.680	2.576	-11.257	-3.513	-7.236	0.213	0.896	1.071	1.722
Zn	O	ZB	0.102	-10.136	1.081	-6.217	-1.194	1.099	1.547	2.254	-16.433	-3.006	-9.197	2.541	0.462	0.427	2.219
Zn	S	ZB	0.275	-10.136	1.081	-6.217	-1.194	1.099	1.547	2.254	-11.795	-2.845	-7.106	0.642	0.742	0.847	2.366
Zn	Se	ZB	0.259	-10.136	1.081	-6.217	-1.194	1.099	1.547	2.254	-10.946	-2.751	-6.654	1.316	0.798	0.952	2.177
Zn	Te	ZB	0.241	-10.136	1.081	-6.217	-1.194	1.099	1.547	2.254	-9.867	-2.666	-6.109	0.099	0.945	1.141	1.827
Ga	N	ZB	0.433	-5.818	-0.108	-2.732	0.130	0.994	1.330	2.163	-13.585	-1.867	-7.239	3.057	0.539	0.511	1.540
Ga	P	ZB	0.341	-5.818	-0.108	-2.732	0.130	0.994	1.330	2.163	-9.751	-1.920	-5.596	0.183	0.826	0.966	1.771
Ga	As	ZB	0.271	-5.818	-0.108	-2.732	0.130	0.994	1.330	2.163	-9.262	-1.839	-5.341	0.064	0.847	1.043	2.023
Ga	Sb	ZB	0.158	-5.818	-0.108	-2.732	0.130	0.994	1.330	2.163	-8.468	-1.847	-4.991	0.105	1.001	1.232	2.065
Ge	Ge	ZB	0.202	-7.567	-0.949	-4.046	2.175	0.917	1.162	2.373	-7.567	-0.949	-4.046	2.175	0.917	1.162	2.373
Rb	F	RS	-0.136	-4.289	-0.590	-2.360	-0.705	2.240	3.199	1.960	-19.404	-4.273	-11.294	1.251	0.406	0.371	1.428
Rb	Cl	RS	-0.161	-4.289	-0.590	-2.360	-0.705	2.240	3.199	1.960	-13.902	-3.971	-8.700	0.574	0.679	0.756	1.666
Rb	Br	RS	-0.164	-4.289	-0.590	-2.360	-0.705	2.240	3.199	1.960	-12.650	-3.739	-8.001	0.708	0.749	0.882	1.869
Rb	I	RS	-0.169	-4.289	-0.590	-2.360	-0.705	2.240	3.199	1.960	-11.257	-3.513	-7.236	0.213	0.896	1.071	1.722
Sr	O	RS	-0.221	-6.032	0.343	-3.641	-1.379	1.911	2.548	1.204	-16.433	-3.006	-9.197	2.541	0.462	0.427	2.219
Sr	S	RS	-0.369	-6.032	0.343	-3.641	-1.379	1.911	2.548	1.204	-11.795	-2.845	-7.106	0.642	0.742	0.847	2.366
Sr	Se	RS	-0.375	-6.032	0.343	-3.641	-1.379	1.911	2.548	1.204	-10.946	-2.751	-6.654	1.316	0.798	0.952	2.177
Sr	Te	RS	-0.381	-6.032	0.343	-3.641	-1.379	1.911	2.548	1.204	-9.867	-2.666	-6.109	0.099	0.945	1.141	1.827
Ag	F	RS	-0.156	-8.058	-1.667	-4.710	-0.479	1.316	1.883	2.968	-19.404	-4.273	-11.294	1.251	0.406	0.371	1.428
Ag	Cl	RS	-0.044	-8.058	-1.667	-4.710	-0.479	1.316	1.883	2.968	-13.902	-3.971	-8.700	0.574	0.679	0.756	1.666
Ag	Br	RS	-0.030	-8.058	-1.667	-4.710	-0.479	1.316	1.883	2.968	-12.650	-3.739	-8.001	0.708	0.749	0.882	1.869
Ag	I	ZB	0.037	-8.058	-1.667	-4.710	-0.479	1.316	1.883	2.968	-11.257	-3.513	-7.236	0.213	0.896	1.071	1.722
Cd	O	RS	-0.087	-9.581	0.839	-5.952	-1.309	1.232	1.736	2.604	-16.433	-3.006	-9.197	2.541	0.462	0.427	2.219
Cd	S	ZB	0.070	-9.581	0.839	-5.952	-1.309	1.232	1.736	2.604	-11.795	-2.845	-7.106	0.642	0.742	0.847	2.366
Cd	Se	ZB	0.083	-9.581	0.839	-5.952	-1.309	1.232	1.736	2.604	-10.946	-2.751	-6.654	1.316	0.798	0.952	2.177
Cd	Te	ZB	0.113	-9.581	0.839	-5.952	-1.309	1.232	1.736	2.604	-9.867	-2.666	-6.109	0.099	0.945	1.141	1.827

A	B	DFT	ΔE	IP(A)	EA(A)	HOMO(A)	LUMO(A)	rs(A)	rp(A)	rd(A)	IP(B)	EB(B)	HOMO(B)	LUMO(B)	rs(B)	rp(B)	rd(B)
		Classification															
In	N	ZB	0.150	-5.537	-0.256	-2.697	0.368	1.134	1.498	3.108	-13.585	-1.867	-7.239	3.057	0.539	0.511	1.540
In	P	ZB	0.170	-5.537	-0.256	-2.697	0.368	1.134	1.498	3.108	-9.751	-1.920	-5.596	0.183	0.826	0.966	1.771
In	As	ZB	0.122	-5.537	-0.256	-2.697	0.368	1.134	1.498	3.108	-9.262	-1.839	-5.341	0.064	0.847	1.043	2.023
In	Sb	ZB	0.080	-5.537	-0.256	-2.697	0.368	1.134	1.498	3.108	-8.468	-1.847	-4.991	0.105	1.001	1.232	2.065
Sn	Sn	ZB	0.016	-7.043	-1.039	-3.866	0.008	1.057	1.344	2.030	-7.043	-1.039	-3.866	0.008	1.057	1.344	2.030
B	Sb	ZB	0.581	-8.190	-0.107	-3.715	2.248	0.805	0.826	1.946	-8.468	-1.847	-4.991	0.105	1.001	1.232	2.065
Cs	F	RS	-0.112	-4.006	-0.570	-2.220	-0.548	2.464	3.164	1.974	-19.404	-4.273	-11.294	1.251	0.406	0.371	1.428
Cs	Cl	RS	-0.152	-4.006	-0.570	-2.220	-0.548	2.464	3.164	1.974	-13.902	-3.971	-8.700	0.574	0.679	0.756	1.666
Cs	Br	RS	-0.158	-4.006	-0.570	-2.220	-0.548	2.464	3.164	1.974	-12.650	-3.739	-8.001	0.708	0.749	0.882	1.869
Cs	I	RS	-0.165	-4.006	-0.570	-2.220	-0.548	2.464	3.164	1.974	-11.257	-3.513	-7.236	0.213	0.896	1.071	1.722
Ba	O	RS	-0.095	-5.516	0.278	-3.346	-2.129	2.149	2.632	1.351	-16.433	-3.006	-9.197	2.541	0.462	0.427	2.219
Ba	S	RS	-0.326	-5.516	0.278	-3.346	-2.129	2.149	2.632	1.351	-11.795	-2.845	-7.106	0.642	0.742	0.847	2.366
Ba	Se	RS	-0.350	-5.516	0.278	-3.346	-2.129	2.149	2.632	1.351	-10.946	-2.751	-6.654	1.316	0.798	0.952	2.177
Ba	Te	RS	-0.381	-5.516	0.278	-3.346	-2.129	2.149	2.632	1.351	-9.867	-2.666	-6.109	0.099	0.945	1.141	1.827
Ge	C	ZB	0.808	-7.567	-0.949	-4.046	2.175	0.917	1.162	2.373	-10.852	-0.872	-5.416	1.992	0.644	0.630	1.631
Sn	C	ZB	0.450	-7.043	-1.039	-3.866	0.008	1.057	1.344	2.030	-10.852	-0.872	-5.416	1.992	0.644	0.630	1.631
Ge	Si	ZB	0.264	-7.567	-0.949	-4.046	2.175	0.917	1.162	2.373	-7.758	-0.993	-4.163	0.440	0.938	1.134	1.890
Sn	Si	ZB	0.136	-7.043	-1.039	-3.866	0.008	1.057	1.344	2.030	-7.758	-0.993	-4.163	0.440	0.938	1.134	1.890
Sn	Ge	ZB	0.087	-7.043	-1.039	-3.866	0.008	1.057	1.344	2.030	-7.567	-0.949	-4.046	2.175	0.917	1.162	2.373

TABLE II. Values related to 82 AB binaries: total energy difference between Rock-Salt and Zinc-Blende ($\Delta E = E^{RS} - E^{ZB}$) (calculated using DFT) and seven atomic properties of corresponding A and B atom. All the data are taken from ref.[3]. IP, EA, HOMO, LUMO stand for Ionization Potential, Electron Affinity, Highest Occupied Molecular Orbital and Lowest Unoccupied Molecular Orbital, respectively. rs, rp, rd denote distances where the radial probability density reaches the maximum for s, p, d electronic shells, respectively.

III. ALLOY SUPERCELL

Figure-S.4 shows the alloy supercell used in DFT calculations.

IV. REFERENCES

[1] W. Kirch, ed., "Pearson's correlation coefficient," in *Encyclopedia of Public Health* (Springer Netherlands, Dordrecht, 2008) pp. 1090–1091.

[2] D. Rajnarayan and D. Wolpert, "Bias-variance trade-offs: Novel applications," in *Encyclopedia of Machine Learning*, edited by C. Sammut and G. I. Webb (Springer US, Boston, MA, 2010) pp. 101–110.

[3] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big Data of Materials Science: Critical Role of the Descriptor," Phys. Rev. Lett. **114**, 105503 (2015).

[4] F. Gao and L. Han, "Implementing the Nelder-Mead simplex algorithm with adaptive parameters," Comput Optim Appl **51**, 259–277 (2012).

[5] G. H. Golub and C. F. Van Loan, *Matrix computations*, fourth edition ed., Johns Hopkins studies in the mathematical sciences (The Johns Hopkins University Press, Baltimore, 2013).

[6] C. G. Broyden, "The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations," IMA Journal of Applied Mathematics **6**, 76–90 (1970), https://academic.oup.com/imamat/article-pdf/6/1/76/2233756/6-1-76.pdf.

[7] R. dembo, S. Eisenstat, and T. Steihaug, "Inexact Newton Methods," SIAM J. Numer. Anal. **19**, 400–408 (1982).

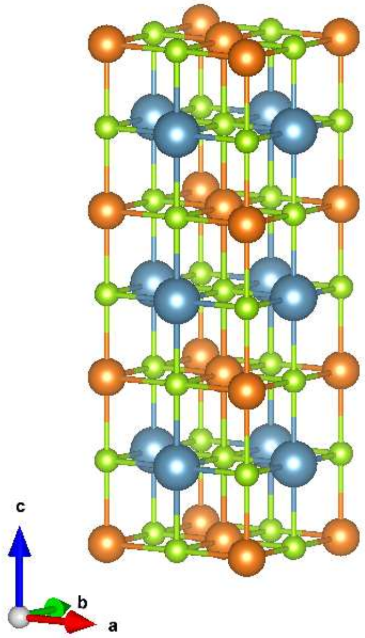

FIG. S.4. $Mg_{0.5}Ca_{0.5}Se$ rocksalt supercell: Mg is reported in orange, Ca in blue and Se in green. The supercell is obtained alternating layers of Mg and Ca in the cation sub-lattice along the c primitive vector.