Electronic Supplementary Material (ESI) for Digital Discovery.

Supplemental Information for:

Grouped representation of interatomic distances as a similarity measure for crystal structures

Ruizhi Zhang[1], Sohan Seth[2] and James Cumby[1]*
[1]*School of Chemistry, University of Edinburgh, Edinburgh, EH9 3FJ*
[2]*School of Informatics, University of Edinburgh, EH8 9AB*

Email: james.cumby@ed.ac.uk

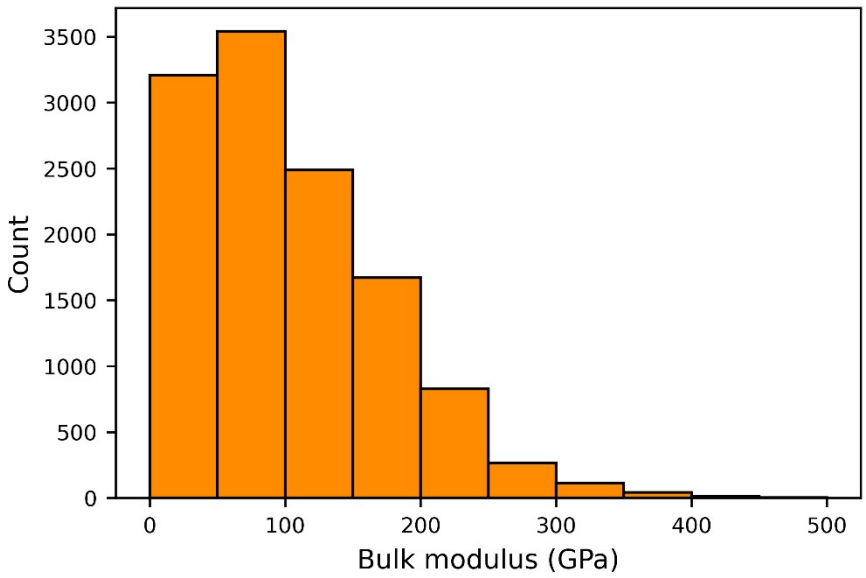


Figure S1. Distribution of bulk moduli for structures extracted from the materials project.

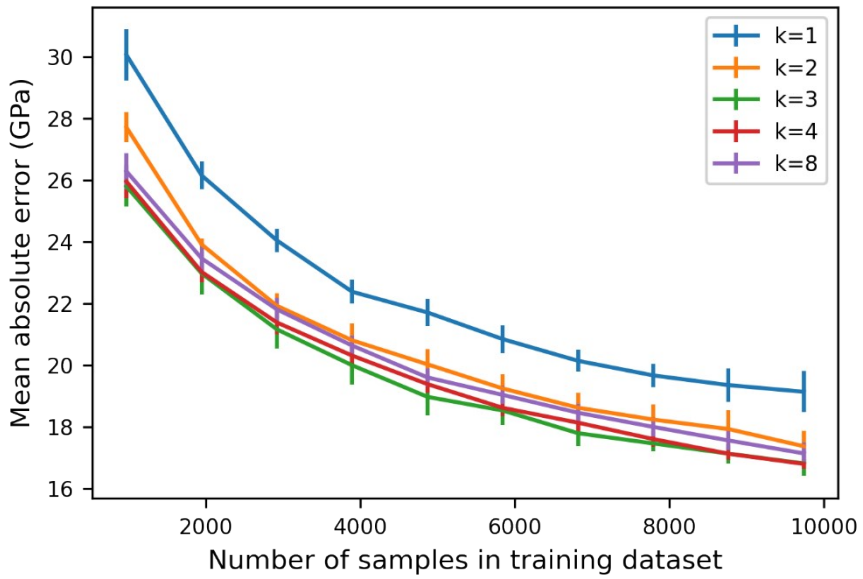


Figure S2. Mean absolute error as a function of the number of training samples at different *k* values in kNN model.

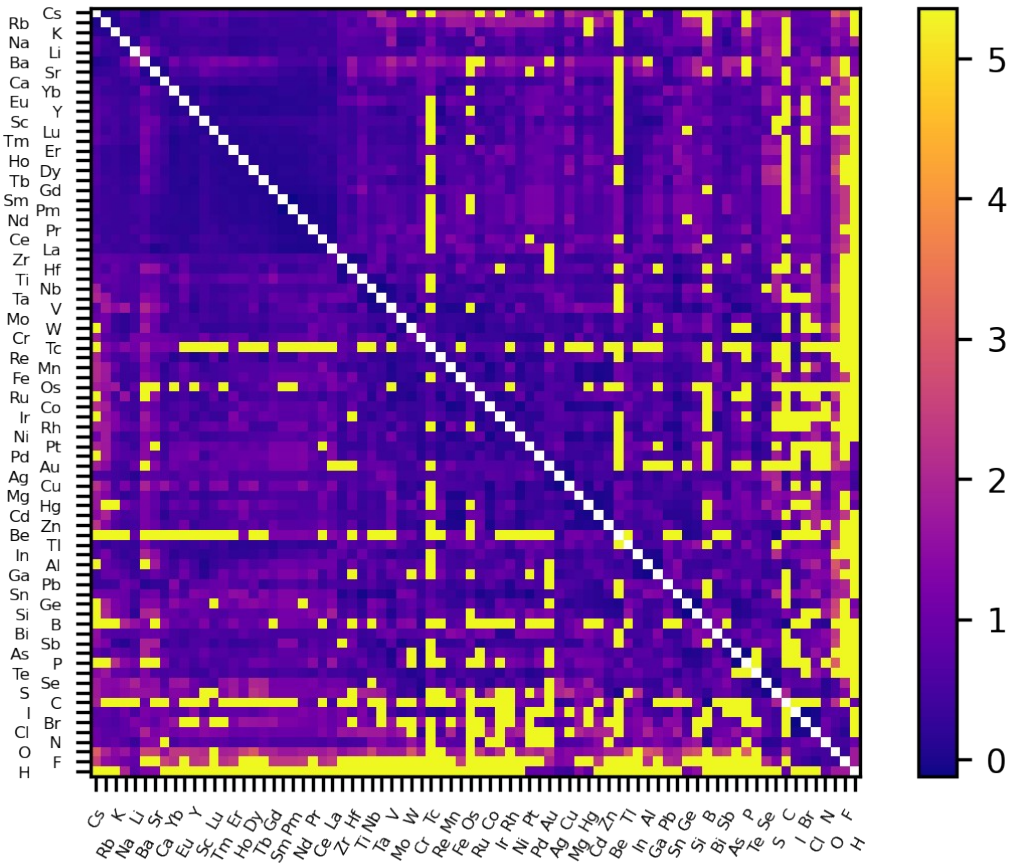


Figure S3. Log value of elemental distance matrix, i.e. log(diss(A,B)) – see definition in method section. Diss(A,B) values are used to calculate compositional similarity, based on a modification of work by Hautier *et al.*[1]

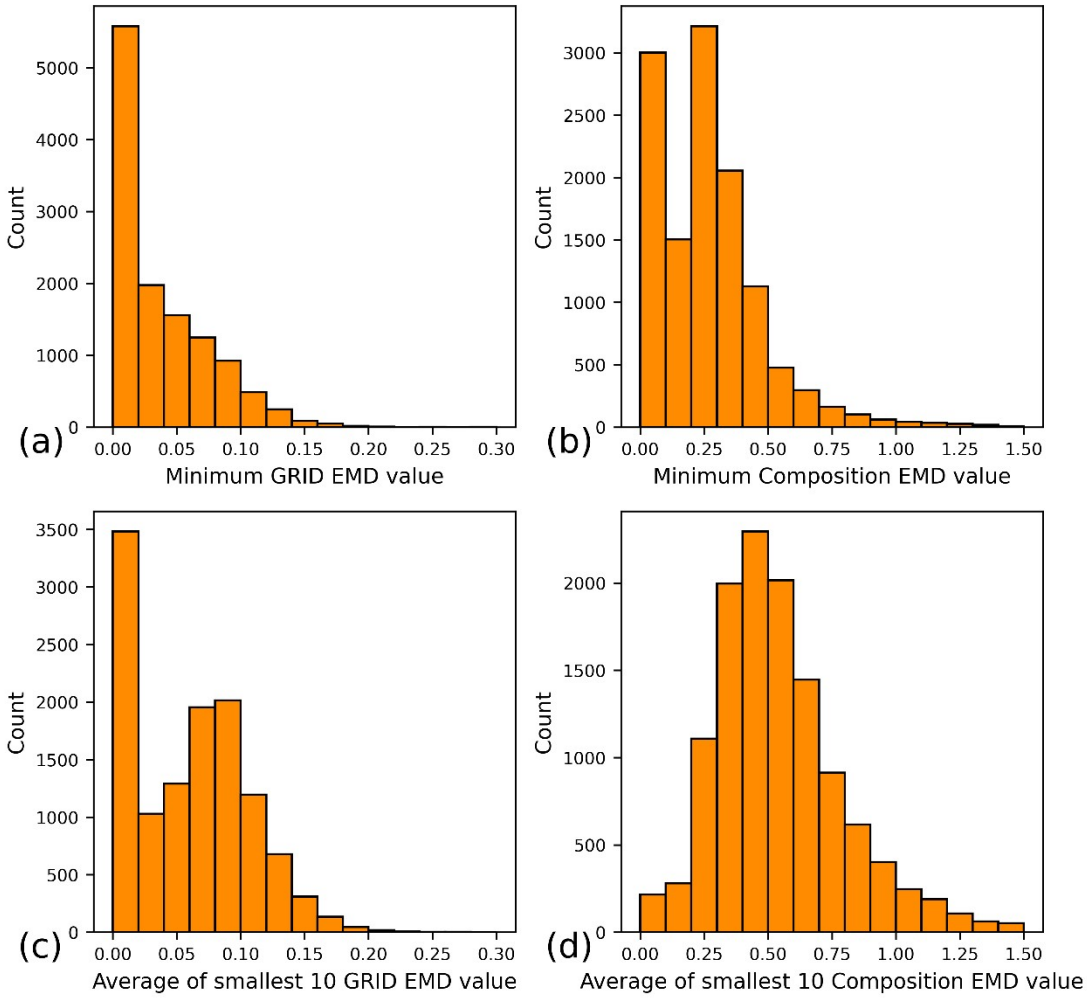


Figure S4. Distribution of EMD values of all the 12,178 compounds in the datasets. As only the smallest EMD values are used in the kNN model, the minimum EMD values for each compound is found and plotted for (a) GRID and (b) composition, respectively. The average values of ten smallest EMD values are also calculated and shown for (c) GRID and (d) composition, respectively.

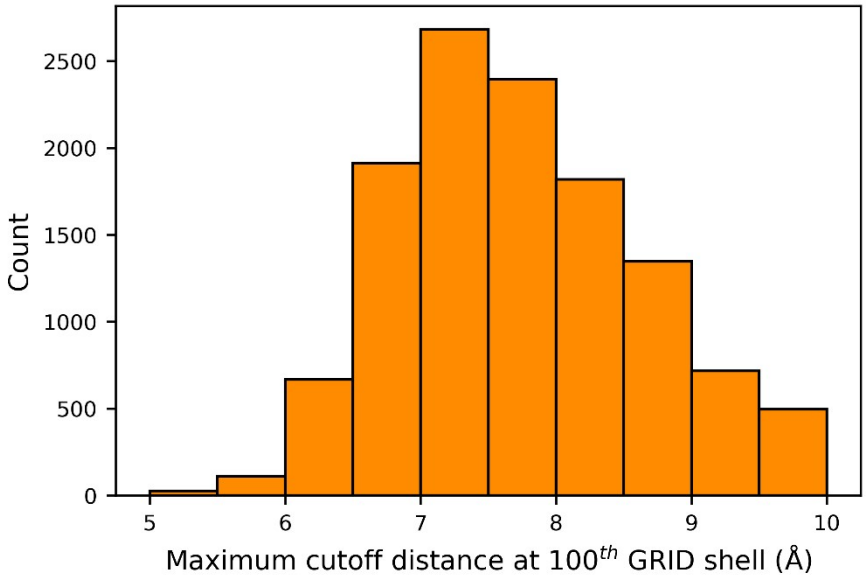


Figure S5. A histogram of the maximum distance recorded by 100 GRID groups for all materials in the bulk modulus dataset. Distances are those prior to exponential smoothing.

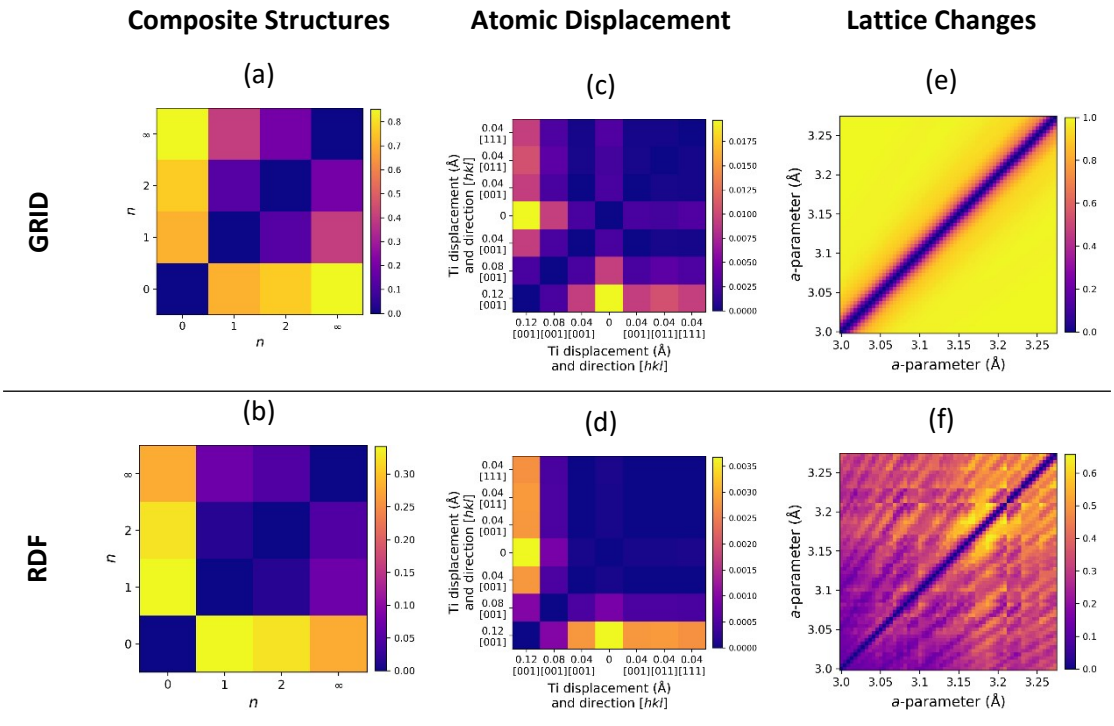


Figure S6. Pairwise cosine distance using both GRID and RDF for (a,b) composite Ruddlesden-Popper structures, (c,d) distorted $BaTiO_3$ perovskites generated through Ti displacements and (e,f) cubic perovskite with linearly varying lattice parameters.

Table S1. Simulated distorted structures of $BaTiO_3$ (from the Materials Project, mp-2998) generated through Ti atom displacements (directions refer to cubic unit cell, a = 4.036 Å).

Displacement direction	Displacement (fractional unit)	Space group after distortion	Ti atomic position (fractional coordinates *x,y,z*)
--	0.00	Pm-3m	0.50, 0.50, 0.50
[001]	0.01	P4mm	0.50, 0.50, 0.51
[001]	0.02	P4mm	0.50, 0.50, 0.52
[001]	0.03	P4mm	0.50, 0.50, 0.53
[011]	0.01	Amm2	0.50, 0.507, 0.507
[111]	0.01	$P6_3/mmc$	0.506, 0.506, 0.506

Table S2: Selected literature reports of machine-learned bulk modulus prediction. Underlying datasets are obtained from the materials project (MP), Thermoelectric design lab (TE) or Automatic-FLOW (AFLOW).

Dataset (size) ref	**Important features**	**Training algorithm**	**Bulk Modulus**		**Shear Modulus**		**E_F**	**Band Gap**	
			MAE	**RMSE**	**MAE**	**RMSE**	**MAE**	**MAE**	**RMSE**
MP (1940) 2	Volume per atom; row number; cohesive energy; electronegativity	local polynomial regression	--	0.0750 (log(GPa))	-	0.1378 (log(GPa)	-	-	-
TE (1805) 3	Molar volume; atomic radius	Random forest	13.56 (GPa)	22.65 (GPa)	-	-	-	-	-
AFLOW (~3000) 4	Property-labeled material fragments	gradient boosting decision trees	8.68 (GPa)	14.25 (GPa)	10.62 (GPa)	18.43 (GPa)	-	0.35 (eV)	0.51 (eV)
MP (3248) 5	Cohesive energy; volume per atom; density	Support Vector Machine Regression	--	17.2 (GPa)	-	16.5 (GPa)	-	-	-
MP (3402) 6	Crystal Graph	Convolutional Neural Networks	0.054 (log(GPa))	--	0.087 (log(GPa))	-	0.039 (eV)	0.388 (eV)	-
MP (4664) 7	Crystal Graph	Convolutional Neural Networks	0.050 (log(GPa))	--	0.079 (log(GPa))	-	0.028 (eV / atom)	0.33 (eV)	-
MP (6975) 8	Distance / adjacency matrix; Atomic properties	Neural Networks	10.05 (GPa)	15.16 (GPa) 0.079 (log (GPa))	9.92 (GPa)	14.09 (GPa) 0.123 (log(GPa))			

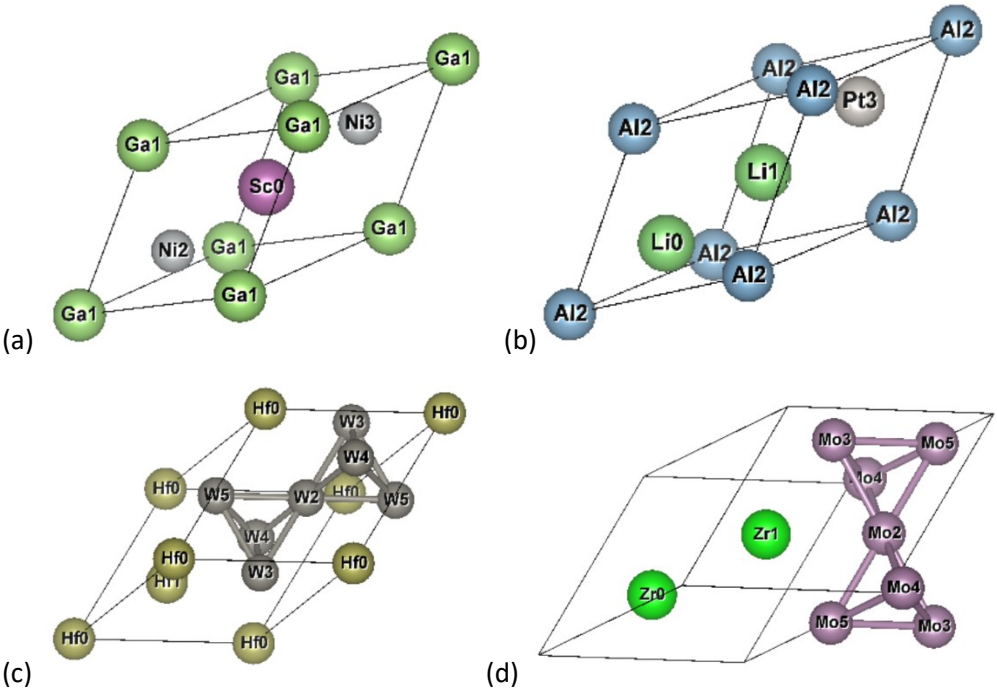


Figure S7. Example structure pairs with different composition but similar structures which yield identical GRIDs: (a) $ScGaNi_2$ (mp-11400) and (b) $PtAlLi_2$ (mp-30818); and (c) HfW_2 (mp-1400) and $ZrMo_2$ (mp-2049).

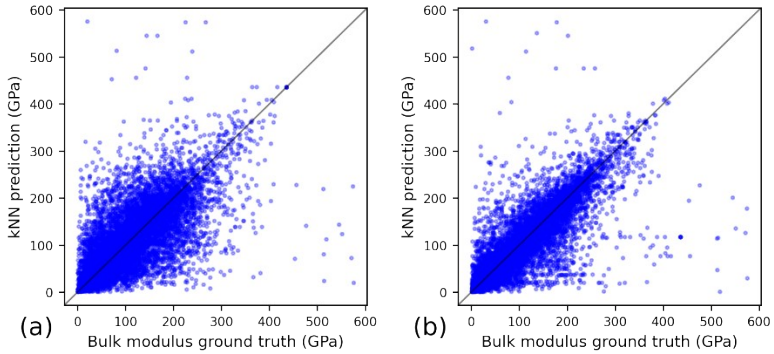


Figure S8. Predicted bulk modulus vs. materials project values using EMD dissimilarity based on (a) GRID alone; and (c) composition alone.

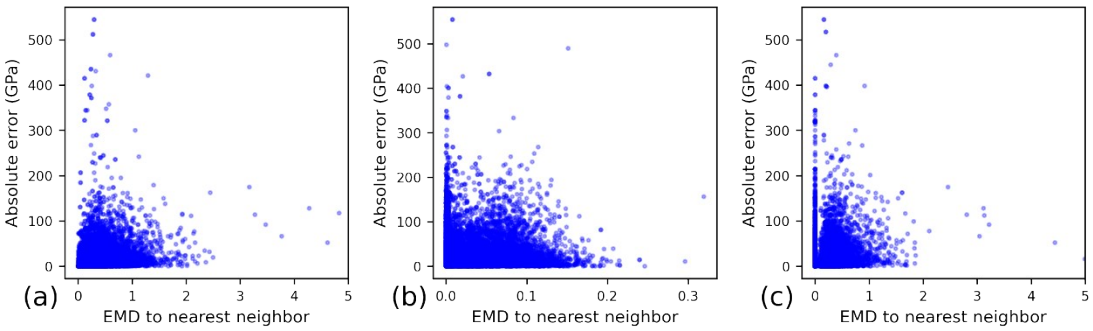
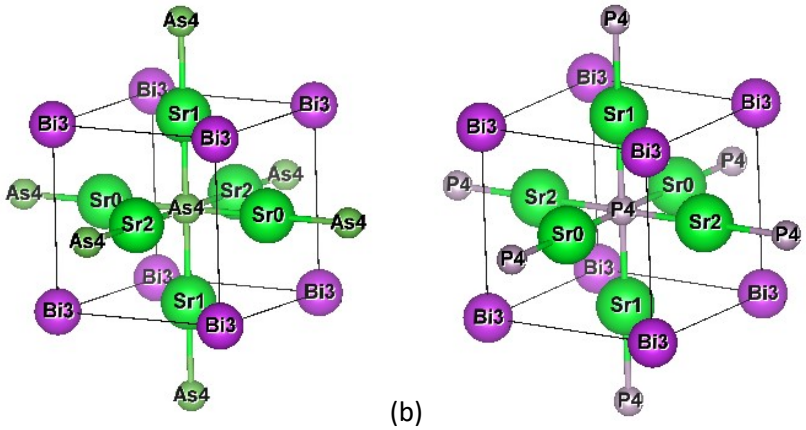


Figure S9. Absolute error in bulk modulus prediction *vs.* EMD to the nearest neighbour of each material using (a) GRID; (b) composition; and (c) combined GRID and composition.



(a) (b)

Figure S10. Crystal structures of the two materials showing maximum prediction error in bulk modulus, both of which have very similar GRID representations. (a) $BiAsSr_3$ has a lattice constant of 5.880 Å and bulk modulus is 575 GPa; (b) $BiPSr_3$ has a lattice constant of 5.777 Å and bulk modulus of 30 GPa (fractional coordinates are identical).

References

1. Hautier, G.; Fischer, C.; Ehrlacher, V.; Jain, A. and Ceder, G., Data Mined Ionic Substitutions for the Discovery of New Compounds, *Inorg. Chem.*, **2011**, 50, 656–663.
2. de Jong, M.; Chen, W.; Notestine, R.; Persson, K.; Ceder, G.; Jain, A.; Asta, M. and Gamst, A., A Statistical Learning Framework for Materials Science: Application to Elastic Moduli of k-nary Inorganic Polycrystalline Compounds, *Sci. Rep.*, **2016**, 6, 34256.
3. Furmanchuk, A.; Agrawal, A. and Choudhary, A., Predictive Analytics for Crystalline Materials: Bulk Modulus, *RSC Adv.*, **2016**, 6, 95246–95251.
4. Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S. and Tropsha, A., Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals, *Nat. Commun.*, **2017**, 8, 15679.
5. Tehrani, A. M.; Oliynyk, A. O.; Parry, M.; Rizvi, Z.; Couper, S.; Lin, F.; Miyagi, L.; Sparks, T. D. and Brgoch, J., Machine Learning Directed Search for Ultraincompressible, Superhard Materials, *J. Amer. Chem. Soc.*, **2018**, 140, 9844–9853.
6. Xie, T. and Grossman, J. C., Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties, *Phys. Rev. Lett.*, **2018**, 120, 145301.
7. Chen, C.; Ye, W.; Zuo, Y.; Zheng, C. and Ong, S. P. , Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals, *Chem. Mater.*, **2019**, 31, 3564–3572.
8. Zeng, S.; Li, G.; Zhao, Y.; Wang, R. and Ni, J., Machine Learning-Aided Design of Materials with Target Elastic Properties, *J. Phys. Chem. C*, **2019**, 123, 5042–5047.