luMOD: A lightweight universal Mid-range atomic Orientational Descriptor

Yanhao Deng,[†,‡] Tianle Jiang,[†] Zeyu Deng,[*,‡] and Yanming Wang[*,†,¶]

†*University of Michigan - Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University, 800 Dongchuan Rd., Minhang District, Shanghai, 200240, P. R. China*

‡*Department of Materials Science and Engineering, National University of Singapore, Singapore, Singapore*

¶*Global Institute of Future Technology, Shanghai Jiao Tong University, 800 Dongchuan Rd., Minhang District, Shanghai, 200240, P. R. China*

E-mail: msedz@nus.edu.sg; yanming.wang@sjtu.edu.cn

Dataset Construction

Table S1: The materials and potentials chosen for each crystal structure. For structures without potentials, they are tattled with a random displacement ($< 0.1\mathring{A}$) instead of running MD simulations.

Structure	Composition	Potential	Reference
FCC	Au	EAM	[1]
BCC	Fe	EAM/FS	[2]
HCP	Mg	EAM/FS	[3]
DC	Si	SW	[4]
NaCl	NaCl	buck/coul/long	[5]
CsCl	CsCl	EIM	[6]
CaF_2	CaF_2	born/coul/long	[7]
TiO_2	TiO_2 /	/	
ABX_3	$CaTiO_3$	/	/
P2	$LiCoO_2$	/	/
$FePO_4$	$FePO_4$ /	/	
$LiFePO_4$	$CaTiO_3$	/	/

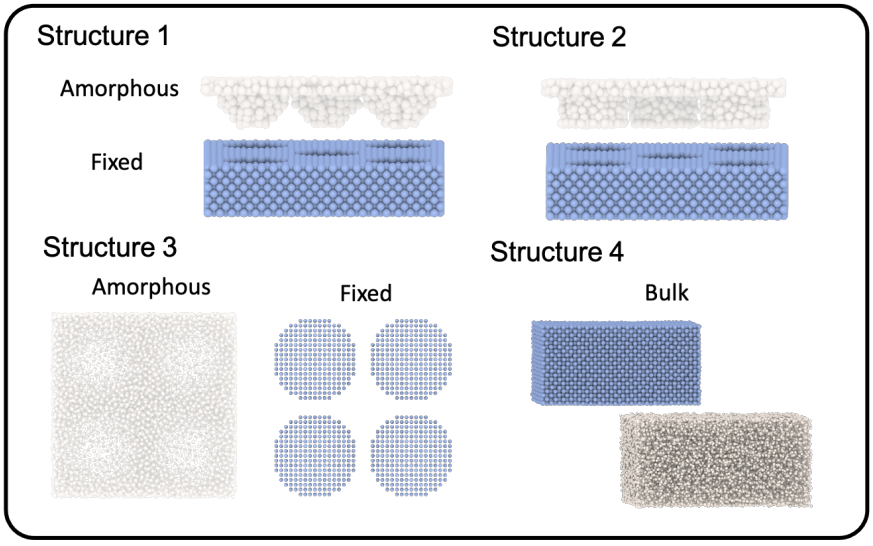
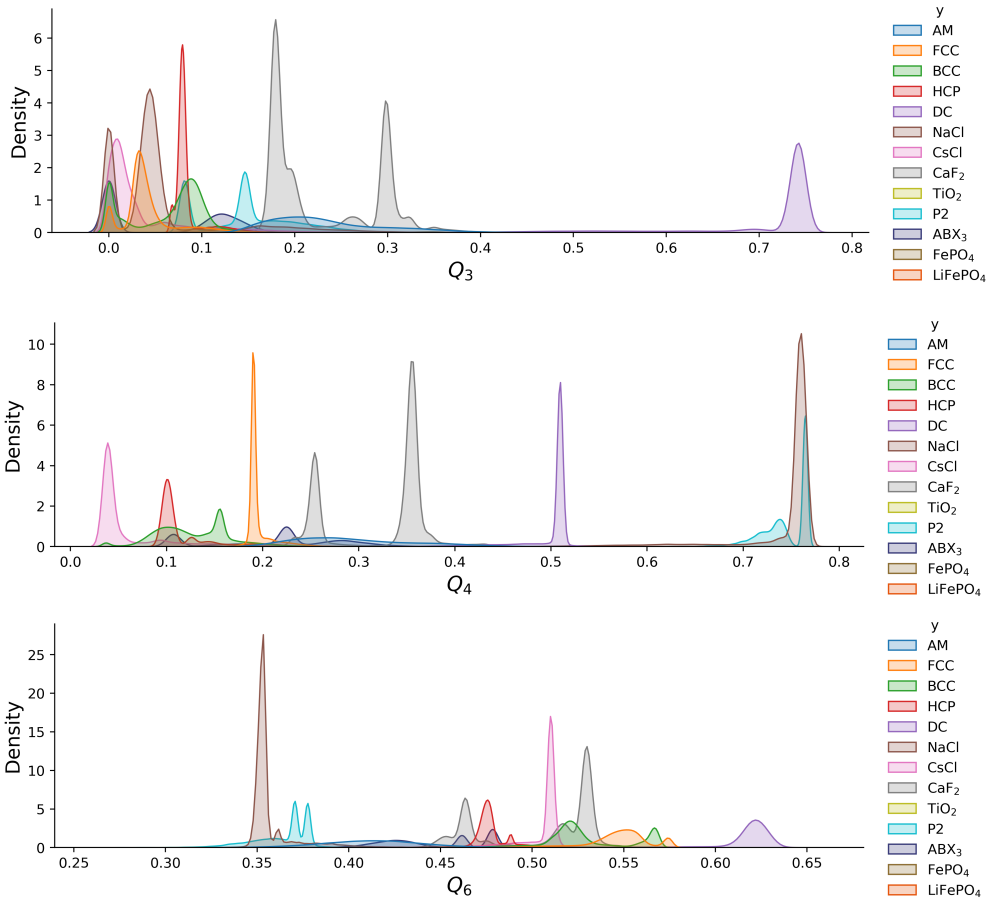


Figure S1: Four representative structures designed to generate the labeled data for each crystal structure.

These atomic configurations were generated by ATOMSK[8]. Molecular Dynamics simulations were performed using the LAMMPS code for all structures in the database. All MD simulations were computed with a time step of 1 fs using the Nosé-Hoover thermostat in a NPT

ensemble for 10 ps.[9] For each crystal structure, as introduced in,[10] we designed the different structures to simulate different interfaces, and only the fixed parts are included as training data. For the amorphous class, we simulate with bulk structures at a temperature above the melting point. 20% of this dataset is used as a validation set to obtain the accuracy.

Feature Distribution



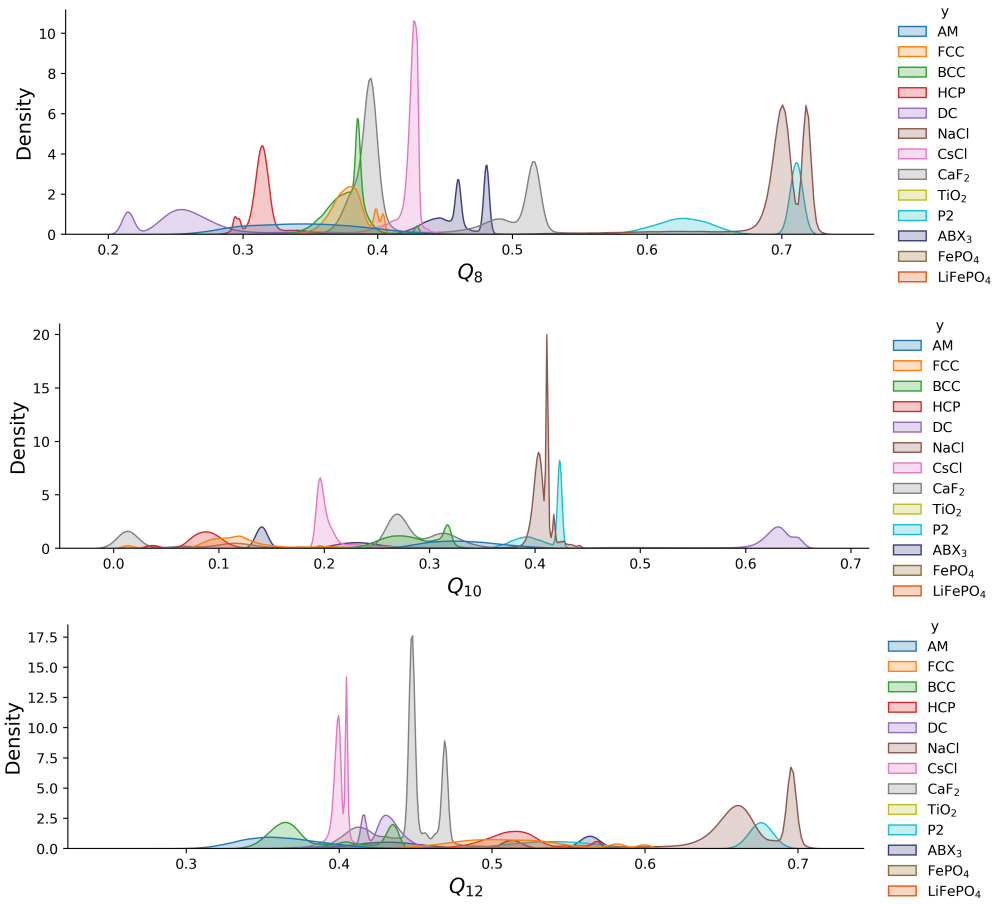
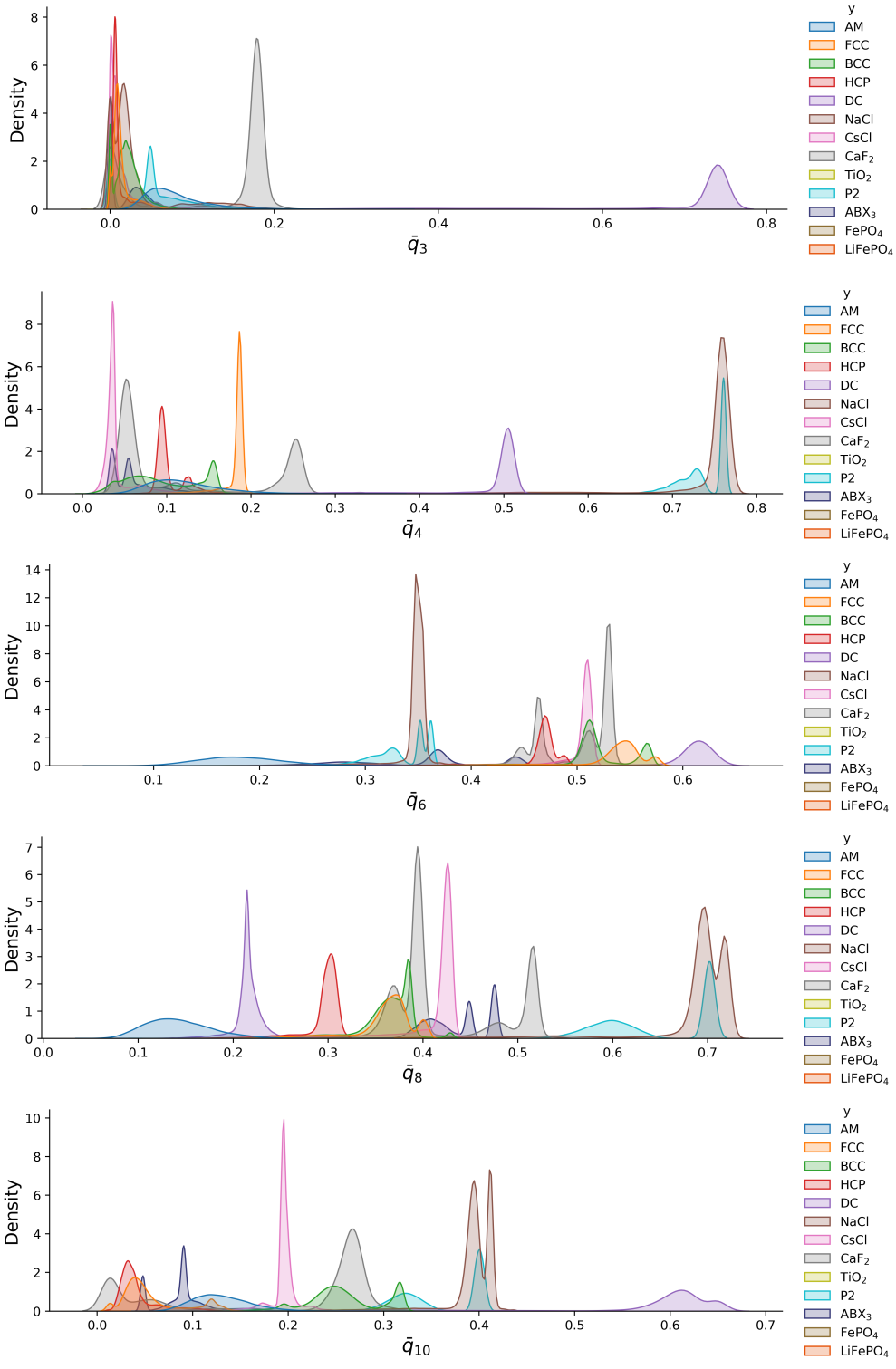


Figure S2: The distributions of q_l calculated with different crystal structures.


Density
$\bar{q}_3$
$\bar{q}_4$
$\bar{q}_6$
$\bar{q}_8$
$\bar{q}_{10}$
y
AM
FCC
BCC
HCP
DC
NaCl
CsCl
CaF_2
TiO_2
P2
ABX_3
$FePO_4$
$LiFePO_4$

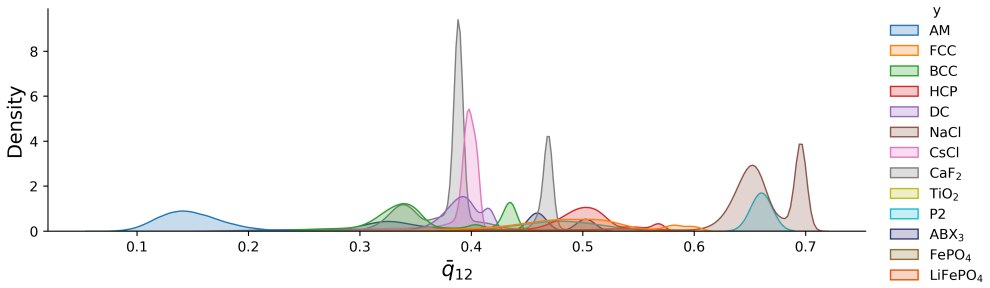
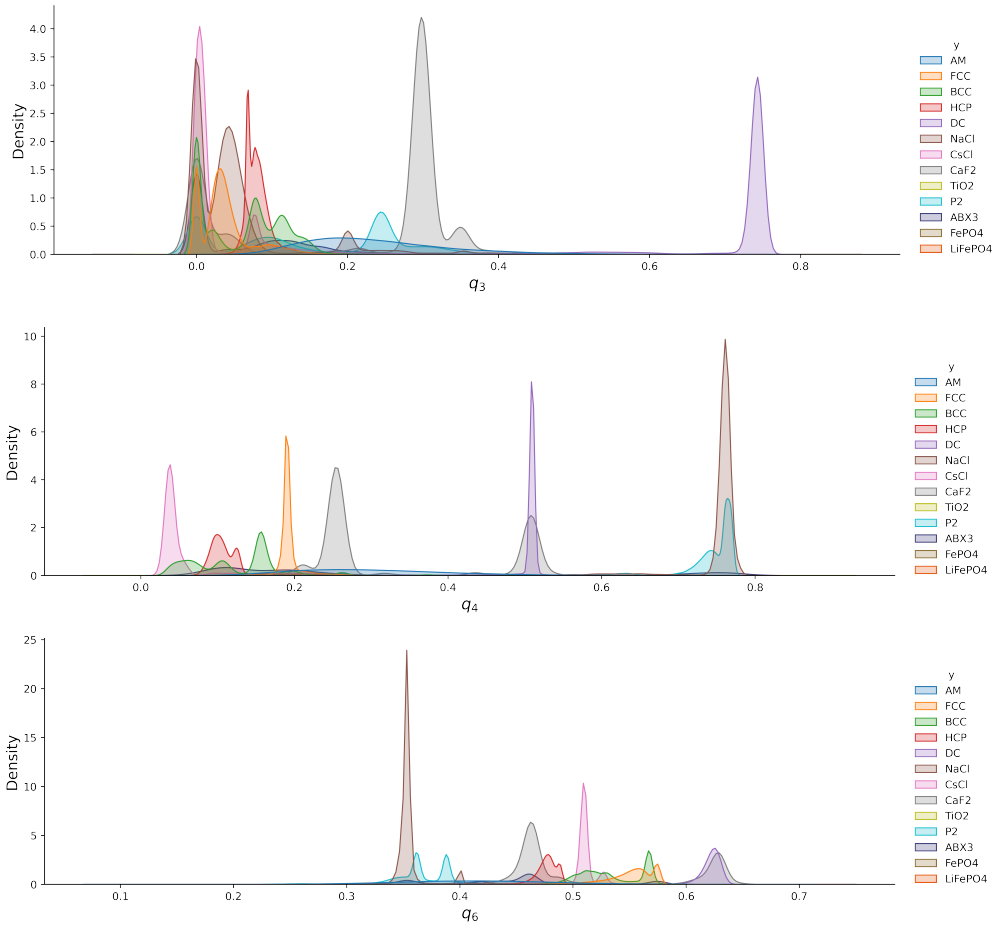


Figure S3: The distributions of $\bar{q}_l$ calculated with different crystal structures.



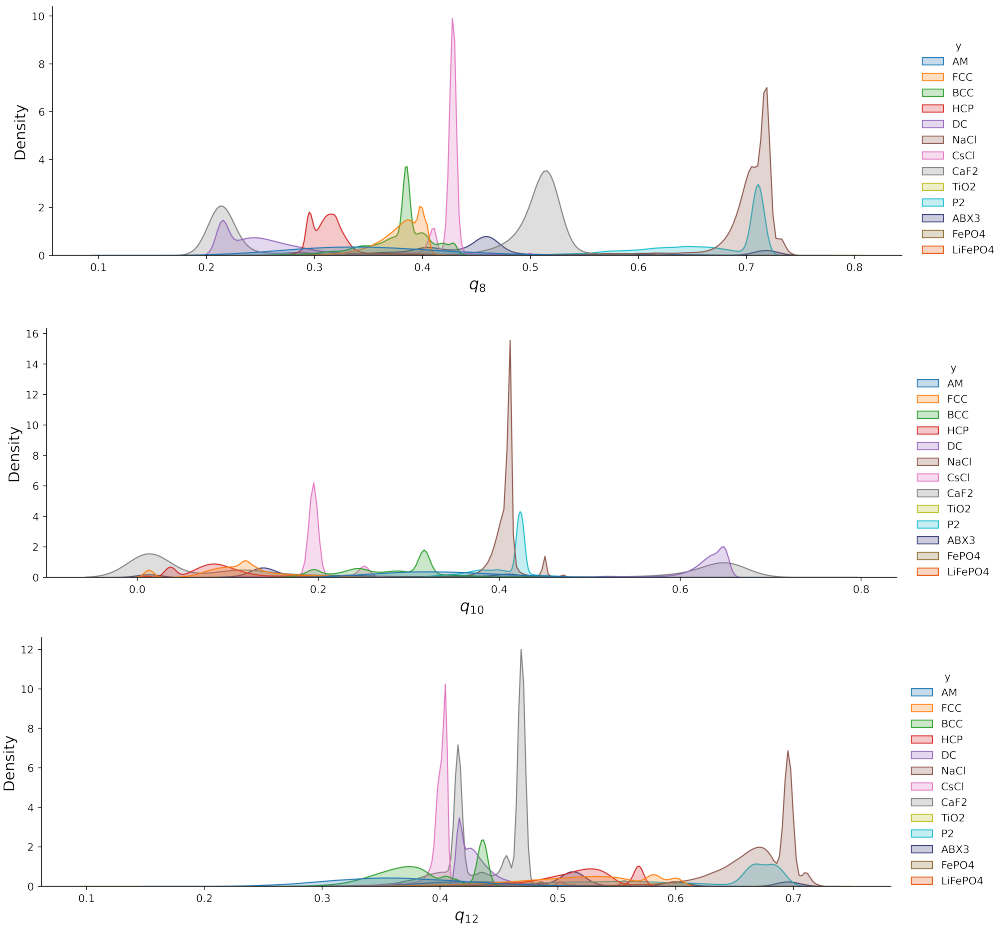


Figure S4: The distributions of Q_l calculated with different crystal structures.

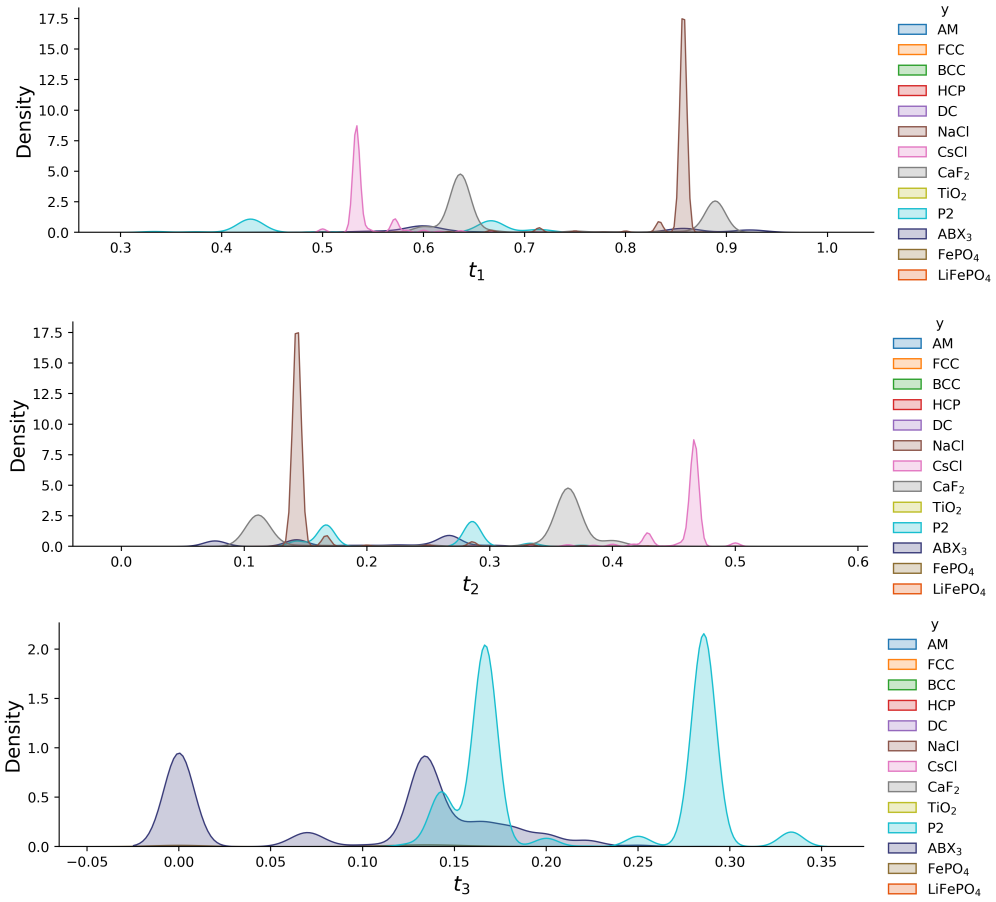


Figure S5: The distributions of type proportion t_n calculated with different crystal structures.

Calculation of SOAP and ACSF

We use the DScribe package to calculate these two descriptors.[11,12] In this work, luMOD calculates the neighbors within the first 2 neighbor layers (often considered as within $5\mathring{A}$), with $l = 3, 4, 6, 8, 12$ in spherical harmonics. For a consistent comparison, we chose similar settings for ACSF and SOAP, as well as maintaining a low dimension.

For SOAP, **r_max** is 6, **l_max** is 3, and **species** includes all the elements in the training data. As luMOD does not use a radial basis equivalent,**n_max** is set to 2.

For ACSF, **r_cut** is 6.0. For two-body interaction, **g2_params** is set to `[[1, 1], [1, 2]]`. For three-body interaction, **g4_params** is set to `[[1, 1,1], [1, 2,1]`.

Classification Performance

We adopted an ensemble machine-learning method called stacking to take advantage of different models, as each basic machine-learning model may perform well but fail in certain corner cases. Our framework consists of Level-0 Models (4 based models) and Level-1 Models (a mega model). The former fits the training data and generates predictions to be compiled. The latter learns how to best combine the predictions of the base models and gives the final structure classification results. The details of the ensemble model and training process can be found in.[10]

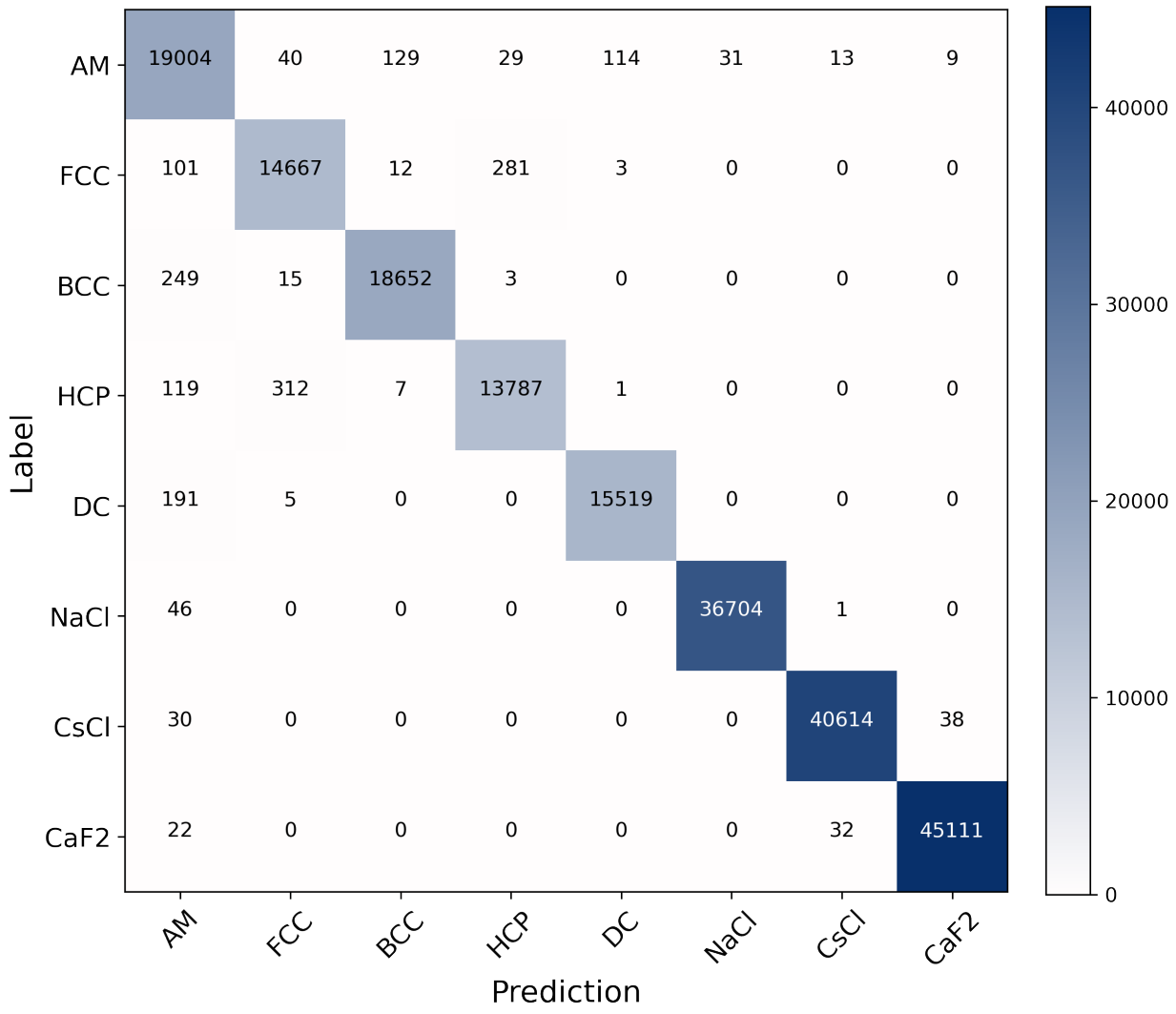


Figure S6: Performance of the previous LESS model by the 12-dimension feature. Crystal classes including $CaTiO_3$, $LiCoO_2$, and $CaTiO_3$ are not shown because this previous model can not identify them.

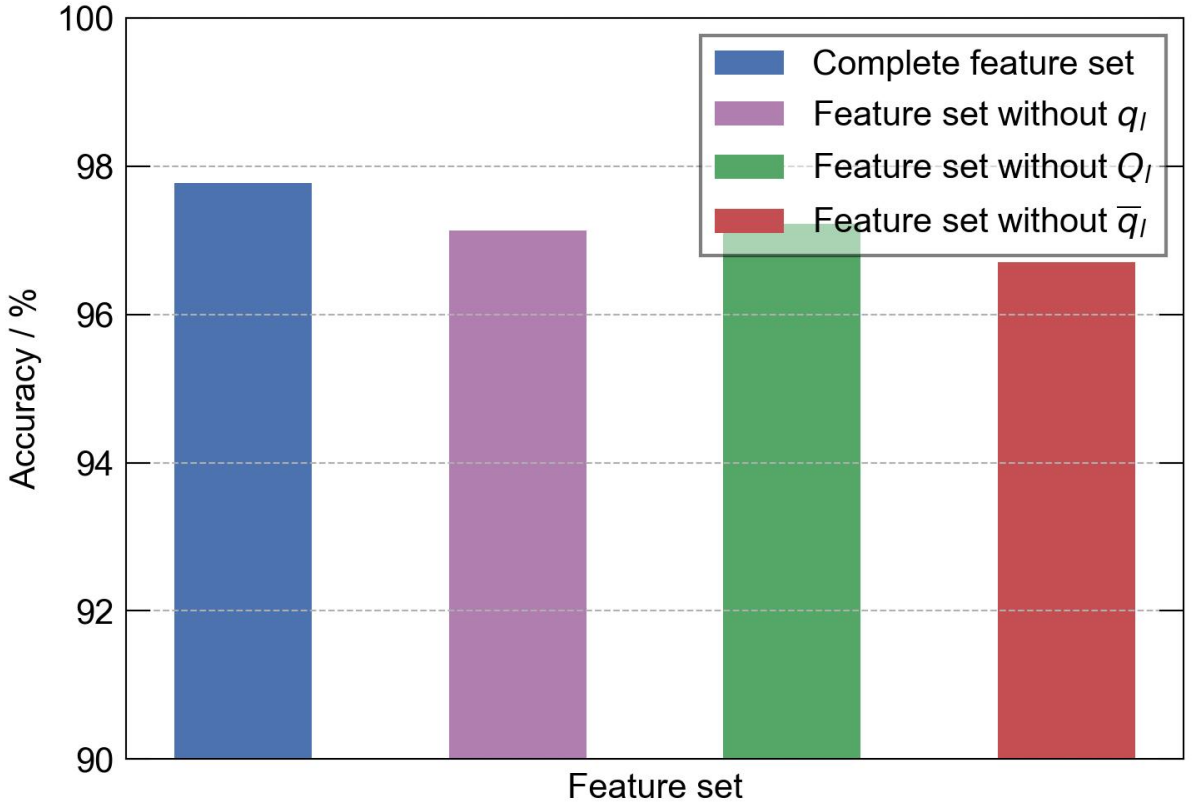


Figure S7: Performance for each part of the dataset.

References

(1) Foiles, S.; Baskes, M.; Daw, M. S. Embedded-atom-method functions for the fcc metals Cu, Ag, Au, Ni, Pd, Pt, and their alloys. *Physical review B* **1986**, *33*, 7983.

(2) Mendelev, M.; Han, S.; Srolovitz, D.; Ackland, G.; Sun, D.; Asta, M. Development of new interatomic potentials appropriate for crystalline and liquid iron. *Philosophical magazine* **2003**, *83*, 3977–3994.

(3) Sun, D.; Mendelev, M.; Becker, C.; Kudin, K.; Haxhimali, T.; Asta, M.; Hoyt, J.; Karma, A.; Srolovitz, D. J. Crystal-melt interfacial free energies in hcp metals: A molecular dynamics study of Mg. *Physical Review B* **2006**, *73*, 024116.

(4) Stillinger, F. H.; Weber, T. A. Computer simulation of local order in condensed phases of silicon. *Physical review B* **1985**, *31*, 5262.

(5) Khrapak, S. A.; Chaudhuri, M.; Morfill, G. E. Freezing of Lennard-Jones-type fluids. *The Journal of chemical physics* **2011**, *134*, 054120.

(6) Zhou, X.; Doty, F.; Yang, P. Atomistic simulation study of atomic size effects on B1 (NaCl), B2 (CsCl), and B3 (zinc-blende) crystal stability of binary ionic compounds. *Computational materials science* **2011**, *50*, 2470–2481.

(7) Dixon, M.; Gillan, M. Computer simulation of fast ion transport in fluorites. *Le Journal de Physique Colloques* **1980**, *41*, C6–24.

(8) Hirel, P. Atomsk: A tool for manipulating and converting atomic data files. *Computer Physics Communications* **2015**, *197*, 212–219.

(9) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comp. Phys. Comm.* **2022**, *271*, 108171.

(10) Deng, Y.; Wang, Y.; Xu, K.; Wang, Y. Lightweight Extendable Stacking Framework for Structure Classification in Atomistic Simulations. *Journal of Chemical Theory and Computation* **2023**, *19*, 8332–8339.

(11) Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications* **2020**, *247*, 106949.

(12) Laakso, J.; Himanen, L.; Homm, H.; Morooka, E. V.; Jäger, M. O.; Todorović, M.;

Rinke, P. Updates to the DScribe library: New descriptors and derivatives. *The Journal of Chemical Physics* **2023**, *158*.