Supplementary Information for

Uncertainty-Quantified Hybrid ML/DFT High Throughput Screening Method for Crystals

Juhwan Noh, Geun Ho Gu, Sungwon Kim, and Yousung Jung*

Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291, Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

*ysjn@kaist.ac.kr

S.1 DFT calculations

We performed spin-polarized PBE[1]+*U* calculations and PAW[2]-PBE pseudopotentials as implemented in the *ab initio* package, VASP,[3] and we used 3.9 as *U*-value for Mn taken from Materials Project.[4] We relaxed both atomic positions and unit cell parameters using conjugate gradient descent method with convergence criteria of 1.0e-5 for energy and 0.05 $\mathrm{eV/\AA}$ for force with 500 eV cut-off energy. Brillouin zone is used with *k*-point densities at or larger than 1000 k-points per atoms using the *Pymatgen* Ver. 2019.1.24 package.[5] All the other detail for computing E_{hull} and $\Delta G^{\mathrm{min}}_{\mathrm{pbx,1.5V}}$ are the same as our previous work.[6]

S.2 Additional figures and tables

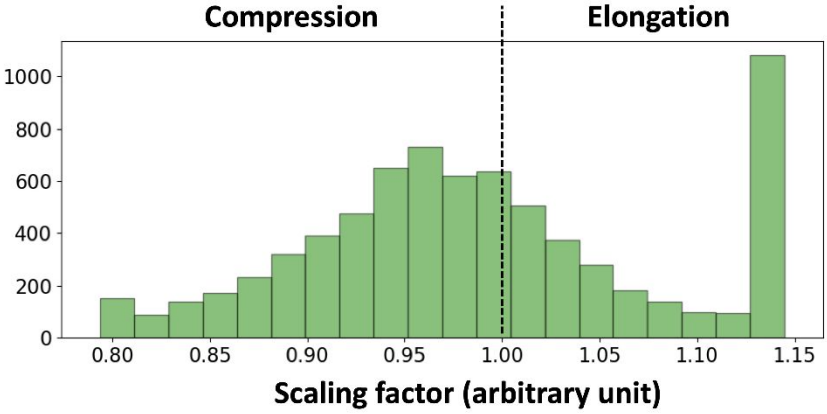

Figure S1. Distribution of scaling factor values proposed by Chu et al.[7] implemented in *Pymatgen*.[5] Here, scaling factor lower than 1.00 means compressed cell from the initial geometry and scaling factor larger than 1.00 means elongated cell from the initial geometry.

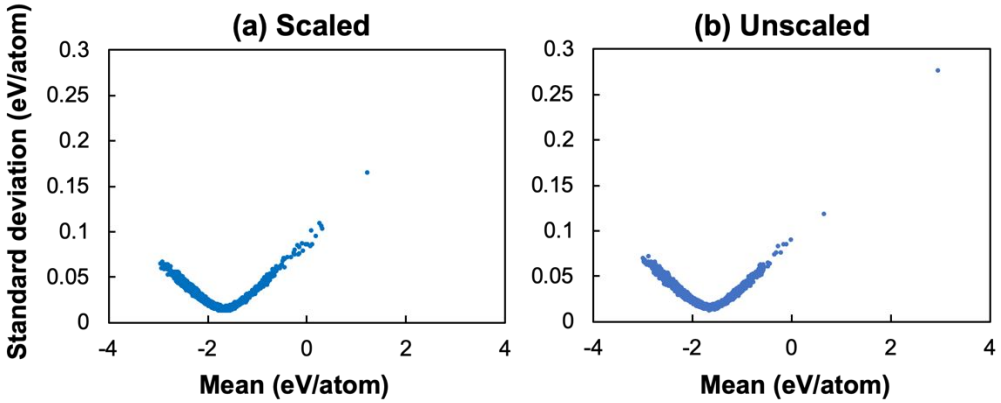

Figure S2. Results of the computed mean and standard deviation of the 200-randomly sampled data for (a) lattice scaling and (b) substitution preprocessing method.

Table S1. Screening results for various models and formation energy estimation method (CGCNN, CGCNN-H, CGCNN-HD (μ), CGCNN-HD ($\mu - \sigma$) and CGCNN-HD ($\mu - 2\sigma$)). *The number of required DFT* represents materials satisfying criteria before DFT calculations, and *screening criteria met* represents materials satisfying criteria after DFT calculations. *The number of reproduced data* represents materials reproduced by ML-HTS among the 74 materials selected from the DFT-HTS. DFT-HTS data is listed for reference, and also the computed MAE (eV/atom) values are listed in the parenthesis.

Model Name	# of required DFT	Screening criteria met	# of reproduced data
DFT-HTS	7,356	74	-
CGCNN	37 (0.014)	24	22 (0.027)

CGCNN-H	48 (0.002)	33	29 (0.025)
CGCNN-HD (μ)	65 (0.017)	43	35 (0.028)
CGCNN-HD ($\mu - \sigma$)	88 (0.042)	53	42 (0.037)
CGCNN-HD ($\mu - 2\sigma$)	110 (0.069)	61	50 (0.050)

References

1. Perdew, J. P.; Burke, K.; Ernzerhof, M., Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.
2. Blöchl, P. E., Projector augmented-wave method. *Phys. Rev. B* **1994**, *50*, 17953.
3. Kresse, G.; Furthmüller, J., Software VASP, vienna (1999). *Phys. Rev. B* **1996**, *54*, 169.
4. Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G., Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002.
5. Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G., Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013**, *68*, 314-319.
6. Noh, J.; Kim, S.; ho Gu, G.; Shinde, A.; Zhou, L.; Gregoire, J. M.; Jung, Y., Unveiling new stable manganese based photoanode materials via theoretical high-throughput screening and experiments. *Chem. Commun.* **2019**, *55*, 13418-13421.
7. Chu, I.-H.; Roychowdhury, S.; Han, D.; Jain, A.; Ong, S. P., Predicting the volumes of crystals. *Comput. Mater. Sci.* **2018**, *146*, 184-192.