Virp: neural network-accelerated prediction of physical properties in site-disordered materials

Andy Paul Chen*[a], **Martin Hoffmann Petersen**[b], **Kedar Hippalgaonkar**[a]

[a] *School of Materials Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Republic of Singapore* andypaul.chen@ntu.edu.sg,kedar@ntu.edu.sg

[b] *Department of Energy Storage and Conversion, Technical University of Denmark, Anker Engelunds Vej 301, DK-2800 Kongens Lyngby, Denmark* mahpe@dtu.dk

* Presenting author

1. Introduction

Site-disordered materials are defined by crystal structures in which at least one crystallographic site is partially occupied by atoms of different elements. This category is diverse and encompasses materials from metal alloys such as CoFe[1], ordered vacancy compounds such as $CuIn_3Se_5$[2], and correlated disorder materials such as water ice[3]. Stoichiometric tuning and doping in materials synthesis also work by substituting one element for another at specific crystallographic sites, making many synthetic compounds site-disordered[4]. Point defects are a particular expression of site disorder. Naturally occurring minerals are predominantly determined to exhibit site disorder. As such, site-disordered materials are ubiquitous in materials science.

Table 1: Prevalence of site-disordered materials (SDM) in experimental databases.

	Ordered	**SDM**	Error
ICSD[5]	122517 (53.4%)	**106970 (46.6%)**	4966 (2.1%)
AMCSD[6]	10655 (50.6%)	**9153 (43.5%)**	1246 (5.9%)

First-principles simulation methods, especially density functional theory (DFT), have been instrumental in the past few decades in exploring the link between crystal structure and material properties. Large DFT computational databases have been compiled, including Materials Project (MP)[7] and the Open Quantum Materials Database (OQMD)[8]. Despite their abundance in real life, site-disordered materials are conspicuously absent from these databases. This originates from the inability of DFT software to treat atomic sites with site disorder.

Numerous strategies exist to bridge this gap between theory and experiment. Cluster expansion and special quasirandom structures (SQS)[9, 10] are used to simulate, as closely as possible, a random distribution of elements at the disordered sites in a quasirandom or virtual cell. The coherent potential approximation (CPA) method [11] is used to simulate an effective medium potential created by the mixture of elements in a disordered system. The applicability of these methods is limited to simple disordered materials, especially metal alloys, and SQS tends to be computationally expensive. In 2017, the software `Supercell`[12] was the first of its kind developed that can generate quickly a large number of virtual supercells for materials that do not have correlated disorder. More recently, in 2023, `aflow++`[13] has emerged, using a batch of virtual supercells and DFT to predict the physical properties of a site-disordered material using Boltzmann averaging.

At this juncture, we are far from solving the problem of computational expense. The presence of site disorder turns one simple computational routine into many heavy routines, depending on the supercell size and the size of the batch. In this work, we present the `virp` code, which employs current neural network-derived methods and sampling theory to circumvent computational complexity issues in materials property prediction with site disorder.

2. Substantial section

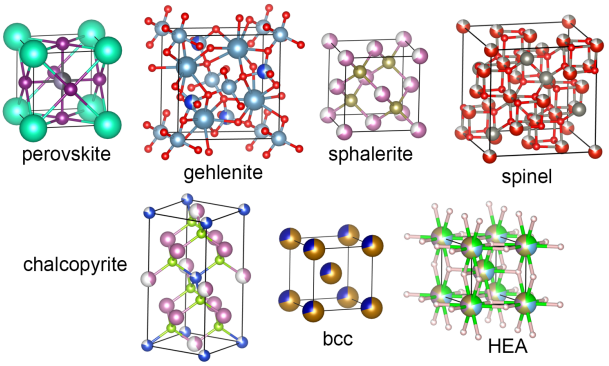

Fig. 1: Cells used in trial demonstrations

2.1 Permutative fill and enumeration

Database building operations are performed on a set of trial cells with site disorder. For each trial cell, a set of up to 700 virtual cells are generated, structurally optimised using `CHGNET`, and assigned a predicted band gap using `matgl`[14]. Each batch operation can be completed in the space of one week.

When `virp` treats a site-disordered unit cell, it first creates a supercell by replicating the unit cell a

number of times along the crystallographic axes; the number chosen should be large enough to minimise periodic boundary effects.

Following the method of the `Supercell` code, the instances of each disordered site in the supercell is randomly assigned an atom (or lack thereof) according to the proportional occupancy of the elements in the site as specified in the crystallographic data. The continuous cumulative proportions are mapped onto a discrete array (a "snap") by rounding. Each atomic species is guaranteed at least one site, and the anti-biasing feature maps a cumulative proportion by rounding up or down with equal probabilities if it lies exactly in the middle of two integers in the snap. The total number of distinct virtual cells N_v, discounting symmetry operations, is thus:

$$N_v = \prod_s \sum_{\mathbf{x}} \frac{m(s)!}{\prod_i x_i!}. \tag{1}$$

Here, s denotes a crystallographic site, $m(s)$ is the multiplicity of the site in the supercell, and $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{N}^n$ denotes a snap.

2.2 *Boltzmann averaging and sampling*

Similar to `aflow++`, the Boltzmann-averaged expectation value $\langle P \rangle$ of a certain property P from the calculated or predicted values p_i of each virtual cell (of energy E_i) in the sample set.

$$\langle P \rangle = \frac{\sum_i p_i e^{\frac{-E_i}{k_B T}}}{\sum_j e^{\frac{-E_j}{k_B T}}} \tag{2}$$

The Yamane sampling regime is recommended for when the target quantity is continuous[15]. Despite its population size dependence, the sample size levels off for larger populations. As such, a sample size of 400 is sufficient to maintain a margin of error of under 5% for Boltzmann-averaged quantities. Contrary to what Ohkotnikov et al.[12] and Oses et al.[13] may suggest, a complete sampling of the configuration space is not necessary.

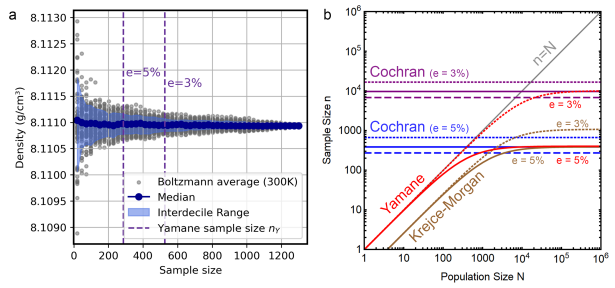

Fig. 2: (a) Boltzmann-averaged density of $Co_{0.3}Fe_{0.7}$ (bcc) based on sampling a population of 1300 virtual cells; (b) Yamane, Cochran, and Krejce-Morgan sample sizes ($p = 0.5$) against population size. For comparison, the measured density of $Co_{0.3}Fe_{0.7}$ is 8.017 g/cm^3[16].

2.3 *Symmetrical equivalence of virtual cells*

Accurate approximation of disordered structure require large supercells. This also means that the probability that two generated virtual cells are symmetrically equivalent is small, at 0-6%. Redundant cells can be identified by the CHGNET total energies of their un-relaxed structures without the need for symmetry resolution, which is computationally expensive. This is distinct from the approach of `Supercell`, which requires symmetry resolution and practically limits the size of the supercell one can choose.

Table 2: Size of configuration space (N) and redundancy in a set of 700 generated virtual structures from the trial set.

	N	Repeat	(%)
Perovskite	1.9×10^{15}	41	5.9%
Gehlenite	6.4×10^{27}	20	2.9%
Sphalerite	5.8×10^{28}	2	0.29%
Spinel	$\gg 10^{308}$	6	0.86%
Chalcopyrite	3.2×10^{16}	18	2.6%
bcc	5.0×10^{32}	34	4.9%
HEA	$\gg 10^{308}$	22	3.1%

2.4 *Implementation and code availability*

The `virp` program is available as a package on the PyPI repository. The source code can also be accessed on GitHub (`https://github.com/andypaulchen/virp`). Results in this abstract are generated by version `v1.2.1`.

Acknowledgments

We acknowledge Savyasanchi Aggarwal for his contribution in our fruitful discussions. APC and KH acknowledge support from Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1, Sponsor Award ID RG138/23. Calculations are performed on the `Khompute` server in the School of Materials Science and Engineering, Nanyang Technological University, with generous assistance from Nong Wei.

References

[1] Shalabh Srivastava, Andy Paul Chen, Tanmay Dutta, Rajagopalan Ramaswamy, Jaesung Son, Mohammad S. M. Saifullah, Kazutaka Yamane, Kangho Lee, Kie-Leong Teo, Yuan Ping Feng, and Hyunsoo Yang. Effect of $(Co_xFe_{1-x})_{80}B_{20}$ Composition on the Magnetic Properties of the Free Layer in Double-Barrier Magnetic Tunnel Junctions. *Physical Review Applied*, 10(2):024031, August 2018. Publisher: American Physical Society.

[2] Takeshi Hanada, Azusa Yamana, Yoshio Nakamura, Osamu Nittono, and Takahiro Wada.

Crystal Structure of $CuIn_3Se_5$ Semiconductor Studied Using Electron and X-ray Diffractions. *Japanese Journal of Applied Physics*, 36(Part 2, No. 11B):L1494–L1497, November 1997.

[3] David A. Keen and Andrew L. Goodwin. The crystallography of correlated disorder. *Nature*, 521(7552):303–309, May 2015.

[4] Rohit Prasanna, Aryeh Gold-Parker, Tomas Leijtens, Bert Conings, Aslihan Babayigit, Hans-Gerd Boyen, Michael F. Toney, and Michael D. McGehee. Band Gap Tuning via Lattice Contraction and Octahedral Tilting in Perovskite Materials for Photovoltaics. *Journal of the American Chemical Society*, 139(32):11117–11124, August 2017.

[5] G Bergerhoff, R Hundt, R Sievers, and I D Brown. The Inorganic Crystal Structure Data Base. *Journal of Chemical Information and Computer Sciences*, 23(2):66–69, 1983. Publisher: ACS Publications.

[6] Robert T Downs and Michelle Hall-Wallace. The American Mineralogist crystal structure database. *American Mineralogist*, 88:247–250, 2003.

[7] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, July 2013.

[8] Scott Kirklin, James E. Saal, Bryce Meredig, Alex Thompson, Jeff W. Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1(September), 2015. Publisher: Nature Publishing Group.

[9] Dominik Gehringer, Martin Friák, and David Holec. Models of configurationally-complex alloys made simple. *Computer Physics Communications*, 286:108664, May 2023.

[10] Mattias Ångqvist, William A. Muñoz, J. Magnus Rahm, Erik Fransson, Céline Durniak, Piotr Rozyczko, Thomas H. Rod, and Paul Erhart. ICET – A Python Library for Constructing and Sampling Alloy Cluster Expansions. *Advanced Theory and Simulations*, 2(7):1900015, July 2019.

[11] Levente Vitos. The EMTO-CPA Method. In *Computational Quantum Mechanics for Materials Engineers*, pages 83–94. Springer London, 2007.

[12] Kirill Okhotnikov, Thibault Charpentier, and Sylvian Cadars. Supercell program: a combinatorial structure-generation approach for the local-level modeling of atomic substitutions and partial occupancies in crystals. *Journal of Cheminformatics*, 8(1):17, December 2016.

[13] Corey Oses, Marco Esters, David Hicks, Simon Divilov, Hagen Eckert, Rico Friedrich, Michael J. Mehl, Andriy Smolyanyuk, Xiomara Campilongo, Axel Van De Walle, Jan Schroers, A. Gilad Kusne, Ichiro Takeuchi, Eva Zurek, Marco Buongiorno Nardelli, Marco Fornari, Yoav Lederer, Ohad Levy, Cormac Toher, and Stefano Curtarolo. aflow++: A C++ framework for autonomous materials design. *Computational Materials Science*, 217:111889, January 2023.

[14] Chi Chen, Yunxing Zuo, Weike Ye, Xiangguo Li, and Shyue Ping Ong. Learning properties of ordered and disordered materials from multi-fidelity data. *Nature Computational Science*, 1(1):46–53, January 2021.

[15] Taro Yamane. *Statistics: An Introductory Analysis*. A Harper International Edition. Harper & Row, 1967.

[16] M. Hocine, A. Guittoum, M. Hemmous, D. Martínez-Blanco, P. Gorria, B. Rahal, J. A. Blanco, J. J. Sunol, and A. Laggoun. The role of silicon on the microstructure and magnetic behaviour of nanostructured $(Fe_{0.7}Co_{0.3})_{100-x}Si_x$ powders. *Journal of Magnetism and Magnetic Materials*, 422:149–156, 2017.