# Atomic Sudoku: Stochastic approaches for correlated disorder materials

**Leck Boon Keng** [1] **Andy Paul Chen** [1] **Kedar Hippalgaonkar** [1]

[1] *School of Materials Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Republic of Singapore*. Correspondence to: Andy Paul Chen andypaul.chen@ntu.edu.sg, Kedar Hippalgaonkar kedar@ntu.edu.sg.

## 1. Introduction

Solid materials are identified by their crystalline nature: barring grain boundaries or other defects, the atoms are arranged in a regular manner in relation to one another. As such, even with an indefinite number of atoms, the overall structure can be described by a manageable set of symmetry operations. This is in contrast to the disorder inherent in amorphous materials such as glass. However, crystalline solids can exhibit site-disorder, where one or more Wyckoff sites are not represented by a single element, but a mix of different elements or even a lack thereof. Site-disorder occurs in about half of all documented materials in experimental databases such as the Crystallography Open Database[1] and the Inorganic Crystal Structure Database[2]. We can consider site disorder in their varying levels of complexity and intractability, the sources of which we list as follows:

1. multiple disordered Wyckoff sites, as seen in certain high-entropy alloys[3] or complex ionic materials[4];
2. short-range order[5], where an atom has a preference in which atomic species occupy its immediate environment;
3. correlated disorder, where the resolution of one site can affect the filling of sites further away, or even throughout the cell[6].

SQS implementations such as AFLOW-CCE[7] are tailored to short-range order, but are too computationally expensive to generate larger supercells. Random filling algorithms such as `Supercell`[8] are flexible enough to generate large virtual supercells for multiple disordered Wyckoff sites, but they ignore correlation between disordered sites. Algorithms can be developed to generate supercells for individual correlated disorder systems, for example GenIce[9], but these approaches are built on an *a priori* implementation of site-filling rules which precludes application to more than one material system, in this case water ice. To facilitate high-throughput analyses on large datasets, it might be necessary to develop a generalised approach – large supercells, correlated sites, *and* applicable broadly across most (if not all) materials systems.

While widespread, site disorder crystal structures are underrepresented in computational material databases such as Materials Project[10] and the Open Quantum Materials Database[11]. This is due to the additional complexities of having to represent a unit cell with a virtual supercell, as well as the large number of atoms present in the supercell. When these databases are used to train generative models and machine-learned interatomic potentials, the contribution of half of all experimentally observed materials could be ignored, leading to biased and underperforming models. Current efforts such as Dis-GEN[12] focus on generative design procedures and aim to predict potential materials including site-disordered ones. Thus, developing a computationally efficient pipeline that lends site disordered materials amenable to first principles calculations is of particular importance and urgency for our time.

## 2. Substantial section

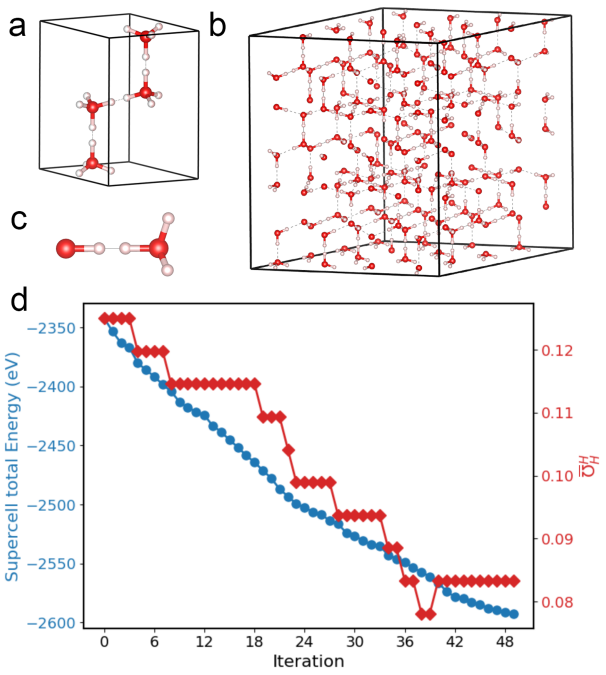


Fig. 1: Implementation of a simple stochastic procedure for refining a virtual supercell for water ice: a) The unit cell of hexagonal water ice $I_h$, showing half-occupied H sites; b) a virtual supercell generated by random filling of H sites; c) pathological case where two adjacent H sites are both occupied; d) after 50 iterations of energy-directed site-shuffling, the occurrence of H-H coordination (indicated by $\overline{\upsilon_H^H}$) also decreases.

A valid virtual supercell for a correlated disorder material should reflect both the original stoichiometry of the source unit cell (Fig. 1 a) and the underlying logic governing site filling. This logic is not encoded

in the structural data. As such, random filling in a water ice supercell (Fig. 1 b) can include many local features where the filling logic is violated (Fig. 1 c).

We approach the probability of valid site filling in water ice from the point of view of local coordination of atoms. First, we define the coordination number of an atomic species $A$ to a certain instance $B$ of a Wyckoff site as $\mho_B^A$. The average coordination number $\overline{\mho_B^A}$ can be evaluated across the virtual supercell.

We understand that filling rules in water ice forbids the filling of two adjacent H sites at the same time. Refining the virtual cell thus means bringing its $\overline{\mho_H^H}$ to 0. In our example case, we attempt to achieve this using an iterative method, in which for each iterative step, 10 H-site swaps are performed, the energy of the result cells evaluated with the MACE-MPA potential, and the lowest-energy result fed into the next iteration. Fig. 1 d shows that cell energy and $\overline{\mho_H^H}$ decrease in tandem, demonstrating that machine-trained interatomic potential MACE-MPA[13] can be suitable for aiding virtual cell refinement. In addition, more sophisticated stochastic approaches, such as those based on Metropolis Monte-Carlo[14], are the subject of our ongoing studies.

**Acknowledgments**

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. We acknowledge Pjotrs Žguns for his contributions in our fruitful conversations.

**References**

[1] Saulius Gražulis, Daniel Chateigner, Robert T. Downs, A. F. T. Yokochi, Miguel Quirós, Luca Lutterotti, Elena Manakova, Justas Butkus, Peter Moeck, and Armel Le Bail. Crystallography Open Database – an open-access collection of crystal structures. *Journal of Applied Crystallography*, 42(4):726–729, August 2009.

[2] 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.

[3] Karoline Stolze, Jing Tao, Fabian O. Von Rohr, Tai Kong, and Robert J. Cava. Sc–Zr–Nb–Rh–Pd and Sc–Zr–Nb–Ta–Rh–Pd High-Entropy Alloy Superconductors on a CsCl-Type Lattice. *Chemistry of Materials*, 30(3):906–914, February 2018.

[4] Andy Paul Chen, Wei Nong, Maung Thway, Jose Recatala-Gomez, Haiwen Dai, Wenhao Zhai, D. V. Maheswar Repaka, and Kedar Hippalgaonkar. Augmented chalcopyrites: A search for new Cu-In-Te phases. *Physical Review Materials*, 8(8):083801, August 2024.

[5] Killian Sheriff, Yifan Cao, Tess Smidt, and Rodrigo Freitas. Quantifying chemical short-range order in metallic alloys. *Proceedings of the National Academy of Sciences*, 121(25):e2322962121, June 2024.

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

[7] Rico Friedrich and Stefano Curtarolo. AFLOW-CCE for the thermodynamics of ionic materials, October 2023. arXiv:2310.18187 [cond-mat].

[8] 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.

[9] Masakazu Matsumoto, Takuma Yagasaki, and Hideki Tanaka. GenIce: Hydrogen-Disordered Ice Generator. *Journal of Computational Chemistry*, 39(1):61–64, January 2018.

[10] 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.

[11] 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):15010, 2015.

[12] Martin Hoffmann Petersen, Ruiming Zhu, Haiwen Dai, Savyasanchi Aggarwal, Nong Wei, Andy Paul Chen, Arghya Bhowmik, Juan Maria Garcia Lastra, and Kedar Hippalgaonkar. Dis-GEN: Disordered crystal structure generation, July 2025. arXiv:2507.18275 [cond-mat].

[13] Ilyes Batatia, Dávid Péter Kovács, Gregor N C Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 11423–11436. Curran Associates, Inc., New Orleans, 2022.

[14] Pjotrs A. Žguns, Andrei V. Ruban, and Natalia V. Skorodumova. Ordering and phase separation in Gd-doped ceria: a combined DFT, cluster expansion and Monte Carlo study. *Phys. Chem. Chem. Phys.*, 19(39):26606–26620, 2017.