A Noise-Robust Data Assimilation Method for Crystal Structure Prediction Using Powder Diffraction Intensity

Crystal structure determination is one of the most fundamental problems in materials science. The theoretical structure prediction is the multidimentional global optimization problem of the potential energy. Because of the high computational costs of the potential energy and the complexity of the multidimensional search space, various methods to systematically and efficiently find stable structures have been researched and developed. In the previous researches[1, 2] and this study, we are developing the structure prediction method to incorporate the experimental data into the theoretical structure search based on data assimilation.

In this method, we define the "difference" between the experimental data and the calculated value for the structure as the penalty function. Instead of the potential energy, we minimize the cost function F defined as follows,

$$F(R; I_{\mathrm{ref}}) = E(R) + \alpha N D[I_{\mathrm{ref}}, I_{\mathrm{calc}}(R)], \tag{1}$$

where R is the crystal structure, I_{ref} is the reference experimental data, I_{calc} is the calculated one, E is the potential energy, D is the penalty function, N is the number of atoms, and α is the control parameter. That is, we solve the multi-objective optimization problem for the potential energy and the penalty function.

Our method is explained with a schematic diagram as shown in Fig. 1. The correct structure is ideally the global minimum of both the potential energy and the penalty function. Since the local minima of them are different, the correct structure is more emphasized in the cost function than in the potential energy. Therefore, by optimizing the cost function, we can find the correct structure faster than by optimizing only the potential energy. The control parameter α adjusts how strongly the search space is restricted by the penalty function to match the experimental data.
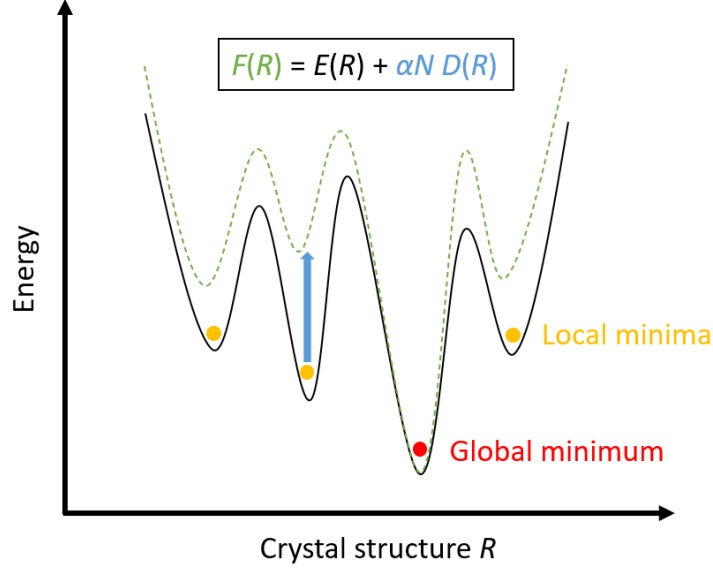
Fig. 1 Schematic diagram of the cost function. The horizontal axis represents the crystal structure R, which is actually a multidimensional space, such as cell parameters and atomic coordinates. The vertical axis represents the energy, which is optimized in structure search. The black line represents the potential energy E, and the red and yellow circles are its global minimum and local minima, respectively. The blue arrow represents the penalty function $\alpha N D$ added to the potential energy, and the green dotted line represents the cost function F.

In the previous research[1], they use the powder X-ray diffraction (XRD) pattern as the reference experimental data, and introduce "crystallinity"-type penalty function defined as follows,

$$D(R) = 1 - \frac{\int_{\theta=\theta_{\mathrm{obs}}} I_{\mathrm{calc}}(R)\mathrm{d}\theta}{\int_{\theta_{\mathrm{min}}}^{\theta_{\mathrm{max}}} I_{\mathrm{calc}}(R)\mathrm{d}\theta}, \tag{2}$$

where I_{calc} is the calculated diffraction intensity, θ is the diffraction angle, $[\theta_{\mathrm{min}}, \theta_{\mathrm{max}}]$ is the reference angle range, and θ_{obs} is the peak position observed in the reference experimental diffraction pattern I_{ref}. This penalty function does not depend on the experimental intensity information, and represents the coincidence ratio of peak positions between experimental and calculated diffraction patterns. By minimizing this penalty function, we can restrict the search space to satisfy the same extinction rule as in the experiment. Due to experimental errors such as the effect of preferred orientation, the peak intensity ratio in experiments deviates from the ideal calculated one. This penalty function works well even with unreliable experimental intensity ratio. On the other hand, it does not work well when the target structure has low symmetry without the extinction law, or when it is difficult to determine peak positions due to noise.

In this study, in order to improve the success rate and robustness of our method, we introduce a new penalty function of the correlation coefficient between I_{calc} and I_{ref} as follows,

$$D(R) = 1 - \frac{\int_{\theta_{\mathrm{min}}}^{\theta_{\mathrm{max}}} (I_{\mathrm{calc}}(R) - \overline{I}_{\mathrm{calc}}(R))(I_{\mathrm{ref}} - \overline{I}_{\mathrm{ref}})\mathrm{d}\theta}{\sqrt{\int_{\theta_{\mathrm{min}}}^{\theta_{\mathrm{max}}} (I_{\mathrm{calc}}(R) - \overline{I}_{\mathrm{calc}}(R))^2\mathrm{d}\theta}\sqrt{\int_{\theta_{\mathrm{min}}}^{\theta_{\mathrm{max}}} (I_{\mathrm{ref}} - \overline{I}_{\mathrm{ref}})^2\mathrm{d}\theta}}, \tag{3}$$

where $\overline{I}$ is the average intensity among the reference angle range $[\theta_{\min}, \theta_{\max}]$. The crystallinity-type penalty function emphasizes peaks not found in the experiment, whereas this one emphasizes large peaks. This penalty function explicitly includes the experimental intensity information, and restricts the search space more to match the experimental data. Therefore, it is expected that this penalty function can improve the search efficiency more. Instead, due to the difference between experimental and calculated peak intensities, this penalty function does not become zero for the correct structure. Even if the correct structure does not correspond to the global minimum of the penalty function, the penalty function method can support the optimization of the potential energy. We can find stable structures by performing the structure relaxation with only the potential energy after the optimization of the cost function.

In our code, we implement Simulated Annealing (SA)[3] using Molecular Dynamics (MD) for the optimization of the cost function. In test calculations of the structure search for known structures, we use model potentials as the potential energy in order to reduce the computational costs. In addition, instead of actual experimental data, we use the calculated XRD pattern for the correct structure as the reference data in the penalty functions (see Eqs. 2 and 3). We use VASP[4, 5] for first-principles calculations of the potential energy. We also use LAMMPS[6, 7] to perform model potential calculations.

For comparison of the correlation-coefficient-type and crystallinity-type penalty functions, we perform the structure prediction based on data assimilation for SiO_2 coesite. Coesite is suitable for this comparison due to its large primitive cell and relatively low symmetry. In these test clculations, the correlation-coefficient-type penalty function significantly improves the search efficiency over the crystallinity-type one. By adding artificial noise to the reference diffraction pattern, we confirm the noise-robustness of our method. The correlation-coefficient-type penalty function works well even if we use the reference data with noise as it is.

Since hydrogen contributes little to the XRD pattern, the penalty function of XRD hardly restricts hydrogen positions. We perform the structure prediction based on data assimilation for ϵ-$Zn(OH)_2$ in order to verify its effectiveness in the system containing hydrogen. In these test calculations, although the XRD pattern is nearly independent of hydrogen positions, the penalty function of XRD is effective to some extent. The penalty function restrict the positions of other atoms, and the potential energy determines hydrogen positions associated with them. If we have the reference neutron diffraction (ND) pattern detecting hydrogen positions, the hybrid penalty function of XRD and ND defined as the following formula is more effective,

$$D = \frac{1}{2}(D_{\mathrm{XRD}} + D_{\mathrm{ND}}). \tag{4}$$

The diffraction pattern cannot distinguish the atomic species corresponding to the peaks, and we can improve the search efficiency by assimilating the reference data of different probes at the same time.

Generally, in order to improve the search efficiency by our method, it is necessary to set an appropriate value of the control parameter α in Eq. 1. The penalty function restricts the search space to match the reference data, but narrows the diffusion range of atoms. By performing isothermal pre-simulations with different α, we estimate an appropriate range of α such that the penalty function is kept low and the atoms move sufficiently.

The heavier the atoms, the greater the contributions to the XRD pattern and the stronger the forces from the penalty function. For the system containing the atomic species with

significantly different weights from each other, if we set an appropriate α for heavier atoms to move sufficiently, the diffusion range of lighter atoms becomes too large. In such a case, we can improve the search efficiency by setting different bath temperatures for each atomic species so that the diffusion ranges are about the same.

As a practical application, we perform the structure search in Al-Ca-H system based on first-principles calculations and actual experimental XRD pattern. This is a collaborative research with Dr. Hiroyuki Saitoh (QST) and Prof. Shin-ichi Orimo (Tohoku Univ., WPI-AIMR/IMR). They discover a new hydride in high-pressure synthesized Al-Ca-H system, and propose us the structure prediction for it with their experimental data. This new hydride is a candidate hydrogen storage materials, and the composition ratio has not yet been determined. We search for Al-Ca structures that match the experimental data based on data assimilation using a model potential as the potential energy. Then, we add hydrogen atoms in a obtained Al-Ca structure, and perform the structure relaxation based on first-principles calculations. As a result, we discover new $Al_{12}Ca_{20}H_{76}$ structure that is almost as stable as known structures and somewhat consistent with the experiment.

Bibliography

[1] N. Tsujimoto, D. Adachi, R. Akashi, S. Todo, and S. Tsuneyuki. Crystal structure prediction supported by incomplete experimental data. *Physical Review Materials*, 2:053801, 2018.

[2] D. Adachi, N. Tsujimoto, R. Akashi, S. Todo, and S. Tsuneyuki. Search for common minima in joint optimization of multiple cost functions. *Computer Physics Communications*, 241:92, 2019.

[3] S. Kirkpatrick, C. D. Gelatt Jr, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671, 1983.

[4] G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B*, 354:11169, 1996.

[5] G. Kresse and D. Joubert. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical Review B*, 59:1758, 1999.

[6] S. Plimpton. Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*, 117:1, 1995.

[7] Lammps molecular dynamics simulator. `https://lammps.sandia.gov/`.