The prediction of material structure from chemical composition has been a long-standing challenge in natural science. Although there have been various methodological developments and successes with computer simulations[1–8], the prediction of crystal structures comprising more than several tens of atoms in the unit cell still remains difficult due to the many degrees of freedom, which increase exponentially with the number of atoms. Here we show that when some experimental data is available, even if it is totally insufficient for conventional structure analysis, it can be utilized to support and substantially accelerate structure simulation. In particular, we formulate a cost function based on a weighted sum of interatomic potential energies and a penalty function referred to as “crystallinity”, which is defined using limited X-ray diffraction data. This method is applied to well-known polymorphs of SiO$_2$ with up to 96 atoms in the simulation cell to find that it reproduces the correct structures efficiently with a very limited number of diffraction peaks. The penalty function is confirmed to destabilize the local minima of the potential energy surface, which facilitates finding the correct structure. This method opens a new avenue for determining and predicting structures that are difficult to determine by conventional methods, such as surface, interface, glass, and amorphous structures.

Since the invention of the X-ray diffraction method about a century ago, various diffraction techniques have been developed to determine the atomistic structure of objects, and they have served as essential tools over a broad range of research fields, such as solid state physics and biochemistry. However, structure determination becomes unreliable if the experimental diffraction data is incomplete, for example when the diffraction intensity is weak, the resolution is low, or there is a large amount of background noise. This weak point is sometimes a serious obstacle for modern materials science. The recent trend respects pursuit of novel functionalities of materials in nanoscale and/or under extreme condition such as high pressure and temperature. However, it is mostly impossible to obtain evident experimental data for such systems.

A way unaffected by these experimental complexities is theoretical structure prediction[9]. The modeled interatomic potential or first-principles electronic structure calculations based on density functional theory are used to calculate the total energy of the system $E$, as a function of atomic configuration. Using the chemical composition of the entire system as an input enables realistic and energetically stable structures to be derived through the optimization of $E$. Thanks to recent development of various efficient optimization methods such as random sampling[1], metadynamics[2], simulated annealing[3, 4], particle-swarm optimization[5, 6] and genetic algorithms[7, 8], theory-based methods are rapidly becoming practical, even obtaining predictive power. A highly successful case was a study on hydrogen sulfide under pressure. A high critical temperature ($T_c$) superconducting phase was predicted[10, 11] prior to the experimental observation[12], and the structure of the phase in a pressure cell was later confirmed by comparison of the experimentally observed and predicted diffraction data[13]. However, the applicability of such theoretical schemes is also severely limited by the complexity of the target material. As a system becomes more complex, the number of trivial minima in the multidimensional energy landscape grows significantly and the computational cost of reaching the relevant minima increases. A breakthrough is thus required to meet the needs in the modern materials science.

Here, we propose a new direction of theoretical development for the structure determination problem that paves the way to a systematic improvement in the ability to reach stable structures. The central idea is to utilize incomplete experimental data to improve the efficiency of energy optimization, i.e., joint optimization of the total energy and auxiliary functions implemented with experimental data is performed instead. Such an approach has been pursued before in the experimental community. Direct space methods[14] and data-assisted structure searches[15] have been used to optimize the atomic configuration so that experimental diffraction data is reproduced. There have been attempts to use an additional energy cost function to support efficient fitting; however, this has included the assumption that the observed data is sufficiently accurate[16–18]. Instead, we develop an inverse approach in which the theoretical optimization is supported by partial experimental data. The cost function is represented as:

$$F(R) = E(R) + \alpha N D[g(R), g_{\text{obs}}],$$  (1)

where $E$ is the potential energy calculated for atomic position $R$, $g$ denotes an observable function of $R$, $D$ denotes the penalty function that represents ‘distance’ or the metric functional, which becomes minimum when $g = g_{\text{obs}}$, $\alpha$ is a control parameter, and $N$ is the number of atoms. The correct structure that reproduces the experimental data remains stable, while other experimentally irrelevant structures become unstable. It thus becomes easy to predict the correct structure, as illustrated in Fig. 1. The value of $\alpha$ is tuned so that the energy barrier in $E$ is overcome (see Methods for details). Any of the optimization methods mentioned above can be applied to...
FIG. 1. Schematic image of method of utilizing available experimental data for crystal structure prediction. The aim of crystal structure prediction is to find the global minimum (orange sphere) of a potential energy surface $E(R)$ (blue line), where $R$ denotes atomic coordinates. Although accurate calculation of the potential energy is now available, finding the lowest energy is still difficult because the search space $R$ is high dimensional and the potential energy surface $E$ has many local minima (light blue sphere). To fill the trivial local minima, a penalty function $D$, defined with available experimental data is introduced. We expect that it is easier to find the global minimum of the cost function $F$ (red line), which is the sum of $E$ and $D$, than the optimization using only $E$.

locate the global minimum of the cost function $F$.

In this work, we demonstrate an efficient implementation of a structural search with $F$, assuming an example situation where observed X-ray diffraction data is incomplete. A part of the X-ray powder diffraction data is used to formulate the penalty function $D$:

$$D(R) = 1 - \lambda(R),$$  

(2)

where $\lambda$ denotes the “crystallinity” defined using experimental data for a specific crystal as

$$\lambda = \frac{\sum_{\theta_{\text{obs}}} \int_{\theta_{\text{obs}}-\Delta}^{\theta_{\text{obs}}} I_{\text{calc}}(\theta)d\theta}{\int I_{\text{calc}}(\theta)d\theta}.$$  

(3)

Here, $I_{\text{calc}}(\theta)$ denotes the intensity of the calculated diffraction pattern, $\theta_{\text{obs}}$ denotes the peak positions in the experimentally observed (referenced) diffraction pattern and $\Delta$ denotes the diffraction angle resolution. In the numerator, the intensities of calculated diffraction patterns, $I_{\text{calc}}$, are summed up only at the observed peak positions, $\theta_{\text{obs}}$. The important point is that the intensities of the observed diffraction peaks are not considered, but only the peak positions, which are insufficient to determine the structure by themselves, are used. Therefore, intensity modifications due to the texture of the sample, i.e., the preferred orientation effect, do not need to be corrected. The crystallinity $\lambda$ equals 1 when the structure is correct.

This method was applied to the well-known polymorphs of SiO$_2$. The potential energy surface for SiO$_2$ has numerous local minima; therefore, it is difficult to reach the experimentally observed crystal structures based only on knowledge about the chemical composition. Thus, it is a good material to demonstrate the effectiveness of the proposed method. Three polymorphs of SiO$_2$ were adopted as target structures: coesite, low cristobalite and low quartz. Si and O atoms are represented by blue and red spheres, respectively. Only information of peak positions from 20° to 45° (the green squares) is used to define the penalty function $D$.

FIG. 2. Crystal structures of three polymorphs of SiO$_2$ and the corresponding powder X-ray diffraction patterns. Crystallographic unit cells and diffraction patterns for a, coesite, b, low cristobalite, and c, low quartz. Si and O atoms are represented by blue and red spheres, respectively. Only information of peak positions from 20° to 45° (the green squares) is used to define the penalty function $D$. To compare the efficiency of the scheme with $F$ and $E$, a large number of structure search trials were conducted with random atomic configurations as inputs. A typical method
set to the melting temperature, which is plausible to overcome the energy minima. By changing the annealing temperature $T_0$, the success rate can be further improved (panel a, lower). The upward shift of the final energy distributions with increasing $\alpha$ indicates that escape from the energy minima is generally facilitated at the expense of the increase in the median of the energy. Panel d of Fig. 4 represents the trajectories of $E$ and $F$ during optimization with $F$ from an experimentally irrelevant metastable structure to the target structure. It is clear that the minimum for the $E$ surface corresponds to the middle of the slope for the $F$ surface, and the basin of energy minimum for $E$ is buried. Therefore, optimization with the $F$ surface easily yields the target structure. The additional penalty function, although formulated with very limited experimental data, thus makes it significantly easier to find the realistic structures.

We have demonstrated that utilizing incomplete diffraction data can significantly improve the efficiency of computational structure prediction. A structure that cannot be experimentally determined due to poor diffraction data can now be determined with this method. Any experimental data can be adopted with this method to formulate the penalty function $D$; therefore, a way is opened to determine the structure of not only crystals but also other difficult to determine targets, such as surfaces, interfaces, and glass.

This work was supported by Elements Strategy Initiative to Form Core Research Center in Japan, Japan Society for the Promotion of Science KAKENHI Grant No. 17H02930 and “Materials research by Information Integration” Initiative (MFI) project of the Support Program for Starting Up Innovation Hub from Japan Science and Technology Agency (JST). N.T. and D.A. were supported by the Japan Society for the Promotion of Science through the Program for Leading Graduate Schools (MERIT). The authors thank Y. Tanaka for providing (simulated annealing with potential energy $E$) failed to find the target structure out of 500 samples, which reflects the complexity of the SiO$_2$ potential surface. In contrast, with $F$, the target structures were correctly reached with high efficiency; 12, 109 and 148 samples out of 500 reached the global minimum for coesite, low cristobalite and low quartz, respectively. Fig. 3 shows the time evolution of “crystallinity” during the simulation for typical trials. When $F$ is formulated with the “crystallinity” from the diffraction data for coesite (panel a), the corresponding structure is correctly reproduced, as represented by the “crystallinity” (solid line) reaching 1.0. Similar results also apply when $F$ is defined by the “crystallinity” with the peak positions of low cristobalite and low quartz (panels b and c). Note that the “crystallinity” is kept far smaller than 1.0 by a typical simulation with $E$ (dashed lines), which indicates that the configuration $R$ is trapped around experimentally irrelevant energy minima.

Here we show that the structure of the hypersurface of the function $F$ is significantly modified from $E$ by increasing the mixing factor $\alpha$, which leads to a substantial increase in the probability of determining the correct structures. Fig. 4 summarizes the potential energies for the final configuration reached by the optimization of $F$ by simulated annealing at low temperature (500 K). None of the three target structures was reached within 2,500 (500) trials for coesite (low cristobalite and low quartz) by the optimization with $\frac{\Delta E}{\Delta T} = 0.0$ K (Fig. 4a–c, uppermost panels). In contrast, the number of successful trials increases with increasing $\alpha$ (Fig. 4a–c, red bars). Note that the success probability decreases when $\frac{\Delta E}{\Delta T}$ is too large, which corresponds to the neglect of the potential energy. The optimum value of the factor $\frac{\Delta E}{\Delta T}$ is in the order of the melting temperature, which is plausible to overcome the energy minima.

FIG. 3. Finding correct structure supported by incomplete X-ray diffraction patterns. a, The penalty function $D_{\text{coesite}}$ is defined using the diffraction data for coesite. The green (red, blue) solid line represents the time evolution of crystallinity for coesite (low cristobalite, low quartz) during optimization of the cost function. The green dashed line shows the time evolution of crystallinity for coesite during optimization using only the potential energy. b, c The penalty function is defined using the diffraction data for low cristobalite and low quartz. The parameter $\alpha$ is set to $\frac{\Delta E}{\Delta T} = 20,000$ K in all cases.
FIG. 4. $\alpha$ dependence of the potential energy $E$ distributions for optimized structures and change in $E$ and the cost function $F$ during successful optimization of $F$. a, The target material is the unit cell of coesite, which contains 48 atoms. 2,500 (500) samples were optimized for $4\alpha k_B = 0$ K (other values) by simulated annealing at low temperature (500 K). The most stable structure corresponds to the red bars. Simulated annealing at high temperature (20,000 K) was also performed (lower panel). b, The target material is a $2 \times 2 \times 1$ supercell of low cristobalite, which contains 48 atoms. c, The target material is a $2 \times 2 \times 1$ supercell of low quartz, which contains 36 atoms. d, Starting from the metastable structure shown in the upper left, the correct structure of coesite as shown in the upper right is found by optimization of the cost function, as depicted by the red line. The potential energy $E$ is extracted from the cost function $F$, as shown by the blue line. The potential barrier between the two structures, which makes it difficult to locate the global minimum, disappears with the penalty function. $4\alpha k_B = 20,000$ K in this simulation.

the programs for X-ray diffraction patterns analysis.

METHODS

Details of tuning the control parameter $\alpha$. The efficiency of this method depends on the value of $\alpha$. Here we estimate a suitable value. By adding the value of the penalty function corresponding to the melting temperature to the local minima of the energy surface, atoms are not trapped in local minima and can explore the energy surface to locate the global minimum. Therefore, taking the degree of freedom into account, the order of $\alpha$ is estimated as:

$$\alpha N \overline{D} \gtrsim \frac{3}{2} N k_B T_m,$$

where $N$ is the number of atoms in the simulation cell, $\overline{D}$ denotes the typical value of the penalty function $D$, $k_B$ is the Boltzmann constant and $T_m$ is the melting temperature of the system. The melting temperature for three target materials was roughly estimated as $T_m \sim 5000$ K by small-scale molecular dynamics simulations. Based on the value of “crystallinity”, of which the average is, for example, approximately $\frac{1}{3}$ for coesite, the typical value of $D$ can be estimated as $\frac{2}{3}$. Thus, for the target systems in this letter a suitable $\alpha$ can be roughly estimated as

$$\frac{4\alpha}{9k_B} \gtrsim T_m.$$  

(5)

Details of potential energy calculation. The potential energy and the force are calculated using the LAMMPS[19] package. To calculate the potential energy $E$, a pairwise
Diffraction data is the first 4,500 steps and then quenched to 0 K for the latter 700 steps. The conventional cost steps. MD visualization was created using VMD software[24]. To obtain the calculation results shown in Fig. 4, the temperature was set at 500 K for the first 4,500 steps and then quenched to 0 K for the latter 700 steps. MD visualization was created using VMD software[24].

Details of techniques for optimization. Simulated annealing for molecular dynamics was used for global optimization. The temperature of the system was controlled by the Nosé-Hoover method[21–23]. The integration time step was 1 fs. For simplicity, the cell parameters were fixed, although they can be optimized simultaneously in the principle. To obtain the calculation results shown in Fig. 3, the temperature was initially set at 5,000 K and then quenched to 0 K in the latter 3,900 steps. To obtain the calculation results shown in Fig. 4, the temperature was set at 500 K for the first 4,500 steps and then quenched to 0 K for the latter 700 steps. MD visualization was created using VMD software[24].

Details of penalty function. The conventional cost function used in the structure determination from powder diffraction data is the $R$-factor[25]:

$$R = \frac{\sum_{\theta} |I_{\text{obs}}(\theta) - I_{\text{calc}}(\theta)|}{\sum_{\theta} |I_{\text{obs}}(\theta)|}.$$  \hspace{1cm} (6)

In contrast to the “crystallinity” used in our method (eq.3), in the $R$-factor, all information from observed diffraction patterns is taken into account, and it is thus sensitive to background noise.

In the simulation, the X-ray wavelength was set to 1.540593 Å, which corresponds to that for Cu Kα radiation. The atomic scattering factors given in international tables for structural parameters used in this work (Supplementary Table I.) are the same as those presented in the paper[20], in which the potential model was proposed.

### SUPPLEMENTARY INFORMATION

#### Structural Parameters

Structural parameters used in this work (Supplementary Table I.) are the same as those presented in the paper[20], in which the potential model was proposed.

---

1. Pickard, C. J. & Needs, R. J. *Ab initio* random structure searching. *J. Phys.: Condens. Matter* **23**, 053201 (2011).
2. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12562 (2002).
3. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by Simulated Annealing. *Science* **220**, 671 (1983).
4. Pannetier, J., Bassas-Alsina, J., Rodriguez-Carvajal, J. & Caignaert, V. Prediction of crystal structures from crystal chemistry rules by simulated annealing. *Nature* **346**, 343 (1990).
5. Wang, Y., Lv, J., Zhu, L. & Ma, Y. Crystal structure prediction via particle-swarm optimization. *Phys. Rev. B* **82**, 094116 (2010).
6. Wang, Y., Lv, J., Zhu, L. & Ma, Y. CALYPSO: A method for crystal structure prediction. *Comput. Phys. Commun.* **183**, 2063 (2012).
7. Oganov, A. R. & Glass, C. W. Crystal structure prediction using *ab initio* evolutionary techniques: Principles and applications. *J. Chem. Phys.* **124**, 244704 (2006).
8. Glass, C. W., Oganov, A. R. & Hansen, N. USPEX–Evolutionary crystal structure prediction. *Comput. Phys. Commun.* **175**, 713 (2006).
9. Woodley, S. M. & Catlow, R. Crystal structure prediction from first principles. *Nat. Mater.* **7**, 937 (2008).
10. Li, Y., Hao, J., Liu, H., Li, Y. & Ma, Y. The metallization and superconductivity of dense hydrogen sulfide. *J. Chem. Phys.* **140**, 174712 (2014).
11. Duan, D., Liu, Y., Tian, F., Li, D., Huang, X., Zhao, Z., Yu, H., Liu, B., Tian, W. & Cui, T. Pressure-induced metallization of dense (H$_2$S)$_2$H$_2$ with high-$T_c$ superconductivity. *Sci. Rep.* **4**, 06968 (2014).
12. Drozdov, A. P., Eremets, M. I., Troyan, I. A., Ksenofontov, V. & Shlylin, S. I. Conventional superconductivity at 203 kelvin at high pressures in the sulfur hydride system. *Nature* **525**, 73 (2015).
13. Einaga, M., Sakata, M., Ishikawa, T., Shimizu, K., Eremets, M. I., Drozdov, A. P., Troyan, I. A., Hirao, N. & Ohishi, Y. Crystal structure of the superconducting phase of sulfur hydride. *Nat. Phys.* **12**, 835 (2016).
14. Černý, R. & Favre-Nicolin, V. Direct space methods of structure determination from powder diffraction: principles, guidelines and perspectives. *Z. Kristallogr.* **222**, 105 (2007).
15. Gao, P., Tong, Q., Lv, J., Wang, Y. & Ma, Y. X-ray diffraction data-assisted structure searches. *Comput. Phys. Commun.* **213**, 40 (2017).
16. Putz, H., Schön, J. C. & Jansen, M. Combined method for *ab initio* structure solution from powder diffraction data. *J. Appl. Cryst.* **32**,
Lanning, O. J. et al. Definition of a ‘guiding function’ in global optimization: a hybrid approach combining energy and $R$-factor in structure solution from powder diffraction data. Chem. Phys. Lett. 317, 296 (2000).

Turner, G. W., Tedesco, E., Harris, K. D. M., Johnston, R. L. & Kariuki, B. M. A method for understanding characteristics of multi-dimensional hypersurfaces, illustrated by energy and powder profile $R$-factor hypersurfaces for molecular crystals. Z. Kristallogr. 216, 187 (2001).

Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. J. Comput. Phys. 117, 1 (1995).

Tsuneyuki, S., Tsukada, M., Aoki, H. & Matsui, Y. First-Principles Interatomic Potential of Silica Applied to Molecular Dynamics. Phys. Rev. Lett. 61, 869 (1988).

Nosé, S. A molecular dynamics method for simulations in the canonical ensemble. Mol. Phys. 52, 255 (1984).

Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. J. Chem. Phys. 81, 511 (1984).

Hooover, W. G. Canonical dynamics: Equilibrium phase-space distributions. Phys. Rev. A 31, 1695 (1985).

Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual Molecular Dynamics. J. Mol. Graph. 14, 33 (1996).

Harris, K. D. M. & Tremayne, M. Crystal Structure Determination from Powder Diffraction Data. Chem. Mater. 8, 2554 (1996).

Prince, E. International Tables for Crystallography, Vol. C. 1st online edn (International Union of Crystallography, 2006).

Levien, L. & Prewitt, C. T. High-pressure crystal structure and compressibility of coesite. Am. Mineral. 66, 324 (1981).

Peacor, D. R. High-temperature single-crystal study of the cristobalite inversion. Z. Kristallogr. 138, 274 (1973).

Levien, L., Prewitt, C. T. & Weidner, D. J. Structure and elastic properties of quartz at pressure. Am. Mineral. 65, 920 (1980).