

ARTICLE OPEN

Fine-grained optimization method for crystal structure prediction

Kei Terayama^{1,2,3}, Tomoki Yamashita^{4,5}, Tamio Oguchi^{4,5} and Koji Tsuda^{1,2,4}

Crystal structure prediction based on first-principles calculations is often achieved by applying relaxation to randomly generated initial structures. Relaxing a structure requires multiple optimization steps. It is time consuming to fully relax all the initial structures, but it is difficult to figure out which initial structure leads to the optimal solution in advance. In this paper, we propose a optimization method for crystal structure prediction, called Look Ahead based on Quadratic Approximation, that optimally assigns optimization steps to each candidate structure. It allows us to identify the most stable structure with a minimum number of total local optimization steps. Our simulations using known systems Si, NaCl, Y₂Co₁₇, Al₂O₃, and GaAs showed that the computational cost can be reduced significantly compared to random search. This method can be applied for controlling all kinds of local optimizations based on first-principles calculations to obtain best results under restricted computational resources.

npj Computational Materials (2018)4:32; doi:10.1038/s41524-018-0090-y

INTRODUCTION

In designing new crystalline materials, crystal structure prediction (CSP) for a given chemical composition is the most fundamental task. A large number of physical properties can be predicted using first-principles calculations with a given atomic configuration. However, CSP is quite a difficult problem due to the exponential increase of the number of potential energy minima with respect to the system size.¹ A great deal of effort has been devoted to overcoming this problem. To date, several searching algorithms in CSP have been successfully developed such as random search,^{2–4} simulated annealing,^{5,6} basin hopping,⁷ minima hopping,^{8,9} evolutionary algorithm (EA),^{10–12} particle-swarm optimization (PSO),^{13,14} and Bayesian optimization (BO).¹⁵ Ab initio random structure searching by Pickard and Needs is quite simple but still one of the most efficient approaches even today. EA and PSO are also popular and efficient algorithms as implemented in USPEX^{10–12} and CALYPSO.^{13,14}

Even now, CSP is quite a time-consuming problem: to find the most stable structure, in existing methods, local optimizations of all the structures are performed to evaluate their energies. First-principles density-functional-theory (DFT) codes such as VASP¹⁶ and QUANTUM ESPRESSO¹⁷ calculate the force of each atom for a given structure. Relaxation is performed with the forces by using gradient methods such as the quasi Newton or conjugate gradient methods. Figure 1a shows how to obtain optimized energies by relaxing all the generated structures until convergence to local minima in structure space. However, if the final relaxed energy of a structure can be predicted in the middle of its calculation, we can skip unpromising calculations by stopping their optimizations and reduce the total CSP calculation by giving preference to promising structures as shown in Fig. 1b.

In this paper, we propose a method to very roughly estimate the final energy during local optimization, and we show that the total computational cost can be reduced by controlling the local optimization step based on that estimated energy. Generally, it is a hard to predict the energy after relaxation during local optimization. We obtain a rough estimate of the final energy with a quadratic approximation from current energy and force, focusing on the fact that, not only the energy but also the force on each atom, can be calculated by DFT calculations. In the proposed method, called Look Ahead based on Quadratic Approximation (LAQA), we first generate a large number of candidate structures, then select promising structures based on the energy estimation method, and proceed to calculate them preferentially. When performing CSP, if the calculation is judged insufficient, we may not find a stable structure, and need to perform additional calculation for newly generated candidate structures. To deal with such a procedure, we also propose a method to increase the number of candidate structures gradually, called sequential LAQA (sLAQA).

To show the effectiveness of the proposed methods, we conducted CSP simulations, using 7 typical systems: Si (8 and 16 atoms in unit cell), NaCl (16 and 32 atoms), ferromagnetic Y₂Co₁₇ (19 atoms), Al₂O₃ (10 atoms), and GaAs (16 atoms). We randomly generated hundreds of candidate structures in these systems and calculated all the local optimization steps. We then investigated how much of a computational cost reduction can be obtained with our proposed methods, compared with the widely used random searching approach, which repeatedly generate random structures and performs full local optimizations. The total local optimization steps using LAQA are reduced by a factor ranging from 2.02 up to 21.4, compared to random searching.

¹Graduate school of Frontier Sciences, the University of Tokyo, 5-1-5, Kashiwa-no-ha, Kashiwa, Chiba, Japan; ²RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan; ³Graduate school of Medicine, Kyoto University, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-8507, Japan; ⁴Center for Materials research by Information Integration, Research and Services, Division of Materials Data and Integrated System, National Institute for Materials, Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan and ⁵Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan
Correspondence: Kei Terayama (terayama@cbms.ku-tokyo.ac.jp) or Koji Tsuda (tsuda@k.u-tokyo.ac.jp)

Received: 26 January 2018 Revised: 14 June 2018 Accepted: 24 June 2018

Published online: 10 July 2018

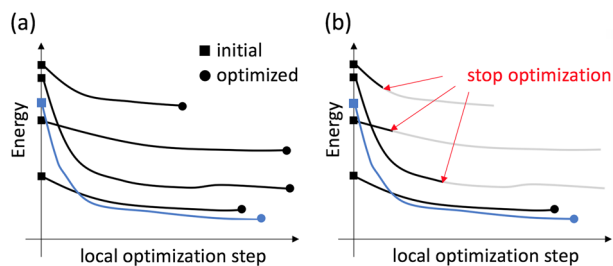


Fig. 1 Basic idea of acceleration of CSP by controlling local optimization steps. All the local optimizations are performed for generated structures in existing approaches **a**. We can reduce total local optimization steps by controlling each calculation of a structure based on the prediction of the finally relaxed energy of the structure **b**

RESULTS

Local optimization controlled by Greedy, LAQA, and sLAQA

In CSP, we compare the total energies evaluated by first-principles DFT calculations among a number of generated structures, and finally predict the minimum energy structure as the global minimum structure of the system. We propose approaches to efficiently perform CSP by preferentially relaxing structures that are predicted to have a low-final energy such as stable structures based on the basic idea shown in Fig. 1.

One of the simplest approaches of local optimization control is to progress relaxation in accordance with the order of energies. In many cases, if a structure is close to its stable state, its energy is considered to be relatively low. Therefore, if local optimization is done from a structure with low energy, it is expected that a stable structure will be obtained at a relatively early stage. We refer to this simplest idea based CSP method as greedy method (Greedy), as such an approach is called greedy algorithm in the field of reinforcement learning.¹⁸ In this method, we first generate a sufficient number of structures, N , that are expected to contain at least one structure that will relax to the most stable structure and then calculate the first local optimization step for them. We search a stable structure by performing local optimization in accordance with the order of their calculated energies.

However, the Greedy method does not always work well because an initial energy of a structure that will relax to a stable structure is not necessarily low as the blue local optimization step in Fig. 1a. On the other hand, during such a local optimization step from initial high to final low energy, it is expected that a relatively large force may be applied to each atom. This is because it is thought that each atom moves greatly until it is fully relaxed. It is expected that we can preferentially select structures that have final low energies, not only using just energy as Greedy, but also other information such as the above-mentioned force and stress applied to a cell.

In this paper, we propose a method (LAQA) to control local optimization based on the following score as the simplest and the most versatile method using energy and force. Figure 2a shows the flowchart of LAQA. First, we generate a sufficient number of structures, N , that are expected to contain at least one structure that will relax to the most stable structure. We calculate the first local optimization step for all the structures (Initialization). Next, we calculate rough estimates of the final energies to select a structure on which to proceed with local optimization (Scoring). For each structure i and local optimization step t , we denote by $E_{i,t}$ the total energy of the structure divided by formula unit (f.u.), and by $F_{i,t}$ the sum of forces on atoms divided by formula unit. $F_{i,t}$ is calculated by averaging the norms of the force applied to each atom. We calculate the score $L_{i,t}$ of each structure i after T steps as

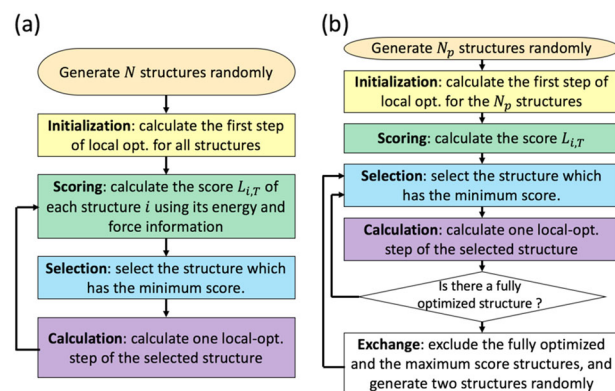


Fig. 2 Flowcharts of CSP using LAQA **a** and sLAQA **b**

follows:

$$L_{i,T} = \begin{cases} \min_{1 \leq t \leq T} E_{i,t} - \frac{F_{i,t}^2}{2\Delta F_{i,T}} & \text{(not optimized)} \\ \infty & \text{(fully optimized)} \end{cases} \quad (1)$$

where $\Delta F_{i,T} = |F_{i,T} - F_{i,T-1}|$. We fix $\Delta F_{i,T} = 1$ for $T = 1$ and $\Delta F_{i,T} = 10^{-6}$ if the values of $F_{i,T}$ and $\Delta F_{i,T-1}$ are the same. In our approach, we preferentially perform local optimization of structures with low values of the score. As can be seen from the first term in Eq. (1), we give priority to structures with low energy. The reason why it is taking minimum value of E is that a spike-like exceptionally large value is often obtained in the calculation process. According to the second term in Eq. (1), the total score decreases as the force applied to each atom increases. Additionally, if the force does not change significantly even if step changes, that is, if it can be thought that large structural change is continuing, the score decreases. For a structure that has been fully optimized, infinity is assigned as its score to avoid additional calculation for it. This score can be regarded as a rough estimation of the fully optimized energy of the system with quadratic approximation. According to this score, the structure which has the minimum score is selected (Selection) and calculation of one local optimization step is performed (Calculation). Note that fully optimized structures, having a score of $+\infty$, are not selected.

Since holding large amount of candidate structures simultaneously when performing LAQA may be costly, we also propose an algorithm to gradually increase the number of optimization structures. We call this approach sequential LAQA (sLAQA). Figure 2b shows the flowchart of sLAQA. We first fix the pooling number N_p and generate N_p structures randomly. The structure to optimize locally is chosen from the pooled structures. The initialization, scoring, selection, and calculation steps are the same as LAQA. After the calculation step, if the calculated structure is fully optimized, we exclude it. Without elimination of unpromising structures, the proportion they occupy in the pool would keep increasing, we therefore simultaneously exclude the structure with the highest score, i.e., the most unpromising structure. Then, we generate two new structures randomly to keep the number of pooled structures N_p , and calculate the first local optimization step for the new structures as initialization. Finally, we add them in the structure pool. In LAQA, if the set of initial structures is the same, the selected structure to calculate based on the L score is the same. Therefore, the number of steps required to obtain a stable structure does not change for the same set. On the other hand, in sLAQA, even if the set of initial structures is the same, since a structure to be added to the pool is newly chosen randomly, the number of steps required to obtain the stable structure changes.

For actual CSP of unknown systems, we need to stop calculation at a certain stage. It is difficult to identify the most stable structure,

but when multiple relaxed structures with the lowest energy are obtained, the structure is considered to be the most stable structure with a high probability. Therefore, in LAQA, it is considered to be practical to stop a search when structures above mentioned are obtained. If such structures are not obtained and final energies of fully relaxed structures become high, it is considered effective to generate a new dataset of initial structures and repeat the calculation using LAQA until such structures are found. In sLAQA, it is considered practical to stop calculation if such structures are obtained.

CSP using RS and BO

In order to examine the effectiveness of LAQA and sLAQA, we introduce other approaches of CSP: random search (RS)² and Bayesian optimization based structure selection.¹⁵

In the RS approach, initial structures are randomly generated and fully relaxed. Here, we first generate a number of initial structure candidates, and perform RS-based CSP by randomly selecting a structure to relax them. While RS is a simple and widely used method, it may take a long time to find a stable structure.

BO is widely used as one of the global optimization methods,¹⁹ and in recent years its usefulness is also shown in the field of materials science.²⁰ In the CSP using BO,¹⁵ a stable structure is searched by repeating a selection of an initial structure among candidate initial ones and relaxation of it. Unlike LAQA, sLAQA, and Greedy, a selected structure is fully optimized. In structure selection, a structure that is expected to have lower final energy is chosen by using the framework of BO based on structure data and their relaxed energies previously calculated. To perform the framework of BO, we adopted the fingerprint of Oganov and Valle²¹ as the descriptor of structures in the previous work.¹⁵ A fingerprint is a vector representation of a structure. The vector is calculated by a fingerprint function that is invariant with respect to shifts in the coordinate system, rotations, and reflections. The fingerprint of Oganov and Valle²¹ was designed to map similar structures to similar vectors. Compared with RS, it is expected to speed up the calculation by efficiently searching for low energy structures by BO-based search. In this paper, we repeated this CSP trials 400 times for each system, and calculated the performance of BO by calculating the average of structures to find a stable structure.

Initial structure generation and local optimization

We randomly generated initial structures with specific space groups, using CrySPY,²² as described by Yamashita et al.¹⁵ Once the space group is specified, some lattice parameters are fixed by the symmetry. The remaining unfixed lattice parameters are taken at random. A combination of the Wyckoff positions corresponding to the space group is randomly selected. The atoms are arranged according to the selected Wyckoff positions under the constraint of the minimum interatomic distance.

Total energy calculations and structure optimizations were carried out using DFT with the projector augmented wave method²³ as implemented in VASP code.¹⁶ The internal atomic coordinates as well as the cell parameters were fully optimized (see [Methods](#) section for the details of DFT calculation).

Tested systems

We performed test simulations of CSP for five typical systems: Si (8 and 16 atoms in unit cell), NaCl (16 and 32 atoms), Y₂Co₁₇ (19 atoms), Al₂O₃ (10 atoms), and GaAs (16 atoms). We denote these systems by Si₈, Si₁₆, Na₈Cl₈, Na₁₆Cl₁₆, Y₂Co₁₇, Al₄O₆, and Ga₈As₈. The most stable structures of Si, NaCl, Y₂Co₁₇, Al₂O₃, and GaAs are the rocksalt, diamond, Th₂Zn₁₇-type, corundum (α -Al₂O₃), and zinc blend structure, respectively. Y₂Co₁₇ is a ferromagnetic intermetallic compound and its space group is $R\bar{3}m$.²⁴ For GaAs, the

Table 1. Basic information for tested systems. The number in parentheses indicates the number of wurtzite structures for Ga₈As₈

System	Generated structures	Optimized to stable structures	Avg. of local opt. steps
Si ₈	500	7	68.91
Si ₁₆	700	3	92.81
Na ₈ Cl ₈	500	18	121.7
Na ₁₆ Cl ₁₆	500	17	177.7
Y ₂ Co ₁₇	700	1	130.6
Al ₄ O ₆	1000	2	83.49
Ga ₈ As ₈	1000	2 (2)	113.4

wurtzite structure whose energy is quite close to the one of the zinc blend structure is also taken into consideration as a stable structure.

We randomly prepared 500, 700, 500, 500, 700, 1000, and 1000 candidate structures for Si₈, Si₁₆, Na₈Cl₈, Na₁₆Cl₁₆, Y₂Co₁₇, Al₄O₆, and Ga₈As₈ as listed in Table 1 and evaluated their total energy. When generating the candidate structures, we imposed constraints on interatomic distance. We used a 1.8 Å distance for Si₁₆, Na₈Cl₈, Na₁₆Cl₁₆, Y₂Co₁₇, Al₄O₆, and Ga₈As₈ and 2.0 Å for Si₈. We compared generated initial structures by calculating the fingerprint.²¹ For each system, almost all of them are different each other. Although only two structures in Na₈Cl₈ are the same, we performed local optimization for them as different structures. Figure 3 shows the result of the locally optimized energies for all the generated initial structures of the Si₈ (a), Si₁₆ (b), Na₈Cl₈ (c), Na₁₆Cl₁₆ (d), Y₂Co₁₆ (e), Al₄O₆ (f), and Ga₈As₈ (g) systems. In each system, the differences of the optimized energies (eV/f.u.) from the energy of the most stable structure are plotted. For Al₄O₆, we show the enlarged result (eV/atom) in Fig. 3h. This result corresponds to the one of the black square region in Fig. 3f. The basic information for these systems is listed in Table 1. Eventually, we optimized 7, 3, 18, 17, 1, and 2 structures to stable structure, for Si₈, Si₁₆, Na₈Cl₈, Na₁₆Cl₁₆, Y₂Co₁₇, and Al₄O₆, respectively. For Ga₈As₈, two stable (zinc blend) structures and two wurtzite structures were obtained. We also collected the total energy and the force on atoms in each local optimization step. The average steps of local optimization for all structures are also listed in Table 1.

Simulation results

To show the effectiveness of the proposed method, we performed CSP simulations on the calculated dataset of the Si₈, Si₁₆, Na₈Cl₈, Na₁₆Cl₁₆, Y₂Co₁₇, Al₄O₆, and Ga₈As₈ systems. In the CSP trials, local optimization is controlled by different algorithms: random sampling (baseline method), LAQA, sLAQA, Greedy, and BO. Instead of actually generating the structure randomly in each CSP trial, we evaluated their performances by randomly choosing from prepared structures. Figure 4 illustrates the control and reduction of local optimization steps by using LAQA for the Y₂Co₁₇ system. Each line shows energy as a function of local optimization step from an initial structure. Bold black lines show the local optimization steps needed to reach the stable structure. Figure 4a shows all the local optimization steps of the Y₂Co₁₇ system. The number of total steps is 91,423. LAQA controls local optimization steps as shown in Fig. 4b for the same system. Stable structure is already optimized after only 3300 steps (~3.6%) and most unpromising structures are not optimized. Figure 4c, d are enlarged results of Fig. 4a, b.

Figure 5a shows the average number of total local-optimization steps required to find the stable structure with random search (blue), LAQA (green), sLAQA (orange, red, purple), Greedy (light

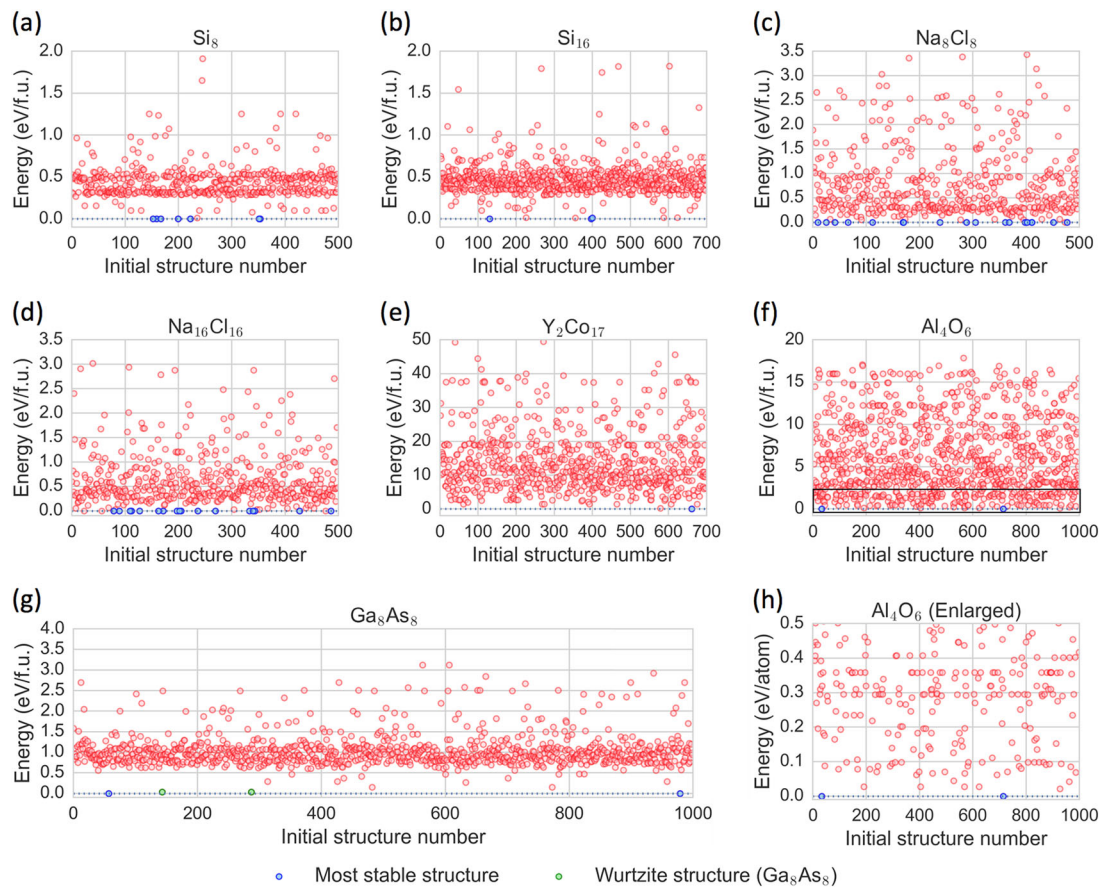


Fig. 3 Locally optimized total energies from randomly generated structures for the Si_8 **a**, Si_{16} **b**, Na_8Cl_8 **c**, $\text{Na}_{16}\text{Cl}_{16}$ **d**, Y_2Co_{17} **e**, Al_4O_6 **f**, and Ga_8As_8 **g** systems. In each system, the differences of the optimized energies (eV/f.u.) from the energy of the most stable structure are plotted. F.u. is the abbreviation for formula unit. For Al_4O_6 , we show the enlarged result (eV/atom) in **h**, where several states exist less than 0.1 eV/atom above the ground-state structure. This result corresponds to the one of the black square region in **f**. Red circles show total energies of fully optimized structures other than stable structures. Blue and green circles show total energies of stable and wurtzite structures for Ga_8As_8 , respectively. The number of stable structures and wurtzite structures for each system are listed in Table 1

blue), and BO (yellow). For Ga_8As_8 , we calculated performances by using these algorithms to find the stable structure (Ga_8As_8 (stable)), a wurtzite structure (Ga_8As_8 (wurtzite)) and at least one of both stable structures and wurtzite structures (Ga_8As_8 (both)). The values of total steps for LAQA and Greedy are calculated using all the prepared structures as generated ones. We calculated the average numbers for sLAQA by changing the number of pooling N_p . We estimated them by repeating CPS trials 2000 times and averaging their total local-optimization steps. To compare the performance of RS and BO with the others, we converted averaged number of structures to find a stable structure and a wurtzite structure into local optimization steps based on the average of local optimization steps for each system in Table 1. We also show the frequency distribution of number of trials required to find the most stable structure using random search, LAQA, and sLAQA in Fig. 5b. The result of random search obtained from the expected values in frequency theory is shown as blue circles. The total number of frequencies is normalized to 100. Green bars and orange histograms show the result of LAQA and sLAQA. The number of steps to find a stable structure in sLAQA is not constant. This is because the set of initial structures with a certain pool number changing in each trial and a structure that newly enters the pool is randomly chosen. The distributions of the result using sLAQA are biased to smaller numbers of trials and the number of local optimization steps can be reduced.

The total steps required to find the stable structure using LAQA were reduced by a factor of 4.23, 10.1, 3.79, 4.78, 14.0, 2.02, and

6.85 compared to the result of random search for Si_8 , Si_{16} , Na_8Cl_8 , $\text{Na}_{16}\text{Cl}_{16}$, Y_2Co_{17} , Al_4O_6 , and Ga_8As_8 , respectively. For Ga_8As_8 (wurtzite) and Ga_8As_8 (both), the total steps were reduced by a factor of 21.4 and 9.60. Since the proposed methods basically calculates preferentially structures expected to have lower final energies, not only stable structures but also wurtzite structures for Ga_8As_8 were also obtained at an early stage. The results using Greedy were almost the same as using LAQA for Si_{16} and Na_8Cl_8 , but were worse for Si_8 , $\text{Na}_{16}\text{Cl}_{16}$, Y_2Co_{17} , Al_4O_6 , and Ga_8As_8 . These differences result from the acceleration effect of local optimization for stable structures by using LAQA. Figure 6 shows the changes of fully optimized order of structures when using Greedy or LAQA. Each dot shows the orders of full optimization for an initial structure, i.e., how many structures were fully optimized before it, using these algorithms. Dark blue and green dots are stable and wurtzite structures. If a structure is plotted below the diagonal, it shows that optimization order is accelerated by using LAQA compared to Greedy. The blue dots in the green squares indicate stable structures whose local optimizations were preferentially made, using LAQA. The structure in the orange circle of Si_{16} was not preferentially relaxed, compared to Greedy. While the initial energy of this structure was relatively high, the force exerted on each atom was small, and as a result it was fully relaxed gradually with relatively long steps (166 steps). This is longer than the average one of Si_{16} (92.81 steps) and that of the structure in the black circle (44 steps). Therefore, the score of LAQA became relatively high for the structure of the orange circle, so its

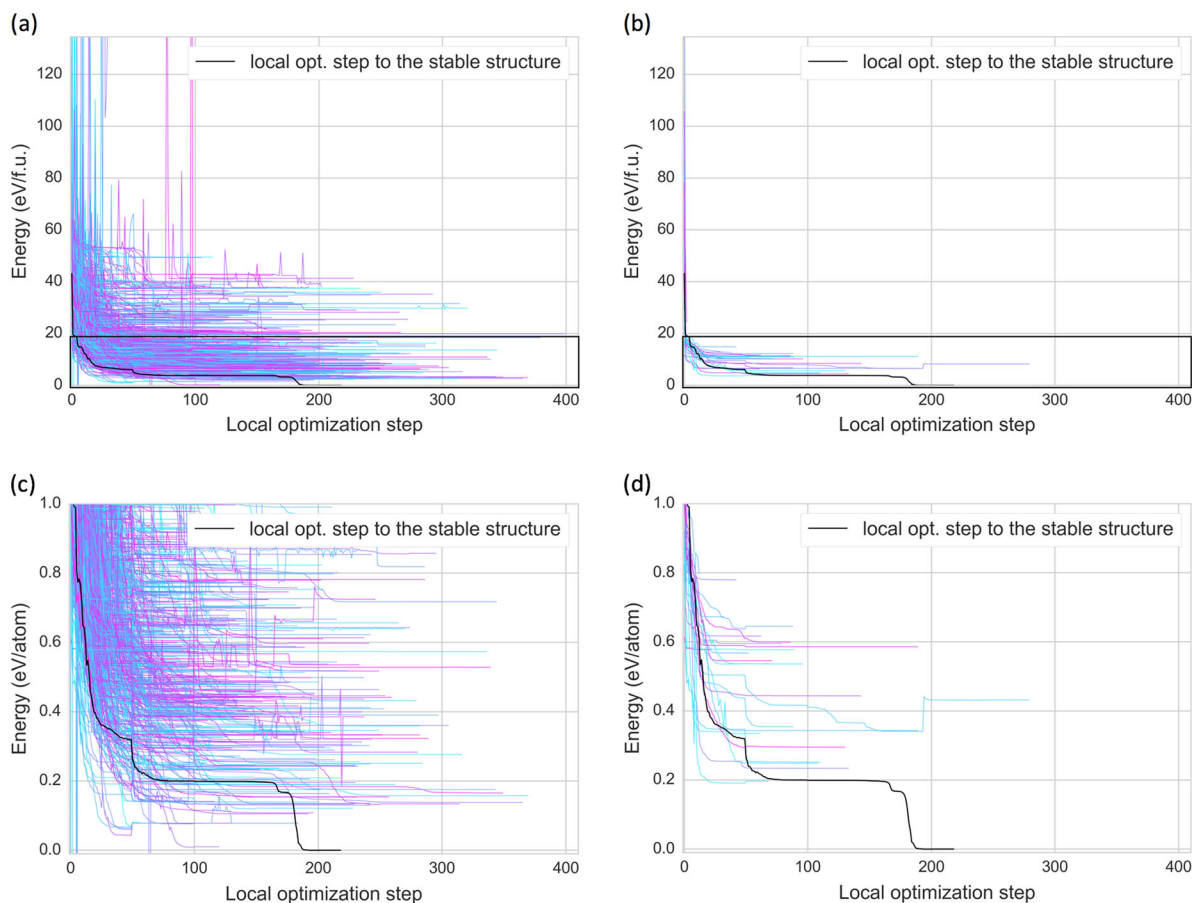


Fig. 4 Control of local optimization steps using LAQA for CSP. Each line shows energy as a function of local optimization step. Bold black lines show the local optimization steps to the stable structure. The other thin lines indicate local optimization steps that were not relaxed to the stable one. F.u. is the abbreviation for formula unit. **a** All local optimization steps for all initial structures (Y_2Co_{17}). The total number of steps is 91,423. **b** Controlled local optimization using LAQA. A structure is already optimized to the stable structure (bold black line) after 3300 steps (~3.6%). **c** and **d** are enlarged results of **a** and **b**. **c** and **d** correspond to the ones of the black squares in **a** and **b**, respectively

relaxation did not proceed preferentially. From these results, the total steps to find the stable structure and wurtzite structures for Si_8 , $Na_{16}Cl_{16}$, Y_2Co_{17} , Al_4O_6 , and Ga_8As_8 can be reduced by LAQA compared to Greedy. For Si_{16} and Na_8Cl_8 in Fig. 6, there are structures that are optimized immediately by Greedy and LAQA (structures in black circles). Since LAQA requires some exploratory calculations to find promising structures using the LAQA score, the results of Greedy were better than those of LAQA. However, even for these systems, the total steps using LAQA were reduced by a factor of 10.1 and 3.79 compared to those of random search. These results suggest the efficiency of the local optimization control and the rough estimation of fully optimized energy in LAQA.

DISCUSSION

For most of the systems, the average steps using sLAQA are decreased as the pooling number N_p is increased as shown in Fig. 5. This is because the larger N_p , the higher probability of including stable structures from the beginning of optimization. Note that even if the pooling number is increased, the performance of sLAQA is different from that of LAQA because sLAQA adds new structures. The performances of sLAQA for Si N_p are relatively poor. This result is considered to be derived from the facts that there were only two stable structures and the stable one in the orange circle in Fig. 6 was not optimized by using LAQA score

preferentially. Although performances depend on the systems, from these results, it was suggested that by using sufficient N_p in sLAQA, we can achieve performance far better than random search.

The results of LAQA and sLAQA showed better performance than those of BO in most cases, while the reduction effects of calculation using BO were not very large compared to RS. It is thought that the relatively low performance of BO is due to the problem that an initial structure and the structure after relaxation are different. In general, in BO based search, it is assumed that a point (structure) in a search space and the value corresponding to that point are in one-to-one correspondence. In CPS, however, there are two structures initial and relaxed ones. Although a calculated energy corresponds to a relaxed structure, we need to choose a structure to calculate next from among candidates of initial structures. This structural differences may adversely affect the performance of BO. On the other hand, our approach in this paper is completely different from BO, and the above-mentioned problem does not become a problem.

In the present study, we have proposed LAQA, and its variants: a novel approach for CSP that controls local optimization steps. Results on seven crystalline systems have demonstrated that LAQA and sLAQA can significantly reduce the total local optimization steps, compared to random search. In future work, we plan to combine our approach with effective structure generation or selection methods such as EA, PSO, or BO. We

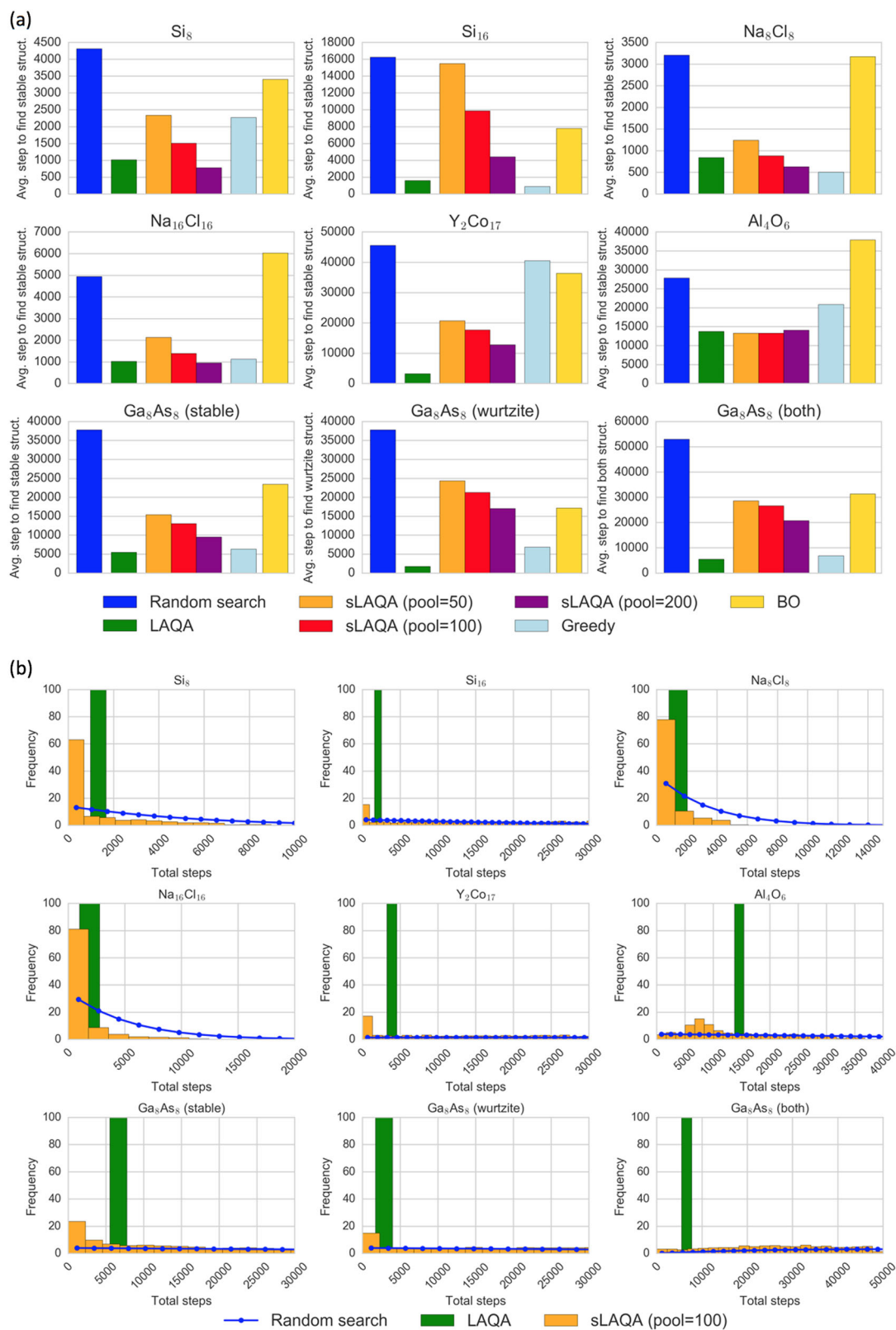


Fig. 5 Reduction result, from using LAQA. **a** Reduction of total local optimization steps required to find the most stable structure using random search, LAQA, sLAQA, Greedy, and BO. For Ga_8As_8 (wurtzite) and Ga_8As_8 (both), the average steps are shown to find a wurtzite structure and at least one of both stable and wurtzite structures, respectively. **b** Frequency distribution of number of trials required to find the stable structure and a wurtzite structure with random search, LAQA, and sLAQA for Si_8 , Si_{16} , Na_8Cl_8 , $\text{Na}_{16}\text{Cl}_{16}$, Y_2Co_{17} , Al_4O_6 , and Ga_8As_8 .

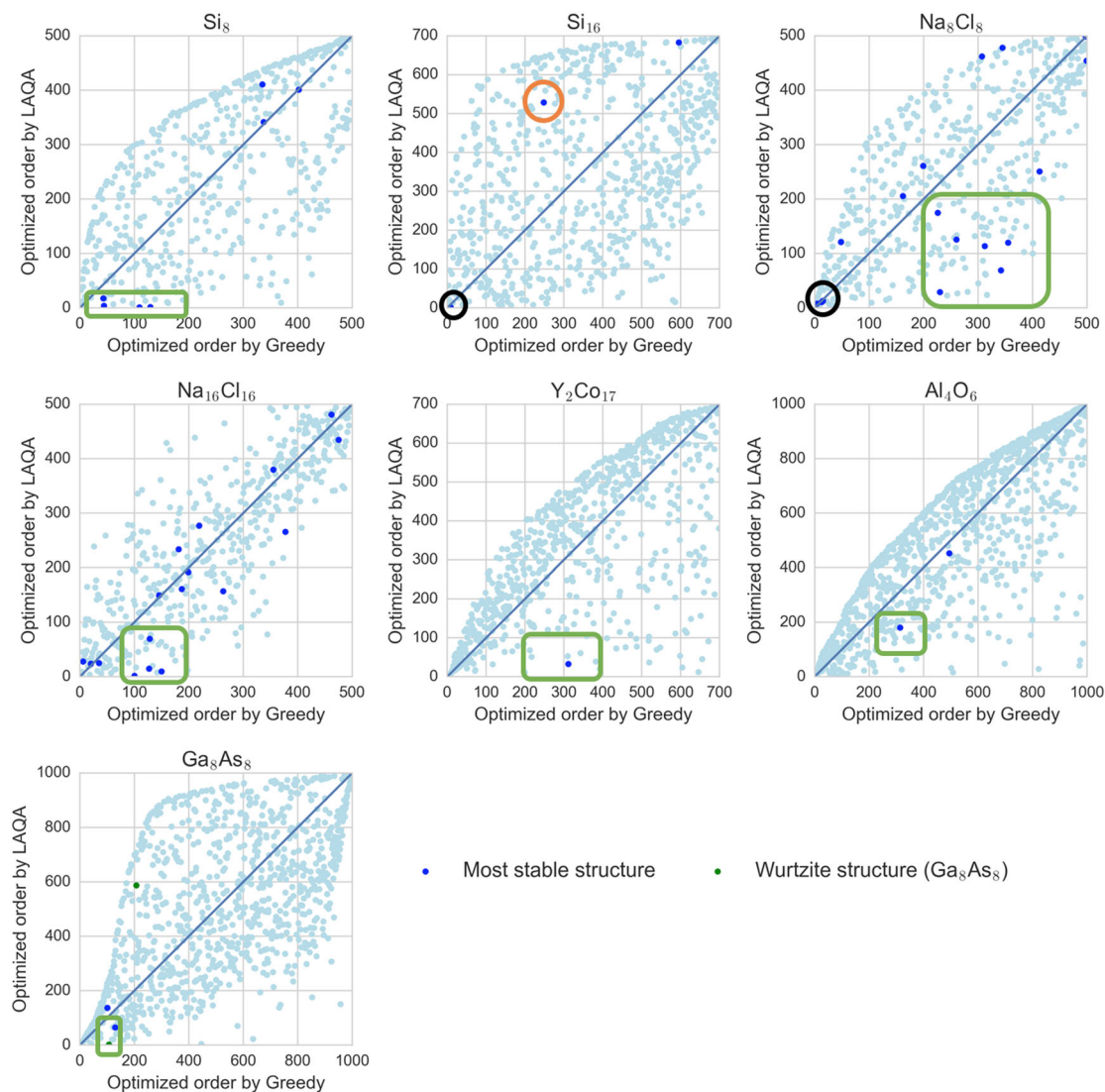


Fig. 6 Acceleration of local optimization for the stable structures and wurtzite structures by using LAQA. Each dot shows how many initial structures were fully optimized before an initial structure by using Greedy and LAQA. Dark blue and green dots show stable structures and wurtzite structures for Ga_8As_8 . The other dots are structures that were not optimized to stable structures and wurtzite structures for Ga_8As_8 . For example, if an initial structure were fully optimized n th and m th using Greedy and LAQA and $n = m$, then the position of a dot is (n, n) on the diagonal. If $n > m$, then its position is (n, m) under the diagonal and it means that the optimization for the structure is accelerated by using LAQA compared to Greedy. See main text for orange and black circles, and green squares

would like to apply this approach to other tasks such as identification of low-energy conformers of molecules.

METHODS

Details of DFT calculation

We employed the generalized gradient approximation by Perdew, Burke, and Ernzerhof²⁵ for exchange-correlation functional. The internal atomic coordinates as well as the cell parameters were fully optimized until forces acting on every atom became less than $0.01 \text{ eV}/\text{\AA}$. The k -point meshes were automatically generated using pymatgen.²⁶ For Si, a cutoff energy of 307 eV for the plane-wave expansion of the wave function and k -point mesh density of 80 \AA^{-3} for reciprocal cells were used. For NaCl, a cutoff energy of 328 eV and k -point mesh density of 80 \AA^{-3} were used. Y_2Co_{17} was treated as a ferromagnet, and 335 eV and 100 \AA^{-3} were employed for the cutoff energy and k -point mesh density, respectively. For Al_2O_3 , a cutoff energy of 500 eV and k -point mesh density of 100 \AA^{-3} were used. For GaAs, a cutoff energy of 261 eV and k -point mesh density of 100 \AA^{-3} were used.

Data availability

Initial structures and calculated data are available at http://www.tsudalab.org/files/csp_dataset.zip. Our implementation is available on Github at <http://github.com/Tomoki-YAMASHITA/CrySPY>.

ACKNOWLEDGMENTS

We thank Assist. Prof. David duVerle, Dr. Zhufeng Hou, and Dr. Masato Sumita for helpful comments and suggestions. This work was supported by the 'Materials research by Information Integration' Initiative (MI2I) project and Core Research for Evolutional Science and Technology (CREST) [Grant number JPMJCR1502] from Japan Science and Technology Agency (JST). It was also supported by Grant-in-Aid for Scientific Research on Innovative Areas "Nano Informatics" [Grant number 25106005] from the Japan Society for the Promotion of Science (JSPS). In addition, it was supported by Ministry of Education, Culture, Sports, Science and Technology (MEXT) as 'Priority Issue on Post-K computer' (Building Innovative Drug Discovery Infrastructure Through Functional Control of Biomolecular Systems).

AUTHOR CONTRIBUTIONS

K. Terayama and K. Tsuda proposed the idea for the optimization strategy. K. Terayama and T.Y. carried out the first principles calculations and analyzed data. T.O. advised the strategy of the research. All authors discussed the results and wrote the manuscript.

ADDITIONAL INFORMATION

Competing interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Stillinger, F. H. Exponential multiplicity of inherent structures. *Phys. Rev. E* **59**, 48–51 (1999).
2. Pickard, C. J. & Needs, R. J. High-pressure phases of silane. *Phys. Rev. Lett.* **97**, 045504 (2006).
3. Pickard, C. J., & Needs, R. J. Structure of phase III of solid hydrogen. *Nat. Phys.* **3**, 473–476 (2007).
4. Pickard, C. J. & Needs, R. J. Ab initio random structure searching. *J. Phys. Condens. Matter* **23**, 053201 (2011).
5. Kirkpatrick, S. et al. Optimization by simulated annealing. *Science* **220**, 671–680 (1983).
6. Pannetier, J., Bassas-Alsina, J., Rodriguez-Carvajal, J. & Caignaert, V. Prediction of crystal structures from crystal chemistry rules by simulated annealing. *Nature* **346**, 343–345 (1990).
7. Wales, D. J. & Doye, J. P. K. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *J. Phys. Chem. A* **101**, 5111–5116 (1997).
8. Goedecker, S. Minima hopping: an efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.* **120**, 9911–9917 (2004).
9. Amsler, M. & Goedecker, S. Crystal structure prediction using the minima hopping method. *J. Chem. Phys.* **133**, 224104 (2010).
10. Oganov, A. R. & Glass, C. W. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *J. Chem. Phys.* **124**, 244704 (2006).
11. Oganov, A. R., Lyakhov, A. O. & Valle, M. How evolutionary crystal structure prediction works and why. *Acc. Chem. Res.* **44**, 227–237 (2011).
12. Lyakhov, A. O., Oganov, A. R., Stokes, H. T. & Zhu, Q. New developments in evolutionary structure prediction algorithm uspx. *Comput. Phys. Commun.* **184**, 1172–1182 (2013).
13. Wang, Y., Lv, J., Zhu, Li. & Ma, Y. Crystal structure prediction via particle-swarm optimization. *Phys. Rev. B* **82**, 094116 (2010).
14. Zhang, Y., Wang, H., Wang, Y., Zhang, L. & Ma, Y. Computer-assisted inverse design of inorganic electrides. *Phys. Rev. X* **7**, 011017 (2017).
15. Yamashita, T. et al. Crystal structure prediction accelerated by bayesian optimization. *Phys. Rev. Mater.* **2**, 013803 (2018).
16. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169 (1996).
17. Stefano, P. G. et al. Quantum espresso: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21**, 395502 (2009).
18. Sutton, R. S. & Barto, A. G. *Reinforcement learning: an introduction*. Vol. 1 (MIT press, Cambridge, 1998).
19. Jones, D. R., Schonlau, M. & Welch, W. J. Efficient global optimization of expensive black-box functions. *J. Glob. Optim.* **13**, 455–492 (1998).
20. Seko, A. et al. Prediction of low-thermal conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization. *Phys. Rev. Lett.* **115**, 205901 (2015).
21. Oganov, A. R. & Valle, M. How to quantify energy landscapes of solids. *J. Chem. Phys.* **130**, 104504 (2009).
22. CrySPY is available at <https://github.com/Tomoki-YAMASHITA/CrySPY>.
23. Blöchl, P. E. Projector augmented wave method. *Phys. Rev. B* **50**, 17953 (1994).
24. Ostertag, W. Crystallographic data for yco3 and y2co17. *Acta Crystallogr.* **19**, 150–151 (1965).
25. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
26. Ong, S. P. et al. Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018