# Classification and enumeration of solid-solid phase transition mechanisms

Fang-Cheng Wang [1], Qi-Jun Ye [1,2,*], Yu-Cheng Zhu [1], and Xin-Zheng Li [1,2,3,†]

[1] *State Key Laboratory for Artificial Microstructure and Mesoscopic Physics, Frontier Science Center for Nano-optoelectronics, School of Physics, Peking University, Beijing 100871, People's Republic of China*

[2] *Interdisciplinary Institute of Light-Element Quantum Materials, Research Center for Light-Element Advanced Materials, and Collaborative Innovation Center of Quantum Matter, Peking University, Beijing 100871, People's Republic of China*

[3] *Peking University Yangtze Delta Institute of Optoelectronics, Nantong, Jiangsu 226010, People's Republic of China*

(Received 5 June 2025; accepted 30 September 2025; published 28 October 2025)

A crystal-structure match (CSM), the atom-to-atom correspondence between two crystalline phases, is used extensively to describe solid-solid phase transition (SSPT) mechanisms. However, existing computational methods cannot account for all possible CSMs. Here, we propose a formalism to classify all CSMs into a tree structure, which is independent of the choices of unit cell and supercell. We rigorously proved that only a finite number of noncongruent CSMs are of practical interest. By representing CSMs as integer matrices, we introduce the CRYSTMATCH method to *exhaustively* enumerate them, which uncontroversially solves the CSM optimization problem under *any* geometric criterion. For most SSPTs, CRYSTMATCH can reproduce all known deformation mechanisms and CSMs within 10 CPU minutes, while also revealing thousands of new candidates. The resulting database can be further used for comparing experimental phenomena, high-throughput energy barrier calculations, or machine learning.

DOI: 10.1103/lw16-4gny

## I. INTRODUCTION

In solid-solid phase transition (SSPT), atoms in one crystal structure are rearranged to form another crystal structure. This process establishes an atom-to-atom correspondence between the initial and final structures, which is called a crystal-structure match (CSM) [1–3]. Every transition path on the potential energy surface (PES) determines a CSM, while the latter describes the SSPT more concisely and flexibly. Consequently, it is common for researchers to identify the CSM, rather than the transition path, to determine which SSPT mechanism is involved [4–8]. This is especially true in nucleation simulations [7–10], where the term "mechanism" is almost synonymous with "CSM", which serves as a common language for both nucleation and concerted mechanism studies [3].

It has long been believed that SSPTs, like chemical reactions, occur along a minimum energy path (MEP) with the lowest energy barrier [11–13]. Just as each transition path determines a CSM, each CSM can also yield an MEP through methods such as the solid-state nudged elastic band (NEB) [12]. However, NEB-like methods can only produce an MEP with the user-specified CSM, ignoring all other CSMs and their MEPs, as illustrated in Fig. 1. Hence, despite their great success, NEB-like methods may still fail to give the lowest-barrier MEP, which has been an open question in the SSPT mechanism research for the past decade [13–15]. The main difficulty is that unlike the atom-to-atom correspondence in a chemical reaction, possible CSMs in an SSPT are extremely numerous—due to the large number of identical atoms and possible ways of lattice deformation—and thus hard to determine by human intuition [1–3].

To investigate what CSMs are favored by the SSPT, efforts have been made on two fronts: variable-CSM MEP calculation [14,15] and pure geometric CSM optimization [13]. The former includes the stochastic surface walking (SSW) [15–17] and PALLAS [14], where intensive energy calculations hinder the enumeration of CSMs. The latter relies on a specific optimization criterion, such as minimum strain [18], minimum number of broken bonds [1], or minimum total distance traveled by the atoms [2,19]. However, geometrically optimized CSMs may not have the lowest energy barrier, and some are even suboptimal under their own criteria [3]. It would be useful to list *all* possible CSMs, both to guide MEP calculations and to identify the best CSM under *any* geometric criterion. To this end, last year we introduced the CRYSTMATCH method to enumerate lattice deformations and representative CSMs in a Ref. [3].

In this paper, we propose a mathematical formalism that fully classifies all CSMs. Building on this, the CRYSTMATCH method is improved and generalized, whose completeness as an exhaustive algorithm is also proven. In Sec. II, we defined in a *cell-independent* manner the deformation gradient, sublattice match (SLM), multiplicity, shuffle distance and congruence relation of the CSM. These concepts organize all CSMs into a tree structure, as shown in Fig. 2, whose pruning scheme are discussed in Sec. III. To traverse this tree, three algorithms are detailed in Sec. IV. Their application is demonstrated in Sec. V using the B1–B2 and graphite-to-diamond transitions as examples.

*Contact author: qjye@pku.edu.cn
†Contact author: xzli@pku.edu.cn

 ©2025 American Physical Society

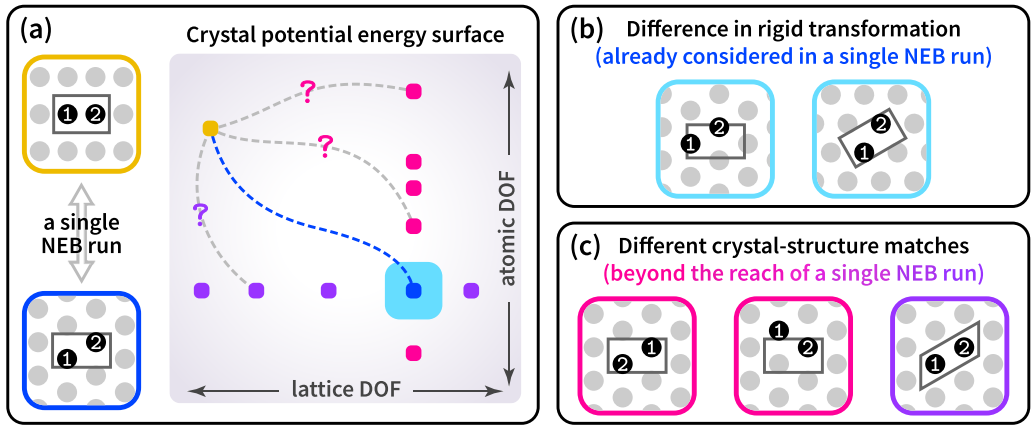


FIG. 1. A schematic diagram showing that given the initial (orthorhombic) and final (hexagonal) crystal structures, the SSPT may proceed in drastically different ways. (a) After specifying the yellow and blue endpoints, the NEB method can find the MEP as indicated by the blue dashed line. However, completing the SSPT does not require the lattice and atoms to end up in the blue-box configuration. Other possible endpoints (cyan, pink, and purple) spread out in both lattice and atomic degrees of freedom (DOF) on the PES, whose MEPs (gray dashed lines) are inaccessible in a single NEB run. (b) Endpoints that differ in rigid transformations (cyan) can be easily taken into account by the NEB method [20]. (c) Endpoints that differ in CSMs, including those with the same SLM (pink) and different SLMs (purple), which we use CRYSTMATCH to exhaust.

It should be noted that the molecular dynamics (MD) has also been used extensively to study SSPTs [21]. Among different transition paths, the one with the lowest energy barrier is most likely to be obtained by MD. This bypasses the CSM problem, but it takes an average time $\bar{t} \propto \exp(\beta\Delta^{\ddagger})$ to see a rare event with energy barrier $\Delta^{\ddagger}$ to happen. Such $\Delta^{\ddagger}$ increases linearly with the supercell volume until it can accommodate a critical nucleus [8–10,12], making the MD simulation very hard—if not infeasible. Fortunately, enhanced sampling techniques like the metadynamics (MetaD) [22,23] can reinforce the occurrence of rare events by adding a history-dependent bias potential, making it possible to simulate nucleation in SSPTs [8,9]. Nevertheless, the bias potential may elevate different energy barriers *unequally*, thus altering the lowest-barrier path and CSM. We do not discuss here how CRYSTMATCH can serve MetaD, but simply show that the enumeration results of the former contain the CSM yielded by the latter.

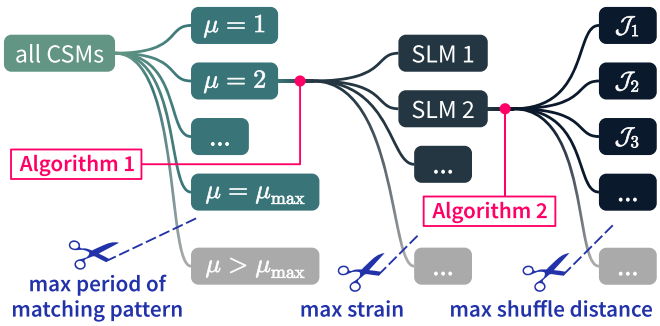


FIG. 2. Given the initial and final crystal structures, all CSMs between them can be classified by multiplicity $\mu$ and SLM, which describe the periodicity and deformation of each CSM. Empirically, the number of atoms in one period of the matching pattern does not exceed the order of $\sim 10^1$; see Table II. Thus we may assume $\mu \leqslant \mu_{\max}$. Also, CSMs with too much strain or shuffle distance are deemed nonphysical. After these pruning steps, the tree contains only a finite number of CSMs (leaf nodes), which can be obtained via Algorithms 1 and 2.

## II. CELL-INDEPENDENT FORMALISM

Conventionally, CSMs are described by a pair of initial and final supercells, as shown in Fig. 3. However, the multivaluedness of this description causes severe inconveniences for both the theory and the enumeration of CSMs [2,3,24]. The purpose of this section is to establish a *cell-independent* formalism, making it self-evident that multiplicity and SLM—the main subjects in Fig. 2—are general,

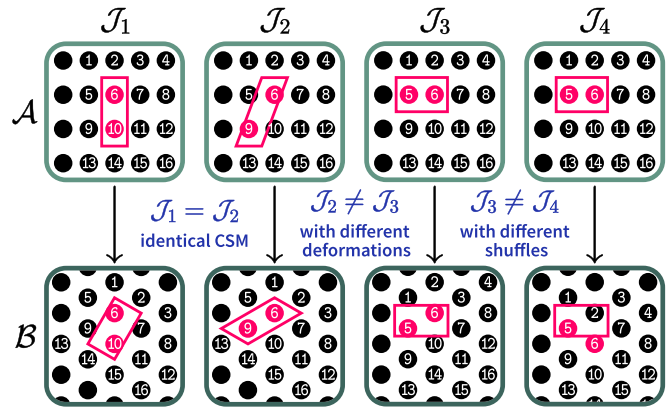


FIG. 3. Four example CSMs $\mathcal{J}_1$, $\mathcal{J}_2$, $\mathcal{J}_3$, $\mathcal{J}_4$ between crystal structures $\mathcal{A}$ and $\mathcal{B}$. The atom labels 1–16 indicate how the CSMs map each atom in $\mathcal{A}$ to its corresponding atom in $\mathcal{B}$. Conventionally, each CSM is described by the correspondences between supercell vectors (pink frames) and between motifs (pink atoms). However, the atom labels show that $\mathcal{J}_1$ and $\mathcal{J}_2$ are identical, though their conventional descriptions differ. One can also see that $\mathcal{J}_2$ and $\mathcal{J}_3$ have different deformations, while $\mathcal{J}_3$ and $\mathcal{J}_4$ have identical deformation and different shuffles.

TABLE I. Frequently used symbols in Sec. II.

| Symbol | Meaning |
|---|---|
| $\mathcal{A}, \mathcal{B}$ | Crystal structure (atomic structure) |
| $S$ | Deformation gradient (linear transformation) |
| $+\mathbf{t}$ | Translation map $\mathbf{a} \mapsto \mathbf{a}+\mathbf{t}$, where $\mathbf{t} \in \mathbb{R}^3$ |
| $(S, \mathbf{t})$ | Affine transformation $\mathbf{a} \mapsto S\mathbf{a}+\mathbf{t}$ |
| $S\mathcal{A}+\mathbf{t}$ | Crystal structure transformed by $(S, \mathbf{t})$ |
| $L_A$ | Lattice of a crystal structure $\mathcal{A}$ |
| $\tilde{L}_A$ | Sublattice of a crystal structure $\mathcal{A}$ |
| $Z_A$ | Number of atoms in a motif of $\mathcal{A}$ |
| $\mathcal{J}$ | CSM (atom-to-atom correspondence) |
| $(\tilde{L}_A, \tilde{L}_B, S)$ | SLM from $\mathcal{A}$ to $\mathcal{B}$ |
| $\tilde{Z}$ | Period of an SLM or a CSM |
| $\mu$ | Multiplicity of an SLM or a CSM |

singlevalued, intrinsic properties of CSMs. We also define congruence relations among CSMs and SLMs, which are crucial for the efficiency of CRYSTMATCH.

To ensure the unambiguousness of the narration and proofs, we define fundamental terms (e.g., "lattice" and "crystal structure") using the language of set theory. All lemmas are proved in Appendix A, while frequently used symbols are summarized in Table I. Commutative diagrams, as a tool for depicting the composition of homomorphisms, are also adopted to simplify the exposition. Let us begin with a set of points in $\mathbb{R}^3$, where each point is endowed with an atomic species; we denote by $\mathbb{X}$ the set of all atomic species.

*Definition 1 (Atomic structure).* An atomic structure is a nonempty set $\mathcal{A} \subset \mathbb{R}^3$ together with a mapping $\chi_A : \mathcal{A} \to \mathbb{X}$.

*Definition 2 (Atom-to-atom correspondence).* Let $\mathcal{A}$ and $\mathcal{B}$ be atomic structures. We say that a mapping $\mathcal{J} : \mathcal{A} \to \mathcal{B}$ is an atom-to-atom correspondence if (1) $\mathcal{J}$ is bijective, and (2) $\mathcal{J}$ preserves atom species, i.e., makes the following diagram commute:

$$\begin{array}{ccc} \mathcal{A} & \xrightarrow{\mathcal{J}} & \mathcal{B} \\ & {\scriptstyle\chi_A}\searrow \quad \swarrow{\scriptstyle\chi_B} & \\ & \mathbb{X} & \end{array} . \tag{1}$$

The above diagram is essentially a directed graph, whose vertices are sets and arrows are mappings. Paths composed of multiple arrows naturally represent a composite mapping, from its starting point to its endpoint. Commutativity of Eq. (1) means that any two paths with the same starting point and endpoint are equal, i.e.,

$$\chi_A = \chi_B \circ \mathcal{J}. \tag{2}$$

One can also interpret Eq. (1) as follows: For any $\mathbf{a} \in \mathcal{A}$, it is mapped along different paths to

$$\chi_A(\mathbf{a}) = \chi_B(\mathcal{J}(\mathbf{a})), \tag{3}$$

which means that the species of atom $\mathbf{a}$ is the same as that of its counterpart $\mathcal{J}(\mathbf{a})$.

Commutative diagrams are primarily used in category theory and homological algebra, where the arrows represent homomorphisms. In this spirit, we redefine for all subsequent diagrams: Arrows between two atomic structures *always* represent atom-to-atom correspondences, i.e., species-preserving bijections. Consequently, the symbols $\chi_A$ and $\mathbb{X}$ no longer appear in most diagrams. This approach handles polyatomic SSPTs once and for all. We hope the reader keeps in mind that $\mathcal{A} \xrightarrow{\cdots} \mathcal{B}$ denotes an "isomorphism" between atomic structures.

### A. Lattices and crystal structures

*Definition 3 (Lattice).* We say that a vector set $L \subset \mathbb{R}^3$ is a lattice if it (1) is a group under vector addition, (2) is full rank (has three linearly independent elements), and (3) has a positive lower bound $\lambda$ on the distance between any two of its elements, i.e.,

$$\forall \mathbf{t}_1, \mathbf{t}_2 \in L, \quad \mathbf{t}_1 \neq \mathbf{t}_2 \Rightarrow |\mathbf{t}_1 - \mathbf{t}_2| \geqslant \lambda. \tag{4}$$

*Definition 4 (Crystal structure).* Let $\mathcal{A}$ be an atomic structure. We say that a vector $\mathbf{t} \in \mathbb{R}^3$ is a translation element of $\mathcal{A}$ if the mapping $\mathbf{a} \mapsto \mathbf{a}+\mathbf{t}$ is an atom-to-atom correspondence from $\mathcal{A}$ to itself, i.e.,

$$\mathcal{A} \xrightarrow{+\mathbf{t}} \mathcal{A}. \tag{5}$$

We say that an atomic structure $\mathcal{A}$ is a crystal structure if all its translation elements form a lattice, denoted as $L_A$. We refer to $L_A$ as the lattice of $\mathcal{A}$, and sublattices of $L_A$ sublattices of $\mathcal{A}$.

Let $\tilde{L}_A$ be a sublattice of a crystal structure $\mathcal{A}$. The relation $\sim$ on $\mathcal{A}$ defined as

$$\mathbf{a} \sim \mathbf{a}' \iff \exists \mathbf{t} \in \tilde{L}_A, \quad \mathbf{a}' = \mathbf{a}+\mathbf{t} \tag{6}$$

is an equivalence relation, whose reflexivity, symmetry, and transitivity are ensured by the identity element, invertibility, and closure of the group $(\tilde{L}_A, +)$, respectively. We say that $\mathbf{a}, \mathbf{a}' \in \mathcal{A}$ are $\tilde{L}_A$-equivalent, and denote the $\tilde{L}_A$ equivalence class of $\mathbf{a}$ by

$$\mathbf{a} + \tilde{L}_A = \{\mathbf{a}+\mathbf{t} | \mathbf{t} \in \tilde{L}_A\}, \tag{7}$$

and the quotient set by

$$\mathcal{A}/\tilde{L}_A = \{\mathbf{a} + \tilde{L}_A \,|\, \mathbf{a} \in \mathcal{A}\}. \tag{8}$$

For any crystal structure $\mathcal{A}$, we denote $Z_A = |\mathcal{A}/L_A|$, which equals the number of atoms in a motif of $\mathcal{A}$. In practice, crystal structures always have a finite $Z_A$; otherwise, their unit cells would contain infinitely many atoms. We will assume this henceforth, but the reader can also modify Definition 1 to make it rigorous; see Lemma 9.

*Definition 5 (CSM).* A crystal structure match (CSM) is an atom-to-atom correspondence between two crystal structures. We denote the set of all CSMs from crystal structure $\mathcal{A}$ to crystal structure $\mathcal{B}$ by $\mathrm{CSM}(\mathcal{A}, \mathcal{B})$.

### B. Decomposing crystal-structure matches

Now we aim at defining the deformation gradient of a CSM. Our approach is to first define the *shuffle*, a type of CSM that is considered to have no deformation. Then, we will define the deformation of a general CSM as the difference between that CSM and a shuffle, as shown in Fig. 4(a). In the following text, the composition of mappings (including CSMs, linear transformations, and translations) will appear

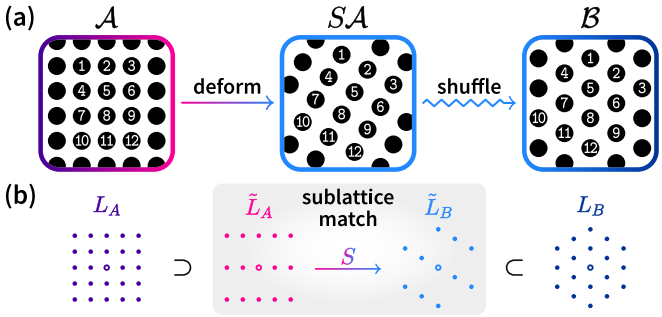


FIG. 4. (a) An example of $\mathcal{J} \in \mathrm{CSM}(\mathcal{A}, \mathcal{B})$ being decomposed into a deformation gradient $S \in \mathrm{CSM}(\mathcal{A}, S\mathcal{A})$ and a shuffle $\mathcal{J}S^{-1} \in \mathrm{CSM}(S\mathcal{A}, \mathcal{B})$. (b) In the above decomposition, a sublattice $\tilde{L}_A \subset L_A$ is deformed to the shuffle lattice $\tilde{L}_B \subset L_B$. We call the triplet $(\tilde{L}_A, \tilde{L}_B, S)$ the SLM of $\mathcal{J}$.

more frequently. When there is no risk of confusion, we will denote $f \circ g$ simply as $fg$.

*Definition 6 (Shuffle).* Let $\mathcal{A}$ and $\mathcal{B}$ be crystal structures. We say that a vector $\mathbf{t} \in L_A \cap L_B$ is a translation element of $\mathcal{J} \in \mathrm{CSM}(\mathcal{A}, \mathcal{B})$ if we have the commutative diagram:

$$\begin{array}{ccc} \mathcal{A} & \xrightarrow{\mathcal{J}} & \mathcal{B} \\ {\scriptstyle +\mathbf{t}}\downarrow & & \downarrow{\scriptstyle +\mathbf{t}} \\ \mathcal{A} & \xrightarrow[\mathcal{J}]{} & \mathcal{B} \end{array}, \tag{9}$$

i.e., $\mathcal{J}(\mathbf{a}+\mathbf{t}) = \mathcal{J}(\mathbf{a}) + \mathbf{t}$ for all $\mathbf{a} \in \mathcal{A}$. If all translation elements of $\mathcal{J}$ form a lattice, we refer to it as the shuffle lattice of $\mathcal{J}$ and say that $\mathcal{J}$ is a shuffle. We use a squiggly arrow $\mathcal{A} \dashrightarrow \mathcal{B}$ to denote a shuffle.

In actual SSPTs, $\mathcal{A}$ and $\mathcal{B}$ in different phases generally have no common translation elements (except $\mathbf{0}$), in which case $\mathcal{J} \in \mathrm{CSM}(\mathcal{A}, \mathcal{B})$ can never be a shuffle. However, it is possible to deform $\mathcal{A}$ such that the deformed structure shares more translation elements with $\mathcal{B}$, making the deformed CSM a shuffle. To formalize this idea, we denote by $S\mathcal{A} = \{S\mathbf{a} \mid \mathbf{a} \in \mathcal{A}\}$ the crystal structure deformed from $\mathcal{A}$ via $S \in \mathrm{GL}(3, \mathbb{R})$. It has a canonical atom-species mapping $\chi_A S^{-1}$, which is the only mapping that makes the following diagram commutes:

$$\begin{array}{ccc} \mathcal{A} & \xrightarrow{S} & S\mathcal{A} \\ {\scriptstyle \chi_A}\downarrow & \swarrow_{\chi_A S^{-1}} & \\ \mathbb{X} & & \end{array}, \tag{10}$$

thus making $S \in \mathrm{CSM}(\mathcal{A}, S\mathcal{A})$. Now we draw the commutative diagram corresponding to Fig. 4(a).

*Definition 7 (Deformation gradient).* We say that $S \in \mathrm{GL}(3, \mathbb{R})$ is a deformation gradient of $\mathcal{J} \in \mathrm{CSM}(\mathcal{A}, \mathcal{B})$ if we have the commutative diagram:

$$\begin{array}{ccc} \mathcal{A} & \xrightarrow{\mathcal{J}} & \mathcal{B} \\ & {\scriptstyle S}\searrow \quad \nearrow & \\ & S\mathcal{A} & \end{array}, \tag{11}$$

i.e., $\mathcal{J}S^{-1} : S\mathcal{A} \to \mathcal{B}$ is a shuffle.

### C. Classifying crystal-structure matches

Next, we will show the uniqueness of the decomposition in Fig. 4(a). This implies that for the CSMs of interest, the tuple $(\tilde{L}_A, \tilde{L}_B, S)$ in Fig. 4(b) is single-valued, which we refer to as the SLM. Here, $\tilde{L}_A$ and $\tilde{L}_B$ describe the spatial periods of the matching pattern in $\mathcal{A}$ and $\mathcal{B}$, respectively. The deformation gradient $S$ must be an isomorphism between $\tilde{L}_A$ and $\tilde{L}_B$, yet it is *not* fully determined by $(\tilde{L}_A, \tilde{L}_B)$.

*Theorem 1.* The deformation gradient of a CSM is either unique or nonexistent.

*Proof.* Assume that $\mathcal{J} \in \mathrm{CSM}(\mathcal{A}, \mathcal{B})$ has two deformation gradients $S_1$ and $S_2$. By Definition 7, we have the commutative diagram:

$$\begin{array}{ccc} \mathcal{A} & \xrightarrow{S_2} & S_2\mathcal{A} \\ {\scriptstyle S_1}\downarrow & \searrow^{\mathcal{J}} & \downarrow{\scriptstyle \mathcal{J}S_2^{-1}} \\ S_1\mathcal{A} & \xrightarrow[\mathcal{J}S_1^{-1}]{} & \mathcal{B} \end{array}. \tag{12}$$

We denote the shuffle lattices of $\mathcal{J}S_1^{-1}$ and $\mathcal{J}S_2^{-1}$ by $\tilde{L}_1$ and $\tilde{L}_2$, respectively. For each $\mathbf{t} \in \tilde{L}_1 \cap \tilde{L}_2$, we have the commutative diagram:

$$\begin{array}{ccccc} S_1\mathcal{A} & \xrightarrow{\mathcal{J}S_1^{-1}} & \mathcal{B} & \xleftarrow{\mathcal{J}S_2^{-1}} & S_2\mathcal{A} \\ {\scriptstyle +\mathbf{t}}\downarrow & & {\scriptstyle +\mathbf{t}}\downarrow & & \downarrow{\scriptstyle +\mathbf{t}} \\ S_1\mathcal{A} & \xrightarrow[\mathcal{J}S_1^{-1}]{} & \mathcal{B} & \xleftarrow[\mathcal{J}S_2^{-1}]{} & S_2\mathcal{A} \end{array}. \tag{13}$$

Combining Eqs. (12) and (13), we obtain

$$\begin{array}{ccccc} & & \mathcal{A} & \xrightarrow{S_2} & S_2\mathcal{A} \\ & {\scriptstyle S_1}\swarrow & \searrow^{\mathcal{J}} & & \downarrow{\scriptstyle \mathcal{J}S_2^{-1}} \\ S_1\mathcal{A} & & \xrightarrow[\mathcal{J}S_1^{-1}]{} & \mathcal{B} & \\ {\scriptstyle +\mathbf{t}}\downarrow & & & {\scriptstyle +\mathbf{t}}\downarrow & \downarrow{\scriptstyle +\mathbf{t}} \\ & & \mathcal{A} & \xrightarrow{S_2} & S_2\mathcal{A} \\ & {\scriptstyle S_1}\swarrow & \searrow^{\mathcal{J}} & & \downarrow{\scriptstyle \mathcal{J}S_2^{-1}} \\ S_1\mathcal{A} & & \xrightarrow[\mathcal{J}S_1^{-1}]{} & \mathcal{B} & \end{array}, \tag{14}$$

from which we can extract the subdiagram:

$$\begin{array}{ccccc} S_1\mathcal{A} & \xleftarrow{S_1} & \mathcal{A} & \xrightarrow{S_2} & S_2\mathcal{A} \\ {\scriptstyle +\mathbf{t}}\downarrow & & & & \downarrow{\scriptstyle +\mathbf{t}} \\ S_1\mathcal{A} & \xleftarrow[S_1]{} & \mathcal{A} & \xrightarrow[S_2]{} & S_2\mathcal{A} \end{array}. \tag{15}$$

The commutativity of Eq. (15) means that $S_1^{-1}\mathbf{t} = S_2^{-1}\mathbf{t}$. By Definitions 6, both $\tilde{L}_1$ and $\tilde{L}_2$ are sublattices of $L_B$ so that $\tilde{L}_1 \cap \tilde{L}_2$ is full rank (Lemma 10). Hence, there exist three linearly independent vectors $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3$ such that

$$S_1^{-1}[\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3] = S_2^{-1}[\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3]. \tag{16}$$

Since $[\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3]$ is invertible, we have $S_1 = S_2$. ■

*Theorem 1.* enables us to classify CSMs by their deformation and shuffle. Now, we merge the deformation gradient and the shuffle lattice into a single concept—the SLM, whose utility will become evident in Secs. III and IV.

*Definition 8 (SLM).* Let $\mathcal{A}$, $\mathcal{B}$ be crystal structures, $\tilde{L}_A$, $\tilde{L}_B$ their respective sublattices, and $S \in \mathrm{GL}(3, \mathbb{R})$. We say

TABLE II. Some previously proposed CSMs (also known as mechanisms or paths). Congruent CSMs (e.g., the Bain mechanism and many shear mechanisms of the FCC-to-BCC transition [25]) are only displayed once. See Ref. [26] for more details.

| SSPT prototype | CSM[a] | $\mu$ | $\tilde{Z}$ | RMSS[b] | RMSD[b] | Discovered via | Experiment/simulation evidence[c] |
|---|---|---|---|---|---|---|---|
| A1–A2 (FCC-to-BCC) | Bain | 1 | 1 | 15.9% | 0 | Inference [28] | MD [29] |
| | Therrien-Stevanović | 6 | 6 | 9.0% | 0.713 Å | P2PTRANS [25] | Pitsch OR [30] |
| | WYZL $\mathcal{J}_2$ | 6 | 6 | 9.0% | 0.884 Å | CRYSTMATCH [3] | Nishiyama-Wassermann OR [31] |
| | WYZL $\mathcal{J}_1$ | 36 | 36 | 4.3% | 0.961 Å | CRYSTMATCH [3] | Kurdjumov-Sachs OR [32] |
| B1–B2 (NaCl-to-CsCl) | Buerger | 1 | 2 | 27.5% | 0 | Inference [33] | MD [34,35] |
| | WTM | 2 | 4 | 17.4% | 1.135 Å | Inference [36] | Watanabe-Blaschko OR [36,37] |
| | TKED | 4 | 8 | 17.4% | 1.605 Å | Inference [38] | MD [7], MetaD [8] |
| | TGS B1–B2 | 6 | 12 | 10.9% | 1.654 Å | P2PTRANS [2] | Very low strain |
| B1–B4 (NaCl-to-wurtzite) | TAS | 1 | 4 | 27.1% | 0.500 Å | Inference [39,40] | MD [4,41], MetaD [9] |
| | ZCS II | 1 | 4 | 18.4% | 0.822 Å | PALLAS [14] | See Ref. [14] |
| | Shimojo *et al.* II | 2 | 8 | 15.0% | 0.865 Å | MD [5] | MD [5,6] |
| | TGS B1–B4 | 3 | 12 | 8.5% | 1.084 Å | P2PTRANS [2] | Very low strain |

[a]The abbreviations stand for Wang-Ye-Zhu-Li (WYZL), Watanabe-Tokonami-Morimoto (WTM), Tolédano-Knorr-Ehm-Depmeier (TKED), Therrien-Graf-Stevanović (TGS), Tolbert-Alivisatos-Sowa (TAS), and Zhu-Cohen-Strobel (ZCS).
[b]The CSMs in the table are defined independently of specific materials. However, to calculate the root-mean-square strain (RMSS) and root-mean-square displacement (RMSD), lattice parameters must be specified. Here, lattice parameters of Fe, CsCl, and ZnO are used.
[c]The orientation relationship (OR) of a CSM is defined in the rotation-free manner, which is also adopted by Refs. [3,5,25,27].

that $(\tilde{L}_A, \tilde{L}_B, S)$ is a sublattice match (SLM) from $\mathcal{A}$ to $\mathcal{B}$ if

$$|\mathcal{A}/\tilde{L}_A| = |\mathcal{B}/\tilde{L}_B|, \tag{17}$$

$$SL_A = L_B. \tag{18}$$

We refer to the value of Eq. (17) as the period of the SLM, denoted by $\tilde{Z}$, and say that the SLM has multiplicity

$$\mu = \frac{\tilde{Z}}{\mathrm{lcm}(Z_A, Z_B)}. \tag{19}$$

Denote by $\mathrm{SLM}(\mathcal{A}, \mathcal{B})$ the set of all SLMs from $\mathcal{A}$ to $\mathcal{B}$, and those with multiplicity $\mu$ by $\mathrm{SLM}(\mathcal{A}, \mathcal{B}; \mu)$.

*Theorem 2.* If $\mathcal{J} \in \mathrm{CSM}(\mathcal{A}, \mathcal{B})$ has a deformation gradient $S$, then $(\tilde{L}_A, \tilde{L}_B, S) \in \mathrm{SLM}(\mathcal{A}, \mathcal{B})$, where $\tilde{L}_B$ is the shuffle lattice of $\mathcal{J}S^{-1}$ and $\tilde{L}_A = S^{-1}\tilde{L}_B$.

*Proof.* Equation (18) is already satisfied and we only need to show Eq. (17). Consider a mapping $p : \mathcal{A}/\tilde{L}_A \to \mathcal{B}/\tilde{L}_B$ such that the following diagram commutes:

$$\begin{array}{ccc} \mathcal{A} & \xrightarrow{\ \mathcal{J}\ } & \mathcal{B} \\ {\scriptstyle +\tilde{L}_A}\downarrow & & \downarrow{\scriptstyle +\tilde{L}_B}, \\ \mathcal{A}/\tilde{L}_A & \xrightarrow[\ p\ ]{} & \mathcal{B}/\tilde{L}_B \end{array} \tag{20}$$

i.e., $p : \mathbf{a} + \tilde{L}_A \mapsto \mathcal{J}(\mathbf{a}) + \tilde{L}_B$. To show that $p$ is well-defined, we must prove its singlevaluedness. For any $\mathbf{a}_1$ and $\mathbf{a}_2$ in $\mathcal{A}$, we have

$$\mathbf{a}_1 + \tilde{L}_A = \mathbf{a}_2 + \tilde{L}_A \tag{21}$$

$$\iff \exists \mathbf{t} \in \tilde{L}_A, \quad \mathbf{a}_1 = \mathbf{a}_2 + \mathbf{t} \tag{22}$$

$$\iff \exists \mathbf{t} \in \tilde{L}_A, \quad S\mathbf{a}_1 = S\mathbf{a}_2 + S\mathbf{t} \tag{23}$$

$$\iff \exists \mathbf{t}' \in \tilde{L}_B, \quad S\mathbf{a}_1 = S\mathbf{a}_2 + \mathbf{t}' \tag{24}$$

$$\iff \exists \mathbf{t}' \in \tilde{L}_B, \quad \mathcal{J}(\mathbf{a}_1) = \mathcal{J}(\mathbf{a}_2) + \mathbf{t}' \tag{25}$$

$$\iff \mathcal{J}(\mathbf{a}_1) + \tilde{L}_B = \mathcal{J}(\mathbf{a}_2) + \tilde{L}_B, \tag{26}$$

where the shuffle property of $\mathcal{J}S^{-1} : S\mathcal{A} \rightsquigarrow \mathcal{B}$ is used to derive Eq. (25). The equivalence of Eqs. (21) and (26) means that $p$ is injective, while $p$ is surjective since $\mathcal{J}$ is surjective. So far, we can see that $p$ is bijective and thus Eq. (17) holds. ■

*Definition 9 (Multiplicity).* If $\mathcal{J} \in \mathrm{CSM}(\mathcal{A}, \mathcal{B})$ has a deformation gradient, we say that the $(\tilde{L}_A, \tilde{L}_B, S)$ defined in Theorem 2 is the SLM of $\mathcal{J}$. In such case, we refer to the period and multiplicity of the SLM as those of $\mathcal{J}$. If the deformation gradient of $\mathcal{J}$ does not exist, we say that the period and multiplicity of $\mathcal{J}$ are $\infty$.

The period $\tilde{Z}$ of a CSM can be viewed as the number of atoms in a "motif" of the CSM. Since $\tilde{Z}$ must be a common multiple of $Z_A$ and $Z_B$ (Lemma 4), the multiplicity $\mu$ always takes values in positive integers. It must be emphasized that $\mu$ and $\tilde{Z}$ are properties of the CSM, which are conceptually distinct from the number of atoms used in SSPT simulations. For example, MetaD calculations using up to $\sim 10^5$ atoms often yield CSMs with $\mu$ and $\tilde{Z}$ less than $\sim 10^1$; see Table II. Practically, we focus on solely those CSMs with finite multiplicity.

### D. Congruence relations

Applying rigid transformations to $\mathcal{A}$ and $\mathcal{B}$ yields "congruent" yet set-theoretically different CSMs, as illustrated in Fig. 5. Such CSMs are equivalent as inputs to NEB-like methods, so we only need to enumerate all "noncongruent" CSMs. We denote by $S\mathcal{A} + \mathbf{t} = \{S\mathbf{a} + \mathbf{t} \mid \mathbf{a} \in \mathcal{A}\}$ the crystal structure transformed from $\mathcal{A}$ via an affine transformation $(S, \mathbf{t})$. In the same sense as in Eq. (10), we have $(S, \mathbf{t}) \in \mathrm{CSM}(\mathcal{A}, S\mathcal{A} + \mathbf{t})$.

*Definition 10 (Congruent CSM).* We say that a CSM $\mathcal{J}'$ is congruent to $\mathcal{J} \in \mathrm{CSM}(\mathcal{A}, \mathcal{B})$ if there exist proper rigid transformations $(R_A, \mathbf{t}_A)$ and $(R_B, \mathbf{t}_B)$ such that

$$\begin{array}{ccc} \mathcal{A} & \xrightarrow{\ \mathcal{J}\ } & \mathcal{B} \\ {\scriptstyle (R_A,\mathbf{t}_A)}\downarrow & & \downarrow{\scriptstyle (R_B,\mathbf{t}_B)}, \\ R_A\mathcal{A} + \mathbf{t}_A & \xrightarrow[\ \mathcal{J}'\ ]{} & R_B\mathcal{B} + \mathbf{t}_B \end{array} \tag{27}$$

i.e., $\mathcal{J}' = (R_B, \mathbf{t}_B)\mathcal{J}(R_A, \mathbf{t}_A)^{-1}$.

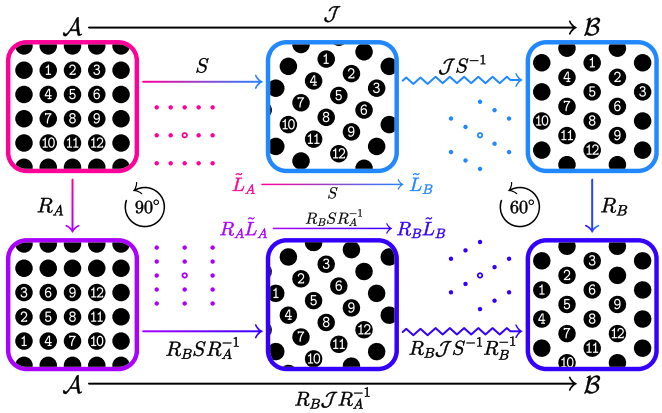


FIG. 5. An example of $R_A$ (a rotation by 90°) and $R_B$ (a rotation by 60°) being symmetry operations of $\mathcal{A}$ (tetragonal) and $\mathcal{B}$ (hexagonal), respectively. For each $\mathcal{J} \in \mathrm{CSM}(\mathcal{A}, \mathcal{B})$, there exists a congruent $R_B\mathcal{J}R_A^{-1} \in \mathrm{CSM}(\mathcal{A}, \mathcal{B})$. Their respective SLMs are $(\tilde{L}_A, \tilde{L}_B, S)$ and $(R_A\tilde{L}_A, R_B\tilde{L}_B, R_BSR_A^{-1})$.

In general, two congruent CSMs have different domains and images, i.e., we have $\mathcal{J}' \notin \mathrm{CSM}(\mathcal{A}, \mathcal{B})$ in Eq. (27), unless $(R_A, \mathbf{t}_A)$ maps $\mathcal{A}$ onto itself and $(R_B, \mathbf{t}_B)$ does the same. In this section, we specifically focus on the case where the proper rigid transformations are symmetry operations of the crystal structures, which define a congruence relation on $\mathrm{CSM}(\mathcal{A}, \mathcal{B})$. We denote by $G_A$ the group formed by all proper rigid transformations $(R, \mathbf{t})$ such that

$$\mathcal{A} \xrightarrow{(R,\mathbf{t})} \mathcal{A}, \tag{28}$$

and define the point group

$$G'_A = \{R \,|\, (R, \mathbf{t}) \in G_A\}. \tag{29}$$

It should be noted that $G'_A$ is *not* a subgroup of $G_A$ unless the latter is symmorphic [42], i.e., $G_A = G'_A \ltimes L_A$. We will see that $G'_A$ and $G'_B$ induce an equivalence relation on $\mathrm{SLM}(\mathcal{A}, \mathcal{B})$, which is inherited from Definition 10.

*Lemma 1.* If $\mathcal{J} \in \mathrm{CSM}(\mathcal{A}, \mathcal{B})$ is a shuffle, then for any $\mathbf{t}_A, \mathbf{t}_B \in \mathbb{R}^3$, the composite mapping $(+\mathbf{t}_B) \circ \mathcal{J} \circ (-\mathbf{t}_A)$ is also a shuffle from $\mathcal{A} + \mathbf{t}_A$ to $\mathcal{B} + \mathbf{t}_B$, whose shuffle lattice is the same as $\mathcal{J}$.

*Theorem 3.* For any $(R_A, \mathbf{t}_A) \in G_A$, $(R_B, \mathbf{t}_B) \in G_B$, and $\mathcal{J} \in \mathrm{CSM}(\mathcal{A}, \mathcal{B})$, the CSM $\mathcal{J}'$ defined by Eq. (27) is also in $\mathrm{CSM}(\mathcal{A}, \mathcal{B})$. If $\mathcal{J}$ has an SLM $(\tilde{L}_A, \tilde{L}_B, S)$, then $\mathcal{J}'$ has the SLM $(R_A\tilde{L}_A, R_B\tilde{L}_B, R_BSR_A^{-1})$.

*Proof.* When $(R_A, \mathbf{t}_A) \in G_A$ and $(R_B, \mathbf{t}_B) \in G_B$, the bottom line of Eq. (27) becomes $\mathcal{A} \xrightarrow{\mathcal{J}'} \mathcal{B}$. To show that $\mathcal{J}'$ has the SLM $(R_A\tilde{L}_A, R_B\tilde{L}_B, R_BSR_A^{-1})$, we only need to show that the deformed CSM

$$[(R_B, \mathbf{t}_B)\mathcal{J}(R_A, \mathbf{t}_A)^{-1}]\left(R_BSR_A^{-1}\right)^{-1} \tag{30}$$

$$= (R_B, \mathbf{t}_B)\mathcal{J}R_A^{-1} \circ (-\mathbf{t}_A) \circ R_AS^{-1}R_B^{-1} \tag{31}$$

$$= (R_B, \mathbf{t}_B)\mathcal{J}S^{-1}R_B^{-1} \circ \left(-R_BSR_A^{-1}\mathbf{t}_A\right) \tag{32}$$

is a shuffle with shuffle lattice $R_B\tilde{L}_B$. By Definition 6, a vector $\mathbf{t} \in \mathbb{R}^3$ is in $\tilde{L}_B$ if and only if we have:

$$\begin{array}{ccc} S\mathcal{A} & \xrightarrow{\mathcal{J}S^{-1}} & \mathcal{B} \\ {\scriptstyle +\mathbf{t}}\downarrow & & \downarrow{\scriptstyle +\mathbf{t}} \\ S\mathcal{A} & \xrightarrow[\mathcal{J}S^{-1}]{} & \mathcal{B} \end{array} \tag{33}$$

Applying $(R_B, \mathbf{t}_B)$ to each crystal structure, we obtain

$$\begin{array}{ccccc} R_BS\mathcal{A}+\mathbf{t}_B & & \xrightarrow{(R_B,\mathbf{t}_B)\mathcal{J}S^{-1}(R_B,\mathbf{t}_B)^{-1}} & & R_B\mathcal{B}+\mathbf{t}_B \\ & {\scriptstyle (R_B,\mathbf{t}_B)}\nwarrow & & \nearrow{\scriptstyle (R_B,\mathbf{t}_B)} & \\ {\scriptstyle +R_B\mathbf{t}}\downarrow & & S\mathcal{A} \xrightarrow{\mathcal{J}S^{-1}} \mathcal{B} & & \downarrow{\scriptstyle +R_B\mathbf{t}\cdot} \\ & {\scriptstyle (R_B,\mathbf{t}_B)}\swarrow & {\scriptstyle +\mathbf{t}}\downarrow \quad\quad \downarrow{\scriptstyle +\mathbf{t}} & \searrow{\scriptstyle (R_B,\mathbf{t}_B)} & \\ & & S\mathcal{A} \xrightarrow[\mathcal{J}S^{-1}]{} \mathcal{B} & & \\ R_BS\mathcal{A}+\mathbf{t}_B & & \xrightarrow[(R_B,\mathbf{t}_B)\mathcal{J}S^{-1}(R_B,\mathbf{t}_B)^{-1}]{} & & R_B\mathcal{B}+\mathbf{t}_B \end{array} \tag{34}$$

Taking the outer loop of the above diagram and using $(R_B, \mathbf{t}_B) \in G_B$, we can see that $\mathbf{t} \in \tilde{L}_B$ if and only if

$$\begin{array}{ccc} R_BS\mathcal{A}+\mathbf{t}_B & \xrightarrow{(R_B,\mathbf{t}_B)\mathcal{J}S^{-1}(R_B,\mathbf{t}_B)^{-1}} & \mathcal{B} \\ {\scriptstyle +R_B\mathbf{t}}\downarrow & & \downarrow{\scriptstyle +R_B\mathbf{t},} \\ R_BS\mathcal{A}+\mathbf{t}_B & \xrightarrow[(R_B,\mathbf{t}_B)\mathcal{J}S^{-1}(R_B,\mathbf{t}_B)^{-1}]{} & \mathcal{B} \end{array} \tag{35}$$

which implies that $(R_B, \mathbf{t}_B)\mathcal{J}S^{-1}(R_B, \mathbf{t}_B)^{-1}$ is a shuffle with shuffle lattice $R_B\tilde{L}_B$. Equation (32) differs from this shuffle by only a translation and is therefore also a shuffle with shuffle lattice $R_B\tilde{L}_B$ according to Lemma 1. ■

*Definition 11 (Congruent SLM).* We say that an SLM is congruent to $(\tilde{L}_A, \tilde{L}_B, S) \in \mathrm{SLM}(\mathcal{A}, \mathcal{B})$ if it can be written as $(R_A\tilde{L}_A, R_B\tilde{L}_B, R_BSR_A^{-1})$ for some $R_A \in G'_A$, $R_B \in G'_B$.

Before describing how to exhaustively enumerate CSMs, we need to clarify what kind of CSM set is *complete*. We expect that they can, via NEB-like methods which allow rigid transformations, yield all SSPT mechanisms we are concerned with. Hence, there is no need to enumerate two or more congruent CSMs, and we say that a subset of $X$ comprising CSMs or SLMs is complete as long as it contains an element of each congruence class in $X$.

*Theorem 4.* If $X$ is a complete subset of $\mathrm{SLM}(\mathcal{A}, \mathcal{B})$, then $\bigcup_{x\in X} Y_x$ is a complete subset of $\mathrm{CSM}(\mathcal{A}, \mathcal{B})$ as long as $Y_x$ is a complete subset of those CSMs with SLM $x$.

*Proof.* As long as $\mathcal{J} \in \mathrm{CSM}(\mathcal{A}, \mathcal{B})$ has an SLM, denoted by $(\tilde{L}_A, \tilde{L}_B, S)$, the completeness of $X$ ensures that there exist $R_A \in G'_A$ and $R_B \in G'_B$ such that

$$\left(R_A\tilde{L}_A, R_B\tilde{L}_B, R_BSR_A^{-1}\right) \in X. \tag{36}$$

Denote by $x$ the left-hand side of Eq. (36). Take $\mathbf{t}_A, \mathbf{t}_B$ such that $(R_A, \mathbf{t}_A) \in G_A$ and $(R_B, \mathbf{t}_B) \in G_B$. Since $\mathcal{J}$ is congruent to $(R_B, \mathbf{t}_B)\mathcal{J}(R_A, \mathbf{t}_A)^{-1}$ whose SLM is $x$ (Theorem 3), the transitivity of congruence means that $\mathcal{J}$ is also congruent to some CSM in $Y_x$. ■

Note that $\bigcup_{x\in X} Y_x$ is also a complete subset of

$$\bigcup_{\substack{R_1\in \mathrm{SO}(3),\boldsymbol{\tau}_1\in\mathbb{R}^3 \\ R_2\in \mathrm{SO}(3),\boldsymbol{\tau}_2\in\mathbb{R}^3}} \mathrm{CSM}(R_1\mathcal{A}+\boldsymbol{\tau}_1, R_2\mathcal{B}+\boldsymbol{\tau}_2), \tag{37}$$

TABLE III. Frequently used symbols in Sec. III.

| Symbol | Meaning |
|---|---|
| $\mathbb{F}^{n\times m}$ | All $n \times m$ matrices over the ring $\mathbb{F}$ |
| $C(\mathbb{Z}^3)$ | Lattice generated by nonsingular $C \in \mathbb{R}^{3\times3}$ |
| $C_A$ | Base matrix of $L_A$ (primitive-cell vectors) |
| $\mathrm{GL}(n, \mathbb{Z})$ | All $n \times n$ unimodular matrices |
| $\mathrm{hnf}(M)$ | HNF of the integer matrix $M$ |
| $\mathrm{HNF}(k)$ | All $3 \times 3$ HNFs with determinant $k$ |
| $(H_A, H_B, Q)$ | IMT representation of an SLM |
| $\mathrm{IMT}(\mu)$ | All IMTs with multiplicity $\mu$ |
| $p$ | Permutation on the set $\{1, \ldots, \tilde{Z}\}$ |
| $(p, \mathbf{t}_1, \ldots, \mathbf{t}_{\tilde{Z}})$ | PCT representation of a shuffle |
| $w(S)$ | Estimated strain energy density of $S$ |
| $\sigma_i(S)$, $s_i$ | $i$-th largest singular value of $S$ |
| $\tilde{\mathcal{J}}$ | Standard shuffle of $\mathcal{J}$ (generally $\tilde{\mathcal{J}} \neq \mathcal{J}S^{-1}$) |
| $\tilde{\mathcal{A}}, \tilde{\mathcal{B}}$ | Half-distorted structures with sublattice $\tilde{L}$ |
| $\tilde{L}$ | Shuffle lattice of $\tilde{\mathcal{J}}$ |
| $\theta_i$ | Normalized weight of the $i$th atom |
| $\ell$ | Type of norm used to define shuffle distance |
| $\hat{d}(\mathcal{J})$ | Shuffle distance of $\mathcal{J}$ |
| $d(\mathcal{J})$ | Minimum $\hat{d}$ among all CSMs congruent to $\mathcal{J}$ |

which incorporates all possible initial and final structures of an SSPT. Hence, the CSMs in $\bigcup_{x\in X} Y_x$ are sufficient as inputs to NEB-like methods. We will demonstrate how to compute $X$ and $Y_x$ in Sec. IV.

## III. REPRESENTATION AND PRUNING

Before we describe the CRYSTMATCH method for exhaustively enumerating all noncongruent CSMs, it is necessary to represent CSMs as *matrices* for computer processing. We also need to introduce some physical constraints to make $\mathrm{CSM}(\mathcal{A}, \mathcal{B})$ finite. In this section, we will show how to represent SLM and shuffle separately, and discuss the general form of pruning criteria. Afterwards, the finiteness of candidate CSMs will be proved. Symbols frequently used in this section are summarized in Table III.

### A. Representing crystal-structure matches

As elaborated in Sec. II, a CSM with finite multiplicity can always be decomposed into an SLM and a shuffle, both of which are cell-independent. However, just as an element of a vector space acquires a coordinate representation only after a basis is chosen, to represent CSMs as matrices, we also need to select a pair of "primitive cells" for $\mathcal{A}$ and $\mathcal{B}$.

#### *1. Representing sublattice matches*

*Definition 12 (Base matrix).* Let $C \in \mathbb{R}^{3\times3}$ be a nonsingular matrix. We say that $C(\mathbb{Z}^3) = \{C\mathbf{k} \mid \mathbf{k} \in \mathbb{Z}^3\}$ is the lattice generated by $C$, and $C$ is the base matrix of $C(\mathbb{Z}^3)$.

*Lemma 2.* Any lattice has a base matrix.

*Lemma 3.* Let $C, \tilde{C} \in \mathbb{R}^{3\times3}$ be nonsingular matrices. $\tilde{C}(\mathbb{Z}^3)$ is a sublattice of $C(\mathbb{Z}^3)$ if and only if there exists an $M \in \mathbb{Z}^{3\times3}$ such that $\tilde{C} = CM$. In such case, the index of subgroup $\tilde{C}(\mathbb{Z}^3)$ in $C(\mathbb{Z}^3)$ is $|\det M|$.

*Lemma 4.* Let $\mathcal{A}$ be a crystal structure and $\tilde{L}_A$ its sublattice. Denote by $k$ the index of subgroup $\tilde{L}_A$ in $L_A$, which is finite by Lemmas 2 and 3. We have $|\mathcal{A}/\tilde{L}_A| = kZ_A$ as long as $Z_A = |\mathcal{A}/L_A|$ is finite.

Take $C_A, C_B \in \mathbb{R}^{3\times3}$ that generate $L_A$ and $L_B$, respectively. For each SLM $(\tilde{L}_A, \tilde{L}_B, S)$, there exists an $M_A \in \mathbb{Z}^{3\times3}$ such that $C_A M_A(\mathbb{Z}^3) = \tilde{L}_A$. From Eq. (18), we can see that $SC_A M_A(\mathbb{Z}^3) = \tilde{L}_B$, so $SC_A M_A$ is a base matrix of $\tilde{L}_B$ which equals $C_B M_B$ for some $M_B \in \mathbb{Z}^{3\times3}$. This provides a way to represent an SLM $(\tilde{L}_A, \tilde{L}_B, S)$ by a pair of integer matrices $(M_A, M_B)$ satisfying

$$C_A M_A(\mathbb{Z}^3) = \tilde{L}_A, \tag{38}$$

$$C_B M_B(\mathbb{Z}^3) = \tilde{L}_B, \tag{39}$$

$$SC_A M_A = C_B M_B, \tag{40}$$

$$Z_A |\det M_A| = Z_B |\det M_B|, \tag{41}$$

where Eq. (41) is derived from Eq. (17) using Lemmas 3 and 4. We can also see that any nonsingular $3 \times 3$ integer matrix pair $(M_A, M_B)$ satisfying Eq. (41) determines an SLM according to Eqs. (38)–(40). This representation of SLM is intuitive and extensively used [12,18,19,24], but is not suitable for SLM enumeration since many different $(M_A, M_B)$ can yield the same SLM [3,24]. To eliminate such representation redundancy, one may define a *canonical base matrix* for each lattice. The concepts of unimodular matrices and Hermite normal form will be useful, just as they play a crucial role in the Hart-Forcade theory [43].

*Definition 13 (Unimodular matrix).* Let $Q \in \mathbb{Z}^{n\times n}$ be a nonsingular integer matrix. We say that $Q$ is unimodular if any of the following equivalent propositions holds

$$Q(\mathbb{Z}^n) = \mathbb{Z}^n, \tag{42}$$

$$Q^{-1} \in \mathbb{Z}^{n\times n}, \tag{43}$$

$$\det Q \in \{+1, -1\}. \tag{44}$$

We denote the matrix group formed by all unimodular matrices as $\mathrm{GL}(n, \mathbb{Z})$.

*Lemma 5.* Two nonsingular matrices $C, C' \in \mathbb{R}^{3\times3}$ generate the same lattice if and only if $C^{-1}C' \in \mathrm{GL}(3, \mathbb{Z})$.

*Definition 14 (HNF).* Let $H \in \mathbb{Z}^{n\times n}$ be a nonsingular matrix. We say that $H$ is in Hermite normal form (HNF) if (1) $H$ is lower triangular and (2) each off-diagonal element of $H$ is nonnegative and *strictly* less than the diagonal element in its row. In other words, the elements of $H$ satisfy

$$\begin{cases} H_{ii} > H_{ij} \geqslant 0, & \text{if } j < i \\ H_{ij} = 0, & \text{if } j > i \end{cases}. \tag{45}$$

*Lemma 6.* Let $M \in \mathbb{Z}^{n\times n}$ be a nonsingular matrix. There exists a unique $H \in \mathbb{Z}^{n\times n}$ in HNF and a unique $Q \in \mathrm{GL}(n, \mathbb{Z})$ such that

$$\exists Q \in \mathrm{GL}(n, \mathbb{Z}), \quad M = HQ. \tag{46}$$

We say that $H$ is the HNF of $M$.

Let $\mathrm{hnf}(M)$ denote the HNF of $M$, and $\mathrm{HNF}(k)$ denote all $3 \times 3$ HNFs with determinant $k$. Lemma 5 implies that

$(M_A, M_B)$ and $(M'_A, M'_B)$ represent the same SLM if and only if $(M'_A, M'_B) = (M_A Q', M_B Q')$ for some $Q' \in \mathrm{GL}(3,\mathbb{Z})$. To make the representation of SLM unique, one may use the mapping $(M_A, M_B) \mapsto (H_A, H_B, Q)$, where

$$H_A = \mathrm{hnf}(M_A), \tag{47}$$

$$H_B = \mathrm{hnf}(M_B), \tag{48}$$

$$Q = \left(H_B^{-1} M_B\right)\left(H_A^{-1} M_A\right)^{-1}. \tag{49}$$

It maps the entire $\{(M_A Q', M_B Q') \mid Q' \in \mathrm{GL}(3,\mathbb{Z})\}$ to a *single* integer-matrix triplet (IMT) [3].

*Definition 15 (IMT).* Let $\mathcal{A}$, $\mathcal{B}$ be crystal structures and $\mu$ a positive integer. We say that

$$\mathrm{IMT}(\mu) = \mathrm{HNF}\left(\frac{\mu\,\mathrm{lcm}(Z_A, Z_B)}{Z_A}\right) \times \mathrm{HNF}\left(\frac{\mu\,\mathrm{lcm}(Z_A, Z_B)}{Z_B}\right) \times \mathrm{GL}(3,\mathbb{Z}) \tag{50}$$

is the set of all IMTs with multiplicity $\mu$.

*Theorem 5.* The mapping $(H_A, H_B, Q) \mapsto (\tilde{L}_A, \tilde{L}_B, S)$ with

$$\tilde{L}_A = C_A H_A(\mathbb{Z}^3), \tag{51}$$

$$\tilde{L}_B = C_B H_B(\mathbb{Z}^3), \tag{52}$$

$$S = (C_B H_B Q)(C_A H_A)^{-1} \tag{53}$$

is a bijection from $\mathrm{IMT}(\mu)$ to $\mathrm{SLM}(\mathcal{A}, \mathcal{B}; \mu)$.

*Proof.* $(\tilde{L}_A, \tilde{L}_B, S)$ is an SLM since $S\tilde{L}_A = \tilde{L}_B$ and

$$|\mathcal{A}/\tilde{L}_A| = Z_A \det H_A \tag{54}$$

$$= \mu\,\mathrm{lcm}(Z_A, Z_B) \tag{55}$$

$$= Z_B \det H_B \tag{56}$$

$$= |\mathcal{B}/\tilde{L}_B|, \tag{57}$$

where Lemmas 3 and 4 are used to derive Eqs. (54) and (57). Definition 15 ensures Eqs. (55) and (56), so that the multiplicity of $(\tilde{L}_A, \tilde{L}_B, S)$ is $\mu$. Conversely, for each $(\tilde{L}_A, \tilde{L}_B, S) \in \mathrm{SLM}(\mathcal{A}, \mathcal{B}; \mu)$, $\tilde{L}_A$ has a unique base matrix $C_A H_A$, where $H_A$ is in HNF (Lemmas 5 and 6) with determinant $\tilde{Z}/Z_A = \mu$ (Lemmas 3 and 4), and so does $\tilde{L}_B$. Since $H_A$ and $H_B$ are determined, $S$ further determines $Q$ by Eq. (53). Therefore every SLM is associated with a unique IMT. ■

Theorem 5 allows us to focus on the enumeration of $\mathrm{IMT}(\mu)$ from now on, which is far more intuitive than the enumeration of SLMs.

### 2. Representing shuffles

*Definition 16 (Motif).* Let $\mathcal{A}$ be a crystal structure and $\tilde{L}_A$ a sublattice of $\mathcal{A}$. We say that $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\} \subset \mathcal{A}$ is an $\tilde{L}_A$ motif of $\mathcal{A}$ if $\mathbf{a}_1, \ldots, \mathbf{a}_n$ are $\tilde{L}_A$-inequivalent and $n = |\mathcal{A}/\tilde{L}_A|$.

*Theorem 6.* Let $(\tilde{L}_A, \tilde{L}_B, S)$ be an SLM with period $\tilde{Z}$, $\{\mathbf{a}_1, \ldots, \mathbf{a}_{\tilde{Z}}\}$ an $\tilde{L}_A$-motif of $\mathcal{A}$, and $\{\mathbf{b}_1, \ldots, \mathbf{b}_{\tilde{Z}}\}$ an $\tilde{L}_B$-motif of $\mathcal{B}$. If $\mathcal{J} \in \mathrm{CSM}(\mathcal{A}, \mathcal{B})$ has the SLM $(\tilde{L}_A, \tilde{L}_B, S)$, there exist a unique permutation $p \in \mathrm{Sym}(\tilde{Z})$ and unique $\mathbf{t}_1, \ldots, \mathbf{t}_{\tilde{Z}} \in \tilde{L}_B$ such that

$$\forall i \in \{1, \ldots, \tilde{Z}\}, \quad \mathcal{J}(\mathbf{a}_i) = \mathbf{b}_{p(i)} + \mathbf{t}_i. \tag{58}$$

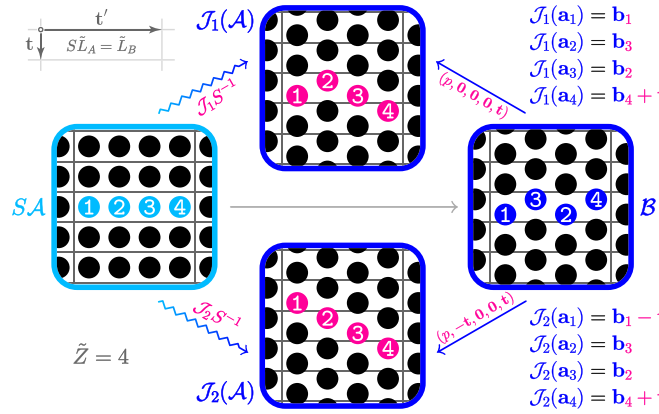


FIG. 6. Two PCTs representing $\mathcal{J}_1, \mathcal{J}_2 \in \mathrm{CSM}(\mathcal{A}, \mathcal{B})$ with the same shuffle lattice (gray grids), respectively. The motifs of $S\mathcal{A}$ (cyan) and $\mathcal{B}$ (blue), denoted by $\{S\mathbf{a}_i\}$ and $\{\mathbf{b}_i\}$, are labeled by 1–4. The shuffle $\mathcal{J}S^{-1}$ maps $S\mathbf{a}_i$ to $\mathcal{J}(\mathbf{a}_i)$ (pink), which is an element of $\mathcal{B}$ and can be written as $\mathbf{b}_{p(i)} + \mathbf{t}_i$. Therefore the PCT representation of a shuffle can be viewed as the unique map $\mathbf{b}_i \mapsto \mathbf{b}_{p(i)} + \mathbf{t}_i$ that makes the triangle in the figure commute, where the horizontal gray arrow represents $S\mathbf{a}_i \mapsto \mathbf{b}_i$ (the canonical bijection between motifs). The pink text shows the specific PCTs representing $\mathcal{J}_1$ and $\mathcal{J}_2$, where $p = \left(\begin{smallmatrix}1 & 2 & 3 & 4\\ 1 & 3 & 2 & 4\end{smallmatrix}\right)$.

Furthermore, if both $\mathcal{J}_1$ and $\mathcal{J}_2$ have $(\tilde{L}_A, \tilde{L}_B, S)$, then $\mathcal{J}_1 = \mathcal{J}_2$ if and only if they have the same $(p, \mathbf{t}_1, \ldots, \mathbf{t}_{\tilde{Z}})$.

*Proof.* In the proof of Theorem 2, we have already shown that $\mathbf{a} + \tilde{L}_A \mapsto \mathcal{J}(\mathbf{a}) + \tilde{L}_B$ is a bijection from $\mathcal{A}/\tilde{L}_A$ to $\mathcal{B}/\tilde{L}_B$. Therefore the images of different $\mathbf{a}_1, \ldots, \mathbf{a}_{\tilde{Z}}$ under $\mathcal{J}$ are in different $\mathbf{b}_1 + \tilde{L}_B, \ldots, \mathbf{b}_{\tilde{Z}} + \tilde{L}_B$. This determines $p$ which is a permutation on $\{1, \ldots, \tilde{Z}\}$, and $\mathbf{t}_1, \ldots, \mathbf{t}_{\tilde{Z}}$ is also uniquely determined by Eq. (58). Furthermore, since each element in $\mathcal{A}$ can be uniquely decomposed as $\mathbf{a} = \mathbf{a}_i + \mathbf{t}$ for some $i \in \{1, \ldots, \tilde{Z}\}$ and $\mathbf{t} \in \tilde{L}_A$, we have

$$\forall \mathbf{a}_i + \mathbf{t} \in \mathcal{A}, \quad \mathcal{J}(\mathbf{a}_i + \mathbf{t}) = \mathbf{b}_{p(i)} + \mathbf{t}_i + S\mathbf{t}, \tag{59}$$

which is derived from Eq. (58) and the following commutative diagram:

$$\begin{array}{ccccc} & & \overset{\mathcal{J}}{\frown} & & \\ \mathcal{A} & \xrightarrow{S} & S\mathcal{A} & \overset{\mathcal{J}S^{-1}}{\rightsquigarrow} & \mathcal{B} \\ {\scriptstyle +\mathbf{t}}\downarrow & & {\scriptstyle +S\mathbf{t}}\downarrow & & \downarrow{\scriptstyle +S\mathbf{t}} \\ \mathcal{A} & \xrightarrow{S} & S\mathcal{A} & \overset{\mathcal{J}S^{-1}}{\rightsquigarrow} & \mathcal{B} \\ & & \underset{\mathcal{J}}{\smile} & & \end{array} \tag{60}$$

From Eq. (59), we can see that $(p, \mathbf{t}_1, \ldots, \mathbf{t}_{\tilde{Z}})$ completely determines a CSM with the SLM $(\tilde{L}_A, \tilde{L}_B, S)$. ■

Given the SLM, Theorem 6 uniquely represents each CSM as a permutation $p$ and $\tilde{Z}$ elements of $\tilde{L}_B$. The underlying reason is that each $(p, \mathbf{t}_1, \ldots, \mathbf{t}_{\tilde{Z}})$ actually represents a shuffle, which we refer to as the "permutation with class-wise translations" (PCT) representation, as illustrated in Fig. 6. We denote by $\mathcal{J}_{p,\mathbf{t}_1,\ldots,\mathbf{t}_{\tilde{Z}}}$ the CSM determined by Eq. (59).

*Definition 17 (PCT).* Given a mapping

$$(\tilde{L}_A, \tilde{L}_B, S) \mapsto (\mathbf{a}_1, \ldots, \mathbf{a}_{\tilde{Z}}, \mathbf{b}_1, \ldots, \mathbf{b}_{\tilde{Z}}), \tag{61}$$

that determines an $\tilde{L}_A$-motif of $\mathcal{A}$ and an $\tilde{L}_B$-motif of $\mathcal{B}$ for each SLM in SLM($\mathcal{A}$, $\mathcal{B}$). We say that

$$\mathrm{PCT}(\tilde{L}_A, \tilde{L}_B, S) = \big\{(p, \mathbf{t}_1, \ldots, \mathbf{t}_{\tilde{Z}}) \in \mathrm{Sym}(\tilde{Z}) \times \tilde{L}_B^{\tilde{Z}} \,\big|\, \forall i,\ \chi_A(\mathbf{a}_i) = \chi_B(\mathbf{b}_{p(i)})\big\} \tag{62}$$

is the set of all PCTs with SLM ($\tilde{L}_A$, $\tilde{L}_B$, $S$), which depends only on the SLM and the mapping.

Note that $p$ can be stored as a permutation matrix, and $\tilde{L}_B$ is isomorphic to $\mathbb{Z}^3$. In this sense, all the ingredients of a CSM can be represented as integer matrices. Given $C_A$, $C_B$, and a deterministic way to choose the $\tilde{L}_A$- and $\tilde{L}_B$-motifs for each SLM, one can reproduce any CSM from its associated IMT and PCT.

### B. Pruning criteria

As a one-to-one correspondence between infinite sets, the CSM has infinitely many possibilities. Fortunately, we are only concerned with the CSMs of those MEPs with low energy barriers. Hence, many CSMs can be excluded, leaving only a finite number of candidates.

#### 1. Strain energy density

From an energetic perspective, the constraint on strain is necessary, since excessive deformation would lead to a large strain energy. Let $w(S)$ denote the strain energy density associated with the deformation gradient $S$. We are only concerned with those CSMs whose deformation gradients satisfy

$$w(S) \leqslant w_{\max}, \tag{63}$$

where $w_{\max}$ is the pruning threshold specified by the user. For example, if one is searching for an MEP with energy barrier lower than that of any known mechanism, $w_{\max}$ can be set to the minimum energy-barrier density among all known mechanisms. This improves the root-mean-square strain (RMSS) pruning criterion we previously used [3], enabling a more reasonable quantification of strains in highly anisotropic materials, e.g., graphite.

Before detailing the definition of $w(S)$, we first outline its rationale. The commutative diagram in Eq. (65) plays a key role: it decomposes the CSM $\mathcal{J}$ into two positive-definite linear transformations, a pure rotation, and a shuffle $\tilde{\mathcal{J}}$. Among these, the pure rotation $UV^{\mathrm{T}}$ makes no contribution to the energy barrier. The other parts respectively determine $w(S)$ and the shuffle distance $d(\tilde{\mathcal{J}})$ to be introduced in the next section. Now we formalize this idea.

*Lemma 7 (SVD).* If $S \in \mathbb{R}^{3\times3}$ has a positive determinant, there exist $U, V \in \mathrm{SO}(3)$ and unique $s_1 \geqslant s_2 \geqslant s_3 > 0$ such that $S = U\Sigma V^{\mathrm{T}}$, where $\Sigma = \mathrm{diag}(s_1, s_2, s_3)$.

Lemma 7 is known as the singular value decomposition (SVD) and $s_i$ is referred to as the $i$-th singular value of $S$ [44], usually denoted by $\sigma_i(S)$. A direct corollary is the polar decomposition (PD), which states that a deformation gradient can always be *uniquely* factorized into a positive-definite transformation and a rotation, as

$$S = (UV^{\mathrm{T}})(V\Sigma V^{\mathrm{T}}) = (S\sqrt{S^{\mathrm{T}}S}^{-1})\sqrt{S^{\mathrm{T}}S}. \tag{64}$$

Note that the rotation $UV^{\mathrm{T}}$ does not contribute to the strain energy, and the positive-definite $V\Sigma V^{\mathrm{T}}$ simply scales the space by $s_i$ along $\mathbf{v}_i$ for $i = 1, 2, 3$, where $\mathbf{v}_i$ denotes the $i$th column of $V$. We estimate the strain energy according to the diagram:

$$\begin{array}{ccc} \mathcal{A} & \xrightarrow{\quad\mathcal{J}\quad} & \mathcal{B} \\ {\scriptstyle\sqrt{V\Sigma V^{\mathrm{T}}}}\downarrow & & \uparrow{\scriptstyle\sqrt{U\Sigma U^{\mathrm{T}}}} \\ V\Sigma^{\frac{1}{2}}V^{\mathrm{T}}\mathcal{A} & \overset{\tilde{\mathcal{J}}}{\rightsquigarrow} & V\Sigma^{-\frac{1}{2}}U^{\mathrm{T}}\mathcal{B} \xrightarrow[UV^{\mathrm{T}}]{} U\Sigma^{-\frac{1}{2}}U^{\mathrm{T}}\mathcal{B} \end{array} \tag{65}$$

where $\tilde{\mathcal{J}}$ is defined as the unique shuffle that makes the diagram commute, which we refer to as the *standard shuffle* of $\mathcal{J}$. Note that the two vertical arrows are pure deformations, while all horizontal arrows in the second row involve no strain. Based on this, $w$ is estimated as the sum of the strain energies associated with the deformation of $\mathcal{A}$ by $\sqrt{V\Sigma V^{\mathrm{T}}}$ and the deformation of $\mathcal{B}$ by $\sqrt{U\Sigma U^{\mathrm{T}}}^{-1}$, namely,

$$w(S) = \frac{1}{2}\sum_{i,j,k,l=1}^{3}\Big\{Y^A_{ijkl}\big[V\big(\Sigma^{\frac{1}{2}} - I\big)V^{\mathrm{T}}\big]_{ij}\big[V\big(\Sigma^{\frac{1}{2}} - I\big)V^{\mathrm{T}}\big]_{kl} + Y^B_{ijkl}\big[U\big(\Sigma^{-\frac{1}{2}} - I\big)U^{\mathrm{T}}\big]_{ij}\big[U\big(\Sigma^{-\frac{1}{2}} - I\big)U^{\mathrm{T}}\big]_{kl}\Big\} \tag{66}$$

$$= \frac{1}{2}\sum_{a,b=1}^{3}\sum_{i,j,k,l=1}^{3}\left[(\sqrt{s_a} - 1)(\sqrt{s_b} - 1)Y^A_{ijkl}V_{ia}V_{ja}V_{kb}V_{lb} + \left(\frac{1}{\sqrt{s_a}} - 1\right)\left(\frac{1}{\sqrt{s_b}} - 1\right)Y^B_{ijkl}U_{ia}U_{ja}U_{kb}U_{lb}\right], \tag{67}$$

where $Y^A_{ijkl}$ and $Y^B_{ijkl}$ denote the elastic (stiffness) tensors of $\mathcal{A}$ and $\mathcal{B}$, respectively.

Substituting Eq. (67) into Eq. (63) yields the inequality that $S$ must satisfy, which depends solely on $Y^A_{ijkl}$, $Y^B_{ijkl}$, and $w_{\max}$. It is worth noting that $w$ is a *congruence-class function*, i.e.,

$$\forall R_A \in G'_A, \quad \forall R_B \in G'_B, \quad w\big(R_B S R_A^{-1}\big) = w(S). \tag{68}$$

This can be seen from Lemma 7, where $R_A$ and $R_B$ only contributes to $V$ and $U$, respectively. As symmetry operations of $\mathcal{A}$ and $\mathcal{B}$, the elastic tensors $Y^A_{ijkl}$ and $Y^B_{ijkl}$ are invariant under $R_A$ and $R_B$, respectively. Hence, SLMs are pruned by congruence classes under Eq. (63).

#### 2. Shuffle distance

For each SLM, there are still infinitely many possible CSMs—even a single atom $\mathbf{a} \in \mathcal{A}$ has infinitely many choices of its counterpart $\mathcal{J}(\mathbf{a})$. An intuitive constraint is that $\mathcal{J}(\mathbf{a})$ should not be too far from $\mathbf{a}$. This is not an energetic consideration, but rather an empirical one, as discussed in Ref. [25], where the average distance traveled by the atoms is shown to

be a key indicator. However, denoting $\mathcal{A}$ as $\{\mathbf{a}_1, \mathbf{a}_2, \dots\}$, we can always find that the average distance diverges as

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} |\mathcal{J}(\mathbf{a}_i) - \mathbf{a}_i| = \infty, \tag{69}$$

unless the deformation gradient of $\mathcal{J}$ equals $I$, i.e., $\mathcal{J}$ is a shuffle [2]. Therefore the appropriate way to define the "average distance" is to consider a shuffle associated with $\mathcal{J}$. Here, we use the standard shuffle $\tilde{\mathcal{J}}$ defined in Eq. (65) instead of $\mathcal{J}S^{-1}$ to ensure the metric symmetry between the initial and final structures. As shorthands, denote by $\tilde{L}$ the shuffle lattice of $\tilde{\mathcal{J}}$ and let $\tilde{\mathcal{A}} = V\Sigma^{\frac{1}{2}}V^{\mathrm{T}}\mathcal{A}$ and $\tilde{\mathcal{B}} = V\Sigma^{-\frac{1}{2}}U^{\mathrm{T}}\mathcal{B}$, which are all uniquely determined by the SLM.

*Definition 18 (Shuffle distance).* Let $\mathcal{J} \in \mathrm{CSM}(\mathcal{A}, \mathcal{B})$ be a CSM with period $\tilde{Z}$. We say that the shuffle distance of $\mathcal{J}$ is

$$\hat{d}(\mathcal{J}) = \left( \sum_{i=1}^{\tilde{Z}} \theta_i |\tilde{\mathcal{J}}(\tilde{\mathbf{a}}_i) - \tilde{\mathbf{a}}_i|^{\ell} \right)^{1/\ell}, \tag{70}$$

where $\ell \geqslant 1$ specifies the norm, $\{\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_{\tilde{Z}}\}$ is an $\tilde{L}$-motif of $\tilde{\mathcal{A}}$, and $\theta_i > 0$ is a normalized weight that only depends on the atomic species of $\tilde{\mathbf{a}}_i$ (e.g., a constant, or atomic mass). One may adopt $\ell = 2$ for the root-mean-square distance (RMSD), or $\ell = 1$ for the arithmetic mean distance.

Equation (9) ensures the value of Eq. (70) to be independent of the choice of the motif so that $\hat{d}$ is well-defined. However, congruent CSMs may not necessarily have the same $\hat{d}$, which is undesirable since they are equivalent as inputs to NEB-like methods. To address this issue, we define

$$d(\mathcal{J}) = \min_{\boldsymbol{\tau}\in\mathbb{R}^3} \left( \sum_{i=1}^{\tilde{Z}} \theta_i |\tilde{\mathcal{J}}(\tilde{\mathbf{a}}_i) + \boldsymbol{\tau} - \tilde{\mathbf{a}}_i|^{\ell} \right)^{1/\ell}, \tag{71}$$

which equals the minimum $\hat{d}$ among all CSMs of the form $(+\boldsymbol{\tau}_B) \circ \mathcal{J} \circ (-\boldsymbol{\tau}_A) \in \mathrm{CSM}(\mathcal{A} + \boldsymbol{\tau}_A, \mathcal{B} + \boldsymbol{\tau}_B)$. Similar to the proof of Theorem 3, one can show that any $\mathcal{J}'$ congruent to $\mathcal{J}$ has a standard shuffle that differs from $(R, \mathbf{t})\tilde{\mathcal{J}}(R, \mathbf{t})^{-1}$ by just a translation, implying that

$$d(\mathcal{J}') = d[(R, \mathbf{t})\tilde{\mathcal{J}}(R, \mathbf{t})^{-1}] = d(\tilde{\mathcal{J}}) = d(\mathcal{J}), \tag{72}$$

where we have used the fact that $(R, \mathbf{t})$ is an isometry. Equation (72) means that congruent CSMs have the same $d$, so that as long as $d(\mathcal{J})$ exceeds the threshold $d_{\max}$, we can exclude all CSMs congruent to $\mathcal{J}$.

### 3. Finiteness of candidate crystal-structure matches

As shown in Fig. 2, the enumeration includes a CSM $\mathcal{J}$ with multiplicity $\mu$ and deformation gradient $S$ if

$$\mu \leqslant \mu_{\max}, \tag{73}$$

$$w(S) \leqslant w_{\max}, \tag{74}$$

$$d(\mathcal{J}) \leqslant d_{\max}. \tag{75}$$

They have all been detailed earlier, except for Eq. (73), which can be regarded as an empirical constraint inferred from Table II—despite MD and MetaD simulations using far more than $10^3$ atoms, the intricacy of the CSMs they produce remain $\tilde{Z} \sim 10^1$. We emphasize that $w_{\max}$ can depend on $\mu$, and $d_{\max}$ can depend on the SLM of $\mathcal{J}$. For example, one might want to slacken the constraints on $w_{\max}$ and $d_{\max}$ when $\mu$ is small. One may also set $d_{\max}$ for each SLM to 110% of the lowest $d$ value among all CSMs with that SLM. In a word, different subtrees in Fig. 2 can be pruned with different thresholds. We denote by $\mathrm{CSM}^{\star}(\mathcal{A}, \mathcal{B})$ the set of all CSMs in $\mathrm{CSM}(\mathcal{A}, \mathcal{B})$ that satisfy Eqs. (73)–(75), and their SLMs by $\mathrm{SLM}^{\star}(\mathcal{A}, \mathcal{B})$.

*Theorem 7.* $\mathrm{SLM}^{\star}(\mathcal{A}, \mathcal{B}; \mu)$ is finite regardless of the values of $\mu$ and $w_{\max}$.

*Proof.* There exists a $\sigma_{\max}$ such that $\sigma_1(S) < \sigma_{\max}$ implies Eq. (74); otherwise, infinite strain would leads to infinite strain energy. We only need to prove that only a finite number of $(H_A, H_B, Q) \in \mathrm{IMT}(\mu)$ satisfy $\sigma_1(S) < \sigma_{\max}$. Since $\mathrm{HNF}(k)$ is finite, the number of possible values of $(H_A, H_B)$ is already finite. It is sufficient to show that given $(H_A, H_B)$, the number of possible values of $Q$ is finite. This is due to the fact that the absolute value of any matrix element of $Q$ does not exceed

$$\sigma_1(Q) = \sigma_1\left(H_B^{-1}C_B^{-1}SC_AH_A\right) \tag{76}$$

$$\leqslant \sigma_1\left(H_B^{-1}\right)\sigma_1\left(C_B^{-1}\right)\sigma_1(C_A)\sigma_1(H_A)\,\sigma_{\max}, \tag{77}$$

where Eq. (76) is derived from Eq. (53) ■

*Theorem 8.* For any given SLM and $d_{\max}$, the number of noncongruent CSMs that have this SLM and satisfy Eq. (75) is finite.

*Proof.* We assume by contradiction that there are infinitely many noncongruent CSMs in $\mathrm{CSM}^{\star}(\mathcal{A}, \mathcal{B})$ that have $(\tilde{L}_A, \tilde{L}_B, S)$. Among them, each $\mathcal{J}_{p,\mathbf{t}_1,\mathbf{t}_2,\dots,\mathbf{t}_{\tilde{Z}}}$ is congruent to a $\mathcal{J}_{p,\mathbf{0},\mathbf{t}'_2,\dots,\mathbf{t}'_{\tilde{Z}}}$, where $\mathbf{t}'_i = \mathbf{t}_i - \mathbf{t}_1$. Since $p$ can take only a finite number of values, there are infinitely many PCTs having the same $p$ and different $(\mathbf{0}, \mathbf{t}'_2, \dots, \mathbf{t}'_{\tilde{Z}})$ that satisfy

$$d\left(\mathcal{J}_{p,\mathbf{0},\mathbf{t}'_2,\dots,\mathbf{t}'_{\tilde{Z}}}\right) \leqslant d_{\max}. \tag{78}$$

These $(\mathbf{0}, \mathbf{t}'_2, \dots, \mathbf{t}'_{\tilde{Z}})$ form an infinite subset of $\tilde{L}_B^{\tilde{Z}}$, so there exist $j$ such that $|\mathbf{t}'_j|$ is unbounded; otherwise, the subset would be finite. Let

$$y = |\tilde{\mathbf{b}}_{p(j)} + \tilde{\mathbf{t}}'_j - \tilde{\mathbf{b}}_{p(1)} - \tilde{\mathbf{a}}_j + \tilde{\mathbf{a}}_1|, \tag{79}$$

which is also unbounded. We have

$$\left[d\left(\mathcal{J}_{p,\mathbf{0},\mathbf{t}'_2,\dots,\mathbf{t}'_{\tilde{Z}}}\right)\right]^{\ell}$$

$$= \min_{\boldsymbol{\tau}\in\mathbb{R}^3} \sum_{i=1}^{\tilde{Z}} \theta_i |\tilde{\mathbf{b}}_{p(i)} + \tilde{\mathbf{t}}'_i + \boldsymbol{\tau} - \tilde{\mathbf{a}}_i|^{\ell} \tag{80}$$

$$\geqslant \min_{\boldsymbol{\tau}\in\mathbb{R}^3} (\theta_1 |\tilde{\mathbf{b}}_{p(1)} + \boldsymbol{\tau} - \tilde{\mathbf{a}}_1|^{\ell} + \theta_j |\tilde{\mathbf{b}}_{p(j)} + \tilde{\mathbf{t}}'_j + \boldsymbol{\tau} - \tilde{\mathbf{a}}_j|^{\ell}) \tag{81}$$

$$= \min_{\boldsymbol{\tau}\in\mathbb{R}^3} (\theta_1 |\tilde{\mathbf{b}}_{p(1)} + \boldsymbol{\tau} - \tilde{\mathbf{a}}_1|^{\ell} + \theta_j |(\tilde{\mathbf{b}}_{p(1)} + \boldsymbol{\tau} - \tilde{\mathbf{a}}_1) + (\tilde{\mathbf{b}}_{p(j)} + \tilde{\mathbf{t}}'_j - \tilde{\mathbf{b}}_{p(1)} - \tilde{\mathbf{a}}_j + \tilde{\mathbf{a}}_1)|^{\ell}) \tag{82}$$

$$\geqslant \min_{\boldsymbol{\tau}\in\mathbb{R}^3} (\theta_1 |\tilde{\mathbf{b}}_{p(1)} + \boldsymbol{\tau} - \tilde{\mathbf{a}}_1|^{\ell} + \theta_j ||\tilde{\mathbf{b}}_{p(1)} + \boldsymbol{\tau} - \tilde{\mathbf{a}}_1| - |\tilde{\mathbf{b}}_{p(j)} + \tilde{\mathbf{t}}'_j - \tilde{\mathbf{b}}_{p(1)} - \tilde{\mathbf{a}}_j + \tilde{\mathbf{a}}_1||^{\ell}) \tag{83}$$

$$= \min_{x\geqslant 0} (\theta_1 x^{\ell} + \theta_j |x - y|^{\ell})$$

$$= \begin{cases} y \min\{\theta_1, \theta_j\}, & \text{if } \ell = 1 \\ y^{\ell}\left(\theta_1^{-\ell+1} + \theta_j^{-\ell+1}\right)^{-\ell+1}, & \text{if } \ell > 1 \end{cases}. \tag{84}$$

ALGORITHM 1. Listing all noncongruent IMTs.

**Input:** $C_A$, $C_B$, $Z_A$, $Z_B$, $\mu$, $w_{\max}$
initialize myimts $\leftarrow \{\}$; $i \leftarrow 0$
**while** $i \leq i^*$ **do**
  generate a random $S_0 \in \Omega_{S_0}$
  **foreach** $H_A \in \mathrm{HNF}\left(\frac{\mu \operatorname{lcm}(Z_A,Z_B)}{Z_A}\right)$ **do**
    **foreach** $H_B \in \mathrm{HNF}\left(\frac{\mu \operatorname{lcm}(Z_A,Z_B)}{Z_B}\right)$ **do**
      $Q \leftarrow \mathrm{round}(H_B^{-1}C_B^{-1}S_0C_AH_A)$
      $S \leftarrow C_BH_BQH_A^{-1}C_A^{-1}$
      **if** $Q \in \mathrm{GL}(3,\mathbb{Z})$ **and** $w(S) \leq w_{\max}$ **then**
        $x \leftarrow$ `Standardize`$(H_A, H_B, Q)$
        **if** $x \notin$ myimts **then**
          add $x$ to myimts
          $i \leftarrow 0$
        **else** $i \leftarrow i+1$
**return** myimts

Since $y$ is unbounded, Eq. (84) can be arbitrarily large, which contradicts Eq. (78). ■

*Theorem 9.* The number of noncongruent CSMs in $\mathrm{CSM}^\star(\mathcal{A},\mathcal{B})$ is finite regardless of $\mu_{\max}$, $w_{\max}$, and $d_{\max}$.

*Proof.* Replace the root node "all CSMs" in Fig. 2 with "all noncongruent CSMs." Theorems 7 and 8 ensure that the breadth of the pruned tree is finite at each level. ■

## IV. ALGORITHMS

To enumerate all noncongruent CSMs in $\mathrm{CSM}^\star(\mathcal{A},\mathcal{B})$, Theorem 4 allows us to enumerate all noncongruent SLMs in $\mathrm{SLM}^\star(\mathcal{A},\mathcal{B})$ first, and then all noncongruent CSMs within each SLM. With Theorems 5 and 6, we only need to generate a complete subset of

$$\mathrm{IMT}^\star(\mu) = \left\{(H_A,H_B,Q) \in \mathrm{IMT}(\mu) \,\middle|\, w\left(C_BH_BQH_A^{-1}C_A^{-1}\right) \leqslant w_{\max}\right\} \tag{85}$$

for $\mu = 1,\ldots,\mu_{\max}$, and a complete subset of

$$\mathrm{PCT}^\star(\tilde{L}_A,\tilde{L}_B,S) = \left\{(p,\mathbf{t}_1,\ldots,\mathbf{t}_{\tilde{Z}}) \in \mathrm{PCT}(\tilde{L}_A,\tilde{L}_B,S) \,\middle|\, d\left(\mathcal{J}_{p,\mathbf{t}_1,\ldots,\mathbf{t}_{\tilde{Z}}}\right) \leqslant d_{\max}\right\} \tag{86}$$

for each $(\tilde{L}_A,\tilde{L}_B,S) \in \mathrm{SLM}^\star(\mathcal{A},\mathcal{B};\mu)$.

In this section, we introduce Algorithms 1–3 that form the core of CRYSTMATCH. Their objectives and performance are summarized in Table IV. These algorithms do not involve electronic structure or atomic force field calculations, i.e., they are purely geometric. This enables a quick overview of all possible deformations and CSMs of an SSPT. The pseudocodes below are presented in an easy-to-understand form, but the actual implementation can be more efficient, e.g., all loops in Algorithm 1 can be parallelized.

### A. Enumerating sublattice matches

Let's first consider how to produce the entire $\mathrm{IMT}^\star(\mu)$ rather than its complete subset. It is easy to enumerate all possible values of $H_A$ and $H_B$, so the difficulty lies solely in

ALGORITHM 2. Listing all noncongruent PCTs.

**Input:** $\mathcal{A}$, $\mathcal{B}$, $(\tilde{L}_A,\tilde{L}_B,S)$, $d_{\max}$
initialize mypcts $\leftarrow \{\}$; myconstraints $\leftarrow$ empty queue
enqueue $\{\}$ to myconstraints
**while** myconstraints is not empty **do**
  dequeue constraint from myconstraints
  $d,(p,\mathbf{t}_1,\ldots,\mathbf{t}_{\tilde{Z}}) \leftarrow$ `OptimizePCT(constraint)`
  **if** $d \leq d_{\max}$ **then**
    **foreach** $(\mathbf{t}'_2,\ldots,\mathbf{t}'_{\tilde{Z}}) \in$ `Fill`$(p)$ **do**
      add $(p,\mathbf{0},\mathbf{t}'_2,\ldots,\mathbf{t}'_{\tilde{Z}})$ to mypcts
    **foreach** newcons $\in$ `Split(constraint,`$p$`)` **do**
      enqueue newcons to myconstraints
**return** mypcts

computing all possible $Q$ for a given HNF pair $(H_A,H_B)$, i.e., producing $\{Q \,|\, (H_A,H_B,Q) \in \mathrm{IMT}^\star(\mu)\}$, which we denote by $\{Q\}$ for simplicity. To this end, we first sample the 9D region

$$\Omega_{S_0} = \{S_0 \in \mathbb{R}^{3\times3} \,|\, w(S_0) \leqslant w_{\max}\}, \tag{87}$$

producing a finite set consisting of *trial* matrices, denoted as $\{S_0\} \subset \Omega_{S_0}$. Then, for each trial matrix $S_0$, we compute the closest IMTs by (1) substituting $S_0$ into Eq. (53), (2) solve for $Q$ for each pair $(H_A,H_B)$, and (3) round each matrix element of $Q$ to the nearest integer. In other words, we apply the linear mapping

$$\phi_{H_A,H_B} : \mathbb{R}^{3\times3} \to \mathbb{R}^{3\times3}, \tag{88}$$

$$S_0 \mapsto H_B^{-1}C_B^{-1}S_0C_AH_A, \tag{89}$$

to the set $\{S_0\}$. As long as $\{S_0\}$ is sampled dense enough, its image under $\phi_{H_A,H_B}$ will intersect each $Q$'s "preimage of rounding"

$$\mathrm{round}^{-1}(Q) = \left\{Q+X \,\middle|\, X \in \left(-\tfrac{1}{2},\tfrac{1}{2}\right)^{3\times3}\right\}, \tag{90}$$

a 9D hypercube centered at $Q$ [Fig. 7(b)]. In that case, we can extract $\{Q\}$ from $\mathrm{round}[\phi_{H_A,H_B}(\{S_0\})]$.

The only remaining problem is how to tell whether the set $\{S_0\}$ is dense *enough* such that $\{Q\}$ is covered by its image under $\phi_{H_A,H_B}$. Our approach is to generate $S_0$ one by one until $i^*$ consecutive iterations fail to produce any new IMT, as described in Algorithm 1. The value of $i^*$ can be derived by hypothesis testing. Whenever the algorithm starts or a new IMT is produced, we initiate a test of the two opposing hypotheses, namely

$H_0$ :The current myimts is complete,

$H_1$ :The current myimts is incomplete.

Then, when we perform $i$ iterations without producing any new IMT, the conditional probability of this event under $H_0$ is 1, whereas under $H_1$ it is only

$$\Pr(i \,|\, H_1) \leqslant (1-\eta)^i, \tag{91}$$

where $\eta$ denotes the minimum probability of a given IMT being generated in a single iteration. If one would reject $H_1$ when $\Pr(i \,|\, H_1) < \epsilon$, it is sufficient to set

$$i^* = \frac{\ln\epsilon}{\ln(1-\eta)}. \tag{92}$$

ALGORITHM 3. Optimizing PCT (minimizing shuffle distance) under given constraint.

**Input:** constraint, and all inputs of Algorithm 2
$\mathbf{a}_1^{(1)},\dots,\mathbf{a}_{\tilde{Z}_1}^{(1)},\mathbf{a}_1^{(2)},\dots,\mathbf{a}_{\tilde{Z}_2}^{(2)},\dots,\mathbf{a}_1^{(n)},\dots,\mathbf{a}_{\tilde{Z}_n}^{(n)},\mathbf{b}_1^{(1)},\dots,\mathbf{b}_{\tilde{Z}_1}^{(1)},\mathbf{b}_1^{(2)},\dots,\mathbf{b}_{\tilde{Z}_2}^{(2)},\dots,\mathbf{b}_1^{(n)},\dots,\mathbf{b}_{\tilde{Z}_n}^{(n)} \leftarrow$ `Motifs`$(\tilde{L}_A,\tilde{L}_B,S)$
**def** `ShuffleDistance`$(\boldsymbol{\tau},$ `returnpct` $= false)$
  **for** $\alpha=1$ **to** $n$ **do**
    **for** $i=1$ **to** $\tilde{Z}_\alpha$ **do**
      **for** $j=1$ **to** $\tilde{Z}_\alpha$ **do**
        $\mathbf{t}_{ij}^{(\alpha)} \leftarrow \operatorname{argmin}_{\mathbf{t}\in\tilde{L}_B}\left|\tilde{\mathbf{b}}_j^{(\alpha)}+\tilde{\mathbf{t}}+\boldsymbol{\tau}-\tilde{\mathbf{a}}_i^{(\alpha)}\right|^\ell$; $\quad D_{ij}^{(\alpha)} \leftarrow \left|\tilde{\mathbf{b}}_j^{(\alpha)}+\tilde{\mathbf{t}}_{ij}^{(\alpha)}+\boldsymbol{\tau}-\tilde{\mathbf{a}}_i^{(\alpha)}\right|^\ell$
    $p^{(\alpha)} \leftarrow$ `BestAssignment`$(D^{(\alpha)},$ `constraint`$)$; $\quad$ **for** $i=1$ **to** $\tilde{Z}_\alpha$ **do** $\mathbf{t}_i^{(\alpha)} \leftarrow \mathbf{t}_{i,p^{(\alpha)}(i)}^{(\alpha)}$
  $d \leftarrow \left(\sum_{\alpha=1}^n\sum_{i=1}^{\tilde{Z}_\alpha}\theta_i^{(\alpha)}D_{i,p^{(\alpha)}(i)}^{(\alpha)}\right)^{1/\ell}$
  **if** `returnpct` **then return** $d,\left((p^{(1)}|p^{(2)}|\cdots|p^{(n)}),\mathbf{t}_1^{(1)},\dots,\mathbf{t}_{\tilde{Z}_1}^{(1)},\mathbf{t}_1^{(2)},\dots,\mathbf{t}_{\tilde{Z}_2}^{(2)},\dots,\mathbf{t}_1^{(n)},\dots,\mathbf{t}_{\tilde{Z}_n}^{(n)}\right)$
  **else return** $d$
$\boldsymbol{\tau}^* \leftarrow \operatorname{argmin}_{\boldsymbol{\tau}\in\mathbb{R}^3}$ `ShuffleDistance`$(\boldsymbol{\tau})$; $\quad$ **return** `ShuffleDistance`$(\boldsymbol{\tau}^*,$ `returnpct` $= true)$

Note that the probability $\epsilon$ of incorrectly asserting completeness decreases exponentially with increasing $i^*$. In this sense, we claim that Algorithm 1 is essentially an *exhaustive* search algorithm.

Recall that our goal is to list all noncongruent elements in IMT$^\star(\mu)$. We can treat each IMT as its congruence class—whenever an IMT is generated, if it is congruent to some element already in myimts, it is considered not new and $i$ increases. However, a better way might be using a `Standardize()` function to convert each IMT into the representative of its congruence class (e.g., the lexicographically smallest one), as shown in the pseudocode. Either way will replace $1-\eta$ in Eq. (92) with the minimum probability that no IMT from a given congruence class is generated in a single iteration. A conservative estimate of $i^*$ is provided in Appendix B, while in practice, $i^*$ can be dynamically adjusted to accelerate the algorithm. When the number of SLMs $N_{\text{SLM}}$ is large, most of the runtime comes from discovering new SLMs (which resets $i$ to zero), rather than from the $i^*$ iteration after all SLMs have been found. Therefore Algorithm 1 has the same time complexity as the coupon collector's problem, i.e., $\mathcal{O}(N_{\text{SLM}}\log N_{\text{SLM}})$.

### B. Enumerating shuffles

To list all noncongruent PCTs, Algorithm 2 involves three functions: (1) `OptimizePCT()` for finding the PCT with the minimum $d$, (2) `Fill()` for generating all noncongruent PCTs with a given $p$, and (3) `Split()` for excluding previously enumerated values of $p$. All we need to do is to iteratively invoke them, as shown in Fig. 8. Next, we will introduce the specific algorithms for implementing these three functions separately.

We use `OptimizePCT(`constraint`)` to generate the PCT with the smallest $d$ under the constraints specified by constraint. In each iteration step, the overall translation $\boldsymbol{\tau}$ is fixed, and the best translation $\mathbf{t}_{ij}\in\tilde{L}_B$ for each atom pair $(\mathbf{a}_i,\mathbf{b}_j)$ is computed. Then, the best permutation $p$ under this $\boldsymbol{\tau}$ is determined using `BestAssignment()` (e.g., the Jonker-Volgenant algorithm [45,46]), so that $\mathbf{t}_1,\dots,\mathbf{t}_{\tilde{Z}}$ and

$$\hat{d}_0(\boldsymbol{\tau}) = \min_{p,\mathbf{t}_1,\dots,\mathbf{t}_{\tilde{Z}}} \hat{d}\left[(+\boldsymbol{\tau})\circ\mathcal{J}_{p,\mathbf{t}_1,\dots,\mathbf{t}_{\tilde{Z}}}\right] \tag{93}$$

are also obtained. This process is iterated with different $\boldsymbol{\tau}$ values until the global minimum of $\hat{d}_0(\boldsymbol{\tau})$ is found, where strategies such as basin-hopping are employed [47]. The final iteration uses $\boldsymbol{\tau}^*=\operatorname{argmin}_{\boldsymbol{\tau}}\hat{d}_0(\boldsymbol{\tau})$, at which point the PCT given by `BestAssignment()` is

$$\operatorname*{argmin}_{p,\mathbf{t}_1,\dots,\mathbf{t}_{\tilde{Z}}} \hat{d}\left[(+\boldsymbol{\tau}^*)\circ\mathcal{J}_{p,\mathbf{t}_1,\dots,\mathbf{t}_{\tilde{Z}}}\right] \tag{94}$$

$$= \operatorname*{argmin}_{p,\mathbf{t}_1,\dots,\mathbf{t}_{\tilde{Z}}} \min_{\boldsymbol{\tau}\in\mathbb{R}^3} \hat{d}\left[(+\boldsymbol{\tau})\circ\mathcal{J}_{p,\mathbf{t}_1,\dots,\mathbf{t}_{\tilde{Z}}}\right] \tag{95}$$

$$= \operatorname*{argmin}_{p,\mathbf{t}_1,\dots,\mathbf{t}_{\tilde{Z}}} d\left(\mathcal{J}_{p,\mathbf{t}_1,\dots,\mathbf{t}_{\tilde{Z}}}\right), \tag{96}$$

which is exactly what we need.

TABLE IV. Summary of the core algorithms. $\tilde{Z}_\alpha$ denotes the number of atoms with the $\alpha$-th species in a shuffle period.

| | | Algorithm 1 | Algorithm 2 | Algorithm 3 |
|---|---|---|---|---|
| Objective | | Enumerate all SLMs | Enumerate all shuffles | Optimize the shuffle |
| Time complexity | | $\mathcal{O}(N_{\text{SLM}}\log N_{\text{SLM}})$ | $\mathcal{O}(N_{\text{shuffle}}\sum_\alpha\tilde{Z}_\alpha^3)$ | $\mathcal{O}(\sum_\alpha\tilde{Z}_\alpha^3)$ |
| Runtime[a] | Fig. 9 | 1 min 11 s (788 SLMs) | – | 13 s (393 shuffles) |
| | Fig. 10 | 2 h 7 min 18 s (9987 SLMs) | – | 1 h 10 min 36 s (5432 shuffles) |
| | Fig. 12 | 1 min 57 s (573 SLMs) | 3 h 12 min 26 s (320 757 shuffles) | – |
| | Fig. 14 | 9 min 29 s (1380 SLMs) | – | 37 s (881 shuffles) |

[a]All runtimes were measured on a single core of a Mac M1 chip. Due to implementation details, the performance of Algorithm 1 differs when using $w(S)\leqslant w_{\max}$ and $\overline{\varepsilon}(S)\leqslant\overline{\varepsilon}_{\max}$.

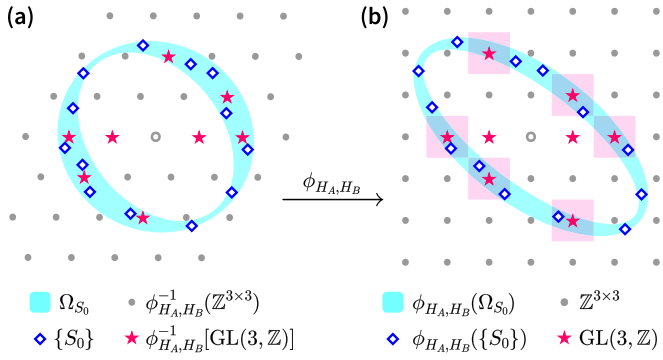


FIG. 7. A schematic diagram showing how Algorithm 1 works, where 9D regions are drawn as 2D shapes for intuitive understanding. (a) $\Omega_{S_0}$ (cyan) is a neighborhood of SO(3) determined solely by $Y^A_{ijkl}$, $Y^B_{ijkl}$ and $w_{\max}$. The trail matrices $\{S_0\}$ (blue diamonds) are generated in $\Omega_{S_0}$, approximating the exact deformation gradients (red stars). (b) $\phi_{H_A,H_B}$ is applied to $\{S_0\}$, and then each $\phi_{H_A,H_B}(S_0)$ is rounded to the nearest integer matrix (gray dots). Each element of $\{Q\}$ (red stars in cyan region) is produced if and only if a $S_0$ lies within its "preimage of rounding" (pink squares). Currently, the algorithm should continue to generate a denser $\{S_0\}$ to produce the topmost $Q$ in the figure.

To formalize this idea and handle atomic species, the pseodocode is presented in Algorithm 3. The function `Motifs`$(\tilde{L}_A, \tilde{L}_B, S)$ is a mapping like Eq. (61) but also sorts the atoms by atomic species, where $n$ denotes the number of atomic species in the initial and final structures, $\tilde{Z}_\alpha$ denotes the number of atoms with the $\alpha$-th species, and other variables related to the $\alpha$-th atomic species are denoted with superscripts as $\mathbf{a}_i^{(\alpha)}$. The function `BestAssignment`$(D,$ constraint$)$ returns a permutation $p$ that minimizes $\sum_i D_{i,p(i)}$, where $D$ is the cost matrix and constraint is a set of the form

$$\{(i,j),(i',j'),\cdots;\overline{(k,l)},\overline{(k',l')},\cdots\}, \tag{97}$$

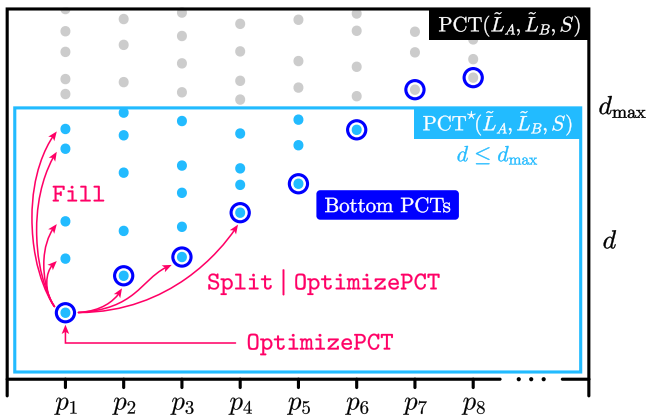


FIG. 8. A schematic diagram showing how Algorithm 2 works, where each point represents a PCT. First, `OptimizePCT()` is invoked to obtain the PCT with the smallest $d$. Then, `Fill()` generates all PCTs that share the same permutation with it, while `Split()` produces a set of constraints to disable this permutation in subsequent computations. By passing these constraints as parameters to `OptimizePCT()`, other bottom PCTs (dots in blue circles) can be generated.

where $(i,j)$ means enforcing $p(i)=j$, and $\overline{(k,l)}$ means requiring $p(k)\neq l$ [48]. The correspondence of each atomic species is optimized separately and then combined into a single PCT, where $(p^{(1)}|p^{(2)}|\cdots|p^{(n)})\in\mathrm{Sym}(\tilde{Z})$ represents performing $n$ permutations separately within $n$ atomic species. Consequently, the time complexity of Algorithm 3 is $\mathcal{O}(\sum_\alpha \tilde{Z}_\alpha^3)$ [46], which also determines the time complexity of Algorithm 2.

We say that $(p,\mathbf{t}_1,\ldots,\mathbf{t}_{\tilde{Z}})$ is a bottom PCT if its $d$ is the smallest among all PCTs with the same permutation. Whenever we obtain a bottom PCT, we invoke `Fill`$(p)$ to list all PCTs of the form $(p,\mathbf{0},\mathbf{t}'_2,\ldots,\mathbf{t}'_{\tilde{Z}})$ that satisfy Eq. (78). Here, we have let $\mathbf{t}'_1=\mathbf{0}$ to make the number of PCTs finite without loss of completeness (see the proof of Lemma 8). Since $\mathbf{v}_i=\tilde{\mathbf{b}}_{p(i)}-\tilde{\mathbf{a}}_i$ is constant during each invocation of `Fill()`, one may simply use a flood fill algorithm to produce

$$\left\{(\mathbf{t}'_2,\ldots,\mathbf{t}'_{\tilde{Z}})\in\tilde{L}_B^{\tilde{Z}-1}\,\middle|\,\min_{\boldsymbol{\tau}\in\mathbb{R}^3}\sum_{i=1}^{\tilde{Z}}\theta_i|\mathbf{v}_i+\tilde{\mathbf{t}}_i+\boldsymbol{\tau}|^\ell\leqslant d_{\max}^\ell\right\}. \tag{98}$$

When $\ell=2$, Lemma 8 implies that $(\mathbf{t}_2,\ldots,\mathbf{t}'_{\tilde{Z}})$ takes values in a known hyperellipsoid so that one may employ a more efficient algorithm.

The last function `Split`(constraint, $p$) excludes the solution $p$ by splitting the assignment problem that satisfies constraint into several subproblems, which was originally proposed in Ref. [48]. It incorporates "excluding $p$" into constraint by returning $\tilde{Z}-1$ mutually disjoint sets

$$\text{constraint}\cup\{\overline{(1,p(1))}\}, \tag{99}$$

$$\text{constraint}\cup\{(1,p(1));\overline{(2,p(2))}\}, \tag{100}$$

$$\text{constraint}\cup\{(1,p(1)),(2,p(2));\overline{(3,p(3))}\}, \tag{101}$$

$$\cdots$$

$$\text{constraint}\cup\{(1,p(1)),\ldots,(\tilde{Z}-2,p(\tilde{Z}-2));\ \overline{(\tilde{Z}-1,p(\tilde{Z}-1))}\}. \tag{102}$$

This partitions all feasible solutions that satisfy constraint and are not $p$. If a set contains incompatible requirements, such as $(i,j)$ and $\overline{(i,j)}$, it will be discarded. In Algorithm 2, whenever we obtain a bottom PCT, at most $\tilde{Z}-1$ sets like Eq. (99) are enqueued to myconstraints, each of which can yield a different bottom PCT via `OptimizePCT()`.

It should be noted that mypcts produced by Algorithm 2 is overcomplete—its elements are not noncongruent but rather appear in equivalence classes. This is because for any $\mathbf{t}_A\in L_A$ and $\mathbf{t}_B\in L_B$, we have $(+\mathbf{t}_B)\circ\mathcal{J}\circ(-\mathbf{t}_A)$ congruent to $\mathcal{J}$, and their PCTs generally have different permutations unless $\mathbf{t}_B-S\mathbf{t}_A\in\tilde{L}_B$. One may eliminate this redundancy by additionally invoking `Split()` to all "congruent" permutations, which is omitted here for brevity. Another point to note is that a PCT generated by Algorithm 2 or 3 may represent a CSM whose SLM is not $(\tilde{L}_A,\tilde{L}_B,S)$, but rather $(\tilde{L}'_A,\tilde{L}'_B,S)$, where $\tilde{L}'_A\supset\tilde{L}_A$ and $\tilde{L}'_B\supset\tilde{L}_B$ are finer lattices. This can only happen when $S\tilde{L}_A=\tilde{L}_B$ is a *proper* subset of $SL_A\cap L_B$, and is a direct consequence of the fact that the PCT representation does not forbid shuffles with shuffle lattices finer than $\tilde{L}_B$. However,

this feature is useful in practical applications (see Sec. V A) and does not affect the completeness of the enumeration.

## V. APPLICATION

In this section, we demonstrate the applications of Algorithms 1–3 through case studies of the B1–B2 and graphite-to-diamond transitions. The enumerated SLMs and CSMs easily covers and goes beyond previously proposed SSPT mechanisms. Here, we focus on the overall characteristics of the resulting database. For NEB-like computations and detailed transition-path analysis using CRYSTMATCH, see Refs. [3,49,50]. For implementation details, see our PYTHON package CRYSTMATCH [51], which utilizes SPGLIB for symmetry detection [52].

### A. Orientation relationship analysis

As one of the simplest SSPTs in compounds, the B1–B2 transition is similar to the FCC-to-BCC transition, but involves two atomic species. Multiple distinct mechanisms—i.e., CSMs—have been proposed [27], including the purely distortive ($d = 0$) Buerger one [33], the Watanabe-Tokonami-Morimoto (WTM) one which can explain the observed orientation relationship (OR) [36], and the Tolédano-Knorr-Ehm-Depmeier (TKED) one which is supported by MetaD simulations involving $\sim 10^4$ atoms [8,38]. Recently, the Therrien-Graf-Stevanović (TGS) mechanism was identified using the P2PTRANS method [25], which has a significantly lower strain. These CSMs and their evidence are summarized in Table II.

A CSM can be decomposed into an SLM and a PCT, where the former consists of a deformation gradient $S$ and a shuffle lattice. Among these components, the $S$ alone determines the OR and the habit plane, and can thus be compared directly with experimental observations. In contrast, the shuffle lattice and the PCT describe microscopic details that are, currently, accessible only through simulations. In this sense, listing all possible $S$ is of broader interest, which can be achieved using Algorithm 1. To further characterize $S$ from a CSM perspective, we used Algorithm 3 to compute the *minimal multiplicity* and *minimal shuffle distance* among all CSMs with deformation gradient $S$, denoted by $\mu_0(S)$ and $d_0(S)$, respectively. For simplicity, we let $\ell = 2$ and assign equal shuffle-distance weights to Cs and Cl atoms, i.e., set $\theta_i$ as a constant. A more physical choice of weights will be discussed in Sec. V B.

As a preliminary investigation, we enumerated all SLMs for the B1–B2 transition with multiplicity $\mu \leqslant 6$ (i.e., period $\tilde{Z} \leqslant 12$) and RMSS

$$\bar{\varepsilon}(S) = \sqrt{\frac{1}{3}\sum_{i=1}^{3}[\sigma_i(S) - 1]^2} \tag{103}$$

less than 30%. This was achieved by replacing $\Omega_{S_0}$ in Algorithm 1 with $\{S \in \mathbb{R}^{3\times 3} \mid \bar{\varepsilon}(S) \leqslant 0.3\}$. From the enumeration result, we identified the SLMs of all established CSMs above, as shown in Fig. 9(a). A total of 788 noncongruent SLMs with 393 distinct deformation gradients are produced, taking 1 min 11 s on a single core of a Mac M1 chip. The subsequent $d_0$ calculation took only 13 s, producing 393 CSMs with the smallest $d$ among all CSMs sharing their respective deformation gradients, i.e., satisfying

$$d(\mathcal{J}) = d_0(S). \tag{104}$$

We refer to a CSM with this property as a *representative* CSM, whose PCT is necessarily a bottom PCT, though the converse does not hold.

It should be noted that all the 12 mechanisms in Table II are representative CSMs—except for the TKED one. This means that Algorithms 1 and 3 are already capable of reproducing most known CSMs, while Algorithm 2 serves as a last resort for exhaustive enumeration. One could certainly apply Algorithm 2 to every SLM, but a more efficient approach is to filter them using experimentally observed ORs. As detailed in Ref. [27], the most commonly observed ORs in the B1–B2 transition are the Watanabe-Blaschko (WB) OR

$$[001][110]_{B1} \parallel [110][001]_{B2}, \tag{105}$$

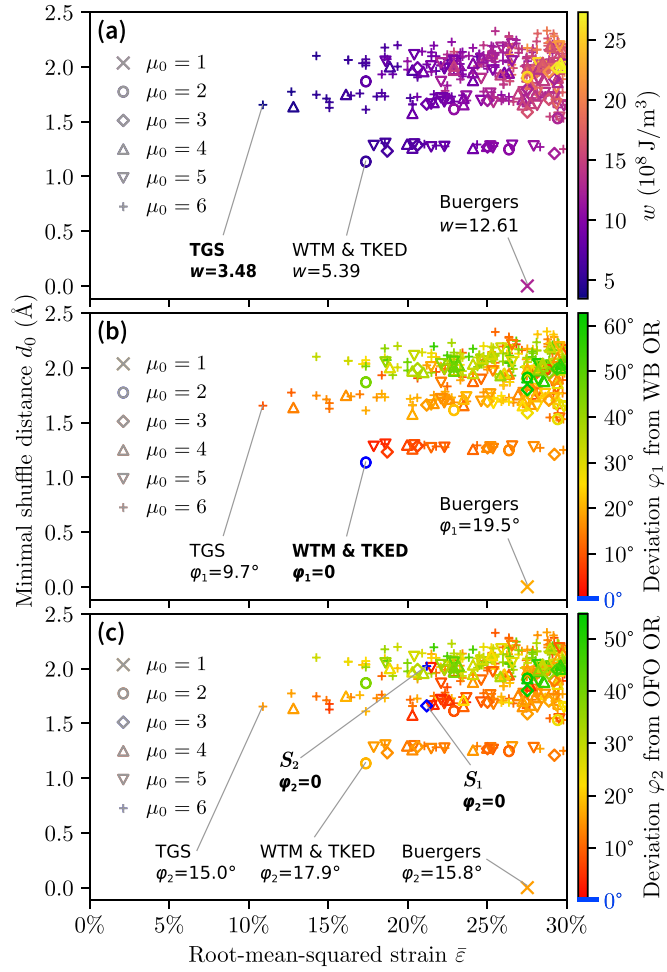


FIG. 9. All deformation gradients $S$ with $\mu_0 \leqslant 6$ and $\bar{\varepsilon} \leqslant 30\%$. Each point represents a unique $S$, whose marker shape indicates its $\mu_0$. (a) Each $S$ is colored according to its estimated strain energy density $w(S)$. Those $S$ of previously proposed mechanisms are pointed out by gray solid lines, where the WTM one and TKED one share the same $S$, and the TGS one has the lowest $w$ among all $S$ enumerated. (b)(c) Each $S$ is colored according to its deviation angles $\varphi_1$ from the WB OR and $\varphi_2$ from the OFO OR. Those $S$ with strictly $\varphi_1 = 0$ or $\varphi_2 = 0$ are highlighted in blue.

and the Okai-Fujiwara-Onodera (OFO) OR

$$[100][111]_{\mathrm{B1}} \parallel [111][100]_{\mathrm{B2}}. \tag{106}$$

For both ORs, we computed the respective deviation angles $\varphi_1$ and $\varphi_2$ of each deformation gradient, defined as the minimal rotation required to produce the OR, as shown in Figs. 9(b) and 9(c). We found that the deformation gradient of the WTM and TKED mechanisms is the only one that can produce the WB OR without rotation, which we denote by $S_{\mathrm{W}}$. On the other hand, the OFO OR can only be produced without rotation by two previously unreported deformation gradients, denoted as $S_1$ and $S_2$. At this point, we can see that CRYSTMATCH provides a systematic way to infer the deformation gradient—and thus the SLM—based on experimental observations.

One naturally wonders whether, as $\mu_{\max}$ increases, there exists a deformation gradient $S$ with lower strain energy $w$ than that of the TGS mechanism, or other $S$ that can reproduce either OR without rotation. The answer to both questions is affirmative as long as we set $\mu_{\max} = 12$, as shown in Fig. 10. Since the Buerger mechanism has $d = 0$, we need not consider mechanisms with higher strain energy than it. Accordingly, we set $w_{\max} = 13 \times 10^8$ J/m$^3$ and performed a more comprehensive enumeration that took about 3 CPU hours, obtaining 9987 noncongruent SLMs and 5432 representative CSMs with different deformation gradients. Despite the total numbers increasing by more than an order of magnitude, only a handful of new cases emerged with $w$ lower than that of the TGS mechanism, $\varphi_1 = 0$ or $\varphi_2 = 0$.

### B. Exhaustive enumeration

Now we assign shuffle-distance weights to Cs and Cl atoms. If equal weights are used, a large number of CSMs will exhibit identical $d$ values, as shown in Fig. 11. A physically reasonable approach is to consider the initial velocity $v$ required to move from the initial state to the transition state on the PES within an average free time, which is proportional to the unweighted $d$. Statistical mechanics tells us that heavier atoms are less likely to acquire velocities, since kinetic energy $T_i = \frac{1}{2} m_i v_i^2$. Therefore each atom's contribution to $d^2$ should be weighted as $\theta_i \propto m_i$, and $\ell = 2$ is thus an appropriate choice. This leads to $\ln(\theta_{\mathrm{Cs}}/\theta_{\mathrm{Cl}}) \approx 1.32$, which is sufficient to lift the degeneracy in $d$.

To investigate the distribution of nonrepresentative CSMs and to reproduce the TKED mechanism, Algorithm 2 is applied to each SLM with $\mu \leqslant 6$ and $w \leqslant 13 \times 10^8$ J/m$^3$, using $d_{\max} = 1.8$ Å. This took about 3 hours, producing 320 757 CSMs, as shown in Fig. 12. However, if one is only interested in the CSMs that exactly conform to the WB OR, it is sufficient to apply Algorithm 2 only to those SLMs with $S_{\mathrm{W}}$, taking merely 7 min 39 s. Figure 13 shows all CSMs with $S_{\mathrm{W}}$, from which one can see that while $S_{\mathrm{W}}$ belongs to the SLM of the WTM mechanism ($\mu = 2$), it also belongs to 7 noncongruent SLMs with $\mu = 4$ and 13 noncongruent SLMs with $\mu = 6$. This is because for a given $(H_A, H_B, Q) \in \mathrm{IMT}(\mu)$, any $M \in \mathbb{Z}^{3\times3}$ with $\det M > 0$ induces an $(H'_A, H'_B, Q') \in \mathrm{IMT}(\mu \det M)$ with the same deformation gradient,

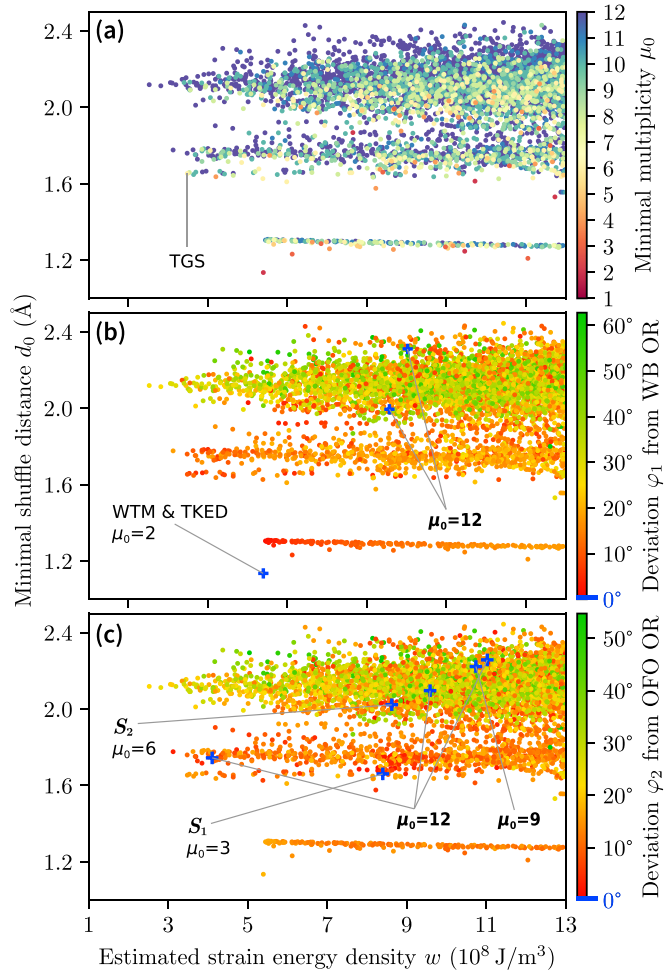


FIG. 10. All deformation gradients $S$ with $\mu_0 \leqslant 12$ and $w \leqslant 13 \times 10^8$ J/m$^3$. Only the Buerger mechanism is omitted, as it has already been shown in Fig. 9. (a) Each $S$ is colored according to $\mu_0$, the minimal multiplicity required for an SLM to have $S$. There are several $S$ with $w$ less than the TGS mechanism, but also higher $d_0$ and $\mu_0$. (b), (c) Each $S$ is colored according to the deviation angles. Blue crosses highlight those $S$ with $\varphi_1 = 0$ or $\varphi_2 = 0$.

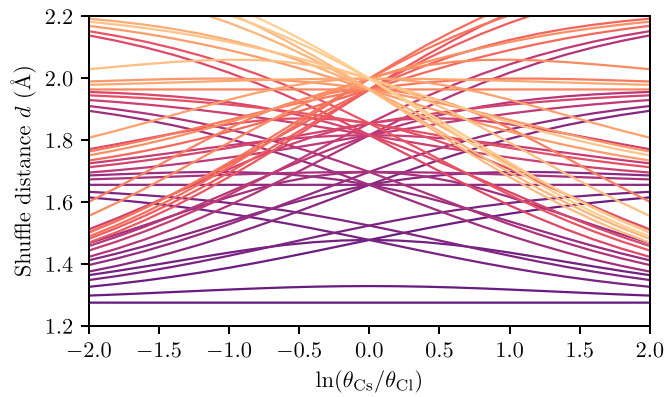


FIG. 11. The shuffle distance $d$ depends on the weighting scheme, i.e., the ratio $\theta_{\mathrm{Cs}}/\theta_{\mathrm{Cl}}$ for SSPTs in CsCl. The figure shows 51 CSMs with the same SLM ($\mu = 6$), whose $d$ has only 10 distinct values when $\theta_{\mathrm{Cs}} = \theta_{\mathrm{Cl}}$. Colors are used to distinguish between different CSMs.

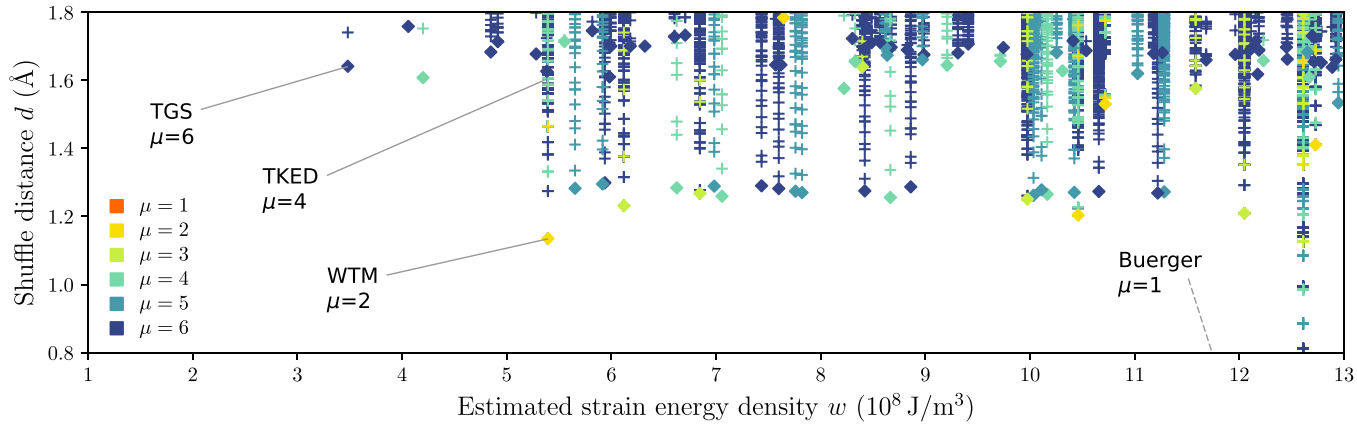


FIG. 12. All CSMs with $\mu \leqslant 6$, $w \leqslant 13 \times 10^8$ J/m$^3$, and $d \leqslant 1.8$ Å. Only the Buerger CSM is ommitted, as it has already been shown in Fig. 9. Diamonds and crosses indicate representative and nonrepresentative CSMs, respectively. Each CSM is colored according to its multiplicity.

where

$$H'_A = \mathrm{hnf}(H_A M), \tag{107}$$

$$H'_B = \mathrm{hnf}(H_B Q M), \tag{108}$$

$$Q' = H'^{-1}_B H_B Q H^{-1}_A H'_A. \tag{109}$$

Overall, the total number of CSMs increases rapidly with both $\mu$ and $d_{\max}$.

Note that $\mathcal{J}_{p,\mathbf{t}_1,\ldots,\mathbf{t}_{\tilde{Z}}}$ satisfying Eq. (59) may have a period smaller than $\tilde{Z}$. Consequently, the WTM CSM $\mathcal{J}_{\mathrm{W}}$ is reproduced in every PCT$^\star(\tilde{L}_A, \tilde{L}_B, S_{\mathrm{W}})$ such that $\tilde{L}_B$ is a sublattice of the shuffle lattice of $\mathcal{J}_{\mathrm{W}} \circ S_{\mathrm{W}}^{-1}$; see Fig. 13. In fact, all CSMs exhibit this repetitive property under the PCT representation. Although one could easily eliminate it by modifying Algorithm 2 and Definition 17, doing so is unnecessary, since the current formalism ensures that PCT$(\tilde{L}_A, \tilde{L}_B, S)$ precisely consists of all possible CSM that a molecular-dynamics simulation employing periodic boundary conditions $\tilde{L}_A$ (initial) and $\tilde{L}_B$ (final), undergoing the deformation $S$, may produce—even when the resulting CSM possesses a smaller intrinsic period.

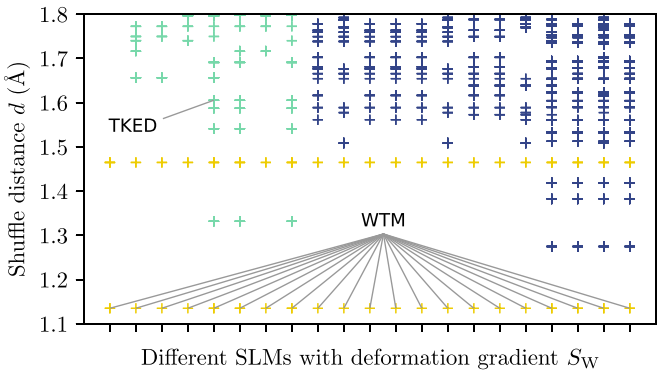


FIG. 13. All CSMs with $\mu \leqslant 6$ and deformation gradient $S_{\mathrm{W}}$. CSMs that can be produced by Algorithm 2 using the same SLM are drawn in the same horizontal coordinate. Yellow, green, and indigo represent $\mu = 2$, 4, and 6, respectively. Gray lines point out the CSMs of the WTM and TKED mechanisms. Note that CSMs with $\mu = 2$ also acquire PCT representations in SLMs with $\mu = 4$ and 6.

### C. Effect of pressure on lattice mismatch

The densities of two solid phases of the same material can differ significantly, making their lattice mismatch and strain energy inevitably large. In such cases, pressure plays a decisive role in the occurrence of SSPTs. For example, at 1500 K, the graphite-diamond coexistence pressure estimated by the Berman-Simon line is ∼5 GPa, while the SSPT is not observed until ∼12 GPa [53,54]. Previous studies also suggest that when the pressure is further increased to 20–50 GPa, both the critical nucleus size and nucleation energy decrease, as shown in Fig. 2 of Ref. [10]. However, lattice mismatch cannot be quantified unless how lattices match is self-evident, or defined in advance—this is precisely what the concept of SLM provides. By enumerating all possible deformation gradients, Algorithm 1 enables a comprehensive analysis of lattice mismatch.

We investigated the graphite-to-diamond transition under pressures $P = 0$, 10, 20, 30, 40, and 50 GPa. The enumeration range was specified as $\mu_{\max} = 6$ (i.e., $\tilde{Z} \leqslant 24$) and $w_{\max} = 5 \times 10^{10}$ J/m$^3$ to ensure inclusion of all established mechanisms, with the initial and final structures as well as elastic tensors calculated at each $P$. The minimal multiplicity $\mu_0$ and minimal shuffle distance $d_0$ for each deformation gradient is also computed as we did in Sec. V A. For the case of $P = 0$, enumerating the SLMs and computing $d_0$ took 9 min 29 s and 37 s, respectively. The joint distribution of $w$ and $d_0$, as well as the spectrum of $w$ are shown in Fig. 14. We can see that as pressure increases, the strain energy of those deformation gradients with small $d_0$ or small $w$ decreases significantly. However, some other deformation gradients—particularly those with larger $\mu_0$ and $d_0$—exhibit opposite or even nonmonotonic trends. This suggests that the actual SSPT mechanism may vary under different pressures. Overall, the "lattice mismatch spectrum" in Fig. 14(b) extends leftward with increasing pressure, consistent with previous experimental and computational studies.

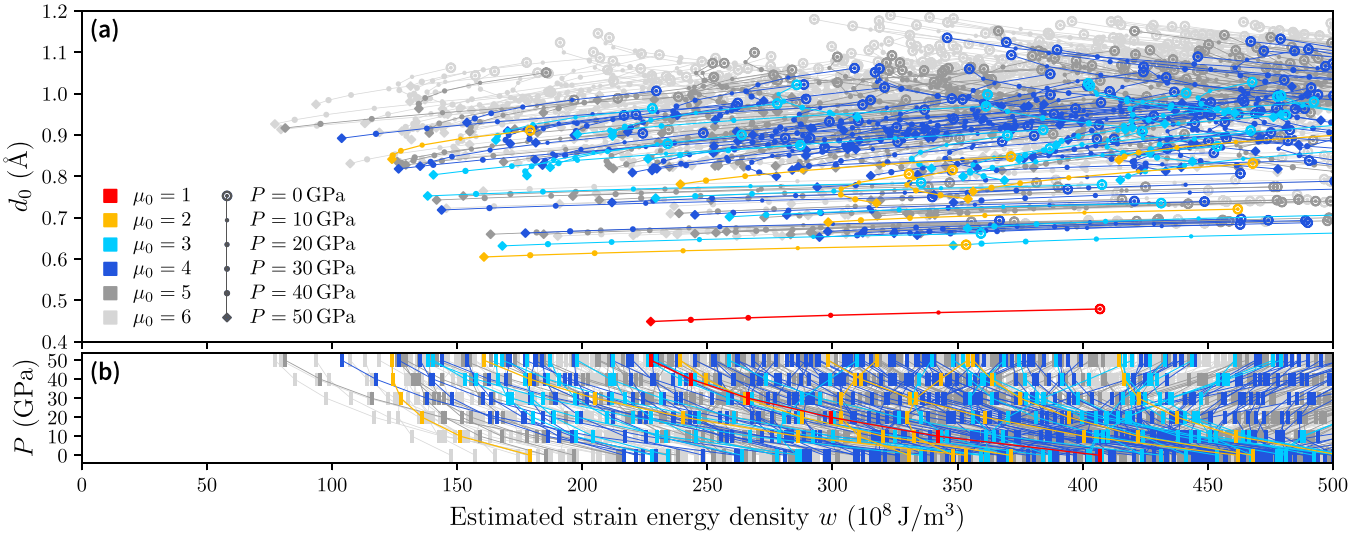


FIG. 14. All deformation gradients $S$ with $\mu_0 \leqslant 6$ and $w \leqslant 500 \times 10^8$ J/m$^3$. Colors are used to distinguish between different multiplicities. (a) The distribution of $w(S)$ and $d_0(S)$ at each pressure $P$. Solid lines connect points representing the same SLM under different $P$, whose marker shapes change with increasing pressure, as indicated in the legend at the lower left. (b) The lattice mismatch spectrum at each $P$. Solid lines show the diverse behaviors of $w$ as a function of $P$ for different SLMs.

## VI. SUMMARY

We have introduced a cell-independent formalism that unequivocally defines the deformation gradient, SLM, and standard shuffle of a CSM, as well as the symmetry-induced congruence relations on CSM($\mathcal{A}, \mathcal{B}$) and SLM($\mathcal{A}, \mathcal{B}$). This maps all possible CSMs of an SSPT onto a tree structure, which is finite up to congruence classes as long as the multiplicity, strain energy and shuffle distance are bounded. Using the IMT and PCT representation, each CSM is encoded in a tuple of integer matrices with minimal redundancy. Based on this, Algorithms 1–3 have been proposed for the exhaustive enumeration of CSMs and SLMs, unveiling novel SSPT mechanisms with unprecedentedly low strain or desired ORs. They lay the groundwork for predicting CSM energy barriers via machine learning and can be seamlessly integrated with widely used methods with negligible computational cost. One may also use them to study defect migration or heteroepitaxy [24,55,56]. All algorithms are implemented in the command-line tool CRYSTMATCH, which is open-sourced and freely available to the community [51].

## ACKNOWLEDGMENTS

We acknowledge helpful discussions with Xu-Yuan Zhou, Lu Yu, Han Wang, Yuze Sun, Wei-Jian Jiang, Yi-Chi Zhang, Jia-Xi Zeng, Qi Liu, Yi Yang, Cong Wang, Qing-Yang Zheng, and Jia-Cheng Gao. We are supported by the National Science Foundation of China under Grants No. 123B2048, No. 12204015, No. 12234001, No. 12404257, No. 12474215, and No. 62321004, and the National Basic Research Program of China under Grants No. 2021YFA1400500 and No. 2022YFA1403500. The computational resources were provided by the supercomputer center at Peking University, China.

## DATA AVAILABILITY

The data that support the findings of this article are openly available [51,57].

## APPENDIX A: LEMMAS AND PROOFS

*Proof of Lemma 1.* $\mathbf{t} \in \mathbb{R}^3$ is a translation element of $\mathcal{J}$ if and only if Eq. (9) holds, which is equivalent to

$$\begin{array}{ccccccc} \mathcal{A}+\mathbf{t}_A & \xleftarrow{+\mathbf{t}_A} & \mathcal{A} & \overset{\mathcal{J}}{\rightsquigarrow} & \mathcal{B} & \xrightarrow{+\mathbf{t}_B} & \mathcal{B}+\mathbf{t}_B \\ {\scriptstyle +\mathbf{t}}\downarrow & & {\scriptstyle +\mathbf{t}}\downarrow & & \downarrow{\scriptstyle +\mathbf{t}} & & \downarrow{\scriptstyle +\mathbf{t}} \\ \mathcal{A}+\mathbf{t}_A & \xleftarrow[+\mathbf{t}_A]{} & \mathcal{A} & \underset{\mathcal{J}}{\rightsquigarrow} & \mathcal{B} & \xrightarrow[+\mathbf{t}_B]{} & \mathcal{B}+\mathbf{t}_B \end{array} \quad \text{(A1)}$$

The outer loop of the above diagram means that $\mathbf{t}$ is a translation element of $(+\mathbf{t}_B) \circ \mathcal{J} \circ (-\mathbf{t}_A)$. The above derivation is reversible, so that the two CSMs have the same translation elements. ■

*Proof of Lemma 2.* We prove the following proposition for each positive integer $n$ by induction. If $L$ is an addition group on $\mathbb{R}^n$ with rank $n$ and

$$\lambda(L) = \inf_{\mathbf{t} \in L \setminus \{\mathbf{0}\}} |\mathbf{t}| > 0, \quad \text{(A2)}$$

then

$$\exists \mathbf{t}_1, \ldots, \mathbf{t}_n, \quad L = \left\{ \sum_{i=1}^{n} k_i \mathbf{t}_i \,\middle|\, k_1, \ldots, k_n \in \mathbb{Z} \right\}. \quad \text{(A3)}$$

Since Eq. (A2) implies Eq. (4), Lemma 2 is nothing but a special case of this proposition for $n = 3$.

*Base case.* For $n = 1$, let $\mathbf{t}_1$ denote an element of $L$ with length $\lambda(L)$. Such an element must exist; otherwise, $L$ would have infinitely many elements with bounded lengths, and hence a cluster point (by the Bolzano-Weierstrass theorem), which contradicts Eq. (4). Since $L$ is an addition group, we

have $\{k_1\mathbf{t}_1 \mid k_1 \in \mathbb{Z}\} \subset L$. There are no other elements in $L$; otherwise, the minimum distance between elements in $L$ would be smaller than $\lambda(L)$.

*Induction step.* Now we show that for every $m \geqslant 1$, if Eq. (2.3) holds for $n = m$, it also holds for $n = m+1$. Let $\mathbf{t}_1$ denote an element of $L \subset \mathbb{R}^{m+1}$ with length $\lambda(L)$. Any element in $\mathbb{R}^{m+1}$ can be uniquely decomposed as

$$\mathbf{v} = c_1\mathbf{t}_1 + \mathbf{v}', \tag{A4}$$

where $c_1 \in \mathbb{R}$ and $\mathbf{v}' \in \mathrm{span}_{\mathbb{R}}\{\mathbf{t}_1\}^{\perp}$ is orthogonal to $\mathbf{t}_1$. This defines a projection mapping

$$\pi : \mathbb{R}^{m+1} \to \mathrm{span}_{\mathbb{R}}\{\mathbf{t}_1\}^{\perp}, \tag{A5}$$

$$\mathbf{v} \mapsto \mathbf{v}'. \tag{A6}$$

Since $\pi$ is a linear mapping, the image of $L$ under this mapping, $\pi(L)$, is an additive group with rank $m$. In fact, we also have $\lambda(\pi(L)) > 0$; otherwise, $L$ would have infinitely many elements with bounded lengths, and hence a cluster point. Therefore $\pi(L)$ satisfies the proposition for $n = m$, so there must exist $\mathbf{t}_2, \ldots, \mathbf{t}_{m+1} \in L$ whose images under $\pi$ generate

$$\pi(L) = \left\{ \sum_{i=2}^{m+1} k_i \pi(\mathbf{t}_i) \,\middle|\, k_2, \ldots, k_{m+1} \in \mathbb{Z} \right\}. \tag{A7}$$

Also, since $L$ is an addition group, we have

$$\left\{ \sum_{i=1}^{m+1} k_i \mathbf{t}_i \,\middle|\, k_1, \ldots, k_{m+1} \in \mathbb{Z} \right\} \subset L, \tag{A8}$$

and we only need to show that there are no other elements in $L$. Note that any $\mathbf{t} \in L$ can be expressed as

$$\mathbf{t} = c_1\mathbf{t}_1 + \sum_{i=2}^{m+1} c_i\mathbf{t}_i, \tag{A9}$$

for some $c_1, \ldots, c_{m+1} \in \mathbb{R}$. Applying $\pi$ to Eq. (A9) and use Eq. (A7), we can see that $c_2, \ldots, c_{m+1}$ must be integers. This implies that

$$(c_1 - \lfloor c_1 \rfloor)\mathbf{t}_1 \in L, \tag{A10}$$

where $\lfloor c_1 \rfloor$ denotes the greatest integer less than or equal to $c_1$. If $c_1 \notin \mathbb{Z}$, the length of Eq. (A10) would be nonzero and smaller than $\lambda(L)$, which contradicts Eq. (A2). At this point, we can see that all coefficients in Eq. (A9) are integers, so that the proposition holds for $n = m+1$. ■

*Proof of Lemma 3.* If $\tilde{C} = CM$ for some $M \in \mathbb{Z}^{3\times3}$, then $\tilde{C}(\mathbb{Z}^3) = C(M(\mathbb{Z}^3))$ is a sublattice of $C(\mathbb{Z}^3)$ since $M(\mathbb{Z}^3)$ is a sublattice of $\mathbb{Z}^3$. Conversely, if $\tilde{C}(\mathbb{Z}^3)$ is a sublattice of $C(\mathbb{Z}^3)$, there exist $\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3 \in \mathbb{Z}^3$ such that the $i$th column of $\tilde{C}$ equals $C\mathbf{k}_i$, i.e., $\tilde{C} = C[\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3]$. To see that the index of this sublattice is $|\det M|$, we only need to use the elementary divisor theorem [58], which states that the quotient group $C(\mathbb{Z}^3)/\tilde{C}(\mathbb{Z}^3)$ is isomorphic to a $\mathbb{Z}$-module of size $|\det M|$. ■

*Proof of Lemma 4.* Arbitrarily take $\mathbf{a}_1, \ldots, \mathbf{a}_{Z_A} \in \mathcal{A}$ that are $L_A$-inequivalent, and $\mathbf{t}_1, \ldots, \mathbf{t}_k \in L_A$ that are $\tilde{L}_A$-inequivalent. We have

$$\mathcal{A} = \bigsqcup_{i=1}^{Z_A} (\mathbf{a}_i + L_A) \tag{A11}$$

$$= \bigsqcup_{i=1}^{Z_A} \left( \mathbf{a}_i + \bigsqcup_{j=1}^{k} (\mathbf{t}_j + \tilde{L}_A) \right) \tag{A12}$$

$$= \bigsqcup_{i=1}^{Z_A} \bigsqcup_{j=1}^{k} ((\mathbf{a}_i + \mathbf{t}_j) + \tilde{L}_A), \tag{A13}$$

where $\sqcup$ denotes the disjoint union of sets. Since each $(\mathbf{a}_i + \mathbf{t}_j) + \tilde{L}_A$ is a distinct $\tilde{L}_A$-equivalence class, we have $|\mathcal{A}/\tilde{L}_A| = kZ_A$. ■

*Proof of Lemma 5.* Note that

$$C(\mathbb{Z}^3) = C'(\mathbb{Z}^3) \iff C^{-1}C'(\mathbb{Z}^3) = \mathbb{Z}^3 \tag{A14}$$

$$\iff C^{-1}C' \in \mathrm{GL}(3, \mathbb{Z}), \tag{A15}$$

where we have used Eq. (42). ■

*Proof of Lemma 6.* The existence of $H$ and $Q$ as well as the uniqueness of $H$ are proved in Ref. [58]. To see that $Q$ is also unique, note that $Q = H^{-1}M$. ■

*Proof of Lemma 7.* $S^{\mathrm{T}}S$ is positive-definite, so it has three orthogonal eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ (arranged in right-handed order) with eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 > 0$. Let $s_i = \sqrt{\lambda_i}$, $V = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] \in \mathrm{SO}(3)$, and $U = SV\Sigma^{-1}$. We have $U \in \mathrm{SO}(3)$ since $U^{\mathrm{T}}U = I$ and $\det U = 1$. Since $\lambda_i$ is uniquely determined by $S$, so is $s_i$. ■

*Lemma 8.* When $\ell = 2$, the right-hand side of Eq. (71) attains the minimum if and only if $\boldsymbol{\tau} = \overline{\tilde{\mathbf{a}}_i} - \overline{\tilde{\mathcal{J}}(\tilde{\mathbf{a}}_i)}$, where the overline denotes the mean value

$$\overline{\mathbf{v}_i} = \sum_{i=1}^{\tilde{Z}} \theta_i \mathbf{v}_i. \tag{A16}$$

Consequently, a CSM with PCT $(p, \mathbf{t}_1, \ldots, \mathbf{t}_{\tilde{Z}})$ has

$$d\left(\mathcal{J}_{p,\mathbf{t}_1,\ldots,\mathbf{t}_{\tilde{Z}}}\right) = \sqrt{\overline{|\tilde{\mathbf{b}}_{p(i)} + \tilde{\mathbf{t}}_i - \tilde{\mathbf{a}}_i - \overline{\tilde{\mathbf{b}}_j} - \overline{\tilde{\mathbf{t}}_j} + \overline{\tilde{\mathbf{a}}_j}|^2}}, \tag{A17}$$

where $\tilde{\mathbf{b}}_i = V\Sigma^{-\frac{1}{2}}U^{\mathrm{T}}\mathbf{b}_i$ is deformed from the $L_B$-motif that defines the PCT, and so is $\tilde{\mathbf{t}}_i = V\Sigma^{-\frac{1}{2}}U^{\mathrm{T}}\mathbf{t}_i$.

*Proof.* For any $\mathbf{v}_1, \ldots, \mathbf{v}_{\tilde{Z}} \in \mathbb{R}^3$ and $\boldsymbol{\tau} \in \mathbb{R}^3$, we have

$$\overline{|\mathbf{v}_i + \boldsymbol{\tau}|^2} \tag{A18}$$

$$= \overline{|(\mathbf{v}_i - \overline{\mathbf{v}_j}) + (\boldsymbol{\tau} + \overline{\mathbf{v}_j})|^2} \tag{A19}$$

$$= \overline{|\mathbf{v}_i - \overline{\mathbf{v}_j}|^2} + 2\,\overline{(\mathbf{v}_i - \overline{\mathbf{v}_j}) \cdot (\boldsymbol{\tau} + \overline{\mathbf{v}_j})} + \overline{|\boldsymbol{\tau} + \overline{\mathbf{v}_j}|^2} \tag{A20}$$

$$= \overline{|\mathbf{v}_i - \overline{\mathbf{v}_j}|^2} + 2\,\overline{(\mathbf{v}_i - \overline{\mathbf{v}_j})} \cdot (\boldsymbol{\tau} + \overline{\mathbf{v}_j}) + |\boldsymbol{\tau} + \overline{\mathbf{v}_j}|^2 \tag{A21}$$

$$= \overline{|\mathbf{v}_i - \overline{\mathbf{v}_j}|^2} + |\boldsymbol{\tau} + \overline{\mathbf{v}_j}|^2. \tag{A22}$$

We can see that $\overline{|\mathbf{v}_i + \boldsymbol{\tau}|^2}$ attains its minimum $\overline{|\mathbf{v}_i - \overline{\mathbf{v}_j}|^2}$ if and only if $\boldsymbol{\tau} = -\overline{\mathbf{v}_i}$. Hence, when $\ell = 2$, we have

$$d(\mathcal{J}) = \sqrt{\overline{|\tilde{\mathcal{J}}(\tilde{\mathbf{a}}_i) - \tilde{\mathbf{a}}_i - \overline{\tilde{\mathcal{J}}(\tilde{\mathbf{a}}_j)} + \overline{\tilde{\mathbf{a}}_j}|^2}}, \tag{A23}$$

which is identical to Eq. (A17). ■

*Lemma 9.* Let $\mathcal{A}$ be a crystal structure and $L$ a sublattice of $\mathcal{A}$. If we additionally require that $\mathcal{A} \subset \mathbb{R}^3$ has no cluster points, then $|\mathcal{A}/L|$ is finite.

*Proof.* We assume by contradiction that the quotient set $\mathcal{A}/L$ is infinite. Take $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3 \in L$ that satisfy Eq. (A3) and consider the parallelepiped

$$P = \left\{ \sum_{i=1}^{3} x_i \mathbf{t}_i \;\middle|\; x_1, x_2, x_3 \in [0, 1) \right\}. \tag{A24}$$

Any element in $\mathbb{R}^3$ can be uniquely decomposed as

$$\mathbf{v} = \sum_{i=1}^{3} c_i \mathbf{t}_i, \tag{A25}$$

where $c_1, c_2, c_3 \in \mathbb{R}$. Therefore the mapping

$$\pi : \mathbb{R}^3 \to P, \tag{A26}$$

$$\sum_{i=1}^{3} c_i \mathbf{t}_i \mapsto \sum_{i=1}^{3} (c_i - \lfloor c_i \rfloor) \mathbf{t}_i, \tag{A27}$$

is well-defined. Note that the image of each $\mathbf{a} + L \in \mathcal{A}/L$ under $\pi$ is a single element in $P \cap \mathcal{A}$, and

$$\pi(\mathbf{a}_1 + L) = \pi(\mathbf{a}_2 + L) \Rightarrow \mathbf{a}_1 + L = \mathbf{a}_2 + L, \tag{A28}$$

i.e., $\pi : \mathcal{A}/L \to P \cap \mathcal{A}$ is injective. Since $\mathcal{A}/L$ is infinite, its image $\pi(\mathcal{A}/L)$ is also infinite, and so is $P \cap \mathcal{A}$. The Bolzano-Weierstrass theorem then states that $P \cap \mathcal{A}$ must have a cluster point.

*Lemma 10.* Let $L$ be a lattice, and $L_1$, $L_2$ its sublattice. The intersection set $L_1 \cap L_2$ is full rank.

*Proof.* Let $C$, $C_1$, $C_2$ be base matrices of $L$, $L_1$, $L_2$, respectively. We only need to prove that there exist nonsingular $N_1, N_2 \in \mathbb{Z}^{3\times3}$ such that

$$C_1 N_1 = C_2 N_2. \tag{A29}$$

Lemma 3 states that $C_1 = CM_1$ and $C_2 = CM_2$ for some nonsingular $M_1, M_2 \in \mathbb{Z}^{3\times3}$. Therefore Eq. (A29) is equivalent to

$$N_1 = M_1^{-1} M_2 N_2. \tag{A30}$$

Since $M_1^{-1}M_2$ is rational, we can always let $m$ be a common multiple of the denominators of its elements. Then we have nonsingular $N_1 = mM_1^{-1}M_2$ and $N_2 = mI$ that satisfies Eq. (A30). ∎

## APPENDIX B: DETAILS OF ALGORITHM 1

For empirical reasons, we generate $S_0 \in \Omega_{S_0}$ using the SVD $S_0 = U\Sigma V^{\mathrm{T}}$. Here, $U$ and $V$ are independently, uniformly sampled in SO(3), while the generation of $\Sigma = \mathrm{diag}(s_1, s_2, s_3)$ depends on $U$ and $V$. We sample $(s_1, s_2, s_3)$ within the region $w(U\Sigma V^{\mathrm{T}}) \leqslant w_{\max}$ according to the distribution

$$f(s_1, s_2, s_3) \propto \left|s_1^2 - s_2^2\right| \left|s_2^2 - s_3^2\right| \left|s_1^2 - s_3^2\right|, \tag{B1}$$

and then sort them to restore $s_1 \geqslant s_2 \geqslant s_3$. This procedure ensures that the projection of $S_0$ on SO(3), i.e., $UV^{\mathrm{T}}$, is uniformly distributed, while for each $(U, V)$, the conditional distribution of $S_0$ follows the standard measure on $\mathbb{R}^{3\times3}$. The following algorithm analysis is based on this distribution, but one may well use other distributions to improve the performance of Algorithm 1.

We begin by considering the probability that a given IMT $(H_A, H_B, Q)$ is generated in a single iteration. This can be estimated using the intersection between the cyan region and the pink hypercube centered at $Q$ in Fig. 7. Given that strains in SSPTs are usually small, the intersection can be considered as a thin vicinity of the 3D section of Eq. (90) intercepted by SO(3). Therefore the probability of $(H_A, H_B, Q)$ being generated is approximately equal to the probability that $\phi_{H_A,H_B}(UV^{\mathrm{T}})$ is generated within this section. Since the cyan region contains $Q$, the volume of this section can be assumed to be no less than 1. On the other hand, $UV^{\mathrm{T}}$ is uniformly distributed over SO(3), whose volume as a 3D submanifold is $16\sqrt{2}\,\pi^2$. Hence, after the deformation by $\phi_{H_A,H_B}$, the probability density near $Q$ is at least

$$\eta_{H_A,H_B} = \frac{\sigma_7(\phi_{H_A,H_B})\sigma_8(\phi_{H_A,H_B})\sigma_9(\phi_{H_A,H_B})}{16\sqrt{2}\,\pi^2}, \tag{B2}$$

where

$$\phi_{H_A,H_B} = \left(H_B^{-1}C_B^{-1}\right) \otimes \left(H_A^{\mathrm{T}}C_A^{\mathrm{T}}\right) \tag{B3}$$

is treated as a $9 \times 9$ matrix.

Now we consider the probability that no IMT in a given congruence class is generated in a single iteration. Since each IMT corresponds to an SLM, $(H_A', H_B', Q')$ is congruent to $(H_A, H_B, Q)$ if and only if there exist $R_A \in G_A'$ and $R_B \in G_B'$ such that

$$H_A' = \mathrm{hnf}(\hat{R}_A H_A), \tag{B4}$$

$$H_B' = \mathrm{hnf}(\hat{R}_B H_B), \tag{B5}$$

$$H_B' Q' H_A'^{-1} = H_B Q H_A^{-1}, \tag{B6}$$

where $\hat{R}_A = C_A^{-1}R_AC_A$ and $\hat{R}_B = C_B^{-1}R_BC_B$ are unimodular matrices that describe how $R_A$ and $R_B$ change the base matrices of $L_A$ and $L_B$, respectively. Therefore the probability that no IMT congruent to $(H_A, H_B, Q)$ is generated has a lower bound

$$\zeta_{H_A,H_B} = \prod_{R_A \in G_A'} \prod_{R_B \in G_B'} \left(1 - \eta_{H_A',H_B'}\right). \tag{B7}$$

Replacing $(1 - \eta)$ in Eq. (92) with this expression yields a significantly smaller

$$i^* = \frac{\ln \epsilon}{\ln \max_{H_A,H_B} \zeta_{H_A,H_B}}, \tag{B8}$$

which can be computed numerically before Algorithm 1 starts.

[1] V. Stevanović, R. Trottier, C. Musgrave, F. Therrien, A. Holder, and P. Graf, Predicting kinetics of polymorphic transformations from structure mapping and coordination analysis, Phys. Rev. Mater. **2**, 033802 (2018).

[2] F. Therrien, P. Graf, and V. Stevanović, Matching crystal structures atom-to-atom, J. Chem. Phys. **152**, 074106 (2020).

[3] F.-C. Wang, Q.-J. Ye, Y.-C. Zhu, and X.-Z. Li, Crystal-structure matches in solid-solid phase transitions, Phys. Rev. Lett. **132**, 086101 (2024).

[4] D. Zahn, Y. Grin, and S. Leoni, Mechanism of the pressure-induced wurtzite to rocksalt transition of cdse, Phys. Rev. B **72**, 064110 (2005).

[5] F. Shimojo, S. Kodiyalam, I. Ebbsjö, R. K. Kalia, A. Nakano, and P. Vashishta, Atomistic mechanisms for wurtzite-to-rocksalt structural transformation in cadmium selenide under pressure, Phys. Rev. B **70**, 184111 (2004).

[6] J. Cai and N. Chen, First-principles study of the wurtzite-to-rocksalt phase transition in zinc oxide, J. Phys.: Condens. Matter **19**, 266207 (2007).

[7] D. Zahn and S. Leoni, Nucleation and growth in pressure-induced phase transitions from molecular dynamics simulations: Mechanism of the reconstructive transformation of nacl to the cscl-type structure, Phys. Rev. Lett. **92**, 250201 (2004).

[8] M. Badin and R. Martoňák, Nucleating a different coordination in a crystal under pressure: A study of the B1-B2 transition in nacl by metadynamics, Phys. Rev. Lett. **127**, 105701 (2021).

[9] P. A. Santos-Florez, H. Yanxon, B. Kang, Y. Yao, and Q. Zhu, Size-dependent nucleation in crystal phase transition from machine learning metadynamics, Phys. Rev. Lett. **129**, 185701 (2022).

[10] R. Z. Khaliullin, H. Eshet, T. D. Kühne, J. Behler, and M. Parrinello, Nucleation mechanism for the direct graphite-to-diamond phase transition, Nat. Mater. **10**, 693 (2011).

[11] H. Eyring, The activated complex in chemical reactions, J. Chem. Phys. **3**, 107 (1935).

[12] D. Sheppard, P. Xiao, W. Chemelewski, D. D. Johnson, and G. Henkelman, A generalized solid-state nudged elastic band method, J. Chem. Phys. **136**, 074103 (2012).

[13] F. Therrien, E. B. Jones, and V. Stevanović, Metastable materials discovery in the age of large-scale computation, Appl. Phys. Rev. **8**, 031310 (2021).

[14] L. Zhu, R. E. Cohen, and T. A. Strobel, Phase transition pathway sampling via swarm intelligence and graph theory, J. Phys. Chem. Lett. **10**, 5019 (2019).

[15] C. Shang, X.-J. Zhang, and Z.-P. Liu, Stochastic surface walking method for crystal structure and phase transition pathway prediction, Phys. Chem. Chem. Phys. **16**, 17845 (2014).

[16] S.-H. Guan, X.-J. Zhang, and Z.-P. Liu, Energy landscape of zirconia phase transitions, J. Am. Chem. Soc. **137**, 8010 (2015).

[17] X.-J. Zhang and Z.-P. Liu, Variable-cell double-ended surface walking method for fast transition state location of solid phase transitions, J. Chem. Theory Comput. **11**, 4885 (2015).

[18] X. Chen, Y. Song, N. Tamura, and R. D. James, Determination of the stretch tensor for structural transformations, J. Mech. Phys. Solids **93**, 34 (2016).

[19] Q.-S. Huang, Y. Zhang, P. F. Liu, H. G. Yang, X. Zhang, and S.-H. Wei, Interface-confined intermediate phase in tio2 enables efficient photocatalysis, Proc. Natl. Acad. Sci. USA **121**, e2318341121 (2024).

[20] M. J. Field, *A Practical Introduction to the Simulation of Molecular Systems* (Cambridge University Press, Cambridge, New York, 1999), p. 153.

[21] S. Scandolo, M. Bernasconi, G. L. Chiarotti, P. Focher, and E. Tosatti, Pressure-induced transformation path of graphite to diamond, Phys. Rev. Lett. **74**, 4015 (1995).

[22] A. Laio and M. Parrinello, Escaping free-energy minima, Proc. Natl. Acad. Sci. USA **99**, 12562 (2002).

[23] R. Martoňák, A. Laio, and M. Parrinello, Predicting crystal structures: The Parrinello-Rahman method revisited, Phys. Rev. Lett. **90**, 075503 (2003).

[24] Y.-F. Li and Z.-P. Liu, Smallest stable $Si/SiO_2$ interface that suppresses quantum tunneling from machine-learning-based global search, Phys. Rev. Lett. **128**, 226102 (2022).

[25] F. Therrien and V. Stevanović, Minimization of atomic displacements as a guiding principle of the martensitic phase transformation, Phys. Rev. Lett. **125**, 125502 (2020).

[26] C. Capillas, J. Perez-Mato, and M. Aroyo, Maximal symmetry transition paths for reconstructive phase transitions, J. Phys.: Condens. Matter **19**, 275203 (2007).

[27] H. T. Stokes, D. M. Hatch, J. Dong, and J. P. Lewis, Mechanisms for the reconstructive phase transition between the $B_1$ and $B_2$ structure types in nacl and pbs, Phys. Rev. B **69**, 174111 (2004).

[28] E. C. Bain and N. Dunkirk, The nature of martensite, Trans. AIME **70**, 25 (1924).

[29] L. Sandoval, H. M. Urbassek, and P. Entel, The bain versus nishiyama–wassermann path in the martensitic transformation of fe, New J. Phys. **11**, 103027 (2009).

[30] W. Pitsch, The martensite transformation in thin foils of iron-nitrogen alloys, Philos. Mag. **4**, 577 (1959).

[31] Z. Nishiyama, *Martensitic Transformation* (Academic Press, New York, 2012).

[32] G. Kurdjumow and G. Sachs, Über den mechanismus der stahlhärtung, Z. Phys. **64**, 325 (1930).

[33] M. J. Buerger, in *Phase Transformations in Solids*, edited by R. Smouchwski, J. E. Mayers, and W. A. Wehl (Wiley, New York, 1948), pp. 183–211.

[34] Y. A. Nga and C. K. Ong, Mechanism of pressure-induced polymorphic transition in nacl, Phys. Rev. B **46**, 10547 (1992).

[35] D. Zahn, Molecular dynamics investigation of the pressure induced B1 to B2 phase transitions of rbbr, J. Solid State Chem. **177**, 3590 (2004).

[36] M. Watanabe, M. Tokonami, and N. Morimoto, The transition mechanism between the cscl-type and nacl-type structures in cscl, Acta Crystallogr. Sect. A **33**, 294 (1977).

[37] O. Blaschko, G. Ernst, G. Quittner, G. Pépy, and M. Roth, Investigations of the nacl-to-cscl phase transition in rbi by elastic diffuse and inelastic neutron scattering, Phys. Rev. B **20**, 1157 (1979).

[38] P. Tolédano, K. Knorr, L. Ehm, and W. Depmeier, Phenomenological theory of the reconstructive phase transition between the nacl and cscl structure types, Phys. Rev. B **67**, 144106 (2003).

[39] S. H. Tolbert and A. Alivisatos, The wurtzite to rock salt structural transformation in cdse nanocrystals under high pressure, J. Chem. Phys. **102**, 4642 (1995).

[40] H. Sowa, On the transition from the wurtzite to the nacl type, Acta Crystallogr. Sect. A **57**, 176 (2001).

[41] M. Wilson and P. A. Madden, Transformations between tetrahedrally and octahedrally coordinated crystals: The wurtzite-

rocksalt and blende-rocksalt mechanisms, J. Phys.: Condens. Matter **14**, 4629 (2002).

[42] M. S. Dresselhaus, G. Dresselhaus, and A. Jorio, *Group Theory: Application to the Physics of Condensed Matter* (Springer, Berlin, Heidelberg, 2007).

[43] G. L. W. Hart and R. W. Forcade, Algorithm for generating derivative structures, Phys. Rev. B **77**, 224115 (2008).

[44] S. Axler, *Linear Algebra Done Right* (Springer, Berlin, Heidelberg, 2015), p. 233.

[45] R. Jonker and A. Volgenant, A shortest augmenting path algorithm for dense and sparse linear assignment problems, Computing **38**, 325 (1987).

[46] D. F. Crouse, On implementing 2d rectangular assignment algorithms, IEEE Trans. Aerosp. Electron. Syst. **52**, 1679 (2016).

[47] D. J. Wales and J. P. Doye, 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 (1997).

[48] K. G. Murty, An algorithm for ranking all the assignments in order of increasing cost, Oper. Res. **16**, 682 (1968).

[49] H. Chen, D. Zhou, Y. Xu, and Q. Li, High-pressure phase transitions and mechanisms in lanthanum mononitride, Chin. Phys. Lett. **42**, 067301 (2025).

[50] J. Wang, Y.-C. Zhang, Y. Liu, H. Deng, M. Liu, Y. Sun, B. Li, X.-Q. Chen, and P. Liu, Atomistic mechanisms of phase transitions in all-temperature barocaloric material $KPF_6$, arXiv:2508.13862.

[51] See https://www.phy.pku.edu.cn/xzli/RESEARCH.htm, for the documentation and source code of CRYSTMATCH.

[52] K. S. Atsushi Togo and I. Tanaka, Spglib: A software library for crystal symmetry search, Sci. Technol. Adv. Mater., Meth. **4**, 2384822 (2024).

[53] F. P. Bundy, H. P. Bovenkerk, H. M. Strong, and R. H. Wentorf Jr, Diamond-graphite equilibrium line from growth and graphitization of diamond, J. Chem. Phys. **35**, 383 (1961).

[54] T. Irifune, A. Kurio, S. Sakamoto, T. Inoue, and H. Sumiya, Ultrahard polycrystalline diamond from graphite, Nature (London) **421**, 599 (2003).

[55] J. Wei, B. Feng, R. Ishikawa, T. Yokoi, K. Matsunaga, N. Shibata, and Y. Ikuhara, Direct imaging of atomistic grain boundary migration, Nat. Mater. **20**, 951 (2021).

[56] J.-L. Li, Y.-F. Li, and Z.-P. Liu, Silicon Orientations to grow semi-polar aln, Nat. Commun. **16**, 1 (2025).

[57] F.-C. Wang, Q.-J. Ye, Y.-C. Zhu, and X.-Z. Li, Crystal-structure matches for the B1-B2 and graphite-to-diamond transitions enumerated by crystmatch, doi:10.5281/zenodo.17076669 (2025).

[58] H. Cohen, *A course in Computational Algebraic Number Theory*, Graduate Texts in Mathematics, Vol. 130 (Springer, Berlin, Heidelberg, 2011).