

Clustering in atom probe tomography data: Coordination number metric, percolation-based parameter scaling, and size effects

Mykola Lazarev^{1,*} and John Banhart^{2,†}

¹*Akhiezer Institute for Theoretical Physics, National Science Center Kharkiv Institute of Physics and Technology, Akademichna 1, 61108 Kharkiv, Ukraine*

²*Institute of Applied Materials, Helmholtz-Zentrum Berlin für Materialien und Energie GmbH, Hahn-Meitner-Platz 1, 14109 Berlin, Germany*



(Received 1 June 2026; accepted 18 September 2026; published 6 October 2026)

The ability to identify nanometer-scale nuclei of additional phases in atom probe tomography (APT) is often limited by the sensitivity of clustering algorithms to user-defined control parameters. Conventional approaches typically rely on the Euclidean distance metric and consider only solute atoms, thereby discarding the solvent atoms that contain most of the spatial information. Here, we introduce a coordination-number metric based on the composition and apply it to higher-order clustering. Using various metrics, we investigate percolation in typical APT structures. By scaling clustering properties to the corresponding percolation thresholds, we define a self-similar variable that is almost invariant with respect to metrics, clustering parameters, and structural disorder. This variable provides a relevant description of clustering and enables the formal transfer of optimal parameters between clustering methods. We also study the characteristic clustering behavior in small precipitates and quantify how the precipitate-matrix interface alters the composition spectrum and broadens the clustering curve. Finally, using simulations that incorporate limited spatial precision, detection efficiency, and other APT reconstruction artifacts, we show that the coordination-number-based approach effectively compensates for heterogeneous dilutions and outperforms solute-density-based methods across all tested scenarios.

DOI: [10.1103/17ms-c7hr](https://doi.org/10.1103/17ms-c7hr)

I. INTRODUCTION

Understanding the mechanisms of precipitate formation and their evolution is necessary for predicting the behavior of materials under thermal and mechanical loading. Atom probe tomography (APT) is a unique tool for acquiring such information at the atomic level [1]. It combines compositional sensitivity with position resolution, providing a three-dimensional image of the chemical arrangement of up to 10^9 atoms. To reliably analyze the acquired dataset, intrinsic limitations of APT must be considered, including detection efficiency, spatial and chemical inaccuracies, and local magnification effects [2,3].

Beyond material characterization, APT data constitute a distinctive class of disordered point patterns, making them attractive for studies of percolation and cluster statistic. They largely preserve the compositional distribution of the original crystalline state while showing gaslike disorder in atomic positions.

A note regarding terminology is necessary here. When describing the formation of additional phases from a solid solution, the following terms—often not clearly distinguished—are used: nuclei of emerging phases, atom clusters, precipitates, etc. Atom clusters are typically small and thermodynamically unstable, evolving into larger, more stable precipitates. Conversely, in data analysis, atom positions are fixed, and a set of parameters controls cluster identification. From this perspective, a cluster is a network of atoms linked according to some formal criteria. In this paper, connected groups of atoms identified by cluster analysis will be called clusters, and the corresponding task will be referred to as clustering; physical clusters will be referred to as precipitates, and the process will be referred to as precipitation.

Characterization of large precipitates is usually straightforward. In this case, cell-based methods, concentration isosurfaces, and related methods are quite effective [2,3]. However, small precipitates, especially those with complex morphologies, require advanced clustering techniques. Relevant approaches use correlations in the mutual arrangement of solute atoms and include the maximum separation distance (MSD) [4], density-based spatial clustering of applications with noise (DBSCAN) [5], core linkage [6], or Delaunay/Voronoi tessellations [7–9]. These methods use the spatial positioning of only solute atoms, which comprise a small fraction of the total number of detected atoms. Solvent atoms are typically involved in subsequent postprocessing steps, such as enveloping and erosion. Hence, most APT data are disregarded in the initial, crucial step of clustering.

*Contact author: n.lazarev@kipt.kharkov.ua

†Contact author: banhart@helmholtz-berlin.de

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

Defining the linkage between atoms is the cornerstone of cluster analysis. Conventional methods, such as MSD, use the Euclidean distance criterion. Recently, a maximum coordination number (MCN) method, based on local composition, was proposed [10]. Earlier, related approaches were also applied in APT data mining [11–15]. A direct counterpart to MCN in machine learning is the k -nearest neighbors (kNN) classifier, where k plays the role of a coordination number [16].

The coordination number of an atom is essentially the number of its nearest neighbors. It is 12 in fcc and hcp crystals, and 8 in bcc crystals. In glasses and liquids, it is defined by the first peak of the radial distribution function [17]. Effective coordination can be extended to include next-nearest neighbors [18]. For gas-type disorder, we can define local coordination using appropriate proximity criteria [19].

Previously, we described the MCN method, constructed a map of phase compositions accessible for cluster analysis, and formulated a criterion for selecting optimal control parameters by maximizing the contrast between clustering degrees of different phases [10]. Those results focused on sufficiently large precipitates and neglected interfacial contributions. Here, we extend MCN to higher-order DBSCAN-type methods and focus on clustering in small precipitates, where interfaces play the central role.

In what follows, we first describe the approach for simulating APT-like structures, including the necessary parameters to account for precipitate size and shape, limited detection efficiency, spatial inaccuracies, and local magnification effects. We then formulate coordination-number-based clustering methods that complement approaches based on the solute density. Next, we derive exact coordination-number correlation functions for the ideal solid solution and use them as the basis for composition-spectrum analysis. We characterize clustering via the percolation theory and show that the clustering behavior is governed by a self-similar variable that combines composition with the methods' control parameters. Using simulations of small, isolated precipitates, we quantify the influence of the matrix–precipitate interface. Finally, we compare the performance of the clustering methods considered across various multiphase alloy states.

II. SIMULATION SETUP

The technique for simulating APT-like atomic structures has been detailed in [10]. The generation of random atomic coordinates starts from the ideal fcc lattice with the lattice constant L . Periodic boundary conditions are used. Selected regions within the lattice are referred to as precipitates (P), approximated by spheroids with principal axes, $a = b$, and c . The remaining lattice forms the matrix (M). Both regions are assumed to be ideal solid solutions differing only in their solute concentrations c_M and c_P . A solvent-only interfacial layer of thickness L is located between the matrix and the precipitate to mimic the depletion zone surrounding precipitates.

Next, we remove part of the atoms to simulate the limited detection efficiency η . Uniform, uncorrelated atom removal is used in conventional simulation practice [20]. In APT, however, correlated losses can occur near various inhomogeneities on the tip of the specimen [21]. To address this

effect, we remove a randomly selected atom along with its neighbors at the $\chi - 1$ nearest sites, repeating this process until the target value of η is reached. Obviously, at the correlated loss parameter $\chi = 1$ we simulate uncorrelated removal.

To simulate the given precision of atomic coordinates Λ , we add displacement vectors randomized in directions and magnitudes to the coordinates so the standard deviation is $\Lambda/2$ along each coordinate axis [22].

Finally, to simulate local magnification effects in APT [3,23,24] arising from phase heterogeneity and ambiguity of tip curvature, we deform the model crystal by applying a triangular-wave transformation to the coordinates:

$$\xi' = \xi + \delta \frac{\mathcal{L}_\xi}{2\pi} \arcsin\left(\sin \frac{2\pi}{\mathcal{L}_\xi} \xi\right), \quad \xi = x, y, z. \quad (1)$$

Here, $\mathcal{L}_\xi$ are the dimensions of the simulation box and δ is the single strain parameter. This transformation results in piecewise constant, alternating multiaxial strains, forming a heterogeneous distribution of dilation and shear regions.

The APT structures are typically characterized by the detection efficiency $\eta = 0.3 - 0.8$ and atomic coordinate precision $\Lambda/L = 0.25 - 1$, where the lower limit of Λ is reached only along certain directions [3]. In this study, we fix η to 0.5. The parameter Λ is set to 0.5, 1, or 4 in units of the lattice constant L . The value of the parameter Λ close to 0.5 corresponds to the best spatial precision achievable in APT. The value of Λ close to 1 corresponds to the typical APT accuracy for complex alloys [25]. The value of Λ close to 4 corresponds to gas-type disorder. We mostly use $\chi = 1$ and $\delta = 0$, increasing δ up to 0.3 to simulate density heterogeneity.

We define the precipitate size by the average number of solutes n_P remaining after the random removal of atoms. Given the specified precipitate solute concentration c_P , the total average number of atoms in the initial precipitate is $n_P/(\eta c_P)$.

III. METHODS OF CLUSTER ANALYSIS

Atomic links determine whether an atom in a solid solution belongs to a particular cluster. Once the links have been established, cluster identification is straightforward.

A. Maximum separation distance

In the MSD method [4], two solute atoms are linked if their separation does not exceed the fixed cutoff d_c . The connected network of linked solutes forms clusters. Subsequently, solvent atoms within the d_c neighborhood of each solute may be assigned to the same cluster.

B. Maximum coordination number

Let us consider the entire set of atoms, including both the solute and the solvent ones. For each atom i , we form an ordered neighbor list of all other atoms j , sorted by the proximity criterion, e.g., by the Euclidean distance. The coordination number m_{ji} is defined as the sequence number of the atom j in the neighbor list of the atom i and can be regarded as the separation between atoms j and i . However, numbers m_{ji} and m_{ij} often differ in disordered structure, being unidirectional distances. That is, numbers m_{ij} constitute a

quasimetric space [26]. To obtain the true metric, we define the coordination distance as follows:

$$M_{ij} = \frac{m_{ij} + m_{ji}}{2}. \quad (2)$$

Let atoms i and j be linked if $M_{ij} \leq z_0$, where z_0 is a fixed cutoff. We refer to z_0 as the MCN, like d_c in the MSD method. In an ideal crystal, z_0 can take values according to the number of atoms in coordination shells (e.g., $z_0 = 12, 18, 42$, etc. in fcc), and each atom has exactly z_0 linked neighbors. However, in a disordered system, at any z_0 , the local number of neighbors varies from atom to atom. The average number of neighbors z depends on the parameter z_0 and the degree of structural disorder. Fortunately, the value of z is only slightly less than z_0 for the metric (2). Thus, the control parameter z_0 gives the average number of neighbors in MCN directly. The link structure generated by MCN is identical to the k -nearest-neighbor graph [19] when k is set equal to z_0 . The distributions of local neighbor numbers for different metrics are detailed in Appendix A.

Mathematically, MCN differs from MSD only in the metric: it replaces the Euclidean distance d_{ij} between atoms i and j with the coordination distance M_{ij} . Unlike the Euclidean distance, which depends only on the coordinates of two atoms, the coordination distance is environment-dependent because all surrounding atoms determine it.

Applying the MCN algorithm requires the following steps:

- (1) Set a fixed cutoff parameter z_0 .
- (2) For each solute atom i , find a list of the $2z_0$ nearest atoms j (both solute and solvent), sorted in ascending order of the Euclidean distance d_{ij} . The rank of atom j in this list is the coordination number m_{ji} . Store the coordination numbers of the solute atoms only for subsequent use.
- (3) Calculate the coordination distances (2) and set the links between solute atoms i and j for which $M_{ij} \leq z_0$.
- (4) Identify clusters of linked atoms using a relevant depth- or breadth-first search procedure (see, e.g., [27]).

To establish the correspondence between the main parameters of MCN and MSD, let us write the average neighbor number within the distance d_c as follows:

$$z(d_c) = \rho_0 \int_0^{d_c} g_{\text{rdf}}(r) 4\pi r^2 dr, \quad (3)$$

where $g_{\text{rdf}}(r)$ is the radial distribution function, $\rho_0 = N/V$ is average atomic density, N is the total number of atoms, and V is the volume of the system.

C. Coordination numbers based on the Voronoi diagram

The atomic neighborhood can be determined not only by the Euclidean distance criterion but also by the geometric relationship of atoms. Let us show how coordination numbers can be derived from the Voronoi diagram. Accordingly, two atoms are neighbors if they are separated by the common face of touching Voronoi polyhedra. The neighbors of an atom constitute the first shell. Each subsequent shell contains the outer neighbors of the previous shell. The first five Voronoi's shells contain 12, 42, 92, 162, 252 atoms in the fcc lattice. In a disordered APT-like structure, each atom has from four to several tens of nearest neighbors, and the average numbers

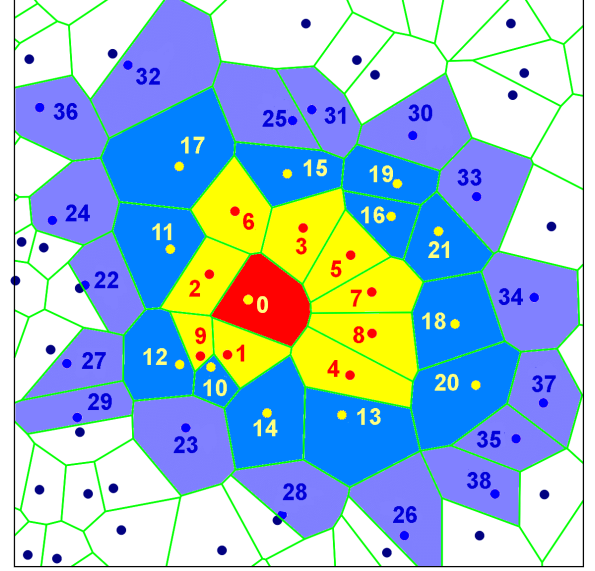


FIG. 1. A two-dimensional example illustrating the definition of coordination numbers based on the Voronoi diagram, shown as thin lines. Neighbors in the first shell of atom 0 are enumerated from 1 to 9 according to the values of the corresponding viewing angles of the polygon edges. Neighbors in the second and third shells are numbered according to the sum of such angles along the optimal path.

of atoms on the first five shells are about 15.5, 70, 180, 360, and 620 [28]. Figure 1 demonstrates the definition of shells in a planar random structure, which was generated with a relatively large spatial disorder parameter $\Lambda = 2$. The Voronoi diagram was calculated using the software described in [29].

To enumerate the atoms in each separate shell, we take the solid angles related to the faces of polyhedral as the relevant proximity criterion. This way, the neighbors in the first shell are numbered in descending order of the corresponding solid angles. As an illustration in Fig. 1, the atoms on the first shell are numbered according to the edges of the Voronoi polygon.

For each atom in subsequent shells, we find the optimal path that maximizes the sum of the solid angles along the allowed paths, including only links between atoms in neighboring shells. The method of coordination numbers based on the Voronoi diagram will be referred to as VCN. Representative maps of links obtained with different methods are given in Appendix A.

D. Maximum coordination number of the k th order

The method of DBSCAN groups densely packed atoms using the criterion of a minimum of k points in the epsilon neighborhood of the given point [5]. DBSCAN was initially applied to APT data as a three-parameter core-linkage clustering [6]. Here, we formulate this method in the coordination number representation.

Using higher-order correlations requires at least a two-step procedure. First, we define the so-called core solute atoms. A solute atom p is a core one if within the coordination distance z_0 this atom has no less than $k - 1$ neighboring solutes. Two parameters, k and z_0 , define core solutes.

In the second step, we find the connectivity of the core atoms: two core atoms, p and q , are linked if their coordination distance obeys $M_{pq} \leq z_L$, where z_L is the given linkage coordination distance. The usual clustering search is performed to identify the cluster backbones.

Finally, all remaining solutes whose coordination distance does not exceed z_L from any core atom join the nearest core atom. The core-linkage coordination-number method has three control parameters: k , z_0 , and z_L . In this paper, we restrict the method to the two-parameter version, setting $z_L = z_0$ as in the initial DBSCAN formulation [5]. It is easy to verify that for $k = 1$ the method reduces to the first-order MCN method, in which all solutes can be treated as core atoms. We refer to the generalized method as kMCN because it originates from MCN. The k -order method based on the Voronoi diagram is therefore denoted kVCN.

The popular isopositional method (IPM) for the clustering of APT data, see, e.g., [15,30], being a modification of the DBSCAN algorithm, employs a local concentration threshold criterion during the core atom selection step and subsequently applies the d_c criterion when merging core atoms into clusters, as in the MSD method. Thus, IPM utilizes a mixed metric that combines approaches based on both composition and density. It can be expected that the characteristics and performance of the IPM will fall between those of standard DBSCAN and kMCN.

Notice that in the original DBSCAN, the ε vicinity of a core point p contains at least k points, including p , meaning the core point has no fewer than $k - 1$ neighbors. For compatibility, we numerate the clustering order parameter k accordingly. On the other hand, the parameter k is the sequence number of a neighboring solute atom in the correlation function formalism. Thus, the correlation function of the k th order corresponds to the kMCN clustering method of the $(k + 1)$ th order.

IV. CORRELATION FUNCTIONS AND COMPOSITION SPECTRUM ANALYSIS

A. Correlation functions

The pair correlation functions of the solute atoms complement the corresponding cluster analysis method; see, e.g., [31]. A negative binomial distribution describes the correlation function of the k th solute atom on the coordination numbers m in an ideal solid solution of a given concentration c [10,32]:

$$B(m|k, c) = \frac{(m-1)!}{(k-1)!(m-k)!} (1-c)^{m-k} c^k, \quad m \geq k. \quad (4)$$

To clarify the physical meaning of (4), we consider a related correlation function in the coordinate representation, i.e., the probability of finding the k th neighbor at the distance r in a gaslike system [6,33–35],

$$P(r|k, \alpha) = \frac{3\alpha r^2 (\alpha r^3)^{k-1}}{(k-1)!} \exp(-\alpha r^3), \quad (5)$$

where $\alpha = 4\pi\rho/3$, $\rho = N_B/V$, and N_B is the number of solute atoms in the volume V .

A change of variables, $m' = (4\pi\rho_0/3)r^3$, reduces (5) to the gamma distribution [32],

$$G(m'|k, c) = \frac{c(m')^{k-1}}{(k-1)!} \exp(-cm'), \quad (6)$$

where $\rho_0 = N/V$ is the atomic density, $c = N_B/N$ is the concentration of solutes, and $N = N_A + N_B$ is the total number of atoms.

The relation (6) was obtained for a system with a gaslike coordinate disorder. However, from a formal point of view, we can consider the variable m' in (6) as a quantity determined by (3), where z and d_c are replaced by m' and r , respectively. Since (3) is valid for an arbitrary homogeneous structure, the applicability of (6) extends far beyond the validity of (5). The approximation (6) together with (3) is much like the correlation function formalism used for an interacting fluid [36]. As a result, the two-parametric (k, c) discrete distribution (4) resembles the gamma distribution (6) with respect to the continuous variable m' , and the correlation function (4) converges to (6) at $c \rightarrow 0$.

Consider the direct calculation of the correlation function in a simulated or experimentally obtained system. Given the atom coordinates, we build ordered lists of all neighbors for each atom i . Let m_{ik} be the coordination number of the k th solute atom in the neighbor list of the i th atom. Then the local k th-order correlation function of atom i is equal to 1 at $m = m_{ik}$ and 0 otherwise, i.e., $Q(i, k|m) = \delta_{mm_{ik}}$, where δ_{ij} is the Kronecker delta. The desired distribution function is then found by averaging over the reference atoms i . Averaging can be done over either the solvent i_A or solute i_B atoms. Both methods give the same result for a single-phase solid solution. However, for a heterogeneous system, we prefer averaging over solute atoms to achieve compatibility with coordinate-based correlation functions and more reliable phase separation:

$$Q_B(m|k) = \frac{1}{N_B} \sum_{i_B} Q(i_B, m|k). \quad (7)$$

B. Composition spectrum analysis

A method for determining the local density spectrum of solute atoms was previously proposed [33,37]. The k -order correlation functions (5) were used as basis functions. A similar composition spectrum analysis can be formulated in the coordination-number representation. The idea is to expand the measured correlation function (7) into a series of basis functions (4):

$$Q(m|k) = \sum_{j=1}^n g_j(k) B(m|k, c_j). \quad (8)$$

Here c_j is the concentration and g_j is the fraction of solute atoms in the j th phase, n is the number of phases. A related density-based series were considered in [6,35].

Generalizing (8) is straightforward. Let n tend to infinity, and then change to the continuum limit,

$$Q(m|k) = \int_0^1 g(k, c) B(m|k, c) dc, \quad (9)$$

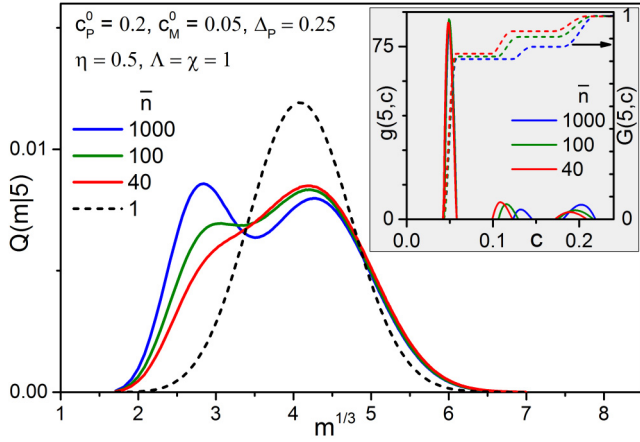


FIG. 2. Pair correlation functions $Q(m|k)$ of the fifth-nearest solute in APT-like structures with different sizes ($\bar{n}$) of precipitates are shown by solid lines. The dashed line represents the corresponding function found in a uniform solid solution of the same average concentration of solutes. The inset shows the calculated composition spectra $g(5, c)$ (solid lines) and the cumulative fractions of solutes $G(5, c) = \int_0^c g(5, c')dc'$ (dashed lines).

where the function $g(k, c)$ can be viewed as the composition spectrum of k th order. The spectral decomposition (9) is the Fredholm integral equation of the first kind with respect to the function $g(k, c)$, and $B(m|k, c)$ is its kernel. This equation is ill posed and requires special methods to solve it [38,39].

One might expect that for all k , the functions $g(k, c)$ should be close to each other, but this is not always the case. The reason is that these functions see the composition at different scales. The smaller the k , the more detailed composition is detected. On the contrary, increasing k effectively suppresses small-scale variations in composition.

To demonstrate the method, we simulate a set of two-phase alloys containing precipitates of various sizes. The solute concentrations in the matrix and precipitates are $c_M = 0.05$ and $c_P = 0.2$, respectively; the fraction of solutes in the precipitates is $\Delta_P = 0.2$; the lattice disorder parameters are $\eta = 0.5$, $\Lambda = \chi = 1$; and the sizes of the precipitates are $\bar{n} = 40, 100$, or 1000 solute atoms. The calculated fifth-order correlation functions are shown as solid lines in Fig. 2. For the alloy with large precipitates ($\bar{n} = 1000$), the phase peaks are separated: the matrix peak is observed near the coordination number $m = 78$, and the precipitate peak is around $m = 23$. As the precipitate sizes decrease, these peaks begin to converge and merge.

The solutions of the integral equation (9) for the considered correlation functions are shown by the solid lines in the inset to Fig. 2. The sharp, high peaks in the composition spectrum around $c = 0.05$ correspond to the matrix and almost coincide at different sizes $\bar{n}$. The right peaks near $c = 0.2$ correspond to the precipitate and the positions and heights of these peaks depend on $\bar{n}$. Peaks in the intermediate region at $c \in [0.1, 0.15]$ correspond to the contribution of the matrix-precipitate interfaces. For clarity, the cumulative (integral) composition spectra are given as dashed lines in the inset to Fig. 2 and directly show the contributions of various alloy regions. For alloys with large precipitates ($\bar{n} = 1000$), the

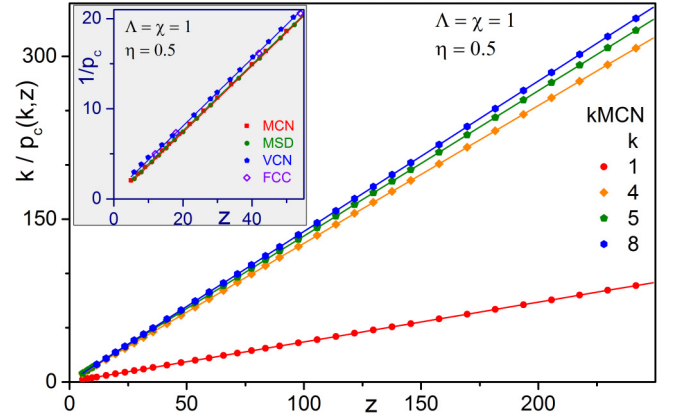


FIG. 3. The correlations between the average coordination number z and the percolation threshold p_c obtained by the kMCN method at fixed k in an APT-like structure. The lines show the two-parameter fitting by (11). The inset displays the same correlation obtained using first-order clustering methods. Filled symbols show data in the APT-like structure. Open symbols depict data in the fcc lattice with extended-range neighborhoods [43], where available coordination numbers z are 12, 18, 42, etc. The lines show fitting by (10).

solute atom fractions in the matrix, interfaces, and the bulk of the precipitates are 0.79, 0.06, and 0.15, respectively. For alloys with small precipitates ($\bar{n} = 40$), these fractions are 0.81, 0.11, and 0.08, respectively. In the latter case, more solute atoms are found in the interface than in the bulk of the precipitates. As a result, the effective solute concentration in small precipitates is significantly lower than the nominal concentration. A more detailed analysis shows that the interface contribution increases with decreasing precipitate size n , increasing concentration c_P , and coordinate disorder Λ .

V. PERCOLATION

Percolation theory describes clustering in random structures and is widely used to understand various physical phenomena such as the order-disorder phase transition in spin glasses, crack propagation in structural materials, solidification of liquids, and others [40,41]. The properties of percolation in APT structures within the MCN method were described in [10]. Here, we provide a comparative study of the percolation observed in other methods.

The inset of Fig. 3 shows the dependencies of the percolation threshold p_c on the average coordination number z for different first-order clustering methods. The Newman-Ziff algorithm [42] was applied. The value of z is calculated for the given z_0 and specified structure. The relationship between z and the MSD clustering parameter d_c is given by (3). This figure also shows the percolation thresholds on regular lattices [43], which are close to those obtained by the VCN method in an APT-like structure.

For all considered methods, the value of $1/p_c$ has a nearly linear dependence on z , so the following relation can approximate the percolation threshold:

$$p_c(z) = \frac{A}{z + a}. \quad (10)$$

TABLE I. Fitting parameters of (11) and (12) obtained by different methods in APT- and gaslike structures ($\Lambda = 0.5, 1, 4; \eta = 0.5$) and the fcc lattice. The parameters were fitted in the range of z_0 from 6 to 242 for disordered structures and from 12 to 248 for the fcc lattice.

Method	$P \setminus k$	1	4	5	6	7	8
kMCN	A	2.731	0.788	0.747	0.731	0.724	0.722
$\Lambda = 0.5$	a	1.05	0.89	0.70	0.56	0.43	0.31
$\Lambda = 1$	A	2.729	0.787	0.747	0.731	0.724	0.721
	a	0.74	0.64	0.50	0.36	0.21	0.12
$\Lambda = 4$	A	2.734	0.789	0.748	0.731	0.724	0.721
	a	0.58	0.50	0.37	0.25	0.11	0.0
DBSCAN	A	2.726	0.786	0.746	0.730	0.723	0.721
$\Lambda = 1$	a	0.25	0.21	0.18	0.15	0.13	0.11
fcc	A	2.730	0.787	0.746	0.730	0.723	0.721
	a	1.90	1.50	1.20	0.90	0.70	0.55
kMCN	A	2.735	0.788	0.748	0.731	0.724	0.722
$\Lambda = 1$	a	1.22	0.90	0.77	0.56	0.40	0.35
Eq. (12)	B	-28.6	-10.1	-9.8	-4.9	-3.1	-5.8
	b	40.2	19.0	17.1	6.2	0.4	4.7

Although the fitting parameters in (10) depend on the properties of disorder and clustering methods, the differences are small.

When discussing percolation in kMCN or DBSCAN, we use the term “percolation in k -order clustering models” here to avoid confusion with the well-known k -core percolation [41,44]. The percolation thresholds in high-order models were obtained using a modified algorithm since percolation occurs only via core atoms. The modified algorithm is to some extent more complicated because at each step it requires verifying whether the added atom and its neighbors are core atoms. The modification increases processing demand, yet it ensures the linearity of computation time with respect to the total number of atoms. See more details in Appendix B.

Since the minimum cluster size is k in kMCN, and percolation occurs only through core atoms, the percolation threshold is expected to be proportional to k . That is indeed the case, but only for large k . Figure 3 shows the correlation of p_c and z for kMCN. We find that the percolation thresholds, $p_c(k, z)$, coincide at $k = 1, 2, 3$. Starting at $k = 4$, the values of $p_c(k, z)$ begin to increase with k . For $k > 5$, the ratio $p_c(k, z)/k$ rapidly becomes weakly dependent on k .

As Fig. 3 suggests, the inverse of the percolation threshold $1/p_c$ exhibits almost perfect linearity with z for all investigated values of k . Thus, the direct generalization of (10) to the higher-order methods reads

$$p_c(k, z) = \frac{kA_k}{z + a_k}. \quad (11)$$

The values of the fitting parameters A_k and a_k are presented in Table I. As mentioned above, $p_c(k) \equiv p_c(1)$ at $k = 2, 3$. Notably, the values of A_k obtained by different clustering methods in different structures are almost identical to the second digit after the decimal point. The values of the shift parameter a_k differ slightly as well.

The correlation (11) is also observed in the ideal fcc lattice with extended-range neighborhoods. In this case, the values of the factors A_k almost coincide with those for the disordered structure, but the parameters a_k appear to be 1.5 to 0.5 larger

when k changes from 1 to 8; see Table I. A larger value of a_k means that the percolation thresholds in an ideal lattice are higher than in a disordered structure for the given z . The same conclusion holds when comparing first-order methods.

To memorize, the slope parameter A_k can be roughly estimated as $A_k \approx e/k$, for $1 \leq k \leq 3$, and $A_k \approx e - 2$ for $k > 5$, where $e = 2.72$ is the Euler’s number. The asymptotes of A_k can be easily joined.

We can now make an estimate of the average number of neighboring solute atoms at the percolation threshold, $N_c^{(s)}$. Since the solute concentration is $N_c^{(s)}/z$, we immediately get from (11) that for large z ,

$$N_c^{(s)} \approx kA_k.$$

This estimation is nearly independent of z and applies to both random structures and regular lattices. It can help to select the optimal parameters for the DBSCAN method involving only solute atoms.

For a more precise description of the correlation between p_c and z , a second-order approximation can be used:

$$p_c(k, z) = \frac{kA_k}{z + a_k + B_k/(z + b_k)}. \quad (12)$$

The fitting parameters of (12) are given in Table I as well. To compare the accuracies of (11) and (12), we consider kMCN with $k = 6$ for an APT-like structure ($\Lambda = 1, \eta = 0.5$). In this case, the relative error of the linear approximation (11) is about 2% at $z = 8$, decreasing to the order of 0.1% at $z > 18$. The second-order approximation (12) yields an error that is an order of magnitude smaller. Its accuracy is even higher than that of the fifth-order polynomial approximation of $1/p_c(z)$ on z . Although the relation (12) has theoretical interest, the linear approximation (11) is sufficient for this work.

The relations (11) and (12) indicate the existence of some self-similar variable composed of the average coordination distance z and the solute concentration c at the fixed k . In a

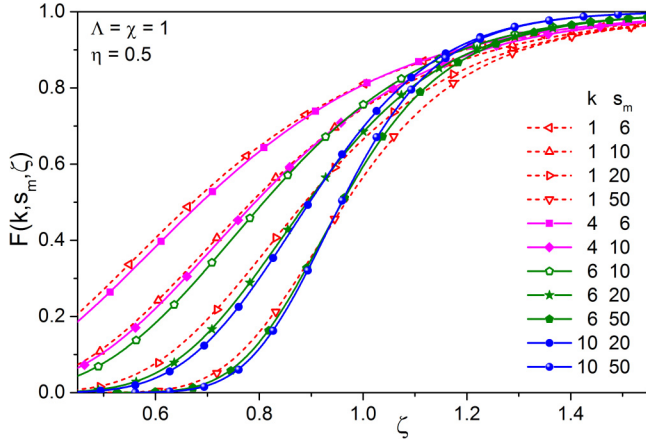


FIG. 4. Clustering curves in the bulk depending on the control parameters of the kMCN method.

similar way to [10], we introduce

$$\zeta(k, z, c) = c/p_c(k, z), \quad (13)$$

where $p_c(k, z)$ is defined by (11). The normalization of ζ is chosen so that it is equal to one at the percolation threshold.

Let $\varphi(s, \zeta)$ be the size (s) distribution of solute clusters in a random solid solution described by the average coordination number z and solute concentration c expressed in terms of the variable ζ defined by (13). We assume that this distribution is normalized to the total number of solute atoms as follows $\sum_{s=1}^{\infty} s\varphi(s, \zeta) = 1$.

The fraction of clustered solutes or the clustering curve is expressed by the complementary cumulative distribution function of cluster size:

$$F(s_m, \zeta) = \sum_{s=s_m}^{\infty} s\varphi(s, \zeta). \quad (14)$$

It provides the probability of finding a solute atom in a cluster of at least s_m solutes. This function helps select the main clustering parameters and compare clustering methods [10].

Figure 4 demonstrates the behavior of $F(s_m, \zeta)$ for kMCN at different control parameters k and s_m . A comparison between the first-order MCN and the fourth-order kMCN shows only minimal differences, suggesting similar performance for these methods in cluster analysis. As the order k increases, the function $F(s_m, \zeta)$ pushes towards zero at $\zeta < 1$, and it rapidly goes towards one at $\zeta > 1$. Due to such behavior, the high- and low-concentration phases can be effectively separated.

Examining the dependence of $F(s_m, \zeta)$ on ζ (Fig. 4) reveals that it resembles a growth curve. The Gompertz function [45] seems to be the best fit here [10].

VI. SIMULATION OF FINITE SIZE PRECIPITATES

So far, we have explored clustering in a homogeneous infinite structure. Real alloys contain several phases. Let the primary phase, called the matrix, occupy most of the alloy with the concentration of solute atoms c_M , and let the precipitates of the second phase be small and contain a higher concentration of solutes c_P . In such a model, all previously

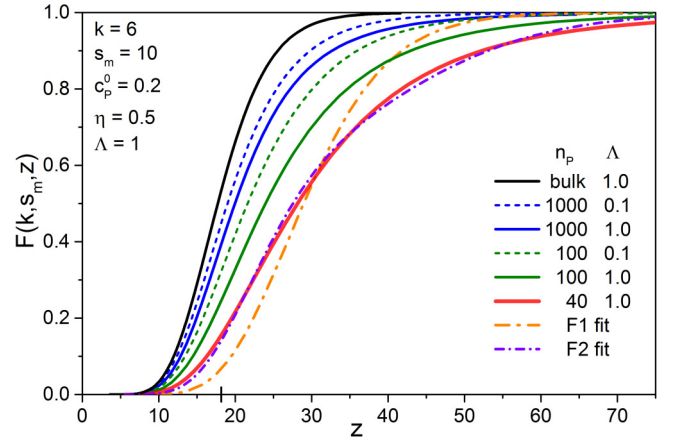


FIG. 5. Fraction of clustered solutes vs average coordination number z for isolated spherical precipitates of different sizes n_p and spatial precision Λ . The designed composition of precipitates c_P^0 and other parameters are indicated in the plot. The dash-dot lines show the fitting of the clustering curve for precipitates of size $n_p = 40$ by one- and two-phase clustering models.

obtained results for the infinite structures are directly applicable to the matrix. Precipitates require additional consideration.

A. Clustering in isolated precipitates

The finite size of a precipitate changes the clustering behavior as the influence of the matrix-precipitate interface increases. To estimate this influence, we apply the formalism of the complementary cumulative distribution function of cluster size (14) for an ensemble of isolated precipitates of a given size, generated using the methodology described in Sec. II, with $c_M = 0$. Then, we perform the cluster analysis using the kMCN method with the fixed k , and vary the control parameter z_0 over a wide range. The number of precipitates exceeds 10^5 in each run to ensure statistically reliable results.

Figure 5 shows the fraction of clustered atoms for different precipitate sizes at the given nominal concentration of solute atoms $c_P^0 = 0.2$. The simulation conditions are indicated on the graph. The leftmost curve corresponds to the infinite system. Since in this case the percolation occurs at about $z = 18$, the clustered fraction becomes above 0.99 for $z > 30$. In finite-size precipitates, the corresponding curves are much flatter. For small precipitates ($n_p = 40$), almost complete clustering ($F(z) > 0.99$) occurs only at $z > 90$, which is three times more than in bulk clustering. This huge difference explains why detecting small precipitates is only possible when the ratio c_P/c_M is high.

The clustering curves for $\Lambda = 0.1$ and 1 are very close in the bulk. However, in finite-size precipitates, these curves differ noticeably, and the difference increases with decreasing precipitate size, indicating a greater contribution of the interface layer as Λ increases.

The behavior of clustering curves in Fig. 5 suggests that the apparent solute concentration is below the given value c_P^0 . We can select the effective concentration c_1 that makes the clustering curve $F(k, s_m, \zeta(z, c_1))$ as close as possible to the observed one. The rightmost solid line shows the clustering curve for precipitates of size $n_p = 40$. The long-dash-dot line

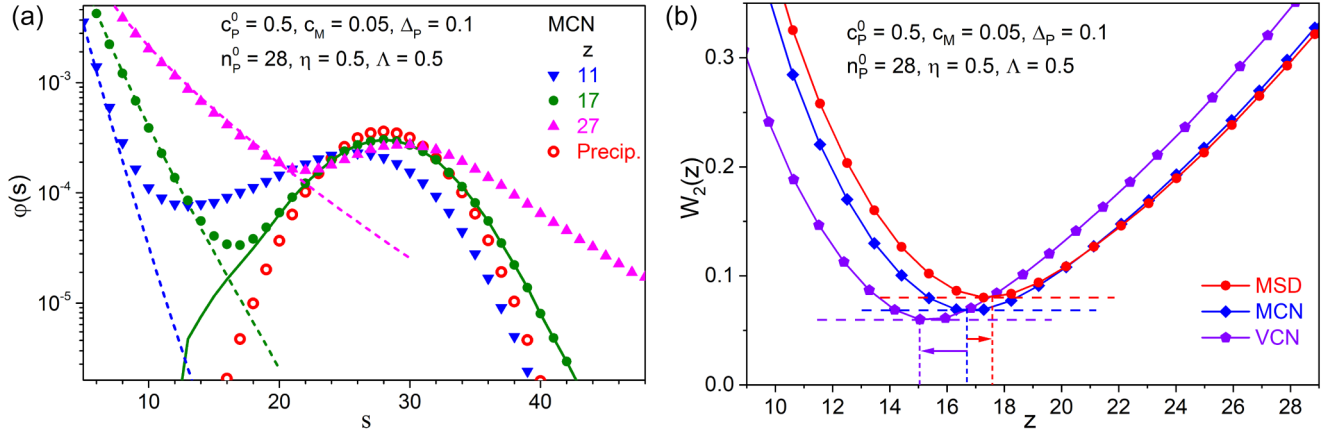


FIG. 6. (a) Cluster size distributions obtained by the MCN method for three values of the parameter z at the concentration ratio $c_p^0/c_M = 10$. Open circles display the given size distribution of precipitates. Filled symbols show calculated cluster distributions; dashed lines show the corresponding distributions in the matrix; the green solid line shows the optimal reduced distribution defined by (16). (b) Comparison of the efficiency of cluster analysis using different first-order clustering methods. The dashed lines show the positions of the $W_2(z)$ minima, which define the optimal $z^{(\text{opt})}$ of the methods.

shows fitting by a single-phase system (F_1) with a concentration of $c_1 = 0.115$. The two-phase fitting (F_2) is significantly better, as shown by the short-dash-dot line. The fitted concentrations are $c_1 = 0.15$ and $c_2 = 0.08$, and the solute fraction in the second phase is $\Delta = 0.3$.

Figures 5 and 2 confirm the validity of the two-phase precipitate model in both composition-spectrum and clustering-curve analyses. However, the fitting parameters c_1 , c_2 , and Δ are effective and depend on the approach used.

B. Optimality criterion for the parameters of the clustering method

Selecting optimal parameters requires an appropriate criterion. It can be directly defined for simulating data, as the properties of precipitates are known in advance. Due to the limited detection efficiency, the sizes of the simulated precipitates are distributed in a wide interval, even if initially they were of the same size. We compare the given size distribution of precipitates $\phi_g(s)$ with that resulting from cluster analysis $\phi(s, \lambda)$, where λ is the set of parameters of a clustering method, e.g., $\lambda \in k, z_0, s_m$ for the kMCN. The closer these distributions are, the better the method's parametrization. Let us consider the following statistical distance:

$$W_p(\lambda) = \left\{ \frac{\sum_{s=s_m}^{\infty} |s\phi(s, \lambda) - s\phi_g(s)|^p}{\sum_{s=s_m}^{\infty} |s\phi(s, \lambda) + s\phi_g(s)|^p} \right\}^{1/p}, \quad (15)$$

where the parameter p takes the values 1 or 2.

The functional (15) can be viewed as a criterion for selecting good clustering control parameters. It is positively definite and symmetric with respect to the distributions $\phi(s, \lambda)$ and $\phi_g(s)$. It vanishes when the distributions coincide and reaches 1 when strongly uncorrelated. The minimum of $W_p(\lambda)$ determines the optimal parameter set λ . The criterion (15) for both values of p gives the same optimal parameter set λ only if the given and calculated distributions are close to each other, implying that $W_p(\lambda) \ll 1$. Note that criteria (15) are only used

to determine the optimal set of parameters λ and to compare different clustering methods for a given $\phi_g(s)$.

C. Clustering in a two-phase system

1. First-order methods

Figure 6(a) shows the cluster-size distributions calculated by the MCN method for the simulated two-phase system with the matrix concentration c_M and relatively high precipitate concentration c_P . About 30 structure realizations were generated, yielding more than 10^5 precipitates. Then, a cluster analysis was performed for the control parameter z , ranging from 8 to 48 in steps of 1. Three of the resulting distributions are shown.

Both criteria (15) yield the optimal parameter $z^{(\text{opt})}$ of about 17 (filled circles). Two distributions obtained using nonoptimal parameters are also shown for comparison. At optimal parametrization, three characteristic intervals of cluster sizes can be distinguished. Background clusters from the matrix $\hat{\phi}(s, z, c_M)$ dominate at small sizes ($s < 15$), as shown by dashed lines in Fig. 6. The matrix contribution can be subtracted,

$$\phi_R(s, z) = \phi(s, z) - (1 - \Delta_P)\hat{\phi}(s, z, c_M), \quad (16)$$

where Δ_P is the fraction of solutes in precipitates. The reduced distributions are shown by the solid line. In the intermediate range ($15 < s < 35$), a good agreement with the precipitate distribution is observed. The tail of the distribution for $s > 35$ is due to the attachment of matrix solutes to clusters in precipitates. This effect is particularly evident when the control parameter is overestimated, as for $z = 27$ in Fig. 6(a). Conversely, an underestimated $z = 11$ does not ensure complete clustering of the precipitate region.

The cluster distributions in Fig. 6(a), obtained using the MCN method, appear similar to those from other clustering methods. To make a quantitative comparison, we use the $W_2(z)$ metric for various methods, with d_c and z related by (3). As shown in Fig. 6(b), the optimal values of the control parameter z differ across the methods under consideration, and the $W_2(z)$

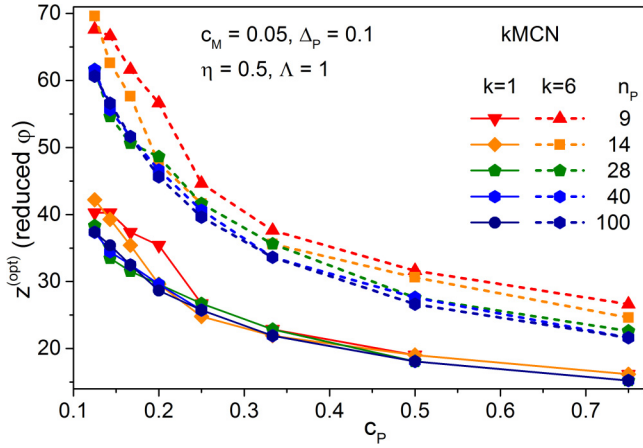


FIG. 7. Dependencies of the optimal control parameter on the composition and size of the precipitates in kMCN methods after reduction of the matrix contribution according to (16).

curves are shifted slightly to the right and left for the MSD and VCN methods, respectively. This behavior reflects the dependencies of the percolation thresholds, shown in Fig. 3.

The optimal value of the coordination number $z^{(\text{opt})}$ depends on the size and composition of precipitates, as seen in Fig. 7. For small precipitates, the optimal $z^{(\text{opt})}$ is slightly increased due to the contribution of the precipitate-matrix interface, see Figs. 2 and 5. However, the dependence on the precipitate size is small.

The compositional dependence $z^{(\text{opt})}(c_p)$ appears to be $\propto 1/\sqrt{c_p}$, as shown in Fig. 7. This behavior can be understood based on the requirement to select the $z^{(\text{opt})}$ value that ensures a high degree of clustering in the precipitate and minimal clustering in the matrix. This degree of clustering is controlled by the self-similar variable, $\zeta \propto zc$, which must be much greater than one in the precipitate and much less than one in the matrix. Given the sharp dependence of $F(\zeta)$ in Fig. 5, we obtain the heuristic relationship $z^{(\text{opt})} \propto (c_p c_M)^{-1/2}$.

2. High-order methods

Figure 8 presents the optimal cluster size distributions for the kMCD method at $k = 6$. At the precipitate composition $c_p = 0.5$, the system is identical to that shown in Fig. 6, where the selection of the optimal control parameter of first-order methods was demonstrated. We see two main differences comparing dependencies in Figs. 8 and 6. The first one is the increase in the optimal control parameter $z^{(\text{opt})}$ by about 1.5. The second is the significant suppression of the low-sized tail in the unreduced distributions. However, the reduced distributions, shown as solid lines in both figures, bear a striking resemblance.

Figure 8 also demonstrates how the quality of the optimal cluster size distributions changes with the c_p^0/c_M ratio. At the ratio of 10 ($c_p^0 = 0.5$), the agreement with the precipitate size distribution is mostly very good. At $c_p^0 = 0.25$, the deviation from the specified precipitate distribution is notable. However, by carefully selecting the control parameters, the reduced distribution yields the correct fraction of clustered solute atoms. Finally, at $c_p^0 = 0.125$, the resulting cluster distribution is far from the specified precipitate distribution, although a

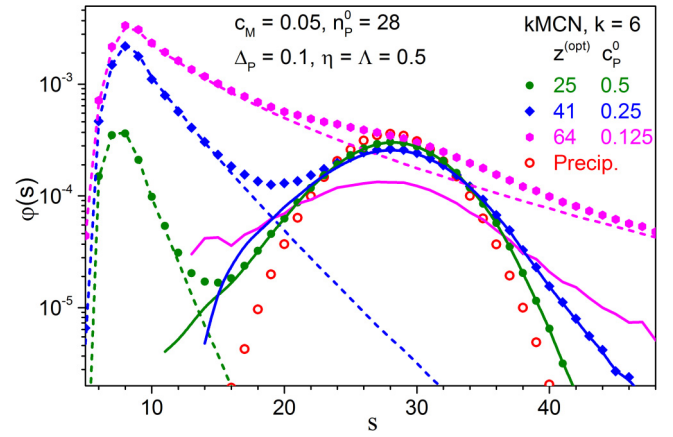


FIG. 8. Cluster size distributions obtained by the kMCN method at $k = 6$ for different precipitate compositions at optimal control parameters. The random structure property and labeling are the same as in Fig. 6(a).

deviation from the randomized arrangement of solute atoms can be detected.

A quantitative comparison shows a slight advantage of high-order methods at the most problematic small c_p/c_M ratios. Surprisingly, in the opposite case of a large c_p/c_M ratio, the first-order methods have a slight advantage according to the $W_2(\lambda)$ criterion at optimal values of z or d_c . This behavior is related to the features of high-order methods resulting in smoothing the local composition inhomogeneities. The same features are observed in the correlation functions (Fig. 8). The interface boundary appears to be effectively broadening. This effect is more pronounced for small precipitates, limiting the applicability of k -order methods to precipitate sizes roughly $n_p > 2k$.

A comparison of the optimal parameters' dependencies, shown in Fig. 7, is also noteworthy. First, we observe a similarity between the results obtained by methods of different k . However, in addition to the obvious scaling to large values of the control parameter z , high-order k methods show a more pronounced dependence on the precipitate size n_p due to the influence of the precipitate-matrix interface.

To describe the scaling of $z^{(\text{opt})}$ values with the method order k quantitatively, we present the dependence of the optimal self-similar variable $\zeta^{(\text{opt})} \equiv \zeta(k, z^{(\text{opt})}, c_M)$ defined by (13) on the precipitate composition, as shown in Fig. 9. We see almost perfect scaling for $1 \leq k \leq 10$. With increasing k and the ratio c_p/c_M , there is some downward deviation of $\zeta^{(\text{opt})}$ values. However, since the optimality region for $\zeta^{(\text{opt})}$ widens with the increase in the c_p/c_M ratio, this deviation remains insignificant.

Combining the data shown in Figs. 7–9, we are convinced that $z^{(\text{opt})}$ is inversely proportional to $(c_p c_M)^{1/2}$ for arbitrary k , at least for the large precipitates; see the inset in Fig. 9.

Figure 10 provides a quantitative comparison of clustering quality as a function of the method order k , as well as the size and composition of precipitates. The MCN method consistently outperforms the MSD method. Although not shown in the figure, DBSCAN is also less efficient than the corresponding kMCN method. The results for small and large

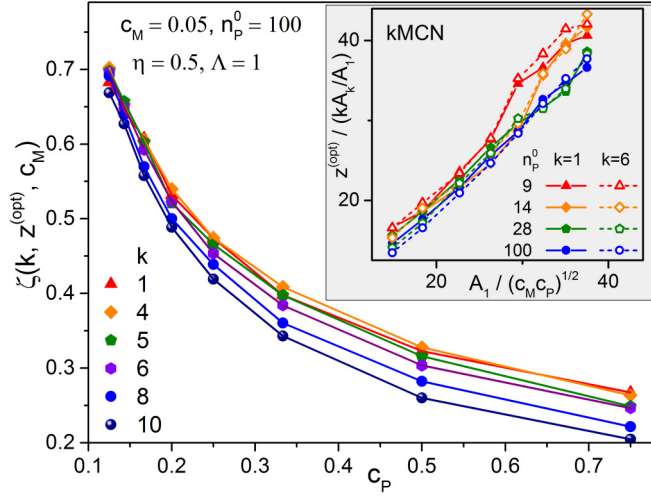


FIG. 9. The dependence of the optimal self-similar variable $\zeta^{(\text{opt})}$ in the matrix for different order parameters k of kMCN methods. The inset figure shows the scaling of the optimal parameter $z^{(\text{opt})}$ with the method order k and the precipitate composition c_P .

precipitates are clearly separated, and the clustering quality for large precipitates is higher (corresponding to a lower W_2 value). The dependence of W_2 on the method order k is more complex. At a low c_P/c_M ratio, increasing the value of k positively affects the clustering quality. However, at high c_P/c_M , a nonmonotonic dependence of $W_2(k)$ is observed, with the minimum near $k = 6$. These observations are valid only for $n_P \gtrsim 2k$.

D. Impact of local magnification effects on the performance of clustering methods

Due to the heterogeneity of the sample composition, distortions of the trajectories of evaporated atoms occur, manifested in local magnification effects [23]. Ambiguity in the tip curvature, crystal defects, and poles can also lead to observed density heterogeneities. An approach to simulating magnification effects by local anisotropic deformation of the precipitate

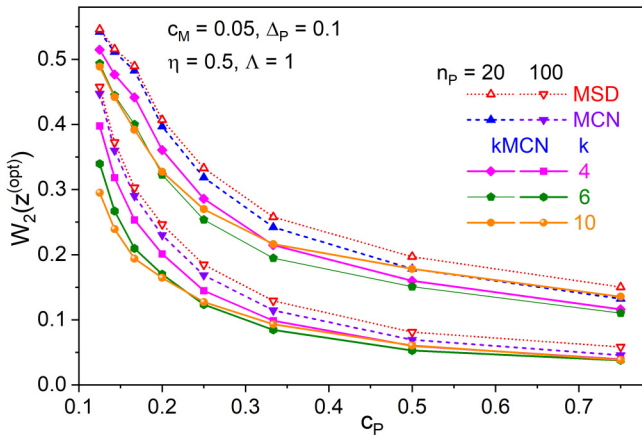


FIG. 10. Comparison of quality metrics (15) obtained by different methods depending on the composition and size of spherical precipitates.

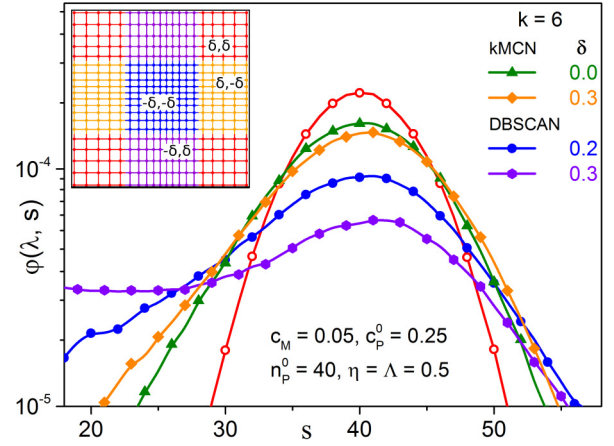


FIG. 11. The impact of heterogeneous strain on the results of cluster analysis obtained by kMCN and DBSCAN. Cluster size distributions at optimal control parameters are shown at several values of the strain parameter δ . The random structure properties are indicated on the plot. Inset: Two-dimensional illustration of piecewise constant deformation (1) of square lattice at the strain parameter $\delta = 0.3$. Circles and lines indicate atom positions and interatomic links, respectively.

region was described in [24]. We employ a similar, yet simplified, technique, applying non-uniform deformation according to (1).

The features of the applied deformation can be illustrated by a two-dimensional lattice, as shown in the inset to Fig. 11, where four types of deformed regions are distinguished, corresponding to diagonal elements of the strain tensor, $\hat{\varepsilon} \equiv (\varepsilon_{xx}, \varepsilon_{yy}) = (\pm\delta, \pm\delta)$. Regions of uniform expansion and compression are observed, as well as regions of shear. Since the two shear regions are equivalent, three types of deformation are represented in this lattice.

The strain tensor in a 3D lattice, as defined in (1), also has only nonzero diagonal elements and provides four types of deformed areas. One-eighth of the crystal is uniformly compressed, and the same part is uniformly expanded. The remaining parts are mixed strained, e.g., $\hat{\varepsilon}^* = \pm(\delta, \delta, -\delta)$, and contain both shear and dilatation components. The shapes of precipitates in regions with mixed strain change from spherical to ellipsoidal.

Figure 11 compares the sensitivity of the methods to non-uniform deformation. Even the smallest strain δ negatively affects the performance of DBSCAN, whereas the kMCN method demonstrates greater robustness to this type of deformation. As δ increase, some degradation in the quality of the cluster analysis using the kMCN method is attributable to the shear component of the strain.

Another advantage of the kMCN method is that the optimal value of the control parameter $z^{(\text{opt})}$, equal to 40 in this case, does not depend on the strain parameter δ . On the other hand, the optimal DBSCAN parameter $d_c^{(\text{opt})}$ decreases from 0.68 to 0.58 nm as δ increases from 0 to 0.3.

The kVCN method, based on Voronoi diagrams, demonstrates even greater resistance against strain-induced distortions. However, these benefits are evident only at low levels of noise in the atomic coordinates Λ . As the noise level

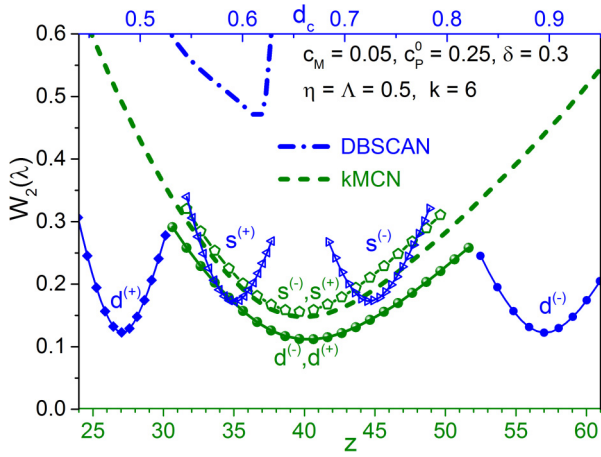


FIG. 12. Dependencies of clustering quality metrics on control parameters separately for each of the four types of regions of a heterogeneously deformed crystal by the field (1) at the strain $\delta = 0.3$. Filled symbols refer to pure dilatation ($d^{(\pm)}$), open symbols to regions of mixed deformation ($s^{(\pm)}$). The dashed and dash-dotted lines show the dependencies for the entire heterogeneously strained crystal.

increases, the advantages of kVCN vanish, and parity is reached between kVCN and kMCN methods.

Effects of pure dilatation and pure shear on clustering

Given the results presented in Fig. 11, the origin of this behavior remains unclear. When applying the strain field (1), some areas undergo pure uniform compression or expansion with local volume changes by factors of $(1 \pm \delta)^3$. The remaining parts of the crystal undergo mixed deformation, in which the local finite strain tensor can be decomposed into dilatation and pure shear components [46]. The decomposition implies dilatations with volume changes by factors of $(1 - \delta)(1 + \delta)(1 \pm \delta)$ and the shear strain of about 2δ .

Figure 12 shows how the clustering quality metrics W_2 depend on the control parameters z and d_c for kMCN and DBSCAN in each of the specified areas individually. The dilatation component of strain has no effect on clustering in kMCN (see green circles). The $W_2(z)$ dependence of mixed deformation areas runs slightly higher (see open green circles), reflecting the contribution of shear deformations. The resulting green dashed line lies in between. An important property of these three dependences is that they all reach their minima at the same value of the control parameter z , in this case around 40.

DBSCAN yields completely different results (see triangular and rhombic symbols in Fig. 12). A clear direct correlation with the densities of the four areas specified above is evident. The position of the local minimum $d_c^{(\text{opt})}$ of the function $W_2(d_c)$ is proportional to $\rho^{-1/3}$, where ρ is the average atomic density in the respective area. Due to variations in atomic densities, the optimal parameter $d_c^{(\text{opt})}$ in one area is not proper for another area. Consequently, the DBSCAN dependence $W_2(d_c)$ averaged over the entire crystal (dash-dotted line) lies significantly above the kMCN line. Therefore, it is the

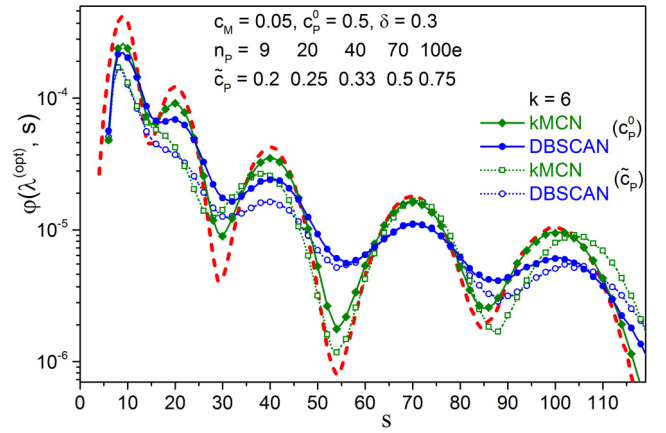


FIG. 13. Comparison of the kMCN and DBSCAN methods applied to heterogeneously deformed structures containing precipitates of different sizes and compositions. The nominal precipitate sizes are $n_p = 9, 20, 40, 70$ for spherical shapes and $n_p = 100$ for prolate spheroids. The red dashed line shows the precipitate size distribution. In a system with precipitates of the same nominal composition ($c_p^0 = 0.5$), the reduced cluster size distributions obtained at the optimal control parameters $z^{(\text{opt})} = 24$ and $d_c^{(\text{opt})} = 0.53$ nm are shown by filled symbols. In a system with precipitates of different compositions ($\tilde{c}_p = 0.2, 0.25, 0.33, 0.5, 0.75$), the optimal control parameters were $z^{(\text{opt})} = 27$ and $d_c^{(\text{opt})} = 0.55$ nm. The corresponding cluster distributions are shown as open symbols.

variation in dilatation deformations that leads to a significant degradation of DBSCAN performance.

Our analysis shows that when using both kMCN and DBSCAN, shear stress alone has little effect on clustering quality across the entire range of precipitate compositions and shear strain magnitudes studied. Clustering quality was assessed using metrics (15), with kMCN consistently outperforming DBSCAN.

E. Superposition of multiple-type precipitates

So far, we have considered clustering in systems with precipitates of a given single size n_p and single composition c_p . Since this study does not allow precipitate overlaps, the integral cluster distributions represent a superposition of the contributions from individual precipitates. Let us assume that during the simulation we have constructed a cluster distribution $\varphi(\lambda, n_p, c_p, s)$ of a given size n_p , and composition c_p for the set of control parameters λ . The cluster distribution in the mixed system is then expressed as a weighted sum of the obtained partial distributions,

$$\bar{\varphi}(\lambda, s) = \sum_{n_p, c_p} \Delta_{n_p c_p} \varphi(\lambda, n_p, c_p, s), \quad (17)$$

where $\Delta_{n_p c_p}$ is the fraction of solute atoms in precipitates of size n_p and solute concentration c_p .

Figure 13 demonstrates optimal cluster distributions in a strained alloy containing precipitates of different compositions, sizes and shapes. The cluster distributions shown were reduced according to (16). Optimal parameters were determined by (15). For the same nominal composition of all precipitates with $c_p/c_M = 10$, kMCN reproduces

the precipitate size distributions well despite the substantial strain value δ ; see filled diamonds in Fig. 13. The cluster size distribution yields the correct average value of the precipitate sizes and the fraction of precipitated solutes. The cluster distribution is very close to the given precipitate distribution for the largest sizes ($n_p > 40$), but the reproduction deteriorates for smaller sizes. This result is partly explained by the nonoptimality of the chosen control parameter $z = 24$ for small precipitates (see Fig. 7). However, the main reason is the dependence of the reproduction quality on the precipitate size, even with optimal parameters (see Fig. 10).

Heterogeneous strain is always challenging for DBSCAN. In the considered system it yields a cluster distribution that qualitatively reproduces the given precipitate distribution, though it underestimates the fraction of precipitated solutes.

Variations in precipitate composition significantly complicate cluster analysis, as shown by the open symbols in Fig. 13. In such a complex system, a single optimal compromise parameter $z^{(\text{opt})}$ is no longer capable of providing comparable efficiency to that in a system with single-phase precipitates. In this case, locally density-adaptive methods such as OPTICS [47,48] and hierarchical density-based cluster analysis [49] can be used. The application of these methods in the coordination-number metrics will be presented elsewhere.

VII. CONCLUSIONS

In this study, we extended the concept of the maximum coordination number [10] to the higher-order kMCN and kVCN methods. The first one ranks neighboring atoms by the Euclidean distance, whereas the second one uses geometric properties of the Voronoi diagram. Although the second approach may prove more robust to structural perturbations, both methods yield similar results when analyzing highly disordered APT structures [see Fig. 6(b)].

The correlation function of solute atoms in the ideal solid solution was exactly described by a negative binomial distribution (4), which constitutes the basis for composition spectrum analysis (9). We estimated the composition and effective fraction of interface layers as a function of structural disorder and precipitate size, as shown in Fig. 2.

Using the formalized approach based on coordination numbers, we investigated the percolation properties of APT structures. As shown in Fig. 3, the percolation thresholds p_c exhibit an approximately linear dependence of $1/p_c$ on the average coordination number z . Moreover, for high-order methods ($k > 4$) the quantity $p_c z/k$ proves to be almost invariant with respect to the clustering method parameters k and z , as well as the structural parameters Λ and η . This scaling behavior motivated the introduction of the self-similar variable (13), which provides a relevant description of clustering and enables the direct transfer of optimal parameters between methods (see Fig. 9). The behavior of clustering curves further clarifies the features of different clustering methods when applied to complex systems (see Fig. 5).

Simulations of cluster distributions in a two-phase system showed that the results depend on the clustering method, as illustrated in Figs. 6–10. Methods based on coordination numbers consistently outperformed density-oriented methods. We identified two main sources of discrepancy between the

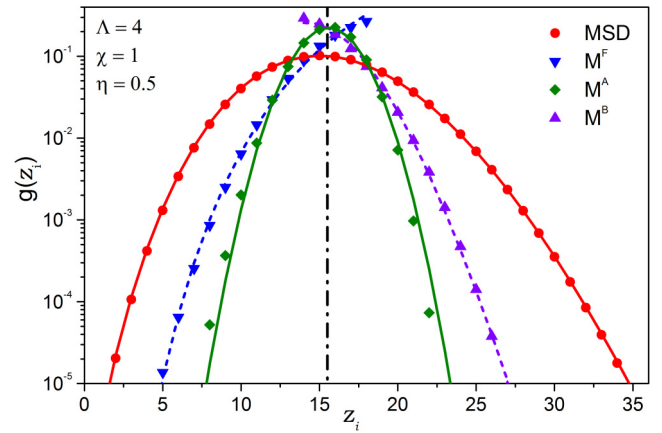


FIG. 14. The nearest-neighbor number distributions in MSD and variants of the MCN metric, using a gaslike random structure. The parameter z_0 was set to 18, 16, and 14 for the M^F , M^A , and M^B methods, respectively. The dash-dot line shows the location of the mean $z \approx 15.5$ for all distributions. Solid lines: Fittings of MSD by the Poisson and M^A by the Gaussian distributions, respectively. Dashed lines: Fittings of M^F and M^B by two-parameter gamma distributions.

calculated cluster distributions and the given precipitate distributions. The first one appears in the small-cluster regime as background clustering in the matrix. The second one arises from the attachment of matrix solutes to clusters within the precipitates (see Fig. 6). These precipitate clusters can be effectively separated from matrix clusters by subtracting the matrix contribution (16).

Regarding the order k of the clustering method, the quality of the cluster analysis improves distinctly as k increases when the c_P/c_M ratio is small. At large c_P/c_M ratios, however, the dependence on k becomes negligible (see Fig. 10).

Simulation of local magnification effects demonstrated a clear advantage of composition-based methods (kMCN) over density-based methods (DBSCAN). Pure shear was found to affect cluster analysis significantly only at high strain levels and to depend on precipitate composition. Under these conditions, kMCN slightly outperformed DBSCAN. The contrast between the two methods becomes especially pronounced under heterogeneous dilatation. kMCN is entirely resistant to dilation-type strains, whereas the performance of DBSCAN degrades rapidly as the amplitude of heterogeneous dilatation increases (see Figs. 11 and 13).

For the cluster analysis of APT data, we recommend replacing methods based on the Euclidean metric (e.g., DBSCAN) with their coordination-number-based counterparts, such as kMCN, which performed better in many cases, never worse, and required only a marginal increase in computational cost.

ACKNOWLEDGMENTS

The authors thank N. Wanderka and C. Abromeit for valuable discussions and critical comments. We also thank F. Vurpillot for clarifying details of the isopositional method. M.L. acknowledges Helmholtz-Zentrum Berlin for financial support and for providing access to the computing cluster used for the extensive numerical simulations.

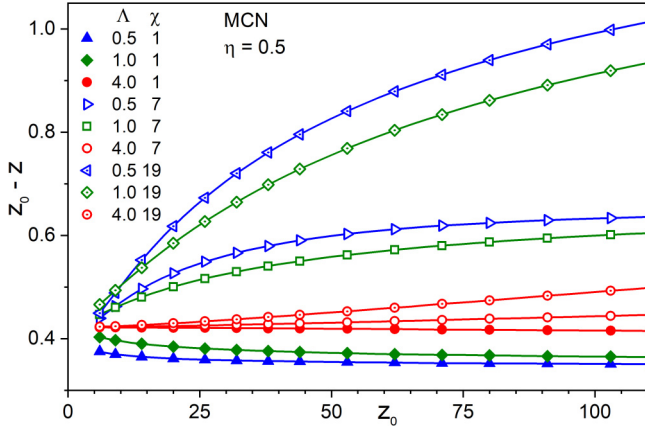


FIG. 15. Dependence of the average number of neighbors z on the MCN parameter z_0 for the metrics (2) at different values of spatial precision Λ and correlated loss χ .

DATA AVAILABILITY

Examples of the original software used to obtain the results are presented in the Supplemental Material [50] for this article. There are no other publicly available research data or software supporting this manuscript. Requests for further information or data should be sent to the authors.

APPENDIX A: THE EFFECT OF THE METRIC ON THE CONFIGURATION OF LINKS

1. Distributions of local neighbor numbers

We chose the arithmetic mean (M^A) of the coordination numbers (2) as the metric in this study; one could otherwise use the geometric (M^G) or harmonic (M^H) mean. Previously, we used marginal metrics based on backward [$M_{ij}^B = \min(m_{ij}, m_{ji})$] and forward [$M_{ij}^F = \max(m_{ij}, m_{ji})$] symmetrizations [10]. It was demonstrated that the metric M^F exhibits certain advantages in regions of abrupt change in atomic density, such as external surface and internal extended defects. Metric (2) preserves this property, as the inequalities $M^F \geq M^A \geq M^G \geq M^H \geq M^B$ hold, and M^A is closest to M^F .

The simplicity of computing metric (2), which uses only integer arithmetic, is also important.

Figure 14 compares the distributions of local neighbor numbers z_i for different metrics in a gas-like system. The number of neighbors within the Euclidean separation d_c follows the Poisson distribution with the variance equal to $z(d_c)$, as defined by (3). The marginal metrics M^F and M^B produce highly skewed distributions. The MCN method using metric (2) yields an almost symmetric distribution with the least variance, which can be roughly estimated by the relation $\sigma^2 \approx 1 + 0.12z$. Apparently, the narrowness of this distribution provides some advantages for the MCN method in cluster analysis.

To further reduce the variance, we can redefine the coordination numbers m_{ji} based on the coordination distances (2) rather than the Euclidean distances used initially. The first iteration of the redefinition reduces the variance by a factor of 2 to 3. Each subsequent iteration narrows the distribution even more. This procedure provides an easy alternative to constructing a topologically disordered lattice with a constant coordination number [52].

The dependencies of the mean z of the neighbor distribution on the parameter z_0 and the properties of structural disorder obtained within MCN and metric (2) are shown in Fig. 15. It appears that z is always slightly less than z_0 . For the uncorrelated atom loss ($\chi = 1$), the difference $z_0 - z$ does not exceed about 0.4 both in the APT-like structure ($\Lambda = 0.5 - 1$) and for the gas-type disorder ($\Lambda > 4$). For the correlated atom loss ($\chi > 1$), it increases but only slightly.

2. Maps of links

To gain a visual understanding of metric properties, it is helpful to compare the maps of links that form in a highly disordered system. To this end, we consider the same fragment of 2D atomic arrangement shown in Fig. 1 as a reference system. Figure 16 demonstrates the maps obtained by three methods with the control parameter $z_0 = 6$, equal to the average number of edges per Voronoi cell for a homogeneous Poisson point process in 2D.

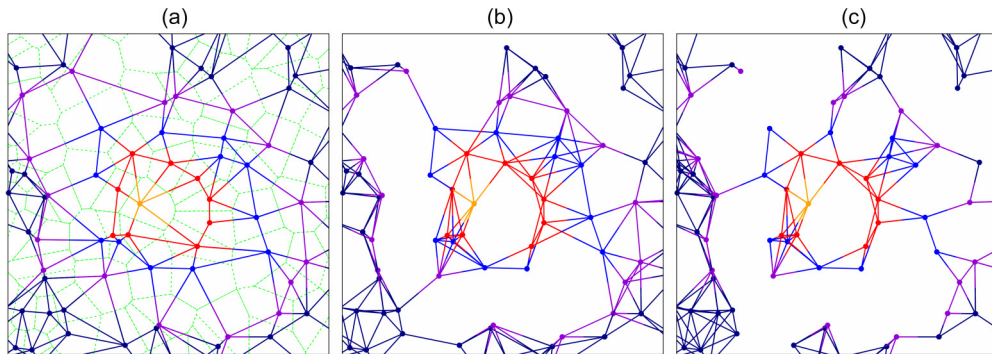


FIG. 16. The maps of links (solid lines) obtained in various metrics for the atomic arrangement shown in Fig. 1. The line colors represent the shells of the central atom. (a) VCN metric: The dashed lines depict the Voronoi diagram. The control parameter z_0 was set to 6, resulting in the average number of links z of about 5.4. (b) MCN metric: The same control parameter z_0 , resulting in $z \approx 5.6$. (c) MSD metric: Parameter d_c corresponds to $z = 6$ according to Eq. (3).

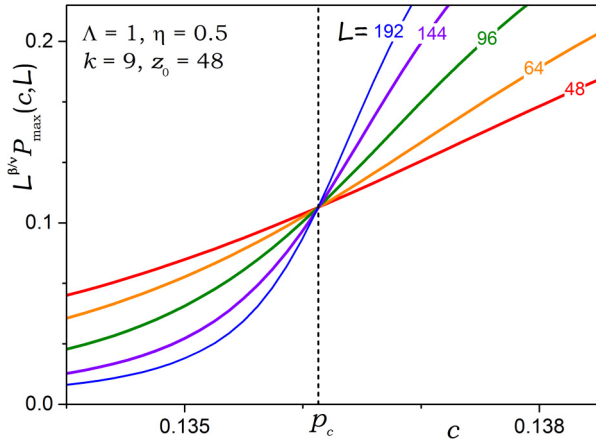


FIG. 17. Dependence of $\mathcal{L}^{\beta/\nu} P_{\max}(c, \mathcal{L})$ on the concentration c and the box size $\mathcal{L}$ for high-order ($k = 9$) percolation. The scaling parameter used is $\beta/\nu = 0.477$. The calculated coordination-number mean is $z = 47.627$ at $z_0 = 48$.

The VCN graph [Fig. 16(a)] appears to be the most balanced. Almost all atoms have 5 to 6 nearest neighbors. In contrast, the MSD graph [Fig. 16(c)] is the most unbalanced one. A distinct crowding of links is clearly visible on the left side and sparseness on the right side. The MCN graph [Fig. 16(b)] appears intermediate with an acceptable balance. It should be noted that the comparison of methods was chosen at the most favorable control parameter value for VCN: $z_0 = 6$. As z_0 increases, the properties of the VCN and MCN graphs converge. In 3D, we observed some advantages of the VCN in the range of about $z_0 \approx (12 - 20)$.

APPENDIX B: SIMULATION OF PERCOLATION THRESHOLDS IN HIGH-ORDER METHODS

To estimate the percolation threshold, the simulation proceeds in three stages: (a) a random site structure is generated

in a box of linear size $\mathcal{L}$ with disorder parameters η , Λ , and χ ; and neighbor lists are constructed for each site using z_0 to identify core points and z_L to define links between them; (b) a modified Newman–Ziff algorithm is then used to calculate the statistical quantity P_n in the microcanonical ensemble, with an additional check after each atom insertion to detect newly formed core points, while clustering is performed only on core points; and (c) after many repetitions of the previous steps, a convolution-type transformation is applied to obtain $P(c)$ in the canonical ensemble [42].

This algorithm yields the size of the largest cluster directly, so we use this quantity, normalized by the total number of sites, as $P(c)$. It is well established that the scaling behavior of the largest cluster size near the percolation threshold is given by [53]

$$P_{\max}(c, \mathcal{L}) \sim \mathcal{L}^{-\beta/\nu} \mathcal{F}((c - p_c)\mathcal{L}^{1/\nu}), \quad (\text{B1})$$

where $\mathcal{F}$ is a universal scaling function for systems with identical configurations and boundary conditions. In an ideal lattice, any box size $\mathcal{L}$ satisfies these requirements. In a random structure, however, $\mathcal{L}$ must be large enough to ensure statistical equivalence among systems of different sizes. For the same reason, a random structure is generated multiple times for sampling. For a typical box size of $\mathcal{L} = 100$, we use at least $N_{\text{conf}} = 250$ structure realizations (stage 1) and more than $N_{\text{iter}} = 10^5$ iterations (stage 2). The ratio of critical exponents β/ν depends only on the dimension of space d [40]. It is equal to $d - d_f$, where d_f is the fractal dimension of the percolating cluster, which is approximately 2.523 for $d = 3$ [54,55].

It follows from (B1) that the curves $\mathcal{L}^{\beta/\nu} P_{\max}(c, \mathcal{L})$ intersect near $c = p_c$ for any valid values of $\mathcal{L}$. Figure 17 illustrates an application of this approach. Under the conditions indicated in the plot, the obtained percolation threshold $p_c(k, z)$ is 0.13613(2). For comparison, formulas (10) and (11), using the parameters listed in Table I, yield values of 0.13622 and 0.13612, respectively.

- [1] B. Gault, A. Chieramonti, O. Cojocaru-Mirédin, P. Stender, R. Dubosq, C. Freysoldt, S. K. Makineni, T. Li, M. Moody, and J. M. Cairney, Atom probe tomography, *Nat. Rev. Methods Primers* **1**, 51 (2021).
- [2] M. K. Miller, *Atom-Probe Tomography: Analysis at the Atomic level* (Springer, New York, NY, 2000).
- [3] W. Lefebvre, F. Vurpillot, and X. Sauvage, *Atom Probe Tomography: Put Theory into Practice* (Academic Press, Oxford, 2016).
- [4] J. M. Hyde and C. A. English, An analysis of the structure of irradiation induced Cu-enriched clusters in low and high nickel welds, *MRS Proc.* **650**, 66 (2000).
- [5] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in *KDD-96 Proceedings* (AAAI Press, Washington, DC, 1996), pp. 226–231.
- [6] L. T. Stephenson, M. P. Moody, P. V. Liddicoat, and S. P. Ringer, New techniques for the analysis of fine-scaled clustering phenomena within atom probe tomography (APT) data, *Microsc. Microanal.* **13**, 448 (2007).
- [7] W. Lefebvre, T. Philippe, and F. Vurpillot, Application of Delaunay tessellation for the characterization of solute-rich clusters in atom probe tomography, *Ultramicroscopy* **111**, 200 (2011).
- [8] P. Felfer, A. Ceguerra, S. Ringer, and J. Cairney, Detecting and extracting clusters in atom probe data: A simple, automated method using Voronoi cells, *Ultramicroscopy* **150**, 30 (2015).
- [9] E. A. Marquis and J. M. Hyde, Applications of atom-probe tomography to the characterisation of solute behaviours, *Mater. Sci. Eng.: R: Rep.* **69**, 37 (2010).
- [10] M. Lazarev and J. Banhart, A concept of local coordination number for the characterization of solute clusters within atom probe tomography data, *Microsc. Microanal.* **30**, 793 (2024).
- [11] T. T. Tsong, S. B. McLane, Jr., M. Ahmad, and C. S. Wu, Field evaporation events as Markov chains: A time-of-flight

- atom-probe study of iridium, Pt-Rh alloys, and metallic glasses, *J. Appl. Phys.* **53**, 4180 (1982).
- [12] R. P. Kolli and D. N. Seidman, Comparison of compositional and morphological atom-probe tomography analyses for a multicomponent Fe-Cu steel, *Microsc. Microanal.* **13**, 272 (2007).
- [13] J. Hyde, E. Marquis, K. Wilford, and T. Williams, A sensitivity analysis of the maximum separation method for the characterisation of solute clusters, *Ultramicroscopy* **111**, 440 (2011).
- [14] A. V. Ceguerra, M. P. Moody, L. T. Stephenson, R. K. Marceau, and S. P. Ringer, A three-dimensional Markov field approach for the analysis of atomic clustering in atom probe data, *Philos. Mag.* **90**, 1657 (2010).
- [15] P. Styman, H. Weekes, J. Hyde, B. Radiguet, F. Vurpillot, G. Da Costa, C. Pareige, A. Etienne, and P. Pareige, APT characterisation of radiation damage in RPV steels: Experimental and data processing uncertainties, NUGENIA-PLUS: Nuclear Generation II and III Association (2016); see also W. Lefebvre-Ulrikson, G. Da Costa, L. Rigutti, and I. Blum, Data mining, in *Atom Probe Tomography* (Academic Press, Oxford, UK, 2016), Chap. 9, pp. 279–317.
- [16] T. Cover and P. Hart, Nearest neighbor pattern classification, *IEEE Trans. Inf. Theory* **13**, 21 (1967).
- [17] P. M. Ossi, *Disordered Materials: An Introduction* (Springer, Berlin, Heidelberg, 2006).
- [18] Y. Waseda, *The Structure of Non-Crystalline Materials: Liquids and Amorphous Solids* (McGraw-Hill, New York, 1980).
- [19] L. Mathieson and P. Moscato, An introduction to proximity graphs, in *Business and Consumer Analytics: New Ideas* (Springer, Cham, 2019), pp. 213–233.
- [20] L. T. Stephenson, M. P. Moody, and S. P. Ringer, Theory of solute clustering in materials for atom probe, *Philos. Mag.* **91**, 2200 (2011).
- [21] F. De Geuser, B. Gault, A. Bostel, and F. Vurpillot, Correlated field evaporation as seen by atom probe tomography, *Surf. Sci.* **601**, 536 (2007).
- [22] L. T. Stephenson, M. P. Moody, B. Gault, and S. P. Ringer, Nearest neighbour diagnostic statistics on the accuracy of APT solute cluster characterisation, *Philos. Mag.* **93**, 975 (2013).
- [23] M. K. Miller and M. G. Hetherington, Local magnification effects in the atom probe, *Surf. Sci.* **246**, 442 (1991).
- [24] J. M. Hyde, G. DaCosta, C. Hatzoglou, H. Weekes, B. Radiguet, P. D. Styman, F. Vurpillot, C. Pareige, A. Etienne, G. Bonny, N. Castin, L. Malerba, and P. Pareige, Analysis of radiation damage in light water reactors: Comparison of cluster analysis methods for the analysis of atom probe data, *Microsc. Microanal.* **23**, 366 (2017).
- [25] B. Gault, F. De Geuser, C. Freysoldt, B. Klaes, and F. Vurpillot, Spatial resolution(s) in atom probe tomography, *Microsc. Microanal.* **32**, ozag064 (2026).
- [26] L. A. Steen and J. A. Seebach, *Counterexamples in Topology* (Springer, New York, 1978).
- [27] K. Binder and D. W. Heermann, *Monte Carlo Simulation in Statistical Physics* (Springer, 2019), p. 258.
- [28] V. M. Worlitzer, G. Ariel, and E. A. Lazar, Pair correlation function based on Voronoi topology, *Phys. Rev. E* **108**, 064115 (2023).
- [29] J. Lu, E. A. Lazar, and C. H. Rycroft, An extension to Voro++ for multithreaded computation of Voronoi cells, *Comput. Phys. Commun.* **291**, 108832 (2023).
- [30] B. M. Jenkins, A. Zakirov, F. Vurpillot, A. Etienne, C. Pareige, P. Pareige, and B. Radiguet, On the iron content of Mn-Ni-Si-rich clusters that form in reactor pressure vessel steels during exposure to neutron irradiation, *Acta Mater.* **281**, 120384 (2024).
- [31] R. K. W. Marceau, Atomic-scale analysis of light alloys using atom probe tomography, *Mater. Sci. Technol.* **32**, 209 (2016).
- [32] M. H. DeGroot and M. J. Schervish, *Probability and Statistics*, 4th ed. (Addison-Wesley, Boston, 2012).
- [33] N. Wanderka, N. Lazarev, C. S. T. Chang, and J. Banhart, Analysis of clustering in Al-Mg-Si alloy by density spectrum analysis of atom probe data, *Ultramicroscopy* **111**, 701 (2011).
- [34] S. Chandrasekhar, Stochastic problems in physics and astronomy, *Rev. Mod. Phys.* **15**, 1 (1943).
- [35] T. Philippe, F. De Geuser, S. Duguay, W. Lefebvre, O. Cojocar-Mirédin, G. Da Costa, and D. Blavette, Clustering and nearest neighbour distances in atom-probe tomography, *Ultramicroscopy* **109**, 1304 (2009).
- [36] B. Bhattacharjee, nth-nearest-neighbor distribution functions of an interacting fluid, *Phys. Rev. E* **67**, 041208 (2003).
- [37] M. Timpel, N. Wanderka, R. Schlesiger, T. Yamamoto, N. Lazarev, D. Isheim, G. Schmitz, S. Matsumura, and J. Banhart, The role of strontium in modifying aluminium-silicon alloys, *Acta Mater.* **60**, 3920 (2012).
- [38] A. V. Morozov, *Methods for Solving Incorrectly Posed Problems* (Springer, New York, 1984).
- [39] T. Roths, M. Marth, J. Weese, and J. Honerkamp, A generalized regularization method for nonlinear ill-posed problems, *Comput. Phys. Commun.* **139**, 279 (2001).
- [40] D. Stauffer and A. Aharony, *Introduction to Percolation Theory* (Taylor & Francis, 1994), p. 192.
- [41] *Complex Media and Percolation Theory*, edited by M. Sahimi and A. G. Hunt (Springer, New York, 2021).
- [42] M. E. J. Newman and R. M. Ziff, Fast Monte Carlo algorithm for site or bond percolation, *Phys. Rev. E* **64**, 016706 (2001).
- [43] Z. Xun, D. Hao, and R. M. Ziff, Site and bond percolation thresholds on regular lattices, *Phys. Rev. E* **105**, 024105 (2022).
- [44] M. Li, R.-R. Liu, L. Lü, M.-B. Hu, S. Xu, and Y.-C. Zhang, Percolation on complex networks: Theory and application, *Phys. Rep.* **907**, 1 (2021).
- [45] C. P. Winsor, The Gompertz curve as a growth curve, *Proc. Natl. Acad. Sci. USA* **18**, 1 (1932).
- [46] E. H. Dill, *Continuum Mechanics: Elasticity, Plasticity, Viscoelasticity* (CRC Press, Boca Raton, 2006).
- [47] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, OPTICS: Ordering points to identify the clustering structure, *ACM SIGMOD Record* **28**, 49 (1999).
- [48] J. Wang, D. K. Schreiber, N. Bailey, P. Hosemann, and M. B. Toloczko, The application of the OPTICS algorithm to cluster analysis in atom probe tomography data, *Microsc. Microanal.* **25**, 338 (2019).
- [49] I. Ghamarian and E. A. Marquis, Hierarchical density-based cluster analysis framework for atom probe tomography data, *Ultramicroscopy* **200**, 28 (2019).
- [50] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/17ms-c7hr> for the complete source code, README documentation, and demonstration datasets used to simulate

- clustering in two-phase systems and to calculate percolation thresholds. Third-party computational tools used within the software package, which includes Refs. [29,42,51].
- [51] S. Harase and T. Kimoto, Implementing 64-bit maximally equidistributed F2-linear generators with Mersenne prime period, *ACM Trans. Math. Software* **44**, 1 (2018).
- [52] M. Schrauth and J. S. E. Portela, Fast algorithm for topologically disordered lattices with constant coordination number, *Phys. Rev. Res.* **1**, 033061 (2019).
- [53] S. Mertens, I. Jensen, and R. M. Ziff, Universal features of cluster numbers in percolation, *Phys. Rev. E* **96**, 052119 (2017).
- [54] Y. Deng and H. W. J. Blöte, Monte Carlo study of the site-percolation model in two and three dimensions, *Phys. Rev. E* **72**, 016126 (2005).
- [55] C. D. Lorenz and R. M. Ziff, Precise determination of the bond percolation thresholds and finite-size scaling corrections for the sc, fcc, and bcc lattices, *Phys. Rev. E* **57**, 230 (1998).