PCCP

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2023, **25**, 658

Molecular cluster analysis using local order parameters selected by machine learning

Kazuaki Z. Takahashi *

Accurately extracting local molecular structures is essential for understanding the mechanisms of phase and structural transitions. A promising method to characterize the local molecular structure is defining the value of the local order parameter (LOP) for each particle. This work develops the Molecular Assembly structure Learning package for Identifying Order parameters (MALIO), a machine learning package that can propose an optimal (set of) LOP(s) quickly and automatically for a huge number of LOP species and various methods of selecting neighboring particles for the calculation. We applied this package to distinguish between the nematic and smectic phases of uniaxial liquid crystal molecules, and selected candidate LOPs that could be used to precisely observe the nematic–smectic phase transition. The LOP candidates were used to observe the nucleation and subsequent percolation transition, and the effect of the choice of LOP species and neighboring particles on the statistics of local molecular structures (clusters) was examined. The procedure revealed the time evolution of the number of clusters and the dependence of the percolation curve on the number of neighboring particles for each LOP species. The LOP species with the lowest dependence on the number of neighboring particles was the best-performing LOP species in the MALIO screening strategy. These results not only show that machine learning can powerfully screen a huge number of LOP species and suggest only a few promising candidates, but also indicate that MALIO can select the best LOP species.

Received 11th August 2022,
Accepted 9th November 2022

DOI: 10.1039/d2cp03696g

rsc.li/pccp

1 Introduction

Phase transitions have been studied not only in condensed matter physics,[1] but also in a wide range of fields including biology and engineering.[2–6] In particular, controlling the rate at which the transition occurs has been a subject of interest for many years, not only because it directly affects the applicability of materials to devices in which the transition is an important part of the mechanism,[7,8] but also because it affects the speed of material generation and processing for higher yields of desirable material structures and prevention of undesirable structuring.[9–12] Regardless of whether the process is a first-order phase transition or phase separation, it is important to focus on local molecular structures to understand the transition phenomena. In phase transitions, the local structures occur in the bulk or at the interface before the transition, and often have post-transitional structural motifs.[1,11,13–21] Attempts to understand transition kinetics from the time evolution of the size and shape statistics of local molecular structures (clusters) have been made for many years and yielded many valuable results.[16,17,22–32] Such cluster analysis is supported by precisely extracting local molecular structures. For simple materials such as noble gases and metals, it is relatively easy to extract the local molecular structure by selecting local density differences.[33,34] However, ordering of materials with high electrostatic and molecular shape anisotropy, such as water, silica, liquid crystals, and polymers, is much more complex, and it is difficult to characterize differences in various ordered structures in terms of local density differences. For such problems, global order parameters have been sometimes used to justify cluster analysis based on slight differences in local density, but the limitations of this strategy have already been reported.[26,35] Recall that the global order parameter, by its definition, gives a simple value that indicates the degree of order in the entire molecular system, and thus the global order parameter alone cannot detect the local structure inside a molecular system. More recently, attempts have been made to define order parameter values for individual atoms and molecules (*i.e.*, local order parameters (LOPs)) and to characterize local molecular structures in more detail. LOPs were first developed to distinguish basic crystal structures such as body-centered cubic, face-centered cubic, hexagonal close-packed, and glass,[36–42] and is now a promising method to distinguish various molecular structures.[26,30,32,35,39,43–48] The advantage of an LOP is that it is possible to quantitatively determine which structure a

Research Center for Computational Design of Advanced Functional Materials, National Institute of Advanced Industrial Science and Technology (AIST), Central 2, 1-1-1 Umezono, Tsukuba, 305-8568, Ibaraki, Japan.
E-mail: kazu.takahashi@aist.go.jp; Tel: +81 29 861 2972

material belongs to for each of its constituents (such as atoms, molecules, and coarse-grained particles). This allows us to reveal local molecular structures and their distributions with very high resolution. In fact, several studies have already demonstrated that LOP can track self-organization dynamics during phase transitions.[26,30,32,35,49] Furthermore, it has been reported that 3D molecular coordinate data can be obtained from experiments and LOP values can be calculated to analyze experimental results.[50] However, information on neighboring particles is essential for calculating an LOP value. Therefore, the amount of information on neighboring particles affects the accuracy of the LOP value and thus the extraction of local molecular structures. As the number of neighboring particles increases, the LOP value may become more robust and the accuracy of structure determination may increase, while the resolution of local molecular structures may decrease. This trade-off between the number of neighboring particles and resolution has often been discussed in calculating LOP values,[35,51] but a unified view has not yet been obtained. Another problem is that the effective LOP species for a particular substance or phase structure is not self-evident, so it is necessary to select the best one from a huge number of LOP candidates.[32,42,45–48]

This work develops the molecular assembly structure learning package for identifying order parameters (MALIO), a machine learning package that can quickly and automatically suggest the optimal LOPs and LOP combinations for various protocols for selecting neighboring particles. MALIO is based on the machine learning-aided local structure analyzer (ML-LSA),[32,45] but attempts to speed up LOP calculations by completely re-coding in Cython[52] and improving related libraries. To evaluate the performance of LOP selection by MALIO, we focus on the nematic–smectic transition of uniaxial liquid crystal molecules. The nematic–smectic transition is a first-order phase transition. It has long been known that nucleation occurs in the early stage of the first-order phase transition, regardless of the substance.[1] Nucleation is a phenomenon in which the phase transition proceeds only when molecular clusters of a certain size are generated.[1] Therefore, the existence of molecular clusters with short-range order (CCs: cybotactic clusters) in the early stages of the liquid crystal phase transition has long been proposed and has been measured using X-rays in uniaxial and bent-core liquid crystal systems.[53,54] However, the shape and role of CCs during phase transitions have long remained unclear. Molecular cluster analysis applying LOP to trajectories of phase transition dynamics obtained from molecular dynamics simulations has revealed for the first time the shape and role of CCs during the phase transition.[32] A similar analysis has also been attempted for the isotropic-nematic transition.[55] MALIO is applied to distinguish between nematic and smectic phases of uniaxial liquid crystal molecules, and candidate LOPs are selected for each protocol for selecting neighboring particles. The selected LOP candidates are used to observe the nucleation and subsequent percolation transition, and the impact of both the LOP species and the protocols on cluster statistics are investigated. Note that although defect formation and destruction occur during the liquid crystal phase transition,[56,57] it is difficult to observe defects on the scale of an optical microscope in molecular simulations. Especially in the case of the smectic phase, nothing can be said about layer defects unless it is possible to identify which molecular assemblies constitute a single smectic layer. Also, from a molecular point of view, the formation and destruction of defects is likely to be a phenomenon that occurs at least further after the percolation transition that occurs after nucleation. Therefore, it is outside the scope of this work.

2 Methodology

2.1 Molecular dynamics simulations

Molecular dynamics (MD) simulations of the liquid crystal phase transition were performed using the Soft-Core Gay-Berne (SCGB) model in the large-scale atomic/molecular massively parallel simulator (LAMMPS).[58] The pairwise interaction potential of the SCGB model, U_{SCGB}, is expressed by the equations

$$U_{\mathrm{SCGB}} = (1 - f)U_{\mathrm{GB}} + fU_{\mathrm{SC}}, \tag{1}$$

$$f = 1/\{1 + \exp[b(\sigma_{\mathrm{a}} - r)]\}, \tag{2}$$

$$U_{\mathrm{GB}} = 4\varepsilon_{\mathrm{a}}\left[\left(\frac{\sigma_{\mathrm{s}}}{r - \sigma_{\mathrm{a}} + \sigma_{\mathrm{s}}}\right)^{12} - \left(\frac{\sigma_{\mathrm{s}}}{r - \sigma_{\mathrm{a}} + \sigma_{\mathrm{s}}}\right)^{6}\right], \tag{3}$$

$$U_{\mathrm{SC}} = a(r - \sigma_{\mathrm{a}}), \tag{4}$$

where f is a sigmoidal logistic function, U_{GB} is the pairwise interaction potential of the ellipsoidal GB particles, ε_{a} denotes the anisotropic energy of an ellipsoidal pair, r is the distance between the centers of mass of a pair of particles, σ_{a} is the anisotropic length of the ellipsoidal pair, σ_{s} is the length of the side-by-side configuration of the ellipsoids, U_{SC} is the soft-core potential energy, a is the potential slope of the soft repulsive energy barrier, and b is the steepness of the sigmoidal logistic seaming function. By introducing the parameter $\kappa = \sigma_{\mathrm{e}}/\sigma_{\mathrm{s}}$, in which σ_{e} denotes the length of the end-to-end configuration of the ellipsoids, we can write the anisotropic energy ε_{a} as

$$\varepsilon_{\mathrm{a}} = \varepsilon(\varepsilon_{\mathrm{a}}^{'})^{\mu}(\varepsilon_{\mathrm{a}}^{''})^{\nu}, \tag{5}$$

where ε denotes the characteristic well depth of the interaction potential, $\varepsilon_{\mathrm{a}}^{'}$ and $\varepsilon_{\mathrm{a}}^{''}$ denote the contributions corresponding to the well depth and configuration anisotropies, and μ and ν are multipliers for determining these two contributions to the pair potential. The factor $\varepsilon_{\mathrm{a}}^{'}$ is characterized by introducing a parameter $\kappa' = \varepsilon_{\mathrm{s}}^{'}/\varepsilon_{\mathrm{e}}^{'}$, where $\varepsilon_{\mathrm{e}}^{'}$ and $\varepsilon_{\mathrm{s}}^{'}$ denote energy contributions from the end-to-end and side-by-side ellipsoid configurations, respectively. The factor $\varepsilon_{\mathrm{a}}^{''}$ is characterized by κ. Therefore, the detailed shape of U_{GB} is determined from the values of the four parameters κ, κ', μ, and ν. We set $\kappa = 3$, $\kappa' = 5$, $\mu = 1$, and $\nu = 3$. Note that the above parameter set has been traditionally used because the physical properties of nematic and smectic B liquid crystal phases are well displayed.[59,60] For the characteristic length, energy, and mass of the SCGB systems, $\sigma = \sigma_{\mathrm{s}}$, $\varepsilon = k_{\mathrm{B}}T$,

and m are each set to 1; here, m is the mass of one SCGB particle. The terms a and b were set to $-70\varepsilon\sigma^{-1}$ and $-100\sigma^{-1}$, respectively, on the basis of previous reports.[60,61] Using the above parameter settings, the nematic–smectic transition is guaranteed to be observed for SCGB systems by quenching from temperature T = 2.4 to 1.8 at a density of $0.3\sigma^{-3}$.[60] Therefore, an ensemble was used having a constant number of particles at a density of $0.3\sigma^{-3}$ contained within a cubic box of constant volume and temperature with full periodic boundary conditions imposed. The initial configurations for the nematic–smectic phase transition trajectory were prepared carefully by being cooled gradually from the isotropic phase at T = 6.0 to the nematic phase at T = 2.4. The configurations were then quenched to T = 1.8 below the nematic–smectic transition temperature $T_{\mathrm{N-Sm}}$ = 2.25. To observe the fast nucleation during the weak first-order phase transition, a velocity Verlet integrator with a fine timestep of $2.0 \times 10^{-5}\tau$ was used for SCGB, where $\tau = (m\sigma/\varepsilon)^{1/2}$ is a time unit. The temperature was controlled using a Nosé–Hoover chain thermostat.[62] The SCGB potential was truncated at 8.0σ to precisely compute the intermolecular interactions during the phase transition.

2.2 Local order parameters

MALIO implements 17 typical definition functions of LOPs that have been developed individually and are introduced below. Note that most of the LOPs have internal parameters, and a huge number of LOPs can be considered by changing these parameters.

The neighborhood parameter A was developed by Honeycutt and Andersen[63,64] and by Radhi and Behdinan[41] to characterize the crystal structure of Lennard-Jones fluid based on the distance between the pair particles and its neighbor particles. A is expressed as

$$A_M^1(i) = \frac{1}{N}\sum_{j\in N_{\mathrm{b}}(i)}\left|\sum_{k\in N_{\mathrm{b}}(i,j)}(r_{ik}+r_{jk})\right|^2 \quad (6)$$

$$A_M^2(i) = \frac{1}{N}\sum_{j\in N_{\mathrm{b}}(i)}\left|\sum_{k\in N_{\mathrm{b}}(i,j)}(r_{ij}+r_{kj})\right|^2 \quad (7)$$

$$A_M^3(i) = \frac{1}{N}\left|\sum_{j\in N_{\mathrm{b}}(i)}\sum_{k\in N_{\mathrm{b}}(i,j)}(r_{ij}+r_{kj})\right|^2, \quad (8)$$

where $N_{\mathrm{b}}(i)$ is an array that stores the identifiers of neighbor beads of bead i in order of decreasing distance from bead i, $N_{\mathrm{b}}(i,j)$ is an array including the identifiers of mutually neighboring particles of particles i and j, N and M denote the numbers of elements in arrays $N_{\mathrm{b}}(i)$ and $N_{\mathrm{b}}(i,j)$, and r_{ij} is a vector from particle j to particle i. We consider M = 1, 2, and 3. MALIO uses local averaging to improve the accuracy of molecular structure classification by LOPs. The "locally averaged" neighborhood parameter $\bar{A}$ is defined as

$$\bar{A}_M^{\mathrm{var}}(i) = \frac{1}{N+1}\sum_{j\in\tilde{N}_{\mathrm{b}}(i)} A_M^{\mathrm{var}}(j), \quad (9)$$

where the superscript "var" indicates the variation of A from 1 to 3, and $\tilde{N}_{\mathrm{b}}(i)$ is an array including the identifiers of neighboring particles and particle i. The local averaging is based on $\tilde{N}_{\mathrm{b}}$ and does not lead to an increase in the effective number of neighboring particles. This process is the same as that used for spherical harmonic functions in the modified bond-orientational order parameters used by Lechner and Dellago,[39] described below. The neighbors of particle i are not the same as those of particle j. From this uniqueness of $\tilde{N}_{\mathrm{b}}$, the LOP values for particles i and j are different regardless of the local averaging process. Importantly, differences in the LOP values determine the resolution at which we can distinguish local molecular structures. Thus, excessively repeating the averaging operation does cause any loss of local information by diminishing the differences.

The bond-angle order parameter B was used by Ackland and Jones[38] to identify dislocation defects in colloidal suspensions based on the bond angle between the central particle and its neighbor particles. B is expressed as

$$B_{n_1,n_2,\phi}(i) = \frac{1}{N(N-1)/2}\sum_{j>\,k\in N_{\mathrm{b}}(i)} f(\theta_{jik}) \quad (10)$$

$$f(\theta_{jik}) = \cos^{n_1}(n_2\theta_{jik}+\phi), \quad (11)$$

where θ_{jik} is the angle between r_{ij} and r_{ik}, n_1 and n_2 are positive integers, and ϕ is an offset angle. We consider n_1 = 1 and 2; n_2 = 1, 2, and 3; and ϕ = 0, $2/3\pi$, $\pi/2$, $\pi/3$, $\pi/4$, $\pi/5$, and $\pi/6$. The locally averaged neighborhood parameter $\bar{B}$ is defined as

$$\bar{B}_{n_1,n_2,\phi}(i) = \frac{1}{N+1}\sum_{j\in\tilde{N}_{\mathrm{b}}(i)} B_{n_1,n_2,\phi}(j). \quad (12)$$

The centrosymmetry parameter C was used by Kelchner and co-workers[37] to analyze dislocations and defects on metal surfaces based on the distance between the central particle and its neighbor particles. C is expressed as

$$C(i) = \sum_{j\in N_{\mathrm{b}}'(i)} |r_{ij}+r_{ik}|^2, \quad (13)$$

where $N_{\mathrm{b}}'(i)$ is an array including the identifiers of neighbor beads of half of $N_{\mathrm{b}}(i)$ in order of nearest neighbor from particle i, N' denotes the number of elements in array $N_{\mathrm{b}}'(i)$, and k satisfies the relation $k = j + N - N'$. The locally averaged neighborhood parameter $\bar{C}$ is defined as

$$\bar{C}(i) = \frac{1}{N+1}\sum_{j\in\tilde{N}_{\mathrm{b}}(i)} C(j). \quad (14)$$

The neighbor distance parameter D was used by Stukowski[40] to identify the crystal structure at grain boundaries by introducing a scale factor related to the distance of neighbor particles.

Table 1 Functions represented by f_α, f_β, and f_γ

Function	Formula
f_1	r
f_2	r^2
f_3	$\sqrt{r}$

D is expressed as

$$D_{f_\alpha,f_\beta,f_\gamma}(i) = \frac{1}{N(N-1)/2} \times \sum_{j>k\in N_b(i)} f_\alpha(r_{ij}) f_\beta(r_{ik}) f_\gamma(r_{jk}), \quad (15)$$

where r is the interparticle distance, and f_α, f_β, and f_γ are the scale factor functions of r. In MALIO, f_α, f_β, and f_γ can be freely defined by users as scalar functions of r. Table 1 shows the functions represented by f_α, f_β, and f_γ in this work. In the following sections, the specific form of f_α, f_β, or f_γ corresponds to the function listed for each subscript of the parameters in Table 1. Note that α, β, and γ correspond to the subscript numbers of the function f in Table 1, respectively. The locally averaged neighborhood parameter $\bar{D}$ is defined as

$$\bar{D}_{f_\alpha,f_\beta,f_\gamma}(i) = \frac{1}{N+1} \sum_{j\in\bar{N}_b(i)} D_{f_\alpha,f_\beta,f_\gamma}(j). \quad (16)$$

The angular Fourier series parameter F was developed by Bartok and co-workers[43,44] to analyze potential energy surfaces in bulk crystals and silicon based on periodic properties of the structure. F is expressed as

$$F_{f_\alpha,f_\beta,a}(i) = \frac{1}{N(N-1)/2} \times \sum_{j>k\in N_b(i)} f_\alpha(\min(r_{ij},r_{ik})) f_\beta(\max(r_{ij},r_{ik})) \times \cos(a\theta_{jik}), \quad (17)$$

where a is an angular factor. We consider a = 1.0, 2.0, 3.0, 4.0, 6.0, 8.0, π/ϕ_0, $2\pi/\phi_0$, $3\pi/\phi_0$, $4\pi/\phi_0$, and $6\pi/\phi_0$, where ϕ_0 = 109.5π/180. The locally averaged neighborhood parameter $\bar{F}$ is defined as

$$\bar{F}_{f_\alpha,f_\beta,a}(i) = \frac{1}{N+1} \sum_{j\in\bar{N}_b(i)} F_{f_\alpha,f_\beta,a}(j). \quad (18)$$

The angle histogram parameter H developed in previous work[45] is expressed as

$$H_\nu(i) = FT_{\text{ampl.}}(h(\theta_{jik}))\delta(\tau-\nu), \quad (19)$$

where $FT_{\text{ampl.}}$ is the amplitude function after a Fourier transform, h is a function representing the histogram, δ is the Dirac delta function, and ν is the frequency of the Dirac delta function. We consider ν = 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, and 12.0. The locally averaged angle histogram parameter $\bar{H}$ is defined as

$$\bar{H}_\nu(i) = \frac{1}{N+1} \sum_{j\in\bar{N}_b(i)} H_\nu(j). \quad (20)$$

The tetrahedral order parameter I was developed by Chau and Hardwick[65,66] to evaluate the tetrahedral configurations of molecules and was applied to water, methane and Lennard-Jones fluids. I is expressed as

$$I(i) = 1 - \frac{3}{8} \sum_{j>k\in N_b(i)} [\cos(\theta_{jik}) + 1/3]^2. \quad (21)$$

The locally averaged neighborhood parameter $\bar{I}$ is defined as

$$\bar{I}(i) = \frac{1}{N+1} \sum_{j\in\bar{N}_b(i)} I(j). \quad (22)$$

The bond-orientational order parameters Q^S and W^S, based on spherical harmonic functions, were originally developed by Steinhardt and co-workers[36] to quantitatively evaluate the orientational order of supercooled liquids and metallic glasses. Q^S and W^S are expressed as

$$Q_l^S(i) = \sqrt{\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |q_{lm}(i)|^2} \quad (23)$$

$$W_l^S(i) = \sum_{m_1+m_2+m_3=0} \begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} q_{lm_1}(i) q_{lm_2}(i) q_{lm_3}(i) \Big/ \left(\sum_{m=-l}^{l} |q_{lm}(i)|^2 \right)^{3/2} \quad (24)$$

$$q_{lm}(i) = \frac{1}{N} \sum_{j\in N_b(i)} Y_{lm}(r_{ij}), \quad (25)$$

where l is an arbitrary positive integer denoting the degree of the harmonic function, m is an integer that runs from $-l$ to $+l$, and Y_{lm} is a spherical harmonic function. We consider l = 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, and 24. It is worth noting that due to the versatility of bond-orientational order parameters, several variations have been devised and applied to the evaluation and analysis of various molecular structures. The locally averaged bond-orientational order parameters $\bar{Q}^S$ and $\bar{W}^S$ are defined as

$$\bar{Q}_l^S(i) = \frac{1}{N+1} \sum_{j\in\bar{N}_b(i)} Q_l^S(j), \quad (26)$$

$$\bar{W}_l^S(i) = \frac{1}{N+1} \sum_{j\in\bar{N}_b(i)} W_l^S(j). \quad (27)$$

The modified bond-orientational order parameters Q^L and W^L developed by Lechner and Dellago[39] demonstrated the ability to distinguish between crystals and supercooled liquids of Lennard-Jones fluids by locally averaging the spherical

harmonic function term q_{lm}. Note that the local averaging process applied to all LOPs in this work was an operation inspired by the above fact. Q^{L} and W^{L} are expressed as

$$Q_l^{\mathrm{L}}(i) = \sqrt{\frac{4\pi}{2l+1}\sum_{m=-l}^{l}|\bar{q}_{lm}(i)|^2} \tag{28}$$

$$W_l^{\mathrm{L}}(i) = \sum_{m_1+m_2+m_3=0}\begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} \bar{q}_{lm_1}(i)\bar{q}_{lm_2}(i)\bar{q}_{lm_3}(i) \Big/ \left(\sum_{m=-l}^{l}|\bar{q}_{lm}(i)|^2\right)^{3/2} \tag{29}$$

$$\bar{q}_{lm}(i) = \frac{1}{N+1}\sum_{j\in\tilde{N}_{\mathrm{b}}(i)} q_{lm}(j). \tag{30}$$

The locally averaged modified bond-orientational order parameters $\bar{Q}^{\mathrm{L}}$ and $\bar{W}^{\mathrm{L}}$ are defined as

$$\bar{Q}_l^{\mathrm{L}}(i) = \frac{1}{N+1}\sum_{j\in\tilde{N}_{\mathrm{b}}(i)} Q_l^{\mathrm{L}}(j), \tag{31}$$

$$\bar{W}_l^{\mathrm{L}}(i) = \frac{1}{N+1}\sum_{j\in\tilde{N}_{\mathrm{b}}(i)} W_l^{\mathrm{L}}(j). \tag{32}$$

The alternative bond-orientational order parameters LQ and LW,[30,67] for which q_{lm} was normalized, were useful in analyzing local molecular structures in ice nucleation, growth, or melting. LQ and LW are expressed as

$$LQ_l(i) = \frac{1}{N} \times \sum_{j\in\tilde{N}_{\mathrm{b}}(i)} \frac{\sum_{m=-l}^{l} q_{lm}(i)q_{lm}^*(j)}{\left|\sum_{m=-l}^{l} q_{lm}(i)q_{lm}^*(j)\right|\left|\sum_{m=-l}^{l} q_{lm}(j)q_{lm}^*(i)\right|} \tag{33}$$

$$LW_l(i) = \sum_{m_1+m_2+m_3=0}\begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} \times lq_{lm_1}(i)lq_{lm_2}(i)lq_{lm_3}(i) \Big/ \left(\sum_{m=-l}^{l}|lq_{lm}(i)|^2\right)^{3/2} \tag{34}$$

$$lq_{lm}(i) = \frac{1}{N}\sum_{j\in\tilde{N}_{\mathrm{b}}(i)} \frac{q_{lm}(i)q_{lm}^*(j)}{|q_{lm}(i)||q_{lm}(j)|}. \tag{35}$$

The locally averaged alternative bond-orientational order parameters $\bar{L}Q$ and $\bar{L}W$ are defined as

$$\bar{L}Q_l(i) = \frac{1}{N+1}\sum_{j\in\tilde{N}_{\mathrm{b}}(i)} LQ_l(j), \tag{36}$$

$$\bar{L}W_l(i) = \frac{1}{N+1}\sum_{j\in\tilde{N}_{\mathrm{b}}(i)} LW_l(j). \tag{37}$$

The modified alternative bond-orientational order parameters LQ^{M} and LW^{M}, locally averaged over the lq_{lm} of LQ and LW, were implemented in MALIO. LQ^{M} and LW^{M} are expressed as

$$LQ_l^{\mathrm{M}}(i) = \frac{1}{N} \times \sum_{j\in\tilde{N}_{\mathrm{b}}(i)} \frac{\sum_{m=-l}^{l} \bar{q}_{lm}(i)\bar{q}_{lm}^*(j)}{\left|\sum_{m=-l}^{l} \bar{q}_{lm}(i)\bar{q}_{lm}^*(j)\right|\left|\sum_{m=-l}^{l} \bar{q}_{lm}(j)\bar{q}_{lm}^*(i)\right|} \tag{38}$$

$$LW_l^{\mathrm{M}}(i) = \sum_{m_1+m_2+m_3=0}\begin{pmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{pmatrix} \times \bar{l}q_{lm_1}(i)\bar{l}q_{lm_2}(i)\bar{l}q_{lm_3}(i) \Big/ \left(\sum_{m=-l}^{l}|\bar{l}q_{lm}(i)|^2\right)^{3/2} \tag{39}$$

$$\bar{l}q_{lm}(i) = \frac{1}{N+1}\sum_{j\in\tilde{N}_{\mathrm{b}}(i)} lq_{lm}(j). \tag{40}$$

The locally averaged modified alternative bond-orientational order parameters $\bar{L}Q^{\mathrm{M}}$ and $\bar{L}W^{\mathrm{M}}$ are defined as

$$\bar{L}Q_l^{\mathrm{M}}(i) = \frac{1}{N+1}\sum_{j\in\tilde{N}_{\mathrm{b}}(i)} LQ_l^{\mathrm{M}}(j), \tag{41}$$

$$\bar{L}W_l^{\mathrm{M}}(i) = \frac{1}{N+1}\sum_{j\in\tilde{N}_{\mathrm{b}}(i)} LW_l^{\mathrm{M}}(j). \tag{42}$$

The Legendre polynomial parameter S considered in this work is expressed as

$$S_n(i) = \frac{1}{N}\sum_{j\in\tilde{N}_{\mathrm{b}}(i)} P_n(u_i\cdot u_j), \tag{43}$$

where P_n is an n-th order Legendre polynomial function, and u is a unit direction vector. S is inspired by Onsager's order parameter,[68] but the molecular system average of S does not match Onsager's order parameter because the orientation direction of particle i is used instead of the average orientation direction. Note that S_2 is used to observe the isotropic-nematic transition of uniaxial liquid crystals.[55] We consider n = 2, 4, and 6. The locally averaged Legendre polynomial parameter $\bar{S}$ is defined as

$$\bar{S}_n(i) = \frac{1}{N+1}\sum_{j\in\tilde{N}_{\mathrm{b}}(i)} S_n(j). \tag{44}$$

The modified Legendre polynomial parameter T in this work is expressed as

$$T_{n,d}(i) = \frac{1}{N}\sum_{j\in\tilde{N}_{\mathrm{b}}(i)} P_n(u_i\cdot u_j)\cos(2\pi z/d), \tag{45}$$

where z is the distance from particle j to the plane perpendicular to the orientation direction of particle i containing the coordinates of particle i, and the parameter d is the distance between parallel layered structures. T is inspired by MacMillan's smectic order parameter,[69] but the molecular system average of T does not match MacMillan's order parameter because the orientation direction of particle i is used instead of the average orientation direction. We consider d = 2.0, 2.25, 2.5, 2.75, 3.0, 3.25, 3.5, and 3.75. The locally averaged Legendre polynomial parameter $\bar{T}$ is defined as

$$\bar{T}_{n,d}(i) = \frac{1}{N+1} \sum_{j \in \bar{N}_b(i)} T_{n,d}(j). \quad (46)$$

Because actual neighbors such as $N_b(i)$ and $N_b(i,j)$ directly influence the LOP values, their selection is an important factor in extracting local molecular structures. Several protocols are possible, such as using particles within the cutoff radius as neighbors,[26,35,42,45–48] predetermining the number of neighboring particles,[32,42,45–48] and using Delaunay triangulation based on Voronoi diagrams.[48,51,70,71] All of the above protocols are implemented in MALIO, and can be selected according to the molecular structure characteristics and the correspondence with various additional analyses after the LOP values are determined. In this work, the number of neighboring particles was fixed on the basis of previous studies[32,45] on liquid crystal molecules. Specifically, N was varied from 6 to 14, and the optimal LOP for each N was searched.

In addition to single LOPs, the classification performances of combinations of two LOPs were also evaluated. Therefore, a total of 2 220 777 ($=({}_{702}C_1 + {}_{702}C_2) \times 9$ neighboring conditions) different combinations of LOPs were considered.

2.3 Machine learning strategy for screening LOPs

It is difficult to examine in detail the ability to extract the local molecular structure in transition for more than 2 million LOP combinations, even with the high efficiency of machine learning. Therefore, MALIO screens LOPs according to the most important and simple question: Can LOPs successfully distinguish between pre- and post-transition molecular structures? This is because, at minimum, good LOPs must be able to distinguish between pre- and post-transition molecular structures with high accuracy to enable comprehensive observation of events during transition. The procedure for performing the above screening is shown in Fig. 1: (i) first, small nematic and smectic molecular structures were entered into MALIO to serve as motifs for the molecular systems before and after the transition. The input structures were created in the same manner as described in the subsection "Molecular Dynamics Simulations", but the number of molecules was set to 1701, which is small enough to create well-defined molecular structures. The number of input structures was set to 200 for each of the two structures. (ii) MALIO extracted each SCGB particle and its neighboring particles from each molecular structure using the neighboring protocol and determined the local particle coordinates L_i. Here, we chose the protocol in which the number of neighboring particles is predetermined. That is, L_i represents the information on particle i and the nearest to N-th nearest neighbor from particle i. (iii) Structural descriptors were determined by calculating 702 LOPs for all 680 400 (= 1701 particles × 200 coordinate data × 2 phases) of L_is. The structure name (nematic or smectic) was tagged onto each L_i and designated as the objective variable. (iv) The structure descriptors and objective variables were then stored in the descriptor array D_s and the structure name vector n, respectively. (v) Finally, the random forest method implemented in Scikit-learn (version 0.20.3)[72] was used to estimate the operator vector w satisfying the relation $D_s w = n$. The number of trees in the forest was set to 100, and the maximum depth of each tree was set to 10. Default values in Scikit-learn were used for the other settings. The quality of w was also checked by 5-fold cross-validation implemented in Scikit-learn. The combination of two LOPs can be considered by arbitrarily selecting two of the elements of w and setting the others to null. This process was carried out using a sequential forward selection algorithm.[73] The classification accuracy of each (set of) LOP(s) was calculated from the correct tagging rate (CTR) as implemented in MALIO. The CTR c is defined as

$$c = \frac{Z_{\text{correct}}}{Z_{\text{total}}}, \quad (47)$$

where Z_{correct} is the number of correct tags obtained from $D_s w$, and Z_{total} is the total number of tags. LOPs were screened using

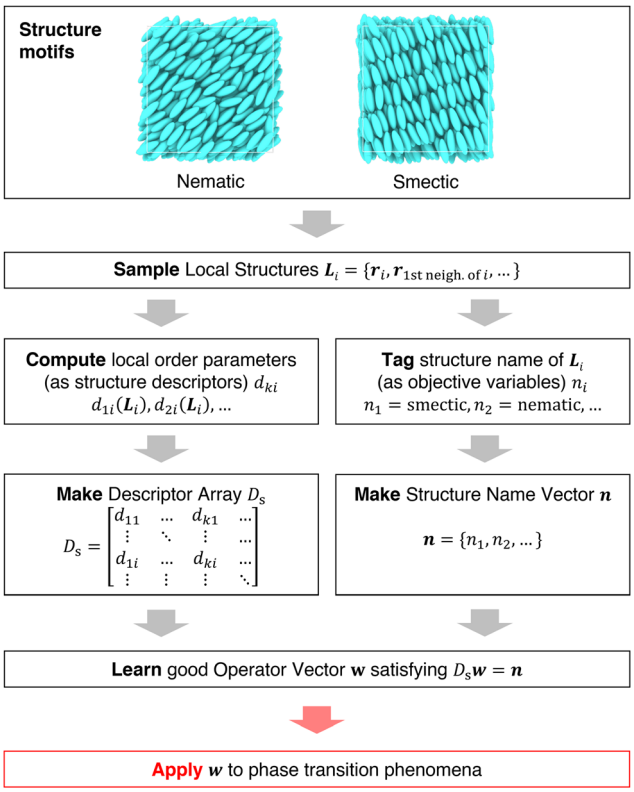

Fig. 1 Screening steps.

the CTR. The (set of) LOP(s) that performed best (*i.e.*, the maximum value of c) for each N was finally selected.

2.4 Application of selected LOPs to the nematic–smectic transition

Once the candidate LOPs were carefully selected using the screening strategy, the LOPs were applied to observe the nematic–smectic phase transition. The following procedure was implemented: (i) candidate LOPs were calculated for the time evolution of the three-dimensional coordinate data of a molecular system in which a monodomain nematic phase with 1 million SCGB particles was quenched to the temperature at which it becomes smectic (temperature $T = 1.8$; see the subsection "Molecular Dynamics Simulations"). (ii) The values of the specific (set of) LOP(s) were stored in the descriptor array $D_{s,q}$. The elements of $D_{s,q}$ corresponding to other LOPs were set to null. Here, $D_{s,q}$ is a function of time corresponding to the coordinates at each time. (iii) We estimated whether each particle belonged to the nematic- or smectic-like structure by applying w from the previous subsection to $D_{s,q}$ (the elements other than the candidate LOPs were set to null). Note that w is an operator vector learned from small nematic and smectic molecular structures and is independent of time. (iv) A cluster analysis was performed for the smectic molecules on the basis of the structure name tagged onto each particle at each time point. In this case, the neighboring conditions used as criteria for particle grouping were identical to those of the LOP neighboring protocol. The above procedure was used to obtain the kinetics of the phase transition phenomena, including nucleation and subsequent percolation, in the form of cluster statistics. The obtained results were compared to examine the influence of the choice of LOP species and neighboring particles on the cluster statistics.

3 Results and discussion

3.1 Evaluation of MALIO's ability to compute local order parameters

MALIO is a package that is an advanced version of ML-LSA with complete re-coding in Cython and improvements in related libraries, and is expected to speed up LOP computation, but it should be quantitatively demonstrated how much speedup is actually achieved. Fig. 2 shows the dependence of LOP computational time on the number of neighboring particles when using (a) ML-LSA and (b) MALIO; the ratio of LOP computational time for ML-LSA to that for MALIO is also plotted in (c). One core of Intel(R) Xeon(R) Gold 6252 CPU @ 2.10 GHz was used to measure the LOP computation time. The LOP computation time plotted in Fig. 2 was measured at the end of the LOP computation for 680,400 of L_is. $\bar{A}_2^1$ is the LOP with relatively high computational cost, $\bar{Q}_2^{\mathrm{L}}$ is the LOP calculated by MALIO's original spherical harmonic function library, and $\bar{S}_2$ is the LOP with relatively low computational cost. In ML-LSA, the computational time for $\bar{A}_2^1$ and $\bar{Q}_2^{\mathrm{L}}$ increased with increasing N, while the computational time for $\bar{S}_2$ hardly increased with increasing N. In MALIO, the computational time was much shorter than in ML-LSA. The computational time for $\bar{A}_2^1$ increased with

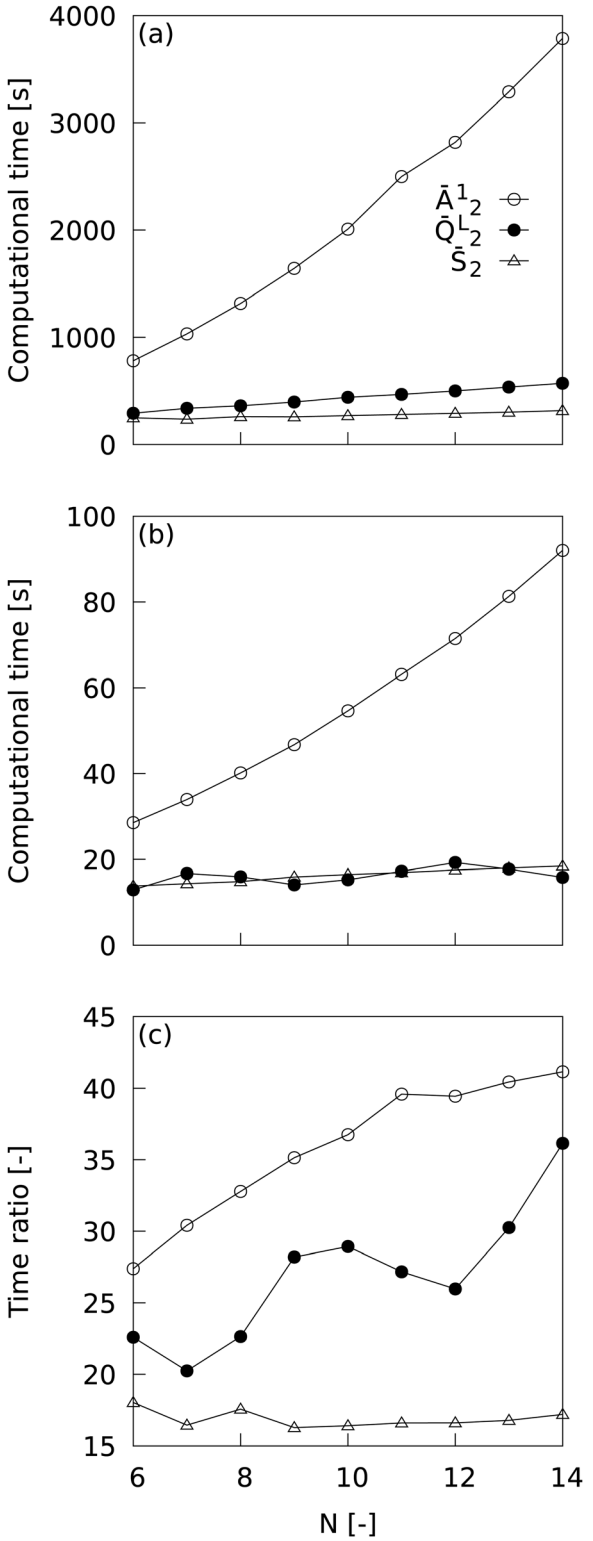

Fig. 2 Dependence of LOP computational time on the number of neighboring particles when using (a) ML-LSA and (b) MALIO. The ratio of LOP computational time for ML-LSA to that for MALIO is also plotted in (c).

increasing N, while the computational time for $\bar{Q}_2^{\mathrm{L}}$ and $\bar{S}_2$ hardly increased with increasing N. Efficient computation of $\bar{Q}_2^{\mathrm{L}}$ by MALIO's original spherical harmonic function calculation library resulted in saturation of $\bar{Q}_2^{\mathrm{L}}$ computational time with respect to N. The ratio of LOP computational time for ML-LSA to that for MALIO reveals a dramatic speedup in LOP computational time with MALIO. A maximum speedup of over 40 times was achieved in $\bar{A}_2^1$, and a maximum speedup of over 35 times was achieved in $\bar{Q}_2^{\mathrm{L}}$. Even in $\bar{S}_2$, where the computational cost is relatively low, a speedup of over 17 times was achieved.

3.2 Screening of local order parameters

Table 2 shows the (set of) LOP(s) that best distinguished nematic and smectic phases. The second- and third-best (sets of) LOPs are also shown. The $\bar{Q}^{\mathrm{L}}$ series was found to be the best single LOP for all N conditions. In particular, $\bar{Q}_2^{\mathrm{L}}$ attained the best CTR for $N \geq 8$ and was the most robust to N. The best set of two LOPs was also found to be the combination of the $\bar{Q}^{\mathrm{L}}$ series (chosen as the best single LOP) and the Q^{L} series for all N conditions. However, the improvement in c due to adding the Q^{L} series was small (less than 0.01 for almost all N), and the increase in the number of LOPs was not commensurate with the CTR increase. Recall that in additional analyses such as drawing free energy landscapes, increasing the number of LOPs leads directly to an exponential increase in computational cost due to the curse of dimensionality, and makes it less easy to interpret results. A single LOP should be used in the above cases. The second- and third-best LOPs also fall into this case.

The local average was used for almost all of the single LOPs from best to third-best. Therefore, to explore other possibilities for the two LOP sets, we examined the CTRs only for LOPs that were not locally averaged. Table 3 shows the (set of) LOP(s) that best distinguished nematic and smectic phases when the locally averaged LOPs were excluded. At $8 \leq N \leq 11$, the best LOP set was the combination of Q_2^{L} and Q_{12}^{L}. Considering both high CTR and robustness to N, subsequent analyses were performed for $\bar{Q}_6^{\mathrm{L}}$, $\bar{Q}_2^{\mathrm{L}}$, $\bar{Q}_{12}^{\mathrm{L}}$, $\bar{A}_2^1$, $\{Q_2^{\mathrm{L}},Q_{12}^{\mathrm{L}}\}$, Q_2^{L}, and Q_{12}^{L}.

Table 3 Peak performance of each (set of) locally unaveraged LOP(s) in distinguishing between nematic and smectic structures. LOP is an abbreviation for local order parameter and CTR is an abbreviation for correct tagging rate

N [–]	LOP	CTR [—]
6	Q_6^{L}	0.954 ± 0.004
7	Q_6^{L}	0.945 ± 0.004
8	Q_{12}^{L}	0.938 ± 0.003
9	Q_{12}^{L}	0.954 ± 0.004
10	Q_2^{L}	0.965 ± 0.007
11	Q_2^{L}	0.973 ± 0.006
12	Q_2^{L}	0.979 ± 0.005
13	$H_{1.0}$	0.984 ± 0.005
14	$H_{1,0}$	0.989 ± 0.004
6	$\{Q_6^{\mathrm{L}},B_{2,3,\pi/2}\}$	0.958 ± 0.005
7	$\{Q_6^{\mathrm{L}},Q_2^{\mathrm{L}}\}$	0.957 ± 0.005
8	$\{Q_{12}^{\mathrm{L}},Q_2^{\mathrm{L}}\}$	0.965 ± 0.005
9	$\{Q_{12}^{\mathrm{L}},Q_2^{\mathrm{L}}\}$	0.976 ± 0.005
10	$\{Q_2^{\mathrm{L}},Q_{12}^{\mathrm{L}}\}$	0.982 ± 0.004
11	$\{Q_2^{\mathrm{L}},Q_{12}^{\mathrm{L}}\}$	0.987 ± 0.004
12	$\{Q_2^{\mathrm{L}},H_{6.0}\}$	0.990 ± 0.003
13	$\{H_{1.0},H_{6.0}\}$	0.992 ± 0.003
14	$\{H_{1.0},H_{6.0}\}$	0.994 ± 0.003

3.3 Impact of a small number of neighboring particles

As long as the LOP can distinguish local structures in transition with high accuracy, it is desirable to have as few neighboring particles as possible. For a small number of neighboring particles (N = 6 and 7), $\bar{Q}_6^{\mathrm{L}}$ is the best LOP. Therefore, we first observed the cluster statistics when $\bar{Q}_6^{\mathrm{L}}$ was used. Fig. 3 shows (a) the time evolution of the number of smectic molecules in the system and in the maximum cluster, and (b) the ratio of the

Table 2 Peak performance of each (set of) LOP(s) in distinguishing nematic and smectic structures. LOP is an abbreviation for local order parameter and CTR is an abbreviation for correct tagging rate

	LOP (CTR [–])		
N [–]	Best	Second-best	Third-best
6	$\bar{Q}_6^{\mathrm{L}}$ (0.978 ± 0.004)	$\bar{A}_2^1$ (0.955 ± 0.006)	Q_6^{L} (0.954 ± 0.004)
7	$\bar{Q}_6^{\mathrm{L}}$ (0.977 ± 0.004)	$\bar{Q}_2^{\mathrm{L}}$ (0.959 ± 0.008)	$\bar{A}_2^1$ (0.958 ± 0.007)
8	$\bar{Q}_2^{\mathrm{L}}$ (0.976 ± 0.005)	$\bar{Q}_{12}^{\mathrm{L}}$ (0.974 ± 0.004)	$\bar{Q}_6^{\mathrm{L}}$ (0.969 ± 0.004)
9	$\bar{Q}_2^{\mathrm{L}}$ (0.985 ± 0.004)	$\bar{Q}_{12}^{\mathrm{L}}$ (0.982 ± 0.004)	$\bar{F}_{f_3,f_3,6.0}$ (0.973 ± 0.004)
10	$\bar{Q}_2^{\mathrm{L}}$ (0.991 ± 0.003)	$\bar{Q}_{12}^{\mathrm{L}}$ (0.987 ± 0.003)	$\bar{A}_2^1$ (0.978 ± 0.006)
11	$\bar{Q}_2^{\mathrm{L}}$ (0.994 ± 0.002)	$\bar{Q}_{12}^{\mathrm{L}}$ (0.990 ± 0.003)	$\bar{A}_2^1$ (0.982 ± 0.005)
12	$\bar{Q}_2^{\mathrm{L}}$ (0.996 ± 0.001)	$\bar{Q}_{12}^{\mathrm{L}}$ (0.992 ± 0.002)	$\bar{A}_2^1$ (0.984 ± 0.005)
13	$\bar{Q}_2^{\mathrm{L}}$ (0.997 ± 0.001)	$\bar{Q}_{12}^{\mathrm{L}}$ (0.994 ± 0.002)	$\bar{B}_{1,3,\pi/6}$ (0.988 ± 0.004)
14	$\bar{Q}_2^{\mathrm{L}}$ (0.998 ± 0.001)	$\bar{Q}_{12}^{\mathrm{L}}$ (0.995 ± 0.002)	$\bar{B}_{1,3,\pi/6}$ (0.991 ± 0.004)
6	$\{\bar{Q}_6^{\mathrm{L}},Q_6^{\mathrm{L}}\}$ (0.988 ± 0.003)	$\{\bar{A}_2^1,\bar{Q}_5^{\mathrm{L}}\}$ (0.968 ± 0.006)	$\{Q_6^{\mathrm{L}},\bar{Q}_5^{\mathrm{L}}\}$ (0.967 ± 0.005)
7	$\{\bar{Q}_6^{\mathrm{L}},Q_6^{\mathrm{L}}\}$ (0.986 ± 0.003)	$\{\bar{Q}_2^{\mathrm{L}},\bar{Q}_{12}^{\mathrm{L}}\}$ (0.981 ± 0.004)	$\{\bar{A}_2^1,\bar{Q}_{12}^{\mathrm{L}}\}$ (0.976 ± 0.005)
8	$\{\bar{Q}_2^{\mathrm{L}},Q_2^{\mathrm{L}}\}$ (0.989 ± 0.003)	$\{\bar{Q}_{12}^{\mathrm{L}},\bar{Q}_4^{\mathrm{L}}\}$ (0.984 ± 0.004)	$\{\bar{Q}_6^{\mathrm{L}},\bar{B}_{1,1,\pi/4}\}$ (0.981 ± 0.004)
9	$\{\bar{Q}_2^{\mathrm{L}},Q_2^{\mathrm{L}}\}$ (0.994 ± 0.002)	$\{\bar{Q}_{12}^{\mathrm{L}},\bar{Q}_4^{\mathrm{L}}\}$ (0.990 ± 0.003)	$\{\bar{F}_{f_3,f_2,6.0},\bar{S}_2\}$ (0.983 ± 0.003)
10	$\{\bar{Q}_2^{\mathrm{L}},Q_2^{\mathrm{L}}\}$ (0.997 ± 0.001)	$\{\bar{Q}_{12}^{\mathrm{L}},Q_{12}^{\mathrm{L}}\}$ (0.993 ± 0.002)	$\{\bar{A}_2^1,\bar{W}_2^{\mathrm{L}}\}$ (0.985 ± 0.004)
11	$\{\bar{Q}_2^{\mathrm{L}},Q_2^{\mathrm{L}}\}$ (0.998 ± 0.001)	$\{\bar{Q}_{12}^{\mathrm{L}},Q_{12}^{\mathrm{L}}\}$ (0.995 ± 0.002)	$\{\bar{A}_2^1,\bar{W}_2^{\mathrm{L}}\}$ (0.988 ± 0.004)
12	$\{\bar{Q}_2^{\mathrm{L}},Q_2^{\mathrm{L}}\}$ (0.999 ± 0.001)	$\{\bar{Q}_{12}^{\mathrm{L}},Q_{12}^{\mathrm{L}}\}$ (0.997 ± 0.001)	$\{\bar{A}_2^1,\bar{W}_2^{\mathrm{L}}\}$ (0.990 ± 0.004)
13	$\{\bar{Q}_2^{\mathrm{L}},Q_2^{\mathrm{L}}\}$ (1.000 ± 0.000)	$\{\bar{Q}_{12}^{\mathrm{L}},Q_{12}^{\mathrm{L}}\}$ (0.998 ± 0.001)	$\{\bar{B}_{1,3,\pi/6},\bar{W}_2^{\mathrm{L}}\}$ (0.994 ± 0.003)
14	$\{\bar{Q}_2^{\mathrm{L}},Q_2^{\mathrm{L}}\}$ (1.000 ± 0.000)	$\{\bar{Q}_{12}^{\mathrm{L}},Q_{12}^{\mathrm{L}}\}$ (0.998 ± 0.001)	$\{\bar{B}_{1,3,\pi/6},\bar{W}_2^{\mathrm{L}}\}$ (0.996 ± 0.002)

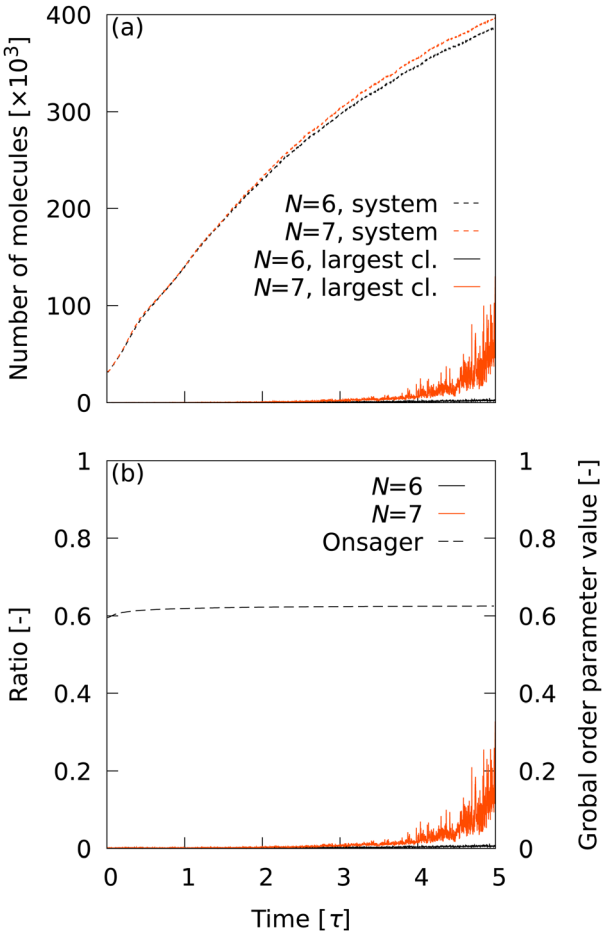

Fig. 3 (a) Time evolution of the number of smectic molecules in the system and in the maximum cluster. (b) Ratio of the number of molecules in the largest cluster to the number of smectic molecules in the system. The time evolution of Onsager's global order parameter was also plotted.

number of molecules in the largest cluster to the number of smectic molecules in the system. The time evolution of Onsager's global order parameter was also plotted in Fig. 3(b). With the CTR of $\bar{Q}_6^{\mathrm{L}}$ for N = 6 and 7 ($c \sim 0.98$), a number of smectic molecules were detected even in the initial structure, which should be completely in the nematic phase. This depends completely on the CTRs in Table 2, implying that extremely high CTRs are required to observe cluster statistics in nucleation. At t = 5.0 τ, the total number of smectic molecules reached about 400 000. However, it is clear from Fig. 3(b) that the largest cluster did not grow at all until t = 5.0 τ. The results indicate that the transition does not progress from nucleation to percolation, even though approximately 40% of the molecules are determined to be smectic molecules. The same trend was observed for other LOP species when N = 6 and 7 were used (data not shown). This unnatural behavior indicates that N is too small and is insufficient to capture the smectic layer. Therefore, in the following analysis, the conditions N = 8–14 were used and $\bar{Q}_6^{\mathrm{L}}$ was excluded. Note that Onsager's global order parameter was also insensitive to increases in the number of smectic molecules.

3.4 Pre-nucleation behavior

The behavior of the number of smectic molecules in the system at the beginning of the transition is important because it affects the subsequent nucleation process. Fig. 4 shows the time evolution of the number of smectic molecules at $0 \leq t \leq 0.2\tau$ in the system observed using (a) $\bar{Q}_2^{\mathrm{L}}$, (b) $\bar{Q}_{12}^{\mathrm{L}}$, (c) $\bar{A}_2^1$, (d) $\{Q_2^{\mathrm{L}}, Q_{12}^{\mathrm{L}}\}$, (e) Q_2^{L}, and (f) Q_{12}^{L}. The number of smectic molecules was divided by the total number of molecules. The initially observed smectic molecules originate from the error in classification using LOPs. With $\bar{Q}_2^{\mathrm{L}}$, the number of initial smectic molecules could be suppressed to less than 0.5% for $N \geq 12$. That the number of molecules hardly increased with time up to $t > 0.1\tau$ independently of N is consistent with the fact that cluster formation is less likely to proceed in the initial nucleation stage.

With $\bar{Q}_{12}^{\mathrm{L}}$, the number of initial smectic molecules could be suppressed to less than 1% for $N \geq 13$, but the observed number of smectic molecules increased immediately afterwards. With $\bar{A}_2^1$, the number of initial smectic molecules could not be suppressed to less than 1% for any of the N conditions, nor could the increase in the number of smectic molecules immediately afterwards be suppressed. With $\{Q_2^{\mathrm{L}}, Q_{12}^{\mathrm{L}}\}$, the number of initial smectic molecules could be suppressed to less than 1% for $N \geq 12$, but the number of smectic molecules increased immediately afterwards. With Q_2^{L}, the number of initial smectic molecules could not be suppressed to less than 1% for any of the N conditions, but its increase with time at $t > 0.1\tau$ was similar to that for $\bar{Q}_2^{\mathrm{L}}$. With Q_{12}^{L}, the number of initial smectic molecules could not be suppressed to less than 1% for any of the N conditions, nor could the increase in number of smectic molecules immediately afterwards be suppressed. The results indicate the superiority of $\bar{Q}_2^{\mathrm{L}}$.

3.5 Percolation transition

The progress from nucleation to percolation is also noteworthy. Fig. 5 shows the time evolution of the ratio of the number of smectic molecules in the largest cluster to the number of smectic molecules in the system, captured by (a) $\bar{Q}_2^{\mathrm{L}}$, (b) $\bar{Q}_{12}^{\mathrm{L}}$, (c) $\bar{A}_2^1$, (d) $\{Q_2^{\mathrm{L}}, Q_{12}^{\mathrm{L}}\}$, (e) Q_2^{L}, and (f) Q_{12}^{L}. The Onsager's order parameter and the all-particle mean of $\bar{Q}_2^{\mathrm{L}}$, $\langle \bar{Q}_2^{\mathrm{L}} \rangle$, were also plotted to show the time evolution of global order parameters. A rapid increase in the ratio signifies the onset of percolation, and the onset of convergence to 1 signifies completion of percolation. With $\bar{Q}_2^{\mathrm{L}}$, the curves overlapped for all N conditions, indicating that the percolation onset and completion times were independent of N. The completion time ($t \sim 1.5\tau$) agreed perfectly with the results of numerical analysis corresponding to X-ray scattering experiments on liquid crystal molecular systems undergoing phase transition. Fig. 6 shows the time evolution of X-ray scattering intensity calculated for the liquid crystal molecular systems during the phase transition. The time evolution of scattering intensity was calculated using the following procedure: (i) the scattering pattern at each time was calculated on the basis of previous studies.[74,75] (ii) Because the intensity difference between the two phases is larger near the second harmonic

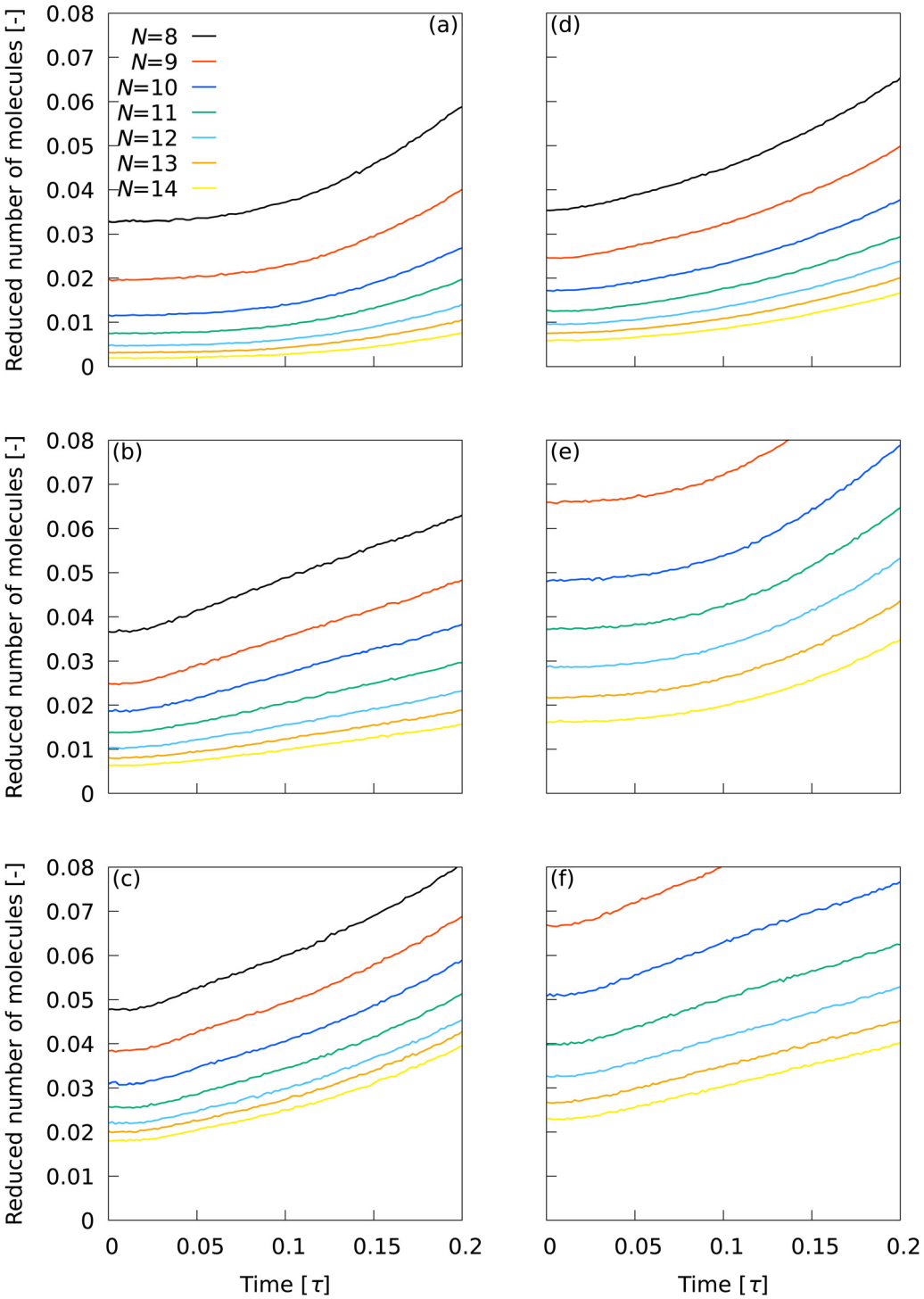

Fig. 4 Time evolution of the number of smectic molecules at $0 \leq t \leq 0.2\tau$ in the system observed using (a) $\bar{Q}_2^{\mathrm{L}}$, (b) $\bar{Q}_{12}^{\mathrm{L}}$, (c) $\bar{A}_2^{\mathrm{T}}$, (d) $\{Q_2^{\mathrm{L}}, Q_{12}^{\mathrm{L}}\}$, (e) Q_2^{L}, and (f) Q_{12}^{L}. The number of smectic molecules was divided by the total number of molecules.

according to previous studies,[32,74,75] the intensity at each time was integrated over a rectangular region cut in the ranges $-3.5\sigma^{-1} \leq q_{\mathrm{a}} \leq 3.5\sigma^{-1}$ and $3.5\sigma^{-1} \leq q_{\mathrm{b}} \leq 6.0\sigma^{-1}$, where q_{a} and q_{b} are the wavenumbers of the components perpendicular and parallel to the liquid crystal orientation direction, respectively. Because the X-ray analysis is performed completely independently from the various analyses using LOPs, it can be used as one indicator to evaluate LOP performance. The X-ray intensity increased almost linearly for $0 \leq t \leq 1.475\tau$. However, for $t > 1.475\tau$, the increasing trend became gradual and intermittent. The results show that the smectic local structures complete percolation at $t = 1.475\tau$, implying the number of smectic molecules in the

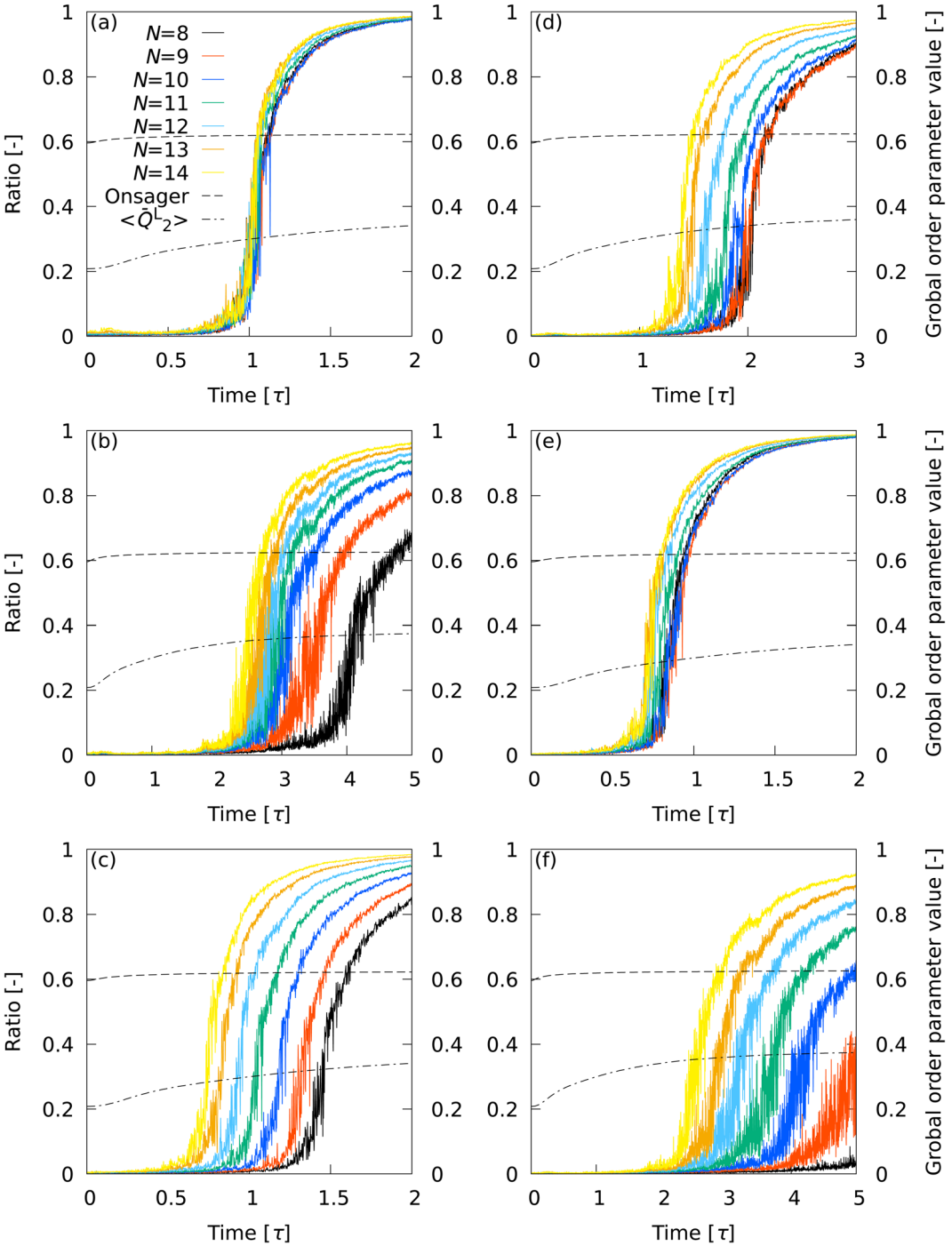

Fig. 5 Time evolution of the ratio of the number of smectic molecules in the largest cluster to the number of smectic molecules in the system, captured by (a) $\bar{Q}_2^{\mathrm{L}}$, (b) $\bar{Q}_{12}^{\mathrm{L}}$, (c) $\bar{A}_2^1$, (d) $\{Q_2^{\mathrm{L}},Q_{12}^{\mathrm{L}}\}$, (e) Q_2^{L}, and (f) Q_{12}^{L}. The Onsager's order parameter and $\langle\bar{Q}_2^{\mathrm{L}}\rangle$ were also plotted to show the time evolution of global order parameters.

largest cluster becomes almost equal to the number of total smectic molecules in the system.

With $\bar{Q}_{12}^{\mathrm{L}}$, the curves for the different N conditions did not overlap each other. The completion time of percolation deviated from the X-ray intensity results for all N conditions, but the two sets of results became closer with increasing N. With $\bar{A}_2^1$, the curves for different N conditions did not overlap each other. Interestingly, curves relatively close to the $\bar{Q}_2^{\mathrm{L}}$ results were obtained at $10 \leq N \leq 12$, where the performance of $\bar{A}_2^1$ was third-best (see Table 2). The curves for different N conditions did not overlap with each other with $\{Q_2^{\mathrm{L}},Q_{12}^{\mathrm{L}}\}$, although the results appeared to be slightly improved over the $\bar{Q}_{12}^{\mathrm{L}}$ results. The percolation completion time tended to be closer to the X-ray intensity results as N was increased. A master curve

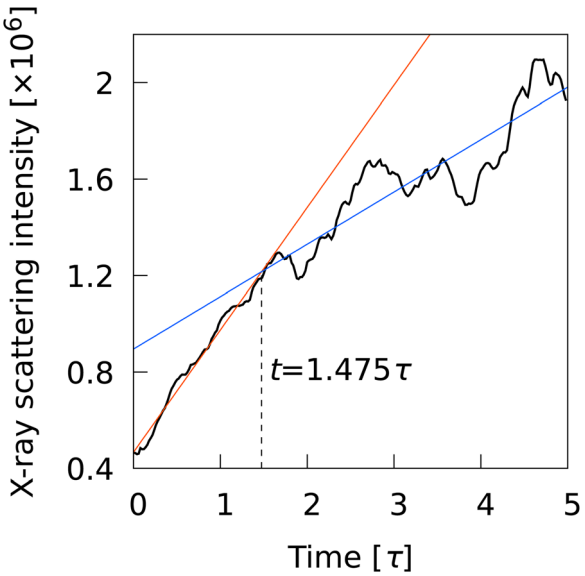

Fig. 6 Time evolution of X-ray scattering intensity calculated for the quenched structures.

similar to the $\bar{Q}_2^L$ results was obtained with Q_2^L, but the overlap between the curves was not as exact as with $\bar{Q}_2^L$. With Q_{12}^L, the results were inferior to the $\bar{Q}_{12}^L$ results.

Overall, the superiority of $\bar{Q}_2^L$ was pronounced in describing percolation transitions. The effectiveness of the LOP local average and the effect of LOP combinations were also evident. The sensitivity of the global order parameters to the percolation transition was extremely low, showing only an almost monotonically increasing trend without any response to the transition time. This fact indicates that it is difficult to detect the details of molecular self-assembly dynamics with global order parameters.

3.6 Cluster statistics from nucleation to percolation

The time evolution of the total number of clusters during nucleation is important, as is the behavior of the clusters before nucleation (the initial stage in Fig. 4) and during the percolation transition (in Fig. 5). Although the number of clusters was affected by the number of neighboring particles, its time evolution was found to have an N-independent master curve for each LOP species. Fig. 7 shows the master curves for the time evolution of the total number of clusters observed using (a) $\bar{Q}_2^L$, (b) $\bar{Q}_{12}^L$, (c) $\bar{A}_2^1$, (d) $\{Q_2^L,Q_{12}^L\}$, (e) Q_2^L, and (f) Q_{12}^L. For the same LOP, when the time at which the number of clusters reaches its maximum is shifted to zero and the number of clusters is normalized using the maximum value, the time evolution of the total number of clusters shows one characteristic curve despite the use of different N values. The results show the nature of each LOP is independent of the number of neighboring particles in the cluster statistics. Although the actual number of clusters varies with N, the time evolution of the cluster number is similar for all N values, demonstrating a strong universality in the way clusters are captured by an LOP.

With $\bar{Q}_2^L$, a master curve was obtained that showed little deviation due to differences in N, except for the initial behavior of the transition. The initial behavior of the cluster number and subsequent systematic convergence to the master curve indicates that the impact of errors in $\bar{Q}_2^L$ on cluster statistics is almost negligible for $N \geq 12$, as implied in Fig. 4.

A discrepancy was observed in $\bar{Q}_{12}^L$ between curves on the long time range due to the N-dependence in the percolation transition. This trend was also observed for $\bar{A}_2^1$ and Q_{12}^L, but appeared to be somewhat improved for $\{Q_2^L,Q_{12}^L\}$. The behavior of Q_2^L was similar to that of $\bar{Q}_2^L$, but the discrepancy due to the difference in N was larger than that of $\bar{Q}_2^L$. These results demonstrate the superiority of $\bar{Q}_2^L$ in describing the nucleation and subsequent percolation transition.

Overall, the outstanding usefulness of $\bar{Q}_2^L$ was demonstrated for all events of the initial stage, nucleation, and percolation transition. The optimal N for $\bar{Q}_2^L$ is $N = 12$, because $N \geq 12$ is desirable for the initial stage and nucleation, and an unnecessarily large N reduces the cluster resolution.

4 Conclusions

In this work, LOPs were selected under an appropriate and efficient screening strategy using MALIO, which enables fast LOP computation and machine learning to evaluate the CTR. MALIO demonstrated superior performance to ML-LSA in LOP computational speed. MALIO achieved a speedup of approximately 17 times even for LOPs with low computational cost, and up to 40 times or more for LOPs with high computational cost. Using the selected LOPs as candidates, the nucleation and subsequent percolation transition were observed to investigate the effect of LOP species and selection of neighboring particles on the cluster statistics. We found the LOP (*i.e.*, $\bar{Q}_2^L$) and appropriate number of neighboring particles ($N = 12$) that accurately describe all events before, during, and after nucleation. The LOP was chosen from among more than 2 million possibilities. The master curve of the time variation of the total number of clusters for each LOP indicates that the LOP is universal in the way it captures the local molecular structures, regardless of the difference in number of neighboring particles. This nature is expected to be confirmed when an LOP is applied to various phase transitions, not only the nematic–smectic transition in this work. Furthermore, the behavior of the percolation transition exhibited by $\bar{Q}_2^L$ clearly indicates that $\bar{Q}_2^L$ is not only an accurate LOP but also a global order parameter in the percolation transition, and is independent of the number of neighboring particles. This nature of $\bar{Q}_2^L$ is expected to be confirmed when exploring the applicability of LOPs to various phase transition phenomena. Interestingly, this nature of $\bar{Q}_2^L$ is already implied by MALIOâ€™s screening results. $\bar{Q}_2^L$ is the LOP with the highest CTR at $8 \leq N \leq 14$, which indicates the accuracy and robustness of $\bar{Q}_2^L$ to a wide range of N. This "signal from machine learning" was substantiated in the actual cluster statistics from nucleation to percolation. This shows that it may be possible to find the

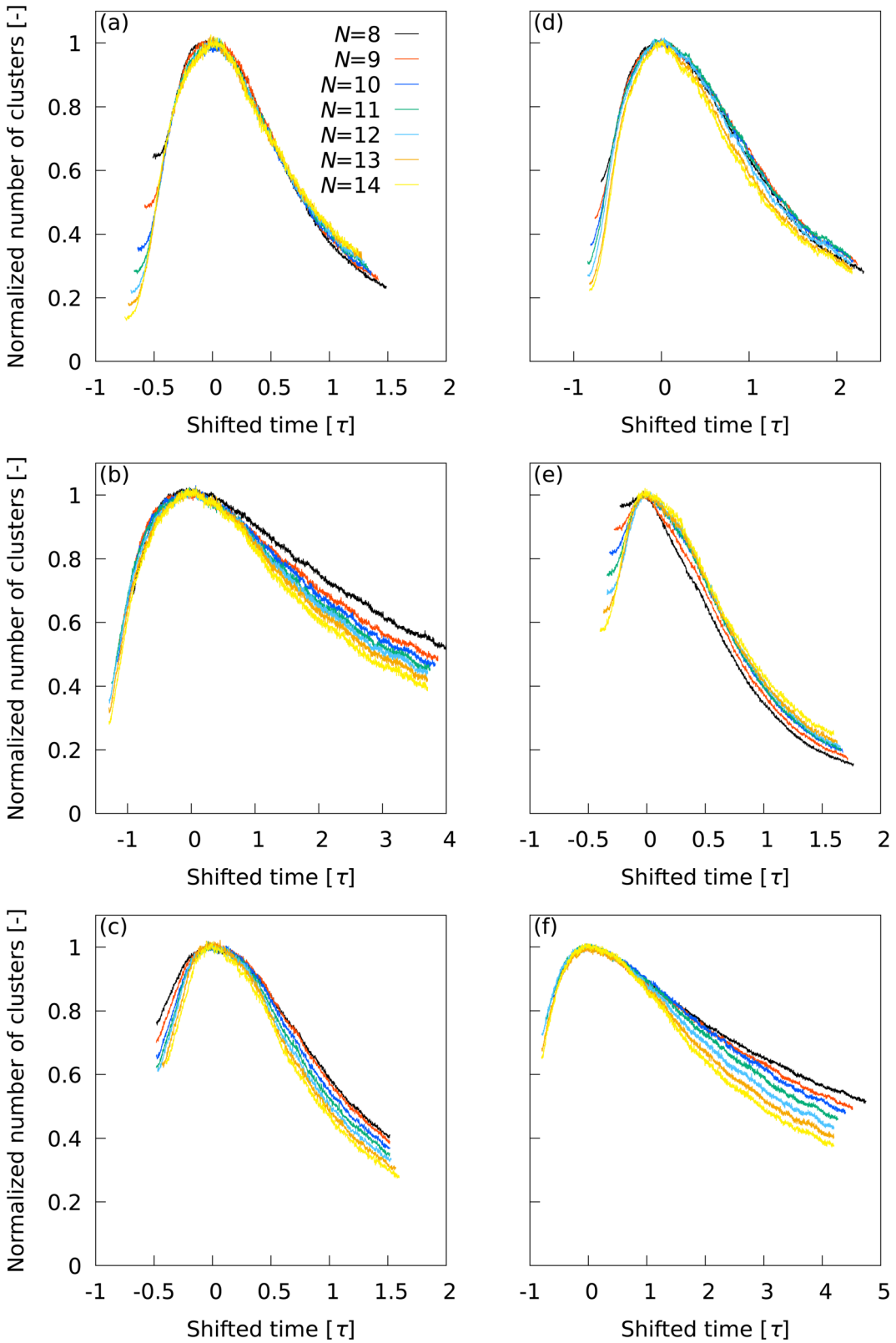

Fig. 7 Master curves for the time evolution of the total number of clusters observed using (a) $\bar{Q}_2^{\mathrm{L}}$, (b) $\bar{Q}_{12}^{\mathrm{L}}$, (c) $\vec{A}_2^1$, (d) $\{Q_2^{\mathrm{L}},Q_{12}^{\mathrm{L}}\}$, (e) Q_2^{L}, and (f) Q_{12}^{L}.

optimal LOP directly from the screening strategy used in this work. Of course, this possibility needs to be demonstrated by applying LOPs to a variety of phase transition phenomena, including confined and multicomponent systems, and is a

topic for future research. In addition, although MALIO's protocols for neighboring particle selection can in principle be modified to suit the crystal and simulation box geometry, it may be necessary to use only certain protocols for some problems.

Overall, these results show that machine learning can powerfully screen a huge number of LOP species and suggest only a few promising candidates. We also established a reliable guideline for selecting the number of neighboring particles for LOPs, which has often been a controversial issue. In the future, selecting appropriate LOPs for various phase transitions will lead to deeper understanding of their phenomena and exploration of their potential applications.[76]

Code availability

MALIO and other codes for analysis are available from the corresponding author upon reasonable request, based on the publication protocol of the developed codes as permitted by a project, JPNP18016, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

Conflicts of interest

The author declares no conflict of interest.

Acknowledgements

The author thanks Dr Ryuji Sakamaki and Dr Osamu Hino of X-Ability Co., Ltd. for their comprehensive support in developing MALIO. This article is based on results obtained from JPNP18016 commissioned by NEDO, and JST, PRESTO Grant Number JPMJPR22O6, Japan.

Notes and references

1 P. M. Chaikin, T. C. Lubensky and T. A. Witten, *Principles of condensed matter physics*, Cambridge university press, Cambridge, 1995, vol. 10.
2 T. E. Strzelecka, M. W. Davidson and R. L. Rill, *Nature*, 1988, **331**, 457–460.
3 J. SantaLucia, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 1460–1465.
4 A. A. Hyman and K. Simons, *Science*, 2012, **337**, 1047–1049.
5 A. A. Hyman, C. A. Weber and F. Jülicher, *Annu. Rev. Cell Dev. Biol.*, 2014, **30**, 39–58.
6 S. L. Ilca, X. Sun, K. El Omari, A. Kotecha, F. de Haas, F. DiMaio, J. M. Grimes, D. I. Stuart, M. M. Poranen and J. T. Huiskonen, *Nature*, 2019, **570**, 252–256.
7 T. B. Saw, W. Xi, B. Ladoux and C. T. Lim, *Adv. Mater.*, 2018, **30**, 1802579.
8 J. Zhao, U. Gulan, T. Horie, N. Ohmura, J. Han, C. Yang, J. Kong, S. Wang and B. B. Xu, *Small*, 2019, **15**, 1900019.
9 J. S. Langer, *Rev. Mod. Phys.*, 1980, **52**, 1.
10 A. Pimpinelli and J. Villain, *Physics of Crystal Growth Authors*, Cambridge University Press, 1998, DOI: **10.1017/CBO9780511622526I**.
11 D. Kashchiev, *Nucleation*, Elsevier, 2000.
12 E. D. Sloan Jr and C. A. Koh, *Clathrate hydrates of natural gases*, CRC press, 2007.
13 D. Erdemir, A. Y. Lee and A. S. Myerson, *Acc. Chem. Res.*, 2009, **42**, 621–629.
14 P. G. Vekilov, *Cryst. Growth Des.*, 2010, **10**, 5007–5019.
15 P. G. Vekilov, *Nanoscale*, 2010, **2**, 2346–2357.
16 W. Qi, Y. Peng, Y. Han, R. K. Bowles and M. Dijkstra, *Phys. Rev. Lett.*, 2015, **115**, 185701.
17 M. Salvalaglio, C. Perego, F. Giberti, M. Mazzotti and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, E6–E14.
18 G. C. Sosso, J. Chen, S. J. Cox, M. Fitzner, P. Pedevilla, A. Zen and A. Michaelides, *Chem. Rev.*, 2016, **116**, 7078–7116.
19 J. F. Lutsko, *Sci. Adv.*, 2019, **5**, eaav7399.
20 D. Kashchiev, *J. Cryst. Growth*, 2020, **530**, 125300.
21 J. De Yoreo, *Crystallization via Nonclassical Pathways: Nucleation, Assembly, Observation & Application*, ACS Publications, vol. 1, 2020, pp. 1–17.
22 S. Auer and D. Frenkel, *Nature*, 2001, **409**, 1020–1023.
23 C. Desgranges and J. Delhommelle, *Phys. Rev. Lett.*, 2007, **98**, 235502.
24 M. R. Walsh, C. A. Koh, E. D. Sloan, A. K. Sum and D. T. Wu, *Science*, 2009, **326**, 1095–1098.
25 R. Demichelis, P. Raiteri, J. D. Gale, D. Quigley and D. Gebauer, *Nat. Commun.*, 2011, **2**, 590.
26 E. Sanz, C. Vega, J. Espinosa, R. Caballero-Bernal, J. Abascal and C. Valeriani, *J. Am. Chem. Soc.*, 2013, **135**, 15008–15017.
27 P. J. Smeets, A. R. Finney, W. J. Habraken, F. Nudelman, H. Friedrich, J. Laven, J. J. De Yoreo, P. M. Rodger and N. A. Sommerdijk, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, E7882–E7890.
28 K. Henzler, E. O. Fetisov, M. Galib, M. D. Baer, B. A. Legg, C. Borca, J. M. Xto, S. Pin, J. L. Fulton and G. K. Schenter, *et al.*, *Sci. Adv.*, 2018, **4**, eaao6283.
29 H. Niu, P. M. Piaggi, M. Invernizzi and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.*, 2018, **115**, 5348–5352.
30 M. Fitzner, G. C. Sosso, S. J. Cox and A. Michaelides, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 2009–2014.
31 L. Li, J. Zhong, Y. Yan, J. Zhang, J. Xu, J. S. Francisco and X. C. Zeng, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**, 24701–24708.
32 K. Z. Takahashi, T. Aoyagi and J.-I. Fukuda, *Nat. Commun.*, 2021, **12**, 5278.
33 K. Yasuoka and M. Matsumoto, *J. Chem. Phys.*, 1998, **109**, 8451–8462.
34 Y. Shibuta, K. Oguchi, T. Takaki and M. Ohno, *Sci. Rep.*, 2015, **5**, 1–9.
35 A. Reinhardt, J. P. Doye, E. G. Noya and C. Vega, *J. Chem. Phys.*, 2012, **137**, 194504.
36 P. J. Steinhardt, D. R. Nelson and M. Ronchetti, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1983, **28**, 784.
37 C. L. Kelchner, S. Plimpton and J. Hamilton, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1998, **58**, 11085.

38 G. Ackland and A. Jones, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2006, **73**, 054104.
39 W. Lechner and C. Dellago, *J. Chem. Phys.*, 2008, **129**, 114707.
40 A. Stukowski, *Modell. Simul. Mater. Sci. Eng.*, 2012, **20**, 045021.
41 A. Radhi and K. Behdinan, *Comput. Mater. Sci.*, 2017, **126**, 182–190.
42 H. Doi, K. Z. Takahashi and T. Aoyagi, *J. Chem. Phys.*, 2020, **152**, 214501.
43 A. P. Bartók, R. Kondor and G. Csányi, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2013, **87**, 184115.
44 A. Seko, A. Togo and I. Tanaka, *Nanoinformatics*, Springer, Singapore, 2018, pp. 3–23.
45 H. Doi, K. Z. Takahashi, K. Tagashira, J.-I. Fukuda and T. Aoyagi, *Sci. Rep.*, 2019, **9**, 16370.
46 H. Doi, K. Z. Takahashi and T. Aoyagi, *J. Chem. Phys.*, 2021, **154**, 164505.
47 H. Doi, K. Z. Takahashi and T. Aoyagi, *J. Comput. Chem.*, 2021, **42**, 1720–1727.
48 H. Doi, K. Z. Takahashi and T. Aoyagi, *J. Phys. Chem. A*, 2021, **125**, 9518–9526.
49 M. Fitzner, P. Pedevilla and A. Michaelides, *Nat. Commun.*, 2020, **11**, 1–9.
50 Z. Wang, F. Wang, Y. Peng and Y. Han, *Nat. Commun.*, 2015, **6**, 6942.
51 W. Mickel, S. C. Kapfer, G. E. Schröder-Turk and K. Mecke, *J. Chem. Phys.*, 2013, **138**, 044501.
52 S. Behnel, R. Bradshaw, C. Citro, L. Dalcin, D. S. Seljebotn and K. Smith, *Comput. Sci. Eng.*, 2011, **13**, 31–39.
53 G. Albertini, M. Corinaldesi, S. Mazkedian, S. Melone, M. Ponzi-Bossi and F. Rustichelli, *Solid State Commun.*, 1977, **24**, 433–437.
54 O. Francescangeli, F. Vita and E. T. Samulski, *Soft Matter*, 2014, **10**, 7685–7691.
55 T. Nozawa, P. E. Brumby, S. Ayuba and K. Yasuoka, *J. Chem. Phys.*, 2019, **150**, 054903.
56 I. Chuang, N. Turok and B. Yurke, *Phys. Rev. Lett.*, 1991, **66**, 2472.
57 I. Chuang, R. Durrer, N. Turok and B. Yurke, *Science*, 1991, **251**, 1336–1342.
58 S. Plimpton, *J. Comput. Phys.*, 1995, **117**, 1–19.
59 R. Berardi, A. P. J. Emerson and C. Zannoni, *J. Chem. Soc., Faraday Trans.*, 1993, **89**, 4069–4078.
60 R. Berardi, J. S. Lintuvuori, M. R. Wilson and C. Zannoni, *J. Chem. Phys.*, 2011, **135**, 134119.
61 R. Berardi, C. Zannoni, J. S. Lintuvuori and M. R. Wilson, *J. Chem. Phys.*, 2009, **131**, 174107.
62 K. Mochizuki, M. Matsumoto and I. Ohmine, *Nature*, 2013, **498**, 350–354.
63 J. D. Honeycutt and H. C. Andersen, *J. Phys. Chem.*, 1987, **91**, 4950–4963.
64 E. Maras, O. Trushin, A. Stukowski, T. Ala-Nissila and H. Jonsson, *Comput. Phys. Commun.*, 2016, **205**, 13–21.
65 P.-L. Chau and A. Hardwick, *Mol. Phys.*, 1998, **93**, 511–518.
66 E. Duboué-Dijon and D. Laage, *J. Phys. Chem. B*, 2015, **119**, 8406–8418.
67 E. B. Moore, E. De La Llave, K. Welke, D. A. Scherlis and V. Molinero, *Phys. Chem. Chem. Phys.*, 2010, **12**, 4124–4134.
68 L. Onsager, *Ann. N. Y. Acad. Sci.*, 1949, **51**, 627–659.
69 W. L. McMillan, *Phys. Rev. A: At., Mol., Opt. Phys.*, 1971, **4**, 1238.
70 B. Boots, K. Sugihara, S. N. Chiu and A. Okabe, *Spatial tessellations: concepts and applications of Voronoi diagrams*, John Wiley & Sons, 2009.
71 W. F. Reinhart and A. Z. Panagiotopoulos, *Soft Matter*, 2018, **14**, 6083–6089.
72 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
73 A. Jain and D. Zongker, *IEEE transactions on pattern analysis and machine intelligence*, 1997, **19**, 153–158.
74 M. A. Bates and G. R. Luckhurst, *J. Chem. Phys.*, 2003, **118**, 6605–6614.
75 G. Skačej and C. Zannoni, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 10193–10198.
76 K. E. Blow, D. Quigley and G. C. Sosso, *J. Chem. Phys.*, 2021, **155**, 040901.

