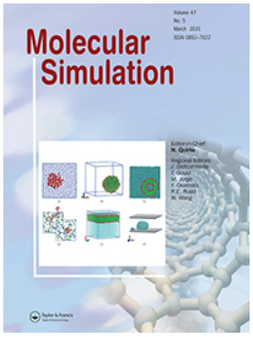

Molecular Simulation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gmos20

On the distance between A and B in molecular configuration space

Sun-Ting Tsai & Pratyush Tiwary

To cite this article: Sun-Ting Tsai & Pratyush Tiwary (2021) On the distance between A and B in molecular configuration space, Molecular Simulation, 47:5, 449-456, DOI: 10.1080/08927022.2020.1761548

To link to this article: https://doi.org/10.1080/08927022.2020.1761548

MOLECULAR SIMULATION
2021, VOL. 47, NO. 5, 449–456
https://doi.org/10.1080/08927022.2020.1761548

Taylor & Francis
Taylor & Francis Group

On the distance between A and B in molecular configuration space

Sun-Ting Tsai[a] and Pratyush Tiwary[b]

[a]Department of Physics and Institute for Physical Science and Technology, University of Maryland, College Park, MD, USA; [b]Department of Chemistry and Biochemistry and Institute for Physical Science and Technology, University of Maryland, College Park, MD, USA

ABSTRACT

In this brief review, we discuss the question of calculating the distance between different molecular configurations. We focus on the theoretical basis of different existing methods for this problem, including ones which are more physics-driven and ones which are more data-driven. We explain the key ideas behind these methods, and conclude with what we see as the central challenges. We hope this review will be helpful to someone aiming to understand these methods, many of which are new and have not been compared and reviewed together. We also hope it will help develop new methods that address the challenges we identify and others.

ARTICLE HISTORY
Received 7 February 2020
Accepted 21 April 2020

KEYWORDS
Kinetics; reaction coordinate; molecular dynamics; sampling; machine learning

1. Introduction

In this review, we discuss the concept of kinetically relevant distances (henceforth, distance in short) between two configurations of a given molecular system. This distance can be formally defined in many ways, but approximately it is a measure that is proportional to the average round-trip interconversion time between two configurations, while following the other mathematical properties expected of a distance measure such as symmetry, non-negativity and, though not always, the triangle inequality. As an analogy consider the distance between any two points on the earth's surface, which is best not measured through an Euclidean measure, but by taking into account the surface of the earth, elevation, other geological and even geopolitical aspects such as borders between nations. In other words, defining a 'kinetically relevant' distance measure is not a simple geometric operation, but one that requires awareness and quantification of numerous factors. Similarly for molecular systems, it is far from trivial to define distance metrics that are true to underlying physical mechanisms at work and capture average interconversion times. Such distance metrics matter in molecular simulations for at least two reasons: making sense of high-dimensional data, and designing a progress coordinate for enhanced sampling in rare event systems. In this brief review, we collect and examine such distance measures that have been proposed over the years for comparing molecular ensembles and structures to each other.

Many of these metrics are rooted in a physical perspective, and these ones are reviewed in Section 2. In Section 3, we review more data-driven and physics-agnostic approaches. Of course, the demarcation between physics-based and physics-agnostic is somewhat arbitrary, and every approach here does require data and is thus data-driven in some sense. As such this demarcation between themes is approximate at best.

The approaches discussed here can be applied to different types of data, and can also be linked with each other. For example, the root-mean squared deviation (RMSD) introduced in Section 2.1 can be computed based on either atomic coordinates or system-specific collective variables (CVs) or order parameters (OPs). The RMSD itself can also be used as the input of diffusion map [1] (Section 2.3), kinetic map and commute map [2] (Section 2.4). The kinetic map and commute map distances are derived with only assuming basic properties of reversible Markov process. Thus they can be used with any abstract data type as long as it follows the desired Markovianity. As we discuss in Section 2.5, the path CV [3] approach is usually not applied to high-dimensional data, a set of pre-selected OPs is needed, and therefore does not directly use atomic coordinates as input. On the other hand, the entropy-based metrics [4,5], which we will discuss in Section 2.6 computes the distance between distributions of atomic coordinates. To distinguish various polymorphs, we may need to include the relative coordinates to compare the distributions as well. As dimensionality reduction methods, sketch-map [6] and AMINO [7] can both be used for high-dimensional data. AMINO is designed for analyzing the information carried by CVs, while sketch-map is applied directly on coordinates in configuration space.

Finally in Section 4, we conclude with the status of the problem and challenges ahead, including how machine learning might help. All through this review we focus on the underlying theoretical constructs and do not discuss any applications of these concepts.

2. Physics based approaches

In this section we cover a range of approaches that are rooted in the key physical aspects at work. As mentioned in the introduction, naturally all of these approaches do require access to data, however they can be distinguished from the approaches of Section 3 in that the key principle at work here is rooted in some sort of physical or mechanistic understanding. We start with

CONTACT Pratyush Tiwary ptiwary@umd.edu

one of the simplest possible metrics, namely the root mean squared deviation (RMSD). We cover many other increasingly sophisticated distance metrics that can be derived solely on the basis of configuration space. We also discuss an example of a metric that is rooted in trajectory space per construction in order to deal with glassy systems.

2.1. Root mean squared deviation (RMSD)

We start this section by considering one of the simplest yet widely used distance measures to quantify how far two molecular configurations are from one other. This is the root-mean squared deviation (RMSD) metric from a given reference state [8,9]. As compared to Euclidean distance, this removes the often irrelevant translational and rotational degrees of freedom. The central idea is that conformations that look very different (i.e. high RMSD) must be reachable only through slow interconversion processes, and correspondingly, structures with low relative RMSD should be easily kinetically accessible from each other. However, this is not always true. The RMSD can be defined over the cartesian coordinates of all atoms in the system in general, or for more system-specific subsets of atoms, such as all C_α atoms in a biomolecule or non-solvent molecules in a chemical reaction, or in the space of some other low-dimensional collective variables (CVs). If too many atoms are included in the RMSD, then many very different high-dimensional configurations could naturally be constructed which have the same (within noise) RMSD from any given reference configuration. In other words, as the number of participating atoms is increased, the RMSD becomes increasingly degenerate. Reducing the number of atoms that go into the RMSD calculation in general makes it less noisy and more reliable. However, selecting a small subset of the atoms also means pre-selection of the crucial physics or chemistry aspects that matter. Take for example the case of ion pair dissociation in water. Here it was shown in Ref. [10] that solvent reorganisation and solvent fluctuations play a critical role in this process even in the simple case of sodium and chloride ions. Considering solvent ions would have made RMSD here effectively meaningless, and excluding them would make it useless. This dilemma is quite typical with the use of RMSD as a kinetic measure, and thus necessitates the development of alternate approaches which we discuss in the remaining parts of this review.

2.2. Isocommittor

The concept of committor, though not the term itself, was possibly first used by Onsager in his study of ion recombination in 1938 [11]. It can be contrasted with RMSD in terms of the isocommittor being far more reliable as a kinetic metric (essentially per construction), yet also being far harder to compute. Given a two-state system, the committor for any point in configuration space is defined as the probability for the trajectory started from that point with randomised velocity to reach one boundary state before it reaches another [12,13]. One can then join configuration points with the same committment probability to construct a locus of isocommittor surfaces, and finally by considering the normal direction to these locii, one gets the committor direction. This is an optimal reaction coordinate for this two state system, and correlates monotonically with kinetic distance from the boundary state used to define it [14]. However, calculation such a committor is computationally expensive as it requires an enormous number of simulations [14]. Furthermore it can be hard to implement or interpret in more than two state systems [15]. However, if one did explicitly calculate the isocommittor, distance along the corresponding isoplanes would serve as an accurate kinetic metric between configurations. Unfortunately, a full such calculation does not only require a long unbiased simulation with back-and-forth between the configurations of interest, which itself is hard to obtain for practical systems, but in addition, it needs extra short simulations from multiple points separating the two configurations. Likelihood maximisation schemes have been proposed to mitigate the computational cost of such procedures [16].

2.3. Diffusion map

When dealing with the evolution of a molecular configuration, one is often primarily interested in the slow degrees of freedom, treating other modes of change as fluctuations which do not carry much information about key mechanisms. It is therefore natural to identify the dissimilarity between two molecular configurations using a distance metric defined along an embedding from such a slow interconverting process. The isocommittor introduced in Section 2.2 is one way of doing this. There are many other linear methods such as principal component analysis [17–20] and classical multidimensional scaling [21] which also endeavour to search for this low-dimensional embedding. By linear we mean that these dimensionality reduction methods project high-dimensional data to the low-dimensional space by linear transformations of the high-dimensional data. However, for most systems it may not be possible to define such embeddings using linear methods as the actual pathway of evolution can be highly nonlinear.

Diffusion map is a dimensionality reduction technique [1,8,22] that allows discovering nonlinear low-dimensional embeddings. A central assumption is that the evolution of molecular configurations can be described as a diffusion process, and thus it is more likely that a configuration evolves to a configuration close as per the associated diffusion distance. With this assumption, diffusion map was designed to follow the underlying geometric structure by transforming the data in the configuration space to an embedding space, such that the Euclidean distance in this embedding space approximates the diffusion distance in the original configuration space. Given two configurations, if there is a pathway along which intermediate configurations exist, the diffusion distance between these two configurations would be smaller than between two configurations without such a pathway, even though the latter might have a smaller Euclidean distance separation in the configuration space. Thus, the diffusion map preserves the local similarity between configurations. We now restate this intuitive picture in a more mathematical framework.

The diffusion map method starts from defining a Gaussian kernel which can be used to measure the local similarity:

$$A_{ij} = \exp\left(-\frac{d(\mathbf{x}_i, \mathbf{x}_j)^2}{2\epsilon}\right) \tag{1}$$

where i and j represent two molecular configurations and $d(i, j)$ can be the Euclidean distance or RMSD (see Section 2.1). The tunable parameter ε provides a local length scale which is sometimes referred to as the 'kernel bandwidth'. If two configurations are further apart than ε, the kernel approaches zero quickly. Therefore, applying the Gaussian kernel has the effect of connecting points lying in the neighbourhood roughly of size ε. Note that the kernel is symmetric and positive semi-definite as it should be for constructing a distance metric.

The next step is to normalise each row of the $\mathbf{A} = [A_{ij}]$ matrix to yield a new $\mathbf{M}$ matrix:

$$M_{ij} = \frac{A_{ij}}{\sum_{j=1}^{N} A_{ij}} \quad i, j = 1, \ldots, N. \tag{2}$$

where $\mathbf{M}$ is a right stochastic Markov transition matrix. By doing so we are saying that the process we are dealing with is a Markov process. The transition matrix $\mathbf{M}$ naturally defines a random walk, which has a spectral decomposition:

$$M_{ij}^{(m)} = \sum_k \lambda_k^m \phi_k(\mathbf{x}_i)\phi_k(\mathbf{x}_j) \tag{3}$$

where $M_{ij}^{(m)} = M_{ij}$ can be interpreted as the transition probability $p^{(m)}(\mathbf{x}_j|\mathbf{x}_i)$ of the system to move from configuration $\mathbf{x}_i$ to $\mathbf{x}_j$ in m time units. $\lambda_0 = 1 \geq \lambda_1 \geq \lambda_2 \geq \cdots$ are the eigenvalues and $\phi_k(\mathbf{x})$ are the corresponding eigenfunctions of $\mathbf{M}$. The squared diffusion distance $D_m^2(\mathbf{x}_1, \mathbf{x}_2)$ is then the Euclidean distance in the space of weighted coordinates $\Phi(\mathbf{x})$:

$$D_m^2(\mathbf{x}_1, \mathbf{x}_2) = \|\Phi_m(\mathbf{x}_1) - \Phi_m(\mathbf{x}_2)\|^2 \tag{4}$$

$$= \sum_{k\geq 0} \lambda_k^{2m}(\phi_k(\mathbf{x}_1) - \phi_k(\mathbf{x}_2))^2 \tag{5}$$

The mapping from configuration space $\mathbf{x}$ to the Euclidean space of diffusion distance $\Phi_m(\mathbf{x}) = (\lambda_0^m\phi_0(\mathbf{x}), \lambda_1^m\phi_1(\mathbf{x}), \ldots)$ is then called the diffusion map. Since the eigenvalues λ_k^m of the Markov transition matrix M_{ij}^m corresponds to the time scale of the diffusion process, it can then perform a dimensionality reduction to the low-dimensional embedding of the slow degree of freedom by truncating at certain λ_k, where the $\lambda_{k+1}, \lambda_{k+2}, \ldots$ are all fast fluctuations that we don't want.

2.4. Kinetic map and commute map

Consider a dynamical system with configuration space Ω and propagation of probability density of states $\rho_t(x)$ defined as follows for $x, y \in \Omega$:

$$\rho_{t+\tau}(\mathbf{y}) = \int_{\mathbf{x}\in\Omega} \rho_t(\mathbf{x})p_\tau(\mathbf{y}|\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{6}$$

$$= \mathcal{P} \circ \rho_t(\mathbf{x}) \tag{7}$$

where $\mathcal{P}$ is the dynamical operator for the Markov process, $p_\tau(\mathbf{y}|\mathbf{x})$ is the transition probability density for finding the system at $\mathbf{y}$ after a time duration τ when starting from $\mathbf{x}$ at time 0. In addition, we assume that the system has a unique equilibrium distribution $\pi(\mathbf{x})$ which can be found by solving

$$\pi(\mathbf{x}) = \mathcal{P} \circ \pi(\mathbf{x}) \tag{8}$$

The squared kinetic distance $D_\tau^2(\mathbf{x}_1, \mathbf{x}_2)$ at a lag time τ between two molecular configurations $\mathbf{x}_1$ and $\mathbf{x}_2$ is then defined as [1]:

$$D_\tau^2(\mathbf{x}_1, \mathbf{x}_2) = \|p_\tau(\mathbf{y}|\mathbf{x}_1) - p_\tau(\mathbf{y}|\mathbf{x}_2)\|_{\pi^{-1}}^2 \tag{9}$$

$$= \int_{\mathbf{y}\in\Omega} \frac{|p_\tau(\mathbf{y}|\mathbf{x}_1) - p_\tau(\mathbf{y}|\mathbf{x}_2)|^2}{\pi(\mathbf{y})}\,\mathrm{d}\mathbf{y} \tag{10}$$

Note that here $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{y}$ are not restricted to be position vectors but can also be other state variables in general. In addition, the transition process $p_\tau(\mathbf{y}|\mathbf{x})$ does not have to be a diffusion process. This expression calculates the distance between two probability distributions evolving from delta functions centred at $\mathbf{x}_1$ and $\mathbf{x}_2$ after a duration τ. If the two configurations $\mathbf{x}_1$ and $\mathbf{x}_2$ are separated by a large barrier which cannot be overcome within this time scale τ, this distance would be large.

The expression in Equation (10) can be calculated when the transition density is directly accessible. In practice, a spectral decomposition technique is used in estimating this kinetic distance for more complex systems where the density might be harder to directly calculate. For a metastable Markov process which satisfies detailed balance $\pi(\mathbf{x})p_\tau(\mathbf{y}|\mathbf{x}) = \pi(\mathbf{y})p_\tau(\mathbf{x}|\mathbf{y})$, the transition density can be approximated by

$$p_\tau(\mathbf{y}|\mathbf{x}) = \sum_{j=1}^{n} \lambda_j(\tau)\psi_j(\mathbf{x})\phi_j(\mathbf{y}) + \mathcal{P}^{\text{fast}}(\tau)\rho_t(\mathbf{x}) \tag{11}$$

where ϕ_i, ψ_i are the eigenfunctions of the propagator $\mathcal{P}$ and the backward propagator $\mathcal{T}$ respectively, related by $\phi_i(\mathbf{x}) = \pi(\mathbf{x})\psi_i(\mathbf{x})$. The eigenvalues are sorted in non-increasing norm:

$$\lambda_0 = 1 \geq |\lambda_1| \geq \cdots |\lambda_n| \tag{12}$$

Suppose we operate at a lag time τ such that $|\lambda_{n+1}(\tau)| \approx 0$ or $\mathcal{P}^{\text{fast}}(\tau) \circ \rho_t(\mathbf{x}) \approx 0$ everywhere, then $p_\tau(\mathbf{y}|\mathbf{x}) \approx \sum_{j=0}^{n} \lambda_j(\tau)\psi_j(\mathbf{x})\phi_j(\mathbf{y})$. Plugging this expression in Equation (10), we obtain a form which can be used to calculate the kinetic distance

$$D_\tau^2(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{n} [\lambda_j\psi_j(\mathbf{x}_1) - \lambda_j\psi_j(\mathbf{x}_2)]^2 \tag{13}$$

Although this expression is similar to the diffusion distance of Equation (5) introduced in the diffusion map, it can be used for any reversible Markov dynamics. The weighted coordinates $\tilde{\psi}_j(\mathbf{x}) = \lambda_j\psi_j(\mathbf{x})$ are then defined as the kinetic map:

$$\tilde{\Psi} = (\tilde{\psi}_0, \tilde{\psi}_1, \ldots \tilde{\psi}_n) \tag{14}$$

Note that when the transition density $p_\tau(\mathbf{y}|\mathbf{x})$ is given by a diffusion process, this $\tilde{\Psi}(\mathbf{x})$ is a diffusion map.

The above expression, however, depends on the lag time parameter τ and does not provide a clear physical interpretation. In order to avoid the dependency, the squared commute distance is defined as the integrated version of the kinetic

distance [2]:

$$d^2_{\text{comm}} = \int_0^\infty D^2_\tau(\mathbf{x}_1, \mathbf{x}_2)\, \mathrm{d}\tau \tag{15}$$

$$= \sum_{j=1}^{n} [\psi_j(\mathbf{x}_1) - \psi_j(\mathbf{x}_2)]^2 \int_0^\infty \lambda_j^2(\tau)\, \mathrm{d}\tau \tag{16}$$

$$= \sum_{j=1}^{n} \left[\sqrt{\frac{t_j}{2}} \psi_j(\mathbf{x}_1) - \sqrt{\frac{t_j}{2}} \psi_j(\mathbf{x}_2) \right]^2 \tag{17}$$

where t_j is the relaxation time scale of the jth process. The commute distance is equal to the Euclidean distance in the space of $\tilde{\psi}_j = \sqrt{t_j/2}\psi_j(\mathbf{x})$. This distance metric is then called the commute map. As demonstrated in a simple Markov process [2], this squared commute distance $d^2_{\text{comm}}(\mathbf{x}_1, \mathbf{x}_2)$ between two molecular configurations $\mathbf{x}_1$ and $\mathbf{x}_2$ is equivalent to the averaged commute time between them:

$$t_{\text{comm}} = \frac{t_{12} + t_{21}}{2} \tag{18}$$

where t_{ij} is the mean first passage time from $\mathbf{x}_i$ to $\mathbf{x}_j$.

2.5. Path collective variable

The path CV formalism was introduced in Ref. [3] in the context of enhanced sampling of fluctuations along and orthogonal to a given trial path connecting any two molecular configurations A and B. The first step is to define milestones along a given reference path taking the system from A to B. A variable s, introduced in Ref. [3] in a continuous (i.e. infinitely many milestones) and discrete (finite number of milestones, which is usually the case) quantifies the progress along these milestones and thus along the chosen starting path. These milestones themselves are typically not defined in high-dimensional configuration space but in the space of pre-selected order parameters. For example, this could be in the RMSD space or in the contact map space[23,24]. The value of s then is a natural metric to calculate the distance along the chosen path between any two configurations.

Depending on whether this path is indeed the preferred path of least resistance [25], this distance might or might not be the kinetic distance we seek in this review. In order to discover other possible pathways and quantify distance along them, Ref. [3] introduces a second z variable which is orthogonal to s and measures distance *from* the starting path. Together s and z could be biased in enhanced sampling simulations such as metadynamics to discover new pathways and distances along them[3].

2.6. Entropy-based distance metric

Often one is interested in characterising molecular configurations resulting from the process of phase transitions. While the concepts of this subsection apply to generic phase transitions, here we consider for instance crystallisation, which is a process going spontaneously from disordered to ordered geometric structures. A natural way to characterise such transitions is thus quantifying the entropy, which is qualitatively related to the extent of ordering. It is therefore intuitive to use entropy to quantify how and when a system crystallises. In addition, using entropy to detect the crystallisation in the atomistic simulations has further advantages. Since entropy does not prejudge the system's geometric structure, it can be used to describe a transition in which multiple crystal phases are involved or classify previously unknown ordered phases.

Unfortunately, there is no exact expression for entropy which one could calculate with computational ease during a molecular simulation, to use either as a biasing variable in the context of enhanced sampling or simply as a distance coordinate quantifying the extent of crystallisation. Piaggi et. al. have taken the idea from liquid state theory in Ref. [4], in which the excess entropy per atom is expanded in an infinite series of multiparticle correlation functions [26]. For computational ease, they proposed the use of just the first term in the expansion which only includes the two-body correlation function:

$$S_2 = -2\pi\rho k_B \int_0^\infty [g(r) \ln g(r) - g(r) + 1] r^2\, \mathrm{d}r \tag{19}$$

where $g(r)$ is the radial distribution function and ρ is the density of the system. As mentioned in [4], even though entropy can only have proper thermodynamic meaning when being averaged over an ensemble of states, the use of the instantaneous value as per Equation (19) is proposed. In practice, the radial distribution function is replaced by a modified version:

$$g_m(r) = \frac{1}{4\pi N \rho r^2} \sum_{i \neq j} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(r - r_{ij})^2/(2\sigma^2)} \tag{20}$$

where r_{ij} is the distance between particles i and j, and σ is a broadening parameter. The modification is used in order to ensure the calculation of the derivative of CVs.

Theoretically, it has also been shown there is an interplay between entropy and enthalpy during first-order phase transitions including crystallisation [4]. To deal with this, the authors in Ref. [4] while studying crystallisation of Na and Al, included the enthalpic CV s_H:

$$s_H = \frac{U(\mathbf{R}) + PV}{N} \tag{21}$$

where P is the pressure, V is the volume, $U(\mathbf{R})$ is the potential energy and N is the number of atoms in the system. It was found that while studying the crystallisation of Al, both entropic and enthalpic CVs were needed to distinguish between the possible BCC and FCC crystal phases; while for Na, only using entropic CVs was sufficient, as the only crystalline phase was BCC which entropy could be used to distinguish from the liquid phase.

In subsequent work [27], this method was also extended for dealing with molecular crystals which have significant complexity due to the presence of a large number of polymorphs. A common reason polymorphs can arise is due to various competing orientations of constituent molecules. In order to include the relative orientations between two molecules, one can represent each molecule by the position of its centre of mass and define the orientation with a vector $\mathbf{v}_i$ associated to

the ith molecule. The radial distribution function $g(r, \theta)$ is then different from the previous $g(r)$ in Equation (19), as it involves an angle θ between two molecules, which is defined as $\theta = \arccos\left(\frac{\mathbf{v}_i \cdot \mathbf{v}_j}{|\mathbf{v}_i||\mathbf{v}_j|}\right)$. The entropic CV s itself then, in contrast to the one in Equation (19) is also modified to include this new variable θ:

$$S_\theta = -\pi\rho k_B \int_0^\infty \int_0^\pi \left[g(r, \theta)\ln g(r, \theta) - g(r, \theta) + 1\right] r^2 \sin\theta \, \mathrm{d}r \, \mathrm{d}\theta \tag{22}$$

As mentioned in Ref. [27], instead of considering all Euler angles ϕ, θ, ψ, which makes the calculation of $g(r, \phi, \theta, \psi)$ cumbersome, a single angle θ has been used. This makes the choice of θ crucial when calculating S_θ during a simulation.

Equation (22) can be used to perform enhanced sampling along the entropy coordinate, but is still not sufficient to identify and classify different polymorphs. By observing that Equation (22) can be viewed as the distance between the $g(r, \theta)$ of the present configuration and the $g(r, \theta)$ of an ideal gas; i.e. $g(r, \theta) = 1 \; \forall r, \theta$, [27] defines a new quantity which is similar to Kullback-Leibler (KL) divergence in form:

$$D(g_1 \| g_2) = \int_0^\infty \int_0^\pi \left[g_1(r, \theta) \ln \frac{g_1(r, \theta)}{g_2(r, \theta)} - g_1(r, \theta) + g_2(r, \theta) \right] \times r^2 \sin\theta \, \mathrm{d}r \, \mathrm{d}\theta \tag{23}$$

This quantity is convex and has a minimum when $g_1 = g_2$. To make it a well-defined distance metric for the use, we can use a symmetric form:

$$d(g_1, g_2) = \frac{D(g_1 \| g_2) + D(g_2 \| g_1)}{2} \tag{24}$$

Note that Equations (19) and (21) are used for only crystallisation of simple atomic systems, while Equations (22)–(24) includes the quantities to identify detailed molecular orientation θ.

So far in this sub-section we have described using entropic CVs to define distance metric in order to classify different polymorphs emerging in crystal nucleation. The entropy was approximated by computing the excess entropy per atom using the radial distribution function. A very different approach has been proposed [5] which uses neural networks to directly learn the entropy as a function of temperature and density. A key idea in this approach is to calculate the entropy through the Helmholtz free energy:

$$S_i = \frac{U_i + A^{\mathrm{ML}}(\rho, T)}{T} \tag{25}$$

where S_i is defined as the entropy for each configuration i generated during the (N, P, T) simulation, U_i and V_i are the internal energy and volume of the system, and A^{ML} is the ML prediction of the Helmholtz free energy at the density $\rho = \frac{N}{V_i}$ and temperature T. The internal energy is calculated by:

$$U_i = U_i^{pot} + \frac{3}{2} N k_B T \tag{26}$$

where U_i^{pot} is contributed to by the internal potential energy and internal kinetic energy. In Ref. [5], the Helmholtz free energy was trained by the data points generated from the Johnson, Zollweg, and Gubbins equation of state for the liquid phase [28] and from the van der Hoef equation of state for the face-centred cubic crystal [29]. The authors also used entropy calculated from ML as the reaction coordinate to drive MD simulation of a Lennard-Jones system with the umbrella sampling method [5]. The nucleation process is also shown to be consistent with the previous study using the ML approach.

2.7. s-ensemble and ergodicity based distance metrics

In the previous section, we considered the case of phase transitions such as crystallisation where entropy can be used to define useful distance metrics quantifying the progress of the transition as well as the difference between different competing polymorphs. An even more complicated transformation is that from liquid phase to glass phase [30,31]. While it is reported to be a first-order transition [32], the glassy phase is inherently ill-defined in configuration space due to the presence of dynamical heterogeneities [33], and a more natural way to view this transition is as one occurring in trajectory space, as suggested for example by Chandler [32,34] and others [35,36]. The essential idea is that onset of glassiness corresponds to a loss in ergodicity. To capture this transformation in ergodicity, one introduces a dynamical field s, which drives the system from active or ergodic liquid state to inactive or non-ergodic glass state. In order to study this transition which is driven out of equilibrium, Lester, et. al. [32] applied the transition path sampling method [37] with a perturbed path ensemble distribution:

$$P_s[x(t)] = P_0[x(t)] \exp\{-sK[x(t)]\} \tag{27}$$

where $x(t)$ represents the position of the system in the configuration space, $P_0[x(t)]$ is the equilibrium probability distribution, $K[x(t)]$ is the order parameter which is defined in order to discriminate the system between liquid and glassy states. In the ordinary transition path sampling method, the distribution $P_0[x(t)]$ is sampled by performing a random selection of n independent trajectories each of length t_{obs} in the trajectory space. One can also perform the selection based on the Metropolis rule for the perturbed distribution $P_s[x(t)]$, which means that we accept or reject the selections so as to preserve the weight $P_s[x(t)]$. This way to perform the transition path sampling is called 's-ensemble'. In [32], $K[x(t)]$ is defined as a measure of mean dynamical activity, calculated as the following functional of trajectory $x(t)$ over period $(0, t_{\mathrm{obs}})$:

$$K[x(t)] = \Delta t \sum_{t=0}^{t_{\mathrm{obs}}} \sum_{j=1}^{N} |\mathbf{r}(t + \Delta t) - \mathbf{r}(t)|^2 \tag{28}$$

where N is the number of particles in the system, $\mathbf{r}_j$ refers to position as function of time t, Δt is the time increment such that the number of points summing over time is equal to $t_{\mathrm{obs}}/\Delta t$. The order parameter is integrated over the observation time t_{obs} and therefore depends on the system's history. Note that the $K[x(t)]$ is not normalised with respect to the system size and observation period and so is an extensive quantity,

When the system is in the liquid phase, the particles are mobile and $K[x(t)]$ is large; on the other hand, when the system is in glass phase, the particles are immobile and $K[x(t)]$ is small. It has been shown in the finite-size atomistic systems, $\langle K[x(t)]\rangle_s$ (where subscript denotes sampling using $P_s[x(t)]$) changes abruptly at the critical field $s = s^*$, which is a feature of first-order phase transitions. Just like the density can be used to distinguish liquid and gas, the order parameter $K[x(t)]$ serves as a quantity to distinguish supercooled liquids and glass. As previously mentioned, the difference between liquids and glass cannot be seen readily in the state space since they have similar local structure [38], and that a more prominent difference is that in the ergodicity. The quantity $K[x(t)]$, which involves the squared displacement $|\mathbf{r}(t+\Delta t) - \mathbf{r}(t)|^2$, measures the local dynamical activity and thus corresponds to an effective ergodicity measure. It can also be viewed as the relative distance away from equilibrium in the trajectory space. The order parameter $K[x(t)]$ is also used to detect the dynamical coexistence as it quantifies the order-disorder transition in space time. [34,39] In practice, one weights the trajectories by the factor e^{sK} to favour the nonergodic or immobile state. If then the transition to nonergodic phase happens for nonzero s, there is a dynamical phase coexistence exists at s=0.

3. Data driven and other physics-agnostic approaches

In this section we cover a few approaches that are not inherently based in the physics of the problem (for instance, existence of few slow modes or onset of lack of ergodicity). Like in Section 2 it is inevitable that we cannot cover the huge number of approaches that exist in this domain, and instead we restrict our attention to two emblematic approaches, one recent and one not-so-recent. We also want to highlight that the diffusion map approach of Section 2.3 could very well be considered as a more data-driven than physics-driven approach and be included in this section.

3.1. Sketch-map

Many nonlinear dimensionality reduction algorithms assume that the configuration phase can be mapped to a lower-dimensional space which locally resembles Euclidean space. These nonlinear manifold learning methods include the local linear embedding [40], Isomap [41] and Diffusion map in Section 2.3. However, it has been shown that this assumption can be invalid for data taken from typical atomistic simulations [6]. There is also evidence that the potential energy surfaces of protein can be in a fractal dimension or can have an intrinsically non-Euclidean topology [6,19,42]. A dimensionality reduction method called sketch-map was proposed by Ceriotti et al which rectifies such problems [6]. It is a nonlinear multidimensional scaling method which is able to find a distance metric which preserves the connectivity within and between clusters of data points in the higher-dimensional space.

Sketch-map takes sigmoid transforms of both the high- and low-dimension representations and minimises a stress function between them defined as follows:

$$\chi^2 = \frac{\sum_{j\neq i} w_i w_j \left[F(R_{ij}) - f(r_{ij})\right]^2}{\sum_{j\neq i} w_i w_j} \tag{29}$$

where w_i is the weight of point i and $R_{ij} = |X_i - X_j|_{(D)}$ and $r_{ij} = |x_i - x_j|_{(d)}$ is some generic distance between points i and j in the high- and low-dimensional spaces. The subscripts D and d represent the dimensions of corresponding high- and low-dimensional spaces. F and f are the general sigmoid functions of the form:

$$s_{\sigma,a,b}(r) = 1 - \left[1 + \left(2^{a/b} - 1\right)\left(\frac{r}{\sigma}\right)^a\right]^{-b/a} \tag{30}$$

where $s_{\sigma,a,b}(\sigma) = 1/2$, and the exponent a and b determine the rate at which the function approaches 0 and 1. The functions F and f then become transformations which transform R_{ij} and r_{ij} to values between 0 and 1. As a result, $F(R_{ij}) - f(r_{ij})$ is small when the length scale of R_{ij} and r_{ij} are *both* larger or smaller than σ. In other words, it creates a mapping which preserves the spatial relation between the connections in the low and high-dimensional spaces. In practice, the same value of σ is used in both F and f. Different values of σ simply corresponds to a scaling of coordinates. The value of a and b, however, are chosen differently as a_D and b_D for F and a_d and b_d for f. It has been shown that precise choice of a and b has little effect on the performance of this method.

The computational cost for evaluating Equation (29) grows quadratically with the data points involved [6,43,44]. If one uses the 'point-wise global' optimisation strategy, as the authors used in their first paper [6], the cost could scale even cubically. Therefore, a set of landmark points are often chosen to represent the original trajectory in order to avoid the efficiency problem. The landmark points are either picked up randomly or by using the farthest point sampling strategy (FPS) as mentioned in Ref. [6]. The weights for the landmarks are evaluated by the estimate of free energy, thereby necessitating accurate sampling of the free energy landscape before sketch-map can be performed. After completing the minimisation, one can obtain the low-dimensional representation x through minimising the following function:

$$\delta^2(x) = \frac{\sum_{i=1}^{N} w_i \left[F(|X - X_i|_{(D)}) - f(|x - x_i|_{(d)})\right]^2}{\sum_{i=1}^{N} w_i} \tag{31}$$

where N is the number of landmarks, $|X - X_i|_{(D)}$ is the distance between ith landmark points and the frame that is projected and $|x - x_i|_{(d)}$ is the projected Euclidean distance in the low-dimensional embedding. As stated above, while sketch-map provides a way to express complex high-dimensional clusters and their corresponding connectivity with a low-dimensional representation, it suffers from the efficiency problem and has to choose landmarks as the representative subset. Recently, Lemke et. al. [45] have combined the autoencoders, a popular machine learning technique, with sketch-map to provide another way for efficiently performing the sketch-map algorithm with all data points instead of landmarks, with still acceptable efficiency. They called their method 'Encodermap', which

uses the sketch-map stress function as the cost function in the autoencoder neural network.

3.2. AMINO

The very last approach we discuss in this review is called Automatic mutual information noise omission (AMINO) [7] that constructs distance not between molecular configurations, but instead between the collective variables or order parameters one could use to describe these configurations. AMINO is a data-driven method that allows screening out redundancies from such a large dictionary of CVs, in terms of the physics they represent. AMINO involves taking a biased or unbiased molecular simulation trajectory to calculate the mutual information $I(X, Y)$ between any two order parameters X and Y . The central idea is that if two order parameters or CVs carry completely different or independent information, then $I(X, Y) = 0$. Further, increasingly higher non-zero values of $I(X, Y)$ correspond with X and Y becoming increasingly correlated or dependent. Inspired by this. we define a normalised distance metric $D(X, Y)$ between two OPs as follows:

$$D(X, Y) = 1 - \frac{I(X, Y)}{H(X, Y)} \tag{32}$$

where $H(X, Y)$ is the joint entropy of X and Y. This quantity $D(X, Y)$ then defines the pairwise distances in an abstract information-theoretic space. In this abstract space, order parameters which carry similar amount of information will be close to each other. As a result, we can expect multiple clusters in this abstract space. The goal then becomes defining good representations of such clusters. In [7], the authors used a quantity called distortion jump which allows them to identify the appropriate number of clusters. After that, they could pick the centres of these clusters as the approximated representations. The smaller set of CVs or order parameters returned by AMINO are then directly beneficial to nearly all the methods described in this review in terms of significantly reducing their computational workload.

4. Conclusions and challenges ahead

In this brief review we have surveyed some of the many measures available to determine the distance between two molecular configurations. Such a measure is useful both from the purpose of making sense of high-dimensional and often hard-to-interpret data, and also for the purpose of being used as a progress coordinate for driving enhanced sampling simulations of rare events. Such a kinetic distance could also be useful in the design of coarse-grained models which are accurate in context of dynamics. We do not make any claims on the completeness of this work as this continues to be a field of active research interest. We conclude this review by what we deem to be some of the key challenges ahead. First, it is our view that most of these metrics are still suited to the analysis of trajectories where different regions of the configuration or phase space have been sampled exhaustively. While this is of course useful, the problem of constructing kinetic distances in undersampled situations and from biased simulations is more or less a very open problem. One approach here could be to mix commute maps with ideas such as the SGOOP approach which allow extracting timescale-like quantities from biased simulations [46]. Another challenge is if recent machine learning developments especially in the context of recurrent neural networks, such as Long Short-Term Memory (LSTM) networks [47], reservoir computing [48] or others [49] can be helpful in the endeavour of recovering kinetic distance from data. For example, reservoir computing has been found useful for model-free prediction of spatio-temporal evolution in chaotic systems, which is closely related to the problem discussed here. As mentioned in the introduction, choosing an appropriate set of pre-selected CVs is important for some of the methods discussed in this review, and thus care needs to be exercised in deciding the space in which the distance metric is computed. Finally, many practical problems will have different physics at work in different parts of configuration space or at different moments in time -- crystal nucleation being a classic example where intermediate pathways could involve amorphous glassy-like states [50]. Similar concerns may apply to protein-ligand dissociation and protein conformational change problems [51,52]. Developing kinetic measures for such systems, so that they can be analyzed or their sampling enhanced in an automated manner, will require mixing many of the approaches discussed here as well as developing new ones.

Acknowledgments

Acknowledgment is made to the Donors of the American Chemical Society Petroleum Research Fund for partial support of this research (PRF 60512-DNI6). We thank Zachary Smith for proofreading this manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Acknowledgment is made to the Donors of the American Chemical Society Petroleum Research Fund for partial support of this research (PRF 60512-DNI6).

References

[1] Coifman RR, Lafon S, Lee AB, et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc Natl Acad Sci. 2005;102(21):7426–7431.
[2] Noe F, Banisch R, Clementi C. Commute maps: separating slowly mixing molecular configurations for kinetic modeling. J Chem Theory Comput. 2016;12(11):5620–5630.
[3] Branduardi D, Gervasio FL, Parrinello M. From a to b in free energy space. J Chem Phys. 2007;126(5):054103.
[4] Piaggi PM, Valsson O, Parrinello M. Enhancing entropy and enthalpy fluctuations to drive crystallization in atomistic simulations. Phys Rev Lett. 2017;119(1):015701.
[5] Desgranges C, Delhommelle J. Crystal nucleation along an entropic pathway: teaching liquids how to transition. Phys Rev E. 2018;98(6):063307.
[6] Ceriotti M, Tribello GA, Parrinello M. Simplifying the representation of complex free-energy landscapes using sketch-map. Proc Natl Acad Sci. 2011;108(32):13023–13028.

[7] Ravindra P, Smith Z, Tiwary P. Automatic mutual information noise omission (amino): generating order parameters for molecular systems. Mol Syst Des Eng. 2020;5(1):339–348.
[8] Rohrdanz MA, Zheng W, Maggioni M, et al. Determination of reaction coordinates via locally scaled diffusion map. J Chem Phys. 2011;134(12):03B624.
[9] Cohen FE, Sternberg MJE. On the prediction of protein structure: the significance of the root-mean-square deviation. J Mol Biol. 1980;138(2):321–333.
[10] Geissler PL, Dellago C, Chandler D. Kinetic pathways of ion pair dissociation in water. J Phys Chem B. 1999;103(18):3706–3710.
[11] Onsager L. Initial recombination of ions. Phys Rev. 1938;54(8):554.
[12] Weinan E, Ren W, Vanden-Eijnden E. Transition pathways in complex systems: reaction coordinates, isocommittor surfaces, and transition tubes. Chem Phys Lett. 2005;413(1-3):242–247.
[13] Berezhkovskii AM, Szabo A. Committors, first-passage times, fluxes, Markov states, milestones, and all that. J Chem Phys. 2019;150(5):054106.
[14] Krivov SV. On reaction coordinate optimality. J Chem Theory Comput. 2013;9(1):135–146.
[15] Kells A, Mihálka ZÉ, Annibale A, et al. Mean first passage times in variational coarse graining using Markov state models. J Chem Phys. 2019;150(13):134107.
[16] Peters B, Trout BL. Obtaining reaction coordinates by likelihood maximization. J Chem Phys. 2006;125(5):054108.
[17] García AE. Large-amplitude nonlinear motions in proteins. Phys Rev Lett. 1992;68(17):2696.
[18] Amadei A, Linssen ABM, Berendsen HJC. Essential dynamics of proteins. Proteins. 1993;17(4):412–425.
[19] Hegger R, Altis A, Nguyen PH, et al. How complex is the dynamics of peptide folding? Phys Rev Lett. 2007;98(2):028102.
[20] Zhuravlev PI, Materese CK, Papoian GA. Deconstructing the native state: energy landscapes, function, and dynamics of globular proteins. J Phys Chem B. 2009;113(26):8800–8812.
[21] Cox MAA, Cox TF. Multidimensional scaling. In: Handbook of data visualization. Berlin: Springer; 2008. p. 315–347.
[22] Ferguson AL, Panagiotopoulos AZ, Debenedetti PG, et al. Systematic determination of order parameters for chain dynamics using diffusion maps. Proc Natl Acad Sci. 2010;107(31):13597–13602.
[23] Limongelli V, Bonomi M, Marinelli L, et al. Molecular basis of cyclooxygenase enzymes (coxs) selective inhibition. Proc Natl Acad Sci. 2010;107(12):5411–5416.
[24] Tiwary P, Limongelli V, Salvalaglio M, et al. Kinetics of protein–ligand unbinding: predicting pathways, rates, and rate-limiting steps. Proc Natl Acad Sci. 2015;112(5):E386–E391.
[25] Tiwary P, Berne BJ. Predicting reaction coordinates in energy landscapes with diffusion anisotropy. J Chem Phys. 2017;147(15):152701.
[26] Nettleton RE, Green MS. Expression in terms of molecular distribution functions for the entropy density in an infinite system. J Chem Phys. 1958;29(6):1365–1370.
[27] Piaggi PM, Parrinello M. Predicting polymorphism in molecular crystals using orientational entropy. Proc Natl Acad Sci. 2018;115(41):10251–10256.
[28] Karl Johnson J, Zollweg JA, Gubbins KE. The lennard-jones equation of state revisited. Mol Phys. 1993;78(3):591–618.
[29] Van der Hoef MA. Free energy of the lennard-jones solid. J Chem Phys. 2000;113(18):8142–8148.
[30] Starr FW, Sastry S, Douglas JF, et al. What do we learn from the local geometry of glass-forming liquids? Phys Rev Lett. 2002;89(12):125501.
[31] Charbonneau P, Kurchan J, Parisi G, et al. Fractal free energy landscapes in structural glasses. Nat Commun. 2014;5(1):1–6.
[32] Hedges LO, Jack RL, Garrahan JP, et al. Dynamic order-disorder in atomistic models of structural glass formers. Science. 2009;323(5919):1309–1313.
[33] Kob W, Donati C, Plimpton SJ, et al. Dynamical heterogeneities in a supercooled lennard-jones liquid. Phys Rev Lett. 1997;79(15):2827.
[34] Merolle M, Garrahan JP, Chandler D. Space–time thermodynamics of the glass transition. Proc Natl Acad Sci. 2005;102(31):10837–10840.
[35] Biroli G, Garrahan JP. Perspective: the glass transition. J Chem Phys. 2013;138(12):12A301.
[36] Pinchaipat R, Campo M, Turci F, et al. Experimental evidence for a structural-dynamical transition in trajectory space. Phys Rev Lett. 2017;119(2):028004.
[37] Bolhuis PG, Chandler D, Dellago C, et al. Transition path sampling: throwing ropes over rough mountain passes, in the dark. Annu Rev Phys Chem. 2002;53(1):291–318.
[38] Chandler D, Garrahan JP. Dynamics on the way to forming glass: bubbles in space-time. Annu Rev Phys Chem. 2010;61:191–217.
[39] Giardina C, Kurchan J, Lecomte V, et al. Simulating rare events in dynamical processes. J Stat Phys. 2011;145(4):787–811.
[40] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science. 2000;290(5500):2323–2326.
[41] Das P, Moll M, Stamati H, et al. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. Proc Natl Acad Sci. 2006;103(26):9885–9890.
[42] Piana S, Laio A. Advillin folding takes place on a hypersurface of small dimensionality. Phys Rev Lett. 2008;101(20):208101.
[43] Tribello GA, Ceriotti M, Parrinello M. Using sketch-map coordinates to analyze and bias molecular dynamics simulations. Proc Natl Acad Sci. 2012;109(14):5196–5201.
[44] Ceriotti M, Tribello GA, Parrinello M. Demonstrating the transferability and the descriptive power of sketch-map. J Chem Theory Comput. 2013;9(3):1521–1532.
[45] Lemke T, Peter C. Encodermap: dimensionality reduction and generation of molecule conformations. J Chem Theory Comput. 2019;15(2):1209–1215.
[46] Tiwary P, Berne BJ. Spectral gap optimization of order parameters for sampling complex molecular systems. Proc Natl Acad Sci. 2016;113(11):2839–2844.
[47] Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM; 1999.
[48] Pathak J, Hunt B, Girvan M, et al. Model-free prediction of large spatiotemporally chaotic systems from data: a reservoir computing approach. Phys Rev Lett. 2018;120(2):024102.
[49] Wang Y, Lamim Ribeiro JM, Tiwary P. Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. Nat Commun. 2019;10(1):1–8.
[50] Salvalaglio M, Vetter T, Giberti F, et al. Uncovering molecular details of urea crystal growth in the presence of additives. J Am Chem Soc. 2012;134(41):17221–17233.
[51] Frauenfelder H, Sligar SG, Wolynes PG. The energy landscapes and motions of proteins. Science. 1991;254(5038):1598–1603.
[52] Ferreiro DU, Hegler JA, Komives EA, et al. On the role of frustration in the energy landscapes of allosteric proteins. Proc Natl Acad Sci. 2011;108(9):3499–3503.