A mini-review of clustering algorithms and their theoretical properties, with applications to molecular science

Frédéric Cazals

A mini-review of clustering algorithms and their theoretical properties, with applications to molecular science *

Frédéric Cazals †

March 14, 2024

Abstract

Clustering is a fundamental task, in particular to analyse potential and free energy landscapes in molecular science. In this survey, I review the key properties of three remarkable clustering algorithms (`k-means ++`, persistence-based clustering, and spectral clustering) with a double perspective. The first one is the specification of the main mathematical and algorithmic properties of the algorithms; the second one is the relevance of these methods for structural, thermodynamic, and kinetic analysis. Doing so provides a unique opportunity to mention important connexions between optimization, graph theory, geometry, and theoretical biophysics.

1 Introduction

Levinthal's paradox, complexity, and computer science. The folding of a protein or the formation of a protein complex raises a conundrum known as Levinthal's paradox: despite an exponential number of conformations, Nature finds the *native* states in a time frame incompatible with an exhaustive search. From the biophysical standpoint, Levinthal's paradox asks why and how biomolecules are such fast *structure seekers*; from the computational complexity standpoint, it asks which computational problem is being solved in such a short time scale.

This conundrum was at the heart of the Biogeometry project, which developed in the USA at the turn of the century. Its rationale was to promote collaborations between molecular science (protein science in particular) and computer science. The inception of the project was rooted in seminal works by Richards on atomic packing properties [1] and Janin on the description of protein - protein interfaces [2]. These works were relying on geometric algorithms such as Voronoi diagrams [3] and later α-shapes [4], which proved optimal compute molecular surfaces and volumes. With the participation of M. Levitt, the Biogeometry project also touched on dynamics, raising difficult questions early identified [5]. In 2013, Levitt, Karplus and Warshel were indeed awarded the Nobel Prize in chemistry for *the development of multiscale models for complex chemical systems*, including force fields and associated mathematical models.

Dynamics and Energy Landscapes. It is in this context, rocked between geometric algorithms on the one hand, and applications to molecular science on the other hand, that I attended the 2010 edition of the *Energy landscape* workshop organized in Chemnitz, on the recommendation of Christian Müller, a Swiss colleague by then at ETH [1]. The soccer World Cup was certainly providing opportunities for good discussions–especially regarding the strike of the French team, a state of affairs which fortunately changed in 2018. More importantly, on the scientific agenda, I was immediately struck by this small and vibrant group, whose main concerns were revolving around entropy, free energy, density of states, and the like. I was pushed out of my comfort zone, and *Energy Landscapes* became one of my favorite meetings, progressively

*In honor of the 66th birthday of Christian Schön

†Centre Inria d'Université Côte d'Azur, F-06902 Sophia-Antipolis, FRANCE. Email: Frederic.Cazals@inria.fr

[1]Program still available from `https://www.tu-chemnitz.de/physik/CPHYS/Conferences/EL/EL2010/`.

contributing to strengthen my understanding of the connexions between structure, thermodynamics, and kinetics. By decoupling structure, thermodynamics and kinetics [6, 7], energy landscapes indeed provide a route of choice to get insights in Levinthal's paradox. Also, by splitting complex problems into simpler ones, this framework opens the door to the design of more efficient algorithms and their mathematical analysis.

In these meetings, Christian Schön has been acting as an outstanding (scientific) entertainer, if I may say, with his endless curiosity and questions helping less knowledgeable attendees to further their understanding. This mini-survey is meant to foster this curiosity on one particular topic: clustering and some of its connexions in theoretical biophysics.

Energy landscapes and clustering. A central concept when studying energy landscapes is that of *basin* [8], [6]. From the differential geometry standpoint, given a so-called Morse function, a basin is defined as the stable manifold of the negative gradient vector fields [9], that is, those points *flowing* to the minimum along the integral curves of the negative gradient vector field.

From the computational standpoint, characterizing basins is a fundamental endeavor often undertaken using clustering algorithms. In a nutshell, given a set of data points (also called observations), together with a pairwise distance (or similarity) measure, *clustering* consists of forming homogeneous groups of observations. This short survey reviews three remarkable such algorithms, and discusses connexions with the study of basins, from the structural (Sec. 2), thermodynamic (Sec. 3), and kinetic standpoint (Sec. 4). Finally, Section 5 discusses molecular distances / similarities which can be used for clustering biomolecules.

2 `k-means` and `k-means++`

Intuition. As just said, a natural way to define clusters is in terms of *homogeneous* groups of data points. If one assigns one representative point to each cluster (not necessarily a data point), the *homogeneity* can be assessed by the variance of distances between points in the cluster and the cluster representative. Finding clusters minimizing the sum of cluster variances yields the `k-means` algorithm.

The technique. `k-means` is best explained for observations in the usual d-dimensional Euclidean space. For a predefined number of clusters k, the problem consists of forming a partition of the data set into k clusters $C_1, \ldots, C_k$. Each cluster is associated with a center c_i – the *representative* point in the previous paragraph, and each data sample is assigned to its nearest center. Initially, one chooses the k centers at random amongst the data points, and assigns each data point to its nearest center. This presentation entails that a clustering is implicitly defined by the Voronoi diagram [3] of the centers. (Recall that the Voronoi diagram is the partition of the plane induced by the equivalence relationship *have the same nearest center*.)

To further our understanding, recall that given a finite point set in $\mathbb{R}^d$, the point (which in general is not a data point) minimizing the sum of squared distances to the data points is the so-called center of mass. In `k-means`, the center of each cluster is its center of mass, which is why `k-means` aims at finding the partition minimizing the functional

$$\Phi_k = \sum_{i=1,\ldots,k} \sum_{x_j \in C_i} \|x_j - c_i\|^2 . \tag{1}$$

From the combinatorial standpoint, assigning n points to k clusters yields k^n possibilities, but this gross enumeration contains clusterings with empty clusters. A sharper analysis shows that the number of partitions into k nonempty clusters is given by the so-called Stirling number of second kind, which, for fixed k, behaves as $k^n/k!$. Getting further insights into the complexity of the problem requires discussing whether the values of d and k are fixed or not. (I) Assume d and k are fixed. As noted above, a `k-means` clustering is implicitly defined by the Voronoi diagram of the k centers in the d-dimensional space. By enumerating all such partitions and computing the functional of Eq. (1), one solves `k-means` exactly. It turns out that the maximum number of possible partitions of n data points generated by the Euclidean Voronoi diagram is $O(n^{O(dk)})$ [10], so that `k-means` is solvable in polynomial time. (II) Assume now that n and d are free, but $k = 2$ is fixed. Then, using a reduction constructing instances of 2-means with n points in dimension

$d = 2n$, it has been shown that `k-means` is NP-hard [11]. (III) Finally, with $d = 2$, assume that n and k are free–that is there is no bound on k which can be linear in n. Then planar `k-means` is also NP-hard [12].

From a practical standpoint, the previous insights do not provide a general effective procedure. Forty years back, Lloyd invented the eponymous iterations scheme to solve this problem [13]. Having selected at random an initial set of k centers, the process consists of iterating two steps: (i) ascribe each data point to its nearest center, (ii) recompute the center of mass of each cluster. The process halts when the clusters are stable.

The outcome naturally depends on the initial choice of seeds–it is a random variable–and no information is provided with respect to the optimal value of Φ_k, denoted Φ_{OPT}. In this context, a *tour de force* has been the design of the *smart* seeding strategy which consists of ensuring that the initial centers are correctly placed in the unknown clusters, yielding the `k-means++` algorithm. This is achieved by biasing the choice of initial seeds using the squared distance between the points themselves [14]. (For a review of seeding strategies, see also [15].) The expected outcome is qualified by the following theorem, whose proof requires computing the expectation of the random variable given by Eq. (1) on all possible initial choices of the seeds:

$$\mathbb{E}\left[\Phi\right] \leq 8(\ln k + 2)\Phi_{OPT}. \tag{2}$$

Remarkably, this result is not a consequence of Lloyd iterations, but instead of a choice of the initial random seeds. In practice, state-of-the-art implementations of `k-means++`, e.g. that of Sci-kit-learn run say 10 instances of `k-means`, and retain the best result.

As a final comment, we note that (in practice!) `k-means` or `k-means++` require a predefined number of clusters k. Increasing k yields a decrease of Φ_k. The usual strategy to choose a *good* k consists of plotting the *elbow cure*, namely Φ_k as a function of k. A bent in this curve means that adding clusters is not beneficial anymore.

Connexion to energy landscapes. `k-means` solely focuses on geometric distances to the cluster centers. and promotes rather isotropic clusters, which may not be pertinent to define basins. One possible solution to this problem consists of running `k-means` twice, by varying the number of clusters, and to seek a correspondence between the clusters of these two clusterings–using the D-family-matching matching algorithm developed in [16]. Doing so indeed aggregates clusters into *meta-clusters* whose geometry may be a better approximation of that of basins. For example, one obtains information such as "*the union of clusters A_i and A_j and A_k from clustering one matches the union of clusters B_l and B_m from clustering two*". Nevertheless, with a focus on the geometric consistency, `k-means` is therefore best suited for structural analysis. An important application has been the design of states in Markov state models [17].

Software. `k-means++` is available in a number of software suites, *e.g.* `scikit-learn` `https://scikit-learn.org/stable/`. The implementation in the Structural Bioinformatics Library `https://sbl.inria.fr` is generic and makes it possible to run `k-means++` for unit vectors on the d-dimensional spheres, as required by spectral clustering (see below).

Notes. An original approach to `k-means` consists of seeking the coordinates of the k centers minimizing Φ_k, in which case `k-means` becomes an optimization problem in $\mathbb{R}^{dk}$ [18]. This approach has the advantage of providing insights on the landscape of solutions, encoded in a disconnectivity graph (DG) [6], as intuitively, almost equivalent clusterings are found in *funnels* of the DG. However, I am not aware of any theoretical result in this setting.

As stressed above, `k-means` uses centers of masses (c.o.m.) of the original points, which in general are not data points. This can be an issue as the c.o.m. may not even live in the same space. For example, the c.o.m. of points on a circle is not in general located on this circle. In all generality, the c.o.m. of points in a general metric space (*e.g.* a manifold equipped with a Riemannian metric) is the so-called Fréchet mean [19, 20]. Computing accurately such c.o.m. is in general challenging. The case of angular data, which is of interest in molecular science to model torsion angles, has recently been handled in [21]. The version of `k-means` using data points as representatives (instead of c.o.m.) is known a `k-medoids` [22].

3 Persistence-based clustering

Intuition. A classical problem for geographers is to properly define peaks on a mountain. In principle, any modest rock on a mountain trail could define a peak since it is a local maximum of the elevation, but this lack common sense. To avoid such outcomes, a proper definition of peaks usually involves two criteria [23]: the prominence (the flying distance to a taller local maximum), and the culminance (the elevation drop to a local minimum connected to a higher local maximum). These notions have been used to design clustering methods as well. Intuitively, one can define a probability density estimate in the data space by locally counting the neighboring data points of a point. If properly filtered out using culminance and/or prominence, the *attraction basins* of the retained local maxima of this estimated density define clusters. Let us now review more specifically methods following this spirit.

The popular algorithm `DBSCAN` [24] defines clusters using core points (data point with a predefined number of neighbors within a ball of fixed radius) which are mutually reachable. To relieve these two fixed thresholds, a more general model consists of assuming that the points have an underlying probability density, in which case a cluster can be defined as the *attraction basin* of each local maxima (modes) of this density [25]. The corresponding algorithm is known as *mean shift*, or *mode seeking*, and various variants have been used in molecular science [26]. In the sequel, we present a more elaborate version relying on *topological persistence* [27], a technique developed in the context of geometric and topological inference [28].

The technique. Let us now make two more assumptions about the data. First, we assume that the samples x_i form the vertices of a graph. One typically uses a *nearest neighbor graph* (NNG), connecting a sample to a predefined number of neighbors, or to all samples within a given distance radius. Second, we assume that each sample has an *elevation* or *height*, defining a *landscape*. For clustering, the height is the estimated density at the sample, obtained from say a kernel density estimate. Then, the *attraction basin* of each such local maximum defines a cluster. When samples represent molecular conformations, the height is in general a potential (or free) energy. Then, clusters correspond to the *attraction basin* of local minima. To present a coherent treatment of both cases, negating the estimate of the density, we focus in the sequel on local minima and their attraction basins.

With these two assumptions, the *steepest* edge connecting a sample x_i to one of its neighbors can be used to estimate the (negative) gradient of the height function, called the *pseudo-gradient* in the sequel. Using the pseudo-gradient, one obtains a generalization of the continuous world: a sample whose neighbors in the NNG are all above (resp. below) is termed a local minimum (resp. maximum); under suitable conditions (to avoid degeneracies such as monkey saddles), a sample with neighbors above and below is termed a saddle point. To assign points to basins of local minima, one processes points by increasing elevation, and for each point which is not a saddle point, iteratively follows the pseudo-gradient to end up in a local minimum. The *persistence* of a basin is the elevation rise between the height of its minimum, and that of the lowest saddle connecting to a deeper basin. This is also called the *barrier height* in biophysics. Considering all basins, the 2D scatter plot whose x-axis (resp y-axis) is the birth date (resp. the height of the saddle defining the persistence) is the so-called *persistence diagram* (PD). In the PD, points near the diagonal define *topological noise*, while points away from the diagonal define significant basins. In the context of clustering, the structure of the PD has been characterized under suitable assumptions [27].

Connexion to energy landscapes. The key strength of this method is that the PD provides a natural way to infer a suitable number of persistent local minima (for basins) or local maxima (for clusters). For classical clustering applications, one expects in general a small number of clusters. In molecular science, the barrier heights typically vary continuously [29], in which case one may count the number of basins as a function of persistence, a natural way to assess *ruggedness*.

Once the number of significant basins has been decided, the landscape is easily simplified using the so-called Union-Find algorithm [30]. Moreover, the simplification allows retaining the connexions between the critical points (minima, saddle points) corresponding to persistent basins only [31]. These connexions are of special interest in particular when using Arrhenius' law. A clear limitation of persistence-based clustering,

though, is that the analysis relies solely on energy barriers. Neither the *shape* of basins nor transition probabilities between basins are taken into account.

Software. Persistence-based clustering is available in the Gudhi library for Topological Data Analysis `https://gudhi.inria.fr/index.html`, as well as in the Structural Bioinformatics Library `https://sbl.inria.fr`, which also features the methods presented in [32].

Notes. The computation of a NNG is an interesting problem in itself, if one wishes to avoid the quadratic cost associated to the inspection of every single pair of data points. For Euclidean distance, one can use kd-trees or random projection trees (RP trees) [33]. For metrics not associated to an embedding, that is, when one gets pairwise distances but cannot use coordinates to split the ambient space, the generalization of kd-trees and RP trees are known as metric trees [34].

4 Spectral clustering

Intuition. Spectral clustering is tightly connected to two classical problems in graph theory. The first is community detection: loosely speaking, a *community* in a graph is a set of nodes highly interconnected, but with few connexions to other nodes outside that community. The second one is the notion of cut. Given a graph, a cut is a partition of the vertices into two subsets, obtained by removing selected edges. A *maximum cut* is such that its size is at least the size of any other cut, and a cut is *balanced* if the two subsets have *comparable* sizes. In short, spectral clustering is a set of techniques aiming at finding communities based on spectral properties of the so-called *graph Laplacian*, and the outcome provides provably good approximations of graph cuts [35].

The technique. Assume that the samples x_i form the vertices of a graph, *e.g.* a NNG graph. Let us also assume that edges carry positive weights $w_{ij} \geq 0$. Without any a-priori assumption, uniform weights $w_{ij} = 1$ are used. In a geometric setting, one may use

$$w_{ij} = \exp(-\left\|x_i - x_j\right\|^2 / (2\varepsilon^2)), \tag{3}$$

with ε a scale parameter. The graph is non-directed so that weights are symmetric, and the corresponding matrix is denoted W. Define the generalized degree d_i of a node as the sum over its neighbors of the weights w_{ij}, and let D be the corresponding diagonal matrix. The graph Laplacian is the matrix $L = D - W$. This matrix has remarkable properties. One of them states that if the graph has k connected components, the eigenspace associated with the eigenvalue 0 is the set of indicator vectors for the connected components. Two quantities related to L are the normalized Laplacian $L_{\text{sym.}} = D^{-1/2}LD^{-1/2}$, and the random walk Laplacian $L_{\text{rw}} = D^{-1}L$.

One remarkable spectral clustering algorithm to obtain k clusters is as follows [36]. First, compute the eigenvalues $L_{\text{sym.}}$ and sort them by increasing value. Retain the eigenvectors associated with the k smallest eigenvalues and form the corresponding $n \times k$ matrix by truncating these vectors to their first k coordinates. Normalize the rows of this matrix, so that each row now represents a unit vector on the unit $(k-1)$-dimensional sphere S^{k-1}. (NB: intuitively, each such vector is the indicator vector indicating to which cluster the row/point belongs to.) Then, run `k-means` (or better `k-means++`) on these vectors. Remarkably, for $k = 2$, the result provides an approximation of the NP-hard problem max cut–via a so-called relaxation.

The previous algorithm also has an interpretation in terms of random walk. To see how, define the random walk / Markov chain on the aforementioned weighted graph as follows: starting from a given node, say x_i, the walker moves to one of its neighbors x_j which is chosen according to the probability mass function $\{w_{ij}/d_i\}$. The corresponding transition matrix is $P = D^{-1}W = I - L_{\text{rw}}$. Because of the previous equality, λ is an eigenvalue of L_{rw} with eigenvector for u if and only if $1 - \lambda$ is an eigenvalue of P with eigenvector u. Small eigenvalues of $L_{\text{sym.}}$ or large values of the transition matrix P are thus equivalent to determine clusters. Using this construction, it can be shown that the random walk can be used to define a cut which seldom transitions from one vertex set to its complement [35].

Connexion to energy landscapes. The random walk construction calls for a comment with respect to the mean first passage time (MFPT) m_{ij}, namely the expectation of the time needed to reach node x_j when the random walker starts at x_i. Symmetrizing MFPT yields the commute distance $c_{ij} = m_{ij} + m_{ji}$, namely the expected time to travel from x_i to x_j and then back to x_i. The commute distance can be computed with the generalized inverse of the Laplacian L [37], and $\sqrt{c_{ij}}$ is actually an Euclidean distance [38, 39]. In the limit $n \to \infty$, it has been shown that the commute distance converges to $1/d_i + 1/d_j$, thus depending on the local density around the nodes rather than the structure of the graph [40].

Yet another related construction is diffusion maps (DM) [41, 42, 43]. Consider a weight matrix W defined from a kernel –see Eq. 3. Since $w_{ij} > 0$, the Perron-Frobenius yields eigenvalues $1 = \lambda_1 > \mid \lambda_2 \mid \geq \ldots \mid \lambda_n \mid$. The DM is an embedding of each sample x_i using the eigenvalues and right eigenvectors of P. Also denote $p_t(x_j \mid x_i)$ the probability to reach x_j from x_i in t steps. Denoting $\pi(x_k)$ the probability of node x_k in the stationary distribution of the Markov chain, the diffusion distance is defined as

$$D_t^2(x_i, x_j) = \sum_{x_k} (p_t(x_k \mid x_i) - p_t(x_k \mid x_j))^2 / \pi(x_k). \tag{4}$$

As opposed to the commute time, the diffusion distance depends on the parameter t.

The connexion between diffusion maps and Langevin (and Fokker-Planck) equations has been studied in detail [43]. For data sampled according to the Boltzmann distribution of some potential energy $V(x)$, the eigenvectors/values of the normalized Laplacian $L_{\text{sym.}}$ correspond to a diffusion process with potential $2V(x)$. A proper normalization makes it possible to recover a diffusion with potential $V(x)$. This connexion underlies the application of diffusion maps for the estimation of free energy landscapes [44], and to the design of collective coordinates [45]. As a final comment, we also note the connexion of the discrete random walk to its continuous analogous, generally referred to as master equations [46, 6].

Software. The manipulation of Laplacian requires a linear algebra library, and I recommend the C++ library Eigen `https://eigen.tuxfamily.org/`. The implementation of the spectral clustering exposed above requires a generic implementation of `k-means` accommodating points on the unit sphere. As noted earlier, one such implementation can be found in the Structural Bioinformatics Library.

Notes. A key property of spectral clustering is to be related to a diffusion process sampling Boltzmann distribution. However, from the computational side, choosing the right *bandwidth* is an issue–the parameter ε in Eq. 3, letting alone the manipulation of potentially large matrices. Spectral clustering is related to nonnegative matrix factorization and [47] and `k-means` [48]. In the realm of molecular science, for a graph with unit weights, the Laplacian is also the so-called Kirchhoff matrix. For a molecular model equipped with springs (elastic network model) it is associated with the calculation of isotropic fluctuations, which can be made anisotropic using normal modes [49]. Interestingly, spectral clustering applied to a similarity matrix derived from atomic fluctuations [50] or pairwise atomic contacts [51] yields a very effective strategy to identify (quasi-)rigid domains in proteins.

5 Additional notes

5.1 Molecular distances

As stated above, any clustering algorithm requires a distance / similarity measures. Let us briefly review the main options used for biomolecules. The classical distance used when working in Cartesian coordinates is the *least Root Mean Square Deviation* (lRMSD), computed from the optimal rigid motion superimposing the two structures studied [52]. The lRMSD requires an alignment, depends on the alignment length, and medium values usually convey little information. We note that computing the alignment requires care to handle molecular symmetries, and is a problem in itself for proteins which do not have the same sequence–see

e.g. [53]. To counter the difficulties of the lRMSD, one can combine local measures, e.g. the so-called Binet-Cauchy score [54]. In a similar spirit, for a molecule decomposed into a number of domains, the lRMSD associated to these domains can be combined as described in [55].

The decomposition of a structure into rigid domains is also relevant in the context of energy landscape exploration, *e.g.* to define move sets. The aforementioned Laplacian can also be used to identify globular/rigid domains within molecules using spectral clustering [50].

5.2 Parametric clustering models

The three clustering methods discussed above do not use any analytical model for the clusters. This limitation prompted the development of `k-subspace` clustering methods [56, 57], whose aim is to select a small number of original coordinates (features) so that clusters are clearly identified in those subspaces. Classical subspace techniques use a unique predefined cluster model, and need to cope with overfitting. Since *richer* models better fit the data (*e.g.* plane better fits noisy data distributed along a line than the line itself), controlling the number of model parameters is usually done using criteria such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC) [58], or the Minimum Description Length (MDL) [59]. As an example, using such a regularization, the method from [60] seeks compact clusters in affine spaces of small dimension. Such regularized clustering methods are somewhat similar the removal of noisy features using dimensionality reduction to ease the clustering task [36].

5.3 Comparing clusterings and clustering assessment

Clustering a data set can be done in many ways, by varying the clustering algorithm and its parameters. Methods to assess and compare clusterings are thus of special interest. In a statistical context, the assessment of a clustering can be done resorting to a null model [61, 62]. To compare two clusterings irrespective of a null model, global measures such as Rand index and the mutual information can be used [63, 62]. To go beyond these global measures and find a one-to-one matching between meta-clusters, one can use the aforementioned D-family-matching matching algorithm developed in [16].

6 Outlook

The study of energy landscapes is a fundamental endeavor, with far-reaching connexions between (statistical) physics, complexity theory, and algorithms. Clustering is one algorithmic technique of prime importance, and I hope this mini-survey will contribute to disseminate some of the key ideas underlying three prominent clustering methods, and also to leverage the analysis of (bio-)molecular data.

The many editions of the *Energy Landscapes* I have attended, in Chemnitz (Germany), Obergurgl (Austria), Durham (UK), Kalamata (Greece), Telluride (US), or Porquerolles (France) have provided unique opportunities to ponder on such problems. From an epistemological point of view, it is certainly important to enjoy the output of *Artificial Intelligence/Deep Learning* methods, for molecules in general, and proteins in particular. However, I believe it is equally important to keep moving with physical concepts and algorithms whose complexity is understood, as these are critical contributions to knowledge in terms of interpretable models.

I am certainly looking forward to more discussions with this small, eclectic, and vibrant community, and with Christian Schön in particular.

Acknowledgments. I wish to thank the anonymous reviewers and Christian Schön for their suggestions in improving the manuscript.

References

[1] B. Lee and F.M. Richards. The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology*, 55(3):379–380, 1971.

[2] C. Chothia and J. Janin. Principles of protein-protein recognition. *Nature*, 256:705–708, 1975.

[3] J.-D. Boissonnat and M. Yvinec. *Algorithmic geometry*. Cambridge University Press, UK, 1998. Translated by H. Brönnimann.

[4] H. Edelsbrunner. Weighted alpha shapes. Technical Report UIUCDCS-R-92-1760, Dept. Comput. Sci., Univ. Illinois, Urbana, IL, 1992.

[5] C.L. Brooks, M. Karplus, and B. Montgomery Pettitt. *Advances in Chemical Physics, Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics*. Wiley, 1988.

[6] D. J. Wales. *Energy Landscapes*. Cambridge University Press, 2003.

[7] J.C. Schön. Energy landscapes in inorganic chemistry. In J. Reedijk and K.R. Peoppelmeier, editors, *Comprehensive inorganic chemistry III*. Elsevier, 2023.

[8] J.C. Schön and M. Jansen. Prediction, determination and validation of phase diagrams via the global study of energy landscapes. *Int. J. of Materials Research*, 100(2):135, 2009.

[9] J.W. Milnor. *Morse Theory*. Princeton University Press, Princeton, NJ, 1963. m-mt-63.

[10] Mary Inaba, Naoki Katoh, and Hiroshi Imai. Applications of weighted voronoi diagrams and randomization to variance-based k-clustering. In *Proceedings of the tenth annual symposium on Computational geometry*, pages 332–339, 1994.

[11] S. Dasgupta. The hardness of k-means clustering. Technical Report CS-2007=0890, UCSD, 2008.

[12] Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is NP-hard. *Theoretical Computer Science*, 442:13–21, 2012.

[13] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 1982.

[14] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *ACM-SODA*, page 1035. Society for Industrial and Applied Mathematics, 2007.

[15] M. Celebi, H. Kingravi, and P. Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm/. *Expert systems with applications*, 40, 2013.

[16] F. Cazals, D. Mazauric, R. Tetley, and R. Watrigant. Comparing two clusterings using matchings between clusters of clusters. *ACM J. of Experimental Algorithms*, 24(1):1–42, 2019.

[17] V. Pande, K. Beauchamp, and G.R. Bowman. Everything you wanted to know about markov state models but were afraid to ask. *Methods*, 52(1):99–105, 2010.

[18] Luke Dicks. *K-means landscapes: exploring clustering solution spaces using energy landscape theory*. PhD thesis, Cambridge Univ., Dpt. of Chemistry, Cambridge, UK, 2021.

[19] M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré*, volume 10, pages 215–310, 1948.

[20] Xavier Pennec, Stefan Sommer, and Tom Fletcher. *Riemannian Geometric Statistics in Medical Image Analysis*. Academic Press, 2019.

[21] F. Cazals, B. Delmas, and T. O'Donnell. Fréchet mean and p-mean on the unit circle: decidability, algorithm, and applications to clustering on the flat torus. In D. Coudert and E. Natale, editors, *Symposium on Experimental Algorithms*, Sophia Antipolis, 2021. Lipics.

[22] Leonard Kaufman. Partitioning around medoids (program pam). *Finding groups in data*, 344:68–125, 1990.

[23] C. Thöni. Criteria to define summits in the Swiss alps: prominence and culminance height. *Les Alpes*, 1:26–28, 2003.

[24] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[25] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE PAMI*, 17(8):790–799, 1995.

[26] A. Rodriguez and A. Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

[27] F. Chazal, L. Guibas, S. Oudot, and P. Skraba. Persistence-based clustering in Riemannian manifolds. *J. ACM*, 60(6):1–38, 2013.

[28] Jean-Daniel Boissonnat, Frédéric Chazal, and Mariette Yvinec. *Geometric and topological inference*, volume 57. Cambridge University Press, 2018.

[29] J. Carr, D. Mazauric, F. Cazals, and D. J. Wales. Energy landscapes and persistent minima. *The Journal of Chemical Physics*, 144(5):4, 2016.

[30] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2022.

[31] F. Cazals and D. Cohen-Steiner. Reconstructing 3D compact sets. *Computational Geometry Theory and Applications*, 45(1-2):1–13, 2011.

[32] F. Cazals, T. Dreyfus, D. Mazauric, A. Roth, and C.H. Robert. Conformational ensembles and sampled energy landscapes: Analysis and comparison. *J. Comp. Chem.*, 36(16):1213–1231, 2015.

[33] S. Dasgupta and K. Sinha. Randomized partition trees for exact nearest neighbor search. *JMLR: Workshop and Conference Proceedings*, 30:1–21, 2013.

[34] Peter N Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces. In *ACM SODA*, volume 93, pages 311–321, 1993.

[35] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[36] Donglin Niu, Jennifer Dy, and Michael I Jordan. Dimensionality reduction for spectral clustering. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 552–560. JMLR Workshop and Conference Proceedings, 2011.

[37] L. Lovász. Random walks on graphs: A survey. *Combinatorics, Paul erdos is eighty*, 2(1):1–46, 1993.

[38] Douglas J Klein, José Luis Palacios, Milan Randić, and Nenad Trinajstić. Random walks and chemical graph theory. *Journal of chemical information and computer sciences*, 44(5):1521–1525, 2004.

[39] Francois Fouss, Alain Pirotte, Jean-Michel Renders, and Marco Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on knowledge and data engineering*, 19(3):355–369, 2007.

[40] Ulrike Von Luxburg, Agnes Radl, and Matthias Hein. Hitting and commute times in large random neighborhood graphs. *The Journal of Machine Learning Research*, 15(1):1751–1798, 2014.

[41] Boaz Nadler, Stephane Lafon, Ioannis Kevrekidis, and Ronald Coifman. Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. *Advances in neural information processing systems*, 18, 2005.

[42] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, and S.W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *PNAS*, 102(21):7426–7431, 2005.

[43] B. Nadler, S. Lafon, R. Coifman, and I.G. Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006.

[44] M. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi. Determination of reaction coordinates via locally scaled diffusion map. *J. of Chemical Physics*, 134(12), 2011.

[45] E. Chiavazzo, R. Covino, R. Coifman, C.W. Gear, A. Georgiou, G. Hummer, and I. Kevrekidis. Intrinsic map dynamics exploration for uncharted effective free-energy landscapes. *PNAS*, 114(28):E5494–E5503, 2017.

[46] R.S. Berry and R. Breitengraser-Kunz. Topography and dynamics of multidimensional interatomic potential surfaces. *Physical review letters*, 74(20):3951, 1995.

[47] Chris Ding, Xiaofeng He, and Horst D Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM international conference on data mining*, pages 606–610. SIAM, 2005.

[48] Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135, 2006.

[49] Ali Rana Atilgan, SR Durell, Robert L Jernigan, Melik C Demirel, O Keskin, and Ivet Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal*, 80(1):505–515, 2001.

[50] Luca Ponzoni, Guido Polles, Vincenzo Carnevale, and Cristian Micheletti. SPECTRUS: A dimensionality reduction approach for identifying dynamical domains in protein complexes from limited structural datasets. *Structure*, 23(8):1516–1525, 2015.

[51] F. Cazals, J. Herrmann, and E. Sarti. Simpler protein domain identification using spectral clustering. NA(NA), 2024. https://www.biorxiv.org/content/10.1101/2024.02.10.579762v1.

[52] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976.

[53] D. Ritchie, A. Ghoorah, L. Mavridis, and V. Venkatraman. Fast protein structure alignment using Gaussian overlap scoring of backbone peptide fragment similarity. *Bioinformatics*, 28(24):3274–3281, 2012.

[54] F. Guyon and P. Tufféry. Fast protein fragment similarity scoring using a Binet-Cauchy kernel. *Bioinformatics*, 30(6):784–791, 2014.

[55] F. Cazals and R. Tetley. Characterizing molecular flexibility by combining lRMSD measures. *Proteins: structure, function, and bioinformatics*, 87(5):380–389, 2019.

[56] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *Acm sigkdd explorations newsletter*, 6(1):90–105, 2004.

[57] Dingding Wang, Chris Ding, and Tao Li. K-subspace clustering. 2009.

[58] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of statistical learning : data mining, inference and prediction*. Springer, 2001. htf-esldm-01.

[59] P. D. Grünwald. *The minimum description length principle*. MIT press, 2007.

[60] F. Cazals, L. Goldenberg, and S. Suren. Subspace-embedded spherical clusters: a novel cluster model for compact clusters of arbitrary dimension. *Submitted*, 2024.

[61] Allan D Gordon. Null models in cluster validation. In *From Data to Knowledge: Theoretical and Practical Aspects of Classification, Data Analysis, and Knowledge Organization*, pages 32–44. Springer, 1996.

[62] Alexander J Gates and Yong-Yeol Ahn. The impact of random models on clustering similarity. *Journal of Machine Learning Research*, 18(1–28), 2017.

[63] M. Meilă. Comparing clusterings–an information based distance. *Journal of multivariate analysis*, 98(5):873–895, 2007.