Conformational Ensembles and Sampled Energy Landscapes: Analysis and Comparison

Frédéric Cazals, Tom Dreyfus, Dorian Mazauric, Andrea Roth, Charles Robert

Conformational Ensembles and Sampled Energy Landscapes: Analysis and Comparison

F. Cazals and T. Dreyfus and D. Mazauric and A. Roth and C. H. Robert

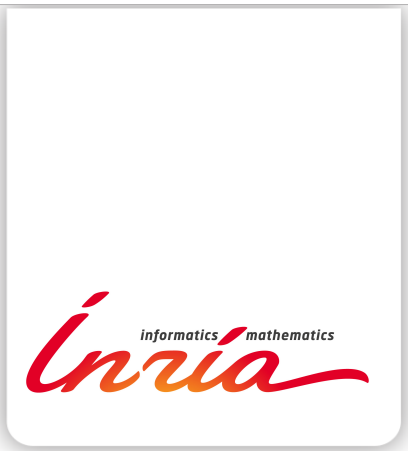
RESEARCH REPORT

N° 8610

October 2014

Project-Team Algorithms-Biology-Structure (ABS)

ISSN 0249-6399 ISRN INRIA/RR--8610--FR+ENG



Conformational Ensembles and Sampled Energy Landscapes: Analysis and Comparison

F. Cazals* and T. Dreyfus† and D. Mazauric‡ and A. Roth §
and C. H. Robert¶

Project-Team Algorithms-Biology-Structure (ABS)

Research Report n° 8610 — version 2 — initial version October 2014 —
revised version March 2015 — 70 pages

Abstract: We present novel algorithms and software addressing four core problems in computational structural biology, namely analyzing a conformational ensemble, comparing two conformational ensembles, analyzing a sampled energy landscape, and comparing two sampled energy landscapes. Using recent developments in computational topology, graph theory, and combinatorial optimization, we make two notable contributions. First, we a present a generic algorithm analyzing height fields. We then use this algorithm to perform density based clustering of conformations, and to analyze a sampled energy landscape in terms of basins and transitions between them. In both cases, topological persistence is used to manage ruggedness. Second, we introduce two algorithms to compare transition graphs. The first is the classical *earth mover distance* metric which depends only on local minimum energy configurations along with their statistical weights, while the second incorporates topological constraints inherent to conformational transitions.
Illustrations are provided on a simplified protein model (BLN69), whose frustrated potential energy landscape has been thoroughly studied.
The software implementing our tools is also made available, and should prove valuable wherever conformational ensembles and energy landscapes are used.

Key-words: Molecular conformations, energy landscapes, sampling, Morse theory, optimal transport, optimization

* Inria ABS. Correspondance: Frederic.Cazals@inria.fr
† Inria ABS
‡ Inria Geometrica
§ Inria ABS
¶ IBPC-LBT/CNRS

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Ensembles de conformations, et paysages énergétiques échantillonnés: analyse et comparaison

Résumé : Nous présentons de nouveaux algorithmes pour quatre problèmes de modélisation d'importance centrale en biologie structurale, à savoir analyser un ensemble de conformations, comparer deux tels ensembles, analyser un paysage énergétique échantillonné, et comparer deux tels paysages. En exploitant certains développements récents en topologie computationnelle, théorie des graphes, et optimisation combinatoire, nous présentons deux contributions principales. D'une part, nous présentons un algorithme générique permettant d'analyser une fonction hauteur. Cet algorithme est ensuite utilisé pour effectuer un clustering basé sur la densité d'états, et aussi pour étudier un paysage énergétique échantillonné. Dans les deux cas, la théorie de la persistence topologique est utilisée pour gérer la *rugosité*. D'autre part, nous introduisons deux algorithmes permettant de comparer deux graphes de transition. Le premier algorithme est la distance classique dite de *transport de masse*, alors que le second incorpore des contraintes de connectivité inhérentes aux transitions observées sur un paysage énergétique.

A titre d'illustration,ces algorithmes sont appliqués à un modèle de protéine simplifiée (BLN69), dont le paysage énergétique frustré a été échantillonné en détail.

Les logiciels implémentant nos algorithmes sont mis à disposition, et devraient s'avérer utiles dans toutes les situations où des ensembles de conformations et des paysages énergétiques echantillonnés sont manipulés.

Mots-clés : Conformations moléculaires, paysages énergétiques, échantillonnage, théorie de Morse, théorie du transport optimal, optimisation

Contents

1 Introduction

1.1 Analyzing conformational ensembles and energy landscapes

colorblack

1.1.1 Energy landscapes

Evaluating the potential energy of a molecular system at every point in its configurational space yields the associated potential energy landscape (PEL), from which thermodynamic and kinetic properties of the associated free energy landscape (FEL) can be derived [Wal03]. Features of the PEL play a crucial role in understanding the system's behavior: in particular, the local minima correspond to (meta-)stable states, whereas a *saddle* connecting two local minima defines a possible transition between the linked meta-stable states. Indeed, decomposition of the PEL into its associated minima and saddle points can be used to model the dynamics of the system. However, obtaining a representative view of the PEL for systems involving from hundreds to tens of thousands of atoms is a major undertaking, as the number of minima increases exponentially with the size of the system. Because an analytical expression of the PEL is in general unknown, explorations are usually conducted via numerical simulations, yielding two broad classes of questions: exploring and sampling the PEL, and exploiting such samplings in the context of deriving thermodynamic and kinetic information.

1.1.2 Exploring landscapes

Exploring a PEL consists of generating a collection of points in the conformational space. We will refer to a collection as a *sampling* (or a *sampled landscape*, or simply *landscape*) in the following. The goals of such explorations are essentially twofold. On the one hand, one wishes to visit basins with low potential energy. On the other hand, one also wishes to sample transitions between basins, as these condition the time evolution of the system. Explorations may in principle be carried out using standard Metropolis algorithm (Monte Carlo, MC) or molecular dynamics (MD) simulations, in both of which cases the sampled data consists of points on the energy surface that in general correspond to neither minima nor saddle point configurations. Such methodologies, standardly applied in the canonical ensemble, are most useful for systems without significant excursions away from the global minimum; otherwise they are susceptible to being trapped in local minima, being unlikely to traverse high energy barriers. In the following we present a brief, necessarily incomplete overview of some of the prototypical sampling algorithms applied to the more general problem of global exploration of rugged PELs associated with high-dimensional systems.

colorblackWe first examine strategies derived from the Metropolis algorithm. For systems with a continuous configuration space, a key Monte Carlo based strategy performing a random walk in the space of local minima is basin hopping [LS87]. In short, once a Monte Carlo sampling strategy/move set has been chosen, the method consists of taking a random step away from the current configuration and minimizing the energy, and then accepting/rejecting the local minimum obtained based on the Metropolis criterion. We note in passing that the aforementioned minimization step, when carried out deterministically, consists in following an integral curve leading to a local minimum. This operation, also called *quenching*, is key to revealing the basins visited and the barriers crossed [EK87, TB95]. colorblackCombining basin hopping with the choice of random configurations confined below a lid yields the *threshold algorithm* [SPJ96], a refinement of the lid algorithm, a method targeting systems for which neighbors of particular configuration can be exhaustively enumerated [SSSA93, SvdPS99]. In a related vein, since many

paths from a high-lying point may lead to quite different local minima, stochastic quenches have also been used [WSM99]. Hierarchical optimization strategies have also been developed, in particular to handle systems with quasi-independent sub-systems. In that case, conformations of lowest energy may be obtained by identifying and combining nearly-rigid blocks from known local minima. This idea is exploited in [Kri02], where candidate conformations are generated by mixing coordinates of known minima.

One may also mention strategies avoiding revisiting known regions, a scheme known as *tabu search* in optimization. For robot motion planning, methods aiming at fostering the exploration of unexplored space have been designed using rapidly-exploring random trees [LKJ00]. Combining these ideas with the Metropolis test yields the *transition rapidly growing random trees* method [JCPC11b]. While such "history-dependent potential" approaches are thus commonly used to enhance exploration of the PEL, the meta-dynamics method combines them with collective coordinates (see below) to sample the FEL [LP02].

Exploration of systems characterized by rugged PELs may also be achieved using enhanced MD or MC exploration. An early development was simulated annealing [KV$^+$83], in which an initally high-temperature MC simulation is slowly cooled (on a prescribed schedule) in order to explore lower and lower energy minima. Generalized ensemble approaches, initially developed in the Monte Carlo context[BN92] (reviewed in [MO09]), rely on non-Boltzmann weighting to achieve a random walk in, e.g., potential energy or temperature space, the latter approach referred to as the expanded ensemble or simulated tempering [LMSVV92]. These approaches, while in principle allowing the system to surmount any energy barrier, shift the burden towards determining appropriate weight factors; however, once this is achieved, canonical (Boltzmann) sampling over a wide range of temperatures can be obtained from a single simulation. The generalized ensemble concept has been implemented in standard and Langevin MD [HOE96]. A related approach, requiring no *a priori* knowledge of the weight factors, is the widely used replica-exchange molecular dynamics (REMD) [SO99] in which configurations from a set of independent, neighboring-temperature canonical simulations– launched in parallel over as wide a temperature range as possible– are exchanged at a specified frequency following a Metropolis-like criterion. MD approaches have also been combined with simulated annealing to better explore rough PELs, notably using REMD starting with a restrained set of replicas at high temperature[KZ09].

As shown by the previous descriptions, exploring a landscape is clearly a technical endeavor, and depends on the type of model used (discrete or continuous system, atomic or coarse grained, etc), the potential energy function, the sampling protocol and associated parameters (temperature, step size, etc). The approaches we develop below can be applied to collections of both types: those consisting essentially of non-critical point configurations, from MD, for example, or else those obtained from extensive mapping of the PEL in terms of optimized critical points; e.g., minima obtained by basin hopping, which may or not be enriched with associated saddle points.

1.1.3 Modeling thermodynamic and kinetic properties

Qualitative insight into the system's dynamics can be obtained directly from the sampling using connectivity information. Of particular interest in this context are tree based representations encoding the connectivity of basins of the PEL. In the context of landscape exploration, the *disconnectivity graph (DG)* was introduced to model diffusion processes using a random walk on the tree connecting the basins [HS88], and to portray the shape and the connectivity of the landscape [BK97, Heu97]. In particular, typical DG were proposed to explain the performances of *fast structure seekers* [WMW98, DWB05, Ber10]. A disconnectivity graph being a tree (or a forest of trees), more precise statistics on the *transition graph* connecting local minima across

saddles were considered (degree distribution, graph diameter, path lengths). In particular, it has been shown that selected systems exhibit *small world* properties [DM05].

Quantitative thermodynamic and kinetic analysis are typically undertaken in two ways.

Dimensionality-reduction based approaches. Sampling of a given system generated by MC or MD is generally assumed to define a thermodynamic ensemble, and can be either undertaken or analyzed in terms of a small number of *collective coordinates* defined to capture the *essential dynamics* of a system. Integrating over the remaining degrees of freedom allows the definition of a free-energy profile, in which case the collective coordinate may be interpreted as an order parameter or reaction coordinate leading from one state of the system to another. Such coordinates are often designed to reflect large-amplitude, correlated movements in the system.

Remarkably, large amplitude - low frequency motions have been investigated at about the same time using three highly related mathematical tools, namely principal components analysis (PCA), multi dimensional scaling (MDS), and least square affine approximation. We briefly review these. Since atomic motions inherently encode correlations between atomic displacements, the covariance matrix of atomic positions is expected to play a key role [ALB93]. Diagonalizing this matrix yields the *principal components analysis* (PCA) of the data, namely the (mutually pairwise orthogonal) directions maximizing the variance [TPK02, KHM$^+$05]. Alternatively, the essential dynamics may be captured by finding a set of directions minimizing in the least square sense to the projections of conformations onto the corresponding affine space [GBHK97], a process actually also using the aforementioned principal directions. Finally, we note that PCA has a dual, known as Multidimensional Scaling (MDS), a quadratic optimization problem aiming at finding a low dimensional embedding best approximating the Gram matrix (the matrix of inner products) of conformations. The resulting low dimensional embedding[AA92, TB95] is identical to that provided by PCA, as is easily seen from a singular value decomposition of the covariance matrix and of the Gram matrix [LV07].

A limitation of PCA and MDS is that they are are linear methods. The nonlinear variant of MDS, called ISOMAP [dSLT00], has also been used [DMS$^+$06]. This strategy uses geodesic distances computed from a nearest neighbor graph rather than distances in the ambient space, whence its non linear nature [LV07]. However, despite improved results from this approach, letting pairwise distances drive the dimensionality reduction process overlooks the relative accessibility of the states linked. This motivated the introduction of dimensionality-reduction methods based on diffusion maps [RZMC11], which are especially relevant since such distances can also be derived for a system at equilibrium from the Fokker-Planck equation [NLCK06].

State based approaches. An alternative approach to studying dynamics consists of using state models. On the one hand, thermodynamic properties can be computed in terms of contributions from different local minima, e.g. using the harmonic superposition approximation (HSA) [WB06]. Adding knowledge of the stationary points connecting (usually grouped) minima allows the use of, e.g., transition-state theory to estimate kinetics for transitions in the configuration space of the system under study [CW08]. Thermodynamics and kinetic properties can be modeled directly using Markov state models [PBB10]. Such methods aim at modeling kinetically relevant states together with the transition rates between them. Once a matrix of transition rates between the states of interest has been built, transition state theory can be used to model the dynamics using the Master equation [BBK95, Kam92]. In this context, a pre-requisite consists of filling the transition matrix. For relatively simple systems, the coarse grain model may be obtained by clustering conformations upon projecting them in a lower dimensional space, using one of the aforementioned methods [NAKH14]. For more complex systems, databases of local minima and saddle points may be used [Wal02]. A database of critical points may have to be coarse

grained to assemble the microstates into macrostates, the inner relaxation within a macrostate being faster than the interaction with the surrounding macrostates [SPJ96, KSH98, PBB10]. Practically, an exhaustive computation of critical points may be circumvented using discrete path sampling [Wal02, DB09]. Using these techniques, a dynamical diagnosis can then be posed for typical disconnectivity graphs, thus bridging the gap between topography and dynamics [DWB05]. In materials sciences, such approaches have recently been proposed with the goal of designing optimal temperature schedules aiming at accessing a specified region of a landscape [HS13].

1.1.4 Contributions

While sampling and modeling thermodynamic and kinetic properties of a molecular system, questions arise concerning the appropriateness of the samplings for the desired purpose. While conventional MC is subject to such concerns as the relevance of the moveset in allowing for satisfactory exploration of the configurational space, MD simulations prompt questions concerning their ergodicity, which involves knowledge of the timescales required for local or global equilibration. Such questions may be addressed by quantitatively analyzing samplings and comparing results obtained using qualitatively different approaches, for example, MD and basin hopping.

Each of these classes of problems deals with large sets of conformations, in the context of the overall potential energy or free energy landscape. For convenience we also refer to these collections as "ensembles", while distinguishing them from an ensemble in the thermodynamic sense as required. From a methodological standpoint, we consequently address four related topics, namely structural analysis of a conformational ensemble, comparing two conformational ensembles, analyzing a sampled energy landscape, and comparing two sampled energy landscapes. These four methodological challenges are the focus of this paper.

We present a suite of approaches to analyzing and comparing sampled energy landscapes, and more generally conformational ensembles, and make the corresponding software available. These methods are coherent in the sense that they exploit a handful of constructions arising recently in computational geometry (nearest neighbor algorithms in metric spaces), computational topology (persistence theory), graph theory (shortest path algorithms), and optimization (a variant of the so-called earth mover distance). More specifically, our contributions are:

- *Analysis of a conformational ensemble.* In addition to various approaches aiming at assessing their structural diversity, we provide a novel hierarchical clustering method for conformational ensembles. The method assigns clusters to selected local maxima of the density estimated from the sampling, which are selected as the most significant ones using topological persistence.
  Associated software: program sbl-conf-ensemble-analysis.exe.

- *Comparison of two conformational ensembles.* We exploit recent development of search structures in metric spaces to compute distances of the Hausdorff type between conformational ensembles.
  Associated software: program sbl-conf-ensemble-comparison.exe.

- *Analysis of a sampled energy landscape.* We present a method colorblackto denoise a landscape using topological persistence (persistence diagrams), that is, to infer the distribution of barrier heights and smooth out the small ones when a clear-cut separation is observed – a property usually referred to as geometric frustration. We use the same algorithm to identify connections between basins, without resorting to optimization methods to find numerical approximations of saddle points. Finally, we propose novel analyses of transition

graphs, notably statistics on paths between selected minima (known as landmarks). Associated software: program sbl-landscape–analysis.exe.

- *Comparison of two sampled energy landscapes.* We provide two novel methods to compare two sampled landscapes once each landscape has been converted into a vertex-weighted transition graph, namely a graph connecting basins, the weight of a vertex in such a graph being an estimate of the size of the basin represented by this vertex. Such comparisons are of interest in various settings, e.g., to assess the coherence of two force fields for a given system (atomic, coarse grained), to compare two related systems (e.g. a wild type and mutant protein), or simply to compare simulations launched with different initial conditions. Associated software: program sbl-landscape-comparison.exe.

As just mentioned, these tools are generic and can be applied in a variety of settings, notably when the sampled conformations consist entirely of representatives of critical points (local minima and saddles), or not ("plain" samples). colorblackPractically, we put these tools into action on a simplified BLN protein model consisting of pseudo amino-acids of three types (hydrophobic (B), hydrophilic (L) and neutral (N)), each represented by a bead [HT90]. Selected such models, with appropriate sequences, are known to fold into a structure with a hydrophobic core, mimicking real proteins to a certain extent. Using this model, we use our methods to model the PEL of BLN69, and to assess its exploration. Note that the first scenario for this application is solely concerned with critical points, the second is carried out with plain samples.

1.2 Concepts and terminology

We consider a macromolecular system m involving s atoms, the ith-atom being denoted $m[i]$. The conformational space of the system is denoted $\mathcal{C}$, and its dimension d. The distance between two conformations is denoted $d_{\mathcal{C}}(\cdot,\cdot)$. We shall in general use the least root mean square deviation (lRMSD), colorblacknamely the square root of the average squared distance deviation in atom positions, minimized over rigid motions of the system. A *conformation* or *sample* refers to a conformation of the system. We note in passing that two conformations may be separated by relatively small distances, yet separated by large or even insurmountable barriers.

The potential energy of a conformation c is denoted $V(c)$. If the gradient of the potential energy vanishes at c, the conformation is called a *critical point*. Practically, we shall deal with two types of critical points: local minima, and index one saddles (saddles for short). If c is not a critical point, *quenching* c consists of (numerically) following the negative gradient $-\nabla V$ until a local minimum $q(c)$ is found. In this case, there exists an integral curve of the gradient vector field $-\nabla V$ joining c to $q(c)$, and one says that c *flows* to $q(c)$.

A *conformational ensemble*, also called *sampling* C is a set of n conformations, that is $C = \{c_1, \ldots, c_n\}$. The associated set of quenched conformations is denoted C_q, that is $C_q = \{q(c_1), \ldots, q(c_n)\}$. If Cartesian coordinates are used, the conformations in C can also be aligned on a reference conformation, say c_1. Aligning the i-th conformation onto c_1 results in the representation of that conformation denoted $\tilde{c}_i$, and the set of such conformations is denoted $\tilde{C}$. Once a one-to-one correspondence between the atoms of c_i and c_1 has been set, aligning c_i onto c_1 requires computing the rigid motion yielding the least root mean square deviation between c_1 and c_i.

A *nearest neighbor graph* (NNG) of conformations is a graph whose vertices are conformations, with edges joining selected pairs. That is, a NNG connects conformations in the configuration space $\mathcal{C}$ of the system studied. We use NNG in two guises. First, we build a NNG by connecting

a sample to its k nearest neighbors. Second, we build a NNG by connecting a sample to all samples of the ensemble within a given distance r. Practically, the distance used is the lRMSD.

A *height function* is a mapping from the conformational space $\mathcal{C}$ to the real numbers, and a conformational space equipped with such a function is called a *landscape*. Given a fixed elevation h, the portion of the landscape located below (resp. above) the elevation h is called a *sublevel set* (resp. *super level set*). Naturally, the landscape obtained when the height function is the potential energy is called the potential energy landscape (PEL). A *lifted sample* is a sample equipped with a real number, called its *height*. When this number represents the potential energy, a collection of such samples is called a *sampled energy landscape*. Using the connectivity of a NNG to connect lifted samples results in a *lifted NNG*. For example, if the height is the potential energy, the lifted NNG defines a network on the PEL.

Of special interest on a landscape are the local minima and the transition paths connecting them. In the smooth setting, a transition between two local minima corresponds to the existence of an integral curve joining the saddle to the minimum. More specifically, in Morse theory, these curves define the so-called *unstable manifold* of the saddle. Generically, two such curves are found for each saddle — note however that they may end up in the same local minimum, a situation we refer to as a *bump transition* (Fig. 4). We encode such transitions in a *transition graph* (TG), namely a graph whose nodes are minima and saddles, with one edge between the minimum m_1 and the saddle σ if there exists a direct transition path (m_1, σ, m_2) with m_2 another local minimum. We also define a *compressed transition graph* as a TG where the two edges emanating from a saddle are merged. That is, such a graph contains only minima and contains an edge between two of them if and only if these minima are connected through a saddle in the TG. As we shall see, such compressed graphs are useful for the comparison of landscapes.

2 Algorithms

2.1 Pre-requisites

2.1.1 Nearest-neighbor graphs

Rationale. We shall derive various parameters on conformational ensembles and sampled landscapes from proximity relationships between conformations, that is, from *nearest neighbor graphs* (NNG).

Algorithms. In computing a NNG, the demanding operation consists of retrieving the relevant nearest neighbors of a conformation (the k nearest neighbors or the samples within distance r). While this operation can always be performed by a linear scan over the ensemble, various data structures aiming at pruning the search exist [Sam06]. When the distance used between conformations is the plain Euclidean distance, spatial partitions can be used, so as to limit the search for the neighbors of a point to cells of the partition located nearby the point. One prominent such data structure is the spatial partition used in the *approximate nearest neighbors* algorithm [AMN$^+$98]. The situation is however more involved for searching neighbors in a general metric space, i.e. when the distance function used does not come with an embedding of the samples in a Euclidean space. This situation is precisely that faced when using the lRMSD. Fortunately, efficient algorithms also exist in this case, and we use *proximity forests* [OD13].

2.1.2 Topological analysis of height functions

Rationale. Consider a landscape. We present a method to analyze a lifted NNG on such a landscape, borrowing concepts and tools from Morse theory [Mil63, FK97] and persistence theory [EH10]. Our incentives are twofold. When the height function is an estimated density of states, the method yields a clustering algorithm generalizing that presented in [HPD$^+$01] (section 2.2.3). In that case, the clustering algorithm requires scanning the landscape downward with a horizontal plane, so as to discover when connected components of *super level sets* appear and merge. On the other hand, when the height function is the potential energy, the method is used to study the basins of a sampled PEL, in a manner related to disconnectivity graph (DG) [BK97, Wal03] (section 2.4.2). Studying merges between energy basins then requires scanning the landscape upwards with a horizontal plane, so as to discover when connected components of *sublevel sets* appear and merge [PBC10].

In fact, all three methods, namely ours and those from [HPD$^+$01] and [PBC10] exploit the same construction from Morse theory [Mil63], which consists of studying the evolution of sub (or super-) level sets of a height function, discretized using a nearest neighbor graph. But our approach goes beyond the previous two in two ways. First, we handle multiple scales in the barrier height distribution using topological persistence, which allows in particular handling full disconnectivity graphs, a difficulty mentioned in [PBC10]. Second, our method is general enough to exploit all the information available, including the knowledge of exact local minima obtained by quenching, as we shall see in section 2.4.2.

Persistence diagrams for super level sets. Local maxima of a height function are specific points, two properties of which are of crucial importance in the following development. The first property is the decomposition of a landscape into *attraction basins* induced by the maxima of a landscape. Indeed, the basin of a maximum is the loci of points of the landscape flowing to that maximum when following the gradient of the height function. To state the second property, recall the well known fact from Morse theory [Mil63], which states that the topology of super level sets changes precisely at critical points of the height function. For example, when sweeping the landscape downward with an horizontal plane, a connected component of super level sets is created when the plane hits a new maximum, and two such components merge when the plane hits a critical point of index $d-1$.

Practically, a difficulty consists of dealing with large numbers of local maxima, possibly non significant ones. We explain how to focus on the most prominent ones, borrowing the terminology from geography.

In topography, the *prominence* of a local maximum is the closest distance to the nearest local maximum with higher elevation, while its *culminance* is the least elevation drop to a saddle leading to a higher local maximum. (Note that the culminance is related to the *barrier height* used in biophysics.) Using these notions, the *persistence diagram* (PD) of super level sets is the 2D plot with one point for each local maximum (excepting the global one), defined as follows: the abscissa of the point is the elevation of the maximum, while its ordinate is the elevation of the saddle defining its culminance.

The following remarks are in order:

- By construction, the highest peak is not paired with any saddle. That is, if the height field has m local maxima, its PD contains $m-1$ points.
- A point of the PD corresponds to a pairing involving one maximum and one saddle. Mathematically, these points are critical points of the height field: the maximum is an index d c.p. (its Hessian has d negative eigenvalues), while the saddle is an index $d-1$ c.p. (its Hessian has $d-1$ negative eigenvalues).

- All points of the PD are below the diagonal $y = x$, and the distance from a point to the diagonal is a measure of the stability of the corresponding local maximum. For example, if the mountain range contains p well identified peaks, and q *tiny/spurious* peaks, then, the PD contains q points near the diagonal, and p points far away from the diagonal. It can also be shown that PD are stable when the underlying landscape incurs changes [CSEH05].
- Consider a case where a PD has two sets of clearly separated points, one set near the diagonal (non persistent max), and one set far away. If the former are not significant, their basins can be assigned to significant maxima. Intuitively, this operation consists of performing merges between these basins, using the adjacency encoded in the saddles joining them.

Persistence diagrams for sub-level sets. The analysis presented above for local maxima and super level sets applies directly for local minima and sublevel sets, yielding a persistence diagram whose points code local minima of the height field. The main changes are the following ones:

- By construction, the lowest local minimum is not paired with any saddle. That is, if the height field has m local minima, its PD contains $m - 1$ points.
- A point of the PD corresponds to a pairing involving one local minimum and one saddle. Mathematically, these points are critical points of the height field: the minimum is an index 0 c.p. (its Hessian has zero negative eigenvalues), while the saddle is an index one c.p. (its Hessian has one negative eigenvalue).
- All points of the PD are above the diagonal $y = x$, and the distance from a point to the diagonal is a measure of the stability of the corresponding local minimum.
- A PD can also be simplified by re-assigning the basins of non significant minima. That is, if the point of the PD are clearly separated, those near the diagonal can be removed by merging each such basin into the deeper basin it is connected to. Note that this simplification removes the least persistent basins first, as opposed to merging the saddles in a given energy slice. Note also that this simplification cancels one (saddle, minimum) pair at a time.

This notion is illustrated on Fig. 1 for a 2D elevation function used in optimization benchmarks, known as the Himmelblau function. On this example, the clear separation of the points of the PD of noisy Himmelblau call for its simplification. Upon performing the aforementioned merges, one is left with a PD identical to that of the noiseless function, i.e. involving three points only.

colorblackPractically, exploiting a PD typically requires two parameters: a maximum energy of interest $E_{\max}$, and a persistence threshold ΔE. Upon setting these values, the PD ends up partitioned into five regions defined by three lines, so that a local minimum m gets qualified with respect to three criteria (Fig. 2):

- Selected/rejected: m is selected provided that its birth date occurs before $E_{\max}$.
- Persistent/canceled: m is persistent if its persistence is larger than a user defined threshold ΔE.
- Filtered/un-filtered: the basin of m is filtered if the death date of m is larger than the threshold $E_{\max}$.

The possible combinations, also illustrated on Fig. 2, are:

- $m \in R_1$: rejected. Such a local minimum is rejected, since its energy is larger than $E_{\max}$.
- $m \in R_2$: selected / canceled / un-filtered. Such a local minimum is selected, yet canceled by persistence. Because m dies before $E_{\max}$, all conformations found in its basin shall be part of another basin.
- $m \in R_3$: selected / canceled / filtered. A local minimum which is selected, yet canceled by persistence. However, because m dies after $E_{\max}$, only the portion of its basin below $E_{\max}$ shall be found in another basin.
- $m \in R_4$: selected / persistent / un-filtered. Such a local minimum is selected, and is not canceled by persistence. Because m dies before $E_{\max}$, all the samples found in its basin are considered.
- $m \in R_5$: selected / persistent / filtered. This combination is similar to the previous case, except that samples whose energy higher than $E_{\max}$ are discarded. Note that the cluster associated with the global minimum belongs to this region even though it is not found on the PD, as the global minimum never dies.

Algorithm. colorblackWe now sketch algorithm `PLeSE`, for *Persistent Level Set Extraction*. Consider a discrete set of lifted points on a landscape, and assume that we wish to analyze its minima and sublevel sets — the method is valid mutatis mutandis for maxima. The algorithm, which borrows upon [CadOS11] and [CCS11], (i) identifies samples playing the role of minima and saddles and computes the pairings defining the persistence diagram, (ii) assigns basins to minima by performing a *discrete* quench, and (iii) offers the possibility to cancel non significant minima (the samples associated with basins canceled get transferred into the basins of minima which survive.

Denote $\{(c_i, h_i)\}$ the lifted samples processed, with h_i the elevation of c_i. Also assume that a lifted NNG connecting the samples has been built, and that the negative gradient of the height function has been estimated at each sample, e.g. by selecting the edge $\mathbf{c_i c_j}$ maximizing the ratio $(h_i - h_j)/d_{\mathcal{C}}(c_i, c_j)$.

Using this graph, we process samples by increasing height. colorblackDefine the *lower star* of a sample as the subset of its neighbors in the NNG, with a smaller elevation. A sample whose lower star is empty defines a local minimum. If a sample is not a local minimum, we perform a *discrete quench* using the NNG, that is, we iteratively follow the estimated gradient to end up at a local minimum, and assign it to the basin of that minimum. colorblackOnce the discrete quench has been done, we check whether this sample c_i has a neighbor c_j in its lower star, flowing to a different local minimum. If so, the sample c_i termed a *transition* or a *pseudo-saddle*, and the pair $(q(c_i), c_i)$ defines a point of the persistence diagram. Note that c_i is termed a pseudo-saddle since it may actually be far from the real saddle while located in the neighborhood of its stable manifold (Fig. 3). Despite this limitation and since we are only dealing with conformations provided by sampling methods, as opposed to numerical methods tracking saddle points [HJJ02], we will call such samples saddles.

This algorithm is used in section 2.2.3 for clustering and in section 2.4.2 to build DG.

Remark 1 *colorblackThe previous algorithm can be implemented in two ways, depending on whether or not one wishes to run a single query (setting the thresholds E_{max} and ΔE once for all), or multiple queries (one varies the two thresholds).*

In the former case, the one-pass algorithm developed in the context of clustering suffices [CadOS11]. This algorithm works on samples (conformations), and performs the assignment of samples to basins and the persistence based simplification on the fly.

In the latter case, one should use the simplification algorithm proposed in [CCS11]. Once the samples have been assigned to basins, the algorithm simplifies the transition graph in a recursive manner, so that sequence of increasing ΔE thresholds can be accommodated. Furthermore, for a given simplification threshold, multiple sublevel set queries can be carried out, with a single query performing the selection of relevant sample from the critical points rather than from the samples themselves, for the basins whose death date is less than the prescribed threshold E_{max}. Phrased differently, the algorithm from [CCS11] is well suited to handle landscapes involving a large number of critical points and/or conformations. For example, the one dealt with in section 3 contains hundreds of thousands.

Remark 2 *colorblackA method for regrouping states separated by free energy barriers below a specified threshold was presented in [CW08, Section 2.2]. Assume that one knows how to compute the free energy of two groups of local minima (e.g., using harmonic superposition), as well as that of the inter-group transition state ensemble. The method consists of iteratively performing binary merge if the forward and backward free energy barriers are less than a user specified threshold.*

The recursive simplification scheme of the transition graph sketched above and detailed in [CCS11] is different in two respects. First, in [CCS11], the ordering in which simplifications are done is imposed by the height of barriers taken in increasing order – as opposed to performing simplifications bottom-up as in [CW08, Section 2.2]. Note that in both cases, the simplification halts when a prescribed barrier height is reached. Second, in [CCS11], the simplification criterion is asymmetric since while processing a saddle, only the smallest height drop plays a role. This is related to the fact that in Morse theory, the focus is on sublevel sets of the function studied. Instead, forward and backward barriers are used in [CW08, Section 2.2].

Remark 3 *colorblackTechnically, Morse theory and Morse homology are concerned with smooth function defined on manifolds, that is, singularities are excluded. Practically, such singularities could be present – for example a Lennard-Jones potential is singular at the origin. However, as we deal with finite energy samples, we build discrete structures compatible with the hypotheses required to apply Morse theory. In particular, since we deal with 0 and 1 homology (i.e. connected components associated with basins, and cycles defined by transitions), we only require the boundary operator of a one-dimensional cell complex, which is merely a graph in our case.*

2.2 Conformational ensembles: analysis

In the sequel, we provide various statistics to assess the diversity of a conformational ensemble generated by a sampling algorithm.

2.2.1 Sampling diversity

Rationale. Currently we use two measures for the sampling diversity, in order to assess the extent to which a molecule deforms within an ensemble C. The first is the standard method for estimating the root-mean squared atom fluctuations ($RMSF$). Denoting $\overline{\tilde{c}[k]}$ the average position of the k-th atom in the aligned samples $\tilde{C}$. The RMSF for atom k is defined as

$$RMSF_k = \sqrt{\frac{1}{n} \sum_{i=1,\ldots,n} \left\| \tilde{c}_i[k] \; - \; \overline{\tilde{c}[k]} \right\|^2}.$$

The second measure is the bounding box (in Cartesian coordinates) of the ensemble $\tilde{C}$ of aligned structures.

Algorithms. Computing these measures requires performing a registration of each conformation into a unique coordinate system. Using the first conformation c_1 as reference, one can transform every other conformation by applying to it the rigid transform defining the lRMSD between c_i and c_1.

2.2.2 Sampling sparsity via spanning trees

Rationale. A conformational ensemble may contain *clusters*, i.e. groups of conformations such that pairwise distances within a cluster are smaller than distances between conformations across clusters. We investigate such properties using graphs.

Assume that a connected NNG G has been built over the samples, and denote $V[G]$ (resp. $E[G]$) its vertex set (resp. its edge set). Assume that to each edge $e = \{c_i, c_j\} \in E[G]$ is attached the quantity $d_{\mathcal{C}}(c_i, c_j)$. we compute a minimum spanning tree MST(G) of G. If the conformational ensemble contains n conformations, this tree involves $n-1$ edges. For an edge $e = \{c_i, c_j\}$ of this graph, the *edge length* is the distance between the conformations c_i and c_j, that is $d_{\mathcal{C}}(c_i, c_j)$. Denoting E the edge set of the MST, and $e = \{c_i, c_j\}$ a particular edge, we then report the following statistical summary for G:

$$\min_{e \in E} d_{\mathcal{C}}(c_i, c_j), \operatorname*{median}_{e \in E} d_{\mathcal{C}}(c_i, c_j), \max_{e \in E} d_{\mathcal{C}}(c_i, c_j). \tag{1}$$

To capture the significance of this summary, consider the situation where the ensemble has a cluster structure. In that case, most of the edges are short ones, only those connecting clusters being long ones, which reads plainly from the statistical summary.

Algorithms. Extracting the MST out of a connected graph is a classical problem in computer science. We use Prim's algorithm, which iteratively attaches one node not connected yet, namely that using the shortest edge available which does not create a cycle.

2.2.3 Persistence based clustering

Rationale. When an ensemble features clusters, as e.g. seen from the statistical summary from the edge lengths found in a MST, the next step consists of finding these clusters. Upon estimating the sample density at each sample from the ensemble, a three stage strategy consists of associating one cluster to each local maximum of this density [Che95, HPD+01], and to filter out spurious (i.e. small) ones. We now briefly review these three steps.

The first step is the density estimation step. Assume that a NNG has been built, so that each sample is linked to a number of its nearest neighbors, say its k nearest neighbors. Intuitively, the density about a sample is inversely proportional to the distances to the nearest neighbors. Formally, denoting k the number of nearest neighbors and V_d the volume of the unit ball in $\mathbb{R}^d$, the local density at sample c_i can be estimated as [BCCS+11]:

$$\hat{f}_n(c) = \frac{1}{n\, V_d} \left(\frac{\sum_{j=1}^{k_n} j^{2/d}}{\sum_{j=1}^{k_n} (d_{\mathcal{C}}(c, n^j(c)))^2} \right)^{d/2} \tag{2}$$

For the second step, consider the lifted NNG obtained by endowing each sample with the previous estimated density. Since each cluster is associated with a local maximum, we resort to

the PLeSE algorithm from section 2.1 to identify the samples which are local maxima, and to assign the remaining samples to their attraction basins.

Finally, the third step consists of filtering out the spurious local maxima using topological persistence, as also explained in section 2.1.

Remark 4 *Due to the high dimensionality of configuration spaces of molecular systems, the estimate of Eq. (2) is typically small. For such cases, PD are best plotted in log-log scale.*

2.3 Conformational ensembles: comparisons

Having presented methods to assess one conformational ensemble, we now address the problem of comparing two such ensembles.

2.3.1 Hausdorff distance between ensembles

Rationale. A standard problem in designing sampling algorithms consists of assessing the ability of an algorithm to detect local minima of interest, or points nearby such minima.

To perform such an assessment, consider two sets of conformations, a reference set R (typically local minima), and an ensemble C produced by a sampling algorithm. To assess the coverage of the set R by C, we compute for every sample $r \in R$ its nearest neighbor in C:

$$\forall r \in R : d(r, C) = \min_{c \in C} d_{\mathcal{C}}(r, c). \tag{3}$$

We then report the following triple:

$$\min_{r \in R} d(r, R), \text{median}_{r \in R} d(r, R), \max_{r \in R} d(r, R). \tag{4}$$

Note that the third entry in Eq. (4) is the so-called one-sided Hausdorff distance between R and C – the sample from R which is the most isolated from C.

Algorithms. To find the nearest neighbors of a sample, we use the data structures described in section 2.1.1.

colorblack

2.3.2 Minimum spanning forests

Rationale. The Hausdorff distance being a max min type criterion, we also consider a different criterion, involving at least one distance for each conformation from C and R.

We do so by computing a minimal spanning forest of a weighted complete bipartite graph induced by the two sets C and R. More precisely, the nodes represent the conformations and there is an edge between every node representing a conformation $c \in C$ and every node representing a conformation $r \in R$. The weight of an edge is the distance $d_{\mathcal{C}}(r, c)$. The problem then consists of computing a minimum weight subset of edges such that every node of the bipartite graph is incident to at least one edge of this subset. This means that every conformation is associated with at least another one.

Algorithms. It is easily shown that an optimal subset of edges induces a forest; i.e., a subgraph without cycles. A simple algorithm consists of choosing, for every node, the incident edge with the smallest weight. This strategy corresponds to the first step of Borůvka's algorithm for computing a minimum spanning tree in a graph [NMN01]. (See also http://en.wikipedia.org/wiki/Bor%C5%AFvka%27s_algorithm.) Note that if the weights are not all different, a correction step is necessary in order to avoid cycles.

2.4 Energy landscapes and transition graphs: analysis

2.4.1 Transition graphs and disconnectivity graphs from critical points

Rationale. When a database of connected critical points (local minima and index one saddles) is available, a transition graph is readily built. A powerful representation of the energy landscape encoded in the transition graph is in terms of *disconnectivity graph*. As discussed in section 2.1, persistence diagrams provide a quantitative alternative that also allows simplifications.

Algorithms. The persistence diagram of the sub-level sets of the landscape is computed from the transition graph as explained in section 2.1.2. In particular, this PD gives the persistences of all local minima.

2.4.2 Transition graphs and disconnectivity graphs from a conformational ensemble

Rationale. Consider now the case where a plain sampling has been obtained, together with the local minimum associated with each sample (obtained by quenching) [SW82, NP06]. Let us also assume that the transitions themselves have not been computed. In the sequel, we therefore provide a variation on algorithm `PLeSE`, using samples and the associated local minima only, and computing a transition graph as well as the associate DG and PD. Note in particular that this algorithm does not involve any transition path sampling algorithm.

Algorithms. Assume that one is given a conformational ensemble C together with the corresponding quenched points C_q. We build a NNG on the conformations $C \cup C_q$ as explained in section 2.1.1, and in addition endow each sample c_i with an edge to $q(c_i)$. Note that this connection directly links each sample to its local minimum, thus avoiding having to perform the discrete quench operation using estimated gradients as explained in section 2.1.2. In running the algorithm `PLeSE` on this graph, selected samples are detected as transitions. colorblackMore precisely, algorithm `PLeSE` identifies pairs of samples (c_i, c_j), such that (i) c_j belongs to the lower star of c_i (recall that samples are processed by increasing height), and (ii) $q(c_i) \neq q(c_j)$. We also say that these samples witness a *bifurcation*. When such a pair of samples is found, we record the path from $q(c_i)$ to $q(c_j)$ through the samples c_i and c_j. The collection of all such paths defines a transition graph between the vertices from the set C_q.

Remark 5 *The following remarks are in order: colorblack*

- *The bifurcations, as defined above, may involve more than two local minima. This is e.g. the case if a sample has in its lower star two or more samples flowing to distinct minima. This situation may occur in different settings: (i) in case of coarse sampling, the connected samples are located in the basins of several minima, (ii) the samples are located in the neighborhood of a critical point of index > 1, or (iii) they are located in the neighborhood of a degenerate critical point (a so-called monkey saddle). Note that our algorithm, purely based on sampling, does not distinguish between these situations.*
- *The saddles detected may not be close to the real saddles (i.e., critical points of index one of the PEL), but they nevertheless indicate transitions between two basins associated with $q(c_i)$ and $q(c_j)$ (Fig. 3).*
- *Since the* `PLeSE` *algorithm also detects local minima, it is tempting to apply the previous method to the set C directly, i.e. without using the local minima in C_q. However,* `PLeSE` *identifies a point as a minimum if its connected neighbors are all at higher energies. If two*

(or more) such minima are in the same basin of the PEL, all but one are spurious. This situation arises frequently in high-dimensional spaces due to sparse sampling.

Remark 6 *colorblackThe previous analysis can be used to detect entropic bottlenecks. To see how, consider the situation where two marginally rugged plains denoted A and B are connected through a limited region S of configuration space of similar energy and rugosity – which corresponds to an entropic bottleneck.*

For such a task, persistence can be used to quantify the ruggedness of samples lying in a given region $A \cup S \cup B$, and isolate this region from the rest of the sampling. (Note that such a region is a connected component of a sublevel set of the PEL, for an appropriate energy value.) A graph cut technique (e.g., based on spectral clustering) may then be applied to find a balanced cut isolating A and B across S.

2.4.3 Disconnectivity graphs: morphology

The generic situation expected in a DG is that of a binary merge, namely a saddle is linked to two local minima. (Degenerate situations may occur, though: for example, a *monkey saddle* is linked to three local minima.) That is a generic DG is a binary tree: each internal node representing a saddle is associated with at most two children, and each leaf represents a local minimum.

The morphology of a complete binary tree can vary between that of a *path* (in which one child of each internal node is a leaf and the other is an internal node) and that of a perfectly balanced tree (in which each internal node has two children i.e. subtrees of the same size). To assess the structure of such trees, we use the so-called external path length (EPL) [Mah92]. Defining the *depth* of a node as the number of edges required to reach it from the root, the EPL is the sum of depths of the leaves of the tree. If n is the number of internal nodes (index one saddles in our case), the EPL when the tree is a path is $\binom{n}{2} + 2n$, which we will write using the shorthand EPL_{path}, while that of a perfectly balanced tree is $n \log_2 n$. colorblackSince the latter, symmetrical tree is not a realistic reference, we use instead the EPL of a random binary search tree [Mah92], which is asymptotically equivalent to $\text{EPL}_{rand.} \sim 2n \ln n$.

2.4.4 Transition graphs: topology and Betti Numbers

Rationale. The exploration of a complex energy landscape typically requires several runs of an exploration algorithm, with different starting points and parameters. This strategy results in a database of local minima and transitions between them. Connecting these critical points results in a transition graph, which may be connected or not. We report topological and geometric information characterizing this graph.

Topological information. For a given graph, the number of connected components and the number of cycles in this graph are known as the first and second Betti numbers [EH10], and are respectively denoted β_0 and β_1. We report them for the transition graph, and we also compute statistics on a per connected component basis.

The presence of cycles in a transition graph is especially interesting, since such cycles correspond to different paths connecting the same local minima. Cycles involving two edges are also of interest, as they may arise in two situations. The first corresponds to the situation in which a loop around a saddle is anchored in a single local minimum (Fig. 4). The second is due to post-sampling processing techniques that may be used to group minima. For instance, in the BLN model, enantiomeric left and right handed helices may be considered as identical in certain contexts; identifying them as such results in the creation of a cycle.

Geometric information. The simplest geometric information embodied in a transition graph is the length of its edges, and can be assessed with the statistical summary of Eq. (1).

In particular, the presence of long edges hints at challenging landscapes: even when the $d_{\mathcal{C}}$ between all pairs of local minima is small, one may have to follow a long valley to reach a saddle connecting two minima.

Algorithms. Computing the first and second Betti number of a graph respectively require running a traversal of the graph and a union-find algorithm [EH10, CLRSon].

2.4.5 Transition graphs: paths

Rationale. Consider a transition graph, and assume that selected minima called *landmarks* have been singled out. For example, one may select the lowest local minima or the most persistent ones (section 2.1).

For a given landmark, the distribution of distance from that landmark to its connected saddles provides an estimate on the size of the basin of that landmark [DM05]. Also, a large value for the smallest such distance provides information on the distance to be traveled to escape from the basin of that minimum. Similarly, the relative position of all landmarks is assessed by the statistical summary of all pairwise distances.

Algorithms. Computing the distances between landmarks can be done using a variant of Dijkstra's shortest path algorithm, where one walks on the graph provided, but records and relaxes distances between landmarks only [CLRSon].

2.4.6 Transition graphs: embeddings

Rationale. To thoroughly explore an energy landscape, one usually launches a series of explorations with different starting points. These explorations typically produce new local minima which are stored in a database, and transition path sampling algorithms can be used to connect them. To launch subsequent explorations, it is useful to visualize the relative positions of the (significant) minima discovered.

Algorithms. Assume that one has selected a set of p local minima, say the p lowest ones. To embed them in 2D or 3D space in order to visualize their relative positions, we use dimensionality reduction algorithms such as Multi-dimensional scaling (MDS) [LV07]. We do so in two guises. In the first setting, we compute the symmetric matrix of all pairwise distances between all pairs of minima, using the metric $d_{\mathcal{C}}$ from the conformational space. Then, MDS is called to produce a 2D or 3D embedding best respecting these distances. However, the distance $d_{\mathcal{C}}$ ignores the topography of the landscape. In the second setting, we therefore replace it by the distance of the shortest path computed on a TG connecting local minima and saddles. More precisely, a path connecting two local minima consists of a sequence of connections *minimum - saddle* and *saddle - minimum*. We add up the length (computing using the distance $d_{\mathcal{C}}$) of all such edges, and select the minimum length obtained. We call this distance the *cumulative edge distance* d_{ced}. Note that computing these lengths is done with the variant of Dijsktra's algorithm mentioned in the previous section.

2.5 Energy landscapes and transition graphs: comparison

PEL and vertex weighted TG. colorblackThe importance of recognizing archetypal energy landscapes has long been acknowledged [WMW98], in particular in the context of assessing the

dynamics of the systems studied [DWB05]. More generally, the problem of comparing two PEL arises in various context, e.g. to evaluate the (lack of) coherence between two force fields, to compare the properties of two similar molecules or clusters, or to check whether two runs of an exploration method visited the same regions of the conformational space. To clarify things, by *comparing PEL*, we actually refer to the problem of comparing two *vertex weighted compressed transition graphs*, namely a graph connecting local minima across index one saddles, each vertex being in addition endowed with the probability of its basin (see section 1.2 for compressed transition graphs). In general, obtaining such a graph for the full landscape is beyond reach. Instead, one deals with a graph associated with a database of critical points, or inferred from a conformational ensemble. In both cases, the graph processed is a subgraph of the vertex weighted compressed transition graph of the whole landscape. For the sake of conciseness, we still refer to the problem of comparing two such graphs as a *PEL comparison*.

In designing PEL comparisons, two categories of criteria are of interest. The first one deals with features of the basins, namely the exact local minimum associated with each basin, and the probability of that basin. The second deals with transitions between significant basins.

In the following we provide two comparison methods: the first one deals with features of the basins only, while the second one additionally exploits the information on transitions.

2.5.1 Earth mover distance

Rationale. Consider the basin B of a local minimum. Abusing notations, the footprint of the basin in the conformational space $\mathcal{C}$ is also denoted B. Denoting k_B is the Boltzmann constant and Z the partition function, the probability of the basin is given by integrating the Boltzmann factor over the basin region, namely:

$$w(B) = \frac{1}{Z} \int_B \exp \frac{-V(c)}{k_B T} dc. \tag{5}$$

If minima are found by optimization (e.g., basin hopping), the $w(B)$ can be estimated using the eigenvalues of the Hessian (curvature) matrix evaluated at those points [Wal03]. If, on the other hand, the samples are obtained from a thermodynamic ensemble, as is typically the case for molecular dynamics and Monte Carlo procedures, Boltzmann weighting is automatically satisfied; the basin weight can correspondingly be estimated from the number of points falling in the basin region

$$w(B) = \frac{n_B}{n}. \tag{6}$$

Consider now two landscapes, called the source landscape PEL_s and the demand landscape PEL_d for the sake of exposure, respectively involving n_s and n_d basins. colorblackThe minimum associated with the ith source basin B_i is denoted s_i, and that associated with the jth destination basin B_j is denoted d_j. We also assume that the weights of the basins have been computed. Denoting these $w_i^{(s)}$ and $w_j^{(d)}$ for a source and demand basin respectively, we define the sum of weights $W_s = \sum_i w_i^{(s)}$ and $W_d = \sum_j w_j^{(d)}$. Finally, we also assume that a metric $d_{\mathcal{C}}$ is available to compare two conformations, so that the distance between the two aforementioned local minima is $d_{\mathcal{C}}(s_i, d_j)$.

To compare the two landscapes, we use the *earth mover distance* [RTG00], which is a particular case of mass transportation [Vil03]. Intuitively, the technique fills basins in the target (aka demand) landscape using mass from basins of the source landscape. A basin from PEL_s can be

split into several parts, and equivalently, a basin from PEL_d can be filled from several basins from PEL_s (Fig. 5). Denote f_{ij} the mass from $B_i \in \mathrm{PEL}_s$ moved into $B_j \in \mathrm{PEL}_d$. The cost of moving f_{ij} units of mass depends linearly on the distances between the minima s_i associated with B_i and d_j associated with B_j. A *transport plan* from PEL_s to PEL_d is defined by triples (s_i, d_j, f_{ij}), with $f_{ij} > 0$. Note that there are at most $n_s \times n_d$ such triples.

Finding the optimal i.e. least cost transport plan amounts to solving the following linear program (LP):

$$LP \begin{cases} \mathrm{Min} \sum_{i=1,\ldots,n_s, j=1,\ldots,n_d} f_{ij} \times d_{\mathcal{C}}(s_i, d_j) & \\ \sum_{i=1,\ldots,n_s} f_{ij} = w_j^{(d)} & \forall j \in 1,\ldots,n_d, \\ \sum_{j=1,\ldots,n_d} f_{ij} \leq w_i^{(s)} & \forall i \in 1,\ldots,n_s, \\ f_{ij} \geq 0 & \forall i \in 1,\ldots,n_s, \forall j \in 1,\ldots,n_d. \end{cases} \tag{7}$$

The first equation is the linear functional to be minimized, while the remaining ones define linear constraints. In particular, the second one expresses the fact that every basin from PEL_d need to be filled, while the third one indicates that a basin from PEL_s cannot provide more than it contains. colorblack(To simplify matters, we have assumed that $W_s \geq W_d$. Handling the case $W_s < W_d$ poses no difficulty, and the reader is referred to [RTG00].)

Based on this linear program, we introduce the *total number of edges*, the *total flow*, the *total cost*, and their ratio, known as the *earth mover distance* [RTG00]:

$$\begin{cases} M_{\mathrm{EMD}} = \sum_{i,j|f_{ij}>0} 1, \\ F_{\mathrm{EMD}} = \sum_{i,j} f_{ij}, \\ C_{\mathrm{EMD}} = \sum_{i,j} f_{ij} d_{\mathcal{C}}(s_i, d_j), \\ d_{\mathrm{EMD}} = C_{\mathrm{EMD}} / F_{\mathrm{EMD}}. \end{cases} \tag{8}$$

Algorithms. Consider a sampling C and the corresponding quenched samples C_q. To compute the weight of a basin associated to a local minimum $m \in C_q$, we use the assignment of the samples from C into the basins of local minima as done by algorithm `PLeSE` (section 2.4.2). The weight of a basin is then obtained using Eqs. (5) and (6).

Once the weights of basins have been computed, solving the linear program of Eq. (7) has polynomial complexity [Kar84]. Practically, various solvers can be used, e.g. the one from the Computational Geometry Algorithms Library [cga], `lp_solve` [1], the `CPLEX` solver from IBM [2], etc. In the following, the algorithm solving the linear program of Eq. (7) is called `Alg-EMD-LP`.

2.5.2 Earth mover distance with connectivity constraints

Rationale. The previous comparison ignores transitions between local minima. To take these connections into account, we modify the method by imposing connectivity constraints to transport plans. To see whether a transport plan is valid, pick *any* connected subgraph S from PEL_s — that is S connects selected local minima in PEL_s. Let D be the set of vertices from PEL_d such that for each vertex d_j in D, there is at least one edge emanating from a vertex s_i of the subgraph S with $f_{ij} > 0$. The transport plan is called valid iff the subgraph D is also connected. Intuitively, this constraint reads as follows: any connected set of basins from PEL_s can only

[1] `http://sourceforge.net/projects/lpsolve/`

[2] `http://www-304.ibm.com/ibm/university/academic/pub/page/ban_ilog_programming`

send mass to a connected set of basins from PEL_d (Fig. 6). In doing so, one compares the two landscapes while preserving connectivity constraints between basins.

It should be noticed that the connectivity preservation may prevent from fully satisfy the demand.

Algorithms. Finding transport plans respecting connectivity constraints turns out to be a hard combinatorial problem. The problem is not in APX, which means that if $\mathbf{P} \neq \mathbf{NP}$ holds, then, no polynomial algorithm with constant approximation factor exist. However, a greedy algorithm providing admissible solutions respecting connectivity constraints, denoted `Alg-EMD-CCC-G` for *earth mover distance with cost and connectivity constraints*, has been reported in [CM15]. colorblackUsing this algorithm, in a manner analogous to Eq. (8), we define the *total number of edges*, the *total flow*, the *total cost*, and their ratio:

$$\begin{cases} M_{\text{Alg-EMD-CCC-G}} = \sum_{i,j|f_{ij}>0} 1, \\ F_{\text{Alg-EMD-CCC-G}} = \sum_{i,j} f_{ij}, \\ C_{\text{Alg-EMD-CCC-G}} = \sum_{i,j} f_{ij} d_{\mathcal{C}}(s_i, d_j), \\ d_{\text{Alg-EMD-CCC-G}} = C_{\text{Alg-EMD-CCC-G}} / F_{\text{Alg-EMD-CCC-G}}. \end{cases} \tag{9}$$

Remark 7 *It can be shown that the Earth Mover Distance as defined by Eq. (7) yields a metric, provided that $d_{\mathcal{C}}$ is itself a metric, and that the sum of weights for the source and the demand graphs are equal [RTG00]. On the other hand, the EMD with connectivity constraints fails to satisfy the triangle inequality [CM15]. colorblackThe EMD with connectivity constraints is not symmetric either, but is easily made so in taking a symmetric function of the one sided quantities. To assess this lack of symmetry, we introduce the following ratios, respectively geared towards the flow and the cost:*

$$\textit{colorblack} \begin{cases} r^F_{sym} = \dfrac{\min(F_{\textit{Alg-EMD-CCC-G}}(A,B), F_{\textit{Alg-EMD-CCC-G}}(B,A))}{\max(F_{\textit{Alg-EMD-CCC-G}}(A,B), F_{\textit{Alg-EMD-CCC-G}}(B,A))}, \\ \\ r^C_{sym} = \dfrac{\min(C_{\textit{Alg-EMD-CCC-G}}(A,B), C_{\textit{Alg-EMD-CCC-G}}(B,A))}{\max(C_{\textit{Alg-EMD-CCC-G}}(A,B), C_{\textit{Alg-EMD-CCC-G}}(B,A))}. \end{cases} \tag{10}$$

3 Results

In this section, we present the test system used, and apply the tools presented above to investigate two scenarios, namely *modeling a landscape*, and *comparing samplings*.

3.1 System used: BLN models and datasets

BLN69 model. We use a 69 residue BLN model protein [BFHG03], whose landscape has been extensively sampled [Wal03, OWJ11]. The BLN model represents each protein residue as one of 3 types of beads, namely hydrophobic(B), hydrophilic(L) and neutral(N). The potential energy for the model is recalled in the supplemental section 6.1. In comparing two BLN69 structures, we extend the least root mean square deviation (lRMSD) by an additional test accomodating chirality of the polymeric structure [OJW12]. That is, when comparing two conformations c_i and c_j, two distances are computed, between c_i and c_j, and between c_i and the enantiomer of c_j; if the latter is smaller, the coordinates of the enantiomer are substituted for the originals.

The fast folding of proteins is usually interpreted in terms of minimal frustration [BW87, BOSW95, OLSW97], and frustration models have been developed both at the local level [TKSW13] and at the whole PEL level [DWB05]. The BLN69 model has a frustrated potential energy surface, with several deep basins close in energy to the global minimum ($V = -105.19$ in reduced energy units) but separated from it by high barriers. Of the top ten lowest energy structures, four lie close to the global minimum, differing only by changes in the turn regions, whereas six are found with different arrangements of the β-strands and are separated from the others by high energy barriers, whence the frustration.

Earlier work by [OWJ11] resulted in a database of 458,082 minima and 378,913 index 1 critical points (saddles) which was furnished to us by the authors, and from which we define:

- BLN69-all: the full database of critical points, obtaining using a combination of basin hopping and path sampling runs.
- BLN69-E_{-100}: the subset BLN69-all featuring all local minima whose energy is less than 100 energy units.
- BLN69-top10: the 10 lowest minima from BLN69-all.

We note in passing that when referring to local minima of the BLN69 model, we use the indices of critical points from BLN69-all.

Generation of conformational ensembles. We also generated conformational ensembles using the transition rapidly exploring growing random tree (`T-RRT`) algorithm, which has been shown to be efficient at sampling energy landscapes [JCPC11a]. We also subsequently quenched these samples. The incentive for using `T-RRT`, which refines the *rapidly growing random trees* exploration method [KL00], is that it uses an extension strategy known as the Voronoi bias, which avoids regions already visited by those that are yet unexplored. More precisely, `T-RRT` builds a tree exploring the conformational space. At each step, one tree node gets extended, each node being chosen with a probability equal to the volume of its Voronoi cell in the Voronoi diagram of the ensemble generated so far. The stepsize and temperature are adjusted throughout in order to facilitate sampling higher energy regions.

Using this algorithm, we launched an exploration generating 10^4 samples for each local minimum from the dataset BLN69-top10 resulting in ten corresponding samplings denoted TRRT-top10-1, -2, etc.. The parameters used in each run were the extension stepsize $\delta = 0.5$ in euclidean distance, the initial reduced temperature $T_{init} = 2$, and the temperature multiplication factor $\lambda = 1.001$. The stepsize δ was chosen based on the observation of the size of the bounding box of BLN69-all. These values were chosen empirically to favor access to lower lying regions of the PEL, and reflect the fact that `T-RRT` itself is not a quench-based method.

When building the NNG from the conformational ensemble, we used the lRMSD distance, with the chirality check discussed above to identify enantiomers of the BLN molecule that were closer in terms of lRMSD.

Finally, to compute the weight of basins associated to such samplings, we used Eq. (6) even though `T-RRT` method itself employs non-thermodynamic sampling in exploring higher-energy regions of the PEL. To challenge the comparison algorithms, we also created a second dataset as follows: for each run of TRRT-top10, we retained the local minima, but assigned a uniform mass to each basin. (That is, a basin of a landscape with p local minima gets a mass of $1/p$.) The ten corresponding datasets are denoted TRRT-top10-U-1, -2, etc, with U indicating *uniform*.

3.2 Modeling a landscape

Topology of transition graphs. We first analyze the overall structure of the transition graph encoded in the database BLN69-all. On this graph, whose nodes are the critical points (minima and saddles), the numbers of connected components (c.c.) and cycles are respectively $\beta_0 = 138010$ and $\beta_1 = 58821$. However, one prominent c.c. involves 163480 vertices and features $\beta_1 = 58039$ cycles. In the following, we focus on the analysis of this giant component. On this giant component, one observes 590 bump transitions (out of 970 in the whole dataset) (Fig. 4). One also finds minima with up to 6 bump transitions. In principle, the presence of possibly multiple bump transitions would hamper sequential exploration algorithms such as basin hopping by reducing the possibility for escape away from such basins[3]; for the same barrier crossing probability a larger step would be required to flow into a different minimum.

Binary and k-ary DG. The complete DG associated with the dataset BLN69-all shown in Fig. 7 is a binary tree (section 2.4.3). This differs from DG representations most often used in energy landscape studies, which group several transitions from a given energy slice into a single node to increase visibility, resulting in a k-ary tree[pel]. Zooming in on the region of the complete DG near the global minimum reveals its binary structure (Fig. 7, left-hand side).

We analyze the morphology of this binary tree using the external path length (section 2.4.3). For the largest connected component of Fig. 7, one obtains $\text{EPL}/\text{EPL}_{path} = 0.28$, but $\text{EPL}/\text{EPL}_{rand.} = 963.46$. This clearly shows that the binary tree associated with BLN69-all is essentially a path. This feature is likely related to the strategy used to find local minima and index one saddles connecting them in the original work on the exploration of the BLN landscape [Wal03, OWJ11].

This path-like tree structure also accounts for the unbalanced shape observed in Fig. 7. Indeed, the classical DG drawing algorithm consists of sorting the children of a node using either the number of leaves in the sub-trees or their lowest energy. For a binary tree reducing to a path, as we have here, the largest subtree always ends up on the same side (left or right). For a k-ary tree, the same procedure typically results in a more symmetric shape.

Persistence of local minima. We next analyze the persistence of local minima in BLN69-all, using the approach explained in section 2.4.1 (Fig. 8). Since minima from BLN69-top10 correspond to an energy $E < -104$, we restrict the PD to this set (Fig. 8(Inset)). This plot shows a clear gap isolating 10 minima (9 points of the diagram plus the global minimum, which never appears, as explained in Methods). Remarkably, out of the 10 most persistent minima, one finds 6 conformations from BLN69-top10 plus 4 new ones, whose lRMSD to local minima from BLN69-top10 is circa 0.5Å(supplemental Table 2). This observation shows the advantage of PD in identifying out the most stable local minima.

As can be seen from Fig. 7 (right-hand side), due to graphical limitations, plotting all the data simultaneously is not particularly useful for interpreting global dynamics of the system. The use of a persistence threshold greatly facilitates interpretation of the complete DG for the frustrated BLN system. For instance, as shown above, the persistence analysis (Fig. 8) shows that many basins are well separated from one another with persistences of up to 70 energy units. Drawing the complete DG using a persistence threshold of 15 ϵ spotlights these deep, meta-stable basins or states (Fig. 9 and 10). As opposed to grouping by transition state energies in fixed increments, as is done in typically reported DG's, such denoising by persistence analysis can be used to simplify the graphical construction without losing the precision in the representation of the transition energies.

[3] On the other hand, self-connexions do not make contributions to the structural dynamics.

We also note in passing that upon simplifying the whole DG with persistence (from Fig. 7 to Fig. 9), its path-like structure becomes even more dominant, since the ratio EPL/EPL$_{path}$ jumps from 0.28 to 0.99.

We may also examine the structure of the landscape in the region of the global minimum by focusing on the sub-level set of the landscape located below -95 energy units. This sub-level set, which we shall denote M, consists of 45408 local minima structured in 9562 c.c.. The particular c.c. containing the global minimum, denoted M_g, involves 22214 local minima. Observing that the median persistence of the top 10 local minima (BLN69-top10) is 6.14ϵ (supplemental Table 5), we simplified M, and thus M_g, using this value as a persistence threshold. In doing so, the number of remaining local minima reduces to 19: the global minimum plus 4 persistent minima from BLN69-top10 plus 14 other persistent basins from M_g. The latter minima are not in the top10, yet are more persistent than four local minima from BLN69-top10.

Transitions. To capture global properties of the transition graph, we first turn to the structure of attraction basins of the local minima in the set BLN69-E_{-100}. In particular, we analyze the *star* of each local minimum, that is, the direct connexions to surrounding saddles (Table 3). Interestingly, selected transitions are characterized by a long distance and a low energy (see e.g. the 5th column for the minimum of index 1). Clearly, such situations are difficulties for exploration algorithms, which need to cover a long distance to a saddle before reaching another basin.

We complement the previous analysis with parameters of the transition graph upon marking the 10 lowest local minima as landmarks, making use of the approaches presented in section 2.4.5. Statistics on the transition paths are compiled in the supplemental Table 4. One observes that a path joining two minima involves up to 5 intermediate minima, or equivalently 10 edges in the graph. It is also seen that there exist paths that are short in terms of nodes, but long distance-wise: for example, the two minima of indices 1 and 311 are separated by one saddle, yet the associated cumulative edge distance is $d_{ced} = 1.182$. These facts illustrate the difficulty faced by sampling algorithms, which may have to discover a long transition path to cross a single saddle — possibly high in energy. Another point of view is provided by the distribution of lengths of edges of a minimum spanning tree spanning the critical points (supplemental Table 1). For the datasets BLN69-E_{-100} and BLN69-all, a significant gap is observed between the median and the maximum edge lengths in the MST. This gap clearly shows that subsets of the ensemble are quite separated.

Finally, it is also seen that 5 local minima (6, 8, 142, 311, 33250) dare connected to the global minimum by a single saddle. When distance d_{ced} between these conformations is small, this explains why sampling algorithms consistently report these five minima during an exploration run (data not shown).

Comparing the distances lRMSD and d_{ced} via the ratio lRMSD/d_{ced} also proves of interest to identify pairs which are close in the lRMSD sense but far away according to the cumulative edge distance. For example, the pair (1974, 7305) yields a ratio of 0.32 and (311, 7305) a ratio of 0.37, while the path between (1, 142) is more direct, with a ratio of 0.92.

Sketching the landscape using non linear dimensionality reduction. To visualize the relative positions of the lowest local minima of the BLN landscape, as explained in section 2.4.6, we apply MDS to the distance matrices between all pairs from BLN69-top10 (Fig. 11). Both 2D projections have a similar global structure. However, as observed above, pairs nearby in the lRMSD sense may be far apart when considering the cumulative edge distance d_{ced} — e.g. the pairs (1974, 7305) and (311, 7305).

3.3 Comparing samplings

Setup. We apply our PEL comparison algorithms on the dataset TRRT-top10. To generate the vertex weighted transition graphs (section 2.5), we proceed in two stages. First, we apply topological persistence to merge the non significant basins into larger ones (section 2.1.2). Practically, this process gets rid of basins with persistence less than 750 energy units. Then, the version of algorithm `PLeSE` for PEL was run (section 2.4.2), resulting in the assignment of each sample to its basin (whence the basin weights) and a list of connections between basins.

Detailed information for each run, namely the DG, the associated persistence diagram, together with the distribution of masses, can be found in the supplemental section 7.2. Two facts are worth noticing. First, the basin weights within each dataset TRRT-top10 are unevenly distributed, since the local minimum from which the `T-RRT` run was started dominates the sampling, a fact related not only to intrinsic concerns about exploration efficiency but more fundamentally to frustration of the landscape itself. This indeed was the motivation for creating the "challenge" dataset TRRT-top10-U, which is the same as TRRT-top10 except that basin weights have been attributed that correspond to a more uniform sampling of the conformational space. Second, the transitions found involve samples with relatively high energies with respect to those of the lowest minima, a fact owing to the hybrid nature of the sampling, which consists of non critical points together with the associated quenched local minima, namely the ensemble $C \cup C_q$ from section 2.4.2. This, however, is not an issue in comparing PEL samplings, since the saddles are used simply to connect basins; the saddle heights themselves are unused.

We performed all pairwise comparisons of the 10 PEL samplings TRRT-top10-1-10 and TRRT-top10-U, computing for each pair (A, B) three transport plans, namely: that associated with the linear program, and the two associated with the earth mover distance with connectivity constraints – since `Alg-EMD-CCC-G` is not symmetric.

Algorithm `Alg-EMD-LP` and connectivity constraints. As algorithm `Alg-EMD-LP` is oblivious to critical point connectivity, we first focus on how well the connectivity is preserved. We do this by computing the fraction of vertices and edges of the input graph inducing, through the flow, a connected subgraph of the demand graph. On the 45 instances of the dataset TRRT-top10, the statistical summaries (min, median, max) of these vertex and edge fractions are respectively of (0.1, 0.62, 1.), and (0.03, 0.89, 1.) (supplemental Tables 6 and 7). For the dataset TRRT-top10-U, the summaries become (0., 0.63, 0.97) and (0.01, 0.4, 1.) (supplemental Tables 8 and 9). Thus, transport plans obtained by solutions of the linear program do disrupt connectivity constraints, especially for difficult cases.

Algorithm `Alg-EMD-CCC-G` and demand satisfaction. For algorithm `Alg-EMD-CCC-G`, the connectivity is preserved by construction, but this may prevent full satisfaction of the demand, which motivated the definition of the total flow Eq. (9).

Consider first the demand satisfaction, assessed by the ratio $\%_F = F_{\text{Alg-EMD-CCC-G}}/W_d$. (Recall that W_d stands for the sum of the weights of the basins of the demand graph.) For the dataset TRRT-top10, a worst-case value of 99.86% is observed (supplemental Table 10), showing that constraints do not provide a serious hindrance to satisfy the demand. Consider now the ratio r^F_{sym} as defined in Eq. (10). It turns out that algorithm `Alg-EMD-CCC-G` does not exhibit significant asymmetry on these instances, since the previous ratio lies in the interval $[0.94, 1]$, with a median of 0.99. colorblackNevertheless, the remarkable behavior of algorithm `Alg-EMD-CCC-G` partly owes to the nature of the instances, for, as observed above, a frustrated landscape results in one basin taking most of the mass. Therefore, the transport plan associated

with two landscapes is dictated by the flow from the dominant minimum of the first landscape to the dominant minimum of the second landscape. In other words, the transport cost is essentially the distance between these two minima.

The situation deteriorates for the dataset TRRT-top10-U (supplemental Table 11). First, the symmetry ratio r^F_{sym} takes values in the interval $[0.07, 0.96]$, with a median of 0.48. Also, satisfying the demand gets more difficult, and one finds both easy and hard instances: across all pairs, $\min\max(\%_F(A,B), \%_F(B,A))$ is as low as 5.87%, while $\max\min(\%_F(A,B), \%_F(B,A))$ is as high as 71.43%. The former pair defines a difficult case – the transportation problem is difficult both ways, while the latter is a simple one – the transportation problem is easy both ways.

Transport costs. In general, connectivity constraints could also deteriorate the transport cost computed by algorithm `Alg-EMD-CCC-G`. For the dataset TRRT-top10 (supplemental Tables 12 and 13), the pairwise correlations among $C_{\text{EMD}}(A,B), C_{\text{Alg-EMD-CCC-G}}(A,B)$ and $C_{\text{Alg-EMD-CCC-G}}(B,A)$ are at least equal to 0.999 (supplemental Fig. 22). As discussed above, this owes to the structure of the graphs modeling the sampling.

Again, the situation changes for the dataset TRRT-top10-U (supplemental Tables 14, 15). On the one hand, the costs become asymmetric, the symmetry ratio r^C_{sym} of Eq. (10) spanning the interval $[0.00, 0.95]$ with a median of 0.43 – as opposed to $[0.99, 1.]$ with a median of 1.. On the other hand, the correlation with the cost of linear transport plans is lost (Pearson correlation coefficients down to 0.25, 0.09, and 0.06; supplemental Fig. 23).

On the hardness of comparing landscapes. As evidenced by the previous analysis, the comparison of landscapes is a difficult problem for which one may favor the satisfaction of the demand (using `Alg-EMD-LP`) at the detriment of the connectivity constraints, or vice-versa. The intrinsic hardness of the problem, as explained in section 2.5.2, makes the presence of such cases inevitable.

3.3.1 BLN69: Summary of novel insights

colorblackAnalyzing the database of critical points for BLN69 and the connectivity information encoded therein, as well as portions of the landscape revealed by different exploration methods, reveals several novel insights.

The topological persistence analysis allows a global view of the entirety of the PEL data for the BLN69 model protein system (Fig. 8) that is not feasible with an unfiltered discontinuity graph (e.g., Fig. 7). In such complex cases discontinuity graphs are often applied to "manually" culled or regrouped subsets of the database in order to produce interpretable diagrams. In contrast, the full persistence diagram immediately shows the presence of high barriers (here up to 70 energy units) separating even relatively low-energy structures, emphasizing the frustrated nature of the BLN69 system and allowing a preliminary identification of kinetic traps, which further presumably underlie the variability in success rates for attaining global minima in this system seen for different exploration methods.

Relatedly, topological persistence also provides a robust way of systematically reporting only those basins interconnected by barriers of a prescribed height, or whose local minima lie in a specific energy slice. In particular, we show that for the typical figures considered, 14 such local minima distinguish themselves based on persistence criteria– while previous work tended to focus on the 10 lowest-energy ones.

The embedding of critical points performed by the `ISOMAP` algorithm also reveals clusters of critical points in the sample database. The presence of such structuration in the database of

samples can be used to shed light on the ability of exploration methods to uncover important local minima.

Calculating the internal path-length statistic for the overall basin connectivity in the BLN69 protein model, as encoded in the disconnectivity graph of the landscape, reveals a rather unbalanced structure. Intuitively, this suggests an underlying funnel nature to the PEL despite the high degree of frustration in this system.

Finally, in using the earth-mover's distance to compare samplings obtained for the BLN69 protein model, with its more than 200 degrees of freedom, we saw that despite high complexity the algorithm is effective in handling comparisons of landscapes with hundreds of minima.

4 Conclusion

This paper revisits classical concepts and present new ones for the analysis and comparison of conformational ensembles and sampled landscapes. From a methodological standpoint, two classes of novel tools are of particular interest.

The first class is concerned with persistence based algorithms providing a coherent framework to handle landscapes presenting multiple scales in the barrier height distribution. When the landscape processed is the potential energy of the system, these algorithms allow automatically selecting minima with specified values in terms of energy and saddle height. When the processed landscape encodes the density of states, the same algorithms yield a clustering method whose output hints at the number of stable states.

The second class is concerned with methods aiming at comparing sampled landscapes. The comparison may be restricted to the persistent basins and their volumes, or may also incorporate connexions between the basins. The possibility of comparing sampled landscapes should prove invaluable in investigating the coherence of two force fields, in distinguishing comparable but distinct macromolecular systems, or simply in comparing simulations launched with different initial conditions.

The software corresponding to these novel tools is made available in two guises. First, executables are provided for various operating systems (Linux, MacOS). Second and possibly most importantly, the C++ will be released within the *Structural Bioinformatics Library* during the first trimester of 2015.

Acknowledgments. The authors wish to thank D. Wales and M. Oakley for providing the database of stationary points of the BLN69 model.

References

[AA92] R. Abagyan and P. Argos. Optimal protocol and trajectory visualization for conformational searches of peptides and proteins. *Journal of molecular biology*, 225(2):519–532, 1992.

[ALB93] Andrea Amadei, Antonius Linssen, and Herman JC Berendsen. Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics*, 17(4):412–425, 1993.

[AMN+98] S. Arya, D.M. Mount, N.S. Netanyahu, R. Silverman, and A.Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)*, 45(6):891–923, 1998.

[BBK95] R.S. Berry and R. Breitengraser-Kunz. Topography and dynamics of multidimensional interatomic potential surfaces. *Physical review letters*, 74(20):3951, 1995.

[BCCS+11] G. Biau, F. Chazal, D. Cohen-Steiner, L. Devroye, and C. Rodriguez. A weighted k-nearest neighbors density estimate for geometric inference. *Electronic Journal of Statistics*, 5(204-237), 2011.

[Ber10] R Stephen Berry. Energy landscapes: topographies, interparticle forces and dynamics, and how they are related. *Theoretical Chemistry Accounts*, 127(3):203–209, 2010.

[BFHG03] Scott Brown, Nicolas J. Fawzi, and Teresa Head-Gordon. Coarse-grained sequences for protein folding and design. *Proc Natl Acad Sci U S A*, 100(19):10712–10717, Sep 2003.

[BH04] A. Banyaga and D. Hurtubise. *Lectures on Morse Homology*. Kluwer, 2004.

[BK97] O. Becker and M. Karplus. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *The Journal of Chemical Physics*, 106(4):1495–1517, 1997.

[BN92] B.A Berg and T. Neuhaus. Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Physical Review Letters*, 68(1):9, 1992.

[BOSW95] J.D. Bryngelson, J.N. Onuchic, N. Socc, and P.G. Wolynes. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Structure, Function, and Bioinformatics*, 21(3):167–195, 1995.

[BW87] J. Bryngelson and P.G. Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences*, 84(21):7524–7528, 1987.

[CadOS11] F. Chazal, L.J. Guibas an dS.Y. Oudot, and P. Skraba. Persistence-based clustering in riemannian manifolds. In *ACM SoCG*, pages 97–106, 2011.

[CCS11] F. Cazals and D. Cohen-Steiner. Reconstructing 3D compact sets. *Computational Geometry Theory and Applications*, 45(1-2):1–13, 2011.

[cga] CGAL, Computational Geometry Algorithms Library. http://www.cgal.org.

[Che95] Yizong Cheng. Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(8):790–799, 1995.

[CLRSon] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction to algorithms*. MIT press, 2009 (3rd edition).

[CM15] F. Cazals and D. Mazauric. Mass transportation problems with connectivity constraints, with applications to energy landscape comparison, 2015. Inria report 8611.

[CSEH05] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. In *ACM Symp. Comp. Geometry*, 2005.

[CW08] J.M. Carr and D.J. Wales. Folding pathways and rates for the three-stranded β-sheet peptide beta3s using discrete path sampling. *The Journal of Physical Chemistry B*, 112(29):8760–8769, 2008.

[DB09] Christoph Dellago and Peter G Bolhuis. Transition path sampling and other advanced simulation techniques for rare events. In *Advanced computer simulation approaches for soft matter sciences III*, pages 167–233. Springer, 2009.

[DM05] Jonathan PK Doye and Claire P Massen. Characterizing the network topology of the energy landscapes of atomic clusters. *The Journal of chemical physics*, 122(8):084105, 2005.

[DMS+06] P. Das, M. Moll, H. Stamati, L. Kavraki, and C. Clementi. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *PNAS*, 103(26):9885–9890, 2006.

[dSLT00] V. de Silva, J.C. Langford, and J.B. Tenenbaum. Graph approximations to geodesics on embedded manifolds. Technical report, 2000.

[DWB05] Florin Despa, David J Wales, and R Stephen Berry. Archetypal energy landscapes: Dynamical diagnosis. *The Journal of chemical physics*, 122(2):024103, 2005.

[EH10] H. Edelsbrunner and J. Harer. *Computational topology: an introduction*. AMS, 2010.

[EK87] R. Elber and M. Karplus. Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science*, 235(4786):318–321, 1987.

[FK97] A.T. Fomenko and T.L. Kunii. *Topological Modeling for visualization*. Springer, 1997.

[GBHK97] A.E. García, R. Blumenfeld, G. Hummer, and J.A. Krumhansl. Multi-basin dynamics of a protein in a crystal environment. *Physica D: Nonlinear Phenomena*, 107(2):225–239, 1997.

[Heu97] A. Heuer. Properties of a glass-forming system as derived from its potential energy landscape. *Physical review letters*, 78(21):4051, 1997.

[HJJ02] Graeme Henkelman, Gísli Jóhannesson, and Hannes Jónsson. Methods for finding saddle points and minimum energy paths. In *Theoretical Methods in Condensed Phase Chemistry*, pages 269–302. Springer, 2002.

[HOE96] U. Hansmann, Y. Okamoto, and F. Eisenmenger. Molecular dynamics, langevin and hydrid monte carlo simulations in a multicanonical ensemble. *Chemical physics letters*, 259(3):321–330, 1996.

[HPD+01] F. Hamprecht, C. Peter, X. Daura, W. Thiel, and W.F. van Gunsteren. A strategy for analysis of (molecular) equilibrium simulations: Configuration space density estimation, clustering, and visualization. *The Journal of Chemical Physics*, 114(5):2079–2089, 2001.

[HS88] K.H. Hoffmann and P. Sibani. Diffusion in hierarchies. *Physical Review A*, 38(8):4261, 1988.

[HS13] K.H. Hoffmann and J.C. Schön. Controlled dynamics on energy landscapes. *The European Physical Journal B*, 86(5):1–10, 2013.

[HT90] J.D. Honeycutt and D. Thirumalai. Metastability of the folded states of globular proteins. *Proceedings of the National Academy of Sciences*, 87(9):3526–3529, 1990.

[JCPC11a] L. Jaillet, F.J. Corcho, J-J. Pérez, and J. Cortés. Randomized tree construction algorithm to explore energy landscapes. *Journal of computational chemistry*, 32(16):3464–3474, 2011.

[JCPC11b] L. Jaillet, F.J. Corcho, J.J. Pérez, and J. Cortés. A randomized tree construction algorithm to explore energy landscapes. *J. of Comp. Chem.*, 2011.

[Kam92] N.G. Van Kampen. *Stochastic processes in physics and chemistry*, volume 1. Elsevier, 1992.

[Kar84] N. Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311. ACM, 1984.

[KHM+05] T. Komatsuzaki, K. Hoshino, Y. Matsunaga, G. Rylance, R. Johnston, and D.J. Wales. How many dimensions are required to approximate the potential energy landscape of a model protein? *The Journal of chemical physics*, 122(8):084714, 2005.

[KL00] James J Kuffner and Steven M LaValle. Rrt-connect: An efficient approach to single-query path planning. In *Robotics and Automation, 2000. Proceedings. ICRA'00. IEEE International Conference on*, volume 2, pages 995–1001. IEEE, 2000.

[Kri02] S.V. Krivov. Hierarchical global optimization of quasiseparable systems: Application to lennard-jones clusters. *Physical Review E*, 66(2):025701, 2002.

[KSH98] T. Klotz, S. Schubert, and K.H. Hoffmann. Coarse graining of a spin-glass state space. *Journal of Physics: Condensed Matter*, 10(27):6127, 1998.

[KV+83] Scott Kirkpatrick, MP Vecchi, et al. Optimization by simmulated annealing. *science*, 220(4598):671–680, 1983.

[KZ09] S. Kannan and M. Zacharias. Simulated annealing coupled replica exchange molecular dynamics–an efficient conformational sampling method. *Journal of structural biology*, 166(3):288–294, 2009.

[LKJ00] Steven M LaValle and James J Kuffner Jr. Rapidly-exploring random trees: Progress and prospects. 2000.

[LMSVV92] A.P. Lyubartsev, A.A. Martsinovski, S.V. Shevkunov, and P.N. Vorontsov-Velyaminov. New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. *J. Chem. Phys*, 96:1776–83, 1992.

[LP02] A. Laio and M. Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.

[LS87] Z. Li and H.A. Scheraga. Monte carlo-minimization approach to the multiple-minima problem in protein folding. *Proceedings of the National Academy of Sciences*, 84(19):6611–6615, 1987.

[LV07] J.A. Lee and M. Verleysen. *Nonlinear dimensionality reduction*. Springer Verlag, 2007.

[MADWB11] N. Michaud-Agrawal, E.J. Denning, T. Woolf, and O. Beckstein. Mdanalysis: a toolkit for the analysis of molecular dynamics simulations. *Journal of computational chemistry*, 32(10):2319–2327, 2011.

[Mah92] H. Mahmoud. *Evolution of random search trees*. Wiley-Interscience, 1992.

[Mil63] J.W. Milnor. *Morse Theory*. Princeton University Press, Princeton, NJ, 1963.

[MO09] A. Mitsutake and Y. Okamoto. Multidimensional generalized-ensemble algorithms for complex systems. *J Chem Phys*, 130(21):214105, Jun 2009.

[NAKH14] L. Nedialkova, M. Amat, I. Kevrekidis, and G. Hummer. Diffusion maps, clustering and fuzzy markov modeling in peptide folding transitions. *J. Chem. Phys.*, 140, 2014.

[NLCK06] Boaz Nadler, Stéphane Lafon, Ronald R Coifman, and Ioannis G Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006.

[NMN01] Jaroslav Nešetřil, Eva Milková, and Helena Nešetřilová. Otakar borůvka on minimum spanning tree problem (Translation of both the 1926 papers, comments, history). *Discrete Mathematics*, 233(1):3–36, 2001.

[NP06] Naoko Nakagawa and Michel Peyrard. The inherent structure landscape of a protein. *Proc Natl Acad Sci U S A*, 103(14):5279–5284, Apr 2006.

[OD13] Stephen O'Hara and Bruce A Draper. Are you using the right approximate nearest neighbor algorithm? In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 9–14. IEEE, 2013.

[OJW12] M.T. Oakley, R.L. Johnston, and D.J. Wales. The effect of nonnative interactions on the energy landscapes of frustrated model proteins. *J. At. Mol. Opt. Phys.*, 2012.

[OLSW97] J.N. Onuchic, Z. Luthey-Schulten, and P.G. Wolynes. Theory of protein folding: the energy landscape perspective. *Annual review of physical chemistry*, 48(1):545–600, 1997.

[OWJ11] M. T. Oakley, D. J. Wales, and R. L Johnston. Energy landscape and global optimization for a frustrated model protein. *The Journal of Physical Chemistry B*, 115(39):11525–11529, 2011.

[PBB10] 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.

[PBC10] L.B. Pártay, A.P. Bartók, and G. Csányi. Efficient sampling of atomic configurational spaces. *J. Phys. Chem. B*, 114:10502–10512, 2010.

[pel] pele : Python energy landscape explorer. `http://pele-python.github.io/pele/`.

[RTG00] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[RZMC11] 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.

[Sam06] H. Samet. *Foundations of multidimensional and metric data structures*. Morgan Kaufmann, 2006.

[SO99] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters*, 314(1):141–¡151, 1999.

[SPJ96] J.C. Schön, H. Putz, and M. Jansen. Studying the energy hypersurface of continuous systems-the threshold algorithm. *Journal of Physics: Condensed Matter*, 8(2):143, 1996.

[SSSA93] P. Sibani, C. Schön, P. Salamon, and J-O. Andersson. Emergent hierarchical structures in complex-system dynamics. *EPL (Europhysics Letters)*, 22(7):479, 1993.

[SvdPS99] P. Sibani, R. van der Pas, and J.C. Schön. The lid method for exhaustive exploration of metastable states of complex systems. *Computer Physics Communications*, 116(1):17–27, 1999.

[SW82] Frank H. Stillinger and Thomas A. Weber. Hidden structure in liquids. *Phys. Rev. A*, 25:978–989, 1982.

[TB95] J.M. Troyer and F.E. Cohen Becker. Protein conformational landscapes: Energy minimization and clustering of a long molecular dynamics trajectory. *Proteins: Structure, Function, and Bioinformatics*, 23(1):97–110, 1995.

[TKSW13] Ha H Truong, Bobby L Kim, Nicholas P Schafer, and Peter G Wolynes. Funneling and frustration in the energy landscapes of some designed and simplified proteins. *The Journal of chemical physics*, 139(12):121908, 2013.

[TPK02] M.L. Teodoro, G.N. Phillips, and L.E. Kavraki. A dimensionality reduction approach to modeling protein flexibility. In *ACM RECOMB*, 2002.

[Vil03] C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[Wal02] D.J. Wales. Discrete path sampling. *Molecular Physics*, 100(20):3285–3305, 2002.

[Wal03] D.J. Wales. *Energy Landscapes*. Cambridge University Press, 2003.

[WB06] D.J. Wales and T.V. Bogdan. Potential energy and free energy landscapes. *The Journal of Physical Chemistry B*, 110(42):20765–20776, 2006.

[WMW98] David J Wales, Mark A Miller, and Tiffany R Walsh. Archetypal energy landscapes. *Nature*, 394(6695):758–760, 1998.

[WSM99] M. Wevers, J.C. Schön, and M.Jansen. Global aspects of the energy landscape of metastable crystal structures in ionic compounds. *Journal of Physics: Condensed Matter*, 11(33):6487, 1999.

5 Artwork

Figure 1 Height field analysis of the Himmelblau function $f(x,y) = (x^2+y-11)^2+(x+y^2-7)^2$**.** The function has four local minima, four index one saddles, and one local maximum. The left and right columns respectively represent the function as defined in the previous equation, and to a noisy version obtained by adding Gaussian noise (noisy Himmelblau). **(Top row)** Plots of the function. The inset on the right side indicates the noise. **(Middle row)** Disconnectivity graphs. While that of Himmelblau features three saddles corresponding to the merges between the four basin, that of noisy Himmelblau contains more events. **(Bottom row)** Persistence diagram of sub-level sets. That of Himmelblau has three points, corresponding to three pairs (minimum, saddle). Note that the basin of the global minimum does not appear on the PD since it does not merge with a deeper basin. That of noisy Himmelblau contains two sets of points: those near the diagonal correspond to small wiggles on the landscape, while the remaining three are in one-to-one correspondence with those of the left diagram.

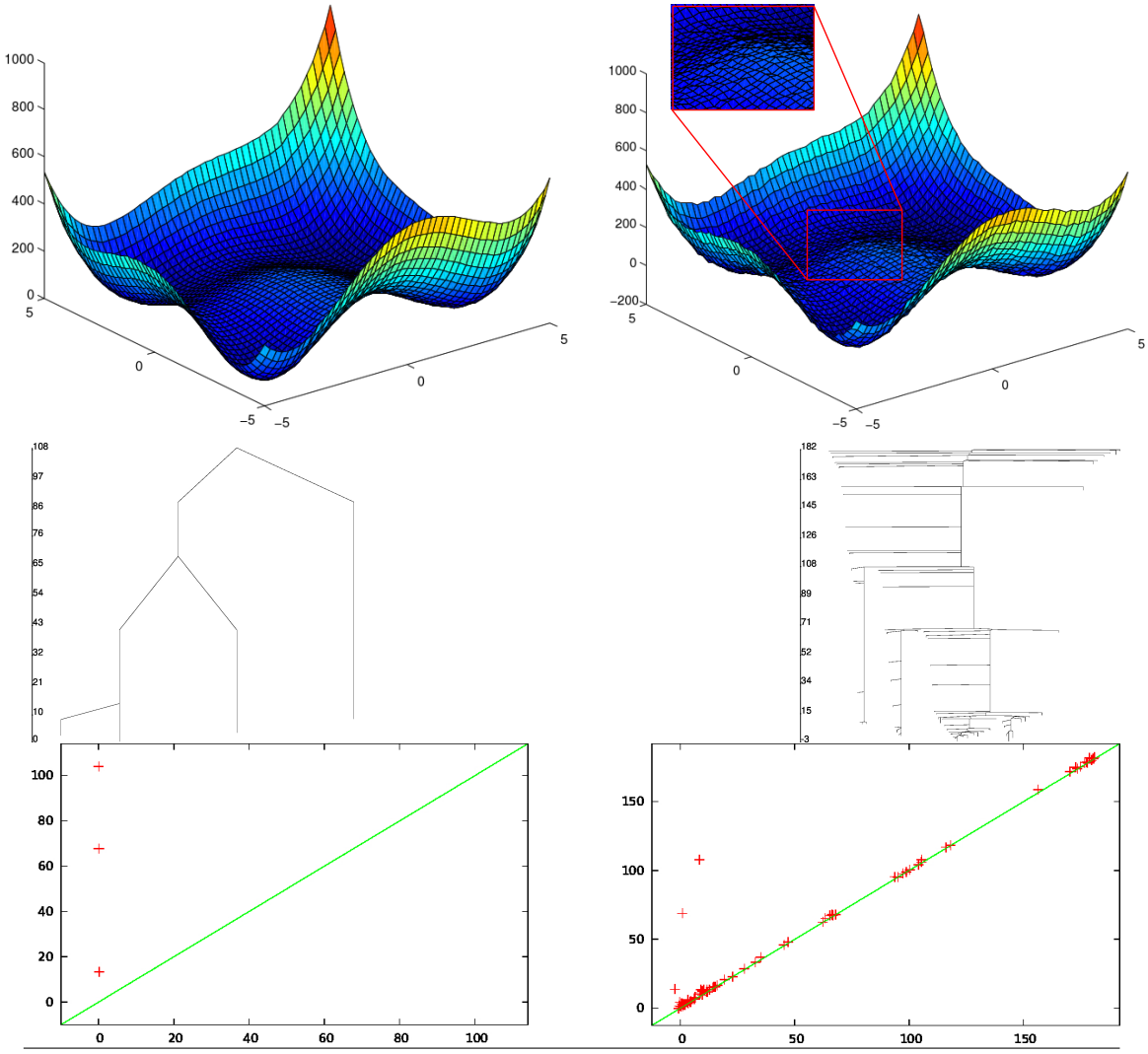


Figure 2 colorblack**Exploiting a persistence diagram: selecting samples belonging to a basin born before a prescribed energy E_{max} and with a persistence higher than a threshold ΔE.** The thresholds $E_{\max}$ and ΔE yield a partition of the PD consisting of five regions $R_1, \ldots, R_5$. See text for details.

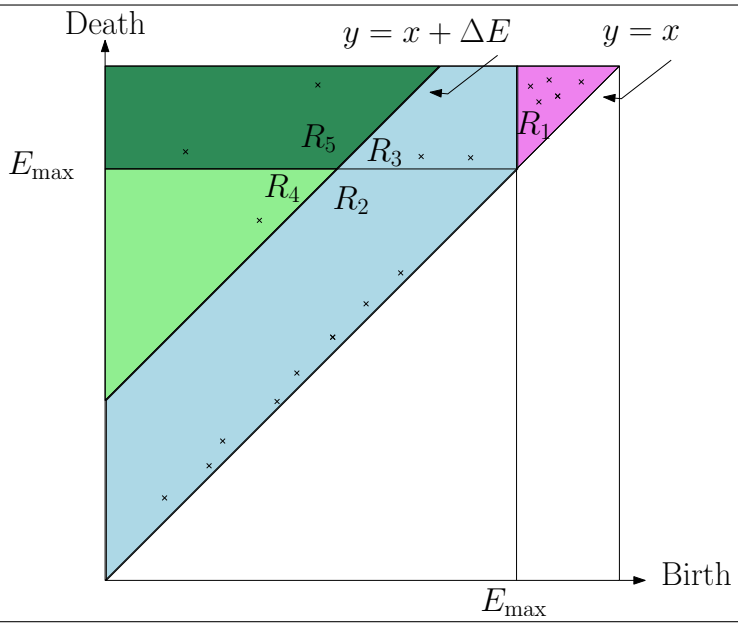


Figure 3 Identifying the connexion between the basins of two local minima by quenching two samples c_i and c_j. (Top) A simple 2D landscape Sample c_i is quenched to $\sigma_i^{(0)}$ while c_j is quenched to $\sigma_j^{(0)}$. The samples c_i and c_j *sit across* the (white) ridge separating the basins of their respective minima. **(Bottom) General case** Portrayed is a generic sketch of the flow lines associated with the negative gradient vector field of the potential energy $-\nabla V$ (see e.g. the Morse homology theorem in [BH04]). Sample c_i is quenched to $\sigma_i^{(0)}$ while c_j is quenched to $\sigma_j^{(0)}$. The two samples define a line segment intersecting transversely the stable manifold in the index one critical point $\sigma_{ij}^{(1)}$ — which is $d-1$ dimensional since the index (number of negative eigenvalues) of the critical point is $d-1$. The unstable manifold of $\sigma_{ij}^{(1)}$ is one-dimensional, and yields the connections with the two local minima $\sigma_i^{(0)}$ and $\sigma_j^{(0)}$.

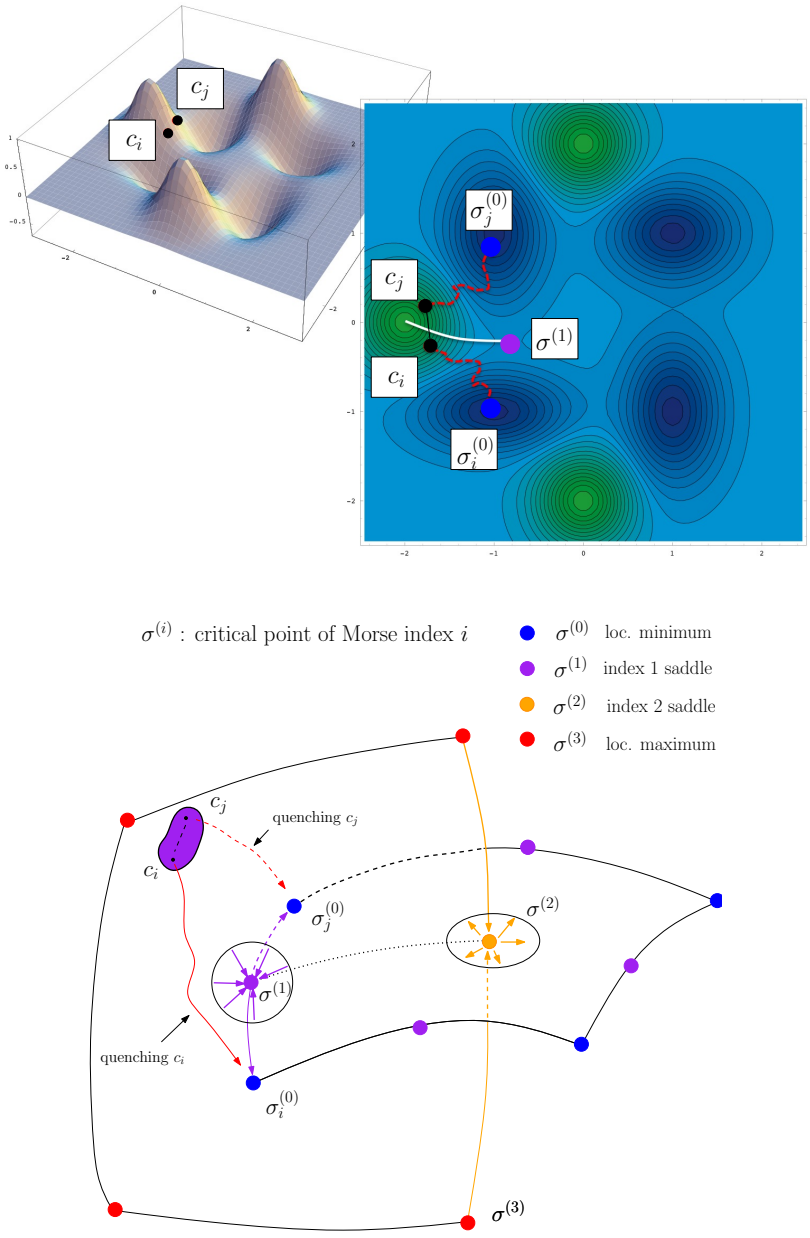


Figure 4 A loop around a saddle s anchored in a single local minimum m — a.k.a bump transition. Note that the dotted path is located behind the bump.

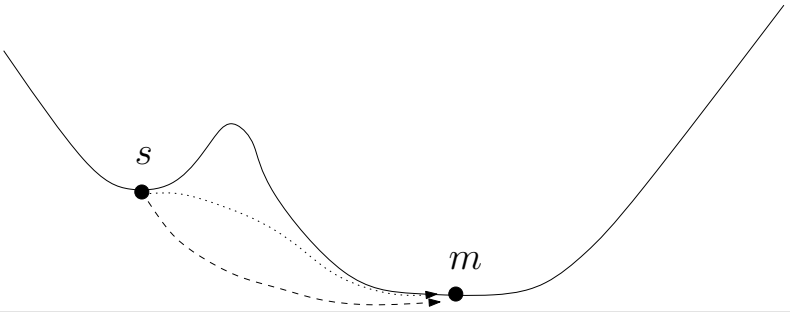


Figure 5 Comparing two energy landscapes $\mathbf{PEL}_s$ and $\mathbf{PEL}_d$. The landscape PEL_s is partitioned into the basins associated to its local minima (two of them on this example), and likewise for PEL_d (four local minima). Comparing the landscapes is phrased as a mass transportation problem on the bipartite graph defined by the two sets of minima. Note that sets of connected basins from the top landscape are mapped to connected basins of the bottom landscape.

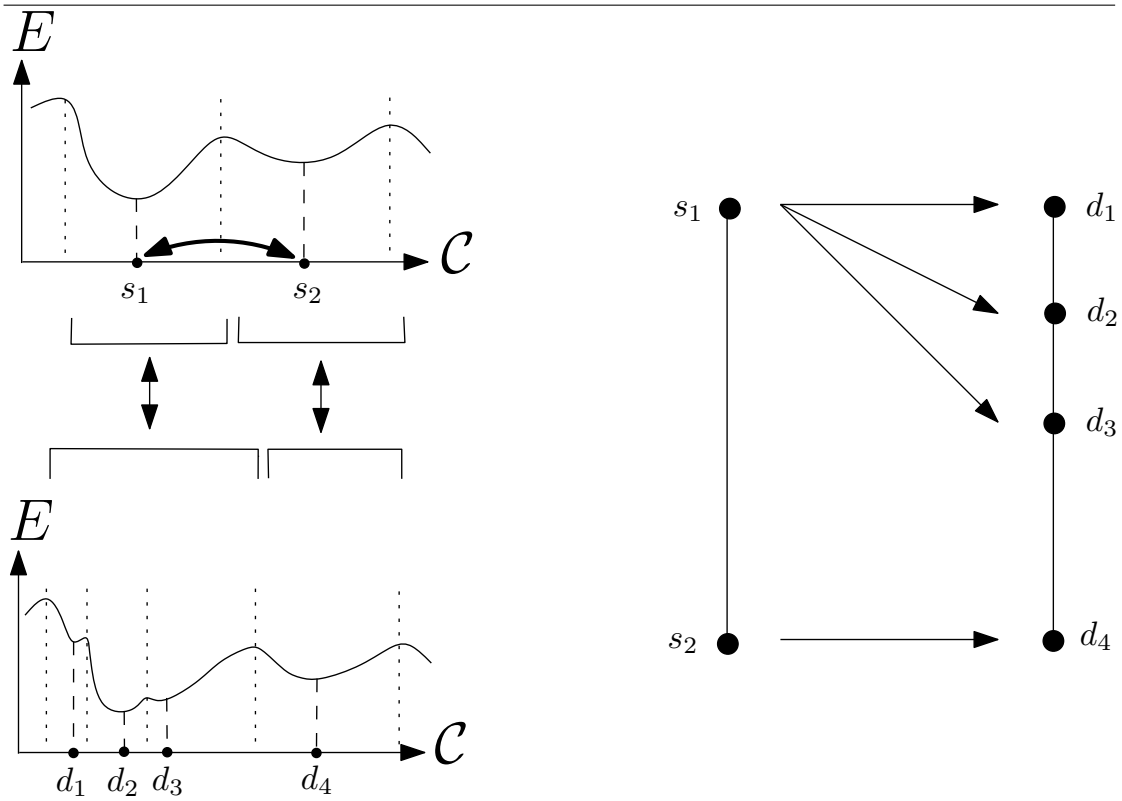


Figure 6 The solution of the linear program may not satisfy connectivity constraints. A transport plan between a source and a demand graph each consisting of a linear chain of four vertices. The vertices of the edge $\{s_1, s_2\}$ of the source graph export towards the vertices d_1 and d_3 of the demand graph. The subgraph of the demand graph induced by these vertices is not connected – there is no edge linking d_1 to d_3.

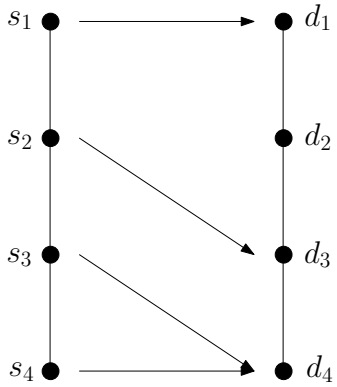


Figure 7 Disconnectivity graph associated with the dataset BLN69-all (see also Fig. 8). The inset is an arbitrary zoom on the tree containing the global minimum for $E < -103$. There are four minima from BLN69-top10 that are in the inset (in-circled).

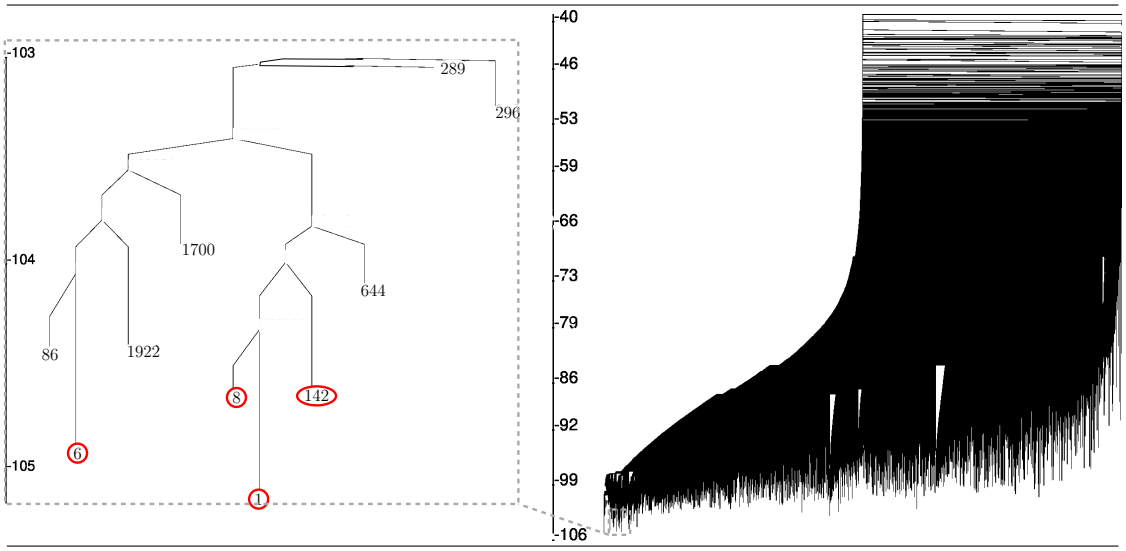


Figure 8 Persistence of the basins in BLN69-all. The main plot features the 458,082 minima, while the inset zooms on the region corresponding to the forty minima in BLN69-all with energy $E < -104$ units. The 10 most persistent minima (9 points of the diagram, plus the global minimum) contained in the ellipsis correspond to 6 conformations from BLN69-top10 and to 4 new ones. See also the supplemental Table 2.

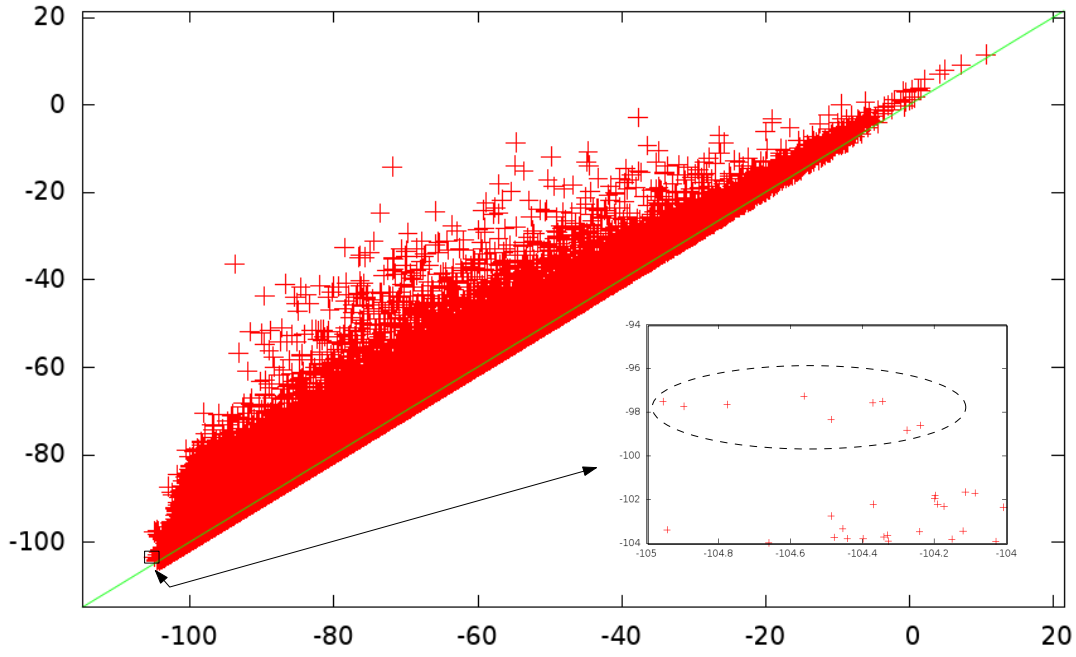


Figure 9 Frustration of BLN69 revealed by topological simplification using persistence. Illustrated is the DG of Fig. 7(right), simplified with a persistence threshold of 15 ϵ, which reveals the frustration of BLN69

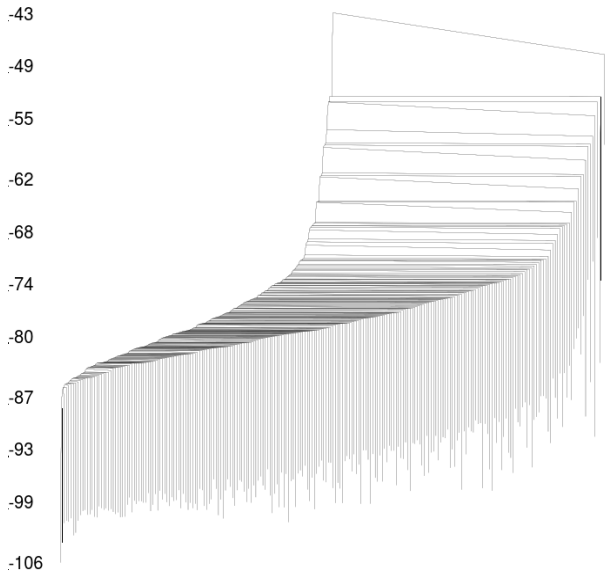


Figure 10 The DG of Fig. 9 plotted after clustering saddles within energy slices of height 0.5ϵ**.** Note that both DG have been plotted by the same algorithm, so that the visual difference stems from the binary versus k-ary structure of the trees.

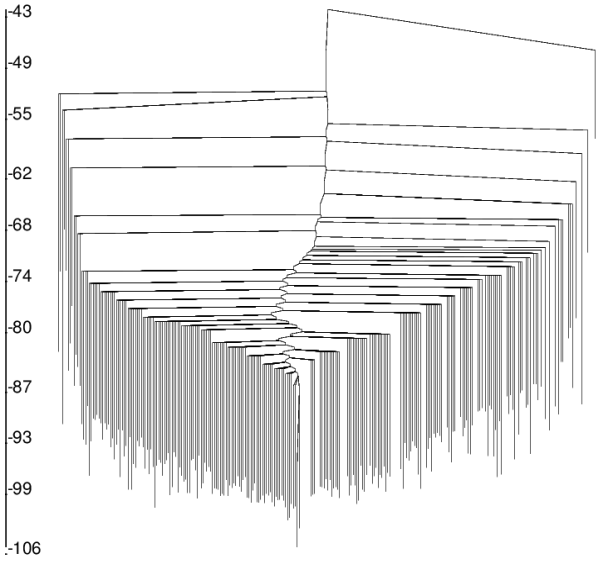
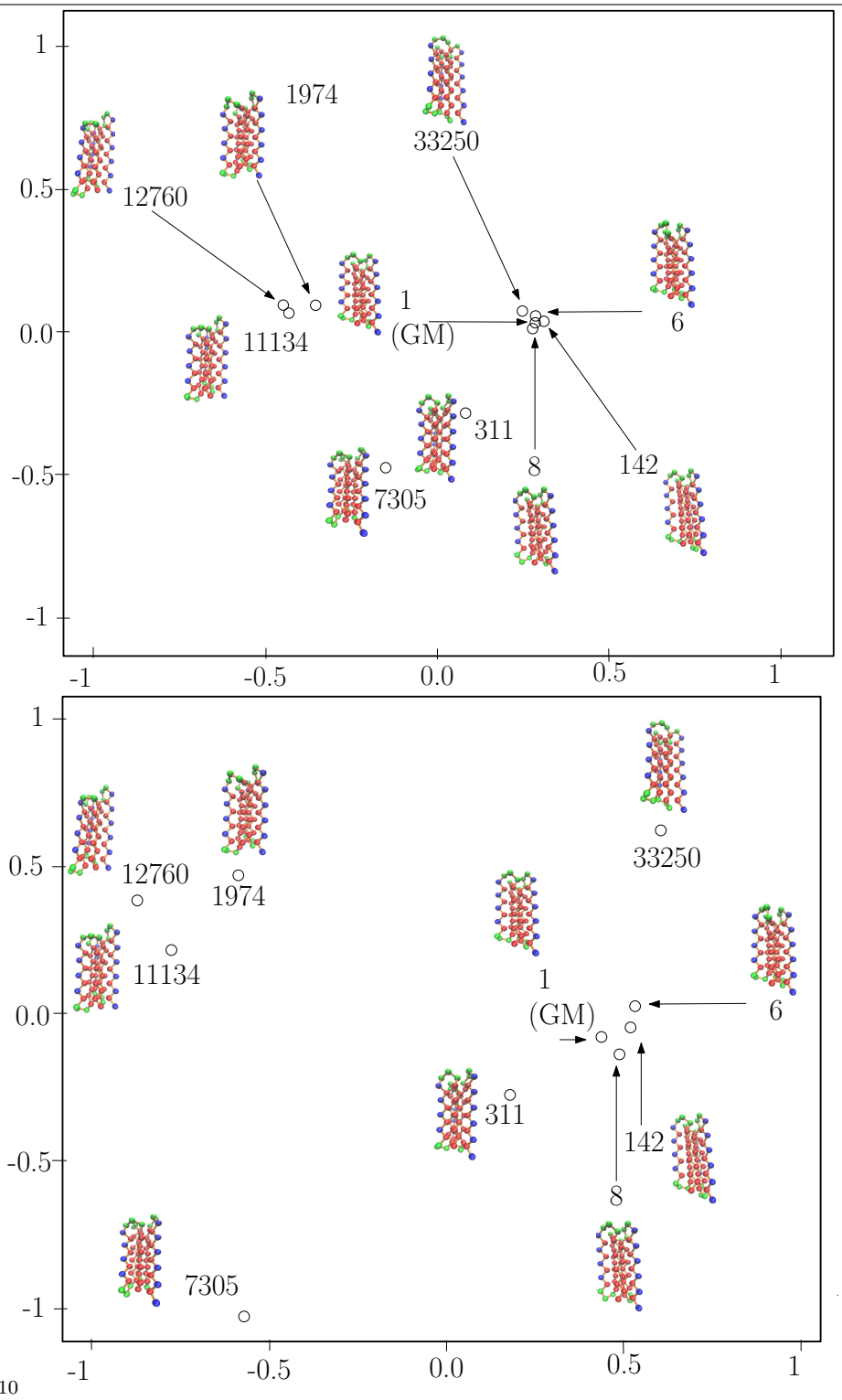


Figure 11 2D sketch of the energy landscape representing BLN69-top10 using multi dimensional scaling (MDS) on a matrix of pairwise cumulative distances. (Top) The distance used is the lRMSD . **(Bottom)** The distance used is the cumulative edge distance defined in section 2.4.6, the landmarks being the 10 lowest local minima.



6 Supplemental: Methods

6.1 BLN

The potential energy of the BLN69 model is given by:

$$V = \frac{1}{2}K_r \sum_{i}^{N-1}(R_{i,i+1} - R_e)^2 + \frac{1}{2}K_\theta \sum_{i}^{N-2}(\theta_i - \theta_e)^2 + \epsilon \sum_{i}^{N-3}[A_i(1+\cos\phi_i) + B_i(1+\cos 3\phi_i)] + 4\epsilon \sum_{i}^{N-2}\sum_{j=i+2}^{N} C_{i,j}[(\frac{\sigma}{R_{i,j}})^{12} - D_{i,j}(\frac{\sigma}{R_{i,j}})^6].$$

Note that the first three terms are bonded terms, while the fourth is the non bonded term (LJ potential). Parameter definitions and values are as specified in [Wal03].

7 Supplemental: Results

7.1 Modeling a Landscape

Table 1 Sampling diversity assessment based on the statistical summary (min, median, max) of edges of a minimum spanning tree spanning a conformational ensemble. The statistics reported are those of Eq. (1), the distance used being the lRMSD .

	min	median	max
BLN69-top10	0.0981464	0.360528	0.625568
BLN69-E_{-100}	0.00046106	0.118135	5.3118
BLN69-all	0.0206142	0.254532	2.42746

Table 2 Novel persistent local minima from Fig. 8(Inset), and the nearest neighbors from the BLN69-top10 dataset

index pers min	index nn in BLN69-top10	lRMSD
6872	311	0.482
2507	311	0.36
31779	11134	0.463
77992	311	0.482

Table 3 Local Analysis of the star of minima from the BLN69-E_{-100} that lie at energy less -104. The star of a given minimum consists of all the saddles directly connected to it in BLN69-E_{-100}. The lRMSD between the minimum and the saddle together with the energy increase ΔE were recorded, whence a pair for each saddle. Sorting these pairs by increasing lRMSD yields the columns 3–5. Sorting these pairs by increasing ΔE yields the columns 6–8.

Index	num. TS	Min dist	Median dist	Max dist	Min Δ_E	Median Δ_E	Max Δ_E
1	340	(0.0659, 1.17)	(0.32, 17.3)	(1.82, 17)	(0.842, 0.0681)	(13.5, 0.22)	(54.3, 1.58)
6	153	(0.0658, 1.16)	(0.307, 13.4)	(3.53, 19.4)	(0.896, 0.0677)	(11.9, 0.586)	(28.7, 0.312)
8	166	(0.0443, 0.348)	(0.297, 9.27)	(1.85, 28.8)	(0.348, 0.0443)	(11.8, 0.289)	(28.8, 1.85)
86	97	(0.0455, 0.416)	(0.298, 7.18)	(2.93, 19.4)	(0.416, 0.0455)	(10.7, 0.396)	(23.6, 0.705)
142	185	(0.0492, 0.667)	(0.303, 20.4)	(1.37, 14.8)	(0.667, 0.0492)	(12.9, 0.836)	(30.5, 1.06)
195	263	(0.0647, 0.896)	(0.341, 11.9)	(2.65, 22.2)	(0.896, 0.0647)	(14.2, 0.358)	(55, 1.72)
311	75	(0.0665, 1.12)	(0.273, 3.62)	(1.64, 27.8)	(1.12, 0.0665)	(10.3, 0.457)	(27.8, 1.64)
383	44	(0.0493, 0.679)	(0.274, 15.2)	(1.3, 25)	(0.679, 0.0493)	(10.5, 0.547)	(25, 1.3)
644	138	(0.0435, 0.334)	(0.312, 12)	(1.41, 28.5)	(0.334, 0.0435)	(11, 0.222)	(46.7, 1.1)
1245	81	(0.0452, 0.734)	(0.506, 21.8)	(1.78, 30.1)	(0.105, 0.0469)	(13.1, 0.9)	(30.1, 1.78)
1255	78	(0.066, 1.17)	(0.262, 26.4)	(1.82, 39)	(0.831, 0.067)	(11.5, 0.306)	(39, 1.82)
1404	69	(0.0763, 2.09)	(0.268, 14.7)	(2.05, 17.6)	(0.376, 0.104)	(9.78, 0.435)	(28.4, 0.278)
1630	147	(0.044, 0.36)	(0.294, 15.9)	(1.53, 21.2)	(0.36, 0.044)	(13.7, 0.339)	(48.5, 0.984)
1922	98	(0.049, 0.655)	(0.282, 6.16)	(1.4, 18.5)	(0.655, 0.049)	(11.1, 0.334)	(27.4, 1.18)
1963	88	(0.0573, 0.163)	(0.257, 12.7)	(1.05, 13.1)	(0.163, 0.0573)	(11.9, 0.345)	(25.2, 0.762)
1974	215	(0.0644, 0.862)	(0.321, 20.3)	(2.92, 19)	(0.799, 0.091)	(14.6, 0.729)	(34.2, 0.268)
2010	99	(0.0722, 2.57)	(0.306, 6.87)	(0.957, 19.7)	(0.345, 0.0996)	(9.86, 0.573)	(28.3, 0.282)
2240	89	(0.0456, 0.593)	(0.277, 14.9)	(2.93, 19.1)	(0.593, 0.0456)	(13.7, 0.468)	(25.9, 1.29)
2507	73	(0.0684, 2.61)	(0.243, 17.7)	(1.25, 14.3)	(0.377, 0.103)	(9.6, 0.264)	(28.3, 0.264)
3618	86	(0.087, 1.83)	(0.283, 11.1)	(1.31, 26.5)	(1.71, 0.131)	(11.4, 0.216)	(31.7, 0.239)
3703	130	(0.0636, 0.844)	(0.311, 17.1)	(2.94, 19.1)	(0.844, 0.0636)	(14.4, 0.431)	(37.6, 2.13)
5876	62	(0.0451, 0.0978)	(0.548, 15.5)	(1.87, 26.6)	(0.0978, 0.0451)	(12.8, 0.87)	(27.2, 1.25)
6327	55	(0.0484, 0.118)	(0.3, 7.42)	(1.77, 26.4)	(0.118, 0.0484)	(8.52, 0.593)	(26.4, 1.77)
6872	93	(0.0718, 2.52)	(0.279, 7.48)	(1.29, 20.9)	(1.68, 0.132)	(10.5, 0.329)	(23.4, 0.239)
7265	97	(0.0437, 0.0881)	(0.493, 19.2)	(1.91, 28.1)	(0.0881, 0.0437)	(12.9, 0.954)	(28.9, 1.76)
7271	75	(0.0618, 0.743)	(0.287, 13.6)	(1.58, 18.8)	(0.322, 0.0771)	(10.4, 0.408)	(30.5, 0.339)
7300	90	(0.0621, 0.769)	(0.28, 7.25)	(1.77, 18.3)	(0.309, 0.0753)	(9.67, 0.637)	(32.7, 1.52)
7305	93	(0.0454, 0.752)	(0.289, 5.92)	(1.08, 18.4)	(0.29, 0.0744)	(10.3, 0.398)	(21.9, 0.276)
7489	75	(0.052, 0.899)	(0.255, 20.1)	(1.62, 18.4)	(0.273, 0.0719)	(10.7, 0.228)	(27.7, 0.377)
9846	256	(0.0667, 0.865)	(0.332, 15.1)	(1.95, 34.8)	(0.865, 0.0667)	(15.4, 0.27)	(37.6, 1.67)
11134	221	(0.0644, 0.86)	(0.326, 20.6)	(2.93, 18.8)	(0.86, 0.0644)	(14.3, 0.491)	(31.5, 1.73)
11434	164	(0.0461, 0.403)	(0.288, 13.9)	(2.94, 18.8)	(0.403, 0.0461)	(14.3, 0.327)	(27.3, 0.208)
11545	158	(0.0471, 0.627)	(0.294, 17.6)	(1.69, 17.3)	(0.627, 0.0471)	(14.2, 0.285)	(25.8, 0.829)
12760	183	(0.064, 0.833)	(0.309, 9.9)	(2.95, 28.8)	(0.833, 0.064)	(14.9, 0.169)	(32.7, 2.06)
31779	105	(0.0688, 0.836)	(0.242, 12.2)	(1.96, 47.6)	(0.836, 0.0688)	(11.1, 0.282)	(47.6, 1.96)
33250	131	(0.0652, 1.15)	(0.286, 13.8)	(1.38, 14.8)	(0.868, 0.0678)	(13.1, 0.459)	(29.7, 0.439)
33308	57	(0.0481, 0.42)	(0.273, 19.8)	(1.38, 14.8)	(0.42, 0.0481)	(11.6, 0.177)	(23.6, 0.505)
34174	68	(0.0527, 0.103)	(0.292, 13.7)	(1.62, 47.9)	(0.103, 0.0527)	(12.7, 0.258)	(47.9, 1.62)
34676	92	(0.05, 0.702)	(0.266, 7.56)	(1.38, 14.9)	(0.702, 0.05)	(11.1, 0.235)	(23.6, 0.801)
77992	72	(0.0645, 1.15)	(0.26, 7.55)	(1.08, 21.3)	(1.15, 0.0645)	(11.3, 0.22)	(28.7, 0.288)

Table 4 Distances between minima from BLN69-top10. The third column (d_{ced}) refers to the cumulative edge distance computed on the transition graph, as defined in section 2.4.6. The fourth column (num. edges) is the number of edges found on the shortest path joining the two minima in the transition graph: a value of $2p$ two means that the path goes through p saddles to connect the two minima of interest.

index m1	index m2	d_{ced}	path: num. edges	lRMSD	lRMSD/d_{ced}
1	6	0.195	2	0.163	0.835
1	8	0.112	2	0.098	0.875
1	142	0.115	2	0.106	0.921
1	311	1.182	2	0.515	0.435
1	1973	1.492	6	0.676	0.453
1	7305	1.425	6	0.655	0.459
1	11134	1.199	4	0.698	0.582
1	12760	1.399	6	0.712	0.508
1	33250	0.697	2	0.360	0.521
6	7	0.307	4	0.191	0.623
6	142	0.310	4	0.194	0.625
6	311	1.272	6	0.547	0.428
6	1974	1.332	6	0.676	0.507
6	7305	1.620	8	0.670	0.413
6	11134	1.395	6	0.715	0.512
6	12760	1.594	8	0.713	0.448
6	33250	0.893	4	0.409	0.459
8	142	0.227	4	0.146	0.643
8	311	1.294	8	0.519	0.401
8	1974	1.522	8	0.671	0.440
8	7305	1.414	6	0.641	0.453
8	11134	1.308	6	0.695	0.531
8	12760	1.504	8	0.707	0.470
8	33250	0.810213	4	0.372	0.459
142	311	1.29774	4	0.519	0.402
142	1974	1.52341	10	0.684	0.450
142	7305	1.54053	8	0.667	0.433
142	11134	1.31497	6	0.708	0.540
142	12760	1.51454	8	0.721	0.477
142	33250	0.812982	4	0.372	0.459
311	1974	1.58991	10	0.633	0.400
311	7305	1.57245	8	0.587	0.373
311	11134	1.52672	8	0.700	0.557
311	12760	1.69276	10	0.709	0.419
311	33250	1.29726	6	0.553	0.428
1974	7305	1.74716	8	0.672	0.386
1974	11134	1.01362	4	0.364	0.360
1974	12760	1.22234	6	0.396	0.324
1974	33250	1.37802	6	0.687	0.501
7305	11134	1.49398	6	0.625	0.419
7305	12760	1.69967	8	0.640	0.376
7305	33250	2.12317	8	0.727	0.342
11134	12760	0.208721	2	0.181	0.905
11134	33250	1.52891	6	0.727	0.478
12760	33250	1.55398	6	0.745	0.479

Table 5 Persistences of local minima from BLN69-top10.

Index	Persistence
1	inf
6	1.55
8	0.35
142	0.67
311	7.29
1974	7.42
7305	6.14
11134	7.15
12760	1.73
33250	7.11

7.2 Monitoring a Sampling Process

7.2.1 Energy Landscape Analysis

Figure 12 Disconnectivity tree, persistence diagram, and histogram of masses for the `T-RRT` run started near the local minimum of 1(gmin). The vertex weighted transition graph contains 21 vertices and 24 edges; it has 1 connected component and 4 cycles.

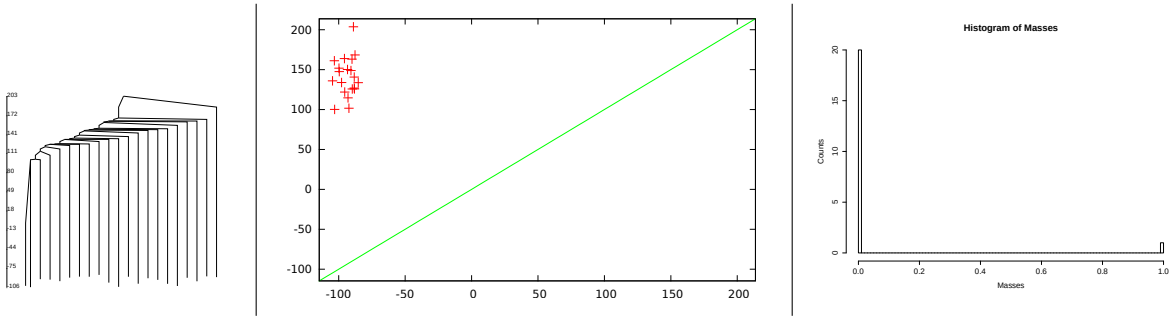


Figure 13 Disconnectivity tree, persistence diagram, and histogram of masses for the `T-RRT` run started near the local minimum of 6 . The vertex weighted transition graph contains 33 vertices and 60 edges; it has 1 connected component and 28 cycles.

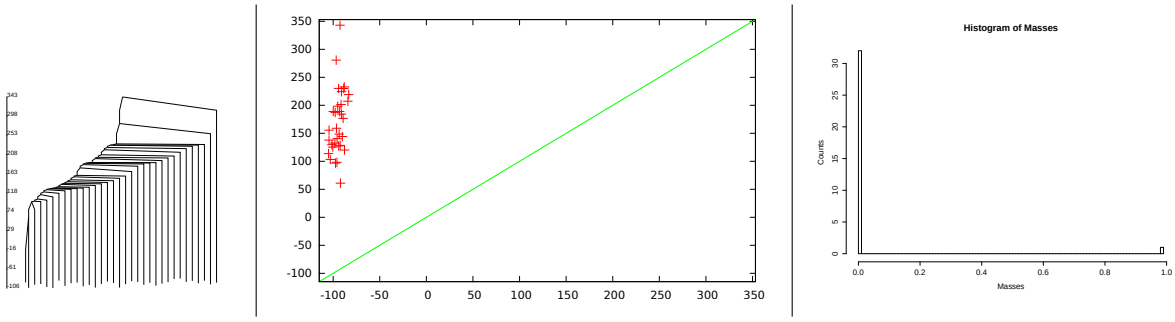


Figure 14 Disconnectivity tree, persistence diagram, and histogram of masses for the `T-RRT` run started near the local minimum of 8 . The vertex weighted transition graph contains 27 vertices and 30 edges; it has 1 connected component and 4 cycles.

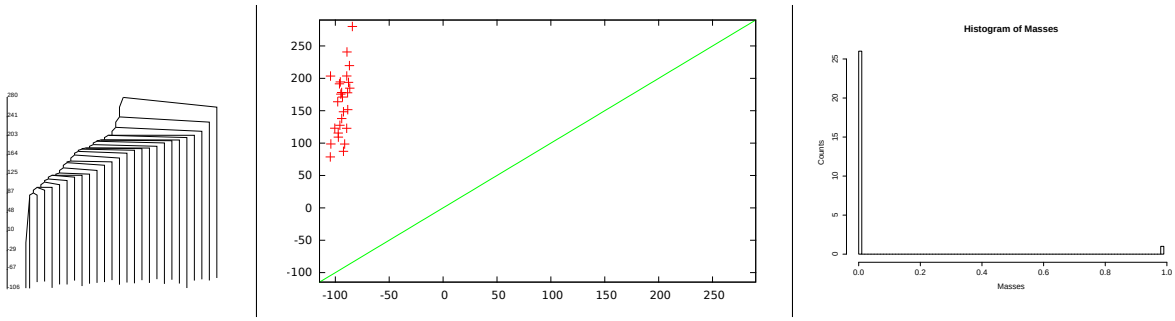


Figure 15 Disconnectivity tree, persistence diagram, and histogram of masses for the `T-RRT` run started near the local minimum of 142 . The vertex weighted transition graph contains 17 vertices and 18 edges; it has 1 connected component and 2 cycles.

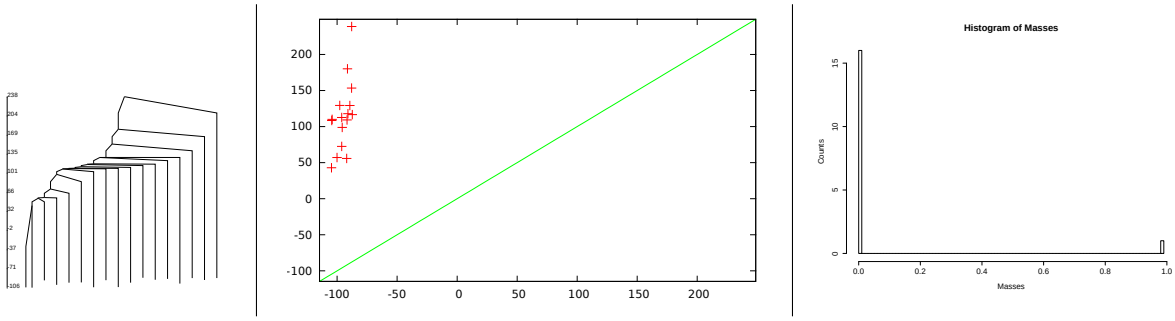


Figure 16 Disconnectivity tree, persistence diagram, and histogram of masses for the `T-RRT` run started near the local minimum of 311 . The vertex weighted transition graph contains 15 vertices and 14 edges; it has 1 connected component and 0 cycles.

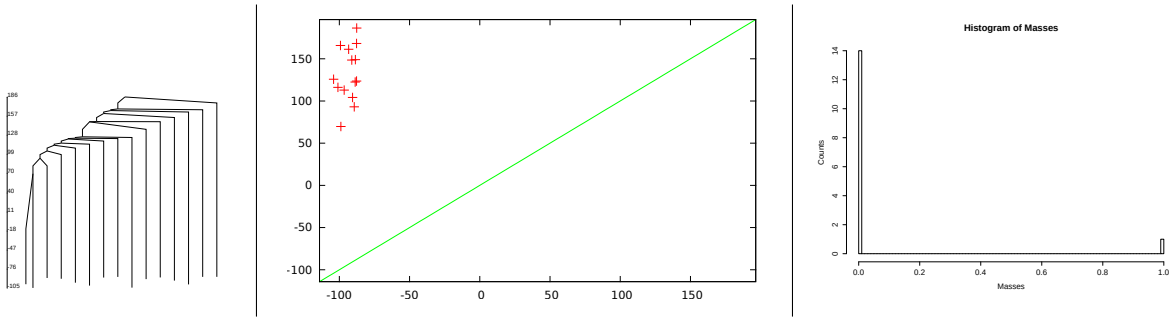


Figure 17 Disconnectivity tree, persistence diagram, and histogram of masses for the `T-RRT` run started near the local minimum of 1974 . The vertex weighted transition graph contains 46 vertices and 77 edges; it has 1 connected component and 32 cycles.

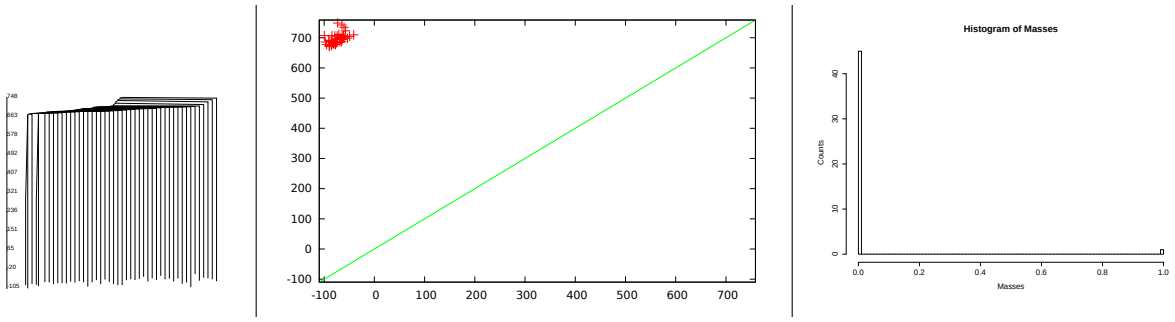


Figure 18 Disconnectivity tree, persistence diagram, and histogram of masses for the `T-RRT` run started near the local minimum of 7305 . The vertex weighted transition graph contains 90 vertices and 156 edges; it has 1 connected component and 67 cycles.

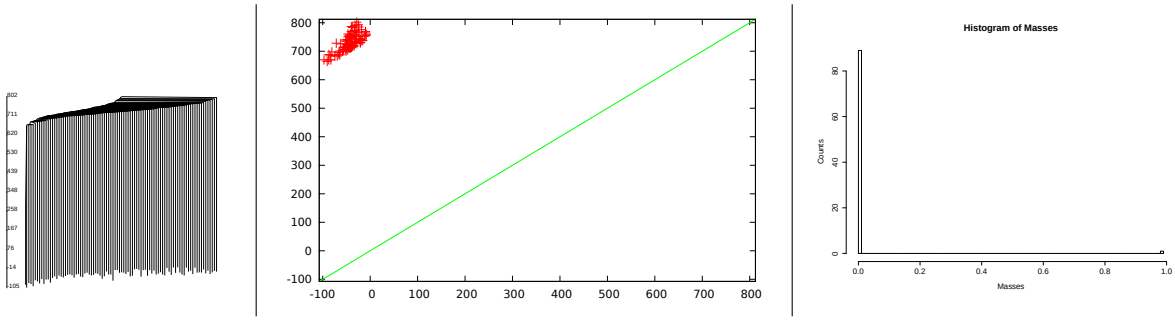


Figure 19 Disconnectivity tree, persistence diagram, and histogram of masses for the `T-RRT` run started near the local minimum of 11134 . The vertex weighted transition graph contains 168 vertices and 360 edges; it has 1 connected component and 193 cycles.

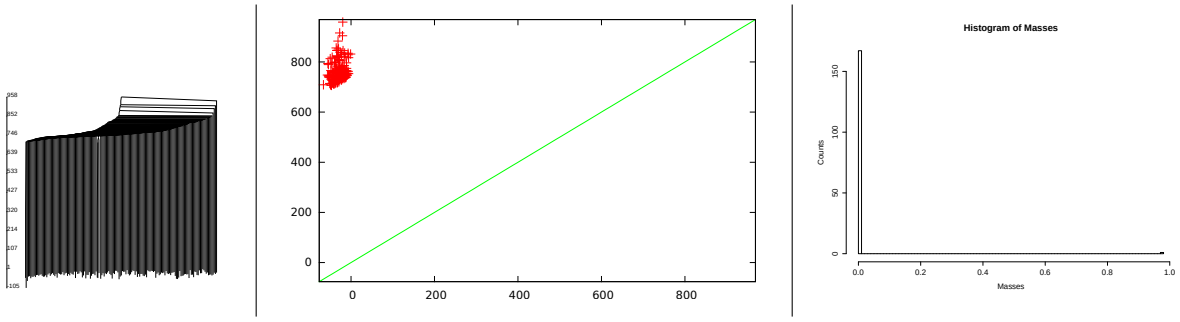


Figure 20 Disconnectivity tree, persistence diagram, and histogram of masses for the `T-RRT` run started near the local minimum of 12760 . The vertex weighted transition graph contains 439 vertices and 1700 edges; it has 1 connected component and 1262 cycles.

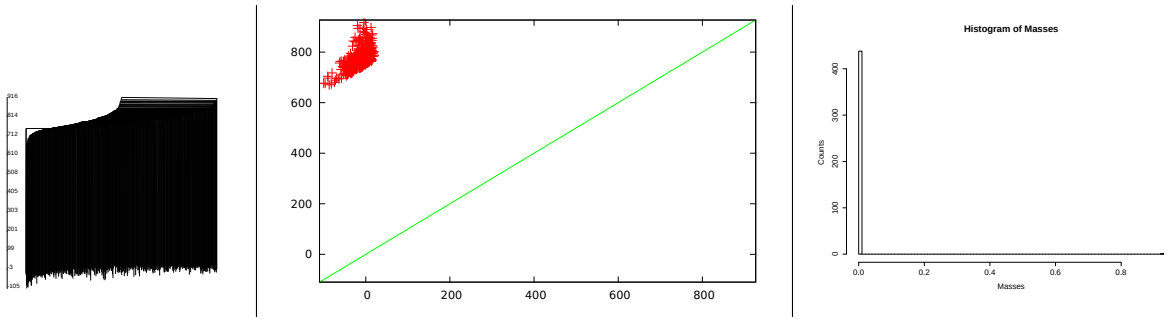
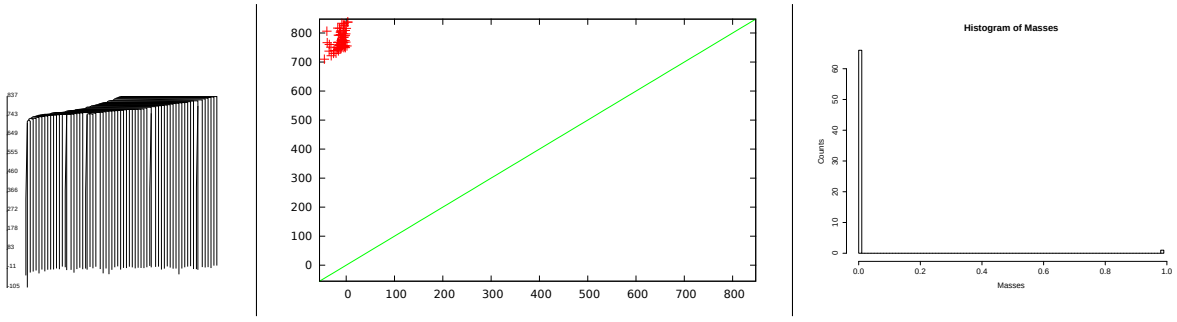


Figure 21 Disconnectivity tree, persistence diagram, and histogram of masses for the `T-RRT` run started near the local minimum of 33250 . The vertex weighted transition graph contains 67 vertices and 247 edges; it has 1 connected component and 181 cycles.



7.2.2 EMD: Connectivity Constraints

Table 6 Algorithm `Alg-EMD-LP`, dataset TRRT-top10: fraction of vertices of the source graph inducing a connected graph on the demand graph. The row (resp. column) index is that of the source (resp. demand) local minimum.

	1(gmin)	6	8	142	311	1974	7305	11134	12760	33250
1(gmin)	1.00	0.64	0.87	0.60	0.86	0.16	0.19	0.10	0.10	0.35
6	1.00	1.00	0.95	1.00	0.92	0.28	0.15	0.21	0.15	0.31
8	0.90	0.71	1.00	1.00	0.93	0.15	0.22	0.26	0.19	0.35
142	0.92	0.75	0.64	1.00	0.57	0.25	0.29	0.24	0.18	0.29
311	0.33	0.55	0.25	0.29	1.00	0.27	0.20	0.40	0.13	0.40
1974	0.93	0.65	0.53	0.71	1.00	1.00	0.27	0.20	0.22	0.36
7305	0.89	0.83	0.60	0.95	0.95	0.80	1.00	0.62	0.21	0.92
11134	1.00	0.93	0.89	0.96	1.00	0.88	0.87	1.00	0.17	0.86
12760	1.00	0.98	0.99	0.96	1.00	0.97	0.89	0.77	1.00	0.90
33250	0.84	0.68	0.81	0.67	0.94	0.74	0.42	0.54	0.21	1.00

Table 7 Algorithm `Alg-EMD-LP`, dataset TRRT-top10: fraction of edges of the source graph inducing a connected graph on the demand graph. The row (resp. column) index is that of the source (resp. demand) local minimum.

	1(gmin)	6	8	142	311	1974	7305	11134	12760	33250
1(gmin)	1.00	1.00	0.82	1.00	1.00	0.82	0.83	0.83	0.83	0.83
6	1.00	1.00	0.97	0.92	1.00	0.64	0.63	0.63	0.63	0.63
8	0.96	0.89	1.00	1.00	1.00	0.90	0.87	0.87	0.87	0.90
142	0.92	0.92	1.00	1.00	1.00	0.94	0.94	0.89	0.89	0.89
311	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1974	1.00	0.70	0.94	0.88	1.00	1.00	0.60	0.58	0.58	0.69
7305	1.00	1.00	1.00	1.00	1.00	0.90	1.00	0.60	0.59	1.00
11134	1.00	1.00	1.00	0.97	1.00	1.00	0.97	1.00	0.03	0.98
12760	1.00	0.99	1.00	1.00	1.00	1.00	0.98	0.90	1.00	0.95
33250	0.96	0.92	0.93	0.95	1.00	0.96	0.47	0.30	0.27	1.00

Table 8 Algorithm `Alg-EMD-LP`, dataset TRRT-top10-U: fraction of vertices of the source graph inducing a connected graph on the demand graph. The row (resp. column) index is that of the source (resp. demand) local minimum.

	1(gmin)	6	8	142	311	1974	7305	11134	12760	33250
1(gmin)	1.00	0.14	0.10	0.52	0.52	0.10	0.05	0.05	0.05	0.05
6	0.58	1.00	0.67	0.64	0.70	0.06	0.06	0.03	0.03	0.06
8	0.63	0.15	1.00	0.67	0.59	0.07	0.04	0.04	0.00	0.07
142	0.18	0.24	0.18	1.00	0.65	0.06	0.06	0.06	0.12	0.06
311	0.13	0.07	0.07	0.13	1.00	0.07	0.07	0.07	0.07	0.07
1974	0.72	0.57	0.54	0.78	0.83	1.00	0.09	0.02	0.02	0.11
7305	0.83	0.72	0.80	0.87	0.96	0.60	1.00	0.06	0.01	0.54
11134	0.93	0.86	0.87	0.92	0.93	0.79	0.55	1.00	0.01	0.69
12760	0.96	0.95	0.95	0.97	0.97	0.92	0.82	0.69	1.00	0.87
33250	0.78	0.70	0.75	0.79	0.85	0.58	0.15	0.03	0.01	1.00

Table 9 Algorithm `Alg-EMD-LP`, dataset TRRT-top10-U: fraction of edges of the source graph inducing a connected graph on the demand graph. The row (resp. column) index is that of the source (resp. demand) local minimum.

	1(gmin)	6	8	142	311	1974	7305	11134	12760	33250
1(gmin)	1.00	0.83	0.83	0.08	0.83	0.83	0.83	0.83	0.79	0.83
6	0.63	1.00	0.58	0.60	0.22	0.10	0.22	0.12	0.08	0.12
8	0.90	0.90	1.00	0.20	0.90	0.87	0.87	0.87	0.20	0.07
142	0.06	0.89	0.17	1.00	0.94	0.11	0.11	0.89	0.72	0.06
311	1.00	0.07	1.00	1.00	1.00	0.07	0.07	1.00	0.64	1.00
1974	0.60	0.10	0.62	0.16	0.12	1.00	0.58	0.58	0.45	0.04
7305	0.65	0.12	0.63	0.17	0.18	0.63	1.00	0.01	0.01	0.59
11134	0.57	0.20	0.57	0.54	0.56	0.49	0.04	1.00	0.43	0.49
12760	0.42	0.17	0.37	0.41	0.42	0.31	0.04	0.27	1.00	0.35
33250	0.40	0.16	0.15	0.13	0.40	0.06	0.27	0.27	0.23	1.00

7.2.3 EMD: Demand Satisfaction

Table 10 Algorithm `Alg-EMD-CCC-G`, dataset TRRT-top10: percentage of the demand satisfaction. The row (resp. column) index is that of the source (resp. demand) local minimum.

	1(gmin)	6	8	142	311	1974	7305	11134	12760	33250
1(gmin)	100.00	100.00	100.00	99.90	99.92	100.00	100.00	99.98	99.89	100.00
6	99.72	100.00	99.79	99.66	99.64	100.00	100.00	99.98	99.88	100.00
8	99.91	100.00	100.00	99.86	99.85	100.00	100.00	100.00	99.89	100.00
142	100.00	100.00	100.00	100.00	99.96	100.00	100.00	99.99	99.89	100.00
311	100.00	100.00	100.00	100.00	100.00	100.00	99.99	99.99	99.87	99.95
1974	99.51	99.62	99.39	99.29	99.43	100.00	100.00	100.00	99.99	100.00
7305	98.55	98.83	98.62	98.49	98.47	99.04	100.00	100.00	100.00	99.41
11134	95.70	95.97	95.76	95.64	95.62	96.17	97.10	100.00	100.00	96.53
12760	93.77	94.03	93.83	93.71	93.69	94.23	95.14	97.98	100.00	94.57
33250	99.14	99.41	99.21	99.08	99.06	99.63	100.00	100.00	99.86	100.00

Table 11 Algorithm `Alg-EMD-CCC-G`, dataset TRRT-top10-U: percentage of the demand satisfaction. The row (resp. column) index is that of the source (resp. demand) local minimum.

	1(gmin)	6	8	142	311	1974	7305	11134	12760	33250
1(gmin)	100.00	72.29	71.43	14.29	69.52	66.05	40.32	40.48	18.91	56.65
6	66.67	100.00	55.89	62.92	59.39	58.63	45.15	29.38	18.00	63.18
8	75.66	80.81	100.00	79.52	84.44	67.87	48.52	44.91	19.13	66.28
142	10.64	65.60	61.44	100.00	70.98	55.88	31.44	36.17	17.96	54.78
311	76.19	64.85	55.56	88.24	100.00	15.36	33.33	31.07	17.15	56.42
1974	19.98	65.42	63.61	71.87	64.20	100.00	48.45	32.69	21.00	55.13
7305	19.21	24.34	11.48	83.59	14.44	77.78	100.00	51.07	1.57	67.94
11134	19.05	6.55	7.74	11.90	13.69	70.68	3.41	100.00	47.69	68.83
12760	21.62	8.66	10.02	14.35	61.41	59.80	5.87	51.48	100.00	56.35
33250	50.04	25.60	11.94	11.94	13.43	7.46	47.91	32.19	19.54	100.00

7.2.4 Transport Costs

Table 12 Algorithm `Alg-EMD-LP`, dataset TRRT-top10: transport costs. The row (resp. column) index is that of the source (resp. demand) local minimum.

	1(gmin)	6	8	142	311	1974	7305	11134	12760	33250
1(gmin)	0.00	0.16	0.10	0.11	0.51	0.67	0.65	0.68	0.69	0.36
6	0.16	0.00	0.19	0.19	0.54	0.67	0.66	0.69	0.70	0.41
8	0.10	0.19	0.00	0.15	0.52	0.67	0.63	0.67	0.69	0.38
142	0.11	0.19	0.15	0.00	0.52	0.68	0.66	0.69	0.70	0.38
311	0.51	0.54	0.52	0.52	0.00	0.63	0.58	0.68	0.69	0.55
1974	0.67	0.67	0.67	0.68	0.63	0.00	0.67	0.36	0.39	0.69
7305	0.65	0.66	0.63	0.66	0.58	0.66	0.00	0.61	0.62	0.72
11134	0.67	0.69	0.67	0.68	0.67	0.36	0.60	0.00	0.05	0.70
12760	0.67	0.69	0.67	0.68	0.67	0.37	0.60	0.05	0.00	0.70
33250	0.36	0.41	0.37	0.37	0.55	0.68	0.72	0.70	0.72	0.00

Table 13 Algorithm `Alg-EMD-CCC-G`, dataset TRRT-top10: transport costs. The row (resp. column) index is that of the source (resp. demand) local minimum.

	1(gmin)	6	8	142	311	1974	7305	11134	12760	33250
1(gmin)	0.00	0.16	0.10	0.11	0.51	0.67	0.65	0.68	0.69	0.36
6	0.16	0.00	0.19	0.19	0.54	0.67	0.66	0.69	0.70	0.41
8	0.10	0.19	0.00	0.15	0.52	0.67	0.63	0.67	0.69	0.38
142	0.11	0.19	0.15	0.00	0.52	0.68	0.66	0.69	0.70	0.38
311	0.51	0.54	0.52	0.52	0.00	0.63	0.58	0.68	0.69	0.55
1974	0.67	0.67	0.67	0.68	0.63	0.00	0.67	0.36	0.39	0.69
7305	0.65	0.66	0.63	0.66	0.58	0.67	0.00	0.61	0.62	0.72
11134	0.68	0.69	0.67	0.69	0.68	0.36	0.61	0.00	0.06	0.70
12760	0.69	0.71	0.69	0.70	0.69	0.39	0.62	0.06	0.00	0.72
33250	0.36	0.41	0.38	0.38	0.55	0.69	0.72	0.70	0.72	0.00

Table 14 Algorithm `Alg-EMD-LP`, dataset TRRT-top10-U: transport costs. The row (resp. column) index is that of the source (resp. demand) local minimum.

	1(gmin)	6	8	142	311	1974	7305	11134	12760	33250
1(gmin)	0.00	0.27	0.22	0.23	0.55	0.87	1.19	1.30	1.61	1.47
6	0.27	0.00	0.27	0.28	0.58	0.87	1.19	1.29	1.59	1.45
8	0.22	0.27	0.00	0.26	0.55	0.86	1.18	1.28	1.60	1.46
142	0.23	0.28	0.26	0.00	0.54	0.87	1.19	1.30	1.61	1.47
311	0.55	0.58	0.55	0.54	0.00	0.85	1.19	1.24	1.58	1.44
1974	0.87	0.87	0.86	0.87	0.85	0.00	1.10	1.13	1.39	1.39
7305	1.19	1.19	1.18	1.19	1.19	1.10	0.00	1.20	1.30	1.37
11134	1.30	1.29	1.28	1.30	1.24	1.13	1.20	0.00	1.28	1.31
12760	1.61	1.59	1.60	1.61	1.58	1.39	1.30	1.28	0.00	1.43
33250	1.47	1.45	1.46	1.47	1.44	1.39	1.37	1.31	1.43	0.00

Table 15 Algorithm `Alg-EMD-CCC-G`, dataset TRRT-top10-U: transport costs. The row (resp. column) index is that of the source (resp. demand) local minimum.

	1(gmin)	6	8	142	311	1974	7305	11134	12760	33250
1(gmin)	0.00	0.17	0.14	0.01	0.37	0.54	0.43	0.47	0.22	0.78
6	0.16	0.00	0.13	0.16	0.33	0.47	0.48	0.33	0.21	0.87
8	0.15	0.20	0.00	0.19	0.46	0.55	0.52	0.52	0.22	0.92
142	0.00	0.15	0.13	0.00	0.37	0.45	0.32	0.42	0.21	0.76
311	0.42	0.37	0.29	0.47	0.00	0.11	0.33	0.34	0.19	0.76
1974	0.14	0.54	0.52	0.60	0.51	0.00	0.49	0.34	0.23	0.72
7305	0.17	0.23	0.09	0.99	0.12	0.85	0.00	0.61	0.01	0.92
11134	0.20	0.06	0.07	0.12	0.13	0.79	0.03	0.00	0.60	0.90
12760	0.26	0.10	0.12	0.17	0.89	0.80	0.06	0.63	0.00	0.78
33250	0.68	0.31	0.13	0.13	0.15	0.08	0.65	0.40	0.25	0.00

Figure 22 Correlation between the transport costs delivered by `Alg-EMD-LP` and `Alg-EMD-CCC-G` (both ways), on runs of `T-RRT` started near the 10 lowest local minima. The Pearson coefficients is at least equal to 0.999 for the three pairs.

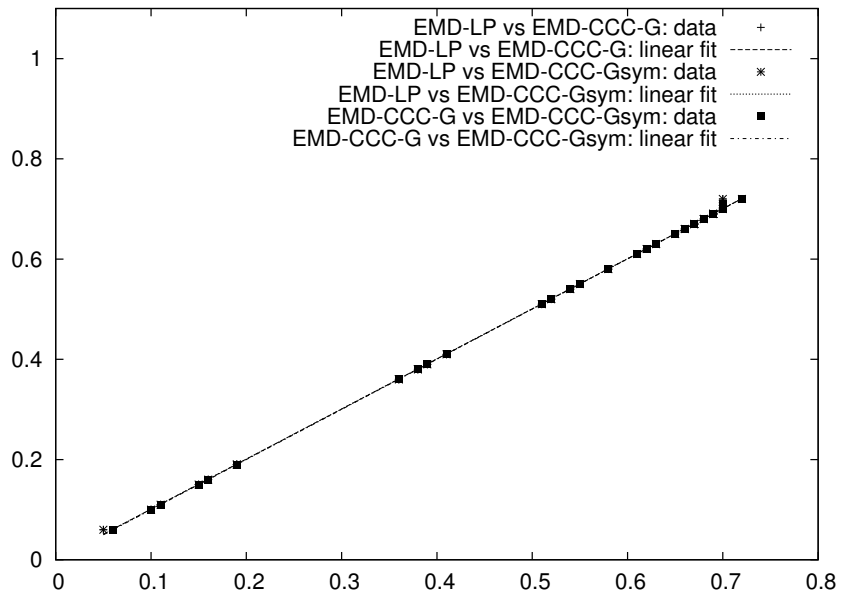
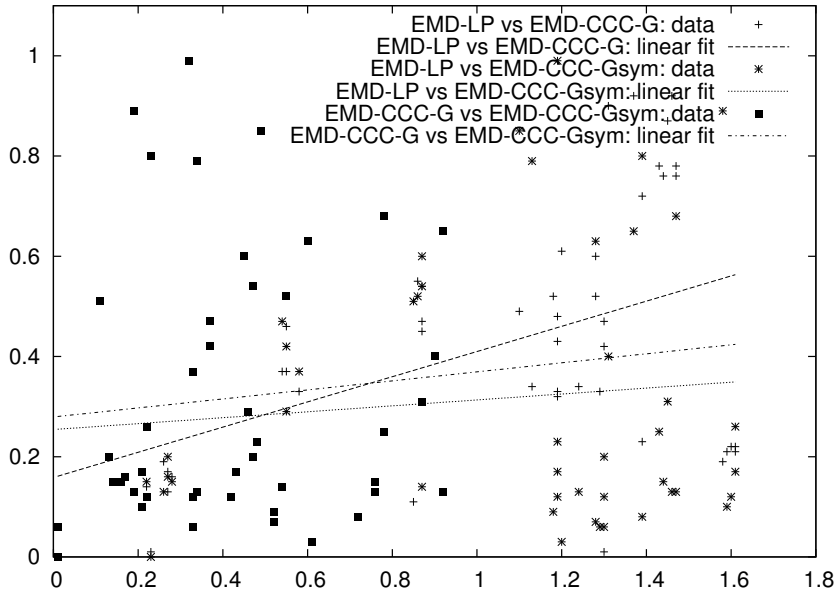


Figure 23 Correlation between the transport costs delivered by `Alg-EMD-LP` and `Alg-EMD-CCC-G` (both ways), on runs of `T-RRT` started near the 10 lowest local minima, with uniform masses (see text for details). The Pearson coefficients are respectively equal to 0.25, 0.09 and 0.06 for the three pairs.



8 Supplemental: Software

Our code is developed in generic C++ within the Structural Bioinformatics Library[4] based upon the Computational Geometry Algorithms Library [cga]. In the following, we describe the file formats used by the applications, and proceed with a concise presentation of each of them.

8.1 Specifications and File Formats

Three objects are manipulated, and we present them in turn.

Conformations and conformational ensembles. We use two formats to store conformations in Cartesian coordinates. The first one is the classical PDB format. The second one is the text Point_d format, where a line consists of an integer indicating the number n of variables, followed by values for these n variables. Using this format, one line codes one conformation, so that a file of such lines codes a conformational ensemble.

As an example, the conformations of a n atom molecules read as:

```
...
3n x_1 y_1 z_1 ... x_n y_n z_n
...
```

[4]Upon publication of the paper, the programs described thereafter will be made available from the following web site: `http://structural-bioinformatics-library.org/`

Remark 8 *The analysis presented in this paper only resort to Cartesian coordinates, whence the aforementioned representations. However, all algorithms relying upon nearest neighbors and nearest neighbor graphs are fully generic, and can be use in conjunction with other distances. Users willing to use alternative representations (in particular internal coordinates) and the associate distances, will only have to instantiate our generic classes to do so.*

Sampled Energy Landscape. A sampled EL is a conformational ensemble such that each conformation is endowed with an energy. To decouple the conformations and the energies, all energies are listed in a separate file whose n-th line corresponds to the n-th conformation in the ensemble. (For this reason, loading an energy file can only be done once a file for conformations has been loaded.) Here is an example such file:

```
-105.2
-104.8
...
```

Transition Graph. A TG is a graph whose nodes are (substitutes for) critical points, namely local minima and index one saddles. An edge in the TG corresponds to the existence of a path between pairs of critical points. As seen in the main text, our analysis use two types of TG:

- *the vertex transition graph:* both minima and transition states are encoded as vertices, and may be decorated with information (energy, volume of the *basin* associated with a minimum). An edge connects one local minimum and one transition state.
- *the compressed transition graph:* the minima are represented by the vertices, and additional pieces of information (transition states and their energy) are associated with edges connecting the minima.

We have seen in the main text that TG can be built either from a database of critical points, or from a conformational ensemble. Both strategies rely on various pieces of information, which are specified as follows:

- If a DB of critical points is used, we expect the pairs of local minima linked to listed in a text file, with indices starting at 0:

  ```
  0 2
  1 10
  ...
  ```

- To decorate local minima with additional information, one uses additional files: one (optional) file for the conformations associated with the local minima; one (optional) file for the energies of the local minima. To map these information to their respective local minima, it is assumed that local minima have indices starting at zero.
- To decorate the edges linking pairs of local minima, one proceeds similarly, with one (optional) file for transition states, and one (optional) file for their energies. To map these pieces of information to the edges, it is assumed that the ordering of edges and that of the associated pieces of information are coherent.

Once a TG has been constructed, it is stored as a compressed transition graph. Practically, the Boost Graph Library and the associated Serialization mechanisms are used, producing an XML file.

Remark 9 *To handle data generated by molecular dynamics packages, the user is referred to packages such as* `mdanalysis`*, see [MADWB11] and* `https://code.google.com/p/mdanalysis/`*.*

8.2 Applications : Executables and Calls

The analysis presented in this work are carried out by six executables:

- the analysis of conformation ensembles : sbl-conf-ensemble-analysis.exe and sbl-conf-ensemble-comparison.exe;
- the generation of transition graphs: sbl-transition-graph-builder-from-DB-of-critical-points.exe and sbl-transition-graph-builder-from-sampled-energy-landscape.exe;
- the analysis of the transition graphs: sbl-landscape–analysis.exe and sbl-landscape-comparison.exe;

In the following, we provide a demo run, then the list of all options obtained when using the option *help* of the corresponding application.

8.2.1 sbl-conf-ensemble-analysis.exe

```
./sbl-conf-ensemble-analysis.exe -v --points-file <path/to/confs>\
 --num-neighbors 10 --nng-connected 15

Various analysis on conformational ensembles: sampling diversity,
density-based clustering, comparison to reference sets of critical
points. The input consists of conformations in PDB or Point_d format.

Copyright Inria / Algorithms - Biology - Structure, 2015.
Version: 1.0.0

General Options:
  -h [ --help ] [=arg(=1)] (=0)         Print this help message.
  --workflow [=arg(=1)] (=0)            Print the workflow of the application
                                        without any run.
  -l [ --log ] [=arg(=1)] (=0)          Put the log in a file.
  -v [ --verbose ] [=arg(=1)] (=0)      Dumps high level statistics.
  -c [ --colored-log ] [=arg(=1)] (=0)  Color the log (red for module and
                                        loader names, green for statistics,
                                        blue for report).
  -d [ --directory-output ] arg (=.)    Output directory containing the output
                                        files (default is current).
  -o [ --output-prefix ] [=arg(=IMPLICIT)]
                                        Prefix to add to all output file names.
Optional Modules:
  --sampling-diversity [=arg(=1)] (=0) Run Sampling Diversity Analysis
  --nng-builder [=arg(=1)] (=0)        Run the NNG Builder
  --mst [=arg(=1)] (=0)                Run Minimal Spanning Tree Analysis
  --mtb [=arg(=1)] (=0)                Run Morse Theory Based Analysis
Conformations loader:
  --pdb-files arg                   File listing the PDB files of one
                                    ensemble.
  --points-file arg                 File listing the conformations in Point_d
                                    format.
  -w [ --water ] [=arg(=1)] (=0)    Load water molecules (default is false).
  --hetatoms [=arg(=1)] (=0)        Load hetero-atoms (default is false).
  --hydrogens [=arg(=1)] (=0)       Load hydrogens (default is false).
  -a [ --alternate ] arg (= )       Alternate coordinates: alternative to be
                                    used (char).
  -p [ --occupancy-policy ] arg (=1) Selection policy for atoms with occupancy
                                    not equal to 1.
                                    Allowed values are:
                                     -p 1 (all, default)
                                     -p 2 (forbbiden)
                                     -p 3 (none)
                                     -p 4 (max)
                                     -p 5 (min)
  -B [ --B-factor-limit ] arg (=80) Threshold for temperature factors (default
```

```
                                    max value is 80).
-m [ --model-number ] arg (=1)      ID of the model to be loaded from the PDB
                                    file (default is 1).
--load-chains arg                   Subset of chains to be loaded.
--save-ensembles [=arg(=1)] (=0)    Save each loaded ensemble in Point_d
                                    format in a plain text file.
```

```
Nearest Neighbors Graph:
  --num-neighbors arg   Target number  of neighbors for each vertex.
  --distance-range arg  Distance range specification.
  --nng-connected arg   Attempt connecting the NNG by iteratively increasing
                        the number of neighbors; halt if NNG is connected, or
                        if the prescribed num of neighbors is reached.
  --nng-file arg        Skip calculations by loading the NNG from a file.

Morse Theory Based Analyzer:
  --persistence-threshold arg (=-1)    Persistence threshold (default -1, i.e
                                       no persistence).
  --sublevelset-threshold arg (=0)     Sublevelset threshold (default 0, i.e no
                                       sublevelset).
  --no-popup [=arg(=1)] (=0)           Prevent the graphical display of the
                                       persistence diagram and the
                                       disconnectivity forest.
  --show-trees arg (=0)                Show only the n trees with the largest
                                       number of leaves in the disconnecitivity
                                       forest (default: show the whole forest).
  --delta-height arg (=-1)             Club the saddles in the displayed
                                       disconnectivity forest within intervals
                                       of the input size.
  --split-basins [=arg(=1)] (=0)       Report each persistent basin into a
                                       separate plain text file (default is: a
                                       single serialized xml file).
  --MSW-filename arg                   Load the MSW chain complex instead of
                                       computing it.
  --normalize-gradient [=arg(=1)] (=0) Normalize the gradient on the NNG by the
                                       distance between the vertices.
```

8.2.2 sbl-conf-ensemble-comparison.exe

```
./sbl-conf-ensemble-analysis.exe -v --points-file <path/to/confs1>\
 --points-file <path/to/confs2>

Compare two conformational ensembles using the Hausdorff distance
between the two ensembles. The input consists of conformations in
PDB or Point_d format.

Copyright Inria / Algorithms - Biology - Structure, 2015.
Version: 1.0.0

General Options:
  -h [ --help ] [=arg(=1)] (=0)         Print this help message.
  --workflow [=arg(=1)] (=0)            Print the workflow of the application
                                        without any run.
  -l [ --log ] [=arg(=1)] (=0)          Put the log in a file.
  -v [ --verbose ] [=arg(=1)] (=0)      Dumps high level statistics.
  -c [ --colored-log ] [=arg(=1)] (=0)  Color the log (red for module and
                                        loader names, green for statistics,
                                        blue for report).
  -d [ --directory-output ] arg (=.)    Output directory containing the output
                                        files (default is current).
  -o [ --output-prefix ] [=arg(=IMPLICIT)]
                                        Prefix to add to all output file names.
Optional Modules:
  --Hausdorff [=arg(=1)] (=0)           Hausdorff distance between two sets of conformations
  --symmetric-difference [=arg(=1)] (=0)
                                        Symmetric difference between two sets
                                        of conformations
Conformations loader:
  --pdb-files arg                   File listing the PDB files of one
                                    ensemble.
  --points-file arg                 File listing the conformations in Point_d
                                    format.
  -w [ --water ] [=arg(=1)] (=0)    Load water molecules (default is false).
  --hetatoms [=arg(=1)] (=0)        Load hetero-atoms (default is false).
  --hydrogens [=arg(=1)] (=0)       Load hydrogens (default is false).
  -a [ --alternate ] arg (= )       Alternate coordinates: alternative to be
                                    used (char).
  -p [ --occupancy-policy ] arg (=1) Selection policy for atoms with occupancy
                                    not equal to 1.
                                    Allowed values are:
                                     -p 1 (all, default)
                                     -p 2 (forbbiden)
                                     -p 3 (none)
                                     -p 4 (max)
                                     -p 5 (min)
  -B [ --B-factor-limit ] arg (=80) Threshold for temperature factors (default
```

```
                                    max value is 80).
-m [ --model-number ] arg (=1)      ID of the model to be loaded from the PDB
                                    file (default is 1).
--load-chains arg                   Subset of chains to be loaded.
--save-ensembles [=arg(=1)] (=0)    Save each loaded ensemble in Point_d
                                    format in a plain text file.
```

```
Hausdorff Distance for Point Clouds:
  --one-sided-hausdorff arg Compute only the one sided Hausdorff Distance

Symmetric Difference:
  --identity-threshold arg (=1.0000000000000001e-05)
                                        Threshold below which two conformations
                                        are considered identical for the
                                        considered distance
(e.g. least-RMSD).
```

8.2.3 sbl-transition-graph-builder-from-DB-of-critical-points.exe

```
./sbl-transition-graph-builder-from-DB-of-critical-points.exe\
--points-file <path/to/confs1> --energies </path/to/energies1>\
--points-file <path/to/confs2> --energies </path/to/energies2>\
--transition-edges <path/to/edges> --discard-loops -v

Compute the compressed transition graph associated to a data base of
critical points. The input consists of five files: the conformations
of minima and transitions in PDB or Point_d format, the energies of
minima and transitions, and the list of pairs of minima linked by
each transition.

Copyright Inria / Algorithms - Biology - Structure, 2015.
Version: 1.0.0

General Options:
  -h [ --help ] [=arg(=1)] (=0)         Print this help message.
  --workflow [=arg(=1)] (=0)            Print the workflow of the application
                                        without any run.
  -l [ --log ] [=arg(=1)] (=0)          Put the log in a file.
  -v [ --verbose ] [=arg(=1)] (=0)      Dumps high level statistics.
  -c [ --colored-log ] [=arg(=1)] (=0)  Color the log (red for module and
                                        loader names, green for statistics, blue for report).
  -d [ --directory-output ] arg (=.)    Output directory containing the output
                                        files (default is current).
  -o [ --output-prefix ] [=arg(=IMPLICIT)]
                                        Prefix to add to all output file names.
Conformations loader:
  --pdb-files arg                   File listing the PDB files of one ensemble.
  --points-file arg                 File listing the conformations in Point_d format.
  -w [ --water ] [=arg(=1)] (=0)    Load water molecules (default is false).
  --hetatoms [=arg(=1)] (=0)        Load hetero-atoms (default is false).
  --hydrogens [=arg(=1)] (=0)       Load hydrogens (default is false).
  -a [ --alternate ] arg (= )       Alternate coordinates: alternative to be used (char).
  -p [ --occupancy-policy ] arg (=1) Selection policy for atoms with occupancy
                                    not equal to 1.
                                    Allowed values are:
                                     -p 1 (all, default)
                                     -p 2 (forbbiden)
                                     -p 3 (none)
                                     -p 4 (max)
                                     -p 5 (min)
  -B [ --B-factor-limit ] arg (=80) Threshold for temperature factors (default
                                    max value is 80).
  -m [ --model-number ] arg (=1)    ID of the model to be loaded from the PDB
                                    file (default is 1).
  --load-chains arg                 Subset of chains to be loaded.
  --save-ensembles [=arg(=1)] (=0)  Save each loaded ensemble in Point_d
```

```
                                  format in a plain text file.
  --energies arg                     File providing one energy per
                                     conformation of an ensemble (txt file).
  --transition-edges arg             File listing the transitions between
                                     minima as edges (min_i, min_j).
```

```
Transition Graph Builder from DB of Critical Points:
  --discard-loops [=arg(=1)] (=0) Discard all the transition states incident to
                                  only one minimum.
```

8.2.4 sbl-transition-graph-builder-from-sampled-energy-landscape.exe

```
./sbl-transition-graph-builder-from-sampled-energy-landscape.exe\
--points-file <path/to/confs> --energies </path/to/energies>\
--samples-to-local-mins-map <path/to/samples-to-local-mins>\
--points-file <path/to/local-mins> --energies </path/to/local-mins>\
--num-neighbors 10 --nng-connected 15 -v

Compute the compressed transition graph associated to a sampled energy
landscape based on nearest neighbors and watershed tranform analysis.
The input consists of conformations in PDB or Point_d format.

General Options:
  -h [ --help ] [=arg(=1)] (=0)         Print this help message.
  --workflow [=arg(=1)] (=0)            Print the workflow of the application
                                        without any run.
  -l [ --log ] [=arg(=1)] (=0)          Put the log in a file.
  -v [ --verbose ] [=arg(=1)] (=0)      Dumps high level statistics.
  -c [ --colored-log ] [=arg(=1)] (=0)  Color the log (red for module and
                                        loader names, green for statistics,
                                        blue for report).
  -d [ --directory-output ] arg (=.)    Output directory containing the output
                                        files (default is current).
  -o [ --output-prefix ] [=arg(=IMPLICIT)]
Conformations loader:
  --pdb-files arg                   File listing the PDB files of one
                                    ensemble.
  --points-file arg                 File listing the conformations in Point_d
                                    format.
  -w [ --water ] [=arg(=1)] (=0)    Load water molecules (default is false).
  --hetatoms [=arg(=1)] (=0)        Load hetero-atoms (default is false).
  --hydrogens [=arg(=1)] (=0)       Load hydrogens (default is false).
  -a [ --alternate ] arg (= )       Alternate coordinates: alternative to be
                                    used (char).
  -p [ --occupancy-policy ] arg (=1) Selection policy for atoms with occupancy
                                    not equal to 1.
                                    Allowed values are:
                                     -p 1 (all, default)
                                     -p 2 (forbbiden)
                                     -p 3 (none)
                                     -p 4 (max)
                                     -p 5 (min)
  -B [ --B-factor-limit ] arg (=80) Threshold for temperature factors (default
                                    max value is 80).
  -m [ --model-number ] arg (=1)    ID of the model to be loaded from the PDB
                                    file (default is 1).
  --load-chains arg                 Subset of chains to be loaded.
  --save-ensembles [=arg(=1)] (=0)  Save each loaded ensemble in Point_d
                                    format in a plain text file.
```

	Prefix to add to all output file names.
--energies arg	File providing one energy per conformation of an ensemble (txt file).
--samples-to-mins arg	File listing the pairs of indices of (sample_i, local_min_i).

```
Nearest Neighbors Graph (NNG):
  --num-neighbors arg   Target number  of neighbors for each vertex.
  --distance-range arg  Distance range specification.
  --nng-connected arg   Attempt connecting the NNG by iteratively increasing
                        the number of neighbors; halt if NNG is connected, or
                        if the prescribed num of neighbors is reached.
  --nng-file arg        Skip calculations by loading the NNG from a file.

Watershed Transform for Minima:
  --persistence-threshold arg (=0)      Persistence threshold (default 0, i.e
                                        no persistence).
  --normalized-gradient [=arg(=1)] (=0) Gradient vector field of the height
                                        function estimated by the normalized
                                        gradient.
  --interactive [=arg(=1)] (=0)         Interactive tunning of the persistence
                                        threshold.
  --no-popup [=arg(=1)] (=0)            Prevent the graphical display of the
                                        persistence diagram and the
                                        disconnectivity forest.
  --delta-height arg (=-1)              Club the saddles located in the height
                                        slice when drawing the disconnectivity
                                        forest.
  --split-basins [=arg(=1)] (=0)        Report each persistent basin into a
                                        separate plain text file (default is: a
                                        single serialized xml file).
```

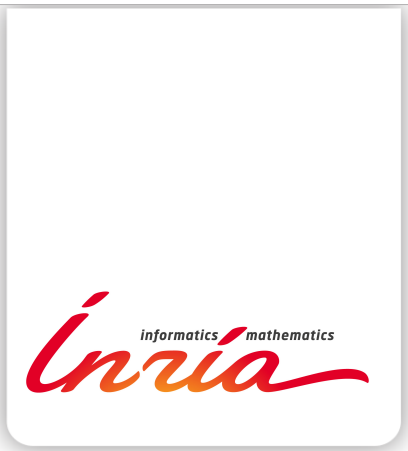
8.2.5 sbl-landscape–analysis.exe

```
./sbl-energy-landscape-analysis.exe -v --transition-graph <path/to/tg>\
 --run-mode t --run-mode p

Various analysis on energy landscapes: graph topology, landmark paths,
watershed transform.  The input consists of a compressed transition
graph.

Copyright Inria / Algorithms - Biology - Structure, 2015.
Version: 1.0.0

General Options:
  -h [ --help ] [=arg(=1)] (=0)         Print this help message.
  --workflow [=arg(=1)] (=0)            Print the workflow of the application
                                        without any run.
  -l [ --log ] [=arg(=1)] (=0)          Put the log in a file.
  -v [ --verbose ] [=arg(=1)] (=0)      Dumps high level statistics.
  -c [ --colored-log ] [=arg(=1)] (=0)  Color the log (red for module and
                                        loader names, green for statistics,
                                        blue for report).
  -d [ --directory-output ] arg (=.)    Output directory containing the output
                                        files (default is current).
  -o [ --output-prefix ] [=arg(=IMPLICIT)]
                                        Prefix to add to all output file names.

Optional Modules:
  --topology [=arg(=1)] (=0)  Run topology analysis
  --landmarks [=arg(=1)] (=0) Run landmark paths analysis
  --Morse [=arg(=1)] (=0)     Run persistence based topographical analysis
Transition Graph Loader:
  --transition-graph arg Load the compressed transition graph from a XML
                         archive.
Morse Theory Based Analyzer:
  --persistence-threshold arg (=-1)    Persistence threshold (default -1, i.e
                                       no persistence).
  --sublevelset-threshold arg (=0)     Sublevelset threshold (default 0, i.e no
                                       sublevelset).
  --no-popup [=arg(=1)] (=0)           Prevent the graphical display of the
                                       persistence diagram and the
                                       disconnectivity forest.
  --keep-largest-ccs arg (=0)          Keep only the n ccs with the largest
                                       number of points (default: keep all ccs.
  --club-saddles arg (=-1)             Club the saddles in the displayed
                                       disconnectivity forest within intervals
                                       of the input size.
  --draw-labels [=arg(=1)] (=0)        Add labels to the leaves of the
                                       disconnectivity forest.
  --split-basins [=arg(=1)] (=0)       Report each persistent basin into a
```

```
                                    separate plain text file (default is: a
                                    single serialized xml file).
  --MSW-filename arg                Load the MSW chain complex instead of
                                    computing it.
  --normalize-gradient [=arg(=1)] (=0) Normalize the gradient on the NNG by the
                                    distance between the vertices.
```

8.2.6 sbl-landscape-comparison.exe

```
 ./sbl-energy-landscape-comparison.exe -v --transition-graph <path/to/tg1>\
 --transition-graph <path/to/tg2>

Compare two energy landscapes through their transition graphs using a
Earth Mover Distance algorithm.  The input consists of two compressed
transition graphs.

Copyright Inria / Algorithms - Biology - Structure, 2015.
Version: 1.0.0

General Options:
  -h [ --help ] [=arg(=1)] (=0)         Print this help message.
  --workflow [=arg(=1)] (=0)            Print the workflow of the application
                                        without any run.
  -l [ --log ] [=arg(=1)] (=0)          Put the log in a file.
  -v [ --verbose ] [=arg(=1)] (=0)      Dumps high level statistics.
  -c [ --colored-log ] [=arg(=1)] (=0)  Color the log (red for module and
                                        loader names, green for statistics,
                                        blue for report).
  -d [ --directory-output ] arg (=.)    Output directory containing the output
                                        files (default is current).
  -o [ --output-prefix ] [=arg(=IMPLICIT)]
                                        Prefix to add to all output file names.

Transition Graph Loader:
  --transition-graph arg Load the compressed transition graph from a XML
                         archive.

Earth Mover Distance (EMD):
  --with-connectivity-constraints [=arg(=1)] (=0)
                                        Run EMD-CC instead of EMD.
  --algorithm arg (=reg)                Algorithm when constrained: reg for
                                        regular or star for star.
  --edge-selection arg (=min-cost)      With connectivity constraints, edge
                                        selection: min-cost or first-found --
                                        default is min-cost.
  --recursion-mode arg (=refined)       With connectivity constraints,
                                        recursion mode: refined or coarse --
                                        default is refined
  --symmetric-mode [=arg(=1)] (=0)      With connectivity constraints,  compute
                                        EMD-CC both ways (N.B: EMD is
                                        symmetric).
  --dot-flow-threshold arg (=0)         Removes edges with smaller flow than
                                        the threhold from the dot file.
```



RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399