Supporting Text

Weight distribution: Analytical Derivation

In the following, it is shown that the energy landscape ($U(\mathbf{x})$) and the weight distribution ($P_t(w)$) of the corresponding CSN are related by an analytical formula. The weight of a node is defined as the number of times the configuration is visited during the simulation. In the continuous approximation and spherical coordinates $P_t(w)$ for $w > 0$ is written as

$$P_t(w) = \frac{1}{V_t} \int_0^\infty dr \int r^{D-1} d\Omega \delta(w(r, \Omega, t) - w), \tag{i}$$

with V_t the volume of the space visited in the simulation and D the dimension. Ω is the solid angle in D-dimensional spherical coordinates and $w(r, \Omega, t)$ the weight of the node at position (r, Ω), at time t. For simplicity spherical symmetry of the energy landscape ($U(\mathbf{x}) = U(r)$) will be assumed.

For large enough t, $w(r, t)$ is proportional to the stationary solution. Taking $U(r)$ in the units of $k_B T$ and $U(0) = 0$: $w(r, t) = w(0, t) \exp(-U(r))$. The property of $\delta(f(x))$ gives [1]

$$\delta(w(r, t) - w) = \delta(w(0, t) \exp(-U(r)) - w) = \sum_{i=1}^{n} \frac{\delta(r - r_i^*)}{|wU'(r_i^*)|} \tag{ii}$$

with $\exp(-U(r_i^*)) = w/w(0, t)$ and n is the number of simple zeros of $w(r, t) - w$ (values of w such that $w(r, t) - w$ has no zeros are excluded and for such w, $P_t(w) = 0$). By introducing Eq. (ii) into Eq. (i) one obtains

$$P_t(w) = \frac{C}{w} \sum_{i=1}^{n} \frac{{r_i^*}^{D-1}}{|U'(r_i^*)|} \tag{iii}$$

with C the appropriate normalizing factor including the time dependence (since we restrict ourselves to the stationary state, the time dependence will be dropped in the following). The first important remark is the w^{-1} factor in Eq. (iii). This factor does not depend on the particular shape of the energy landscape, neither on the dimension D. Thus any weight distribution is expected to have a power-law $P(w) \sim w^{-1}$ multiplied by a modulating factor.

Simple Energy Landscape Models

Quadratic well

The quadratic well in spherical coordinates is defined by the equation (see also Fig. 5)

$$U(r) = \alpha r^2 \qquad \alpha > 0. \tag{iv}$$

---

[1] For a given function $f(x)$ with n simple zeros $f(x_i^*) = 0, f'(x_i^*) \neq 0, i = 1 \ldots n$: $\delta(f(x)) = \sum_{i=1}^{n} \frac{\delta(x - x_i^*)}{|f'(x_i^*)|}$.

In this case, r^* of Eq. (ii) (i.e., the values of r which yield the peaks of the delta function) can be analytically calculated as

$$r^* = \left(-\frac{1}{\alpha} \ln \left(\frac{w}{w(0)} \right) \right)^{\frac{1}{2}}, \tag{v}$$

which, in turn, gives

$$P(w) = \frac{C}{w} \left(\ln \left(\frac{w(0)}{w} \right) \right)^{\frac{D}{2}-1}. \tag{vi}$$

The node weight distribution $P(w)$ follows a power-law of exponent -1 with a logarithmic correction for $D > 2$. This is verified numerically as shown in Fig. 6A. The analytical curves are obtained from Eq. (vi) with the fitting parameters $w(0)$ as the weight of the heaviest node visited during the simulation and C the appropriate normalization. $P(w)$ follows exactly a power-law for $D = 2$. The logarithmic correction becomes more and more significant as D increases, such that some of the distributions seem to follow a power-law with an exponent smaller than -1. Fig. 6B shows a rescaling of $P(w)$ where straight lines are linear regressions. Eq. (vi) shows that the slope should be equal to $D/2 - 1$. The Monte Carlo results are close to the expected ones and the small differences comes from finite size effects and discretization. Though this rescaling seems appealing to extract the effective dimension of a system, we observed that the values of the slopes are rather sensitive to finite sampling, especially for large w. Therefore this analysis has to be used with great care to infer the dimensionality of the system.

Square well

The infinite square well with spherical symmetry is defined by (see also Fig. 5):

$$U(r) = \begin{cases} 0 & \text{if } r \leq 1; \\ \infty & \text{if } r > 1. \end{cases} \tag{vii}$$

In this case, all the sites with $r \leq 1$ have an equal probability of being visited resulting in a flat stationary solution for $w(r)$ and $P(w)$ becomes a delta function. In the simulation a gaussian distribution is obtained, due to the finite size of the sampling. Fig. 7 shows a comparison between $P(w)$ obtained for the quadratic well and the square well in $D = 2$ dimensions. The gaussian has a mean value around $\bar{w} = 258$, which is close to the expected value (the number of snapshots $(2 \cdot 10^6)$ divided by the number of possible sites: $\pi \cdot (1/a)^2 \approx 7850$, $\Rightarrow \bar{w}_{exp} = 255$, where $a = 0.02$ is the distance between two neighbor sites in the Monte Carlo simulation (see also Methods).

Mexican-Hat landscape

The two cases studied above illustrate two different types of energy landscapes. In the quadratic well, the dynamics of the system is mainly driven by the gradient of the energy landscape, and the energy basin is enthalpic. Whereas for the square well, the energy basin is entropic. It is important to note that $P(w)$ bears the signature of the difference between these two cases. An example containing both kinds of energy landscapes, i.e., entropic and enthalpic, is the so called "Mexican-Hat" landscape model. The energy function has a spherical symmetry and is defined by (see also Fig. 8A):

$$U(r) = 40(r^6 - 1.95r^4 + r^2). \tag{viii}$$

In $D > 1$ dimensions, the model has two energy basins. The central basin has a minimum at $r = 0$. The surrounding basin is a shell centered at $r = 0.97$. The two minima are intrinsically different. The central minimum is well defined and point-like, therefore is enthalpic. The second minimum corresponds to $r^* \approx 0.97$. It is not point-like and has an entropic part along Ω, where Ω is the solid angle in D dimensions, plus an enthalpic part along r. The two energy basins are separated by a maximum at $\hat{r} \approx 0.59$. On a log-log plot, $P(w)$ of the central basin ($r < \hat{r}$) has a broad tail, which is typical for an enthalpic well (Fig. 8B). On the other hand, $P(w)$ of the surrounding basin ($r > \hat{r}$) shows a rather flat region followed by an exponential decay. This decay is typical of entropic basins such as the square well. We do not obtain a gaussian shape since the basin has also an enthalpic component along r. As it can be seen in Fig. 8C, the nodes with a low weight are either close to $\hat{r}$ or have $r > 1.05$. The nodes of the surrounding basin with a large weight are close to the minimum r^*. As expected the heaviest node of the network is at $r = 0$.

Cluster Detection in Multiple Minima Energy Landscape Models

Multidimensional double well

The multidimensional double well is defined by the energy function:

$$U(\mathbf{x}) = 5\sum_{i=1}^{D}(x_i^4 - 2x_i^2 - \epsilon x_i + 1). \tag{ix}$$

Along each dimension it has the shape of a double well with a slight asymmetry as shown in Fig. 9A. In the following the case $D = 5$ and $\epsilon = 0.05$, already seen in the main text is described in details. A simulation with a=0.2 and $N = 3 \cdot 10^6$ was performed to sample the energy surface.

124'156 different sites have been visited. In order to reduce the complexity we first grouped the sites according to a distance criterion. We divide the space into boxes of size $(3a)^D$. All the 3^5 sites within one box form one node of the network. The MCL algorithm with $p = 1.2$ finds the expected 32 clusters. In Fig. 9B, an histogram of the sizes of the different clusters is shown. There are 6 groups of nodes equally distributed along the x-axis (note the logarithmic scale), reflecting the fact that the free energy of the minima can have 6 different values. The system is characterized by 2,815 nodes (Fig. 9C). The fact that most of the clusters are connected to all the other ones make the display of the network a difficult task. In Fig. 9D the network where each node represents a cluster is displayed. The size of the nodes correspond to the total weight of the cluster and the size of the edges to the total weight of edges connecting the two clusters (i.e. it is inversely proportional to the height of the energy barrier). As expected, there is one heaviest node (red node) corresponding to the sites with only positive coordinates, five nodes of almost the same size (blue nodes) corresponding to the sites with one negative coordinate, ten nodes (green nodes) to the sites with two negative coordinates, ten nodes (pink nodes) to the sites with three negative coordinates, five nodes to those with four negative coordinates (yellow) and one last node to the sites with only negative coordinates. A detailed analysis of the coordinates of the nodes of the clusters shows that only 3.4% are "miss-classified", in the sense that they belong to a cluster that does not correspond to their actual basin.

Mexican-Hat landscape

In Fig. 10A the cluster structure of the Mexican-Hat model is shown (see also main text). The network is obtained from the simulation on the landscape defined by Eq. (viii) and $D = 2$ dimensions. Clusters are found applying the MCL algorithm with $p = 1.2$. From the picture the central cluster (green nodes) and the surrounding cluster (red nodes) are clearly visible. In Fig. 10B the distribution of the radial coordinate r for each of the two clusters is shown. Less than 6% of the nodes are "misclassified", i.e. either belongs to the central cluster and have $r > 0.59$, or belongs to the surrounding cluster and have $r < 0.59$.

Alanine Dipeptide: Network Clusterization Details

In this section all the details for the clusterization of the configuration space network of the alanine dipeptide for the MCL, Potts Hamiltonian and Q optimization methods are reported (see the main text for a description of the three algorithms).

In the MCL case (1; 2), large values of the parameter p results in an increased number of communities. For $p = 1.5$, 19 communities are detected but only six of them have a statistically significant population. They include the expected C_{7eq}, α_R, C_{7ax}, α_L basins, the C_5 basin which relaxes very fast to C_{7eq}, and the poorly sampled transition region between C_{7ax} and α_L, namely $T_{\phi>0}$. As the value of p is decreased first the C_5 basin is absorbed to C_{7eq} and than the $T_{\phi>0}$ region to C_{7ax}. For $p = 1.2$ the physically relevant four basins of the landscape are correctly identified (no other clusters are detected). A further decrease of p results in the grouping of the $\phi > 0$ region into one community (see Figs. 11 and 13). For $p = 1$ the landscape is considered as one community by definition. Interestingly, the decrease of the parameter p nicely anti-correlates with the minimum activation barrier between the detected free-energy basins. Values of $p = 1.5, 1.4, 1.3, 1.2$ and 1.1 reflects minimum activation barriers of $\Delta F^{\ddagger} = 1.5, 2.2, 2.2, 2.8$ and 3.6 kcal/mol, respectively (see Fig. 11 and Table 1). This result indicates that MCL algorithm preserves the barriers between free-energy basins and for this reason is suitable for the analysis of networks generated by a dynamical process.

For the case of the Potts model algorithm (3) the number of possible spin states (i.e. maximal number of clusters) was set to 10. The parameter γ enforces that all the communities have an outer link density smaller than γ. For $\gamma = 0.00001$ two main clusters are detected which correspond to the $\phi > 0$ and $\phi < 0$ regions of the (ϕ, ψ) space. While increasing γ at first basins α_L and C_{7ax} are split. For $\gamma = 0.001$ a physically reasonable clusterization is obtained with all the 10 spin states populated. The four main basins are detected as well as the $T_{\phi>0}$ region (see text above). This partition is very similar to the one obtained with MCL with $p = 1.3 - 1.4$. Increasing the parameter γ to 0.01 results in the detection of 10 communities. Among them four clusters are defined by dihedral angles corresponding to the four basins of attraction of the dipeptide. Unfortunately other spin states, which detect extensive transition regions between the four main basins, are also quite populated (see Fig. 12). For $\gamma = 0.1$, the Potts Hamiltonian is dominated by the antiferromagnetic term. Five spin states are equally populated but without any relation to the real partition (see Fig. 12). As it is shown in Figs. 12 and 13, while decreasing the value of γ, $C_{7_{eq}}$ and α_R are grouped together before $C_{7_{ax}}$ is merged to α_L. This behavior indicates that the decrease of the parameter γ does not result in the grouping of basins which are separated by higher activation barriers (see Table I in the main text) as it would have been desirable. This observation reflects the fact that the algorithm forces the outer link density between clusters, which depends on the number of direct transitions, to be smaller than γ (fewer transitions are observed between $C_{7_{ax}}$ and α_L with respect to $C_{7_{eq}}$ and α_R).

Finally, Q optimization algorithm focuses on the detection of clusters of nodes with an intra-cluster link density higher than random grouping (4). As shown in Fig. 12, this results in an unphysical picture of the community structure of the alanine dipeptide configuration space: the C_{7eq} basin, which is the region of the network characterized by the highest link weights, is split in four parts and $C_{7_{ax}}$ is merged with α_L. Even if this algorithm has proven to perform efficiently on networks of other data types, it fails to recover the correct partition into free-energy basins of the dipeptide landscape. This is probably due to the fact that link densities can be very different in the four densely connected regions of the network while the modularity Q is a useful clustering parameter under the (not explicit) assumption of homogeneous intra-cluster link density.

Alanine Dipeptide: Further Configuration Space Discretizations

In the main text the configuration space of the alanine dipeptide was discretized by dihedral angles for a total of $n \times n$ cells, where $n = 50$. In the following, it is investigated how the MCL clusterization changes for different values of n and for a completely different discretization of the configuration space based on inter-atomic distances.

Dihedral angles

A more fine/coarse division of the dihedral space results in an increasing/decreasing number of nodes and links. A discretization of the (ϕ, ψ) space in 20×20 cells results in a network of 348 nodes. MCL algorithm with $p = 1.2$ detects four communities which are shown in Fig. 14A. The community structure found in this case is very much the same as the one presented in the main text where $n = 50$. Similar results are found for higher values of n which indicates that the community structure is robust upon different discretizations of the (ϕ, ψ) space.

Interatomic distances

A more stringent test on the robustness of free-energy basins detection is carried out considering a completely different discretization of the configuration space based on interatomic distances. This approach is rather crude and represents a challenging test on the dependence of the community detection procedure upon the definition of network nodes. Each cell of the configuration space is defined by an array of inter-atomic distances of the atoms of the central alanine residue , e.g., (d1, d2, d3, d4, d5, d6, d7, d8, d9, d10) (see Table 2). During the simulation the 10 distances have

fluctuations between 0.5 Å to more than 2 Å. Interatomic distance values are discretized using bins of $M = 0.3$ Å which results in a network of 10713 nodes (which is one order of magnitude larger than a dihedral discretization). Interestingly, four communities are found by the MCL algorithm with $p = 1.2$ which is consistent with the community structure found when cells are defined according to dihedral angles (see Fig. 14B). However, for larger values of the discretization parameter M the inter-basin transition probabilities are larger than the one found by the (ϕ, ψ) discretization, reflecting the presence of many misplaced nodes near the saddles. This behavior suggests that crucial for a good clusterization is structural homogeneity inside the cells of the configuration space.

References

[1] Van Dongen, S. (2000) Ph.D. thesis (University of Utrecht, The Netherlands).

[2] Gfeller, D., Chappelier, J.C. & De Los Rios, P. (2005) *Phys. Rev. E* **72**, 56135.

[3] Reichardt, J. & Bornholdt, S. (2004) *Phys. Rev. Lett.* **93**, 218701.

[4] Clauset, A., Newman, M.E.J. & Moore, C. (2004) *Phys. Rev. E* **70**, 066111.

TABLE 2 List of the 10 interatomic distances for the atoms belonging to the central alanine residue

distance name	interacting atoms
d1	N : C_β
d2	N : C
d3	N : O
d4	H : C_α
d5	H : C_β
d6	H : C
d7	H : O
d8	C_α : O
d9	C_β : C
d10	C_β : O

FIG. 5 The quadratic (dotted line) and square well (blue line) energy functions in $D = 1$.

FIG. 6 **(A)** Node weight distribution for the quadratic well energy landscape ($\alpha = 5$) in four different dimensions. The red curve shows the analytical estimate as detailed in *Methods and Models* with $w(0)$ the weight of the heaviest node and C the appropriate normalization. The parameters are: $D = 2$, $a = 0.02$, $N = 2 \cdot 10^6$; $D = 4$, $a = 0.1$, $N = 2 \cdot 10^6$; $D = 6$, $a = 0.2$, $N = 2 \cdot 10^6$; $D = 10$, $a = 0.3$, $N = 5 \cdot 10^6$. The parameter a is the distance between two neighbor sites in the Monte Carlo simulation (see *Methods and Models* for details). **(B)** Rescaled node weight distribution. For $D = 2, 4, 6, 10$, the slopes computed by linear regressions are $0, 1.01, 2.01, 4.08$, respectively, and thus are close to the expected values. Note that for large weights the fluctuations due to discretization become important and the data do not follow anymore a straight line. For clarity, the data for $D = 2$ have been shifted in order not to overlap.

FIG. 7 Comparison between the node weight distribution in $D = 2$ dimensions for the square well and the quadratic well. We have used $a = 0.02$ and $N = 2 \cdot 10^6$ in both cases. The distribution for the square well fits a gaussian.

FIG. 8 **(A)** Energy function of the Mexican-Hat model along the radial coordinate r. **(B)** Normalized node weight distribution for the two basins of the Mexican-Hat model in $D = 3$ dimensions. The nodes have been attributed to the basins according to the coordinate r. **(C)** Position of every node as a function of the node weight. $r = 0.97$ corresponds to the minimum in the surrounding basin (entropic). $r = 0.59$ corresponds to the maximum separating the two basins. $a = 0.05$, $N = 10^6$.

FIG. 9 **(A)** Double well landscape model energy function in one dimension. The model is characterized by 2^D minima. **(B)** Histogram of the size of the clusters found by the MCL algorithm with $p = 1.2$. The color code is the same as in **(D)**. **(C)** Network representation of the configuration space model in $D = 5$ dimensions. Both the shape and the color of the nodes represent clusters. **(D)** Network of the clusters. The size of the nodes is proportional to the weight of the clusters and the size of the edges is proportional to the number of transitions from one cluster to another one visited during the simulation. $\epsilon = 0.05$, $a = 0.2$, $N = 3 \cdot 10^6$

FIG. 10 **(A)** Network representation of the Mexican-Hat landscape model in $D = 2$ dimensions. Nodes are color coded according to the two clusters detected by the MCL algorithm with $p = 1.2$, $a = 0.02$, $N = 10^5$. **(B)** Distribution of the coordinate r of the nodes of the network for each of the two clusters.

FIG. 11 Ramachandran-map representation of the alanine dipeptide network clusterization with MCL and $p = 1.1, 1.2, 1.3, 1.4, 1.5$. Oval boxes and arrows show the smallest activation energy $\Delta F^{\ddagger}$ to cross for an inter-basin transition. As the parameter p increases smaller activation barriers are detected by the algorithm.

FIG. 12 Ramachandran-map representation of the alanine dipeptide network clusterization. **(Red box)** Modularity optimization cluster structure. **(Black box)** Potts Hamiltonian cluster structure for 5 values of the ferromagnetic-antiferromagnetic coupling parameter γ and 10 spin states (see *Methods and Models*).

FIG. 13 Schematic cluster splittings as the p (MCL) and γ (Potts) granularity-related parameters are changed. The sequence of events shows that MCL consistently finds basins with smaller activation barriers as p is increased. This is not the case for the Potts Hamiltonian approach which searches for optimal outer link densities.

FIG. 14 Discretization of the configuration space of the alanine dipeptide and community detection. **(A)** Ramachandran-map representation of the dihedral discretization in 20×20 cells. **(B)** Ramachandran-map representation of the inter-atomic distance discretization with $M = 0.3$ Å (see Table 2). In both cases nodes are color coded according to the communities detected by the MCL algorithm with $p = 1.2$.