Supplemental Material

Mutational paths in protein-sequence landscapes: from sampling to low-dimensional characterization

Eugenio Mauri, Simona Cocco, Remi Monasson

April 21, 2022

1 Path sampling algorithm

1.1 Detailed description

We introduce below the details of the sampling procedure we use to obtain paths connecting two fixed sequences in a landscape described by the probability distribution P_{model}. Starting from a path $\{\mathbf{v}_t\}$, we look at intermediate sequences (starting from $t = 1$) and propose a mutation with the constraint that the Hamming distance between $\mathbf{v}_{t-1}$ and $\mathbf{v}_{t+1}$ is not greater than 1. We accept this move with a probability fixed to ensure detail balance. Different cases have to be considered, depending on the Hamming distance D_H between the new attempted sequence and existing ones:

- $D_H(\mathbf{v}_{t-1}, \mathbf{v}_{t+1}) = 0$. In this case the new sequence $\mathbf{v}'_t$ can have a single mutation at any site, compared with the two adjacent sequences along the path. Hence, we first draw a random site i, then we propose a new sequence $\hat{\mathbf{v}}_{t-1}$ amino-acids for that site $\hat{v}^i_t$ drawn from the distribution $\propto P^{\beta}_{model}(\cdot|\mathbf{v}^{\setminus i}_{t-1})$, where the amino acids are fixed on all the sites different from i. Then if the old sequence $\mathbf{v}_t$ already had a mutation with respect to $\mathbf{v}_{t-1}$ at given site j, we accept the new mutated sequence $\hat{\mathbf{v}}_t$ (which is equal to $\mathbf{v}_{t-1}$ apart from the amino acid at site i) with a probability

$$p_{acc}(\mathbf{v}_t \to \hat{\mathbf{v}}_t) = \\ = \min\left(1, \frac{\pi_{tr}(\mathbf{v}_{t-1}, \hat{\mathbf{v}}_t)^{2\beta} \sum_z P^{\beta}_{model}(M^z_i \mathbf{v}_{t-1})}{\pi_{tr}(\mathbf{v}_{t-1}, \mathbf{v}_t)^{2\beta} \sum_z P^{\beta}_{model}(M^z_j \mathbf{v}_{t-1})}\right), \tag{1}$$

  where M^z_i indicates the mutation z at site i. If $\mathbf{v}_t$ or $\hat{\mathbf{v}}_t$ are equal to $\mathbf{v}_{t-1}$, then the acceptance probability is $p_{acc}(\mathbf{v}_t \to \hat{\mathbf{v}}_t) = \min(1, \pi_{tr}(\mathbf{v}_{t-1}, \hat{\mathbf{v}}_t)^{2\beta}/\pi_{tr}(\mathbf{v}_{t-1}, \mathbf{v}_t)^{2\beta})$.

- $D_H(\mathbf{v}_{t-1}, \mathbf{v}_{t+1}) = 1$. In this case the new sequence $\hat{\mathbf{v}}_t$ can have a single mutation only at the site i where $\mathbf{v}_{t-1}$ and $\mathbf{v}_{t+1}$ are different. At that site, we propose a new mutation from the distribution $\propto P^{\beta}_{model}(\cdot|\mathbf{v}^{\setminus i}_{t-1})$ and accept it with probability $p_{acc} = \exp[-\Lambda\beta(D_H(\hat{\mathbf{v}}_t, \mathbf{v}_{t-1}) + D_H(\hat{\mathbf{v}}_t, \mathbf{v}_{t+1}) - D_H(\mathbf{v}_t, \mathbf{v}_{t-1}) - D_H(\mathbf{v}_t, \mathbf{v}_{t+1}))]$.

- $D_H(\mathbf{v}_{t-1}, \mathbf{v}_{t+1}) = 2$. In this case the previous and subsequent sequence present two mutations at site i and j. The new sequence $\hat{\mathbf{v}}_t$ can be of two forms: it can have the same mutation of $\mathbf{v}_{t+1}$ (with respect to $\mathbf{v}_{t-1}$) at site i or at site j. Hence, we extract one of the two possibilities with a probability weighted accordingly with $P^{\beta}_{model}(\hat{\mathbf{v}}_t)$.

1.2 Proof of detailed balance

To prove that dynamics given by our algorithm of path sampling converges to the target distribution we have to prove that it respects detailed balance, i.e. the reversibility of each Markov step. We consider the transition from a path $\{\mathbf{v}_t\}$ to a new path that differ only by one sequence $\mathbf{v}'_t$ at time t. we write the detailed balance condition as

$$\mathcal{P}_{path}(\{\mathbf{v}_t\})p_{trans}(\mathbf{v}_t \to \mathbf{v}'_t) = \mathcal{P}_{path}(\{\mathbf{v}'_t\})p_{trans}(\mathbf{v}'_t \to \mathbf{v}_t)$$
$$\pi_\lambda(\mathbf{v}_{t-1}, \mathbf{v}_t)\pi_\lambda(\mathbf{v}_t, \mathbf{v}_{t+1})P_{model}(\mathbf{v}_t)p_{trans}(\mathbf{v}_t \to \mathbf{v}'_t) = \pi_\lambda(\mathbf{v}_{t-1}, \mathbf{v}'_t)\pi_\lambda(\mathbf{v}'_t, \mathbf{v}_{t+1})P_{model}(\mathbf{v}'_t)p_{trans}(\mathbf{v}'_t \to \mathbf{v}_t) \tag{2}$$

If $D_H(\mathbf{v}_{t-1}, \mathbf{v}_{t+1}) = 0$, the new sequence can have a mutation at any site i compared to its neighbour $\mathbf{v}_{t-1}$, while $\mathbf{v}_t$ will have the mutation at another site j (note that i and j can be equal. Hence the transition probability in this case will be

$$p_{trans}(\mathbf{v}_t \to \mathbf{v}'_t) = \frac{1}{N}\frac{P_{model}(\mathbf{v}'_t)}{\sum_{z=1}^{Q} P_{model}(M_i^z \mathbf{v}_{t-1})}p_{acc}(\mathbf{v}_t \to \mathbf{v}'_t),\ p_{trans}(\mathbf{v}'_t \to \mathbf{v}_t) = \frac{1}{N}\frac{P_{model}(\mathbf{v}_t)}{\sum_{z=1}^{Q} P_{model}(M_j^z \mathbf{v}_{t-1})}p_{acc}(\mathbf{v}'_t \to \mathbf{v}_t). \tag{3}$$

By substituting everything in the detailed balance condition we obtain the acceptance probability described in the main paper. This will hold similarly when $D_H(\mathbf{v}_{t-1}, \mathbf{v}_{t+1}) = 1$ (with $i = j$). For $D_H(\mathbf{v}_{t-1}, \mathbf{v}_{t+1}) = 2$, the sequences $\mathbf{v}_t$ and $\mathbf{v}'_t$ can either be equal to $\mathbf{v}_{t+1}$ or $\mathbf{v}_{t-1}$, from which the condition presented in the paper descends.

2 Lattice Proteins

To benchmark the performances of this MC procedure to find good transition path between two sequences, we test it on Lattice Proteins [LD89], a well known toy-model for protein structure. We consider a protein sequence of 27 amino acids folding into a 3D structure specified as a self-avoiding path over a 3x3x3 lattice where each amino acid occupies one node. The probability of a sequence $\mathbf{v}$ to fold into a specific structure $\mathcal{S}$ is given by the interaction energies between amino acids in contact in the structure (i.e. those who occupy neighbouring nodes of the lattice, but are not adjacent in the protein sequence). In particular, the total energy of a sequence with respect to a given structure is given by

$$\mathcal{E}_{LP}(\mathbf{v}|\mathcal{S}) = \sum_{i<j} c_{ij}^{\mathcal{S}} E_{MJ}(v_i, v_j) \tag{4}$$

where $c^{\mathcal{S}}$ is the contact map ($c_{ij}^{\mathcal{S}} = 1$ if sites are in contact and 0 otherwise), while the pairwise energy $E_{MJ}(v_i, v_j)$ represents the amino-acid physico-chemical interactions given by the the Miyazawa-Jernigan knowledge-based potential [MJ96]. The probability to fold into a specific structure is written as

$$p_{nat}(\mathcal{S}|\mathbf{v}) = \frac{e^{-\mathcal{E}_{LP}(\mathbf{v}|\mathcal{S})}}{\sum_{\mathcal{S}'} e^{-\mathcal{E}_{LP}(\mathbf{v}|\mathcal{S}')}}, \tag{5}$$

where the sum is over the entire set of self-avoiding path in the cubic lattice.
The function p_{nat} represents a suitable landscape that maps each sequence to a score measuring the quality of its folding. To study in more detail this landscape, we will consider an alignment of sequences folding into a specific structure (that we will call $\mathcal{S}$) sampled from a low temperature MC sampling using $-\beta \log p_{nat}(\cdot|\mathcal{S})$ (with $\beta = 10^3$) as effective energy [JGS+16].

3 Restricted Boltzmann Machines and training parameter

To study the problem of transition paths we first need a model to infer a landscape from our sequence data set. At this scope, we are going to use Restricted Boltzmann Machines, an unsupervised energy-based model able to learn representations of the data in a two-layer bipartite graph [FI12]. The first "visible" layer represents the protein sequence $\mathbf{v} = \{v_1, ..., v_N\}$ where each unit takes one out of 21 possible states (20 amino acids + 1 alignment gap). The second is the "hidden" layer which displays the real-valued representations $\mathbf{h} = \{h_1, ..., h_M\}$. The joint probability distribution for $\mathbf{v}$ and $\mathbf{h}$ is

$$P_{RBM}(\mathbf{v}, \mathbf{h}) \propto \exp\left[\sum_{i=1}^{N} g_i(v_i) + \sum_{i,\mu} w_{i\mu}(v_i)h_\mu - \sum_{\mu} \mathcal{U}(h_\mu)\right] \tag{6}$$

up to a normalization constant. visible and hidden units are coupled through the matrices $w_{i\mu}$ and the value of each unit is biased by the local fields g_i and $\mathcal{U}_\mu$. In [TCM19] it has been shown that this model is able to recover statistically relevant

sequence motifs playing crucial roles in the structure and functionality of different protein families. Following their approach, we choose $\mathcal{U}_\mu$ to be double Rectified Linear Unit (dReLU) potentials of the form

$$\mathcal{U}_\mu(h) = \frac{1}{2}\gamma_{\mu,+}h_+^2 + \frac{1}{2}\gamma_{\mu,-}h_-^2 + \theta_{\mu,+}h_+ + \theta_{\mu,-}h_- \,, \text{ where } h_+ = \max(h,0),\ h_- = \min(h,0), \tag{7}$$

where we have defined the hyper-parameters $\gamma_{\mu,\pm}$, $\theta_{\mu,\pm}$.
At this point, we need a learning procedure to infer the hyper-parameters that best fit our data. We decided to use a Persistent Contrastive Divergence algorithm [Tie08] which has been shown to be sufficiently good and robust under cautious choice of the regularization hyper-parameters [Tub18a]. The code and the data used to train our RBMs for Lattice Proteins and WW domains can be found in [Tub18b]. The hyper-parameters used for learning are the following:

- For WW (N=31):
  - M = 50
  - Batch size = 100
  - Number of epochs = 500
  - Learning rate = 5×10^{-3} (which has a decay rate of 0.5 after 50% of iterations)
  - $L_1 b$ regularization = 0.25
  - Number of MC step between each update = 10
- For Lattice Protein (N=27):
  - $M = 100$
  - Batch size = 100
  - Number of epochs = 100
  - Learning rate = 5×10^{-3} (which has a decay rate of 0.5 after 50% of iterations)
  - $L_1 b$ regularization = 0.025
  - Number of MC step between each update = 5

4 Statistic of paths sampled in LP

Here we show the relation between the sequences sampled using the sampling procedure described above (using the RBM as model) with the mean-field solution described in the main paper. Looking at Figure S1 We see, consistently with the mean-field solution, that global solutions prefers to activate input 14 more than the direct solution. Furthermore, statistically this input is more likely to be reduced before activating input 4, which is consistent with the mean-field solution.
Analysing the difference between the direct and global paths along each inputs we notice that global solutions maintain high scores in terms of p_{nat} by exploiting the interactions between amino-acids at sites 5,6,11 and 22 (see Figure 2 in the main paper). Global paths are divided in two classes corresponding to the chemical nature of the interaction used to bind these amino-acids. One cluster (shown in maroon in Figure 2 of the main paper and in Figure S1) uses Cys-Cys bridges to establish this interactions. Instead the most populated cluster (shown in red in the same figures) exploits electrostatic interactions that are not initially present in the target sequences (hence forbidden along direct paths). Some of these interactions are shown in Figure S2. This explains why the Principal Component Analysis shown in Figure 2 of the main paper does not discriminate between direct paths and most of the global ones (*i.e.* red cluster).

To deeper characterize the behaviour of global solutions, we also plot in Figure S3 the average distance from the direct space as a function of time obtained from the paths sampled using $P_{model} \propto P_{RBM}$, the likelihood from the trained RBM model.

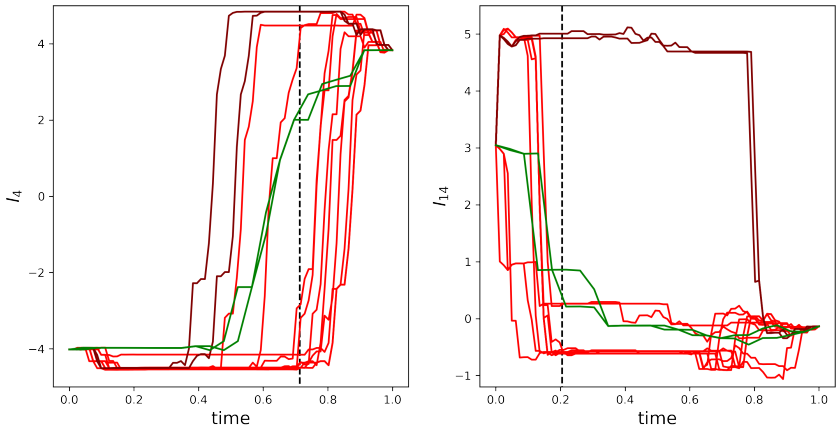

Figure S1: Plot of some relevant inputs as function of time sampled with our Monte-Carlo procedure ($\Lambda = 2$ and target $\beta = 3$). The colors respect those presented in Figure 2 in the main paper. Black lines correspond to the average time at which the input switch value (>0 for I_4 and <2 for I_{14}).

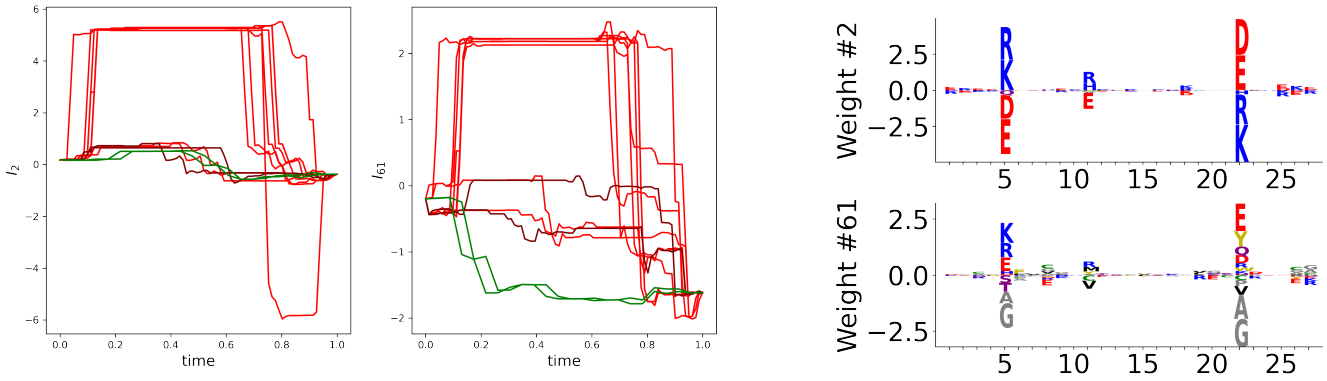

Figure S2: Plot of some relevant inputs (and their respective weights logos) as function of time sampled with our Monte-Carlo procedure ($\Lambda = 2$ and target $\beta = 3$) exploiting relevant electrostatic interactions forbidden in the direct space. The colors respect those presented in Figure 2 in the main paper.

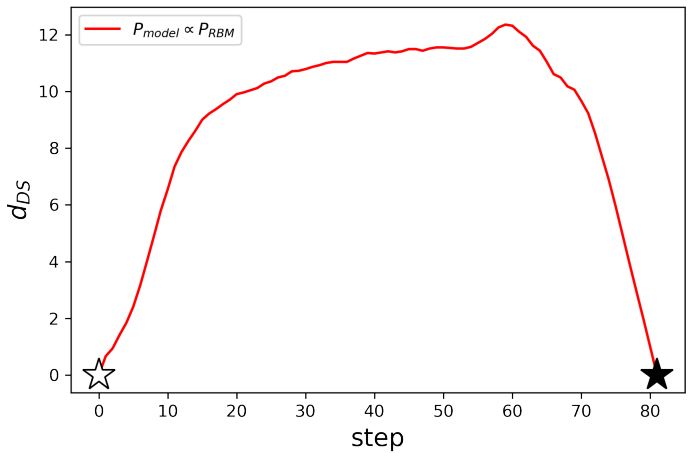

Figure S3: Average value of the distance from direct space as a function of the step along the path connecting the two modes in the LP landscape. Black and White stars refer to the same sequences as in Figure 2 of the main text.

5 Additional information about mutational paths of the WW domain

5.1 Lists of the tested sequences

Here we present the reference sequences sampled with the MC algorithm and tested using AlphaFold (note that the first sequence of each list represent the YAP1 wild-type sequence from [ES99], while the last is the natural wild-type specific for each class):

- From YAP1 to wild-type of class II/III:
  - `LPAGWEMAKTSS-GQRYFLNHIDQTTTWQDP`
  - `LPAGWEMARTSD-GQVYFINHNTQTTTWQDP`
  - `LPPGWQEARTPD-GRVYYINHNTKTTTWTKP`
  - `-PPEWQEARTPD-GRVYYYNHNTKTTTWTKP`
  - `---EWQEARTPD-GRVYYYNHNTKQTTWTKP`
  - `---EWQEAKTPD-GRVYYYNKNTKQTTWEKP`
  - `---EWQEFKTPA-GKKYYYNKNTKQSRWEKP`
- From YAP1 to wild-type of class IV:
  - `LPAGWEMAKTSS-GQRYFLNHIDQTTTWQDP`
  - `LPPGWEVRYTRS-GRPYFVNHNTKTTTWEDP`
  - `LPPGWEVRYSRSKNRPYFVNHNTKTTTWEDP`
  - `LPPGWEVRHSRSKNRPYFFNHNTKTTTWEPP`
  - `LPPGWEVRHSRSKNRPYFFNHNTKESTWEPP`
  - `LPPPWEVRISRSKNRPYFFNTETKESLWEPP`
  - `LPKPWIVKISRSRNRPYFFNTETHESLWEPP`
- From YAP1 to wild-type of class I:
  - `LPAGWEMAKTSS-GQRYFLNHIDQTTTWQDP`
  - `LPAGWEMAKTSE-GQRYFINHNTQTTTWQDP`
  - `LPPGWEMAYTPE-GERYFINHNTKTTTWLDP`

– LPPGWEMGITRG-GRVFFINHETKSTTWLDP

– LPRSWTYGITRG-GRVFFINHEAKSTTWLHP

– LPRSWTYGITRG-GRVFFINEEAKSTTWLHP

5.2 details on the TM-score

To compare the inferred structures of the sampled sequences along the path with those of the target natural sequences we used the Template Modelling (TM) score developed in [ZS04] and represents a variation of the Levitt–Gerstein (LG) score [LG98]. Compared to other similarity score (like root-mean-square deviation (RMSD)) it gives a more accurate measure since it relies more on the global similarity of the full sequence rather than the local similarities.
Practically, we consider a target sequence of length L_{target} and a template one whose structure has to be compared with. First, we align the two sequences and we take the L_{common} pairs of residues that commonly appear aligned. Then the score is computed as

$$\text{TM-score} = \max_{\{d_i\}} \left[\frac{1}{L_{\text{target}}} \sum_{i=1}^{L_{\text{common}}} \frac{1}{1 + \left(\frac{d_i}{d_0(L_{\text{target}})} \right)^2} \right] , \tag{8}$$

where d_i is the distance between the ith pair of residues between the template and the target structures after alignment, and $d_0(L_{\text{target}}) = 1.24(L_{\text{target}} - 15)^{1/3} - 1.8$ is a distance scale that normalizes distances. This formula gives a score between 0 and 1. if TM-score< 0.2, the two sequences are totally uncorrelated, while they can be considered to have the same structure if TM-score> 0.5. In Figure S4 we show the tables of the TM-scores associated with the sequences tested above.

	YAP1	wt class I
YAP1	1	0.88
1	0.95	0.83
2	0.94	0.86
3	0.91	0.82
4	0.84	0.98
wt class I	0.88	1

	YAP1	wt class IV
YAP1	1	0.57
1	0.94	0.55
2	0.82	0.65
3	0.74	0.71
4	0.69	0.84
5	0.67	0.84
wt class IV	0.57	1

	YAP1	wt class II/III
YAP1	1	0.71
1	0.94	0.74
2	0.91	0.75
3	0.75	0.77
4	0.74	0.66
5	0.75	0.7
wt class II/III	0.71	1

Figure S4: Table of the TM-scores measured between the structures inferred from AlphaFold.

5.3 Visualization of mean-field paths

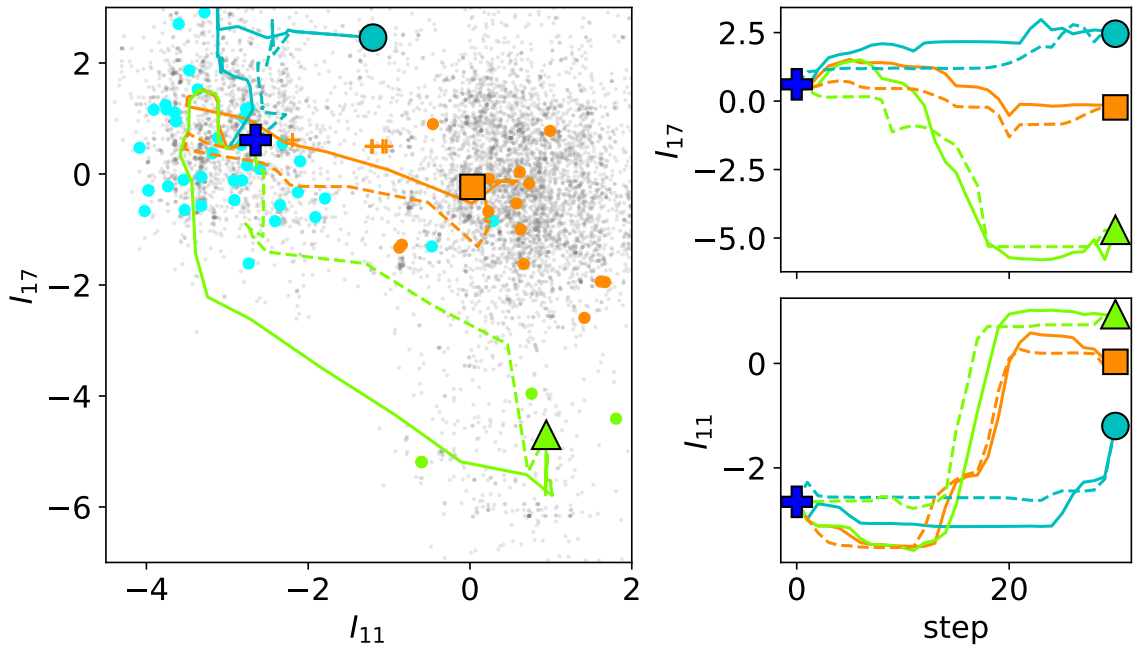

Figure S5: Mean-field paths for the RBM model trained over WW domain dataset. Solid and dashed lines stand, respectively, for global and direct solutions. Here $\beta = 3$, $\alpha = 1$, $\lambda = 1$, $\gamma = 2$ and $T = 30$. Same edge sequences as in Figure 3 of the main text.

6 Derivation of mean-field equations for path sampling with a RBM model

To exploit the nature of the RBM as a mean-field model we rewrite the probability distribution of the model as

$$P_{RBM}(\mathbf{v}) = \frac{1}{Z_{RBM}} \int \prod_{\mu} \mathrm{d}h_\mu \exp\left(\sum_i g_i(v_i) + \sum_{i,\mu} w_{i\mu}(v_i) h_\mu - \sum_\mu \mathcal{U}_\mu(h_\mu) \right) \tag{9}$$

$$= \frac{1}{Z_{RBM}} \exp\left(\sum_i g_i(v_i) + N \sum_\mu \Gamma_\mu\left(\frac{1}{N} I_\mu\right) \right) , \tag{10}$$

where γ is defined in the main text.

By introducing the order parameters $m_t^\mu = I_{\mu,t}/N$, $q_t = \frac{1}{N}\sum_i \delta_{v_{i,t},v_{i,t+1}}$ as well as the overlap potential Φ, we can write the probability for a path in the order parameter space as (in the large N limit)

$$P_{path}(\{\mathbf{m}_t, q_t\}_{t=1}^T; \beta) \propto \sum_{\{\mathbf{v}_t\}} \prod_{t,\mu} \delta\left(\frac{1}{N}\sum_i w_{i\mu}(v_{i,t}) - m_t^\mu \right) \prod_t \delta\left(\frac{1}{N}\sum_i \delta_{v_{i,t},v_{i,t+1}} - q_t \right) P_{RBM}^\beta(\mathbf{v}_t) \pi^\beta(\mathbf{v}_t, \mathbf{v}_{t+1}) \tag{11}$$

$$= \exp\left[\beta N \left(\sum_{t,\mu} \Gamma(m_t^\mu) - \sum_t \Phi(q_t) \right) \right] \times \tag{12}$$

$$\times \int \left(\prod_{\mu,t} \mathrm{d}\hat{m}_t^\mu \prod_t \mathrm{d}\hat{q}_t \right) \sum_{\{\mathbf{v}_t\}} \exp\left[\sum_{i,t} g_i(v_{i,t}) + N \sum_{t,\mu} \hat{m}_t^\mu \left(\frac{1}{N}\sum_i w_{i\mu}(v_{i,t}) - m_t^\mu \right) + N \sum_t \hat{q}_t \left(\frac{1}{N}\sum_i \delta_{v_{i,t},v_{i,t+1}} - q_t \right) \right] \tag{13}$$

$$\approx \exp\left[\beta N \left(\sum_{t,\mu} \Gamma(m_t^\mu) - \sum_t \Phi(q_t) \right) + N\mathcal{S}(\{\mathbf{m}_t, q_t\}_{t=1}^T) \right] = \exp\left(-N\beta f_{path}(\{\mathbf{m}_t, q_t\}) \right) , \tag{14}$$

where

$$\mathcal{S}(\{\mathbf{m}_t, q_t\}_{t=1}^T) = \min_{\{\hat{\mathbf{m}}_t, \hat{q}_t\}} \frac{1}{N} \sum_i \log Z_i(\{\hat{\mathbf{m}}_t, \hat{q}_t\}) - \sum_{t,\mu} m_t^\mu \hat{m}_t^\mu - \sum_t q_t \hat{q}_t \tag{15}$$

and

$$Z_i(\{\hat{\mathbf{m}}_t, \hat{q}_t\}) = \sum_{v_1,..,v_t,...,v_T} \exp\left[\sum_t g_i(v_t) + \sum_{t,\mu} \hat{m}_t^\mu w_{i\mu}(v_t) + \sum_t \hat{q}_t \delta_{v_t,v_{t+1}}\right]. \tag{16}$$

Under minimization we find the result shown in the main paper. To obtain numerically the set of magnetizations and overlap that minimize the free energy, we note that the saddle point equation for f_{path} leads to the following self-consistent equation:

$$m_t^\mu = \frac{1}{N} \sum_i \frac{1}{Z_i} \sum_{v_1,..,v_t,...,v_T} w_{i\mu}(v_t) \exp\left[\sum_t g_i(v_t) + \sum_{t,\mu} \hat{m}_t^\mu w_{i\mu}(v_t) + \sum_t \hat{q}_t \delta_{v_t,v_{t+1}}\right] \tag{17}$$

$$q_t = \frac{1}{N} \sum_i \frac{1}{Z_i} \sum_{v_1,..,v_t,...,v_T} \delta_{v_t,v_{t+1}} \exp\left[\sum_t g_i(v_t) + \sum_{t,\mu} \hat{m}_t^\mu w_{i\mu}(v_t) + \sum_t \hat{q}_t \delta_{v_t,v_{t+1}}\right], \tag{18}$$

where $\hat{q}_t = -\beta\Phi'(q_t)$ and $\hat{m}_t^\mu = \beta\Gamma'_\mu(m_t^\mu)$. We solve this set of equations using gradient descent. To compute the LHS we first compute the partition functions Z_i using the transfer matrix method and then we take their gradient using automatic differentiation technique built in the Python library JAX [BFH+18].
To obtain the average distance from the direct space, we need to compute at each time at each site the probability of a specific state $a = 1, ..., A$. This can be computed as

$$f_{i,t}(a|\{\mathbf{m}_t, q_t\}) = \frac{1}{Z_{path}(\{\mathbf{m}_t, q_t\})} \sum_{\{\mathbf{v}_t\}} \delta_{v_{i,t},a} \prod_{t,\mu} \delta\left(\frac{1}{N}\sum_i w_{i\mu}(v_{i,t}) - m_t^\mu\right) \prod_t \delta\left(\frac{1}{N}\sum_i \delta_{v_{i,t},v_{i,t+1}} - q_t\right) \tag{19}$$

$$= \frac{1}{Z_{path}} \int \left(\prod_{\mu,t} \mathrm{d}\hat{m}_t^\mu \prod_t \mathrm{d}\hat{q}_t\right) \exp\left(-N\sum_{\mu,t} \hat{m}_t^\mu m_t^\mu - N\sum_t \hat{q}_t q_t\right) \sum_{\{\mathbf{v}_t\}} \delta_{v_{i,t},a} \exp\left(\sum_{i,t,\mu} \hat{m}_t^\mu w_{i\mu}(v_{i,t}) + \sum_{t,i} \hat{q}_t \delta_{v_{i,t},v_{i,t+1}}\right) = \tag{20}$$

$$= \frac{1}{Z_{path}(\{\mathbf{m}_t, q_t\})} \partial_{g_{i,a}} \exp\left[N \min\left(-\sum_{\mu,t} \hat{m}_t^\mu m_t^\mu - \sum_t \hat{q}_t q_t + \frac{1}{N}\sum_i \log Z_i^{g_{i,a}}\right)\right] = \tag{21}$$

$$= \partial_{g_{i,a}} \log \sum_{v_1,...,v_{T-1}} \exp\left[\sum_{t,a} g_{i,a}\delta_{v_t,a} + \sum_{t,\mu} \hat{m}_t^\mu w_{i\mu}(v_t) + \sum_t \hat{q}_t \delta_{v_t,v_{t+1}}\right]\Bigg|_{\mathbf{g}_i=0} \tag{22}$$

Once $f_{i,t}$ is computed, we obtain the average distance from direct space at time t through

$$d_{DS}(t) = \frac{1}{N} \sum_i \sum_{a \neq \{v_i^{start}, v_i^{end}\}} f_{i,t}(a). \tag{23}$$

Finally, up to now we have note defined Φ in this section in order to leave the computation as general as possible. Below we will define the potential as

$$\Phi(q) = \lambda |q - (1 - \gamma/T)|^{-\alpha} T^{-\alpha-1} \tag{24}$$

(with $\gamma > 1$) to ensure a good scaling as $T \to \infty$. We note that this is a generalisation of the potential shown in the main paper where we have fixed the exponential $\alpha = 1$ and $\lambda = 1$.

7 Analysis of the Hopfield-Potts model

7.1 The model

To deeper explore the relation between direct and global path we consider an Hopfield-Potts model with $M = 2$ patterns and $A \geq 3$ states per site (called **a**, **b** and **c** and so on) . We will consider the thermodynamic limit $N \to \infty$. each pattern $\mathbf{w}_\mu$ is constructed as follows:

$$\begin{aligned} w_{1i}(v_i) &= \delta_{v_i,\mathbf{a}} + \omega\delta_{v_i,\mathbf{c}} \\ w_{2i}(v_i) &= \delta_{v_i,\mathbf{b}} + \omega\delta_{v_i,\mathbf{c}} \end{aligned} \tag{25}$$

The energy of the model is given by

$$E(\mathbf{v}) = -\frac{1}{2}\sum_\mu \sum_{i,j} w_{i\mu}(v_i) w_{j\mu}(v_j)\,. \tag{26}$$

This model is equivalent to a RBM with M hidden units and quadratic local potentials:

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i,\mu} w_{i\mu}(v_i) h_\mu + \frac{1}{2}\sum_\mu h_\mu^2\,. \tag{27}$$

Hence we can compute the optimal path between the target sequences using the mean field approach described above. Denoting $m_t^\mu(\mathbf{v})$ (with $\mu = 1, 2, 3$) the projection along the vectors $\delta_{v_i,\mathbf{a}}$, $\delta_{v_i,\mathbf{b}}$, $\omega\delta_{v_i,\mathbf{c}}$ and q_t the usual overlap defined above, we rewrite the free energy of the path as

$$f_{path}(\{\mathbf{m}_t\},\{q_t\}) = \sum_t ((m_t^1)^2/2 + (m_t^2)^2/2 + (m_t^3)^2 + m_t^3(m_t^1 + m_t^2)) + \sum_t (\Phi(q_t) - q_t\Phi'(q_t)) - \frac{1}{\beta}\log Z_{1\mathrm{D}}\,, \tag{28}$$

where

$$Z_{1\mathrm{D}} = \sum_{\{v_t\}} \exp\left[\beta \sum_t m_t^1(\delta_{v_t,\mathbf{a}} + \omega\delta_{v_t,\mathbf{c}}) + m_t^2(\delta_{v_t,\mathbf{b}} + \omega\delta_{v_t,\mathbf{c}}) + m_t^3(\delta_{v_t,\mathbf{a}} + \delta_{v_t,\mathbf{a}} + 2\omega\delta_{v_t,\mathbf{c}}) - \Phi'(q_t)\delta_{v_t,v_{t+1}}\right] \tag{29}$$

As boundary condition, we set the target sequences to be $v_0 = \{\mathbf{a}\}^N$ and $v_{T+1} = \{\mathbf{b}\}^N$.

As we shall see, this model undergoes a first order phase transition at $\omega = \omega_c$ in the limit $\beta \times T \to \infty$. In this limit, when $\omega < \omega_c$, the minimum of the free energy corresponds to the direct solution from v_0 to v_{T+1} that one obtains by restricting the sum in $Z_{1\mathrm{D}}$ over the first two colors only. We will refer to this solution as $\#_2$. When $\omega > \omega_c$, this solution is no longer a minimum of the free energy, and the latter is minimized by global paths introducing novel mutations at intermediate steps with non zero value of m_t^3. Moreover, we will see how ω_c will depend on the parameters of Φ in (24).

7.2 Mimimization of the path free energy in the direct subspace

To do so, we first have to find a solution of the direct problem $\#_2 = \{m_t^1, m_t^2, q_t\}$. The direct solution is given by solving the following coupled equations:

$$m_t^1 = \frac{1}{Z_{1\mathrm{D}}^{dir}} \sum_{\{v_t=\mathbf{a},\mathbf{b}\}^{T-2}} \delta_{v_t,\mathbf{a}} \exp\left[\beta \sum_t m_t^1\delta_{v_t,\mathbf{a}} + m_t^2\delta_{v_t,\mathbf{b}} - \Phi'(q_t)\delta_{v_t,v_{t+1}}\right] \tag{30}$$

$$q_t = \frac{1}{Z_{1\mathrm{D}}^{dir}} \sum_{\{v_t=\mathbf{a},\mathbf{b}\}^{T-2}} \delta_{v_t,v_{t+1}} \exp\left[\beta \sum_t m_t^1\delta_{v_t,\mathbf{a}} + m_t^2\delta_{v_t,\mathbf{b}} - \Phi'(q_t)\delta_{v_t,v_{t+1}}\right]\,, \tag{31}$$

where the partition function $Z_{1\mathrm{D}}^{dir}$ is the same as in (29) but with the sum running over the first two colors **a** and **b** only. Moreover we have $m_t^2 = 1 - m_t^1$.

First we guess that the direct solution is of the form:

$$m_t^{1,dir} = \begin{cases} 1 & \text{for } t/T < \hat{x} \\ 1 - \frac{t/T-\hat{x}}{1-2\hat{x}} + \eta(t/T) & \text{for } t/T \in (\hat{x}, 1-\hat{x}) \\ 0 & \text{for } t/T > 1-\hat{x} \end{cases} , \; q_t^{dir} = \begin{cases} 1 & \text{for } t/T < \hat{x} \\ 1 + \frac{1}{T}\left(-\frac{1}{1-2\hat{x}} + \eta'(t/T)\right) & \text{for } t/T \in (\hat{x}, 1-\hat{x}) \\ 1 & \text{for } t/T > 1-\hat{x} \end{cases} , \tag{32}$$

where we impose $q^{dir}_{\tau=t/T} = 1+\partial_\tau m^{1,dir}_\tau/T$, while η is a perturbation of the order of $1/T^{1/(\alpha+1)}$. We first note that $\hat{x}$ is related to the value of Θ shown in the main paper through $\Theta = 1-2\hat{x}$. We then inject this Ansatz into the equation for m^1 and try to find η and $\hat{x}$ which closes the equation at the zeroth order in T. Plugging this Ansatz into the definition of the partition function $Z^{dir}_{\rm 1D}$ we notice that $-\Phi'(q^{dir}(\tau)) = \lambda\alpha|\gamma - 1/(1-2\hat{x}) + \eta'(\tau)|^{-\alpha-1}$. By rewriting $\gamma - 1/(1-2\hat{x}) + \eta'(\tau) = \xi(\tau)/T^{\frac{1}{\alpha+1}}$, we have $-\Phi'(q^{dir}(\tau)) \sim T$. The linear term in T in the coupling interactions of our 1D model forces the partition function to be dominated by the configurations $v_t = \mathbf{a}$ for $t < \hat{x}T$ and $v_t = \mathbf{b}$ for $t > \hat{x}T$. Hence the partition function can be rewritten as follows:

$$Z^{dir}_{\rm 1D} = T\int_0^1 {\rm d}\tau \exp\left[\beta T\left(\int_0^\tau {\rm d}y m_y^{1,dir} + \int_\tau^1 {\rm d}y(1-m_y^{1,dir}) - \frac{\lambda\alpha}{|\xi(\tau)|^{\alpha+1}}\right)\right] . \tag{33}$$

We can neglect the first order correction in T and maximize the argument in the exponential to obtain the leading term of the partition function. This amounts to solving the following differential equation in $\tau \in (\hat{x}, 1-\hat{x})$:

$$-2\frac{\tau-\hat{x}}{1-2\hat{x}} + 1 + \frac{\lambda\alpha(\alpha+1)\xi'(\tau)}{\xi(\tau)^{\alpha+2}} = 0\,. \tag{34}$$

Solving this differential equation leads to

$$\xi(\tau) = \left[\frac{1}{\xi(\hat{x})^{\alpha+1}} - \frac{\tau^2 - \hat{x}^2 - (\tau-\hat{x})}{\lambda\alpha(1-2\hat{x})}\right]^{\frac{-1}{\alpha+1}} . \tag{35}$$

In order to ensure the continuity of $\Phi'(q_\tau)$ in $\tau = \hat{x}$, we fix $\xi(\hat{x}) = \gamma T^{1/(\alpha+1)}$. By definition of ξ we have $\xi/T^{1/(\alpha+1)} = \gamma - 1/(1-2\hat{x}) + \eta'$. Rewriting $\hat{x} = 1/2 - 1/(2\gamma) + \zeta/T^{1/(\alpha+1)} + o(T^{1/(\alpha+1)})$ we can fix η at the first order as

$$\eta(\tau) - \eta(\hat{x}) = \frac{1}{T^{1/(\alpha+1)}}\left[\int_{\hat{x}}^\tau \xi(y){\rm d}y + 2\zeta\gamma^2(\tau-\hat{x})\right] + o\left(\frac{1}{T^{1/(\alpha+1)}}\right) . \tag{36}$$

By imposing the boundary condition $\eta(\hat{x}) = \eta(1-\hat{x}) = 0$ we can also fix ζ at first order. It is easy to check that (30),(31) are fulfilled at zeroth order by this solution.

7.3 The direct-to-global phase transition

We now write the first derivative of the free energy along the third magnetization m_t^3:

$$\left.\frac{\partial f_{path}}{\partial m_t^3}\right|_{\#_2} = 1 - \left.\langle \delta_{v_t,\mathbf{a}} + \delta_{v_t,\mathbf{b}} + 2\omega\delta_{v_t,\mathbf{c}}\rangle_{\rm 1D}\right|_{\#_2} . \tag{37}$$

By studying this derivatives we will show the existence of a critical ω_c discriminating a regime where all the derivatives vanish ($\omega < \omega_c$) and a regime with negative derivatives ($\omega > \omega_c$).

As above, the average in the RHS of (37) is dominated, for $T \to \infty$, by the ground state path. Two classes of competing configurations must be considered: the usual direct configurations that start in $v_0 = \mathbf{a}$ and turn into $\mathbf{b}$ at some point t such that $t/T \in (\hat{x}, 1-\hat{x})$; another one starting in $\mathbf{a}$ then changing to $\mathbf{c}$ at some point $t = xT$ ($x \in (0, 1/2)$) and then turning into $\mathbf{b}$ when $t = (1-x)T$. The energy of the first set of configurations (for $T \gg 1$) is given by:

$$E_1 = -T\left(\hat{x} + \frac{1}{2}\right) + \frac{\lambda\alpha}{\gamma^{\alpha+1}}\,, \tag{38}$$

while the second configuration has energy

$$E_2(x) = \begin{cases} -T\left(2x + \omega(1-2x)\right) + \frac{2\lambda\alpha}{\gamma^{\alpha+1}} & \text{for } x \le \hat{x} \\ -T\left(2\hat{x} + 2\int_{\hat{x}}^x {\rm d}y\,(1 - \frac{y-\hat{x}}{1-2\hat{x}}) + \omega(1-2x) - \frac{2\lambda\alpha}{|\xi(x)|^{\alpha+1}}\right) & \text{for } x \in (\hat{x}, 1/2) \end{cases} \tag{39}$$

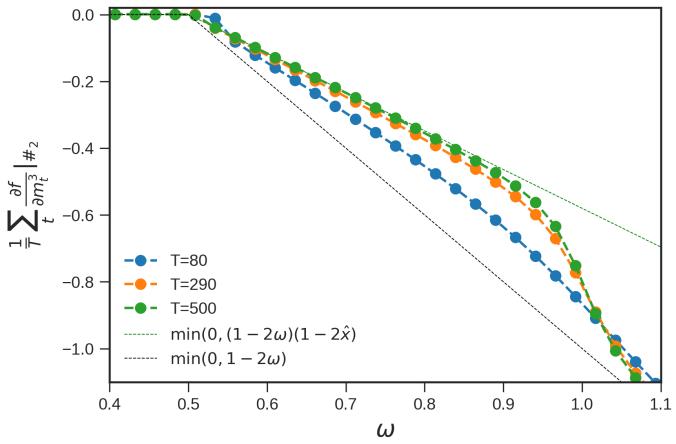

Figure S6: Average value of the free energy derivative along m_t^3. Here $\beta = 3$, $\gamma = 2$, $\alpha = 1$ and $\lambda = 3$.

which is minimized for $x = \hat{x}$ when $\omega \in (1/4, 1)$ and at $x = 0$ when $\omega > 1$. Here the condition $E_2 < E_1$ leads to a phase transition at the critical value $\omega_c = 1/2 + \lambda\alpha/(T\gamma^{\alpha+1}(1-2\hat{x}))$.

The last argument holds until $\hat{x} > 0$. For finite T we could be in a regime where $\hat{x} = 0$. This leads to a different solution for the phase transition. To study this regime we consider $\lambda = \hat{\lambda}T^{(\alpha+2)/(\alpha+1)}$. Injecting it into the expression for the partition function $Z_{\rm 1D}^{dir}$ and expanding the force in terms of $\eta' = \hat{\eta}'/T^{1/(\alpha+1}$, we have

$$Z_{\rm 1D}^{dir} = T\int_0^1 {\rm d}\tau \exp\left[\beta T\left(\int_0^\tau {\rm d}y m_y^{1,dir} + \int_\tau^1 {\rm d}y(1-m_y^{1,dir}) - T^{1/(\alpha+1)}\frac{\hat{\lambda}\alpha}{|\gamma-1|^{\alpha+1}} + \frac{\hat{\lambda}\alpha(\alpha+1)\hat{\eta}'(\tau)}{|\gamma-1|^{\alpha+2}}\right)\right]. \tag{40}$$

Maximization leads to the following solution for η:

$$\eta(\tau) = \frac{|\gamma-1|^{\alpha+2}}{T^{1/(\alpha+1)}\hat{\lambda}\alpha(\alpha+1)}\left[\frac{\tau^3}{3} - \frac{\tau^2}{2} + \frac{\tau}{6}\right], \tag{41}$$

where we have imposed the boundary condition $\eta(0) = \eta(1) = 0$. The condition $-\Phi'(q) > \frac{\lambda\alpha}{\gamma^{\alpha+1}}$ leads to the condition

$$\lambda\alpha > \frac{T}{6}\left(\frac{1}{|\gamma-1|^{\alpha+1}} - \frac{1}{\gamma^{\alpha+1}}\right)^{-1}, \tag{42}$$

which is valid in the case $\lambda = \hat{\lambda}T^{(\alpha+2)/(\alpha+1)}$ and $T \to \infty$. The derivative of the free energy in eq. (37) can be done as above by considering the two classes of relevant configurations. The energy of the first class is

$$E_1 = -\frac{2T}{3} + \frac{\lambda\alpha}{|\gamma-1|^{\alpha+1}} \tag{43}$$

while the second class corresponds to energy

$$E_2 = -T\left(\omega + \frac{1}{3}\right) + \frac{\lambda\alpha}{|\gamma-1|^{\alpha+1}}. \tag{44}$$

The condition $E_2 < E_1$ leads to a new critical value $\omega_c = 1/3 + \lambda\alpha/(T|\gamma-1|^{\alpha+1})$. Merging together the two regimes studied above, we find that the critical line at the first order in $1/T$ is given by:

$$\omega_c = \max\left(\frac{1}{2} + \frac{\lambda\alpha}{T\gamma^{\alpha+1}\Theta}, \frac{1}{3} + \frac{\lambda\alpha}{T|\gamma-1|^{\alpha+1}}\right), \tag{45}$$

where $\Theta = 1 - 2\hat{x}$. In Figure S6 we plot the behavior of the derivative of the free energy for different value of ω, showing the instability at the corresponding critical ω_c.

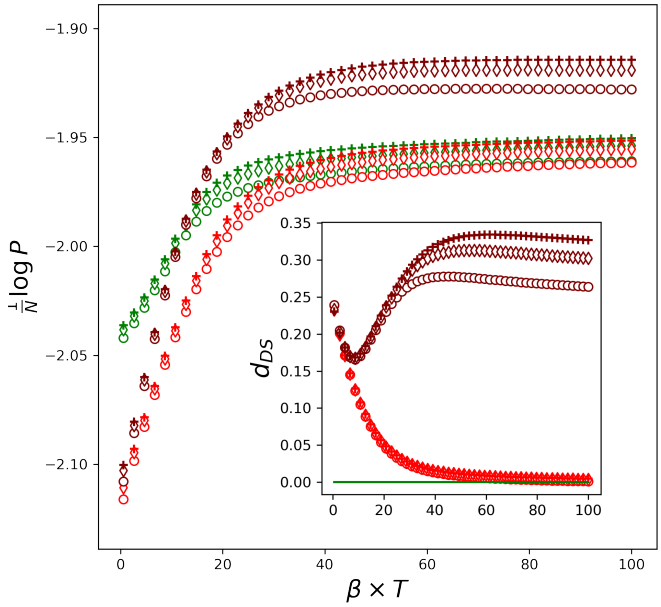

Figure S7: Average log-likelihood and distances to direct space (inset) of intermediate sequences as a function of $\beta \times T$. Symbols stands for different T (circles for T=20, diamonds for T=30 and pluses for T=40). Green symbols represents direct solutions (which are of course independent of ω), Red symbols represents global solutions with $\omega = 0.4$ and maroon symbols represent global solutions for $\omega = 0.7$. The other parameters are set to $\alpha = 1$, $\gamma = 2$, $\lambda = 1$.

The crossover is visible in Figure S7, where can observe the coincidence of the average log-likelihoods of intermediate sequences along direct and global paths at large T for small ω, and the higher quality of global paths for large ω. Notice that these results are valid when T is sent to large values while keeping β fixed. If β is small, e.g. of the order of $\frac{1}{T}$, the domination of global paths on direct paths is due to the larger entropy of the former. Figure S7 shows that, for small $\beta \times T$, global paths are indeed of lesser quality (probability) than their direct counterparts, even at high ω.

References

[BFH+18] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. Available online at: `http://github.com/google/jax`.

[ES99] Xavier Espanel and Marius Sudol. A single point mutation in a group i ww domain shifts its specificity to that of group ii ww domains. *Journal of Biological Chemistry*, 274(24):17284–17289, 1999.

[FI12] Asja Fischer and Christian Igel. An Introduction to Restricted Boltzmann Machines. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 14–36. Springer, Berlin, Germany, September 2012.

[JGS+16] Hugo Jacquin, Amy Gilson, Eugene Shakhnovich, Simona Cocco, and Rémi Monasson. Benchmarking inverse statistical approaches for protein structure and design with exactly solvable models. *PLOS Computational Biology*, 12(5):1–18, 05 2016.

[LD89] Kit Fun Lau and Ken A. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22(10):3986–3997, October 1989.

[LG98] Michael Levitt and Mark Gerstein. A unified statistical framework for sequence comparison and structure comparison. *Proceedings of the National Academy of sciences*, 95(11):5913–5920, 1998.

[MJ96] Sanzo Miyazawa and Robert L. Jernigan. Residue – Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. *J. Mol. Biol.*, 256(3):623–644, March 1996.

[TCM19] Jérôme Tubiana, Simona Cocco, and Rémi Monasson. Learning protein constitutive motifs from sequence data. *eLife*, 8:e39397, March 2019.

[Tie08] Tijmen Tieleman. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 1064–1071. Association for Computing Machinery, New York, NY, USA, July 2008.

[Tub18a] Jérôme Tubiana. *Restricted Boltzmann machines : from compositional representations to protein sequence analysis.* PhD thesis, Université Paris sciences et lettres, Paris, France, November 2018.

[Tub18b] Jerome Tubiana. Probabilistic graphical models (pgm), 2018. Available online at: `https://github.com/jertubiana/PGM`.

[ZS04] Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

8 Appendix A: Weights Logo for the WW domain

link to the RBM trained on WW domain data.

9 Appendix B: Weights Logo for the Lattice Proteins

link to the RBM trained on LP domain data.