JCTC Journal of Chemical Theory and Computation

Interpretable Machine Learning of Amino Acid Patterns in Proteins: A Statistical Ensemble Approach

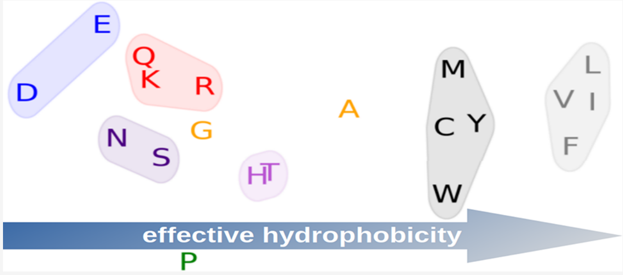
Anna Braghetto, Enzo Orlandini, and Marco Baiesi*

Cite This: *J. Chem. Theory Comput.* 2023, 19, 6011−6022

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Explainable and interpretable unsupervised machine learning helps one to understand the underlying structure of data. We introduce an ensemble analysis of machine learning models to consolidate their interpretation. Its application shows that restricted Boltzmann machines compress consistently into a few bits the information stored in a sequence of five amino acids at the start or end of α-helices or β-sheets. The weights learned by the machines reveal unexpected properties of the amino acids and the secondary structure of proteins: (i) His and Thr have a negligible contribution to the amphiphilic pattern of α-helices; (ii) there is a class of α-helices particularly rich in Ala at their end; (iii) Pro occupies most often slots otherwise occupied by polar or charged amino acids, and its presence at the start of helices is relevant; (iv) Glu and especially Asp on one side and Val, Leu, Iso, and Phe on the other display the strongest tendency to mark amphiphilic patterns, i.e., extreme values of an *effective hydrophobicity*, though they are not the most powerful (non)hydrophobic amino acids.



1. INTRODUCTION

Various machine learning (ML) methods are applied to proteins.[1−25] For example, outstanding advancements have shown how ML can boost the prediction of protein native states[13−15] and complexes[16,17] based only on amino acid sequences. However, the aim of several approaches is not to achieve a reliable (black box) tool for protein structure prediction but to get informative knowledge from the big data available for protein sequences and structures.

Interpretable ML[26,27] focuses on understanding the cause of a model's decision and enhancing human capability to consistently predict the model's result. Interpretable ML versions are more complex and informative than standard statistical analysis and can improve our understanding of proteins.[1,2,4−10] In particular, they might detect patterns not emerging naturally from studying the abundance and correlations of amino acids in secondary structures. Among the well-known patterns, for instance, there is the amphiphilic structure of several α-helices and β-sheets,[28,29] which are mostly (charged or) polar ($\mathbb{P}$) on one side and nonpolar ($\mathbb{N}$) on the other side. In an α-helix, with pitch of ∼3.6 residues, the typical (non)polarity switch occurs every two residues. On the other hand, in a β-sheet, the three-dimensional alternation of the side chains takes place at every step. Hence, an amphiphilic sequence would be, for example, $\mathbb{PNPNP}$.

In this work, we use a simple form of interpretable unsupervised ML, restricted Boltzmann machines (RBMs),[30−40] which allow extraction of deep, nontrivial insight without losing the most transparent information on data statistics encoded in local biases. Conveniently, the weights and biases learned by RBMs can be visualized and easily interpreted. This established approach has already revealed correlated amino acids within protein families,[1] drug−target interactions,[2] and correlations within DNA sequences.[41,42]

A novelty of our work is a statistical ensemble approach to unsupervised ML, which improves the robustness of the findings. By training RBMs of the same size but with different weight initializations, we checked whether they all converge to the same final set of learned weights. The maximally complex RBMs preserving this ensemble coherence are optimal, as they perform encoding of the correlations within data samples while providing stable and transparent information on the data.

We show that our optimal RBMs perform extreme information compression to two or three bits, encoding the essential correlations between amino acids at the beginning or end of α-helices and β-sheets. In addition to recovering the expected amphiphilic structures, this approach (i) discovers more subtle yet relevant amino acid patterns in each portion of the secondary structure and (ii) provides evidence of similarity

Received: April 5, 2023
Published: August 8, 2023

https://doi.org/10.1021/acs.jctc.3c00383

between amino acids' roles in these structures, including some surprising ones.

RBMs distinguish two classes of amino acids, which we map to ℙ and ℕ, as shown in Table 1. Contrary to standard classification,[43] but similar to some partitionings (see references collected by Stephenson and Freeland[44]), Tyr belongs to the class ℕ of hydrophobic amino acids. Pro is mostly ℙ, as discussed in detail below. Some surprising subclasses emerge, especially by looking at the results in α-helices, where it turns out that Thr and His play a similar weak role in the amphiphilic patterns. RBMs classify Asp and Glu on the one side and Val, Leu, Iso, and Phe on the other as the most diverse amino acids. However, Trp has the highest experimental hydrophobicity, while Arg and Lys have the lowest values.[45] To explain this finding, we argue that RBMs detect a kind of *effective hydrophobicity*, emphasizing how deeply amino acids play a hydrophobic or hydrophilic role in the amphiphilic alternation in α-helices and β-sheets. These findings only partially overlap with those expressed by known diagrams of consensus amino acid similarity.[44]

Table 1. RBM-Based Classification of Amino Acids[a]

ℙ	charged	Arginine	Arg	R
		Lysine	Lys	K
		Histidine	His	H
		Aspartic acid	Asp	D
		Glutamic acid	Glu	E
	polar	Asparagine	Asn	N
		Glutamine	Gln	Q
		Serine	Ser	S
		Threonine	Thr	T
ℕ		Tyrosine	Tyr	Y
	nonpolar	Tryptophan	Trp	W
		Valine	Val	V
		Leucine	Leu	L
		Isoleucine	Iso	I
		Phenilalanine	Phe	F
		Methionine	Met	M
		Cysteine	Cys	C
∼ ℙ		Proline	Pro	P
		Glycine	Gly	G
∼ ℕ		Alanine	Ala	A

[a] The first column labels which amino acids can be classified as polar/hydrophilic (ℙ) and nonpolar/hydrophobic (ℕ) according to the weights of our RBMs. The second column shows the textbook classification of amino acids.[43] According to RBMs, Tyr behaves as a nonpolar amino acid, Pro behaves mostly as a polar one (∼ℙ), and Ala is slightly ℕ only in β-sheets. Gly is neither clearly ℙ nor ℕ. The color of each amino acid symbol in the last column (and the offset) follows the subgrouping we introduce based on the weights learned by the RBMs applied to α-helices.

2. METHODS

Here we describe the datasets and the methods used to analyze them. Some more technical details are reported in Supporting Information (SI) section S1.

2.1. Data. To each protein sequence stored in the reduced CATH ensemble of 31 884 natural proteins,[46] we apply the DSSP algorithm[47,48] to determine the secondary structure to which every amino acid belongs (see SI section S2 for comparison with results obtained with the STRIDE algorithm[49]). Then we collect all sequences within α-helices and β-sheets long enough to contain $\Gamma = 5$ amino acids. We then build four sets: one with the first $\Gamma = 5$ amino acids in α-helices (following the standard orientation from the N terminus to the C terminus of the protein), one with the last Γ amino acids in α-helices, and the same for two more sets at the start and end of β-sheets. The two sets referring to the first and last $\Gamma = 5$ amino acids in α-helices contain 129300 samples each, while the other two sets, concerning the start and end of β-sheets, include 101 382 sequences each.

We used one-hot encoding to represent amino acids. That is, the kth amino acid is stored as a sequence $\mathbf{v}^k = (-1, -1, ..., +1, ..., -1, -1)$ of 20 integers where only a +1 element is present at position k. This encoding is how an RBM reads the amino acid in a portion of its visible units. A sequence of Γ amino acids $(k_1, ..., k_\Gamma)$ is thus translated into one-hot encoding stacked as $\mathbf{v} = (\mathbf{v}_1^{k_1}, \mathbf{v}_2^{k_2}, ... \mathbf{v}_\Gamma^{k_\Gamma})$ giving a total of $N_v = 20\Gamma$ digits in a data sample.

To monitor the training of each RBM, we randomly split the data into a training set (80%) and a validation set (20%). The training set is used to optimize the RBM parameters and compute the pseudo-log-likelihood (PLL) function, which measures the quality of data reconstruction by RBMs.[50] The PLL of the validation set is then used to check the performance of the RBM in reproducing the statistics of new data. Note that in principle this procedure could cause differences in the results. However, the cases we find in the ensemble of RBMs, as explained below, reveal when variability is small and highlight general patterns.

2.2. Restricted Boltzmann Machines. The RBM is an unsupervised machine learning method based on a simple neural network architecture. It aims to reproduce the empirical distribution of data samples by encoding the correlations between their elements, the *visible units* v_i $(1 \le i \le N_v)$. This encoding uses a set of parameters and a layer of hidden (or latent) variables, h_j $(1 \le j \le N_h)$. The parameters defining the method are the weights w_{ij} in an $N_v \times N_h$ matrix connecting visible to hidden units and the local biases that act on both the visible (a_i) and hidden (b_j) units. Figure 1 shows a sketch of an RBM. The statistical weight of a $(\mathbf{v}, \mathbf{h})$ configuration is given by

$$e^{-E(\mathbf{v},\mathbf{h})} = \exp\left(\sum_{i=1}^{N_v}\sum_{j=1}^{N_h} v_i w_{ij} h_j + \sum_{i=1}^{N_v} a_i v_i + \sum_{j=1}^{N_h} b_j h_j\right) \quad (1)$$

It resembles a Boltzmann weight with energy $E(\mathbf{v}, \mathbf{h})$, for which we will use "spin" variables $v_i = \pm 1$, $h_j = \pm 1$. Since this version generates a finite number of hidden states (2^{N_h}), it facilitates an interpretation of the structure of weights between hidden and visible units and of local biases. Initially, weights and biases of untrained RBMs are drawn randomly from chosen distributions.

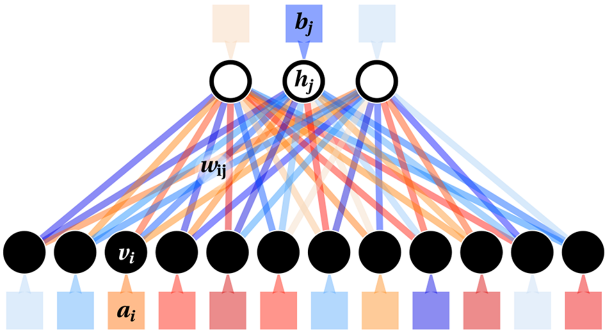


Figure 1. Sketch of an RBM with $N_v = 12$ visible units (black circles, where data are given as an input) and $N_h = 3$ hidden units (white circles). Red and blue shadings indicate positive and negative values of single weights (plotted as lines joining units in the two layers) and biases (boxes next to units).

The bipartite structure of the RBM allows easy generation of **h** from **v**. This step should encode the correlations within data sequences in N_h hidden units for a trained RBM. When $N_h \ll N_v$, the RBM acts as an *information bottleneck* enforcing such a simple model, with its small resources, to capture the crucial properties of the analyzed data. The $\mathbf{v} \to \mathbf{h}$ step selects each h_i independently with probability

$$p(h_j|\mathbf{v}) \sim \exp\left[h_j\left(b_j + \sum_{i=1}^{N_v} w_{ij}v_i\right)\right] \tag{2}$$

Similarly, one generates **v** if **h** is known through $\mathbf{v} \sim p(\mathbf{v}|\mathbf{h})$. Each of the Γ blocks $\mathbf{v}_\gamma$ is generated independently. The indices $i \in I(\gamma)$ of the 20 weights w_{ij} pointing to segment $\mathbf{v}_\gamma$ are those relevant for its sampling. By remapping these indices i to the interval $k = 1, ..., 20$, we pick an amino acid k with probability

$$p(k|\mathbf{h}) \sim e^{2\phi_k(\mathbf{h})} \tag{3}$$

with

$$\varphi_k(\mathbf{h}) = a_k + \sum_{j=1}^{N_h} w_{kj}h_j \tag{4}$$

where $\varphi_k(\mathbf{h})$ is the local field on the site k.

The core of the training of an RBM consists of sampling values in visible and hidden units through an algorithm termed contrastive divergence with n Monte Carlo steps (CD-n) .[51,52] We alternatively sample from conditional distributions, starting from a data sample $\mathbf{v}_0$ at $t = 0$ up to $t = n$ steps: $\mathbf{h}_{t+1} \sim p(\mathbf{h}|\mathbf{v}_t)$ and $\mathbf{v}_{t+1} \sim p(\mathbf{v}|\mathbf{h}_{t+1})$. The statistics of the sampled configurations allow the estimation of the gradient of the data log-likelihood according to the Boltzmann weight (eq 1) in the directions of all parameters w_{ij}, a_i, and b_j. Then we apply a standard gradient ascent algorithm (in our case, Adam[52]). In addition to CD-n, we use persistent CD-n (PCD-n), a variant that should better sample the configurational space.[52,53]

Once trained, the same sampling procedure may generate realistic amino acid sequences (k_1, ..., k_Γ). Through eq 3, an RBM decides how to decode the hidden units **h** and generate a sequence.

2.3. Ensemble of RBMs. A novelty of this work is using a statistical ensemble analysis of machine learning. As a trade-off between computational costs and statistical relevance, we run $R = 30$ independent realizations of RBMs with the same size N_v, N_h and the same number of training epochs (we found that typically 50 or 100 epochs are enough to train the model). Note that each realization differs by the initialization of the weights and the random splitting of the data into training and validation sets.

From now on, let us assume that each hidden unit j in a single RBM is characterized by its set of weights w_{ij} ($1 \le i \le N_v$). All hidden units from RBM realizations are then collected in an ensemble of hidden units to study their properties and check whether common patterns exist. For every pair of hidden units j, m, the Euclidean distance $d_{jm} = \sqrt{\sum_i |w_{ij} - w_{im}|^2}$ estimates their similarity. When used in a clustering algorithm, it identifies *groups* of hidden units (see SI section S1 for more details).

By averaging the weights w_{ij} within each group and the biases a_i and b_j of each RBM, we build their average RBM (aRBM): this is supposed to represent the best summary of the relevant information learned by the ensemble. First, we use the aRBM to compute the probability of the 2^{N_h} possible hidden states given that **v** are all points in a dataset (eq 2). Then, from hidden states weighted with their probabilities, we use eq 3 to verify the ability of the aRBM to faithfully reproduce the statistics of the original dataset in the visible space.

2.4. Selecting the Number of Hidden Units. The main aim is to find simple patterns representative of the redundant, generic correlations in amino acid sequences (at the start of α-helices, etc.) while neglecting specific patterns of single sequences with RBMs. The key to achieving this goal is the information bottleneck obtained by setting a small number of hidden units N_h.

We monitor how many groups of hidden units emerge by increasing N_h (Figure 2a for the CD-1 training of RBMs and Figure 2b for PCD-10). Generally, the ratio of groups to hidden units stays maximal up to $N_h = 3$ for α-helices and $N_h = 2$ for β-sheets. For these values of N_h, almost all RBM realizations have the same palette of hidden units. Only in a few cases does the clustering algorithm (see SI section S1) classify units as *noise* due to their significant diversity from all other ones. Note that beyond these values of N_h, there is no clear one-to-one correspondence between hidden units in an RBM and groups, and uniformity in the ensemble of RBMs is lost.

To evaluate the performance of the RBMs, we compute the PLL as a function of N_h (see Figure 3). From $N_h = 1$, for CD-1 and PCD-10, the PLL quickly reaches a plateau around $N_h \approx 3$. By adding more hidden units ($N_h > 3$), one does not significantly improve the performance of the RBM. Moreover, for CD-1 we can go up to $N_h = 30$, finding a decreasing trend of the PLL for large N_h values in all cases. Hence, more complex RBMs are heterogeneous and suboptimally trained by the oversimplified CD-1 algorithm.

All considered, we show the results for $N_h = 3$ for α-helices and $N_h = 2$ for β-sheets. From now on, we will discuss only the results from PCD-10. Those from CD-1 are similar.

3. RESULTS

3.1. How to Read Weight Patterns. We summarize the properties of the ensemble of RBMs via a set of plots, as reported, for instance, in Figure 4 for the starting strand of α-helices. We average the values for weights in a group or biases from all RMBs in the ensemble. Thus, the displayed values represent the aRBM.

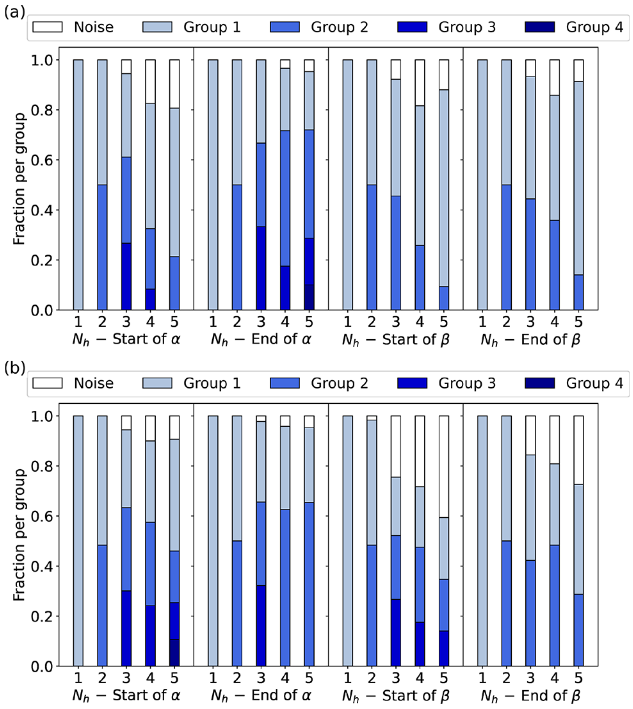


Figure 2. For (a) CD-1 and (b) PCD-10, we show the number and relative size of groups emerging from clustering hidden units in the ensemble of RBMs for every position of the secondary structure that we study. In both cases, we conclude the following: for α-helices, N_h = 3 is the optimal number of hidden units, while for β-sheets it is N_h = 2. These are the maximum values where the number of groups matches the number of hidden units and the noise is still tiny, i.e., where each RBM in the ensemble has learned the same set of hidden units.

Figure 4a reports the table of distances between any two hidden units of the ensemble of RBMs. The units are sorted and collected into the groups (colored squares with light internal colors along the diagonal) detected by the clustering algorithm. Some units, marked as "noise", are not assigned to any group.

Figure 4b shows the visible bias a_i of the aRBM, reshaped to a 20 × Γ matrix for better readability. The exponential of this bias is a good indicator of the mean probability of finding an amino acid at a specific position in the sequence. For instance, it shows that Glu (E) has a high chance of appearing at positions 2 and 3. For each of the groups, we show in Figure 4c−e the weights w_{ij} of the corresponding hidden units in the aRBM. In addition, the N_h biases b_j of the aRBM are specified in the caption.

A table such as the one reported in Figure 4c may be read as follows: a hidden unit in group 1 "pushes" a random pattern of amino acids biased by the weights toward a particular sequence, depending on its stored value, $h_1 = \pm 1$. According to eqs 3 and 4, $h_1 = 1$ raises the probability of picking amino acids with weights w_{i1} to a value significantly larger than zero (red shades). For instance, one can notice that D is chosen more frequently at position 3 and I, L, V, and F at positions 1 and 5. The opposite happens if $h_1 = -1$.

All other random choices are possible with a gradually lower probability. The unit does not (de)select any particular amino acid when weights have values close to zero (light colors in the table). Instead, another unit may be the one that drives the sequence selection at that position. For instance, in Figure 4c,d we see that units in group 1 and group 2 have a strong set of weights at positions 1, 3, and 5, which are complementary to those of group 3 (stronger at positions 2 and 4; see Figure 4e). Therefore, the units in different groups may take care of different alternating slots in the sequence.

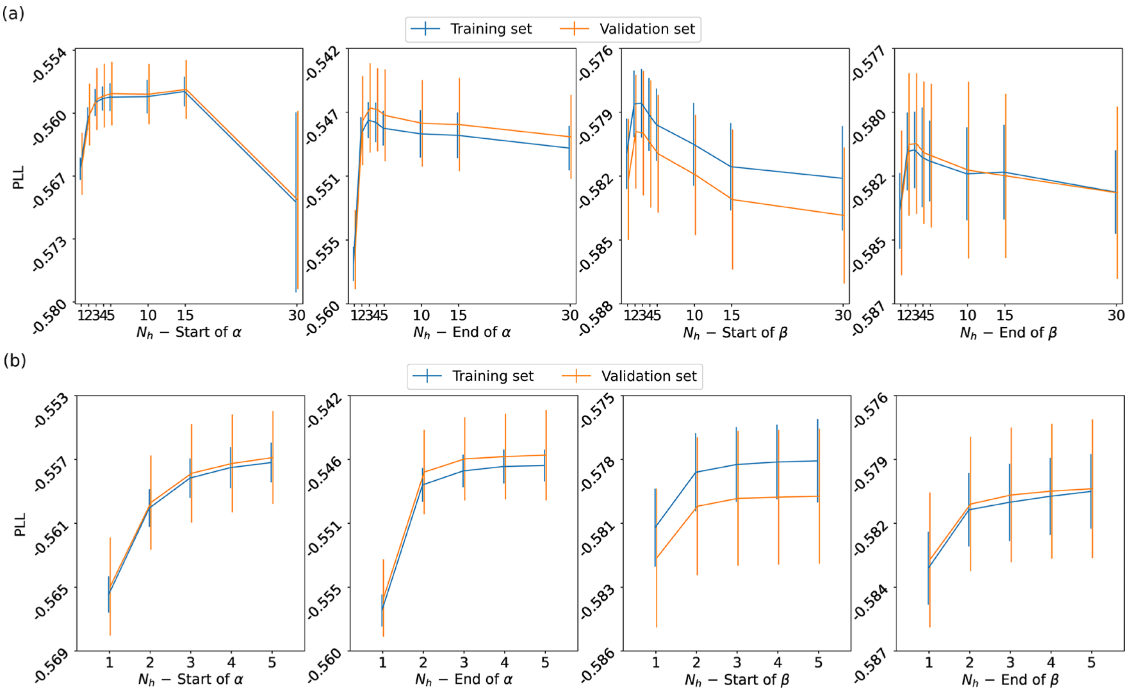


Figure 3. Pseudo-log-likelihood as a function of the number of hidden units for RBMs trained with (a) CD-1 and (b) PCD-10, shown for each of the four segments of secondary structure that we study. The PLLs for the training and validation sets are compatible, showing that the RBMs have achieved robust training.

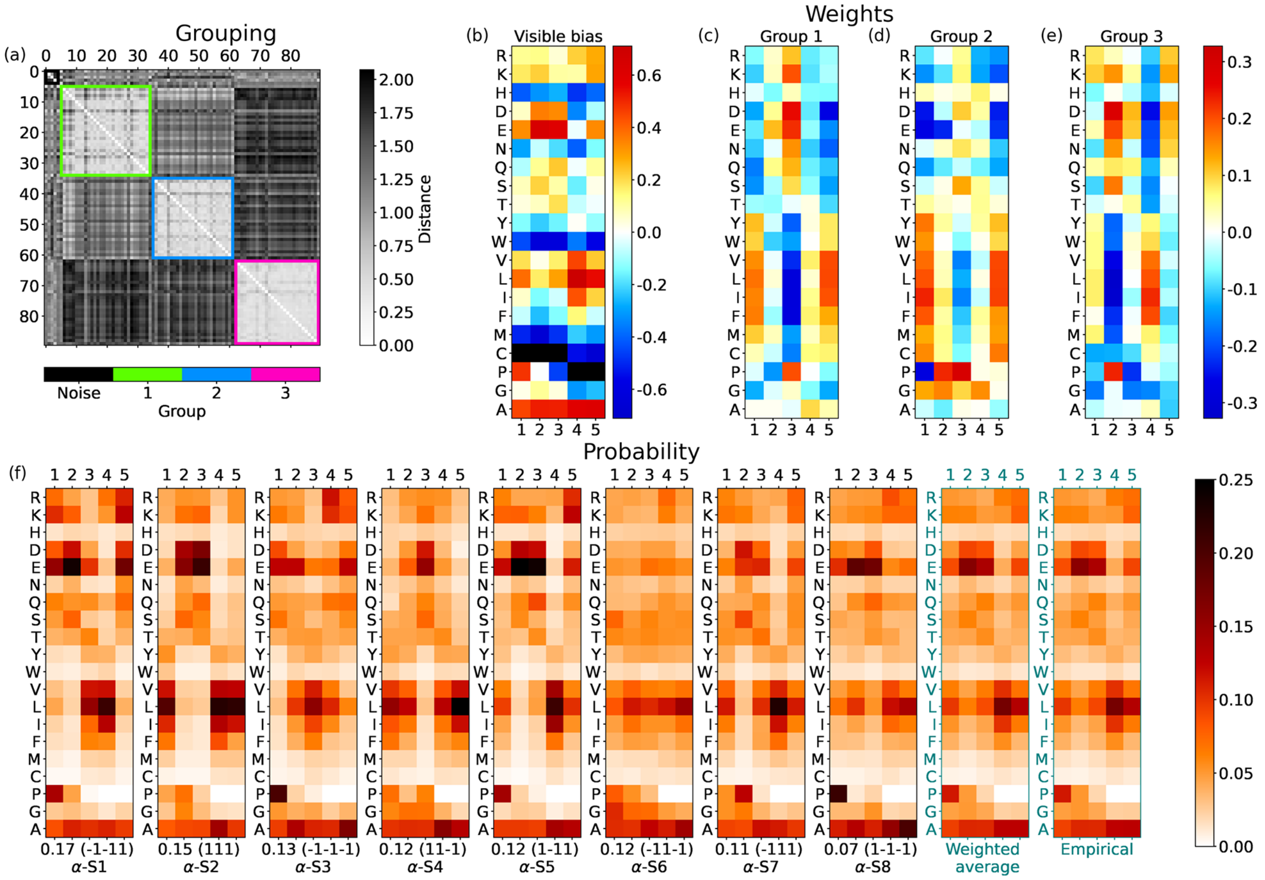


Figure 4. For the start of α-helices, with $N_h = 3$ hidden units: (a) Matrix with gray shade indicating the distance d_{jm} between the weights w_{ij} and w_{im} of different hidden units j and m; the colored boxes highlight the groups found by the DBSCAN clustering. (b) Average biases a_i learned by the ensemble of RBMs, reshaped from an array with $20\Gamma = 100$ entries to a $20 \times \Gamma$ table, in which each column corresponds to a given encoding $\mathbf{v}_\gamma$ and each row to a given amino acid (a similar scheme is used in (c−e)). Values more negative than the lower threshold in the scale are marked with black squares (in this case for Cys and Pro, which essentially leads to the negligible probability of finding these amino acids in those positions). (c−e) Average weights of units in groups 1, 2, and 3. (f) The shade of each slot in each panel shows the probability of picking a specific amino acid at a given position. Hence columns are normalized to 1. The first $2^{N_h} = 8$ panels show the probabilities for every hidden state (the sequence of ±1's in the parentheses at the bottom, where it follows the value of its empirical frequency). Hidden states are labeled and ranked with decreasing frequency; e.g., α-S1 is the most probable hidden state at the start of α-helices. The last two panels show the average of RBM α-S states weighted according to their frequency and the actual probability of amino acids at the $\Gamma = 5$ initial positions of α-helices. In practice, the prescription of the RBM for reconstructing meaningful sequences would be (i) to pick a hidden state at random according to its frequency and (ii) according to probabilities in its table, for every position $\gamma \leq \Gamma$, to pick an amino acid at random. The values of the hidden bias in the aRBM for each group are $b_1 = -1.129$, $b_2 = 1.270$, and $b_3 = -1.496$.

The first $2^{N_h} = 8$ panels in Figure 4f represent probabilities (eq 3) to choose amino acids at every slot $\gamma \leq \Gamma$ (normalized in columns at fixed γ) if the aRBM is in a given state $\mathbf{h}$. With eq 2, we compute the probability of each of the $2^{N_h} = 8$ hidden states $\mathbf{h}$ from the biases b_j, weights w_{ij}, and $\mathbf{v}$ in the dataset. This is indicated below each panel in Figure 4f (e.g., 0.18). After this, we also specify the state $\mathbf{h}$ (e.g., $(1\ -1\ -1)$) and a chosen label (e.g., α-S1). We rank the $\mathbf{h}$ states in Figure 4f from the most frequent to the least likely.

Each of the first 2^{N_h} panels of Figure 4f thus displays a typical correlation of probabilities followed by the aRBM to build a sequence of amino acids. The last two panels are the average of the first 2^{N_h} panels, weighted with their frequencies, and the empirical average in the dataset.

In the following, we specify the discussion of the four regions of the secondary structure analyzed in this work.

3.2. Start of α-Helices.

The first training set we study contains stretches of the first $\Gamma = 5$ positions in all (long enough) α-helices in proteins of the CATH database. The corresponding set of trained RBMs with $N_h = 3$ yields three significant groups of hidden units (see Figure 4a). For $N_h = 1$, 2, 3, we see groups 1, 3, and finally 2, respectively. By including additional hidden units, we continue to observe these three groups, confirming that RBMs encode the main patterns within the analyzed sequences with three hidden units.

Figure 4b shows the bias a_i of the aRBM. It is quite structured compared to other cases, shown later as the end of α-helices and β-sheets. This structure denotes a tendency of amino acids to appear more frequently at specific positions. Notice the pattern of Pro, with high intensity (red) at the first position, which sensibly decreases in the next positions (the black color means that a_i is below the lower level of the scale), in agreement with the known abundance of Pro at the start of helices.[54] Notably, at position $\gamma = 4$, there stands out a peculiar behavior: a high intensity for nonpolar amino acids (in particular Val (V), Leu (L), and Iso (I)) aligns with a low intensity for polar amino acids (especially Asp (D), Glu (E), and Asn (N)). Consistently, an average depletion of a polar

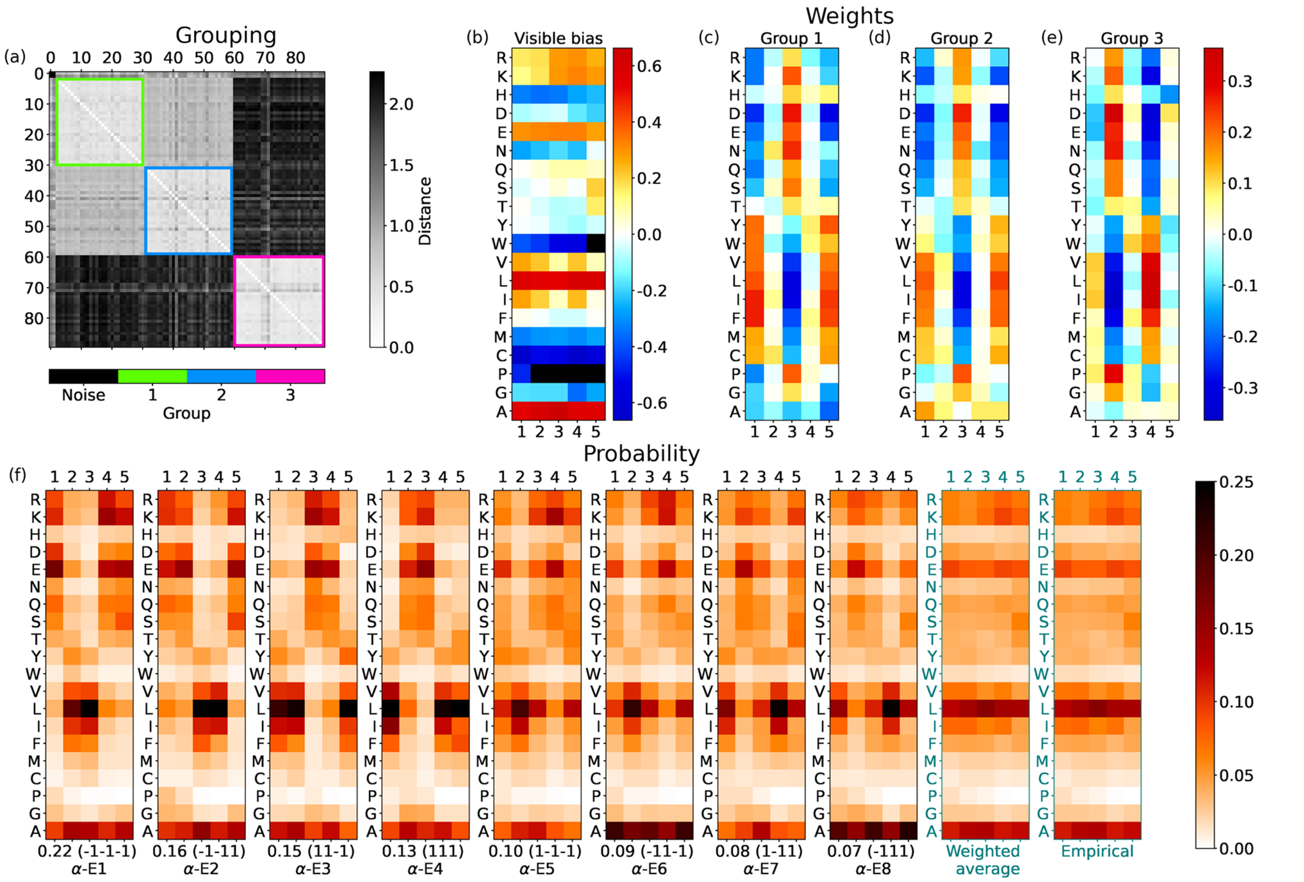


Figure 5. For the end of α-helices with three hidden units, the same scheme as in Figure 4. Hidden bias: $b_1 = 0.541$, $b_2 = -0.410$, and $b_3 = -0.191$.

amino acid at position $\gamma = 4$ at the start of α-helices is visible in the empirical statistics, shown in the last panel of Figure 4f.

In addition to the average trend dictated by the bias, the aRBM, thanks to the hidden units, can modulate the correlations among amino acids in single sequences. Hidden units in group 1 (Figure 4c) address anticorrelations between $\mathbb{P}$ and $\mathbb{N}$ amino acids at positions $\gamma = 1, 3, 5$. For instance, $h_1 = 1$ promotes the pattern $\mathbb{N}$-$\mathbb{P}$-$\mathbb{N}$ while $h_1 = -1$ promotes $\mathbb{P}$-$\mathbb{N}$-$\mathbb{P}$. Group 3 (Figure 4e) instead mainly encodes the correlations among amino acids at positions $\gamma = 2, 4$. Group 2 (Figure 4d) is similar to group 1 but also displays a set of large weights for Pro. This set adds significant insight into the correlations between Pro as a starter of helices and its following amino acids (the bias did not show such a rich structure): for example, weights in group 2 suggest that P, E, and D are interchangeable at the position $\gamma = 1$ and that they are strongly correlated with D and E at $\gamma = 5$ and anticorrelated with P at $\gamma = 3$.

Given the aRBM, we check the states in the hidden space in Figure 4f, allowing us to merge the information from biases and weights. Different configurations appear, but almost all show a repeated scheme with polar and nonpolar amino acid alternation with blocks of about two elements, consistent with an amphiphilic structure in α-helices. More interestingly, states α-S1, α-S3, α-S5, and α-S8 (sharing $h_2 = -1$ that promotes Pro in group 2) include the activation of Pro at the start of the sequence, paired with Glu in the second position (for this subset of α-helices, we notice that Glu's activation is not fixed only at the second position but is active also at the first or third position). This pattern provides two main classes of amino acid alternation: (Pro)$\mathbb{PNNP}$ or (Pro)$\mathbb{PPNN}$. In this context, Pro behaves as polar, with a higher frequency (α-S1, α-S3), or as nonpolar, with a lower frequency (α-S5, α-S8).

We have thus shown that training led the RBMs to automatically detect and decompose the start of α-helices into eight nontrivial modes. The reverse, trivial process of averaging their probabilities leads to the average behavior shown in the second-to-last panel of Figure 4f, which matches the empirical probabilities (last panel). Notably, the RBM decomposition would not be accessible a priori by standard statistical tools. Moreover, the discovered heterogeneous eight modes *generate* synthetic sequences, each with its own probabilistic pattern.

3.3. End of α-Helices.

The results from RBMs with $N_h = 3$ for the last $\Gamma = 5$ amino acids of the α-helices are displayed in Figure 5. Again, three groups of hidden units emerge from clustering their weights. For $N_h = 4$, the values would remain the same. However, by increasing N_h from $N_h = 1$, we note that groups 1 and 2 are represented by their averaged version for $N_h \leq 2$, while they split for $N_h = 3$. This splitting is convincing: indeed, the PLL slightly increases in the $N_h = 2 \rightarrow 3$ step (Figure 3), and above all, the division into separate groups by the clustering is clear (see Figure 5a).

Groups 1 and 2 determine the alternation of $\mathbb{P}$ and $\mathbb{N}$ at positions $\gamma = 1, 3, 5$. What distinguishes them is the weight pattern of Ala, which flips its sign from one group to the other (see Figure 5c,d. Group 3 instead fixes the alternation of $\mathbb{P}$ and $\mathbb{N}$ at positions $\gamma = 2, 4$.

The visible bias in Figure 5b shows that amino acids distribute almost uniformly at different positions at the end of

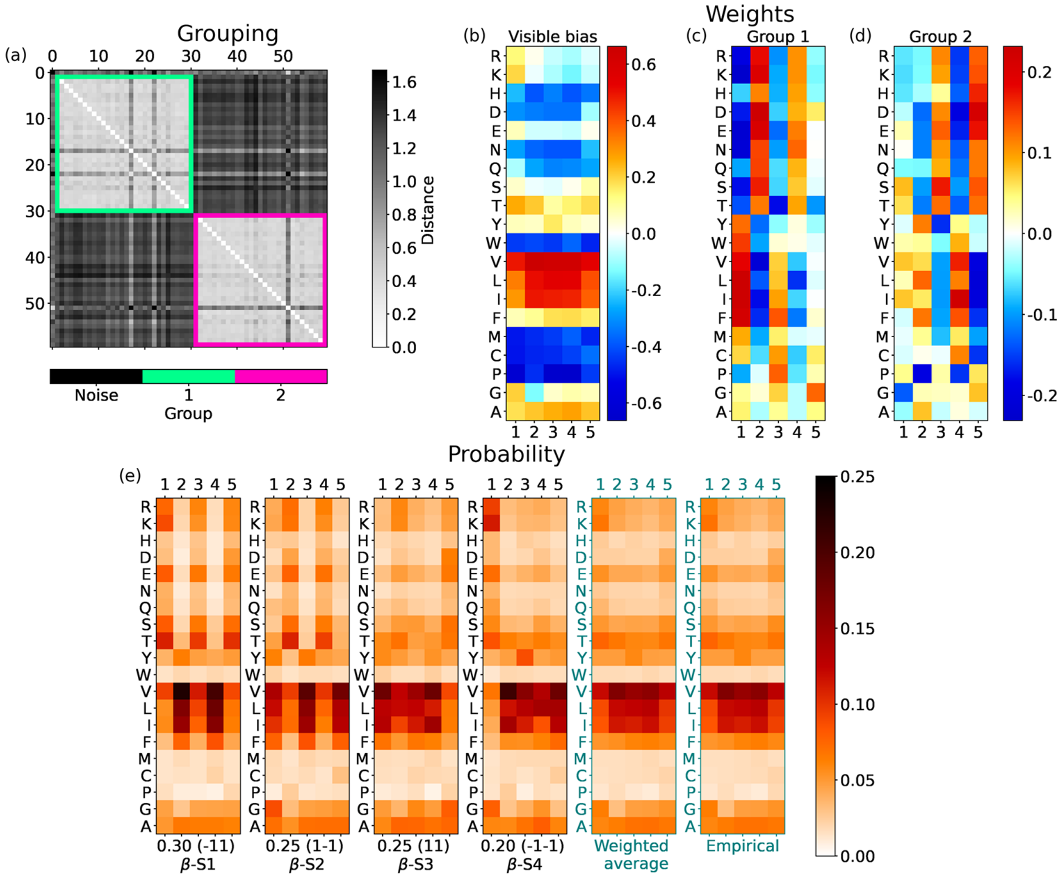


Figure 6. For the start of β-sheets with two hidden units, the same scheme as in Figure 4. Hidden bias: b_1 = 0.458 and b_2 = −0.002.

α-helices, with a significantly high bias toward Leu (L) and Ala (A). However, some slight deviations from the general behavior are visible. For example, at the last position of the helix (γ = 5), some polar amino acids are more probable (see N, S, and T), while some nonpolar ones become less likely (see V, I). Note also the low bias of Gly at the next-to-last position (γ = 4).

In the hidden space, the aRBM reproduces, on average, the visible statistics (see Figure 5f). As observed at the start of α-helices, some states (α-E1, α-E2, α-E3, and α-E4) report the polar and nonpolar alternation with period ∼2. More interestingly, in states α-E6 and α-E8, the behavior of Ala spikes with a high probability in every position. As known, Ala is a strong helix stabilizer.[55,56] Consistently, Ala has a high bias a_i in the RBM and thus can act as a wild card: its placement in a typical sequence at the end of an α-helix is relatively free, and it fits even at the specific positions of charged and polar amino acids. This high bias was also visible at the start of α-helices (Figure 4b), where no weight pattern induces the splitting of hidden units into separate groups based on Ala. The boosted probability of Ala in states α-E6 and α-E8 reveals a subclass of α-helix endings (15% of the cases) richer in Ala than typical α-helices. We have verified a posteriori that AAAAA is among the 10 most frequent sequences at the end of α-helices. Hence, polyalanine[56] is a characterizing feature of the terminal part of α-helices.

Finally, note that the patterns of the states α-E1, α-E2, α-E3, and α-E4 are similar but somehow shifted. Our explanation is that α-helices may end with different "phases" for exposure to the solvent. In some cases, it is convenient for the last amino acids of a helix to be polar; in others, it is the opposite. Patterns in α-E5, α-E6, α-E7, and α-E8 show shifts of polarity that satisfy different needs.

3.4. Start of β-Sheets.

An alternating sequence $\mathbb{PNPN}$ ... of polar and nonpolar amino acids may allow β-sheets to expose side chains of the same kind at each of their two sides, making them amphiphilic. For N_h = 1, we find that the single hidden unit has weights of alternating signs with γ and opposite polarity for $\mathbb{P}$ and $\mathbb{N}$, which would often lead to generating amphiphilic sequences. However, not all β-sheet stretches follow this simple amphiphilic scheme. For N_h = 2, two groups emerge from clustering. The three hidden unit groups emerging for N_h = 3 instead invalidate the analysis based on the aRBM for two reasons. First, many units are considered noise by the clustering algorithm; second, within single RBMs, we find high heterogeneity in the combination of groups. Therefore, we choose N_h = 2 as the optimal number of hidden units leading to the most consistent yet complex aRBM. In support of this choice, note that the most significant increase in the PLL occurs from N_h = 1 to N_h = 2 (Figure 3).

The weights of the two groups preserve the $\mathbb{PN}$ alternation only at the beginning (group 1, Figure 6c) or at the end (group 2, Figure 6d). These will yield a hidden state **h** compatible with the amphiphilic pattern of weights if combined with the proper signs of h_1 and h_2: the probability of amino acids for mode β-S1 (Figure 6e) promotes the $\mathbb{PNPNP}$ alternation, while that for mode β-S2 promotes the $\mathbb{NPNPN}$ pattern. They cover 55% of the cases.

However, the remaining 45% of combinations of hidden states suppress the $\mathbb{PN}$ alternation, and β segments $\mathbb{NNNNP}$

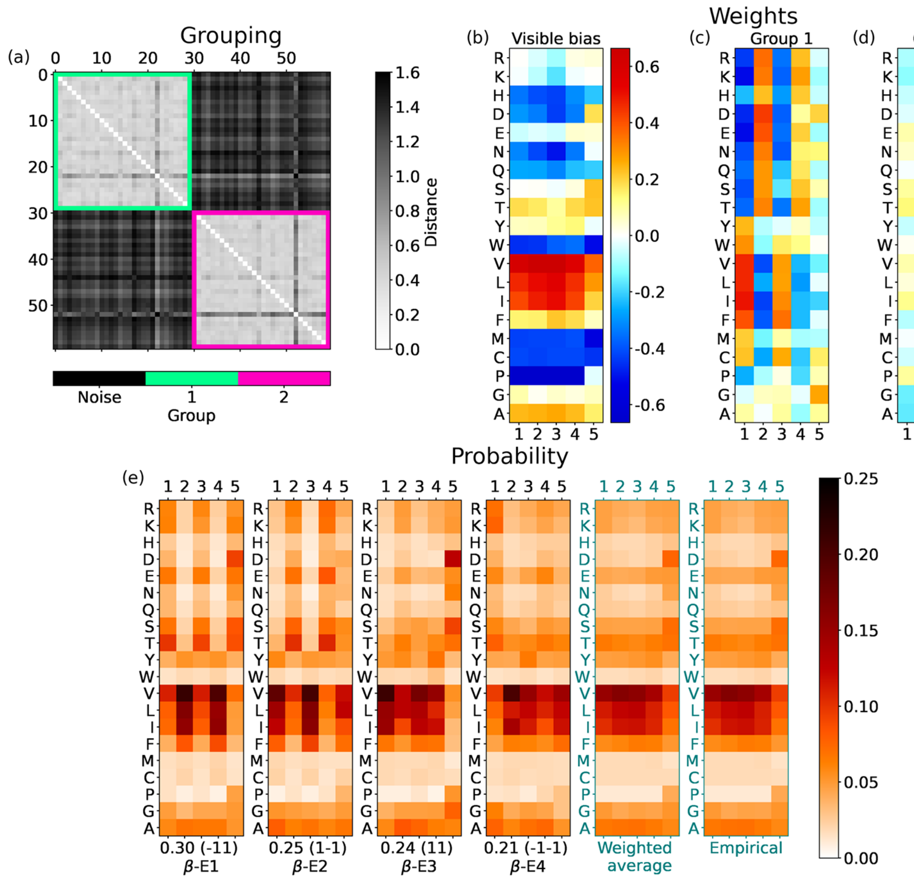


Figure 7. For the end of β-sheets with two hidden units, the same scheme as in Figure 4. Hidden bias: $b_1 = -0.349$ and $b_2 = -0.226$.

(mode β-S3) and ℙℕℕℕℕ (β-S4) are more likely to be generated by RBMs. In particular, β-S4 shows a strong activation of polar amino acids in the first position of β-sheets in comparison to the aliphatic ones, which are instead very favored in the next four positions.

The weighted average of probabilities for β-S1⋯4, as before for α-helices, matches the empirical distributions (last panels of Figure 6e). The RBM-learned decomposition thus splits the start of β-sheets into four modes: the first two modes promote amphiphilic patterns, and the last two modes favor uniform stretches of four ℕ's (mostly I, L, V) capped by a different type of amino acid. This decomposition somewhat joins previous results in which the amphiphilic alternation of β-sheets was seen by different works, with more straightforward statistical tools, either over-represented[57] or under-represented.[58]

The bias a_i at the start of the β-sheets shows a uniform distribution of amino acids at different positions of the chain (see Figure 6b). For instance, aliphatic amino acids show a high bias. However, for a small subset of amino acids, there emerges variability. For example, Arg (R) and Lys (K) have a decreasing bias from the first position in the β strand to the following ones. Perhaps the most interesting behavior is observed for Gly, with a high bias except at the second position ($\gamma = 2$), suggesting that Gly is not likely to appear there.

3.5. End of β-Sheets.

Generally, the analysis of the end of β-sheets retraces the start of β-sheets. Thus, on average, the ensemble of RBMs can capture only patterns of little complexity in β-sheets compared with those of α-helices. We take $N_h = 2$ also for the end of β-sheets, and again we observe two groups similar to those at the start of β-sheets (Figure 7c,d).

Visible biases (Figure 7b) show a uniform distribution of amino acids at different positions close to the ends of β-sheets. However, there is a significant increase in the bias at the last position ($\gamma = 5$) for many small amino acids. Furthermore, many of these are polar (Asp (D), Asn (N), Ser (S), Thr (T)), and there is also Pro (P). In the next section, we will stress that Pro is often positively correlated with polar amino acids. The biases shown in Figure 7b are different from those found at the start of β-sheets (Figure 6b). As a consequence, the probabilities in Figure 7e diverge slightly from those in Figure 6e. In particular, mode β-E3 promotes sequences such as ℕℕℕℕℙ, ℕℕℕℕ(Pro), or ℕℕℕℕ(Gly).

One may notice some similarity between modes β-S in Figure 6e and modes β-E in Figure 7e. This is due to an overlap of 20% of sequences between the two datasets corresponding to β-sheets with five residues. This is not seen for α-helices, which are on average longer, with only 3% having a length of five residues.

3.6. Amino Acid Similarities.

The abundance or absence of a given amino acid in α-helices or β-sheets is primarily encoded in the visible biases a_i. One can check that they correlate with results from standard statistical analysis.[59] However, these biases are not directly related to the polarity or size of amino acids. Hence, they do not provide complete information about the amino acid patterns in secondary structures.

The refined information on amino acid similarities is given by the weights shown in panels (c), (d), and eventually (e) of

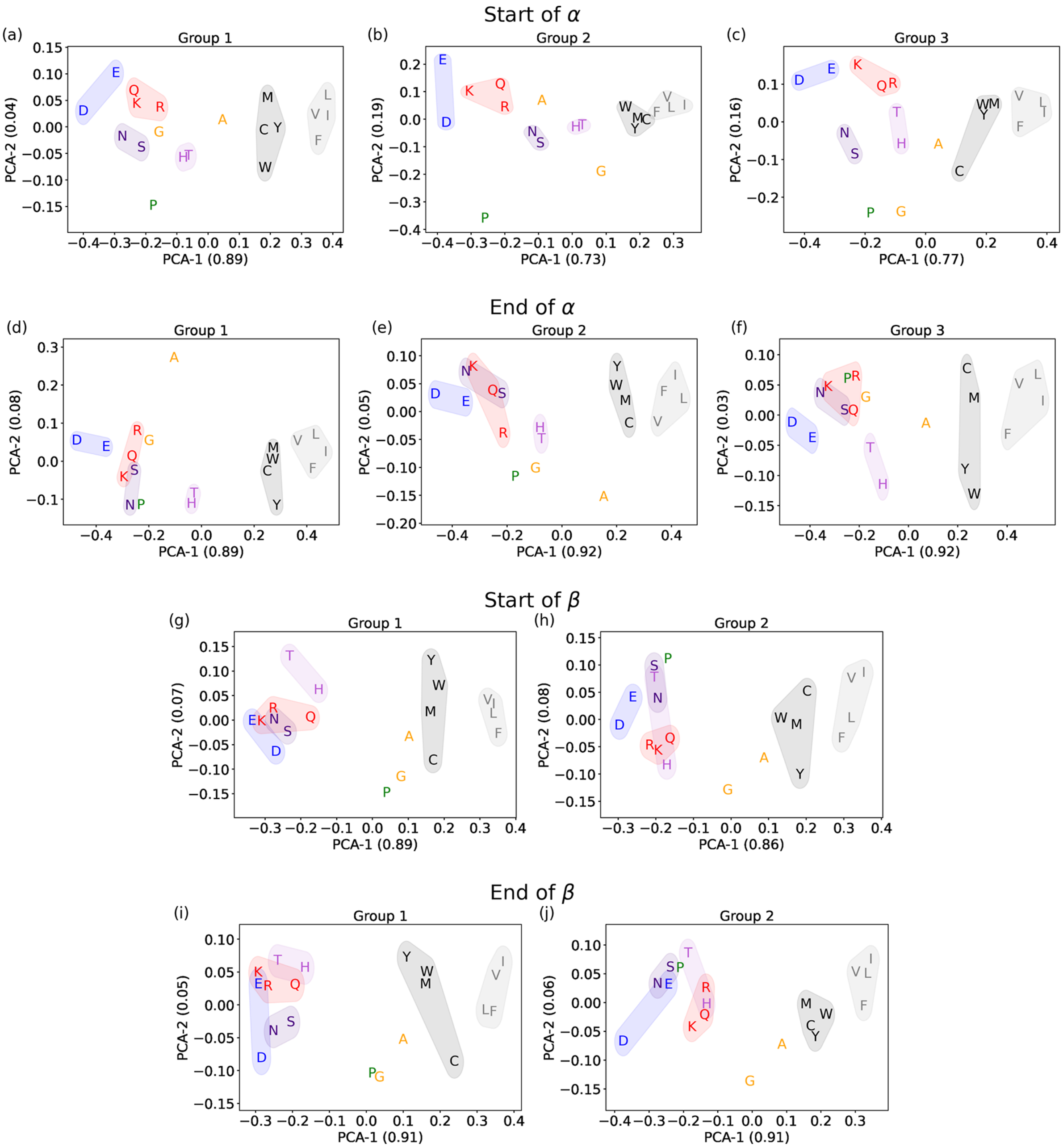


Figure 8. Principal component analysis of amino acid weights. Each panel shows the first two components of the PCA for each amino acid in a hidden-unit group for a given part of the secondary structure. The numbers in the axis labels in the parentheses are the average explained variances. Color-shaded ensembles and single amino acids are discussed in the text. (a−c) Groups at the start of α-helices, shown in Figure 4. (d−f) Groups at the end of α-helices (Figure 5). (g, h) Groups at the start of β-sheets (Figure 6). (i, j) Groups at the end of β-sheets (Figure 7).

Figures 4−7. Each row in a panel shows the weights of a given amino acid in that group. The similarity of amino acids in a group emerges when their weights are interchangeable; i.e., the $\Gamma = 5$ weights appearing in the row of a given amino acid can be swapped with the other ones in a row of an equivalent amino acid without a significant change in the whole set of weights w_{ij} of the corresponding hidden unit j.

In our unsupervised machine learning approach, the salient traits of amino acids' similarity emerge from principal component analysis (PCA). For each hidden unit, we compute the PCA of the $\Gamma = 5$ weights associated with each amino acid. Then, for all groups shown in Figures 4−7, we show these two PCA components in Figure 8 to check amino acid similarities. The number in each axis label represents the average variance *explained* by each PCA component, measuring its relevance. In all cases, the first component of PCA, PCA-1, explains the major part of the variance and is related to the polarity of the amino acids.

Families of interchangeable amino acids emerge, as highlighted in all panels of Figure 8. Let us discuss our interpretation of these plots by collecting similar amino acids in small coherent groups. We define this by looking primarily at their PCA components in α-helices, where there is a clearer subdivision. Our amino acid cataloging was anticipated in Table 1.

Aspartic Acid (D) and Glutamic Acid (E). These negatively charged amino acids are always at the left boundary of the PCA-1 component. Looking at the general arrangement of amino acids in the panels of Figure 8, we interpret this as a signal of a strong hydrophilic tendency that stands out even among charged and polar amino acids. Indeed, in all cases, the Pearson coefficient between PCA-1 and the hydrophobicity is ~0.9. Although Asp and Glu seem mostly similar, for the end of β-sheets, in Figure 8j, we observe that Asp (D) stands away from other polar amino acids. This indicates that Asp has a special role at the end of the β-sheets, which cannot be implemented even by Glu.

Asparagine (N) and Serine (S). The PCA always shows these two small polar amino acids very close to each other. Moreover, they are placed between the pair of Asp and Glu and the central part of the PCA-1 component. This should be related to their lower hydrophilic tendency.

Lysine (K), Arginine (R), and Glutamine (Q). These amino acids have positively charged (K, R) or polar (Q) long side chains. They appear similar, and for them, we can retrace the comments just made for Asn and Ser.

Histidine (H) and Threonine (T). Histidine is a weakly positively charged, large (>150 Da) amino acid, while Thr is a polar, small (<120 Da) amino acid. Thus, it is surprising to find them very well paired in the PCA plots for α-helices, where they sit in a middle region and are not very close to those of other hydrophilic amino acids. Hence, His and Thr display a similar weak tendency to contribute to the amphiphilic pattern in α-helices. In β-sheets, instead, they are not so correlated and are more overlapped with other polar amino acids.

Tyrosine (Y), Tryptophan (W), Methionine (M), and Cysteine (C). These amino acids always have very similar PCA values on the right side of the panels. This quartet comprises a duo of aromatic amino acids (Y, W) and a duo of nonpolar amino acids with sulfur (M, C). In particular, Cys is a small, unique amino acid that can form disulfide bonds. Yet, the PCA correctly places it in the mild hydrophobic region (i.e., with positive but not extreme PCA-1 values).

Valine (V), Leucine (L), Isoleucine (I), and Phenylalanine (F). Three aliphatic amino acids (V, L, and I) and Phe are always equivalently set on the rightmost side of the PCA-1 component. Our analysis with RBMs thus reveals that these four amino acids should be regarded as the strongest hydrophobic amino acids.

Alanine (A). Ala shows neither a clear hydrophobic tendency nor a clear hydrophilic tendency in the PCA plots of α-helices (Figure 8a–f). Nevertheless, we find a peculiar isolation of Ala from the other amino acids in groups 1 and 2 of the end of α-helices (Figure 8d,e), with an opposite sign of the PCA-2 component in the two cases. As discussed above, this is related to the unique role of Ala in helices, in particular at their end, where stretches of five Ala are not rare. However, in β-sheets, Ala shows a mild tendency to cluster with the $\mathbb{N}$ group and thus behave as hydrophobic (Figure 8g–j).

Glycine (G). Even if Gly is a nonpolar amino acid, in α-helices it is mainly found in the region populated by hydrophilic amino acids. However, this is not the case in β-sheets, where Gly is not affiliated with other groups.

Proline (P). Similarly to Gly, Pro is not polar but is often aligned with polar amino acids along PCA-1. However, P displays several extreme values of PCA-2, which isolate it from the other amino acids. The most striking case is in group 2 at the start of α-helices (Figure 8b), which RBMs use to highlight the importance of Pro in this portion of the secondary structure.

Before concluding, we note that our PCA plots are similar to the embeddings learned by much more complex neural networks using Transformers.[12] That analysis showed that the machine catalogs amino acids based on their biological properties.

4. CONCLUSIONS

We introduce and showcase how an ensemble analysis of (unsupervised) machine learning models, based on restricted Boltzmann machines (RBMs) and with an information bottleneck in encoding data correlations, offers a relatively easy reading of precise yet unexpected similarities between amino acids and emphasizes essential features for building secondary structures. Besides recovering a way to promote the frequent amphiphilic design of α-helices and β-sheets, RBMs discover that there are relevant motifs that, to the best of our knowledge, are not known.

The most diverse scenario is at the start of α-helices. RBMs recover the known relative abundance of Pro in their first positions and promote it to the role of a highly relevant feature in addition to amphiphilicity. Moreover, RBMs add information on correlations between Pro and other amino acids, particularly Asp and Glu, which lead to two typical types of helices starting with Pro. Our complete analysis reveals a frequent alignment of Pro with polar amino acids.

At the end of α-helices, there emerges a particular behavior of Ala, which is the distinguishing amino acid between two otherwise similar amphiphilic patterns. This bimodality implies that in nature there is a class of α-helices closed by stretches richer in Ala than in typical helices.

Moreover, our analysis allows refining of the separation between polar and nonpolar amino acids, highlighting intriguing subclasses. The most unexpected is the coupling of His and Thr in α-helices, where they do not contribute to the amphiphilic patterns. Then, for instance, we found the coupling of Phe with the aliphatic amino acids or the alignment of Trp with Tyr, Met, and Cys.

The first component of our PCA (PCA-1) is strongly correlated but does not precisely follow the hydrophobicity ranking reported in the literature. Nevertheless, PCA-1 explains most of the fluctuations of weights in the RBM. Hence, it is crucial to unveil its meaning. We conjecture that PCA-1, the main feature learned by RBMs to reproduce realistic alternations of polarity in secondary structures, expresses a form of *effective hydrophobicity*. In other words, it reveals how much an amino acid, in α-helices and β-sheets, is mainly focused on the role of being either hydrophobic or hydrophilic. For example, even if it is not the most hydrophilic amino acid, Asp most often displays the strongest negative PCA-1 value (and has a special role in closing β-sheets).

One may wonder whether results similar to ours could emerge from a standard approach based, for instance, on two-site correlations from the sequence data, which requires the computation and parallel visualization of many matrices (see SI section S3). This procedure makes recognizing and interpreting some meaningful patterns possible but far from being naturally summarized in a simple set of ranked-by-relevance multisite correlations, as achieved by an analysis of RBM weights with an increasing number of hidden units. Note, moreover, that RBMs can generate sequences, which is not possible with correlation matrices.

To conclude, the RBM is a simple unsupervised machine learning method that retrieves known results and enriches previous knowledge. Moreover, the RBM's architecture is readable and, with some effort, interpretable, yielding nontrivial information that is inaccessible by standard statistical tools. For example, we have provided an interpretation of the RBM weights in our study of amino acid patterns and similarities in secondary structures. However, the richness of the results may allow the reader to notice additional details of the arrangement of amino acids in the secondary structures.

■ ASSOCIATED CONTENT

Data Availability Statement

The code used in this work is available upon reasonable request.

SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.3c00383.

> Details on the implementation and training of RBMs and on the clustering of hidden units (section S1), comparison between DSSP and STRIDE algorithms for α-helices (section S2), and matrices of correlations between amino acids (section S3) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Marco Baiesi − *Department of Physics and Astronomy, University of Padova, 35131 Padua, Italy; INFN, Sezione di Padova, 35131 Padua, Italy;* orcid.org/0000-0002-4513-9191; Email: marco.baiesi@unipd.it

Authors

Anna Braghetto − *Department of Physics and Astronomy, University of Padova, 35131 Padua, Italy; INFN, Sezione di Padova, 35131 Padua, Italy;* orcid.org/0009-0008-7039-3811

Enzo Orlandini − *Department of Physics and Astronomy, University of Padova, 35131 Padua, Italy; INFN, Sezione di Padova, 35131 Padua, Italy;* orcid.org/0000-0003-3680-9488

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.3c00383

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Aurelien Decelle, Emanuele Locatelli, Lorenzo Rosset, and Antonio Trovato for the useful discussions and feedback on the text. M.B. and E.O. were supported by research grants BAIE_BIRD2021_01 and ORLA_BIRD2020_01 of the University of Padova.

■ REFERENCES

(1) Tubiana, J.; Cocco, S.; Monasson, R. Learning protein constitutive motifs from sequence data. *eLife* **2019**, *8*, No. e39397.

(2) Wang, Y.; Zeng, J. Predicting drug-target interactions using restricted Boltzmann machines. *Bioinformatics* **2013**, *29*, i126−i134.

(3) Malbranke, C.; Bikard, D.; Cocco, S.; Monasson, R.; Tubiana, J. Machine learning for evolutionary-based and physics-inspired protein design: Current and future synergies. *Curr. Opin. Struct. Biol.* **2023**, *80*, 102571.

(4) Karimi, M.; Wu, D.; Wang, Z.; Shen, Y. DeepAffinity: interpretable deep learning of compound−protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* **2019**, *35*, 3329−3338.

(5) Karimi, M.; Wu, D.; Wang, Z.; Shen, Y. Explainable deep relational networks for predicting compound−protein affinities and contacts. *J. Chem. Inf. Model.* **2021**, *61*, 46−66.

(6) Rodríguez-Pérez, R.; Bajorath, J. Feature importance correlation from machine learning indicates functional relationships between proteins and similar compound binding characteristics. *Sci. Rep.* **2021**, *11*, 14245.

(7) Rube, H. T.; Rastogi, C.; Feng, S.; Kribelbauer, J. F.; Li, A.; Becerra, B.; Melo, L. A.; Do, B. V.; Li, X.; Adam, H. H.; et al. Prediction of protein−ligand binding affinity from sequencing data with interpretable machine learning. *Nat. Biotechnol.* **2022**, *40*, 1520−1527.

(8) Cai, H.; Vernon, R. M.; Forman-Kay, J. D. An interpretable machine-learning algorithm to predict disordered protein phase separation based on biophysical interactions. *Biomolecules* **2022**, *12*, 1131.

(9) Ali, S. D.; Tayara, H.; Chong, K. T. Interpretable machine learning identification of arginine methylation sites. *Comput. Biol. Med.* **2022**, *147*, 105767.

(10) Tubiana, J.; Schneidman-Duhovny, D.; Wolfson, H. J. ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nat. Methods* **2022**, *19*, 730−739.

(11) Mataeimoghadam, F.; Newton, M.; Dehzangi, A.; Karim, A.; Jayaram, B.; Ranganathan, S.; Sattar, A. Enhancing protein backbone angle prediction by using simpler models of deep neural networks. *Sci. Rep.* **2020**, *10*, 19430.

(12) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R.; et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, No. e2016239118.

(13) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W.; Bridgland, A.; et al. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706−710.

(14) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.

(15) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871−876.

(16) Humphreys, I. R.; Pei, J.; Baek, M.; Krishnakumar, A.; Anishchenko, I.; Ovchinnikov, S.; Zhang, J.; Ness, T. J.; Banjade, S.; Bagde, S. R.; et al. Computed structures of core eukaryotic protein complexes. *Science* **2021**, *374*, No. eabm4805.

(17) Drake, Z. C.; Seffernick, J. T.; Lindert, S. Protein complex prediction using Rosetta, AlphaFold, and mass spectrometry covalent labeling. *Nat. Commun.* **2022**, *13*, 7846.

(18) Torrisi, M.; Pollastri, G.; Le, Q. Deep learning methods in protein structure prediction. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1301−1310.

(19) Iuchi, H.; Matsutani, T.; Yamada, K.; Iwano, N.; Sumi, S.; Hosoda, S.; Zhao, S.; Fukunaga, T.; Hamada, M. Representation learning applications in biological sequence analysis. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 3198−3208.

(20) Wu, L.; Huang, Y.; Lin, H.; Li, S. Z. A Survey on Protein Representation Learning: Retrospect and Prospect. *arXiv (Computer Science.Machine Learning)*, December 31, 2022, 2301.00813, ver. 1. https://arxiv.org/abs/2301.00813 (accessed 2023-07-26).

(21) Hermosilla, P.; Schäfer, M.; Lang, M.; Fackelmann, G.; Vázquez, P. P.; Kozlíková, B.; Krone, M.; Ritschel, T.; Ropinski, T. Intrinsic-extrinsic convolution and pooling for learning on 3D protein structures. *arXiv (Computer Science.Machine Learning)*, April 19, 2021,

2007.06252, ver. 2. https://arxiv.org/abs/2007.06252 (accessed 2023-07-26).

(22) Wang, Y.; You, Z.-H.; Yang, S.; Li, X.; Jiang, T.-H.; Zhou, X. A high efficient biological language model for predicting protein−protein interactions. *Cells* **2019**, *8*, 122.

(23) Ding, X.; Zou, Z.; Brooks, C. L., III. Deciphering protein evolution and fitness landscapes with latent space models. *Nat. Commun.* **2019**, *10*, 5644.

(24) Weigt, M.; White, R. A.; Szurmant, H.; Hoch, J. A.; Hwa, T. Identification of direct residue contacts in protein−protein interaction by message passing. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 67−72.

(25) Gligorijević, V.; Renfrew, P. D.; Kosciolek, T.; Leman, J. K.; Berenberg, D.; Vatanen, T.; Chandler, C.; Taylor, B. C.; Fisk, I. M.; Vlamakis, H.; et al. Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **2021**, *12*, 3168.

(26) Molnar, C. *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable*, 2020.

(27) Kamath, U.; Liu, J. *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*; Springer, 2021.

(28) Xiong, H.; Buckwalter, B. L.; Shieh, H.-M.; Hecht, M. H. Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92*, 6349−6353.

(29) Kamtekar, S.; Schiffer, J. M.; Xiong, H.; Babik, J. M.; Hecht, M. H. Protein design by binary patterning of polar and nonpolar amino acids. *Science* **1993**, *262*, 1680−1685.

(30) Smolensky, P. Information processing in dynamical systems: Foundations of harmony theory. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*; MIT Press, 1986; Vol. *1*, Chapter 6, pp 194−281.

(31) Hinton, G. E.; Sejnowski, T. J. Learning and relearning in Boltzmann machines. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*; MIT Press, 1986; Vol. *1*, Chapter 7, pp 282−317.

(32) Hinton, G. E. A Practical Guide to Training Restricted Boltzmann Machines. In *Neural Networks: Tricks of the Trade*; Montavon, G., Orr, G. B., Müller, K.-R., Eds.; Lecture Notes in Computer Science, Vol. *7700*; Springer, 2012; pp 599−619.

(33) Tubiana, J.; Monasson, R. Emergence of compositional representations in restricted Boltzmann machines. *Phys. Rev. Lett.* **2017**, *118*, 138301.

(34) Decelle, A.; Fissore, G.; Furtlehner, C. Thermodynamics of restricted Boltzmann machines and related learning dynamics. *J. Stat. Phys.* **2018**, *172*, 1576−1608.

(35) Roussel, C.; Cocco, S.; Monasson, R. Barriers and dynamical paths in alternating Gibbs sampling of restricted Boltzmann machines. *Phys. Rev. E* **2021**, *104*, 034109.

(36) Fernandez-de-Cossio-Diaz, J.; Cocco, S.; Monasson, R. Disentangling representations in Restricted Boltzmann Machines without adversaries. *arXiv (Computer Science.Machine Learning)*, March 8, 2023, 2206.11600, ver. 4. https://arxiv.org/abs/2206.11600 (accessed 2023-07-26).

(37) Decelle, A.; Furtlehner, C.; Seoane, B. Equilibrium and non-equilibrium regimes in the learning of restricted Boltzmann machines. *J. Stat. Mech.: Theory Exp.* **2022**, *2022*, 114009.

(38) Decelle, A.; Rosset, L.; Seoane, B. Unsupervised hierarchical clustering using the learning dynamics of RBMs. *arXiv (Computer Science.Machine Learning)*, June 9, 2023, 2302.01851, ver. 3. https://arxiv.org/abs/2302.01851 (accessed 2023-07-26).

(39) Herzog, B.; Casier, B.; Lebègue, S.; Rocca, D. Solving the Schroödinger Equation in the Configuration Space with Generative Machine Learning. *J. Chem. Theory Comput.* **2023**, *19*, 2484−2490.

(40) Iyengar, S. S.; Kais, S. Analogy between Boltzmann machines and Feynman path integrals. *J. Chem. Theory Comput.* **2023**, *19*, 2446−2454.

(41) Si, Z.; Yu, H.; Ma, Z. Learning deep features for DNA methylation data analysis. *IEEE Access* **2016**, *4*, 2732−2737.

(42) Di Gioacchino, A.; Procyk, J.; Molari, M.; Schreck, J. S.; Zhou, Y.; Liu, Y.; Monasson, R.; Cocco, S.; Šulc, P. Generative and interpretable machine learning for aptamer design and analysis of in vitro sequence selection. *PLoS Comput. Biol.* **2022**, *18*, e1010561.

(43) Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P.; et al. *Molecular Biology of the Cell*; Garland Science USA: New York, 2002; Vol. *4*.

(44) Stephenson, J. D.; Freeland, S. J. Unearthing the root of amino acid similarity. *J. Mol. Evol.* **2013**, *77*, 159−169.

(45) Meiler, J.; Müller, M.; Zeidler, A.; Schmäschke, F. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Mol. Model. Annu.* **2001**, *7*, 360−369.

(46) *CATH S40 Database*, 2019. http://download.cathdb.info/cath/releases/all-releases/v4_3_0/non-redundant-data-sets/.

(47) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.

(48) Joosten, R. P.; Te Beek, T. A.; Krieger, E.; Hekkelman, M. L.; Hooft, R. W.; Schneider, R.; Sander, C.; Vriend, G. A series of PDB related databases for everyday needs. *Nucleic Acids Res.* **2011**, *39*, D411−D419.

(49) Frishman, D.; Argos, P. Knowledge-based protein secondary structure assignment. *Proteins: Struct., Funct., Bioinf.* **1995**, *23*, 566−579.

(50) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press, 2016.

(51) Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural computation* **2002**, *14*, 1771−1800.

(52) Mehta, P.; Bukov, M.; Wang, C.-H.; Day, A. G.; Richardson, C.; Fisher, C. K.; Schwab, D. J. A high-bias, low-variance introduction to machine learning for physicists. *Phys. Rep.* **2019**, *810*, 1−124.

(53) Tieleman, T. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th International Conference on Machine Learning*; Association for Computing Machinery, 2008; 1064−1071.

(54) Kim, M. K.; Kang, Y. K. Positional preference of proline in α-helices. *Protein Sci.* **1999**, *8*, 1492−1499.

(55) Zhuang, Y.; Bureau, H. R.; Lopez, C.; Bucher, R.; Quirk, S.; Hernandez, R. Energetics and structure of alanine-rich α-helices via adaptive steered molecular dynamics. *Biophys. J.* **2021**, *120*, 2009−2018.

(56) Mier, P.; Elena-Real, C. A.; Cortés, J.; Bernadó, P.; Andrade-Navarro, M. A. The sequence context in poly-alanine regions: structure, function and conservation. *Bioinformatics* **2022**, *38*, 4851−4858.

(57) Mandel-Gutfreund, Y.; Gregoret, L. M. On the significance of alternating patterns of polar and non-polar residues in beta-strands. *J. Mol. Biol.* **2002**, *323*, 453−461.

(58) Broome, B. M.; Hecht, M. H. Nature disfavors sequences of alternating polar and non-polar amino acids: implications for amyloidogenesis. *J. Mol. Biol.* **2000**, *296*, 961−968.

(59) Malkov, S. N.; Živković, M. V.; Beljanski, M. V.; Hall, M. B.; Zarić, S. D. A reexamination of the propensities of amino acids towards a particular secondary structure: classification of amino acids based on their chemical structure. *J. Mol. Model.* **2008**, *14*, 769−775.