**Supplementary Information**

# WARIO: Weighted families of contact maps to characterize conformational ensembles of (highly-)flexible proteins

Javier González-Delgado[1,2], Pau Bernadó[3], Pierre Neuvial[2] and Juan Cortés[1]

[1] *LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France.*

[2] *Institut de Mathématiques de Toulouse, Université de Toulouse, CNRS, Toulouse, France.*

[3] *Centre de Biologie Structurale, Université de Montpellier, INSERM, CNRS, Montpellier, France.*

The Supplementary Information is organized as follows:

- Section S1 details the methodology defining the complete clustering pipeline.
  - Section S1.1 starts by relaxing the threshold-based contact definition for Euclidean distances through the introduction of sequence-dependent contact intervals.
  - The role of relative orientation in short-range contacts is addressed in Section S1.2. This section also explains how the orientation can be combined with the Euclidean distance to define a metric accounting for residue-residue interactions. The precise form of this combination is determined through empirical analysis of the interactions between amino acids. This is presented in Section S1.3.
  - Then, Section S1.4 defines the contact function for amino acid pairs as a decreasing function of their relative pose distance, whose form is once again empirically calibrated.
  - Section S1.5 presents the dimensionality reduction and clustering algorithms that complete the pipeline.
- In Section S2, we demonstrate that refining the contact definition by removing arbitrary thresholds and incorporating relative orientation significantly improves the performance of the method.
- Section S3 includes supplementary figures complementing the comparative analysis of WARIO with other clustering approaches.
- Finally, Section S4 presents the complete characterizations of the conformational ensembles for the three proteins studied in this work.

## S1 Methods

This section details the main components of WARIO's methodology.
To calibrate the *contact function*, we made use of a set of 15177 experimentally-determined high-resolution structures of protein domains extracted from the SCOPe 2.07 release [1]. Throughout this section, this set will be referred to as the *structural database*.

### S1.1 Contact intervals for the Euclidean distance

Contact between amino acid residues is usually defined by setting universal thresholds to the Euclidean distance between their positions [2]. By universal, we mean that these thresholds are fixed independently of the amino acid identities or their distance along the sequence. However, when looking at how

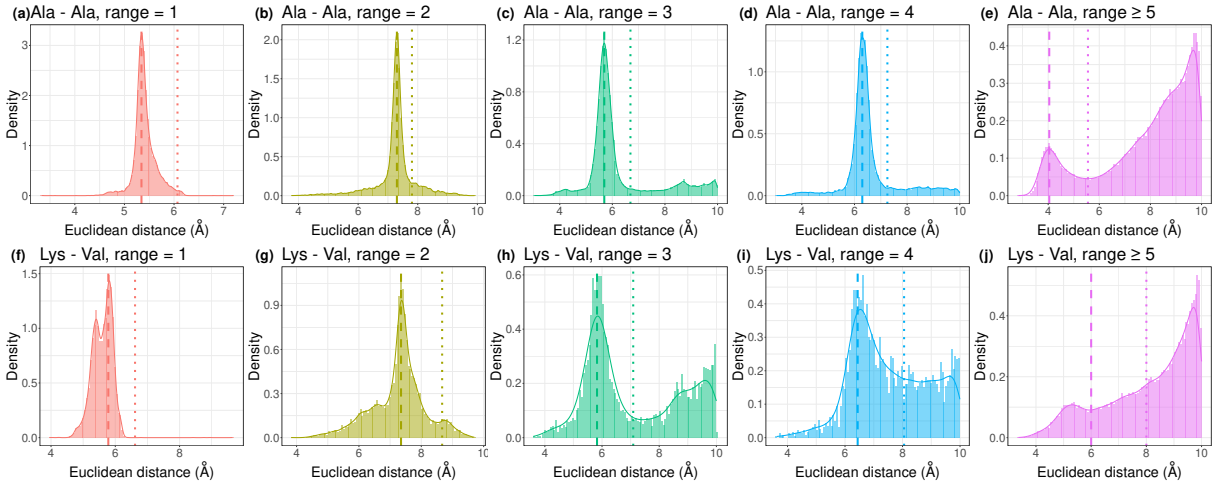



Figure S1: Empirical distribution of the Euclidean distance between (a-e) Ala-Ala and (f-j) Lys-Val residues in the structural database, stratified by range groups. Distributions are depicted through a histogram and a kernel density estimate. Vertical dashed and dotted lines indicate the lower and upper limits of the contact intervals for the Euclidean distance respectively.

contact distances distribute in nature, we directly observe that residue-residue interactions concentrate around distance values that change according to these parameters. To account for this, we computed the Euclidean distance between every pair of residues in the structural database, and represented their empirical distribution stratifying their identities and range (the distance between both residues along the sequence, in number of residues). Figure S1 presents the encountered distributions truncated to the interval $[0\mathring{A}, 10\mathring{A}]$ for two pairs of residues at ranges in $\{1, 2, 3, 4\}$ and $[5, \infty)$.

Figure S1 illustrates the fact that the residue-residue Euclidean distance truncated to $[0\mathring{A}, 10\mathring{A}]$ is not identically distributed across amino acid identities and ranges. Distance values concentrate around sequence-dependent maxima with sequence-dependent variance. Therefore, contact descriptors computed from Euclidean distance should take this information into account and avoid universal thresholds that contradict the empirical behavior. The sequence-specific distance distributions presented in Figure S1 allow us to relax the threshold-based definition of contact for Euclidean distances. Let $A_i$, $A_j$ denote a pair of amino acid types and $S_{ij} = 1, 2, \ldots$ denote a sequence range. Let $f_{ij}^{\mathbb{R}^3}$ denote the density function of the Euclidean distance distribution for $A_i$-$A_j$ pairs at range $S_{ij}$ estimated from the structural database and truncated to the interval $[0\mathring{A}, 10\mathring{A}]$. The *Euclidean contact interval* for $A_i$-$A_j$ pairs at range $S_{ij}$ is defined as the real interval

$$C_{ij}^{\mathbb{R}^3}(A_i, A_j, S_{ij}) = C_{ij}^{\mathbb{R}^3} = [\Delta_{a;i,j}^{\mathbb{R}^3}, \Delta_{b;i,j}^{\mathbb{R}^3}], \tag{S1}$$

where $\Delta_{a;i,j}^{\mathbb{R}^3}$ is the abscissa smaller than $8\mathring{A}$ presenting the highest maximum of $f_{ij}^{\mathbb{R}^3}$ and $\Delta_{b;i,j}^{\mathbb{R}^3}$ is the closest abscissa from the right to $\Delta_{a;i,j}^{\mathbb{R}^3}$ presenting a minimum of $f_{ij}^{\mathbb{R}^3}$. Both limits are depicted in Figure S1 with dashed and dotted lines respectively. For low maximum prominences (as in Figure S1(j)), the Euclidean contact interval is set to $[6\mathring{A}, 8\mathring{A}]$ by default. Note that, as distance distributions are not markedly different when varying $S_{ij} \geq 5$, we are setting $C_{ij}^{\mathbb{R}^3}(A_i, A_j, S) = C_{ij}^{\mathbb{R}^3}(A_i, A_j, S')$ for every $S, S' \geq 5$. The complete list of contact intervals and the counterparts of Figure S1 for every amino acid pair and range class are available at https://gitlab.laas.fr/moma/WARIO.

The intervals (S1) allow a continuous description of residue-residue interactions by removing binary contact classifications. The upper limit of (S1) represents the distance value at which the interaction probability starts to be substantial, and continuously increases until reaching the lower limit of (S1),

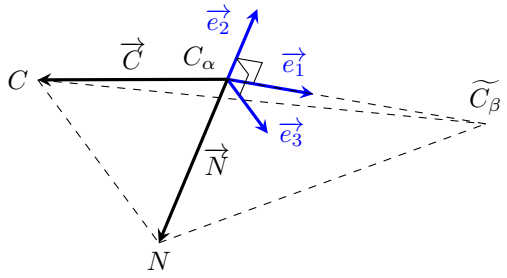



Figure S2: Figure S1 in [8]. The three vectors $\{\overrightarrow{e_1}, \overrightarrow{e_2}, \overrightarrow{e_3}\}$ defining the reference frame, built from the virtual atom $\widetilde{C_\beta}$ and vectors $\overrightarrow{C}$ and $\overrightarrow{N}$.

beyond which interaction occurs with high probability. Replacing thresholds by intervals is the key idea to define continuous functions accounting for a contact strength that increases smoothly as the interaction probability starts to be significant. Their explicit applicability in this work is detailed in the following sections.

### S1.2 Distance to ideal orientations

Relative orientation plays a determinant role in residue-residue interactions [3–6]. This idea was already incorporated in RING [5, 7], where contact thresholds were defined by integrating the values of backbone angles mediating multiple types of interactions. Here, we propose to capture this effect through a meaningful representation of the spatial pose of each amino acid. This can be achieved by defining a residue-specific reference frame at each $C_\beta$ atom ($C_\alpha$ for glycines) as it was done in [8]. The detailed construction of the reference system is explained in [8, Section S1.1]. An outline of its definition is presented here. To encode the angular configuration of the backbone at the residue level, we first define a virtual atom $\widetilde{C_\beta}$, which exists also for glycines. The position of $\widetilde{C_\beta}$ is an estimate of the position of the true $C_\beta$ when it exists, but it is defined for every residue using only the coordinates of the $C_\alpha$, $C$ and $N$ atoms. We denote as $\overrightarrow{C}$ and $\overrightarrow{N}$ the vectors going from $C_\alpha$ to $C$ and $N$ atoms, respectively, and we define $\overrightarrow{CN} = \overrightarrow{N} - \overrightarrow{C}$. The residue-specific reference frame is defined as follows:

$$\begin{cases} \overrightarrow{e_1} = \overrightarrow{\widetilde{C_\beta}}/\|\overrightarrow{\widetilde{C_\beta}}\| \\ \overrightarrow{e_2} = \overrightarrow{CN}/\|\overrightarrow{CN}\| \times \overrightarrow{e_1} \\ \overrightarrow{e_3} = \overrightarrow{e_1} \times \overrightarrow{e_2}, \end{cases} \tag{S2}$$

where $\times$ denotes the cross-product. An illustration of (S2) is presented in Figure S2. Note that the third basis vector $\overrightarrow{e_3}$ is parallel to $\overrightarrow{CN}$ under the hypothesis that the atoms $C$, $N$, $C_\alpha$ and $\widetilde{C_\beta}$ form a perfect tetrahedron. Let $L$ denote the sequence length and $i \in \{1, \dots, L\}$ the position of the $i$-th residue. Denoting as $\mathcal{F}_i = \{\overrightarrow{e_{1,i}}, \overrightarrow{e_{2,i}}, \overrightarrow{e_{3,i}}\}$ the reference system (S2) built on the $i$-th residue, its relative orientation with respect to another residue at position $j \neq i$ will be measured by considering the angles between the first and third basis vectors:

$$\theta_{1;i,j} = \arccos\langle \overrightarrow{e_{1,i}}, \overrightarrow{e_{1,j}} \rangle, \qquad \theta_{3;i,j} = \arccos\langle \overrightarrow{e_{3,i}}, \overrightarrow{e_{3,j}} \rangle, \tag{S3}$$

where $\langle \cdot\,, \cdot \rangle$ denotes the inner product in $\mathbb{R}^3$. The reason why the angles (S3) were chosen to capture the role of orientation in residue-residue interactions is that they present preferred configurations in nature. This was observed in the structural database for short range contacts i.e. for $S_{ij} = |i - j| < 5$.

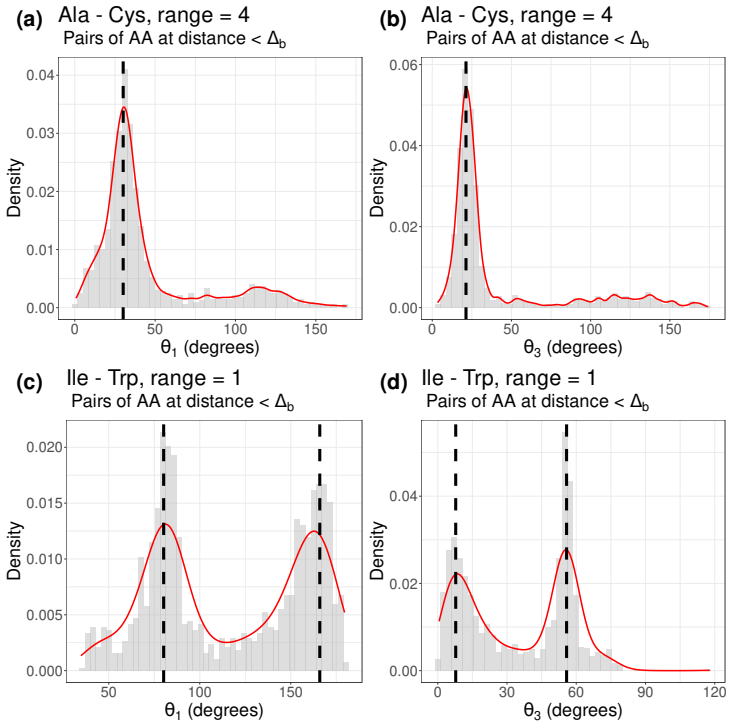



Figure S3: Empirical distribution of angles (S3) computed from the empirical database for pairs of residues at Euclidean distance smaller than $\Delta_b^{\mathbb{R}^3}(A_i, A_j, S_{ij})$, stratified by amino acid types and range. In (a-b), distributions for Ala-Cys pairs at range 4 and, in (c-d), distributions for Ile-Trp pairs at range 1. The non-negligible maxima of the kernel density estimates (red curves) are marked with a dashed black line.

Figure S3 depicts an example of the empirical distribution of (S3) when the Euclidean distance between both amino acids has crossed the upper limit of the contact interval (S1) i.e. it is smaller than $\Delta_{b;i,j}^{\mathbb{R}^3}$. Indeed, residues beyond the upper limit present preferred relative orientations that are specific to their identities and range. These preferred orientations might not be unique, as in Figure S3(c-d) and represent the contact poses with highest probability in nature. For each pair of amino acid residue types and range class, we took up to three maxima from the density estimates of the empirical distributions of (S3). The maximum with the highest density value was always kept, and the subsequent maxima were kept if their prominence with respect to the first maximum was not negligible. We refer to these maxima as the *ideal orientations* for $A_i$-$A_j$ pairs at range $S_{ij}$, and we denote them as

$$\theta_{1;i,j}^* = \theta_{1;i,j}^*(A_i, A_j, S_{ij}) \quad \text{and} \quad \theta_{3;i,j}^* = \theta_{3;i,j}^*(A_i, A_j, S_{ij}) \tag{S4}$$

for the angles between the first and third basis vectors respectively. Note that (S4) are non-empty subsets of $[0°, 180°]$ containing up to three values. The complete list of $\theta_{1;i,j}^*$ and $\theta_{3;i,j}^*$ sets and their corresponding counterparts of Figure S3 are available at https://gitlab.laas.fr/moma/WARIO.

Following the fact that the angles (S3) concentrate around a set of sequence-specific ideal orientations when both amino acid residues interact, it is possible to define how close to the ideal contact setting is the relative orientation of a pair of residues. For two residues at positions $i \neq j$ in the sequence, this is

done by considering the *distance between the pair* $\{\mathcal{F}_i, \mathcal{F}_j\}$ *and its ideal orientation*:

$$d^2_{\theta^*}(\{\mathcal{F}_i, \mathcal{F}_j\}) = \frac{1}{4}\, h\left(\min_{\theta\in\theta^*_{1;i,j}} |\theta_{1;i,j} - \theta|\right)^2 + \frac{1}{4}\, h\left(\min_{\theta\in\theta^*_{3;i,j}} |\theta_{3;i,j} - \theta|\right)^2, \tag{S5}$$

where $h$ is a monotonic function on $[0°, 180°]$ defined by $h(x) = \sin(x)$ if $x \leq 90°$ and $h(x) = 1 - \cos(x)$ otherwise. Note that the quantity $d_{\theta^*}(\{\mathcal{F}_i, \mathcal{F}_j\})$ in (S5) takes values in $[0, 1]$, with $d_{\theta^*}(\{\mathcal{F}_i, \mathcal{F}_j\}) = 0$ being a perfect match to the ideal orientation and $d_{\theta^*}(\{\mathcal{F}_i, \mathcal{F}_j\}) = 1$ the strongest disagreement with such setting. Remark also that we have omitted the explicit dependence of (S5) on $A_i$, $A_j$ and $S_{ij}$ to lighten notation. As we mentioned before, preferred orientations were only found when $S_{ij} = |i - j| < 5$. We refer to this setting as short-range and to the case $S_{ij} \geq 5$ as long-range. Consequently, the relative orientation of the residue pair will only be considered for short-range interactions. In that case, we need to find a suitable strategy to combine distance and orientation information to correctly account for contact. This is addressed in the following section.

### S1.3 Relative pose distance

We have sought to define an appropriate combination of Euclidean and orientation distances that correctly acts as proxy for the interaction between residue pairs. Let $i \neq j$ denote two sequence positions and $\mathcal{F}_i$, $\mathcal{F}_j$ the $i$-th and $j$-th reference frame defined in (S2). We denote by $d_{\mathbb{R}^3}(\mathcal{F}_i, \mathcal{F}_j)$ the Euclidean distance between the positions of both residues. We propose to combine $d_{\mathbb{R}^3}(\mathcal{F}_i, \mathcal{F}_j)$ with (S5) as

$$(1 - \omega_{\theta^*})^2\, d^2_{\mathbb{R}^3}(\mathcal{F}_i, \mathcal{F}_j) + \omega^2_{\theta^*}\, d^2_{\theta^*}(\{\mathcal{F}_i, \mathcal{F}_j\}), \tag{S6}$$

where the weight $\omega_{\theta^*} \in [0, 1]$ governs the distance-orientation balance. Of course, the main problem here is the suitable choice of $\omega_{\theta^*}$. This should be done by considering the following guidelines:

- $(i)$ relative orientation must only be considered for short-range interactions,
- $(ii)$ relative orientation must only be considered when both residues are close in Euclidean distance, i.e. closer than the upper limit of their Euclidean contact interval (S1),
- $(iii)$ relative orientation must significantly enhance the contact strength if it is close to the ideal setting, and remain ineffective otherwise.

The first conclusion that can be extracted is that for $\omega_{\theta^*}$ to satisfy $(i - iii)$ it must be a function of the pair of frames, the amino acid residue types and the sequence range $\omega_{\theta^*} = \omega_{\theta^*}(\mathcal{F}_i, \mathcal{F}_j, A_i, A_j, S_{ij})$. To lighten notation, we will omit the explicit dependence on range and residue identities and write only $\omega_{\theta^*} = \omega_{\theta^*}(\mathcal{F}_i, \mathcal{F}_j)$. The first point $(i)$ can be easily guaranteed by asking $\omega_{\theta^*} = 0$ if $S_{ij} \geq 5$. For long-range interactions, contact will be exclusively encoded by the Euclidean distance between both residues. Ensuring $(ii)$ remains to ask $\omega_{\theta^*}$ to be a decreasing function of $d^2_{\mathbb{R}^3}(\mathcal{F}_i, \mathcal{F}_j)$, whose smooth decay concentrates in the Euclidean contact interval (S1). Finally, satisfying $(iii)$ demands that $\omega_{\theta^*}$ is also decreasing with $d_{\theta^*}(\{\mathcal{F}_i, \mathcal{F}_j\})$. Note that the word *significantly* has been added in $(iii)$. In other words, $\omega_{\theta^*}$ needs to balance distance and orientation in a way that they are comparable beyond the Euclidean contact interval when orientation plays a non-negligible role. This can be ensured if the following relation holds:

$$(1 - \omega_{\theta^*})^2\, d^2_{\mathbb{R}^3}(\mathcal{F}_i, \mathcal{F}_j) \sim \omega^2_{\theta^*}\, d^2_{\theta^*}(\{\mathcal{F}_i, \mathcal{F}_j\}) \quad \text{for all } d_{\mathbb{R}^3}(\mathcal{F}_i, \mathcal{F}_j) \leq \Delta^{\mathbb{R}^3}_{a;i,j}, \tag{S7}$$

where $\Delta^{\mathbb{R}^3}_{a;i,j}$ is the lower limit of the Euclidean contact interval for $A_i$-$A_j$ pairs at range $S_{ij}$, defined in (S1). All these conditions are verified if the following functional form is chosen to define $\omega_{\theta^*}$:

$$\omega_{\theta^*}(\mathcal{F}_i, \mathcal{F}_j) = \begin{cases} 1 - \tanh\left[4\left(d^2_{\theta^*}\left(\{\mathcal{F}_i, \mathcal{F}_j\}\right) + g_{ij}\left(d^2_{\mathbb{R}^3}\left(\{\mathcal{F}_i, \mathcal{F}_j\}\right)\right)\right)^2\right] & \text{if} \quad S_{ij} < 5, \\ 0 & \text{otherwise,} \end{cases} \tag{S8}$$

where $g_{ij}$ is a function defined by

$$g_{ij}(x) = \frac{1}{2}\left(\frac{x}{\Delta^{\mathbb{R}^3}_{b;i,j}}\right)^{\frac{d^{\mathbb{R}^3}_{ij}}{2}} \quad \text{for all } x \geq 0 \quad \text{and} \quad d^{\mathbb{R}^3}_{ij} = \frac{\log\left(\text{argtanh}\left(1/\Delta^{\mathbb{R}^3}_{a;i,j}\right)\right)}{\log\left(\text{argtanh}\left(\Delta^{\mathbb{R}^3}_{a;i,j}/\Delta^{\mathbb{R}^3}_{b;i,j}\right)\right)}, \tag{S9}$$

where $\Delta^{\mathbb{R}^3}_{a;i,j}$ (resp. $\Delta^{\mathbb{R}^3}_{b;i,j}$) is the lower (resp. upper) limit of the Euclidean contact interval for $A_i$-$A_j$ pairs at range $S_{ij}$, defined in (S1). With this, it is possible to define the *relative pose distance* between the pair of residues $A_i$-$A_j$ with frames $\mathcal{F}_i$, $\mathcal{F}_j$ at range $S_{ij}$ in the sequence as the function

$$d_{\text{RP}}\left(\{\mathcal{F}_i, \mathcal{F}_j\}\right) = \left(1 - \omega_{\theta^*}\left(\{\mathcal{F}_i, \mathcal{F}_j\}\right)\right)^2 d^2_{\mathbb{R}^3}(\mathcal{F}_i, \mathcal{F}_j) + \omega^2_{\theta^*}\left(\{\mathcal{F}_i, \mathcal{F}_j\}\right) d^2_{\theta^*}\left(\{\mathcal{F}_i, \mathcal{F}_j\}\right), \tag{S10}$$

where we have omitted the dependence on reside types and range for simplicity and $\omega^2_{\theta^*}\left(\{\mathcal{F}_i, \mathcal{F}_j\}\right)$ is defined in (S8). A clear visualization of the orientation weight (S8) and the relative pose distance (S10) is presented in Figure S4 for Ala-Ala pairs at range 3. The curves in Figure S4 show that definition (S8) satisfies conditions $(i - iii)$. Note first that for Euclidean distances greater than the upper limit of the Euclidean contact interval, orientation is not considered to describe interaction. Its contribution smoothly increases when crossing the Euclidean contact interval from right to left, becoming comparable to the one of the Euclidean distance after crossing the lower limit. The rise of $\omega_{\theta^*}$ is stronger when the relative orientation of $\{\mathcal{F}_i, \mathcal{F}_j\}$ gets closer to its ideal setting, and weaker otherwise. Indeed, orientation has no effect in the worst scenario $d_{\theta^*}(\{\mathcal{F}_i, \mathcal{F}_j\}) = 0$. In other words, the role of orientation is to enhance the contact strength defined by the Euclidean distance when it is close to the ideal setting. It is important to remark that the dependence of $\omega_{\theta^*}$ on Euclidean distance and orientation occurs smoothly in all directions. This is possible thanks to the definition of Euclidean contact intervals (S1), which allows to concentrate the smooth variation of $\omega_{\theta^*}$ within a sequence-dependent range of values extracted in agreement to the observed empirical behavior.

### S1.4 Contact function definition

This section is devoted to define contact between amino acids as a continuous function taking values in $[0, 1]$ and correctly acting as an indicator of their interaction strength. In other words, contact will be defined as a decreasing function of the relative pose distance:

$$\omega^C_{ij}(\{\mathcal{F}_i, \mathcal{F}_j\}) = t_{ij}\left(d_{\text{RP}}(\{\mathcal{F}_i, \mathcal{F}_j\})\right) \quad \text{with } t_{ij} : [0, \infty) \longrightarrow [0, 1] \quad \text{decreasing.} \tag{S11}$$

The contact function $\omega^C_{ij}$ will take values close to 1 (resp. 0) when the relative pose distance between residues at positions $i \neq j$ is close to (resp. far from) 0. Once again, we will ask the contact function to decrease smoothly with $d_{int}$ and to concentrate its decay inside an empirically determined interval. To calibrate its functional form, we proceeded analogously to the previous sections, and started by computing the empirical distribution of the relative pose distance (S10) for every pair of amino acids at ranges in

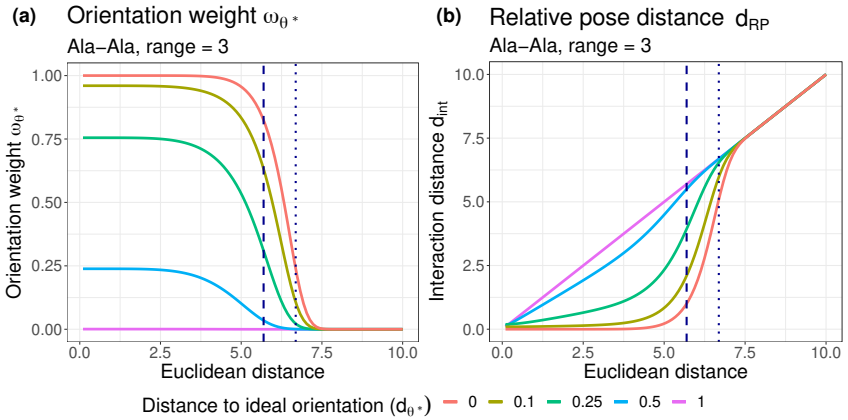



Figure S4: For Ala-Ala pairs at range 3, (a) the weight function (S8) and (b) the relative pose distance (S10). Both quantities are depicted as a function of the Euclidean distance between residue positions and stratified by distance to the ideal orientation. Dashed and dotted vertical lines indicate respectively the lower and upper limit of the Euclidean contact interval.

$\{1, 2, 3, 4\}$ extracted from the structural database. The results for Ala-Ala and Lys-Val pairs are presented in Figure S5.

Figure S5 illustrates the effect of incorporating (S5) to the Euclidean distance when accounting for short-range residue-residue interactions. If we compare panels in Figure S5 with their counterparts of Figure S1, we see how the relative pose distance (S10) enhances -by translating their Euclidean distance value to the left- those residue pairs whose relative orientation is close to the ideal one. This translation is very clear for pairs at range 1, for which the uni-modal distributions of Figure S1 become bi-modal in Figure S5, but it also appreciable for ranges higher than one, where the probability mass moves to smaller distance values thanks to the residue pairs with low values of (S5). Note that the shift is more visible for contacts at range 1 due to the high concentration of the distance distribution around its mean. Indeed, distances and orientations between consecutive residues are very physically restricted. For longer ranges, the shift is equally present but less appreciable through Figure S5 due to the higher variance of the distance distributions. To conclude, defining (S10) allows us to filter residue-residue interactions that, besides corresponding to amino acids close in Euclidean distance, present ideal relative orientations. We introduce now the analogous contact interval of (S1) for the relative pose distance (S10). Let $A_i$, $A_j$ denote a pair of amino acid identities and $S_{ij} = 1, 2, \ldots$ denote a sequence range. Let $f_{ij}^{\mathrm{RP}}$ denote the density function of the relative pose distance distribution for $A_i$-$A_j$ pairs at range $S_{ij}$ estimated from the structural database and truncated to the interval $[0\mathring{A}, 10\mathring{A}]$. The *contact interval* for $A_i$-$A_j$ pairs at range $S_{ij}$ is defined as the real interval

$$C_{ij}^{\mathrm{RP}}(A_i, A_j, S_{ij}) = C_{ij}^{\mathrm{RP}} = [\Delta_{a;i,j}^{\mathrm{RP}}, \Delta_{b;i,j}^{\mathrm{RP}}], \tag{S12}$$

where $\Delta_{a;i,j}^{\mathrm{RP}}$ is the smaller abscissa presenting a maximum of $f_{ij}^{\mathrm{RP}}$ and $\Delta_{b;i,j}$ is the closest abscissa from the right to $\Delta_{a;i,j}^{\mathrm{RP}}$ presenting a minimum of $f_{ij}^{\mathrm{RP}}$. As the relative pose distance (S10) corresponds to the Euclidean one for $S_{ij} \geq 5$, we have

$$\Delta_{a;i,j}^{\mathrm{RP}} = \Delta_{a;i,j}^{\mathbb{R}^3}, \quad \Delta_{b;i,j}^{\mathrm{RP}} = \Delta_{b;i,j}^{\mathbb{R}^3} \quad \text{for all } S_{ij} \geq 5.$$

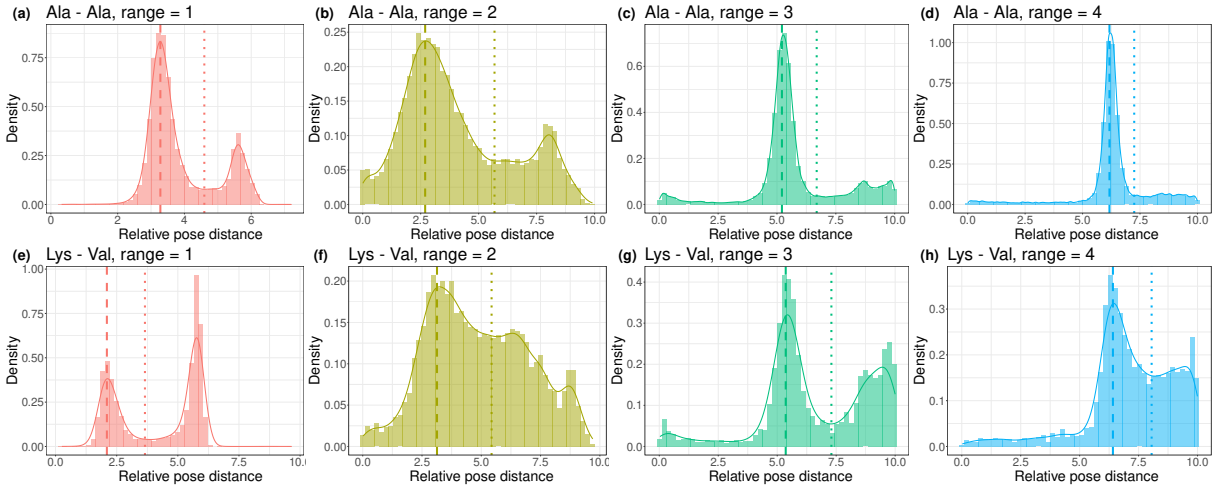



Figure S5: Empirical distribution of the relative pose distance (S10) between (a-e) Ala-Ala and (f-j) Lys-Val residues in the structural database, stratified by range groups. Distributions are depicted through a histogram and a kernel density estimate. Dashed and dotted vertical lines indicate respectively the lower and upper limit of the contact interval.

Then, we choose the decreasing function $t_{ij}$ in (S11) to concentrate its smooth decay in (S12). This can be done by choosing

$$t_{ij}(x) = 1 - \tanh\left[\left(\frac{x}{\Delta^{\mathrm{RP}}_{b;i,j}}\right)^{d^{\mathrm{RP}}_{ij}}\right] \quad \text{for all } x \geq 0 \quad \text{and} \quad d^{\mathrm{RP}}_{ij} = \frac{\log\left(\operatorname{argtanh}\left(1/\Delta^{\mathrm{RP}}_{a;i,j}\right)\right)}{\log\left(\operatorname{argtanh}\left(\Delta^{\mathrm{RP}}_{a;i,j}/\Delta^{\mathrm{RP}}_{b;i,j}\right)\right)}. \tag{S13}$$

The curve of $t_{ij}$ is illustrated in Figure S6 for Ala-Ala pairs at range 3. It shows how the contact function (S11) represents a relaxation of the classical step function based on a universal threshold. Here, contact is described by a continuous function whose transition from low to high values is smooth and concentrated inside an empirically determined sequence-specific interval.

### S1.5 UMAP and HDBSCAN algorithms

The Uniform Manifold Approximation and Projection (UMAP) algorithm was introduced in the very technical work [9], together with a more accessible and fully detailed documentation [10]. UMAP is a graph layout algorithm incorporating several theoretical foundations that provide it with a robust and well-established framework. Succintly, the UMAP algorithm builds a graph in the high dimensional space and then performs an optimization step to find the most similar graph in a lower dimension. UMAP begins by building balls centered at each point and connecting points whose corresponding balls overlap. This yields the representation of the dataset as a simplicial complex, that captures many of the main topological properties of the high-dimensional space [11]. To deal with the arbitrariness of the radius choice, the connections are made probabilistic and the edges of the graph are weighted. The resulting graphical representation is projected into a lower-dimensional space via a force-directed graph layout algorithm. The optimization procedure is similar to the one of t-SNE [12], but it effectively preserves a more substantial amount of global structure [13]. UMAP has found numerous applications in various domains, such as genetics [14,15], single-cell genomics [16,17] or neuroimaging [18]. Besides, its popularity is steadily increasing due to its demonstrated empirical efficiency, especially in enhancing the performance of clustering methods [19].

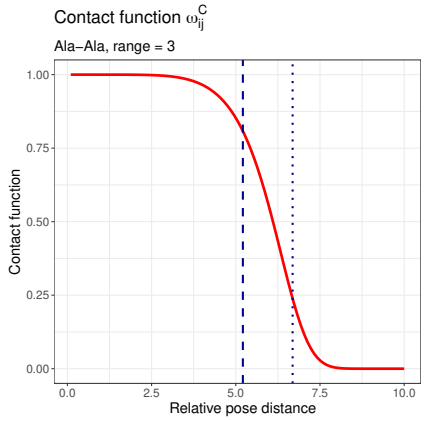



Figure S6: Contact function (S11) as a function of relative pose distance (S10) for Ala-Ala pairs at range 3. Vertical dashed and dotted lines indicate respectively the lower and upper limit of the contact interval.

HDBSCAN [20] is a hierarchical version of the DBSCAN [21] clustering algorithm. It is a density-based method, so it performs better than classical distance-based techniques like $k$-means when clusters have arbitrary shapes and sizes, or in the presence of noise or outliers. The algorithm initially follows a similar approach to that of DBSCAN. It involves a density-based transformation of the space, akin to DBSCAN, and subsequently performs single linkage clustering on the transformed space. However, an alternative strategy is carried out to avoid the use of an epsilon value to define a cutoff level for the dendrogram, enabling the identification of the more stable or persistent clusters. Instead of the cutoff parameter, HDBSCAN needs the choice of the minimum cluster size, which is more intuitive and interpretable in practical scenarios. This, together with its remarkable computational efficiency, has made HDBSCAN a very popular algorithm often implemented in combination with dimensionality reduction techniques [15, 22, 23]. For a complete explanation of the algorithm details, we refer to the HDBSCAN documentation [24].

## S2 The importance of refining contact definition

We assessed whether the effort made in Section S1 to define contact as a continuous function that integrates sequence and geometrical information is worth it to characterize ensembles. To do so, we kept the same strategy of characterizing an ensemble by a weighted family of contact maps, but starting from the classical contact definition i.e. by considering the matrix

$$\mathbf{C} = \begin{pmatrix} c_{12;1} & \cdots & c_{ij;1} & \cdots & c_{(L-1)L;1} \\ c_{12;2} & \cdots & c_{ij;2} & \cdots & c_{(L-1)L;2} \\ \vdots & & \vdots & & \vdots \\ c_{12;n} & \cdots & c_{ij;n} & \cdots & c_{(L-1)L;n} \end{pmatrix}, \tag{S14}$$

with $c_{ij,k} = 1$ if $d_{\mathbb{R}^3}(\mathcal{F}_i^k, \mathcal{F}_j^k) \leq 8\mathring{A}$ and $c_{ij,k} = 0$ otherwise, where $\mathcal{F}_i^k$ denotes the $i$-th reference frame (S2) for the $k$-th conformation. As the entries of (S14) are binary, we chose the Jaccard distance to project the data into the 10-dimensional UMAP space. Then, the clustering was performed using the

Euclidean distance between points in the low-dimensional space. Of course, using the classical contact definition based on thresholds imposes the need of metrics that are well-defined for this type of data. The choice of such metric is not straightforward and neither is its suitability in the low-dimensional projection. Whether we can correctly compare points in the UMAP space with the Euclidean distance when the high-dimensional space is $\{0,1\}^p$ is not a trivial question to address. Moving to the continuous scenario might ease the interpretation of the low-dimensional projection when using exclusively the Euclidean distance between points.

After implementing the previously described analysis to the CHCHD4 ensemble, we observed a substantial disagreement between methods when looking at the number of classified conformations. When using (1), WARIO classified the 78% of conformations. This proportion decreased to 65% when using (S14). The number of retrieved clusters was almost the same as in Section 3.1, where WARIO found 23 classes versus the 22 retrieved here, using the same value for the minimum cluster size: the 1% of $n$. When looking at the cluster-specific contact maps, we find similar contact trends when comparing both approaches. We can identify groups of conformations similarly classified with both approaches by detecting visually matching contact maps. Three examples are presented in Figure S7. This was expected as the definition proposed in Section S1 is a refinement of the classical one, and no extreme disagreements should appear.

However, remarkable differences appear when diving into short-range contacts, for which relative orientation played a role in the relative pose distance (S10). To illustrate this, we focus on the last row of Figure S7. Both contact maps seem to indicate the presence of helical motifs near the C-terminal. We already showed it in Figure 3(e) for the continuous contact definition. Conformations belonging to the corresponding cluster exhibit $\alpha$-helix structure at residues 21-24, which is confirmed by the DSSP propensities presented in Figure S8(a). However, despite the visual similarity of panels (e) and (f) in Figure S7, we can appreciate that values for the continuous contact function (panel (e)) are slightly higher at the C-terminal than the ones for the binary definition (panel (f)). This means that residues 21-24 are *closer in relative pose distance* (S10) *than in Euclidean distance*. In other words, taking relative orientation into account enhances contact identification when it is close to the preferred behavior observed in nature. Indeed, the proportion of $\alpha$-helix structures at 21-24 in cluster 6 for the binary contact clustering (see Figure S8(b)) is considerably smaller. This can be alternatively illustrated by looking at the conformations from such cluster, shown in Figure S8(c), which differ from the structured behavior depicted in Figure 4(e). Consequently, redefining contact as a continuous function (S11) that integrates sequence information and relative orientation is crucial to make the classification coherent in terms of secondary structure.

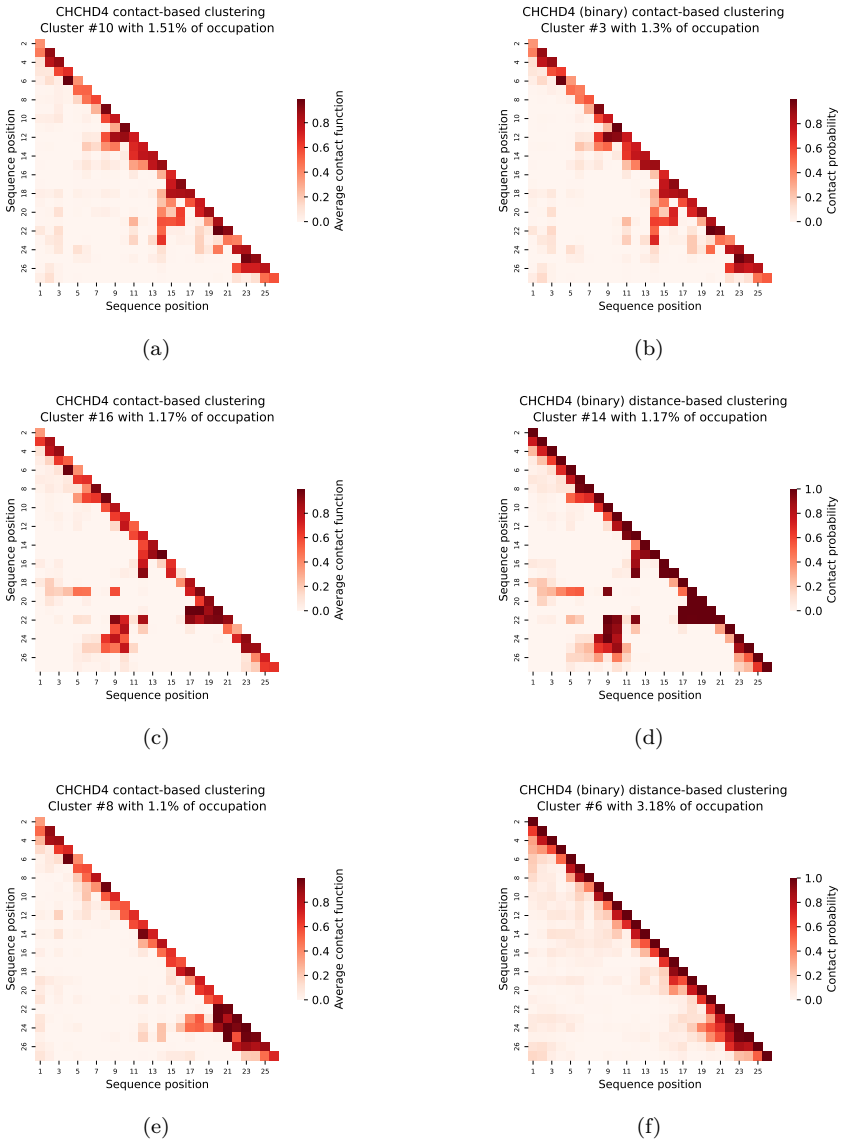
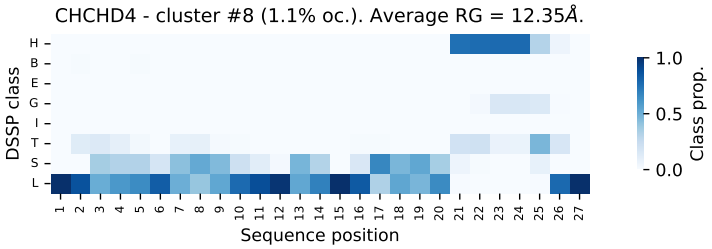
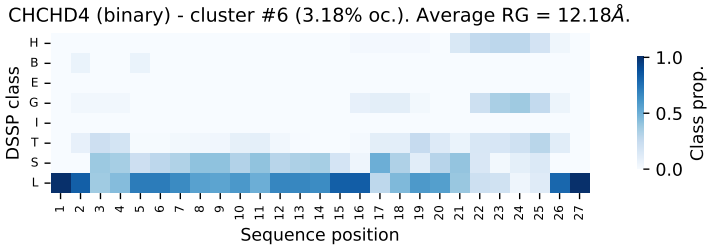


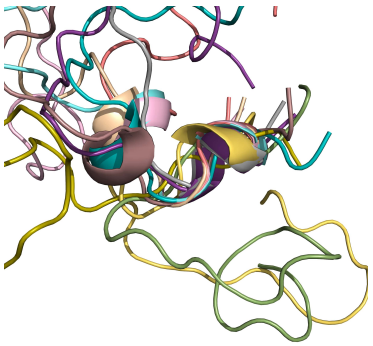

Figure S7: Left (resp. right) column: cluster-specific $\omega$-contact maps (resp. average contact maps) for CHCHD4 after performing the UMAP+clustering pipeline on (1) (resp. (S14)). Maps in the same row are those who visually match each other among both classification techniques.




(a) UMAP+HDBSCAN on (1).




(b) UMAP+HDBSCAN on (S14).



(c) Random conformers of cluster 6.

Figure S8: (a-c): Average DSSP secondary structure propensities across cluster conformations after performing the UMAP+HDBSCAN pipeline on (1), for cluster 8 (a) and on (S14), for cluster 6 (b), for the CHCHD4 ensemble. (c): 10 conformations randomly selected from cluster 6 in the previously detailed conditions.

## S3 Additional figures using other clustering approaches

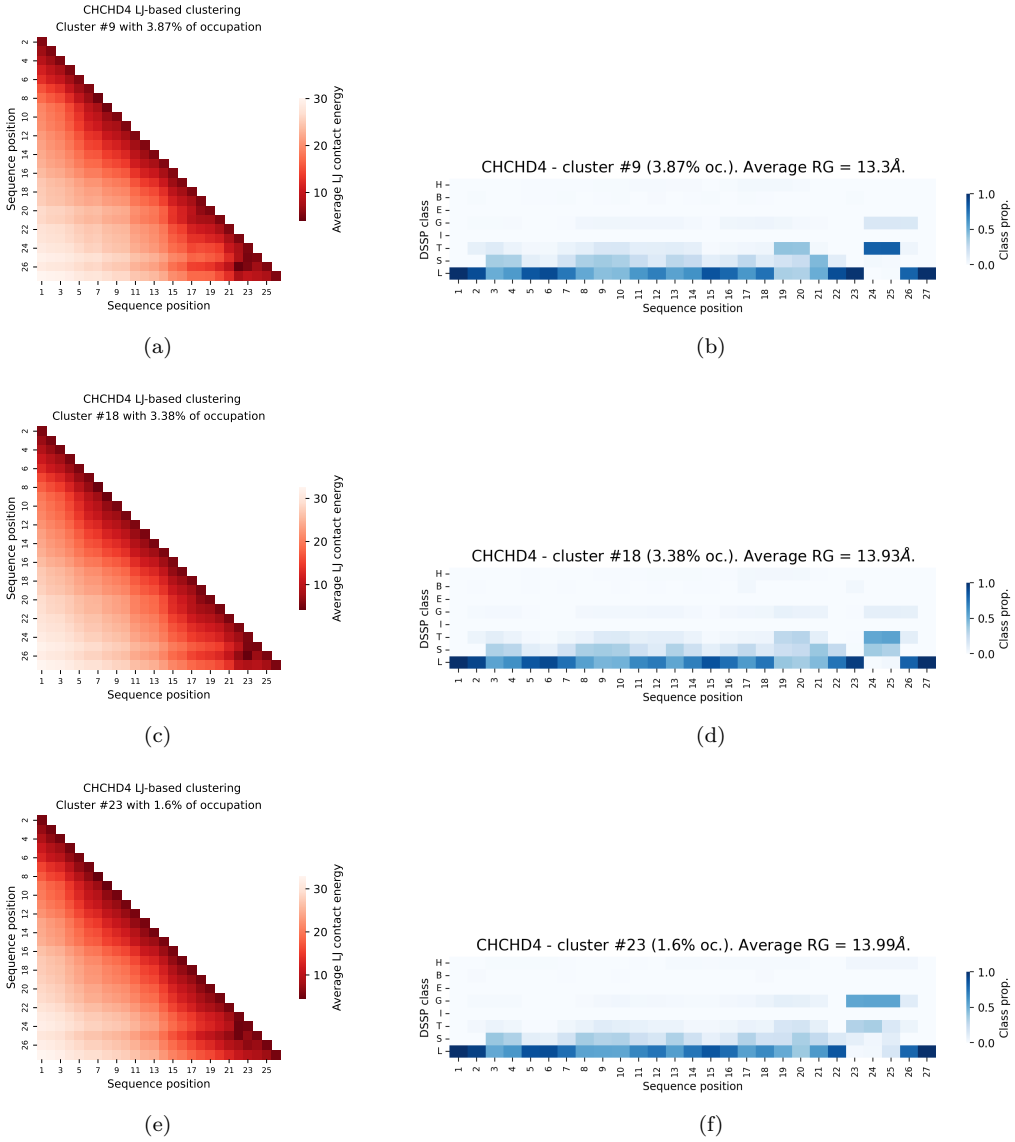


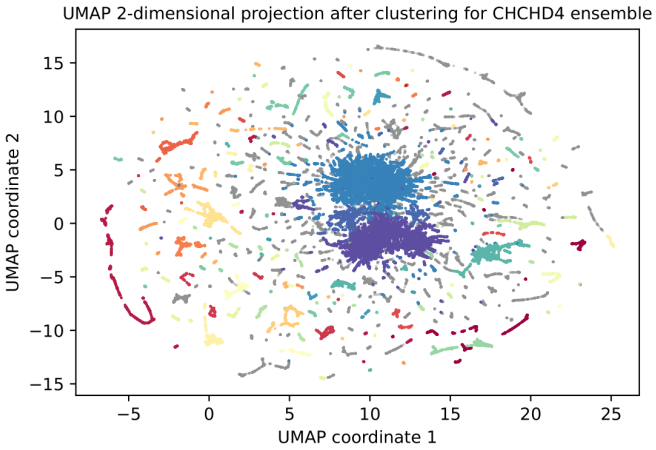

Figure S9: Left column: CHCHD4 cluster-specific Lennard-Jones contact maps after implementing the UMAP+clustering pipeline to the set of all inter-residue LJ potentials. Right column: average DSSP secondary structure propensities across cluster conformations corresponding to the cluster in the same row, left column.

## S4 Additional results

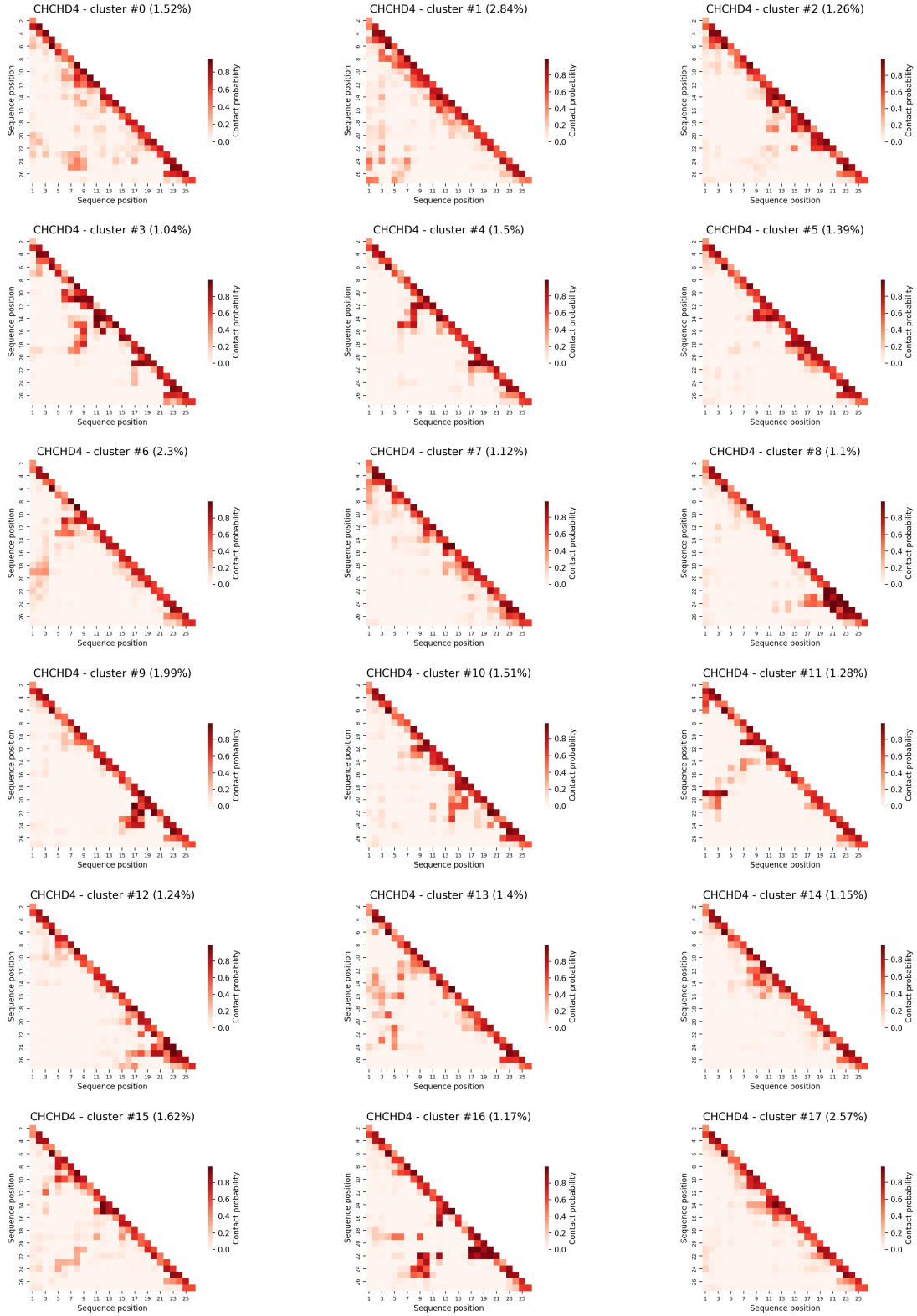
This Section presents the complete characterizations of the conformational ensembles for the three proteins studied in this work. For each one of the examples, we first present the two-dimensional UMAP embedding of conformations featured by contact functions. Points are colored according to the HDBSCAN classification, illustrating the overall distribution of the cluster occupancies. Then, the complete family of weighted $\omega$-contact maps is presented for the ensemble. Finally, we show the secondary structure propensities for the conformations of each cluster, together with their average radius of gyration. For each one of the latter plots, DSSP classes are depicted in decreasing ordinates as: H (alpha helix), B (residue in isolated beta-bridge), E (extended strand, participates in beta ladder), G (3-helix (3/10 helix)), I (5-helix (pi helix)), T (hydrogen bonded turn), S (bend) and L (loops and irregular elements).

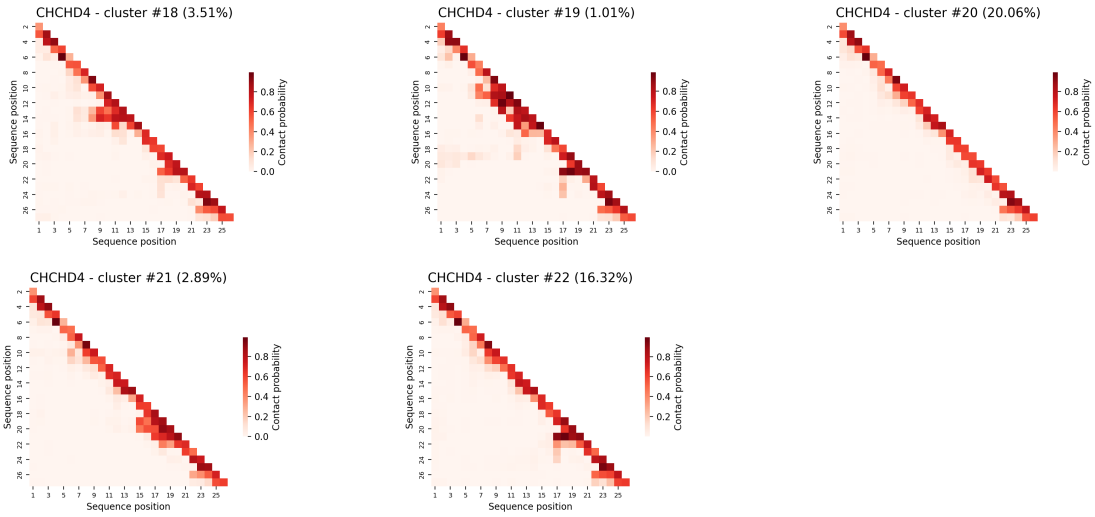
### S4.1 Complete characterization of CHCHD4

#### *S4.1.1 Two-dimensional UMAP projection*

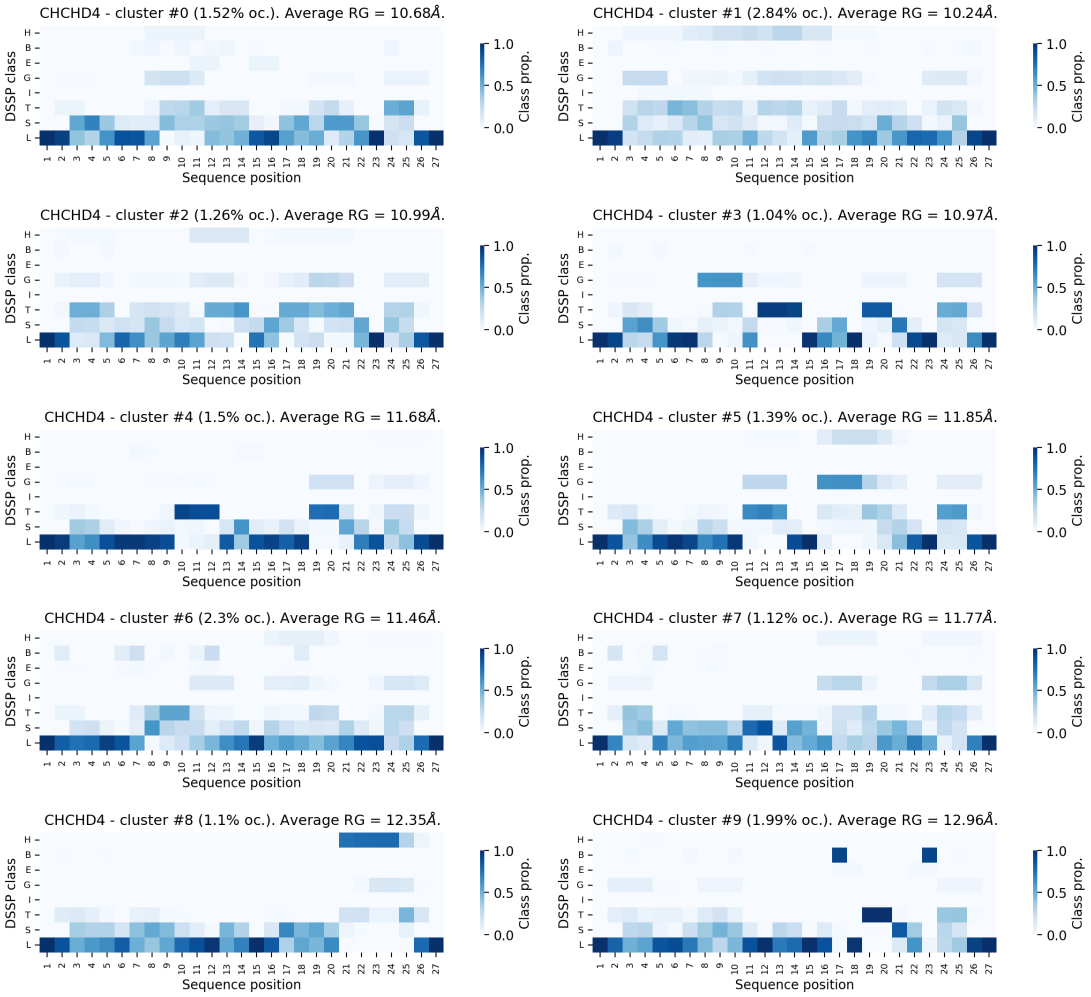


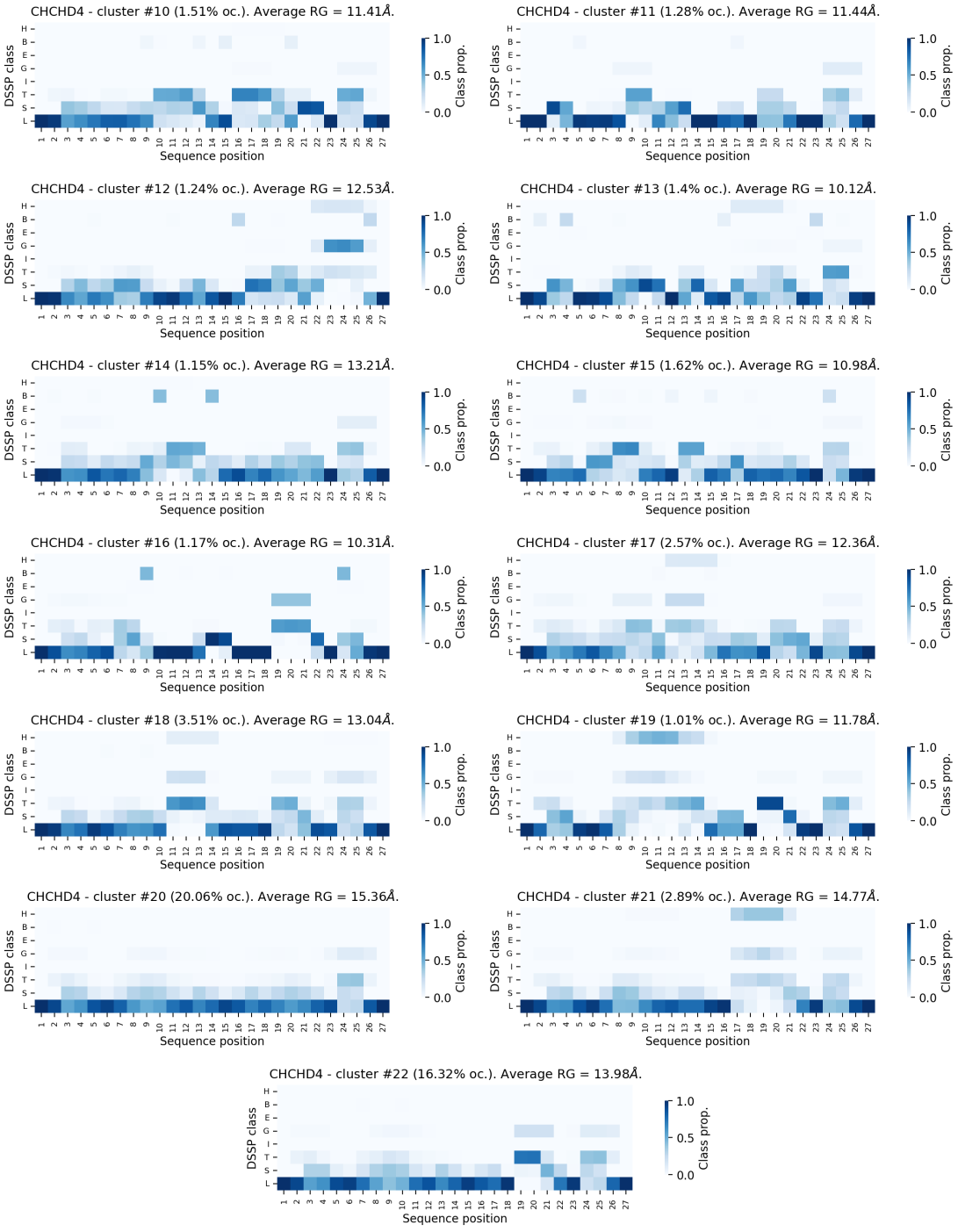
### *S4.1.2 Complete family of weighted ω-contact maps*





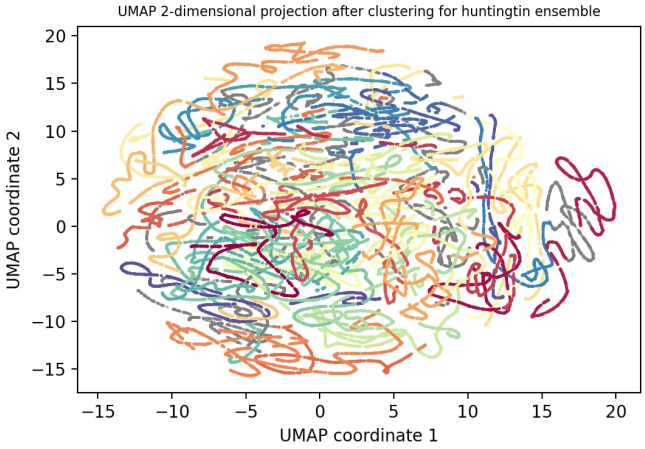

### *S4.1.3 Secondary structure propensities and average radii of gyration*



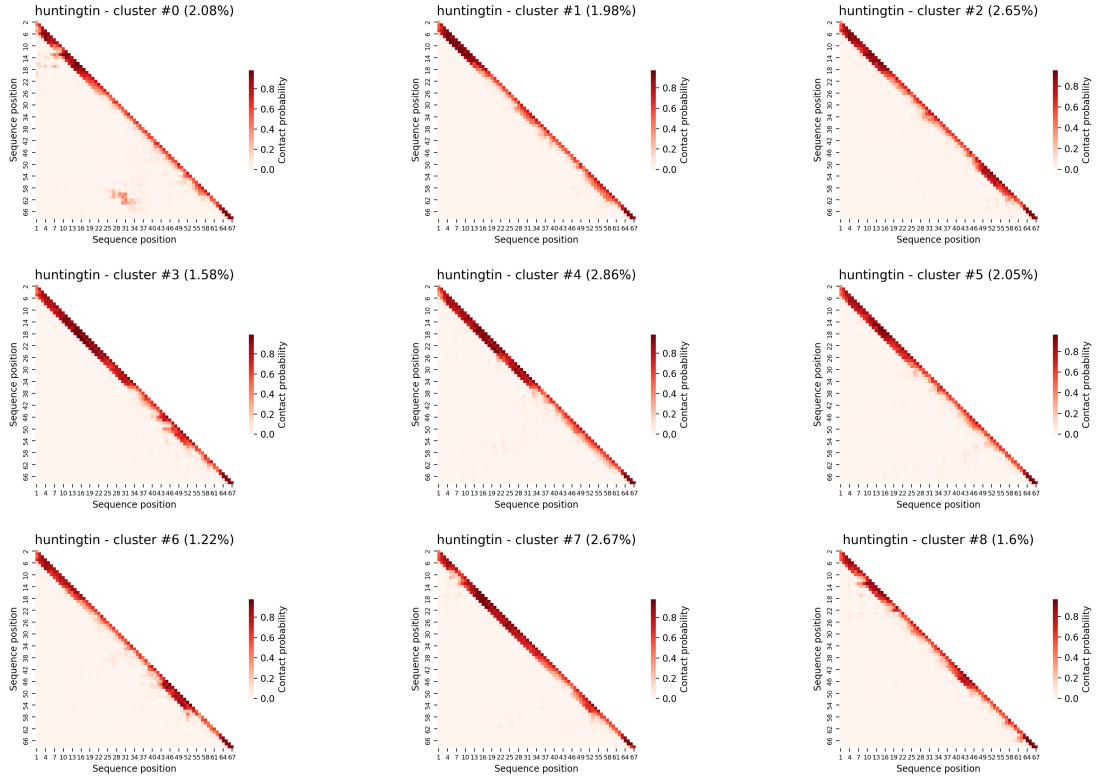

CHCHD4 - cluster #10 (1.51% oc.). Average RG = 11.41Å.
CHCHD4 - cluster #11 (1.28% oc.). Average RG = 11.44Å.
CHCHD4 - cluster #12 (1.24% oc.). Average RG = 12.53Å.
CHCHD4 - cluster #13 (1.4% oc.). Average RG = 10.12Å.
CHCHD4 - cluster #14 (1.15% oc.). Average RG = 13.21Å.
CHCHD4 - cluster #15 (1.62% oc.). Average RG = 10.98Å.
CHCHD4 - cluster #16 (1.17% oc.). Average RG = 10.31Å.
CHCHD4 - cluster #17 (2.57% oc.). Average RG = 12.36Å.
CHCHD4 - cluster #18 (3.51% oc.). Average RG = 13.04Å.
CHCHD4 - cluster #19 (1.01% oc.). Average RG = 11.78Å.
CHCHD4 - cluster #20 (20.06% oc.). Average RG = 15.36Å.
CHCHD4 - cluster #21 (2.89% oc.). Average RG = 14.77Å.
CHCHD4 - cluster #22 (16.32% oc.). Average RG = 13.98Å.
DSSP class
Sequence position
Class prop.

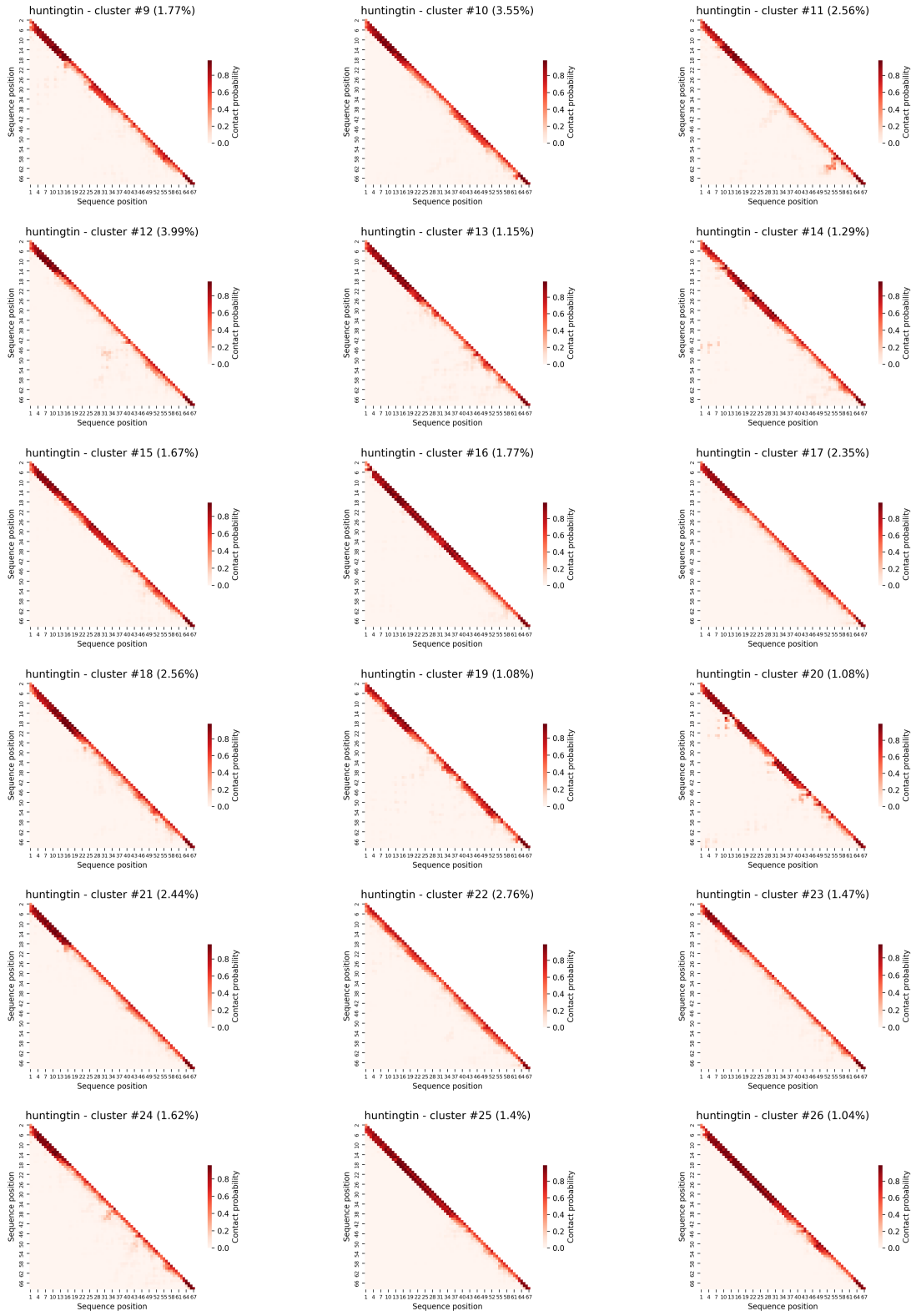
## S4.2 Complete characterization of Huntingtin

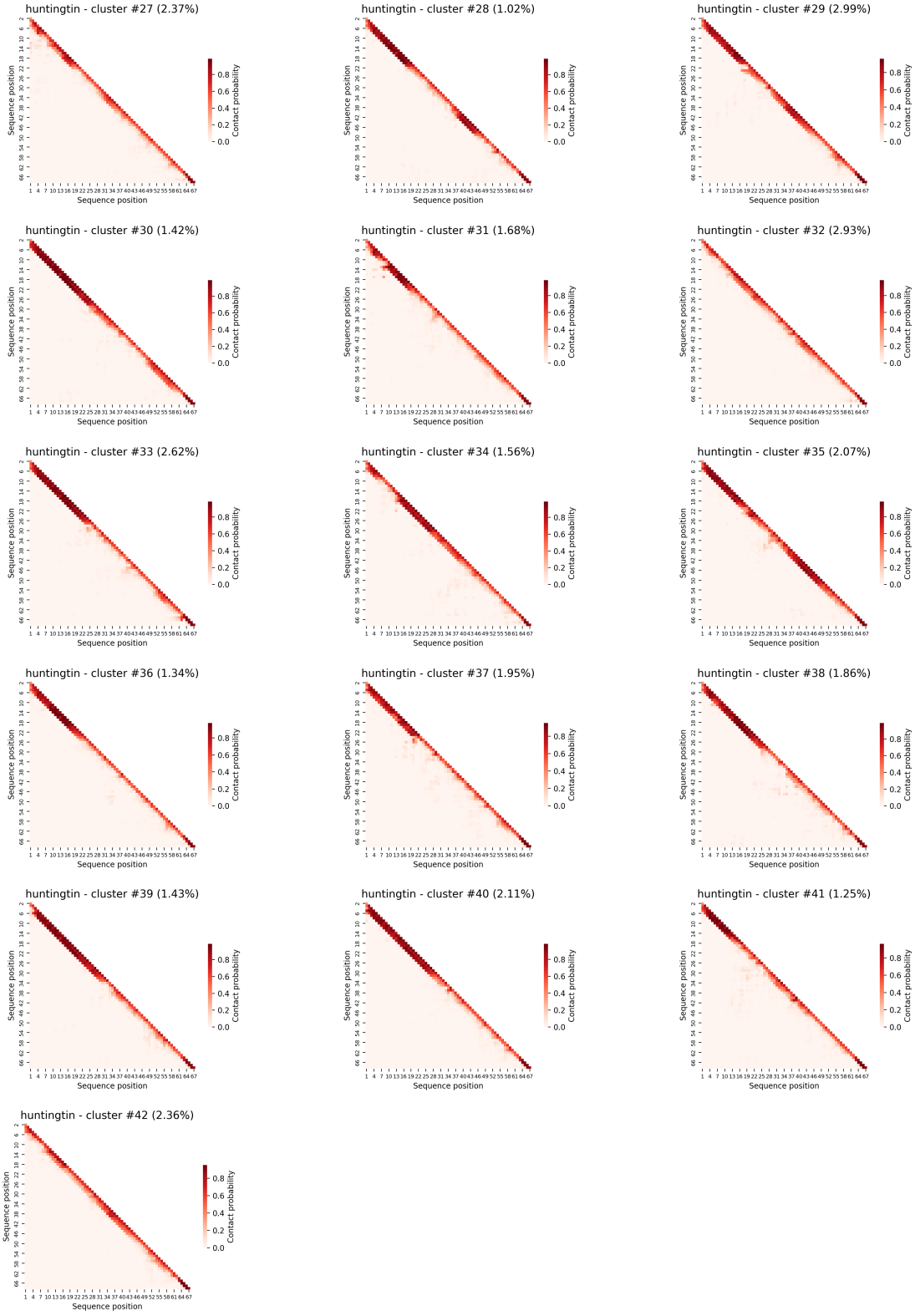
### *S4.2.1 Two-dimensional UMAP projection*



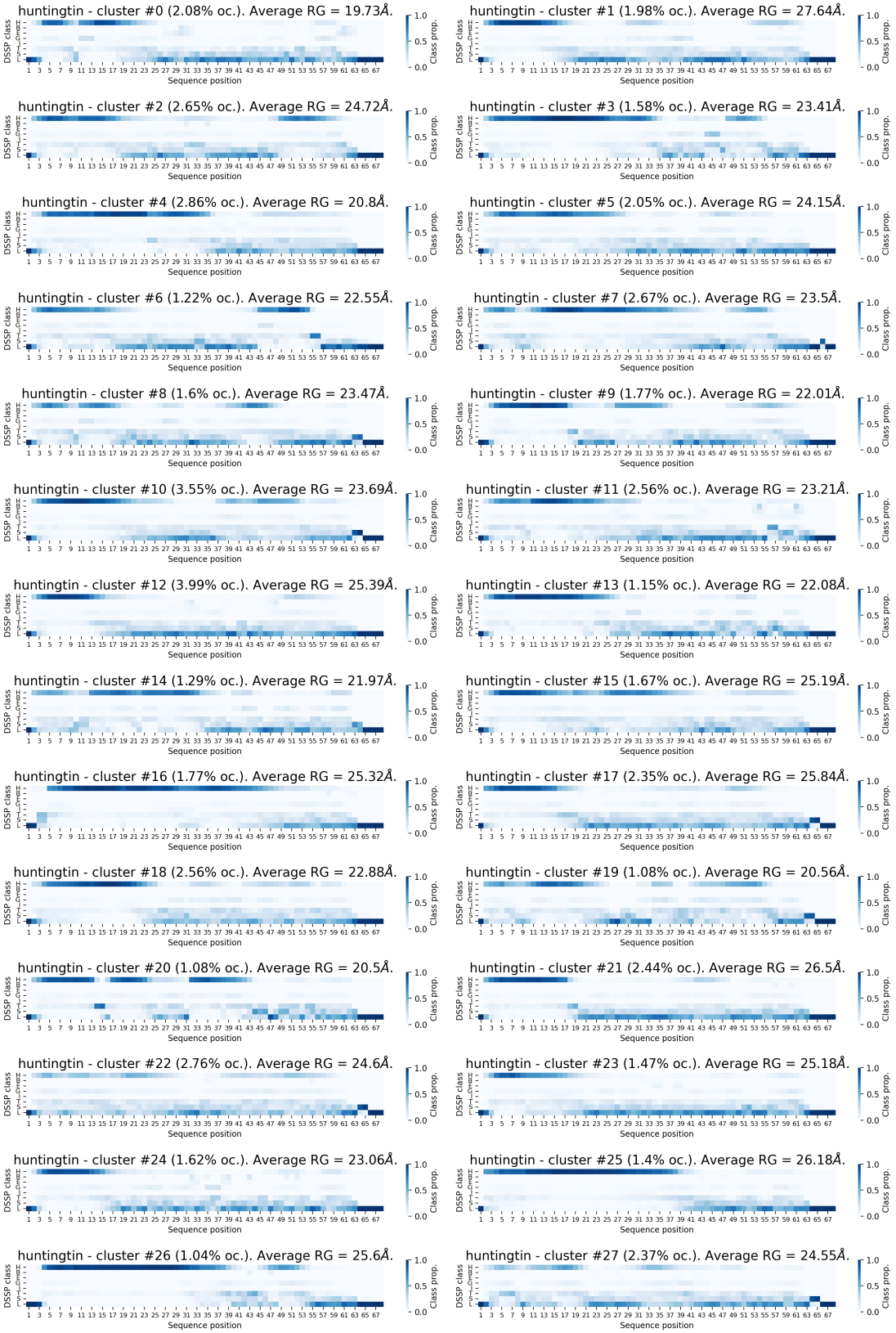

### *S4.2.2 Complete family of weighted ω-contact maps*

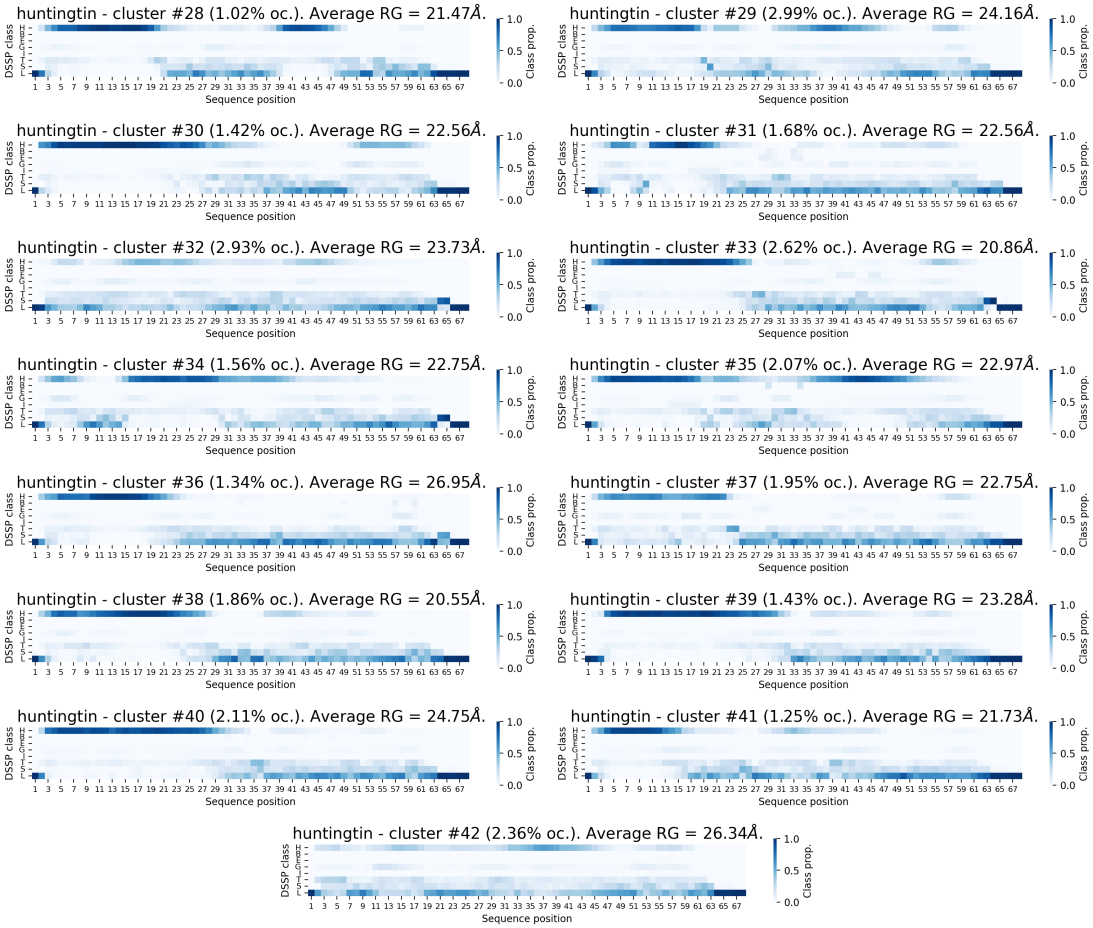



huntingtin - cluster #9 (1.77%)
huntingtin - cluster #10 (3.55%)
huntingtin - cluster #11 (2.56%)
huntingtin - cluster #12 (3.99%)
huntingtin - cluster #13 (1.15%)
huntingtin - cluster #14 (1.29%)
huntingtin - cluster #15 (1.67%)
huntingtin - cluster #16 (1.77%)
huntingtin - cluster #17 (2.35%)
huntingtin - cluster #18 (2.56%)
huntingtin - cluster #19 (1.08%)
huntingtin - cluster #20 (1.08%)
huntingtin - cluster #21 (2.44%)
huntingtin - cluster #22 (2.76%)
huntingtin - cluster #23 (1.47%)
huntingtin - cluster #24 (1.62%)
huntingtin - cluster #25 (1.4%)
huntingtin - cluster #26 (1.04%)
Sequence position
Contact probability


huntingtin - cluster #27 (2.37%)
huntingtin - cluster #28 (1.02%)
huntingtin - cluster #29 (2.99%)
huntingtin - cluster #30 (1.42%)
huntingtin - cluster #31 (1.68%)
huntingtin - cluster #32 (2.93%)
huntingtin - cluster #33 (2.62%)
huntingtin - cluster #34 (1.56%)
huntingtin - cluster #35 (2.07%)
huntingtin - cluster #36 (1.34%)
huntingtin - cluster #37 (1.95%)
huntingtin - cluster #38 (1.86%)
huntingtin - cluster #39 (1.43%)
huntingtin - cluster #40 (2.11%)
huntingtin - cluster #41 (1.25%)
huntingtin - cluster #42 (2.36%)
Sequence position
Contact probability

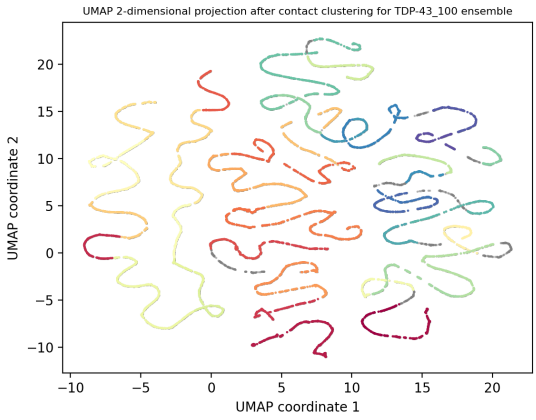
### *S4.2.3 Secondary structure propensities and average radii of gyration*



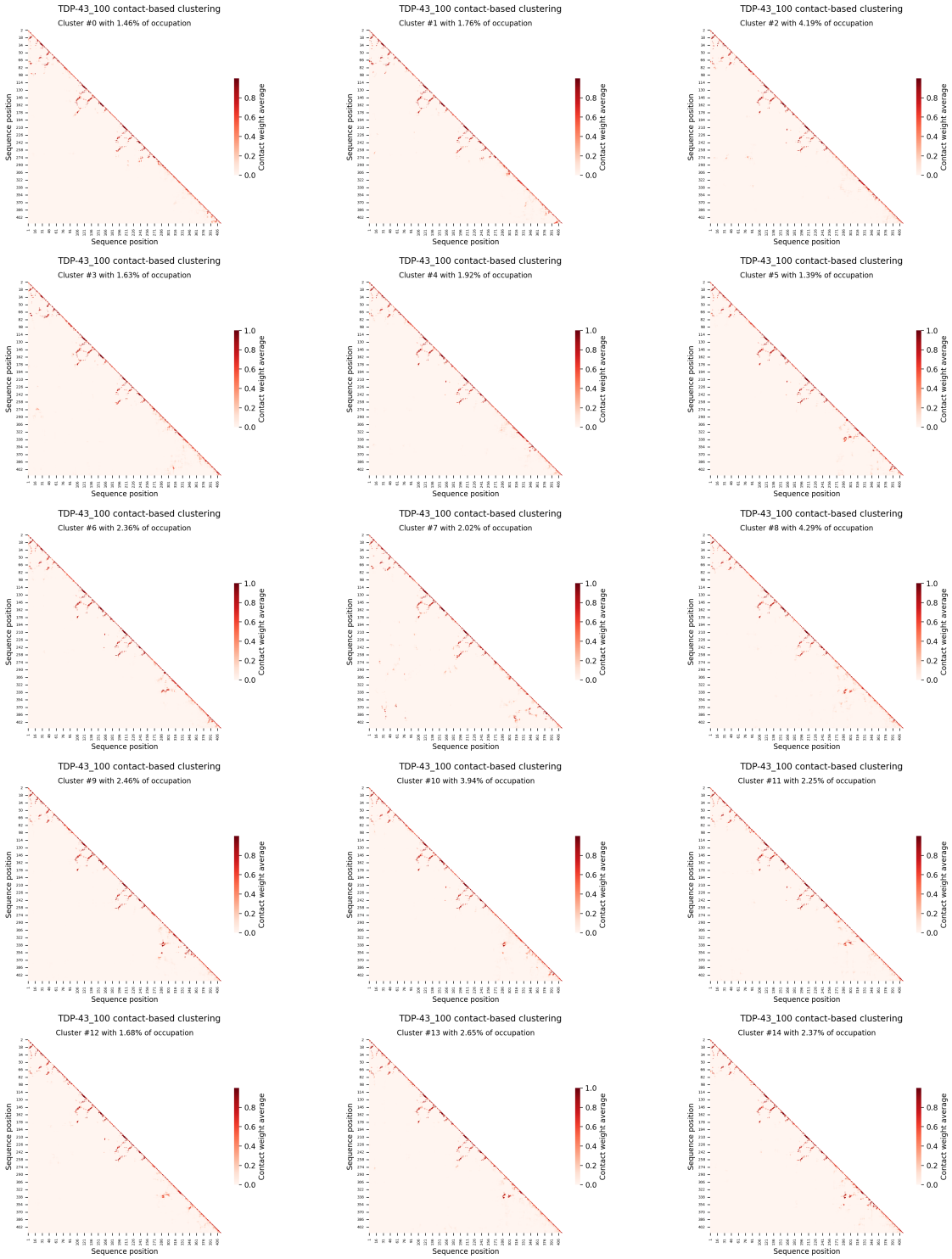


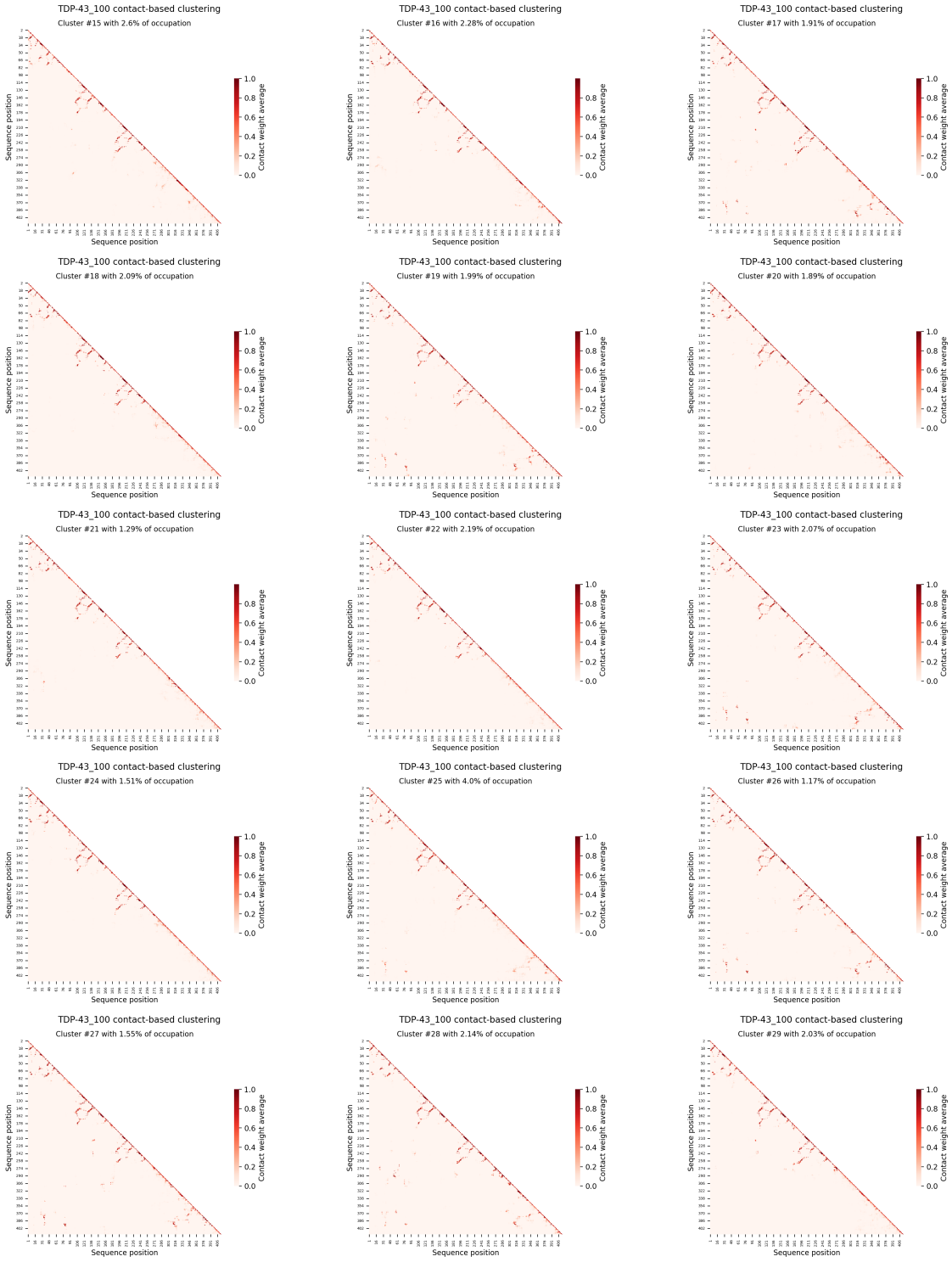

## S4.3 Complete characterization of TDP-43 at 100 mM NaCl

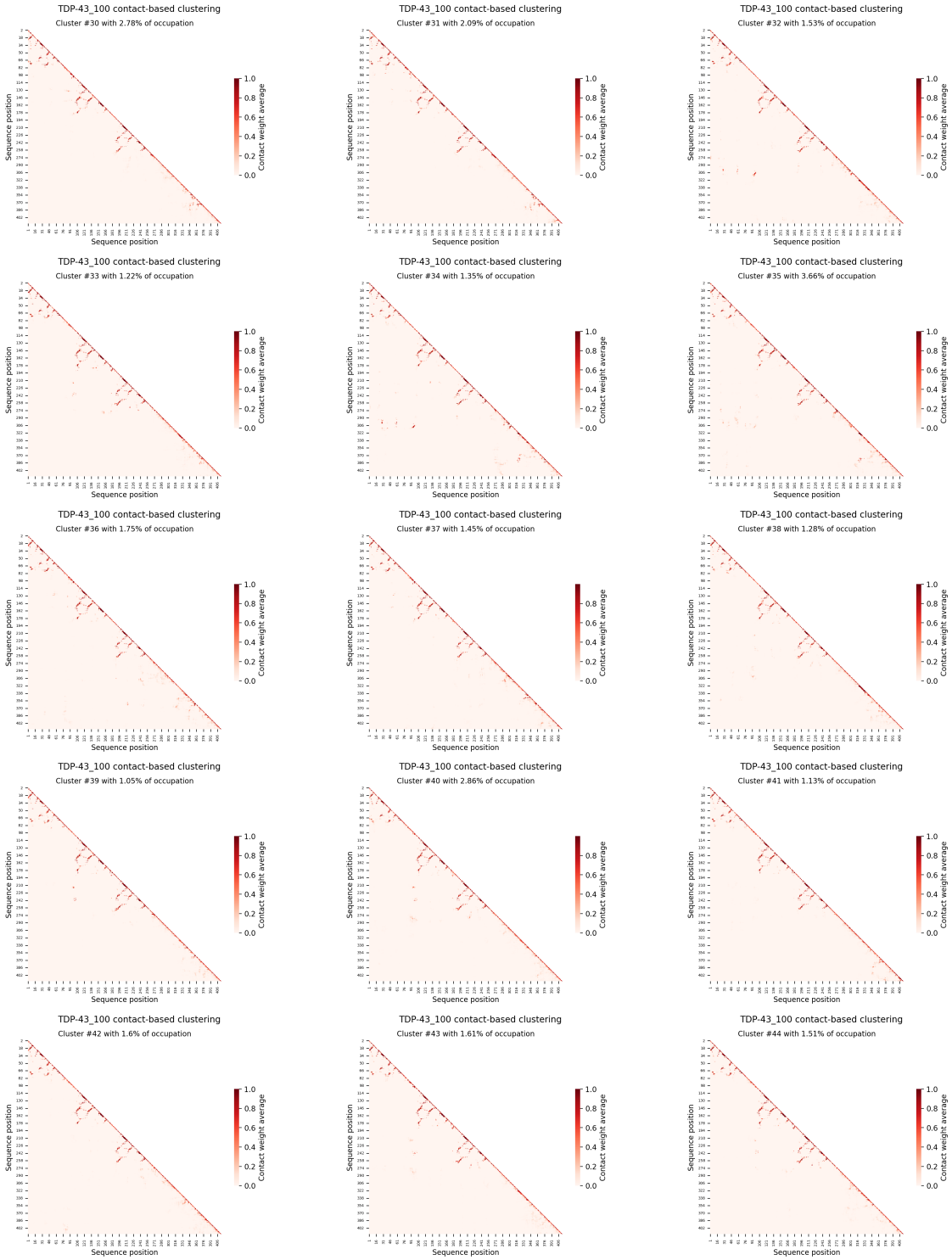
### *S4.3.1 Two-dimensional UMAP projection*



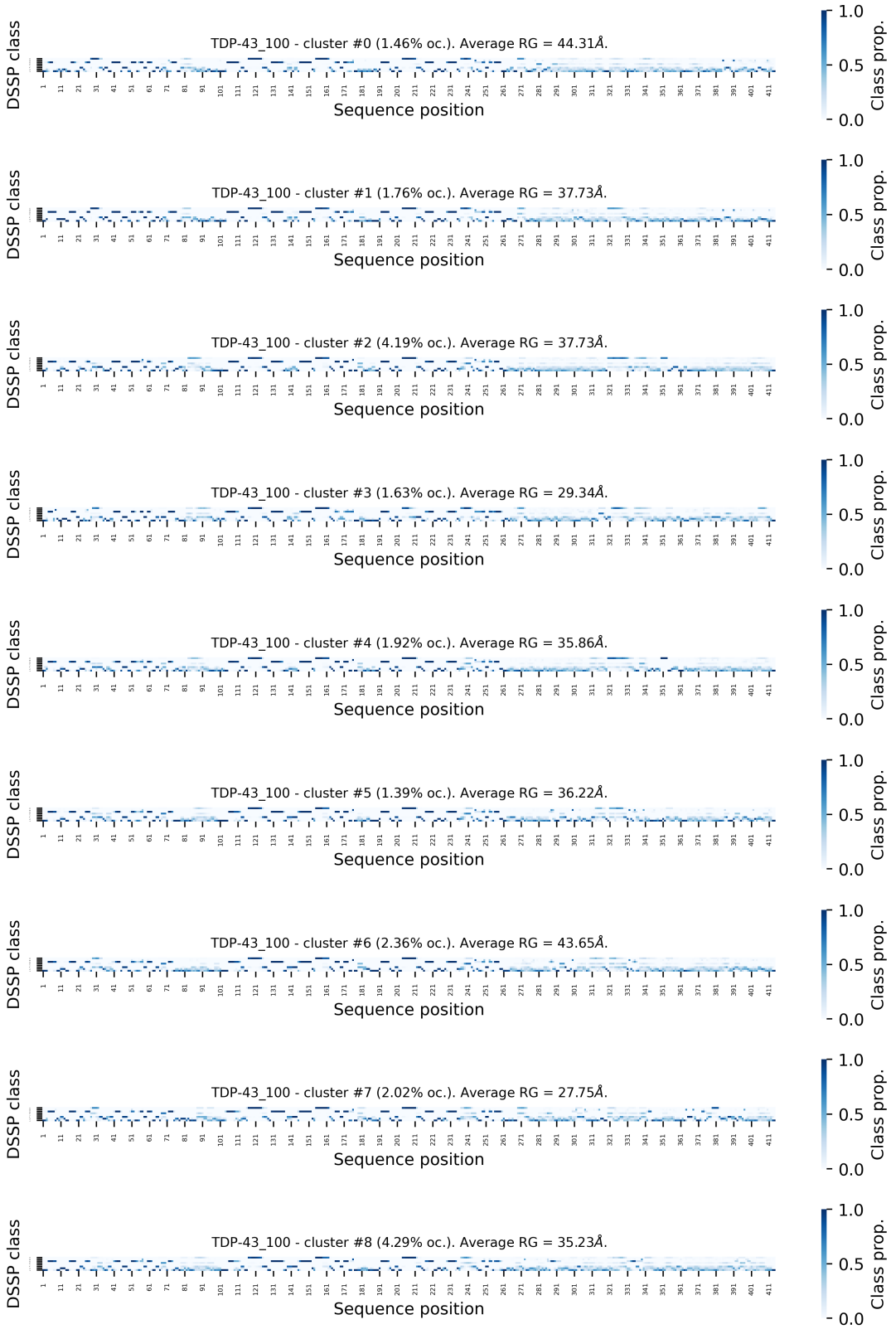
### *S4.3.2 Complete family of weighted ω-contact maps*

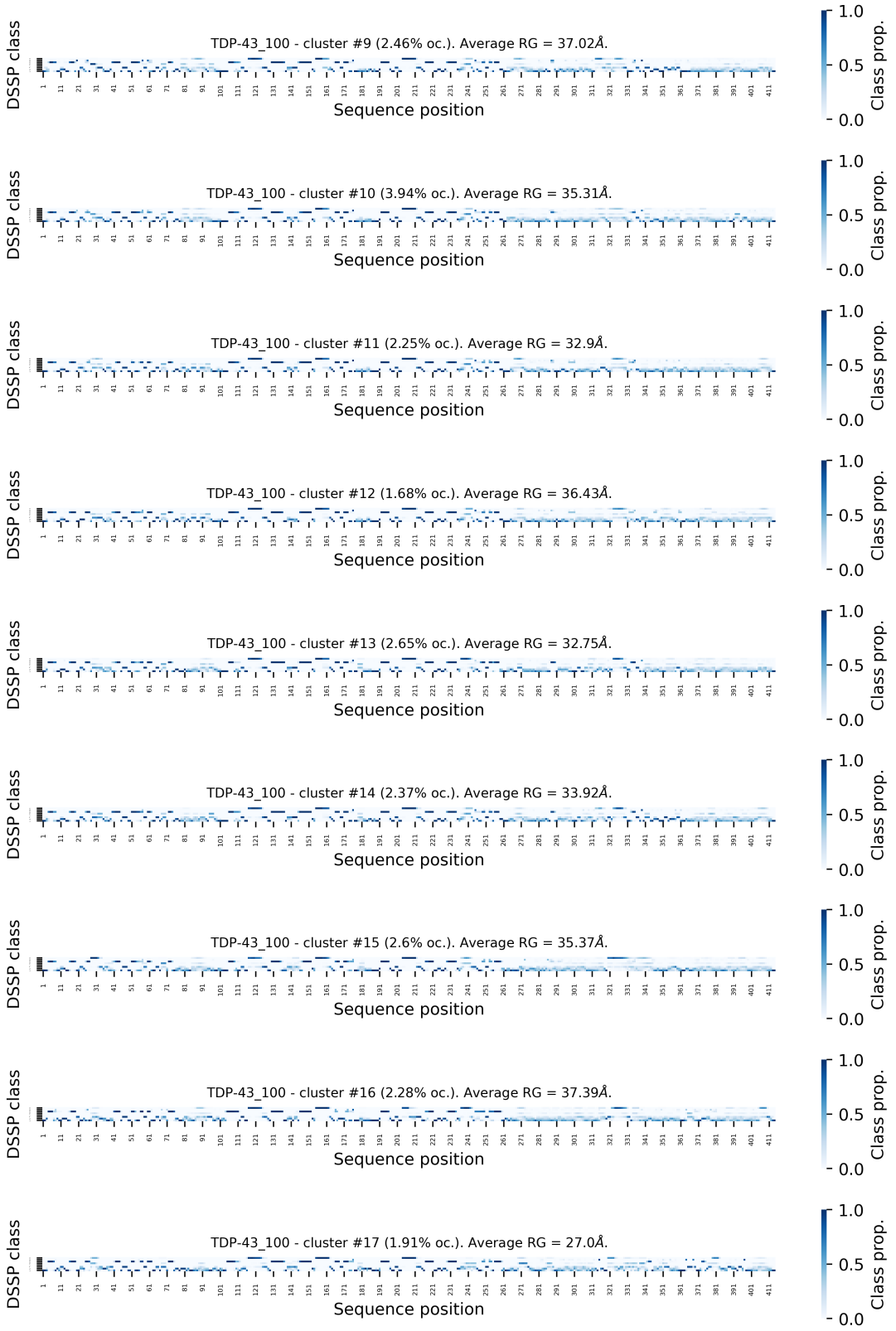


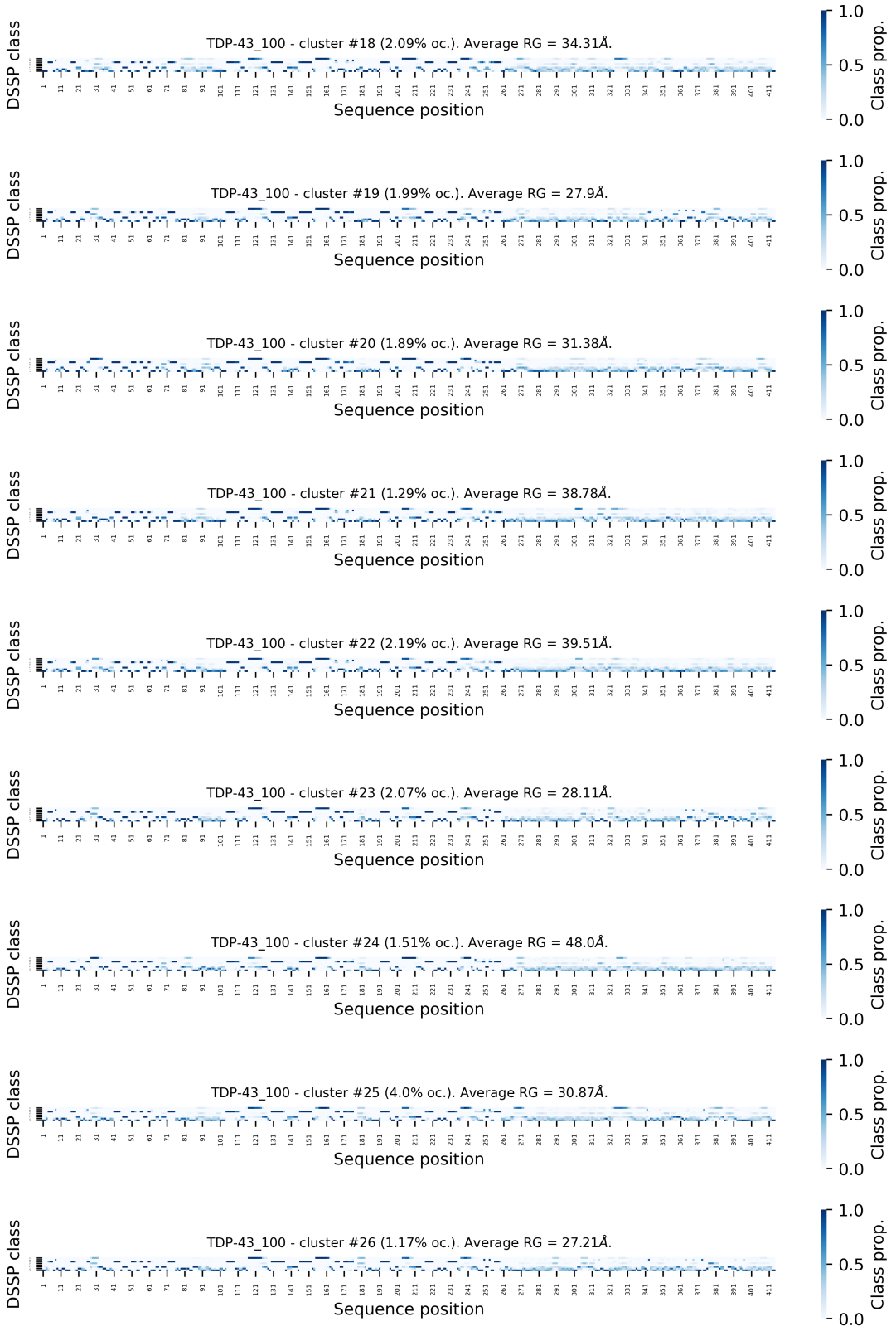

TDP-43_100 contact-based clustering
Cluster #15 with 2.6% of occupation
TDP-43_100 contact-based clustering
Cluster #16 with 2.28% of occupation
TDP-43_100 contact-based clustering
Cluster #17 with 1.91% of occupation
TDP-43_100 contact-based clustering
Cluster #18 with 2.09% of occupation
TDP-43_100 contact-based clustering
Cluster #19 with 1.99% of occupation
TDP-43_100 contact-based clustering
Cluster #20 with 1.89% of occupation
TDP-43_100 contact-based clustering
Cluster #21 with 1.29% of occupation
TDP-43_100 contact-based clustering
Cluster #22 with 2.19% of occupation
TDP-43_100 contact-based clustering
Cluster #23 with 2.07% of occupation
TDP-43_100 contact-based clustering
Cluster #24 with 1.51% of occupation
TDP-43_100 contact-based clustering
Cluster #25 with 4.0% of occupation
TDP-43_100 contact-based clustering
Cluster #26 with 1.17% of occupation
TDP-43_100 contact-based clustering
Cluster #27 with 1.55% of occupation
TDP-43_100 contact-based clustering
Cluster #28 with 2.14% of occupation
TDP-43_100 contact-based clustering
Cluster #29 with 2.03% of occupation
Sequence position
Contact weight average

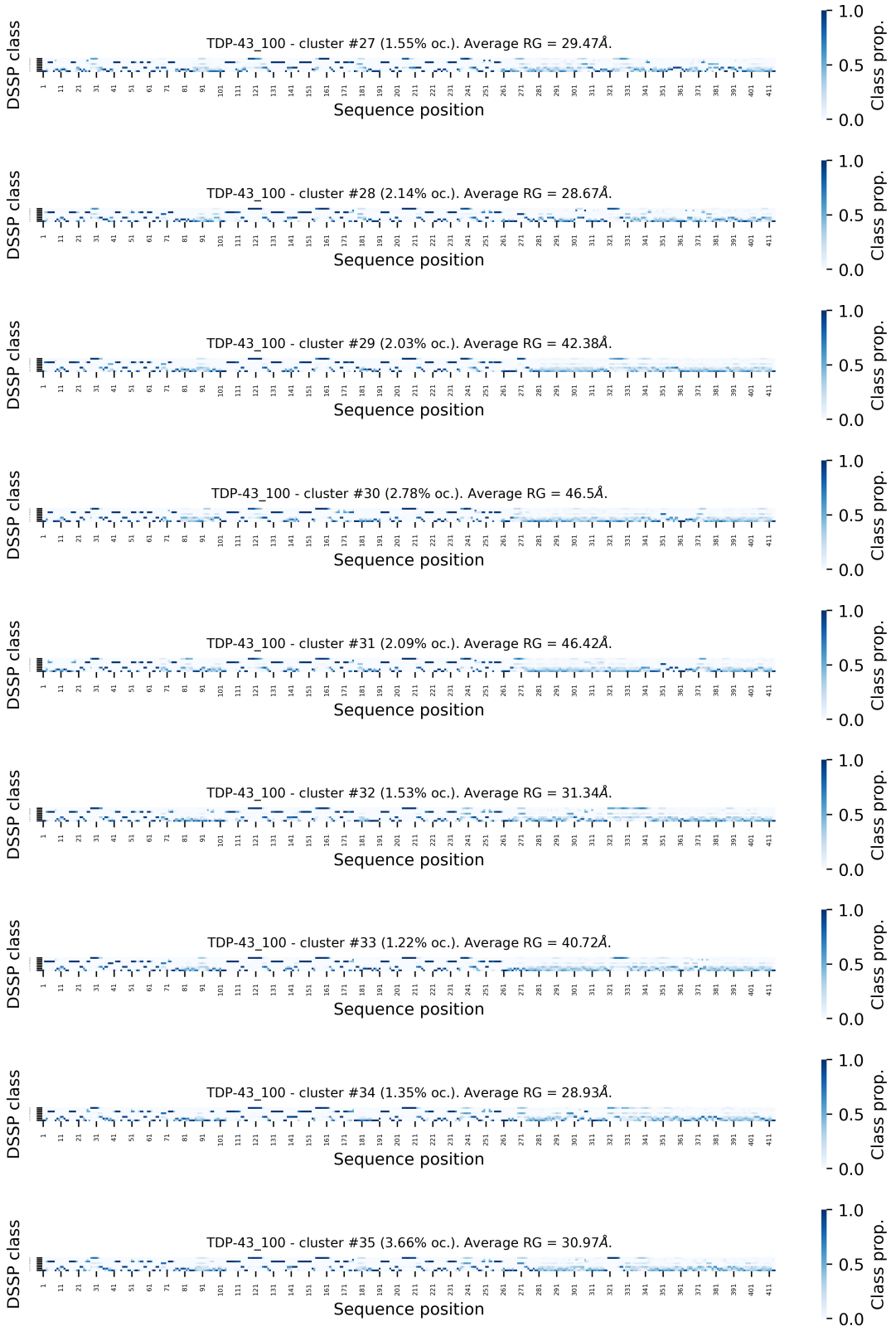

TDP-43_100 contact-based clustering
Cluster #30 with 2.78% of occupation
TDP-43_100 contact-based clustering
Cluster #31 with 2.09% of occupation
TDP-43_100 contact-based clustering
Cluster #32 with 1.53% of occupation
TDP-43_100 contact-based clustering
Cluster #33 with 1.22% of occupation
TDP-43_100 contact-based clustering
Cluster #34 with 1.35% of occupation
TDP-43_100 contact-based clustering
Cluster #35 with 3.66% of occupation
TDP-43_100 contact-based clustering
Cluster #36 with 1.75% of occupation
TDP-43_100 contact-based clustering
Cluster #37 with 1.45% of occupation
TDP-43_100 contact-based clustering
Cluster #38 with 1.28% of occupation
TDP-43_100 contact-based clustering
Cluster #39 with 1.05% of occupation
TDP-43_100 contact-based clustering
Cluster #40 with 2.86% of occupation
TDP-43_100 contact-based clustering
Cluster #41 with 1.13% of occupation
TDP-43_100 contact-based clustering
Cluster #42 with 1.6% of occupation
TDP-43_100 contact-based clustering
Cluster #43 with 1.61% of occupation
TDP-43_100 contact-based clustering
Cluster #44 with 1.51% of occupation
Sequence position
Contact weight average

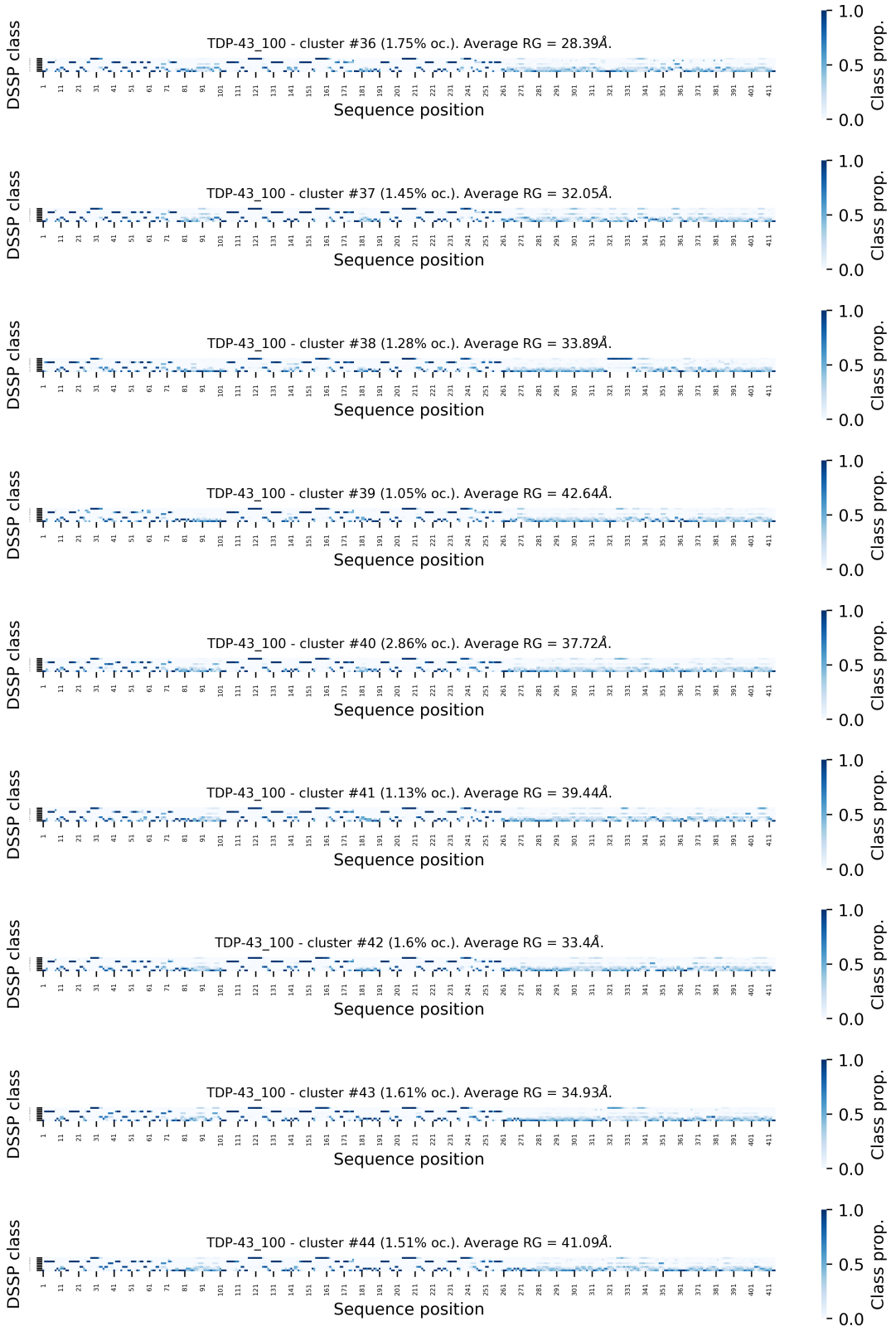
### *S4.3.3 Secondary structure propensities and average radii of gyration*




TDP-43_100 - cluster #9 (2.46% oc.). Average RG = 37.02Å.
TDP-43_100 - cluster #10 (3.94% oc.). Average RG = 35.31Å.
TDP-43_100 - cluster #11 (2.25% oc.). Average RG = 32.9Å.
TDP-43_100 - cluster #12 (1.68% oc.). Average RG = 36.43Å.
TDP-43_100 - cluster #13 (2.65% oc.). Average RG = 32.75Å.
TDP-43_100 - cluster #14 (2.37% oc.). Average RG = 33.92Å.
TDP-43_100 - cluster #15 (2.6% oc.). Average RG = 35.37Å.
TDP-43_100 - cluster #16 (2.28% oc.). Average RG = 37.39Å.
TDP-43_100 - cluster #17 (1.91% oc.). Average RG = 27.0Å.
DSSP class
Sequence position
Class prop.
1.0
0.5
0.0

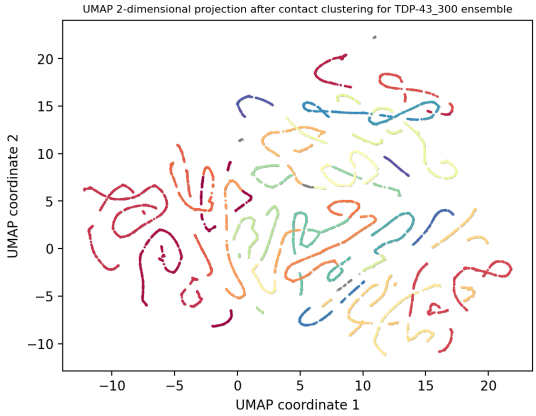

TDP-43_100 - cluster #18 (2.09% oc.). Average RG = 34.31Å.
TDP-43_100 - cluster #19 (1.99% oc.). Average RG = 27.9Å.
TDP-43_100 - cluster #20 (1.89% oc.). Average RG = 31.38Å.
TDP-43_100 - cluster #21 (1.29% oc.). Average RG = 38.78Å.
TDP-43_100 - cluster #22 (2.19% oc.). Average RG = 39.51Å.
TDP-43_100 - cluster #23 (2.07% oc.). Average RG = 28.11Å.
TDP-43_100 - cluster #24 (1.51% oc.). Average RG = 48.0Å.
TDP-43_100 - cluster #25 (4.0% oc.). Average RG = 30.87Å.
TDP-43_100 - cluster #26 (1.17% oc.). Average RG = 27.21Å.
DSSP class
Sequence position
Class prop.
1.0
0.5
0.0


TDP-43_100 - cluster #27 (1.55% oc.). Average RG = 29.47Å.
TDP-43_100 - cluster #28 (2.14% oc.). Average RG = 28.67Å.
TDP-43_100 - cluster #29 (2.03% oc.). Average RG = 42.38Å.
TDP-43_100 - cluster #30 (2.78% oc.). Average RG = 46.5Å.
TDP-43_100 - cluster #31 (2.09% oc.). Average RG = 46.42Å.
TDP-43_100 - cluster #32 (1.53% oc.). Average RG = 31.34Å.
TDP-43_100 - cluster #33 (1.22% oc.). Average RG = 40.72Å.
TDP-43_100 - cluster #34 (1.35% oc.). Average RG = 28.93Å.
TDP-43_100 - cluster #35 (3.66% oc.). Average RG = 30.97Å.
DSSP class
Sequence position
Class prop.
1.0
0.5
0.0

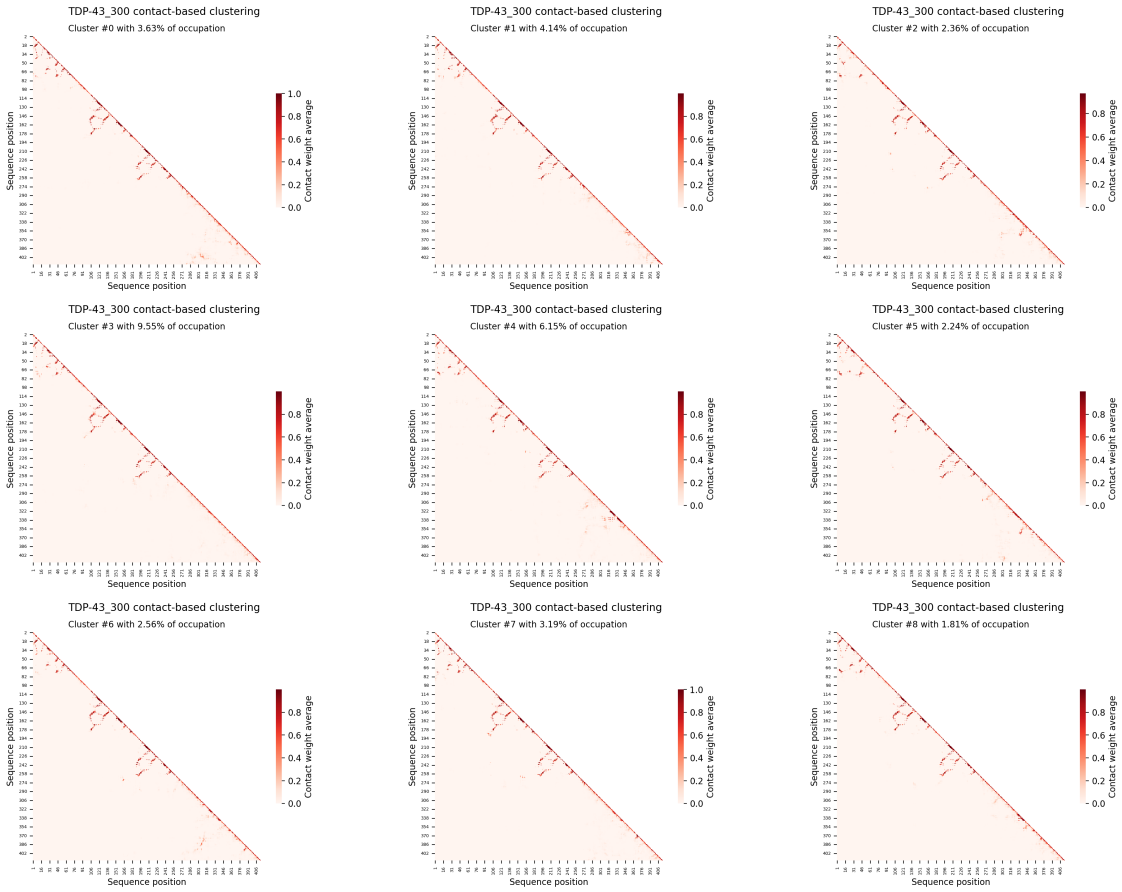

TDP-43_100 - cluster #36 (1.75% oc.). Average RG = 28.39Å.
TDP-43_100 - cluster #37 (1.45% oc.). Average RG = 32.05Å.
TDP-43_100 - cluster #38 (1.28% oc.). Average RG = 33.89Å.
TDP-43_100 - cluster #39 (1.05% oc.). Average RG = 42.64Å.
TDP-43_100 - cluster #40 (2.86% oc.). Average RG = 37.72Å.
TDP-43_100 - cluster #41 (1.13% oc.). Average RG = 39.44Å.
TDP-43_100 - cluster #42 (1.6% oc.). Average RG = 33.4Å.
TDP-43_100 - cluster #43 (1.61% oc.). Average RG = 34.93Å.
TDP-43_100 - cluster #44 (1.51% oc.). Average RG = 41.09Å.
DSSP class
Sequence position
Class prop.
1.0
0.5
0.0

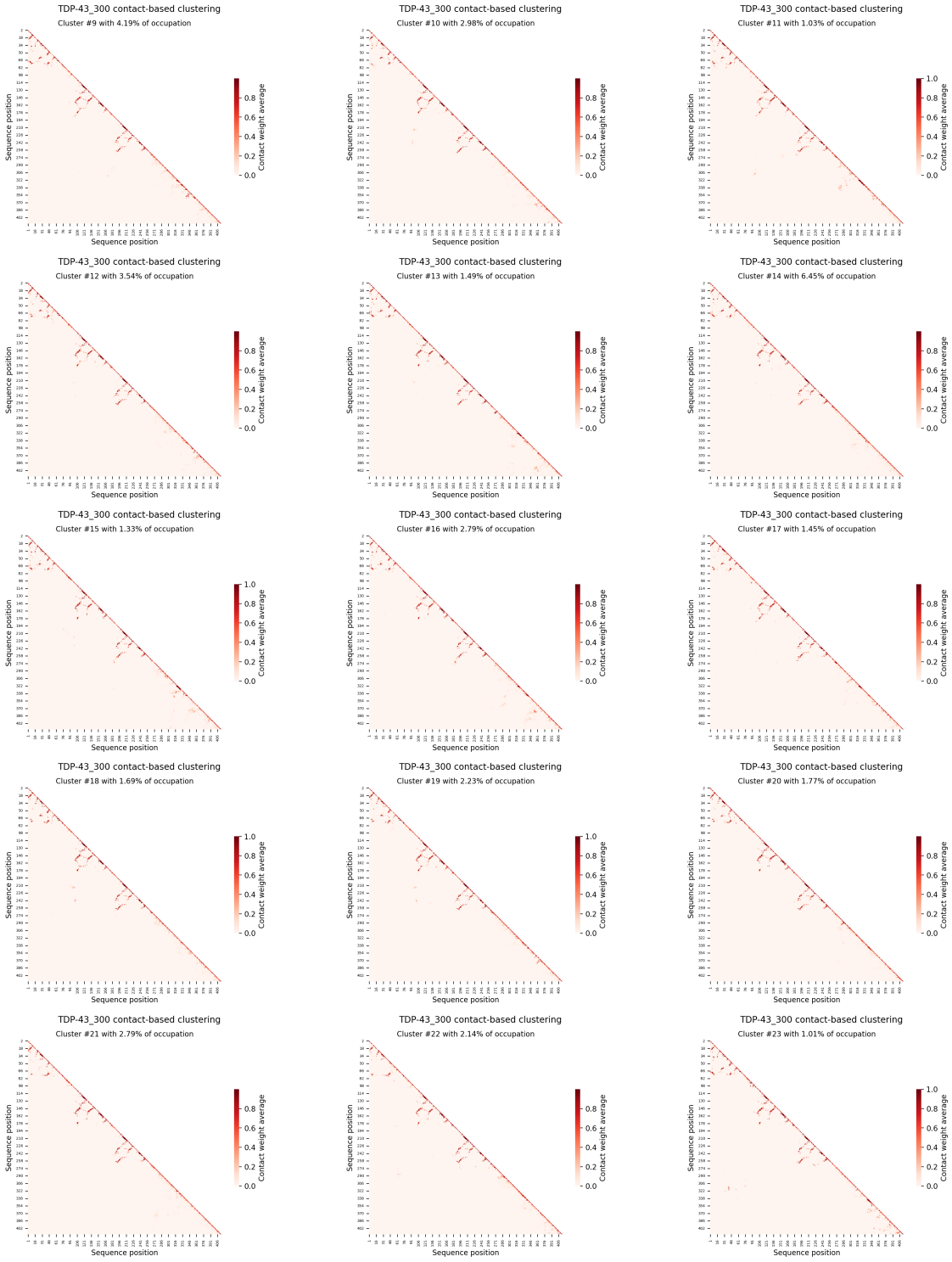
## S4.4 Complete characterization of TDP-43 at 300 mM NaCl

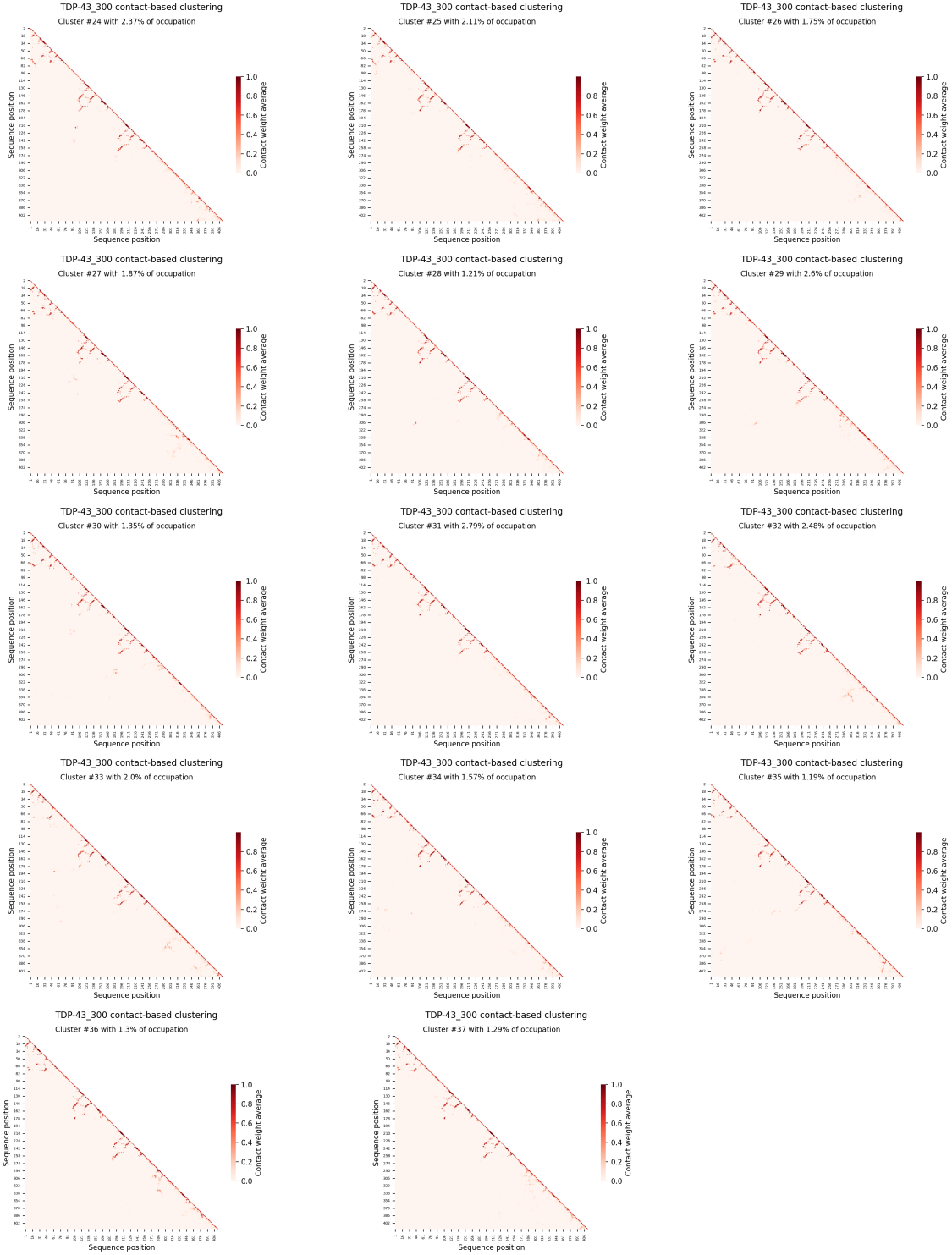
### *S4.4.1 Two-dimensional UMAP projection*



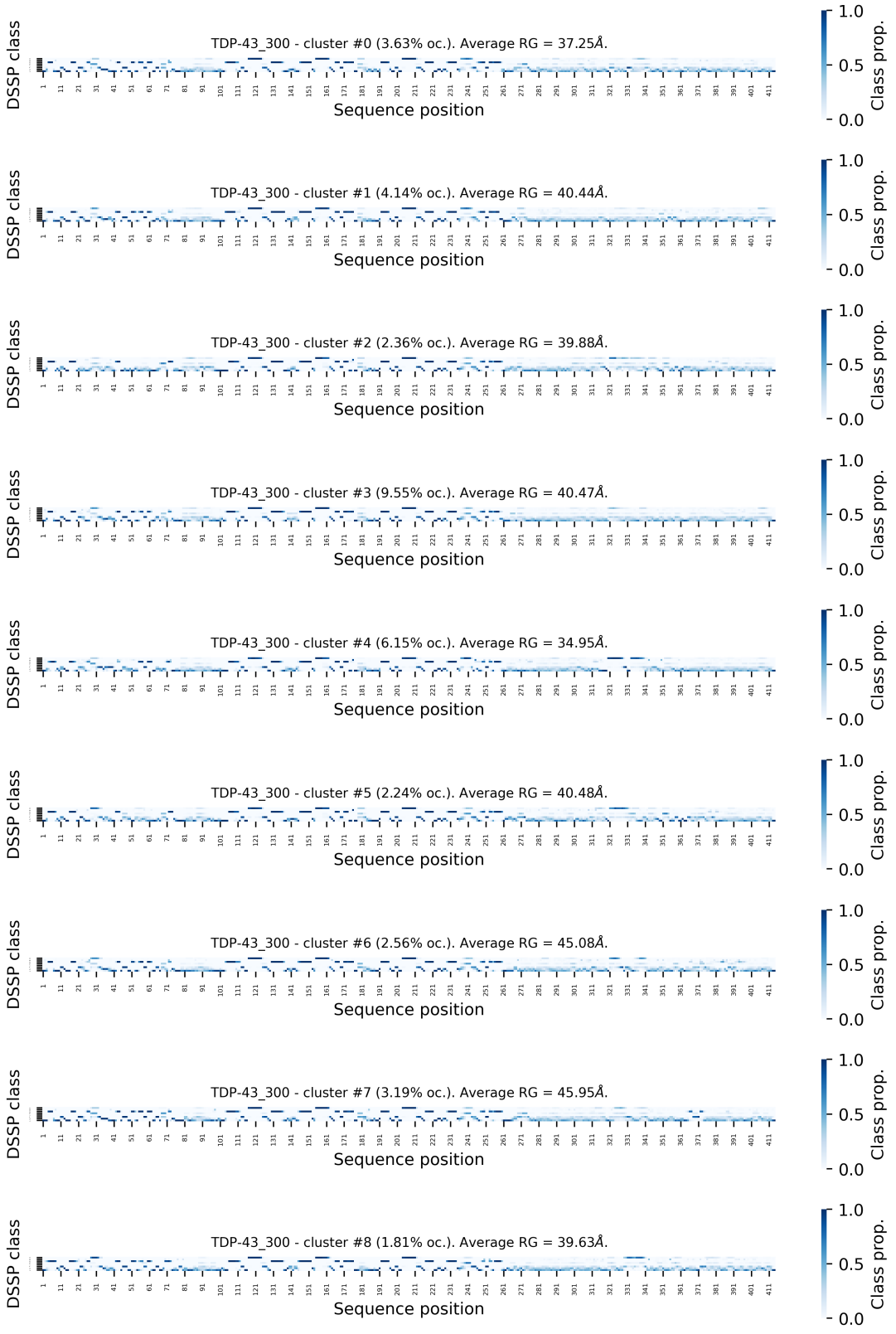

### *S4.4.2 Complete family of weighted ω-contact maps*

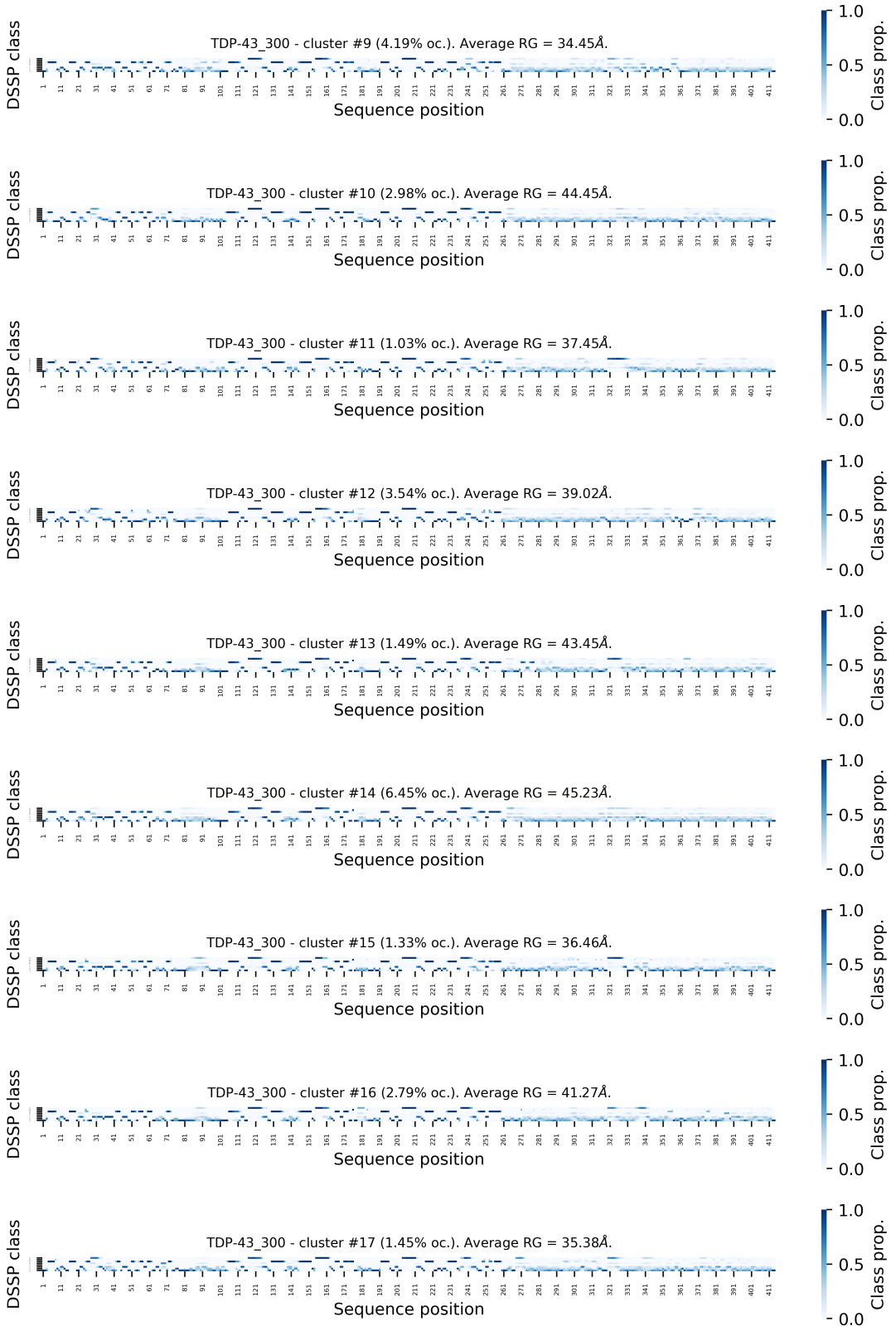


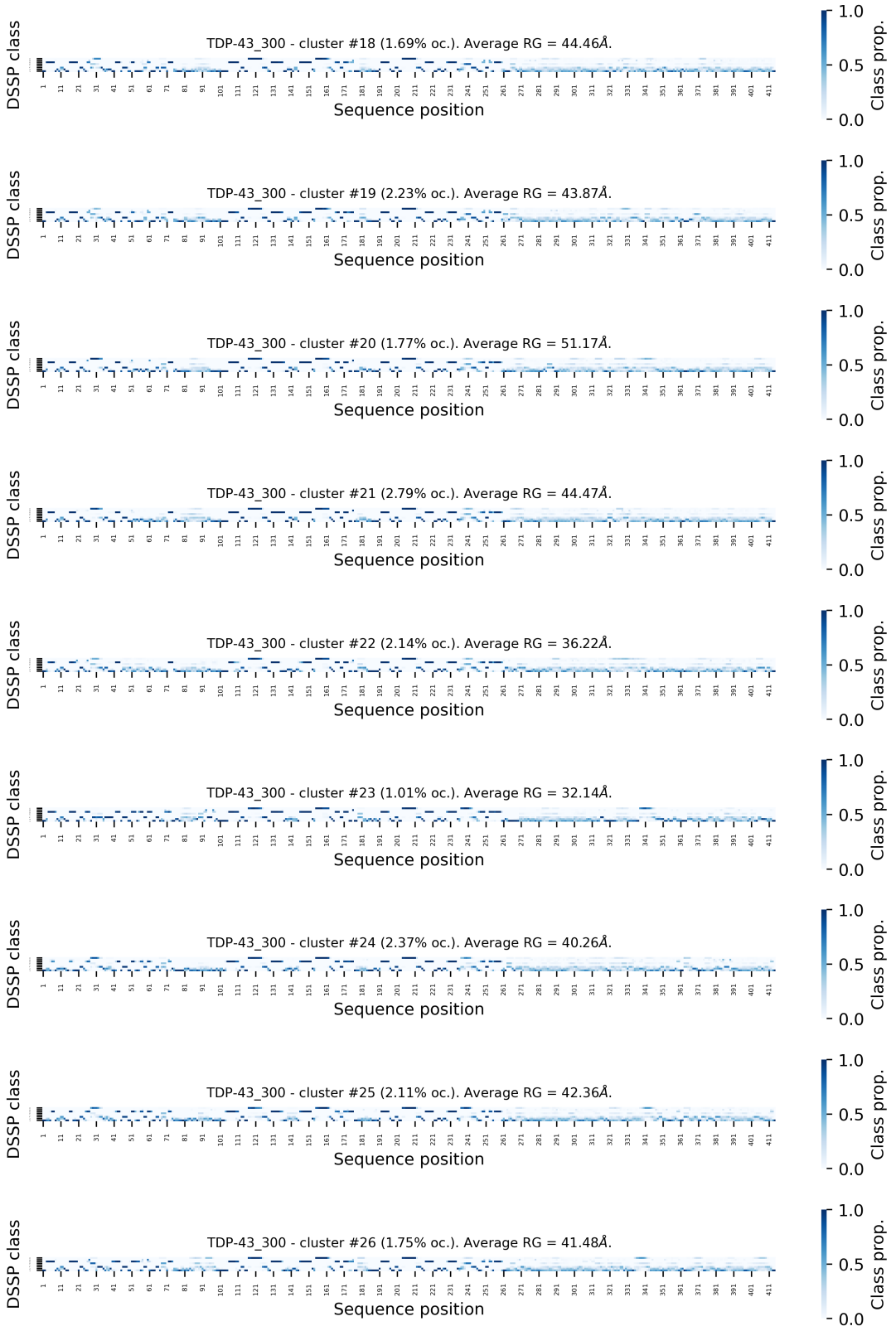

TDP-43_300 contact-based clustering
Cluster #9 with 4.19% of occupation
TDP-43_300 contact-based clustering
Cluster #10 with 2.98% of occupation
TDP-43_300 contact-based clustering
Cluster #11 with 1.03% of occupation
TDP-43_300 contact-based clustering
Cluster #12 with 3.54% of occupation
TDP-43_300 contact-based clustering
Cluster #13 with 1.49% of occupation
TDP-43_300 contact-based clustering
Cluster #14 with 6.45% of occupation
TDP-43_300 contact-based clustering
Cluster #15 with 1.33% of occupation
TDP-43_300 contact-based clustering
Cluster #16 with 2.79% of occupation
TDP-43_300 contact-based clustering
Cluster #17 with 1.45% of occupation
TDP-43_300 contact-based clustering
Cluster #18 with 1.69% of occupation
TDP-43_300 contact-based clustering
Cluster #19 with 2.23% of occupation
TDP-43_300 contact-based clustering
Cluster #20 with 1.77% of occupation
TDP-43_300 contact-based clustering
Cluster #21 with 2.79% of occupation
TDP-43_300 contact-based clustering
Cluster #22 with 2.14% of occupation
TDP-43_300 contact-based clustering
Cluster #23 with 1.01% of occupation
Sequence position
Contact weight average

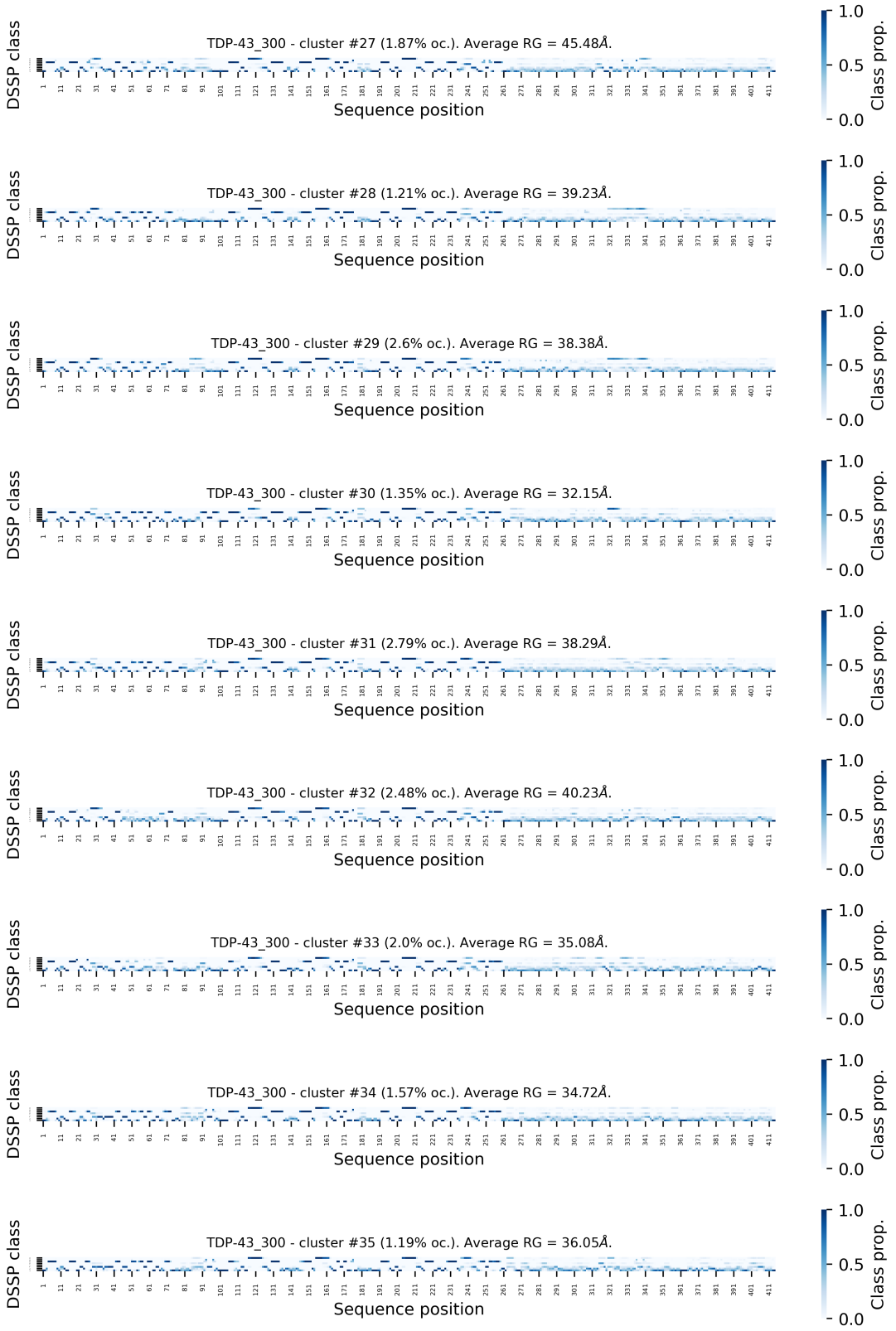

TDP-43_300 contact-based clustering
Cluster #24 with 2.37% of occupation
TDP-43_300 contact-based clustering
Cluster #25 with 2.11% of occupation
TDP-43_300 contact-based clustering
Cluster #26 with 1.75% of occupation
TDP-43_300 contact-based clustering
Cluster #27 with 1.87% of occupation
TDP-43_300 contact-based clustering
Cluster #28 with 1.21% of occupation
TDP-43_300 contact-based clustering
Cluster #29 with 2.6% of occupation
TDP-43_300 contact-based clustering
Cluster #30 with 1.35% of occupation
TDP-43_300 contact-based clustering
Cluster #31 with 2.79% of occupation
TDP-43_300 contact-based clustering
Cluster #32 with 2.48% of occupation
TDP-43_300 contact-based clustering
Cluster #33 with 2.0% of occupation
TDP-43_300 contact-based clustering
Cluster #34 with 1.57% of occupation
TDP-43_300 contact-based clustering
Cluster #35 with 1.19% of occupation
TDP-43_300 contact-based clustering
Cluster #36 with 1.3% of occupation
TDP-43_300 contact-based clustering
Cluster #37 with 1.29% of occupation
Sequence position
Contact weight average

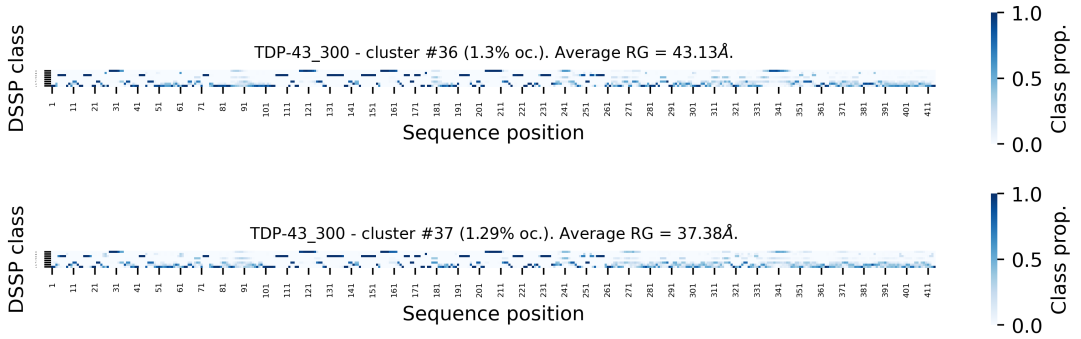
### *S4.4.3 Secondary structure propensities and average radii of gyration*




TDP-43_300 - cluster #9 (4.19% oc.). Average RG = 34.45Å.
TDP-43_300 - cluster #10 (2.98% oc.). Average RG = 44.45Å.
TDP-43_300 - cluster #11 (1.03% oc.). Average RG = 37.45Å.
TDP-43_300 - cluster #12 (3.54% oc.). Average RG = 39.02Å.
TDP-43_300 - cluster #13 (1.49% oc.). Average RG = 43.45Å.
TDP-43_300 - cluster #14 (6.45% oc.). Average RG = 45.23Å.
TDP-43_300 - cluster #15 (1.33% oc.). Average RG = 36.46Å.
TDP-43_300 - cluster #16 (2.79% oc.). Average RG = 41.27Å.
TDP-43_300 - cluster #17 (1.45% oc.). Average RG = 35.38Å.
DSSP class
Sequence position
Class prop.
1.0
0.5
0.0


TDP-43_300 - cluster #18 (1.69% oc.). Average RG = 44.46Å.
TDP-43_300 - cluster #19 (2.23% oc.). Average RG = 43.87Å.
TDP-43_300 - cluster #20 (1.77% oc.). Average RG = 51.17Å.
TDP-43_300 - cluster #21 (2.79% oc.). Average RG = 44.47Å.
TDP-43_300 - cluster #22 (2.14% oc.). Average RG = 36.22Å.
TDP-43_300 - cluster #23 (1.01% oc.). Average RG = 32.14Å.
TDP-43_300 - cluster #24 (2.37% oc.). Average RG = 40.26Å.
TDP-43_300 - cluster #25 (2.11% oc.). Average RG = 42.36Å.
TDP-43_300 - cluster #26 (1.75% oc.). Average RG = 41.48Å.
DSSP class
Sequence position
Class prop.
1.0
0.5
0.0


TDP-43_300 - cluster #27 (1.87% oc.). Average RG = 45.48Å.
TDP-43_300 - cluster #28 (1.21% oc.). Average RG = 39.23Å.
TDP-43_300 - cluster #29 (2.6% oc.). Average RG = 38.38Å.
TDP-43_300 - cluster #30 (1.35% oc.). Average RG = 32.15Å.
TDP-43_300 - cluster #31 (2.79% oc.). Average RG = 38.29Å.
TDP-43_300 - cluster #32 (2.48% oc.). Average RG = 40.23Å.
TDP-43_300 - cluster #33 (2.0% oc.). Average RG = 35.08Å.
TDP-43_300 - cluster #34 (1.57% oc.). Average RG = 34.72Å.
TDP-43_300 - cluster #35 (1.19% oc.). Average RG = 36.05Å.
DSSP class
Sequence position
Class prop.
1.0
0.5
0.0




## References


[1] J.-M. Chandonia, N. K. Fox, and S. E Brenner. SCOPe: classification of large macromolecular structures in the structural classification of proteins—extended database. *Nucleic Acids Research.*, 47(D1):D475–D481, 2018.

[2] M.A. Hakim Newton, Julia Rahman, Rianon Zaman, and Abdul Sattar. Enhancing protein contact map prediction accuracy via ensembles of inter-residue distance predictors. *Computational Biology and Chemistry*, 99:107700, 2022.

[3] José Ramón López-Blanco and Pablo Chacón. KORP: knowledge-based 6D potential for fast protein and loop modeling. *Bioinformatics*, 35(17):3013–3019, 2019.

[4] Rémi Zallot, Nils Oberg, and John A. Gerlt. The EFI web resource for genomic enzymology tools: Leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. *Biochemistry*, 58(41):4169–4182, 2019.

[5] Damiano Clementel, Alessio Del Conte, Alexander Miguel Monzon, Giorgia F Camagni, Giovanni Minervini, Damiano Piovesan, and Silvio C E Tosatto. RING 3.0: fast generation of probabilistic residue interaction networks from structural ensembles. *Nucleic Acids Researchearch*, 50(W1):W651–W656, 2022.

[6] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[7] Alberto J. M. Martin, Michele Vidotto, Filippo Boscariol, Tomàs Di Domenico, Ian Walsh, and Silvio C. E. Tosatto. RING: networking interacting residues, evolutionary information and energetics in protein structures. *Bioinformatics*, 27(14):2003–2005, 2011.

[8] Javier González-Delgado, Amin Sagar, Christophe Zanon, Kresten Lindorff-Larsen, Pau Bernadó, Pierre Neuvial, and Juan Cortés. Wasco: A wasserstein-based statistical tool to compare conformational ensembles of intrinsically disordered proteins. *Journal of Molecular Biology*, 435(14):168053, 2023. Computation Resources for Molecular Biology.

[9] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[10] Leland McInnes. How umap works, 2018. Accessed: 2023-07-21.

[11] Samuel Eilenberg and Norman Steenrod. *Foundations of Algebraic Topology*. Princeton University Press, 1952.

[12] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[13] Ruizhi Xiang, Wencan Wang, Lei Yang, Shiyuan Wang, Chaohan Xu, and Xiaowen Chen. A comparison for dimensionality reduction methods of single-cell rna-seq data. *Frontiers in Genetics*, 12, 2021.

[14] Alex Diaz-Papkovich, Luke Anderson-Trocmé, Chief Ben-Eghan, and Simon Gravel. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLOS Genetics*, 15(11):e1008432, 2019.

[15] Alex Diaz-Papkovich, Shadi Zabad, Chief Ben-Eghan, Luke Anderson-Trocmé, Georgette Femerling, Vikram Nathan, Jenisha Patel, and Simon Gravel. Topological stratification of continuous genetic variation in large biobanks, 2023.

[16] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W H Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1):38–44, 2018.

[17] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 2019.

[18] Benson Mwangi, Tian Siva Tian, and Jair C. Soares. A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, 12(2):229–244, 2013.

[19] Mebarka Allaoui, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study. In *Lecture Notes in Computer Science*, pages 317–325. Springer International Publishing, 2020.

[20] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172. Springer Berlin Heidelberg, 2013.

[21] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[22] Rajeswari Appadurai, Jaya Krishna Koneru, Massimiliano Bonomi, Paul Robustelli, and Anand Srivastava. Clustering heterogeneous conformational ensembles of intrinsically disordered proteins with t-distributed stochastic neighbor embedding. *Journal of Chemical Theory and Computation*, 2023.

[23] Anja Conev, Mauricio Menegatti Rigo, Didier Devaurs, André Faustino Fonseca, Hussain Kalavadwala, Martiela Vaz de Freitas, Cecilia Clementi, Geancarlo Zanatta, Dinler Amaral Antunes, and Lydia E Kavraki. EnGens: a computational framework for generation and analysis of representative protein conformational ensembles. *Briefings in Bioinformatics*, 24(4):bbad242, 2023.

[24] Leland McInnes. How hdbscan works, 2016. Accessed: 2023-07-21.