NAR Genomics and Bioinformatics, 2024, **6**, 1–17
https://doi.org/10.1093/nargab/lqae002
Advance access publication date: 29 January 2024
Standard Article

Analysis of the inter-domain orientation of tandem RRM domains with diverse linkers: connecting experimental with AlphaFold2 predicted models

Joel Roca-Martínez [1,2], Hyun-Seo Kang[3,4], Michael Sattler [3,4] and Wim Vranken [1,2,*]

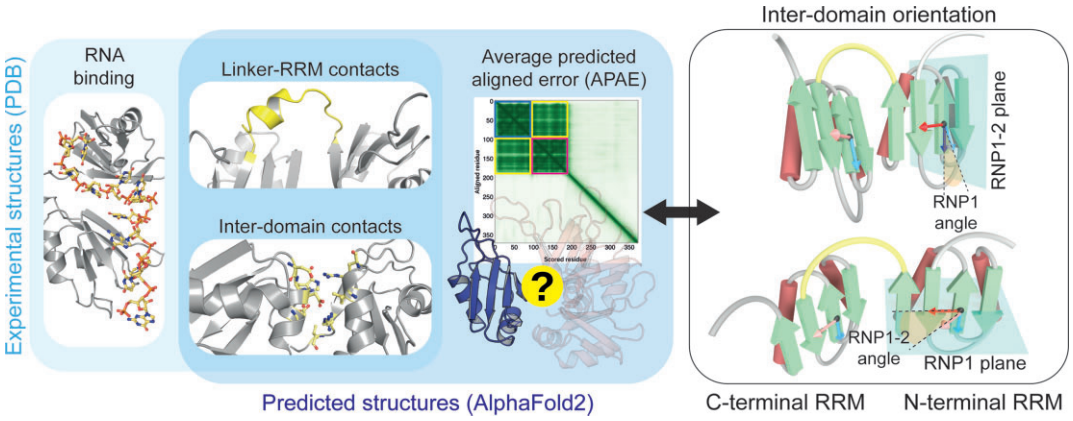
[1]Interuniversity Institute of Bioinformatics in Brussels, VUB/ULB, Brussels 1050, Belgium
[2]Structural Biology Brussels, Vrije Universiteit Brussel, Brussels 1050, Belgium
[3]Helmholtz Munich, Molecular Targets and Therapeutics Center, Institute of Structural Biology, 85764 Neuherberg, Germany
[4]Technical University of Munich, TUM School of Natural Sciences, Department of Bioscience, Bavarian NMR Center, 85747 Garching, Germany

[*]To whom correspondence should be addressed. Tel: +32 2 6291952; Email: wim.vranken@vub.be

Abstract

The RNA recognition motif (RRM) is the most prevalent RNA binding domain in eukaryotes and is involved in most RNA metabolism processes. Single RRM domains have a limited RNA specificity and affinity and tend to be accompanied by other RNA binding domains, frequently additional RRMs that contribute to an avidity effect. Within multi-RRM proteins, the most common arrangement are tandem RRMs, with two domains connected by a variable linker. Despite their prevalence, little is known about the features that lead to specific arrangements, and especially the role of the connecting linker. In this work, we present a novel and robust way to investigate the relative domain orientation in multi-domain proteins using inter-domain vectors referenced to a stable secondary structure element. We apply this method to tandem RRM domains and cluster experimental tandem RRM structures according to their inter-domain and linker-domain contacts, and report how this correlates with their orientation. By extending our analysis to AlphaFold2 predicted structures, with particular attention to the inter-domain predicted aligned error, we identify new orientations not reported experimentally. Our analysis provides novel insights across a range of tandem RRM orientations that may help for the design of proteins with a specific RNA binding mode.

Graphical abstract



Introduction

The RNA Recognition Motif (RRM) is the most frequently observed RNA-binding domain across all species, being particularly prevalent in eukaryotes where it plays a key role in post-transcriptional regulation processes (1). The canonical RRM fold has a conserved β1α1β2β3α2β4 topology, with an approximate length of 90 amino acid residues. RRMs are highly versatile proteins and many variations and extensions of the canonical fold where the β-sheet provides the main binding platform, are observed. These subfamilies can bind RNA in various ways (1). A single RRM domain recognizes a short stretch of RNA, 2–5 nucleotides (2), thus limiting its specificity and affinity. Therefore, to overcome this, RRMs are often observed in conjunction with other domain or RNA binding proteins (RBPs)(3, 4), adding an avidity effect for a stronger binding. In humans, around 47% of all the RRM-containing proteins have two or more of these domains, with up to 6 observed (5).

Received: August 16, 2023. **Revised:** December 7, 2023. **Editorial Decision:** January 3, 2024. **Accepted:** January 9, 2024

The relative orientation between two RRM domains affects how the RNA is recognized, as well as its binding kinetics, hence affecting their biological role and mode of action (6,7). There are several factors that affect the tandem arrangement, the main ones being (i) RRM–RRM contacts (also referred to as inter-domain contacts), (ii) linker-domain contacts, (iii) the linker length and (iv) RNA binding. These factors result in a broad range of stable tandem RRM arrangements. Connecting linkers play an especially important role in the overall tandem behavior. These linkers are often reported to regulate the RNA-binding activity of RRM domains by interacting with the RRM domains, providing additional RNA-binding interfaces, or defining the distance between the recognized target RNA motifs (8,9). RNA binding can also contribute to the final arrangement, influencing the RRMs to adopt specific orientations for effective binding (3,10).

Therefore, characterizing the orientation of two RRM domains with respect to each other is crucial to understand how they collectively bind RNA. Despite the wide biological implications of tandem RRMs, there is very limited structural information available. There are at the moment only 39 proteins with structures deposited in PDB with two or more RRMs, from which 26 have at least 1 domain bound to the RNA and only 17 have two or more domains bound.

In spite of the data scarcity on multi-RRM containing proteins, some efforts have been made to classify proteins with multiple RRMs based on their binding mode (11,12). In their free form, tandem RRMs can be independent from each other, e.g. Hrp1 (13), or have a pre-established contacts interface that keeps the domains in close proximity with a particular arrangement. In the latter case, the β-sheets can remain exposed for RNA binding, e.g. hnRNP A1 (14), or form inter-domain contacts between each other preventing RNA binding, which has been reported for U2AF2 as an autoinhibitory mechanism (10). Upon RNA binding, different arrangements are observed that can be categorized in 3 main groups;

i) Adjacent RRMs, where both RRMs bind to a continuous RNA stretch to achieve higher affinity and specificity than single RRMs (4). Slightly different RRM orientations are observed within this binding mode, ranging from an extended β-sheet to a 'closed' conformation where both β-sheets face each other surrounding the RNA. In this arrangement, the connecting linker often becomes rigid upon binding (11), as reported for Sex-lethal (15) (Figure 1A), Hrp1 (13), Nucleolin (16) and HuR (17), but not always, for example, the connecting linker in U2AF2 remains disordered in the bound form (10,18).
ii) RNA-looping RRMs already have a pre-formed arrangement stabilized by multiple interactions (contacts between the RRM domains or between the linker and the domains) in such a way that prevents a continuous RNA stretch to bind both domains simultaneously. This has been argued to be linked to splicing repression as it might force the RNA to loop to bind both domains. A well-characterized example is PTB RRM3-4 (19) (Figure 1B).
iii) Independent RRMs, in some cases RRMs that are independent in their free form may remain this way upon RNA binding. These RRMs can recognize distant RNA motifs, usually thanks to a long and flexible connecting linker. This allows to quickly scan long RNA sequences to find suitable binding sites. PTB RRM1-2 (19) is a clear representative of this group. Often a third RRM domain is flexibly connected to tandem RRM domains and may contribute to RNA binding (20,21) or even mediate protein-protein interactions (20).

This classification covers some of the most studied tandem RRMs but it is largely based on a qualitative observation of the experimental structures, lacking a concise description of their orientation and which are the driving features leading to each tandem arrangement. Moreover, due to the lack of data it is very likely that there are binding modes for which no experimental structure is yet available, as well as unknown binding mechanisms.

In our study, we provide a novel computational analysis to study domain orientations that can be applied to other multi-domain proteins, based on protein sequence alignment and the identification of conserved secondary structure elements that enable the definition of a stable intramolecular vector. We apply this method to tandem RRM domains and cluster all the experimental tandem RRMs according to their interactions, both inter-domain and with the linker, and report how that correlates with their orientation. By extending our analysis to AlphaFold2 (AF2)(21) predicted structures, with particular attention to the inter-domain predicted aligned error (PAE), we identify new orientations that have not yet been reported experimentally and broaden the sequence and structure space for already known arrangements.

Materials and methods

RRMs structural data collection

Experimental tandem RRM structures

We retrieved the experimentally solved structures for multiple RRMs from Inter3Mdb, a curated database that incorporates UniProt (22), PDB (23) and PFAM (24) as its primary sources of information. 727 PDB entries were retrieved from the database, accounting for 217 different proteins. Out of that set, 178 proteins contain a single RRM domain, while 39 contain at least 2 RRMs. After excluding proteins with missing linkers or where the RRM pair available was not consecutive (e.g. RRM1-3 or RRM2-4), 33 different proteins remained, accounting for 35 tandem RRMs, as two of them contained 4 RRM domains. The resulting sequences were aligned to the RRM master alignment, a previously published carefully curated alignment (25) (Supplementary Dataset S1).

AlphaFold2-predicted tandem RRM structures

To gather all the RRMs with at least one predicted structure available, we used hmmsearch (hidden Markov model search) using the HMMER software (http://hmmer.org - version (3.3.2) against UniRef90 (26), and then retrieved all the available predicted structures on the AlphaFold2 (AF2)(21) database. For the search we also used the forementioned master alignment (25). With this query, 163 982 proteins were identified containing from 1 to a maximum of 14 RRMs each, from which 47 144 proteins had at least 2 RRMs. Each single domain was then aligned to the master alignment. After removing the sequences that were nor properly fit into the alignment, (i.e. sequences introducing gaps in any of the key secondary structure elements) 33 549 proteins containing at least 2 RRMs remained (Supplementary Dataset S2).

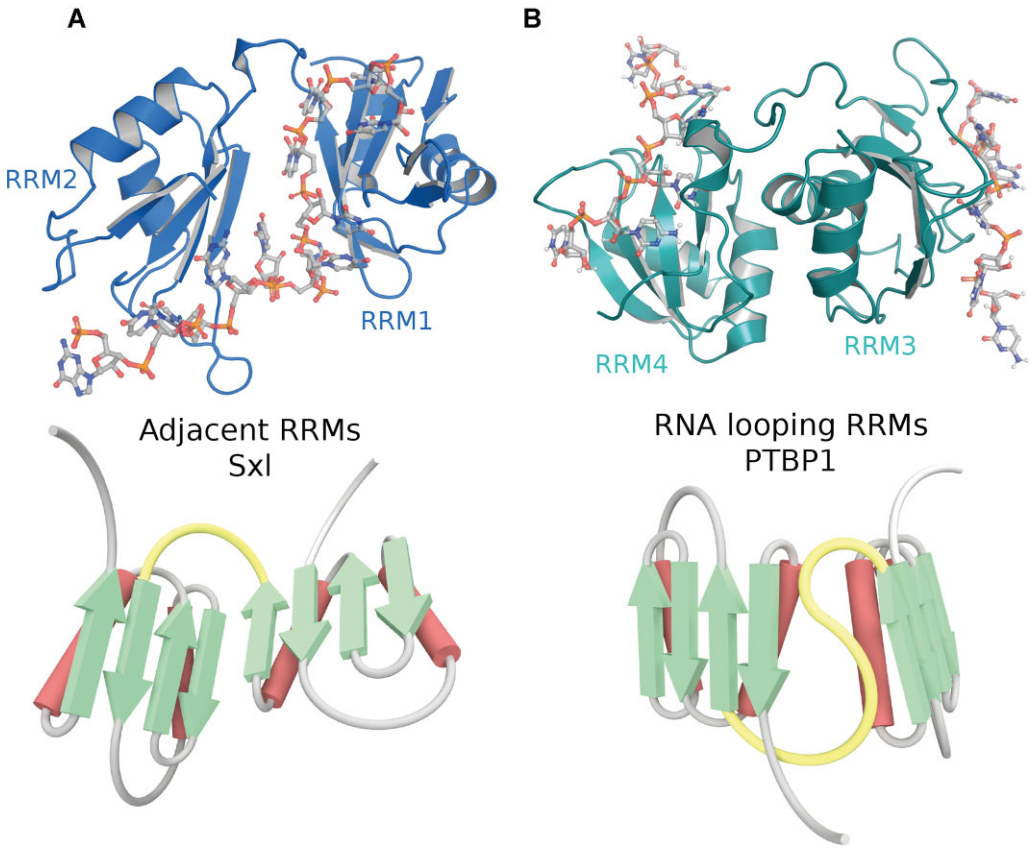


Figure 1. Top, representative tandem RRMs for the two commonly observed arrangements: (**A**) adjacent, Sxl (PDB Id: 1B7F) and (**B**) RNA looping, PTBP1 (PDB Id: 2ADC). Bottom, schematic representation of the tandem relative orientations.

Inter-domain predicted aligned error calculation

The predicted aligned error (PAE) is one of the two main output confidence metrics of AF2. It indicates the model's confidence on the relative position between any pair of residues within the structure. It is measured in angstroms (Å) and is not perfectly symmetric, i.e. the PAE between a residue in position X and a residue in position Y is different depending on which residue is taken as reference in the alignment.

This metric is particularly useful in our study where we defined an average predicted aligned error (APAE) as a measure of confidence for the relative position of multiple RRMs in a predicted tandem RRM structure. Based on the master RRM alignment we extracted the RRM core domain boundaries. We then calculated the APAE for the residues that align one domain against the other considering its asymmetry (Equation 1). The APAE among all the identified tandem RRM domains in the AF2 database was stored in a JSON file (Supplementary Dataset S3) for downstream analysis. We labelled as high confidence models any tandem RRM with an APAE lower than 10 Å, resulting in a set with 7080 tandem RRM structures.

Equation (1): Equation to calculate the inter-domain APAE between two domains. N and M are the number of residues in the N-terminal and C-terminal RRM domains, respectively, with i and j the residue positions for each.

$$Average\ interdomain\ PAE = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(PAE_{i,j} + PAE_{j,i}\right)}{2{*}N{*}M} \quad (1)$$

Experimental-AlphaFold2 tandem RRMs assessment

We combined all the tandem RRMs for which we had experimental and AF2 structures available, resulting in 22 protein pairs. We performed a structural alignment between those 22 experimental tandem RRMs and their AF2 counterparts to correlate the APAE and the RMSD between the structures. Both bound and unbound states of the experimental structures were compared separately to their AF2 models to assess the 'preferred' orientation. The RRM domain limits were identified based on the sequence alignment, and using the region from the first residue of the RRM1 β1 to the last residue of the RRM2 β4. The RMSD is calculated selecting the Cα of both structures and using the align feature from the MDAnalysis Python library (27).

Tandem RRM orientation

To make a large-scale assessment of the tandem RRM domain-orientation for both the experimental and AF2 models, we selected three highly conserved positions in stable secondary structure elements, two in RNP1 (β3) and one in RNP2 (β1). Those residues often include aromatic side chains in canonical RRMs and have a crucial role stacking the RNA nucleotide bases. They are located in the core of the β-sheet, and due to the high structural similarity of the core RRM domain, their respective position is conserved among different RRMs.

Two vectors were defined for each of the RRM domains, one of the vectors was defined between the positions β3–1 and β3–3 (RNP1 vector) and the other one between β3–3 and

β1–3 (RNP1–2 vector). We used them to define the RNP1 and RNP1–2 angles, which capture the two main rotation axes of the domains. The vectors were defined from Cα to Cα using the coordinates in the PDB file. To assess the domain orientation with a 360° range, we defined the RNP1 and RNP1–2 planes using the RNP1 and RNP1–2 vectors as the normal vector of the planes, respectively, and the RNP1 central position as a point of the plane. The planes are defined for the N-terminal RRM and the vectors for both domains are normalized and projected to it. This is analogous to define a plane for the C-terminal RRM β-sheet (referred to as the β-sheet plane) and calculating the intersection angle with the N-terminal RNP1 and RNP1–2 planes. By projecting the vector instead, we capture the same angle while also keeping track of the directionality, thus allowing a comprehensive 360° determination of the angle.

To determine the angle we compute the arctangent between each pair of projected and normalised vectors (Equation 2). This produces a value ranging from $-\pi$ radians ($-180°$) to π radians (180°). The same equation applies for both the RNP1 and RNP1–2 angles. We used SymPy (28) to define the plane equations and operate with the vectors, which is a python library for symbolic computation.

Equation (2): Equation to calculate the arctangent of the angle between two vectors. The RRM1 and RRM2 vectors represent the N-terminal and C-terminal vector for RNP1 or RNP1–2 vectors, depending on the calculated angle. The N vector represents the normal vector of the plane (nonzero vector orthogonal to the plane). The cross and asterisk denote the cross and dot product of the vectors, respectively.

$$arctan\,(angle) = \left(\left(\overrightarrow{RRM1}\ X\ \overrightarrow{RRM2},\right) * \vec{N}\right) * \overrightarrow{RRM1} * \overrightarrow{RRM2} \quad (2)$$

Inter-domain and linker-domain contacts analysis

Based on the RRM sequence alignment, we defined the RRM core region for each domain from the first β1 position to the last β4 position. Linker, C- and N-terminal regions were excluded from this selection to limit the inter-domain contacts to the core RRM domain. We computed the inter-domain contacts using a distance threshold of 6 Å among the Cβ atoms of all included residues. We used Biopython (29) to handle the structures and define the contacts. The identified interactions are available as Supplementary information in Supplementary Datasets S4 and S5, for the experimental tandem RRMs and AF2 predicted structures, respectively.

Similar to the RRM core region definition, the linker region connecting two RRM domains was defined based on the RRM sequence alignment. The inter-domain linkers are predominantly disordered and highly variable in both sequence and length, to align them we arranged the linker residues to connect to the closest fixed secondary structure element, so placing any gaps in the middle of the alignment. With this procedure (sequence squeezing), the position of the linker residue in the multiple sequence alignment gives more insights into its role in protein function in relation to the overall fold. The linker alignments are available as Supplementary information in Supplementary Datasets S6 and S7, for the experimental and the high confidence AF2-predicted tandem RRMs, respectively.

Tandem RRM clustering with *k*-means

We labelled the tandem RRMs with the inter-domain and linker-domain contacts data from the previous analysis. For the experimental structures dataset with 188 tandem RRM instances (184 PDB structures), 381 binary features (contacts) were defined, 271 inter-domain contacts and 110 linker-domain contacts. We assigned a value of 1 if the contact is present and 0 otherwise. For the AF2 set we defined 1425 features for the 7080 high confidence models analysed, 478 being inter-domain contacts and 947 linker-domain contacts. The increased number of features already shows we are exploring a larger spectrum of RRM orientations (Table 1).

Using *k*-means we clustered the structures and visualized them in the context of the RNP1 and RNP1–2 angles. *k*-means is an unsupervised algorithm that clusters the entries in an arbitrary number of clusters based on the mean values of the clusters' features. To select an appropriate number of clusters for each dataset (experimental and predicted structures), we calculated the inertia, i.e. value that measures how well a dataset is clustered by *k*-means, from 1 to 50 clusters. We used the elbow method to find the optimal number of clusters, the elbow point is the number of clusters from which the inertia decrease begins to slow. For the selected number of clusters, we analysed the feature importance. We used the scikit-learn (30) implementation of the *k*-means algorithm for both the cluster generation and feature importance calculation.

To determine the feature importance we computed the absolute difference between the centroid of each variable and cluster to the overall means of that variable. The bigger the difference, the higher the importance of that variable to separate the entries in that cluster from the rest. We chose this simple method for a better interpretability. As we work with binary features representing presence/absence of contacts, a big difference among the variable centroids implies an overrepresentation of a specific contact in the cluster. Contacts with a difference over 0.5 (present in more than half of the structures in the clusters) are labelled as key contacts.

Angles and contacts cluster comparison

To assess the similarities among the experimental and AF2 generated clusters, we made pairwise comparisons among all the clusters based on the inter-domain angles and the key contacts stabilizing such orientations.

Inter-domain angles

To evaluate the angle similarity we generated kernel density estimates (KDE) for all the AF2 clusters using the RNP1 and RNP1–2 angles. We then calculated the average density of all the experimental tandem RRM clusters against each of the cluster-KDEs. The higher the density obtained implies a larger overlapping among the observed angles in the clusters. We used the kernel density implementation of scikit-learn (30) to generate the KDEs and to score the experimental clusters against them. For the KDE generation we used a bandwidth of 20, consistent with the angle calculation robustness (Supplementary Figure S1), further discussed in the inter-domain orientation results section.

Key contacts

To compare the key contacts among the clusters, we labelled the clusters with the secondary structure elements that participate in the interactions (e.g. N-terminal RRM β2 interacts

Table 1. Summary of the number of structures, contacts and predicted clusters in the experimental and AF2 tandem RRM sets

	Experimental tandem RRMs	AF2 tandem RRMs
Tandem structures	188	7080
Unique tandem RRMs	35	7080
Inter-domain contacts	271	478
Linker-domain contacts	110	947

The experimental and AF2 sets are analysed independently and then compared.

with C-terminal RRM β4). To map the key contact positions to the secondary structure elements we used the experimental and AF2 alignments (Supplementary Datasets S1 and S2, respectively). The linker is just taken as whole for this analysis. With both experimental and AF2 clusters featurized, we computed the pairwise Jaccard index among them. This raises a value that ranges from 0 when there is not any shared interaction between the clusters, to 1 when all the interactions are the same.

Gene ontology analysis

We performed an exploratory gene ontology (GO) enrichment analysis of the AF2 clusters to connect the observed orientations with specific molecular functions and biological processes. We used the 33 549 identified tandem RRMs from the AF2 database hmmsearch as the population set. Then we used the goatools python library (31) to identify the enriched GO terms for each of the clusters using the default parameters but with the Bonferroni correction to only keep the most relevant terms (P-value = 0.05).

Results

RRMs structural data

The tandem RRM structural data consists of two main datasets. The first is an experimental dataset with all the structures available from the PDB in June 2023, totalling 184 PDB structures comprising 35 unique tandem RRMs (Figure 2A). There is a varying number of structures available per tandem RRM, ranging from a single PDB entry to 35 entries for the most studied one (U2AF2). Similarly, we gathered all the all the available predicted structures in the AlphaFold2 (AF2)(21) database, 163 982 proteins containing from 1 to a maximum of 14 RRMs each. We selected only proteins with at least 2 RRM domains (47 144 entries), and then we filtered out the ones that could not be properly aligned to our master alignment (see Materials and methods for details), resulting in 33549 aligned proteins (Figure 2A). We retrieved the last version of the AF2 models for each of those proteins.

Filtering AlphaFold2 tandem RRM structures

To assess the confidence of the AF2 tandem RRM structures regarding the domain's relative position, we employed the predicted aligned error (PAE) from each AF2 model. PAE informs on the accuracy of the relative position across all the residue pairs in the AF2-predicted structures. This metric is provided in angstroms and ranges from 0 to 31.75 Å, for the highest to the lowest possible confidence, respectively. We used this data to estimate the error between the relative position of the tandem RRM domains and assess to which extent we can rely on each AF2 model. A low average predicted aligned error (APAE) value between two RRM domains means that AF2 predicts the relative position between the domains with high confidence, while higher PAE values have the opposite interpretation (Figure 3).

To ensure we only analysed high confidence AF2 models regarding the tandem RRMs orientation, we investigated the inter-domain APAE for the 33549 aligned tandem RRMs as detailed in the materials and methods section. We observed that most of the structures have an APAE over 15 Å while only 1.6% (660 entries) had an APAE under 5 Å (Supplementary Figure S2A). We defined as high confidence models all the tandem RRMs with an inter-domain APAE below 10 Å. A 5 Å threshold in APAE should have a minimal impact on inter-domain angles. It is broadly accepted as good resolution at the residue level in structural biology, and it is also the distance threshold used for defining contacts (between heavy atoms) in state-of-the-art protein complex predictors such as AlphaFold-Multimer (32). In contrast, a larger error of up to 10 Å could significantly affect our analysis. Therefore, for the models with an APAE less than 10 Å, we excluded loops from the APAE calculations, as these regions are dynamic but not relevant for how precisely the orientation of the structured regions between the domains is defined. After this recalculation, the majority of models have an APAE between 4 and 7 Å (Supplementary Figure S2B), which should not significantly affect the inter-domain angle either.

Furthermore, we observed that many tandem RRMs are still very well predicted in this range when compared with the experimental structures (Supplementary Figures S3 and S4), despite being a relatively large distance. Using an APAE threshold of 10Å, 7080 entries remained (21% of the tandem RRMs), which were further characterized using the inter-domain and linker-domain contacts data, which we refer as high confidence dataset (Supplementary Dataset S5). Tandem RRMs showing an APAE higher than 10Å are not considered in our analysis. The high APAE indicates that there is no evolutionary information to define any interaction between the domains, suggesting these are tandem RRMs that only show transient interactions, whose orientation is highly RNA-dependent, or that they bind the RNA independently.

A single protein can contain several tandem RRMs with different APAE values. A clear example is the human PTB, with APAE values of 26.8 and 6.9 Å for the RRM1-2 and RRM3-4 pairs, respectively. This agrees with the structural evidence which suggests that RRMs 3 and 4 bind RNA cooperatively while the other RRM domains act independently (19).

Experimental structures versus their AlphaFold2 models

A recurrent question after the AF2 release is whether it predicts the active or inactive state of proteins, which in our case translates to the bound or unbound conformation. The RRM domains are connected by a flexible linker, which allows the domains to rearrange upon binding. Therefore, understanding

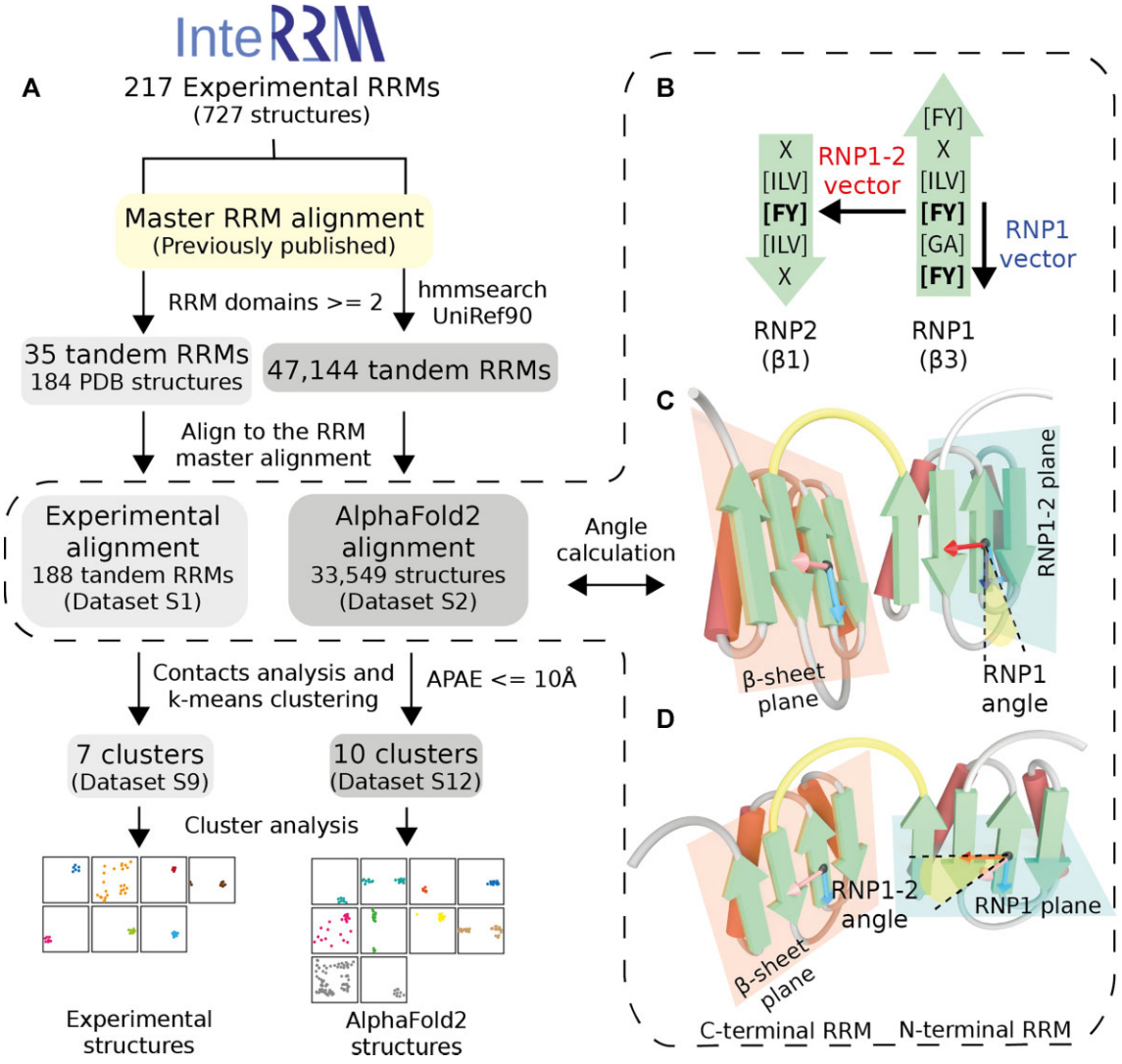


Figure 2. Data processing and analysis workflow for the experimental (**A**, left) and AlphaFold2 structures (A, right) and a schematic representation of the interdomain angle calculation, with three selected positions in RNP1 and RNP2 defining the RNP1 (blue) and RNP1–2 (red) vectors (**B**). The RNP1 angle is analogous to the intersection angle between the RNP1–2 plane (blue, N-terminal RRM) and the β-sheet plane (orange, C-terminal RRM) (**C**), while the RNP1–2 angle representation is analogous to the intersection angle between the RNP1 plane (blue, N-terminal RRM) and the β-sheet plane (orange, C-terminal RRM) (**D**).

which state AF2 is predicting and whether it does so consistently, is an intriguing question, as it might enable a better assessment of the likely tandem RRM orientation, for which we do not have an experimental reference. This is especially relevant as the co-evolutionary information driving AF2 predictions might capture relevant inter-domain contacts essential for the RRM orientation.

We aligned all the experimental tandem RRM structures to their AF2 counterparts and calculated their RMSD, and focused on the tandem proteins from which we have both bound and unbound structures available. We identified 5 proteins with the two domains bound to RNA and also an unbound structure, hnRNP A1, Sxl, U2AF2, PTBP1 (RRM3-4) and HuR, and compared their RMSD with the inter-domain APAE from the corresponding AF2 model (Figure 4). In all the proteins the AF2 models tend to be closer to the bound conformation, being particularly evident for HuR, Sxl and PTBP1, with an RMSD lower than 3 Å whilst their unbound structures have an RMSD of 7, 14.5 and 18–19 Å, respectively. The stable orientation of the hnRNP A1 RRMs (33) is again reflected on the low RMSD observed for both their bound and unbound structures, which AF2 confidently predicts as noted on the low inter-domain APAE of 6.4 Å. Finally, a large variability is observed with U2AF2, the unbound structures always present a high RMSD (>15.5 Å) but the bound cases are still very variable with RMSD values ranging from 0.0 Å (e.g. PDB Id 6XLW_A) to over 15 Å (e.g. PDB Id 3VAF_B), proving the high conformational dynamics of this tandem RRM (10,34).

We observed similar patterns between the half-bound structures and their unbound counterparts, when the RRM1 (Supplementary Figure S3A) or the RRM2 (Supplementary Figure S3B) are bound to the RNA. Finally, we also studied the correlation between the RMSD and the APAE for any type of bound structure (both or single domains) and unbound structures independently, therefore including the 22 tandem RRMs present in both the experimental and the AF2 sets. Entries with lower APAE values tend to be closer to their actual experimen-

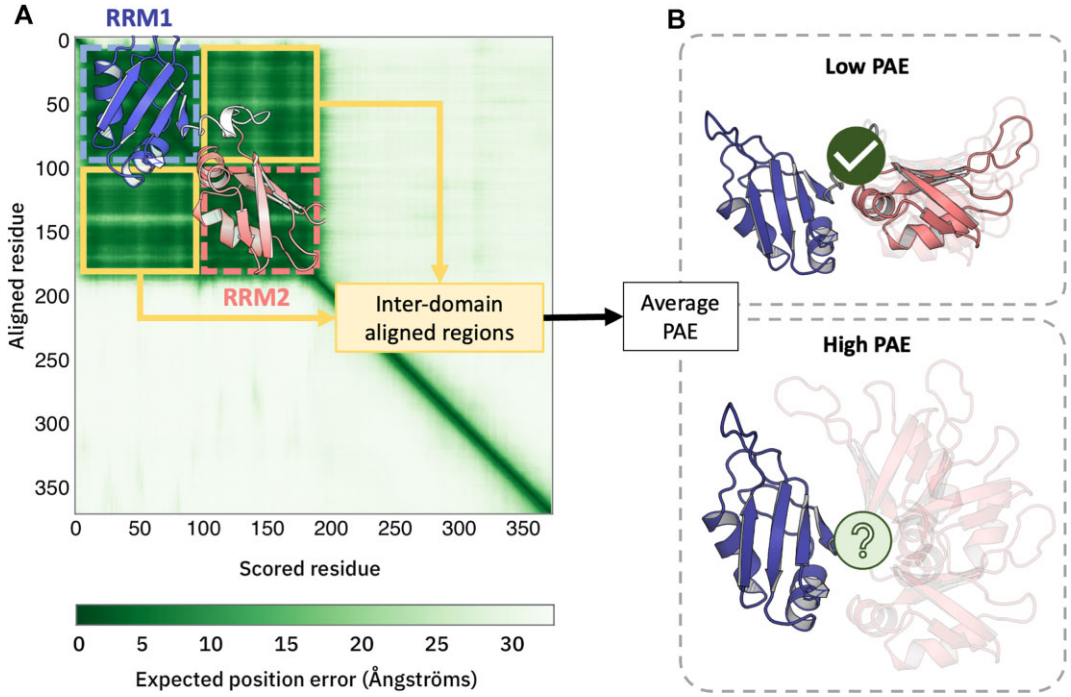


Figure 3. Predicted aligned error (PAE) interpretation for a tandem RRM example (HNRNP A1, UniProt Id: P09651). (**A**) The RRM1 and RRM2 are represented in blue and salmon on top of their approximate positions in the PAE plot. The regions where both domains are aligned are highlighted in yellow, informing on the expected error in their relative orientations (disordered region for residues 200–370 is not represented for simplicity). (**B**) Simplistic representation for the PAE interpretation between two RRM domains.

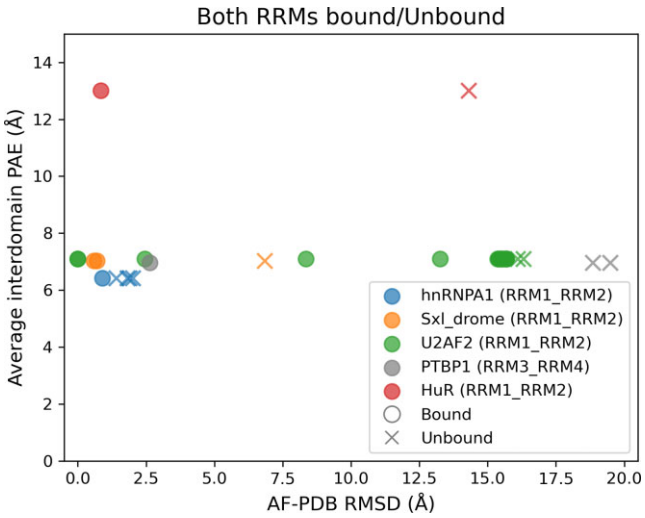


Figure 4. RMSD and APAE values between the experimental tandem structures and their AF2 models. All the available structures in PDB are represented for hnRNP A1 (blue) Sxl, (orange) U2AF2 (green), PTBP1 (grey) and HuR (red) with its bound state depicted as a circle (bound) or an X (unbound).

tal structures in both their bound (Supplementary Figure S4A) and unbound forms (Supplementary Figure S4B). The APAE and RMSD values among all the PDB structures and their corresponding AF2 models are available as Supplementary information (Supplementary Dataset S8).

This comparison between the experimental structures and their AF2 predicted structures revealed that the models tend to capture the RNA bound conformation. This could be argued as a consequence of a bias in the data towards bound complexes, but the tandem RRM search over PDB retrieved more unbound domains than bound ones. Despite this finding, the experimental data available is limited and we cannot assume that it is a generally applicable principle.

Inter-domain orientation

The respective orientation of the domains in tandem RRMs has been shown to be crucial for understanding their functional implications (7). To robustly characterize the inter-domain orientation between RRMs we introduced two inter-domain vectors referenced to stable secondary structure elements (Figure 2B). This allows us to compute the RNP1 (Figure 2C) and RNP1–2 (Figure 2D) angles that capture the two main rotation axes to describe the tandem RRM orientation in a simplified manner. The angles range from −180° to 180° (Figure 5A). To verify the robustness of the selected positions and the derived vectors, we calculated the angle between the RNP1 and the RNP1–2 vectors across all the individual domains in the dataset, observing that for the vast majority of the structures the angle ranges from 60° to 80°, with an average and standard deviation of 72° and 6°, respectively (Supplementary Figure S1).

Experimental tandem RRMs

We analysed all available PDB structures containing tandem RRMs to understand possible domain orientations, as they are often intrinsic to the RNA binding process. For proteins containing 4 RRM domains (yeast PABP and PRP24), the RRM1-2 and RRM3-4 pairs are analysed individually.

The identified 35 tandem RRMs, with 188 tandem structures, were split into 4 different categories based on which domains are bound to RNA (Figure 5B). There are 16 proteins with both RRMs bound to RNA (58 structures), 6 with

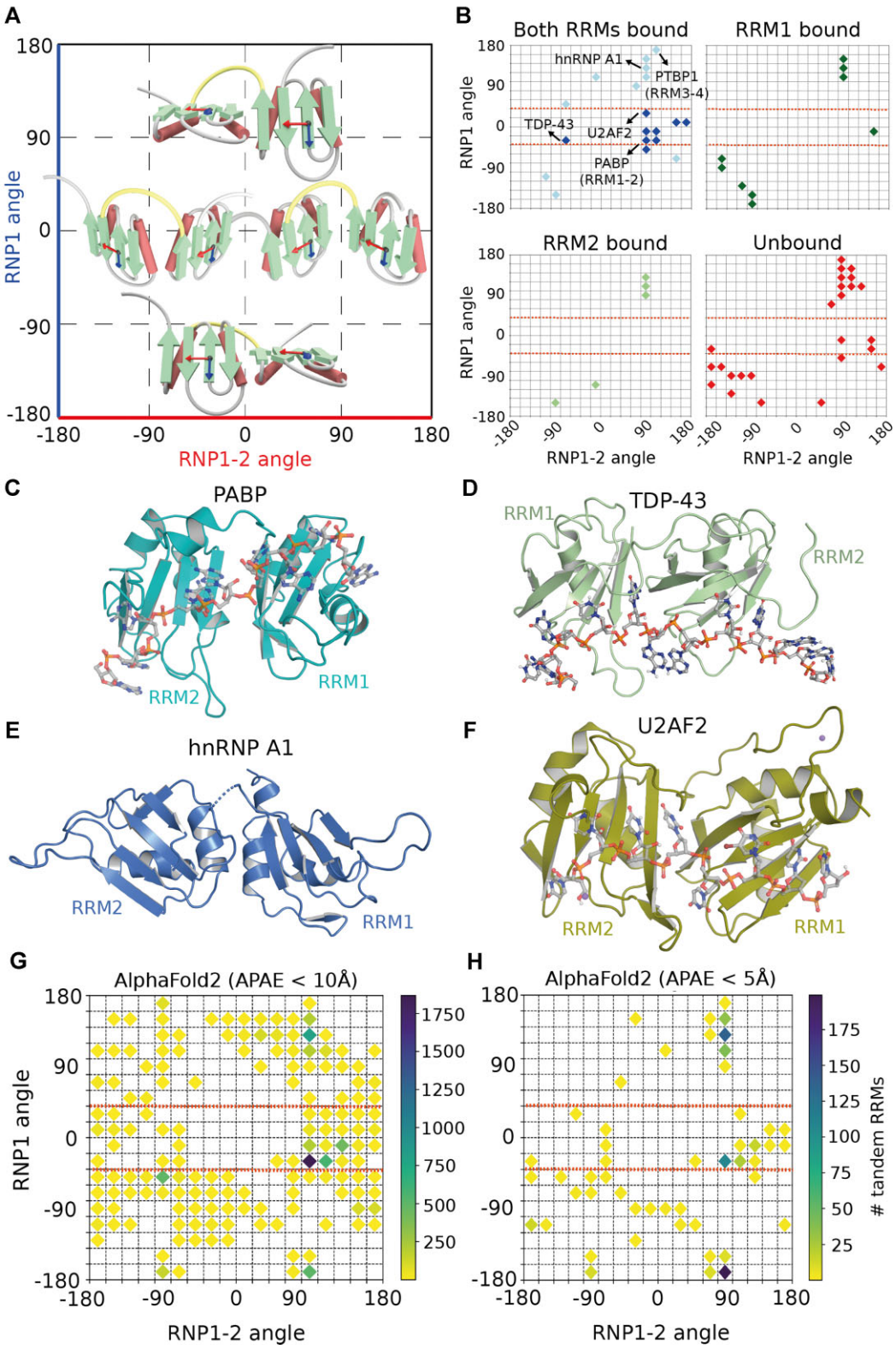


Figure 5. Tandem RRM orientations across experimental and AF2 structures. (**A**) Schematic representation for the RNP1 (blue) and RNP1–2 (red) angle interpretation for tandem RRMs. (**B**) RNP1 and RNP1–2 angles for the experimental tandem RRMs. The proteins are divided based on the RNA bound state; both domains bound (top-left, 16 proteins, bound to continuous and discontinuous RNAs in dark blue and light blue, respectively), only N-terminal RRM bound to RNA (top-right, 6 proteins, dark green), only C-terminal RRM bound to RNA (bottom-left, 5 proteins, light green), unbound RRMs (bottom right, 21 proteins, red). Some well-studied tandem RRMs are labelled. Experimental structures for (**C**) PABP (PDB Id: 1CVJ), (**D**) TDP-43 (PDB Id: 4BS2), (**E**) hnRNP A1 (PDB Id: 1UP1), (**F**) U2AF2 (PDB Id: 5EV1). Distribution of the RNP1 and RNP1–2 angles for the 7080 AlphaFold predicted tandem RRMs with an inter-domain APAE lower than 10 Å (**G**) and the 660 models with an APAE lower than 5 Å (**H**). The orange dashed lines in panels B, G and H delimit the 40 to –40 RNP1 angle range.

only the N-terminal RRM bound (40 structures), 5 with the C-terminal RRM bound (12 structures) and 21 in the free form (79 structures). The angle values are shown in ranges of 20º for a clearer depiction and generalization of the data, and consistent with the angle robustness. Regions with a data point on them can contain one or more structures sharing the same orientation, and proteins with several solved structures can also fall in different regions when showing different orientations.

When both RRMs are bound to a continuous RNA stretch there is a clear restriction on the possible RNP1 and RNP1–2 angles (dark blue data points in Figure 5B). This means that the β-sheet of both RRMs are approximately parallel (RNP1 angle values from −40º to 40º, delimited with an orange dashed line in Figure 5B) and form a cleft with a variable RNP1–2 angle that ranges from 90 degrees in the most closed conformations to almost 180º when the β-sheets of both RRMs are completely extended next to each other. A representative structure of this category is PABP RRM1-2 (Figure 5C). The only exception to this rule is TDP-43 (Figure 5D) with RNP1 and RNP1–2 angles of −25.2º and −63.0º, respectively. Notably, this tandem RRM is the only one in the dataset (bound to a continuous RNA stretch) where the N-terminal RRM binds to the 5′ end of the RNA and the C-terminal RRM to the 3′ end (35). The distinct RNA directionality requires a different tandem arrangement with both RRMs forming an extended β-sheet, but inverted with respect to more common arrangements like PABP RRM1-2. Proteins in which both domains are bound to different RNA fragments exhibit a larger variability on the RNP1 and RNP1–2 angles (light blue data points in Figure 5B). One of the well-studied cases is PTBP1 RRM3-4 (Figure 1B). This orientation is stabilized by an extensive network of inter-domain contacts that remains upon RNA binding (19,36).

For the cases where only one of the RRMs is bound to the RNA, or unbound, we did not observe any clear pattern. On the free RRMs (Figure 5B, bottom-right), more orientations are explored making evident the flexibility of the inter-domain linker allowing a broad range of arrangements. Some orientations seem to be preferred, especially the regions with RNP1 and RNP1–2 angles around 100º and 120º, respectively (e.g. hnRNP A1, Figure 5E). However, this could be just due to the limited availability of tandem RRM structures. The vector values for all the studied entries, alongside with its bound state, is available as Supplementary material (Supplementary Dataset S9).

To further explore the effect of RNA binding on RRM domain rearrangements, we selected the 6 tandem RRMs with both bound and unbound structures available, hnRNP A1, Sxl, U2AF2, PTBP1, HuR and RBM45. There are clear domain rearrangements in half of them, denoted by the RNP1 and RNP-2 angle changes for Sxl, U2AF2 and elav1/HuR (Supplementary Figure S5). These proteins show flexible linkers of different lengths, ranging from around 10 residues in HuR and Sxl, to over 30 in U2AF2, and their rearrangement upon binding have already been shown experimentally. In HuR the conformational changes upon binding also induce contacts between the linker and the RNA (17). A similar behavior is observed in Sxl (15). However, U2AF2's rearrangement is more complex, involving an equilibrium among multiple conformations and a self-inhibiting role of the linker that prevents non-specific RNA binding (34). Contrarily, hnRNP A1, PTBP1 RRM3-4 and RBM45 show little to no change in the domain's orientation upon binding (Supplementary Figure S5). The β-sheets from PTBP1 RRM3-4 and RBM45 adopt a conformation that prevents a continuous stretch of nucleotides to bind both domains simultaneously.

AF2 tandem RRMs

Following the same procedure as in the experimental tandem RRMs, and based on the AF2 tandem RRMs alignment to the master RRM alignment, we calculated the RNP1 and RNP1–2 angles (see materials and methods section for details). We investigated the distribution of the RRMs orientation for two different subsets, depending on the inter-domain average predicted aligned error (APAE). The larger set consists of 7080 structures with an APAE < 10 Å (Figure 5G), and the smaller set includes 660 structures with an APAE <5 Å (Figure 5H). Many new orientations are explored in comparison to the experimental set, revealing new RNP1 and RNP1–2 angles. The most represented arrangements will be discussed in detail in the clustering section. However, some orientations are much more preferred than others, particularly those with an RNP1–2 angle of approximately 90º. This angle indicates a tendency for RRM domains to create a cleft at a right angle, facilitating RNA binding (similar to U2AF2 bound form, Figure 5F). Despite fewer tandem orientations being captured within the highest confidence set (APAE < 5 Å), a similar pattern is observed, with several angle combinations that were not observed in the 35 analysed experimental tandem RRMs (Figure 5B and H).

Contacts analysis

Experimental tandem RRMs

To characterize the interplay between the RRMs in all the experimental tandem RRM structures, we analysed the number of contacts between the RRM domains (inter-domain contacts), and between the connecting linker and the domains (linker-domain contacts). Both the presence and absence of contacts provide relevant information on how the tandem RRMs can bind the RNA. Clearly defined and conserved RRM interfaces such as in PTB RRM3-4 (19,36) fix the orientation between the domains and force the protein to interact with RNA motifs at a specific distance and adopt a specific topology. Other tandem RRMs lack such contacts and might act independently from each other, thus binding the RNA with a different mechanism. It has been proposed that the latter type allows the protein to quickly scan the RNA sequences to identify potential binding sites (11).

We analysed the number of inter-domain contacts and their position in the context of the RRM alignment, showing the expected heterogeneity that ultimately leads to the broad range of possible binding modes. A contact is defined by a distance lower than the threshold of 6 Å between the Cβ atoms of two residues. From the 35 analysed experimental tandem RRMs, 14 show at least 1 inter-domain interaction in either the bound or unbound form (Supplementary Figure S6), to a maximum of 45 contacts for the RRM1-2 pair of PRP24 (PDB Id 6ASO), where the RRM1 β-sheet is occluded by RRM2, so creating an extensive network of contacts and preventing the RRM1 β-strands from binding RNA. Despite differences in the number of contacts between the bound and unbound structures in specific proteins, no general trend is observed.

The positions involved in the inter-domain contacts also change drastically among different proteins. We observed that

many of the involved positions are in the structured elements (Supplementary Figure S7), mainly between the RRM1 β4 and the RRM2 β2, which are often observed in contact when the tandem RRMs form an extended β-sheet (adjacent RRMs). Other tandem RRM arrangements are stabilized by different interactions and will be discussed in the tandem clustering section.

The contacts with the linker were similarly determined based on the master RRM alignment to identify the first and last linker residues (Supplementary Dataset S6). The linker length is highly variable, ranging from 4 residues in yeast PRP24 RRM1-2, to over 45 in the human PTB RRM3-4 pair. We created a sequence alignment for the linker region by simply pushing the residues to the sides and introducing the gaps in the centre, also referred to as sequence squeezing. Notably, we observed that the linker residues often interact with both RRM domains. Out of the 35 experimental structures, in 30 of them there is at least one interaction between the linker and the C-terminal RRM (e.g. RRM2 in a RRM1/2 tandem), while 21 interact with the N-terminal RRM (e.g. RRM1 in a RRM1/2 tandem), on either their bound and/or unbound forms (Supplementary Figure S8). (Supplementary Figure S9).

The 188 tandem RRM structures for the 35 experimental tandem RRMs were binary-labelled for all the contacts observed in the set, where '1' denotes presence of contacts, otherwise as '0'. The resulting dataset has a dimension of 188 entries by 381 binary features (contacts), of which 271 are inter-domain contacts and 110 linker-domain contacts (Supplementary Dataset S4).

AF2 tandem RRMs

We determined the contacts for the 7080 high confidence AF2 models following the same procedure as with the experimental set, and then mapping them to the AF2 RRM alignment. The inter-domain contacts showed a similar pattern to the experimental set, with most of the contacts involving the structured elements of the domain (Supplementary Figure S10). We compared the APAE for all the selected AF2 models (33 549 structures) with the number of inter-domain contacts, and we observed that the lack of contacts correlates with a higher average APAE (Supplementary Figure S11). This is likely connected to the presence of co-evolutionary signals in the multiple sequence alignment used by AF2, which are required to define inter-domain contacts that can then predict the tandem RRM orientation with a certain confidence.

The linker contacts with both RRM domains were also determined following the same procedure as with the experimental set. The linker interacts significantly more with the C-terminal RRM, with 62% (4416 structures) exhibiting at least one contact, while 23% (1633 structures) have one or more interactions with the N-terminal RRM. However, as already observed in the experimental set, the linker/N-terminal RRM contacts are more variable than the linker/C-terminal RRM contacts, where the N-terminal RRM contacts are observed almost all over the structured region while the C-terminal RRM contacts are limited to the β3 and α2 (Supplementary Figure S12).

As with the experimental tandem RRMs, the 7080 predicted structures were binary-labelled with all the observed contacts, a total of 1425 features (478, 489 and 458 for the inter-domain, linker/N-terminal RRM and linker/C-terminal RRM contacts, respectively).

Contact-based clustering

The inter-domain contact information for both the experimental and AF2 sets is used to cluster the tandem RRMs and compare which contacts (features) correlate with which specific orientation (RNP1 and RNP1–2 angles). Therefore, the clustering is performed using the inter-/linker-domain contacts as the only input.

Experimental tandem RRMs

In the unbound form the tandem RRMs orientation is driven by two main factors, the inter-domain contacts between the RRMs that may stabilize a specific orientation, and the length and flexibility of the linker domain, that may ultimately be involved in RRM interactions. Based on the RRM domain and linker alignments, we could robustly identify all the interacting positions in the dataset and compare them each other, a total of 381 binary contact features. Using *k*-means we generated clusters of 188 tandem RRM structures based on those binary features, and then visualizing their orientations. This allows to identify the correlations of RRM domain arrangements that are dependent on inter-/linker-domain contacts.

Using the *k*-means assessment of the model's inertia (see clustering section in materials & methods for details) we determined an informative number of clusters for the experimental dataset. The inertia informs on how well the *k*-means is performing and it decreases as the number of clusters increases. We tested from 1 to 50 clusters and determined the 'elbow' of the inertia values at 7 clusters (Supplementary Figure S13). Notably, the resulting clusters grouped structures with very similar orientations but often with just multiple structures from the same protein or related ones, in the same or different bound states (Figure 6A). Those clusters are still informative as it allows to identify which contacts lead to which orientations, and whether the presence/absence of RNA plays a role. The only exception is the unresolved cluster 1, which contains 28 proteins and an approximate average of 1 and 2 inter-domain and linker-domain contacts per structure, respectively. The cluster assignments for each tandem RRM are available as Supplementary data (Supplementary Dataset S10). An additional figure with bound and/or unbound representative structures from each cluster is available as Supplementary data (Supplementary Figure S14).

Clusters 2, 3, 4 and 5 are populated only by several structures of the same proteins, PRP24 RRM1-2 (yeast), PUF60, IF2B3 and U2AF2, respectively. Clusters 2, 3 and 4 contain unbound or half-bound structures (the RNA is bound to one of the 2 domains), but we do not observe any relevant change in the orientation upon binding. Alternatively, cluster 5 contains 7 structures of U2AF2 with both domains bound to the RNA. As other unbound structures for U2AF2 are available in the dataset, but not grouped in this cluster, this agrees on the RNA dependence of this tandem arrangement already discussed in literature (10,18).

Cluster 0 is populated by several structures from the hnRNP family, hnRNP A1 and hnRNP A2/B1, in either their bound or unbound forms with no effect on the domains' orientation. Similarly, cluster 6 contains several structures from the human and yeast RRM1-2 PABP orthologs, in all cases with both domains RNA bound.

We performed a feature importance analysis to identify the interactions that help discriminating different clusters better. Ultimately, this reveals the main contacts that drive spe-

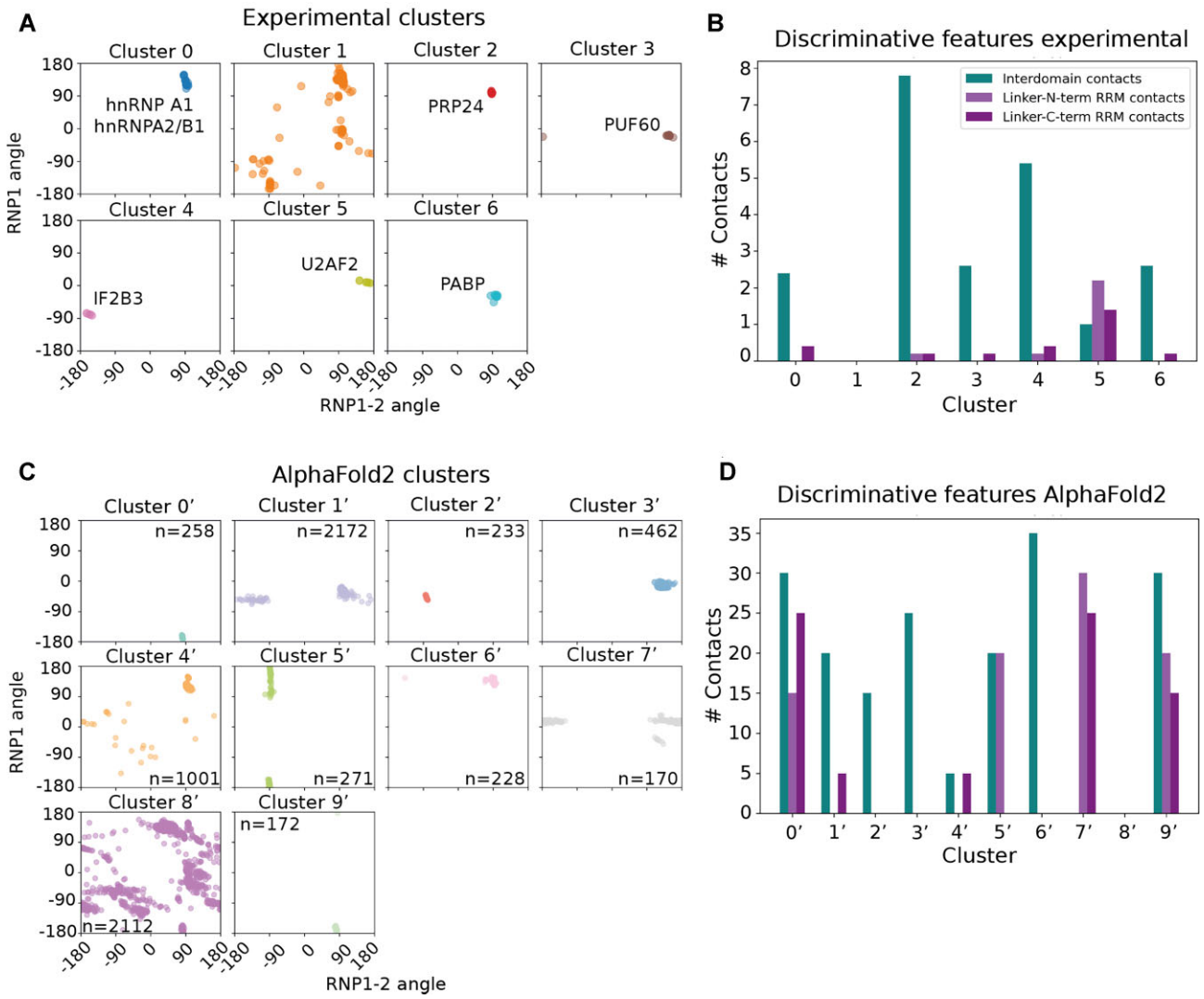


Figure 6. Contacts-based clustering and feature importance analysis. Experimental tandem RRMs: (**A**) RNP1 and RNP1–2 angles for the 7 clusters obtained with *k*-means. Each data point represents a PDB structure including multiple chains, and the proteins within each cluster are labelled except for cluster 1 that contains 21 different proteins. (**B**) Number of features with an importance higher than 0.5 (key contacts) separated by cluster and tandem parts involved in the interaction. AF2 tandem RRMs (7080 proteins): (**C**) RNP1 and RNP1–2 angles for the 10 clusters obtained with *k*-means. The number of proteins per cluster is labelled. (**D**) Number of features with an importance higher than 0.5 separated by cluster and tandem parts involved in the interaction.

cific tandem RRM orientations and potentially help stabilizing RNA-bound orientation. For the discriminative features, we selected any contact with a feature importance higher than 0.5, and labelled them as key contacts. Because of the binary nature in our features (presence/absence of a contact), a threshold of 0.5 means that at least half of the structures in the cluster have that interaction. The variability among the clusters comes from the different key contacts involved in the arrangement stabilization (Supplementary Figures S15–S17), which vary in number and parts involved (Figure 6B). We observed that in clusters 0 (hnRNPs, Figure 5E), 2 (PRP24), 3 (PUF60), 4 (IF2B3) and 6 (PABP, Figure 5C) most of the interactions with the highest importance are inter-domain contacts (from 12 to 39 contacts) further stabilized by 1 or 2 linker-domain contacts, often with the C-terminal RRM. The location of the linker contacts is quite conserved among these clusters, most of them occurring between the last linker residues and the β3-loop-α2 C-terminal RRM region, but also with the last residues of β4 from that same RRM (Supplementary Figure S17). On the other side, the RRM positions involved in the inter-domain contacts are highly variable (Supplementary Figure S15).

In contrast, we observe a completely different pattern in cluster 5 (U2AF2, Figure 5F) where most of the key contacts (feature importance > 0.5) involve the linker with either of the RRM domains. This cluster is populated by seven U2AF2-bound structures with a clearly defined orientation (Figure 6, cluster 5). This orientation is stabilized through a large network of interactions, with an average of 5 inter-domain contacts between the RRM1 α2-loop-β4 and the RRM2 α1 (Supplementary Figure S15), and over 15 linker-domain contacts across the 7 structures (Supplementary Figures S16 and S17). The top 50 features for each cluster including the most discriminative contacts (key contacts, importance > 0.5) are provided as Supplementary material (Supplementary Dataset S11).

In cluster 1 (unresolved cluster) there are no key contacts due to its high heterogeneity and being mostly populated by protein structures lacking inter-/linker-domain contacts, with a few exceptions that should be studied independently. The general trend in the lack of conserved interacting positions is clearly reflected in the broader range of angles observed in the contained structures (Figure 6, cluster 1).

The seven defined clusters revealed distinct orientations and interaction patterns, indicating the significance of inter-/linker-domain contacts and, in some cases, the RNA binding in shaping the RRM arrangements (e.g. U2AF2 cluster 5). Despite the clustering of these groups with the same or very related proteins, this establishes as reference and comparison point for the following AF2 clustering and analysis.

AF2 tandem RRMs

We clustered the selected 7080 entries using the *k*-means method. We calculated the inertia of the models from 1 to 50 hypothetical clusters to choose the most informative number (Supplementary Figure S18) (see clustering section in materials & methods for details). Using the elbow method we chose to split the data in 10 clusters, as there is a clear drop in the inertia's slope from that point onwards. The cluster assignments for each entry in the dataset are available as Supplementary data (Supplementary Dataset S12). There is a varying number of entries per cluster that ranges from 170 proteins (cluster 7′) to 2172 proteins (cluster 1′) (We refer to the AF2 clusters with a prime mark to easily distinguish them from the experimental clusters). As observed in the experimental clusters, the inter-domain and linker-domain contact labels are enough to cluster a varying number of tandem RRMs, as 8 out of the 10 generated clusters show a clear conservation in their orientation at least in one of the studied angles (Figure 6C). The only clear overlapping occurs between clusters 0′ and 9′ which essentially capture the same tandem arrangement, but stabilized by slightly different contacts. A representative structure for each of the well-defined clusters is available as Supplementary material (Supplementary Figure S19).

We also analysed the most discriminative contacts for each of the clusters following the same procedure as with the experimental tandem RRMs. Features with an importance higher than 0.5 (i.e. key contacts observed in more than 50% of the structures in the cluster) are counted and split in the three different contact types investigated, inter-domain, linker/N-terminal RRM and linker/C-terminal RRM contacts (Figure 6D). There is a large variation on the type and number of the key contacts among clusters, ranging from 3 contacts in cluster 2′ (excluding the unresolved clusters 4′ and 8′), to 14 contacts in cluster 0′. The 50 features with the highest importance for each cluster are available as Supplementary data (Supplementary Dataset S13).

The inter-domain contacts are the most discriminative features among the different clusters, being the most represented type of contact in 6 out of the 8 clusters with a conserved tandem orientation. The amino acid positions involved in such contacts are also quite variable (Supplementary Figure S20) agreeing with the different orientations observed (Figure 6C). Notably, both the linker/N-terminal and linker/C-terminal contacts are also identified as relevant contacts in 6 of the clusters, despite in some cases the contact positions are quite similar among clusters. The selected contacts involving the linker and the N-terminal RRM are quite discriminative among clusters 0′, 5′, 7′ and 9′ (Supplementary Figure S21). Contrarily, contacts involving the linker and the C-terminal RRM are quite similar in most of the cases, despite being selected as relevant interactions in clusters 0′, 1′, 4′, 7′ and 9′. The last residues of the linker often interact with the β3-loop-α2 regions, with the exception of cluster 7′ where the linker also interacts with the β2 strand (Supplementary Figure S22).

In 8 out of 10 AF2 clusters, the inter-/linker-domain contacts are discriminative enough to identify conserved arrangements. Despite the most prevalent key contacts are still inter-domain interactions as in the experimental set, both linker-N-terminal and linker-C-terminal contacts were relevant in multiple clusters, with some variations in their positions.

Experimental and AF2 clusters comparison

Comparing experimental and predicted clusters is essential for discovering new potential orientations that might not have been observed experimentally. Additionally, this process helps us understand the allowed sequence variations within the already characterized orientations. Interestingly, we have identified similarities among some of the experimental clusters and the 10 AF2 clusters. To quantify these similarities, we conducted two analyses, (i) We scored the experimentally observed angles against the AF2 cluster Kernel Density Estimations (KDEs)(Figure 7A) and (ii) we also analysed the shared interactions among the clusters using the Jaccard index (Figure 7B). In both cases, a higher value implies a higher level of similarity among the clusters. A representation of the KDE regions generated from the AF2 angles distributions is available as Supplementary data (Supplementary Figure S23).

Despite the angle comparison suggesting multiple matches among certain clusters, the contact analysis resolves most of the cases. For example, the experimental cluster 0 has a similar orientation to both AF2 clusters 4′ and 6′, with 0.87 and 0.85 density values, respectively. But regarding the contacts comparison, the Jaccard indexes obtained for cluster 4′ and 6′ are 0.4 and 0, respectively, clearly matching the experimental cluster 0 (hnRNPs, Figure 5E) with the AF2 cluster 4′. Moreover, hnRNP A1 is present in both the experimental cluster 0 and its AF2-predicted structure in the AF2 cluster 4′, reinforcing the similarity between the clusters. This also remarks the fact that despite observing similar orientations as shown by the angle distributions, the interactions stabilizing such orientations can vary significantly.

This analysis reveals intriguing similarities between certain experimental and AF2 clusters and potentially stable but unexplored tandem orientations. By quantifying the overlap and shared interactions among the clusters, we gain valuable insights into their structural relationships. However, it is important to note that despite similar angle distributions, the stabilizing interactions for these orientations can still differ. Notably, AF2 clusters 2′ and 6′ show a conserved orientation that has not been reported experimentally stabilized by those contacts. In both cases the arrangements are stabilized by different discriminative contacts compared to the ones observed for experimental clusters (Figure 7B). This can be illustrated by the models for the human protein RBM46 and plant protein MEI2-like2, for respectively the AF2 clusters 2′ and 6′. The contact maps between the RRM domains illustrate which positions stabilize these 'uncommon' arrangements confidently predicted by AF2. In cluster 2′, there is a 'V-shaped' arrangement between the β-sheets, whereas cluster 6′ displays a particular extended β-sheet configuration with both sheets point-

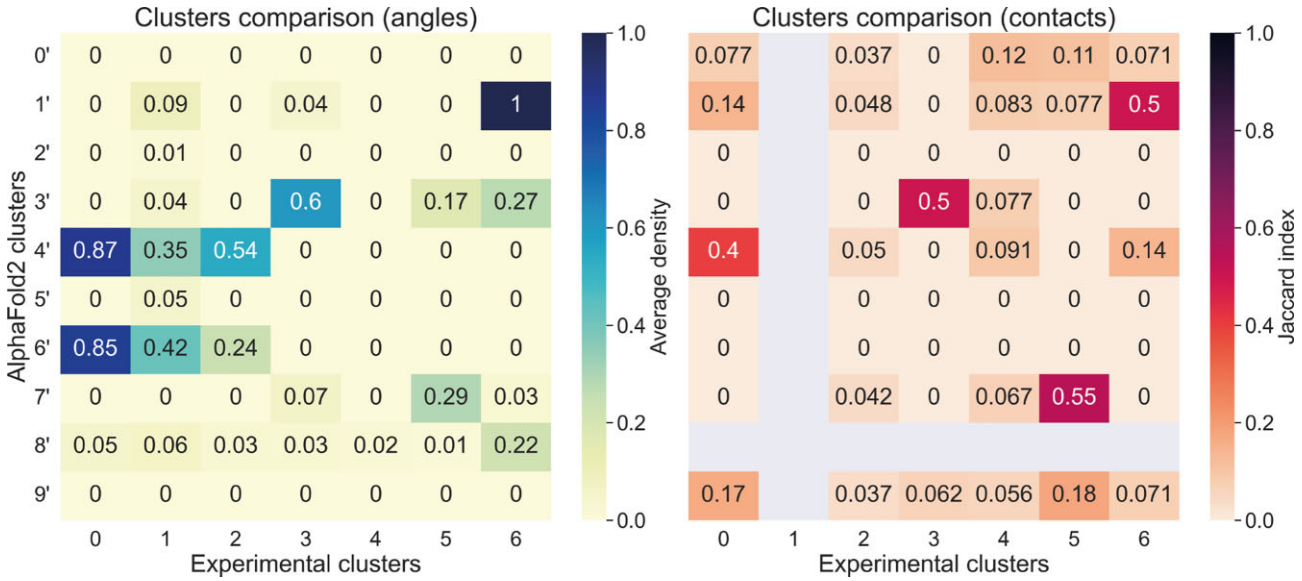


Figure 7. Experimental and AF2 cluster comparison based on the tandem angles (left) and contacts (right).

ing to opposite directions instead of lying side by side as in PABP or TDP-43. The contacts are extracted from the 233 and 288 structures that populate clusters 2′ and 6′, respectively (Figure 8).

AlphaFold2 clusters variability

To evaluate the variability captured within each AF2 cluster we studied both the sequence variability and amino acid composition of all the tandem RRMs.

Sequence variability

To evaluate the sequence variability of the 10 defined AF2 clusters, we examined the pairwise sequence identity within each cluster (Supplementary Figure S24). The observed sequence identity distributions differ slightly among clusters, with the exception of the unresolved cluster 8′, which exhibits a notably low average sequence identity of around 20%. For the remaining clusters, sequence identity values mostly predominantly fall within the range of 40% to 80% identity, highlighting their sequence homology.

To assess the presence of paralogs within each cluster, we examined the number of tandem RRMs belonging to the same species. All well-defined clusters show a varying number of potential paralogs, ranging from 19 (cluster 7′ containing 170 tandem RRMs) to 356 (Cluster 1′ containing 2172 tandem RRMs) (Supplementary Table S1). This underscores the widespread presence of paralogs (and orthologs by exclusion) in all AF2 clusters, aligning with previous research on duplication events involving RRMs (2).

Amino acid composition within the AF2 clusters

The different orientations observed in tandem RRMs are ultimately a consequence of the presence of different amino acids in specific positions. Those residues stabilize those arrangements by establishing interactions between the domains, or between the linker and either of the domains. We computed the fraction of the 20 amino acids across the 2 RRM domains of each tandem (excluding the linker) for the 10 AF2 clusters (Supplementary Figure S25). We observed that very often different clusters show different amino acid compositions, both involving different residue types but also within specific types. This suggests that detailed analyses and comparisons between the orientations captured among different clusters, validated by experimental observations, will help to better understand how the different arrangements are stabilized.

Discussion

Understanding the binding mode of tandem RRMs is a crucial step in elucidating their functional implications and biological roles (6,7). The inter-domain interactions, and even the interactions between the linker and either of the RRM domains, help stabilizing specific tandem RRM orientations, leading to different binding modes. In tandem RRMs such as hnRNP A1 and PTB RRM3-4, the network of inter-domain contacts is strong enough to establish an orientation that persists upon binding (33,37). Many efforts have been made to solve the structure of specific RRMs and investigate their functional roles, often focusing on therapeutically relevant human RRMs (38,39) and/or proteins involved in crucial cell processes such as splicing (10,40). Previous work to categorize tandem RRMs based on their binding mode were limited to available experimental data at the time, and purely based on structure visualization (11,12). Nowadays, even for human multiple RRMs, there is still an enormous gap between the 25 proteins with available structures in PDB (23) and the 107 we found via HMM search (Supplementary Figure S26A). This gap is much larger across all species, where a large-scale search against UniRef90 retrieved over 47000 proteins containing at least 2 RRMs, and up to 14 (Supplementary Figure S26B).

In this work we implemented a novel methodology to robustly characterize the orientation of multiple domain proteins, and applied it to tandem RRMs (Figure 2). Based on a carefully curated alignment, we identified 3 key positions that allowed us to define two vectors for each RRM domain. The vectors are used to calculate two angles that capture the two main rotation axes between the domains (Figure 2B-D). The experimental tandem RRMs analysis revealed the importance of the inter-/linker-domain interactions to stabilize cer-

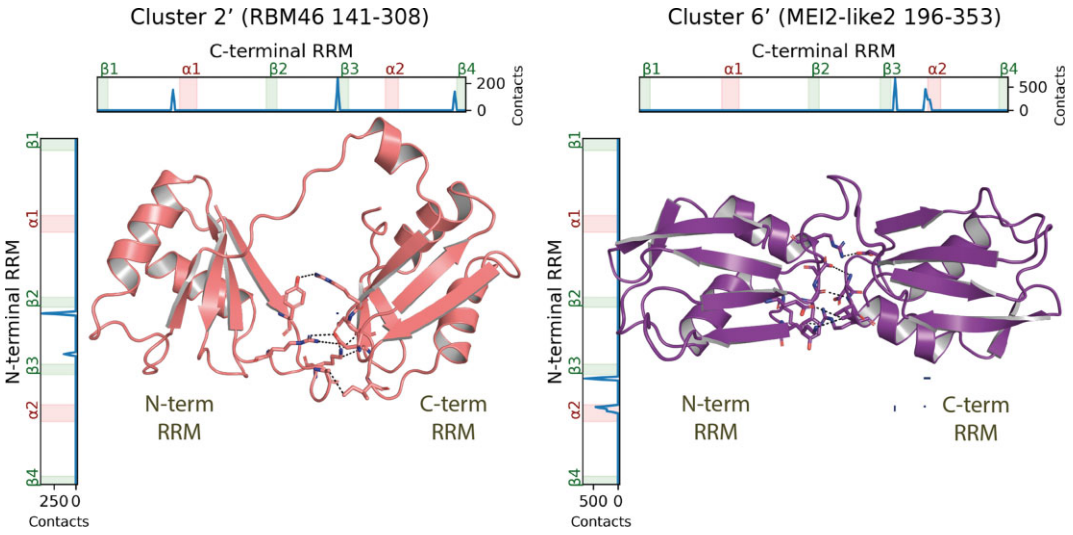


Figure 8. Contact maps and representative structures for cluster 2′ (left, UniProt ID: Q8TBY0) and cluster 6′ (right, UniProt ID: Q6ZI17). Both clusters lack an experimental counterpart and represent potentially new tandem arrangements.

tain tandem RRM arrangements. Despite all the well-defined experimental clusters only contain a single protein or a pair of related proteins, it works as our ground truth and comparison point for the AF2 analysis. Moreover, the feature importance analysis highlights the relevance of the inter-domain contacts, but also of the linker-domain contacts, with at least one contact observed in all the well-defined clusters with either of the RRM domains. Moreover, we also gain knowledge from all the tandem RRMs that do not show any interactions and that are grouped in the mixed cluster 1, whose tandem arrangement is likely to be RNA driven, or there is not enough structural evidence to generate an informative cluster.

To further investigate the AF2 clusters we performed an exploratory gene ontology (GO) enrichment analysis for each cluster using all the tandem RRMs identified on UniRef90 as the population set (Supplementary Figure S27). This allows to identify the overrepresented biological processes and molecular functions linked to each cluster (we excluded subcellular location from the analysis), and define likely orientation-function relationships. The complete lists with the significantly enriched and depleted terms for each cluster are available as Supplementary material (Supplementary datasets S15–S23).

Equivalent experimental-AF2 clusters

The experimental clusters 3 (PUF60), 5 (U2AF2, Figure 5F) and 6 (PABPs, Figure 5C) show a high degree of similarity with the AF2 clusters 3′, 7′ and 6′, respectively, as observed in both angles and contacts comparison (Figure 7). The experimental cluster 3 contains 13 structures of the human PUF60 tandem RRM in different bound states. This protein is involved in the 3′ splice site recognition (41), and it is also included in the AF2 cluster 3′. The contacts analysis showed that this arrangement is stabilized by multiple inter-domain contacts, with 13 and 5 key contacts for the experimental and AF2 clusters, respectively (Supplementary Figures S14 and S19). In this case the contacts stabilize an orientation where the RRM2 is prevented to bind RNA as its β-sheet is occluded by the RRM1 (Supplementary Figures S14 and S15). This agrees with the available experimental structures where the RRM2 has not been reported bound to RNA. The gene ontology analysis also supports the similarities within these clusters as most of the enriched terms for the AF2 cluster 3′ are related to RNA splicing and mRNA splice site recognition (Supplementary Figure S27).

The experimental cluster 5 is exclusively populated by a particular arrangement of human U2AF2 RRM1-2 structures when both domains are RNA bound. The arrangement is stabilized mostly by linker-domain contacts, with a total of 17 contacts versus 5 inter-domain contacts (Figure 6B). This tandem RRM is also observed in other experimental clusters, manifesting the highly dynamic nature of this complex (34), partly granted for an almost 30-residues long flexible linker. Notably, the AF2 cluster 7′ shows a similarly conserved tandem orientation (Figure 7A) purely stabilized by linker-domain contacts, with 6 and 5 linker-RRM1 and linker-RRM2 contacts, respectively (Figure 6D). The contacts analysis highlights the similarity between the experimental and AF2 clusters 5 and 7′ (Figure 7B). In this arrangement, multiple residues from the linker interact with the α2-loop-β4 region of the RRM1, and also with part of the β2, β3 and α2 elements of the RRM2, bringing the 2 RRMs together while also participating in RNA interactions, as already observed in several tandem RRM-RNA complexes (42). Transient linker-RRM2 contacts are observed in the unbound state, preventing binding to weak pyrimidine-tract RNAs, and proving crucial to regulate RNA binding selectivity via RNA proofreading (43). Therefore, the function of the linker is of ultimate importance for this particular RRM arrangement, and plays and important role on the 3′ splice site signal recognition. The GO analysis also supports the similarity between the clusters as there are several splicing related enriched terms (Supplementary Figure S27).

The last experimental cluster with a clear AF2 cluster counterpart is cluster 6, populated by the human and yeast RRM1-2 pair of PABP. The experimental cluster shows a conserved orientation over the 10 available bound structures, forming an extended β-sheet that recognizes around 8 nucleotides (44,45). The respective AF2 models for these tandems are grouped in cluster 1′, which contains a total of 2172 pre-

dicted models with a highly conserved RNP1 angle around −30º, while the RNP1–2 angle remains more dynamic. The angle and contacts comparison between the clusters suggest a high similarity between them (Figure 7) This is further supported by the GO analysis, which results on several enriched terms related to mRNA translation and poly(A) binding (Supplementary Figure S27), as those are some of the main roles of PABP. This suggests a strong correlation between this binding mode and its biological function.

Experimental clusters without an AF2 equivalent

The experimental tandem RRMs in clusters 2 (PRP24 RRM1-2) and cluster 4 (IF2B3) are stabilized by 37 and 27 inter-domain contacts, respectively. This large network of contacts is fixing a specific arrangement that does not change upon RNA binding, clearly observed on the angle values for different bound states of these proteins. These experimental clusters do not share clear similarities with any of the AF2 generated clusters, especially regarding their contacts both obtaining very low Jaccard indexes with any of the clusters (Figure 7).

The experimental cluster 1 is very heterogenous, showing a broad distribution of orientations. The cluster is populated by 28 different proteins, but only a few of them are included in the AF2 analysis. The remaining entries have an inter-domain APAE higher than 10 Å, meaning AF2 cannot find a strong enough signal to accurately predict the orientation of the domains. This indicates that those are rather flexible tandem RRMs and its arrangement is highly dependent on the bound RNA rather than a pre-established orientation stabilized by inter-domain contacts. However, a deeper analysis within these clusters may still reveal other stable arrangements, but that do not generate a cluster on their own due to lack of data and the limited number of clusters we created.

Unexplored AF2 clusters

The AF2 analysis revealed several clusters that do not match with any of the experimental ones, but that include many different proteins and a conserved orientation stabilized by multiple inter-/linker-domain interactions. Interestingly the AlpaFold2 clusters 0′ and 9′ show a very similar orientation stabilized by essentially the same inter-domain contacts (Supplementary Figure S20). The only difference between the clusters are the linker contacts with the N-terminal RRM (Supplementary Figure S21), which again highlights the linker importance that may affect the binding kinetics and regulation of the proteins within these two clusters. Notably, the 258 entries within cluster 0′ are RRM3-4 pairs, including the well-studied PTB (19). On the other hand, 171 out of 172 entries in cluster 9′ are RRM2-3 pairs, and one RRM4-5 pair, as identified by the hmmsearch. However, this change on the RRM's position can be due to enough mutations on some of the RRMs that prevents it to be identified by hmmsearch, or that cannot be aligned to the RRM master alignment. The GO analysis revealed similar enriched terms between these clusters, mostly regarding RNA splicing and mRNA processing, in agreement with PTB main roles.

The AF2 clusters 2′, 5′ and 6′ do not contain any tandem RRM from which we have experimental structures available, but they still show a clear conservation in their orientation (Figure 6C). The captured arrangements within these clusters are mostly stabilized by inter-domain interactions, but also by linker-domain contacts in cluster 5′ (Figure 6D). The human RBM46 RRM1-2 tandem is part of cluster 2′, showing a high conservation in both the RNP1 and RNP1–2 angles. The RRM1 is partially occluded by the RRM2 in this arrangement (Supplementary Figure S19), but not as clearly as in the PRP24 experimental cluster 2 (Supplementary Figure S14), with its β-sheet completely packed against RRM2 with extensive inter-domain contacts (46). Clusters 5′ and 6′ also show unique orientations not reported experimentally, but they do not contain any human protein. To further study their orientations we chose PABP-interacting protein 8 for cluster 5′ (CID8, UniProt Id Q9C8M0, *Arabidopsis thaliana*) and MEI2-like2 protein for cluster 6′ (ML2, UniProt Id Q6ZI17, *Schizosaccharomyces pombe*) as representative structures with experimental evidence at the protein and transcript levels, respectively (Supplementary Figure S19). The RRM domains on CID8 form an asymmetric cleft due to an atypical relative rotation of the RRM domains when compared with the extended RRM arrangement observed on human and yeast PABPs (AF2 cluster 1′). This is also clearly observed on the different RNP1 angles between these clusters (Figure 6C). The β-sheets for both domains are still accessible to the RNA, despite slightly occluded by the inter-domain linker that interacts extensively with N-terminal RRM (Cluster 5′ in Supplementary Figure S16). On the other hand, ML2 presents a symmetrical arrangement with both RRM domains interacting with each other with the 'bottom' part of the domain (Cluster 6′ in Supplementary Figure S19) with both β-sheets completely exposed for presumably RNA binding. However, experimental data regarding both their structures and RNA binding capabilities is crucial to better understand the functional implications of these tandem arrangements.

The linker length is also conserved

An interesting finding of this work is the relevance of the linker contacts with any of the RRM domains to stabilize a broad range of tandem RRM orientations, which seems to be the rule rather than the exception. In most of the experimental and AF2 clusters there is at least 1 relevant contact that involves the linker, and up to 18 in the experimental cluster 5, containing multiple U2AF2 structures. Consequently, we analysed whether the linker length is also conserved, alongside the observed contacts, within the AF2 clusters. Notably, we observed that except for the highly heterogenous cluster 8′, the rest of the clusters show a clear conservation of the linker length (Supplementary Figure S28).

Conclusions

With this work we have robustly characterized the orientation of the experimental tandem RRM structures, and clustered them based on their inter-domain and linker-domain contact positions to unravel the main features leading to such orientations. We expanded our analysis by including AF2 models and a careful assessment of their inter-domain relative position. For this task we used the PAE values, conveniently provided for each AF2 model as well. This allows us to assess how confident the model is on the relative orientation between domains. The GO exploratory analysis also allowed us to look at the relationships between certain arrangements and the biological processes they are or might be involved in from an evolutionary perspective. A deeper analysis on the residue-residue contacts conservation will provide valuable insights to design specific tandem RRMs with a specific arrangement. This, coupled with the recent advances on affinity-enhanced RNA bind-

ing domains (47) may provide the right tools to start studying many relevant tandem RRMs that remain unexplored to this date.

Notably, our novel approach to study multi-domain arrangements can also be potentially applied to any other pair of adjacent protein domains. The only requirement is the identification of structurally conserved positions in stable secondary structure elements that define the core fold of the protein. As illustrated here, these enable the definition of a unique vector for each domain, allowing the computation of their relative orientations. The number, position and direction of the vectors defined must be in accordance with the rotation axes to be studied. As proteins often have more than 1 folded domain, this new approach to quantitatively compare inter-domain orientations can be broadly applied.

Limitations

Our AF2 analysis is limited to tandem RRMs that show a consistent network of interactions, as it is in those cases where AF2 predicts the RRMs orientations with certain confidence (Supplementary Figure S11). Most of the predicted models are therefore excluded, but this also indicates that the arrangement of such domains is likely to be highly RNA dependent, as the unbound domains may tumble independently from each other. Moreover, AF2 is mostly trained on crystal structures, which in unbound tandem RRMs can be affected by crystal packing and lead to inaccurate arrangements that would bias the AF2 predictions. Simplifying the tandem RRM orientation to two different angles was a trade-off between data generalisation and visualization and real structure accuracy. For certain subtle rearrangements it remains important to analyse the actual protein structures. Finally, it is essential to note that our analysis is based on a static representation of protein structures. Specifically, tandem RRMs are highly dynamic entities, and as a result, their behavior in solution can vary significantly.

Data availability

All the datasets and code required to run the main analysis/plots for this work are publicly available at https://bitbucket.org/bio2byte/tandem_rrm_analysis/src/master/.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

We express our gratitude to Dr David Bickel for his help in the mathematical interpretation of inter-domain angles.

Funding

Marie Skłodowska-Curie Innovative Training Network (MSCA-ITN) RNAct supported by European Union's Horizon 2020 research and innovation programme [813239 to W.V. and M.S.]; European Regional Development Fund and Brussels-Capital Region-Innoviris within the framework of the Operational Programme 2014–2020 [ERDF-2020 project ICITY-RDI.BRU to W.V.].

Conflict of interest statement

None declared.

References

1. Cléry,A., Blatter,M. and Allain,F.H.-T. (2008) RNA recognition motifs: boring? Not quite. *Curr. Opin. Struct. Biol.*, **18**, 290–298.
2. Tsai,Y.S., Gomez,S.M. and Wang,Z. (2014) Prevalent RNA recognition motif duplication in the human genome. *RNA*, **20**, 702–712.
3. Hennig,J., Militti,C., Popowicz,G.M., Wang,I., Sonntag,M., Geerlof,A., Gabel,F., Gebauer,F. and Sattler,M. (2014) Structural basis for the assembly of the Sxl–Unr translation regulatory complex. *Nature*, **515**, 287–290.
4. Stitzinger,S.H., Sohrabi-Jahromi,S. and Söding,J. (2023) Cooperativity boosts affinity and specificity of proteins with multiple RNA-binding domains. *NAR Genomics Bioinformatics*, **5**, lqad057.
5. Agarwal,A. and Bahadur,R.P. (2023) Modular architecture and functional annotation of human RNA-binding proteins containing RNA recognition motif. *Biochimie*, **209**, 116–130.
6. Maris,C., Dominguez,C. and Allain,F.H.-T. (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J.*, **272**, 2118–2131.
7. Daubner,G.M., Cléry,A. and Allain,F.H.-T. (2013) RRM–RNA recognition: NMR or crystallography…and new findings. *Curr. Opin. Struct. Biol.*, **23**, 100–108.
8. Corley,M., Burns,M.C. and Yeo,G.W. (2020) How RNA-binding proteins interact with RNA: molecules and mechanisms. *Mol. Cell*, **78**, 9–29.
9. Lunde,B.M., Moore,C. and Varani,G. (2007) RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.*, **8**, 479–490.
10. Mackereth,C.D., Madl,T., Bonnal,S., Simon,B., Zanier,K., Gasch,A., Rybin,V., Valcárcel,J. and Sattler,M. (2011) Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. *Nature*, **475**, 408–411.
11. Cléry,A. and Allain,F.H.-T. (2013) From structure to function of RNA binding domains Landes. *Madame Curie Bioscience Database*.
12. Mackereth,C.D. and Sattler,M. (2012) Dynamics in multi-domain protein recognition of RNA. *Curr. Opin. Struct. Biol.*, **22**, 287–296.
13. Pérez-Cañadillas,J.M. (2006) Grabbing the message: structural basis of mRNA 3′UTR recognition by Hrp1. *EMBO J.*, **25**, 3167–3178.
14. Kooshapur,H., Choudhury,N.R., Simon,B., Mühlbauer,M., Jussupow,A., Fernandez,N., Jones,A.N., Dallmann,A., Gabel,F., Camilloni,C., *et al.* (2018) Structural basis for terminal loop recognition and stimulation of pri-miRNA-18a processing by hnRNP A1. *Nat. Commun.*, **9**, 2479.
15. Handa,N., Nureki,O., Kurimoto,K., Kim,I., Sakamoto,H., Shimura,Y., Muto,Y. and Yokoyama,S. (1999) Structural basis for recognition of the tra mRNA precursor by the sex-lethal protein. *Nature*, **398**, 579–585.
16. Johansson,C., Finger,L.D., Trantirek,L., Mueller,T.D., Kim,S., Laird-Offringa,I.A. and Feigon,J. (2004) Solution structure of the complex formed by the two N-terminal RNA-binding domains of nucleolin and a pre-rRNA target. *J. Mol. Biol.*, **337**, 799–816.
17. Wang,H., Zeng,F., Liu,Q., Liu,H., Liu,Z., Niu,L., Teng,M. and Li,X. (2013) The structure of the ARE-binding domains of Hu antigen R (HuR) undergoes conformational changes during RNA binding. *Acta Cryst. D*, **69**, 373–380.
18. Huang,J., Warner,L.R., Sanchez,C., Gabel,F., Madl,T., Mackereth,C.D., Sattler,M. and Blackledge,M. (2014) Transient electrostatic interactions dominate the conformational equilibrium sampled by multidomain splicing factor U2AF65: a combined NMR and SAXS study. *J. Am. Chem. Soc.*, **136**, 7068–7076.

19. Oberstrass,F.C., Auweter,S.D., Erat,M., Hargous,Y., Henning,A., Wenter,P., Reymond,L., Amir-Ahmady,B., Pitsch,S., Black,D.L., *et al.* (2005) Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science*, **309**, 2054–2057.
20. Kielkopf,C.L., Lücke,S. and Green,M.R. (2004) U2AF homology motifs: protein recognition in the RRM world. *Genes Dev.*, **18**, 1513–1526.
21. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
22. The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
23. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
24. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J., *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
25. Roca-Martínez,J., Dhondge,H., Sattler,M. and Vranken,W.F. (2023) Deciphering the RRM-RNA recognition code: a computational analysis. *PLoS Comput. Biol.*, **19**, e1010859.
26. Suzek,B.E., Huang,H., McGarvey,P., Mazumder,R. and Wu,C.H. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**, 1282–1288.
27. Michaud-Agrawal,N., Denning,E.J., Woolf,T.B. and Beckstein,O. (2011) MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.*, **32**, 2319–2327.
28. Meurer,A., Smith,C.P., Paprocki,M., Čertík,O., Kirpichev,S.B., Rocklin,M., Kumar,A.., Ivanov,S., Moore,J.K., Singh,S., *et al.* (2017) SymPy: symbolic computing in Python. *PeerJ Comput. Sci.*, **3**, e103.
29. Cock,P.J.A., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B., *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
30. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V., *et al.* (2011) Scikit-learn: machine learning in Python. *JMLR*, **12**, 2825–2830.
31. Klopfenstein,D.V., Zhang,L., Pedersen,B.S., Ramírez,F., Warwick Vesztrocy,A., Naldi,A., Mungall,C.J., Yunes,J.M., Botvinnik,O., Weigel,M., *et al.* (2018) GOATOOLS: a Python library for gene ontology analyses. *Sci. Rep.*, **8**, 10872.
32. Evans,R., O'Neill,M., Pritzel,A., Antropova,N., Senior,A., Green,T., Žídek,A., Bates,R., Blackwell,S., Yim,J., *et al.* (2022) Protein complex prediction with AlphaFold-multimer. bioRxiv doi: https://doi.org/10.1101/2021.10.04.463034, 04 October 2021, preprint: peer not reviewed.
33. Barraud,P. and Allain,F.H.-T. (2013) Solution structure of the two RNA recognition motifs of hnRNP A1 using segmental isotope labeling: how the relative orientation between RRMs influences the nucleic acid binding topology. *J. Biomol. NMR*, **55**, 119–138.
34. Voith von Voithenberg,L., Sánchez-Rico,C., Kang,H.-S., Madl,T., Zanier,K., Barth,A., Warner,L.R., Sattler,M. and Lamb,D.C. (2016) Recognition of the 3′ splice site RNA by the U2AF heterodimer involves a dynamic population shift. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E7169–E7175.
35. Lukavsky,P.J., Daujotyte,D., Tollervey,J.R., Ule,J., Stuani,C., Buratti,E., Baralle,F.E., Damberger,F.F. and Allain,F.H.-T. (2013) Molecular basis of UG-rich RNA recognition by the human splicing factor TDP-43. *Nat. Struct. Mol. Biol.*, **20**, 1443–1449.
36. Lamichhane,R., Daubner,G.M., Thomas-Crusells,J., Auweter,S.D., Manatschal,C., Austin,K.S., Valniuk,O., Allain,F.H.-T. and Rueda,D. (2010) RNA looping by PTB: evidence using FRET and NMR spectroscopy for a role in splicing repression. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 4105–4110.
37. Chen,X., Yang,Z., Wang,W., Qian,K., Liu,M., Wang,J. and Wang,M. (2021) Structural basis for RNA recognition by the N-terminal tandem RRM domains of human RBM45. *Nucleic Acids Res.*, **49**, 2946–2958.
38. De Silva,N.I.U., Fargason,T., Zhang,Z., Wang,T. and Zhang,J. (2022) Inter-domain flexibility of Human ser/arg-rich splicing factor 1 allows variable spacer length in cognate RNA's bipartite motifs. *Biochemistry*, **61**, 2922–2932.
39. Cléry,A., Krepl,M., Nguyen,C.K.X., Moursy,A., Jorjani,H., Katsantoni,M., Okoniewski,M., Mittal,N., Zavolan,M., Sponer,J., *et al.* (2021) Structure of SRSF1 RRM1 bound to RNA reveals an unexpected bimodal mode of interaction and explains its involvement in SMN1 exon7 splicing. *Nat. Commun.*, **12**, 428.
40. Cho,S., Hoang,A., Chakrabarti,S., Huynh,N., Huang,D.-B. and Ghosh,G. (2011) The SRSF1 linker induces semi-conservative ESE binding by cooperating with the RRMs. *Nucleic Acids Res.*, **39**, 9413–9421.
41. Hsiao,H.-H.T., Crichlow,G.V., Murphy,J.W., Folta-Stogniew,E.J., Lolis,E.J. and Braddock,D.T. (2020) Unraveling the mechanism of recognition of the 3′ splice site of the adenovirus major late promoter intron by the alternative splicing factor PUF60. *PLoS One*, **15**, e0242725.
42. Agrawal,A.A., Salsi,E., Chatrikhi,R., Henderson,S., Jenkins,J.L., Green,M.R., Ermolenko,D.N. and Kielkopf,C.L. (2016) An extended U2AF65–RNA-binding domain recognizes the 3′ splice site signal. *Nat. Commun.*, **7**, 10950.
43. Kang,H.-S., Sánchez-Rico,C., Ebersberger,S., Sutandy,F.X.R., Busch,A., Welte,T., Stehle,R., Hipp,C., Schulz,L., Buchbender,A., *et al.* (2020) An autoinhibitory intramolecular interaction proof-reads RNA recognition by the essential splicing factor U2AF2. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 7140–7149.
44. Schäfer,I.B., Yamashita,M., Schuller,J.M., Schüssler,S., Reichelt,P., Strauss,M. and Conti,E. (2019) Molecular basis for poly(A) RNP architecture and recognition by the Pan2-Pan3 deadenylase. *Cell*, **177**, 1619–1631.
45. Deo,R.C., Bonanno,J.B., Sonenberg,N. and Burley,S.K. (1999) Recognition of polyadenylate RNA by the poly(A)-binding protein. *Cell*, **98**, 835–845.
46. Montemayor,E.J., Curran,E.C., Liao,H.H., Andrews,K.L., Treba,C.N., Butcher,S.E. and Brow,D.A. (2014) Core structure of the U6 snRNP at 1.7 Å resolution. *Nat. Struct. Mol. Biol.*, **21**, 544–551.
47. Chaves-Arquero,B., Collins,K.M., Abis,G., Kelly,G., Christodoulou,E., Taylor,I.A. and Ramos,A. (2023) Affinity-enhanced RNA-binding domains as tools to understand RNA recognition. *Cell Rep. Methods*, **3**, 100508.

Received: August 16, 2023. **Revised:** December 7, 2023. **Editorial Decision:** January 3, 2024. **Accepted:** January 9, 2024