Supporting Information Appendix

All-Atom Simulations Disentangle the Functional Dynamics Underlying Gene Maturation in the Intron Lariat Spliceosome

Lorenzo Casalino[a], Giulia Palermo[b], Angelo Spinello[c], Ursula Rothlisberger[d] and Alessandra Magistrato[c*]

a. Molecular and Statistical Biophysics, International School for Advanced Studies (SISSA), via Bonomea 265, 34136 Trieste, Italy.

b. Department of Bioengineering, University of California Riverside, Riverside, CA 92507, USA.

c. CNR-IOM-Democritos National Simulation Center c/o SISSA, via Bonomea 265, 34136 Trieste, Italy.

d. Laboratory of Computational Chemistry and Biochemistry, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland.

Table of contents:

***Corresponding author:** Dr. Alessandra Magistrato (alessandra.magistrato@sissa.it)

1. Supporting Methods

1.1 Structural models: ILS-1 and ILS-2. Our molecular dynamics (MD) simulations were based on the *Schizosaccharomyces Pombe* (*S. pombe*) spliceosome reconstructed with cryo-electron microscopy (cryo-EM) at the average resolution of 3.6 Å (PDB entry 3JB9) (1, 2). The two model systems, ILS-1 (Fig. S1) and ILS-2 (Fig. S2) were built on this structure, which captured the spliceosome at a late stage of the splicing cycle, namely the intron lariat spliceosome complex. Indeed, this structure well defines the intron lariat (IL), while showing a weak EM density for the exon, which was either already released as pre-mRNA or lost during the purification. In addition to the 5'-exon, other regulatory factors characterizing the C and C* complex (i.e., Slu7, Prp18, Prp22) are missing, suggesting that this structure most likely corresponds to the post-splicing intron lariat spliceosome (ILS) complex (1, 3). The deposited PDB structure shows an asymmetric morphology which exceeds 300 Å in its longest dimension. The core proteins and RNAs have a resolution ranging from 2.9 Å to 3.6 Å (up to 5 Å in some cases), while the most peripheral regions exhibit a poor EM-density (with a resolution larger than 5 Å). Importantly, this model provided for the first time precious near-atomic details (exceeding 3.2 Å for some proteins in the core region) on the intact catalytic site architecture and on four multicomponent subcomplexes (i.e., U5 snRNP, U2 snRNP, NTC and NTR) comprising a total of 37 proteins, 3 snRNAs and the IL. In particular, nearly complete atomic models for some crucial U5 snRNPs proteins like the central Spp42 (Prp8 in *S. cerevisiae*) and Cwf10 (Snu114 in *S. cerevisiae*) were defined along with a first glimpse of some NTC and NTR proteins. With the aim of studying the functional dynamics of the most important, central and conserved SPL components, we

considered only the core of this structure (Fig. S3). In particular, we built two model systems, namely 'ILS-1' (Fig. S1), counting 721'089 atoms, and 'ILS-2' (Fig. S2), counting 914'099 atoms. The ILS-1 model consists of 16 proteins (i.e., (i) Spp42, Cwf10, Cwf17 and the 7 Sm-ring chains of U5 snRNP, (ii) Cwf2 and Cwf15 of the NTC core, and (iii) Prp5, Cwf5, Cwf19, Cwf14 from NTR), 3 snRNAs (U5, U6 and U2 snRNA) and the IL. In the ILS-2 model we included additional domains of Spp42 (endonuclease and RNase H-like domains, 498 aa in total) and Cwf10 (domain I, II, III, IV and V, 571 aa in total) and two extra proteins (i.e., Prp45 of NTR and Prp17). A guanosine diphosphate (GDP) molecule was also included in the ILS-2 as solved in the original PDB structure.

Both the models were embedded in a 14 Å layer of TIP3P (4) water molecules, thus leading to a box size of 168 · 193 · 249 $Å^3$ for ILS-1 and of 212 · 189 · 256 $Å^3$ for ILS-2, containing also the four catalytic Mg^{2+} ions, 7 Zn^{2+} and 202/194 (ILS-1/ILS-2) Na^+ counter ions. The final atomic systems were generated using the coordinates provided in the original PDB entry (1, 2). Importantly, chains A (Spp42), E (Sm-B), G (Sm-D2) and L (Cwf17) of the PDB structure contained small gaps due to unresolved residues (from 1 up to 12). *De novo* model building, as implemented in Modeller 9v16 (5), was used to reconstruct these missing loops, which were further refined through the loop refinement procedure (6, 7). The generated loops were first selected among 50 models according to the DOPE score (8) and subsequently evaluated through an accurate visual inspection. We remark that Modeller has been shown to be very accurate for small loops modeling (9).

1.2 Molecular Dynamics simulations. The two models were subjected to MD simulations with the Gromacs 5 (10) software package. The AMBER-ff12SB force field

(FF) was adopted for proteins (11), while the ff99+bsc0+χOL3 FF was used for RNAs (12, 13), since these are the most validated and recommended force fields for protein/RNA systems (14). Mg^{2+} ions were described with the non-bonded fixed point charge FF due to Åqvist (15) as it was shown to properly describe binuclear sites (16). Na^{+} ions parameters were taken from Joung et al. (17) while Zn^{2+} ions were modelled with the cationic dummy atoms approach developed by Pang (18). The GDP molecule was described using the parameters developed by Meagher at al. (19). The RESP charges of the BP adenosine (A501) were calculated according to the Merz-Singh-Kollman (MK) scheme (20) and derived on the structure of the A501-G100 dinucleotide upon an optimization with Gaussian 09 (21) program at Hartree-Fock level of theory with the 6-31g* basis set, followed by a fitting on the electrostatic potential with the antechamber module of ambertools13 (22). The topologies were built with ambertools 13 and were subsequently converted in a GROMACS format using the software acpype (23).

MD simulations were performed on the isothermal-isobaric ensemble (NPT) using periodic boundary conditions. Temperature control at 300 K was achieved by stochastic velocity rescaling thermostat (24), while pressure control was accomplished by coupling the systems to a Parrinello-Rahman barostat with a reference pressure of 1 bar (25, 26). LINCS algorithm (27) was used to constrain the bonds involving hydrogen atoms and the particle mesh Ewald method (28) to account for long-range electrostatic interactions with a cutoff of 12 Å. Four replicas, three for ILS-1 and one for ILS-2, were run using an integration time step of 2 fs, reaching an overall simulation time of 3.25 μs (3 × 0.75 μs for ILS-1 and 1 × 1 μs for ILS-2).

In all simulations, we have used a very careful and slow equilibration protocol as recommended in the literature for protein/RNA MD simulations (14). Namely, the systems were initially put through a soft minimization using a steepest descent algorithm with a force convergence criterion set to 1000 kJ mol^{-1} nm^{-1}. Then, the models were smoothly annealed from 0 to 300 K with a temperature gradient of 50 K every 2 ns and for a total of 12 ns. In this phase, only water molecules and Na^+ ions were allowed to move, while the rest was subjected to harmonic position restraints with a force constant of 1000 kJ/mol nm^2. Once the temperature was raised up to 300 K, 20 ns of NPT simulations were conducted to stabilize the pressure to 1 bar by coupling the systems to a Berendsen barostat (29) and imposing the same restraints used in the heating phase. Subsequently, the barostat was switched to Parrinello-Rahman and the position restraints on proteins and RNAs were restricted only to the backbone atoms. These were gradually decreased in three consecutive steps of 30, 10, 10 ns each, during which the force constant was set to 1000, 250, 50 kJ/mol nm^2, respectively. Finally, after an attentive equilibration protocol of ~80 ns, all the restraints were released and the production runs were performed for ~670 ns (for a total of ~750 ns) for each of the ILS-1 replicas, while for ILS-2 replica the production run was conducted for ~920 ns (for a total of 1 μs).

1.3 Data analysis. The snapshots were collected every 50 ps of MD trajectories and were subsequently visualized with the VMD software (30). Analyses of the root mean square (RMSD) deviation and radius of gyration (R_g) have been performed with the *cpptraj* module of Ambertools 16 (31) (Fig. S4).

Principal Component Analysis (PCA). PCA was performed on the stripped trajectories (1 frame each 100 ps) with *cpptraj* module of Ambertools 16 (31) to extract

the ‘essential dynamics’ of the ILS complex. Indeed, PCA can report on the large-scale, collective motions occurring in biological macromolecules undergoing MD simulations, thus providing valuable information on major conformational changes occurring along MD trajectories (32, 33). Here, the essential motions of proteins and RNAs have been captured starting from the mass-weighted covariance matrix of the Cα and P atoms, respectively. The covariance matrices were constructed from the atoms position vectors upon an RMS-fit to the reference starting configuration of the MD production run in order to remove the rotational and translational motions. Each element in the covariance matrix is the covariance between atoms *i* and *j,* defining the *i,j* position of the matrix. The covariance C_{ij} is defined as:

$$C_{ij} = \langle (\vec{r}_i - \langle \vec{r}_i \rangle)(\vec{r}_j - \langle \vec{r}_j \rangle) \rangle \quad (1)$$

where $\vec{r}_i$ and $\vec{r}_j$ are the position vectors of atoms i and j, and the brackets denote an average over the sampled time period. For ILS-1 the matrix was calculated on 3833 Cα and 255 P atoms over 6700 frames, corresponding to last 670 ns of the MD simulations. For ILS-2 the matrix was derived from 5207 Cα and 255 P atoms over 9200 frames, corresponding to last 920 ns of the MD production run. The two terms in Eq. 1 represent the displacement vectors for atoms *i* and *j*. A positive sign of this product indicates that the two atoms move in a correlated manner, otherwise, a negative value points to an anti-correlated motion between the two atoms. If the product is zero, then it evinces that the atoms displacements are independent of each other. The covariance matrix was then diagonalized, leading to a complete set of orthogonal collective eigenvectors, each associated to a corresponding eigenvalue. The eigenvalues denote how much each eigenvector is representative of the system dynamics, thus giving a measure of the

contribution of each eigenvector to the total variance. Indeed, the eigenvectors with the largest eigenvalues correspond to the most relevant motions. By projecting the displacements vectors of each atom along the trajectory onto the eigenvectors (i.e., by taking the dot product between the two vectors at each frame), the Principal Components (PC) were then obtained. A total of 6700 and 9200 frames were used for the PCA of ILS-1 and ILS-2, respectively, with the maximum number of eigenvalues given by *min* (3 × n°-of-atoms, n°-of-frames) = 6700 PCs (ILS-1) and 9200 PCs (ILS-2). The cumulative variance accounted by all the PCs was calculated both for ILS-1 and ILS-2 (Fig. S9). Subsequently, for each replica, PC1 was plotted against PC2 to generate the scatter plot displaying how the conformational space defined by the first two modes is sampled through the MD simulations (Fig. S8). The Normal Mode Wizard plugin (34) of VMD was used to visualize the essential dynamics along the principal eigenvectors and to draw the arrows highlighting their direction.

Correlation scores (*CSs*). The cross-correlation matrices (or normalized covariance matrices) based on the Pearson's correlation coefficients (CC_{ij}) were calculated with the *cpptraj* module of Ambertools 16 (22) from the covariance matrices previously obtained (Figs. S11-S14). The cross-correlation analysis offers the possibility to qualitatively capture the linear coupling of the motions between two residues over the entire trajectory. Each element of the cross-correlation matrix in the *i,j* position corresponds to a Pearson's CC_{ij}, i.e. the normalized covariance between atoms *i* and *j* (Cα atoms in case of proteins and P atoms in case of RNAs), calculated with the formula:

$$CC_{ij} = \frac{\langle (\vec{r}_i - \langle \vec{r}_i \rangle)(\vec{r}_j - \langle \vec{r}_j \rangle) \rangle}{\left[(\langle \vec{r}_i^{\,2} \rangle - \langle \vec{r}_i \rangle^2)(\langle \vec{r}_j^{\,2} \rangle - \langle \vec{r}_j \rangle^2) \right]^{1/2}} \qquad (2)$$

where the normalization factor at the denominator is the product between the standard deviations of the two position vectors. CC_{ij} range from a value of -1, which indicates a totally anti-correlated motion between two atoms, and a value of +1, which instead means a linearly correlated lockstep motion. In line with other studies (35-37), in order to make the correlation matrices more explicit, allowing a prompt interpretation of the most important functional motions taking place in our simulations, we have calculated the correlation scores (*CSs*) between each SPL component and all the others (35). In particular, each CS_{IJ} between a protein/RNA *I* and a protein/RNA *J* was calculated as:

$$CS_{IJ} = \sum_{\substack{i \in I \\ j \in J}}^{N} CC_{ij} \tag{3}$$

which sums the CC_{ij} between the residues/nucleotides *i* belonging to the protein/RNA *I* and the residues/nucleotides *j* belonging to the protein/RNA *J*. When I = J, the *CS* is intended as an *intra*-correlation score, while in the case of I ≠ J the *CS* is meant as an *inter*-correlation score. Importantly, due to its large size and to better characterize its critical role, we separately treated the Spp42 domains and some peculiar motifs of the N-terminal domain (N-t). As such, for each component (i.e., proteins, Spp42 domains/N-t motifs, RNAs) one *intra*- and ($M - 1$) *inter*-correlation scores were computed, where *M* is the number of SPL components. Importantly, in ILS-1 the values $-0.6 < CC_{ij} < +0.6$ were discarded in the reckoning of the scores, in order to eliminate the noise due to uncorrelated motions. In ILS-2 we applied a less strict criterion (i.e., $-0.4 < CC_{ij} < +0.4$) since this model has shown less evident coupled motions. Our aim, indeed, was to spotlight from these matrices only the most relevant correlated and anti-correlated motions between two SPL components to further inspect a possible biological function

linked with their dynamics. All the *CSs* obtained for each SPL component have been normalized by the highest score (in absolute value) registered for that specific component. This has reduced all the scores to values ranging from -1 to +1, getting rid of the bias due to the different sizes of the macromolecule considered (i.e., larger macromolecule, higher score). Subsequently, the normalized scores were plotted in a histogram showing the correlation/anti-correlation motions between each pair of SPL components (Main text, Fig. 1b for ILS-1 and Fig. 5b for ILS-2).

Electrostatic calculations. Electrostatic calculations were performed on the proteins included in ILS-1 and ILS-2 considering the cryo-EM model and configurations harvested at different times along the simulations with the Adaptive Poisson-Boltzmann Solver (APBS1.4) software (38). APBS evaluates the electrostatic properties of large biomolecules by efficiently solving the Poisson-Boltzmann electrostatic equation (PBE) (38). The selected geometries were first converted to the pqr format with pdb2pqr software (39, 40) with unvaried protonation state and by using the same force field employed in the MD simulations. Subsequently, following previous applications (36), APBS calculations were carried out using the Linearized Poisson-Boltzmann Equation (LPBE) with a grid spacing of ~0.7 Å, at 298 K and 150 mM as ionic strength for monovalent ions. The external dielectric constant was set to 78.0 to reproduce the aqueous medium, while the internal dielectric constant was fixed at 2.0 to mimic the non-polar environment of the solute.

2. Supporting Figures

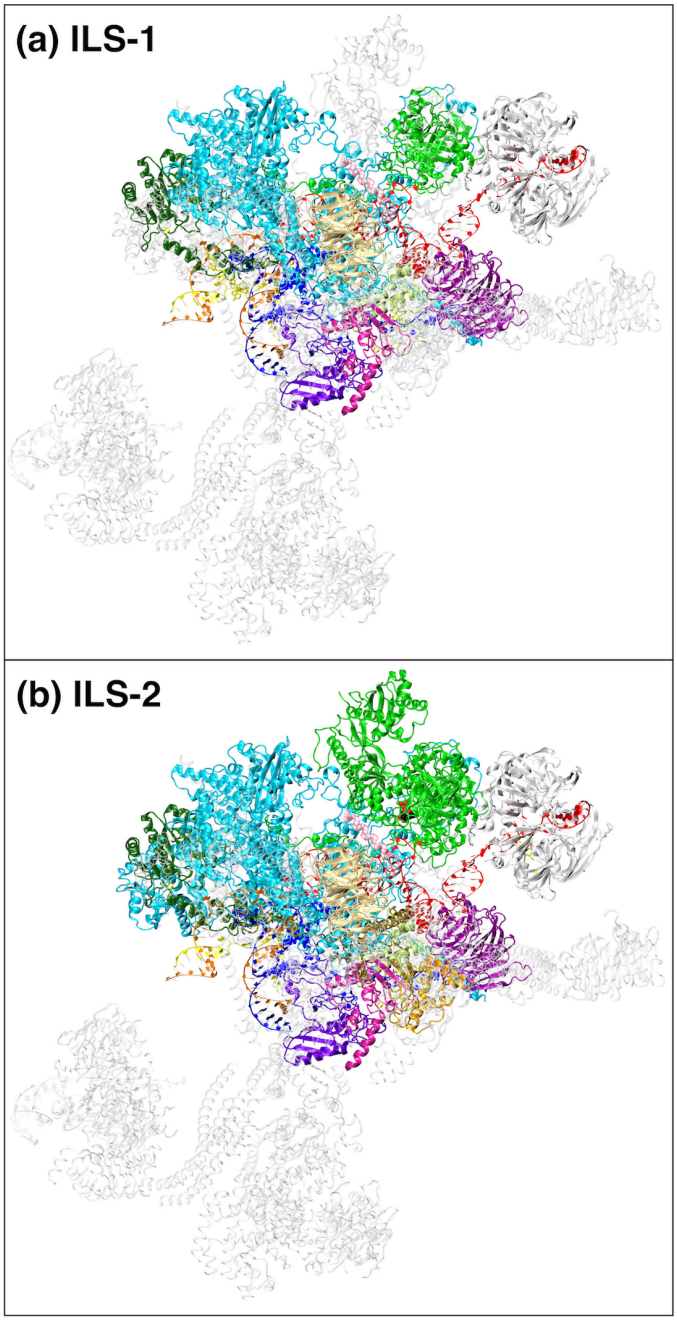


Fig. S1. The core region of the intron lariat spliceosome (ILS) complex. The core region of the spliceosome ILS complex solved from the yeast *S. pombe* (PDB 3JB9) (1, 2) has been considered for this study. Here, our models **(a)** ILS-1 and **(b)** ILS-2 are shown with a colored opaque cartoon (proteins) and ribbons (RNAs) representation, while the low-resolution portions of 3JB9, not included in our models, are depicted with a grey transparent representation.

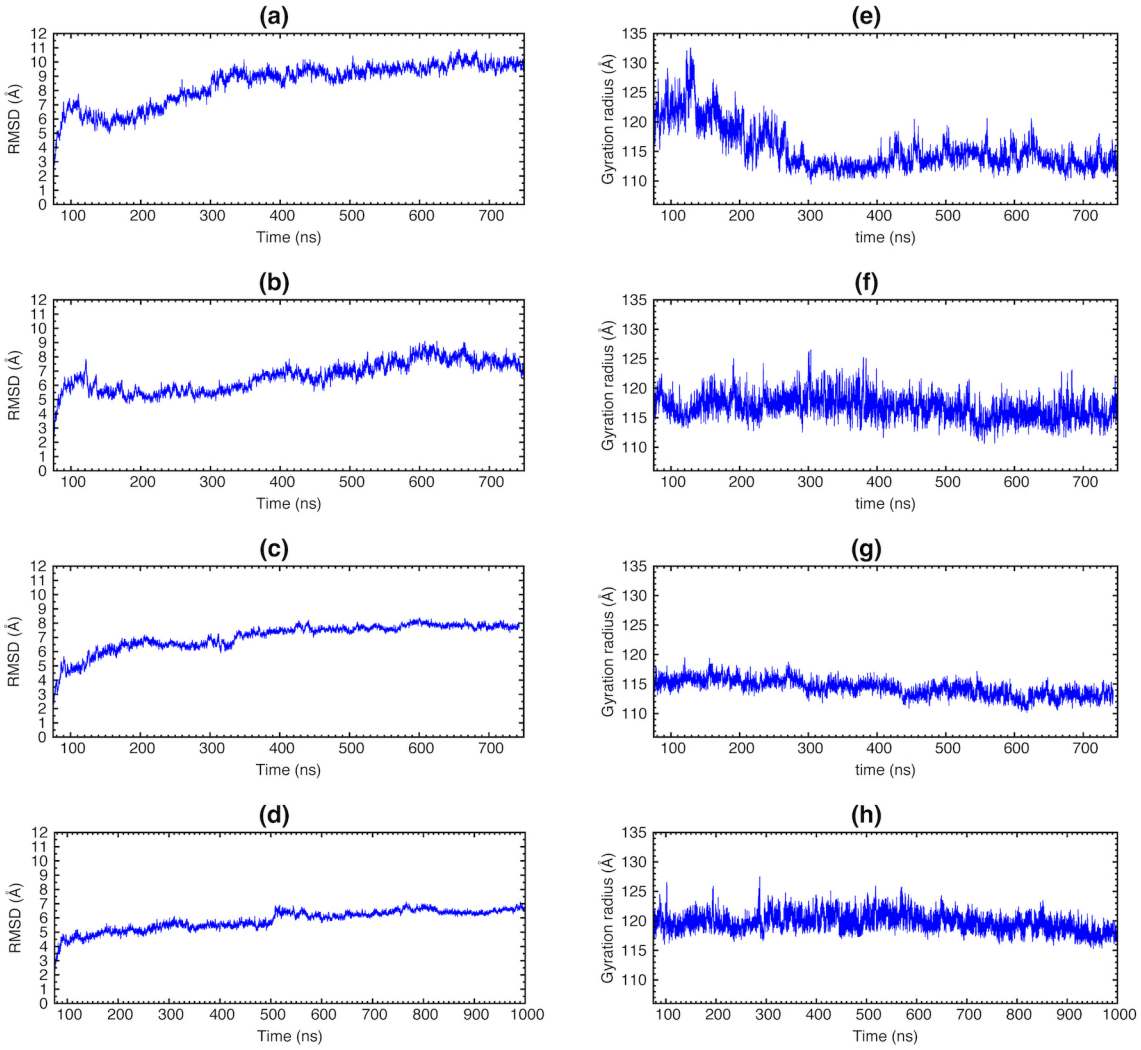


Fig. S2. Root Mean Square Displacement (RMSD) and Radius of Gyration (R_g) profiles. Time evolution (ns) of RMSD **(a, b, c)** and R_g **(e, f, g)** obtained from Molecular Dynamics (MD) replicas of ILS-1 model. In panels **(d)** and **(h)** the time evolution (ns) of RMSD and R_g obtained from the MD replica of ILS-2 model is respectively shown. All the profiles are obtained including all the proteins and RNAs in the analyses.

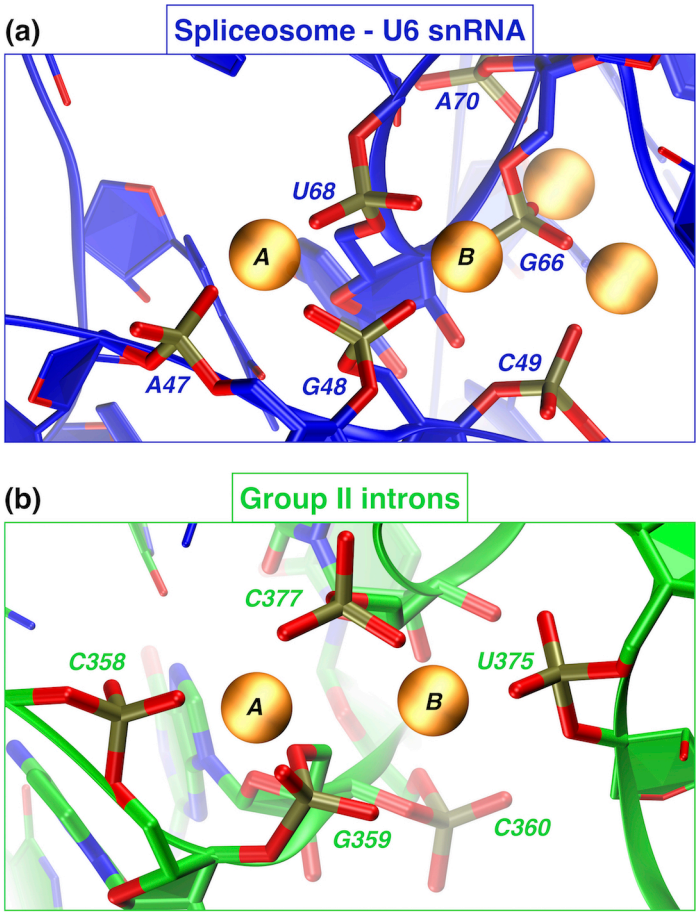


Fig. S3. The hallmark catalytic site upon MD simulations. Comparison between the catalytic site of spliceosome **(a)** and group II introns **(b)** after MD simulations. **(a)** Snapshot representing the active site at the end of MD simulation of ILS-1 model replica #1. Mg^{2+} ions are depicted with orange spheres and those involved in the splicing reaction are indicated with A and B. U6 snRNA is shown as blue ribbons, while the phosphate groups directly involved in the coordination of the Mg^{2+} ions are highlighted with licorice representation. **(b)** Snapshot representing the active site of a group IIC intron, shown as green ribbons, obtained from our recent QM/MM MD simulations (41). Mg^{2+} ions and phosphate groups are shown with orange spheres and licorice representation, respectively.

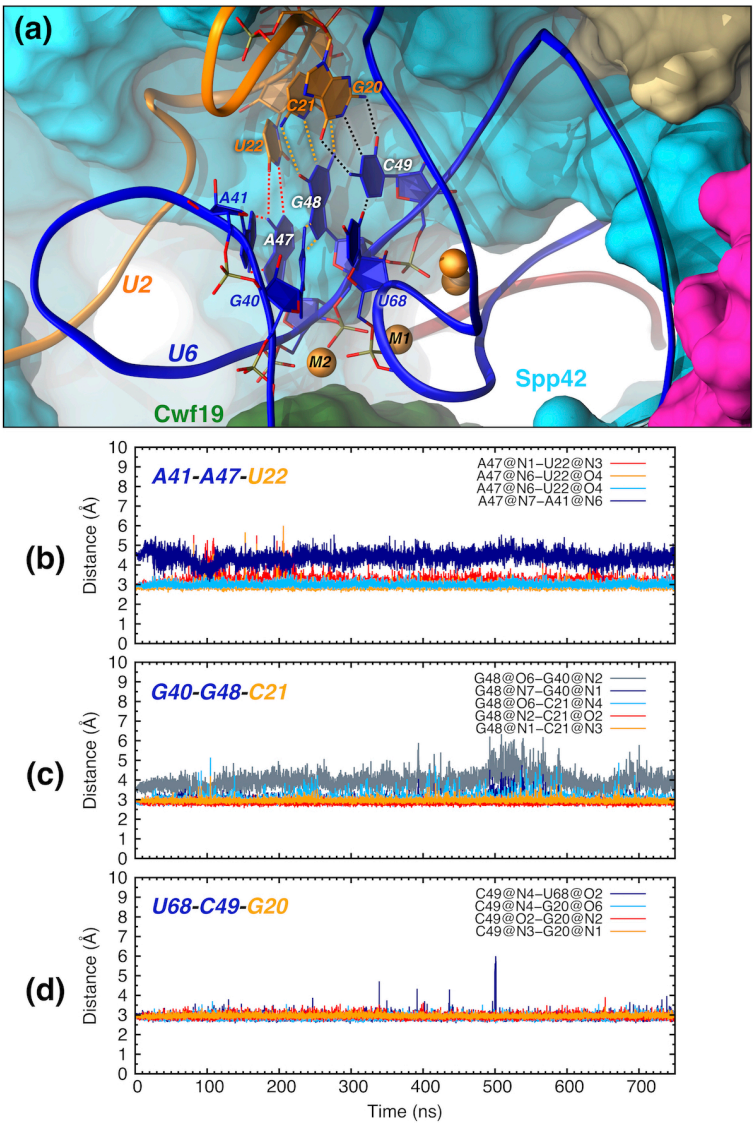


Fig. S4. The triple-helix motif within the spliceosomal catalytic core. (a) Snapshot representing the catalytic site of the ILS-1 model after 750 ns of MD simulations. U2 and U6 snRNAs are represented as orange and blue tubes, respectively. The nucleotides involved in the triple-helix are depicted with licorice representation. Mg^{2+} are represented as orange spheres. The proteins forming the catalytic cavity are shown with a surface representation and highlighted with different colors. Panels **(b)**, **(c)** and **(d)** monitor the base pairs between the nucleotides involved in the triple-helix, i.e. A41-A47-U22, G40-G48-C21 and U68-C49-G20, respectively. The time (ns) evolutions of the hydrogen bonds distances (Å) along the MD replica #1 of ILS-1 model are highlighted with different colors.

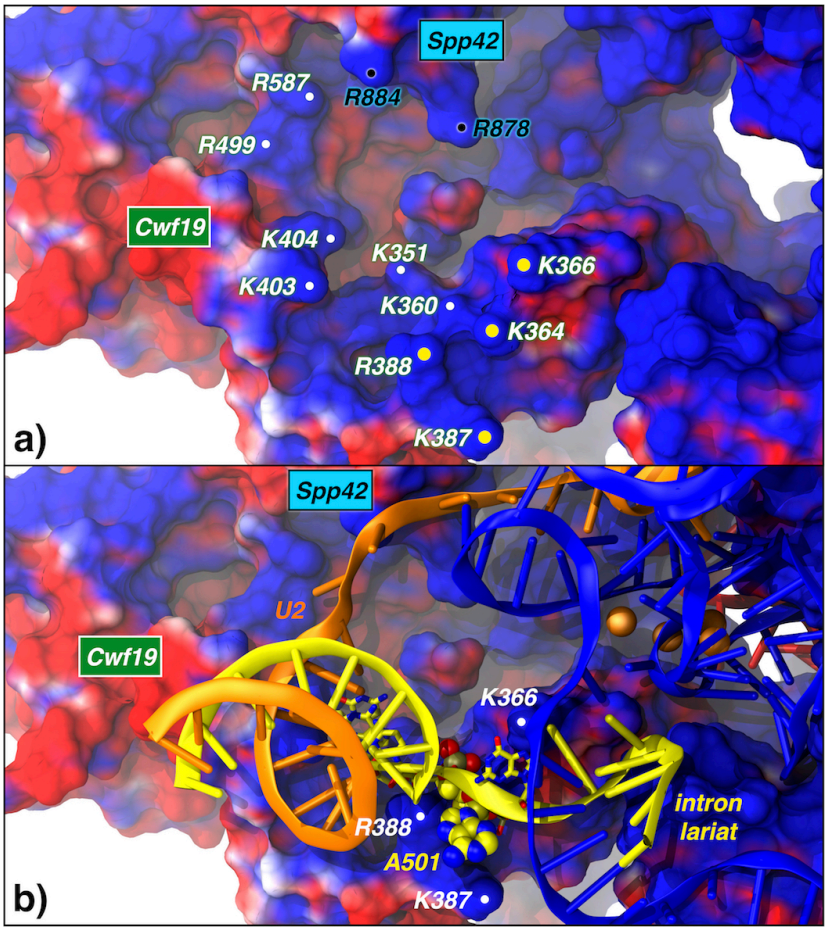


Fig. S5. The positively charged pocket formed by Cwf19 and Spp42. (a) Positively charged cavity formed by Cwf19 and Spp42 as in the cryo-EM structure, represented with the electrostatic surface, with blue and red colors representing positive and negative charges, respectively. The most important positively charged residues are indicated with white (Cwf19) and black (Spp42) labels. K364, K366, K387, R388 are highlighted with a yellow dot as they are in proximity of the branching adenosine. **(b)** U2, U6 snRNAs (orange and blue), and intron lariat (yellow) are also represented, with the branching A501 depicted with van der Waals spheres.

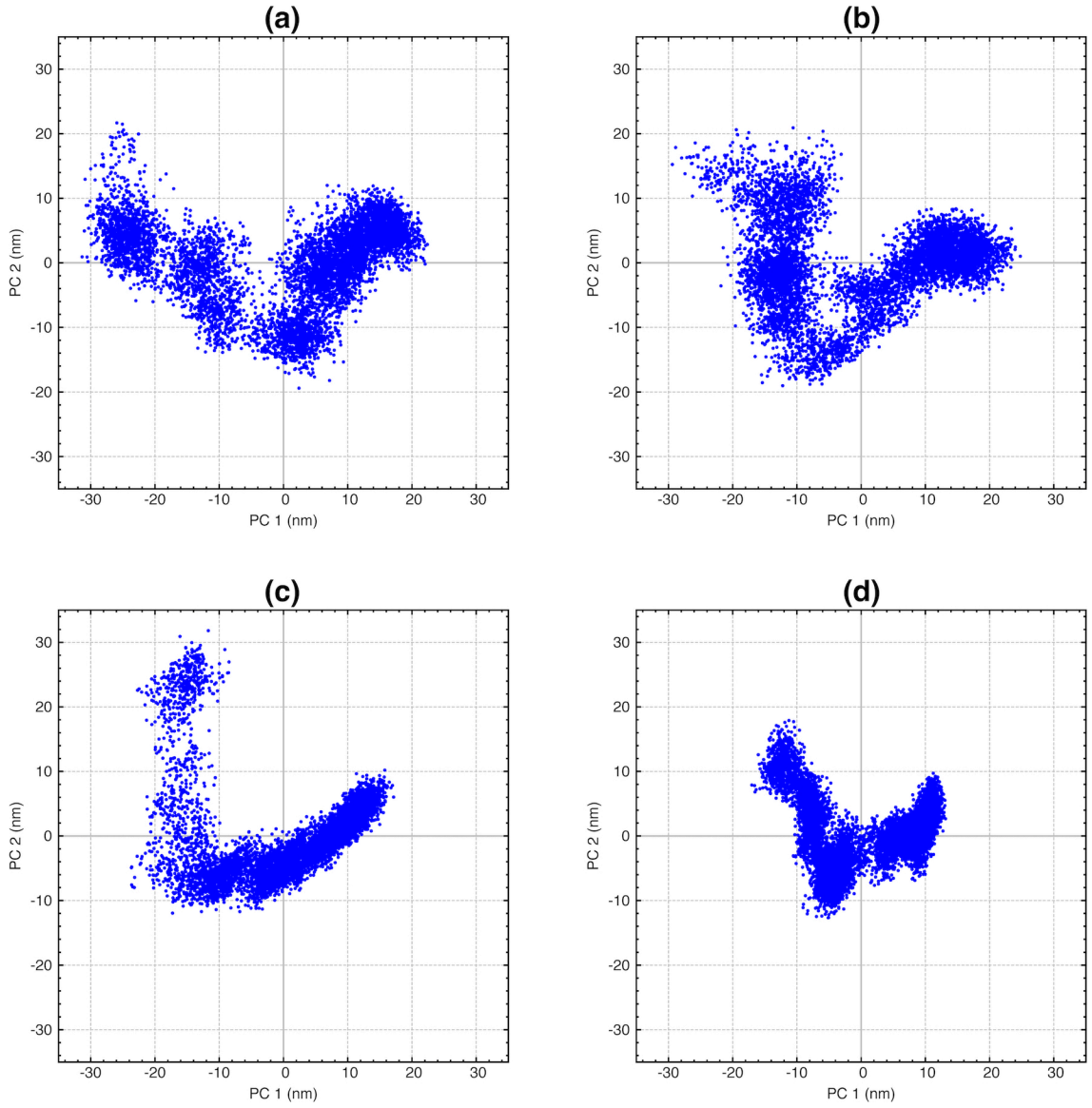


Fig. S6. PCA scatter plots. Scatter plots (PC1 vs. PC2) representing the projections of the Cα and P displacements along the trajectory onto the first principal eigenvector, PC1 (x-axis), vs the projections onto the second principal eigenvector, PC2 (y-axis), as derived from MD replicas of spliceosome ILS-1 **(a, b, c)** and ILS-2 **(d)** models.

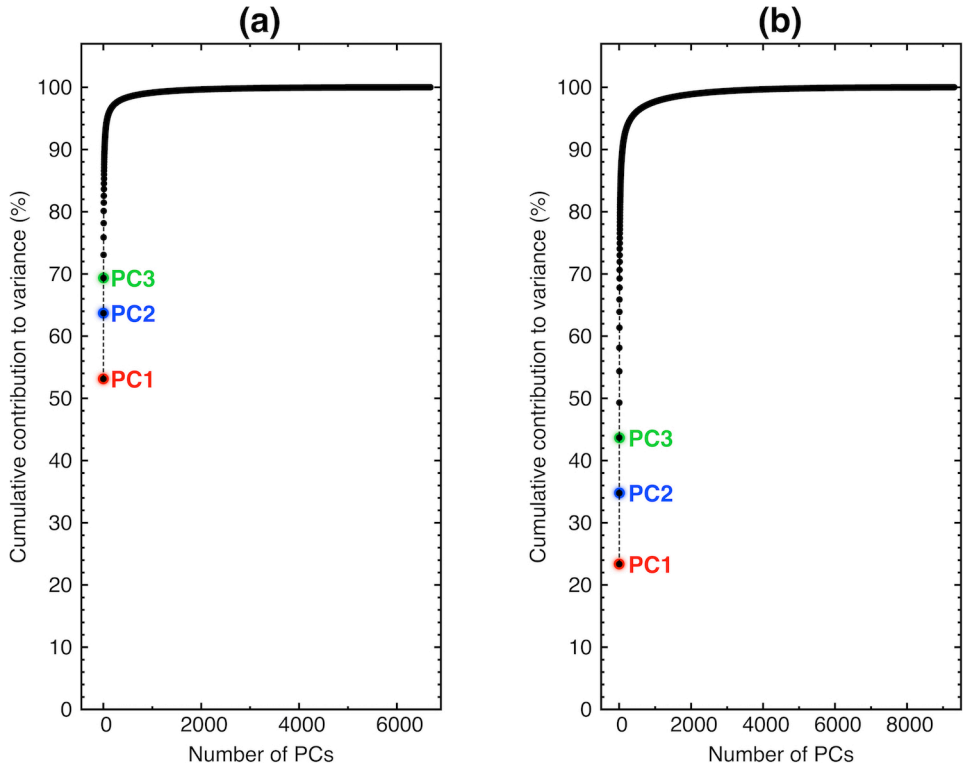


Fig. S7. PCs cumulation contribution to variance. Cumulative contribution (%, y-axis) of all the principal components (PCs, x-axis) to the variance of the overall spliceosome (SPL) motion calculated upon Principal Component Analysis on the SPL ILS-1 **(a)** and ILS-2 **(b)** models. The contribution from the first three PCs are highlighted in red, blue and green, respectively.

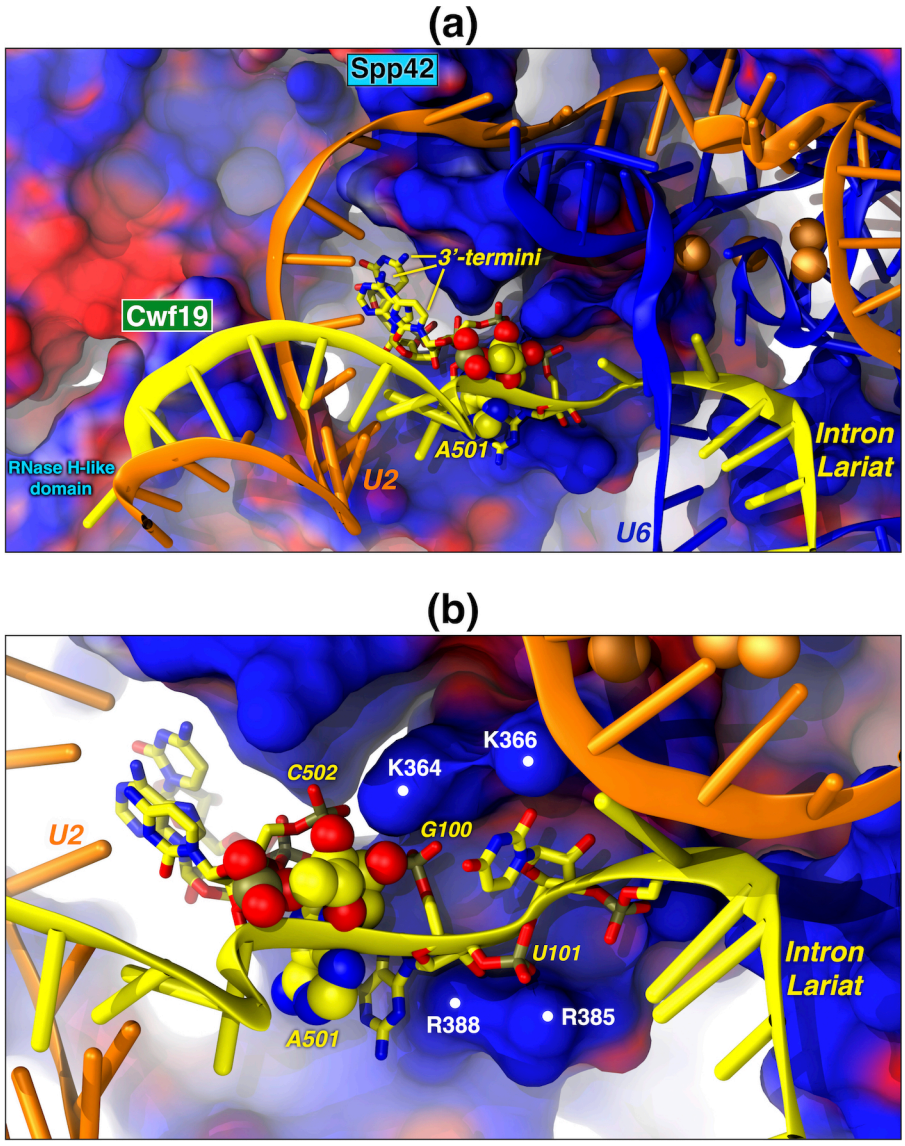


Fig. S8. The polar tweezers anchoring the branch point region in ILS-2. (a) The polar tweezers (magnified view in **(b)**) formed by K364, K366, R388 and also R385 of Cwf19 in ILS-2 model along the MD simulations. Cwf19 and Spp42 are shown with the electrostatic surface (blue and red for positive and negative charges, respectively). U2, U6 snRNAs and the intron lariat (IL) are depicted in orange, blue and yellow cartoons. A501 is shown with van der Waals spheres, and the IL 3′-termini in licorice.

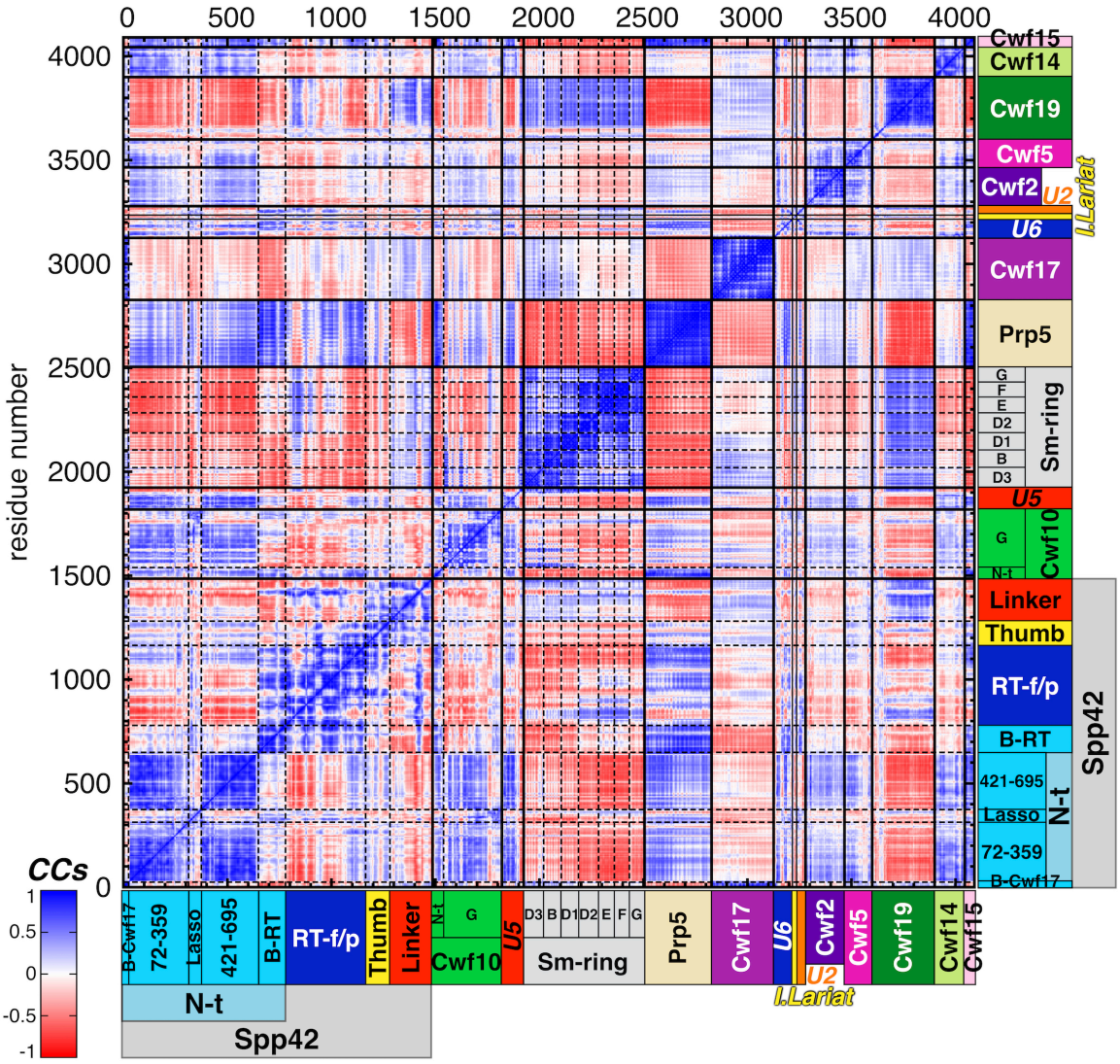


Fig. S9. Pearson's coefficients cross-correlation matrix of ILS-1, replica #2. Pearson's coefficients (*CCs*) cross-correlation matrix derived from the mass-weighted covariance matrix constructed over the last 670 ns of MD simulations of ILS-1 (replica #2) for Cα and P atoms. The Pearson's coefficients are comprised between -1 (anti-correlation, red) and +1 (correlation, blue). Spliceosome components names (proteins and RNAs) are highlighted with different colors and listed. Abbreviations: N-t D, N-terminal Domain; RT-f/p, retro-transcriptase finger/palm.

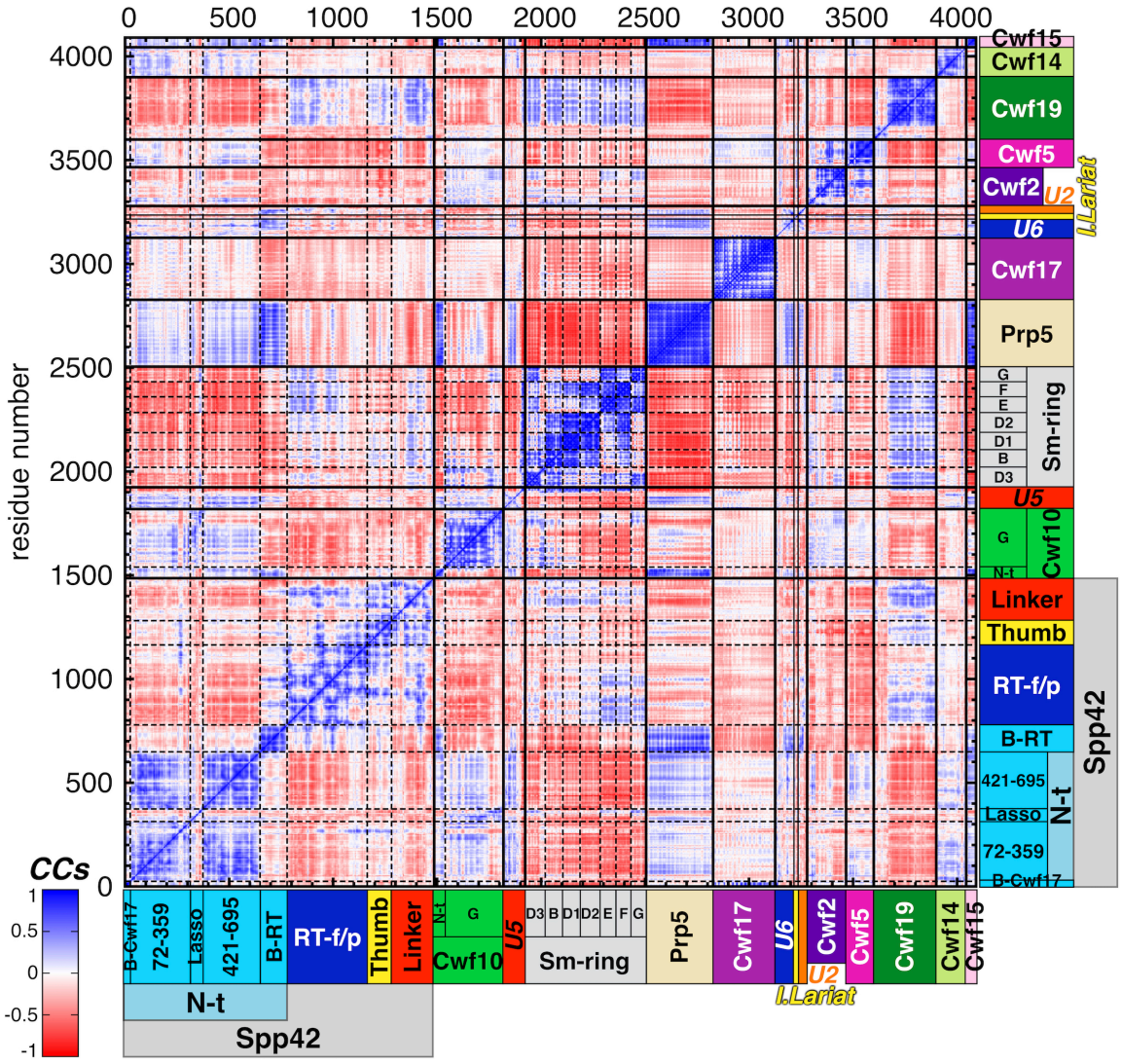


Fig. S10. Pearson's coefficients cross-correlation matrix of ILS-1, replica #3. Pearson's coefficients (*CCs*) cross-correlation matrix derived from the mass-weighted covariance matrix constructed over the last 670 ns of MD simulations of ILS-1 (replica #3) for Cα and P atoms. The Pearson's coefficients are comprised between -1 (anti-correlation, red) and +1 (correlation, blue). Spliceosome components names (proteins and RNAs) are highlighted with different colors and listed. Abbreviations: N-t D, N-terminal Domain; RT-f/p, retro-transcriptase finger/palm.

3. Supporting Tables

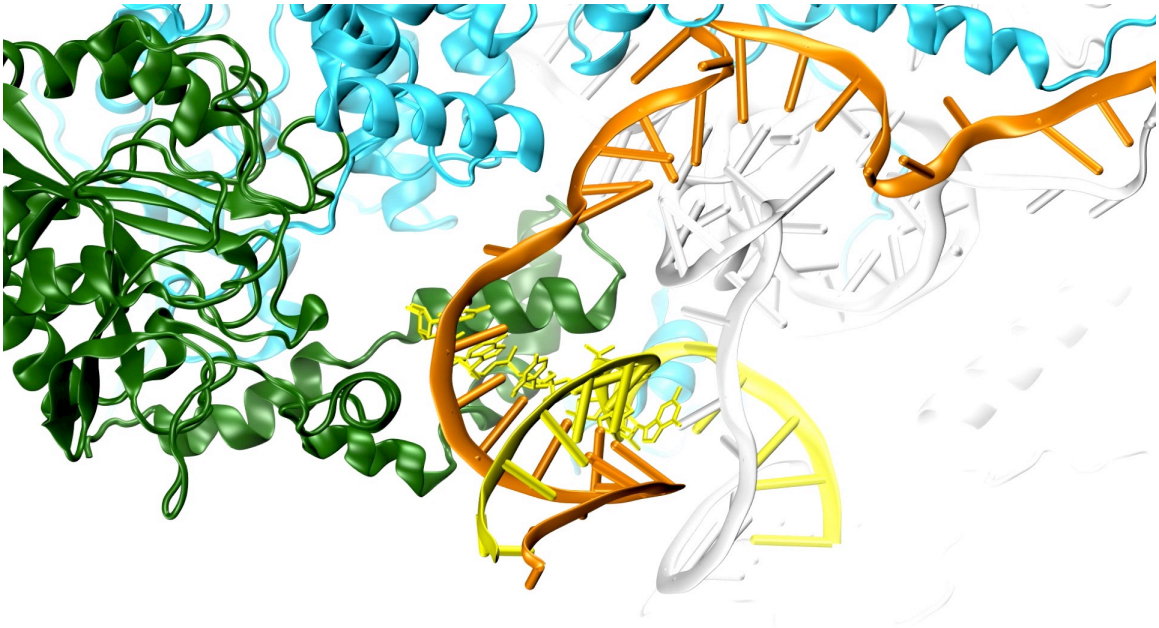
SPLICEOSOME ILS-1 MODEL						
Total number of atoms (water included) = 721089			**Cryo-EM 3.6 Å (3jb9)**	**Organism:** *Schizosaccharomyces Pombe*		
Solute atoms: 70190 atoms, 36501 heavy atoms			**Protein force field:** ff12SB	**RNA force field**: ff99+bsc0+χOL3		
CHAIN	**MOLECULE**	**CONSIDERED**	**MODELLED**	**RESOLUTION**	**VMD RESIDUE**	**N° of RES**
A	**Spp42 (*Prp8*)**	47 - 1532	303 to 313	2.9 ~ 3.6	0 to 1485	1486
B	**Cwf10 (*Snu114*)**	68 - 400	/	2.9 ~ 3.8	1486 to 1818	333
C	**U5 snRNA**	7 - 111	/	2.9 ~ 3.6	1819 to 1923	105
D	**SM-D3**	2 - 97	/	3.3 ~ 4.0	1924 to 2019	581
E	**SM-B**	2 - 86	48 to 59	3.3 ~ 4.0	2020 to 2104	
F	**SM-D1**	1 - 82	/	3.3 ~ 4.0	2105 to 2186	
G	**SM-D2**	19 - 115	85 to 86	3.3 ~ 4.0	2187 to 2283	
H	**SM-E**	9 - 84	/	3.3 ~ 4.0	2284 to 2359	
I	**SM-F**	4 - 75	/	3.3 ~ 4.0	2360 to 2431	
J	**SM-G**	3 - 75	/	3.3 ~ 4.0	2432 to 2504	
K	**Prp5**	149 - 470	/	~ 3.4	2505 to 2826	322
L	**Cwf17**	42 - 340	81, 147, 250 to 253	3.3 ~ 4.0	2827 to 3125	299
N	**U6 snRNA**	1 - 90	/	2.9 ~ 4.5	3126 to 3215	90
O + Q	**intron lariat**	100 - 107 + 492 - 504	"GA2" = A501 bonded to G100	/	3216 to 3235	20
P	**U2 snRNA**	1 - 43	/	2.9 ~ 4.5	3236 to 3278	43
Y	**Cwf2**	49 - 235	/	3.3 ~ 5.0	3279 to 3465	187
a	**Cwf5**	18 - 151	/	3.3 ~ 4.0	3466 to 3599	134
c	**Cwf19**	334 - 633	/	3.4 ~ 4.0	3600 to 3899	300
e	**Cwf14**	3 - 146	/	~ 3.4	3900 to 4043	144
h	**Cwf15**	24 - 70	/	3.0 ~ 4.0	4044 to 4090	47
	Mg+	# 4	Aqvist force field	/	4091 to 4094	4
	ZNB (Zn2+)	# 7 (35 atoms)	Pang dummy cations force field	/	4095 to 4101	7
	Na+	# 202	Joung & Cheatham force field	/	4102 to 4303	202
	Wat	# 216886	TIP3P	/	4304 to 221189	216886

Table S1. System details of the spliceosome model ILS-1. CHAIN refers to the chain name as reported in the original PDB structure 3JB9 (1, 2). MOLECULE refers to the names of all the included spliceosome components, ions and other molecules (in parenthesis the corresponding name for *S. cerevisiae* are reported). CONSIDERED indicates the regions of the cryo-EM structure that were included in ILS-1 model, with the residue number as in the original PDB. MODELLED lists the residues that were *de-novo* modelled by us using Modeller 9v16 (5). RESOLUTION reports the resolutions by which the considered molecules were reconstructed in the original PDB. VMD RESIDUE indicates the numeration adopted for visualization and analysis with Visual Molecular Dynamics software (VMD). N° of RES lists the number of residues/nucleotides for each specific spliceosome component included in our ILS-1 model. The force fields used for proteins, RNAs, ions and water molecules are listed.

SPLICEOSOME ILS-2 MODEL						
Total number of atoms (water included) = 914099			**Cryo-EM 3.6 Å (3jb9)**	***Organism:*** *Schizosaccharomyces Pombe*		
Solute atoms: 92276 protein and RNA atoms, 47510 heavy atoms			**Protein force field**: ff12SB	**RNA force field**: ff99+bsc0+χOL3		
CHAIN	**MOLECULE**	**CONSIDERED**	**MODELLED**	**RESOLUTION**	**VMD RESIDUE**	**N° of RES**
A	**Spp42 (*Prp8*)**	47 - 2030	303-313/1533-1538/1781-1783	2.9 ~ 3.6	0 to 1983	1984
				~4.0 (Rnase H)		
B	**Cwf10 (*Snu114*)**	68 - 971	/	2.9 ~ 3.8	1984 to 2887	904
C	**U5 snRNA**	7 - 111	/	2.9 ~ 3.6	2888 to 2992	105
D	**SM-D3**	2 - 97	/	3.3 ~ 4.0	2993 to 3088	581
E	**SM-B**	2 - 86	48 to 59	3.3 ~ 4.0	3089 to 3173	
F	**SM-D1**	1 - 82	/	3.3 ~ 4.0	3174 to 3255	
G	**SM-D2**	19 - 115	85 to 86	3.3 ~ 4.0	3256 to 3352	
H	**SM-E**	9 - 84	/	3.3 ~ 4.0	3353 to 3428	
I	**SM-F**	4 - 75	/	3.3 ~ 4.0	3429 to 3500	
J	**SM-G**	3 - 75	/	3.3 ~ 4.0	3501 to 3573	
K	**Prp5**	149 - 470	/	~ 3.4	3574 to 3895	322
L	**Cwf17**	42 - 340	81, 147, 250 to 253	3.3 ~ 4.0	3896 to 4194	299
N	**U6 snRNA**	1 - 90	/	2.9 ~ 4.5	4195 to 4284	90
O + Q	**intron lariat**	100 - 107 + 492 - 504	"GA2" = A501 bonded to G100	/	4285 to 4304	20
P	**U2 snRNA**	1 - 43	/	2.9 ~ 4.5	4305 to 4347	43
Y	**Cwf2**	49 - 235	/	3.3 ~ 5.0	4348 to 4534	187
a	**Cwf5**	18 - 151	/	3.3 ~ 4.0	4535 to 4668	134
c	**Cwf19**	334 - 633	/	3.4 ~ 4.0	4669 to 4968	300
e	**Cwf14**	3 - 146	/	~ 3.4	4969 to 5112	144
h	**Cwf15**	24 - 70	/	3.0 ~ 4.0	5113 to 5159	47
M	**Prp45**	100 - 271	/	3.0 ~ 4.5	5160 to 5331	172
g	**Prp17**	29 - 161	/	3.3 ~ 4.5	5332 to 5464	133
	Mg+	# 4	Aqvist force field	/	5465 to 5468	4
	ZNB (Zn2+)	# 7 (35 atoms)	Pang dummy cations force field	/	5469 to 5475	7
	GDP	#1 (40 atoms)	Meagher KL force field	/	5476	1
	Na+	# 194	Joung & Cheatham force field	/	5477 to 5670	194
	Water	# 273850	TIP3P	/	5671 to 279520	273850

Table S2. System details of the spliceosome model ILS-2. CHAIN refers chain name as in the original PDB 3JB9 (1, 2). MOLECULE refers to the names of all the included spliceosome components, ions and other molecules are listed (in parenthesis the corresponding name for *S. cerevisiae*). CONSIDERED indicates the regions of the cryo-EM structure that were included in ILS-2 model, with the residue number as in the original PDB. MODELLED lists the residues that were *de-novo* modelled by us using Modeller 9v16 (5). RESOLUTION reports the resolutions by which the considered molecules were reconstructed in the original PDB. VMD RESIDUE indicates the numeration adopted for visualization and analysis with Visual Molecular Dynamics software (VMD). N° of RES lists the number of residues/nucleotides for each specific spliceosome component included in our ILS-2 model. The force fields used for proteins, RNAs, ions, GDP and water molecules are listed.

4. Supporting Movie



Movie S1. Principal Component Analysis applied to Molecular Dynamics trajectories of discloses the displacement of the branch helix formed by the intron lariat (yellow, cartoon representation) and the U2 snRNA (orange, cartoon representation). The motion along the first eigenvector is shown as observed in the ILS-1 model, with Cwf19 (green, cartoon representation) and Spp42 (cyan, cartoon representation) playing a crucial role.

5. References

1. Yan C, *et al*. (2015) Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science* 349(6253):1182-1191.
2. Hang J, Wan R, Yan C, & Shi Y (2015) Structural basis of pre-mRNA splicing. *Science* 349(6253):1191-1198.
3. Nguyen TH, *et al*. (2016) CryoEM structures of two spliceosomal complexes: starter and dessert at the spliceosome feast. *Curr Opin Struct Biol* 36:48-57.
4. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, & Klein ML (1983) Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys* 79(2):926-935.
5. Sali A & Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234(3):779-815.
6. Fiser A, Do RK, & Sali A (2000) Modeling of loops in protein structures. *Protein Sci* 9(9):1753-1773.
7. Fiser A & Sali A (2003) ModLoop: automated modeling of loops in protein structures. *Bioinformatics* 19(18):2500-2501.
8. Shen MY & Sali A (2006) Statistical potential for assessment and prediction of protein structures. *Protein Sci* 15(11):2507-2524.
9. Jamroz M & Kolinski A (2010) Modeling of loops in proteins: a multi-method approach. *BMC Struct Biol* 10:5.
10. Van der Spoel D, *et al*. (2005) GROMACS: Fast, flexible, and free. *J Comput Chem* 26(16):1701-1718.
11. Maier JA, *et al*. (2015) ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* 11(8):3696-3713.
12. Perez A, *et al*. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys J* 92(11):3817-3829.
13. Zgarbova M, *et al*. (2011) Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J Chem Theory Comput* 7(9):2886-2902.
14. Sponer J, *et al*. (2017) How to understand atomistic molecular dynamics simulations of RNA and protein-RNA complexes? *Wiley Interdisciplinary Reviews: RNA* 8(3).
15. Aqvist J (1990) Ion Water Interaction Potentials Derived from Free-Energy Perturbation Simulations. *J Phys Chem* 94(21):8021-8024.
16. Casalino L, Palermo G, Abdurakhmonova N, Rothlisberger U, & Magistrato A (2017) Development of Site-Specific Mg2+-RNA Force Field Parameters: A Dream or Reality? Guidelines from Combined Molecular Dynamics and Quantum Mechanics Simulations. *J Chem Theory Comput* 13(1):340-352.
17. Joung IS & Cheatham TE, III (2008) Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J Phys Chem B* 112(30):9020-9041.
18. Pang YP (1999) Novel zinc protein molecular dynamics simulations: Steps toward antiangiogenesis for cancer treatment. *J Mol Model* 5(10):196-202.

19. Meagher KL, Redman LT, & Carlson HA (2003) Development of polyphosphate parameters for use with the AMBER force field. *J Comput Chem* 24(9):1016-1025.
20. Singh UC & Kollman PA (1984) An Approach to Computing Electrostatic Charges for Molecules. *J Comput Chem* 5(2):129-145.
21. Frisch MJ, *et al*. (2009) Gaussian 09 (Gaussian, Inc., Wallingford, CT, USA).
22. Case DA, *et al*. (2012) AMBER 12 (University of California, San Francisco, San Francisco, CA).
23. Sousa da Silva AW & Vranken WF (2012) ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Res Notes* 5:367.
24. Bussi G, Donadio D, & Parrinello M (2007) Canonical sampling through velocity rescaling. *J Chem Phys* 126(1).
25. Parrinello M & Rahman A (1980) Crystal Structure and Pair Potentials: a Molecular-Dynamics Study. *Phys Rev Lett* 45(14):1196-1199.
26. Parrinello M & Rahman A (1981) Polymorphic Transitions in Single-Crystals - a New Molecular-Dynamics Method. *J Appl Phys* 52(12):7182-7190.
27. Hess B, Bekker H, Berendsen HJC, & Fraaije JGEM (1997) LINCS: A linear constraint solver for molecular simulations. *J Comput Chem* 18(12):1463-1472.
28. Darden T, York D, & Pedersen L (1993) Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J Chem Phys* 98(12):10089-10092.
29. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, & Haak JR (1984) Molecular dynamics with coupling to an external bath. *J Chem Phys* 81(8):3684-3690.
30. Humphrey W, Dalke A, & Schulten K (1996) VMD: Visual molecular dynamics. *Journal of Molecular Graphics & Modelling* 14(1):33-38.
31. Case DA, *et al*. (2016) AMBER 2016 (University of California, San Francisco, San Francisco, CA).
32. David CC & Jacobs DJ (2014) Principal component analysis: a method for determining the essential dynamics of proteins. *Methods Mol. Biol*. 1084:193-226.
33. Amadei A, Linssen ABM, & Berendsen HJC (1993) Essential Dynamics of Proteins. *Proteins: Structure, Function, and Genetics* 17(4):412-425.
34. Bakan A, Meireles LM, & Bahar I (2011) ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics* 27(11):1575-1577.
35. Palermo G, Miao Y, Walker RC, Jinek M, & McCammon JA (2016) Striking Plasticity of CRISPR-Cas9 and Key Role of Non-target DNA, as Revealed by Molecular Simulations. *ACS Cent Sci* 2(10):756-763.
36. Palermo G, *et al*. (2017) Protospacer Adjacent Motif-Induced Allostery Activates CRISPR-Cas9. *J Am Chem Soc* 139(45):16028-16031.
37. Pavlin M, *et al*. (2018) A Computational Assay of Estrogen Receptor α Antagonists Reveals the Key Common Structural Traits of Drugs Effectively Fighting Refractory Breast Cancers. *Sci Rep* 8(1):649.
38. Baker NA, Sept D, Joseph S, Holst MJ, & McCammon JA (2001) Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc Natl Acad Sci USA* 98(18):10037-10041.

39. Dolinsky TJ, *et al*. (2007) PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res* 35:W522-W525.
40. Dolinsky TJ, Nielsen JE, McCammon JA, & Baker NA (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res* 32:W665-W667.
41. Casalino L, Palermo G, Rothlisberger U, & Magistrato A (2016) Who Activates the Nucleophile in Ribozyme Catalysis? An Answer from the Splicing Mechanism of Group II Introns. *J Am Chem Soc* 138(33):10374-10377.