Chapter 1

Appendix E: Supplementary Material on the SARS-CoV-2 nucleocapsid protein

This chapter is adapted from the following publication:

Cubuk, J., Alston, J.J., Incicco, J.J., Singh, S., Stuchell-Brereton, M.D., Ward, M.D., Zimmerman, M.I., Vithani, N., Griffith, D., Wagoner, J.A., Bowman, G.R., Hall, K.B., Soranno, A., The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA, Available on Biorxiv: https://doi.org/10.1101/2020.06.17.158121 [1]

1.1 Supplementary Methods

1.1.1 Sequence analysis

Disorder prediction was performed using IUPred, with additional analysis and sequence parsing done with localCIDER and protfasta, respectively [2–4]

Amino acid sequence of the N protein used in simulations. Highlighted regions delineate folded domains. Underline bolded residues identify the sites of dyes for single-molecule fluorescence experiments.

```
1   MSDNGPQNQR NAPRITFGGP SDSTGSNQNG ERSGARSKQR RPQGLPNNTA
51  SWFTALTQHG KEDLKFPRGQ GVPINTNSSP DDQIGYYRRA TRRIRGGDGK
101 MKDLSPRWYF YYLGTGPEAG LPYGANKDGI IWVATEGALN TPKDHIGTRN
151 PANNAAIVLQ LPQGTTLPKG FYAEGSRGGS QASSRSSSRS RNSSRNSTPG
201 SSRGTSPARM AGNGGDAALA LLLLDRLNQL ESKMSGKGQQ QQGQTVTKKS
251 AAEASKKPRQ KRTATKAYNV TQAFGRRGPE QTQGNFGDQE LIRQGTDYKH
301 WPQIAQFAPS ASAFFGMSRI GMEVTPSGTW LTYTGAIKLD DKDPNFKDQV
351 ILLNKHIDAY KTFPPTEPKK DKKKKADETQ ALPQRQKKQQ TVTLLPAADL
401 DDFSKQLQQS MSSADSTQA
```

1.1.2 Simulation Methods

Monte Carlo simulations

All simulations were performed at 330 K and at 15 mM NaCl, as have been used previously in a variety of systems [2, 5, 5–7]. Simulation analysis was performed with MDTraj and camparitraj (http://ctraj.com/) [**?**, 8]. For IDR

only simulations all degrees of freedom were fully sampled (backbone and sidechain dihedral angles and rigid-body positions) as is standard in CAMPARI Monte Carlo simulations [9]. For simulations of IDRs in the context of folded domains, the backbone dihedral angles of the folded domains were held fixed while all sidechains were fully sampled, as were backbone dihedral angles for the disordered regions, as applied previously [10]. The folded state starting structures were obtained from PDB structures (listed below).

For IDR-only simulations 30-40 independent simulations were run generating final ensembles of 40-60 K conformations. For simulations of IDRs in the context of folded domains, the number of independent simulations and the length of the simulation varied. For the NTD-RBD simulations 400 independent simulations were run, with 2 independent simulations per starting seed from MD simulations (see methods below) leading to a final ensemble of $\sim$400 K conformations (24 M steps per simulation). For the RBD-LINK-dimerization construct, ten independent simulations were run for a final ensemble of 32 K conformers (66 M steps per simulation). For the dimerization-CTD construct 40 independent simulations were run providing a final ensemble of 40 K conformations (66 M steps per simulation). For a complete description of simulation details see Table 1.4.

For the NTD-RB construct, we used a sequential sampling approach in which long timescale MD simulations of the RBD in isolation performed on the Folding@home distributed computing platform were first used to generate hundreds of starting conformations [11, 12]. Those RBD conformations were then used as starting structures for independent all-atom Monte Carlo simulations. Monte Carlo simulations were performed with the ABSINTH forcefield in which the RBD backbone dihedral angles are held fixed but the NTD is fully sampled, as are RBD sidechains. The RBD starting structure used was extracted from the 6VYO PDB crystal structure, which is equivalent to the 6YI3 NMR structure.

For RBD-Link-dimerization domain simulations, we opted to use a single starting seed structure for the folded domains based on the NMR and crystal-structure conformations for the RBD and dimerization domains, respectively. To generate the monomeric starting structure of the dimerization domain, we first built a homology model of the SARS-CoV-2 dimerization dimer from the NMR structure of the SARS dimerization structure (PDB: 2jw8) using SWISS-MODEL [13, 14]. We chose this strategy because at the time, no dimerization structure existed, a situation that has since resolved itself [15]. Nevertheless, the SARS and SARS-CoV-2 dimerization domains are essentially identical, such that this is a minor detail.

For dimerization domain-CTD simulations, a single starting structure for the dimerization domain was again used, selected after MD simulations. Having generated a homology model, we extracted a single protomer from the dimeric structure and ran molecular dynamics simulations to identify equilibrated starting structures. In running initial simulations we discovered that as a monomer, the first 21 residues appear disordered, in agreement with sequence predictions (Fig. 1.1A) but in contrast to their behavior in the dimeric structure (Fig. 1.1C). As a result, we choose to also model these residues as fully disordered. A single starting seed conformation was used for all dimerization-CTD simulations.

Excluded volume (EV) simulations were performed using the same setup, but with a modified Hamiltonian under which solvation, attractive Lennard-Jones, and polar (charge) interactions are scaled to zero, as described previously [16].

Molecular dynamics simulations

All molecular dynamics simulations of SARS-CoV-2 nucleoprotein were performed with Gromacs 2019 using the AMBER03 force field with explicit TIP3P solvent [17–19]. Simulations were prepared by placing the starting structure (PDB ID: 6VYO) in a dodecahedron box that extends 1.0 Å beyond the protein in any dimension. The system was then solvated (29125 atoms), and energy minimized with a steepest descents algorithm until the maximum force fell below 100 kJ/mol/nm using a step size of 0.01 nm and a cutoff distance of 1.2 nm for the neighbor list, Coulomb interactions, and van der Waals interactions. For production runs, all bonds were constrained with the LINCS algorithm and virtual sites were used to allow a 4 fs time step [20, 21]. Cutoffs of 1.1 nm were used for the neighbor list with 0.9 for Coulomb and van der Waals interactions. The Verlet cutoff scheme was used for the neighbor list. The stochastic velocity rescaling (v-rescale) thermostat was used to hold the temperature at 300 K [22]. Conformations were stored every 20 ps.

The FAST algorithm was used to enhance conformational sampling and quickly explore the dominant motions of nucleoprotein [23, 24]. FAST-pocket simulations were run for 6 rounds, with 10 simulations per round, where each simulation was 40 ns in length (2.4 μs aggregate simulation). The FAST-pocket ranking function favored restarting

simulations from states with large pocket openings. Additionally, a similarity penalty was added to the ranking to promote conformational diversity in starting structures, as has been described previously [25]. The FAST dataset was clustered using a k-centers algorithm based on RMSD between frames using backbone heavy atoms (C, Cα, Cβ, N, O) to generate 1421 discrete states, which were then launched on the distributed computing platform Folding@home [11, 12].

Furthering conformational sampling and enhancing statistics, Folding@home produced 500 μs of aggregate simulation. A final k-centers clustering was performed with the combined Folding@home and FAST data using Enspara (https://github.com/bowman-lab/enspara) [26]. This clustering was performed the same as described above and generated 200 discrete states that capture maximal diversity in nucleoproteins' conformational ensemble. These states were then used as the basis for CAMPARI simulations.

Sequential MD/MC sampling approach

The NTD and RBD combined are 173 residues of folded and disordered protein, which raises a significant challenge for all-atom sampling. To address this we leveraged a novel approach in which we first ran several microsecond of all-atom molecular dynamics simulations of RBD alone using the Folding@Home platform and the FAST approach for enhanced conformational sampling [11, 12, 23]. We then identified 200 conformationally distinct states based on these simulations which we used as "seeds" for the RBD. Using these seeds, we reconstructed the previously missing NTD and ran all-atom Monte Carlo simulations in which the NTD was fully sampled, the RBD sidechains are fully sampled, but the RBD backbone dihedral angles are held fixed. Multiple replicas of each starting conformation were run, giving us a total ensemble of $\sim$400 K conformations. In parallel, we also ran simulations of the NTD in isolation, enabling an assessment of the impact of the folded domain.

Coarse-grained Polymer Simulations

Coarse-grained simulations were performed using the PIMMS software package [6, 27]. PIMMS is a Monte Carlo lattice-based simulation engine in which each bead engages in anisotropic interactions with every adjacent lattice site. Moves used here were cluster translation/rotation moves and single-bead perturbation moves. Specifically, every simulation step, each bead in the system is sampled to move to adjacent sites in random order 503 of times multiplied by a factor that reflects the length of the chain. Every 100 moves (on average) a cluster of chains is randomly selected and translated or rotated, where a cluster reflects a collection of two or more chains in direct contact. This moveset provides changes to the system that reflect physical movements expected in a dynamical system, allowing us to - for equivalently sized systems - compare the apparent dynamics of assembly, as has been done previously [28–31]. We repeated the simulations presented using a range of different movesets and, while convergence varied from set-to-set, we always observed analogous results.

All simulations were performed in a 70 x 70 x 70 lattice-site box using period boundary conditions. The results reported are averaged over the final 20% of the simulation to give average values after equivalent numbers of MC steps. The "polymer" is represented as a 61-residue polymer with either a central high-affinity binding site or not. The binder is a 2-bead species. Every simulation was run for 20 x 10^9 Monte Carlo steps, with four independent replicas. Simulations were run with 1,2,3,4 or 5 polymers and 50, 75, 100, 125, 150, 175, 200, 250, 300, 400 binders.

If our simulations are run in a way deliberately designed to rapidly reach equilibrium using enhanced sampling approaches eventually all single-polymer condensates coalesce into one large multi-polymer condensate. Hence, our simulations are deliberately designed explore a regime in which single-polymer condensates are metastable.

1.1.3 Protein expression, purification, and labeling.

Plasmid construct design.

SARS-CoV2 Nucleocapsid protein (NCBI Reference Sequence: YP_009724397.2) including an N term extension containing His9-HRV 3C protease site – CATCATCACCATCATCATCATCACCACCTCGAAGTTCTGTTCCAAG-
GCCCGATGAGTGATAACGGTCCCCAGAATCAACG
GAATGCGCCCAGAATCACGTTCGGCGGTCCAAGCGACAGTACAGGTTCGAATCAGAATGGTGAACGCTCTGGGGCCCGA
AGCAAACAGCGTCGTCCACAGGGTTTGCCGAACAATACGGCTAGCTGGTTCACTGCGCTGACGCAGCACGGAAAAGAAG

ACTTAAAATTTCCGCGAGGCCAGGGGGTCCCGATTAATACTAACTCCTCCCCTGACGATCAAATTGGTTATTATCGTCG
TGCAACCCGCCGTATCCGCGGCGGAGACGGTAAAATGAAAGATCTGTCACCGCGCTGGTATTTTTACTACCTGGGAACA
GGTCCTGAAGCAGGCTTGCCGTATGGCGCTAACAAAGATGGCATTATCTGGGTGGCTACCGAGGGTGCCCTTAATACGC
CGAAAGATCATATTGGAACCCGTAACCCAGCCAATAACGCAGCAATCGTACTGCAGCTGCCGCAGGGGACAACCCTGCC
GAAAGGCTTTTATGCGGAAGGGAGTCGTGGCGGCAGCCAAGCCAGCTCCCGTAGCTCCTCGCGCTCTCGCAACTCCTCG
CGGAATAGTACACCGGGTTCATCACGCGGCACCTCGCCGGCACGCATGGCTGGCAACGGGGGGGATGCGGCTTTGGCGT
TACTTTTACTGGATAGGCTTAACCAGTTGGAAAGTAAAATGAGCGGTAAAGGCCAGCAGCAGCAGGGTCAGACTGTGAC
CAAAAAGAGCGCGGCAGAGGCGTCGAAAAAACCTAGACAAAAGCGTACTGCGACCAAAGCCTACAATGTTACGCAGGCA
TTCGGCCGGCGCGGTCCGGAACAAACCCAGGGCAACTTTGGTGACCAGGAGCTGATTCGTCAGGGAACCGATTACAAAC
ACTGGCCACAGATCGCGCAATTTGCCCCCTCGGCGTCAGCCTTTTTTGGTATGTCTCGCATTGGGATGGAGGTAACCCC
GTCTGGCACGTGGCTGACGTACACGGGCGCTATAAAGCTGGATGATAAAGATCCGAACTTCAAAGACCAGGTGATCTTA
CTGAACAAACATATTGACGCCTATAAAACGTTCCCCCCTACTGAACCTAAGAAAGATAAAAAAAAAAAAGGCCGATGAAA
CCCAAGCGCTACCACAACGCCAGAAAAAGCAGCAGACCGTCACCCTCCTGCCGGCAGCGGACCTCGACGATTTTTCTAA
GCAACTGCAACAAAGCATGTCAAGCGCCGATAGTACACAGGCGTAA - was cloned into the BamHI EcoRI sites
in the MCS of pGEX-6P-1 vector (GE Healthcare) to express the protein product:
GST-LEVLFQGPLGS**HHHHHHHHHH**
LEVLFQGPMSDNGPQNQRNAPRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNNTASWFTALTQHGKEDLKFPRGQGV
PINTNSSPDDQIGYYRRATRRIRGGDGKMKDLSPRWYFYYLGTGPEAGLPYGANKDGIIWVATEGALNTPKDHIGTRNP
ANNAAIVLQLPQGTTLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGDAALALLLLDRLNQL
ESKMSGKGQQQQGQTVTKKSAAEASKKPRQKRTATKAYNVTQAFGRRGPEQTQGNFGDQELIRQGTDYKHWPQIAQFAP
SASAFFGMSRIGMEVTPSGTWLTYTGAIKLDDKDPNFKDQVILLNKHIDAYKTFPPTEPKKDKKKKADETQALPQRQKK
QQTVTLLPAADLDDFSKQLQQSMSSADSTQA. Site-directed mutagenesis was performed on the His9-SARS-CoV2
Nucleocapsid pGEX vector to create M1C R68C, Y172C T245C, and F363C A419C variant N protein constructs. All
cloning and site-directed mutagenesis steps were performed by Genewiz and sequences were verified using sanger
sequencing.

Protein Expression and Purification.

Both GST-His9-SARS-CoV2 M1C-R68C and Y172C-T245C Nucleocapsid variants were expressed recombinantly in BL21 Codon-plus pRIL cells (Agilent). 4L cultures were grown in LB medium containing carbenicillin (100 ug/mL) to OD600 $\sim$ 0.6 and induced with 0.2 mM IPTG for 12 hours at 16ºC. Harvested cells were lysed with sonication at 4ºC in lysis buffer (50mM Tris pH 8, 500 mM NaCl, 10% glycerol, 10 mg/mL lysozyme, 5 mM BME, cOmplete™ EDTA-free Protease Inhibitor Cocktail (Roche), DNAse I (NEB), RNAse H (NEB)). The supernatant was cleared by centrifugation (37000 rpm for 1 hr) and bound to an HisTrap FF column (GE Healthcare) in buffer A (50 mM Tris pH 8, 500 mM NaCl, 10% glycerol, 20mM imidazole, 5 mM BME). GST-His9-N protein fusion was eluted with buffer B (buffer A + 500 mM imidazole) and dialyzed into cleavage buffer (50 mM Tris pH 8, 50 mM NaCl, 10% glycerol, 1 mM DTT) with HRV 3C protease, thus cleaving the GST-His9-N fusion yielding FL N protein with two additional N-term residues (GlyPro). FL N protein was then bound to an SP sepharose FF column (GE Healthcare) and eluted using a gradient of 0-100% buffer B (buffer A: 50mM Tris pH 8, 50mM NaCl, 10% glycerol, 5 mM BME, buffer B: buffer A + 1 M NaCl) over 100 min. Purified N protein variants were analyzed using SDS-PAGE and verified by electrospray ionization mass spectrometry (LC-MS). Concentrations were determined spectroscopically in 50mM Tris (pH 8.0), 500mM NaCl, 10% (v/v) glycerol using an extinction coefficient = 42530 $M^{-1}cm^{-1}$

Both GST-His9-SARS-CoV2 wild-type and F363C A419C Nucleocapsid variants were expressed recombinantly in Gold BL21(DE3) cells (Agilent). 4 L cultures were grown in LB medium with carbenicillin (100 ug/mL) to OD600 $\sim$ 0.6 and induced with 0.2 mM IPTG for 12 hours at 16ºC. Harvested cells were lysed with sonication at 4ºC in lysis buffer (listed above). The supernatant was cleared by centrifugation (37000 rpm for 1 hr) and the pellet was resuspended in 50 mM Tris pH 8, 500 mM NaCl, 10% glycerol, 6 M Urea, 5 mM BME and incubated at 4ºC for one hour. The resuspension was cleared by centrifugation (37000 rpm for 1hr) and the GST-His9-N protein in the supernatant was bound to a FF HisTrap column (GE Healthcare) in buffer A (50 mM Tris pH 8, 500 mM NaCl, 10% glycerol, 20 mM imidazole, 5 mM BME) containing 6 M Urea. The column was then washed with buffer A allowing the protein to refold on the column. The GST-His9-N protein fusion was then eluted with buffer B (buffer A containing 500 mM imidazole) and dialyzed into cleavage buffer (50 mM Tris pH8, 50 mM NaCl, 10% glycerol, 1 mM DTT)

containing HRV 3C protease. FL N protein was then bound to an SP sepharose FF column (GE Healthcare) and eluted using a gradient of 0-100% buffer B (buffer A: 50mM Tris pH 8, 50 mM NaCl, 10% glycerol, 5 mM BME, buffer B: buffer A + 1 M NaCl) over 100 min. Purified N protein variants were analyzed using SDS-PAGE and verified by electrospray ionization mass spectrometry (LC-MS). Protein concentrations of stock solutions were determined spectroscopically in 50mM Tris (pH 8.0), 500mM NaCl, 10% (v/v) glycerol using an extinction coefficient = 42530 $M^{-1}cm^{-1}$

Fluorescent Dye Labeling.

All Nucleocapsid variants were labeled with Alexa Fluor 488 maleimide (Molecular Probes) under denaturing conditions in buffer A (50mM Tris pH8, 50mM NaCl, 10% glycerol, 6M Urea, 1mM DTT) at a dye/protein molar ratio of 0.7/1 for 2 hrs at room temperature. Single labeled protein was isolated via ion-exchange chromatography (Mono S 5/50 GL, GE Healthcare - protein bound in buffer A and eluted with 0-100% buffer B (buffer A + 1 M NaCl) gradient over 100 min) and UV-Vis spectroscopic analysis to identify fractions with 1:1 dye:protein labeling. Single labeled Alexa Fluor 488 maleimide labeled N protein was then subsequently labeled with Alexa Fluor 594 maleimide at a dye/protein molar ratio of 1.3/1 for 2 hrs at room temperature. Double labeled (488:594) protein was then further purified via ion-exchange chromatography (Mono S 5/50 GL, GE Healthcare - see above).

1.1.4 Single Molecule Spectroscopy

Experimental setup and procedure.

Single-molecule fluorescence measurements were performed with a Picoquant MT200 instrument (Picoquant, Germany). For single-molecule FRET measurements, a diode laser (LDH-D-C-485, PicoQuant, Germany) was synchronized with a supercontinuum laser (SuperK Extreme, NKT Photonics, Denmark), filtered by a z582/15 band pass filter (Chroma) and pulsed at 20 MHz for pulsed interleaved excitation (PIE) (Müller et al., 2005) of labeled molecules. Emitted photons were collected with a 60x1.2 UPlanSApo Superapochromat water immersion objective (Olympus, Japan), passed through a dichroic mirror (ZT568rpc, Chroma, USA), and filtered by a 100 μm pinhole (Thorlabs, USA). Photons are counted and accumulated by a HydraHarp 400 TCSPC module (Picoquant, Germany). For FRET-FCS measurements, the same diode laser was used in continuous-wave mode to excite the donor dye. Photons emitted from the sample were collected by the objective, and scattered light was suppressed by a filter (HQ500LP, Chroma Technology) before the emitted photons passed the confocal pinhole (100 mm diameter). The emitted photons were then distributed into four channels, first by a polarizing beam splitter and then by a dichroic mirror (585DCXR, Chroma) for each polarization. Donor and acceptor emission was filtered (ET525/50m or HQ642/80m, respectively, Chroma Technology) and then focused on SPAD detectors (Excelitas, USA). The arrival time of every detected photon was recorded with a HydraHarp 400 TCSPC module (PicoQuant, Germany).

FRET experiments were performed by exciting the donor dye with a laser power of 100 μW (measured at the back aperture of the objective). For pulsed interleaved excitation experiments, the power used for exciting the acceptor dye was adjusted to match a total emission intensity after acceptor excitation to the one observed upon donor excitation (between 50 and 70 mW). Single-molecule FRET efficiency histograms were acquired from samples with protein concentrations between 50 pM and 100 pM. Trigger times for excitation pulses (repetition rate 20 MHz) and photon detection events were stored with 16 ps resolution.

For fluorescence correlation spectroscopy (FCS) experiments, acceptor-donor labeled samples with a concentration of 100 pM were excited by either the 485 nm diode laser or the supercontinuum laser at the powers indicated above. However, in the experiments on protein oligomerization, due to an increase in the fluorescence background upon addition of unlabeled protein above 1 μM, only the correlations corresponding to direct acceptor excitation (582 nm) have been considered reliable for the analysis.

For nsFCS, FRET samples of acceptor-donor labeled protein with a concentration of 100 pM were excited by the same diode laser but in continuum wavelength mode.

All measurements were performed in 50 mM Tris pH 7.32, 143 mM β-mercaptoethanol (for photoprotection), 0.001% Tween 20 (for surface passivation) and GdmCl at the reported concentrations. All measurements were performed in uncoated polymer coverslip cuvettes (Ibidi, Wisconsin, USA), which significantly decrease the fraction of

protein adhering to the surface (compared to normal glass cuvettes) under native conditions. For comparison, experiments have been performed also in glass cuvette coated with PEG, which provided analogous results to the polymeric cuvette.

Each sample was measured for at least 30 min at room temperature (295 $\pm$ 0.5 K).

FRET efficiency histograms.

Fluorescence bursts from individual molecules were identified by time-binning photons in bins of 1 ms and retaining the burst if the total number of photons detected after donor excitation was larger than 30. Transfer efficiencies for each burst were calculated according to $E = nA/(nA + nD)$ where nD and nA are the number of donor and and receptor photons, respectively. Corrections for background, acceptor direct excitation, channel crosstalk, differences in detector efficiencies, and quantum yields of the dyes were applied (Schuler et al., 2012). The labeling stoichiometry ratio S was computed accordingly to

$$S = \frac{I_D}{\gamma_{PIE}(I_A + I_D)} \tag{1.1}$$

where $I_D$and I_A represent the total intensities observed after donor and acceptor excitation and γ_{PIE} provides a correction factor to account for differences in the detection efficiency and laser intensities. Bursts with stoichiometry corresponding to 1:1 donor:acceptor labeling (in contrast to donor and acceptor only populations) were selected and finally from the selected bursts a histogram of transfer efficiencies is constructed. Variations in the selection criteria for the stoichiometry ratio do not impact significantly the observed mean transfer efficiency (within experimental errors).

To estimate the mean transfer efficiency and deconvolve multiple populations (e.g for the NTD construct) from the transfer efficiency histograms, each population was approximated with a Gaussian peak function. For fitting more than one peak, the histogram was analyzed with a sum of Gaussian peak functions. For the conversion of transfer efficiency to distances, we used the value of the Förster radius for Alexa488 and Alexa594 previously determined and reported in literature, R_0 = 5.4 nm [32]. We further correct the value accounting for the dependence of the Förster radius on the solution refractive index. The changes in refractive index caused by increasing concentrations of GdmCl or KCl were measured with an Abbe refractometer (Bausch and Lomb, USA).

Finally, we estimated a systematic error on transfer efficiency of $\pm$ 0.03, based on the variation of transfer efficiency of the same reference samples after different calibrations of the instrument over the last two years. Standard deviation of the transfer efficiency for multiple repeats of the NTD, LINK, and CTD constructs is equal or less than $\pm$ 0.01. Since we aim for a comparison with simulations, here we consider the systematic error as the larger source of error and we propagate the corresponding effect on all the calculated distances.

Fluorescence lifetimes and anisotropies analysis.

A quantitative interpretation of this transfer efficiency in terms of distance distribution requires the investigation of protein dynamics. A first method to assess whether the transfer efficiency reports about a rigid distance (e.g. structure formation or persistent interaction with the RBD) or is the result of a dynamic average across multiple conformations is the comparison of transfer efficiency and fluorescence lifetime.The interdependence of these two factors is expected to be linear if the protein conformations are identical on both timescales (nanoseconds as detected by the fluorescence lifetime, milliseconds as computed from the number of photons in each burst). Alternatively, protein dynamics give rise to a departure from the linear relation and an analytical limit can be computed for configurations rearranging much faster than the burst duration. The dependence of the fluorescence lifetimes on transfer efficiencies determined for each burst was compared with the behavior expected for fixed distances and for a chain sampling a broad distribution of distances. For a fixed distance, R, the mean donor lifetime in the presence of acceptor is given by

$$t_D(R) = t_{D0}(1 - E(R)) \tag{1.2}$$

where t_D is the lifetime in the absence of acceptor, and

$$E(R) = \frac{1}{1 + \frac{R^6}{R_0^6}} \tag{1.3}$$

For a chain with a dye-to-dye distance distribution $P(R)$, the donor lifetime is

$$t_D = \frac{\int tI(t)dt}{\int I(t)dt} \tag{1.4}$$

where

$$I(t) = I_0 P(R) \exp\left(\frac{-t}{tD(R)}\right)dR \tag{1.5}$$

is the time-resolved fluorescence emission intensity following donor excitation. A similar calculation can be carried out for describing the acceptor lifetime [33] delay given by

$$\frac{t_A(R) - t_{A0}}{t_{D0}} \tag{1.6}$$

Donor and acceptor lifetimes at different concentrations of GdmCl were analyzed by fitting subpopulation-specific time-correlated photon counting histograms after donor and acceptor excitation, respectively.

Multiparameter detection allows also excluding possible artifacts, such as insufficient rotational averaging of the fluorophores or quenching of the dyes. Subpopulation-specific anisotropies were determined for both donor and acceptor of all three constructs for NTD, LINK, and CTD, and values were found to vary between 0.1 and 0.2 for the donor and between 0.1 and 0.2 for the acceptor, sufficiently low to assume as a good approximation for the orientational factor $\kappa^2 = 2/3$.

Fluorescence Correlation Spectroscopy (FCS) analysis.

In order to determined changes in the hydrodynamic radius (Rh) of the protein, FCS correlations were analyzed assuming 3D diffusion of the molecule across a three dimensional Gaussian profile of the confocal volume (Rigler, Eur Blophys J (1993) 22:169-175). For 1 diffusing species, and in the absence of photophysical transitions in the time scale of the lag times analyzed, this formalism amounts to the following time autocorrelation function.

$$g(\tau) = t + \frac{1}{N}\left(1 + \frac{\tau}{\tau_D}\right)^{-1}\left(1 + \frac{\tau}{\alpha^2\tau_D}\right)^{-1/2} \tag{1.7}$$

where N is the average number of molecules in the confocal volume, τD is the diffusion time along the xy plane, α is the eccentricity of the three dimensional Gaussian observational volume.

$$\tau_D = \frac{\omega_{xy}^2}{4D} \tag{1.8}$$

where D is the 3D translational diffusion coefficient and ωxy is the radius from the center of the laser beam at which the light intensity decreases e^2 times from its maximum value at the center. $\alpha = \omega z/\omega xy$.

Additionally, in order to account for contributions of the photophysics of the fluorophore to the correlation observed in the μs timescale, we added two triplet terms multiplying the diffusion correlation term (see for example Krichevsky, Rep. Prog. Phys. 65 (2002) 251–297). The overall equation that we fit to the FCS traces is then

$$g(\tau) = 1 + (g_D(\tau) - 1)\left(1 + c_{T1}\exp\left(-\frac{\tau}{\tau_{T1}}\right)\right)\left(1 + c_{T2}\exp\left(-\frac{\tau}{\tau_{T2}}\right)\right) \tag{1.9}$$

where τ_{T1}, $\tau_{T2}, c_{T1}, and c_{T1}$, denotes the characteristic times and amplitudes of the contributions of two triplet states to$g(\tau)$. Parameters $\tau_D, \tau_{T1}, \tau_{T2}, c_{T1}, c_{T2} and N$ were fitted by least square nonlinear regression analysis for each concentration of unlabeled protein tested (Fig. 1.13A-B), while α was fixed at a value of 6 determined independently from analysis of fluorescence intensity profiles of fluorescent nanobeads.

Making use of the definition of τ_D and the Stokes-Einstein equation, we have, for each concentration of unlabeled protein

$$\frac{\tau_D}{\tau_{D0}} = \frac{R_h}{R_{h0}} \tag{1.10}$$

where τ_{D0} and R_{h0} are the diffusion time and hydrodynamic radius in the absence of unlabeled protein, respectively. Error bars in Fig. 1.13 B are the standard errors of Rh / Rh0 estimated from propagation of the standard errors across multiple measurements of the diffusion times obtained from the fit.

Nanosecond Fluorescence Correlation Spectroscopy.

Autocorrelation curves of acceptor and donor channels and cross-correlation curves between acceptor and donor channels were calculated with the methods described previously [34, 35]. All samples have been measured at a concentration of 100 pM and bursts with a transfer efficiency between 0.3 and 0.8 have been selected to eliminate the contribution of donor only to the correlation amplitude. Finally, the correlation was computed over a time window of 5 μs and characteristics timescales were extracted according to:

$$g_{ij} = 1 + \frac{1}{N}(1 - c_{AB}\exp[-\frac{\tau - \tau_0}{\tau_{AB}}])(1 + c_{CD}\exp[-\frac{\tau - \tau_0}{\tau_{CD}})(1 + c_T\exp[-\frac{\tau - \tau_0}{\tau_T}) \tag{1.11}$$

where N is the mean number of molecules in the confocal volume and i and j indicate the type of signal (either from the Acceptor or Donor channels). The three multiplicative terms describe the contribution to amplitude and timescale of photon antibunching (AB), chain dynamics (CD), and triplet blinking of the dyes (T). τ_{CD} is then converted in the reconfiguration time of the interdye distance τ_r correcting for the filtering effect of FRET as described previously [36]. An additional multiplicative CD term has been added only for the donor-donor correlations to describe the fast decay observed at very short time. Such a decay is not found in the correlations of other disordered proteins measured on the instrument and we associate the fast decay with the rotational motion of the overall protein. A fit to this fast decay is about 2 ns.

Polymer models of distance distributions.

Conversion of mean transfer efficiencies for fast rearranging ensembles requires the assumption of a distribution of distances. Here, we compared the results of two distinct polymer models: the Gaussian model and a Self-Avoiding Walk (SAW) model that accounts for changes in the excluded volume [37]. This second model has been shown to provide a better description of chain distribution and scaling exponent when compared to distance distributions from MD simulations [38]. Importantly, both models rely only on one single fitting parameter, the root mean square interdye distance $r = \langle R^2\rangle^{1/2}$ for the Gaussian chain and the scaling exponent ν for the SAW model.

Estimates of these parameters are obtained by numerically solving:

$$\langle E\rangle = \int_0^{l_c} P(R)E(R)dr \tag{1.12}$$

where R is the interdye distance, l_c is the contour length of the chain, $P(r)$ represents the chosen distribution, and $E(R)$is the Förster equation for the dependence of transfer efficiency on distance R and Förster radius:

$$E(R) = \frac{R_0^6}{R_0^6 + R^6} \tag{1.13}$$

The Gaussian chain distribution is given by:

$$P_{FJC}(R, r) = 4\pi R^2(\frac{3}{2\pi r^2})^{3/2}\exp(-\frac{3R^2}{2r^2}) \tag{1.14}$$

The SAW model can be expressed as:

$$P_{SAW}(R, \nu) = A_1\frac{4\pi}{b_0N^\nu}(\frac{R}{b_0N^\nu})^{2+g}\exp(-A_2(\frac{R}{b_0N^\nu})^\delta) \tag{1.15}$$

where

$$A_1 = \frac{\delta}{4\pi}\frac{\Gamma[5 + \frac{g}{\delta}\frac{3+g}{2}]}{\Gamma[3 + \frac{g}{\delta}\frac{5+g}{2}]}, A_2 = (\frac{\Gamma[5 + \frac{g}{\delta}]}{\Gamma[5 + \frac{g}{\delta}]})^\delta, g = \frac{\gamma - 1}{\nu}, \delta = \frac{1}{1 - \nu} \tag{1.16}$$

γ =1.1615, and Γ is the Euler Gamma Function, b_0=0.55nm is an empirical prefactor [38], N is the number of residues between the fluorophores, and ν is the scaling exponent.

Finally, when converting the distance from transfer efficiencies, to account for the length of dye linkers and compare the experimental data with simulations, the root-mean-squared interdye distance r was rescaled according to

$$r_{m,n} = |m-n|^{0.5} I_{dye} |m-n+2I_{dye}|^{0.5} \tag{1.17}$$

with I_dye=4.5 (Aznauryan et al., 2016; Hoffmann et al., 2007). Finally, the persistence length is computed using the Gaussian conversion $r^2 = 2l_p l_c$ [39].

Binding of denaturant and folding.

As in previous works [40–42], we model the chain expansion with the denaturant in terms of a simple binding model:

$$r_c = r_0(1 + \rho \frac{Kc}{1+Kc}) \tag{1.18}$$

Where r_0 is the mean square interdye distance at zero denaturant,ρ is a term the captures the extent of chain expansion with the denaturant compared to r_0, and the K is the binding constant, and c is the concentration of denaturant.

In presence of folded domains, we can imagine the folding/unfolding of the domains can affect the overall size of the chain because of an increase or decrease of excluded volume due to the surrounding folded domains (which screen part of the available conformations) or because of the folding or unfolding of elements in the region between the fluorophores. To account for this effect, as in the case of the NTD, we weighed the effect of denaturant on the chain for the fraction folded f_f and unfolded f_u accordingly to:

$$r_c = (r_0 f_f + r_{0u} f_u)(1 + \rho \frac{Kc}{1+Kc}) \tag{1.19}$$

where r_{0f} and r_{0u} are the root mean square interdye distance in presence of folded or unfolded domains in native buffer,

$$f_f = \frac{\exp[-m(c-c_m)]}{1+\exp[-m(c-c_m)]} \tag{1.20}$$

and $f_u = 1 - f_f$, where c_m the midpoint concentration and m the denaturant m value, representing the dependence of free energy on denaturant concentration. The stability parameter ΔG_0 can be computed as $\Delta G_0 = mc_m$.

Polymer model of electrostatic interactions.

The disordered regions of the N protein are enriched in positive and negative charges. To provide a term of comparison in the interpretation of protein conformations as function of salt concentration, we use the polymer theory for polyampholyte solutions developed by Higgs and Joanny [40, 43], which has been shown previously to capture quantitatively the conformational changes of unstructured proteins. Briefly, the root mean square interdye distance is equal to $r = N^{0.5} * l_0 * \alpha$ where N is the number of monomers in the disordered region, l_0 is the length of elementary segment (here 0.36 nm) and α is the ratio between l and l_0, with l being a rescaled segment that accounts for excluded volume and electrostatic interactions.

$\alpha$is computed according to the equation proposed by Higgs and Joanny [40, 43]:

$$\alpha^5 - \alpha^3 = \frac{4}{3}(\frac{3}{2\pi})^{1.5} N^{0.5} v^* \tag{1.21}$$

where v^* is an effective excluded volume given by the sum of three terms:

$$v^* b^3 = v b^3 + \frac{4\pi l_B (f-g)^2}{k^2} - \frac{\pi l_B^2 (f-g)^2}{k} \tag{1.22}$$

Here, v is the excluded volume (accounting for physical excluded volume and positive and attractive interactions that are not due to electrostatics), f and g are the fraction of positive and negative residue respectively for considered segment of the protein, k is the Debye screening length, and l_b is the Bjerrum length.

Importantly, when accounting for the fraction of negative charges, we also account for the contribution of the -2 net charge of each dye at pH 7.3.

Salt dependence of NTD, LINK, and CTD conformations.

In addition to studying the conformations under native buffer conditions, we investigate how salt affects the conformations of the three disordered regions. We started by testing the effects of electrostatic interactions on the NTD conformational ensemble. Moving from buffer conditions and increasing concentration of KCl, we observed a small but noticeable shift toward lower transfer efficiencies, which represents an expansion of the NTD due to screening of electrostatic interactions. This can be rationalized in terms of the polyampholyte theory of Higgs and Joanny [40, 43] (see Table 1.2), where the increasing concentration of ions screens the interaction between oppositely charged residues (see Fig. 1.10).

We then analyzed for comparison the LINK construct. Interestingly, we find a negligible effect of salt screening on the root mean square distance r172-245 as measured by FRET (see Fig. 1.10). Predictions of the Higgs and Joanny theory for the content of negative and positive charges within the LINK construct indicates a variation of interdye distance dimension that is comparable with the measurement error. It has to be noted that in this case the excluded volume term in the Higgs and Joanny theory will empirically account not only for the excluded volume of the amino acids in the chain, but also for the excluded volume occupied by the two folded domains.

Finally, we test if the addition of salt can provide similar effects than those obtained by GdmCl on the conformations of the CTD: interestingly, we do not observe any significant variation either in transfer efficiency or distribution width (Fig. 1.10), suggesting that the broadening of the population observed for the CTD does not originate from electrostatic interactions.

1.1.5 Testing protein oligomerization.

NativePAGE experiments were performed to verify that purified recombinantly expressed SARS-CoV-2 N protein is capable of forming dimers and oligomers, in analogy to SARS-CoV N protein, and as shown in more recent work for SARS-CoV-2 [14, 44, 45]. Indeed, NativePAGE experiments reveal the existence of multiple bands (Fig. 1.13C-D). However, since the lowest band in the NativePAGE corresponds to an apparent molecular weight of $\sim$70-80 kDa, we wanted to verify the oligomeric state of this band.

To test whether the apparent mass is due to a slow mobility of the protein because of its high positive charge, we performed crosslinking experiments. These experiments confirm the formation of dimers, tetramers, and high oligomeric species, as a function of protein concentration above 500 nM (Fig. 1.13E-F). These oligomeric species are in equilibrium with the monomer, the smallest species on the denaturing SDS PAGE (which has the expected molecular weight of $\sim$45 kDa). It has to be noted that, because of the slow reactivity of the crosslinking agent (see Methods below), the crosslinking experiments do not represent the population of monomeric and oligomeric species at equilibrium. However, the comparison between the NativePAGE and the crosslinking experiments supports the fact that the smallest band in the NativePAGE is indeed the monomer protein.

We finally turned to Fluorescence Correlation Spectroscopy (FCS) to test whether labeled protein can form dimers. We measured the CTD construct that carries one labeling position at the end of the oligomerization domain. When increasing the concentration of unlabeled protein, we observe a systematic increase in the hydrodynamic radius when compared to the hydrodynamic radius under native conditions (Fig. 1.13A-B). This suggests that the labeled protein can form higher oligomeric species in a concentration regime comparable to the one observed in NativePAGE and SDS PAGE experiments and that at 100 pM (the concentration used in single-molecule experiments), no oligomer is formed. Caution must be used in the interpretation of the oligomeric bound species observed in FCS experiments, since labeling mutation may have affected the affinity of the dimerization domain. Future experiments will address the role of mutation on dimerization. Finally, all experiments have been performed at two different time points, after 1 hour and after 24 hours of incubation of the labeled sample with unlabeled protein to test any kinetic effect on the measured value. No significative difference has been observed.

Taken together, NativePAGE crosslinking experiments verify that in smFRET and FCS experiments we are in fact monitoring the behavior of the monomeric SARS-CoV-2 N protein.

1.1.6 Protein Crosslinking Methods.

50 mM disuccinimidyl suberate (DSS) (Thermo Scientific) stock solution was prepared (10 mg into 540 uL of anhydrous DMSO (Sigma)). All protein samples were prepared in 20 mM NaPi pH 7.4 (with and without 200 mM NaCl) at the following concentrations: 0.1, 0.5, 1, 5, 10 and 20uM. DSS stock solution was added to each sample to a final concentration of 1.25 mM. Samples were incubated for 1 hour at room temperature. Samples were then quenched to a final concentration of 200mM Tris pH 7.4 and allowed to incubate for 15 minutes. Crosslinked proteins were then analyzed using SDS PAGE and Coomassie staining.

1.1.7 NativePAGE Methods.

All protein samples were prepared in 20 mM NaPi pH 7.4 (with and without 200 mM NaCl) at the following concentrations: 0.05, 0.1, 0.5, 1, 5, 10 and 20 uM. Samples were subjected to NativePAGE (Invitrogen) and protein mobility was analyzed with Coomassie staining.

Development of turbidity in solutions of N protein and poly(rU) was followed through measurements of absorbance at 340 nm in a microvolume spectrophotometer (NanoDrop, Thermo, USA). Mixtures were prepared in 500 μl plastic reaction tubes by adding 4 μl protein solution into 3 μl of poly(rU) and absorbance was recorded 45 s – 75 s after mixing. Working solutions were kept at room temperature during experiments.

Reaction media was 50 mM Tris, pH 7.5 (HCl), 0.002 % v/v Tween20, and NaCl as indicated in Results.

poly(rU) (Midland Certified Reagent Company, TX, USA, lot number 011805) was reconstituted into this media from stocks dissolved in RNAse free water. According to the manufacturer, the size of poly(rU) molecules is mostly less than 250 nucleotides (nt.) and longer than 200 nt.

Protein stocks (in 50 mM Tris pH 8.0, 500 mM NaCl, 10% v/v glycerol) were buffer exchanged into the desired buffer through size exclusion chromatography in Zeba Spin 7 k MWCO desalting columns (Thermo, USA). poly(rU) concentrations in working dilutions were assessed through the absorbance at 260 nm employing an extinction coefficient of 9.4 $mM^{-1}cm^{-1}$ (Michelson, 1959). Protein concentrations were assessed through the absorbance at 280 nm employing an extinction coefficient of 42.53 $mM^{-1}cm^{-1}$, computed according to the method proposed by Pace et al. (Pace et al., 1995).

The limiting concentrations of nucleic acid across which an increase in turbidity was detected were estimated through interpolation of the data. To this end, an empirical equation, describing the trends observed at all concentrations, was fitted to the data and then was solved to extract the poly(rU) concentrations at which turbidity reaches a limit value above the background signal. We used a limiting absorbance value of 0.005 units (340 nm, 1 mm path length). We found that an appropriate function for this end is an exponential of a Gaussian distribution function $F(x)$:

$$F(x) = A(1 - \exp[-\beta\gamma(x)]) \tag{1.23}$$

where

$$\gamma(x) = \frac{1}{(2\pi)^{0.5}\sigma} \exp\left[\frac{-(x-\mu)^2}{2\sigma}\right] \tag{1.24}$$

where x denotes poly(rU) concentration and $A, \mu, \sigma and \mu$ are parameters fitted through weighted minimum least squares for each protein concentration (solid lines in Fig. **??**A-B and limiting value points in panels C-D). To characterize the observed global trends of turbidity, as a function of both RNA and protein concentration, we determined approximate functional forms of the dependence on protein concentration of the individually fitted parameters ($A(p), \beta(p), \sigma(p) and \mu(p)$, where p is protein concentration). The observed dependencies were increasing linearly for $\mu(p)$ and quadratic for $\beta(p)$ and $\sigma(p)$. A was the worst defined parameter and thus displayed the least clear trend. For the results in absence of added salt we employed an increasing power function with exponent as a fitting parameter (best fit value was < 1), whereas for the results in presence of 50 mM NaCl the trend of $A(p)$ was better described by a decreasing exponential function.

We thus used the functional forms $A(p), \beta(p), \sigma(p)$ and $\mu(p)$ to construct a global function dependent on both protein and RNA concentration. Global fitting of this equation to the whole set of turbidity titration curves provided the turbidity contour plots shown in Fig. **??**C-D (solid lines). Contour lines were computed at 1, 10, 20, 50 and 100 times the limiting value employed ($A_{340nm,1mm}$= 0.005).

Table 1.1: Fit parameters to denaturant binding model.

	ρ	K (M^{-1})	r_0 (Å)
NTD (1 pop)	1.3± 0.2	0.36 ± 0.05	50 ± 2 (fixed)
NTD (2 pop)	(shared parameter)	(shared parameter)	36 ± 3
LINK	1.1 ± 0.03	0.06 ±0.03	57 ± 2 (fixed)
CTD	0.47 ± 0.02	0.36 ± 0.1	50 ± 2 (fixed)

Table 1.2: Fit parameters of Higgs & Joanny theory

	v
NTD	4.4 ± 0.1
LINK	5.5 ± 0.3
CTD	8.4 ± 0.9

1.2 Supplementary Figures

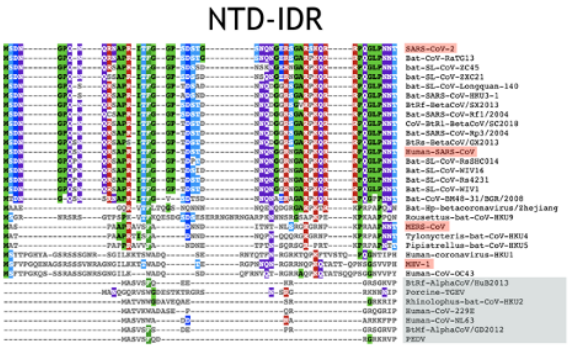

Figure 1.1: Sequence alignment of the coronavirus N-terminal domain (NTD).

Table 1.3: Scaling exponents

	νSAW	νsimulation
NTD	0.520 ± 0.009	0.52
LINK	0.538 ± 0.008	0.58
CTD	0.546 ± 0.008	0.49

Table 1.4: All-atom simulation summary

System	No. sims	Total steps per sim (M).	Prod. steps per sim.(M)	Config. output	Ensemble size
NTD-RBD	400	24	20	20,000	399,000
RBD-LINK-DIM	10	66	60	20,000	31,113
DIM-CTD	40	24	20	20,000	40,000
NTD	40	71	66	30,000	64,000
LINK	30	101	80	30,000	66,660
CTD	40	71	66	30,000	64,000

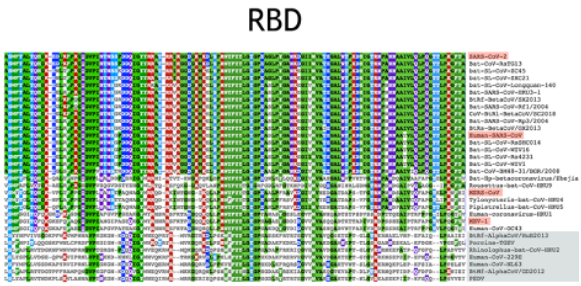

Figure 1.2: Sequence alignment of the coronavirus RNA binding domain (RBD).

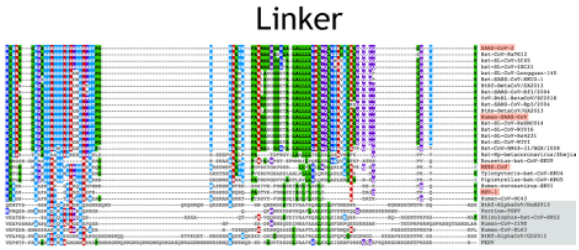

Figure 1.3: Sequence alignment of the coronavirus linker (LINK).

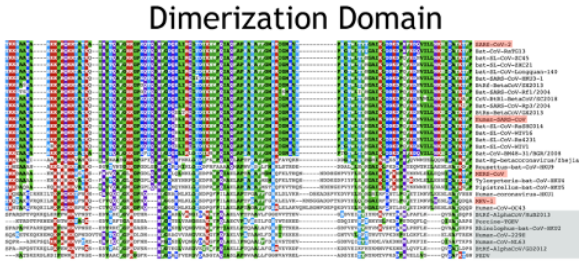

Figure 1.4: Sequence alignment of the coronavirus dimerization domain.

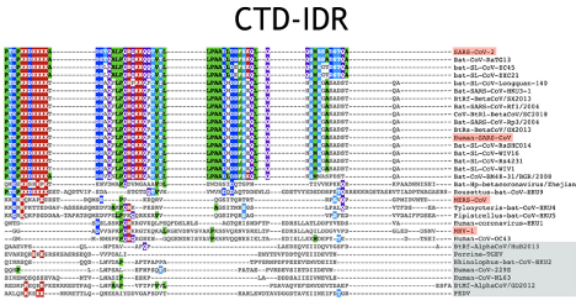

Figure 1.5: Sequence alignment of the coronavirus C-terminal domain (CTD)

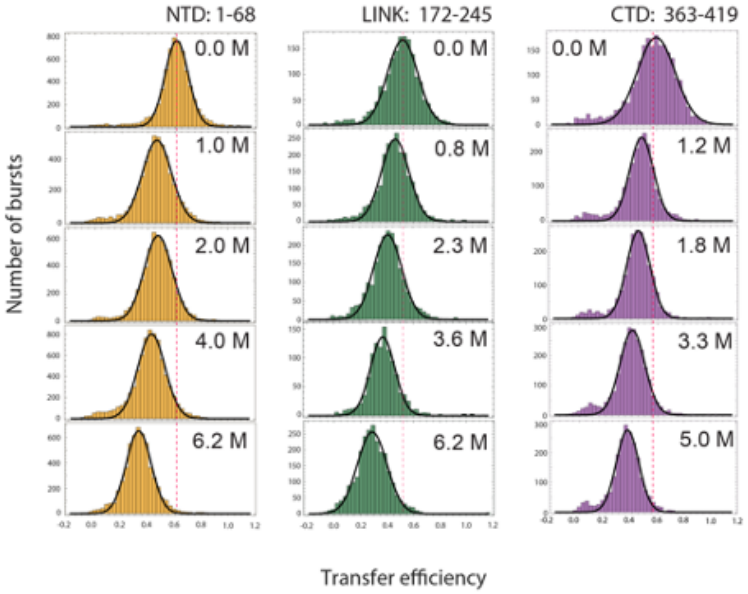

Figure 1.6: Histograms of transfer efficiency distributions across denaturant concentrations for NTD, LINK, and CTD constructs.

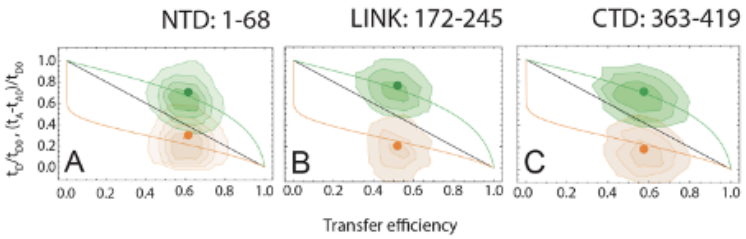

Figure 1.7: Dependence of fluorescence lifetime on transfer efficiency. **A.** NTD construct. **B.** LINK construct. **C.** CTD construct. Black line: linear dependence expected for a rigid molecule. Green line: the donor lifetime (normalized by the donor lifetime in absence of FRET: tD/tD0) in the limit of dynamics much faster than the burst duration but slower than the fluorophore lifetime. Orange line: the acceptor lifetime delay (normalized by the donor lifetime in absence of FRET. The green and orange contour plots represent the corresponding distributions of donor lifetime and acceptor lifetime delay as observed in single-molecule experiments under native conditions The green and orange dots represent the mean value of the measured distributions. A larger overlap between donor and acceptor lifetime populations is observed for the NTD and CTD, hinting to possible static conformations.

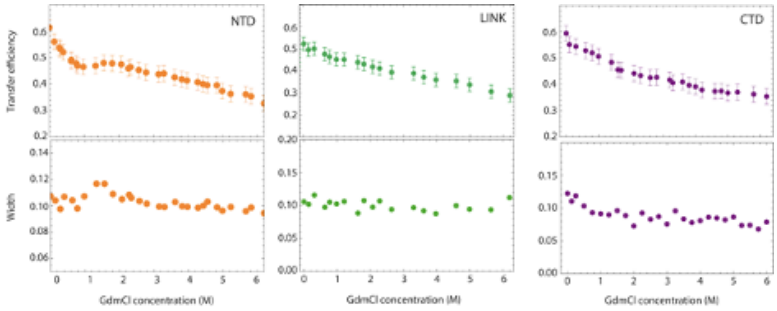

Figure 1.8: Mean transfer efficiency and width of NTD, LINK, and CTD across denaturant. The mean transfer efficiency of the NTD domain exhibits a plateau between 1 and 2 M; at the same concentration we observe a small but systematic increase in the amplitude of the transfer efficiency distribution hinting to the coexistence of two populations in slow exchange with very similar transfer efficiencies. The CTD width also shows a small increase in the width of the transfer efficiency distribution that may reflect the formation of local structure under native conditions (e.g. the putative helical binding motif).

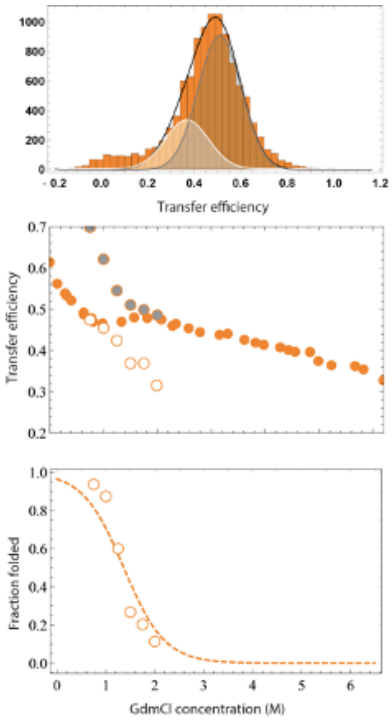

Figure 1.9: Fit of NTD construct with two populations. To address the change in amplitude that occurs from the NTD construct between 1 and 2 M GdmCl, we attempt a fit of the same data using two populations with a fixed distance equal to average width outside the 1-2 M GdmCl region (see for comparison Fig. 1.8). **Upper panel:** fit of the transfer efficiency histogram at 1.5 M GdmCl. The white- and gray- shaded areas reflect fits to the "folded RBD" population and to the "unfolded RBD" population. **Central panel:** Comparison of transfer efficiencies with a single fit (solid orange circles, compare Fig. 1.8) and from the two populations: gray solid circles for the "unfolded RBD" population and unfilled circles for the "folded RBD" population. **Lower panel:** Fraction folded estimated from the fit with Eq. S7 compared to the fraction of "folded RBD" obtained from computing the ratio between the area under "folded RBD" species and the total histogram area.

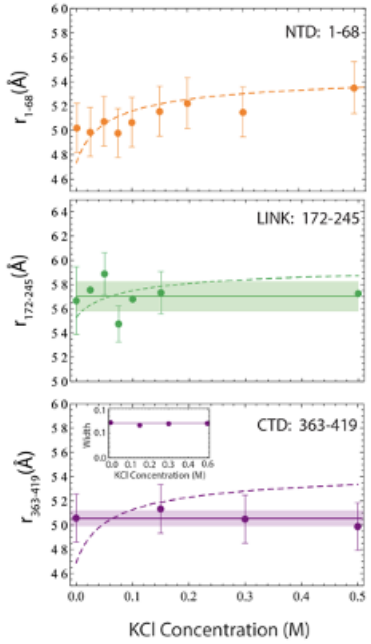

Figure 1.10: Interdye distances of NTD, LINK, CTD in presence of salt (KCl). **Upper panel:** root mean square interdye distance between position 1 and 68. Dashed line: fit according to the Higgs and Joanny model (Eq. 1.21-1.22) predicts a comparable change to the one observed. Central panel: root mean square interdye distance between position 172 and 245. Dashed line: fit according to the Higgs and Joanny model (Eq. 1.21-1.22) predicts a comparable change to the one observed. Solid line and shaded area: average value of the root-mean-square interdye distance across all salt conditions and corresponding standard deviation. The standard deviation is comparable to the measurement error. **Lower panel:** root mean square interdye distance between position 363 and 419. Dashed line: fit according to the Higgs and Joanny model (Eq. 1.21-1.22) does not capture the observed trend. This can be possibly explained considering the significant predicted population of helical conformations in the CTD. Solid line and shaded area: average value of the root-mean-square interdye distance across all salt conditions and corresponding standard deviation. Inset: no variation in the width of the transfer efficiency population is observed upon addition of KCl.

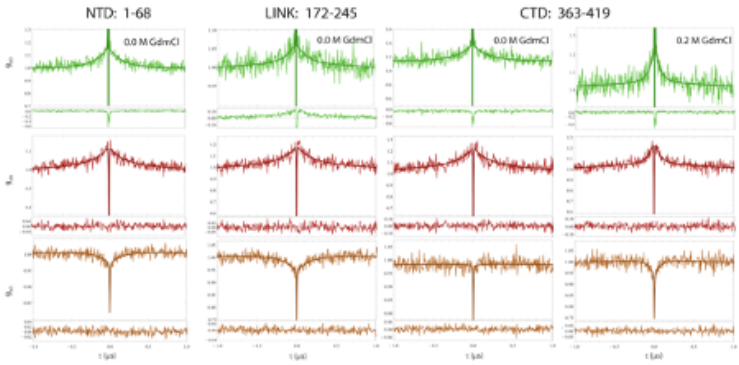

Figure 1.11: Chain dynamics measured via ns-FCS. Nanosecond FCS measurements for the NTD, LINK, and CTD constructs provide a measure of the dynamics on the nanosecond timescale. The donor-donor (green), acceptor-acceptor (red), and donor-acceptor (orange) correlation are fitted to a global model that accounts for antibunching, FRET dynamic populations, and triplet. The acceptor-donor correlation shows a clear anticorrelated change for NTD and LINK in the signal that reflects the anticorrelated nature of the donor-acceptor energy transfer as a function of distance: an increase in acceptor reflects a decrease in donor. The CTD cross-correlation exhibits a flat behavior. Occurrence of a correlation in the donor-donor and acceptor-acceptor autocorrelations corresponding with a characteristic time t_b= 190 $\pm$ 30 ns suggests the presence of chain dynamics, either through FRET or Photo-induced Electron Transfer (PET) [46, 47]. Addition of 0.2 M GdmCl, which causes a small decrease in the transfer efficiency width (Fig. 1.8) leads to an anticorrelation in the cross-correlation of CTD. The correlation decay appears also faster with a $t_C D$= 70 $\pm$ 15 ns. All measurements are normalized to the value measured at 1 μs for highlighting the amplitude relative to the reconfiguration term.

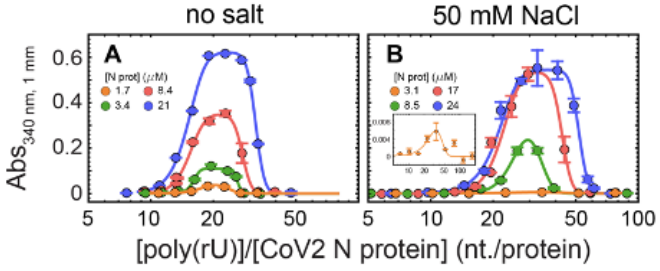

Figure 1.12: Turbidity experiments plotted against RNA/protein ratio. Representative turbidity titrations with poly(rU) in 50 mM Tris, pH 7.5 (HCl) at room temperature, in absence of added salt **(A)** and in presence of 50 mM NaCl **(B)**, at the indicated concentrations of N protein. On the x-axis, the concentration of poly(rU) is rescaled for the protein concentration. Points and error bars represent the mean and standard deviation of 2-4 consecutive measurements from the same sample. Solid lines are simulations of an empirical equation fitted individually to each titration curve. An inset is provided for the titration at 3.1 μM N protein in 50 mM NaCl to show the small yet detectable change in turbidity on a different scale. Interestingly, within the experimental error, we observe a clear alignment of the turbidity curves with a maximum at 20 nucleotides per protein in the absence of added salt **(A)** and 30 nucleotides per protein in the presence of 50 mM NaCl **(B)**.

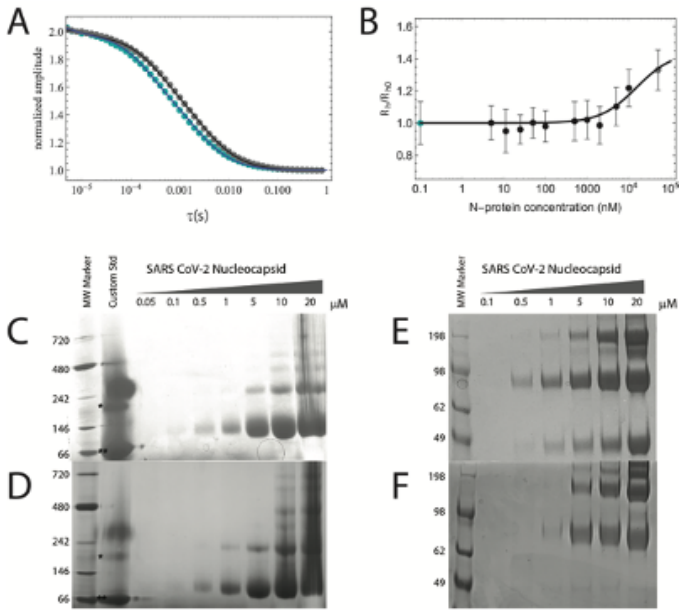

Figure 1.13: Testing SARS-CoV-2 N protein oligomerization. **(A-B)** Fluorescence Correlation Spectroscopy (FCS) of full-length SARS-CoV-2 N protein as a function of protein concentration. (A) FCS traces of 100pM Alexa 488/Alexa 594 N protein labeled at positions 363 and 419 in the absence (blue dots) and the presence (gray dots) of 50 μM unlabeled N protein. (B) Hydrodynamic radius of SARS-CoV-2 N protein obtained from FCS trace analysis (blue dot: 100pM labeled N protein; gray dot: 100pM labeled N protein + 50μM unlabeled N protein). **(C-D)** NativePAGE of full-length SARS-CoV-2 N protein in 20 mM NaPi pH 7.4 as a function of protein concentration in the presence of 200 mM NaCl (C) and in the absence of added salt (D). 'Custom Std' lane contains Alcohol Dehydrogenase (* , 150 kDa) and Bovine Serum Albumin (** , 66 kDa). (E-F) SDS PAGE of crosslinked full-length SARS-CoV-2 N protein in 20 mM NaPi pH 7.4, 1.25mM DSS as a function of protein concentration in the presence of 200mM NaCl (E) and in the absence of added salt (F).

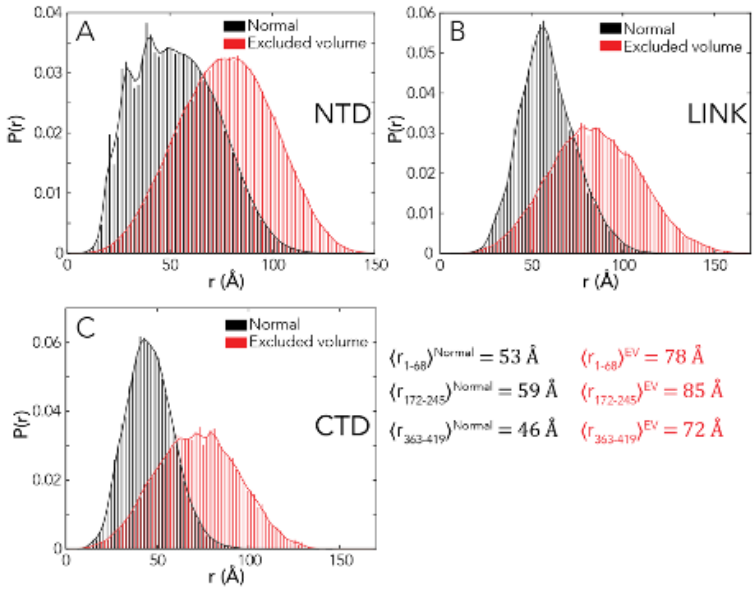

Figure 1.14: Distributions of inter-residue distance from ABSINTH simulations (black) vs. excluded volume simulations (red). Comparison of simulations with the full ABSINTH Hamiltonian ('normal', black) against simulations performed in the excluded volume (EV, red) limit for **A.** NTD, **B.** LINK and **C.** CTD. In all cases the EV simulations report substantially larger average distances than the ABSINTH simulations, as expected given the absence of any attractive intramolecular interactions. The distances reported from the EV simulations are also slightly more expanded than under fully denatured conditions, consistent with systems studied previously (see [2, 48].

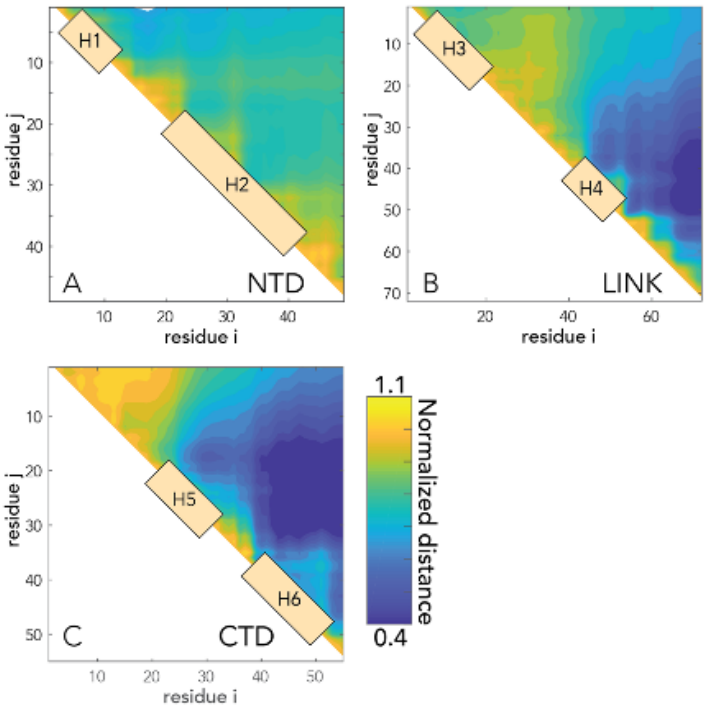

Figure 1.15: Scaling maps for IDR-only simulations. Scaling maps report on the normalized distance between pairs of residues, where normalization is done by the distance expected if the IDRs behaved as self-avoiding chains in the excluded-volume limit. Scaling maps for IDR-only simulations of the **A.** NTD, **B.** LINK and **C.** CTD. For each sequence, transient helices are annotated on the scaling maps. Note that in the LINK we observe interaction between the C-terminal region of the LINK and H4, while H3 does not interact with any parts of the sequence. Similarly, in CTD we see extensive intramolecular interactions between H5 and H6.

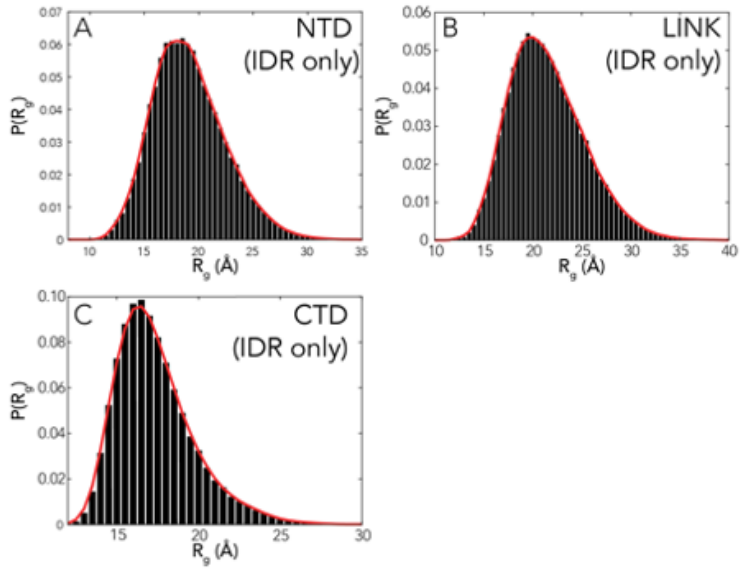

Figure 1.16: Distributions for the radius of gyration (R_g) of for IDR-only simulations. R_g distributions for **A.** NTD, **B.** LINK and **C.** CTD. Average R_g for each IDR in isolation is 19.1 Å (NTD), 21.4 Å (LINK), and 17.1 Å (CTD).

Bibliography

[1] Jasmine Cubuk, Jhullian J. Alston, J. Jeremías Incicco, Sukrit Singh, Melissa D. Stuchell-Brereton, Michael D. Ward, Maxwell I. Zimmerman, Neha Vithani, Daniel Griffith, Jason A. Wagoner, Gregory R. Bowman, Kathleen B. Hall, Andrea Soranno, and Alex S. Holehouse. The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. preprint, Biophysics, June 2020.

[2] Alex S Holehouse and Shahar Sukenik. Controlling structural bias in intrinsically disordered proteins using solution space scanning. *J. Chem. Theory Comput.*, 16(3):1794–1805, March 2020.

[3] Zsuzsanna Dosztányi, Veronika Csizmok, Peter Tompa, and István Simon. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–3434, August 2005.

[4] Alex S Holehouse, Rahul K Das, James N Ahad, Mary O G Richardson, and Rohit V Pappu. CIDER: Resources to analyze Sequence-Ensemble relationships of intrinsically disordered proteins. *Biophys. J.*, 112(1):16–21, January 2017.

[5] Rahul K Das, Yongqi Huang, Aaron H Phillips, Richard W Kriwacki, and Rohit V Pappu. Cryptic sequence features within the disordered protein p27kip1 regulate cell cycle signaling. *Proc. Natl. Acad. Sci. U. S. A.*, 113(20):5616–5621, May 2016.

[6] Erik W Martin, Alex S Holehouse, Ivan Peran, Mina Farag, J Jeremias Incicco, Anne Bremer, Christy R Grace, Andrea Soranno, Rohit V Pappu, and Tanja Mittag. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science*, 367(6478):694–699, February 2020.

[7] Kathryn P Sherry, Rahul K Das, Rohit V Pappu, and Doug Barrick. Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the notch receptor. *Proc. Natl. Acad. Sci. U. S. A.*, 114(44):E9243–E9252, October 2017.

[8] Robert T McGibbon, Kyle A Beauchamp, Matthew P Harrigan, Christoph Klein, Jason M Swails, Carlos X Hernández, Christian R Schwantes, Lee-Ping Wang, Thomas J Lane, and Vijay S Pande. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal*, 109(8):1528–1532, October 2015.

[9] Andreas Vitalis and Rohit V Pappu. ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.*, 30(5):673–699, April 2009.

[10] Esther Ortega, Srinivasan Rengachari, Ziad Ibrahim, Naghmeh Hoghoughi, Jonathan Gaucher, Alex S Holehouse, Saadi Khochbin, and Daniel Panne. Transcription factor dimerization activates the p300 acetyltransferase. *Nature*, 562(7728):538–544, October 2018.

[11] M Shirts and V S Pande. COMPUTING: Screen Savers of the World Unite! *Science*, 290(5498):1903–1904, December 2000.

[12] Maxwell I. Zimmerman, Justin R. Porter, Michael D. Ward, Sukrit Singh, Neha Vithani, Artur Meller, Upasana L. Mallimadugula, Catherine E. Kuhn, Jonathan H. Borowsky, Rafal P. Wiewiora, Matthew F. D. Hurley, Aoife M Harbison, Carl A Fogarty, Joseph E. Coffland, Elisa Fadda, Vincent A. Voelz, John D. Chodera, and Gregory R. Bowman. Citizen scientists create an exascale computer to combat covid-19. *bioRxiv*, 2020.

[13] Andrew Waterhouse, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T Heer, Tjaart A P de Beer, Christine Rempfer, Lorenza Bordoli, Rosalba Lepore, and Torsten Schwede. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46(W1):W296–W303, July 2018.

[14] Mitsuhiro Takeda, Chung-Ke Chang, Teppei Ikeya, Peter Güntert, Yuan-Hsiang Chang, Yen-Lan Hsu, Tai-Huang Huang, and Masatsune Kainosho. Solution structure of the c-terminal dimerization domain of SARS coronavirus nucleocapsid protein solved by the SAIL-NMR method. *J. Mol. Biol.*, 380(4):608–622, July 2008.

[15] Luca Zinzula, Massimiliano Orsini Nagy, and Andreas Bracher. 1.45 angstrom resolution crystal structure of c-terminal dimerization domain of nucleocapsid phosphoprotein from SARS-CoV-2 (PDB: 6YUN). *Protein Data Bank*, May 2020.

[16] Alex S Holehouse, Kanchan Garai, Nicholas Lyle, Andreas Vitalis, and Rohit V Pappu. Quantitative assessments of the distinct contributions of polypeptide backbone amides versus side chain groups to chain expansion via chemical denaturation. *J. Am. Chem. Soc.*, 137(8):2984–2995, March 2015.

[17] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1–2:19–25, 2015.

[18] Yong Duan, Chun Wu, Shibasish Chowdhury, Mathew C Lee, Guoming Xiong, Wei Zhang, Rong Yang, Piotr Cieplak, Ray Luo, Taisung Lee, James Caldwell, Junmei Wang, and Peter Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of computational chemistry*, 24(16):1999–2012, December 2003.

[19] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, July 1983.

[20] K A Feenstra, B Hess, and HJC Berendsen. Improving e$_{c}iencyoflargetimescalemoleculardynamicssimulationsofhydrogen$
$richsystems.JComputChem, 20(8) : 786 - -798, June1999.$

[21] Berk Hess. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *Journal of Chemical Theory and Computation*, 4(1):116–122, January 2008.

[22] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):014101, January 2007.

[23] Maxwell I. Zimmerman and Gregory R. Bowman. Fast conformational searches by balancing exploration/exploitation trade-offs. *Journal of Chemical Theory and Computation*, 11(12):5747–5757, December 2015.

[24] Maxwell I Zimmerman, Justin R Porter, Xianqiang Sun, Roseane R Silva, and Gregory R Bowman. Choice of Adaptive Sampling Strategy Impacts State Discovery, Transition Probabilities, and the Apparent Mechanism of Conformational Changes. *Journal of Chemical Theory and Computation*, q-bio.BM:acs.jctc.8b00500, October 2018.

[25] Maxwell I. Zimmerman, Kathryn M. Hart, Carrie A. Sibbald, Thomas E. Frederick, John R. Jimah, Catherine R. Knoverek, Niraj H. Tolia, and Gregory R. Bowman. Prediction of new stabilizing mutations based on mechanistic insights from markov state models. *ACS Central Science*, 3(12):1311–1321, December 2017.

[26] J. R. Porter, M. I. Zimmerman, and G. R. Bowman. Enspara: Modeling molecular ensembles with scalable data structures and parallel computing. *The Journal of Chemical Physics*, 150(4):044108, 2019.

[27] Alex S Holehouse and Rohit V Pappu. PIMMS (0.24 pre-beta), December 2019.

[28] Noah S Bieler, Tuomas P J Knowles, Daan Frenkel, and Robert Vácha. Connecting macroscopic observables and microscopic assembly events in amyloid formation using coarse grained simulations. *PLoS Comput. Biol.*, 8(10):e1002692, October 2012.

[29] Steven Boeynaems, Alex S Holehouse, Venera Weinhardt, Denes Kovacs, Joris Van Lindt, Carolyn Larabell, Ludo Van Den Bosch, Rhiju Das, Peter S Tompa, Rohit V Pappu, and Aaron D Gitler. Spontaneous driving forces give rise to protein-RNA condensates with coexisting phases and complex material properties. *Proc. Natl. Acad. Sci. U. S. A.*, 116(16):7889–7898, April 2019.

[30] Kristen A Fichthorn and W H Weinberg. Theoretical foundations of dynamical monte carlo simulations. *J. Chem. Phys.*, 95(2):1090–1096, July 1991.

[31] An**?**ela Šarić, Alexander K Buell, Georg Meisl, Thomas C T Michaels, Christopher M Dobson, Sara Linse, Tuomas P J Knowles, and Daan Frenkel. Physical determinants of the self-replication of protein fibrils. *Nat. Phys.*, 12(9):874–880, July 2016.

[32] Benjamin Schuler, Everett A Lipman, and William A Eaton. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*, 419:743, October 2002.

[33] Hoi Sung Chung, John M Louis, and Irina V Gopich. Analysis of fluorescence lifetime and energy transfer efficiency in Single-Molecule photon trajectories of Fast-Folding proteins. *J. Phys. Chem. B*, 120(4):680–699, February 2016.

[34] Daniel Nettels, Irina V Gopich, Armin Hoffmann, and Benjamin Schuler. Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proc. Natl. Acad. Sci. U. S. A.*, 104(8):2655–2660, February 2007.

[35] Daniel Nettels, Armin Hoffmann, and Benjamin Schuler. Unfolded protein and peptide dynamics investigated with single-molecule FRET and correlation spectroscopy from picoseconds to seconds. *J. Phys. Chem. B*, 112(19):6137–6146, May 2008.

[36] Irina V Gopich, Daniel Nettels, Benjamin Schuler, and Attila Szabo. Protein dynamics from single-molecule fluorescence intensity correlation functions. *J. Chem. Phys.*, 131(9):095102, September 2009.

[37] Lothar Schäfer. *Excluded Volume Effects in Polymer Solutions: as Explained by the Renormalization Group*. Springer Science & Business Media, December 2012.

[38] Wenwei Zheng, Gül H Zerze, Alessandro Borgia, Jeetain Mittal, Benjamin Schuler, and Robert B Best. Inferring properties of disordered chains from FRET transfer efficiencies. *J. Chem. Phys.*, 148(12):123329, March 2018.

[39] M Rubinstein and Ralph H Colby. *Polymer Physics*. Oxford University Press, New York, 2003.

[40] Sonja Müller-Späth, Andrea Soranno, Verena Hirschfeld, Hagen Hofmann, Stefan Rüegger, Luc Reymond, Daniel Nettels, and Benjamin Schuler. Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 107(33):14609–14614, August 2010.

[41] Andrea Soranno, Brigitte Buchli, Daniel Nettels, Ryan R Cheng, Sonja Müller-Späth, Shawn H Pfeil, Armin Hoffmann, Everett A Lipman, Dmitrii E Makarov, and Benjamin Schuler. Quantifying internal friction in unfolded and intrinsically disordered proteins with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.*, 109(44):17800–17806, October 2012.

[42] Andrea Soranno, Andrea Holla, Fabian Dingfelder, Daniel Nettels, Dmitrii E Makarov, and Benjamin Schuler. Integrated view of internal friction in unfolded proteins from single-molecule FRET, contact quenching, theory, and simulations. *Proc. Natl. Acad. Sci. U. S. A.*, 114(10):E1833–E1839, March 2017.

[43] Paul G Higgs and Jean-françois Joanny. Theory of polyampholyte solutions. *J. Chem. Phys.*, 94(2):1543–1554, January 1991.

[44] Chung-Ke Chang, Yen-Lan Hsu, Yuan-Hsiang Chang, Fa-An Chao, Ming-Chya Wu, Yu-Shan Huang, Chin-Kun Hu, and Tai-Huang Huang. Multiple nucleic acid binding sites and intrinsic disorder of severe acute respiratory syndrome coronavirus nucleocapsid protein: implications for ribonucleocapsid protein packaging. *J. Virol.*, 83(5):2255–2264, March 2009.

[45] Weihong Zeng, Guangfeng Liu, Huan Ma, Dan Zhao, Yunru Yang, Muziying Liu, Ahmed Mohammed, Changcheng Zhao, Yun Yang, Jiajia Xie, Chengchao Ding, Xiaoling Ma, Jianping Weng, Yong Gao, Hongliang He, and Tengchuan Jin. Biochemical characterization of SARS-CoV-2 nucleocapsid protein. *Biochem. Biophys. Res. Commun.*, 527(3):618–623, June 2020.

[46] Markus Sauer and Hannes Neuweiler. PET-FCS: probing rapid structural fluctuations of proteins and nucleic acids by single-molecule fluorescence quenching. *Methods Mol. Biol.*, 1076:597–615, 2014.

[47] Dominik Haenni, Franziska Zosel, Luc Reymond, Daniel Nettels, and Benjamin Schuler. Intramolecular distances and dynamics from the combined photon statistics of single-molecule FRET and photoinduced electron transfer. *J. Phys. Chem. B*, 117(42):13015–13028, October 2013.

[48] Wenli Meng, Nicholas Lyle, Bowu Luan, Daniel P Raleigh, and Rohit V Pappu. Experiments and simulations show how long-range contacts can form in expanded unfolded proteins with negligible secondary structure. *Proc. Natl. Acad. Sci. U. S. A.*, 110(6):2123–2128, February 2013.