Chapter 1

The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA.

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., Holehouse, A.S., 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 Abstract

The SARS-CoV-2 nucleocapsid (N) protein is an abundant RNA binding protein critical for viral genome packaging, yet the molecular details that underlie this process are poorly understood. Here we combine single-molecule spectroscopy with all-atom simulations to uncover the molecular details that contribute to N protein function. N protein contains three dynamic disordered regions that house putative transiently-helical binding motifs. The two folded domains interact minimally such that full-length N protein is a flexible and multivalent RNA binding protein. N protein also undergoes liquid-liquid phase separation when mixed with RNA, and polymer theory predicts that the same multivalent interactions that drive phase separation also engender RNA compaction. We offer a simple symmetry-breaking model that provides a plausible route through which single-genome condensation preferentially occurs over phase separation, suggesting that phase separation offers a convenient macroscopic readout of a key nanoscopic interaction.

1.2 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an enveloped, positive-strand RNA virus that causes the disease COVID-19 (Coronavirus Disease-2019) [2]. While coronaviruses typically cause relatively mild respiratory diseases, COVID-19 is on course to kill half a million people in the first six months since its emergence in late 2019 [2–4]. Given the timeframe for vaccine development is on the order of months to years, alternative therapeutic approaches are sought to ameliorate viral morbidity and mortality [5].

A challenge in identifying candidate drugs is our relatively sparse understanding of the molecular details that underlie the function of SARS-CoV-2 proteins. As a result, there is a surge of biochemical and biophysical exploration of these proteins, with the ultimate goal of identifying proteins that are suitable targets for disruption, ideally with insight into the molecular details of how disruption could be achieved [6, 7].

While much attention has been focused on the Spike (S) protein, many other SARS-CoV-2 proteins play equally critical roles in viral physiology, yet we know relatively little about their structural or biophysical properties [8–11].

Here we performed a high-resolution structural and biophysical characterization of the SARS-CoV-2 nucleocapsid (N) protein, the protein responsible for genome packaging [12]. A large fraction of N protein is predicted to be intrinsically disordered, which constitutes a major barrier to conventional structural characterization 13. To overcome these limitations, we combined single-molecule spectroscopy with all-atom simulations to build a residue-by-residue description of all three disordered regions in the context of their folded domains The combination of single-molecule spectroscopy and simulations to reconstruct structural ensembles has been applied extensively to uncover key molecular details underlying disordered protein regions 14–19. Our goal here is to provide biophysical and structural insights into the physical basis of N protein function.

In exploring the molecular properties of N protein, we discovered it undergoes phase separation with RNA, as was also reported recently 20–22. Given N protein underlies viral packaging, we reasoned phase separation may in fact be an unavoidable epiphenomenon that reflects physical properties necessary to drive compaction of long RNA molecules. To explore this principle further, we developed a simple physical model, which suggested symmetry breaking through a small number of high-affinity binding sites can organize anisotropic multivalent interactions to drive single-polymer compaction, as opposed to multi-polymer phase separation. Irrespective of its physiological role, our results suggest that phase separation provides a macroscopic readout (visible droplets) of a nanoscopic process (protein:RNA and protein:protein interaction). In the context of SARS-CoV-2, those interactions are expected to be key for viral packaging, such that assays which monitor phase separation of N protein with RNA may offer a convenient route to identify compounds that will also attenuate viral assembly.

1.3 Results

Coronavirus nucleocapsid proteins are multi-domain RNA binding proteins that play a critical role in many aspects of the viral life cycle 12,23. The SARS-CoV-2 N protein shares a number of sequence features with other nucleocapsid proteins from coronaviruses (Fig. **??**-**??**). Work on N protein from a range of model coronaviruses has shown that N protein undergoes both self-association, interaction with other proteins, and interaction with RNA, all in a highly multivalent manner.

The SARS-CoV-2 N protein can be divided into five domains; a predicted intrinsically disordered N-terminal domain (NTD), an RNA binding domain (RBD), a predicted disordered central linker (LINK), a dimerization domain, and a predicted disordered C-terminal domain (CTD) (Fig. 1.1). While SARS-CoV-2 is a novel coronavirus, decades of work on model coronaviruses (including SARS coronavirus) have revealed a number of features expected to hold true in the SARS-CoV-2 N protein. Notably, all five domains are predicted to bind RNA 25–31, and while the dimerization domain facilitates the formation of well-defined stoichiometric dimers, RNA-independent higher-order oligomerization is also expected to occur 30,32–34. Importantly, protein-protein and protein-RNA interaction sites have been mapped to all three disordered regions.

Despite recent structures of the RBD and dimerization domains from SARS-CoV-2, the solution-state conformational behavior of the full-length protein remains elusive 35–37. Understanding N protein function necessitates a mechanistic understanding of the flexible predicted disordered regions and their interplay with the folded domains. A recent small-angle X-ray study shows good agreement with previous work on SARS, suggesting the LINK is relatively extended, but neither the structural basis for this extension nor the underlying dynamics are known 25,38.

Here, we address these questions by probing three three full-length constructs of the N protein with fluorescent labels (Alexa 488 and 594) flanking the NTD, the LINK, and the CTD (see Fig. 1.1A). These constructs allow us to probe conformations and dynamics of the disordered regions in the context of the full-length protein using single-molecule Förster Resonance Energy Transfer (FRET) and Fluorescence Correlation Spectroscopy (FCS) (see SI for details). In parallel to the experiments, we performed all-atom Monte Carlo simulations of each of the three IDRs in isolation and in context with their adjacent folded domains.

1.3.1 The NTD is disordered, flexible, and transiently interacts with the RBD.

We started our analysis by investigating the NTD conformations. Under native conditions, single-molecule FRET measurements revealed the occurrence of a single population with a mean transfer efficiency of 0.61 ± 0.03 (Fig. 1.2A and Fig. **??**). To assess whether this transfer efficiency reports about a rigid distance (e.g. structure formation

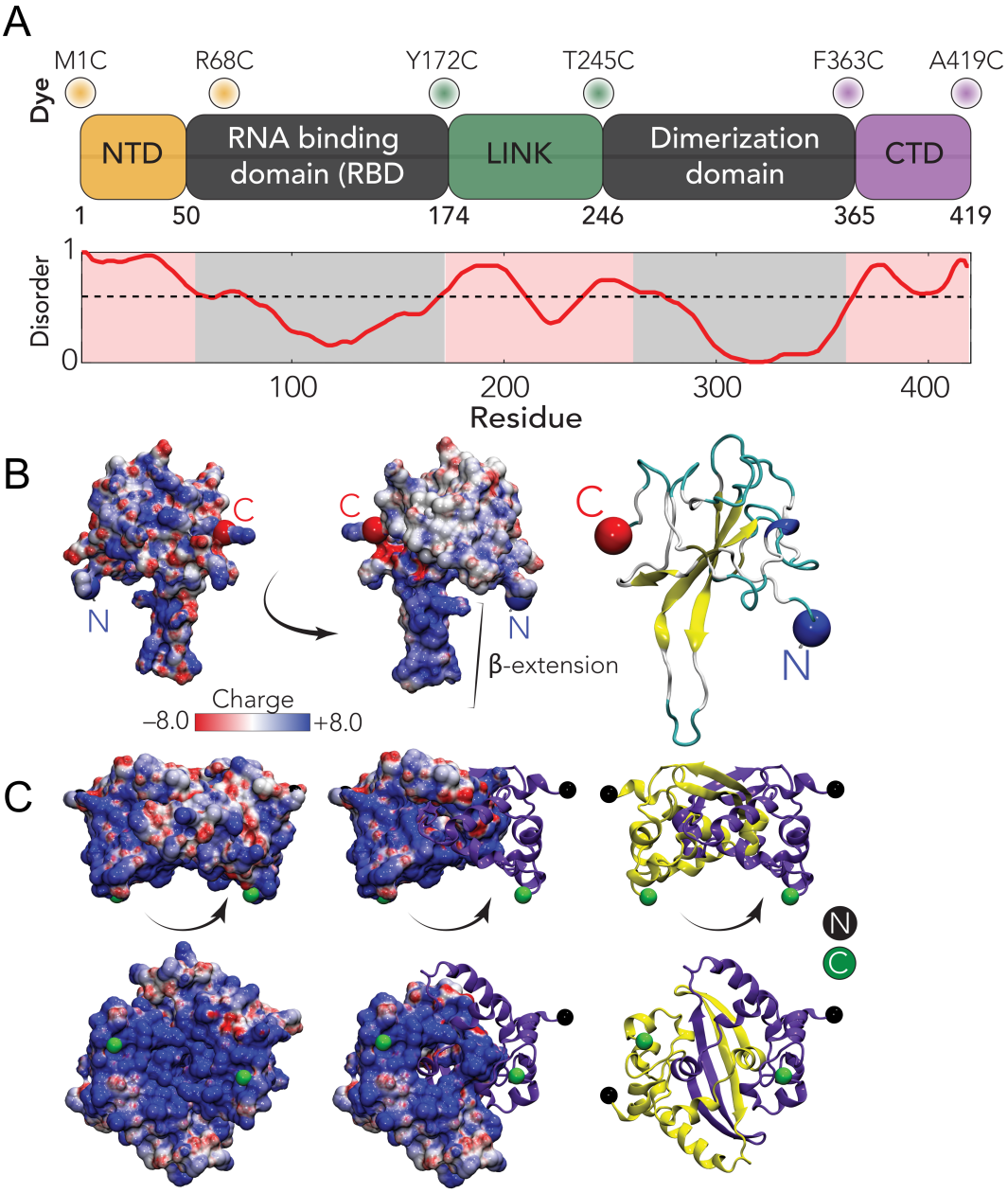

Figure 1.1: Sequence and structural summary of N protein . **A.** Domain architecture of the SARS-CoV-2 N protein. Dye positions used in this study are annotated across the top, disorder prediction calculated across the bottom. The specific positions were selected such that fluorophores are sufficiently close to be in the dynamic range of FRET measurements. Labeling was achieved using cysteine mutations and thiol-maleimide chemistry. **B.** Structure of the SARS-CoV-2 RNA binding domain (RBD) (PDB: 6yi3). Center and left: coloured based on surface potential calculated with the Adaptive Poisson Boltzmann Method 24, revealing the highly basic surface of the RBD. Right: ribbon structure with N- and C-termini highlighted. **C.** Dimer structure of the SARS-CoV-2 dimerization domain (PDB: 6yun). Center and left: coloured based on surface potential, revealing the highly basic surface. Right: ribbon structure with N- and C-termini highlighted.

or persistent interaction with the RBD) or is a dynamic average across multiple conformations, we first compare the lifetime of the fluorophores with transfer efficiency. Under native conditions, the donor and acceptor lifetimes for the NTD construct lie on the line that represents fast conformational dynamics (Fig. **??**A). To properly quantify the timescale associated with these fast structural rearrangements, we leveraged nanoseconds FCS. As expected for a dynamic population 39,40, the cross-correlation of acceptor-donor photons for the NTD is anticorrelated (Fig. 1.2B and **??**). A global fit of the donor-donor, acceptor-acceptor, and acceptor-donor correlations yields a reconfiguration time τ_r = 170 $\pm$ 30 ns. This is longer than reconfiguration times observed for other proteins with a similar persistence length and charge content 40–43, hinting at a large contribution from internal friction due to rapid intramolecular contacts (formed either within the NTD or with the RBD) or transient formation of short structural motifs 44.

As a next step, we assessed the stability of the folded RBD and its influence on the conformations of the NTD by studying the effect of a chemical denaturant on the protein. The titration with guanidinium chloride (GdmCl) reveals a decrease of transfer efficiencies when moving from native buffer conditions to 1 M GdmCl, followed by a plateau of the transfer efficiencies at concentrations between 1 M and 2 M and a subsequent further decrease at higher concentrations (Fig. **??** and **??**). This behavior can be understood assuming that the plateau between 1 M and 2 M GdmCl represents the average of transfer efficiencies between two populations in equilibrium that have very close transfer efficiency and are not resolved because of shot noise. Indeed, this interpretation is supported by a broadening in the transfer efficiency peak between 1 M and 2 M GdmCl, which is expected if two overlapping populations react differently to denaturant. Besides the effect of the unfolding of the RBD, the dimensions of the NTD are also modulated by change in the solvent quality when adding denaturant (Fig. 1.2C, **??**, **??**) and this contribution to the expansion of the chain can be described using an empirical binding model 45–49. A fit of the interdye root-mean-square distances to this model and the extracted stability of RBD (midpoint: 1.25 $\pm$ 0.2 M; ΔG_0 = (3 $\pm$ 0.6) RT) is presented in Fig. 1.2C A comparative fit of the histograms with two populations yields an identical result in terms of RBD stability and protein conformations (Fig. **??**).

These observations provide two important insights. Firstly, the RBD is completely folded under native conditions (Fig. 1.2C). Secondly, the RBD contributes significantly to the conformations of the measured NTD construct, mainly by reducing the accessible space of the disordered tail and favoring expanded configurations, as shown by the shift in transfer efficiency when the RBD is unfolded.

To better understand the sequence-dependent conformational behavior of the NTD we turned to all-atom simulations of an NTD-RBD construct. We used a novel sequential sampling approach that integrates long timescale MD simulations performed using the Folding@home distributed computing platform with all-atom Monte Carlo simulation performed with the ABSINTH forcefield to generate an ensemble of almost 400,000 distinct conformations (see methods) 50,51. We also performed simulations of the NTD in isolation.

We observed excellent agreement between simulation and experiment for the equivalent inter-residue distance (Fig. 1.2D). The peaks on the left side of the histogram reflect specific simulations where the NTD engages more extensively with the RBD through a fuzzy interaction, leading to local kinetic traps 52. We also identified several regions in the NTD where transient helices form, and using normalized distance maps found regions of transient attractive and repulsive interaction between the NTD and the RBD. In particular, the basic beta-strand extension from the RBD (Fig. 1.1B) repels the arginine-rich C-terminal region of the NTD, while a phenylalanine residue (F17) in the NTD engages with a hydrophobic face on the RBD (Fig. 1.2G). Finally, we noticed the arginine-rich C-terminal residues (residues 31 - 41) form a transient alpha helix projecting three of the four arginines in the same direction (Fig. 1.2H). These features provide molecular insight into previously reported functional observations (see Discussion).

1.3.2 The linker is highly dynamic and there is minimal interaction between the RBD and the dimerization domain.

We next turned to the linker (LINK) construct to investigate how the disordered region modulates the interaction and dynamics between the two folded domains. Under native conditions (50 mM Tris buffer), single-molecule FRET reveals a narrow population with mean transfer efficiency of 0.52 $\pm$ 0.03. Comparison of the fluorescence lifetime and transfer efficiency indicates that, like the NTD, the transfer efficiency represents a dynamic conformational ensemble sampled by the LINK (Fig. **??**B). ns-FCS confirms fast dynamics with a characteristic reconfiguration time τ_r of 120 $\pm$ 20 ns (Fig. 1.3B and **??**). This reconfiguration time is compatible with high internal friction effects, as observed for other unstructured proteins 40,41, but may also account for the drag of the surrounding domains. The root-mean-

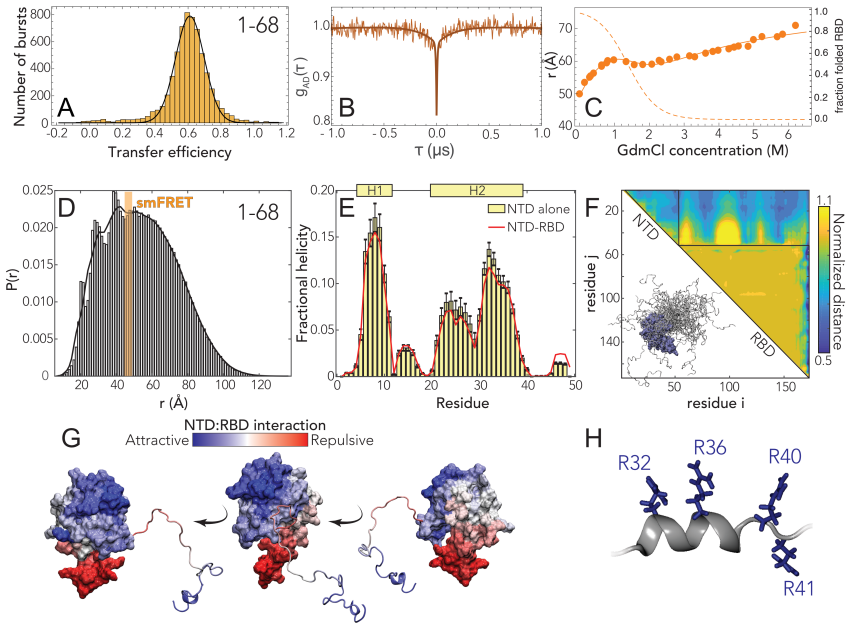

Figure 1.2: The N-terminal domain (NTD) is disordered with residual helical motifs. **A.** Histogram of the transfer efficiency distribution measured across the labeling positions 1 and 68 in the context of the full-length protein, under native conditions (50 mM Tris buffer). **B.** Donor-acceptor cross-correlation measured by ns-FCS (see SI). The observed anticorrelated rise is the characteristic signature of FRET dynamics and the timescale associated is directly related to the reconfiguration time of the probed segment. **C.** Interdye distance as extracted from single-molecule FRET experiments across different concentrations using a Gaussian chain distribution, examining residues 1-68 in the context of the full length protein. The full line represents a fit to the model in Eq. **??**, which accounts for denaturant binding (see Table **??**) and unfolding of the folded RBD. The dashed line represents the estimate of folded RBD across different denaturant concentrations based on Eq. **?? D.** All-atom simulations of the NTD in the context of RBD reveal good agreement with smFRET-derived average distances. The peaks on the left shoulder of the histogram are due to persistent NTD-RBD interactions in a small subset of simulations. **E.** Transient helicity in the NTD in isolation or in the context of the RBD. Perfect profile overlap suggests interaction between the NTD and the RBD does not lead to a loss of helicity. **F.** Normalized distance maps (scaling maps) quantify heterogeneous interaction between every pair of residues in terms of average distance normalized by distance expected for the same system if the IDR had no attractive interactions (the "excluded volume" limit 53). Both repulsive (yellow) and attractive (blue) regions are observed for NTD-RBD interactions. **G.** Projection of normalized distances onto the folded domain reveals repulsion is through electrostatic interaction (positively charged NTD is repelled by the positive face of the RBD, which is proposed to engage in RNA binding) while attractive interactions are between positive, aromatic, and polar residues in the NTD and a slightly negative and hydrophobic surface on the RBD (see Fig. 1.1B, center). **H.** The C-terminal half of transient helicity in H2 encodes an arginine-rich surface.

square interdye distance for the LINK $r_{172-245}$ is equal to 57 $\pm$ 2 Å (l_p= 5.8 $\pm$ 0.4 Å) when assuming a Gaussian Chain distribution and 55 $\pm$ 2 Å (l_p= 5.4 $\pm$ 0.4 Å) when using a SAW model (see SI).

Next, we addressed whether the LINK segment populates elements of persistent secondary structure or forms stable interaction with the RBD or dimerization domains. Addition of the denaturant shows a continuous shift of the transfer efficiency toward lower values (Fig. **??**,**??**), that corresponds to an almost linear expansion of the chain (see Fig. 1.3C). These observations support a model in which LINK is unstructured and flexible and do not reveal a significant fraction of folding or persistent interactions with or between folded domains. Overall, our single-molecule observations report a relatively extended average inter-domain distance, suggesting a low number of interactions between folded domains. To further explore this conclusion, we turned again to Monte Carlo simulations.

As with the NTD, all-atom Monte Carlo simulations provide atomistic insight that can be compared with our spectroscopic results. Given the size of the system an alternative sampling strategy to the NTD-RBD construct was pursued here that did not include MD simulations of the folded domains, but we instead ran simulations of a construct that included the RBD, LINK and dimerization domain. In addition, we also performed simulations of the LINK in isolation.

We again found good agreement between simulations and experiment (Fig. 1.3D). The root mean square inter-residue distance between simulated positions 172 and 245 is 59.1 Å, which is within the experimental error of the single-molecule observations. Normalized distance map shows a number of regions of repulsion, notably that the RBD repels the N-terminal part of the LINK and the dimerization domain repels the C-terminal part of the LINK (Fig. 1.3E). We tentatively suggest this may reflect sequence properties chosen to prevent aberrant interactions between the LINK and the two folded domains. In the LINK-only simulations we identified two regions that form transient helices at low populations (20-25%), although these are much less prominent in the context of the full length protein (Fig. 1.3F). Those helices encompass a serine-arginine (SR) rich region known to mediate both protein-protein and protein-RNA interaction, and leads to the alignment of three arginine residues along one face of a helix. The second helix (H4) is a leucine/alanine-rich hydrophobic helix which may contribute to oligomerization, or act as a helical recognition motif for other protein interactions (notably as a nuclear export signal for Crm1, see Discussion).

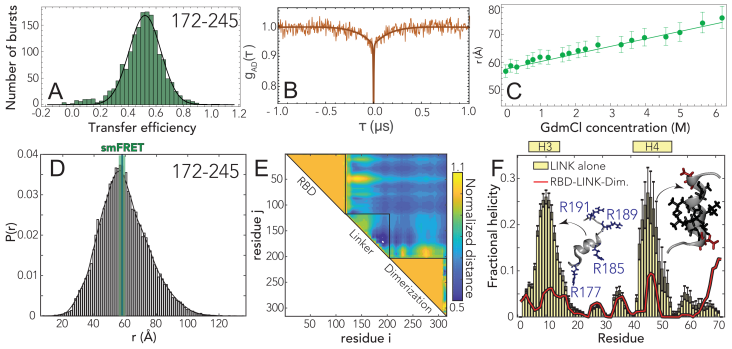

Figure 1.3: The RNA binding domain (RBD) and dimerization domains do not significantly interact and are connected by a disordered linker (LINK) **A.** Histogram of the transfer efficiency distribution measured across the labeling positions 172 and 245 in the context of the full-length protein, under native conditions (50 mM Tris buffer). **B.** Donor-acceptor cross-correlation measured by ns-FCS (see SI). The observed anticorrelated rise is the characteristic signature of FRET dynamics and the timescale associated is directly related to the reconfiguration time of the probed segment. **C.** Interdye distance as extracted from single-molecule FRET experiments across different denaturant concentrations. The full line represents a fit to the model in Eq. **??**, which accounts for denaturant binding. **D** Inter-residue distance distributions calculated from simulations (histogram) show good agreement with distances inferred from single-molecule FRET measurements (green bar). **E.** Scaling maps reveal repulsive interactions between the N- and C-terminal regions of the LINK with the adjacent folded domains. We also observe relatively extensive intra-LINK interactions around helix H4 (see Fig. 1.3F). **F.** Two transient helices are observed in the linker. The N-terminal helix H3 overlaps with part of the SR-region and orientates three arginine residues in the same direction, analogous to behavior observed for H2 in the NTD. The C-terminal helix H4 overlaps with a Leu/Ala rich motif which we believe is a conserved nuclear export signal (see Discussion).

1.3.3 The CTD engages in transient but non-negligible interactions with the dimerization domain.

Finally, we turned to the CTD. Single-molecule FRET experiments again reveal a single population with a mean transfer efficiency of 0.59 $\pm$ 0.03 (Fig. 1.4A) and the denaturant dependence follows the expected trend for a disordered region, with a shift of the transfer efficiency toward lower values (Fig. **??**,**??**), from 0.59 to 0.35. Interestingly, when studying the denaturant dependence of the protein, we noticed that the width of the distribution increases while moving

toward native conditions. This suggests that the protein may form transient contacts or adopt local structure. To investigate this aspect, we turned to the investigation of the dynamics. Though the comparison of the fluorophore lifetimes against transfer efficiency (Fig. **??**c) appears to support a dynamic nature underlying this population, nanosecond FCS reveals a flat acceptor-donor cross-correlation on the ns timescale (Fig. 1.4B). However, inspection of the donor-donor and acceptor-acceptor autocorrelations reveal a correlated decay with a characteristic time of 240 $\pm$ 50 ns. This is different from that expected for a completely static system such as polyprolines 54, where the donor-donor and acceptor-acceptor autocorrelation are also flat. An increase in the autocorrelations can be observed for static quenching of the dyes with aromatic residues.Interestingly, donor dye quenching can also contribute to a positive amplitude in the donor-acceptor correlation 55,56. Therefore, a plausible interpretation of the flat cross-correlation data is that we are observing two populations in equilibrium whose correlations (one anticorrelated, reflecting conformational dynamics, and one correlated, reflecting quenching due contact formation) compensate each other.

To further investigate the possible coexistence of these different species, we performed ns-FCS at 0.2 M GdmCl, where the width of the FRET population starts decreasing and the mean transfer efficiency is slightly shifted to larger values, under the assumption that the decreased width of the population reflects reduced interactions. Indeed, the cross-correlation of ns-FCS reveals a dynamic behavior with a reconfiguration time τ_r= 70 $\pm$ 15 ns (Fig. **??**). Based on these observations, we suggest that a very similar disordered population to the one observed at 0.2 M is also present under native conditions, but in equilibrium with a quenched species that forms long-lived contacts. Under the assumption that the mean transfer efficiency still originates (at least partially) from a dynamic distribution, the estimate of the inter-residue root-mean-square distance is $r_{363-419}$ = 51 $\pm$ 2 Å (l_p= 6.1 $\pm$ 0.4 Å) for a Gaussian chain distribution and $r_{363-419}$ = 49 $\pm$ 2 Å (l_p= 5.6 $\pm$ 0.4 Å) for the SAW model (see SI). However, some caution should be used when interpreting these numbers since we know there is some contribution from fluorophore quenching, which may in turn contribute to an underestimate of the effective transfer efficiency 57.

We again obtained good agreement between all-atom Monte Carlo simulations and experiment (Fig. 1.4D). We identified two transient helices, one (H5) is minimally populated but the second (H6) is more highly populated in the IDR-only simulation and still present at $\sim$20% in the folded state simulations (Fig. 1.4E). The difference reflects the fact that several of the helix-forming residues interact with the dimerization domain, leading to a competition between helix formation and intramolecular interaction. Scaling maps reveal extensive intramolecular interaction by the residues that make up H6, both in terms of local intra-IDR interactions and interaction with the dimerization domain (Fig. 1.4F). Mapping normalized distances onto the folded structure reveals that interactions occur primarily with the N-terminal portion of the dimerization domain (Fig. 1.4G). As with the LINK and the NTD, a positively charged set of residues immediately adjacent to the folded domain in the CTD drive repulsion between this region and the dimerization domain. H6 is the most robust helix observed across all three IDRs, and is a perfect amphipathic helix with a hydrophobic surface on one side and charged/polar residues on the other (Fig. 1.4H). The cluster of hydrophobic residues in H6 engage in intramolecular contacts and offer a likely physical explanation for the complex lifetime data.

1.3.4 N protein undergoes phase separation with RNA.

Over the last decade, biomolecular condensates formed through phase separation have emerged as a new mode of cellular organization 58–61. Given the high interaction valency and the presence of molecular features similar to other proteins we had previously studied, we anticipated that N protein would undergo phase separation with RNA 62–64.

In line with this expectation, we observed robust droplet formation with homopolymeric RNA (Fig. 1.5A-B) under native buffer conditions (50 mM Tris) and at higher salt concentration (50 mM NaCl). Turbidity assays at different concentrations of protein and poly(rU) (200-250 nucleotides) demonstrate the classical reentrant phase behavior expected for a system undergoing heterotypic interaction (Fig. 1.5C-D). It is to be noted that turbidity experiments do not exhaustively cover all the conditions for phase separation and are only indicative of the low-boundary concentration regime explored in the current experiments. In particular, turbidity experiments do not provide a measurement of tie-lines, though they are inherently a reflection of the free energy and chemical potential of the solution mixture 65. Interestingly, phase separation occurs at relatively low concentrations, in the low μM range, which are compatible with physiological concentration of the protein and nucleic acids. Though increasing salt concentration results in an upshift of the phase boundaries, one has to consider that in a cellular environment this effect might be counteracted by cellular crowding.

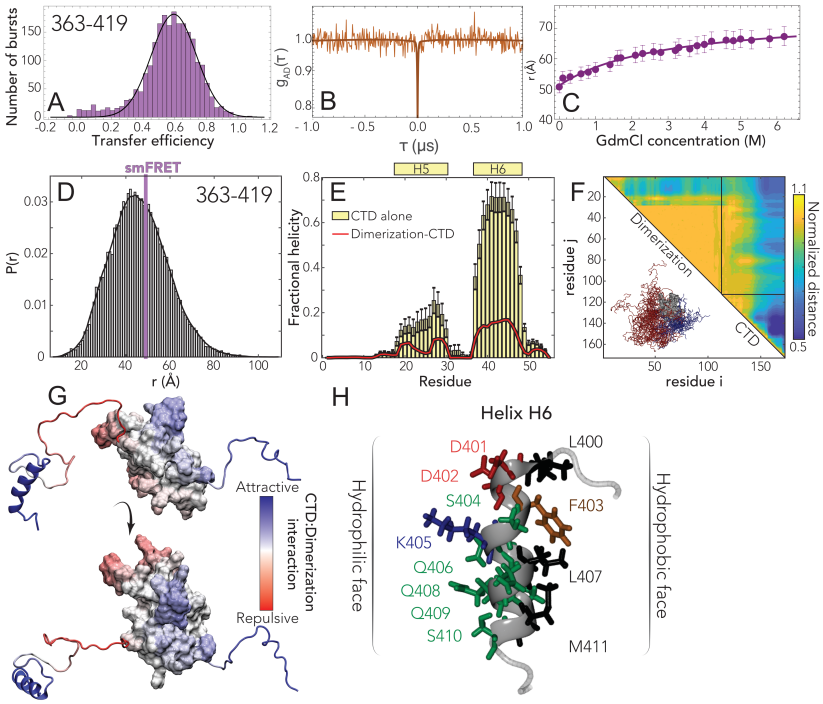

Figure 1.4: The C-terminal domain (CTD) is disordered, engages in transient interaction with the dimerization domain, and contains a putative helical binding motif. **A.** Histogram of the transfer efficiency distribution measured across the labeling positions 363 and 419 in the context of the full-length protein, under native conditions (50 mM Tris buffer). **B.** Donor-acceptor cross-correlation measured by ns-FCS (see SI). The flat correlation indicates a lack of dynamics in the studied timescale or the coexistence of two populations in equilibrium whose correlations (one correlated and the other anticorrelated) compensate each other. **C.** Interdye distance as extracted from single-molecule FRET experiments across different denaturant concentrations. The full line represents a fit to the model in Eq. **??**, which accounts for denaturant binding. **D.** Inter-residue distance distributions calculated from simulations (histogram) show good agreement with distances inferred from single-molecule FRET measurements (purple bar). **E.** Two transient helices (H5 and H6) are observed in the CTD. Both show a reduction in population in the presence of the dimerization domain at least in part because the same sets of residues engage in transient interactions with the dimerization domain. **F.** Normalized contacts maps describe the average inter-residue distance between each pair of residues, normalized by the distance expected if the CTD behaved as a self-avoiding random coil. H6 engages in extensive intra-CTD interactions and also interacts with the dimerization domain. We observe repulsion between the dimerization domain and the N-terminal region of the CTD. **G.** The normalized distances are projected onto the surface to map CTD-dimerization interaction. The helical region drives intra-molecular interaction, predominantly with the N-terminal side of the dimerization domain. **H.** Helix H6 is an amphipathic helix with a polar/charged surface (left) and a hydrophobic surface (right).

One peculiar characteristic of our measured phase-diagram is the narrow regime of conditions in which we observe phase separation of nonspecific RNA at a fixed concentration of protein. This leads us to hypothesize that the protein may have evolved to maintain a tight control of concentrations at which phase separation can (or cannot) occur. Interestingly, when rescaling the turbidity curves as a ratio between protein and RNA, we find all the curve maxima aligning at a similar stoichiometry, approximately 20 nucleotides per protein in absence of added salt and 30 nucleotides when adding 50 mM NaCl (Fig. **??**). These ratios are in line with the charge neutralization criterion proposed by Banerjee et al., since the estimated net charge of the protein at pH 7.4 is +24 66. Finally, given we observed phase separation with poly(rU), the behavior we are observing is likely driven by relatively nonspecific protein:RNA interactions. In agreement, work from the Gladfelter 20, Fawzi 21, Zweckstetter 22, and Yildiz (unpublished) labs have also established

this phenomenon across a range of solution conditions and RNA types.

Having established phase separation through a number of assays, we wondered what -if any- physiological relevance this may have for the normal biology of SARS-CoV-2.

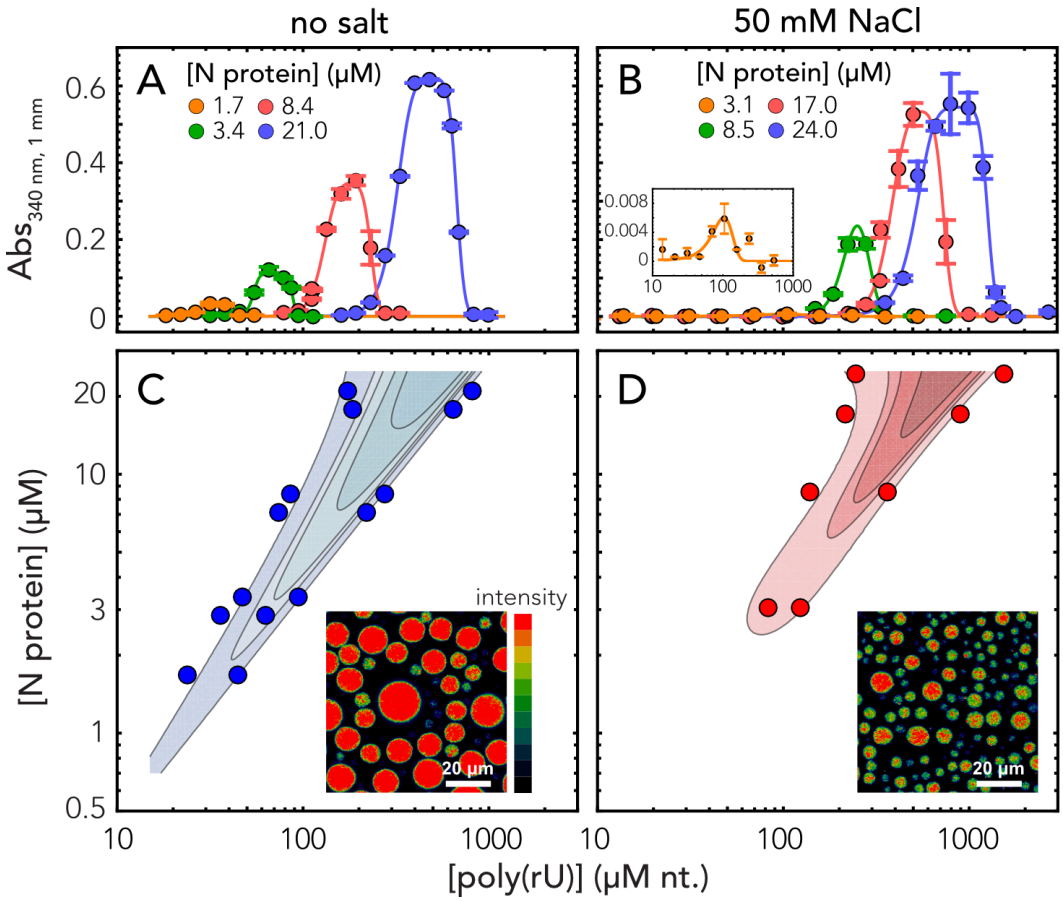

Figure 1.5: Nucleocapsid protein undergoes phase separation with RNA. **A-B.** Appearance of solution turbidity upon mixing was monitored to determine the concentration regime in which N protein and poly(rU) undergo phase separation. Representative turbidity titrations with poly(rU) in 50 mM Tris, pH 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. 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 (see SI). 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. **C-D.** Projection of phase boundaries for poly(rU) and N protein mixtures highlights a re-entrant behavior, as expected for phase-separations induced by heterotypic interactions. Turbidity contour lines are computed from a global fit of all titrations curves (see Appendix E)

1.3.5 A simple polymer model shows symmetry-breaking can facilitate multiple metastable single-polymer condensates instead of a single multi-polymer condensate.

Why might phase separation of N protein with RNA be advantageous to SARS-CoV-2? One possible model is that large, micron-sized cytoplasmic condensates of N protein with RNA form through phase separation and play a role in genome packaging. These condensates may act as molecular factories that help concentrate the components for pre-capsid assembly (where we define a pre-capsid here simply as a species that contains a single copy of the genome

with multiple copies of the associated N protein), a model that has been proposed in other viruses 67.

However, given that phase separation is unavoidable when high concentrations of multivalent species are combined, we propose that an alternative interpretation of our data is that in this context, phase separation is simply an inevitable epiphenomenon that reflects the inherent multi-valency of the N protein for itself and for RNA. This poses questions about the origin of specificity for viral genomic RNA (gRNA), and, of focus in our study, how phase separation might relate to a single genome packaging through RNA compaction.

Given the expectation of a single genome per virion, we reasoned SARS-CoV-2 may have evolved a mechanism to limit phase separation with gRNA (i.e. to avoid multi-genome condensates), with a preference instead for single-genome packaging (single-genome condensates). This mechanism may exist in competition with the intrinsic phase separation of the N protein with other nonspecific RNAs (nsRNA).

One possible way to limit phase separation between two components (e.g. gRNA/nsRNA and N protein) is to ensure the levels of these components are held at a sufficiently low total concentration such that the phase boundary is never crossed. While possible, such a regulatory mechanism is at the mercy of extrinsic factors that may substantially modulate the saturation concentration 68–70. Furthermore, not only must phase separation be prevented, but gRNA compaction should also be promoted through the binding of N protein. In this scenario, the affinity between gRNA and N protein plays a central role in determining the required concentration for condensation of the macromolecule (gRNA) by the ligand (N protein).

Given a defined valence of the system components, phase boundaries are encoded by the strength of interaction between the interacting domains in the components. Considering a long polymer (e.g. gRNA) with proteins adsorbed onto that polymer as adhesive points ("stickers"), the physics of associative polymers predicts that the same interactions that cause phase separation will also control the condensation of individual long polymers 62,71–75. With this in mind, we hypothesized that phase separation is reporting on the physical interactions that underlie genome compaction.

To explore this hypothesis, we developed a simple computational model where the interplay between compaction and phase separation could be explored. Our setup consists of two types of species: long multivalent polymers and short multivalent binders (Fig. 1.6A). All interactions are isotropic and each bead is inherently multivalent as a result. In the simplest instantiation of this model, favourable polymer:binder and binder:binder interactions are encoded, mimicking the scenario in which a binder (e.g. a protein) can engage in nonspecific polymer (RNA) interaction as well as binder-binder (protein-protein) interaction.

Simulations of binder and polymer undergo phase separation in a concentration-dependent manner, as expected (Fig. 1.6B,C). Phase separation gives rise to a single large spherical cluster with multiple polymers and binders (Fig. 1.6D, 1.6H). For a homopolymer, the balance of chain-compaction and phase separation is determined in part through chain length and binder K_d. In our system the polymer is largely unbound in the one-phase regime (suggesting the concentration of ligand in the one-phase space is below the K_d) but entirely coated in the two-phase regime, consistent with highly-cooperative binding behavior. In the limit of long, multivalent polymers with multivalent binders, the sharpness of the coil-to-globule transition is such that an effective two-state description of the chain emerges, in which the chain is either expanded (non-phase separation-competent) OR compact (coated with binders, phase separation competent).

In light of these observations, we wondered if a break in the symmetry between intra- and inter-molecular interactions would be enough to promote single-polymer condensation in the same concentration regime over which we had previously observed phase separation. Symmetry breaking in our model is achieved through a single high-affinity binding site (Fig. 1.6A). We choose this particular mode of symmetry-breaking to mimic the presence of a packaging signal -a region of the genome that is essential for efficient viral packaging- an established feature in many viruses (including coronaviruses) although we emphasize this is a general model, as opposed to trying to directly model gRNA with a packaging signal 76–78.

We performed identical simulations to those in Fig. 1.6C-D using the same system with polymers that now possess a single high affinity binding site (Fig. 1.6E). Under these conditions we did not observe large phase separated droplets (Fig. 1.6F). Instead, each individual polymer undergoes collapse to form a single-polymer condensate (Fig. 1.6E). Collapse is driven by the recruitment of binders to the high-affinity site, where they "coat" the chain, forming a local cluster of binders on the polymer. This cluster is then able to interact with the remaining regions of the polymer through weak "nonspecific" interactions, the same interactions that drove phase separation in Fig. 1.6B,C,D. Symmetry breaking is achieved because the local concentration of binder around the site is high, such that intramolecular

interactions are favoured over intermolecular interaction. This high local concentration also drives compaction at low binder concentrations. As a result, instead of a single multi-polymer condensate, we observe multiple single-polymers condensates, where the absolute number matches the number of polymers in the system (Fig. 1.6G).

Our results can also be cast in terms of two distinct concentration (phase) boundaries - one for binder:high affinity site interaction (c_1), and a second boundary for "nonspecific" binder:polymer interactions (c_2) at a higher concentration. c_2 reflects the boundary observed in Fig. 1.6C that delineated the one and two-phase regimes. At global concentrations below c_2, (but above c_1) the clustering of binders at a high affinity site raises the apparent local concentration of binders above c_2, from the perspective of other beads on the chain. In this way, a local high affinity binding site can drive "local" phase separation of a single polymer.

The high affinity binding site polarizes the single-polymer condensate, such that they are organized, recalcitrant to fusion, and kinetically stable. A convenient physical analogy is that of a micelle, which are non-stoichiometric stable assemblies. Even for micelles that are far from their optimal size, fusion is slow because it requires substantial molecular reorganization and the breaking of stable interactions 79,80.

Finally, we ran simulations under conditions in which binder:polymer interactions were reduced, mimicking the scenario in which non-specific protein:RNA interactions are inhibited (Fig. 1.6L). Under these conditions no phase separation occurs for polymers that lack a high-affinity binding site, while for polymers with a high-affinity binding site no chain compaction occurs (in contrast to when binder:polymer interactions are present, see Fig. 1.6J). This result illustrates how phase separation offers a convenient readout for molecular interactions that might otherwise be challenging to measure.

We emphasize that our conclusions from simulations are subject to the parameters in our model. We present these results to demonstrate an example of "how this single-genome packaging could be achieved", as opposed to the much stronger statement of proposing "this is how it is" achieved. Recent elegant work by Ranganathan and Shakhnovich identified kinetically arrested microclusters, where slow kinetics result from the saturation of stickers within those clusters 81. This is completely analogous to our results (albeit with homotypic interactions, rather than heterotypic interactions), giving us confidence that the physical principles uncovered are robust and, we tentatively suggest, quite general. Future simulations are required to systematically explore the details of the relevant parameter space in our system. However, regardless of those parameters, our model does establish that if weak multivalent interactions underlie the formation of large multi-polymer droplets, those same interactions cannot also drive polymer compaction inside the droplet

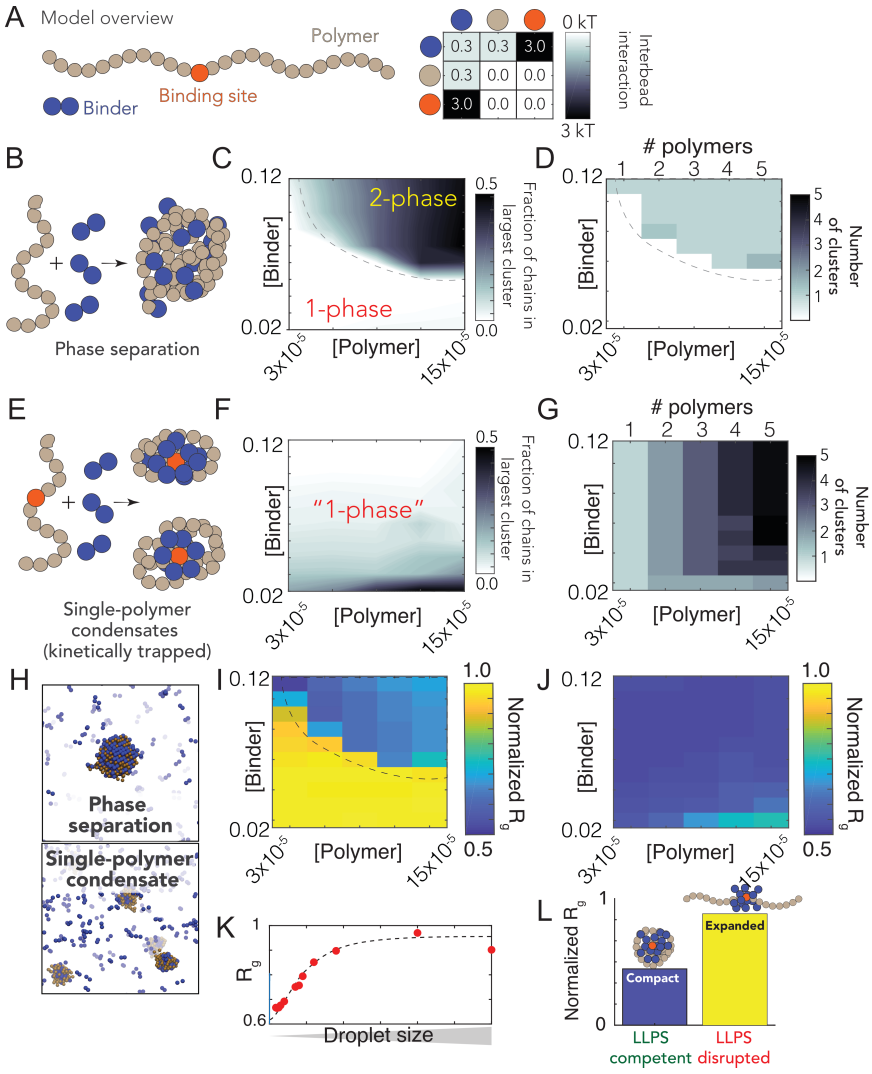

Figure 1.6: A simple polymer suggests symmetry breaking can promote single-polymer condensates over multi-polymer assemblies. **A.** Summary of our model setup, which involves long 'polymers' (61 beads) or short 'binders' (2 beads). Each bead is multivalent and can interact with every adjacent lattice site. The interaction matrix to the right defines the pairwise interaction energies associated with each of the bead times. **B.** Concentration dependent assembly behavior for polymers lacking a high-affinity binding site. **C.** Phase diagram showing the concentration-dependent phase regime - dashed line represents the binodal (phase boundary) and is provided to guide the eye. **D.** Analysis in the same 2D space as C assessing the number of droplets at a given concentration. When phase separation occurs a single droplet appears in almost all cases. **E.** TConcentration dependent assembly behavior for polymers with a high-affinity binding site. **F.** No large droplets are formed in any of the systems, although multiple polymer:binder complexes form. **G.** The number of clusters observed matches the number of polymers in the system - i.e. each polymer forms an individual cluster. **H.** Simulation snapshots from equivalent simulations for polymers with (top) or without (bottom) a single high-affinity binding site. **I.** Polymer dimensions in the dense and dilute phase (for the parameters in our model) for polymers with no high-affinity binding site. Note that compaction in the dense phase reflects finite-size effects, as addressed in panel K, and is an artefact of the relatively small droplets formed in our systems (relative to the size of the polymer). The droplets act as a bounding cage for the polymer, driving their compaction indirectly. **J.** Polymer dimensions across the same concentration space for polymers with a single high-affinity binding site. Across all concentrations, each individual polymer is highly compact. **K.** Compaction in the dense phase (panel I) is due to small droplets. When droplets are sufficiently large we observe chain expansion, as expected from standard theoretical descriptions. **L.** Simulations performed under conditions in which nonspecific interactions between binder and polymer are reduced (interaction strength = 0 kT). Under these conditions phase separation is suppressed. Equivalent simulations for polymers with a high-affinity site reveal these chains are no longer compact. As such, phase separation offers a readout that - in our model - maps to single-polymer compaction.

1.4 Discussion

The nucleocapsid (N) protein from SARS-CoV-2 is a multivalent RNA binding protein critical for viral replication and genome packaging 11,12. To better understand how the various folded and disordered domains interact with one another, we applied single-molecule spectroscopy and all-atom simulations to perform a detailed biophysical dissection of the protein, uncovering several putative interaction motifs. Furthermore, based on both sequence analysis and our single-molecule experiments, we anticipated that N protein would undergo phase separation with RNA. In agreement with this prediction, and in line with work from the Gladfelter and Yildiz groups working independently from us, we find that N protein robustly undergoes phase separation in vitro with model RNA under a range of different salt conditions. Using simple polymer models, we propose that the same interactions that drive phase separation may also drive genome packaging into a dynamic, single-genome condensate. The formation of single-genome condensates (as opposed to multi-genome droplets) is influenced by the presence of one (or more) symmetry-breaking interaction sites, which we tentatively suggest could reflect packaging signals in viral genomes.

1.4.1 All three IDRs are highly dynamic

Our single-molecule experiments and all-atom simulations are in good agreement with one another, and reveal that all three IDRs are extended and highly dynamic. Simulations suggest the NTD may interact transiently with the RBD, which offers an explanation for the slightly slowed reconfiguration time measured by nanosecond FCS. The LINK shows rapid rearrangement, demonstrating the RBD and dimerization domain are not interacting. Finally, we see more pronounced interaction between the CTD and the dimerization domain, although these interactions are still highly transient.

Single-molecule experiments and all-atom simulations were performed on monomeric versions of the protein, yet N protein has previously been shown to undergo dimerization and form higher-order oligomers in the absence of RNA 32. To assess the formation of oligomeric species, we use a combination of nativePAGE, crosslinking and FCS experiments (see Fig. **??**). These experiments also verified that under the conditions used for single-molecule experiments the protein exists only as a monomer.

1.4.2 Simulations identify multiple transient helices

We identified a number of transient helical motifs which provide structural insight into previously characterized molecular interactions. Transient helices are ubiquitous in viral disordered regions and have been shown to underlie molecular interactions in a range of systems 67,82–84.

Transient helix H2 (in the NTD) and H3 (in the LINK) flank the RBD and organize a set of arginine residues to face the same direction (Fig. 1.2E). Both the NTD and LINK have been shown to drive RNA binding, such that we propose these helical arginine-rich motifs (ARMs) may engage in both both nonspecific binding and may also contribute to RNA specificity, as has been proposed previously 25,85,86. The serine-arginine SR-region (which includes H3) has been previously identified as engaging in interaction with a structured acidic helix in Nsp3 in the model coronavirus MHV, consistent with an electrostatic helical interaction 87,88. Recent NMR data also shows excellent agreement with our results, identifying a transient helix that shows 1:1 overlap with H3 22.The SR-region is necessary for recruitment to replication-transcription centers (RTCs) in MHV, and also undergoes phosphorylation, setting the stage for a complex regulatory system awaiting exploration 89,90.

Transient helix H4 (Fig. 1.3H), was previously predicted bioinformatically and identified as a conserved feature across different coronaviruses 25. Furthermore, the equivalent region was identified in SARS coronavirus as a nuclear export signal (NES), such that we suspect this too is a classical Crm1-binding leucine-rich NES 91.

Transient helix H6 is an amphipathic helix with a highly hydrophobic face (Fig. 1.4H). Recent hydrogen-deuterium exchange mass spectrometry also identified H6 37. Residues in this region have previously been identified as mediating M-protein binding in other coronaviruses, such that we propose H6 underlies that interaction 92–94. Recent work has also identified amphipathic transient helices in disordered proteins as interacting directly with membranes, such that an additional (albeit entirely speculative) role could involve direct membrane interaction, as has been observed in other viral phosphoproteins 95,96.

1.4.3 The physiological relevance of nucleocapsid protein phase separation in SARS-CoV-2 physiology

Our work has revealed that SARS-CoV-2 N protein undergoes phase separation with RNA when reconstituted in vitro. The solution environment and types of RNA used in our experiments are very different from the cytoplasm and viral RNA. However, similar results have been obtained in published and unpublished work by several other groups under a variety of conditions, including via in cell experiments (Yildiz group, unpublished) 20–22. Taken together, these results demonstrate that N protein can undergo bona fide phase separation, and that N protein condensates can form in cells. Nevertheless, the complexity introduced by multidimensional linkage effects in vivo could substantially influence the phase behavior and composition of condensates observed in the cell 70,74,97. Of note, the regime we have identified in which phase separation occurs (Fig. 1.5) is remarkably relatively narrow, a prerequisite for the assembly of virion particles containing a single viral genome.

Does phase separation play a physiological role in SARS-CoV-2 biology? Phase separation has been invoked or suggested in many different viral contexts to date 98–102. In SARS-CoV-2, one possible model suggests phase separation may drive recruitment of components to viral replication sites, although how this dovetails with the fact that replication occurs in double-membrane bound vesicles (DMVs) remains to be explored 22,103. An alternative (and non-mutually exclusive) model is one in which phase separation catalyzes nucleocapsid polymerization, as has been proposed in elegant work on measles virus 67. Here, the process of phase separation is decoupled from genome packaging, where gRNA condensation occurs through association with a helical nucleocapsid. If applied to SARS-CoV-2, such a model would suggest that (1) initially N protein and RNA phase separate in the cytosol, (2) some discrete pre-capsid state forms within condensates and, (3) upon maturation, the pre-capsid is released from the condensate and undergoes subsequent virion assembly by interacting with the membrane-bound M, E, and S structural proteins at the ER-Golgi intermediate compartment (ERGIC). While this model is attractive it places a number of constraints on the physical properties of this pre-capsid, not least that the ability to escape the "parent" condensate dictates that the assembled pre-capsid must interact less strongly with the condensate components than in the unassembled state. This requirement introduces some thermodynamic complexities: how is a pre-capsid state driven to assemble if it is necessarily less stable than the unassembled pre-capsid, and how is incomplete or abortive pre-capsid formation avoided if – as assembly occurs – the pre-capsid becomes progressively less stable?

A phase separation and assembly model raises additional questions, such as the origins of specificity for recruitment of viral proteins and viral RNA, the kinetics of pre-capsid-assembly within a large condensate, and preferential packaging of gRNA over sub-genomic RNA. None of these questions are unanswerable, nor do they invalidate this model, but they should be addressed if the physiological relevance of large cytoplasmic condensates is to be further explored in the context of virion assembly.

Our preferred interpretation is that N protein has evolved to drive genome compaction for packaging (Fig. 1.7). In this model, a single-genome condensate forms through N protein gRNA interaction, driven by a small number of high-affinity sites. This (meta)-stable single-genome condensate undergoes subsequent maturation, leading to virion assembly. In this model, condensate-associated N proteins are in exchange with a bulk pool of soluble N protein, such that the interactions that drive compaction are heterogeneous and dynamic. Our model provides a physical mechanism in good empirical agreement with data for N protein oligomerization and assembly 104–106. Furthermore, the resulting condensate is then in effect a multivalent binder for M protein, which interacts with N directly, and may drive membrane curvature and budding in a manner similar to that proposed by Bergeron-Sandoval and Michnick (though with a different directionality of the force) and in line with recent observations from cryo electron tomography (cryoET) 103,107–109

An open question pertains to specificity of packaging gRNA while excluding other RNAs. One possibility is for two high-affinity N-protein binding sites to flank the 5' and 3' ends of the genome, whereby only RNA molecules with both sites are competent for compaction A recent map of N protein binding to gRNA has revealed high-affinity binding regions at the 5' and 3' ends of the gRNA, in good agreement with this qualitative prediction 20. Alternatively only gRNA condensates may possess the requisite valency to drive virion budding through interaction with M at the cytoplasmic side of the ERGIC, offering a physical selection mechanism for budding.

Genome compaction through dynamic multivalent interactions would be especially relevant for coronaviruses, which have extremely large single-stranded RNA genomes. This is evolutionarily appealing, in that as the genome grows larger, compaction becomes increasingly efficient, as the effective valence of the genome is increased 72,73.

The ability of multivalent disordered proteins to drive RNA compaction has been observed previously in various contexts 14,110. Furthermore, genome compaction by RNA binding protein has been proposed and observed in other viruses 106,111,112, and the SARS coronavirus N protein has previously been shown to act as an RNA chaperone, an expected consequence of compaction to a dynamic single-RNA condensate that accommodates multiple N proteins with a single RNA 14,113. Furthermore, previous work exploring the ultrastructure of phase separated condensates of G3BP1 and RNA through simulations and cryoET revealed a beads-on-a-string type architecture, mirroring recent results for obtained from cryoET of SARS-CoV-2 virions 63,103.

N protein has been shown to interact directly with a number of proteins studied in the context of biological phase separation which may influence assembly in vivo 5,21,62,69,114. In particular, G3BP1-an essential stress-granule protein that undergoes phase separation-was recently shown to co-localize with overexpressed N protein 22,63,69,115. G3BP1 interaction may be part of the innate immune response, leading to stress-granule formation, or alternatively N protein may attenuates the stress response by sequestering G3BP1, depleting the cytosolic pool, and preventing stress granule formation, as has been shown for HIV-1 102.

Our model is also in good empirical agreement with recent observations made for other viruses116. Taken together, we speculate that viral packaging may -in general- involve an initial genome compaction through multivalent protein:RNA and protein:protein interactions, followed by a liquid-to-solid transition in cases where well-defined crystalline capsid structures emerge. Liquid-to-solid transitions are well established in the context of neurodegeneration with respect to disease progression 117–119. Here we suggest nature is leveraging those same principles as an evolved mechanism for monodisperse particle assembly.

Regardless of if phase separated condensates form inside cells, all available evidence suggests phase separation is reporting on a physiologically important interaction that underlies genome compaction (Fig. 1.6L). With this in mind, from a biotechnology standpoint, phase separation may be a convenient readout for in vitro assays to interrogate protein:RNA interaction. Regardless of which model is correct, N protein:RNA interaction is key for viral replication. As such, phase separation provides a macroscopic reporter on a nanoscopic phenomenon, in line with previous work 62,72,120,121. In this sense, we believe the therapeutic implications of understanding and modulating phase separation here (and elsewhere in biology) are conveniently decoupled from the physiological relevance of actual, large phase separated "liquid droplets", but instead offer a window into the underlying physical interactions that lead to condensate formation.

1.4.4 The physics of single polymer condensates

Depending on the molecular details, single-polymer condensates may be kinetically stable (but thermodynamically unstable, as in our model simulations) or thermodynamically stable. Delineation between these two scenarios will depend on the nature, strength, valency and anisotropy of the interactions. It is worth noting that from the perspective of functional biology, kinetic stability may be essentially indistinguishable from thermodynamic stability, depending on the lifetime of a metastable species.

It is also important to emphasize that at higher concentrations of N protein and/or after a sufficiently long time period we expect robust phase separation with viral RNA, regardless of the presence of a symmetry-breaking site. Symmetry breaking is achieved when the apparent local concentration of N protein (from the "perspective" of gRNA) is substantially higher than the actual global concentration. As effective local and global concentrations approach one another, the entropic cost of intra-molecular interaction is outweighed by the availability of inter-molecular partners. On a practical note, if the readout in question is the presence/absence of liquid droplets, a high-affinity site may be observed as a shift in the saturation concentration which, confusingly, could either suppress or enhance phase separation. Further, if single-genome condensates are kinetically stable and driven through electrostatic interactions, we would expect a complex temperature dependence, in which larger droplets are observed at higher temperature (up to some threshold). Recent work is showing a strong temperature-dependence of phase separation is consistent with these predictions 20.

Finally, we note no reason to assume single-RNA condensates should be exclusively the purview of viruses. RNAs in eukaryotic cells may also be processed in these types of assemblies, as opposed to in large multi-RNA RNPs. The role of RNA:RNA interactions both here and in other systems is also of particular interest and not an aspect explored in our current work, but we anticipate may play a key role in the relevant biology.

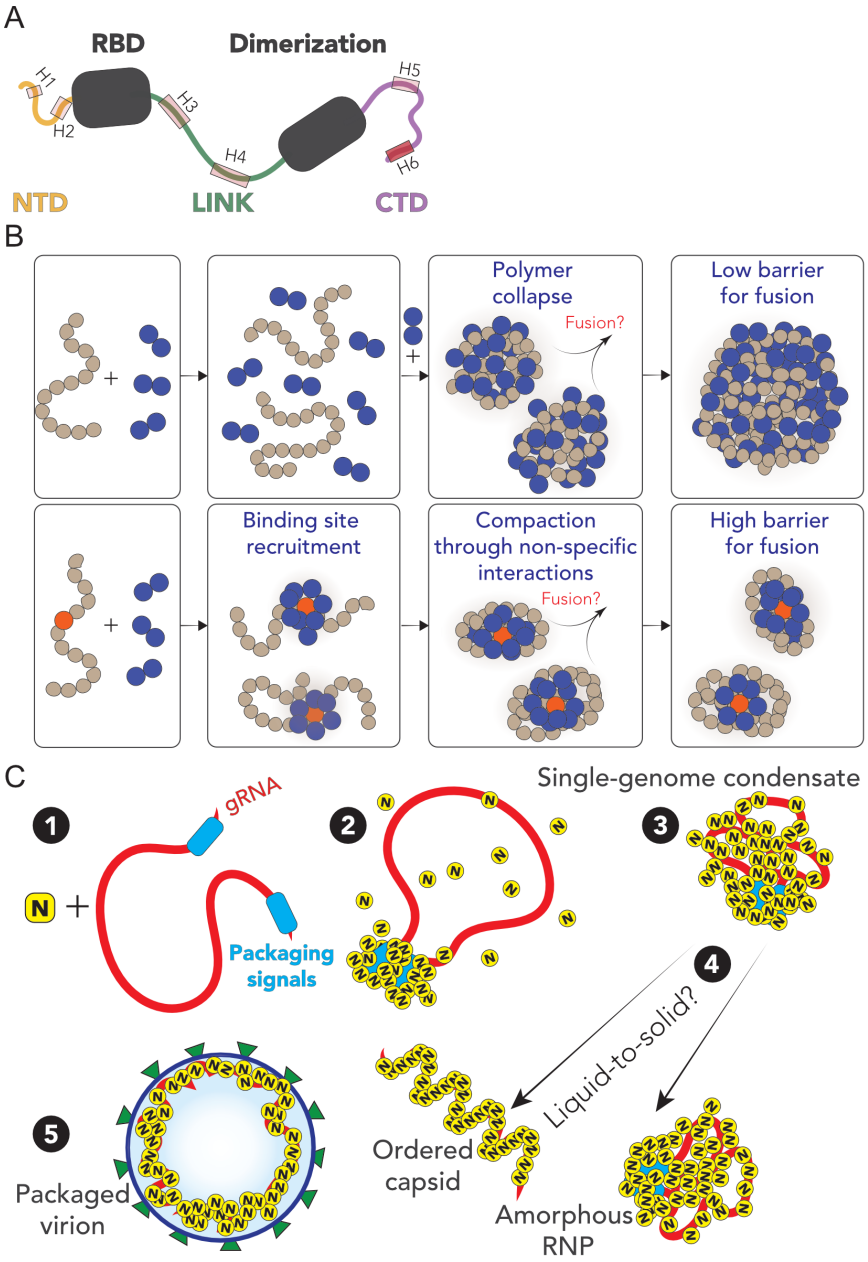

Figure 1.7: Summary and proposed model for N protein behavior. **A.** Summary of results from single-molecule spectroscopy experiments and all-atom simulations. All three predicted IDRs are disordered, highly flexible, and house a number of putative helical binding regions which overlap with subregions identified previously to drive N protein function. **B.** Overview of general symmetry breaking model. For homopolymers, local collapse leads to single-polymer condensates with a small barrier to fusion, rapidly assembling into large multi-polymer condensates. When one (or a small number of) high-affinity sites are present, local clustering of binders at a lower concentration organize the polymer such that single-polymer condensates are kinetically stable. **C.** Proposed model for SARS-CoV-2 genome packaging. **(1)** Simplified model of SARS-CoV-2 genome with a pair of packaging region at the 5' and 3' end of the genome **(2)** N protein preferentially binds to packaging signal regions in the genome, leading to a local cluster of N protein at the packaging signal RNA. **(3)** The high local concentration of N protein drives condensation of distal regions of the genome, forming a stable single-genome condensate. **(4)** Single-genome condensates may undergo subsequent maturation through a liquid-to-solid (crystallization) transition to form an ordered crystalline capsid, or solidify into an amorphous ribonuclear particle (RNP), or some combination of the two. While in some viruses an ordered capsid clearly forms, we favour a model in which the SARS-CoV-2 capsid is an amorphous RNP. Compact single-genome condensates ultimately interact with E, S and M proteins at the membrane, whose concerted action leads to envelope formation around the viral RNA and final virion packaging.

1.5 Methods

1.5.1 All atom simulations

All-atom Monte Carlo simulations were performed with the ABSINTH implicit solvent model and CAMPARI simulation engine (http://campari.sourceforge.net/) 50,122 with the solution ion parameters of Mao et al.123. Simulations were performed using movesets and Hamiltonian parameters as reported previously 62,124. All simulations were performed in sufficiently large box sizes to prevent finite size effects (where box size varies from system to system). For simulations with IDRs in isolation all degrees of freedom available in CAMPARI are sampled. For simulations with folded domains with IDRs, the backbone dihedral angles in folded domains are not sampled, such that folded domains remain structurally fixed (although sidechains are fully sampled). The IDR has backbone and sidechain degrees of freedom sampled.

All-atom molecular dynamics simulations were performed using GROMACS, using the FAST algorithm in conjunction with the Folding@home platform 51,125,126. Post-simulation analysis was performed with Enspara 127. For additional simulation details see the supplementary information.

1.5.2 Coarse-grained Polymer Simulations

Coarse-grained Monte Carlo simulations were performed using the PIMMS simulation engine 128. All simulations were performed in a 70 x 70 x 70 lattice-site box. The results averaged over the final 20% of the simulation to give average values at equivalent states. 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 109 Monte Carlo steps, with four independent replicas. Bead interaction strengths were defined as shown in Fig. 1.6A. For additional simulation details see the supplementary information.

1.5.3 Protein Expression, purification, and labeling.

SARS-CoV-2 Nucleocapsid protein (NCBI Reference Sequence: YP_009724397.2) including an N term extension containing His9-HRV 3C protease site was cloned into the BamHI EcoRI sites in the MCS of pGEX-6P-1 vector (GE Healthcare). Site-directed mutagenesis was performed on the His9-SARS-CoV-2 Nucleocapsid pGEX vector to create M1C R68C, Y172C T245C, and F363C A419C variant N protein constructs and sequences were verified using Sanger sequencing. All variants were expressed recombinantly in BL21 Codon-plus pRIL cells (Agilent) or Gold BL21(DE3) cells (Agilent) and purified using a FF HisTrap column. The GST-His9-N tag was then cleaved using HRV 3C protease and further purified to remove the cleaved tag. Finally, purified N protein variants were analyzed using SDS-PAGE and verified by electrospray ionization mass spectrometry (LC-MS). Activity of the protein was assessed by testing whether the protein is able to bind and condense nucleic acids (see phase-separation experiments) as well as to form dimers (see oligomerization in SI).

All Nucleocapsid variants were labeled with Alexa Fluor 488 maleimide and Alexa Fluor 594 maleimide (Molecular Probes) under denaturing conditions following a two-step sequential labeling procedure (see SI).

1.5.4 Single-molecule fluorescence spectroscopy.

Single-molecule fluorescence measurements were performed with a Picoquant MT200 instrument (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 of donor and acceptor, the power used for exciting the acceptor dye was adjusted to match the acceptor emission intensity to that of the donor (between 50 and 70 mW). Single-molecule FRET efficiency histograms were acquired from samples with protein concentrations between 50 pM and 100 pM and the population with stoichiometry corresponding to 1:1 donor:acceptor labeling was selected. Trigger times for excitation pulses (repetition rate 20 MHz) and photon detection events were stored with 16 ps resolution. For FRET-FCS, samples of double-labeled protein with a concentration of 100 pM were excited by either the diode laser or the supercontinuum laser at the powers indicated above.

All samples were prepared in 50 mM Tris pH 7.32, 143 mM β-mercaptoethanol (for photoprotection), 0.001% Tween 20 (for limiting surface adhesion) 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) (see appendix E).

1.6 Acknowledgements

We thank Amy Gladfelter, Christiane Iserman, Christine Roden, Ahmet Yildiz, Amanda Jack, Luke Ferro, Steve Michnick, Pascale Legault, and Jim Omichinski for sharing data and extensive discussion. We also thank Rohit Pappu for placing our groups in contact with one another.

We thank the labs of John Cooper, Carl Frieden, and Silvia Jansen for providing some of the reagents we have used in this work. We thank Ben Schuler and Daniel Nettels for developing, maintaining, and sharing with us the software package used to analyze the single-molecule data.

J.C. and J.J.A are supported by NIGMS R25 IMSD Training Grant GM103757. We are grateful to the citizen-scientists of Folding@home for donating their computing resources. G.R.B holds an NSF CAREER Award MCB-1552471, NIH R01GM12400701, a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, and a Packard Fellowship for Science and Engineering from The David and Lucile Packard Foundation. []

A.S.H. is a scientific consultant with Dewpoint Therapeutics.

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] Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, Peihua Niu, Faxian Zhan, Xuejun Ma, Dayan Wang, Wenbo Xu, Guizhen Wu, George F Gao, Wenjie Tan, and China Novel Coronavirus Investigating and Research Team. A novel coronavirus from patients with pneumonia in china, 2019. *N. Engl. J. Med.*, 382(8):727–733, February 2020.

[3] Victor M Corman, Doreen Muth, Daniela Niemeyer, and Christian Drosten. Chapter eight - hosts and sources of endemic human coronaviruses. In Margaret Kielian, Thomas C Mettenleiter, and Marilyn J Roossinck, editors, *Advances in Virus Research*, volume 100, pages 163–188. Academic Press, January 2018.

[4] Max Roser, Hannah Ritchie, Esteban Ortiz-Ospina, and Joe Hasell. Coronavirus pandemic (COVID-19). *Our World in Data*, 2020.

[5] Nicole Lurie, Melanie Saville, Richard Hatchett, and Jane Halton. Developing covid-19 vaccines at pandemic speed. *N. Engl. J. Med.*, 382(21):1969–1973, May 2020.

[6] David E Gordon, Gwendolyn M Jang, Mehdi Bouhaddou, Jiewei Xu, Kirsten Obernier, Kris M White, Matthew J O'Meara, Veronica V Rezelj, Jeffrey Z Guo, Danielle L Swaney, Tia A Tummino, Ruth Huettenhain, Robyn M Kaake, Alicia L Richards, Beril Tutuncuoglu, Helene Foussard, Jyoti Batra, Kelsey Haas, Maya Modak, Minkyu Kim, Paige Haas, Benjamin J Polacco, Hannes Braberg, Jacqueline M Fabius, Manon Eckhardt, Margaret Soucheray, Melanie J Bennett, Merve Cakir, Michael J McGregor, Qiongyu Li, Bjoern Meyer, Ferdinand Roesch, Thomas Vallet, Alice Mac Kain, Lisa Miorin, Elena Moreno, Zun Zar Chi Naing, Yuan Zhou, Shiming Peng, Ying Shi, Ziyang Zhang, Wenqi Shen, Ilsa T Kirby, James E Melnyk, John S Chorba, Kevin Lou, Shizhong A Dai, Inigo Barrio-Hernandez, Danish Memon, Claudia Hernandez-Armenta, Jiankun Lyu, Christopher J P Mathy, Tina Perica, Kala B Pilla, Sai J Ganesan, Daniel J Saltzberg, Ramachandran Rakesh, Xi Liu, Sara B Rosenthal, Lorenzo Calviello, Srivats Venkataramanan, Jose Liboy-Lugo, Yizhu Lin, Xi-Ping Huang, Yongfeng Liu, Stephanie A Wankowicz, Markus Bohn, Maliheh Safari, Fatima S Ugur, Cassandra Koh, Nastaran Sadat Savar, Quang Dinh Tran, Djoshkun Shengjuler, Sabrina J Fletcher, Michael C O'Neal, Yiming Cai, Jason C J Chang, David J Broadhurst, Saker Klippsten, Phillip P Sharp, Nicole A Wenzell, Duygu Kuzuoglu, Hao-Yuan Wang, Raphael Trenker, Janet M Young, Devin A Cavero, Joseph Hiatt, Theodore L Roth, Ujjwal Rathore, Advait Subramanian, Julia Noack, Mathieu Hubert, Robert M Stroud, Alan D Frankel, Oren S Rosenberg, Kliment A Verba, David A Agard, Melanie Ott, Michael Emerman, Natalia Jura, Mark von Zastrow, Eric Verdin, Alan Ashworth, Olivier Schwartz, Christophe d'Enfert, Shaeri Mukherjee, Matt Jacobson, Harmit S Malik, Danica G Fujimori, Trey Ideker, Charles S Craik, Stephen N Floor, James S Fraser, John D Gross, Andrej Sali, Bryan L Roth, Davide Ruggero, Jack Taunton, Tanja Kortemme, Pedro Beltrao, Marco Vignuzzi, Adolfo García-Sastre, Kevan M Shokat, Brian K Shoichet, and Nevan J Krogan. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, April 2020.

[7] James M Sanders, Marguerite L Monogue, Tomasz Z Jodlowski, and James B Cutrell. Pharmacologic treatments for coronavirus disease 2019 (COVID-19): A review. *JAMA*, April 2020.

[8] Alexandra C Walls, Young-Jun Park, M Alejandra Tortorici, Abigail Wall, Andrew T McGuire, and David Veesler. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*, 181(2):281–292.e6, April 2020.

[9] Markus Hoffmann, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S Schiergens, Georg Herrler, Nai-Huei Wu, Andreas Nitsche, Marcel A Müller, Christian Drosten, and Stefan Pöhlmann. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell*, 181(2):271–280.e8, April 2020.

[10] Jian Shang, Gang Ye, Ke Shi, Yushun Wan, Chuming Luo, Hideki Aihara, Qibin Geng, Ashley Auerbach, and Fang Li. Structural basis of receptor recognition by SARS-CoV-2. *Nature*, 581(7807):221–224, May 2020.

[11] Jun Lan, Jiwan Ge, Jinfang Yu, Sisi Shan, Huan Zhou, Shilong Fan, Qi Zhang, Xuanling Shi, Qisheng Wang, Linqi Zhang, and Xinquan Wang. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, 581(7807):215–220, May 2020.

[12] Paul S. Masters. Coronavirus genomic RNA packaging. *Virology*, 537:198–207, November 2019.