Conserved molecular recognition by an intrinsically disordered region in the absence of sequence conservation

Alex Holehouse

alex.holehouse@wustl.edu

Washington University in St. Louis https://orcid.org/0000-0002-4155-5729

Jhullian Alston

Boston Children's Hospital

Andrea Soranno

Washington University in St. Louis https://orcid.org/0000-0001-8394-7993

Article

Keywords:

Posted Date: June 3rd, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4477977/v1

Additional Declarations: There is **NO** Competing Interest.

Conserved molecular recognition by an intrinsically disordered region in the absence of sequence conservation

Jhullian J. Alston[1,2,3], Andrea Soranno[1,2], Alex S. Holehouse[1,2,✉]

[1]**Department of Biochemistry and Molecular Biophysics**, Washington University School of Medicine, St. Louis, MO 63110, USA
[2]**Center for Biomolecular Condensates,** Washington University in St. Louis, St. Louis, MO, USA
[3]Present Address, **Program In Cellular and Molecular Medicine (PCMM),** Boston Children's Hospital, Boston, MA, USA

✉ Corresponding author, e-mail: alex.holehouse@wustl.edu

ABSTRACT

Intrinsically disordered regions (IDRs) are critical for cellular function yet often appear to lack sequence conservation when assessed by multiple sequence alignments. This raises the question of if and how function can be encoded and preserved in these regions despite massive sequence variation. To address this question, we have applied coarse-grained molecular dynamics simulations to investigate non-specific RNA binding of coronavirus nucleocapsid proteins. Coronavirus nucleocapsid proteins consist of multiple interspersed disordered and folded domains that bind RNA. Here, we focus on the first two domains of coronavirus nucleocapsid proteins: the disordered N-terminal domain (NTD) and the folded RNA binding domain (RBD). While the NTD is highly variable across evolution, the RBD is structurally conserved. This combination makes the NTD-RBD a convenient model system for exploring the interplay between an IDR adjacent to a folded domain and how changes in IDR sequence can influence molecular recognition of a partner. Our results reveal a surprising degree of sequence-specificity encoded by both the composition and the precise order of the amino acids in the NTD. The presence of an NTD can – depending on the sequence – either suppress or enhance RNA binding. Despite this sensitivity, large-scale variation in NTD sequences is possible while certain sequence features are retained. Consequently, a conformationally-conserved dynamic and disordered RNA:protein complex is found across nucleocapsid protein orthologs despite large-scale changes in both NTD sequence and RBD surface chemistry. Taken together, these insights shed light on the ability of disordered regions to preserve functional characteristics despite their sequence variability.

Introduction

The classical structure-function paradigm states that sequence dictates structure, and structure dictates function [1]. This understanding has driven extensive study of protein structure and dynamics. Understanding the 3D structures that proteins adopt provides insight into their normal function. It also allows us to interpret how and why mutations that disrupt those structures and/or dynamics impair function [2–4]. However, in recent years, there has been a growing focus on understanding if and how disordered regions can contribute to cellular function [5–9]. Intrinsically disordered regions (IDRs) are poorly described by a single 3D structure; instead, they exist as a collection of structurally distinct interconverting conformations known as an ensemble [9–11]. Despite lacking a defined 3D structure, IDRs play critical roles in many aspects of cellular function [9]. Consequently, emerging work suggests that just as folded domains follow a sequence-structure-function relationship, IDRs can follow an analogous sequence-ensemble-function relationship [9,12,13]. Given the importance that structure-function analysis has played in understanding the molecular basis for cellular function, there is a promising and analogous opportunity to understand IDR function through the lens of ensembles [9,14–18].

A major goal of modern molecular biology is to accurately predict protein function directly from amino acid sequence. Rooted in the general assumption that similar protein sequences will exhibit similar molecular behavior, one strategy is to compare the sequence of a protein of interest to those of other known proteins [19–22]. In many cases, multiple sequence alignment of orthologous folded domains reveals high sequence conservation and, therefore, conserved protein function [19,23,24]. This relationship enables us to predict structures of previously unsolved protein structures and infer function by aligning the sequences of an uncharacterized protein against sequences of functionally-characterized folded domains [25–28]. In sum, applying evolutionary information, directly and indirectly, is a central pillar in our modern toolkit for protein sequence analysis.

While IDR sequences can be aligned, their conservation at the residue level is typically lower than their structured counterparts [29–31]. However, even without strict sequence conservation, the presence of disordered regions in a protein is often conserved across orthologs [14,15,31–34]. Assuming orthologous proteins provide equivalent functions, this presents a question: "Can apparently divergent IDRs confer the same molecular functions?". For some IDRs, the only feature that matters may be the existence of Short Linear Motifs (SLiMs), such that a large IDR may appear poorly conserved, yet functional conservation is maintained as long as a few short (5-15 residue) regions are present [35–37]. More recent work has shown that retaining specific physicochemical properties in disordered regions can be sufficient to preserve function [14,15,29,31–34,38–42]. Ultimately, the absence of a specific 3D structure serves to loosen the relationship between sequence and function.

Viruses provide good test systems for exploring evolutionary conservation in IDRs. Eukaryotic viruses use IDRs extensively, and their rapid evolutionary rates – driven by a combination of fast replication times, massive numbers, and strong fitness selection – mean that even between serotypes

of the same virus, substantial divergence in IDRs is often observed [43–48]. For viruses that infect the same host, it is reasonable to expect equivalent selective pressures and equivalent protein function. As such, viral IDRs offer a convenient opportunity to explore how large-scale variation in IDR sequence enables similar functional output.

In this work, we investigated the relationship between IDR sequence and RNA interaction by performing coarse-grained molecular dynamics simulations of coronavirus Nucleocapsid (N) proteins [49,50]. Coronaviruses are positive-sense single-stranded RNA viruses with relatively large (~30 Kb) genomes [50–53]. They typically consist of four major structural proteins: spike (S), envelope (E), membrane (M), and the N protein. The N protein is the most abundant viral protein and drives genomic RNA condensation and packaging during virion assembly, but has also been implicated in the evasion of the host immune system [54–57]. Given its abundance and importance, the N protein is a tractable model system for exploring variation in sequence and function.

Coronavirus N proteins consist of five domains: two folded domains (the RNA Binding Domain [RBD] and dimerization domain) and three IDRs (the N-Terminal Domain [NTD], linker, and C-Terminal Domain [CTD]) (**Fig. 1A**) [58]. Our prior work systematically characterized full-length SARS-CoV-2 (SCO2) N protein using a combination of all-atom simulations, single-molecule Förster Resonance Energy Transfer (smFRET) spectroscopy, and nanosecond Fluorescence Correlation Spectroscopy (ns-FCS) [55]. This work confirmed the disordered nature of the three IDRs and characterized their ensemble behavior in the context of the full-length protein. Importantly, this work revealed minimal interaction between the NTD-RBD and the remainder of the protein.

Given the relatively autonomous behavior of the NTD-RBD domains compared to the rest of the protein, our more recent experimental and computational work focussed on assessing the interaction of a minimal NTD-RBD construct with RNA [57]. While the RBD alone binds $(rU)_{25}$ with a binding affinity of ~0.6 μM^{-1}, the addition of the NTD enhances this affinity around 30-fold. This work also established our ability to obtain near quantitative agreement between coarse-grained molecular dynamics simulations and single-molecule RNA binding experiments in the context of non-specific binding across a range of RNA lengths and in response to small perturbations in the NTD sequence. While we cannot exclude other potential roles for the NTD, our work to date suggests that one of its functions is to enhance N-protein:RNA interactions, presumably to facilitate genome packaging. Despite our prior progress, many questions regarding the molecular details surrounding NTD-RBD:RNA interaction remain.

While the NTD is highly variable in sequence and length across N protein orthologs, a disordered NTD of some type is always present (**Fig. 1B**) [55]. In contrast, the RBD is extremely structurally conserved among orthologs, exhibiting a characteristic right-handed fist structure. This is formed by a four-strand antiparallel β-fold core and a protruding β-hairpin, which we refer to as the β3 extension [59]. Despite this structural conservation, RBD sequences vary across coronaviruses, leading

to changes in surface chemistry (**Fig. 1C**). As such, despite its pivotal role in coronavirus replication, N protein NTD-RBD sequences vary substantially across different coronaviruses.

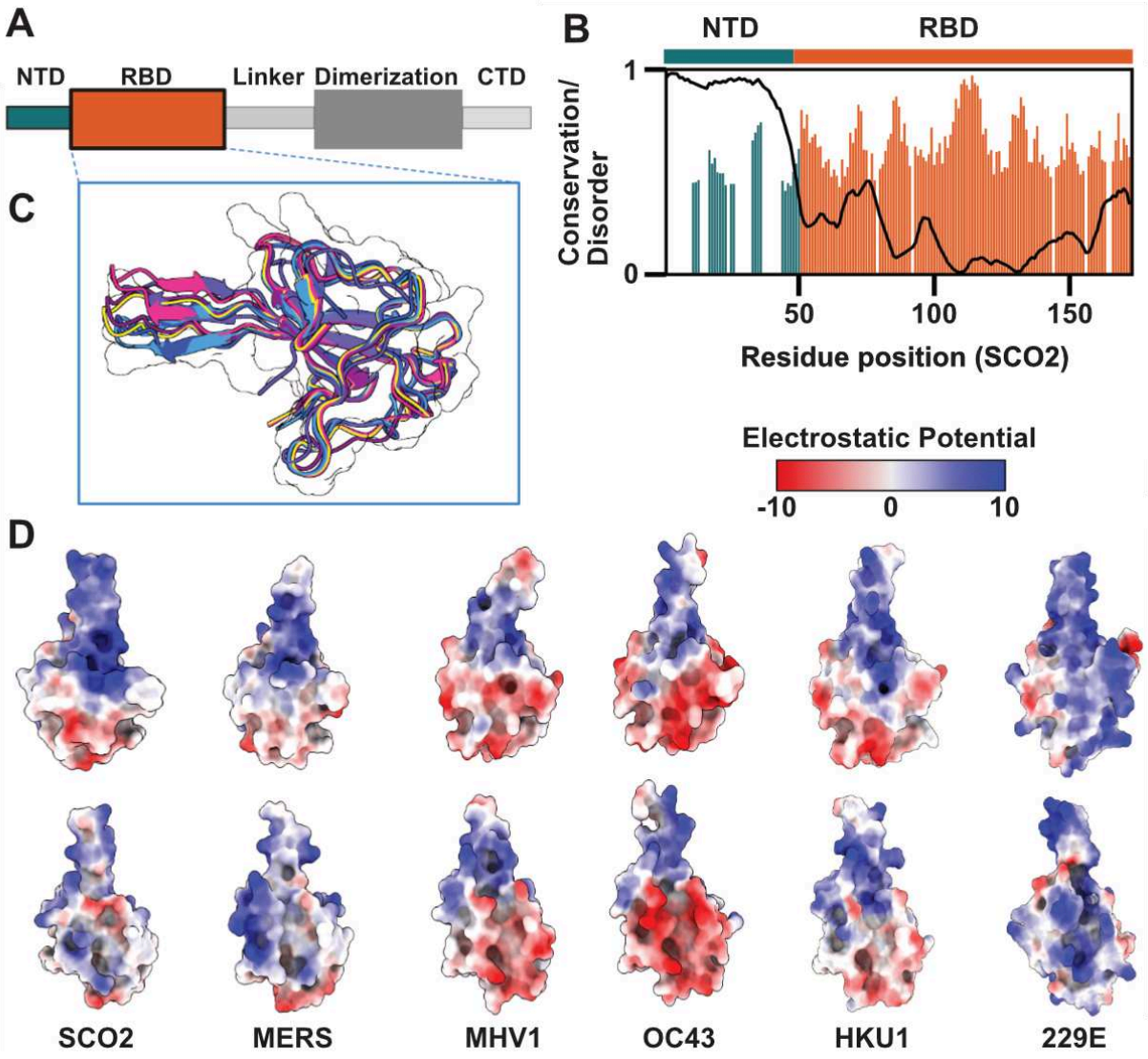

Figure 1. Coronavirus nucleocapsid proteins possess a disordered, poorly-conserved N-terminal domain (NTD) and a more well-conserved folded RNA binding domain (RBD). A. Schematic showing full-length nucleocapsid protein architectures from coronaviruses. The nucleocapsid protein contains three IDRs (NTD, Linker, CTD) and two folded domains (RBD, and Dimerization domains). **B.** Per-residue conservation calculated over 45 orthologous NTD-RBD constructs, including SCO2, MERS, OC43, HKU1, 229E, and MHV1. Conservation is calculated based on the positional Shannon entropy, with values shown only for residues where 80% or more of orthologs possess a residue. The NTD contains many gaps in a relatively poor alignment, while the RBD is almost uniformly populated with relatively highly conserved residues. Disorder propensity is calculated using metapredict. **C.** Overlay of RBD structures for SCO2, MERS, OC43, HKU1, 229E, and MHV1, revealing a high degree of structural conservation in the RBD fold.

D. (Coulomb potential scale in kcal·mol^{-1}·e^{-1}) Surface charge properties of the six RBD structures overlaid in panel C, highlighting differences in surface charge properties despite the conservation of the overall fold.

Given the structurally similar RBDs but differing NTDs, we wondered whether different coronavirus NTD-RBDs bind single-stranded RNA (ssRNA) in the same way or whether they have distinct modes of interaction. Naively, given the large variation in NTD sequence, one might expect fundamentally different modes of recognition. However, recent work has shown that the conservation of IDR ensemble properties is possible despite large changes in IDR sequence [14,60,61]. More broadly, the molecular basis for how the NTD provides a 30-fold increase in binding affinity remains unclear, especially given the NTD-RBD binds RNA almost 60-fold more tightly than the NTD in isolation [57].

To address these questions, we performed coarse-grained molecular dynamics (MD) simulations of NTD-RBD constructs with poly-$(rU)_{25}$ to assess how changes in NTD sequence influence RNA binding. Using this approach, we sought to understand how the sequence properties of an RNA binding domain and flanking disordered region enable them to cooperate to bind nucleic acids and achieve specific binding affinities. Our findings demonstrate that the ability of the SCO2 nucleocapsid protein NTD to potentiate ssRNA binding is determined by a combination of sequence composition and the relative positioning of positively charged amino acids. Our work supports a model in which the NTD and RBD are two halves of a single RNA binding domain, where the two halves make up either side of a conserved RNA binding groove. The disordered nature of the NTD substantially relaxes evolutionary constraints on the NTD, allowing many different sequences to form structurally equivalent bound-state conformations. We suggest that such bi-partite binding domains – made up of both folded and disordered regions – may be a common mode of evolutionarily labile molecular recognition. Our study highlights that disordered regions can enable the conservation of specific binding modes, even in the absence of precise sequence conservation.

Results

"Inert" Intrinsically Disordered Regions Diminish RNA Binding

Our previous work used coarse-grained MD simulations paired with smFRET-based RNA binding experiments to characterize the ability of the SCO2 NTD-RBD to bind ssRNA[57]. These simulations using the Mpipi forcefield were able to qualitatively recapitulate the conformational behavior of the NTD-RBD in the presence and absence of RNA, as well as capture with semi-quantitative accuracy the binding affinity observed for the RBD and NTD-RBD with ssRNAs of differing lengths [57,62]. Simulations and experiments showed that the addition of the disordered NTD_{SCO2} to the folded RBD resulted in a 30-fold increase in the binding affinity for $(rU)_{25}$ compared to the RBD alone. Importantly, this work identified a subregion in the NTD (residues 30-50) that is predicted to interact directly with RNA.

We first sought to establish the relationship between the NTD and RNA binding. We hypothesized that substituting the NTD_{SCO2} with an inert IDR that interacts negligibly with RNA would result in a binding affinity similar to that of the RBD alone. To our surprise, our simulations showed this was not the case.

In the Mpipi model, glycine and serine residues have negligible interactions with RNA or other amino acids. This agrees with prior experimental work that suggests GS-repeat sequences behave as relatively inert Gaussian-like chains [63–65]. We took advantage of this and replaced the 50-residue NTD_{SCO2} with a length-matched GS repeat – $(GS)_{25}$ – and performed simulations with this $(GS)_{25}$-RBD_{SCO2} chimera (**Fig. 2A**) [25,26,66]. Our simulations revealed repeated association and dissociation events between $(rU)_{25}$ and the $(GS)_{25}$-RBD constructs (**Fig. 2B)**, enabling us to calculate an apparent binding association constant, K_A , as done previously (see Methods for details) [57]. For convenience, we normalize this apparent binding affinity by the binding affinity associated with wildtype NTD-RBD binding $(rU)_{25}$, reporting this normalized binding affinity as K_A^*. $K_A^* > 1$ reflects tighter binding than wildtype, while $K_A^* < 1$ reflects weaker binding.

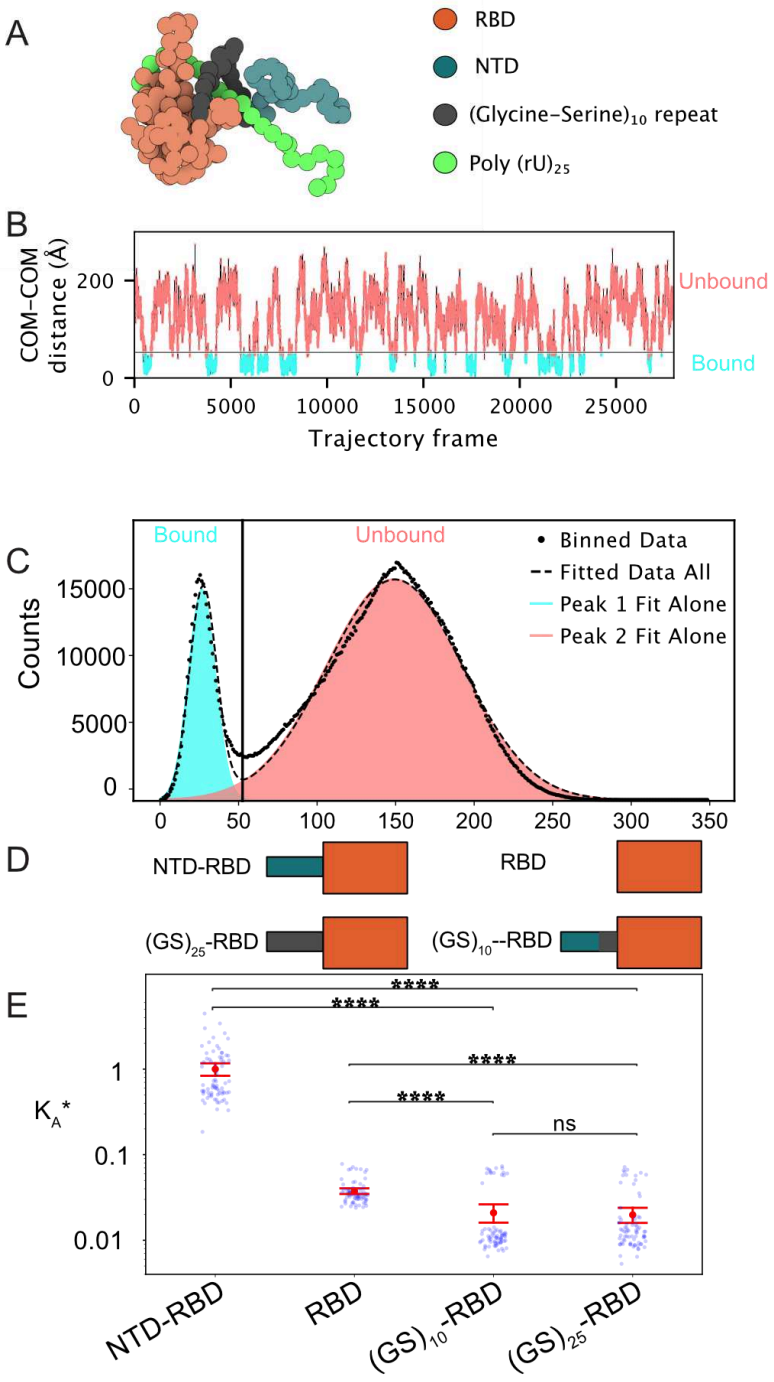

Figure 2. An inert disordered region can suppress a folded domain's RNA binding ability. A. A snapshot of the bound state from a $(GS)_{25}$-RBD + $(rU)_{25}$ simulation trajectory. Simulations utilize the Mpipi forcefield [62]. The model represents both amino acids and nucleotides as single beads with specific amino acid-amino acid and amino acid-nucleotide interactions. Folded domains are rigid, and both disordered regions and nucleic acids are dynamic. **B.** The distances between the COM of the $(GS)_{25}$-RBD and $(rU)_{25}$ are plotted over the course of the simulation. A distance threshold (black line) is determined in C (see also

Methods) and plotted to delineate the bound and unbound frames. **C.** COM-COM distances from B are plotted as a histogram and show a bimodal distribution that correlates with the bound and unbound states of the protein. The distributions are fitted with dual Gaussians. A distance threshold, which separates bound and unbound frames, is determined by minimizing the overlap of the two populations**. D.** Schematic of the four constructs shown in current "D" + $(rU)_{25}$. **E.** An apparent binding affinity (K_A) is calculated by utilizing the fraction of bound and unbound frames and Eq. 1. This is then converted to a relative apparent binding affinity (K_A*) by normalizing all values by dividing by the K_A calculated from the SCO2 NTD-RBD + $(rU)_{25}$ simulations. Blue points represent each individual simulation K_A*, while the red point is the mean of all of the replicate simulations for a given construct. The error bars are the ratio propagated standard error of the mean calculated using Eq. 2. Significance is determined by a Mann-Whitney-Wilcoxon test two-sided with Bonferroni correction. p-value annotation legend: (ns: 5.00e-02 < p <= 1.00e+00), (*: 1.00e-02 < p <= 5.00e-02), (**: 1.00e-03 < p <= 1.00e-02), (***: 1.00e-04 < p <= 1.00e-03), (****: p <= 1.00e-04).

To our surprise, the $(GS)_{25}$-RBD construct bound half as tightly as the RBD alone ($(GS)_{25}$-RBD $K_A^* = 0.020 \pm 0.003$, RBD $K_A^* = 0.037 \pm 0.004$) **(Fig. 2D)**. This result is driven by an entropic excluded volume effect, whereby the $(GS)_{25}$ impedes the ability of RNA molecules to interact with the RBD by occupying space adjacent to positively charged residues on the RBD. Importantly, this result suggests that the tighter binding affinity associated with NTD_{SCO2}-RBD_{SCO2} compared to RBD_{SCO2} alone is due to a cooperative interplay between the NTD and the RBD with RNA [57,67].

Next, we sought to understand how the NTD_{SCO2} enhanced the binding affinity. Given our prior work identified residues 30-50 in the NTD_{SCO2} as an RNA interacting region, we replaced this region with a $(GS)_{10}$ linker. While we anticipated a reduction in binding affinity compared to wildtype, we expected this construct to be stronger than that of the RBD alone. In actuality, we again observed weaker RNA binding compared to the RBD alone with a $K_A^* = 0.021 \pm 0.003$ (**Fig. 2D)**, statistically indistinguishable from the $(GS)_{25}$-RBD construct. With this in mind, our results suggest residues 30-50 are critical for robust RNA binding.

It is widely known that sequence composition and patterning govern the properties adopted by intrinsically disordered regions[9,68]. However, for IDRs adjacent to RNA binding domains and their binding interfaces, our results illustrate that sequence properties can either enhance or diminish RNA binding affinity, depending on the specific IDR sequence. Taken together, our results suggest that the sequence of the N-terminal IDR adjacent to coronavirus RBDs needs to be relatively specific and is most likely conserved, albeit not in the traditional sense of direct sequence alignment; otherwise, without specific residues, the IDR could interfere with RNA binding to the extent of diminishing binding affinity.

Coronavirus Nucleocapsid Protein NTDs have Conserved Sequence Composition

While NTD's in coronavirus nucleocapsid proteins appear to always be disordered, their absolute sequence conservation is poor (**Fig. 1B, Supplementary Fig. 3**). If NTDs exist to enhance RNA binding affinity, and disordered NTDs can diminish RNA binding if the 'wrong' sequence is present, then how do coronavirus NTDs ensure tight RNA binding is conserved despite large scale variation in sequence?

The decrease in binding affinity caused by $(GS)_{10}$ and $(GS)_{25}$ mutant NTDs indicates that any enhancement in RNA binding provided by the NTD_{SCO2} is sequence-dependent. This conclusion is consistent with our prior work, in which we found even small changes in NTD sequence had measurable effects on RNA binding affinity as measured both by single-molecule experiments and by simulations [57].

Operating under the assumption that the NTD_{SCO2} has a role in enhancing RNA binding affinity of the RBD (**Supplementary Fig. 4)**, we reasoned there may be some selective pressure towards NTD sequences that result in a consistent macroscopic RNA binding affinity for the NTD-RBD. Additionally, while RBD structures are highly conserved across coronaviruses, their charged surface residues vary (**Fig. 1D**)[69]. As such, we also wondered if there may be a co-evolutionary coupling between the NTD sequence and the RBD surface. Thus, despite the diverging surface charge of the RBDs, conserved interactions between the NTDs and their respective RBDs could lead to a consistent macroscopic RNA binding affinity.

To investigate this hypothesis, in addition to the NTD-RBD taken from SCO2, we examined NTD-RBD constructs from five other coronaviruses: human coronaviruses OC43, HKU1, and 229E, the Middle East Respiratory Syndrome Coronavirus (MERS), and the Mouse Hepatitis Virus (MHV1). We reasoned that focusing on coronaviruses that predominantly infect the same host would ensure host selective pressures are consistent, thereby minimizing this as a confounding factor to explain differences in RNA binding affinities.

We first examined NTD physicochemical properties that are routinely used to describe IDRs (**Supplementary Table S3-S6**). Despite the large variation in NTD length, all NTDs possess a net positive charge, with the least positive NTD possessing a net charge per residue of +0.056. Expanding this analysis to 45 different coronavirus NTDs, we found no examples in which the net charge was lower than +0.056 (**Supplementary Fig. 5**). This is consistent with RNA binding proteins typically binding RNA through positive electrostatic surfaces that interact with negatively charged RNA [70].

Next, we examined solvent-accessible residues on the RBD surface. We generated five RBD structures for each of the coronaviruses using AlphaFold2, and then took the average of our calculated properties across the five structures [66]. The net charge per residue (NCPR) of the RBD surface residues stratified into three categories: relatively positively charged (229E = 0.126, SCO2 =0.066, MERS = 0.052,) neutral (HKU1 = 0.0, MHV1 = -0.011), and negatively charged (OC43 = -0.053). However, in all case we found that the β3 extension surface was positively charged, albeit to different extents (**Fig. 1D)**.

In summary, while the surface charge of the RBD domains appears more variable, our analysis suggests two key features conserved across coronavirus N proteins: (1) a net positive NTD and (2) a

positive charge on the structurally conserved β3 extension. Compositional conservation in the NTD (i.e., the retention of specific physicochemical features, such as net charge) could enable conserved interactions despite the lack of absolute sequence conservation. We next sought to determine if composition was sufficient or if other sequence properties were important to determine RNA binding.

Sequence Composition Alone Does Not Determine NTD Contribution to Binding Affinity

One possible interpretation of our analysis is that the only factor that matters for NTD function is a net positive charge. To test if composition is the only thing that matters, we designed sequence variants that moved residues 30-50 (which contain several positively charged residues) to different locations across the NTD. [57,67]. We placed residues 30-50 at positions 1, 6, 11, 16, 21, 26 (referred to as mutants T1, T6, T11, T16, T21, T26) and 31 (wildtype) of the NTD_{SCO2} **(Fig. 3A)**. We then performed simulations with $(rU)_{25}$ and calculated apparent binding affinities of each variant. These sequences maintain the same sequence composition but rearrange the amino acids, which allows us to determine whether there are positional contributions to RNA binding or if sequence composition alone is sufficient to achieve RNA binding.

To our surprise, the relative position of residues 30-50 has a significant impact on the apparent binding affinity **(Fig. 3B)**. Two mutants showed wild-type-like binding affinities, yet the others bound RNA more weakly. This suggests that the relative location of positive charge with respect to the RBD tunes RNA binding affinity.

Why do the T6 and T11 variants show wild-type-like binding? Our results thus far suggest that placing a cluster of positively charged residues either directly adjacent (as is the case in the wild-type sequence) or ~30-40 residues (as is the case in the T11 and T6 variants) from the RBD are optimal for tight binding. Indeed, in the wild-type sequence, a pair of arginine residues is found around residues 10-14. However, why such a pattern matters for RNA binding was initially unclear.

To further test how the relative position of positively charged residues impacts RNA binding, we generated 172 scrambled NTD_{SCO2} sequences in which the sequence composition is identical, yet the order of the amino acids has been changed. These scrambles were generated in four ways: The first by randomly shuffling the NTD_{SCO2}; the second by shuffling the NTD_{SCO2} while also making each amino acid change as chemically different from the wild-type sequence as possible in terms of charge and aromaticity; third, by shuffling the NTD_{SCO2} while forcing positively charged residues from falling in the 30-50 residue region; and fourth, by shuffling the NTD_{SCO2} while restricting the majority of charged residues to the 30-50 region or a region spanning residues 4-17. Using these scrambled sequences, we performed coarse-grained MD simulations and calculated K_A^* with $(rU)_{25}$.

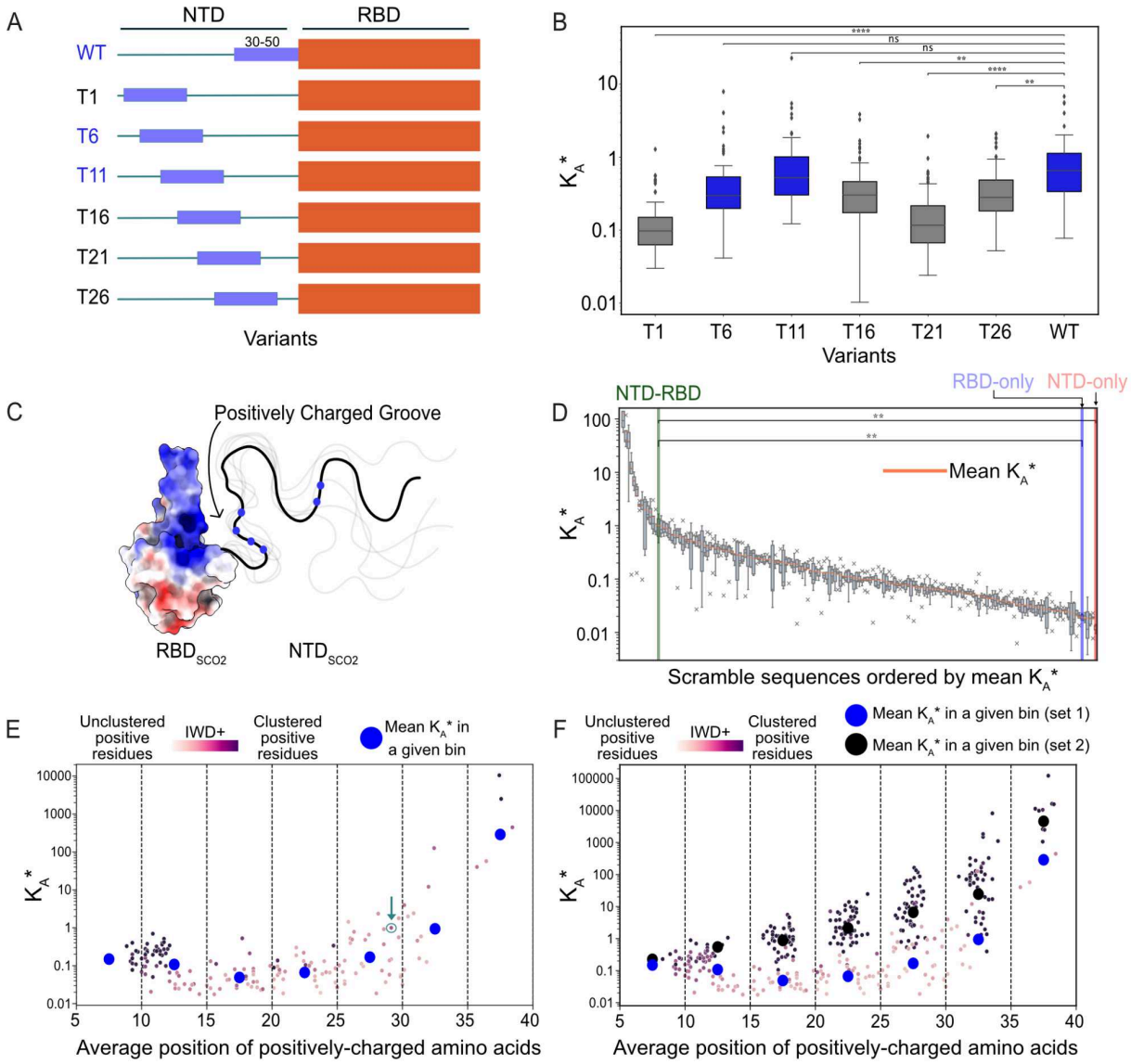

Figure 3. Clusters of positively charged residues determine the affinity enhancement provided by the NTD on RNA binding A. Schematic showing the wild type and mutants that systematically reposition residues 30-50 from the wild-type sequence. **B.** Binding affinity for mutants schematized in panel A. Mutant T6 and T11 show wildtype-like binding affinity, whereas all other variants show binding affinity less than the wild type. **C.** Graphical schematic highlighting the positively-charged and dynamic 'groove" that can form upon RNA binding between the positively-charged β3 extension on the RBD and the cluster of positively charged residues on the NTD. In the RBD positively charged surfaces are colored blue, negatively charged surfaces are colored red, and neutral surfaces are colored white. A representative NTD is drawn with the blue circles representing the relative positions of the positively charged residues. **D.** Binding affinities for 172 scramble variants. Orange bars within each plotted box represent the value of the mean KA* of each scrambled sequences replicate simulations. Each variant reports on the binding affinity for an NTD-RBD construct, where for each variant the NTD sequence was randomly scrambled. Despite having an identical amino acid composition, sequence order enables a four-order-of-magnitude change in binding affinity, highlighting the importance of sequence in dictating binding affinity. **E.** Scramble sequences plotted with

binding affinity vs. the average position of positively charged residues distributed across the sequence. For positional bins, average binding affinity is shown as a blue circle. Individual points are colored based on the IWD+ score, which reports on the clustering of positively charged residues (darker colors = more highly clustered). The wildtype NTD-RBD sequence is shown with both a green arrow and green circle around its data point. Bins that spanned residues 15-20 and 20-25 were each significantly different from the wild-type bin ($p = 0.00013$ and 0.016, respectively) **F.** Same data as shown in E, with an additional set of scrambles designed to cluster positively charged residues. The average binding affinity of this second set is shown as black circles.

Binding affinities were calculated for each of the scrambled sequences and compared with one another (**Fig 3D, Supp. Table 7**). The dynamic range of K_A^* observed here spans five orders of magnitude, demonstrating the dramatic impact relative amino acid position can have on binding affinity. However, for the majority of the scrambled sequences, the binding affinity is fairly similar, and, importantly, this "average" binding affinity is almost an order of magnitude weaker than the wild-type NTD-RBD.

Taken together with our simulations that shifted the 30-50 amino acid region around the NTD_{SCO2}, these results suggest composition is not the sole determinant of how the NTD_{SCO2} influences RNA binding. While 172 scrambled sequences are only a fraction of the total number of possible sequence shuffles that could be generated for the $NTD_{SCO2,}$ the observation that the wild-type NTD_{SCO2} sequence is among those with the highest apparent affinity suggests that the ordering of the residues in the NTD_{SCO2} is specific.

Disordered Region Residue Sequence Positioning Dictates RNA Binding Capacity

While most scrambled sequences had similar binding affinities that were much weaker than the wild-type sequence, we identified a subset of sequences that had binding affinities equal to or greater than that of the wild-type sequence. Based on our simulations testing positioning of the 30-50 amino acid region, we reasoned that the relative position of positively charged residues might underlie the increased binding affinity of these select sequences, highlighting regions of the NTD that are more binding-competent.

To assess how the position of positively charged residues correlates with binding affinity, we plotted binding affinity versus the average position of all positively charged residues in each scrambled sequence that we initially tested (**Fig 3E,** blue circles are the binned means of each sequence). The average position is calculated as the mean of the location of the arginine and lysine residues in the linear sequence of the NTD_{SCO2}. This analysis revealed a correlation between strong binders and the average position of positively charged residues. When the average position of positive residues is around residues 30-40, binding affinity is drastically increased in comparison to the other regions. This same region is relatively positively charged in the wild-type NTD_{SCO2}.

The importance of the position of positively charged residues offers a 'structural' explanation for the enhanced binding affinity afforded by the wild-type NTD. Charged residues within this region

enable the formation of a positively charged 'groove.' One-half of this groove is made of the positively charged surface of the RBD β3 extension, while the other half comes from the disordered NTD. This charged groove enables simultaneous multivalent interactions between the NTD_{SCO2} and the RBD_{SCO2} with RNA and, thus, tight RNA binding (**Fig 3C**). To further explore if a disordered charged groove underlies high-affinity NTD binding, we examined the relationship between charge clustering and RNA binding.

The average position of positive charges along the NTD does not capture the clustering of positively charged residues. To address this, we used the inverse weighted distance (IWD+) metric to calculate the clustering of positively charged residues[71,72]. Our initial set of scrambles showed relatively similar charge clustering, although in almost all cases, sequences with a greater degree of positively charged residue clustering bound more tightly than those where residues were less clustered (**Fig. 3E**).

To more systematically investigate the impact of positive charge clustering, we designed a second library of 214 additional scrambles. In this library, sequences were designed such that all positively charged residues were locally clustered at a specific location (**Fig. 3F**). Sequences with clusters of positive charge generally exhibited increased binding affinities. Moreover, sequences where positively charged residues were clustered towards the C-terminus of the NTD showed – in general – tighter binding than those where positively charged clusters were N-terminal. These results confirm that the presence of a positively charged cluster on the NTD adjacent to the RBD provides the highest affinity binding interface.

Our results thus far are consistent with a model in which the local density of positively charged residues forms one-half of a positively charged binding grove (**Fig. 3C**). While we conventionally think of binding clefts as forming between two folded domains, here we propose a binding interface that straddles the folded RBD surface and the disordered NTD, akin to a flexible thumb and a structured hand. This disordered binding groove model makes several predictions.

First, this model predicts that the NTD should remain disordered upon binding RNA. This prediction is supported by recent nanosecond FCS experiments in which no loss of conformational heterogeneity was seen upon RNA binding [57]. Second, very different sequences should be compatible with RNA binding, a prediction supported by results from our scrambles, which show that if appropriate sequence constraints are met, there are many NTDs with wild-type-like binding (**Fig. 3F**). Third, this model predicts that across different coronaviruses we should expect the mode of RNA binding by the NTD-RBD to be conserved. In other words, even as the surface and sequence of the RBD and NTD vary, we should expect the conformational features of the bound-state ensemble to be preserved. To test this prediction, we next performed simulations of five additional orthologous NTD-RBD constructs with $(rU)_{25}$

NTD-RBD:RNA Behavior in the Bound State is Conserved Across Orthologs

Our scrambles confirm that the NTD sequence has a substantial impact on NTD-RBD RNA binding affinity. We therefore asked if natural NTD sequences encode a similar positively charged "groove" binding mode despite seemingly large-scale variation in NTD sequence and RBD surface chemistry. In this model, specific subregions of the NTD come into closer proximity to the RBD driven by favorable NTD-RNA interactions on one side and RBD-RNA interactions on the other (**Fig. 3C**). To test this, we performed simulations of each of the six ortholog NTD-RBD constructs with $(rU)_{25}$ and assessed the bound-state conformational ensemble of the NTD.

Bound-state ensembles were quantified using scaling maps. Scaling maps capture the average inter-residue distance between all pairs of residues for RNA-bound conformers, and offer a way to quantify the conformational ensemble of an IDR [55,73–75]. Here, scaling map values are calculated as the inter-residue distance measured in the RNA-bound state normalized by the inter-residue distance of sequence-matched NTD-RBD simulations performed in the absence of RNA (**Fig. 4A**). Shades of purple reflect distances that are closer together in the bound state, while shades of green denote regions that are further apart in the bound state. In this way, the scaling map provides a quantitative description of the RNA-bound ensemble of the NTD.

For SCO2, this analysis identified two regions in the NTD that are closer to the RBD in the bound state ensemble centered around residues 10-20 and residues 30-50, similar to our simulations that shifted the 30-50 amino acid region around the $NTD_{SCO2}$and as reported previously [57]. This analysis can be done selectively for one of the residues in the NTD to visualize where it increases RBD interactions when bound to RNA by mapping its distances across the entire NTD-RBD construct with RBD residues colored with respect to NTD distance (**Fig. 4B**). Doing so shows that in the bound state, the NTD moves closer to the positively charged RBD β3 extension, highlighting the formation of a positively charged groove between the positive β3 extension and the positive region spanning amino acids 30-50, as well as contributions from the region spanning amino acids 10-20 in the NTD_{SCO2}. This positive groove effectively envelopes RNA, facilitating a specific bound-state ensemble.

We repeated this analysis for the remaining five orthologs, as well as the $(GS)_{10}$-RBD and $(GS)_{25}$-RBD constructs, to determine if these NTDs also move closer to the RBD. This analysis reveals that the same two specific subregions within the NTD come closer to the RBD across coronavirus orthologs. Despite large-scale variation in both folded-domain surface charge and NTD sequence, the bound-state ensemble (and hence RNA binding mode) appears to be largely conserved across the six coronavirus NTD-RBD constructs examined. However, for the GS mutant NTDs this conformational conservation is lost (**Supplementary Fig. 6)**, highlighting the sequence dependence of these interactions.

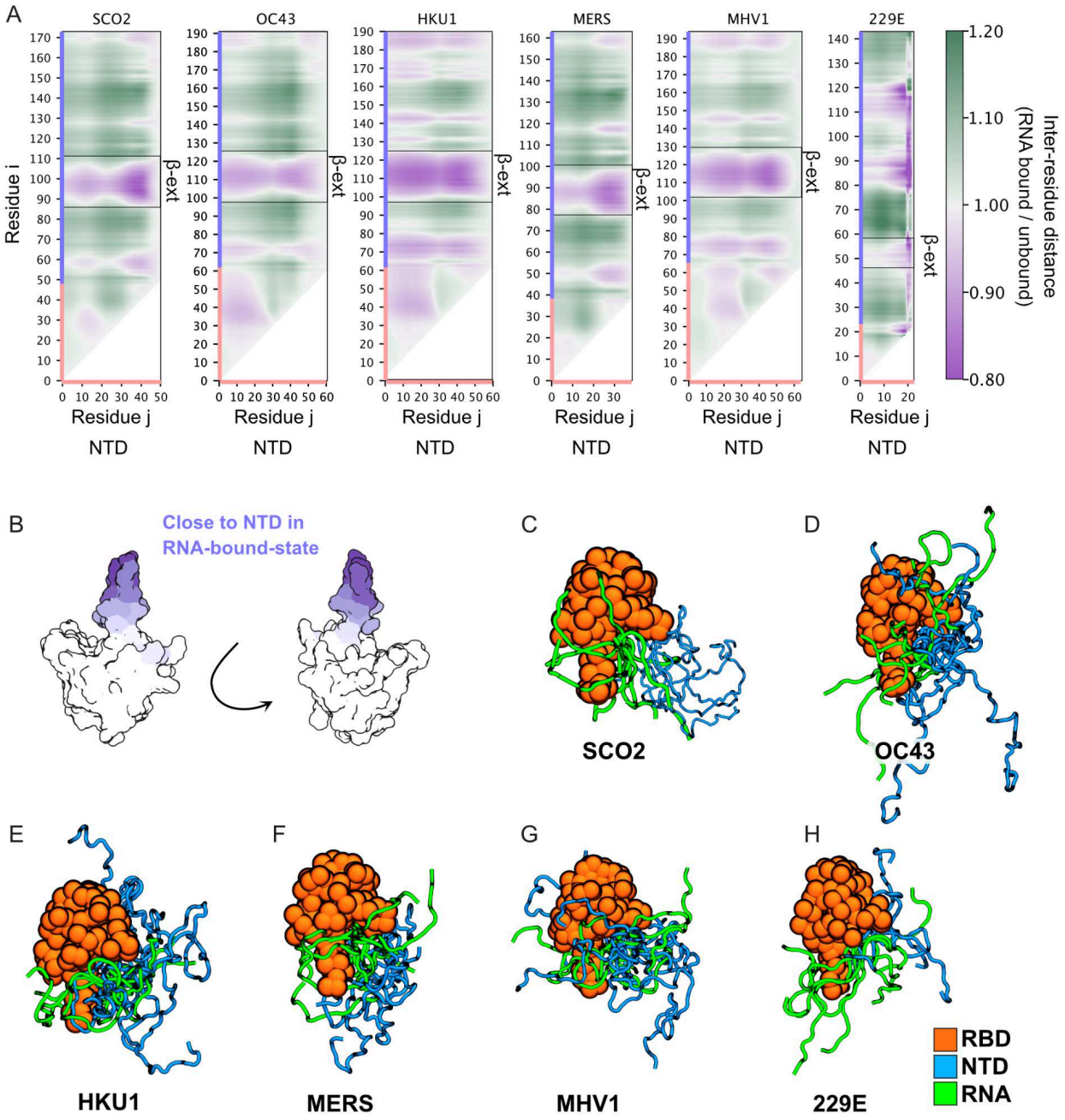

Figure 4. Orthologous nucleocapsid proteins show similar bound-state ensembles despite variations in RBD surface charge residues and NTD sequence. A. Scaling maps quantify the average inter-residue distance between NTD residues (X-axis, colored pink) and NTD or RBD residues (Y-axis, colored pink and light blue respectively) in the bound state. Heatmap values are calculated by calculating the average inter-residue distance in the RNA-bound state and dividing that distance by the average inter-residue distance in the RNA-unbound state. Purple colors report on inter-residue distances that are closer together in the bound state while green colors report on inter-residue distances that are further apart in the unbound state. In all six orthologs, the NTD is closer to the β3 extension in the bound state, reporting on the formation of a positively charged groove in the bound state. **B.** Regions closer to the NTD in the RNA-bound state are

highlighted on the SCO2 RBD structure in shades of purple with more intense purple signifying closer on average. **(C-H)** Representative snapshots from RNA-bound-state ensembles. In all cases, RBD configuration is aligned in the same way, enabling conservation of binding mode to be directly visualized across six distinct orthologs.

To better visualize this result, we generated structural models of the bound-state ensemble with six conformers each (**Fig. 4C**). While these make up a tiny fraction of the bound-state frames, it should be clear that in all cases, RNA binding occurs through a conserved bound-state ensemble, whereby RNA lies along a disordered groove generated between the β3-extension on one side and the NTD on the other. Taken together, our work suggests that co-evolution of the NTD-RBD occurs at the level of preserving a bound-state ensemble, as opposed to sequence or conformational properties in the unbound state.

Discussion and Conclusion

Intrinsically disordered proteins and protein regions are prevalent across eukaryotic, prokaryotic, and viral proteomes[9]. They play a wide variety of essential roles yet – perhaps paradoxically – often appear to be relatively poorly conserved sequences by alignment[29–31]. In this study, we sought to understand how a specific molecular function (RNA binding) could be conserved despite large-scale changes in amino acid sequence. We utilized two domains of various coronavirus nucleocapsid protein orthologs as a convenient model that contains both a disordered region (NTD) and a folded domain (RBD) that binds RNA. Despite poor sequence conservation assessed by alignment across NTDs, we found that the orthologs were compositionally conserved. That is, the orthologs have similar charge properties in both the NTD and portions of the RBD. Specifically, NTDs harbor a net positive charge, while RBDs retain specific positively charged regions on a specific region of their surface. Despite this conservation, the length and sequence of N protein NTDs vary dramatically, and while RBDs maintain the same 3D structure, orthologous RBDs showed a diverse set of surface properties, including negatively charged patches and changes in positive regions.

To assess how the sequence composition of the disordered NTDs influences interactions with the RBDs and impacts RNA binding, we performed coarse-grained molecular dynamics simulations of coronavirus nucleocapsid proteins with single-stranded RNA. These simulations enabled us to interrogate the role of sequence composition and residue positioning in coronavirus NTDs ability to increase binding affinity of the NTD-RBD. We first showed that RNA binding could be enhanced (NTD-RBD) or suppressed ($(GS)_{25}$-RBD) compared to the RBD in isolation depending on the IDR sequence. Further, by testing hundreds of different sequences with the same overall composition, we determined that composition alone does not dictate RNA binding affinity. Instead, our simulations highlight the importance of clusters of positively charged residues, and that the relative position of positive clusters along the NTD also matter. Specifically, our simulations reveal the mode of binding occurs via a disordered, positively charged grove that forms between the NTD and the positive surface of the RBD (specifically the β3 extension). In this way a 'structural' basis for RNA binding emerges, despite the fact the bound state is highly heterogeneous (a result we previously confirmed

via ns-FCS experiments) [57]. Moreover, this specific binding mode is conserved across five additional orthologous NTD-RBD constructs, despite largescale variation in sequence.

This charged groove and the dynamic nature of the RNA-protein interaction is potentially similar to the high affinity yet highly dynamic interactions that have been observed for polyelectrolyte complexes formed by charged polymers or the H1-Prothymosin alpha interaction, and for other IDR:RNA interactions[76–78]. Here the NTD is able to remain highly dynamic and disordered yet still maintain relatively tight binding affinity. Our rationally-designed sequences suggest tighter binding is certainly possible, but whether tighter binding would be functionally advantageous for viral replication is unclear.

Our work here implicates synergistic cooperation between a folded domain and a disordered region to enable high-affinity binding. The exceptional structural conservation of RBDs across coronaviruses may reflect their crucial role in virion structural stability, perhaps enabled via a network of stacked aromatic residues in the RBD core. While RNA binding domains often posses binding clefts, our work here suggests that such clefts need not be fully structuctured, and that a partially disordered binding groove can also enable evolutionarily-labile RNA binding.

Recent work identified arginine-rich motifs within disordered regions adjacent to DNA binding domains across transcription factors, implicating these regions as mediating RNA binding in concert with the DNA binding domain [79]. Given the conserved binding mode uncovered in our work here, we speculate that while defining RNA/DNA binding domains in terms of their folded domains is convenient, the full 'domain' could in some cases be extended to include flanking IDRs that potentiate and/or regulate binding. In particular, we have explicit examples in which adjacent IDRs enhance[80], supress[81,82], or have no effect[83] on DNA binding affinity. These observations dovetail with our own work that suggests the amino acid chemistry of IDRs adjacent to nucleic acid binding domains impacts the macroscopic binding affinity. Furthermore, highly charged flanking IDRs can lubricate interactions between folded domains and nucleic acids by competing with the folded domain for nucleic acid interaction, or nucleic acids for folded domain interactions, a model proposed by the Levy lab almost fifteen years ago [84–88].

The cooperative effect of NTD and RBD binding with loose structural coupling opens the door to compensatory changes in either domain. While identifying such couplings is inherently challenging, we note that the ortholog with the least prominent NTD:RBD interaction profile (229E; **Fig. 4A**) also has the most positively charged RBD, pointing to a potential mechanism to compensate for a 'weaker' (less positively charged) NTD.

Our simulations also hint at the presence of a second RNA binding region in the NTD, centered around residues 12 in SCO2. This is highlighted by the appearance of two local subregions that are close to the RBD in the bound state – one around residues 30-50 but a second around residues 5-15 (**Fig. 4A**). This region is clearly insufficient to enable RNA binding in isolation, because replacing

residues 30-50 in NTD_{SCO2} with a $(GS)_{10}$ yielded a binding affinity indistiguishable from one where the entire NTD was replaced by $(GS)_{25}$ (**Fig. 2E**). Nevertheless, designs that repositioned residues $NTD_{SCO2}^{30\text{-}50}$ to the location of this potential second hotspot (T6 and T11) recovered wildtype-like affinity, suggesting this relative position from the RBD may also be well-poised for RNA binding (**Fig. 3**). We speculate this may reflect an optimal distance between loop-closure entropy, electrostatic repulsion between binding regions, and effective concentration; i.e. that at this number of residues away from the RBD, the NTD can 'fold' back on itself and interact with RNA that is bound to the RBD surface. While they lack absolute sequence conservation, the conserved nature of these hotspots across five of the six orthologs implicates these regions as potentially playing an auxiliary regulatory role in RNA binding.

Recent work has suggested that small-molecules that target specific IDR ensembles may provide a route for sequence-specific pharmacological interventions [89,90]. Given the essential role N protein:RNA interaction has in coronavirus lifecycles, our work here hints at principles to enable the rational design of bivalent molecules that might enable specific NTD-RBD inhibition by outcompeting with RNA to bind in a conformationally-conserved manner. If conventional antiviral structure-guided drug design focusses on conserved structural features, targeting conserved conformational features offers an alternative but conceptually analogous route to pharmacological intervention against regions traditionally considered 'undruggable'[91,92].

While this study focused on the NTD-RBD from coronavirus nucleocapsid proteins, we expect the insights gleaned here will be widely applicable to a range of disordered nucleic acid-binding proteins. While absolute sequence conservation may not be present, there is still the possibility of conserved behavior encoded into diverging sequences. Rather than solely focusing on sequence alignments to provide information on conservation and important residues, quantitatively describing the ensemble that a disordered region takes on and assessing how it behaves with and without its ligand(s) may provide better insight into the residues that are important and sequence features that need to be maintained to ensure proper biological function.

Methods

Molecular Dynamics Simulations

All simulations were performed using the LAMMPS simulation engine[93]. We performed molecular dynamics simulations in the NVT ensemble using the default parameters of the physics-driven coarse-grained force-field Mpipi developed by Joseph et al. [62] The model represents both amino acid residues and nucleotides as chemically unique singular beads and was parameterized to recapitulate the behavior of disordered proteins in isolation as well as their ability to undergo phase separation with and without RNA. Inter-bead interactions consist of a combination of short-range contributions from a Wang-Frenkel potential, which captures a combination of Van der Waals, cation-pi, and pi-pi interactions, and a long-range Coulombic potential for amino acids with net charge and RNA nucleotides. The ability of the Mpipi force field to recapitulate disordered protein

dimensions has been previously shown [62,94]. Simulations were performed under an effective ionic strength of 50 mM NaCl, conditions we previously found to engender good agreement between simulation and experiment when comparing with experimentally-measured RNA binding affinities using single-molecule experiments [57].

We also assessed the ability of the Mpipi forcefield to recapitulate single-stranded RNA (ssRNA) dimensions by comparing simulations of $(rU)_{40}$ with scattering data from small-angle X-ray (SAXS) experiments for the same construct[95]. This comparison revealed excellent agreement across the full scattering curve and in terms of the scattering-derived radius of gyration; using the Molecular Form Factor approach of Riback et al., R_g^{sim} = 30.9 ± 0.1 Å while R_g^{exp} = 30.2 ± 0.3 Å) (**Supplementary Fig. 1**) [96].

Simulations were performed in a 30 nm^3 simulation box with periodic boundary conditions. Protein and RNA are allowed to diffuse freely throughout the box. Disordered regions and ssRNA behave as dynamic flexible polymers, sampling an ensemble of conformations [62]. However, as done previously, folded domains were made rigid, and residues buried within folded domains experienced downscaled non-bonded interactions [57,62]. Unless otherwise specified, all simulations were run for 300 million steps per replicate. The exceptions are the 'scrambled' simulations, which were run for 100 million steps per replicate. Protein and RNA configurations were saved every 10,000 steps, and the first 0.2% was removed for equilibration. Visualization of protein-RNA complexes was done with Protein Imager and VMD [97,98]. Simulations were analyzed using SOURSOP and MDTraj [74,99]. Small angle X-ray scattering was analyzed using the Molecular Form Factor (MFF) (http://sosnick.uchicago.edu/SAXSonIDPs), while synthetic scattering data for simulations were generated using FOXS default settings [96,100].

We performed simulations of the NTD-RBD, NTD, and RBD of six coronavirus orthologs. Specifically, we examined five coronaviruses that infect humans: SARS-CoV-2 (SCO2), Middle Eastern Respiratory Syndrome virus (MERS), Human Coronaviruses OC43, Human Coronavirus HKU1, and Human Coronavirus 229E, as well as Murine Hepatitis Virus (MHV1). Sequence alignments were compared to determine a region of the RBD that was well conserved between all orthologs to delineate the start and end positions of the NTD and RBD's of each ortholog [58,101–103]. For simulations with ssRNA, all simulations were done using $(rU)_{25.}$

To capture conformational heterogeneity in an artificially rigid structure, we utilized Colabfold to generate five different starting structures for each coronavirus orthologous RBD [25,66]. For simulations of wild-type versions of each ortholog's NTD-RBD all five starting structures are used, to enable conclusions to be less biased by a specific starting conformation. As expected, certain RBD conformers bind RNA better than others, but in all cases where different NTDs are compared, the same sets of RBD conformers are used, such that any RBD conformation-specific biases are consistent across the set (**Supplementary Fig. 2**). For the large scrambled library, 1

conformation for the SCO2 RBD is used. All simulations were run with multiple replicates per starting RBD structure, with a minimum of five replicates per RBD conformation.

Limitations of Coarse-Grained Simulations

Our use of the Mpipi model should not be taken to imply that RNA or proteins are faithfully represented at one bead per residue/nucleotide resolution. Both proteins and RNA are complex biomolecules with many degrees of freedom, a chemically heterogeneous structure, and can engage in a variety of sequence and structure-specific interactions that are not captured by a simplified coarse-grain model. Our goal in using a simplified coarse-grain model is to enable high-throughput biophysical assessment in a system that, based on prior work, we have good reason to believe is semi-quantitative in terms of relative accuracy [57,62]. While we refer to the molecules in our simulations as protein and RNA, in reality, they are better thought of as RNA- and protein-flavored polymers. The simplicity of this model enables us to address questions that would be intractable using either higher-resolution simulation approaches or experiments. Despite this, we are under no illusion regarding the simplifying assumptions made for a coarse-grain model.

Calculating Apparent Association Constants From Simulations

We determined apparent association constants (K_A) by using an updated version of our previous center of mass (COM) calculations that were able to qualitatively recapitulate SCO2 NTD-RBD single-stranded RNA binding [57]. To do this, post-equilibration simulation frames were divided into bound and unbound states. This delineation was achieved by first taking the intermolecular center-of-mass distances between the protein and the RNA and plotting the distribution of distances. The histogram of intermolecular distances follows a bimodal distribution that reports on the bound and unbound states, and can be fit with two Gaussians (**Fig. 2C**). We then determined the intersection that minimizes the overlap of the two distributions to define a cutoff distance. The cutoff distance varies based on the size of the protein and RNA. Finally, as done previously, we classify frames as bound or unbound by assessing the linear intermolecular COM distance trajectory and delineating frames as bound when five or more frames are below the cutoff distance. This minimum number of consecutive frames allows us to distinguish between transient random interactions between protein and RNA vs. encounters with a reasonable "lifetime", implying direct and continuous interaction. The distributions and distance cutoffs are calculated for every set of NTD_a-RBD_b + $(rU)_n$ simulations, where *a* and *b* represent specific NTD or RBD sequences and *n* the length of the single-stranded (rU), allowing us to determine protein-RNA specific distance thresholds for each simulation.

The resultant fraction of bound frames is used to calculate an apparent K_D with the equation:

$$K_D = \frac{(1-f_{bound})^2}{N_A V f_{bound}} \quad \text{(Eq. 1)}$$

Here f_{bound} refers to the fraction of frames where the protein and RNA are determined to be in the bound state from our COM-COM distribution analysis. N_A refers to Avagodro's constant, and V is the simulation box volume in liters, which returns a K_D in mol/L. K_A is then calculated using the expression $K_A = 1/K_D$. While we determine if two molecules are bound or unbound in a different manner, this approach is analogous to that of Tesei *et al.* [104].

It is important to note that the K_As determined from these simulations are not meant to represent absolute values that would be comparable to those determined from experiment. Our prior work has shown that K_As calculated from Mpipi simulations for this system lack absolute agreement with experimentally measured values. Despite this, when experiment and simulation-derived K_A values are normalized by an internally consistent reference (i.e., the K_A obtained from NTD-RBD binding $(rU)_{25}$), we see good agreement between simulations and experiment, both as a function of RNA length and as a function of the presence/absence of the NTD [57]. To that end, binding affinity here is reported as K_A*, a normalized binding affinity we define as the ratio of the apparent K_A of a given protein + RNA simulation divided by the corresponding K_A for the analogous SCO2 NTD-RBD binding to $(rU)_{25}$. This enables the SCO2 NTD-RBD + $(rU)_{25}$ simulations binding affinity to be a reference point with which to understand the strength of interactions of other orthologs. All K_A* values are thus greater than 1 (stronger binding than the SCO2 NTD-RBD + $(rU)_{25}$) or less than 1 (weaker binding than the SCO2 NTD-RBD + $(rU)_{25}$).

Error is propagated for our ratio (K_A*) using:

$$\frac{R_{error}}{R} = \sqrt{\left(\frac{A_{error}}{A}\right)^2 + \left(\frac{B_{error}}{B}\right)^2} \qquad \text{(Eq. 2)}$$

R and R_{error} here represent the ratio and the error of the ratio. *A* and *B* represent the numerator and denominator of our ratios, respectively, and A_{error} and B_{error} are their associated errors (standard error of the mean).

Disorder prediction

Disorder prediction is done using metapredict (V2-FF)[60,105].

Calculating Charge Clustering in Disordered Regions

Charge clustering is quantified by the inverse weighted distance (IWD), a metric that has been applied to study amino acid clustering in several systems [71,72,106,107]. Unlike the patterning parameters κ ("kappa") or sequence charge decoration (SCD), which quantify the patterning of oppositely charged residues with respect to one another, here our interest is on the clustering of positive residues only [68,108]. The IWD score allows us to quantify the clustering of a specific subset of residues. When residues are clustered together, the IWD score is high, whereas when residues are

evenly distributed, the IWD score is low. IWD scores were calculated using sparrow (https://github.com/idptools/sparrow).

Statistical Analysis

Every simulation has a minimum of five independent replicates, and calculated values are presented as 95% confidence intervals (box plots, with medians marked), mean and standard error of the mean, or geometric mean and geometric standard deviation (clarified in text below figures). Fitting of Gaussian distributions was done in Python using scipy.optimize.curve_fit [109].

Data Availability and Software

Analysis code and data (calculated distance distributions and contact map information) are deposited at https://github.com/holehouse-lab/supportingdata/tree/master/2023/alston_2023. For further information on using the code, please refer to the deposited Jupyter notebooks.

Acknowledgments

We thank members of the Soranno lab and Holehouse lab for many useful discussions over the years. We particularly thank Dan Griffith, who has provided useful insights for data visualization with Python. We thank Dr. Emery Usher for help in performing Guinier analysis of $(rU)_{40}$ scattering data. We would like to thank Dr. Jerelle Joseph for the parameterization of the Mpipi force field, which enabled us to do this work. We thank Dr. Lois Pollack and Dr. Steve Meisburger for sharing scattering data for $(rU)_{40}$. Funding for this work was provided by the National Institute of Allergy and Infectious Diseases with R01AI163142 to A.S.H. and A.S., by the Human Frontiers in Science Program (HFSP RGP0015/2022) to A.S.H, and by the National Cancer Institute with an F99CA264413 to J.J.A.

Competing interests

No authors have any competing interests.

References

1. Branden, C. & Tooze, J. Introduction to protein structure, Garland Pub. *Inc., New York* (1991).
2. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. & Sunyaev, S. R. A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).
3. Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., Kosmicki, J. A., Rehnström, K., Mallick, S., Kirby, A., Wall, D. P., MacArthur, D. G., Gabriel, S. B., DePristo, M., Purcell, S. M., Palotie, A., Boerwinkle, E., Buxbaum, J. D., Cook, E. H., Jr, Gibbs, R. A., Schellenberg, G. D., Sutcliffe, J. S., Devlin, B., Roeder, K., Neale, B. M. & Daly, M. J. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46,** 944–950 (2014).
4. Matthews, B. W. Structural and genetic analysis of protein stability. *Annu. Rev. Biochem.* **62,** 139–160 (1993).
5. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293,** 321–331 (1999).
6. van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D. T., Kim, P. M., Kriwacki, R. W., Oldfield, C. J., Pappu, R. V., Tompa, P., Uversky, V. N., Wright, P. E. & Babu, M. M. Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114,** 6589–6631 (2014).
7. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. R., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C. H., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, M., Garner, E. C. & Obradovic, Z. Intrinsically disordered protein. *J. Mol. Graph. Model.* **19,** 26–59 (2001).
8. Uversky, V. N. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.*

11, 739–756 (2002).

9. Holehouse, A. S. & Kragelund, B. B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nat. Rev. Mol. Cell Biol.* **25,** 187–211 (2024).
10. Forman-Kay, J. D. & Mittag, T. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure* **21,** 1492–1499 (2013).
11. Mao, A. H., Lyle, N. & Pappu, R. V. Describing sequence–ensemble relationships for intrinsically disordered proteins. *Biochem. J* **449,** 307–318 (2013).
12. Das, R. K., Ruff, K. M. & Pappu, R. V. Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **32,** 102–112 (2015).
13. Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O. G. & Pappu, R. V. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys. J.* **112,** 16–21 (2017).
14. González-Foutel, N. S., Glavina, J., Borcherds, W. M., Safranchik, M., Barrera-Vilarmau, S., Sagar, A., Estaña, A., Barozet, A., Garrone, N. A., Fernandez-Ballester, G., Blanes-Mira, C., Sánchez, I. E., de Prat-Gay, G., Cortés, J., Bernadó, P., Pappu, R. V., Holehouse, A. S., Daughdrill, G. W. & Chemes, L. B. Conformational buffering underlies functional selection in intrinsically disordered protein regions. *Nat. Struct. Mol. Biol.* **29,** 781–790 (2022).
15. Martin, E. W., Holehouse, A. S., Peran, I., Farag, M., Incicco, J. J., Bremer, A., Grace, C. R., Soranno, A., Pappu, R. V. & Mittag, T. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* **367,** 694–699 (2020).
16. Borcherds, W., Theillet, F.-X., Katzer, A., Finzel, A., Mishall, K. M., Powell, A. T., Wu, H., Manieri, W., Dieterich, C., Selenko, P., Loewer, A. & Daughdrill, G. W. Disorder and residual helicity alter p53-Mdm2 binding affinity and signaling in cells. *Nat. Chem. Biol.* **10,** 1000–1002 (2014).

17. Wiggers, F., Wohl, S., Dubovetskyi, A., Rosenblum, G., Zheng, W. & Hofmann, H. Diffusion of a disordered protein on its folded ligand. *Proc. Natl. Acad. Sci. U. S. A.* **118,** (2021).

18. Stuchell-Brereton, M. D., Zimmerman, M. I., Miller, J. J., Mallimadugula, U. L., Incicco, J. J., Roy, D., Smith, L. G., Cubuk, J., Baban, B., DeKoster, G. T., Frieden, C., Bowman, G. R. & Soranno, A. Apolipoprotein E4 has extensive conformational heterogeneity in lipid-free and lipid-bound forms. *Proc. Natl. Acad. Sci. U. S. A.* **120,** e2215371120 (2023).

19. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39,** W29–37 (2011).

20. Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J. & Punta, M. Pfam: the protein families database. *Nucleic Acids Res.* **42,** D222–30 (2014).

21. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536–540 (1995).

22. Orengo, C. A., Pearl, F. M., Bray, J. E., Todd, A. E., Martin, A. C., Lo Conte, L. & Thornton, J. M. The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res.* **27,** 275–279 (1999).

23. McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32,** W20–5 (2004).

24. Olmea, O., Rost, B. & Valencia, A. Effective use of sequence correlation and conservation in fold recognition. *J. Mol. Biol.* **293,** 1221–1239 (1999).

25. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D.,

Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589 (2021).

26. Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., Potapenko, A., Ballard, A. J., Romera-Paredes, B., Nikolov, S., Jain, R., Clancy, E., Reiman, D., Petersen, S., Senior, A. W., Kavukcuoglu, K., Birney, E., Kohli, P., Jumper, J. & Hassabis, D. Highly accurate protein structure prediction for the human proteome. *Nature* **596,** 590–596 (2021).
27. Humphreys, I. R., Pei, J., Baek, M., Krishnakumar, A., Anishchenko, I., Ovchinnikov, S., Zhang, J., Ness, T. J., Banjade, S., Bagde, S. R., Stancheva, V. G., Li, X.-H., Liu, K., Zheng, Z., Barrero, D. J., Roy, U., Kuper, J., Fernández, I. S., Szakal, B., Branzei, D., Rizo, J., Kisker, C., Greene, E. C., Biggins, S., Keeney, S., Miller, E. A., Fromme, J. C., Hendrickson, T. L., Cong, Q. & Baker, D. Computed structures of core eukaryotic protein complexes. *Science* **374,** eabm4805 (2021).
28. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A. I., Cowie, A., Figurnov, M., Fuchs, F. B., Gladman, H., Jain, R., Khan, Y. A., Low, C. M. R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E. D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D. & Jumper, J. M. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* (2024). doi:10.1038/s41586-024-07487-w
29. Langstein-Skora, I., Schmid, A., Emenecker, R. J., Richardson, M. O. G., Götz, M. J., Payer, S.

K., Korber, P. & Holehouse, A. S. Sequence- and chemical specificity define the functional landscape of intrinsically disordered regions. *bioRxiv* 2022.02.10.480018 (2022). doi:10.1101/2022.02.10.480018

30. Brown, C. J., Johnson, A. K., Dunker, A. K. & Daughdrill, G. W. Evolution and disorder. *Curr. Opin. Struct. Biol.* **21,** 441–446 (2011).
31. Zarin, T., Strome, B., Nguyen Ba, A. N., Alberti, S., Forman-Kay, J. D. & Moses, A. M. Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *Elife* **8,** (2019).
32. Zarin, T., Tsai, C. N., Nguyen Ba, A. N. & Moses, A. M. Selection maintains signaling function of a highly diverged intrinsically disordered region. *Proc. Natl. Acad. Sci. U. S. A.* **114,** E1450–E1459 (2017).
33. Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V. & Mittag, T. Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. *Nat. Chem.* **14,** 196–207 (2022).
34. Cohan, M. C., Shinn, M. K., Lalmansingh, J. M. & Pappu, R. V. Uncovering Non-random Binary Patterns Within Sequences of Intrinsically Disordered Proteins. *J. Mol. Biol.* **434,** 167373 (2022).
35. Kumar, M., Michael, S., Alvarado-Valverde, J., Mészáros, B., Sámano-Sánchez, H., Zeke, A., Dobson, L., Lazar, T., Örd, M., Nagpal, A., Farahi, N., Käser, M., Kraleti, R., Davey, N. E., Pancsa, R., Chemes, L. B. & Gibson, T. J. The Eukaryotic Linear Motif resource: 2022 release. *Nucleic Acids Res.* **50,** D497–D508 (2022).
36. Sangster, A. G., Zarin, T. & Moses, A. M. Evolution of short linear motifs and disordered proteins Topic: yeast as model system to study evolution. *Curr. Opin. Genet. Dev.* **76,** 101964 (2022).

37. Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H. & Gibson, T. J. Attributes of short linear motifs. *Mol. Biosyst.* **8,** 268–281 (2012).

38. Lu, A. X., Lu, A. X., Pritišanac, I., Zarin, T., Forman-Kay, J. D. & Moses, A. M. Discovering molecular features of intrinsically disordered regions by using evolution for contrastive learning. *PLoS Comput. Biol.* **18,** e1010238 (2022).

39. Zarin, T., Strome, B., Peng, G., Pritišanac, I., Forman-Kay, J. D. & Moses, A. M. Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. *Elife* **10,** e60220 (2021).

40. Gutierrez, J. I., Brittingham, G. P., Karadeniz, Y., Tran, K. D., Dutta, A., Holehouse, A. S., Peterson, C. L. & Holt, L. J. SWI/SNF senses carbon starvation with a pH-sensitive low-complexity sequence. *Elife* **11,** e70344 (2022).

41. Shinn, M. K., Cohan, M. C., Bullock, J. L., Ruff, K. M., Levin, P. A. & Pappu, R. V. Connecting sequence features within the disordered C-terminal linker of *Bacillus subtilis* FtsZ to functions and bacterial cell division. *Proc. Natl. Acad. Sci. U. S. A.* **119,** e2211178119 (2022).

42. Beh, L. Y., Colwell, L. J. & Francis, N. J. A core subunit of Polycomb repressive complex 1 is broadly conserved in function but not primary sequence. *Proc. Natl. Acad. Sci. U. S. A.* **109,** E1063–71 (2012).

43. Xue, B., Dunker, A. K. & Uversky, V. N. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.* **30,** 137–149 (2012).

44. Mihalič, F., Simonetti, L., Giudice, G., Sander, M. R., Lindqvist, R., Peters, M. B. A., Benz, C., Kassa, E., Badgujar, D., Inturi, R., Ali, M., Krystkowiak, I., Sayadi, A., Andersson, E., Aronsson, H., Söderberg, O., Dobritzsch, D., Petsalaki, E., Överby, A. K., Jemth, P., Davey, N.

E. & Ivarsson, Y. Large-scale phage-based screening reveals extensive pan-viral mimicry of host short linear motifs. *Nat. Commun.* **14,** 2409 (2023).

45. Dyson, H. J. Vital for Viruses: Intrinsically Disordered Proteins. *J. Mol. Biol.* 167860 (2022).
46. Domingo, E., Escarmís, C., Sevilla, N., Moya, A., Elena, S. F., Quer, J., Novella, I. S. & Holland, J. J. Basic concepts in RNA virus evolution. *FASEB J.* **10,** 859–864 (1996).
47. Hendrix, R. W., Lawrence, J. G., Hatfull, G. F. & Casjens, S. The origins and ongoing evolution of viruses. *Trends Microbiol.* **8,** 504–508 (2000).
48. Koonin, E. V., Dolja, V. V. & Krupovic, M. Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* **479-480,** 2–25 (2015).
49. McBride, R., van Zyl, M. & Fielding, B. C. The coronavirus nucleocapsid is a multifunctional protein. *Viruses* **6,** 2991–3018 (2014).
50. Masters, P. S. The molecular biology of coronaviruses. *Adv. Virus Res.* **66,** 193–292 (2006).
51. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17,** 181–192 (2019).
52. Fehr, A. R. & Perlman, S. Coronaviruses: an overview of their replication and pathogenesis. *Methods Mol. Biol.* **1282,** 1–23 (2015).
53. Masters, P. S. Coronavirus genomic RNA packaging. *Virology* **537,** 198–207 (2019).
54. Carlson, C. R., Asfaha, J. B., Ghent, C. M., Howard, C. J., Hartooni, N., Safari, M., Frankel, A. D. & Morgan, D. O. Phosphoregulation of Phase Separation by the SARS-CoV-2 N Protein Suggests a Biophysical Basis for its Dual Functions. *Mol. Cell* **80,** 1092–1103.e4 (2020).
55. 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. *Nat. Commun.* **12,** 1936 (2021).

56. Alston, J. J. & Soranno, A. Condensation goes viral: a polymer physics perspective. *J. Mol. Biol.* 167988 (2023).

57. Cubuk, J., Alston, J. J., Incicco, J. J., Holehouse, A. S., Hall, K. B., Stuchell-Brereton, M. D. & Soranno, A. The disordered N-terminal tail of SARS-CoV-2 Nucleocapsid protein forms a dynamic complex with RNA. *Nucleic Acids Res.* (2023). doi:10.1093/nar/gkad1215

58. Chang, C.-K., Hou, M.-H., Chang, C.-F., Hsiao, C.-D. & Huang, T.-H. The SARS coronavirus nucleocapsid protein--forms and functions. *Antiviral Res.* **103,** 39–50 (2014).

59. Wu, W., Cheng, Y., Zhou, H., Sun, C. & Zhang, S. The SARS-CoV-2 nucleocapsid protein: its role in the viral life cycle, structure and functions, and use as a potential target in the development of vaccines and diagnostics. *Virol. J.* **20,** 6 (2023).

60. Lotthammer, J. M., Ginell, G. M., Griffith, D., Emenecker, R. J. & Holehouse, A. S. [No title]. (2024). doi:10.1038/s41592-023-02159-5

61. Tesei, G., Trolle, A. I., Jonsson, N., Betz, J., Knudsen, F. E., Pesce, F., Johansson, K. E. & Lindorff-Larsen, K. Conformational ensembles of the human intrinsically disordered proteome. *Nature* (2024). doi:10.1038/s41586-023-07004-5

62. Joseph, J. A., Reinhardt, A., Aguirre, A., Chew, P. Y., Russell, K. O., Espinosa, J. R., Garaizar, A. & Collepardo-Guevara, R. Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Nat Comput Sci* **1,** 732–743 (2021).

63. Sørensen, C. S. & Kjaergaard, M. Effective concentrations enforced by intrinsically disordered linkers are governed by polymer physics. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 23124–23131 (2019).

64. Moses, D., Yu, F., Ginell, G. M., Shamoon, N. M., Koenig, P. S., Holehouse, A. S. & Sukenik, S. Revealing the Hidden Sensitivity of Intrinsically Disordered Proteins to their Chemical Environment. *J. Phys. Chem. Lett.* **11,** 10131–10136 (2020).

65. Moses, D., Guadalupe, K., Yu, F., Flores, E., Perez, A. R., McAnelly, R., Shamoon, N. M., Kaur, G., Cuevas-Zepeda, E., Merg, A. D., Martin, E. W., Holehouse, A. S. & Sukenik, S. Structural biases in disordered proteins are prevalent in the cell. *Nat. Struct. Mol. Biol.* 1–10 (2024).

66. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S. & Steinegger, M. ColabFold: making protein folding accessible to all. *Nat. Methods* **19,** 679–682 (2022).

67. Pontoriero, L., Schiavina, M., Korn, S. M., Schlundt, A., Pierattelli, R. & Felli, I. C. NMR Reveals Specific Tracts within the Intrinsically Disordered Regions of the SARS-CoV-2 Nucleocapsid Protein Involved in RNA Encountering. *Biomolecules* **12,** (2022).

68. Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 13392–13397 (2013).

69. Taneja, I. & Holehouse, A. S. Folded domain charge properties influence the conformational behavior of disordered tails. *Curr Res Struct Biol* **3,** 216–228 (2021).

70. Shazman, S. & Mandel-Gutfreund, Y. Classifying RNA-binding proteins based on electrostatic properties. *PLoS Comput. Biol.* **4,** e1000146 (2008).

71. Holehouse, A. S., Ginell, G. M., Griffith, D. & Böke, E. Clustering of aromatic residues in prion-like domains can tune the formation, state, and organization of biomolecular condensates. *Biochemistry* **60,** 3566–3581 (2021).

72. Jankowski, M. S., Griffith, D., Shastry, D. G., Pelham, J. F., Ginell, G. M., Thomas, J., Karande, P., Holehouse, A. S. & Hurley, J. M. The formation of a fuzzy complex in the negative arm regulates the robustness of the circadian clock. *bioRxiv* 2022.01.04.474980 (2022). doi:10.1101/2022.01.04.474980

73. Martin, E. W., Holehouse, A. S., Grace, C. R., Hughes, A., Pappu, R. V. & Mittag, T. Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to

and upon Multisite Phosphorylation. *J. Am. Chem. Soc.* **138,** 15323–15335 (2016).

74. Lalmansingh, J. M., Keeley, A. T., Ruff, K. M., Pappu, R. V. & Holehouse, A. S. SOURSOP: A Python package for the analysis of simulations of intrinsically disordered proteins. *bioRxiv* (2023). doi:10.1101/2023.02.16.528879

75. Gomes, G.-N. W., Krzeminski, M., Namini, A., Martin, E. W., Mittag, T., Head-Gordon, T., Forman-Kay, J. D. & Gradinaru, C. C. Conformational ensembles of an intrinsically disordered protein consistent with NMR, SAXS, and single-molecule FRET. *J. Am. Chem. Soc.* **142,** 15697–15710 (2020).

76. Borgia, A., Borgia, M. B., Bugge, K., Kissling, V. M., Heidarsson, P. O., Fernandes, C. B., Sottini, A., Soranno, A., Buholzer, K. J., Nettels, D., Kragelund, B. B., Best, R. B. & Schuler, B. Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **555,** 61–66 (2018).

77. Srivastava, S. & Tirrell, M. V. in *Advances in Chemical Physics* 499–544 (John Wiley & Sons, Inc., 2016).

78. Holmstrom, E. D., Liu, Z., Nettels, D., Best, R. B. & Schuler, B. Disordered RNA chaperones can enhance nucleic acid folding via local charge screening. *Nat. Commun.* **10,** 2453 (2019).

79. Oksuz, O., Henninger, J. E., Warneford-Thomson, R., Zheng, M. M., Erb, H., Vancura, A., Overholt, K. J., Hawken, S. W., Banani, S. F., Lauman, R., Reich, L. N., Robertson, A. L., Hannett, N. M., Lee, T. I., Zon, L. I., Bonasio, R. & Young, R. A. Transcription factors interact with RNA to regulate genes. *Mol. Cell* **83,** 2449–2463.e13 (2023).

80. Baughman, H. E. R., Narang, D., Chen, W., Villagrán Suárez, A. C., Lee, J., Bachochin, M. J., Gunther, T. R., Wolynes, P. G. & Komives, E. A. An intrinsically disordered transcription activation domain increases the DNA binding affinity and reduces the specificity of NFκB p50/RelA. *J. Biol. Chem.* **298,** 102349 (2022).

81. He, F., Borcherds, W., Song, T., Wei, X., Das, M., Chen, L., Daughdrill, G. W. & Chen, J.

Interaction between p53 N terminus and core domain regulates specific and nonspecific DNA binding. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 8859–8868 (2019).

82. Krois, A. S., Dyson, H. J. & Wright, P. E. Long-range regulation of p53 DNA binding by its intrinsically disordered N-terminal transactivation domain. *Proc. Natl. Acad. Sci. U. S. A.* **115,** E11302–E11310 (2018).
83. Bjarnason, S., McIvor, J. A. P., Prestel, A., Demény, K. S., Bullerjahn, J. T., Kragelund, B. B., Mercadante, D. & Heidarsson, P. O. DNA binding redistributes activation domain ensemble and accessibility in pioneer factor Sox2. *Nat. Commun.* **15,** 1445 (2024).
84. Vuzman, D. & Levy, Y. DNA search efficiency is modulated by charge composition and distribution in the intrinsically disordered tail. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 21004–21009 (2010).
85. Vuzman, D. & Levy, Y. Intrinsically disordered regions as affinity tuners in protein–DNA interactions. *Mol. Biosyst.* (2012). at <https://pubs.rsc.org/en/content/articlehtml/2012/mb/c1mb05273j>
86. Wang, X., Greenblatt, H. M., Bigman, L. S., Yu, B., Pletka, C. C., Levy, Y. & Iwahara, J. Dynamic Autoinhibition of the HMGB1 Protein via Electrostatic Fuzzy Interactions of Intrinsically Disordered Regions. *J. Mol. Biol.* **433,** 167122 (2021).
87. Bigman, L. S. & Levy, Y. Protein Diffusion Along Protein and DNA Lattices: Role of Electrostatics and Disordered Regions. *Annu. Rev. Biophys.* **52,** 463–486 (2023).
88. Wang, X., Bigman, L. S., Greenblatt, H. M., Yu, B., Levy, Y. & Iwahara, J. Negatively charged, intrinsically disordered regions can accelerate target search by DNA-binding proteins. *Nucleic Acids Res.* **51,** 4701–4712 (2023).
89. Zhu, J., Salvatella, X. & Robustelli, P. Small molecules targeting the disordered transactivation domain of the androgen receptor induce the formation of collapsed helical states. *Nat. Commun.*

13, 6390 (2022).

90. Basu, S., Martínez-Cristóbal, P., Frigolé-Vivas, M., Pesarrodona, M., Lewis, M., Szulc, E., Bañuelos, C. A., Sánchez-Zarzalejo, C., Bielskutė, S., Zhu, J., Pombo-García, K., Garcia-Cabau, C., Zodi, L., Dockx, H., Smak, J., Kaur, H., Batlle, C., Mateos, B., Biesaga, M., Escobedo, A., Bardia, L., Verdaguer, X., Ruffoni, A., Mawji, N. R., Wang, J., Obst, J. K., Tam, T., Brun-Heath, I., Ventura, S., Meierhofer, D., García, J., Robustelli, P., Stracker, T. H., Sadar, M. D., Riera, A., Hnisz, D. & Salvatella, X. Rational optimization of a transcription factor activation domain inhibitor. *Nat. Struct. Mol. Biol.* **30,** 1958–1969 (2023).

91. Plavec, Z., Pöhner, I., Poso, A. & Butcher, S. J. Virus structure and structure-based antivirals. *Curr. Opin. Virol.* **51,** 16–24 (2021).

92. Wang, H., Xiong, R. & Lai, L. Rational drug design targeting intrinsically disordered proteins. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **13,** (2023).

93. Thompson, A. P., Aktulga, H. M., Berger, R., Bolintineanu, D. S., Brown, W. M., Crozier, P. S., in 't Veld, P. J., Kohlmeyer, A., Moore, S. G., Nguyen, T. D., Shan, R., Stevens, M. J., Tranchida, J., Trott, C. & Plimpton, S. J. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **271,** 108171 (2022).

94. Lotthammer, J. M., Ginell, G. M., Griffith, D., Emenecker, R. J. & Holehouse, A. S. Direct Prediction of Intrinsically Disordered Protein Conformational Properties From Sequence. *bioRxiv* 2023.05.08.539824 (2023). doi:10.1101/2023.05.08.539824

95. Chen, H., Meisburger, S. P., Pabit, S. A., Sutton, J. L., Webb, W. W. & Pollack, L. Ionic strength-dependent persistence lengths of single-stranded RNA and DNA. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 799–804 (2012).

96. Riback, J. A., Bowman, M. A., Zmyslowski, A. M., Knoverek, C. R., Jumper, J. M., Hinshaw, J. R., Kaye, E. B., Freed, K. F., Clark, P. L. & Sosnick, T. R. Innovative scattering analysis shows

that hydrophobic disordered proteins are expanded in water. *Science* **358,** 238–241 (2017).

97. Tomasello, G., Armenia, I. & Molla, G. The Protein Imager: a full-featured online molecular viewer interface with server-side HQ-rendering capabilities. *Bioinformatics* **36,** 2909–2911 (2020).

98. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph. Model.* **14,** 33–8, 27–8 (1996).

99. McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., Schwantes, C. R., Wang, L.-P., Lane, T. J. & Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109,** 1528–1532 (2015).

100. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* **105,** 962–974 (2013).

101. Gussow, A. B., Auslander, N., Faure, G., Wolf, Y. I., Zhang, F. & Koonin, E. V. Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. *Proc. Natl. Acad. Sci. U. S. A.* **117,** 15193–15199 (2020).

102. Peng, Y., Du, N., Lei, Y., Dorje, S., Qi, J., Luo, T., Gao, G. F. & Song, H. Structures of the SARS-CoV-2 nucleocapsid and their perspectives for drug design. *EMBO J.* **39,** e105938 (2020).

103. Terry, J. S., Anderson, L. B., Scherman, M. S., McAlister, C. E., Perera, R., Schountz, T. & Geiss, B. J. Development of a SARS-CoV-2 nucleocapsid specific monoclonal antibody. *Virology* **558,** 28–37 (2021).

104. Tesei, G., Schulze, T. K., Crehuet, R. & Lindorff-Larsen, K. Accurate model of liquid-liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proc. Natl. Acad. Sci. U. S. A.* **118,** (2021).

105. Emenecker, R. J., Griffith, D. & Holehouse, A. S. Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophys. J.* **120,** 4312–4319 (2021).

106. Boeynaems, S., Ma, X. R., Yeong, V., Ginell, G. M., Chen, J.-H., Blum, J. A., Nakayama, L., Sanyal, A., Briner, A., Van Haver, D., Pauwels, J., Ekman, A., Schmidt, H. B., Sundararajan, K., Porta, L., Lasker, K., Larabell, C., Hayashi, M. A. F., Kundaje, A., Impens, F., Obermeyer, A., Holehouse, A. S. & Gitler, A. D. Aberrant phase separation is a common killing strategy of positively charged peptides in biology and human disease. *bioRxiv* (2023). doi:10.1101/2023.03.09.531820

107. Yang, Z., Deng, X., Liu, Y., Gong, W. & Li, C. Analyses on clustering of the conserved residues at protein-RNA interfaces and its application in binding site identification. *BMC Bioinformatics* **21,** 1–14 (2020).

108. Sawle, L. & Ghosh, K. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *J. Chem. Phys.* **143,** 085101 (2015).

109. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. & SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17,** 261–272 (2020).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SINTDRNA2023CLEAN.pdf