PNAS
www.pnas.org

Main Manuscript for

A Modular Approach to Map Out the Conformational Landscapes of Unbound Intrinsically Disordered Proteins

Thinh D.N. Luong[‡,1,2], Suhani Nagpal[‡,1,3], Mourad Sadqi[1,4] & Victor Muñoz[1,2,3,4,*]

[1]NSF-CREST Center for Cellular and Biomolecular Machines (CCBM), University of California at Merced, Merced, 95343 CA

[2]Chemistry and Chemical Biology Graduate Program, University of California at Merced, Merced, 95343 CA

[3]Bioengineering Graduate Program, University of California at Merced, Merced, 95343 CA

[4]Department of Bioengineering, University of California at Merced, Merced, 95343 CA

Email: vmunoz3@ucmerced.edu

Author Contributions: The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. [‡]These authors contributed equally.

Competing Interest Statement: No competing interests.

Classification. Biological Sciences. Biophysics and Computational Biology

Keywords: intrinsic disorder, folding upon binding, conformational rheostat, energy landscape, conformational biases

This PDF file includes:

Main Text
Figures 1 to 6
Table 1

Abstract

Intrinsically disordered proteins (IDPs) fold upon binding to select/recruit multiple partners, morph around the partner's structure, and exhibit allostery. However, we do not know whether these properties emerge passively from disorder, or rather are encoded into the IDP's folding mechanisms. A main reason for this gap is the lack of suitable methods to dissect the energetics of IDP conformational landscapes without partners. Here we introduce such an approach that we term molecular LEGO, and apply it to NCBD, a helical, molten-globule-like IDP, as proof of concept. The approach entails the experimental and computational characterization of the protein, its separate secondary structure elements (LEGO building blocks), and their super-secondary combinations. Comparative analysis uncovers specific, yet inconspicuous, energetic biases in the conformational/ folding landscape of NCBD, including: 1) strong local signals that define the three native helices; 2) stabilization of helix-helix interfaces via soft pairwise tertiary interactions; 3) cooperative stabilization of a heterogeneous 3-helix bundle fold; 4) a dynamic exchange between sets of tertiary interactions (native and non-native) that recapitulate the different structures NCBD adopts in complex with various partners. Crucially, a tug of war between sets of interactions makes NCBD gradually shift between structural sub-ensembles as a conformational rheostat. Such conformational rheostatic behavior provides a built-in mechanism to modulate binding and switch/recruit partners that is likely at the core of NCBD's function as transcriptional coactivator. Hence, the molecular LEGO approach emerges as a powerful new tool to dissect the conformational landscapes of unbound IDPs and rationalize their functional mechanisms.

Significance Statement

Intrinsically disordered proteins have the unique ability of morphing in response to multiple partners and thereby process sophisticated inputs and outputs. It is, however, a mystery whether their response is passive, that is, entirely determined by the partner, or controlled via an internal, yet unknown, folding mechanism. Here we introduce a novel approach to examine this key question and demonstrate its potential by dissecting the conformational properties of the partially disordered protein NCBD and obtaining important clues about how it performs its biological function.

Main Text

Introduction

The traditional biochemical paradigm states that protein sequences are encoded to fold into thermodynamically stable 3D structures that define their biologically functional states (1). However, about 40% of the human proteome appears to be composed of protein domains/regions that are intrinsically disordered (IDPs or IDRs)(2, 3). IDPs are paradigm challengers because they are disordered in their resting state (4, 5), fold, completely or partially, upon binding to their biological effectors (6, 7), can bind structurally diverse partners (8, 9), and exhibit allostery without quaternary or even defined tertiary structure (10, 11). IDPs are more abundant in higher-order organisms, in whom they play key regulatory roles for essential biological processes (12). From a physical viewpoint, IDPs have distinct sequence patterns (13), including high net charge, low hydrophobicity, and enriched proline content (2, 14). Some IDPs are devoid of any structure, even after binding to partners (15), but many are partially disordered (IPDP) and morph to accommodate their partners. Hence, efforts have focused on investigating their folding upon binding (6, 10, 16-18). These studies have shown that IPDPs bind partners via conformational selection (fold first and then bind) or induced-fit (bind first and fold while bound) processes. However, what remains a mystery is the role (if any) that the folding mechanism of the IPDP plays in defining its binding/functional properties. For instance,

structural disorder is often considered sufficient to enable the IPDP to morph into any required shape on cue. But if so, how does an IPDP manage to bind specifically, select among partners, and exhibit allostery? In addition, folding upon binding is often interpreted as a binary transition (conformational switch). Such transitions require simultaneous folding and binding (19), which contradicts findings of IPDPs binding via induced-fit (20, 21) or alternating between conformational selection and induced-fit (7, 22). Moreover, to fold upon binding as a conformational switch, IPDPs sequences would need to fully encode all the structures they form in complex with diverse partners.

A possible solution to these puzzles is for IPDPs to fold upon binding as conformational rheostats (CR)(23), a functional mechanism linked to the gradual structural transitions of downhill folding (24). Downhill domains have IDP-like sequences and are mostly stabilized by local interactions, which makes them fold fast but also marginally unstable, and hence partially disordered (23). The key to CR function is a flexible conformational ensemble with built-in energetic biases towards specific (potentially multiple) sub-ensembles. Such biases would provide the driving force for selecting partners and allostery, whereas the gradual conformational transitions can explain how IPDPs morph around diverse partners and combine conformational selection and induced-fit binding (23). The connections between downhill folding and IPDP binding have been explored using computational approaches (19, 25, 26). However, to establish whether the folding mechanism is what controls IPDPs' binding and function, it is essential to resolve the conformational landscapes and energetics of the IPDP in absence of partners. Achieving this by experiment has been a major hurdle. The standard approach to investigate protein conformational ensembles relies on thermodynamic and/or kinetic measurements of the (un)folding transition and their analysis with a two-state model (unfolded and native) to determine the changes in free energy upon folding, unfolding, and in equilibrium (27). When performed on collections of select mutants, these experiments provide local perturbation maps that can be used to infer the folding landscape (28). The analysis requires a cooperative (un)folding transition with well-defined ends from which to determine and extrapolate the properties of the interconverting states. For IDPs, this key requirement is met when folding is induced by binding using the partner's concentration as thermodynamic variable (16, 17), but not in the absence of partner. Even partially structured IPDPs exhibit transitions that are too broad and uncooperative for such an approach (29). As a consequence, the folding landscapes of IDPs without partners have only been accessible via molecular simulations (26, 30-32). Such simulations have led to important insights, but it is essential to crosscheck them by experiment at levels comparable to what has been recently attempted for IDP folding upon binding (33).

In response to this challenge, we introduce here a modular approach that we term molecular LEGO. The approach starts by decomposing an IPDP into its basic secondary structural elements, or LEGO building blocks, and their combinations. The combined elements recapitulate subsets of tertiary interactions, in analogy to the complementary indentations between bricks in the LEGO toy. The molecular LEGO is inspired by work in the early 90s that searched for local folding nuclei on two-state folding proteins (34), and which revealed weak local biases (34) and the need for nearly the entire protein to elicit detectable folding (35). A more recent study on the IDP ACTR has shown similarly weak local conformational biases (36). The dissection of an IDP into structural elements has also been used in molecular simulation studies to facilitate conformational sampling via the much faster dynamics of the fragments (37). The key addition here is the comparative quantitative analysis of hierarchically organized protein segments via experiments and simulations. In this regard, the conformational analysis of the building blocks probes local interactions, but also provides reference ensembles for interpreting the properties of higher-order fragments. Such reference ensembles are essential to reliably detect the subtle biases expected on IPDPs, and to convert them into energetic contributions using simple statistical thermodynamic analysis. We contend that such modular approach can provide new key insights about the tertiary interactions and cooperative energetics that stabilize IPDP folding ensembles in absence of partners. To demonstrate this assertion, we focused on the protein NCBD. NCBD is categorized as IPDP, and there is a wealth of biophysical data available on its folding and binding to compare with, including NMR (29, 38), molecular simulations (25, 31) and single-molecule FRET (39-41). NCBD binds to multiple, structurally diverse partners, including IDPs (e.g., p53-TAD (38) and ACTR (8)) and globular proteins such as IRF3 (42), by adapting its ensemble to the partner's properties. In its free form, NCBD exhibits high α-helical content without defined tertiary structure, but it forms a dynamic three-helix bundle driven by a few mid-range contacts (29). Critically, the (dis)ordering transitions of NCBD are broad and featureless, including its thermal unfolding and stabilization via the cosolvent trifluoroethanol (Fig. S1). All these properties make NCBD ideal for a molecular LEGO proof of concept.

Results

Molecular LEGO Design. The design of the LEGO elements (locations and extension along the sequence) on highly disordered proteins is far from trivial unless there are available structures in complex with partners. IPDPs, however, do have residual structure, which for NCBD was sufficient to enable the determination on an NMR ensemble based on chemical shifts and a few mid-range NOEs (29). We used this NMR ensemble to divide the 59-residue sequence of NCBD into four building blocks that represent its local (secondary) structural segments: helices 1, 2, and 3 (H1, H2, H3) and the C-terminal tail (T). We further refined the limits of the α-helical regions based on predictions of helical propensity from AGADIR (43), which delineate a distinct helix profile (Fig. S2). We then designed four combinations of consecutive building blocks (H1H2, H2H3, H3T, H2H3T) that recapitulate the various sets of "native" pairwise tertiary interactions. Finally, the comparison of LEGO elements with the entire protein is expected to inform on the overall contribution from global cooperativity. The complete molecular LEGO design of NCBD is shown in Fig. 1.

Analysis of Conformational Ensembles. We analyzed NCBD and its LEGO elements by experiment and simulation. Experimentally, we employed far-UV circular dichroism spectroscopy, which reports on the average peptide bond conformation and is particularly sensitive to α-helical structure (NCBD and most IPDPs are, or become upon binding, α-helical). We use the cosolvent 2,2,2-trifluoroethanol (TFE) as structure-promoting agent. TFE is a polar/organic cosolvent that induces local structure in peptides and proteins by strengthening the backbone intramolecular hydrogen bonds (44). TFE has been widely used as helix-promoting agent(45), but is also known to stabilize β-hairpin structures (46, 47) and to promote hydrophobic interactions by changing the hydration shell (48). The TFE CD titration of H1 is given in Fig. 2 (left) as an example. In the absence of TFE, the CD spectrum of H1 indicates ~20% α-helix with the remainder being random coil. TFE addition steadily increases the α-helical content of H1 until it plateaus (beyond 0.3 ϕ_{TFE}). Although quantitatively different, the TFE titrations of all the other LEGO elements and full NCBD share the same features (all data shown in Fig. S3). These results indicate that all these TFE titrations can be analyzed in terms of the helix-coil transition, which describes α-helix formation as the interplay between nucleation (σ) and elongation (s)(49). The effect of TFE on helix formation can be simply described as an enhancement in elongation (larger s) due to stronger hydrogen bonds, and hence as sequence independent. Here we used $s(TFE) = 2.75s(H_2O)$, or a ~1 RT stabilization, for all the molecules. The effective s_* at each TFE volume fraction can be calculated as the weighted average of both s-values according to the composition of the mixed solvent (1-ϕ_{TFE} and ϕ_{TFE}) as shown in the Fig. 2 right equation (see SI). When the polypeptide has sufficiently high σ and s parameters in water, the addition of TFE promotes a cooperative (sigmoidal) transition to α-helical structure (Fig. 2 right). In this case, however, is not appropriate to use a homopolymer helix/coil model because the NCBD sequence is highly heterogeneous (Fig. 1). To describe how such heterogeneity can affect the average helical content as a function of TFE (CD only reports the average peptide bond conformation), we implemented a tripartite helix-coil model based on the original Zimm-Bragg treatment (50). The tripartite model discretizes the helical propensity spectrum of a hetero-polypeptide chain into three types of units (peptide bonds): PH, which are already α-helical without TFE; RC, which are random coil regardless of TFE; and IH, which have residual α-helix population that is enhanced by TFE (Fig. 2 right). The model defines the average number of helical peptide bonds on any peptide/protein with four parameters: the number of PH units, and σ, s, and number of IH units (Fig. 2 right); that is only one more than a standard homopolymer helix-coil model. The tripartite model fits the data of all the NCBD molecules much better than the 3-parameter homopolymer model, with an improved performance that is statistically significant at >99% confidence according to the F-test (see SI).

We also performed atomistic MD simulations in explicit solvent: two independent 12 μs trajectories for NCBD and 2-3 sets of 2 μs trajectories for each LEGO element, as we expected faster conformational dynamics on them. We used the CHARMM22* force field with TIP3P water, which have been found suitable for partially disordered proteins (51, 52). We first examined the MD simulations using the fraction of native contacts (Q) as order parameter (Fig. S5). The LEGO building blocks showed sharp fluctuations in Q (they have few native contacts) that take place in tens of ns. The combined LEGO elements exhibited Q fluctuations of smaller amplitude and slower dynamics, but several transitions were still observable in each 2 μs trajectory (Fig. S5). The behavior of NCBD is similar, although with an additional slowdown: six times longer trajectories produce similar numbers of transitions. The observation of several transitions per trajectory and the consistency between independent trajectories suggest that conformational sampling within these timescales is reasonable. We then computed the fraction helix, and nucleation and elongation parameters, for each peptide bond in each molecule. The agreement between the residue-specific helix populations obtained from independent simulations (Figs. 3-5 and S6) further supports that the

simulated timescales afford reasonable sampling. The fraction helix profiles of the LEGO elements and NCBD are given in Figs. 3-5 and S7.

Conformational Propensities of LEGO Building Blocks. In general, we find that the three regions containing α-helices in the native NMR ensemble have residual helical structure and are highly sensitive to TFE (Fig. 3). H1 has the highest residual helical structure, both in experiments and simulations. The maximal helix lengths (i.e., at the highest ϕ_{TFE}) are just one residue longer than in the NMR ensemble, which indicates that the three NCBD helices are defined by strong local signals. The tail (T) does not have detectable helix, but forms a single helical turn (i.e., 1 hydrogen-bonded unit) at the highest ϕ_{TFE}. The TFE transitions are well reproduced by the tripartite helix-coil model, which reveals that the costs of nucleation (σ) are close to the values for polyalanine-based peptides (53). H1 and H3 are slightly easier to nucleate and hence less cooperative than H2. Elongation is slightly <1 for all the peptides, which explains their residual helix (on an infinitely long helix $s = 1$ results in 50% helix), but also their high sensitivity to TFE. T is disordered but contains a short region that is primed to become helical by stabilizing factors.

The MD simulations are in good agreement with the experimental findings, including the average helix content per molecule (particularly H1 and H3), and the presence of marginal helical propensity in T. They also show non-uniform helix populations, hence further supporting the analysis of the experiments with the tripartite helix-coil model. The helical regions in simulations are also in excellent agreement with the NCBD NMR ensemble, confirming the presence of strong local signals. In contrast, the simulations produce systematically lower nucleation costs (about 5-10-fold larger σ) and less propensity to elongate (smaller *s*). Interestingly, the differences in σ and *s* compensate each other to produce similar helical contents (Fig. 3). The implication is that the force field/water model underestimate the cooperativity of the helix-coil transition, and generally of folding, a result that is consistent with previous comparative studies (54).

Estimating Pairwise Tertiary Interactions. The results of the combined LEGO elements are qualitatively similar: i) residual helical structure in native conditions, ii) strong response to TFE, iii) sigmoidal TFE transitions, and iv) agreement with the helix lengths in the NMR ensemble (Fig. 4). However, the comparison between combined LEGO elements and the compounded effects of their individual building blocks (grey curves) reveals significant contributions from tertiary interactions. For instance, the combined elements exhibit enhanced sensitivity to TFE, as manifested by sharper slopes and reaching plateau at lower ϕ_{TFE}, and hence larger σ and *s*; albeit the experiments do not detect marked net increases of helical structure in water. This indicates that each set of pairwise tertiary interactions is insufficient to significantly increase the helix population on its own. The simulations do show enhanced helical content, possibly owing to their much higher sensitivity and resolution. Another observation is that the thermodynamic coupling between consecutive LEGO building blocks has significant impact on redefining the maximal helix lengths, most notably of H3.

On an individual basis, we find that the interactions between helices 1-2 are stronger than between 2-3. H1H2 does in fact exhibit enhanced helical content also in experiments, in excellent agreement with the simulations (cyan in Fig. 4). The effects on H2H3 are more subdued in simulations and only detectable from the TFE response in experiments. The impact of the tail on helix 3 is interesting, as the extended C-terminal sequence stimulates the growth of the helix beyond that found in the NMR ensemble. Helix extension is clear in experiments (3 more residues) and simulations (see H3T in orange in Fig. 4). In other words, the tail does not nucleate helix structure on its own, but it extends a helix coming from the preceding sequence. The simulations indicate that this effect is purely driven by local interactions (helix-coil cooperativity). The extension of H3 onto the tail is also predicted by AGADIR (Fig. S2), further supporting its local origin.

Pairwise interactions do have distinct effects on defining the length of the helices. For instance, the interactions between helices 1-2 do not change the length of either helix in experiments or simulations. In contrast, experiments on H2H3 indicate a maximal helix of ~23 residues (vs. 25 in the NMR ensemble) and ~28 in the sum of H2 and H3. This difference seems to arise in part from helix capping effects of the region connecting helices 1 and 2, which is absent in H2H3 and H2H3T (Fig. 1). This effect is also evident in the simulations, which show some helix population in that connecting region, as well as the stabilization of the beginning of helix 2 in H2 relative to H2H3 (Figs. 3 vs. 4). The experiments also show that helix 2 impedes the elongation of helix 3 into the tail: H2H3T has a maximum helix of 26, in perfect agreement with the NMR ensemble, whereas H2 and H3T add up to almost 30. The same pattern is observed in simulations, which show a longer third helix in H3T than in H2H3T. Strikingly, the simulations also reveal *"non-native"* effects of the tail, which stabilizes helices 2 and 3 without becoming itself helical (brown vs. orange in Fig. 4). Experiments confirm this observation, showing enhanced elongation (*s*) and reduced helix length of H2H3T vs. H3T. The main discrepancy between experiments and simulations is

quantitative: the helix stabilization induced by the tail is stronger in simulations. Hence, the simulations overestimate the helical content, most particularly for H3T and H2H3T, and to a lesser extent, H2H3.

Global Stabilization of the NCBD Ensemble. The LEGO results provide a reference to interpret the uncooperative (non-sigmoidal) TFE transition of full NCBD, which is, in fact, much broader than those of its elements (Fig. 5). Compounding different LEGO elements, we can establish the behavior expected from only local interactions (grey), or after adding the interactions between helices 1-2 (green), or between helices 2-3 and tail (pink). This comparison demonstrates that NCBD has much higher helical content than the sum of its parts: ~24 helical residues in water relative to 6-7 residues for the three combinations (Fig. 5). Helix-coil analysis indicates that ~15 residues are fully helical (PH) in water, whereas the remainder comes from the partial helical population (~30%) of many other IH residues. Hence, in NCBD, the helix-inducible residues (IH) already have high helical content in water, which enormously facilitates nucleation: 10-fold higher σ relative to the LEGO elements. Elongation (*s*) is, on the other hand, minimally higher. In other words, the low TFE sensitivity of NCBD is not because its conformational ensemble is disordered, but because it is already highly primed towards forming α-helical structure via interactions that can only be formed in the entire protein. The effect of TFE on folded globular proteins is complex: it switches from native-stabilizing at low volume fractions to denaturing as TFE becomes the main solvent (44). In NCBD, we see that the native-stabilizing effect extends further in TFE concentration. Indeed, at 0.5 $\phi_{TFE,}$ NCBD reaches ~41 helical residues, in agreement with the NMR ensemble (dashed line in Fig. 5). However, the helix-coil parameters indicate that helix content keeps growing beyond this point (~4 more residues), hence starting to promote non-native conformations. Such an extended native-stabilizing range for TFE could reflect the fact that NCBD is inherently α-helical and lacks a defined hydrophobic core (44). This property could be common to other IPDPs.

For NCBD, the simulations closely reproduce the main experimental results: helical content in water (Fig. 5), nucleation and elongation (Table S2). The simulations also show that helix 2, which has the lowest intrinsic propensity of the three (Fig. 3), is preferentially stabilized in the full protein (Fig. 5), and engages in frequent interactions with the other two helices. The stabilization of helix 2 in presence of both flanking helices is evident in the NCBD helix profile relative to the H1H2+H3T (green) and H1+H2H3T (pink) compounded profiles. This comparison also highlights that helix 1 is mostly stabilized by 1-2 interactions, and helix 3 is stabilized/delimited by its interplay with helix 2 and tail. The NCBD simulations also show the transient formation of many long-range interactions that were not detected in the NMR ensemble ("non-native"); particularly between the tail and helix 1, and between helices 1-3. These interactions are not native but are still consistent with an antiparallel helix bundle fold. Moreover, they contribute to stabilize the helical structure of the NCBD ensemble. For instance, interactions with helix 1 make the tail regain helix structure that is suppressed by helix 2 (Fig. 5). Transient interactions between helices 1 and 3, which were not found by NMR (29), also contribute to stabilize the three-helix bundled ensemble in the simulations.

Interaction Network and Cooperativity. The left panel of Fig. 6 shows the time-averaged "native" contacts observed in simulations of NCBD (bottom right) and the LEGO elements (top left). These maps reveal that H1H2 and H2H3 reproduce the native interactions present in full NCBD, albeit their contacts are slightly more transient. However, NCBD also engages in many non-native interactions, including interactions that are longer range than the super-secondary structures recapitulated by LEGO elements (Fig. 6 right). These "non-native" interactions emerge as the differential factor in cooperatively biasing the conformational landscape of NCBD.

To estimate the energetic contributions from each set of interactions, we resorted to the helix-coil parameters from the LEGO analysis (Figs. 3-5) to calculate the statistical weight for forming a fully "native" α-helix conformation for each molecule. We then estimated the change in free energy from the ratio between the weight of a given combined LEGO element and the product of the weights of its building blocks (see SI). We performed this calculation for the experimental and simulation data (Table 1). The experiments indicate that each set of pairwise tertiary interactions (helices 1-2 and 2-3) contributes ~5-6 kJ/mol, which is comparable to the mean perturbation induced by single-point mutations on folded proteins(55). The interplay between helices 2, 3 and tail contributes ~3 kJ/mol more. The total NCBD stabilization amounts to ~30 kJ/mol, which is comparable to the chemical denaturation free energies of two-state folding proteins, even though NCBD is an IPDP. However, such comparison is misleading because the 30 kJ/mol for NCBD are referenced to a fully disordered ensemble (building blocks). In contrast, unfolded states have residual local structure (56). In general, the simulations produce much stronger pairwise tertiary interactions.

To estimate the cooperative (non-additive) contributions, we subtracted the pairwise interactions from the NCBD total stabilization. This calculation leads to an experimental estimate of ~17 kJ/mol, and of ~5 kJ/mol for the simulations (Table 1). The much smaller value for simulations is consistent with prior reports of MD simulations

underestimating folding cooperativity (54, 57). As for the source of such cooperativity, it seems to arise from the simultaneous formation of tertiary interactions between helices 1-2 and 2-3, and non-native interactions between helices 1-3 with the tail. The simulations also reveal that these sets of interactions compete with one another, resulting in alternating structural patterns. The conflict between sets of tertiary interactions, jointly with strong local propensities, explains why NCBD forms a highly dynamic ensemble rather than one 3D structure.

Discussion

Since IDPs were first identified, we have faced the challenge of explaining how these proteins integrate intrinsic disorder with the ability to select partners, fold upon binding, bind multiple partners, and switch among them in allosteric fashion. A key barrier has been the lack of methods that can dissect the conformational landscapes of IDPs in the absence of partners. Here we introduce a modular approach that is purposely designed to tackle this challenge (molecular LEGO) and apply it to the IPDP NCBD. The approach enables a direct comparison between experiments and simulations in a synergistic fashion. The molecular LEGO should, in principle, be easily generalizable to other IPDPs and hence it adds a powerful new tool for IDP research. In this regard, we outline some basic rules for its general application to disordered proteins:

1) A key element is the design of the LEGO elements. Ideally, one should use a structural ensemble of the unbound protein determined with one of the existing approaches for generating IDP ensembles from limited experimental restraints (58-60). As alternative, one can use a structure of the IDP in complex with a partner, or even a secondary structure prediction profile (61).

2) Because these proteins are flexible/disordered, is convenient to use a structure-promoting cosolvent as thermodynamic variable, which also facilitates comparison with their folding upon binding behavior. TFE is a good option, particularly for IDPs that form α-helical structure (free or upon binding). Other alternatives are osmolytes, such as betaine and TMAO (62), and salts, given that IDPs have very high net charges (13).

3) The conformational analysis should be carried out with techniques sensitive to the backbone conformation. Residue-averaged information is sufficient to address general mechanistic questions, as we do here with circular dichroism, or alternatively with infrared spectroscopy. NMR is an excellent choice since it provides residue-specific structural information, but it could be too labor-intensive to apply to all the LEGO elements and combinations.

4) It is essential to use a statistical thermodynamic treatment to analyze the experimental data, rather than assuming a two-state transition. Such treatment could be simple but should consider conformational entropy explicitly in terms of ensembles of microstates. Molecular simulations can test the physical significance of the model used to analyze the experiments.

On a second front, the molecular LEGO study presented here sheds much needed light into key mechanistic questions related to the conformational behavior of IDPs in general, and of NCBD in particular. Our results demonstrate that the amino acid sequence of NCBD contains strong local signals that singlehandedly define the secondary structural elements present in the ensemble. This observation supports the hypothesis that the conformational behavior of IPDPs is connected to the energetics of downhill folding (23). The combined LEGO elements demonstrate that the few tertiary contacts observed by NMR in NCBD produce energetic biases that help promote an overall helix bundle fold. However, these energetic contributions are relatively small (~5-6 kJ/mol for each set of pairwise tertiary interactions: helices 1-2, and 2-3). From simulations we find that the native tertiary contacts do form frequently but are transient (Fig. 6). These results explain the puzzling observation of specific long-range NOEs on an otherwise molten-globule-like ensemble (29).

The behavior of full NCBD relative to the LEGO elements provides other important clues about IPDP energetics and folding landscapes. For instance, the tertiary interactions between helices 1-2 and 2-3 cooperate in the stabilization of NCBD's helix-bundle fold (mostly via the stabilization of helix 2). But we find that NCBD is much more ordered than expected from just its local and "native" pairwise tertiary interactions. Specifically, our experimental analysis reveals an extra of ~17 kJ/mol stabilization of the NCBD ensemble. That is, the structural factors used to calculate the NMR structure (local conformation and a few long-range NOEs) amount to less than 50% of the total ensemble energetics (Table 1). We find evidence of several such "non-native" factors. The C-terminal tail, which is fully disordered in the NMR ensemble, turns out to be a major player. The tail alone elongates helix 3, but the interactions of helices 2-3 block such extension and keep the tail disordered (H3T vs. H23T in Fig. 4). The tail can also interact with helix 1, resulting on end-to-end contacts (Fig. 6 right) that stabilize helix 1 and form one helix turn on the tail. This helix turn is disconnected from, and bent relative to, helix 3. The end of helix

1 also interacts with the start of helix 3 in parallel fashion (Fig. 6 right), which involves breaking many of the "native" interactions between helices 1-2 and 2-3. The pivotal role of the tail is highlighted by comparing our results with previous simulations of NCBD in which the tail was truncated (25). We note that all of these "non-native" factors can be inferred from, or are consistent with, the LEGO experiments. They are, however, most evident in the simulations. This synergy highlights the importance of combining experiments and simulations in IDP research.

The picture that emerges from our dissection of the NCBD energy landscape is one of a protein with strong local structural biases and a tug of war between sets of tertiary interactions, each stabilizing a distinct conformational sub-ensemble. Hence, the apparent disorder of NCBD arises from the conflict between competing tertiary interactions, which makes NCBD to dynamically alternate between sub-ensembles with slightly different fold architecture. This behavior is in stark contrast with the usual interpretation of disorder as indicative of absent tertiary interactions. Remarkably, the conformational properties we find on NCBD reveal an internal mechanism for driving its sophisticated, multi-partner, folding upon binding behavior. The 3D structure of NCBD in complex with p53-TAD (38) is fully consistent with the "native" sub-ensemble in which helices 1 and 3 interact with helix 2 but do not with each other, and the tail is disordered. These conformational biases are recapitulated by the LEGO elements H1H2, H2H3, and T. In contrast, ACTR and NCBD form an intertwined complex in which helices 2 and 3 of NCBD are set apart by ACTR and helix 3 elongates onto the tail (8), precisely as we see in H3T and H23T. Finally, the "non-native" interactions of helix 1 with helix 3 and tail are fully consistent with the structure that NCBD forms in complex with the stably folded IRF3 (42).

Summarizing, the NCBD folding landscape has built-in energetic biases that cooperate and compete to stabilize the various conformational sub-ensembles that NCBD forms in complex with structurally diverse partners. This behavior uncovers an internal folding mechanism to select partners and modulate affinity that is likely essential for NCBD's recruiting role as transcription coactivator (12). The mechanism we report for NCBD is indicative of a conformational rheostat. It also demonstrates that the molecular LEGO approach can be used to map out subtle energetic biases on IPDPs, which are possibly essential to their biological function.

Materials and Methods

An extended description of materials and methods is provided in supplementary information.

NCBD and Lego elements. Full NCBD was produced by recombinant means as a His-tag fusion and purified by affinity and reverse phase chromatography. Peptides corresponding to the 8 Lego elements and combinations were chemically synthesized by Bio-Synthesis Inc. (Texas).

Experimental Conformational Analysis. The conformational properties of NCBD and Lego elements were characterized using far-UV circular dichroism spectra as a function of the helix promoting agent TFE. The spectra were analyzed using singular value decomposition (SVD) to determine the average number of helical residues per condition. Each CD spectra vs. TFE dataset was analyzed with a tripartite helix/coil transition model in which the average number of helical residues at any given condition arises from the combination of three types of residues: pre-formed helix (PH), random coil (RC), and the elongation and nucleation of TFE-inducible helix (IH). The effect of TFE was modeled to increase elongation in sequence independent manner as $s_* = s(1 + 1.75\Phi_{TFE})$.

Computational Conformational Analysis. Molecular dynamics simulations in explicit solvent were performed using the GROMACS package, the Charmm22* force field and the TIP3P water model. We obtained a total of 24 μs of simulation time for NCBD, 6 μs for H12 and H23T, and 4 μs tall the other peptides. All trajectories were analyzed to compute dihedral angles, hydrogen bonds, fraction of native contacts, time-averaged contact maps, and residue-specific helix elongation and nucleation parameters.

Acknowledgments

This work was supported by the National Science foundation (NSF-MCB-1616759) and the CREST Center for Cellular and Biomolecular Machines (grant NSF-CREST-1547848). V.M. acknowledges additional support from the W.M. Keck Foundation.

References

1. C. B. Anfinsen, Principles that govern the folding of protein chains. *Science* **181**, 223-230 (1973).
2. C. J. Oldfield, A. K. Dunker, Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. *Annu. Rev. Biochem.* **83**, 553-584 (2014).
3. M. M. Babu, The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem. Soc. Trans.* **44**, 1185-1200 (2016).
4. P. Tompa, Intrinsically disordered proteins: a 10-year recap. *Trends in biochemical sciences* **37**, 509-516 (2012).
5. V. N. Uversky, Unusual biophysics of intrinsically disordered proteins. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1834**, 932-951 (2013).
6. P. E. Wright, H. J. Dyson, Linking folding and binding. *Curr. Opin. Struct. Biol.* **19**, 31-38 (2009).
7. J. Dogan, S. Gianni, P. Jemth, The binding mechanisms of intrinsically disordered proteins. *Phys. Chem. Chem. Phys.* **16**, 6323-6331 (2014).
8. S. J. Demarest *et al.*, Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature* **415**, 549-553 (2002).
9. L. Waters *et al.*, Structural diversity in p160/CREB-binding protein coactivator complexes. *J. Biol. Chem.* **281**, 14787-14795 (2006).
10. V. J. Hilser, E. B. Thompson, Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proc. Natl. Acad. Sci.* **104**, 8311 (2007).
11. H. N. Motlagh, J. O. Wrabl, J. Li, V. J. Hilser, The ensemble nature of allostery. *Nature* **508**, 331-339 (2014).
12. P. E. Wright, H. J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18-29 (2015).
13. M. M. Babu, R. W. Kriwacki, R. V. Pappu, Versatility from Protein Disorder. *Science* **337**, 1460 (2012).
14. V. N. Uversky, A. K. Dunker, Understanding protein non-folding. *Biochim. Biophys. Acta.* **1804**, 1231-1264 (2010).
15. A. Borgia *et al.*, Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **555**, 61-66 (2018).
16. S. L. Shammas, M. D. Crabtree, L. Dahal, B. I. Wicky, J. Clarke, Insights into coupled folding and binding mechanisms from kinetic studies. *Journal of Biological Chemistry* **291**, 6689-6695 (2016).
17. S. Gianni, J. Dogan, P. Jemth, Coupled binding and folding of intrinsically disordered proteins: what can we learn from kinetics? *Current opinion in structural biology* **36**, 18-24 (2016).
18. F. Wiggers *et al.*, Diffusion of a disordered protein on its folded ligand. *Proceedings of the National Academy of Sciences* **118** (2021).
19. R. Sharma, D. De Sancho, V. Muñoz, Interplay between the folding mechanism and binding modes in folding coupled to binding processes. *Phys. Chem. Chem. Phys.* **19**, 28512-28516 (2017).
20. J. Lätzer, G. A. Papoian, M. C. Prentiss, E. A. Komives, P. G. Wolynes, Induced Fit, Folding, and Recognition of the NF-κB-Nuclear Localization Signals by IκBα and IκBβ. *J. Mol. Biol.* **367**, 262-274 (2007).
21. J. M. Rogers *et al.*, Interplay between partner and ligand facilitates the folding and binding of an intrinsically disordered protein. *Proc. Natl. Acad. Sci.* **111**, 15420 (2014).
22. S. Sen, J. B. Udgaonkar, Binding-induced folding under unfolding conditions: Switching between induced fit and conformational selection mechanisms. *Journal of Biological Chemistry* **294**, 16942-16952 (2019).
23. V. Muñoz, L. A. Campos, M. Sadqi, Limited cooperativity in protein folding. *Curr. Opin. Struct. Biol.* **36**, 58-66 (2016).

24. M. M. Garcia-Mira, M. Sadqi, N. Fischer, J. M. Sanchez-Ruiz, V. Muñoz, Experimental identification of downhill protein folding. *Science* **298**, 2191-2195 (2002).
25. A. N. Naganathan, M. Orozco, The Native Ensemble and Folding of a Protein Molten-Globule: Functional Consequence of Downhill Folding. *J. Am. Chem. Soc.* **133**, 12154-12161 (2011).
26. Y. Wang *et al.*, Multiscaled exploration of coupled folding and binding of an intrinsically disordered molecular recognition element in measles virus nucleoprotein. *Proceedings of the National Academy of Sciences* **110**, E3743-E3752 (2013).
27. C. Tanford, Protein denaturation. *Adv. Protein. Chem.* **23**, 121-282 (1968).
28. A. R. Fersht, L. Serrano, Principles of protein stability derived from protein engineering experiments. *Curr. Opin. Struct. Biol.* **3**, 75-83 (1993).
29. M. Kjaergaard, K. Teilum, F. M. Poulsen, Conformational selection in the molten globule state of the nuclear coactivator binding domain of CBP. *Proc. Natl. Acad. Sci.* **107**, 12535-12540 (2010).
30. D. Ganguly, J. Chen, Atomistic details of the disordered states of KID and pKID. Implications in coupled binding and folding. *Journal of the American Chemical Society* **131**, 5214-5223 (2009).
31. M. Knott, R. B. Best, A Preformed Binding Interface in the Unbound Ensemble of an Intrinsically Disordered Protein: Evidence from Molecular Simulations. *PLoS Comput. Biol.* **8**, e1002605 (2012).
32. W. Zhang, D. Ganguly, J. Chen, Residual structures, conformational fluctuations, and electrostatic interactions in the synergistic folding of two intrinsically disordered proteins. *PLoS computational biology* **8**, e1002353 (2012).
33. J. Zou, C. Simmerling, D. P. Raleigh, Dissecting the Energetics of Intrinsically Disordered Proteins via a Hybrid Experimental and Computational Approach. *J. Phys. Chem. B .* **123**, 10394-10402 (2019).
34. H. J. Dyson, P. E. Wright, Peptide conformation and protein folding. *Curr. Opin. Struct. Biol.* **3**, 60-65 (1993).
35. J. L. Neira, L. S. Itzhaki, D. E. Otzen, B. Davis, A. R. Fersht, Hydrogen exchange in chymotrypsin inhibitor 2 probed by mutagenesis11Edited by J.Karn. *J. Mol. Biol.* **270**, 99-110 (1997).
36. M. Kjaergaard *et al.*, Temperature-dependent structural changes in intrinsically disordered proteins: Formation of α–helices or loss of polyproline II? *Protein Science* **19**, 1555-1564 (2010).
37. R. J. Lindsay, R. A. Mansbach, S. Gnanakaran, T. Shen, Effects of pH on an IDP conformational ensemble explored by molecular dynamics simulation. *Biophys. Chem.* **271**, 106552 (2021).
38. C. W. Lee, M. A. Martinez-Yamout, H. J. Dyson, P. E. Wright, Structure of the p53 transactivation domain in complex with the nuclear receptor coactivator binding domain of CREB binding protein. *Biochemistry* **49**, 9964-9971 (2010).
39. A. C. Ferreon, J. C. Ferreon, P. E. Wright, A. A. Deniz, Modulation of allostery by protein intrinsic disorder. *Nature* **498**, 390-394 (2013).
40. J.-Y. Kim, F. Meng, J. Yoo, H. S. Chung, Diffusion-limited association of disordered protein by non-native electrostatic interactions. *Nat. Commun.* **9**, 4707 (2018).
41. F. Sturzenegger *et al.*, Transition path times of coupled folding and binding reveal the formation of an encounter complex. *Nat. Commun.* **9**, 4708 (2018).
42. B. Y. Qin *et al.*, Crystal Structure of IRF-3 in Complex with CBP. *Structure* **13**, 1269-1277 (2005).
43. V. Muñoz, L. Serrano, Elucidating the folding problem of helical peptides using empirical parameters. *Nat. Struct. Biol.* **1**, 399-409 (1994).
44. M. Buck, Trifluoroethanol and colleagues: cosolvents come of age. Recent studies with peptides and proteins. *Q. Rev. Biophys.* **31**, 297-355 (1998).
45. P. Luo, R. L. Baldwin, Mechanism of Helix Induction by Trifluoroethanol: A Framework for Extrapolating the Helix-Forming Properties of Peptides from Trifluoroethanol/Water Mixtures Back to Water. *Biochemistry* **36**, 8413-8421 (1997).
46. F. J. Blanco, L. Serrano, Folding of protein G B1 domain studied by the conformational characterization of fragments comprising its secondary structure elements. *Eur. J. Biochem.* **230**, 634-649 (1995).

47. M. S. Searle, R. Zerella, D. H. Williams, L. C. Packman, Native-like β-hairpin structure in an isolated fragment from ferredoxin: NMR and CD studies of solvent effects on the N-terminal 20 residues. *Protein Eng. Des. Sel.* **9**, 559-565 (1996).
48. H. Reiersen, A. R. Rees, Trifluoroethanol may form a solvent matrix for assisted hydrophobic interactions between peptide side chains. *Protein Eng.* **13**, 739-743 (2000).
49. U. Doshi, V. Muñoz, Kinetics of α-helix formation as diffusion on a one-dimensional free energy surface. *Chem. Phys.* **307**, 129-136 (2004).
50. B. H. Zimm, J. K. Bragg, Theory of the Phase Transition between Helix and Random Coil in Polypeptide Chains. *J. Chem. Phys.* **31**, 526-535 (1959).
51. K. Lindorff-Larsen, N. Trbovic, P. Maragakis, S. Piana, D. E. Shaw, Structure and Dynamics of an Unfolded Protein Examined by Molecular Dynamics Simulation. *J. Am. Chem. Soc.* **134**, 3787-3791 (2012).
52. S. Rauscher *et al.*, Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* **11**, 5513-5524 (2015).
53. J. M. Scholtz, R. L. Baldwin, The mechanism of alpha-helix formation by peptides. *Annu. Rev. Biophys. Biomol. Struct.* **21**, 95-118 (1992).
54. R. B. Best, G. Hummer, Optimized Molecular Dynamics Force Fields Applied to the Helix−Coil Transition of Polypeptides. *J. Phys. Chem. B .* **113**, 9004-9015 (2009).
55. D. De Sancho, V. Muñoz, Integrated prediction of protein folding and unfolding rates from only size and structural class. *Phys. Chem. Chem. Phys.* **13**, 17030-17043 (2011).
56. L. A. Campos, M. Sadqi, V. Muñoz, Lessons about Protein Folding and Binding from Archetypal Folds. *Acc. Chem. Res.* **53**, 2180-2188 (2020).
57. L. Sborgi *et al.*, Interaction networks in protein folding via atomic-resolution experiments and long-time-scale molecular dynamics simulations. *Journal of the American Chemical Society* **137**, 6506-6516 (2015).
58. D. H. Brookes, T. Head-Gordon, Experimental Inferential Structure Determination of Ensembles for Intrinsically Disordered Proteins. *J. Am. Chem. Soc.* **138**, 4530-4538 (2016).
59. Y. He, S. Nagpal, M. Sadqi, E. de Alba, V. Muñoz, Glutton: a tool for generating structural ensembles of partly disordered proteins from chemical shifts. *Bioinformatics* **35**, 1234-1236 (2019).
60. M. R. Jensen, R. W. Ruigrok, M. Blackledge, Describing intrinsically disordered proteins at atomic resolution by NMR. *Curr. Opin. Struct. Biol.* **23**, 426-435 (2013).
61. P. Y. Chou, G. D. Fasman, Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.* **47**, 45-148 (1978).
62. D. W. Bolen, Protein stabilization by naturally occurring osmolytes. *Methods Mol. Biol.* **168**, 17-36 (2001).

Figure Legends

Figure 1. Molecular LEGO design. (Top to bottom) The complete NCBD sequence (ID: 2KKJ) and a diagram showing the 3 α-helices of the NMR ensemble in navy blue. Sequences of the 8 LEGO elements: building blocks in primary colors (H1 green, H2 blue, H3 red, T yellow), and combined elements in secondary colors (H1-H2 cyan, H2-H3 magenta, H3-T orange, and H2-H3-T brown). Sketch showing the structure of each fragment and full NCBD (same color code). The building blocks report on secondary structure propensities, and their combinations on pairwise tertiary interactions: e.g., H1-H2 reports on the interactions between helices 1 and 2. Comparison with the full protein reports on the degree of cooperativity.

Figure 2. Experimental conformational analysis. Left) CD spectra of H1 as a function of TFE volume fraction (ϕ_{TFE}). Right) tripartite helix-coil analysis. The top shows an exemplary peptide with preformed helix (PH), TFE-inducible helix (IH), and random coil (RC) units. TFE increases elongation (*s*) in sequence independent fashion. The average number of helical residues obtained from CD (dark blue) is fit to equation 7 (SI) to obtain σ, *s*, IH and PH. RC is obtained as: $RC = N - IH - PH$.

Figure 3. LEGO building blocks. Colors as in Figure 1. From top left to bottom right) Experimental number of helical residues of H1, H2, H3, and T as a function of ϕ_{TFE}. Error bars indicate 1 S.D. from two experiments. The curves represent fits to equation 7 (SI), fitted parameters and fitting errors (one standard deviation) are given in insets. Dash lines indicate the helix length in the NMR structure. 5th Panel) Number of helical residues as a function of time for one exemplary MD trajectory (all data in Fig. S6). The horizontal grey line indicates the experimental value at Φ_{TFE} = 0. 6th Panel) Helix fraction per residue from MD simulations. NCBD's profile is shown with a thin navy-blue line for reference. Horizontal bars signal the average helix length (consecutive residues with > 0.1 helix). The grey dashed line signals 60%.

Figure 4. LEGO combinations. Colors as in Figure 1. From top left to bottom right) Experimental number of helical residues of H1H2, H2H3, H3T, H2H3T as a function of ϕ_{TFE}. Error bars, curve fits, parameters, fitting errors, and dash lines as in Figure 3. The grey curves show the compounded curves of the relevant building blocks (e.g., H1 and H2 for H1H2). 5th Panel) Number of helical residues as a function of time. 6th Panel) Helix fraction per residue from MD simulations. Error bars, symbols, and lines as in Figure 3.

Figure 5. Full NCBD ensemble. Left) Experimental number of helical residues of full NCBD as a function of ϕ_{TFE}. Error bars, curve fits, parameters, fitting errors, and dash lines as in Figure 3. The grey curve shows the compounded H1, H2, H3 and T curves. Pink is H12 plus H3T and green is H1 plus H23T. Right) Helix fraction per residue (top) and number of helical residues (bottom) as a function of time from simulations. Error bars, symbols, and lines as in Figure 3, pink and green as in left panel.

Figure 6. Residue-residue interaction maps. Time averaged residue-residue contacts in the NCBD ensembles. Left) Native contacts (found by NMR). Top left triangle shows the contacts on the combined LEGO elements (local contacts in the color of the building block), and bottom right on full NCBD. Color intensity reflects contact probability in logarithmic scale: lightest shade for $10^{-4} \geq p < 10^{-3}$ to darkest for $10^{-1} \geq p < 1$. Right) total contacts observed in full NCBD parsed in two levels: dark for $10^{-1} \geq p < 1$ and light for $10^{-2} \geq p < 10^{-1}$. Diagonal red dashed lines signal a sequence separation ≤ i, i+34, equivalent to the longest-range NOE observed by NMR.