Estimation of effective concentrations enforced by complex linker architectures using conformational ensembles.

Magnus Kjaergaard[1-3*]

[1] Department of Molecular Biology and Genetics, Aarhus University

[2] The Danish Research Institute for Translational Neuroscience (DANDRITE), Nordic EMBL Partnership for Molecular Medicine

[3] Center for Proteins in Memory - PROMEMO, Danish National Research Foundation

[*] Correspondence: magnus@mbg.au.dk

Abstract:

Proteins and protein assemblies often tether interaction partners to strengthen interactions, to regulate activity through auto-inhibition or -activation, or to boost enzyme catalysis. Tethered reactions are regulated by the architecture of tether, which define an effective concentration of the interactors. Effective concentrations can be estimated theoretically for simple linkers via polymer models, but there is currently no general method for estimating effective concentrations for complex linker architectures consisting of both flexible and folded domains. We describe how effective concentrations can be estimated computationally for any protein linker architecture by defining a realistic conformational ensemble. We benchmark against prediction from a worm-like chain and values measured by competition experiments, and find minor differences likely due to excluded volume effects. Systematic variation of the properties of flexible and folded segments show that the effective concentration is mainly determined by the combination of the total length of flexible segments and the distance between termini of the folded domains. We show that a folded domain in a disordered linker can increase the effective concentration beyond what can be achieved by a fully disordered linker by focusing the end-to-end distance at the appropriate spacing. This suggest that complex linker architecture may have advantages over simple flexible linker, and emphasize that annotation as a linker should depend on the molecular context.

Introduction:

Many biochemical reactions occur internally in molecules or protein complexes including auto-inhibition, multivalent binding or tethered enzyme reactions. When the reactants are physically connected, reactions are zeroth order and thus concentration independent. Instead, reaction rates depend on how the interacting entities are connected, which enable allosteric regulation of tethered reactions via the linkers.[1] The prototypical tether is a flexible linker in the form of an intrinsically disordered region (IDR), a protein region that does have a fixed tertiary structure but is active as a disordered chain. Disordered linkers are ubiquitous in multidomain proteins, and functional annotations suggest that "linker" is the most common function for IDRs[2]. To understand structure-function relationships for linkers, it is necessary to understand how the properties of the linker determine the contact probability of the entities they link.

Molecular tethers are often more complex than a continuous flexible region. In the auto-phosphorylation of calmodulin-dependent kinase II, the catalytic domain phosphorylates a regulatory region in a neighboring subunit. The tether controlling this reaction consists of an IDR, the hub domain of the kinase, and another IDR.[3] In the bivalent interaction between the PDZ domains of PSD-95 and ionotropic glutamate receptors, the tether consists of the disordered C-terminal domain of one receptor subunit, the ion channel domain, and the disordered C-terminal tail of another receptor subunit.[4] The connections that tether biochemical reactions thus vary beyond what is typically considered linkers and may include domains that have other functions. In the following, the term "linker architecture" will be used to describe complex connections that may consist of several rigid and flexible segments, and the term "linker" will be reserved for a single continuous IDR.

Linker architectures enforce of an effective concentration (C_{eff}) of the tethered ligand, which corresponds to the concentration of free ligand that has the same encounter rate as the tethered ligand. When C_{eff} is known, it is possible to describe a tethered system via equilibrium and rate constants of the corresponding untethered system. Quantitative models of C_{eff} has been used to describe processes such as auto-inhibition,[5,6] avidity of multivalent interactions,[7–13] phase separation[14] and intra-complex enzyme catalysis.[15] In principle, C_{eff} can be measured directly in competition experiments, where a free ligand displaces the tethered ligand.[16,17] Such competition experiments require an experimental read-out of intramolecular binding for

example via fluorescent competitor[16] or a designed FRET biosensor.[17,18] Direct measurement through competition experiments also requires that the system can be reconstituted *in vitro,* and that the tethered ligand can be fully displaced, which typically requires millimolar concentrations of competitor. Therefore, experimental determination of C_{eff} is often impractical, and in many cases it is attractive to model C_{eff} theoretically.[7,8,13,19–21]

C_{eff} is determined by the end-to-end distance distribution of the linker, which defines the probability of a tethered ligand contacting a binding site at a certain spacing from the origin of the linker. Fully disordered linkers can often be modelled well by homopolymer models.[22–24] Though homopolymer models ignore many aspects of IDRs they often provide a good first approximation of an IDR, and models such as for example the worm-like chains (WLC) are currently widely used to predict C_{eff}. C_{eff} is defined by the continuous end-to-end distance probability as follows:[7,25]

$$C_{eff}(r) = \frac{p(r)}{4\pi r^2} \frac{10^{27}\ \text{Å}^3 L^{-1}}{N_A} \qquad (1)$$

Where the first term divides the probability distribution by the surface area of a sphere of radius *r*, and the second term converts to molar concentration. As the WLC can be expressed purely mathematically, C_{eff} enforced by flexible linkers can calculated rapidly for example by the interactive web server "C_{eff} calculator".[26] Purely mathematical descriptions, however, struggle with complex linker architectures composed of e.g. both folded and disordered regions. Simplifying assumptions are thus necessary to handle complex linker architectures. For example, it has been suggested that a folded domain in a WLC can be approximated by rigid spacer with a length corresponding to the distance between the attachment sites irrespective of its position in the linker.[21] Such descriptions do not account for sequence specific contacts or excluded volume, and are difficult to validate without direct experimental comparison. This suggests a potential for improvement of the theoretical prediction of C_{eff} enforced by complex linker architectures.

This work seeks to estimate C_{eff} for complex linker architectures composed of disordered and folded segments. Like the WLC, we aim for a simple method that describes the general geometric principles of a complex linker architectures, but do not attempt to account for sequence specific interactions. C_{eff} can be calculated at different spacings for any linker

architecture where we can define a valid end-to-end distribution, although this has not been used to study protein linkers to the best of our knowledge. We thus aim to simulate a realistic structural ensemble for the linker architecture, which allows determination of the distribution of end-to-end distances and C_{eff} as a function of spacing. Results are compared to results from competition experiments and predictions from the WLC, and subsequently we explore how C_{eff} depends on the properties of folded and flexible segments in complex linker architectures.

Materials and methods:

Ensemble generation: Coarse-grained ensembles that represent each residues as a bead were generated using the RANCH routine of EOM2.0.[27,28] Unless indicated otherwise, the "native chain" setting was used for flexible segments, and the ensemble size was set at 10.000. Input sequences were generated by adding flanking GS-repeats to the sequences of the folded domains. The second disordered segment either started with G or S as needed to disambiguate pseudo-atoms. For the linkers used in the biosensors,[17] the residues corresponding to the flanking restriction sites were also included. Poly-alanine α-helices were built using PyMOL (v. 1.8.0.5, Schrödinger Inc.) with lengths ranging from 7 to 77 residues. Folded domains were all crystal structures with a single chain without missing residues. For models with several copies defined in the asymmetric unit, the first model was selected. The sequences used by EOM were truncated to the region defined in the PDB model, and co-factors were removed. The structured domain were: 1IIU,[29] 1MPD,[30] 1N2A,[31] 1TM1[32] (chains E and I separately), 1W6X,[33] 2GI9,[34] 2IC2, 2R4U,[35] 3D7S,[36] 3GV9,[37] 3H1C,[38] 3UC7,[39] 3MYC,[40] 4F5S,[41] 4KDS,[42] 4O75,[43] 5D13, 5DFT, 5EJ6.[44] The r_{NC} distance was measured between the first and last C_{α}–atoms using PyMOL. For the dimeric α-actinin (PDB:1HCI),[45] the disordered segments were attached to the C-terminus of the folded domain, and a distances were measured between the two C-termini.

Determination of C_{eff}: The distance distribution was measured between the first and last bead in each model using a PyMOL script. The resulting list of distances was imported into Rstudio, and a histogram of end-to-end distances was generated in 1 Å bins. The radial probability distribution p(r) is defined as the fraction of conformers with end-to-end distances within 0.5 Å of the bin center. For each bin, p(r) and r were inserted into (Eq. 2) to calculate C_{eff}. For the WLC, the p(r) function was predicted using the "C_{eff} calculator" web applet (http://ceffapp.chemeslab.org)[26] and C_{eff} was calculated for each point using (Eq. 1).

Results:

We aimed to develop a simple strategy for estimating C_{eff} enforced by complex linker architectures such as mixtures of folded and disordered segments. We considered the case where a tethered ligand binds a site located at a fixed spacing from the origin of the linker (Fig. 1A). An analogous situation occurs for bivalent interactions after binding of the first ligand, which has been the main focus of previous studies of linker architecture due to its importance in drug development. We made the simplifying assumption that the linker architecture can be divided into flexible linkers and rigid domains (Fig 1B). Ensembles were generated by the random chain algorithm of the Ensemble Optimization Method (EOM)[27,28], which models flexible linkers as self-avoiding chains sampling backbone dihedral angles from one of three different coil libraries (Fig. 1C). End-to-end distances were extracted from the ensemble and counted in 1 Å bins, which can be visualized as a probability density distribution histogram (Fig. 1D). For each combination of r and $p(r)$, C_{eff} can be calculated as described below, and visualized as C_{eff} at different spacings from the attachment site (Fig. 1E), which will hereafter be referred to as the C_{eff} profile.

Calculation of C_{eff} *from a conformational ensemble*

We would like to be able to estimate C_{eff} from a conformational ensemble. Eq. 1 was developed for mathematically polymer models, where the distances are precise and continuous. In a discrete conformational ensemble, no conformers will have an end-to-end distance of e.g. precisely 42 Å. Instead, it is practical to consider the probability distribution of end-to-end distances in bins of a certain size such that all conformers with an end-to-end distance between 41.5 and 42.5 Å are counted in the 42 Å bin. Analogous to Eq. 1, define C_{eff} as the ratio of the molar concentration of tethered ligand in each bin, and the volume of the bin:

$$C_{eff} = \frac{p(r)_{bin}/N_A}{V(r)_{bin}}$$

For a 1 Å bin centered at 42 Å, $p(r)_{bin}$ is simply the fraction of conformers where the end-to-end distance falls between 41.5 and 42.5 Å. $V(r)_{bin}$ is the difference in volume between sphere of

radii 42.5 Å and 41.5 Å. Eq. 2 thus estimate C_{eff} from a discrete conformational ensemble for a generalized bin width of *bw* including a term to convert to units of from $Å^3$ to L:

$$C_{eff} = \frac{p(r)_{bin}}{(r+\frac{1}{2}bw)^3-(r-\frac{1}{2}bw)^3} \times \frac{3 \times 10^{27}\ Å^3/_L}{4\pi\ N_A} \qquad (2)$$

Comparison to WLC. We compared the prediction from the EOM ensemble to the worm-like chain for a fully disordered chain. We created ensembles of 10.000 models for $(GS)_{50}$ using the three different coil libraries built into EOM: "Compact", "Native" and "Random". Each ensemble can be generated in minutes on a personal computer. The probability distributions were fitted to the WLC model resulting in persistence lengths of 5.2, 6.4 and 7.6 Å, respectively (Fig. 2A-C). The ensemble-derived C_{eff} profile was noisy at low spacings at an ensemble size of 10.000, which is likely to be inherent to C_{eff} determination from ensembles: The close spacings are only represented by few conformers in the ensemble, and values vary stochastically. The noise can be reduced by increasing the ensemble size although at the expense of computational cost, or by increasing the bin size at the expense of spatial resolution. We judged that the curve defined the shape of the C_{eff} profile well enough at this ensemble size and used it for future simulations. C_{eff} profiles calculated from ensembles and the WLC agreed at spacings above ~20 Å (Fig. 2D-F). For shorter spacings, however, the ensemble derived C_{eff} peaked at what appears as a broad plateau before decreasing when it approaches zero. The C_{eff} calculated from the WLC increased monotonously to a maximum at zero as shown previously.[25] Due to the division by r^2 in Eq. (1), the C_{eff} profile magnifies differences at small spacings that are easily overlooked in the distance distribution. The different C_{eff} profile is likely due to differences in how the models encode self-avoidance. The close spacings report on the subset of the ensemble where the ends of the linkers are close to each other, which increase the importance of steric occlusion. In total, the ensemble derived C_{eff} values quantitative agree with the WLC, which accurately model many disordered regions.

Comparison of length scaling and experimental values. Next, we wanted to compare the ensemble C_{eff} to values determined from competition experiments using a FRET biosensor.[17] Experimental C_{eff} values scaled with linker length according to a polymer scaling law, where the scaling exponent depended on the physical properties of the linkers.[17] EOM does not account for electrostatics which is the dominant factor in IDP compaction, so we only compared the

ensemble values to biosensor containing $(GS)_x$-linkers. For each $(GS)_x$-linker, we generated 5 ensembles of 10.000 structures and calculated the resulting C_{eff} at a spacing of 10 Å (Fig. 3), which is the distance between linker attachment sites in the complex used in the biosensor.[46] The spacing of 10 Å is near the noisy maxima in the C_{eff} profile. While the ensemble size defines the shape of the C_{eff} profile, individual bins had highly variable C_{eff} values, and we thus averaged over 3 Å. The competition experiment produced values that are consistently 2-3-fold lower than predicted by the ensemble and the WLC. It is unclear whether this represents a systematic error in the prediction or experiment (see discussion).

The predicted C_{eff} depends on linker length via a power law as indicated by the linear relation in the double logarithmic plot (Fig. 3B) in agreement with both experiments and predictions from WLC. Ensemble C_{eff} values decreased more rapidly with linker length than both the experiments the WLC (Fig. 3B). When fitted similar to the values obtained here experimental C_{eff} for GS-linkers have a scaling exponent of -1.53, which correspond to a scaling exponents for chain length of $\nu = 0.51$. The slight difference from the value reported previously[17] depends on whether you include the 4 residues from the restriction sites in the linker length. The corresponding values are -1.41 ($\nu = 0.47$) for the WLC ($L_p = 6.4$ Å) and -1.83 ($\nu = 0.61$) for the ensemble. The spacing from the biosensor (10 Å) corresponds to region where a systematic deviation is seen between the EOM ensemble and the WLC (Fig. 2), and the difference in scaling exponent is thus expected. Consistent with excluded volume causing this deviation, the scaling coefficients from the ensemble depended on the spacing, and a scaling exponent of -1.63 ($\nu = 0.54$) was observed a spacing of 20 Å. When restricted to an end-to-end distance of 10 Å, excluded volume effects will likely increase for longer chains, which both explains why scaling exponent depends on the spacing and why predictions differ between the WLC and the ensemble. Other types of biophysical experiments show that IDRs have average ν of ~0.50–0.53[24,47,48], but with a lot of individual variation as highly charged IDRs can have ν up to 0.7.[23] The scaling of C_{eff} values from the EOM ensemble with linker length thus corresponds approximately to that expected for an IDR of average compaction, and is thus appropriate for a first order approximation of C_{eff}.

Folded domains in disordered linkers. Having validated the ensemble method, we wanted to investigate linker architectures that are difficult to describe mathematically. The simplest and

most general type of complex linker architecture is a folded domain flanked by two flexible linkers. Mathematical treatment suggests the length of the two disordered segments can simply be added together and considered as a single long linker.[21] To test this assertion, we generated ensembles of disordered linkers with three different folded domains embedded: A SH3 domain, a PDZ domain and a FN3 domain, which are typical small globular domains found in multidomain proteins. We tested the distance distribution and C_{eff} profile with a total of 100 disordered residues (Fig. 4A-F), where the flexible segments sample a much larger volume than the folded domains, and 30 residues where they sample a volume similar to the dimensions of the folded domain (Fig. 4G-L).

The ensembles suggest that different combinations of folded and disordered segments have similar C_{eff} profiles if the total length of the disordered segments is identical. Fig. 4 compares end-to-end distribution for a folded domain surrounded by flexible linkers with similar length of flexible segments, but distributed differently. Generally, the predicted C_{eff} is slightly higher when the linker is distributed equally around the folded domain corresponding to a higher sampling of close contacts. The likely cause of this deviation is the excluded volume effect of the folded domain, which affects the proximal region of the linkers most strongly. The deviation in the C_{eff} is largest for the linker architecture with short flexible segments surrounding the SH3 domain. The SH3 has a short distance between the N- and C-termini (7.1 Å), which likely enhance their excluded volume effects of the linkers. However, these deviations are generally minor, and for most practical purposes the end-to-end distance distribution and C_{eff} profile depend mostly on the total length of flexible segments, and much less on how they distribute around the folded domain.

Is it valid to approximate folded domains with a stiff rod? Next, we sought to test how a folded domain inserted into a long linker affects C_{eff}. Geometric considerations suggest that a folded domain has two effects: It act as a rigid spacer within the linker, and it adds an excluded volume the flexible segments cannot enter. Previous work suggest that the rigid spacer effect is the dominant effect and that the domain can often be approximated by a stiff rod with a length corresponding to the distance between the N- and C-termini of the domain (r_{NC}).[21] This simplification ignores other effects of the shape and size of the folded domain. To test the rod approximation, we inserted 20 folded domains (Fig. 5A) that are roughly globular and span a range of sizes and r_{NC} values (Fig. 5B) in the middle of 100-residue disordered linker. To

approximate a rod, we built ideal poly-alanine α-helices with r_{NC} in the range from 10-100 Å, and generated corresponding ensembles for each with rod in the middle of a 100-residue GS-linker. The mean end-to-end distance from the globular domains largely follows the same curve as the helical rods (Fig. 5C) although some of the large domains occupy a slightly more expanded ensemble likely due to the excluded volume of the domain. C_{eff} is calculated at spacings just below (50 Å) and above (100 Å) the mean end-to-end distance in the full disordered linker (Fig. 5DE). C_{eff} follows the same trends for the folded domains as the helical rods with a monotonous decrease with increasing r_{NC} at a spacing of 50 Å (Fig. 5D), a maximum around r_{NC} of 60 Å for a spacing of 100 Å (Fig. 5E). This suggests that in most cases, the effects of a folded domain in a linker can be approximated by a rigid rod and that the bulk of the domain plays subsidiary role. Furthermore, it also shows that a domain of a fixed size will thus have a different impact, depending on where the flexible linkers are attached: If the termini are on the same side of the domain the effect will be much smaller than if they are located on the opposite side of a folded domain.

The effect of r_{NC}. Having validated the rod assumption, we further examine the end-to-end distance distribution and C_{eff} profile of linker architectures with rods of increasing r_{NC} (Fig. 6A). When r_{NC} is small, the ensemble dimensions are dominated by the disordered chains, but with increasing r_{NC} the dimensions of the rod start to dominate. The increase in end-to-end distance thus accelerate as the length of the rod increases in agreement with (Fig. 5C). These changes mirror a gradual change in the shape of the C_{eff} scaling profile with insertion of rods of increasing r_{NC} (Fig. 6B). For short rods, C_{eff} is still highest at small spacings, whereas the profile gradually changes to have a non-zero maximum for longer rods. The distance distribution translates into different dependence of C_{eff} on the rod length depending on the spacing: At spacings below the average end-to-end distance of the linker without rods, C_{eff} decreased monotonously with rod length (Fig. 5B). Whereas at long spacings (> ~100 Å), C_{eff} increased with rod length for the range tested here, although this trend will reverse when the rod is much longer than the spacing.

Optimal linker architectures for large spacings. Linkers are usually associated with IDRs, but fully disordered linkers span large distances inefficiently. For fully disordered linkers, the maximal C_{eff} that can be achieved rapidly drops off with increasing spacing (Fig. 3). The highest achievable C_{eff} by a disordered linker of optimal length can be calculated using the WLC

implemented in "C_{eff} calculator".[26] Using a L_p of 6.4 Å to match the "native chain" output of EOM (Fig. 2), the highest achievable C_{eff} at spacing of 25 Å is 9.6 mM for a 10-residue linker and 123 μM at 100 Å for a 200-residue linker. To explore whether insertion of a folded domain can increases the maximal C_{eff}, we focused on a binding site separated by a spacing of 100 Å. Linker ensembles were generated with flexible segments flanking α-helical rod corresponding to r_{NC} values of a small (Fig. 7A, 21 residues, r_{NC} = 27.1Å), a medium-sized (Fig. 7B, 42 residues, r_{NC} = 54.9 Å) and a large domain, where the latter has a r_{NC} that approximates the spacing (Fig. 7C, 70 residues, r_{NC} = 92.3 Å). The longest flexible segments considered (200 residues) match the optimal flexible linker, and was gradually shortened to reveal the dependence on the length of the flexible segments for each rod (Fig. 7D). We benchmark these values against a fully disordered linker of optimal length (200 residues), which was predicted to have a C_{eff} of 147 μM from ensemble at this spacing in reasonable agreement with the WLC (123 μM).

For the short rod, the C_{eff} profile with 200 flexible residues is broadly similar to the fully flexible linker, and shortening of the flexible segments monotonously decrease C_{eff} without exceeding the benchmark (Fig 7D). For the medium length rod, the benchmark C_{eff} was exceeded slightly when the rod is coupled to 100 flexible residues. However, C_{eff} was similar to the benchmark for range of flexible segments, and did not drop below the benchmark until the linker was shortened to 40 residues (Fig 7D).

The linker architectures containing the longest rod start at lower C_{eff} as it is too extended for the spacing. Shortening of the flexible segments increase C_{eff} such that it exceeds the benchmark for fewer than ~100 flexible residues and increases noticeably beyond the benchmark for short flexible segments. The origin of this increase can be visualized in the C_{eff} profile (Fig. 7C). The fully disordered linker spans the large spacing inefficiently because it samples the entire volume between the origin and the binding site. The rigid rod in the linker reduce the sampling of short spacings, and when r_{NC} is matched to the spacing it can increase the highest C_{eff} than can be attained. The inclusion of a folded domain into the linker architecture decouples the spacing that needs to be spanned from the volume that is sampled. When r_{NC} is matched to the spacing, shortening of the flanking linkers focus the ligand on the binding site.

For globular proteins, the combinations of a long r_{NC} and short linkers will result in steric clashes not found in the helical model. This suggests that the optimal linker architecture to span

large distances involves elongated, rather than globular, folded domains. To test the putative C_{eff} enhancement of complex linker architectures in a more realistic case, we studied linker architectures containing the rod domain from the scaffolding protein α-actinin. The rod domain is ~250 Å long with an anti-parallel heterodimeric coiled-coil architecture (Fig. 8A).[45] Linker architectures were modelled by inserting disordered segments of variable length at the two C-termini that are separated by 245 Å. Similar to the α-helical model system, these linker architectures focus the p(r) and C_{eff} profile at spacings similar to the length of the domain (Fig. 8BC). For comparison at a spacing of 250 Å, the ideal fully disordered linker of 1283 residues achieves a C_{eff} of ~7 μM. This C_{eff} is exceeded by a factor of two by α-actinin linkers with a total of 200 residues (Fig. 8C), and rises further as the linkers shorten as above. The model in Fig. 8A suggests that this linker architecture would be sterically able to bridge binding sites in most other proteins or assemblies with this spacing. This suggests that the C_{eff} enhancement from inserting folded domains into disordered linkers will be relevant in many biochemical contexts.

Discussion:

Here we have described a general method for predicting C_{eff} enforced by protein linkers from conformational ensembles. Alternatively, such effective concentrations could be calculated from polymer models such like the WLC or a measured directly through competition experiments. The method presented here is more cumbersome than the purely mathematical ploymer models, but is much faster than competition experiments. Where the WLC only describes fully disordered linkers, the ensemble method can determine C_{eff} from complex linker architectures composed of any combination of flexible and folded segments, and thus expand the scope of linker architectures that can be studied.

C_{eff} values from ensembles generated with EOM largely match those predicted by the WLC, suggesting that these methods are comparable for fully disordered linkers that can be described by a homopolymer model. An exception occurs at small spacings (Fig. 2D-F), where the C_{eff} from ensemble simulations are systematically lower, and the C_{eff} profile has a different shape with a pronounced plateauing. This effect is probably due to the excluded volume, which mainly affects the subset of the ensemble where the ends are close. Since the WLC does not account for excluded volume, the values from the ensemble are likely more accurate.

Both the ensemble and WLC predict C_{eff} values that are consistently ~3-fold higher than those measured by competition experiments using a biosensor (Fig. 3).[17] It is not clear, whether this reflects a real difference between the experiments and the simulations or a systematic error. The experimental C_{eff} values are derived from a very different approach, and it would be surprising if they were fully comparable to theoretical estimations. For example, the biosensor contains both the interaction domains and a pair of fluorescent proteins tethered at the end of the linkers. Tethered folded domains increase the excluded volume further,[49] which might lead to systematically lower C_{eff} values than in the simulations. Furthermore, any structural changes in the binding partners upon binding may change the effective length of the flexible segments and r_{NC} as seen for a two-component biosensor using a different interaction pair, where a different scaling exponent was found for the same linker composition.[9] Finally, the theoretical estimations assume precisely defined beginnings of the flexible linkers and distances to the binding site. In practice, such estimations will necessarily be approximations as there is a gradual transition from rigid to flexible, and proteins change structure upon binding. Therefore, systematic deviations between experiment and the theoretically predicted values are not surprising and the experiments should thus not necessarily be regarded as a gold standard.

The ensemble used to calculate C_{eff} could be generated in many different ways. The current implementation using EOM has the advantage of simplicity and only relies on open tools. The limitation of this method is that it does not account for sequence-specific interactions, sequence-specific sampling of backbone dihedral angles, or interactions between flexible segments and the folded domains. This is acceptable for a study of the general trends in interplay between rigid and flexible segments of linker architectures, but suggests potential future improvements. At the cost of increased computational complexity, sequence-specific behavior could be improved by sampling backbone dihedral angles based on values obtained for e.g. tripeptides.[50]

An alternative strategy would be to determine C_{eff} from molecular dynamics simulation of the linker architecture. Simulations of intrinsically disordered regions can capture many sequence specific interactions in the IDRs[51] and between IDRs and folded domains,[49] however at significantly increased complexity compare to the approach presented here. Recently, C_{eff} values enforced by a family of disordered linkers were predicted based from simulations.[12] This study used the average end-to-end distance in the simulation to derive a sequence-specific

persistence length, that could then be used to calculate C_{eff} via the WLC.[52] This accounts for sequence-specific interactions in the flexible linkers and was critical for developing the idea of "conformational buffering" in linker sequences. Based on the results obtained here, however, it is likely that C_{eff} can be calculated in a model-free way from the end-to-end distance distribution from of the simulation trajectory rather than going via the WLC.

The accuracy of the C_{eff} determination would be improved if the ensemble was selected or validated using experimental data. The most informative types of experiments include NMR,[53,54] small angle X-ray scattering,[47] single-molecule FRET[22] or any combination thereof.[55,56] EOM was made for analysis of SAXS data,[27] and was built on Flexible-Mecanno developed for NMR data.[57] The ensemble generation step can thus directly incorporate refinement of the ensemble using experimental data, but this was not be pursued here. Experimental ensembles vary widely in size, but are often much smaller than the ~10.000 conformations needed to construct a reliable C_{eff} profile. Increasing numbers of such experimentally validated examples are available in protein ensemble database,[58] but most have much fewer member. This suggests that for experimental studies aiming at describing tethered reactions a large ensemble should be sought.

Many intra-molecular or intra-complex reactions are tethered by more complex connections than a simple flexible linker. Nevertheless, the term "linker" is often regarded as synonymous with a continuous IDR. We have shown that for the purpose of linking interacting entities, folded domains are easily tolerated in linker architectures, and may even be beneficial compared to a fully disordered linker. This suggests that linkers architectures with the folded domain embedded could be selected for even in the absence of any other function of the folded domain. This broadens the concept of what can be considered a linker, and complicates functional annotations of linker which typically presume disorder. The semantic difficulties can be solved by introduction of the concept of linker architecture as a functional designation based on the context in multidomain proteins or complexes rather than an intrinsic property of the sequence.

The rise of multivalent protein drugs such as bivalent antibodies or engineered biosensors[59] suggest that optimizing linker properties have direct practical and clinical applications. Long disordered linkers sample a large volume, which is disadvantageous when the binding site is

placed at a specific spacing. The combination of an elongated folded domain with short flexible linkers thus provide the optimal linker architecture. Similar results were obtained in simulations of a rigid DNA molecule decorated with flexible polyethylene glycol linkers.[20] The C_{eff} calculated for the linker architectures with a r_{NC} similar to the spacing increases as the linkers shorten, suggesting that the disordered segments are superfluous. This likely demonstrate the limitation of the simplified model used here as the geometry surrounding the binding site is not considered. The IDRs still play a crucial role by providing the conformational flexibility needed to form complexes with minimal physical strain. As illustrated by the case study of the rod domain from α-actinin, coiled-coil domains are an ideal fold to provide a large spacing with a minimum of steric bulk. This may thus explain why long coiled-coil domains are so common in scaffolding proteins.[60]

Acknowledgements:

This work was supported by grants from the Villum Foundation, the Lundbeck Foundation, the Novo Nordisk Foundation and PROMEMO – Center for Proteins in Memory, a Center of Excellence funded by the Danish National Research Foundation (grant number DNRF133). The author thanks Lucia B. Chemes, Nicolas S. González Foutel and Nathalie Wyss for critical feedback on the manuscript.

References:

(1) Huang, Q.; Li, M.; Lai, L.; Liu, Z. Allostery of Multidomain Proteins with Disordered Linkers. *Current Opinion in Structural Biology* **2020**, *62*, 175–182. https://doi.org/10.1016/j.sbi.2020.01.017.

(2) Hatos, A.; Hajdu-Soltész, B.; Monzon, A. M.; Palopoli, N.; Álvarez, L.; Aykac-Fas, B.; Bassot, C.; Benítez, G. I.; Bevilacqua, M.; Chasapi, A.; Chemes, L.; Davey, N. E.; Davidović, R.; Dunker, A. K.; Elofsson, A.; Gobeill, J.; Foutel, N. S. G.; Sudha, G.; Guharoy, M.; Horvath, T.; Iglesias, V.; Kajava, A. V.; Kovacs, O. P.; Lamb, J.; Lambrughi, M.; Lazar, T.; Leclercq, J. Y.; Leonardi, E.; Macedo-Ribeiro, S.; Macossay-Castillo, M.; Maiani, E.; Manso, J. A.; Marino-Buslje, C.; Martínez-Pérez, E.; Mészáros, B.; Mičetić, I.; Minervini, G.; Murvai, N.; Necci, M.; Ouzounis, C. A.; Pajkos, M.; Paladin, L.; Pancsa, R.; Papaleo, E.; Parisi, G.; Pasche, E.; Barbosa Pereira, P. J.; Promponas, V. J.; Pujols, J.; Quaglia, F.; Ruch, P.; Salvatore, M.; Schad, E.; Szabo, B.; Szaniszló, T.; Tamana, S.; Tantos, A.; Veljkovic, N.; Ventura, S.; Vranken, W.; Dosztányi, Z.; Tompa, P.; Tosatto, S. C. E.; Piovesan, D. DisProt: Intrinsic Protein Disorder Annotation in 2020. *Nucleic Acids Research* **2019**, *48* (D1), D269–D276. https://doi.org/10.1093/nar/gkz975.

(3) Buonarati, O. R.; Miller, A. P.; Coultrap, S. J.; Bayer, K. U.; Reichow, S. L. *Conserved and Divergent Features of Neuronal CaMKII Holoenzyme Structure, Function, and High-Order Assembly*; preprint; Molecular Biology, 2021. https://doi.org/10.1101/2021.01.21.427643.

(4) Christensen, N. R.; Pedersen, C. P.; Sereikaite, V.; Pedersen, J. N.; Vistrup-Parry, M.; Sørensen, A. T.; Otzen, D.; Arleth, L.; Teilum, K.; Madsen, K. L.; Strømgaard, K. Bi-Directional Protein-Protein

Interactions Control Liquid-Liquid Phase Separation of PSD-95 and Its Interaction Partners. *bioRxiv* **2021**, 2021.03.03.433781. https://doi.org/10.1101/2021.03.03.433781.

(5) Borcherds, W.; Becker, A.; Chen, L.; Chen, J.; Chemes, L. B.; Daughdrill, G. W. Optimal Affinity Enhancement by a Conserved Flexible Linker Controls P53 Mimicry in MdmX. *Biophysical Journal* **2017**, *112* (10), 2038–2042. https://doi.org/10.1016/j.bpj.2017.04.017.

(6) Zhou, H.-X. How Often Does the Myristoylated N-Terminal Latch of c-Abl Come Off? *FEBS Letters* **2003**, *552* (2–3), 160–162. https://doi.org/10.1016/S0014-5793(03)00911-6.

(7) Zhou, H.-X. Quantitative Account of the Enhanced Affinity of Two Linked ScFvs Specific for Different Epitopes on the Same Antigen. *Journal of Molecular Biology* **2003**, *329* (1), 1–8. https://doi.org/10.1016/S0022-2836(03)00372-3.

(8) Zhou, H.-X. The Affinity-Enhancing Roles of Flexible Linkers in Two-Domain DNA-Binding Proteins. *Biochemistry* **2001**, *40* (50), 15069–15073. https://doi.org/10.1021/bi015795g.

(9) Sørensen, C. S.; Jendroszek, A.; Kjaergaard, M. Linker Dependence of Avidity in Multivalent Interactions Between Disordered Proteins. *Journal of Molecular Biology* **2019**, *431* (24), 4784–4795. https://doi.org/10.1016/j.jmb.2019.09.001.

(10) Mack, E. T.; Snyder, P. W.; Perez-Castillejos, R.; Whitesides, G. M. Using Covalent Dimers of Human Carbonic Anhydrase II To Model Bivalency in Immunoglobulins. *Journal of the American Chemical Society* **2011**, *133* (30), 11701–11715. https://doi.org/10.1021/ja2038084.

(11) Mack, E. T.; Snyder, P. W.; Perez-Castillejos, R.; Bilgiçer, B.; Moustakas, D. T.; Butte, M. J.; Whitesides, G. M. Dependence of Avidity on Linker Length for a Bivalent Ligand–Bivalent Receptor Model System. *Journal of the American Chemical Society* **2012**, *134* (1), 333–345. https://doi.org/10.1021/ja2073033.

(12) Gonzalez-Foutel, N. S.; Borcherds, W. M.; Glavina, J.; Barrera-Vilarmau, S.; Sagar, A.; Estaña, A.; Barozet, 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*; bioRxiv; preprint; 2021. https://doi.org/10.1101/2021.05.14.444182.

(13) Diestler, D. J.; Knapp, E. W. Statistical Thermodynamics of the Stability of Multivalent Ligand-Receptor Complexes. *Physical Review Letters* **2008**, *100* (17). https://doi.org/10.1103/PhysRevLett.100.178101.

(14) Harmon, T. S.; Holehouse, A. S.; Rosen, M. K.; Pappu, R. V. Intrinsically Disordered Linkers Determine the Interplay between Phase Separation and Gelation in Multivalent Proteins. *eLife* **2017**, *6*, e30294.

(15) Dyla, M.; Kjaergaard, M. Intrinsically Disordered Linkers Control Tethered Kinases via Effective Concentration. *Proceedings of the National Academy of Sciences* **2020**, *117* (35), 21413–21419. https://doi.org/10.1073/pnas.2006382117.

(16) Krishnamurthy, V. M.; Semetey, V.; Bracher, P. J.; Shen, N.; Whitesides, G. M. Dependence of Effective Molarity on Linker Length for an Intramolecular Protein−Ligand System. *Journal of the American Chemical Society* **2007**, *129* (5), 1312–1320. https://doi.org/10.1021/ja066780e.

(17) Sørensen, C. S.; Kjaergaard, M. Effective Concentrations Enforced by Intrinsically Disordered Linkers Are Governed by Polymer Physics. *PNAS* **2019**, *116* (46), 23124–23131. https://doi.org/10.1101/577536.

(18) Sørensen, C. S.; Kjaergaard, M. Measuring Effective Concentrations Enforced by Intrinsically Disordered Linkers. In *Intrinsically Disordered Proteins: Methods and Protocols*; Kragelund, B. B., Skriver, K., Eds.; Springer US: New York, NY, 2020; pp 505–518. https://doi.org/10.1007/978-1-0716-0524-0_25.

(19) Diestler, D. J.; Knapp, E. W. Statistical Mechanics of the Stability of Multivalent Ligand−Receptor Complexes. *The Journal of Physical Chemistry C* **2010**, *114* (12), 5287–5304. https://doi.org/10.1021/jp904258c.

(20) Liese, S.; Netz, R. R. Influence of Length and Flexibility of Spacers on the Binding Affinity of Divalent Ligands. *Beilstein Journal of Organic Chemistry* **2015**, *11*, 804–816. https://doi.org/10.3762/bjoc.11.90.
(21) Van Valen, D.; Haataja, M.; Phillips, R. Biochemistry on a Leash: The Roles of Tether Length and Geometry in Signal Integration Proteins. *Biophysical Journal* **2009**, *96* (4), 1275–1292. https://doi.org/10.1016/j.bpj.2008.10.052.
(22) Schuler, B.; Soranno, A.; Hofmann, H.; Nettels, D. Single-Molecule FRET Spectroscopy and the Polymer Physics of Unfolded and Intrinsically Disordered Proteins. *Annual Review of Biophysics* **2016**, *45* (1), 207–231. https://doi.org/10.1146/annurev-biophys-062215-010915.
(23) Hofmann, H.; Soranno, A.; Borgia, A.; Gast, K.; Nettels, D.; Schuler, B. Polymer Scaling Laws of Unfolded and Intrinsically Disordered Proteins Quantified with Single-Molecule Spectroscopy. *Proceedings of the National Academy of Sciences* **2012**, *109* (40), 16155–16160. https://doi.org/10.1073/pnas.1207719109.
(24) Marsh, J. A.; Forman-Kay, J. D. Sequence Determinants of Compaction in Intrinsically Disordered Proteins. *Biophysical Journal* **2010**, *98* (10), 2383–2390. https://doi.org/10.1016/j.bpj.2010.02.006.
(25) Bertagna, A.; Toptygin, D.; Brand, L.; Barrick, D. The Effects of Conformational Heterogeneity on the Binding of the Notch Intracellular Domain to Effector Proteins: A Case of Biologically Tuned Disorder. *Biochemical Society Transactions* **2008**, *36* (2), 157–166. https://doi.org/10.1042/BST0360157.
(26) Kjaergaard, M.; Glavina, J.; Chemes, L. B. Predicting the Effect of Disordered Linkers on Effective Concentrations and Avidity with the "Ceff Calculator" App. In *Methods in Enzymology*; 2021; Vol. 647, pp 145–171. https://doi.org/10.1016/bs.mie.2020.09.012.
(27) Bernadó, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I. Structural Characterization of Flexible Proteins Using Small-Angle X-Ray Scattering. *J. Am. Chem. Soc.* **2007**, *129* (17), 5656–5664. https://doi.org/10.1021/ja069124n.
(28) Tria, G.; Mertens, H. D. T.; Kachala, M.; Svergun, D. I. Advanced Ensemble Modelling of Flexible Macromolecules Using X-Ray Solution Scattering. *IUCrJ* **2015**, *2* (2), 207–217. https://doi.org/10.1107/S205225251500202X.
(29) Zanotti, G.; Calderone, V.; Beda, M.; Malpeli, G.; Folli, C.; Berni, R. Structure of Chicken Plasma Retinol-Binding Protein. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* **2001**, *1550* (1), 64–69. https://doi.org/10.1016/S0167-4838(01)00268-0.
(30) Shilton, B. H.; Shuman, H. A.; Mowbray, S. L. Crystal Structures and Solution Conformations of a Dominant-Negative Mutant OfEscherichia ColiMaltose-Binding Protein. *Journal of Molecular Biology* **1996**, *264* (2), 364–376. https://doi.org/10.1006/jmbi.1996.0646.
(31) Rife, C. L.; Parsons, J. F.; Xiao, G.; Gilliland, G. L.; Armstrong, R. N. Conserved Structural Elements in Glutathione Transferase Homologues Encoded in the Genome of Escherichia Coli. *Proteins: Structure, Function, and Bioinformatics* **2003**, *53* (4), 777–782. https://doi.org/10.1002/prot.10452.
(32) Radisky, E. S.; Kwan, G.; Karen Lu, C.-J.; Koshland, D. E. Binding, Proteolytic, and Crystallographic Analyses of Mutations at the Protease−Inhibitor Interface of the Subtilisin BPN'/Chymotrypsin Inhibitor 2 Complex,. *Biochemistry* **2004**, *43* (43), 13648–13656. https://doi.org/10.1021/bi048797k.
(33) Massenet, C.; Chenavas, S.; Cohen-Addad, C.; Dagher, M.-C.; Brandolin, G.; Pebay-Peyroula, E.; Fieschi, F. Effects of P47phox C Terminus Phosphorylations on Binding Interactions with P40phox and P67phox: STRUCTURAL AND FUNCTIONAL COMPARISON OF P40phox and P67phox SH3 DOMAINS* ♦. *Journal of Biological Chemistry* **2005**, *280* (14), 13752–13761. https://doi.org/10.1074/jbc.M412897200.
(34) Franks, W. T.; Wylie, B. J.; Stellfox, S. A.; Rienstra, C. M. Backbone Conformational Constraints in a Microcrystalline U-15N-Labeled Protein by 3D Dipolar-Shift Solid-State NMR Spectroscopy. *J. Am. Chem. Soc.* **2006**, *128* (10), 3154–3155. https://doi.org/10.1021/ja058292x.

(35) Sheng, F.; Jia, X.; Yep, A.; Preiss, J.; Geiger, J. H. The Crystal Structures of the Open and Catalytically Competent Closed Conformation of Escherichia Coli Glycogen Synthase*. *Journal of Biological Chemistry* **2009**, *284* (26), 17796–17807. https://doi.org/10.1074/jbc.M809804200.
(36) Stieglitz, K. A.; Xia, J.; Kantrowitz, E. R. The First High PH Structure of Escherichia Coli Aspartate Transcarbamoylase. *Proteins: Structure, Function, and Bioinformatics* **2009**, *74* (2), 318–327. https://doi.org/10.1002/prot.22162.
(37) Teotico, D. G.; Babaoglu, K.; Rocklin, G. J.; Ferreira, R. S.; Giannetti, A. M.; Shoichet, B. K. Docking for Fragment Inhibitors of AmpC β-Lactamase. *Proc Natl Acad Sci USA* **2009**, *106* (18), 7455. https://doi.org/10.1073/pnas.0813029106.
(38) Nurmohamed, S.; Vaidialingam, B.; Callaghan, A. J.; Luisi, B. F. Crystal Structure of Escherichia Coli Polynucleotide Phosphorylase Core Bound to RNase E, RNA and Manganese: Implications for Catalytic Mechanism and RNA Degradosome Assembly. *Journal of Molecular Biology* **2009**, *389* (1), 17–33. https://doi.org/10.1016/j.jmb.2009.03.051.
(39) Scian, M.; Lin, J. C.; Le Trong, I.; Makhatadze, G. I.; Stenkamp, R. E.; Andersen, N. H. Crystal and NMR Structures of a Trp-Cage Mini-Protein Benchmark for Computational Fold Prediction. *Proc Natl Acad Sci USA* **2012**, *109* (31), 12521. https://doi.org/10.1073/pnas.1121421109.
(40) Brown, J. W.; Farelli, J. D.; McKnight, C. J. On Unsatisfied Hydrogen Bonds in the N-Terminal Subdomain of Villin Headpiece. *Journal of Molecular Biology* **2011**, *413* (3), 543–547. https://doi.org/10.1016/j.jmb.2011.08.024.
(41) Bujacz, A. Structures of Bovine, Equine and Leporine Serum Albumin. *Acta Crystallographica Section D* **2012**, *68* (10), 1278–1289. https://doi.org/10.1107/S0907444912027047.
(42) Bager, R.; Johansen, J. S.; Jensen, J. K.; Stensballe, A.; Jendroszek, A.; Buxbom, L.; Sørensen, H. P.; Andreasen, P. A. Protein Conformational Change Delayed by Steric Hindrance from an N-Linked Glycan. *Journal of Molecular Biology* **2013**, *425* (16), 2867–2877. https://doi.org/10.1016/j.jmb.2013.05.007.
(43) Ember, S. W. J.; Zhu, J.-Y.; Olesen, S. H.; Martin, M. P.; Becker, A.; Berndt, N.; Georg, G. I.; Schönbrunn, E. Acetyl-Lysine Binding Site of Bromodomain-Containing Protein 4 (BRD4) Interacts with Diverse Kinase Inhibitors. *ACS Chem. Biol.* **2014**, *9* (5), 1160–1171. https://doi.org/10.1021/cb500072z.
(44) Song, H.; Dong, C.; Qin, M.; Chen, Y.; Sun, Y.; Liu, J.; Chan, W.; Guo, Z. A Thiamine-Dependent Enzyme Utilizes an Active Tetrahedral Intermediate in Vitamin K Biosynthesis. *J. Am. Chem. Soc.* **2016**, *138* (23), 7244–7247. https://doi.org/10.1021/jacs.6b03437.
(45) Ylänne, J.; Scheffzek, K.; Young, P.; Saraste, M. Crystal Structure of the Alpha-Actinin Rod Reveals an Extensive Torsional Twist. *Structure* **2001**, *9* (7), 597–604. https://doi.org/10.1016/s0969-2126(01)00619-0.
(46) Gnanapragasam, M. N.; Scarsdale, J. N.; Amaya, M. L.; Webb, H. D.; Desai, M. A.; Walavalkar, N. M.; Wang, S. Z.; Zhu, S. Z.; Ginder, G. D.; Williams, D. C. P66α–MBD2 Coiled-Coil Interaction and Recruitment of Mi-2 Are Critical for Globin Gene Silencing by the MBD2–NuRD Complex. *Proceedings of the National Academy of Sciences* **2011**, *108* (18), 7487–7492.
(47) Bernadó, P.; Svergun, D. I. Structural Analysis of Intrinsically Disordered Proteins by Small-Angle X-Ray Scattering. *Mol. BioSyst.* **2012**, *8* (1), 151–167. https://doi.org/10.1039/C1MB05275F.
(48) Fuertes, G.; Banterle, N.; Ruff, K. M.; Chowdhury, A.; Mercadante, D.; Koehler, C.; Kachala, M.; Estrada Girona, G.; Milles, S.; Mishra, A.; Onck, P. R.; Gräter, F.; Esteban-Martín, S.; Pappu, R. V.; Svergun, D. I.; Lemke, E. A. Decoupling of Size and Shape Fluctuations in Heteropolymeric Sequences Reconciles Discrepancies in SAXS vs. FRET Measurements. *Proceedings of the National Academy of Sciences* **2017**, *114* (31), E6342–E6351. https://doi.org/10.1073/pnas.1704692114.
(49) Mittal, A.; Holehouse, A. S.; Cohan, M. C.; Pappu, R. V. Sequence-to-Conformation Relationships of Disordered Regions Tethered to Folded Domains of Proteins. *Journal of Molecular Biology* **2018**, *430* (16), 2403–2421. https://doi.org/10.1016/j.jmb.2018.05.012.

(50) Estaña, A.; Barozet, A.; Mouhand, A.; Vaisset, M.; Zanon, C.; Fauret, P.; Sibille, N.; Bernadó, P.; Cortés, J. Predicting Secondary Structure Propensities in IDPs Using Simple Statistics from Three-Residue Fragments. *Journal of Molecular Biology* **2020**, *432* (19), 5447–5459. https://doi.org/10.1016/j.jmb.2020.07.026.
(51) Ruff, K. M.; Pappu, R. V.; Holehouse, A. S. Conformational Preferences and Phase Behavior of Intrinsically Disordered Low Complexity Sequences: Insights from Multiscale Simulations. *Current Opinion in Structural Biology* **2019**, *56*, 1–10. https://doi.org/10.1016/j.sbi.2018.10.003.
(52) Zhou, H.-X. Polymer Models of Protein Stability, Folding, and Interactions. *Biochemistry* **2004**, *43* (8), 2141–2154. https://doi.org/10.1021/bi036269n.
(53) Jensen, M. R.; Ruigrok, R. W.; Blackledge, M. Describing Intrinsically Disordered Proteins at Atomic Resolution by NMR. *Current Opinion in Structural Biology* **2013**, *23* (3), 426–435. https://doi.org/10.1016/j.sbi.2013.02.007.
(54) Jensen, M. R.; Zweckstetter, M.; Huang, J.; Blackledge, M. Exploring Free-Energy Landscapes of Intrinsically Disordered Proteins at Atomic Resolution Using NMR Spectroscopy. *Chem. Rev.* **2014**, *114* (13), 6632–6660. https://doi.org/10.1021/cr400688u.
(55) Aznauryan, M.; Delgado, L.; Soranno, A.; Nettels, D.; Huang, J.; Labhardt, A. M.; Grzesiek, S.; Schuler, B. Comprehensive Structural and Dynamical View of an Unfolded Protein from the Combination of Single-Molecule FRET, NMR, and SAXS. *Proc Natl Acad Sci USA* **2016**, *113* (37), E5389–E5398. https://doi.org/10.1073/pnas.1607193113.
(56) 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.* **2020**, *142* (37), 15697–15710. https://doi.org/10.1021/jacs.0c02088.
(57) Bernado, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W. H.; Blackledge, M. A Structural Model for Unfolded Proteins from Residual Dipolar Couplings and Small-Angle x-Ray Scattering. *Proceedings of the National Academy of Sciences* **2005**, *102* (47), 17002–17007. https://doi.org/10.1073/pnas.0506202102.
(58) Lazar, T.; Martínez-Pérez, E.; Quaglia, F.; Hatos, A.; Chemes, L. B.; Iserte, J. A.; Méndez, N. A.; Garrone, N. A.; Saldaño, T. E.; Marchetti, J.; Rueda, A. J. V.; Bernadó, P.; Blackledge, M.; Cordeiro, T. N.; Fagerberg, E.; Forman-Kay, J. D.; Fornasari, M. S.; Gibson, T. J.; Gomes, G.-N. W.; Gradinaru, C. C.; Head-Gordon, T.; Jensen, M. R.; Lemke, E. A.; Longhi, S.; Marino-Buslje, C.; Minervini, G.; Mittag, T.; Monzon, A. M.; Pappu, R. V.; Parisi, G.; Ricard-Blum, S.; Ruff, K. M.; Salladini, E.; Skepö, M.; Svergun, D.; Vallet, S. D.; Varadi, M.; Tompa, P.; Tosatto, S. C. E.; Piovesan, D. PED in 2021: A Major Update of the Protein Ensemble Database for Intrinsically Disordered Proteins. *Nucleic Acids Research* **2021**, *49* (D1), D404–D411. https://doi.org/10.1093/nar/gkaa1021.
(59) Gräwe, A.; Stein, V. Linker Engineering in the Context of Synthetic Protein Switches and Sensors. *Trends in Biotechnology* **2021**, *39* (7), 731–744. https://doi.org/10.1016/j.tibtech.2020.11.007.
(60) Rose, A.; Meier, I. Scaffolds, Levers, Rods and Springs: Diverse Cellular Functions of Long Coiled-Coil Proteins. *CMLS, Cell. Mol. Life Sci.* **2004**, *61* (16). https://doi.org/10.1007/s00018-004-4039-6.

Figures:

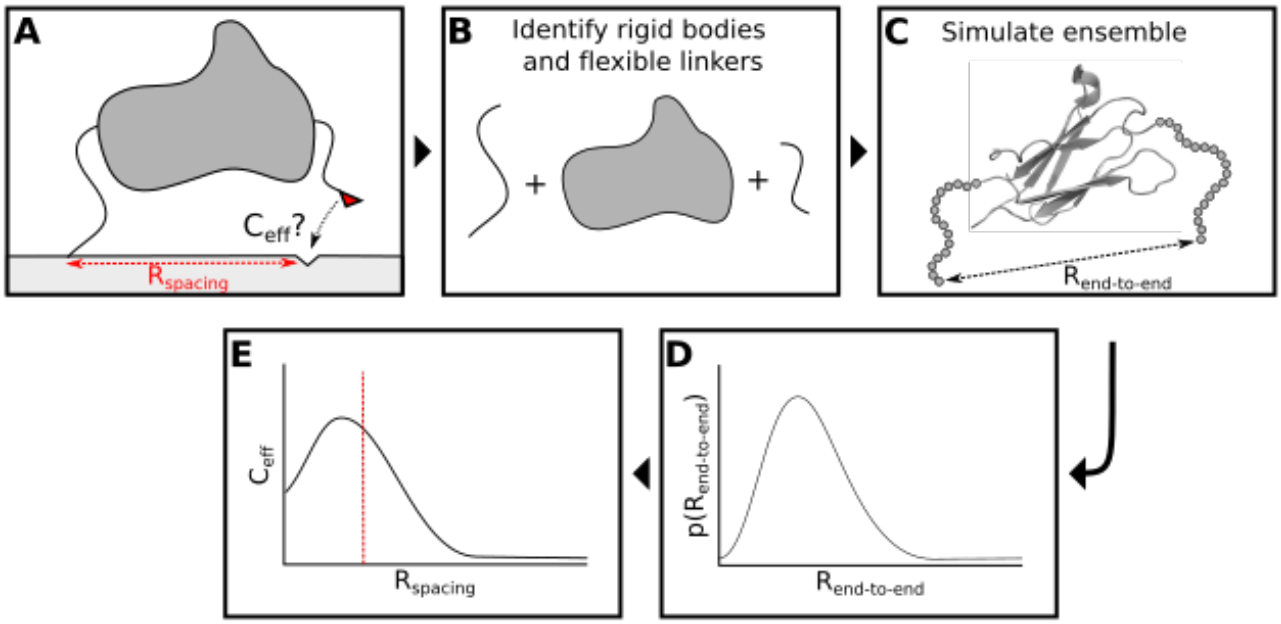

Figure 1: Flowchart for the estimation of C_{eff} for linker architectures composed of folded and disordered segments. First, the linker architecture was divided into parts that will be treated as rigid bodies and parts that will be treated as a disordered chains (B). An ensemble of typically 10.000 conformations was calculated for each linker architecture (C). A histogram of end-to-end distances was constructed (D) and converted into C_{eff} of tethered ligand at different spacings from the initial attachment site (E).

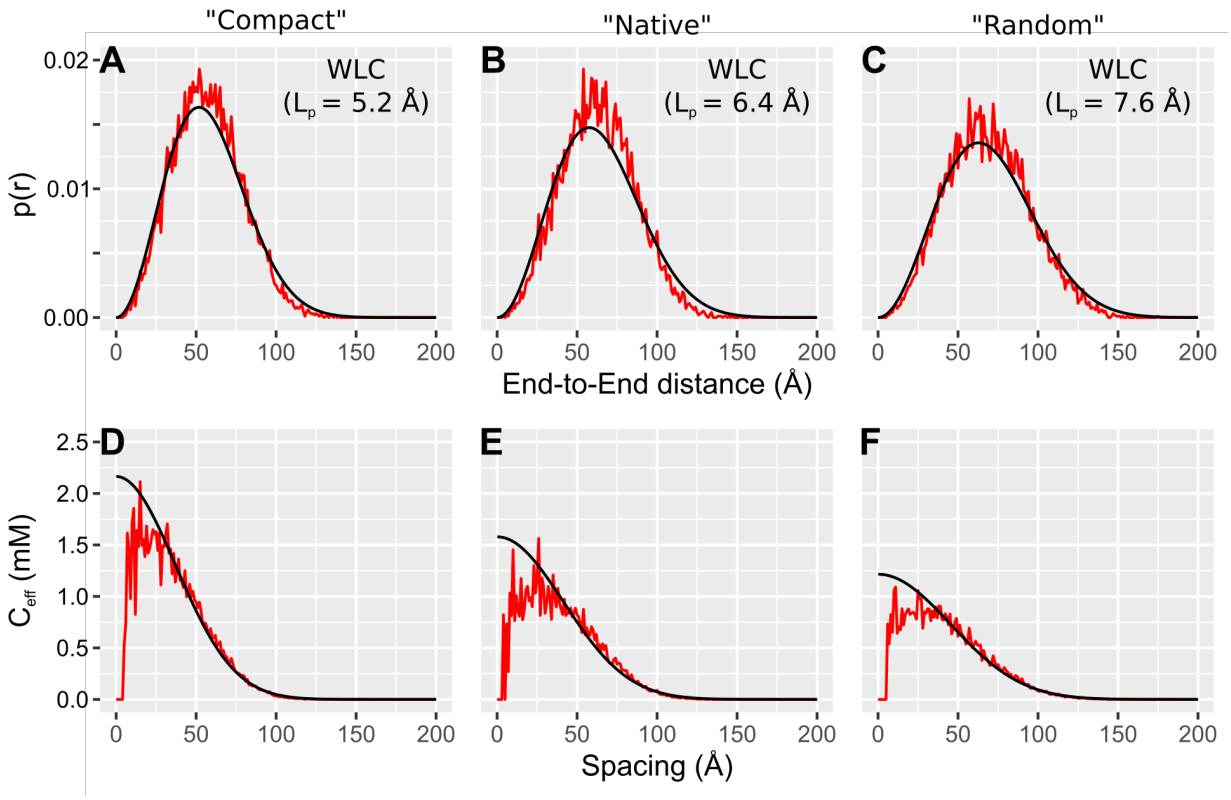

Figure 2: Comparison of ensemble reconstruction and worm-like chain modelling for a fully disordered linker. Ensembles of 10.000 conformations of a $(GS)_{50}$-linker were constructed using the (A, D) "Compact", (B, E) "Native" and (C,F) "Random" coil libraries of EOM.[12] WLC simulations were done using "C_{eff} calculator" using persistence lengths from a fit of the WLC to the p(R), which gave the following values: "Compact": 5.2 Å (5.09 - 5.30 Å). "Native": 6.4 Å (6.27 - 6.63 Å). "Random": 7.6 Å (7.38-7.73 Å). Values in parenthesis represent 95% C.I. of the fitted parameters.

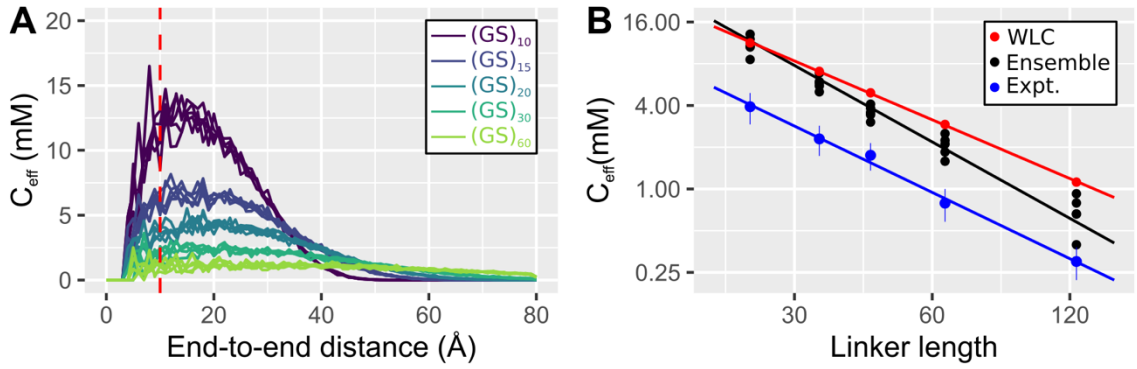

Figure 3: Comparison of predicted C_{eff} for disordered linkers to values measured using a FRET biosensor. (A) Effective concentration landscapes predicted for a series of disordered $(GS)_x$-linkers previously been measured by competition experiments in a FRET biosensor.[25] In addition to the $(GS)_x$-repeat, the linker contains two residues from restriction sites in either end. For each linker, five ensembles of 10.000 conformations were simulated. The dashed line indicates the spacing (10 Å) between the attachment sites in the complex used in the biosensor.[46] (B) Linker length scaling of C_{eff} at 10 Å from prediction and experiments. C_{eff} values from ensemble calculation is averages of three 1 Å bins from 8.5-11.5 Å. Scaling exponents from the fits are WLC: -1.41, ensemble: -1.83 and experiment: -1,59).

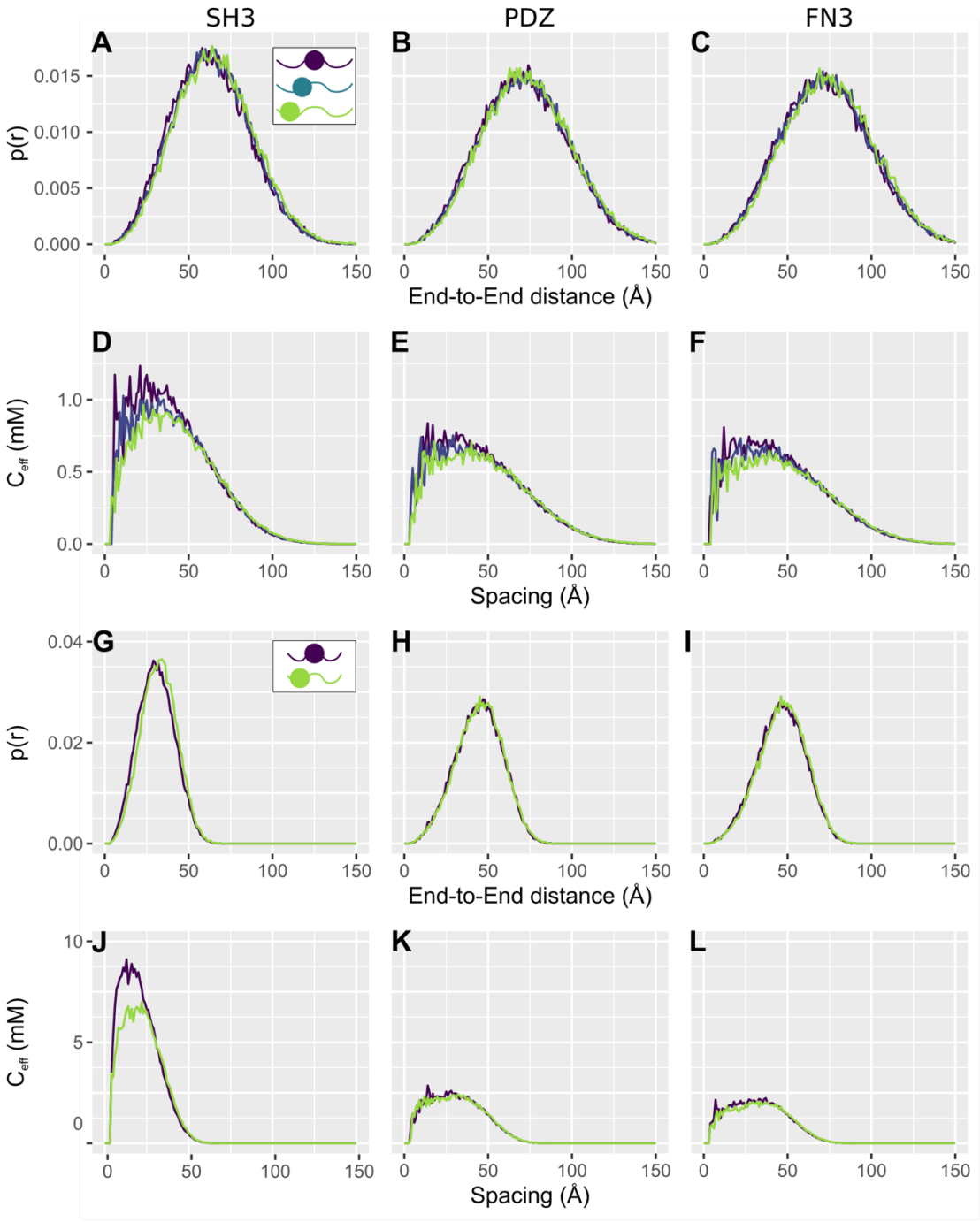

Figure 4: Multiple linkers can mostly be combined into a single longer linker. Predicted end-to-end distance distributions and C_{eff} from 50.000 member ensembles for linkers composed of a folded domain flanked by two disordered segments. (A-F) A total of 100 disordered residues distributed as 50-50, 25-75 and 5-95. (G-L) A total of 30 disordered residues 15-15 and 5-25. Structures used: (A, D, G, J) SH3 domain from (PDB: 1w6x). (B, E, H, K) PDZ3 domain from PSD-95 (PDB: 5d13). (C, F, G, H) 11th FN3 domain from fibronectin (PDB: 5dft).

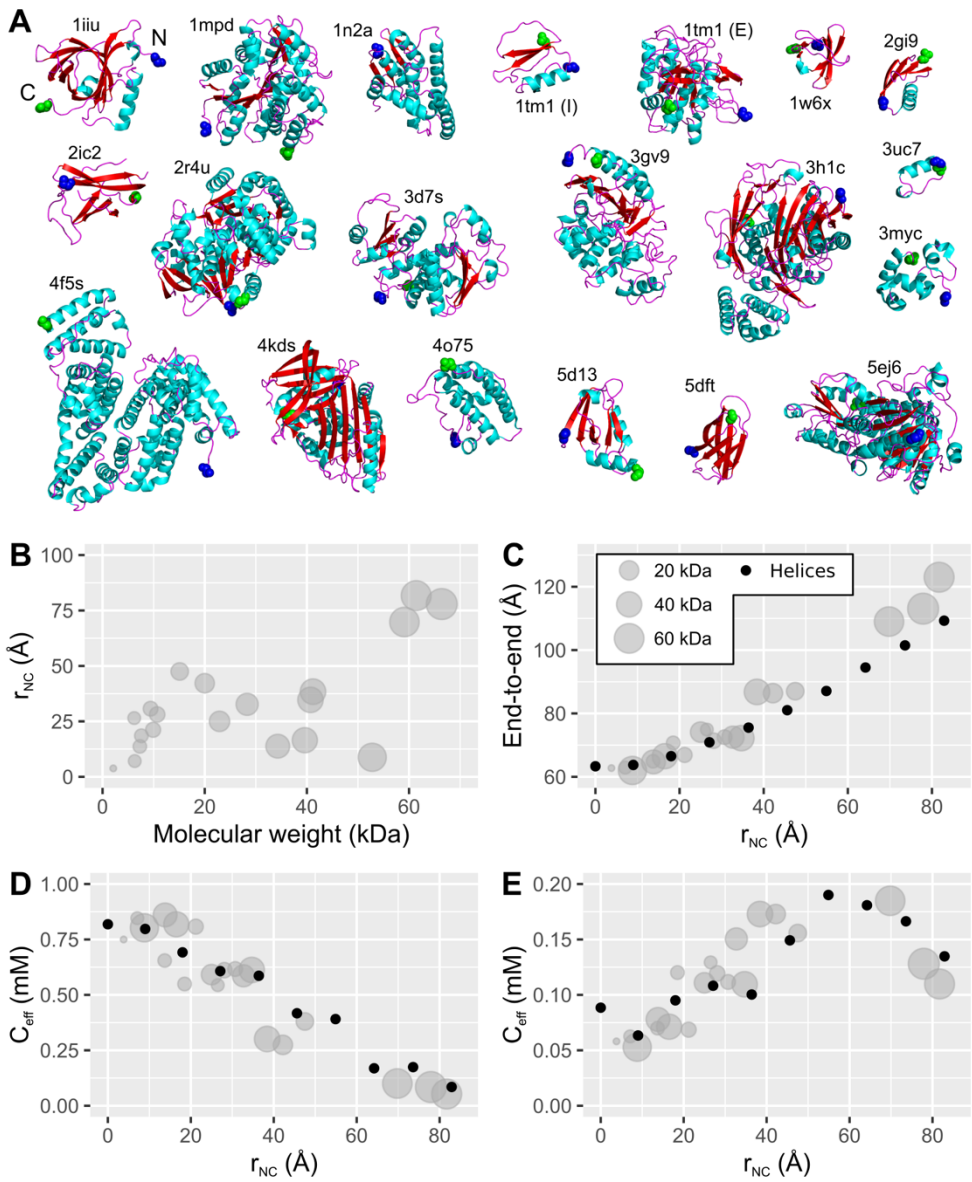

Figure 5: Globular domains in linkers can be approximated by a rigid rod. (A) Structures of the twenty globular domains that were inserted in the middle of a 100-residues GS linker and modelled using EOM. The N-termini (blue) and C-termini (green) are highlighted. (B) Distribution of N-to-C distances of the globular domains (r_{NC}) and molecular weights in the folded domains used. (C) The mean end-to-end distance of the linker containing each of the folded domains compared to the rigid α-helices of variable length. The circle area is proportional to the M_W of the globular domain. (D and E) C_{eff} was calculated at spacings of (D) 50 Å and (E) 100 Å. r_{NC} was measured between C_α of the first and last residue. The ensemble size was 10.000 for all linkers except those used in Fig. 4 (PDB: 1W6X, 5D13, 5DFT), where the ensemble size was 50.000. The data point plotted at r_{NC} = 0 represents the linker without an inserted domain.

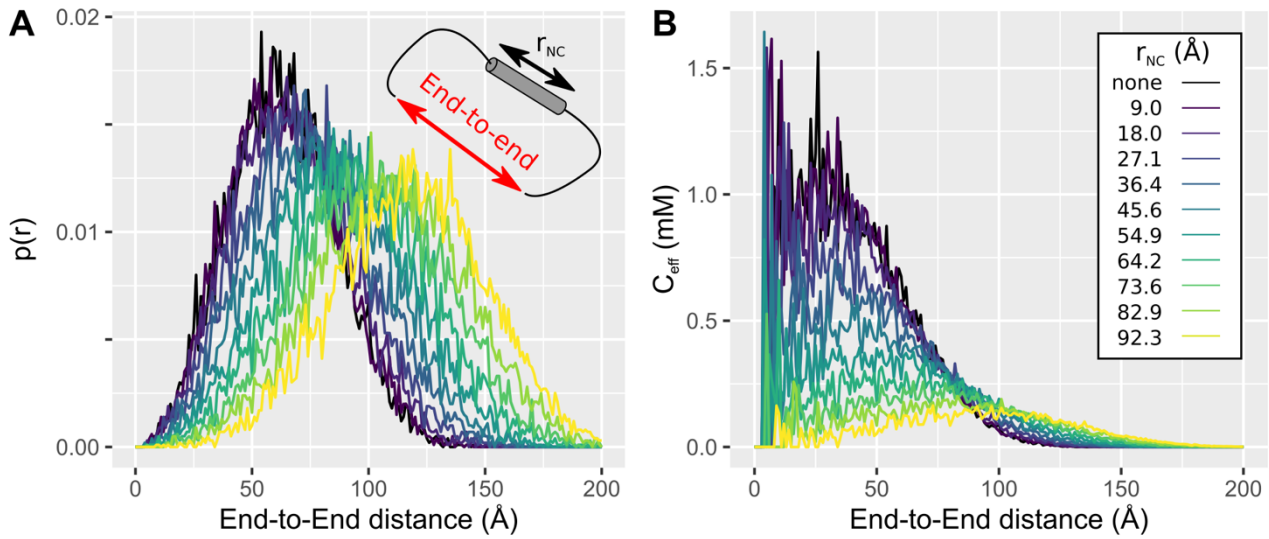

Figure 6: Modelling the effect of folded domain as a stiff α-helix. Rigid poly-alanine α-helices of variable length was inserted in the middle of a 100-residue GS-linker and an ensemble of 10.000 structures was modelled use EOM. For each linker is shown: (A) End-to-end distance distributions, (B) C_{eff} scaling. The length of the α-helix was increased in steps of 7 residues corresponding roughly to 9 Å steps. The r_{NC} of the rigid rod was measured between the C_{α} of the first and last residues of the helix.

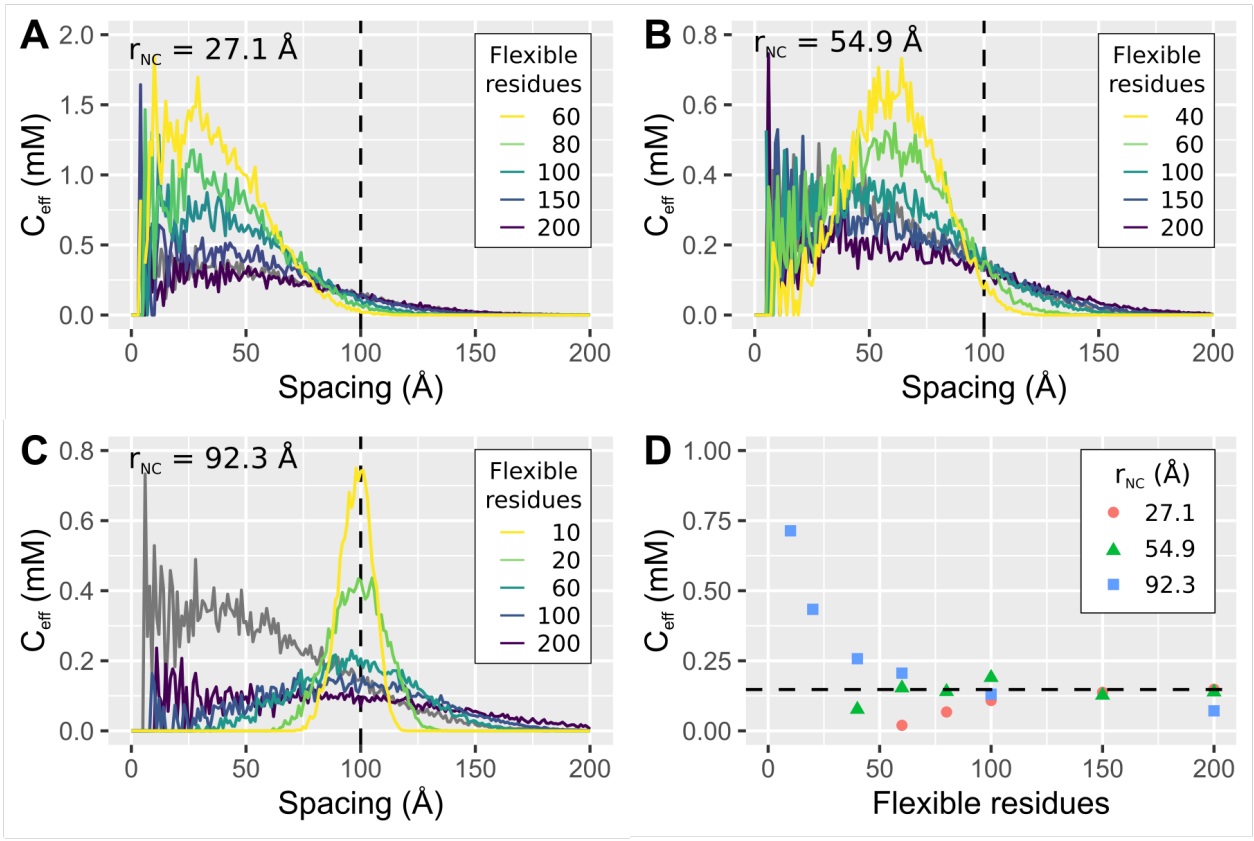

Figure 7: Effective concentrations enforced by linkers composed of a rigid rod flanked by flexible segments of variable length. Ensembles were generated for linkers consisting of flexible segments of variable length with a central α-helix containing (A) 21, (B) 42 and (C) 70 residues corresponding to r_{NC} lengths of 27.1, 54.9 and 92.3 Å. The effective concentrations are compared to the flexible linker that results in the highest C_{eff} at a spacing of 100 Å, i.e. 200-residues. (D) C_{eff} at a spacing of 100 Å for different linker architectures show that a linker containing a folded domain can exceed the maximal value attainable with a fully flexible linker (147µM, dashed line), when the r_{NC} of the folded domain matches the spacing between binding sites.

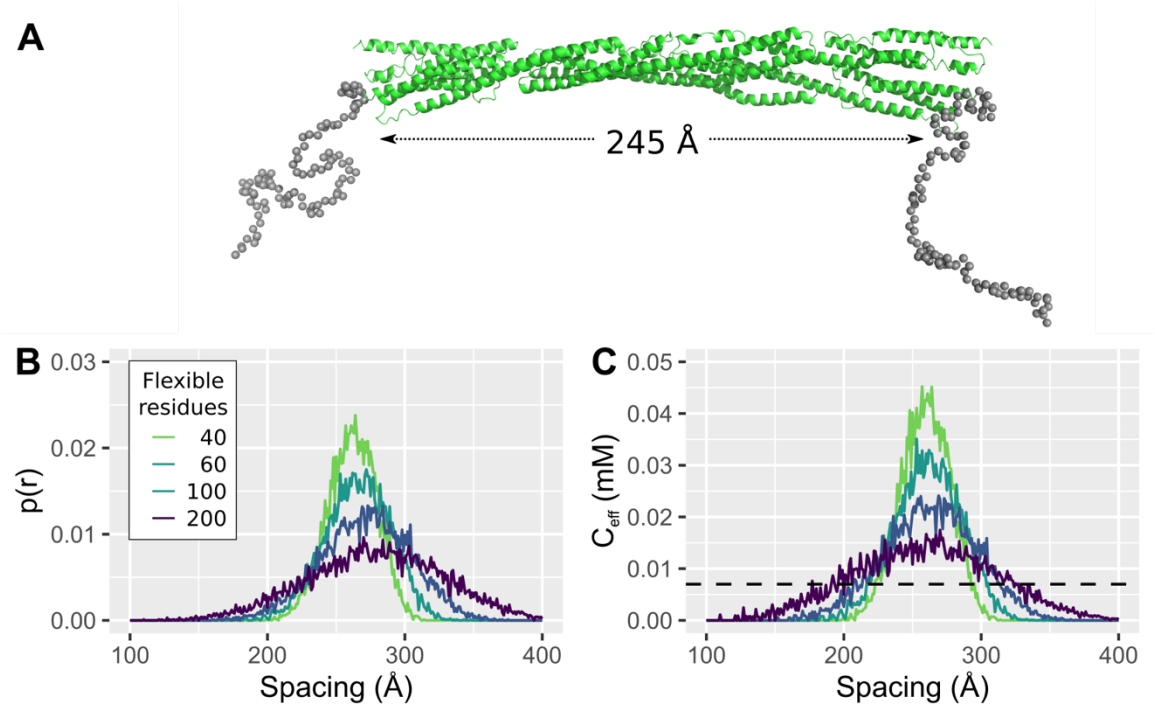

Fig. 8: Modelling of linker ensemble of a disordered linkers containing the rod domain from α-actinin. (A) Represenatative conformation of a linker architecture containing the rod domain of α-actinin (PDB:1HCI)[45] flanked by 100 flexible residues on either side. The rod domain is an anti-parallel homodimer, so the model was made by appending 100 flexible residues to the C-terminus. The Cα-atoms of the C-terminal residues of the two monomers are seperated by 245 Å. (B) End-to-end probability distributions (B) and C_{eff} profiles (C) of linker architectures containing α-actinin flanked disordered segments of variable length. The dashed line in (C) corresponds to the highest achievable C_{eff} (7 μM) by a fully disordered linker (1283 residues) at a spacing of 250 Å modelled by WLC with an L_p of 6.4 Å.