PROGRAMMING CO-FOLDING TO DESIGN BINDERS FOR INTRINSICALLY DISORDERED EPITOPES

Jakub Lála[1,2], **Daniele Visco**[3] & **Stefano Angioletti-Uberti**[1,2,3]
[1] Department of Materials, Imperial College London, London, UK
[2] Thomas Young Centre, Imperial College London, London, UK
[3] Nanograb Ltd, London, UK
sangiole@imperial.ac.uk

ABSTRACT

Due to their lack of a specific structure and dynamical nature, targeting of epitopes that are part of an intrinsically disordered region of a protein is a notoriously difficult task. Here, we describe a computational approach to overcome this problem, based on the use of a protein folding algorithm within a Monte Carlo optimization pipeline to generate peptide binders that bind by co-folding with their epitope. For different protein targets, we show by accurate free-energy calculations that our approach is able to design strong binders, with binding free energies of the order of tens of k_BT (i.e. stronger than $50\,\mathrm{kJ\,mol^{-1}}$), corresponding to K_D values in the nM regime or lower. Direct observation of the molecular structures during the binding process shows the binder and targeted epitope fold upon binding and acquire a structure not presented in their unbound state, suggesting that co-folding is the implied mechanism, and that the latter is correctly described by the protein folding algorithm we employ. Given the ubiquitous presence of unstructured regions in proteins, our results suggest a potential pathway to design drugs targeting a large variety of previously untargetable epitopes and opens new possibilities for therapeutic intervention in diseases where disordered proteins play a key role.

1 INTRODUCTION

Despite lacking a stable structure, intrinsically disordered regions (IDRs) of proteins play critical roles in cellular processes, including cell signaling, regulation, and molecular recognition (Babu et al., 2011; Wright & Dyson, 2014; Holehouse & Kragelund, 2023). These regions are highly dynamic and participate in transient but specific interactions with other biomolecules, making them essential for biological functions. However, it is exactly their dynamical nature that complicates our understanding of their behavior, hindering the rational design of molecules that could selectively bind to specific epitopes in IDRs. Therefore, exploiting them as drug targets remains a challenge.

Traditional structure-based design strategies, which rely on well-defined binding pockets and stable secondary structures, are intrinsically inadequate for IDRs due to their dynamic nature. To understand how proteins fold into their native structures, we hypothesize that protein-folding algorithms must learn how arbitrary amino acid sequences interact and thus lead to co-folding, i.e, a behavior where a binder-target pair undergoes a coordinated transition into a bound stable conformation. Notably, co-folding mediates protein interactions (Kussie et al., 1996; Davidson et al., 1998) and has been proposed as a mechanism for binding specificity and signaling activation via interacting IDRs (Wright & Dyson, 2014). This principle parallels findings in RNA-targeting drugs, where molecules achieve specificity by stabilizing a particular conformation (Ganser et al., 2020; Tong et al., 2024). Based on these premises, we view peptides and small proteins as promising candidates that can bind and induce folding in the target IDRs. Specifically, this paper presents three key contributions:

- We extend the Monte Carlo algorithm to optimize protein sequences by Hie et al. (2022) with new loss function terms that favour binders inducing co-folding and also implement simulated tempering that speeds up the search for an optimal binder.
- Using this protocol, we design novel peptide binders to intrinsically disordered epitopes, opening up the possibility to bind previously considered undruggable protein targets.

- We carry out free-energy calculations with well-tempered metadynamics to estimate the binding affinity *in silico* and thus validate the binders as a proof-of-concept.

For the related works, see Appendix A.

2 METHOD

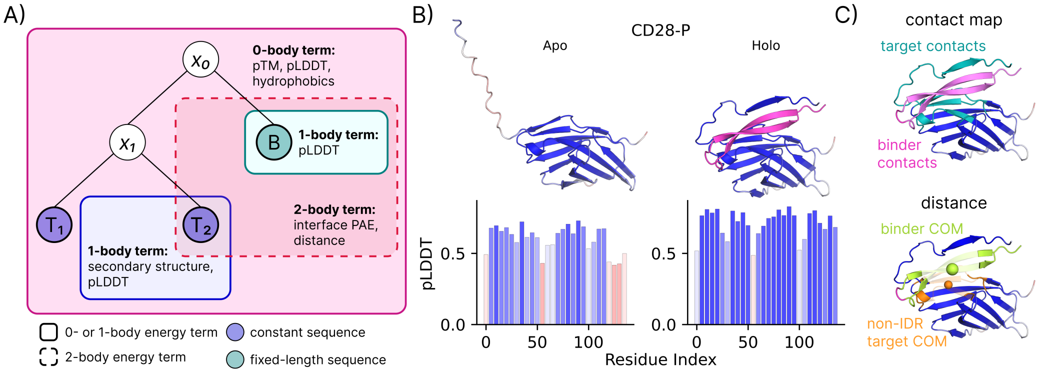

Figure 1: A) Graph representation used in the protocol in Section 2.1 of the energy terms describing the design for a peptide (B) binding to an IDR (T_2). 0-body terms apply to all the residues in the system, 1-body terms apply to point residues, and 2-body terms apply a constraint in a pair-wise fashion. B) Comparison of structures and pLDDT for the protein in its *apo* (unbound) and *holo* (bound) states. The presence of the binder (magenta) increases the pLDDT across the sequence, but especially for residues 120 to 140. The bar plot shows the pLDDT score as an average of 5-residue segments. C) Definition of the collective variables for the free energy calculations described in Section 2.2. Firstly, the contact map c (upper) is used between the IDR (teal) and the binder (light pink). Secondly, the distance r (bottom) is biased between the centres of masses - depicted by the spheres - of the binder (limon) and non-IDR part of the target (orange). Notice a coil on the left of the binder (darker colour) that is not considered a contact, i.e. is not considered to participate in the binding. The bottom visualization shows this clearly with the magenta residues, which are not a part of c, nor of r.

2.1 GUIDED OPTIMIZATION USING SIMULATED TEMPERING

Even for short sequences of 30 amino acids, the number of potential candidates for effective peptide binders that would fold the epitope upon binding is immense, i.e. 20^{30} possibilities. Whereas this large selection increases the possibility that at least one candidate with the right properties exists, it is also an impossibly large space to search without any type of guidance. We solve this issue by using a purpose built loss function to drive the search process, employed within a modified version of a classical Monte Carlo algorithm (Metropolis et al., 1953), called simulated tempering (Marinari & Parisi, 1992), to help sampling of the sequence space in search of the optimal solution.

The practical implementation of our protocol builds on ideas and tools presented by the former FAIR team, as described by Hie et al. (2022). In practice, we used their so-called high-level programming language for proteins that helps us to easily define a loss (energy) function that depends on structural constraints imposed on different parts of the protein-binder complex. We minimize this loss by sampling random sequences and folding them into a structure using a deep learning prediction model, specifically ESMFold (Lin et al., 2023) (see pseudo-code of the protocol in Appendix B.1). Our work differs from Hie et al. (2022) by using simulated tempering instead of simulated annealing, something that we found of extreme importance to achieve low-energy solutions. The loss function contains seven different terms, linearly combined with specific weights. Detailed description of the different terms is reported in Table 1 in Appendix B.1. Apart from the previously used design metrics such as predicted Template Modeling score (pTM) or predicted Local Distance Difference Test (pLDDT), we used interface Predicted Aligned Error (iPAE) and the distance between the target-binder pair to promote binding. More interestingly, we used a secondary structure content energy term and a local pLDDT term with high weighting on the IDR to achieve an effective folding upon binding. Fig. 1A shows the loss function as a graph representation.

2.2 Free-energy calculations of binding using metadynamics

To validate our protocol results, we employed free-energy calculations via the well-tempered metadynamics algorithm (Barducci et al., 2008). We describe the interactions with the Amber ff14SB forcefield (Maier et al., 2015) and an implicit water model (Nguyen et al., 2013). While this solvent choice is suboptimal to describe the energetics compared to an explicit water model, statistical sampling of the collective variable (CV) space and convergence of the free-energy profile is improved. Coarse-graining was necessary to sample bound states with the IDR stretched, otherwise the simulation box would be prohibitively large. Regardless of the several k_bT inaccuracy in binding affinity, our aim is to qualitatively confirm the presence of a deep minima. We experimented with constraining the disordered region to reduce the simulation box size, but that introduced artifacts, such as self-interaction with the neighbouring image, or overestimation of the binding affinity as it artificially folded the IDR, requiring further analytical correction.

We bias the system along two CVs shown in Fig. 1C. First, a contact map $c(\vec{x}_{\text{target}}, \vec{x}_{\text{binder}})$ measuring the number of native contacts, defined by a close-proximity selection of carbon alpha atom positions from the well-structured, high-pLDDT part of the target, $\vec{x}_{\text{target}}$, and the binder, $\vec{x}_{\text{binder}}$. Second, a distance $r(\vec{x}_{\text{target}}, \vec{x}_{\text{binder}})$ between the centres of masses (COMs) of $\vec{x}_{\text{target}}$ and $\vec{x}_{\text{binder}}$. We chose this set of simple CVs to account for both specific and non-specific binding. More importantly, the position of disordered residues is excluded from the computation of $\vec{x}_{\text{target}}$ in order to prevent artifacts introduced by large thermal fluctuations in their position on the COM of $\vec{x}_{\text{target}}$. This would often lie outside of the rigid domain of the target, leading to degeneracies in r. Throughout the simulation we deposited a time-dependent bias, whose magnitude decays with time, following the well-established well-tempered metadynamics. As shown by Barducci et al. (2008), the bias accumulated is directly proportional to the real FES (at convergence), according to:

$$F(r, c) = -\frac{T + \Delta T}{\Delta T} V(r, c) \tag{1}$$

where F is the FES, T is the temperature and ΔT represents the effective excess temperature experienced by the system along the CVs due to the bias introduced. For the purpose of this investigation, we estimate the relative free energy difference by

$$\Delta G = -k_B T \ln \frac{Z_{\text{bound}}}{Z_{\text{unbound}}}, \tag{2}$$

$$\frac{Z_{\text{bound}}}{Z_{\text{unbound}}} = \int_0^{r^*} e^{\frac{-(F(r) - F_{\text{unbound}})}{k_B T}} dr \tag{3}$$

where k_B is the Boltzmann constant and r^* is the cutoff distance defining the boundary between the bound and unbound states. To compute the bound/unbound partition functions, we first marginalized $F(r, c)$ over c to obtain $F(r)$. Then we defined the unbound state *spatially*, as the region of space beyond the transition point r^*, where interaction between the binder and the target are negligible (i.e. $F(r)$ is flat). The average free energy of the system in this region is F_{unbound}, acting as the reference energy in the estimation of ΔG. To avoid excessive sampling of the unbound region, we used an upper wall bias $V_{uw}(r)$ for $r > r_{uw}$. While the use of molecular simulations does not allow us to measure the binding energy at the same level of accuracy that could be attained in experiments, it still provides a very detailed view of the mechanics of the binding process, and thus achieve (albeit *in silico*) a proof-of-concept. Details of the simulation procedure and specific parameters used are in Appendix B.2.

3 Results

As a proof of principle, we report the results of the design protocol on four different proteins: α-synuclein (ASYN), CD28, p53 (P53) and SUMO. We chose these proteins because of the presence of an unstructured region and their potential of being used as drug targets in various therapies (see Appendix A for details). We name the systems after the target, followed by a binder-related tag reflecting its design parameters used in optimization.

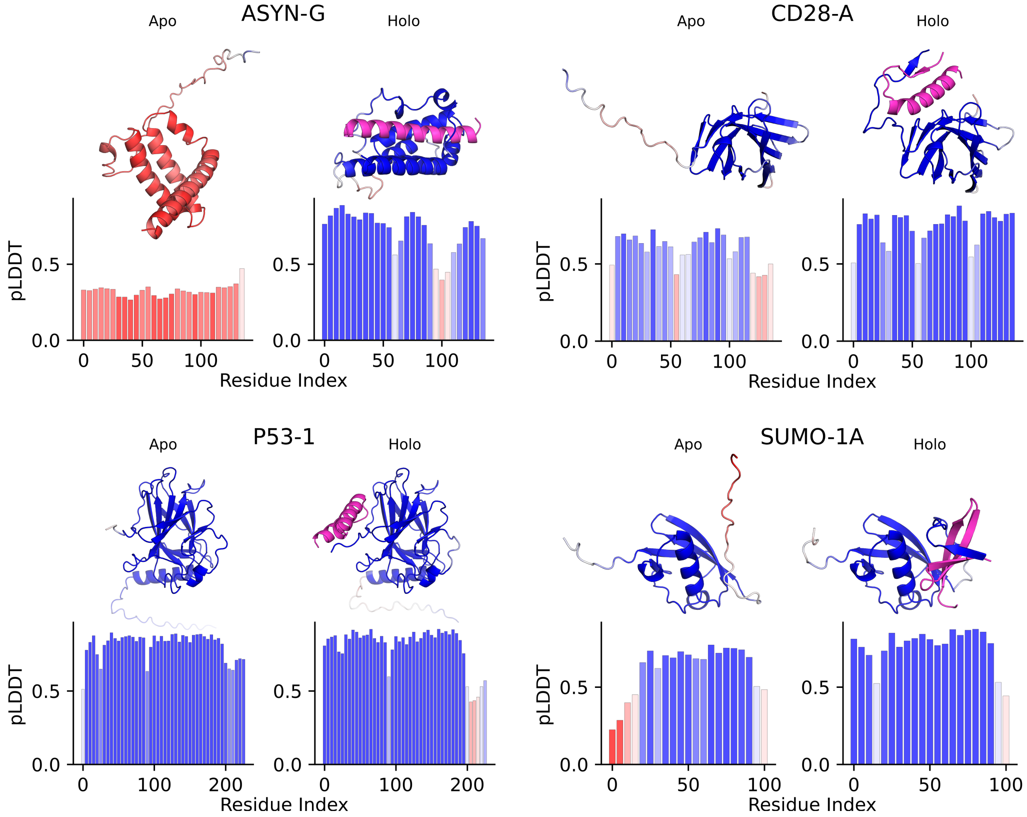

Figure 2: Four examples of peptide-binder (magenta) designs. The structures are colored based on the pLDDT score – with high pLDDT residues being blue, medium pLDDT values being white, and low pLDDT being red. The bottom bar plots for each structure also show the pLDDT as an average of 5-residue segments with the same colour scheme.

Firstly, we present one optimized binder per target in Fig. 2. Notice we are able to find binders for all targets, despite them showing different behaviours. The low pLDDT of α-synuclein across its entire sequence suggests there is no single stable tertiary conformation despite the secondary structure observed. However, after co-folding with the ASYN-G binder, the pLDDT across the entire sequence is increased significantly, potentially correlating with increasing the likelihood of obtaining a single dominating conformation. A similar situation occurs with CD28-A and SUMO-1A binders, yet there one can see that there are very specific disordered regions (residues 120 to 140 for CD28 and residues 1 to 15 for SUMO) for which the binders increase the pLDDT, presumably folding them upon binding. For P53, the binder successfully induces order in residues 1-5, but unexpectedly decreases the pLDDT score in the IDR region (residues 200-230). This effect could arise from either a genuine physical mechanism of induced disorder motivating further investigation, or from limitations in the pLDDT metric itself. Since pLDDT conflates model confidence with structural order, the presence of the binder might might require ESMFold to attempt to predict complexes from sequence space farther away from its training data, making it thus predict structures with lower confidence, particularly in regions already prone to disorder. Lastly, our binders can target the formation of both alpha-helices (ASYN-G) and beta-sheets (CD28-A, SUMO-1A) on the target, while designing beta-sheet binders (SUMO-1A) is something that popular protein design pipelines are known to struggle with (Watson et al., 2023; Pacesa et al., 2024).

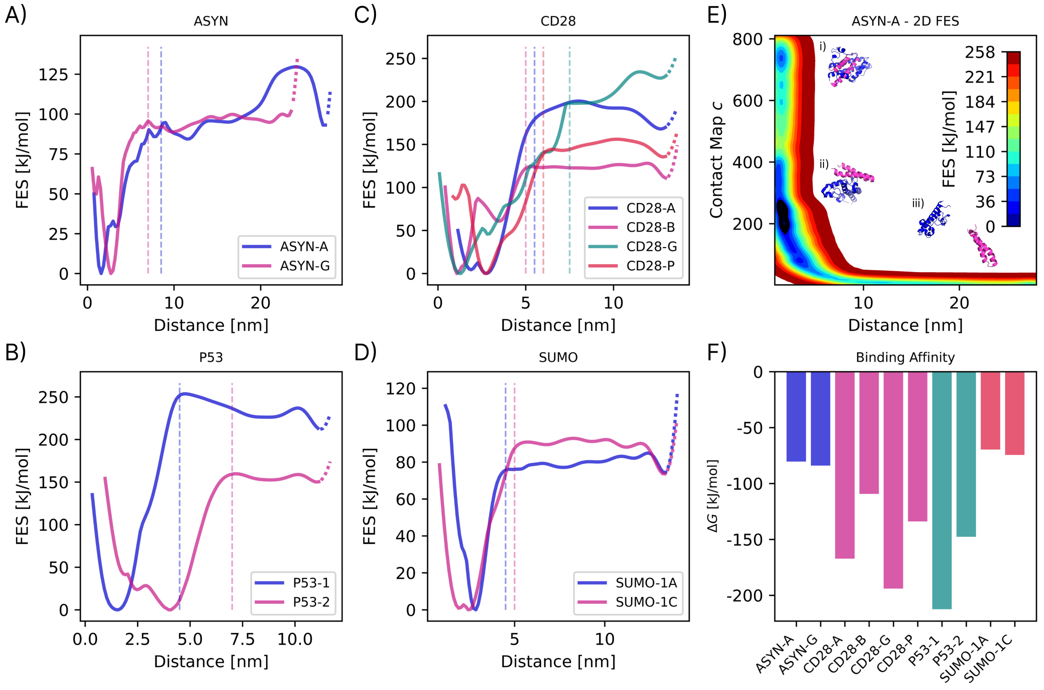

Figure 3: A–D) Free energy profiles along the distance r (as depicted in Fig. 1C). These were computed from the deposited bias using Eq. 1. The dashed vertical lines represent the cutoff r^*. Note that we do not remove the upper wall from $F(r)$ and rather highlight it by using the densely dotted line for $r > r_{uw}$ (see Eq. 6). E) An example of a 2D FES in the CV space for ASYN-A. Example conformations are provided for three of the modes observed: i) high c value, ii) deepest minima at $c \approx 200$ and iii) an unbound state. F) Binding affinities (ΔG) for each system computed from the FES projections in A–D) using Eq. 2, grouped in color by the protein target.

Secondly, we show the results of the metadynamics simulation for 10 binders in Fig. 3 with their estimated binding affinity. Figs. 3A to 3D show clear deep minima at low r across all systems, confirming the presence of a stable binding mode. The variability in the dimension of the bound state across the target-binder pairs can be explained by the IDR moving away from the centre of mass of the rigid part of the protein, while still being bound to the binder. This effect can then be different for each target-binder pair. On the other hand, we need to address a few limitations. First, these simulations come from a lower temperature run (T = 300 K) than the body-relevant temperature (T = 310 K). Second, we observe an artifact due to the upper wall bias $V(r)_{uw}$, which creates an apparent second minima near r_{uw}. It is likely that high stiffness of the wall leads to unphysical forces. We ignore the artifact for the purposes of this paper as it does not significantly contribute to ΔG. We also omit the volume correction due to entropic contributions on the FES at large r. Lastly, many of these simulations end up being stuck in the local basin, suggesting we do not sufficiently sample the CV space, as at convergence we would expect almost-free diffusion through the CV space without any significant energy barriers (see Appendix C for the trajectories). We believe this is due to the bias deposition rate decaying too fast. That leads to an incomplete reconstruction of the FES. We are continuing with these simulations, running with a lower decay on the bias deposition rate, less stiff upper wall, and parallel tempering to address these limitations and yield more accurate estimates for ΔG. Either way, for the purposes of this paper, we believe these preliminary results greatly support our argument of being able to design peptide binders to IDRs with binding affinity of tens of k_BT, an affinity relevant in the clinical biomedical setting.

4 Conclusion

In this paper we have shown a proof-of-concept protein design protocol for peptide binders targeting disordered regions on clinically important protein targets. We have extended a previously developed Monte Carlo optimization technique, which we are currently reformalizing into a standalone package. Our preliminary *in silico* validation shows promising binding affinities that could be later experimentally tested in a lab. Having the ability to design such binders on demand to previously unconsidered domains opens up the range of possibilities across the field of applied biotechnology.

REFERENCES

M Madan Babu, Robin van der Lee, Natalia Sanchez de Groot, and Jörg Gsponer. Intrinsically disordered proteins: regulation and disease. *Current Opinion in Structural Biology*, 21 (3):432–440, June 2011. ISSN 0959-440X. doi: 10.1016/j.sbi.2011.03.011. URL http://dx.doi.org/10.1016/j.sbi.2011.03.011.

Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.*, 100(2):020603, January 2008.

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, October 2022. ISSN 1095-9203. doi: 10.1126/science.add2187. URL http://dx.doi.org/10.1126/science.add2187.

W. Sean Davidson, Ana Jonas, David F. Clayton, and Julia M. George. Stabilization of α-synuclein secondary structure upon binding to synthetic membranes. *Journal of Biological Chemistry*, 273 (16):9443–9449, April 1998. ISSN 0021-9258. doi: 10.1074/jbc.273.16.9443. URL http://dx.doi.org/10.1074/jbc.273.16.9443.

Peter Eastman, Jason Swails, John D. Chodera, Robert T. McGibbon, Yutong Zhao, Kyle A. Beauchamp, Lee-Ping Wang, Andrew C. Simmonett, Matthew P. Harrigan, Chaya D. Stern, Rafal P. Wiewiora, Bernard R. Brooks, and Vijay S. Pande. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, 13(7):e1005659, July 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005659. URL http://dx.doi.org/10.1371/journal.pcbi.1005659.

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, Pushmeet Kohli, John Jumper, and Demis Hassabis. Protein complex prediction with alphafold-multimer. October 2021. doi: 10.1101/2021.10.04.463034. URL http://dx.doi.org/10.1101/2021.10.04.463034.

Laura R. Ganser, Megan L. Kelly, Neeraj N. Patwardhan, Amanda E. Hargrove, and Hashim M. Al-Hashimi. Demonstration that small molecules can bind and stabilize low-abundance short-lived rna excited conformational states. *Journal of Molecular Biology*, 432(4):1297–1304, February 2020. ISSN 0022-2836. doi: 10.1016/j.jmb.2019.12.009. URL http://dx.doi.org/10.1016/j.jmb.2019.12.009.

Casper A. Goverde, Benedict Wolf, Hamed Khakzad, Stéphane Rosset, and Bruno E. Correia. De novo protein design by inversion of the ¡scp¿alphafold¡/scp¿ structure prediction network. *Protein Science*, 32(6), May 2023. ISSN 1469-896X. doi: 10.1002/pro.4653. URL http://dx.doi.org/10.1002/pro.4653.

Casper A. Goverde, Martin Pacesa, Nicolas Goldbach, Lars J. Dornfeld, Petra E. M. Balbi, Sandrine Georgeon, Stéphane Rosset, Srajan Kapoor, Jagrity Choudhury, Justas Dauparas, Christian Schellhaas, Simon Kozlov, David Baker, Sergey Ovchinnikov, Alex J. Vecchio, and Bruno E. Correia. Computational design of soluble and functional membrane protein analogues. *Nature*, 631(8020):449–458, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07601-y. URL http://dx.doi.org/10.1038/s41586-024-07601-y.

Hao-Bo Guo, Alexander Perminov, Selemon Bekele, Gary Kedziora, Sanaz Farajollahi, Vanessa Varaljay, Kevin Hinkle, Valeria Molinero, Konrad Meister, Chia Hung, Patrick Dennis, Nancy Kelley-Loughnane, and Rajiv Berry. Alphafold2 models indicate that protein sequence determines both structure and dynamics. *Scientific Reports*, 12(1), June 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-14382-9. URL http://dx.doi.org/10.1038/s41598-022-14382-9.

Brian Hie, Salvatore Candido, Zeming Lin, Ori Kabeli, Roshan Rao, Nikita Smetanin, Tom Sercu, and Alexander Rives. A high-level programming language for generative protein design. December 2022. doi: 10.1101/2022.12.21.521526. URL http://dx.doi.org/10.1101/2022.12.21.521526.

Alex S. Holehouse and Birthe B. Kragelund. The molecular basis for cellular function of intrinsically disordered protein regions. *Nature Reviews Molecular Cell Biology*, 25(3):187–211, November 2023. ISSN 1471-0080. doi: 10.1038/s41580-023-00673-0. URL http://dx.doi.org/10.1038/s41580-023-00673-0.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, July 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL http://dx.doi.org/10.1038/s41586-021-03819-2.

Antti Kukkula, Veera K. Ojala, Lourdes M. Mendez, Lea Sistonen, Klaus Elenius, and Maria Sundvall. Therapeutic potential of targeting the sumo pathway in cancer. *Cancers*, 13(17):4402, August 2021. ISSN 2072-6694. doi: 10.3390/cancers13174402. URL http://dx.doi.org/10.3390/cancers13174402.

Paul H. Kussie, Svetlana Gorina, Vincent Marechal, Brian Elenbaas, Jacque Moreau, Arnold J. Levine, and Nikola P. Pavletich. Structure of the mdm2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science*, 274(5289):948–953, November 1996. ISSN 1095-9203. doi: 10.1126/science.274.5289.948. URL http://dx.doi.org/10.1126/science.274.5289.948.

Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, September 2002. ISSN 1091-6490. doi: 10.1073/pnas.202427399. URL http://dx.doi.org/10.1073/pnas.202427399.

Andrew Leaver-Fay, Michael Tyka, Steven M. Lewis, Oliver F. Lange, James Thompson, Ron Jacak, Kristian W. Kaufman, P. Douglas Renfrew, Colin A. Smith, Will Sheffler, Ian W. Davis, Seth Cooper, Adrien Treuille, Daniel J. Mandell, Florian Richter, Yih-En Andrew Ban, Sarel J. Fleishman, Jacob E. Corn, David E. Kim, Sergey Lyskov, Monica Berrondo, Stuart Mentzer, Zoran Popović, James J. Havranek, John Karanicolas, Rhiju Das, Jens Meiler, Tanja Kortemme, Jeffrey J. Gray, Brian Kuhlman, David Baker, and Philip Bradley. *Rosetta3*, pp. 545–574. Elsevier, 2011. doi: 10.1016/b978-0-12-381270-4.00019-6. URL http://dx.doi.org/10.1016/B978-0-12-381270-4.00019-6.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. ISSN 1095-9203. doi: 10.1126/science.ade2574. URL http://dx.doi.org/10.1126/science.ade2574.

James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, July 2015. ISSN 1549-9626. doi: 10.1021/acs.jctc.5b00255. URL http://dx.doi.org/10.1021/acs.jctc.5b00255.

E Marinari and G Parisi. Simulated tempering: A new monte carlo scheme. *Europhysics Letters (EPL)*, 19(6):451–458, July 1992. ISSN 1286-4854. doi: 10.1209/0295-5075/19/6/002. URL http://dx.doi.org/10.1209/0295-5075/19/6/002.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953. ISSN 1089-7690. doi: 10.1063/1.1699114. URL `http://dx.doi.org/10.1063/1.1699114`.

Hai Nguyen, Daniel R. Roe, and Carlos Simmerling. Improved generalized born solvent model parameters for protein simulations. *Journal of Chemical Theory and Computation*, 9(4):2020–2034, March 2013. ISSN 1549-9626. doi: 10.1021/ct3010485. URL `http://dx.doi.org/10.1021/ct3010485`.

Martin Pacesa, Lennart Nickel, Christian Schellhaas, Joseph Schmidt, Ekaterina Pyatova, Lucas Kissling, Patrick Barendse, Jagrity Choudhury, Srajan Kapoor, Ana Alcaraz-Serna, Yehlin Cho, Kourosh H. Ghamary, Laura Vinué, Brahm J. Yachnin, Andrew M. Wollacott, Stephen Buckley, Adrie H. Westphal, Simon Lindhoud, Sandrine Georgeon, Casper A. Goverde, Georgios N. Hatzopoulos, Pierre Gönczy, Yannick D. Muller, Gerald Schwank, Daan C. Swarts, Alex J. Vecchio, Bernard L. Schneider, Sergey Ovchinnikov, and Bruno E. Correia. Bindcraft: one-shot design of functional protein binders. October 2024. doi: 10.1101/2024.09.30.615802. URL `http://dx.doi.org/10.1101/2024.09.30.615802`.

Sylvain Peuget, Xiaolei Zhou, and Galina Selivanova. Translating p53-based therapies for cancer into the clinic. *Nature Reviews Cancer*, 24(3):192–215, January 2024. ISSN 1474-1768. doi: 10.1038/s41568-023-00658-3. URL `http://dx.doi.org/10.1038/s41568-023-00658-3`.

Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1–2):141–151, November 1999. ISSN 0009-2614. doi: 10.1016/s0009-2614(99)01123-9. URL `http://dx.doi.org/10.1016/S0009-2614(99)01123-9`.

Giulio Tesei and Kresten Lindorff-Larsen. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. *Open Research Europe*, 2:94, January 2023. ISSN 2732-5121. doi: 10.12688/openreseurope.14967.2. URL `http://dx.doi.org/10.12688/openreseurope.14967.2`.

Yuquan Tong, Jessica L. Childs-Disney, and Matthew D. Disney. Targeting rna with small molecules, from rna structures to precision medicines: Iuphar review: 40. *British Journal of Pharmacology*, 181(21):4152–4173, September 2024. ISSN 1476-5381. doi: 10.1111/bph.17308. URL `http://dx.doi.org/10.1111/bph.17308`.

G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, February 1977. ISSN 0021-9991. doi: 10.1016/0021-9991(77)90121-8. URL `http://dx.doi.org/10.1016/0021-9991(77)90121-8`.

Gareth A Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.*, 185(2):604–613, February 2014.

Robin van der Lee, Marija Buljan, Benjamin Lang, Robert J. Weatheritt, Gary W. Daughdrill, A. Keith Dunker, Monika Fuxreiter, Julian Gough, Joerg Gsponer, David T. Jones, Philip M. Kim, Richard W. Kriwacki, Christopher J. Oldfield, Rohit V. Pappu, Peter Tompa, Vladimir N. Uversky, Peter E. Wright, and M. Madan Babu. Classification of intrinsically disordered regions and proteins. *Chemical Reviews*, 114(13):6589–6631, April 2014. ISSN 1520-6890. doi: 10.1021/cr400525m. URL `http://dx.doi.org/10.1021/cr400525m`.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL `http://dx.doi.org/10.1038/s41586-023-06415-8`.

Carter J. Wilson, Wing-Yiu Choy, and Mikko Karttunen. Alphafold2: A role for disordered protein/region prediction? *International Journal of Molecular Sciences*, 23(9):4591, April 2022. ISSN 1422-0067. doi: 10.3390/ijms23094591. URL http://dx.doi.org/10.3390/ijms23094591.

Peter E. Wright and H. Jane Dyson. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology*, 16(1):18–29, December 2014. ISSN 1471-0080. doi: 10.1038/nrm3920. URL http://dx.doi.org/10.1038/nrm3920.

Sijing Xia, Qin Chen, and Bing Niu. Cd28: A new drug target for immune disease. *Current Drug Targets*, 21(6):589–598, April 2020. ISSN 1389-4501. doi: 10.2174/1389450120666191114102830. URL http://dx.doi.org/10.2174/1389450120666191114102830.

K Yue and K A Dill. Inverse protein folding problem: designing polymer sequences. *Proceedings of the National Academy of Sciences*, 89(9):4163–4167, May 1992. ISSN 1091-6490. doi: 10.1073/pnas.89.9.4163. URL http://dx.doi.org/10.1073/pnas.89.9.4163.

A Related Works

Protein Design. Since the formalization of the inverse protein folding problem in the 1990s Yue & Dill (1992), the field of protein binder design has converged into a workflow where one first samples the protein backbone (RFDiffusion (Watson et al., 2023)) followed by sampling the sequence that would fold into such shape (ProteinMPNN (Dauparas et al., 2022; Goverde et al., 2024)). Afterwards, one can use a filtration step to validate the designs (Rosetta (Leaver-Fay et al., 2011) or AlphaFold2-Multimer (Evans et al., 2021)). Another approach is directly leveraging the power of deep learning algorithms, such as AlphaFold2 (Jumper et al., 2021), and hallucinating the optimal sequence through a gradient-based backward optimization (Goverde et al., 2023). Recently published BindCraft (Pacesa et al., 2024) has formalized such gradient-based workflow with impressive *in vitro* success for finding novel binders.

Intrinsically Disordered Proteins. IDPs are prelevant across the human proteome, accounting for roughly 30% of the human proteome (van der Lee et al., 2014). Their structural flexibility makes them take on multitude of complex functions, including cell signaling, transcriptional regulation, and chromatin remodeling (Babu et al., 2011; Wright & Dyson, 2014; Holehouse & Kragelund, 2023). α-synuclein has been shown to aggregate into toxic species that are implicated in the pathogenesis of Parkinson's disease. CD28 can provide a co-stimulatory signal necessary for T-cell activation and survival, making it relevant in modulating immune responses in various diseases (Xia et al., 2020). The tumor suppressor protein (p53) can have its function restored by inactiving its interaction with negative regulators like MDM2 prevalent in many cancers, resulting in the inhibition of tumor growth (Peuget et al., 2024). Lastly, Small Ubiquitin-like Modifier (SUMO) protein is an attractive drug target because its modification of various substrates is involved in key cellular processes, and dysregulation of SUMOylation has been linked to diseases including cancer (Kukkula et al., 2021).

Given the lack of structure, modeling IDPs is not straightforward. Deep learning models of protein structure, such as AlphaFold2 and ESMFold (Lin et al., 2023), have been shown to predict dynamical nature of residues indirectly through the pLDDT metric. This predicted quantity was originally designed to convey the confidence in the prediction of the underlying model, but it has been shown to also correlate with disorder (Guo et al., 2022; Wilson et al., 2022). One can, however, never decouple these two behaviours, so it remains indecipherable whether a folding model is unconfident or predicts an unstructured region. On the other hand, molecular simulations provide an orthogonal method to study IDPs, including their dynamics and mechanistic behaviour Tesei & Lindorff-Larsen (2023).

Enhanced Sampling Free Energy Calculations. Molecular simulations provide a physics-based method to estimating the binding affinity by sampling the free energy surface. Nevertheless, sampling is computationally prohibitive, hence one has to employ enhanced sampling methods, such as metadynamics (Laio & Parrinello, 2002), umbrella sampling (Torrie & Valleau, 1977), or replica exchange (Sugita & Okamoto, 1999). Some of these methods require manual fine-tuning and a careful choice of collective variables that are biased in order to increase the sampling of the slow mode, i.e. binding and unbinding. Well-tempered metadynamics (Barducci et al., 2008) is an extension of metadynamics that achieves smoother convergence by decaying the amount of deposited bias.

B Detailed Methods

B.1 Monte Carlo Protocol for Guided Optimization

Below we explain the guided optimization algorithm with formal definitions. We used ESMFold as the folding algorithm (Definition B.2), but any structure predictor, regardless of its implementation could be used as a black-box. We formalize the Monte Carlo acceptance criteria according to the (loss) energy function E. We choose a mutation protocol that samples uniformly, meaning $\mathcal{U}$ selects a random residue position i in the sequence s and replaces s_i with a different amino acid, chosen uniformly among the 20 types, i.e. each amino acid is chosen with probability $\frac{1}{20}$. This in turn leads to a symmetric proposal distribution, and thus the transition probabilities $q(s \mid s')$ and $q(s' \mid s)$ cancel out in the Metropolis acceptance probability (Definition B.4). Algorithm 1 shows the complete implementation of the optimization procedure, primarily the simulated tempering, using the definitions below. Note that after each run through a low- and high-temperature regimes we revert to the best state found so far. This means we no longer obey detailed balance, which is on purpose, as we use Monte Carlo as an optimization procedure, and not for sampling distributions and their associated averages.

Definition B.1 (Energy Loss Function) *Let $X \in \mathcal{X}$ be a folded structure. The energy function*

$$E : \mathcal{X} \to \mathbb{R}$$

is defined as a linear combination of loss terms:

$$E(X) = \sum_i w_i \, L_i(X),$$

where $L_i(X)$ denotes the ith loss term and w_i its corresponding weight.

Definition B.2 (Folding Function) *Let $s \in \mathcal{S}$ be a protein sequence. The folding function*

$$\mathcal{F} : \mathcal{S} \to \mathcal{X}$$

maps a sequence s to its corresponding structure $X = \mathcal{F}(s)$.

Definition B.3 (Mutation Operator) *Let $s \in \mathcal{S}$ be a protein sequence, where*

$$\mathcal{S} = \{s = (s_1, s_2, \ldots, s_L) \mid s_i \in \mathcal{A}\}$$

and the alphabet $\mathcal{A}$ consists of 20 amino acid identities. The mutation operator

$$\mathcal{U} : \mathcal{S} \to \mathcal{S}$$

generates a candidate sequence $s' = \mathcal{U}(s)$ according to a specified mutation protocol.

Definition B.4 (Metropolis–Hastings Operator) *Let $s \in \mathcal{S}$ be the current protein sequence, and let $s' = \mathcal{U}(s)$ be a candidate sequence generated by the mutation operator $\mathcal{U}$. The Metropolis–Hastings operator $\mathcal{M}(s;T)$ at the effective temperature $T > 0$ is defined as*

$$\mathcal{M}(s;T) = \begin{cases} s', & \text{with probability } \alpha(s \to s';T), \\ s, & \text{otherwise,} \end{cases}$$

where the acceptance probability is given by

$$\alpha(s \to s';T) = \min\left\{1,\, \exp\left(-\frac{E(X') - E(X)}{T}\right) \frac{q(s \mid s')}{q(s' \mid s)}\right\}.$$

Here, $X = \mathcal{F}(s)$ and $X' = \mathcal{F}(s')$ are the folded structures corresponding to sequences s and s', respectively. The function $E(X)$ gives the energy of a folded structure X, as defined in Definition B.1. Lastly, $q(s' \mid s)$ is the transition probability of generating sequence s' from s under $\mathcal{U}$. Similarly, $q(s \mid s')$ is the probability of proposing the reverse move.

In Table 1, we provide the energy terms of function E for this specific application of designing binders to disordered epitopes. The exact weights w_i need to be carefully considered and tuned to achieve the desired behaviour. We have conceived of the terms and their weights through rational design and trial-and-error.

Algorithm 1 Simulated Tempering for Protein Sequence Design

```
Input:
    • Initial sequence $s_0$
    • Temperatures: $T_{\text{low}}$, $T_{\text{high}}$
    • Number of low-temperature steps $n_{\text{low}}$, high-temperature steps $n_{\text{high}}$
    • Total number of sweeps $\nu_{\text{total}}$
Initialize: $s \leftarrow s_0$, $s^{\text{best}} \leftarrow s_0$, $X^{\text{best}} \leftarrow \mathcal{F}(s^{\text{best}})$
for $\nu = 1$ to $\nu_{\text{total}}$ do
    for $i = 1$ to $n_{\text{low}}$ do                          ▷ Low-temperature optimization
        $s \leftarrow \mathcal{U}(s)$                              ▷ Mutate the sequence
        $X \leftarrow \mathcal{M}(s; T_{\text{low}})$
        if $E(X) < E(X^{\text{best}})$ then
            $s^{\text{best}} \leftarrow s$, $X^{\text{best}} \leftarrow X$          ▷ Update best state if lower energy
        end if
    end for
    for $i = 1$ to $n_{\text{high}}$ do                         ▷ High-temperature exploration
        $s \leftarrow \mathcal{U}(s)$                              ▷ Mutate the sequence
        $X \leftarrow \mathcal{M}(s; T_{\text{high}})$
        if $E(X) < E(X^{\text{best}})$ then
            $s^{\text{best}} \leftarrow s$, $X^{\text{best}} \leftarrow X$          ▷ Update best state if lower energy
        end if
    end for
    $s \leftarrow s^{\text{best}}$, $X \leftarrow X^{\text{best}}$                  ▷ Reset to best state found so far
end for
return $s^{\text{best}}$, $X^{\text{best}}$          ▷ Return the best sequence and corresponding folded structure
```

Table 1: Energy terms of the protein design guided optimization of a binder targeting an IDR.

Term	Type	Metric	Description
L_1	0-body	global pTM	Guides the complex toward high-confidence structures.
L_2	0-body	global pLDDT	Encourages soluble and experimentally verifiable sequences.
L_3	1-body	hydrophobics	Reduces hydrophobic residues on the surface, favouring soluble sequences aiding expression and experimental verification.
L_4	2-body	interface PAE	Minimizes interface predicted alignment error, i.e. having the binder and IDR behave as a single body.
L_5	2-body	average distance	Penalizes large distances between binder and the IDR.
L_6	1-body	secondary structure	Promotes formation of a secondary structure (alpha-helices, beta-sheets) on the IDR.
L_7	1-body	local pLDDT	Enforces further ordering and rigidification of the IDR.

B.2 Free-Energy Calculations Protocol for Binding Affinity

To simulate the system, we used OpenMM (Eastman et al., 2017) with PLUMED (Tribello et al., 2014). The simulations were run for about $1\,\mu\text{s}$, where convergence was checked by relatively low deposition rate of the bias, by visual inspection of the change in the FES, and by converged computation of the binding affinity. After fixing the final structures from ESMFold with `pdbfixer`, we minimize the bound complex with L-BFGS in two phases – first, we restrain the non-hydrogen atoms and minimize the rest of the complex, second, we minimize all atoms together. We run 10 ns of equilibriation in the NVT ensemble. Afterwards, we extract the final structure and compute the CV definition. The contact map c is defined by any carbon alpha atoms closer than $1.5\,\text{nm}$. We used the `RATIONAL` switching function defined as

$$\chi(x) = \frac{1 - \left(\frac{x}{x_0}\right)^6}{1 - \left(\frac{x}{x_0}\right)^8} \tag{4}$$

where the threshold x_0 is set to $0.8\,\mathrm{nm}$. The exponents are chosen to achieve a slow decay for large values of x to avoid a substantial degeneracy of conformations at $c = 0$. To define the distance r, we manually annotate the IDR for each of the proteins based on the experimental structures. We deposit the bias every 500 steps of integration with a timestep of $2\,\mathrm{fs}$. We choose the decay of the bias deposition rate by setting the bias factor to 10. As we used an implicit solvent, we do not utilize the periodic boundary condition anywhere. We impose an upper wall bias on r defined as

$$V_{uw}(r) = \begin{cases} \kappa\,(r - r_{uw})^6 & \text{if } r > r_{uw} \\ 0 & \text{if } r \leq r_{uw} \end{cases} \tag{5}$$

where we set κ to $1000\,\mathrm{kJ\,mol^{-1}}$ and r_{uw} is programmatically defined for each protein target as

$$r_{uw} = 0.38 \times \frac{L_{\text{binder}}}{2} + L_{\text{IDR}} - 1 \tag{6}$$

where L_{binder} is the length of the peptide-binder and L_{IDR} is the length of the IDR in terms of number of residues. All values here are in nanometer units.

C Further Results

We show examples of three trajectories in the CV space in Fig. 4, highlighting some of the issues and behaviour we observe. We primarily focus on r during analysis. For ASYN-A, there is only two binding/unbinding events, which is insufficient to make substantial claims about the results. For CD28-B, similar to ASYN-A, the systems spends the majority of the simulation time very closely bound, but once it fills up the basin to a certain extent, it is able to escape the basin and start to diffuse in the unbound region as well. Lastly, SUMO-1A shows many binding/unbinding events, which can make us more confident in the statistics of the FES reconstruction, but it still spends at least half of the time in the local basin. All these trajectories suggest we need to increase the exploration and sampling speed, even if it is at the cost of larger fluctuations in the estimated ΔG as a function of simulation time. Otherwise our results are not fully conclusive.

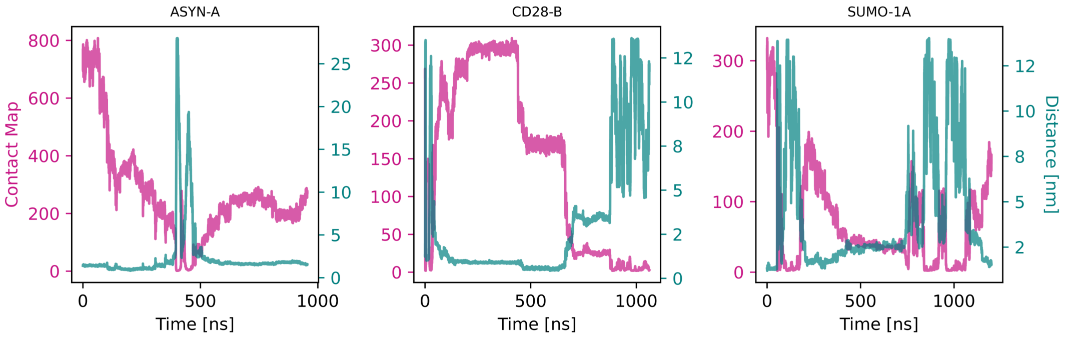

Figure 4: Trajectories of the collective variables for three systems during the well-tempered metadynamics simulation. The vertical left-axis (magenta) shows the contact map c and the vertical right-axis shows the evolution in the distance r. The simulation times differ as these runs were computed for a walltime of 6 days, rather than a pre-defined set of integration steps.