On the Potential of Machine Learning to Examine the Relationship Between Sequence, Structure, Dynamics and Function of Intrinsically Disordered Proteins

Kresten Lindorff-Larsen and Birthe B. Kragelund

Structural Biology and NMR Laboratory & Linderstrøm-Lang Centre for Protein Science, *Department of Biology, University of Copenhagen. Ole Maaløes Vej 5, DK-2200 Copenhagen N, Denmark*

Correspondence to : *lindorff@bio.ku.dk (K. Lindorff-Larsen), bbk@bio.ku.dk (B.B. Kragelund), @LindorffLarsen (K. Lindorff-Larsen), @BBKrage (B.B. Kragelund),*
https://doi.org/10.1016/j.jmb.2021.167196
Edited by Sheena E. Radford

Abstract

Intrinsically disordered proteins (IDPs) constitute a broad set of proteins with few uniting and many diverging properties. IDPs—and intrinsically disordered regions (IDRs) interspersed between folded domains—are generally characterized as having no persistent tertiary structure; instead they interconvert between a large number of different and often expanded structures. IDPs and IDRs are involved in an enormously wide range of biological functions and reveal novel mechanisms of interactions, and while they defy the common structure-function paradigm of folded proteins, their structural preferences and dynamics are important for their function. We here discuss open questions in the field of IDPs and IDRs, focusing on areas where machine learning and other computational methods play a role. We discuss computational methods aimed to predict transiently formed local and long-range structure, including methods for integrative structural biology. We discuss the many different ways in which IDPs and IDRs can bind to other molecules, both via short linear motifs, as well as in the formation of larger dynamic complexes such as biomolecular condensates. We discuss how experiments are providing insight into such complexes and may enable more accurate predictions. Finally, we discuss the role of IDPs in disease and how new methods are needed to interpret the mechanistic effects of genomic variants in IDPs.

Introduction

Intrinsically disordered proteins (IDPs) constitute a broad and relatively heterogeneous class of proteins that have in common that they do not adopt a well-defined three-dimensional structure, at least in the absence of binding partners. This in itself is not a very strict definition because also natively folded proteins are dynamic. Experimentally, disordered proteins are often characterized using a range of biophysical measurements that typically reveal the presence of transiently formed secondary structure elements and occasionally weak, transient longer-range interactions. Analogously, many proteins have regions of intrinsic disorder interspersed within or between folded domains, and in many ways these intrinsically disordered regions (IDRs) behave similarly to IDPs, and in general we will refer to both as IDPs.

The flexibility and dynamics combined with an extended surface area endow IDPs with an ability to adapt, a trait that is often key to their biological function, either because it enables them to bind to multiple different proteins or because the intrinsic dynamics may affect both binding kinetics and thermodynamics. This dynamics, however, also makes it difficult to characterize IDPs both experimentally and computationally.

It was early recognized that the amino acid composition and sequences of IDPs differed in several ways from those of folded proteins. Thus, aided by databases containing experimentally-validated IDPs[1] a large number of prediction methods have been developed to predict protein disorder from sequence alone.[2,3] While overall very successful, such prediction methods inherently need to deal with the heterogeneity in what is considered a disordered protein, including large differences in biological context (complexes, post-translational modifications, etc).

A complementary approach to study IDPs is to characterize the conformational ensembles that they populate. In certain favourable cases, computational methods can on their own predict some conformational properties. Often, however, a detailed and accurate characterization requires integrating one or more types of biophysical experiments with computational methods to collectively derive a collection of structures that represent the conformational heterogeneity[4,5] or dynamics[6] of the protein. A number of such approaches exist and some of the resulting ensembles are collected in the Protein Ensemble Database (PED[7]), by analogy to the Protein Data Bank (PDB[8]), which, however mostly contains more well-defined protein structures or IDPs in complexes.

For folded proteins, Anfinsen's observations[9,10] suggested that it should be possible to predict the three-dimensional structure of a folded protein based only on its primary structure and its interaction with the environment. Over the years, this has led to the field of protein structure prediction, and a plethora of innovative approaches to predict structure from sequence. The accuracy of such methods is evaluated during the biennial critical assessment of structure prediction (CASP) experiment. While there have been continued improvements in the ability to predict structures over the years, the last two installments of CASP (CASP13 (2018) and CASP14 (2020)) have witnessed some substantial and impressive advances in accuracy, in particular in the so-called template-free modelling.[11,12] While a number of developments have contributed to this, we here highlight three. First, in the last decade a number of methods have been developed to extract structural information from multiple sequence alignments (MSAs) e.g. through the analysis of correlated substitutions during evolution.[13–18] Second, there has been an explosion in the number of sequences available making such sequence-based approaches useful and applicable to a wider number of proteins. Finally, various deep-learning approaches have been used to 'learn' the complicated relationship between the amino acid sequence (or MSA) and the three-dimensional structure. Most visible has been the development of the AlphaFold approach[19] in CASP13 and AlphaFold 2 in CASP14,[20] although many other groups have also contributed to these developments including among others Xu[18]; Zheng et al.[21]; Kandathil et al.[22]; AlQuraishi[12,23]; Torrisi et al.[24]; Yang et al.[25]; Baek et al.[26]

Motivated by our own research, this perspectives paper begins by examining whether such methods can be used to predict information about the (highly conformationally heterogeneous) three-dimensional 'structures' and ensembles of IDPs using only the primary structure as input. We also discuss how machine learning methods may aid in integrative modelling of the conformational ensembles of IDPs, referring the reader to a separate paper on this topic more broadly.[27] We discuss the unique properties of IDPs in complexes, both those formed via short linear motifs and in larger assemblies and biomolecular condensates, and how new sources of data may be useful to develop better prediction methods. Finally, we discuss the role of IDPs in human diseases and how an improved understanding of the relationship between sequence, structural properties, formation of complexes and function may help in this area (Figure 1). Overall, we highlight a number of challenges that are particularly relevant for IDPs, and some of the questions that might be addressed by combining machine learning methods with experiments and other computational approaches. For an introduction to deep learning and other machine learning methods we refer the reader to recent detailed reviews.[28–32]

Towards Improved Conformational Ensembles

From sequence to structure

Before discussing potential applications of machine learning approaches to IDPs, we first describe very briefly some of the key steps that have led to improved structure prediction of folded proteins, but note that this description is far from comprehensive. One key ingredient has been the ability to extract structural information, for example in the form of contacts,[13,15] distance distributions[19] or distributions over distances and orientations,[25] from the analysis of MSAs. This work builds on earlier ideas that correlated mutations observed through evolution contain information about the proximity of amino acids in the three-dimensional structure,[33–35] but required more advanced analysis methods, including global analysis methods,[13–18] as well as increased number of sequences to reveal their full potential. The structural information obtained from the MSAs can then be used to guide structure determination using a range of methods to obtain three-dimensional models. Indeed, while many methods have been shown to help improve the accuracy of contact prediction, it is not clear that improved contact prediction always leads to substantially improved models of three-dimensional

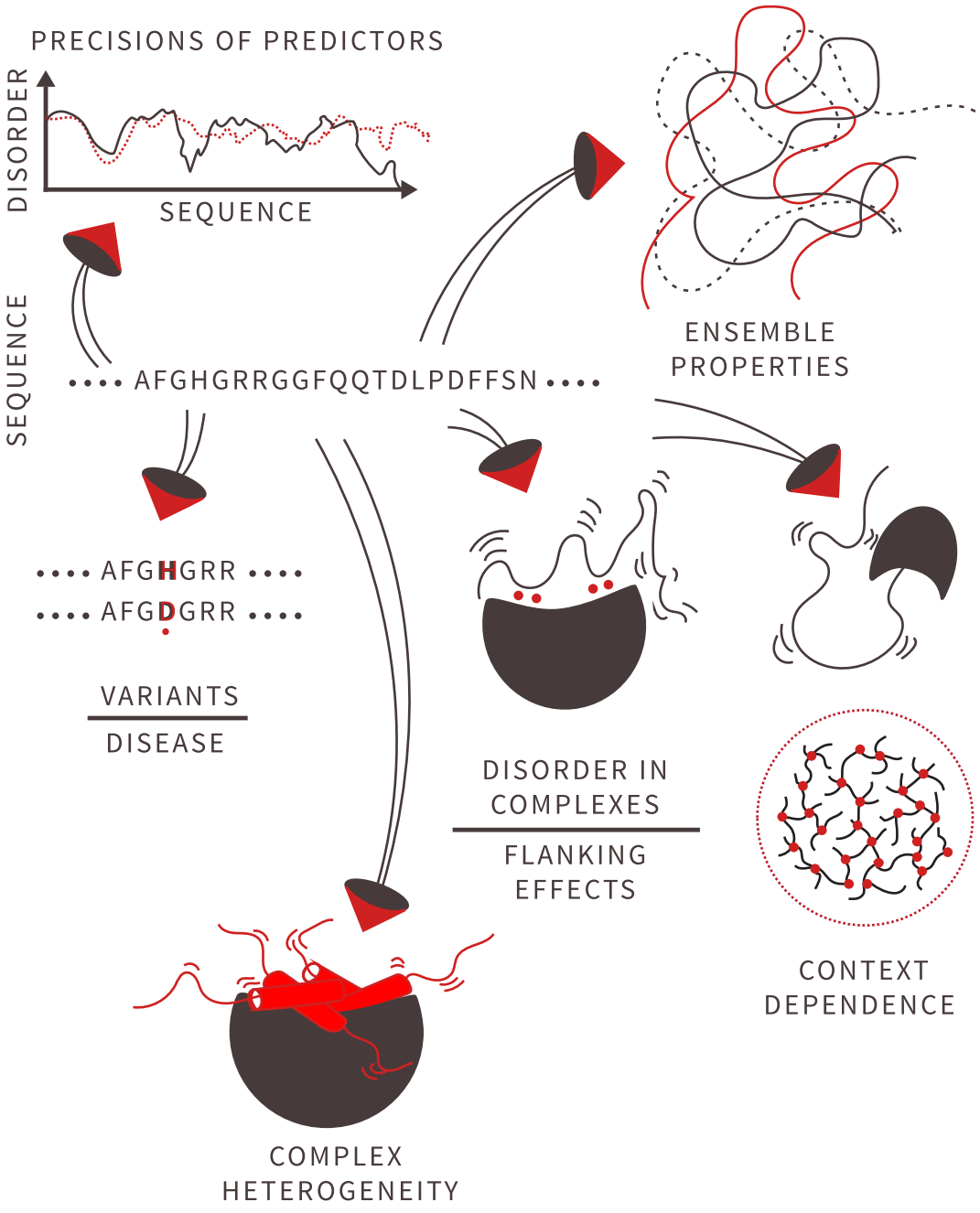

Figure 1. An overview of the relationship between sequence, structure, dynamics and function of IDPs, and how the inherent disorder also in the complexes affects the use of machine learning approaches. Some of the challenges to understand these relationships include improving predictions of disorder, describing ensemble properties, and finding ways to include complex heterogeneity and context effects. Finally, it is still not clear how many disease variants in IDPs lead to disease and by which mechanisms.

structures.[36] More recent improvements in the AlphaFold 2 method use a novel architecture termed the Evoformer to couple structure and sequence information, and predicts the three-dimensional structure by an iterative procedure.[20] AlphaFold 2 and the related RoseTTAFold approach[26] are so-called end-to-end models,[23,37,12] which have been trained to predict structures directly from sequences using large sequence and structure databases.

We return now to the question of how such methods might be applied to IDPs, and note several obstacles that need to be overcome. First, the goal should not generally be to predict a single structure from the sequence, but rather an ensemble of structures that represents the

dynamic nature of IDPs. We here note in passing that some predictions provide multiple structures, but these generally represent the uncertainty of the prediction rather than the intrinsic heterogeneity and dynamics of the structure. Second, it is often difficult to generate high quality and deep MSAs of IDPs, in particular for those of low sequence complexity, or when conserved folded domains used to anchor alignments are lacking. Third, we do not have available a large number of structural ensembles that can be used to benchmark let alone train prediction methods. Thus, in contrast to the case for folded proteins where the PDB contains approximately 155.000 entries with protein structures, the PED contains ca. 200 ensembles.

There is a long history of using structural information gathered from loop regions of folded proteins to predict local conformational properties of disordered peptides and proteins,[38–40] and these approaches are in some ways conceptually related to the use of fragments for protein structure prediction.[41] Such approaches can be generalized and turned into probabilistic models using for example hidden Markov models[42] or dynamic Bayesian networks.[43] Recent developments have used a database of tri-peptide fragments to predict local structural properties in IDPs,[44] or used molecular simulations of peptide fragments to create models of full-length IDPs.[45] Combining such approaches may be a fruitful path towards constructing structural models of IDPs that contain transiently formed local structures.

Approaches are also being developed to predict the relationship between sequence and more global conformational ensembles of IDPs. In one such study, the concept of amino acid co-evolution was applied to predict contacts in IDPs from MSAs.[46] In several of the proteins analysed, the predicted contacts could be shown to coincide with key contacts observed within an IDP when bound in a complex to a folded protein,[46] thus demonstrating that the same principles that have been used so successfully for folded proteins have the potential to provide insight into IDPs, at least when they form complexes. In another study, we used similar sequence analyses of the disordered protein CsgA.[47] Here, we found a strong pattern of predicted contacts that corresponded to a folded amyloid-like state that CsgA forms. In this case, these contacts are preserved by evolution because CsgA forms a functional amyloid that is beneficial to the bacteria. In a recent study, co-evolution between regions in a disordered protein was used to predict long-range interactions that were subsequently supported by NMR measurements.[48] While these and related studies suggest that sequence analysis might contain information that can be extracted to learn about the structures of IDPs, they have so far mostly revealed information about folded states that the IDPs might adopt or, perhaps, local secondary structure in the disordered states.[46] Here we note that there has been a substantial amount of work on predicting local structure in flexible peptides and proteins, but that our focus here is more generally on both local and global structures. An important point is that the inter- and intramolecular interactions that are important for the function(s) of a protein will likely manifest themselves most strongly in the evolutionary record. This in turn means that co-evolving pairs of amino acids or regions will be 'weighted' by their functional importance, so that e.g. a small population of a functionally important state might create a stronger signal than a highly populated state of less importance. Furthermore, certain functionally relevant properties, such as maintaining a preferred orientation or distance between linear motifs, may be encoded in the sequence in ways that are difficult to extract from standard sequence analyses,[49] and may thus be easily overlooked.

One of the limiting factors is also that we do not have a well-developed framework to discuss and quantify the relationship between sequence and ensemble properties. Indeed, as discussed above, protein disorder covers a continuum ranging from almost folded, but flexible and compact globules to chains that appear as statistical random coils. Because these proteins are best described in statistical terms, one approach is to bypass the three-dimensional structure all-together and predict key structural parameters directly from sequence. A number of such studies have focused on discovering the rules that govern the relationship between amino acid composition and patterning and, in particular, compaction of IDPs by parameterizing computational methods using experiments.[50–56] Recently, Cohan et al.[57] developed a conceptual framework to examine such sequence-ensemble relationships more generally. Presumably, such approaches as well as more advanced computational methods and expanded sets of experimental data will be needed to predict structural properties beyond compaction and local structure in IDPs.

Beyond sequence alignments

As discussed above, one of the main sources of information in current protein structure prediction comes from MSAs. While a detailed discussion of methods used to generate MSAs is beyond the scope of this paper, we note here that they are generally constructed based on the assumption of positional homology,[58] i.e. that a specific position in one sequence corresponds to a specific position in a homologous sequence. While this in turn is often the case for folded proteins, the situation in IDPs appears more complicated. Moreover, some IDPs diverge by gene duplication[59] or are only found in some species,[60] and indels (insertion and

deletions) are found to be frequent in IDPs,[61] obscuring alignments further.

Recently, several ideas and methods from natural language processing have been applied to modelling, interpreting and predicting properties from protein sequences.[62–66] While such methods are still difficult to interpret and expensive to train, they might help study proteins for which it is difficult to construct good alignments. Nevertheless, it still appears that MSAs contain substantial information that is not easily extracted from such language models,[63] and indeed combining the two can be advantageous.[20,67]

Initial applications of language models to IDPs suggest that they could be very useful in cases where one cannot construct accurate MSAs.[64] It will be interesting to explore whether these models can be used to discover new rules that govern IDP sequences, and extract structural information from them. For example, as discussed above, properties such as the level of compaction and local structural motifs can to some extent already be predicted from sequence. It will, however, be interesting to explore how such methods can be improved—also for other properties such as long-range interactions or post-translational modifications—by developing new methods to represent sequences that are not based on the positional-conservation dogma that implicitly underlies many structure-prediction methods for folded proteins.[68–70]

Forward models for interpreting experimental data

Many methods for predicting the structures of folded proteins are explicitly or implicitly based on the availability of thousands of labelled sequence-structure pairs in the PDB, used either for parameterization, training or validation. As discussed above, we have much fewer experimentally-derived conformational ensembles available for IDPs, and in this and the following section we discuss how machine learning methods can provide potential advances in modelling IDPs.

While a plethora of methods exist for modelling conformational ensembles of IDPs, they are typically based on either biasing molecular simulations using experimental data or on selecting structures from a pre-generated ensemble to improve agreement with experiments.[4,5,71,72] In these methods it is important that the underlying dynamics and conformational averaging are treated correctly. For folded and rigid proteins one often transforms the experimental measurements into geometric restraints that are then applied during or after simulations. While this is possible for IDPs, a more general approach involves calculating experimental quantities from conformations and ensembles and comparing these to experiments. This calculation relies on so-called 'forward models', i.e. algorithms to calculate experimentally-accessible quantities from conformational ensembles.

To give a concrete example, small-angle X-ray scattering (SAXS) experiments are often used to probe the compaction of an IDP, often quantified by the radius of gyration (R_g). One approach might therefore be to extract the R_g from experimental SAXS data and generate a conformational ensemble with the same (average) R_g. Such an approach would, however ignore solvent contributions to the experimental measurements[73] as well as information from a wider range of scattering angles.[74–76] Also, when multiple sources of experimental data are used it is important to treat errors and ensemble-averaging correctly,[77] and that becomes more difficult when working with quantities that are transformed values of the experimental measurements. Instead of using R_g, the more common approach is to use a forward model to calculate SAXS data from each conformation in an ensemble, and then compare the calculated average with experiments.[78] A number of such forward models exist for SAXS experiments, that differ in how they treat solvent effects, as well as in accuracy and computational efficiency.[79] Importantly, different forward models may give different views of a conformational ensemble,[73,80] because—depending on the relationship between structure and measurement—different ensembles will be needed to agree with the experiments.[81]

There are at least two different approaches to develop forward models, and these approaches can be combined. The first uses the basic physical principles that underlie the experiment to link structure and observable. Again using SAXS as an example, one of the most commonly used methods to calculate SAXS data from experiments (Crysol[82]) calculates SAXS intensities from the scattering amplitude of the protein in vacuum as well as a model for the solvent contribution. The former is in turn based on empirically-derived form factors whereas the solvent contribution is parameterized using two parameters capturing the average solvent displaced by surface atoms and the excess density of the solvation layer. For folded and globular proteins, these two parameters are often fitted based either on a known structure or a model for the structure. Thus, the calculations of SAXS data from structural models may involve combining such a physical model while fitting one or a few empirical parameters against the experimental data.

Other forward models, such as for example methods that are used to calculate protein chemical shifts from protein structures, are also often based on physical principles combined with empirical terms, but have a much larger number of parameters that need to be fit to experiments.[83–86] This in turn is often based on data

for folded proteins for which both high resolution structures and assigned chemical shifts are available. The mathematical function that connects structure and chemical shift is highly complex and has its roots in quantum mechanics. Thus, an alternative approach to express this relationship is to use neural networks.[87–90] One assumption underlying most of these approaches is that the experimental chemical shifts (which are time and ensemble averaged quantities) can be predicted accurately from a single structure. While that may be sufficient to study the rigid regions of folded proteins, other approaches may be needed to deal with more flexible parts.[91–93] For IDPs where the chemical shifts represent a relatively broad ensemble of conformations and with more homogeneous chemical environments, further developments may be required to calculate accurate chemical shifts (relative to the small deviations from random coil values) and to extract structural information from these experiments.[94]

Semi-empirical forward models such as those described above can be extremely difficult to develop for IDPs. This is because we rarely have sets of proteins for which we accurately know the conformational distribution derived independently from the set of measurements that one aims to develop a forward model for. Thus, most structures and ensembles determined for IDPs are implicitly based on forward models trained and validated on folded proteins. In the case of SAXS data this means, for example, that we often make the assumption that the solvation of a disordered protein is similar to that of a natively folded protein, and that the solvation properties are independent of the structure. While this may be true, this is very difficult to validate. One approach towards this goal may be to use more refined forward models to derive the ensembles,[79,95] to reparameterize simplified models using such more refined methods,[73,96] or to refine ensembles and forward models in a self-consistent manner.[81,97,98]

How can machine learning methods aid in the further development of forward models, and thus in our ability to derive conformational ensembles from experimental data? As described above, neural networks have already been used extensively to parameterize a function used to calculate chemical shifts, and we expect such methods will become refined and extended to a wider set of experiments. Machine learning methods have also been developed to extract shape information from SAXS experiments[99] though, to our knowledge, not as forward models. Similarly, a deep neural network based approach has been developed to process and extract structural information from electron paramagnetic resonance (EPR) experiments.[100] Finally, a neural network was recently trained using quantum calculations to predict data from infrared absorption spectroscopy.[101] Circular dichroism (CD) spectroscopy is widely used to study IDPs,[102] yet calculating CD spectra from conformational ensembles of IDPs is difficult and generally based on 'basis spectra' derived from folded proteins often via secondary structure classification.[103] We envisage that machine learning methods can aid in generalizing such approaches towards IDPs.[104] In addition to the improved accuracy potentially afforded by such machine-learning-based forward models, they may also have other advantages such as rapid evaluation and differentiability, both of which can be important when determining conformational ensembles from experimental data.

Improving energy functions for simulating IDPs

Returning to the problem of predicting conformational properties and ensembles of IDPs from sequence we now explore how experiments and machine learning methods may be combined to improve conformational modelling. Ensembles generated either directly from molecular simulations or from integrative modelling using experiments are dependent on the quality of the physical models used in simulations.[71] Thus improved force fields and energy functions both enable more accurate predictions of conformational properties from sequence, but also makes integrative methods more robust.[105–109] While molecular simulations may not be the most computationally efficient approach to predict conformational properties from sequence, it can serve as a benchmark and starting point for developing other approaches.

In recent years there have been substantial improvements in explicit solvent, all-atom force fields used to study the structure and dynamics of IDPs,[110–115] and these improvements have been derived both by better quantum-level calculations and empirical fitting to experimental data. Conformational sampling of IDPs, in particular long IDPs or their complexes, remains a substantial challenge, and therefore implicit solvent or coarse-grained methods are sometimes used.[115] These can in turn be parameterized using either bottom-up (based on more accurate models) or top-down (from experiments) approaches, or indeed a combination of the two.[116]

Some time ago we developed an automated approach to parameterize force fields based on experimental data and applied it to develop a coarse-grained model for IDPs.[117] The basic idea, which had also been explored earlier for force field development,[118–121] is to sample force field parameter space and to optimize the parameters by comparing simulation results against experiments. Using a Bayesian framework it is possible to combine the experiments with other sources of information, and one may use reweighting techniques to speed up parameterization.[117] In some sense, this approach can be considered a machine learning

approach for learning force field parameters from experimental data. Later, similar ideas have been developed and applied to the problem of optimizing all-atom force fields against experimental data.[122–126] The ideas developed by Norgaard et al.[117] have been extended and applied to larger sets of experimental data to construct coarse-grained[127–130] and all-atom[131] models for IDPs. In these approaches, the experimental data are used to refine or parameterize a fixed functional form for the force field and how the protein structure is represented. Recently, a number of machine learning approaches have been developed and used both to construct force fields and to develop coarse-grained representations,[132–136] and we expect such approaches could have a substantial impact on our ability to simulate IDPs at various resolutions.

The methods described above suggest that machine learning methods may be used both to improve our ability to calculate and interpret experimental observables and to parameterize computational models for IDPs directly against experiments. Common to both problems is the focus on interpreting and using the experimental measurements. This is key because the procedure when going from experimental measurements to conformational ensembles involves approximations and loss of information. In the context of folded proteins, this is generally thought to be less of a concern, and the three dimensional coordinates are often a relatively good representation of the system and of the data. This in turn means that structure prediction methods can be trained or benchmarked on the protein structures (coordinates) rather than the experimental measurements used to derive them. We expect that this will not be the case for IDPs, and instead we suggest that machine learning methods for structure prediction should be benchmarked or trained directly on experimental data similarly to the force fields described above. Related, it is still an open question to what extent the complicated models used to predict protein structures from sequence internally represent the physics of proteins[137], and thus training models for structure prediction from experiments may end up being comparable to training molecular force fields.

Towards Predicting Interactions and Complexes

Identifying short linear motifs

An noted above, the primary structures of disordered proteins are generally not very well conserved. Nevertheless, their sequences do carry important information about their function, clues to which can be derived from direct sequence analysis and alignments. Although complicated to perform, and often assisted by manual refinements and adjustments, it is still possible to construct MSAs of disordered proteins and from these alignments identify conservation hotspots in otherwise poorly conserved regions. In such cases, few positions—as little as between two and five—are highly conserved across species and found to be distributed across a confined stretch of approximately a dozen residues. These conserved sequence stretches represent so-called Short Linear Motifs (SLiMs).[138–140] SLiMs are recurrent, and the same SLiM can be identified in different, seemingly unrelated proteins conferring binding to specific partner proteins or other biomolecules. They constitute interactions sites, and the conserved residues are essential contact points that form part of the complex interface, and are thus essential to IDPs and their interactome. Today, more than 2000 SLiMs have been identified and annotated, and more candidate SLiMs reported with many assembled in the Eukaryotic Linear Motif database.[141–143] It is, however, difficult to identify new SLiMs, define SLiM properties and specificity, and to annotate their functions. Below we discuss some areas where new experimental approaches and machine learning method may be integrated to shed further light on these problems.

One problem when applying machine learning methods to predict new instances from known SLiMs is that, typically, only a small number of experimentally verified cases are reported for each individual SLiM. This is mainly because methods for SLiM identification have been low-throughput and have relied mostly on bioinformatics approaches with subsequent biochemical and biological testing,[144] or through integrating computation and medium throughput experiments.[145] More recently however, new high-throughput approaches have been used to define, expand and refine SLiMs. Examples include combining structure-based shape complementarity analysis and proteome-wide affinity purification mass spectrometry[146] and proteomic peptide phage display (ProP-PD), a method for simultaneous proteome-scale identification of SLiM-mediated interactions and foot-printing of the binding region with amino acid resolution.[147,148] Recent work addressed ≈1,000,000 overlapping peptides covering the entire human disorderome in a single binding assay.[149]

The generation of these large data sets provides new possibilities to train various types of prediction methods. Thus, a model has been trained to discriminate experimentally determined 14-3-3-binding SLiMs from non-binding phosphopeptides[150] and a Random Forrest model was trained on a high-throughput phage display data set collected for low-specificity SLiM binding to S100A5 identifying recognition rules based on features of hydrophobicity and shape complementarity as primary determinants.[151] Likewise,

prediction of binding regions in longer IDPs have been aided by the use of a trained bidirectional recurrent neural network, combining sequence, predicted secondary structures, docking scores and predicted disorder to improve the prediction.[152] Thus, machine learning approaches may help identify features that define SLiM binding and specificity, and are often used together with 3D structures, as done e.g. for PDZ binding peptides[153]; a case where also more confident negatives could be included. Similar improvements in the number of reliable true-negatives were achieved in a reevaluation of a high-throughput binding data of SH2-pTyr interactions.[154] Currently such efforts are limited by a relative small number of large data sets, and further that larger scale experiments often address already known SLiMs. Another problem when developing prediction methods is the relatively low number of negative examples in many data sets, which has an impact on the number of false positives provided by the resulting models. Thus, ways to improve this issue are clearly needed. Once addressed, however, machine learning approaches could substantially further our understanding of SLiM-based interactions by enabling extraction of features of interaction that expand our view on sequence properties that determine SLiMs. Such features, which may also relate to conformational features, may help move beyond the expectation and limitation provided by a defined SLiM-sequence space. Indeed, SH2 domains, which are known to bind phospho-tyrosine ligands, have been shown to be able to also accommodate glutamates,[155] which would not be expected solely from the SLiM definition, and therefore not typically included in fragment based database designs for machine learning purposes.[156] Finally, results from machine learning approaches may have the further benefit of contributing to the development of new vocabulary to describe SLiM-based interactions and uncover novel rules for interactions by IDPs.

Annotating function to short linear motifs

Although many SLiMs have been classified, it has been estimated that the human proteome counts more than 100,000 SLiMs, leaving most SLiMs unidentified.[157] Needless to say, each newly discovered potential SLiM in a disordered protein needs experimental verification as well as annotation; a task that remains a huge effort and experimentally highly challenging. So, although identification of their presence may be relatively accessible, and even aided by machine learning approaches, functional annotation of SLiMs remains an obstacle. Current high-throughput approaches for functional annotation have used in vivo SLiM-dependent proximity labeling, and in silico modeling of motif determinants to uncover new interactors,[158] as well as ProP-PD.[147,148]

There are, however, a number of complications that may make it difficult to apply machine learning methods to aid in annotating the function of newly discovered SLiMs. The same SLiM may in one protein be embedded in a sequence that folds to an α-helix when bound, whereas in another protein, the same SLiM may form an extended structure or a β-strand when bound. One example is provided by a set of plant transcription factors that all bind to the $\alpha\alpha$-hub domain RST from RCD1 through the RST-binding SLiM. Here, the transcription factors individually form either a helix, an extended or disordered SLiM structures in the complexes.[144,159] Thus, inherent to SLiMs is a certain plasticity in the position of the key conserved residues that form the critical contacts with the binding partner. Furthermore, the same sequence stretch within a disordered protein can have overlapping SLiMs and form biologically relevant complexes with different partners. There are several examples of this, including the transcriptional activation domains of the tumor suppressor protein p53, which each have many different partners binding to the same overlapping region.[160,161]

Once the target protein is known, additional complications can arise. One is SLiM 'reversibility', in which two proteins with the same SLiM binds in opposite directions to the same partner, as shown for Sap25 and REST binding to Sin3-PAH1[162] and peptide binding to MHC class II molecules.[163] This directly points to the SLiM context as carrying additional functional relevance.[164] Indeed, it has been shown that the context may have both positive and negative effects on binding through charge attraction and repulsion,[165,166] and it may contribute to allosteric regulation.[144,167,168] Thus, the influence of context on SLiM-based interactions is emerging as functionally important and with a large potential relevant to drug targeting.[169] However, these flanking sequences and regions are often not conserved and are not resolved in experimental structures of the protein complexes—or even included in the experiments. Thus, these regions and their potential structural ensembles and conformational preferences cannot be extracted from the PDB and thus they currently constitute a data-gap for training purposes.

As the sequence properties of SLiMs are known only for a small fraction of the predicted SLiMome, there is a strong need for procedures that may enable the identification and annotation of SLiMs without extensive experimental efforts. Combined with the variability in the number of residues separating the key conserved sites within a SLiM, the possibility of being able to predict distance distributions of SLiM-based interactions, in which the possible contact points and special requirements could be mapped, would potentially be an important asset that may help facilitate functional annotation and even pinpoint relevant

binding partners to address. While machine learning seems like a promising approach, the elasticity of the SLiM sequence and the low conservation of the SLiM context would make a purely sequence-based approach difficult. Further information might be obtained from an MSA of the IDP and the binding partner,[14,170,171] although the signal for contacts might be relatively weak and difficult to extract. Very recently, Alpha Fold 2 and RoseTTAFold have been used to predict structures of protein-peptide complexes.[172,173] Another problem that emerges is how to learn from sets of SLiMs that have been characterized in depth, and apply this knowledge to other sequences that have been probed much less.

One recently described approach to learn the rules for protein-peptide interactions is a bespoke machine-learning approach, termed hierarchical statistical mechanical modelling, which can be trained on families with abundant experimental data (structures and sequences).[174] The approach learns a pseudo-energy function for interactions relevant for binding, which can be transferred also to proteins for which less information is available. In this way, the approach provides an elegant example of how machine learning methods can be used to learn general rules of biophysics that enable transfer and predictions on a wider class of problems and systems.

Looking ahead, although many structures have been determined of complexes between folded domains and peptides representing SLiMs from disordered proteins, these structures have in most cases been solved in the absence of the flanking regions. As these regions can be highly relevant for binding specificity and affinity,[175] it is important to develop approaches that take these sequences into account. At the moment, however, the functional and structural properties of flanking regions are poorly understood and rarely studied, making it difficult to develop prediction methods. Initially, it might be fruitful to compare the surface properties of the protein that binds the IDP (e.g. charge patterning and hydrophobicity) to the overall physicochemical properties of the flanking regions. One approach towards such endeavours uses a sequence-based model of charge patterning to relate sequence to function.[69] Eventually, and aided by the generation of data set that include longer peptides or full-length proteins, it may be possible to develop prediction methods that combine local and long-range interactions, perhaps using similar methods as when predicting effects of enhancers in gene regulation.[176,177]

Complexes beyond SLiMs

As most IDPs have large exposed surface areas with high conformational flexibility, they have high potential for binding other proteins.[178,179] In these complexes, IDPs have shown remarkable structural and functional diversity, ranging from complex formation through folding-upon-binding with interfaces of similar composition and properties as to those formed between folded complexes,[180–184] over complexes where the disordered partner remains dynamic to different extents, all of relevance to function.[185–187] At the extreme end of the scale, highly dynamic complexes, which entirely lack the formation of stable secondary or tertiary structures, can form, for example between two highly and oppositely charged IDPs.[188,189] With this diversity, it is relevant to ask how machine learning methods would aid in predicting the structures of these complexes and whether they are able to predict the degree of remaining disorder. If so, would it then be possible to decompose the role of disorder for function and what would be the problems associated with these tasks?

One of the first discoveries from studying disordered protein complexes were that they can fold upon binding, either to an already folded partner through one of two highly discussed mechanisms[182,190,191] or through the occasional mutual folding of two disordered proteins.[192,193] Whereas folding-upon-binding of disordered regions at first may seem highly analogous to the process of protein folding, and hence in principle should be amiable to machine learning approaches to predict the structures of the complexes, there are however a number of obvious caveats to its direct use. For example, even though the binding region may be known, it is not easy to predict from sequence alone, which part of the disordered protein will fold and even less so to predict to which extent the folding occurs.

Further, a continuum of disorder can exist both in the IDP alone and in a complex, and highly disordered complexes, by some termed fuzzy,[194,195] may result in weak and near-stochastic interactions. One such example is the activation domains of transcription factors,[196] whose properties were originally characterized as 'acid blobs and negative noodles'.[197] Recently, a number of multiplexed assays have been used to expand this view and study the functional requirements of the sequence properties of transcriptional activation domains.[198–203] These results confirm the original observations of a requirement for hydrophobic and negatively charged residues and provide additional information about the role of patterning. Further, the data can be used to train various sequence-based machine learning models for activity.[199,200,202,204] The results from these and other studies suggest that most functional variation can be explained solely by amino acid composition, but that there is additional signal from higher-order properties of the amino acid sequence,[49,200] thus highlighting the importance of generating sequence libraries with such properties in mind.[198]

Importantly, the models described above work without explicitly considering the molecular

interactions that drive transcription. More generally, it remains to be seen how much detailed information about the interactions is needed to predict the biological consequences, and vice versa how much mechanistic insight at the molecular level can be gained from such prediction methods. Related to this is the question of how much structural information is contained in the evolutionary record of these complexes. Thus, while one might intuitively think that predicting structures of these complexes would be similar to predicting structures and interactions between folded proteins, this might not turn out to be the case because the relationship between structure and function may be substantially different, and because disorder may persist in the complexes and play functional roles. Indeed, how remaining disorder in complexes can be predicted from sequence is perhaps an even bigger challenge.

Biomolecular condensates

Many IDPs have the capacity to form multivalent interactions that are key for the ability to form so-called biomolecular condensates, either alone, with another IDP or in complex with folded domains or RNA. We refer the reader to recent reviews on the topic,[205–208] and focus here mostly on the role of IDPs in forming such condensates and on a set of problems where machine learning methods might help.

Biomolecular condensates often form via the process of liquid-liquid phase separation (LLPS), and a central requirement for a molecule to form these structures is the ability to form multivalent interactions. In the context of IDPs, this can for example be a protein carrying multiple SLiMs that can bind to folded multidomain proteins[209,210] or a set of amino acid residues within an IDP that can form sufficiently strong interactions between them.[56,211] Key areas for biophysical research include identifying the sequences and interactions that drive phase separation, identifying determinants of specificity in condensate formation, and elucidating the structural and dynamical features in biomolecular condensates. Before examining these questions, we stress that not all IDPs readily undergo LLPS, and that not all condensates involve IDPs.

Given the importance of IDP-IDP and SliM-target interactions, the methods discussed above for characterizing IDPs and SLiMs are also important for studying condensates. One key insight is that—due to the similarity between intramolecular interactions within IDPs and intermolecular interactions between IDPs—there is a correspondence between the propensity of an IDP to sample more compact structures and for it to undergo phase separation.[56,207,208,212–214] Thus, methods to predict compaction of IDPs or to parameterize simulation methods for isolated IDPs will also aid in studying phase separation of IDPs. Similarly, methods to predict SLiMs and their binding partners—and possibly the affinity of pairwise interactions—from sequence or from MSAs will aid in mapping the interactions that drive phase separation in these systems, and help to derive rules and features for their formation.

A number of databases have recently been created to collect information about proteins that undergo phase separation.[215–219] Such databases are now being used to develop prediction methods for phase separation,[220–225] also with the aim of providing insight into the sequences and properties that are important for phase separation. In the same way as prediction methods for protein disorder have played a central role in understanding the role of disorder at the proteome level, such methods have the potential to do the same for biomolecular condensates.[220,221]

Moving ahead, it will be important to extend such databases and prediction methods with additional quantitative information on the propensity to phase separate, and to annotate more broadly what components or features are involved in the formation of condensates. Recently, deep learning methods have been used to analyse large-scale microscopy data for finding candidates for proteins that form condensates.[226] In the same way as many proteins and peptides have been shown to form amyloid structures under some conditions, many proteins will likely undergo LLPS. Thus, in the same way as methods for predicting aggregation propensities have been trained on quantitative measurements of aggregation,[227–229] improvements in our ability to predict the propensity to undergo LLPS will likely involve fitting to or benchmarking against quantitative measurements of phase separation. Such analyses are already being performed with various coarse-grained simulation methods discussed above,[56,130,207,208,230,231] but it may be difficult to scale these methods to proteomewide applications or to scan large numbers of components in heterotypic condensates. The relationship between intra- and inter-molecular interactions and the driving force for phase separation suggests that it might be possible to train sequence-based prediction models on single-chain properties and use these to predict the ability to undergo LLPS. Such methods have already provided a number of general rules about valency and patterning that appear promising for our ability to predict the propensity of proteins to undergo LLPS from their amino acid sequence.[56,231–234] Including the context, such as concentration, crowding and additional partners in heterotypic condensate formation in these models would be an important extension. The conformational landscape of IDPs is also dependent on a richness in protein post-translational modifications such as phosphorylations, methylations, sulfation, and lipidation, and e.g. phosphorylation and arginine methylation has

been shown to affect the formation of condensates.[235–239] Thus, predicting post-translational modifications and their effects on condensates would help provide additional insight into how condensates are regulated.

Intrinsic Disorder and Human Diseases

Given their wide range of biological functions, it is not surprising that IDPs are involved in a number of human diseases[240] including neurodegeneration[241] and in particular in cancer.[242–244] How may machine learning methods help understand the role of IDPs in disease?

While it appears a simple question to ask whether IDPs are enriched in a particular disease, answering this question requires accurate and unbiased predictions of protein disorder.[243] Thus, we need continuous development of databases and quantitative measures of protein disorder and assessment of prediction accuracy, as well as development of new prediction methods.[1,3,245,246]

It is important to gain a better understanding of the molecular mechanisms underlying diseases involving IDPs. The expression of IDPs is tightly regulated, and dysregulation may lead to disease.[247] For folded proteins, it is well established that genetic missense variants may cause disease via a wide range of mechanisms including affecting both protein stability and interactions.[248–250] A substantial number of disease-causing variants are, however, located in regions of predicted disorder and are predicted to affect for example SLiMs.[251] Thus, it is becoming clear that missense variants in IDPs can also lead to disease via perturbed interactions that either cause loss or gain of function,[252–254] including promoting the formation of fibrils and toxic oligomeric species.

Loss of protein stability arising from missense variants and resulting protein degradation is established to be a key mechanism underlying loss of function for many folded proteins,[250,255] and indeed measurements or predictions of protein stability and abundance are useful for predicting loss of function.[256,257] While intrinsic thermodynamic stability of a folded state is not a meaningful quantity for IDPs, missense variants may still affect their cellular abundance. This may for example happen by mutations leading to impaired interactions and degradation, as exemplified by a missense variant in the IDR of the growth hormone receptor; a mutation leading to severe lung cancer.[258] Similarly, missense variants may lead to new interactions by SLiM appearance,[252,259] lack of degradation by interference with degrons, disorder-to-order formation,[251] or changes in long-range interactions.[260] In the latter example, machine learning techniques helped uncover differences in conformational dynamics from molecular dynamics trajectories of β-amyloid and the E22G disease variant, implicating their fibrillation into different morphologies. Thus, we need a better understanding both of the how IDPs are targeted for degradation and the sequence signals that determine cellular abundance,[261] and of how contact remodeling along the chain impacts the ensemble. Disordered regions may act as degradation signals (degrons),[262] and new large-scale experiments are enabling a better understanding of the sequence and structural properties of degrons.[263,264] We expect that such experiments will ultimately enable better predictions of the degradation and abundance of IDPs, and the effects of mutations on these properties.

One particularly important role of IDPs in disease may be in those that are involved in the formation of biomolecular condensates. A number of diseases have been associated with misregulation or formation of such condensates (as recently reviewed by others[265–270]), and thus a better understanding of the sequence properties that drive the formation of condensates will be important for predicting their role in disease[271] as well as for targeting them pharmaceutically.[272]

More generally, in order to better predict how variants in IDPs may cause disease, we need a clearer overview of the relationship between sequence, structural and dynamical properties, binding preferences and function. For folded proteins, analyses of conservation via MSAs are very powerful to predict whether a variant may cause disease,[273,274] but as discussed above, constructing and analysing MSAs provide unique challenges for IDPs. Thus, we need new methods to leverage the increasingly growing sequence databases to predict the effects of sequence variation in IDPs,[275–277] ultimately enabling targeting and drug development for combating diseases related to misregulation and dysfunction of disordered proteins. From an experimental point of view, multiplexed assays of variant effects (also sometimes called deep mutational scans) can provide key insights into both fundamental aspects of protein science[278] and genotype-phenotype relationships and disease.[279] Such experiments are now also beginning to provide a more comprehensive view of the effects of amino acid substitutions in IDPs such as the experiments on activation domains of transcription factors discussed above,[198–203] as well as experiments on a number of aggregation prone disordered proteins.[280–285]

Outlook

IDPs are an enormously broad class of molecules and together with IDRs they are involved in a wide range of biological functions. A key defining feature of IDPs and IDRs is something they do not have, namely a persistent three-dimensional structure. Thus, in many ways they are defined by being different from the globular and membrane

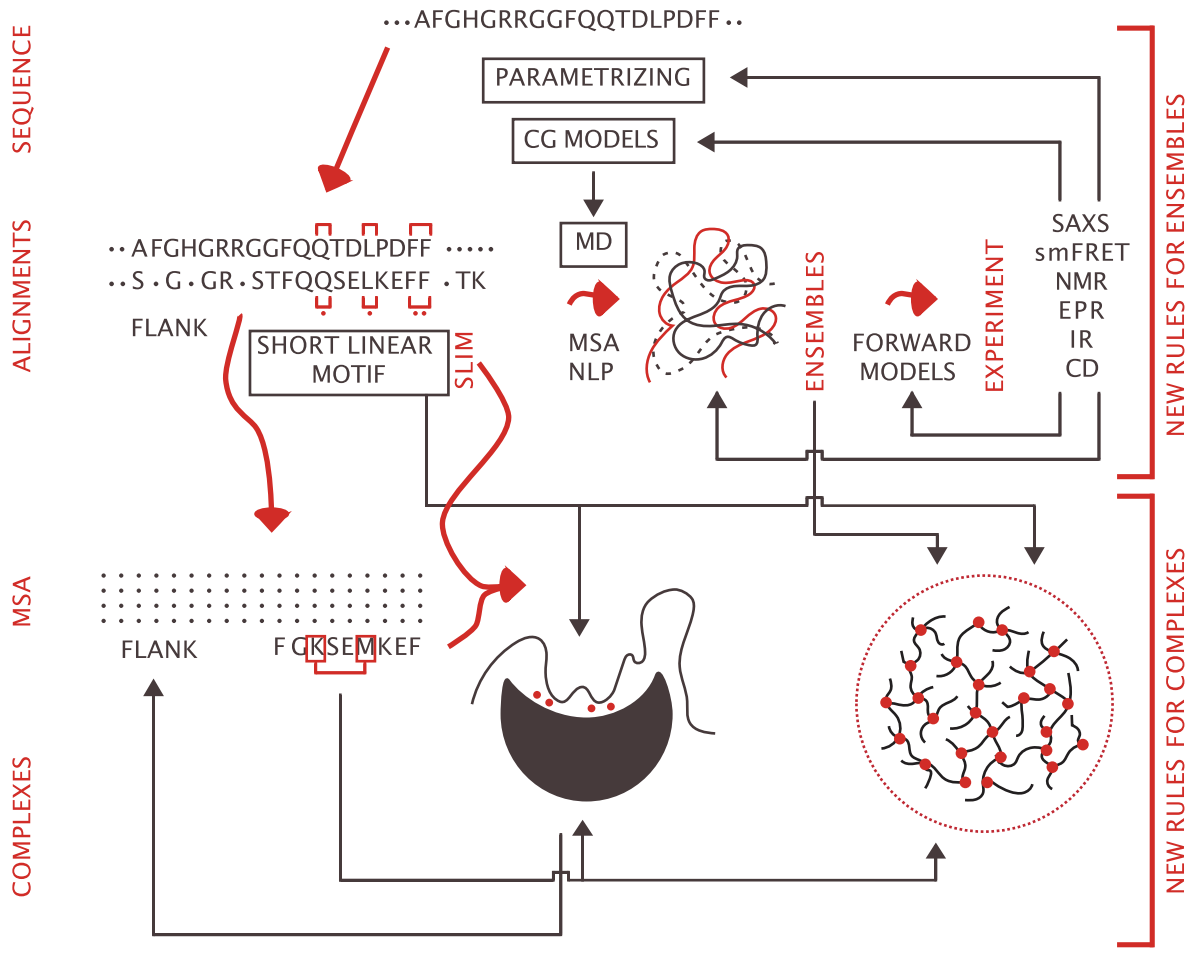

Figure 2. Outlining connections between sequence, structure, dynamics and function of IDPs where implementing machine learning approaches could have a potential (indicated with black connectors). From sequences and sequence alignments, machine learning approaches may help extract conformational properties from poorly defined sequence alignments of IDPs (NLP: natural language processing). Machine learning may be used to improve methods for combining biophysical experiments (here illustrated by SAXS, NMR, smFRET, EPR, IR and CD) and computation for example by deriving better forward models, and helping parameterizing force fields for better coarse grained (CG) models of IDPs. Machine learning may also enable extraction of new SLiMs and annotation of their biological functions, and provide insight into and ability to predict how context and flanking region (flanks) contribute to IDP function. Machine learning may also help predict and understand properties important for the formation of biomolecular condensates, and how context plays roles in their formation and dissolution. Finally, but not illustrated here, the combination of these approaches may help assign pathogenicity to genetic variants of IDPs. United, machine learning in combination with bioinformatics, simulation, theory and experiments can provide new rules for understanding IDP ensembles and IDP function. Jointly, such rules are necessary to enable the important decomposition of how mutations in IDPs may lead to disease states.

proteins that in more than a century have been the central focus of much protein science. Indeed, when the first CASP experiment was performed in 1994[286] only few proteins were recognized as being intrinsically disordered and rarely was the conformational disorder linked to biological function.

In some ways, IDPs and IDRs are simpler than folded proteins because their linear (primary) structure already provides much insight into their chemistry and ability to interact with other molecules. Thus, a number of computational methods have been developed to predict disorder from sequence and to identify local segments of the sequence that can bind to other molecules.

This apparent simplicity, however, can be deceiving. For folded proteins, the necessity to fold into a specific three-dimensional structure puts substantial restraints on the sequence and thus on evolution. Thus, MSAs of folded proteins often provide clues about specific residues and regions that are key to structure and function. Together with a large number of high resolution structures, this has lead to our ability to predict with increasing accuracy the structure of folded

proteins. In contrast, while the sequences of many IDPs are conserved for function, this relationship is complex and different from that governing folded proteins, largely because their function is also coupled to their dynamics. It is interesting to speculate what protein science would have looked like if we had first discovered IDPs, and then later found sequences that fold into specific three-dimensional structures.

Over the last 25 years, we have begun to understand the rules that govern the structural properties of IDPs, their interactions and their biological functions. Like for folded proteins, much insight has come from studying one system at a time, and computational methods are used to consolidate this into rules and predictions. In this perspectives paper we have outlined a number of current problems in studies of IDPs, including our ability to characterize their structural preferences and interactions, and our limitations in describing them. We have highlighted areas where machine learning and other computational methods have already had important impact, and new areas for further exploration (Figure 2). Common to all is the tight interplay between experiment and computation. Particularly important is perhaps the realization that the two need to be developed together, with experiments being designed to inform computational methods, and computational algorithms developed, trained, and benchmarked using experiments.

In addition to developing increasingly more sophisticated experimental and computational methods, the field also needs to collect this information more systematically. Thus, in order for the kinds of machine learning and other computational methods discussed here to continue to have an impact, we need more quantitative, systematically collected and annotated data, organized in easily accessible open databases. We have discussed a number of types of data that can be useful including biophysical data on the structural preferences of IDPs alone and in complexes, the affinities between SLiMs and a larger number of targets, contributions from flanking regions and context, as well as quantitative measurements of the formation of condensates. We look forward to see where these approaches will take the field.

CRediT authorship contribution statement

Kresten Lindorff-Larsen: Conceptualization, Writing – original draft. **Birthe B. Kragelund:** Conceptualization, Writing – original draft.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge many discussions with our colleagues at the Structural Biology and NMR Laboratory and Linderstrøm-Lang Centre for Protein Science, and thank Tanja Mittag for comments on the manuscript. Asta B. Andersen is thanked for graphics support. Our research is supported by the Novo Nordisk Challenge Programmes REPIN (NNF18OC0033926; BBK) and PRISM (NNF18OC0033950; KLL), and the Lundbeck Foundation BRAINSTRUC initiative in structural biology (R155-2015-2666; KLL & BBK).

Received 2 June 2021;
Accepted 4 August 2021;
Available online 12 August 2021

Keywords:
machine learning;
intrinsically disordered protein;
molecular complex;
condensate;
SLiM

References

1. Hatos, A., Hajdu-Soltész, B., Monzon, A.M., Palopoli, N., Álvarez, L., Aykac-Fas, B., Bassot, C., Benítez, G.I., et al., (2020). DisProt: intrinsic protein disorder annotation in 2020. *Nucl. Acids Res.*, **48** (D1), D269–D276.
2. Piovesan, D., Necci, M., Escobedo, N., Monzon, A.M., Hatos, A., Mičetić, I., Quaglia, F., Paladin, L., et al., (2021). MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res.*, **49** (D1), D361–D367.
3. Necci, M., Piovesan, D., Tosatto, S.C., (2021). Critical assessment of protein intrinsic disorder prediction. *Nature Methods*,, 1–10.
4. Mittag, T., Forman-Kay, J.D., (2007). Atomic-level characterization of disordered protein ensembles. *Curr. Opin. Struct. Biol.*, **17** (1), 3–14.
5. Jensen, M.R., Ruigrok, R.W., Blackledge, M., (2013). Describing intrinsically disordered proteins at atomic resolution by NMR. *Curr. Opin. Struct. Biol.*, **23** (3), 426–435.
6. Salvi, N., Abyzov, A., Blackledge, M., (2016). Multi-timescale dynamics in intrinsically disordered proteins from NMR relaxation and molecular simulation. *J. Phys. Chem. Letters*, **7** (13), 2483–2489.
7. Lazar, T., Martínez-Pérez, E., Quaglia, F., Hatos, A., Chemes, L.B., Iserte, J.A., Méndez, N.A., Garrone, N.A., et al., (2021). PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res.*, **49** (D1), D404–D411.
8. wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic acids research. 2019; 47(D1):D520–D528.

9. Anfinsen, C.B., (1973). Principles that govern the folding of protein chains. *Science*, **181** (4096), 223–230.
10. Eisenberg, D.S., (2018). How Hard It Is Seeing What Is in Front of Your Eyes. *Cell*, **174** (1), 8–11.
11. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., Moult, J., (2019). Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Struct. Funct. Bioinformat.*, **87** (12), 1011–1020.
12. AlQuraishi, M., (2021). Machine learning in protein structure prediction. *Curr. Opin. Chem. Biol.*, **65**, 1–8.
13. Lapedes, A., Giraud, B, Jarzynski, C. (2002). Using sequence alignments to predict protein structure and stability with high accuracy. LANL preprint LA-UR-02-4481. http://library.lanl.gov/cgi-bin/getfile?01038177.pdf.
14. Weigt, M., White, R.A., Szurmant, H., Hoch, J.A., Hwa, T., (2009). Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl. Acad. Sci.*, **106** (1), 67–72.
15. Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., Sander, C., (2011). Protein 3D structure computed from evolutionary sequence variation. *PloS One*, **6** (12), e28766.
16. Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J.N., et al., (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci.*, **108** (49), E1293–E1301.
17. Balakrishnan, S., Kamisetty, H., Carbonell, J.G., Lee, S.I., Langmead, C.J., (2011). Learning generative models for protein fold families. *Proteins: Struct. Funct. Bioinformat.*, **79** (4), 1061–1078.
18. Xu, J., (2019). Distance-based protein folding powered by deep learning. *Proc. Natl. Acad. Sci.*, **116** (34), 16856–16865.
19. Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., et al., (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, **577** (7792), 706–710.
20. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Tunyasuvunakool, K., Ronneberger, O., Bates, R., et al., (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*,, 1–11.
21. Zheng, W., Li, Y., Zhang, C., Pearce, R., Mortuza, S., Zhang, Y., (2019). Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins: Struct. Funct. Bioinformat.*, **87** (12), 1149–1164.
22. Kandathil, S.M., Greener, J.G., Jones, D.T., (2019). Recent developments in deep learning applied to protein structure prediction. *Proteins: Struct. Funct. Bioinformat.*, **87** (12), 1179–1189.
23. AlQuraishi, M., (2019). End-to-end differentiable learning of protein structure. *Cell Syst.*, **8** (4), 292–301.
24. Torrisi, M., Pollastri, G., Le, Q., (2020). Deep learning methods in protein structure prediction. Computational and Structural. *Biotechnol. J.*,.
25. Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., Baker, D., (2020). Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci.*, **117** (3), 1496–1503.
26. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., et al., (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*,, eabj8754.
27. Masarati, G., Landau, M., Ben-Tal, N., Lupas, A., Kosloff, M., Kosinski, J., (2021). Integrative structural biology in the era of accurate structure prediction. *J. Mol. Biol.*,, 167127.
28. Min, S., Lee, B., Yoon, S., (2017). Deep learning in bioinformatics. *Briefings Bioinformat.*, **18** (5), 851–869.
29. Jurtz, V.I., Johansen, A.R., Nielsen, M., Almagro Armenteros, J.J., Nielsen, H., Sønderby, C.K., Winther, O., Sønderby, S.K., (2017). An introduction to deep learning on biological sequence data: examples and solutions. *Bioinformatics*, **33** (22), 3685–3690.
30. Xu, Y., Verma, D., Sheridan, R.P., Liaw, A., Ma, J., Marshall, N.M., McIntosh, J., Sherer, E.C., et al., (2020). Deep dive into machine learning models for protein engineering. *J. Chem. Inform. Model.*, **60** (6), 2773–2790.
31. Gao, W., Mahajan, S.P., Sulam, J., Gray, J.J., (2020). Deep learning in protein structural modeling and design. *Patterns*,, 100142.
32. Bepler, T., Berger, B., (2021). Learning the protein language: Evolution, structure, and function. *Cell Syst.*, **12** (6), 654–669.
33. Taylor, W.R., Hatrick, K., (1994). Compensating changes in protein multiple sequence alignments. *Protein Eng. Des. Select.*, **7** (3), 341–348.
34. Neher, E., (1994). How frequent are correlated changes in families of protein sequences? *Proc. Natl. Acad. Sci.*, **91** (1), 98–102.
35. Göbel, U., Sander, C., Schneider, R., Valencia, A., (1994). Correlated mutations and residue contacts in proteins. *Proteins: Struct. Funct. Bioinformat.*, **18** (4), 309–317.
36. Kassem, M.M., Christoffersen, L.B., Cavalli, A., Lindorff-Larsen, K., (2018). Enhancing coevolution-based contact prediction by imposing structural self-consistency of the contacts. *Sci. Rep.*, **8** (1), 1–10.
37. Laine, E., Eismann, S., Elofsson, A., Grudinin, S., (2021). Protein sequence-to-structure learning: Is this the end(-to-end revolution)? *arXiv*,, 2105.07407.
38. Serrano, L., (1995). Comparison between the φ distribution of the amino acids in the protein database and NMR data indicates that amino acids have various φ propensities in the random coil conformation. *J. Mol. Biol.*, **254** (2), 322–333.
39. Smith, L.J., Bolin, K.A., Schwalbe, H., MacArthur, M.W., Thornton, J.M., Dobson, C.M., (1996). Analysis of main chain torsion angles in proteins: prediction of NMR coupling constants for native and random coil conformations. *J. Mol. Biol.*, **255** (3), 494–506.
40. Bernadó, P., Blanchard, L., Timmins, P., Marion, D., Ruigrok, R.W., Blackledge, M., (2005). A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc. Natl. Acad. Sci.*, **102** (47), 17002–17007.
41. Simons, K.T., Kooperberg, C., Huang, E., Baker, D., (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, **268** (1), 209–225.
42. Bystroff, C., Thorsson, V., Baker, D., (2000). HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J. Mol. Biol.*, **301** (1), 173–190.
43. Boomsma, W., Mardia, K.V., Taylor, C.C., Ferkinghoff-Borg, J., Krogh, A., Hamelryck, T., (2008). A generative, probabilistic model of local protein structure. *Proc. Natl. Acad. Sci.*, **105** (26), 8932–8937.

44. Estaña, A., Barozet, A., Mouhand, A., Vaisset, M., Zanon, C., Fauret, P., Sibille, N., Bernadó, P., et al., (2020). Predicting secondary structure propensities in IDPs using simple statistics from three-residue fragments. *J. Mol. Biol.*, **432** (19), 5447–5459.
45. Pietrek, L.M., Stelzl, L.S., Hummer, G., (2019). Hierarchical ensembles of intrinsically disordered proteins at atomic resolution in molecular dynamics simulations. *J. Chem Theory Comput.*, **16** (1), 725–737.
46. Toth-Petroczy, A., Palmedo, P., Ingraham, J., Hopf, T.A., Berger, B., Sander, C., Marks, D.S., (2016). Structured states of disordered proteins from genomic sequences. *Cell*, **167** (1), 158–170.
47. Tian, P., Boomsma, W., Wang, Y., Otzen, D.E., Jensen, M.H., Lindorff-Larsen, K., (2015). Structure of a functional amyloid protein subunit computed using sequence variation. *J. Am. Chem. Soc.*, **137** (1), 22–25.
48. Cordeiro, T.N., Sibille, N., Germain, P., Barthe, P., Boulahtouf, A., Allemand, F., Bailly, R., Vivat, V., et al., (2019). Interplay of protein disorder in retinoic acid receptor heterodimer and its corepressor regulates gene expression. *Structure*, **27** (8), 1270–1285.
49. Gonzalez-Foutel, N.S., Borcherds, W.M., Glavina, J., Barrera-Vilarmau, S., Sagar, A., Estaña, A., Barozet, A., Fernandez-Ballester, G., et al., (2021). Conformational buffering underlies functional selection in intrinsically disordered protein regions. *bioRxiv,*. https://doi.org/10.1101/2021.05.14.444182.
50. Marsh, J.A., Forman-Kay, J.D., (2010). Sequence determinants of compaction in intrinsically disordered proteins. *Biophys. J.*, **98** (10), 2383–2390.
51. Hofmann, H., Soranno, A., Borgia, A., Gast, K., Nettels, D., Schuler, B., (2012). Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci.*, **109** (40), 16155–16160.
52. Das, R.K., Pappu, R.V., (2013). Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci.*, **110** (33), 13392–13397.
53. Sawle, L., Ghosh, K., (2015). A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *J. Chem. Phys.*, **143** (8), 08B615_1.
54. Sørensen, C.S., Kjaergaard, M., (2019). Effective concentrations enforced by intrinsically disordered linkers are governed by polymer physics. *Proc. Natl. Acad. Sci.*, **116** (46), 23124–23131.
55. Zheng, W., Dignon, G., Brown, M., Kim, Y.C., Mittal, J., (2020). Hydropathy patterning complements charge patterning to describe conformational preferences of disordered proteins. *J. Phys. Chem. Letters*, **11** (9), 3408–3415.
56. Martin, E.W., Holehouse, A.S., Peran, I., Farag, M., Incicco, J.J., Bremer, A., Grace, C.R., Soranno, A., et al., (2020). Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science*, **367** (6478), 694–699.
57. Cohan, M.C., Ruff, K.M., Pappu, R.V., (2019). Information theoretic measures for quantifying sequence–ensemble relationships of intrinsically disordered proteins. *Protein Eng. Des. Select.*, **32** (4), 191–202.
58. Bawono, P., Dijkstra, M., Pirovano, W., Feenstra, A., Abeln, S., Heringa, J., (2017). Multiple sequence alignment. *Bioinformatics,*. Springer, pp. 167–189.
59. Lee, Y.D., Wang, J., Stubbe, J., Elledge, S.J., (2008). Dif1 is a DNA-damage-regulated facilitator of nuclear import for ribonucleotide reductase. *Mol. Cell*, **32** (1), 70–80.
60. Rozen, S., Füzesi-Levi, M.G., Ben-Nissan, G., Mizrachi, L., Gabashvili, A., Levin, Y., Ben-Dor, S., Eisenstein, M., et al., (2015). CSNAP is a stoichiometric subunit of the COP9 signalosome. *Cell Rep.*, **13** (3), 585–598.
61. Light, S., Sagit, R., Sachenkova, O., Ekman, D., Elofsson, A., (2013). Protein expansion is primarily due to indels in intrinsically disordered regions. *Mol. Biol. Evol.*, **30** (12), 2645–2653.
62. Alley, E.C., Khimulya, G., Biswas, S., AlQuraishi, M., Church, G.M., (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, **16** (12), 1315–1322.
63. Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., Song, Y.S., (2019). Evaluating protein transfer learning with tape. *Adv. Neural Inform. Process. Syst.*, **32**, 9689.
64. Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., Rost, B., (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformat.*, **20** (1), 1–17.
65. Ofer, D., Brandes, N., Linial, M., (2021). The language of proteins: NLP, machine learning & protein sequences. Computational and Structural. *Biotechnol. J.*,.
66. Weinstein, E.N., Marks, D.S., (2021). A structured observation distribution for generative biological sequence prediction and forecasting. *bioRxiv,*. p. 2020–07.
67. Rao, R., Liu, J., Verkuil, R., Meier, J., Canny, J.F., Abbeel, P., Sercu, T., Rives, A., (2021). Msa transformer. *bioRxiv,*.
68. Pritišanac, I., Vernon, R.M., Moses, A.M., Forman Kay, J. D., (2019). Entropy and information within intrinsically disordered protein regions. *Entropy*, **21** (7), 662.
69. Huihui, J., Ghosh, K., (2021). Intra-chain interaction topology can identify functionally similar Intrinsically Disordered Proteins. *Biophys. J.*,, 1860–1868.
70. Lu, A.X., Lu, A.X., Pritišanac, I., Zarin, T., Forman-Kay, J. D., Moses, A.M., (2021). Discovering molecular features of intrinsically disordered regions by using evolution for contrastive learning. *bioRxiv,*. https://doi.org/10.1101/2021.07.29.454330.
71. Bonomi, M., Heller, G.T., Camilloni, C., Vendruscolo, M., (2017). Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.*, **42**, 106–116.
72. Orioli, S., Larsen, A.H., Bottaro, S., Lindorff-Larsen, K., (2020). How to learn from inconsistencies: Integrating molecular simulations with experimental data. *Prog. Mol. Biol. Translat. Sci.*, **170**, 123–176.
73. Henriques, J., Arleth, L., Lindorff-Larsen, K., Skepö, M., (2018). On the calculation of SAXS profiles of folded and intrinsically disordered proteins from computer simulations. *J. Mol. Biol.*, **430** (16), 2521–2539.
74. Riback, J.A., Bowman, M.A., Zmyslowski, A.M., Knoverek, C.R., Jumper, J.M., Hinshaw, J.R., Kaye, E. B., Freed, K.F., et al., (2017). Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science*, **358** (6360), 238–241.
75. Fuertes, G., Banterle, N., Ruff, K.M., Chowdhury, A., Mercadante, D., Koehler, C., Kachala, M., Girona, G.E.,

et al., (2017). Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements. *Proc. Natl. Acad. Sci.*, **114** (31), E6342–E6351.
76. Zheng, W., Best, R.B., (2018). An extended Guinier analysis for intrinsically disordered proteins. *J. Mol. Biol.*, **430** (16), 2540–2553.
77. Ahmed, M.C., Crehuet, R., Computing, Lindorff-Larsen K., (2020). Analyzing, and Comparing the Radius of Gyration and Hydrodynamic Radius in Conformational Ensembles of Intrinsically Disordered Proteins. *Intrinsically Disordered Proteins*,. Springer, pp. 429–445.
78. Bernadó, P., Mylonas, E., Petoukhov, M.V., Blackledge, M., Svergun, D.I., (2007). Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.*, **129** (17), 5656–5664.
79. Hub, J.S., (2018). Interpreting solution X-ray scattering data using molecular simulations. *Curr. Opin. Struct. Biol.*, **49**, 18–26.
80. Cordeiro, T.N., Pc, Chen, De Biasio, A., Sibille, N., Blanco, F.J., Hub, J.S., Crehuet, R., Bernadó, P., (2017). Disentangling polydispersity in the PCNA-p15PAF complex, a disordered, transient and multivalent macromolecular assembly. *Nucl. Acids Res.*, **45** (3), 1501–1515.
81. Pesce, F., Lindorff-Larsen, K., (2021). Refining conformational ensembles of flexible proteins against small-angle X-ray scattering data. *bioRxiv*,. https://doi.org/10.1101/2021.05.29.446281.
82. Svergun, D., Barberato, C., Koch, M.H., (1995). CRYSOL–a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.*, **28** (6), 768–773.
83. Xu, X.P., Case, D.A., (2001). Automated prediction of 15N, 13Cα, 13Cβ and 13C' chemical shifts in proteins using a density functional database. *J. Biomol. NMR*, **21** (4), 321–333.
84. Shen, Y., Bax, A., (2007). Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J. Biomol. NMR*, **38** (4), 289–302.
85. Kohlhoff, K.J., Robustelli, P., Cavalli, A., Salvatella, X., Vendruscolo, M., (2009). Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *J. Am. Chem. Soc.*, **131** (39), 13894–13895.
86. Han, B., Liu, Y., Ginzinger, S.W., Wishart, D.S., (2011). SHIFTX2: significantly improved protein chemical shift prediction. *J. Biomol. NMR*, **50** (1), 43.
87. Meiler, J., (2003). PROSHIFT: protein chemical shift prediction using artificial neural networks. *J. Biomol. NMR*, **26** (1), 25–37.
88. Shen, Y., Bax, A., (2010). SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR*, **48** (1), 13–22.
89. Li, J., Bennett, K.C., Liu, Y., Martin, M.V., Head-Gordon, T., (2020). Accurate prediction of chemical shifts for aqueous protein structure on Real World data. *Chem. Sci.*, **11** (12), 3180–3191.
90. Yang, Z., Chakraborty, M., Predicting, White AD., (2020). Chemical Shifts with Graph Neural Networks. *bioRxiv*,.
91. Lindorff-Larsen, K., Best, R.B., Vendruscolo, M., (2005). Interpreting dynamically-averaged scalar couplings in proteins. *J. Biomol. NMR*, **32** (4), 273–280.
92. Li, D.W., Brüschweiler, R., (2012). PPM: a side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles. *J. Biomol. NMR*, **54** (3), 257–265.
93. Christensen, A.S., Linnet, T.E., Borg, M., Boomsma, W., Lindorff-Larsen, K., Hamelryck, T., Jensen, J.H., (2013). Protein structure validation and refinement using amide proton chemical shifts derived from quantum mechanics. *PLoS One*, **8** (12), e84123.
94. Crehuet, R., Buigues, P.J., Salvatella, X., Lindorff-Larsen, K., (2019). Bayesian-maximum-entropy reweighting of IDP ensembles based on NMR chemical shifts. *Entropy*, **21** (9), 898.
95. Hermann, M.R., Hub, J.S., (2019). SAXS-restrained ensemble simulations of intrinsically disordered proteins with commitment to the principle of maximum entropy. *J. Chem Theory Comput.*, **15** (9), 5103–5115.
96. Chan-Yao-Chong, M., Deville, C., Pinet, L., van Heijenoort, C., Durand, D., Ha-Duong, T., (2019). Structural characterization of N-WASP domain V using MD simulations with NMR and SAXS data. *Biophys. J.*, **116** (7), 1216–1227.
97. Rieping, W., Habeck, M., Nilges, M., (2005). Inferential structure determination. *Science*, **309** (5732), 303–306.
98. Brookes, D.H., Head-Gordon, T., (2016). Experimental inferential structure determination of ensembles for intrinsically disordered proteins. *J. Am. Chem. Soc.*, **138** (13), 4530–4538.
99. Franke, D., Jeffries, C.M., Svergun, D.I., (2018). Machine learning methods for X-ray scattering data analysis from biomacromolecular solutions. *Biophys. J.*, **114** (11), 2485–2492.
100. Worswick, S.G., Spencer, J.A., Jeschke, G., Kuprov, I., (2018). Deep neural network processing of DEER data. *Sci. Adv.*, **4** (8), eaat5218.
101. Ye, S., Zhong, K., Zhang, J., Hu, W., Hirst, J.D., Zhang, G., Mukamel, S., Jiang, J., (2020). A Machine Learning Protocol for Predicting Protein Infrared Spectra. *J. Am. Chem. Soc.*, **142** (45), 19071–19077.
102. Chemes, L.B., Alonso, L.G., Noval, M.G., de Prat-Gay, G., (2012). Circular dichroism techniques for the analysis of intrinsically disordered proteins and domains. *In: Intrinsically disordered protein analysis*,. Springer, pp. 387–404.
103. Nagy, G., Igaev, M., Jones, N.C., Hoffmann, S.V., Grubmüller, H., (2019). SESCA: predicting circular dichroism spectra from protein molecular structures. *J. Chem Theory Comput.*, **15** (9), 5087–5102.
104. Olamoyesan, A., Ang, D., Rodger, A., (2021). Circular dichroism for secondary structure determination of proteins with unfolded domains using a self-organising map algorithm SOMSpec. *RSC Adv.*, **11** (39), 23985–23991.
105. Lindorff-Larsen, K., Ferkinghoff-Borg, J., (2009). Similarity measures for protein ensembles. *PloS One*, **4** (1), e4203.
106. Camilloni, C., Robustelli, P., Simone, A.D., Cavalli, A., Vendruscolo, M., (2012). Characterization of the conformational equilibrium between the two major substates of RNase A using NMR chemical shifts. *J. Am. Chem. Soc.*, **134** (9), 3968–3971.
107. Tiberti, M., Papaleo, E., Bengtsen, T., Boomsma, W., Lindorff-Larsen, K., (2015). ENCORE: software for quantitative ensemble comparison. *PLoS Comput. Biol.*, **11** (10), e1004415.

108. Larsen, A.H., Wang, Y., Bottaro, S., Grudinin, S., Arleth, L., Lindorff-Larsen, K., (2020). Combining molecular dynamics simulations with small-angle X-ray and neutron scattering data to study multi-domain proteins in solution. *PLoS Comput. Biol.*, **16** (4), e1007870.
109. Ahmed, M.C., Skaanning, L.K., Jussupow, A., Newcombe, E.A., Kragelund, B.B., Camilloni, C., Langkilde, A.E., Lindorff-Larsen, K., (2021). Refinement of α-synuclein ensembles against SAXS data: Comparison of force fields and methods. *Front. Mol. Biosci.*, **8**
110. Best, R.B., (2017). Computational and theoretical advances in studies of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.*, **42**, 147–154.
111. Huang, J., MacKerell Jr, A.D., (2018). Force field development and simulations of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.*, **48**, 40–48.
112. Robustelli, P., Piana, S., Shaw, D.E., (2018). Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci.*, **115** (21), E4758–E4766.
113. Zerze, G.H., Zheng, W., Best, R.B., Mittal, J., (2019). Evolution of all-atom protein force fields to improve local and global properties. *J. Phys. Chem. Letters*, **10** (9), 2227–2234.
114. Mu, J., Liu, H., Zhang, J., Luo, R., Chen, H.F., (2021). Recent Force Field Strategies for Intrinsically Disordered Proteins. *J. Chem. Inform. Model.*, **61** (3), 1037–1047.
115. Shea, J.E., Best, R.B., Mittal, J., (2021). Physics-based computational and theoretical approaches to intrinsically disordered proteins. *Curr. Opin. Struct. Biol.*, **67**, 219–225.
116. Perspective, Noid WG., (2013). Coarse-grained models for biomolecular systems. *J. Chem. Phys.*, **139** (9), 09B201_1.
117. Norgaard, A.B., Ferkinghoff-Borg, J., Lindorff-Larsen, K., (2008). Experimental parameterization of an energy function for the simulation of unfolded proteins. *Biophys. J.*, **94** (1), 182–192.
118. Njo, S.L., van Gunsteren, W.F., Müller-Plathe, F., (1995). Determination of force field parameters for molecular simulation by molecular simulation: An application of the weak-coupling method. *J. Chem. Phys.*, **102** (15), 6199–6207.
119. Norrby, P.O., Liljefors, T., (1998). Automated molecular mechanics parameterization with simultaneous utilization of experimental and quantum mechanical data. *J. Comput. Chem.*, **19** (10), 1146–1166.
120. Groth, M., Malicka, J., Rodziewicz-Motowidło, S., Czaplewski, C., Klaudel, L., Wiczk, W., Liwo, A., (2001). Determination of conformational equilibrium of peptides in solution by NMR spectroscopy and theoretical conformational analysis: Application to the calibration of mean-field solvation models. *Peptide Sci. Original Res. Biomol.*, **60** (2), 79–95.
121. Bathe, M., Rutledge, G.C., (2003). Inverse Monte Carlo procedure for conformation determination of macromolecules. *J. Comput. Chem.*, **24** (7), 876–890.
122. Li, D.W., Brüschweiler, R., (2010). NMR-based protein potentials. *Angew. Chem. Int. Ed.*, **49** (38), 6778–6780.
123. Piana, S., Lindorff-Larsen, K., Shaw, D.E., (2011). How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.*, **100** (9), L47–L49.
124. Di Pierro, M., Elber, R., (2013). Automated optimization of potential parameters. *J. Chem Theory Comput.*, **9** (8), 3311–3320.
125. Wang, L.P., Martinez, T.J., Pande, V.S., (2014). Building force fields: An automatic, systematic, and reproducible approach. *J. Phys. Chem. Letters*, **5** (11), 1885–1891.
126. Cesari, A., Bottaro, S., Lindorff-Larsen, K., Banáš, P., Šponer, J., Bussi, G., (2019). Fitting corrections to an RNA force field using experimental data. *J. Chem Theory Comput.*, **15** (6), 3425–3431.
127. Chen, J., Chen, J., Pinamonti, G., Clementi, C., (2018). Learning effective molecular models from experimental observables. *J. Chem Theory Comput.*, **14** (7), 3849–3858.
128. Latham, A.P., Zhang, B., (2019). Maximum entropy optimized force field for intrinsically disordered proteins. *J. Chem Theory Comput.*, **16** (1), 773–781.
129. Dannenhoffer-Lafage, T., Best, R.B., (2021). A Data-Driven Hydrophobicity Scale for Predicting Liquid-Liquid Phase Separation of Proteins. *J. Phys. Chem. B*,.
130. Tesei, G., Schulze, T.K., Crehuet, R., Lindorff-Larsen, K., (2021). Accurate model of liquid-liquid phase behaviour of intrinsically-disordered proteins from data-driven optimization of single-chain properties. *bioRxiv*,. https://doi.org/10.1101/2021.06.23.449550.
131. Demerdash, O., Shrestha, U.R., Petridis, L., Smith, J.C., Mitchell, J.C., Ramanathan, A., (2019). Using small-angle scattering data and parametric machine learning to optimize force field parameters for intrinsically disordered proteins. *Front. Mol. Biosci.*, **6**, 64.
132. Ruff, K.M., Harmon, T.S., Pappu, R.V., (2015). CAMELOT: A machine learning approach for coarse-grained simulations of aggregation of block-copolymeric protein sequences. *J. Chem. Phys.*, **143** (24), 12B607_1.
133. Husic, B.E., Charron, N.E., Lemm, D., Wang, J., Pérez, A., Majewski, M., Krämer, A., Chen, Y., et al., (2020). Coarse graining molecular dynamics with graph neural networks. *J. Chem. Phys.*, **153** (19), 194101.
134. Gkeka, P., Stoltz, G., Barati Farimani, A., Belkacemi, Z., Ceriotti, M., Chodera, J.D., Dinner, A.R., Ferguson, A.L., et al., (2020). Machine learning force fields and coarse-grained variables in molecular dynamics: application to materials and biological systems. *J. Chem Theory Comput.*, **16** (8), 4757–4775.
135. Giulini, M., Menichetti, R., Shell, M.S., Potestio, R., (2020). An Information-Theory-Based Approach for Optimal Model Reduction of Biomolecules. *J. Chem Theory Comput.*, **16** (11), 6795–6813.
136. Yang, H., Xiong, Z., Zonta, F., (2021). Construction of a neural network energy function for protein physics. *bioRxiv*,.
137. Chowdhury, R., Bouatta, N., Biswas, S., Rochereau, C., Church, G.M., Sorger, P.K., AlQuraishi, M.N., (2021). Single-sequence protein structure prediction using language models from deep learning. *bioRxiv*,. https://doi.org/10.1101/2021.08.02.454840.
138. Neduva, V., Linding, R., Su-Angrand, I., Stark, A., De Masi, F., Gibson, T.J., Lewis, J., Serrano, L., et al., (2005). Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol.*, **3** (12), e405.
139. Van Roey, K., Uyar, B., Weatheritt, R.J., Dinkel, H., Seiler, M., Budd, A., Gibson, T.J., Davey, N.E., (2014). Short linear motifs: ubiquitous and functionally diverse

protein interaction modules directing cell regulation. *Chem. Rev.*, **114** (13), 6733–6778.
140. Jespersen, N., Barbar, E., (2020). Emerging features of linear motif-binding Hub proteins. *Trends Biochem. Sci.*, **45** (5), 375–384.
141. Dinkel, H., Michael, S., Weatheritt, R.J., Davey, N.E., Van Roey, K., Altenberg, B., Toedt, G., Uyar, B., et al., (2012). ELM—the database of eukaryotic linear motifs. *Nucl. Acids Res.*, **40** (D1), D242–D251.
142. Kumar, M., Gouw, M., Michael, S., Sámano-Sánchez, H., Pancsa, R., Glavina, J., Diakogianni, A., Valverde, J.A., et al., (2020). ELM—the eukaryotic linear motif resource in 2020. *Nucl. Acids Res.*, **48** (D1), D296–D306.
143. Gouw, M., Alvarado-Valverde, J., Čalyševa, J., Diella, F., Kumar, M., Michael, S., Van Roey, K., Dinkel, H., et al., (2020). How to Annotate and Submit a Short Linear Motif to the Eukaryotic Linear Motif Resource. *In: Intrinsically Disordered Proteins*,. Springer, pp. 73–102.
144. O'Shea, C., Staby, L., Bendsen, S.K., Tidemand, F.G., Redsted, A., Willemoës, M., Kragelund, B.B., Skriver, K., (2017). Structures and short linear motif of disordered transcription factor regions provide clues to the interactome of the cellular hub protein radical-induced cell death1. *J. Biol. Chem.*, **292** (2), 512–527.
145. Zeke, A., Bastys, T., Alexa, A., Garai, Á., Mészáros, B., Kirsch, K., Dosztányi, Z., Kalinina, O.V., et al., (2015). Systematic discovery of linear binding motifs targeting an ancient protein interaction surface on MAP kinases. *Mol. Syst. Biol.*, **11** (11), 837.
146. Brauer, B.L., Moon, T.M., Sheftic, S.R., Nasa, I., Page, R., Peti, W., Kettenbach, A.N., (2019). Leveraging new definitions of the LxVP SLiM to discover novel calcineurin regulators and substrates. *ACS Chem. Biol.*, **14** (12), 2672–2682.
147. Ivarsson, Y., Arnold, R., McLaughlin, M., Nim, S., Joshi, R., Ray, D., Liu, B., Teyra, J., et al., (2014). Large-scale interaction profiling of PDZ domains through proteomic peptide-phage display using human and viral phage peptidomes. *Proc. Natl. Acad. Sci.*, **111** (7), 2542–2547.
148. Sundell, G.N., Arnold, R., Ali, M., Naksukpaiboon, P., Orts, J., Güntert, P., Chi, C.N., Ivarsson, Y., (2018). Proteome-wide analysis of phospho-regulated PDZ domain interactions. *Mol. Syst. Biol.*, **14** (8), e8129.
149. Benz, C., Ali, M., Krystkowiak, I., Simonetti, L., Sayadi, A., Mihalic, F., Kliche, J., Andersson, E., et al., (2021). Proteome-scale amino-acid resolution footprinting of protein-binding sites in the intrinsically disordered regions of the human proteome. *bioRxiv*,.
150. Madeira, F., Tinti, M., Murugesan, G., Berrett, E., Stafford, M., Toth, R., Cole, C., MacKintosh, C., et al., (2015). 14-3-3-Pred: improved methods to predict 14-3-3-binding phosphopeptides. *Bioinformatics*, **31** (14), 2276–2283.
151. Wheeler, L.C., Perkins, A., Wong, C.E., Harms, M.J., (2020). Learning peptide recognition rules for a low-specificity protein. *Protein Sci.*, **29** (11), 2259–2273.
152. Khan, W., Duffy, F., Pollastri, G., Shields, D.C., Mooney, C., (2013). Predicting binding within disordered protein regions to structurally characterised peptide-binding domains. *PLoS One*, **8** (9), e72838.
153. Kundu, K., Backofen, R., (2014). Cluster based prediction of PDZ-peptide interactions. *BMC Genom.*, **15** (1), 1–11.
154. Ronan, T., Garnett, R., Naegle, K.M., (2020). New analysis pipeline for high-throughput domain–peptide affinity experiments improves SH2 interaction data. *J. Biol. Chem.*, **295** (32), 11346–11363.
155. Wallweber, H.J., Tam, C., Franke, Y., Starovasnik, M.A., Lupardus, P.J., (2014). Structural basis of recognition of interferon-α receptor by tyrosine kinase 2. *Nature Struct. Mol. Biol.*, **21** (5), 443.
156. Plewczyński, D., Tkacz, A., Godzik, A., Rychlewski, L., (2005). A support vector machine approach to the identification of phosphorylation sites. *Cell Mol. Biol. Letters*, **10** (1), 73–89.
157. Tompa, P., Davey, N.E., Gibson, T.J., Babu, M.M., (2014). A million peptide motifs for the molecular biologist. *Mol. Cell*, **55** (2), 161–169.
158. Wigington, C.P., Roy, J., Damle, N.P., Yadav, V.K., Blikstad, C., Resch, E., Wong, C.J., Mackay, D.R., et al., (2020). Systematic discovery of Short Linear Motifs decodes calcineurin phosphatase signaling. *Mol. Cell*, **79** (2), 342–358.
159. Bugge, K., Staby, L., Kemplen, K.R., O'Shea, C., Bendsen, S.K., Jensen, M.K., Olsen, J.G., Skriver, K., et al., (2018). Structure of radical-induced cell death1 hub domain reveals a common αα-scaffold for disorder in transcriptional networks. *Structure*, **26** (5), 734–746.
160. Oldfield, C.J., Meng, J., Yang, J.Y., Yang, M.Q., Uversky, V.N., Dunker, A.K., (2008). Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genom.*, **9** (1), 1–20.
161. Teilum, K., Olsen, J.G., Kragelund, B.B., (2021). On the specificity of protein-protein interactions in the context of disorder. *Biochem. J.*,. in press.
162. Swanson, K.A., Knoepfler, P.S., Huang, K., Kang, R.S., Cowley, S.M., Laherty, C.D., Eisenman, R.N., Radhakrishnan, I., (2004). HBP1 and Mad1 repressors bind the Sin3 corepressor PAH2 domain with opposite helical orientations. *Nature Struct. Mol. Biol.*, **11** (8), 738–746.
163. Günther, S., Schlundt, A., Sticht, J., Roske, Y., Heinemann, U., Wiesmüller, K.H., Jung, G., Falk, K., et al., (2010). Bidirectional binding of invariant chain peptides to an MHC class II molecule. *Proc. Natl. Acad. Sci.*, **107** (51), 22219–22224.
164. Stein, A., Aloy, P., (2008). Contextual specificity in peptide-mediated protein interactions. *PloS One*, **3** (7), e2524.
165. Palopoli, N., González Foutel, N.S., Gibson, T.J., Chemes, L.B., (2018). Short linear motif core and flanking regions modulate retinoblastoma protein binding affinity and specificity. *Protein Eng. Des. Select.*, **31** (3), 69–77.
166. Prestel, A., Wichmann, N., Martins, J.M., Marabini, R., Kassem, N., Broendum, S.S., Otterlei, M., Nielsen, O., et al., (2019). The PCNA interaction motifs revisited: thinking outside the PIP-box. *Cell. Mol. Life Sci.*, **76** (24), 4923–4943.
167. Garcia-Pino, A., Balasubramanian, S., Wyns, L., Gazit, E., De Greve, H., Magnuson, R.D., Charlier, D., van Nuland, N.A., et al., (2010). Allostery and intrinsic disorder mediate transcription regulation by conditional cooperativity. *Cell*, **142** (1), 101–111.
168. Li, J., White, J.T., Saavedra, H., Wrabl, J.O., Motlagh, H. N., Liu, K., Sowers, J., Schroer, T.A., et al., (2017). Genetically tunable frustration controls allostery in an intrinsically disordered transcription factor. *Elife*, **6**, e30688.

169. Bugge, K., Brakti, I., Fernandes, C.B., Dreier, J.E., Lundsgaard, J.E., Olsen, J.G., Skriver, K., Kragelund, B. B., (2020). Interactions by disorder–a matter of context. *Front. Mol. Biosci.*, **7**
170. Skerker, J.M., Perchuk, B.S., Siryaporn, A., Lubin, E.A., Ashenberg, O., Goulian, M., Laub, M.T., (2008). Rewiring the specificity of two-component signal transduction systems. *Cell*, **133** (6), 1043–1054.
171. Burger, L., Van Nimwegen, E., (2008). Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method. *Mol. Syst. Biol.*, **4** (1), 165.
172. Tsaban, T., Varga, J.K., Avraham, O., Aharon, Z.B., Khramushin, A., Schueler-Furman, O., (2021). Harnessing protein folding neural networks for peptide-protein docking. *bioRxiv*,. https://doi.org/10.1101/2021.08.01.454656.
173. Ko, J., Lee, J., (2021). Can AlphaFold2 predict protein-peptide complex structures accurately? *bioRxiv*,. https://doi.org/10.1101/2021.07.27.453972.
174. Cunningham, J.M., Koytiger, G., Sorger, P.K., AlQuraishi, M., (2020). Biophysical prediction of protein–peptide interactions and signaling networks using machine learning. *Nature Methods*, **17** (2), 175–183.
175. Senicourt, L., le Maire, A., Allemand, F., Carvalho, J.E., Guee, L., Germain, P., Schubert, M., Bernadó, P., et al., (2021). Structural Insights into the Interaction of the Intrinsically Disordered Co-activator TIF2 with Retinoic Acid Receptor Heterodimer (RXR/RAR). *J. Mol. Biol.*, **433** (9), 166899.
176. Shlyueva, D., Stampfel, G., Stark, A., (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Rev. Genet.*, **15** (4), 272–286.
177. Avsec, Z., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., et al., (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *bioRxiv*,.
178. Berlow, R.B., Dyson, H.J., Wright, P.E., (2015). Functional advantages of dynamic protein disorder. *FEBS Letters*, **589** (19), 2433–2440.
179. Gao, A., Shrinivas, K., Lepeudry, P., Suzuki, H.I., Sharp, P.A., Chakraborty, A.K., (2018). Evolution of weak cooperative interactions for biological specificity. *Proc. Natl. Acad. Sci.*, **115** (47), E11053–E11060.
180. Rogers, J.M., Wong, C.T., Clarke, J., (2014). Coupled folding and binding of the disordered protein PUMA does not require particular residual structure. *J. Am. Chem. Soc.*, **136** (14), 5197–5200.
181. Sugase, K., Dyson, H.J., Wright, P.E., (2007). Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature*, **447** (7147), 1021–1025.
182. Iešmantavičius, V., Dogan, J., Jemth, P., Teilum, K., Kjaergaard, M., (2014). Helical propensity in an intrinsically disordered protein accelerates ligand binding. *Angew. Chem. Int. Ed.*, **53** (6), 1548–1551.
183. Teilum, K., Olsen, J.G., Kragelund, B.B., (2015). Globular and disordered–the non-identical twins in protein-protein interactions. *Front. Mol. Biosci.*, **2**, 40.
184. Robustelli, P., Piana, S., Shaw, D.E., (2020). Mechanism of coupled folding-upon-binding of an intrinsically disordered protein. *J. Am. Chem. Soc.*, **142** (25), 11092–11101.
185. Brzovic, P.S., Heikaus, C.C., Kisselev, L., Vernon, R., Herbig, E., Pacheco, D., Warfield, L., Littlefield, P., et al., (2011). The acidic transcription activator Gcn4 binds the mediator subunit Gal11/Med15 using a simple protein interface forming a fuzzy complex. *Mol. Cell*, **44** (6), 942–953.
186. Hendus-Altenburger, R., Pedraz-Cuesta, E., Olesen, C. W., Papaleo, E., Schnell, J.A., Hopper, J.T., Robinson, C. V., Pedersen, S.F., et al., (2016). The human Na+/H+ exchanger 1 is a membrane scaffold protein for extracellular signal-regulated kinase 2. *BMC Biol.*, **14** (1), 1–17.
187. Tillu, V.A., Rae, J., Gao, Y., Ariotti, N., Floetenmeyer, M., Kovtun, O., McMahon, K.A., Chaudhary, N., et al., (2021). Cavin1 intrinsically disordered domains are essential for fuzzy electrostatic interactions and caveola formation. *Nature Commun.*, **12** (1), 1–18.
188. Borgia, A., Borgia, M.B., Bugge, K., Kissling, V.M., Heidarsson, P.O., Fernandes, C.B., Sottini, A., Soranno, A., et al., (2018). Extreme disorder in an ultrahigh-affinity protein complex. *Nature*, **555** (7694), 61–66.
189. Schuler, B., Borgia, A., Borgia, M.B., Heidarsson, P.O., Holmstrom, E.D., Nettels, D., Sottini, A., (2020). Binding without folding–the biomolecular function of disordered polyelectrolyte complexes. *Curr. Opin. Struct. Biol.*, **60**, 66–76.
190. Dogan, J., Gianni, S., Jemth, P., (2014). The binding mechanisms of intrinsically disordered proteins. *Phys. Chem. Chem. Phys.*, **16** (14), 6323–6331.
191. Arai, M., Sugase, K., Dyson, H.J., Wright, P.E., (2015). Conformational propensities of intrinsically disordered proteins influence the mechanism of binding and folding. *Proc. Natl. Acad. Sci.*, **112** (31), 9614–9619.
192. Demarest, S.J., Martinez-Yamout, M., Chung, J., Chen, H., Xu, W., Dyson, H.J., Evans, R.M., Wright, P.E., (2002). Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature*, **415** (6871), 549–553.
193. Dogan, J., Schmidt, T., Mu, X., Engström, Å., Jemth, P., (2012). Fast association and slow transitions in the interaction between two intrinsically disordered protein domains. *J. Biol. Chem.*, **287** (41), 34316–34324.
194. Fuxreiter, M., Tompa, P., (2012). Fuzzy complexes: a more stochastic view of protein function. *Fuzziness*,, 1–14.
195. Olsen, J.G., Teilum, K., Kragelund, B.B., (2017). Behaviour of intrinsically disordered proteins in protein–protein complexes with an emphasis on fuzziness. *Cell. Mol. Life Sci.*, **74** (17), 3175–3183.
196. Erkine, A.M., (2018). 'Nonlinear'biochemistry of nucleosome detergents. *Trends Biochem. Sci.*, **43** (12), 951–959.
197. Sigler, P.B., (1988). Acid blobs and negative noodles. *Nature*, **333** (6170), 210–212.
198. Staller, M.V., Holehouse, A.S., Swain-Lenz, D., Das, R.K., Pappu, R.V., Cohen, B.A., (2018). A high-throughput mutational scan of an intrinsically disordered acidic transcriptional activation domain. *Cell Syst.*, **6** (4), 444–455.
199. Ravarani, C.N., Erkina, T.Y., De Baets, G., Dudman, D. C., Erkine, A.M., Babu, M.M., (2018). High-throughput discovery of functional disordered regions: investigation of transactivation domains. *Mol. Syst. Biol.*, **14** (5), e8190.

200. Erijman, A., Kozlowski, L., Sohrabi-Jahromi, S., Fishburn, J., Warfield, L., Schreiber, J., Noble, W.S., Söding, J., et al., (2020). A High-Throughput screen for transcription activation domains reveals their sequence features and permits prediction by deep learning. *Mol. Cell*, **78** (5), 890–902.
201. Tycko, J., DelRosso, N., Hess, G.T., Banerjee, A., Mukund, A., Van, M.V., Ego, B.K., Yao, D., et al., (2020). High-throughput discovery and characterization of human transcriptional effectors. *Cell*,.
202. Sanborn, A.L., Yeh, B.T., Feigerle, J.T., Hao, C.V., Townshend, R.J., Lieberman-Aiden, E., Dror, R.O., Kornberg, R.D., (2021). Simple biochemical features underlie transcriptional activation domain diversity and dynamic, fuzzy binding to Mediator. *Elife*, **10**, e68068.
203. Staller, M.V., Ramirez, E., Holehouse, A.S., Pappu, R.V., Cohen, B.A., (2021). Design principles of acidic transcriptional activation domains. *bioRxiv*,. p. 2020–10.
204. Griffith, D., Holehouse, A.S., (2021). PARROT: a flexible recurrent neural network framework for analysis of large protein datasets. *bioRxiv*,.
205. Banani, S.F., Lee, H.O., Hyman, A.A., Rosen, M.K., (2017). Biomolecular condensates: organizers of cellular biochemistry. *Nature Rev. Mol. Cell Biol.*, **18** (5), 285–298.
206. Peran, I., Mittag, T., (2020). Molecular structure in biomolecular condensates. *Curr. Opin. Struct. Biol.*, **60**, 17–26.
207. Dignon, G.L., Best, R.B., Mittal, J., (2020). Biomolecular phase separation: From molecular driving forces to macroscopic properties. *Ann. Rev. Phys. Chem.*, **71**, 53–75.
208. Choi, J.M., Holehouse, A.S., Pappu, R.V., (2020). Physical principles underlying the complex biology of intracellular phase transitions. *Ann. Rev. Biophys.*, **49**, 107–133.
209. Li, P., Banjade, S., Cheng, H.C., Kim, S., Chen, B., Guo, L., Llaguno, M., Hollingsworth, J.V., et al., (2012). Phase transitions in the assembly of multivalent signalling proteins. *Nature*, **483** (7389), 336–340.
210. Bouchard, J.J., Otero, J.H., Scott, D.C., Szulc, E., Martin, E.W., Sabri, N., Granata, D., Marzahn, M.R., et al., (2018). Cancer mutations of the tumor suppressor SPOP disrupt the formation of active, phase-separated compartments. *Mol. Cell*, **72** (1), 19–36.
211. Wang, J., Choi, J.M., Holehouse, A.S., Lee, H.O., Zhang, X., Jahnel, M., Maharana, S., Lemaitre, R., et al., (2018). A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell*, **174** (3), 688–699.
212. Panagiotopoulos, A.Z., Wong, V., Floriano, M.A., (1998). Phase equilibria of lattice polymers from histogram reweighting Monte Carlo simulations. *Macromolecules*, **31** (3), 912–918.
213. Lin, Y.H., Chan, H.S., (2017). Phase separation and single-chain compactness of charged disordered proteins are strongly correlated. *Biophys. J.*, **112** (10), 2043–2046.
214. Dignon, G.L., Zheng, W., Best, R.B., Kim, Y.C., Mittal, J., (2018). Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. *Proc. Natl. Acad. Sci.*, **115** (40), 9929–9934.
215. Li, Q., Peng, X., Li, Y., Tang, W., Zhu, J., Huang, J., Qi, Y., Zhang, Z., (2020). LLPSDB: a database of proteins undergoing liquid–liquid phase separation in vitro. *Nucl. Acids Res.*, **48** (D1), D320–D327.
216. Li, Q., Wang, X., Dou, Z., Yang, W., Huang, B., Lou, J., Zhang, Z., (2020). Protein Databases Related to Liquid-Liquid Phase Separation. *Int. J. Mol. Sci.*, **21** (18), 6796.
217. Mészáros, B., Erdős, G., Szabó, B., Schád, É., Tantos, Á., Abukhairan, R., Horváth, T., Murvai, N., et al., (2020). PhaSePro: the database of proteins driving liquid–liquid phase separation. *Nucl. Acids Res.*, **48** (D1), D360–D367.
218. You, K., Huang, Q., Yu, C., Shen, B., Sevilla, C., Shi, M., Hermjakob, H., Chen, Y., et al., (2020). PhaSepDB: a database of liquid–liquid phase separation related proteins. *Nucl. Acids Res.*, **48** (D1), D354–D359.
219. Ning, W., Guo, Y., Lin, S., Mei, B., Wu, Y., Jiang, P., Tan, X., Zhang, W., et al., (2020). DrLLPS: a data resource of liquid–liquid phase separation in eukaryotes. *Nucleic Acids Res.*, **48** (D1), D288–D295.
220. Vernon, R.M., Chong, P.A., Tsang, B., Kim, T.H., Bah, A., Farber, P., Lin, H., Forman-Kay, J.D., (2018). Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *elife*, **7** (e31486)
221. Hardenberg, M., Horvath, A., Ambrus, V., Fuxreiter, M., Vendruscolo, M., (2020). Widespread occurrence of the droplet state of proteins in the human proteome. *Proc. Natl. Acad. Sci.*, **117** (52), 33254–33262.
222. van Mierlo, G., Jansen, J.R., Wang, J., Poser, I., van Heeringen, S.J., Vermeulen, M., (2021). Predicting protein condensate formation using machine learning. *Cell Rep.*, **34** (5), 108705.
223. Raimondi, D., Orlando, G., Michiels, E., Pakravan, D., Bratek-Skicki, A., Van Den Bosch, L., Moreau, Y., Rousseau, F., et al., (2021). In-silico prediction of in-vitro protein liquid-liquid phase separation experiments outcomes with multi-head neural attention. *Bioinformatics*,.
224. Saar, K.L., Morgunov, A.S., Qi, R., Arter, W.E., Krainer, G., Knowles, T.P., et al., (2021). Learning the molecular grammar of protein condensates from sequence determinants and embeddings. *Proc. Natl. Acad. Sci.*, **118** (15)
225. Pancsa, R., Vranken, W., Mészáros, B., (2021). Computational resources for identifying and describing proteins driving liquid–liquid phase separation. *Briefings Bioinformat.*,.
226. Yu, C., Shen, B., You, K., Huang, Q., Shi, M., Wu, C., Chen, Y., Zhang, C., et al., (2021). Proteome-scale analysis of phase-separated proteins in immunofluorescence images. *Briefings Bioinformat.*, **22** (3), bbaa187.
227. Chiti, F., Stefani, M., Taddei, N., Ramponi, G., Dobson, C. M., (2003). Rationalization of the effects of mutations on peptide andprotein aggregation rates. *Nature*, **424** (6950), 805–808.
228. Fernandez-Escamilla, A.M., Rousseau, F., Schymkowitz, J., Serrano, L., (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature Biotechnol.*, **22** (10), 1302–1306.
229. Pawar, A.P., Dubay, K.F., Zurdo, J., Chiti, F., Vendruscolo, M., Dobson, C.M., (2005). Prediction of aggregation-prone and aggregation-susceptible regions in proteins associated with neurodegenerative diseases. *J. Mol. Biol.*, **350** (2), 379–392.
230. Dignon, G.L., Zheng, W., Kim, Y.C., Best, R.B., Mittal, J., (2018). Sequence determinants of protein phase behavior from a coarse-grained model. *PLoS Comput. Biol.*, **14** (1), e1005941.

231. Bremer, A., Farag, M., Borcherds, W.M., Peran, I., Martin, E.W., Pappu, R.V., Mittag, T., (2021). Deciphering how naturally occurring sequence features impact the phase behaviors of disordered prion-like domains. *bioRxiv*,.
232. Statt, A., Casademunt, H., Brangwynne, C.P., Panagiotopoulos, A.Z., (2020). Model for disordered proteins with strongly sequence-dependent liquid phase behavior. *J. Chem. Phys.*, **152** (7), 075101.
233. Hazra, M.K., Levy, Y., (2020). Charge pattern affects the structure and dynamics of polyampholyte condensates. *Phys. Chem. Chem. Phys.*, **22** (34), 19368–19375.
234. Amin, A.N., Lin, Y.H., Das, S., Chan, H.S., (2020). Analytical theory for sequence-specific binary fuzzy complexes of charged intrinsically disordered proteins. *J. Phys. Chem. B*, **124** (31), 6709–6720.
235. Nott, T.J., Petsalaki, E., Farber, P., Jervis, D., Fussner, E., Plochowietz, A., Craggs, T.D., Bazett-Jones, D.P., et al., (2015). Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol. Cell*, **57** (5), 936–947.
236. Monahan, Z., Ryan, V.H., Janke, A.M., Burke, K.A., Rhoads, S.N., Zerze, G.H., O'Meally, R., Dignon, G.L., et al., (2017). Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity. *EMBO J.*, **36** (20), 2951–2967.
237. Lu, H., Yu, D., Hansen, A.S., Ganguly, S., Liu, R., Heckert, A., Darzacq, X., Zhou, Q., (2018). Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. *Nature*, **558** (7709), 318–323.
238. Hofwebe, M., Hutten, S., Bourgeois, B., Spreitzer, E., Niedner-Boblenz, A., Schifferer, M., Ruepp, M.D., Simons, M., et al., (2018). Phase separation of FUS is suppressed by its nuclear import receptor and arginine methylation. *Cell*, **173** (3), 706–719.
239. Hofweber, M., Dormann, D., (2019). Friend or foe—Post-translational modifications as regulators of phase separation and RNP granule dynamics. *J. Biol. Chem.*, **294** (18), 7137–7150.
240. Uversky, V.N., Oldfield, C.J., Midic, U., Xie, H., Xue, B., Vucetic, S., Iakoucheva, L.M., Obradovic, Z., et al., (2009). Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. *BMC Genom.*, **10** (1), 1–17.
241. Uversky, V.N., (2015). Intrinsically disordered proteins and their (disordered) proteomes in neurodegenerative disorders. *Front. Aging Neurosci.*, **7**, 18.
242. Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradović, Z., Dunker, A.K., (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.*, **323** (3), 573–584.
243. Deiana, A., Forcelloni, S., Porrello, A., Giansanti, A., (2019). Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell. *PloS One*, **14** (8), e0217889.
244. Mészáros, B., Hajdu-Soltész, B., Zeke, A., Dosztányi, Z., (2021). Mutations of Intrinsically Disordered Protein Regions Can Drive Cancer but Lack Therapeutic Strategies. *Biomolecules*, **11** (3), 381.
245. Nielsen, J.T., Mulder, F.A., (2019). Quality and bias of protein disorder predictors. *Sci. Rep.*, **9** (1), 1–11.
246. Dass, R., Mulder, F.A., Nielsen, J.T., (2020). ODiNPred: Comprehensive prediction of protein order and disorder. *Sci. Rep.*, **10** (1), 1–16.
247. Babu, M.M., van der Lee, R., de Groot, N.S., Gsponer, J., (2011). Intrinsically disordered proteins: regulation and disease. *Curr. Opin. Struct. Biol.*, **21** (3), 432–440.
248. Stefl, S., Nishi, H., Petukh, M., Panchenko, A.R., Alexov, E., (2013). Molecular mechanisms of disease-causing missense mutations. *J. Mol. Biol.*, **425** (21), 3919–3936.
249. Sahni, N., Yi, S., Taipale, M., Bass, J.I.F., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., et al., (2015). Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, **161** (3), 647–660.
250. Stein, A., Fowler, D.M., Hartmann-Petersen, R., Lindorff-Larsen, K., (2019). Biophysical and mechanistic models for disease-causing protein variants. *Trends Biochem. Sci.*, **44** (7), 575–588.
251. Vacic, V., Markwick, P.R., Oldfield, C.J., Zhao, X., Haynes, C., Uversky, V.N., Iakoucheva, L.M., (2012). Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. *PLoS Comput. Biol.*, **8** (10), e1002709.
252. Meyer, K., Kirchner, M., Uyar, B., Cheng, J.Y., Russo, G., Hernandez-Miranda, L.R., Szymborska, A., Zauber, H., et al., (2018). Mutations in disordered regions can cause disease by creating dileucine motifs. *Cell*, **175** (1), 239–253.
253. Li, Y., Zhang, Y., Li, X., Yi, S., Xu, J., (2019). Gain-of-function mutations: an emerging advantage for cancer biology. *Trends Biochem. Sci.*, **44** (8), 659–674.
254. Wong, E.T., So, V., Guron, M., Kuechler, E.R., Malhis, N., Bui, J.M., Gsponer, J., (2020). Protein–protein interactions mediated by intrinsically disordered protein regions are enriched in missense mutations. *Biomolecules*, **10** (8), 1097.
255. Casadio, R., Vassura, M., Tiwari, S., Fariselli, P., Luigi, Martelli P., (2011). Correlating disease-related mutations to their effect on protein stability: A large-scale analysis of the human proteome. *Hum. Mutat.*, **32** (10), 1161–1170.
256. Matreyek, K.A., Starita, L.M., Stephany, J.J., Martin, B., Chiasson, M.A., Gray, V.E., Kircher, M., Khechaduri, A., et al., (2018). Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature Genet.*, **50** (6), 874–882.
257. Cagiada, M., Johansson, K.E., Valančiūtė, A., Nielsen, S. V., Hartmann-Petersen, R., Yang, J.J., Fowler, D.M., Stein, A., et al., (2021). Understanding the origins of loss of protein function by analyzing the effects of thousands of variants on activity and abundance. *Mol. Biol. Evol.*,, msab095.
258. Chhabra, Y., Wong, H.Y., Nikolajsen, L., Steinocher, H., Papadopulos, A., Tunny, K., Meunier, F., Smith, A., et al., (2018). A growth hormone receptor SNP promotes lung cancer by impairment of SOCS2-mediated degradation. *Oncogene*, **37** (4), 489–501.
259. Davey, N.E., Cyert, M.S., Moses, A.M., (2015). Short linear motifs–ex nihilo evolution of protein regulation. *Cell Commun. Signal.*, **13** (1), 1–15.
260. Grazioli, G., Martin, R.W., Butts, C.T., (2019). Comparative exploratory analysis of intrinsically disordered protein dynamics using machine learning and network analytic methods. *Front. Mol. Biosci.*, **6**, 42.
261. van der Lee, R., Lang, B., Kruse, K., Gsponer, J., de Groot, N.S., Huynen, M.A., Matouschek, A., Fuxreiter, M., et al., (2014). Intrinsically disordered segments affect protein half-life in the cell and during evolution. *Cell Rep.*, **8** (6), 1832–1844.

262. Uversky, V.N., (2013). The most important thing is the tail: multitudinous functionalities of intrinsically disordered protein termini. *FEBS Letters*, **587** (13), 1891–1901.
263. Geffen, Y., Appleboim, A., Gardner, R.G., Friedman, N., Sadeh, R., Ravid, T., (2016). Mapping the landscape of a eukaryotic degronome. *Mol. Cell*, **63** (6), 1055–1065.
264. Koren, I., Timms, R.T., Kula, T., Xu, Q., Li, M.Z., Elledge, S.J., (2018). The eukaryotic proteome is shaped by E3 ubiquitin ligases targeting C-terminal degrons. *Cell*, **173** (7), 1622–1635.
265. Aguzzi, A., Altmeyer, M., (2016). Phase separation: linking cellular compartmentalization to disease. *Trends Cell Biol.*, **26** (7), 547–558.
266. Shin, Y., Brangwynne, C.P., (2017). Liquid phase condensation in cell physiology and disease. *Science*, **357** (6357)
267. Elbaum-Garfinkle, S., (2019). Matter over mind: Liquid phase separation and neurodegeneration. *J. Biol. Chem.*, **294** (18), 7160–7168.
268. Boija, A., Klein, I.A., Young, R.A., (2021). Biomolecular condensates and cancer. *Cancer Cell*,.
269. Cai, D., Liu, Z., Lippincott-Schwartz, J., (2021). Biomolecular Condensates and Their Links to Cancer Progression. *Trends Biochem. Sci.*,.
270. Alberti, S., Hyman, A.A., (2021). Biomolecular condensates at the nexus of cellular stress, protein aggregation disease and ageing. *Nature Rev. Mol. Cell Biol.*,, 1–18.
271. Tsang, B., Pritišanac, I., Scherer, S.W., Moses, A.M., Forman-Kay, J.D., (2020). Phase Separation as a Missing Mechanism for Interpretation of Disease Mutations. *Cell*, **183** (7), 1742–1756.
272. Biesaga, M., Frigolé-Vivas, M., Salvatella, X., (2021). Intrinsically disordered proteins and biomolecular condensates as drug targets. *Curr. Opin. Chem. Biol.*, **62**, 90–100.
273. Riesselman, A.J., Ingraham, J.B., Marks, D.S., (2018). Deep generative models of genetic variation capture the effects of mutations. *Nature Methods*, **15** (10), 816–822.
274. Livesey, B.J., Marsh, J.A., (2020). Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol. Syst. Biol.*, **16** (7), e9380.
275. Zarin, T., Strome, B., Ba, A.N.N., Alberti, S., Forman-Kay, J.D., Moses, A.M., (2019). Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *Elife*, **8**, e46883.
276. Zhou, J.B., Xiong, Y., An, K., Ye, Z.Q., Wu, Y.D., (2020). IDRMutPred: predicting disease-associated germline nonsynonymous single nucleotide variants (nsSNVs) in intrinsically disordered regions. *Bioinformatics*, **36** (20), 4977–4983.
277. Zarin, T., Strome, B., Peng, G., Pritišanac, I., Forman-Kay, J.D., Moses, A.M., (2021). Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. *Elife*, **10**, e60220.
278. Fowler, D.M., Fields, S., (2014). Deep mutational scanning: a new style of protein science. *Nature Methods*, **11** (8), 801–807.
279. Starita, L.M., Ahituv, N., Dunham, M.J., Kitzman, J.O., Roth, F.P., Seelig, G., Shendure, J., Fowler, D.M., (2017). Variant interpretation: functional assays to the rescue. *Am. J. Hum. Genet.*, **101** (3), 315–325.
280. Rogers, J.M., Passioura, T., Suga, H., (2018). Nonproteinogenic deep mutational scanning of linear and cyclic peptides. *Proc. Natl. Acad. Sci.*, **115** (43), 10959–10964.
281. Bolognesi, B., Faure, A.J., Seuma, M., Schmiedel, J.M., Tartaglia, G.G., Lehner, B., (2019). The mutational landscape of a prion-like domain. *Nature Commun.*, **10** (1), 1–12.
282. Gray, V.E., Sitko, K., Kameni, F.Z.N., Williamson, M., Stephany, J.J., Hasle, N., Fowler, D.M., (2019). Elucidating the molecular determinants of Aβ aggregation with deep mutational scanning. *G3: Genes Genomes Genet.*, **9** (11), 3683–3689.
283. Newberry, R.W., Leong, J.T., Chow, E.D., Kampmann, M., DeGrado, W.F., (2020). Deep mutational scanning reveals the structural basis for α-synuclein activity. *Nature Chem. Biol.*, **16** (6), 653–659.
284. Newberry, R.W., Arhar, T., Costello, J., Hartoularos, G.C., Maxwell, A.M., Naing, Z.Z.C., Pittman, M., Reddy, N.R., et al., (2020). Robust Sequence Determinants of α-Synuclein Toxicity in Yeast Implicate Membrane Binding. *ACS Chem. Biol.*, **15** (8), 2137–2153.
285. Seuma, M., Faure, A.J., Badia, M., Lehner, B., Bolognesi, B., (2021). The genetic landscape for amyloid beta fibril nucleation accurately discriminates familial Alzheimer's disease mutations. *Elife*, **10**, e63364.
286. Moult, J., Pedersen, J.T., Judson, R., Fidelis, K., (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins*, **23**, ii–iv.