Unifying Coarse-grained Force Fields for Folded and Disordered Proteins

Andrew P. Latham, Bin Zhang

Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA

Abstract

Liquid-liquid phase separation drives the formation of biological condensates that play essential roles in transcriptional regulation and signal sensing. Computational modeling could provide high-resolution structural characterizations of these condensates and help uncover physicochemical interactions that dictate their stability. However, many protein molecules involved in phase separation often contain multiple ordered domains connected with flexible, structureless linkers. Simulating such proteins necessitates force fields with consistent accuracy for both folded and disordered proteins. We provide a critical review of existing coarse-grained force fields for disordered proteins and highlight the challenges in their application to folded proteins. After discussing existing algorithms for force field parameterization, we propose an optimization strategy that should lead to computer models with improved transferability across protein types.

Email address: binz@mit.edu (Bin Zhang)

Introduction

Proteins perform the majority of tasks within the cell. Their proper functions were believed to depend crucially on maintaining unique and stable three-dimensional (3D) structures. The structure-function relationship has led to significant efforts in studying the protein folding problem. Yet, recent studies suggest that a considerable fraction of eukaryotic proteomes is disordered [1, 2]. These intrinsically disordered proteins (IDPs) challenge traditional concepts of the structure-function relationship since their native states do not correspond to unique structures, but consist of an ensemble of heterogeneous conformations [3, 4]. The structural heterogeneity and disorderness could be of functional importance. They may be advantageous for multivalent interactions and mediating the formation of biological condensates through liquid-liquid phase separation [5–7].

The lack of well-defined 3D conformations for IDPs has made their structural characterization difficult. Techniques such as Förster resonance energy transfer (FRET), and small-angle X-ray scattering (SAXS), while offering valuable insights into the conformational ensemble, could not resolve the structural heterogeneity with atomic resolution. The experimental challenges in describing IDPs make computational modeling an attractive alternative. However, many existing force fields were optimized for ordered proteins and struggle to capture the size and flexibility of IDPs [8, 9]. As such, numerous groups have revised existing force fields or created new ones to ensure their accuracy in modeling IDPs [10–16].

Despite the progress in force field development, state-of-the-art computer models still face challenges describing folded and disordered proteins with consistent accuracy. As the same 20 amino acids encode both protein types, it should be possible, in principle, to create a unified force field for their modeling. Such a force field will enjoy a wide variety of applications. It would allow more accurate characterization of the stability of folded and misfolded structures to study disordered proteins that fold upon binding to a partner. In addition, it will enable simulations of proteins that include ordered regions separated by flexible, disordered linkers, a feature commonly seen in those that drive the formation of membraneless organelles.

In this review, we track the progress toward developing force fields applicable to both folded and disordered proteins. We review existing force fields for simulating IDPs, with a focus on

coarse-grained models. We highlight the inherent difficulty for applying force fields optimized for IDPs to study folded proteins or vice versa due to their distinct compositional bias. Force field parameterization algorithms, including both top-down and bottom-up approaches, are then discussed in the context of their applicability for ensuring consistent performance for both protein types. Finally, we discuss an optimization strategy that emphasizes the inclusion of folded and disordered proteins in training set to help improve force field transferability.

Coarse-grained Force Fields for Disordered Proteins

All-atom force fields have been rather successful at studying protein folding and predicting protein structures. Improvements made in the torsion potentials and nonbonded interactions further allowed their application to disordered proteins, as discussed in several recent reviews [14–16]. However, conformational sampling, which is crucial for disordered proteins, can be computationally challenging for single proteins and may become prohibitively costly for studying the aggregation of multiple proteins. Therefore, there is broad interest in developing coarse-grained force fields for simulating IDPs.

Coarse-grained explicit solvent models could strike a good balance between accuracy and efficiency. The MARTINI force field follows a four-to-one mapping strategy to represent four heavy atoms with a single coarse-grained bead and has been used to study the phase behavior of IDPs [17–20]. However, achieving quantitative accuracy often requires further fine-tuning the force field [18–20], including strengthening protein-water interactions or weakening protein-protein interactions. A similarly coarse-grained force field, SIRAH, was introduced by Pantano and coworkers [21]. Unlike MARTINI, SIRAH avoids using artificial constraints to fix secondary structures and could, in principle, predict protein structures de novo. Its well-balanced secondary structure potentials succeed in stabilizing the crystal structure of folded proteins [21] and reproducing the conformational flexibility of IDPs [22].

Numerous groups have also made progress in developing coarse-grained implicit solvent models, which are highly efficient and ideal for large-scale aggregation and phase separation simulations. AWSEM-IDP [23] utilizes the framework from the Associative Memory, Water Mediated, Structure, and Energy Model (AWSEM) introduced by Wolynes and coworkers [24]. To model

IDPs, Wu et al. reduced the strength of secondary structure terms and introduced biasing potentials on the radius of gyration (R_g), the value of which can be obtained from SAXS experiments or all-atom simulations. Baul et al. [25] introduced the self-organized polymer (SOP) coarse-grained model for IDPs by weakening the interaction potential among amino acids from the original SOP model [26]. SOP-IDP succeeded at resolving the sequence-specific heterogeneity between Aβ40 and Aβ42 [27]. Mioduszewski et al. [28] introduced a pseudo-improper dihedral potential to capture backbone and side-chain interactions in a one-bead-per-residue model for IDPs.

The hydrophobicity scale (HPS) model describes the interactions among amino acids with a simplified treatment of electrostatic energy and a short-range contact potential parameterized based on amino acid hydrophobicity [29]. It has been successfully applied to study the liquid-liquid phase separation of low complexity domains [30]. This model has been improved recently to capture temperature-dependent effects on solvent-mediated interactions [31], to account for cation-π interactions [32], and to better reproduce IDP radius of gyrations [33–35]. Latham and Zhang parameterized MOFF-IDP by introducing correctional contact potentials derived from SAXS data to the HPS model [36]. They showed that MOFF-IDP can reproduce R_g for a set of IDPs and succeed at de novo predictions, including the conformational change upon phosphorylation [37].

The various implicit solvent models differ in their resolutions and efficiency, and are suited for investigating different problems. For example, AWSEM-IDP brings along all the benefits of the original model, which uses three beads to represent each amino acid. In particular, AWSEM adopts a sophisticated energy function with many-body potentials and was shown to predict protein structures [24] and protein-protein binding interfaces [38] well. On the other hand, AWSEM-IDP is also computationally more expensive than HPS and related models using only one bead per amino acid. The difference in efficiency can be substantial when simulating large systems, rendering HPS and related models more appealing for phase separation studies.

Inconsistency between Folded and Disordered Protein Force Fields

It's worth noting that the coarse-grained IDP force fields, in general, are not applicable to globular proteins. Many of the force fields were introduced for studying large-scale simulations of liquid-liquid phase separation and used somewhat simplified representations, with only one or two

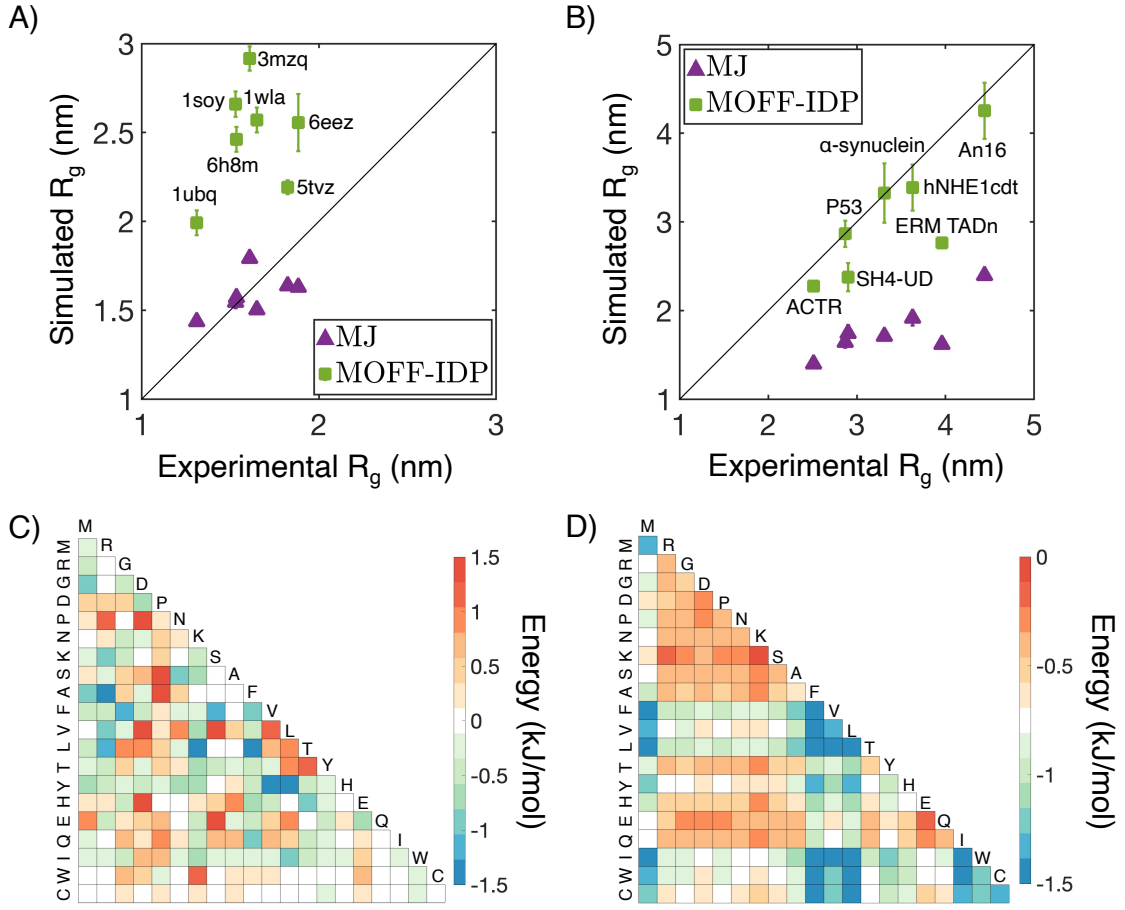

Figure 1: Existing coarse-grained protein force fields are often limited to either folded or disordered proteins, but not both. For example, MJ performs well on folded (A) but not disordered (B) proteins. The opposite trend can be seen for MOFF-IDP, an IDP force field introduced in Ref. [36]. The interaction matrices between amino acids for the two force fields are shown in parts C (MOFF-IDP) and D (MJ). The energy function for MJ can be found in Ref. [39], and interactions among amino acids are based on a scaled Miyazawa-Jernigan (MJ) potential by a factor of 0.6.

beads for the amino acids. The models were not expected to predict tertiary structures of folded proteins with high accuracy at this resolution. However, many of the force fields also over-predict the radius of gyration for globular proteins, indicating too weakened interactions among amino acids (Figure 1A). Similar effects are seen in force fields parameterized for folded proteins, which are often not transferable to IDPs and tend to predict overly collapsed structures (Figure 1B).

The inconsistency between force fields for folded and disordered proteins is not unique to coarse-grained models. All-atom force fields suffer similar issues. Significant efforts have been

devoted to reparameterize the existing force fields to improve their accuracy at modeling disordered proteins [14–16]. However, as pointed out by Shaw and coworkers [40], many atomistic force fields still struggle at achieving consistent accuracy for modeling the size and secondary structure propensities for disordered proteins and the tertiary structures of folded proteins.

The difficulty in parameterizing force fields with consistent accuracy for both protein types can be partly attributed to their distinct sequence composition. Many IDPs are depleted of hydrophobic residues, which promote collapse and folding of globular proteins [41]. Instead, they favor stretches of polar, uncharged residues. Such motifs prevent secondary structure formation but may still drive protein compaction into structureless globules due to the favorable self-solvation [42]. Alternatively, some IDPs possess a higher frequency of charged amino acids and leverage the overall charge composition and patterning for their structural features [43]. Because of the lack of overlap in the sequence space, there is no guarantee that force fields parameterized primarily on one type of protein will be transferable to the other.

Algorithms for Coarse-grained Force Field Parameterization

Could one live with two sets of force fields for folded and disordered proteins, respectively? While intellectually less satisfying, such a solution could still be of practical use. The answer is, unfortunately, no. A survey on human proteins has revealed that a considerable amount of residues (30%) were found to be disordered for a significant fraction (24%) of proteins [42]. Therefore, many proteins contain a mixture of domains with distinct structural features and cannot be classified into the binary category of folded or disordered. To study such proteins, force fields that provide consistent treatment for both protein types must be introduced. Algorithms developed for optimizing force fields of globular proteins [8, 9], which we group into top-down and bottom-up approaches, offer inspirations on how one might achieve such consistency.

Top-down approaches rely on experimental data for force field parameterization. For ordered proteins, the large set of high-resolution structures resolved by X-ray crystallography provide a valuable resource. In addition, the energy landscape theory for protein folding [44–46] offers critical insight into the interactions among amino acids. For a protein to fold reliably into the native state or the crystal structure, the contacts found in the native state must be stronger than the ones

found in unfolded or non-native conformations. This constraint is often described as the folding temperature (T_f) to be higher than the glass transition temperature (T_g), or pictorially, the funneled energy landscape [47]. Numerous algorithms have been introduced to parameterize coarse-grained force fields that satisfy constraints from the energy landscape theory, including optimization of the ratio $T_\mathrm{f}/T_\mathrm{g}$ [48], Z-score optimization [49], maximizing the energy gap between the native and non-native conformation [50], etc.

Bottom-up approaches typically start with an ensemble of structures collected from all-atom simulations, and differ in the specific properties of the ensemble used for force field parameterization. For example, iterative Boltzmann inversion (IBI) [51], and inverse Monte Carlo [52] approaches attempt to match radial distribution functions (RDF) between pairs of particles computed from coarse-grained and atomistic simulations. We note that the ideal target property to be reproduced should be the probability distribution of the coarse-grained structural ensemble. The RDF corresponds to a lower-dimensional projection of this distribution. Due to the loss of information upon projection, even a perfect reproduction of RDF does not guarantee an accurate approximation of the original, high-dimensional conformational distribution [53]. The force matching method [54–56] aims to reproduce the forces acting on coarse-grained sites estimated from the atomistic structural ensemble. Shell introduced the relative entropy algorithm to minimize entropy loss upon coarse-graining [57]. Both methods strive to improve the agreement between conformation distributions estimated from coarse-grained and all-atom models.

Machine learning based methods have gained popularity in recent years [58]. In particular, neural networks provide flexible fitting of complex functions and are ideal for parameterizing high dimensional free energy surfaces and probability distributions from mean forces [59, 60]. Recently, Wang et al. introduced the CGnets method to directly parameterize coarse-grained models [61, 62]. CGnets was shown to out-perform traditional force matching methods in reproducing crucial features of the free energy surface. Alternatively, the free energy surface can be accurately represented by deep generative models, as shown by Noé et al. [63] and others [64–66]. Deep generative models do not require mean forces, and directly parameterize the probability distributions using conformations collected from MD simulations via maximum likelihood optimization. We note that current studies on machine learning based methods have mainly focused on param-

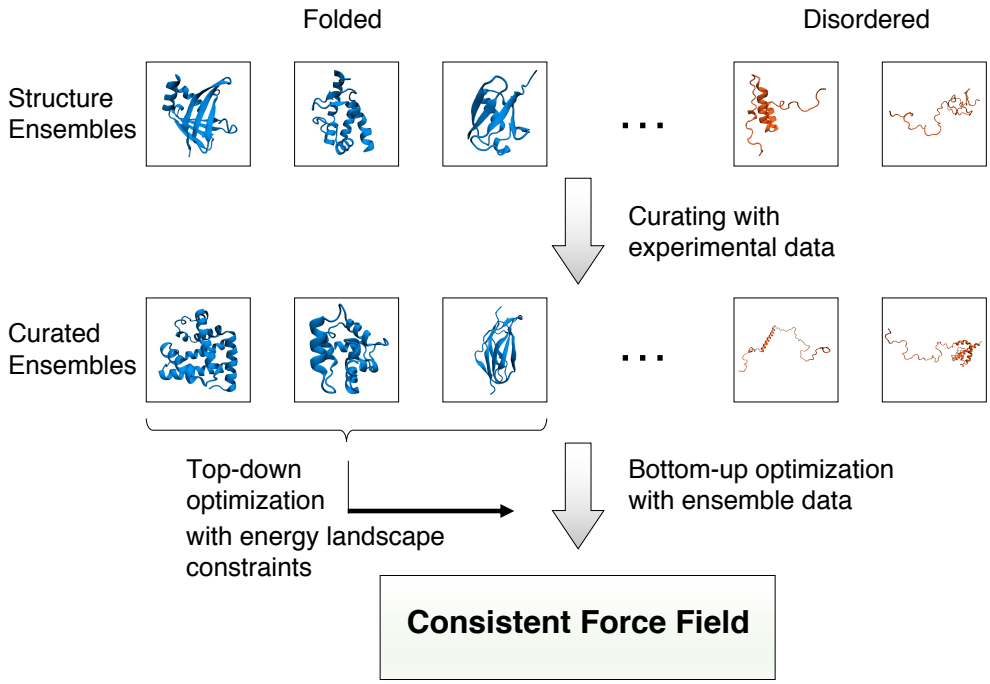

Figure 2: Illustration of the optimization strategy that combines top-down and bottom-up approaches to derive force fields with consistent accuracy. The training data used for parameter optimization consist of structure ensembles collected for a list of folded and disordered proteins. Ideally, the ensembles should be generated from atomistic simulations and further curated with experimental data. Bottom-up approaches such as energy matching or relative entropy minimization can be used to derive coarse-grained force field parameters that best approximate the conformational distribution of each protein. Importantly, top-down approaches can be introduced as additional constraints to enforce lower energy of the native state for folded proteins.

eterizing system-specific models. Additional work is needed to demonstrate their usefulness for deriving transferable force fields.

Strategy for Deriving Unified Force Fields with Consistent Accuracy

The use of experimental data to benchmark force fields as in top-down approaches could help ensure their transferability. Vitalis and Pappu introduced ABSINTH, an atomistic implicit solvent model that describes the solvation free energy using a combination of a direct mean-field interaction (DMFI) and the screening of polar interactions [67]. Parameters of ABSINTH were chosen to stabilize the folded states of two small proteins and reproduce NMR coupling constants and the

polymeric properties of intrinsically disordered peptides. More recently, the force field has been updated with improved dihedral angles [68] and used to study proteins that undergo liquid-liquid phase separation [69]. Ferrie and Petersson introduced a reweighting scheme to switch fragment memory libraries between the two sets that reproduce secondary structure propensities for folded and disordered proteins, respectively [70]. When implemented into Rosetta Modeling Suite, the authors showed that the platform now performs well for predicting the structures for a list of folded and disordered proteins.

Since modern force fields often consist of a large set of parameters, a manual, systematic search of the entire parameter space can be challenging and even infeasible for a large set of experimental data. In that regard, bottom-up methods mentioned in the previous section are advantageous to enable near-autonomous force field optimization. Using a similar functional form as in ABSINTH for the solvation free energy, Bottaro et al. carried out systematic optimizations of parameters in the potential to match explicit solvent simulation data for an $\alpha-$helical peptide and the GB1 hairpin [71]. They found that the resulting force field, EEF1-SB, performs well for unstructured proteins with an increased sampling of expanded conformations while maintaining the native structure of several folded proteins.

Combining top-down and bottom-up approaches may lead to new force field optimization strategies that are particularly helpful at ensuring the consistency between folded and disordered proteins (Figure 2). For example, as in bottom-up approaches, the coarse-grained force field could be parameterized using data collected from all-atom simulations. As all-atom force fields themselves have not yet achieved the desired accuracy, it is crucial to curate the simulated structural ensemble with experimental data, for example, via maximum entropy optimization [72–78]. Since the dataset would inevitably be finite, it is helpful to enforce constraints based on the energy landscape theory for ordered proteins as in top-down approaches to reduce the parameter space further and improve the robustness of force field optimization.

The optimization strategy introduced by Latham and Zhang offers some hints on how to combine different approaches in practice [39]. They constructed the reference structural ensembles using a coarse-grained model parameterized with the Miyazawa-Jernigan (MJ) potential. The ensembles, which included seven folded and sixteen disordered proteins, were corrected with SAXS

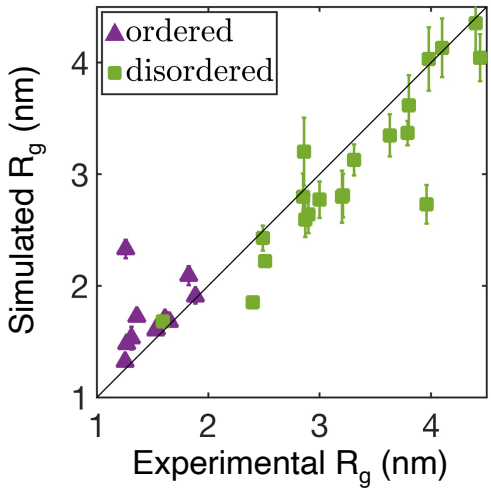

Figure 3: The force field introduced by Latham and Zhang in Ref. [39], MOFF, achieves consistent accuracy in predicting the size of both ordered (purple) and disordered (green) proteins.

data to ensure that the simulated R_g of different proteins match experimental values. Force field parameters were then tuned to reproduce the relative energy, and therefore, the probability density of individual conformations. The particular energy matching scheme introduced by the authors allowed them to ensure that the native states of folded proteins are lower in energy than the unfolded configurations. The resulting force field, termed MOFF, was shown to be transferable across folded and disordered proteins in predicting protein sizes (Figure 3). In favorable cases, the model succeeded in folding globular proteins to their native states.

Generalizing the Latham and Zhang strategy to structural ensembles built from all-atom simulations requires additional research. In particular, since the free energies of protein structures from all-atom simulations are unknown, their energy-matching approach cannot be directly applied. One can, in principle, use the relative entropy minimization approach to derive coarse-grained force fields. However, its computational overhead may become too costly, especially when a large list of proteins is included in building structural ensembles.

Acknowledgement

This work was supported by the National Institutes of Health (Grant R35GM133580) and the National Science Foundation (Grant MCB-2042362). A.L. further acknowledges support by the

National Science Foundation Graduate Research Fellowship Program.

Annotated references

- [22] ** The coarse-grained explicit solvent model, SIRAH, allows de novo prediction of secondary structures. It offers impressive performance in predicting tertiary structure of folded proteins and reproducing the radius of gyration for disordered proteins.

- [25] * A new coarse-grained force field tuned for IDPs provides impressive accuracy in predicting to small-angle X-ray scattering profiles.

- [31] ** The authors introduced a novel way to incorporate temperature effect into the coarse-grained force field. These corrections allowed them to differentiate between upper critical temperature and lower critical temperature when studying the liquid-liquid phase separation of IDPs.

- [33] * The authors improved their hydrophobicity scale model by sampling various hydrophobicity measures and fitting parameters, resulting in a new force field that improves predictions of R_g and phase behavior.

- [34] * The authors derived a data-driven hydrophobicity scale and coarse-grained force field for phase-separating proteins using the force balance method.

- [36] * The authors systematically optimized a coarse-grained force field for IDPs to reproduce experimental radius of gyration via a maximum entropy optimization algorithm.

- [39] ** Latham and Zhang introduced a novel optimization algorithm to parameterize a coarse-grained force field that achieved consistent accuracy for both folded and disordered proteins.

- [61] * The authors introduced a deep learning approach to parameterize the coarse-grained force field via a force-matching scheme.

- [63] ** Deep generative models were used to parameterize complex probability distributions of molecular conformations. Such models offer new methodologies for free energy calculations and force field parameterizations.

- [68] ** The authors introduced grid based terms to improve the dihedral angles of the ABSINTH implicit solvation model and force field paradigm. The resulting model, ABSINTH-C, maintains folded structures of ordered proteins and shows improvements in predicting the secondary structure of homopolypeptides.

- [70] * The authors introduced a reweighting strategy that improved the accuracy of the Rosseta software in predicting the structural ensemble of IDPs without compromising its performance for folded proteins.

- [72] * The authors purpose using Bayesian reweighting to balance sources of error, and apply this method to model NMR chemical shifts of an IDP.

References

[1] Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., Jones, D.T.. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. J Mol Biol 2004;337:635–645.

[2] Habchi, J., Tompa, P., Longhi, S., Uversky, V.N.. Introducing protein intrinsic disorder. Chem Rev 2014;114:6561–6588.

[3] Dyson, H.J., Wright, P.E.. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6:197–208.

[4] Oldfield, C.J., Dunker, A.K.. Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. Annu Rev Biochem 2014;83:553–584.

[5] Banani, S.F., Lee, H.O., Hyman, A.A., Rosen, M.K.. Biomolecular condensates: Organizers of cellular biochemistry. Nat Rev Mol Cell Biol 2017;18:285–298.

[6] Shin, Y., Brangwynne, C.P.. Liquid phase condensation in cell physiology and disease. Science 2017;357:eaaf4382.

[7] Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., Sharp, P.A.. A Phase Separation Model for Transcriptional Control. Cell 2017;169:13–23.

[8] Kar, P., Feig, M.. Recent advances in transferable coarse-grained modeling of proteins. In: Adv. Protein Chem. Struct. Biol.; vol. 96; chap. 5. 2014, p. 143–180.

[9] Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A.E., Kolinski, A.. Coarse-Grained Protein Models and Their Applications. Chem Rev 2016;116:7898–7936.

[10] Chong, S.H., Chatterjee, P., Ham, S.. Computer Simulations of Intrinsically Disordered Proteins. Annu Rev Phys Chem 2017;68:117–134.

[11] Levine, Z.A., Shea, J.E.. Simulations of disordered proteins and systems with conformational heterogeneity. Curr Opin Struct Biol 2017;43:95–103.

[12] Ruff, K.M., Pappu, R.V., Holehouse, A.S.. Conformational preferences and phase behavior of intrinsically disordered low complexity sequences: insights from multiscale simulations. Curr Opin Struct Biol 2019;56:1–10.

[13] Dignon, G.L., Best, R.B., Mittal, J.. Biomolecular Phase Separation : From Molecular Driving Forces to Macroscopic Properties. Annu Rev Phys Chem 2020;71:53–75.

[14] Huang, J., MacKerell, A.D.. Force field development and simulations of intrinsically disordered proteins. Curr Opin Struct Biol 2018;48:40–48.

[15] Nerenberg, P.S., Head-Gordon, T.. New developments in force fields for biomolecular simulations. Curr Opin Struct Biol 2018;49:129–138.

[16] Mu, J., Liu, H., Zhang, J., Luo, R., Chen, H.F.. Recent Force Field Strategies for Intrinsically Disordered Proteins. J Chem Inf Model 2021;61:1037–1047.

[17] Tsanai, M., Frederix, P.W.J.M., Schroer, C.F.E., Souza, P.C.T., Marrink, S.J.. Coacervate formation studied by explicit solvent coarse-grain molecular dynamics with the Martini. Chem 2021;doi:10.1039/d1sc00374g.

[18] Benayad, Z., Von Bülow, S., Stelzl, L.S., Hummer, G.. Simulation of FUS Protein Condensates with an Adapted Coarse-Grained Model. J Chem Theory Comput 2021;17:525–537.

[19] Martin, E.W., Thomasen, F.E., Milkovic, N.M., Cuneo, M.J., Grace, C.R., Nourse, A., et al. Interplay of folded domains and the disordered low-complexity domain in mediating hnRNPA1 phase separation. Nucleic Acids Res 2021;49:2931–2945.

[20] Larsen, A.H., Wang, Y., Bottaro, S., Grudinin, S., Arleth, L., Lindorff-Larsen, K.. Combining molecular dynamics simulations with small-angle X-ray and neutron scattering data to study multi-domain proteins in solution. PLoS Comput Biol 2020;16:e1007870.

[21] Machado, M.R., Barrera, E.E., Klein, F., Sónora, M., Silva, S., Pantano, S.. The SIRAH 2.0 Force Field: Altius, Fortius, Citius. J Chem Theory Comput 2019;15:2719–2733.

[22] Klein, F., Barrera, E.E., Pantano, S.. Assessing SIRAH's Capability to Simulate Intrinsically Disordered Proteins and Peptides. J Chem Theory Comput 2021;17:599–604.

[23] Wu, H., Wolynes, P.G., Papoian, G.A.. AWSEM-IDP: A Coarse-Grained Force Field for Intrinsically Disordered Proteins. J Phys Chem B 2018;122:11115–11125.

[24] Davtyan, A., Schafer, N.P., Zheng, W., Clementi, C., Wolynes, P.G., Papoian, G.A.. AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. J Phys Chem B 2012;116:8494–8503.

[25] Baul, U., Chakraborty, D., Mugnai, M.L., Straub, J.E., Thirumalai, D.. Sequence Effects on Size, Shape, and Structural Heterogeneity in Intrinsically Disordered Proteins. J Phys Chem B 2019;123:3462–3474.

[26] Reddy, G., Thirumalai, D.. Dissecting Ubiquitin Folding Using the Self-Organized Polymer Model. J Phys Chem B 2015;119:11358–11370.

[27] Chakraborty, D., Straub, J.E., Thirumalai, D.. Differences in the free energies between the excited states of Aβ40 and Aβ42 monomers encode their aggregation propensities. Proc Natl Acad Sci U S A 2020;117:19926–19937.

[28] Mioduszewski, Ł., Różycki, B., Cieplak, M.. Pseudo-Improper-Dihedral Model for Intrinsically Disordered Proteins. J Chem Theory Comput 2020;16:4726–4733.

[29] Dignon, G.L., Zheng, W., Kim, Y.C., Best, R.B., Mittal, J.. Sequence determinants of protein phase behavior from a coarse-grained model. PLoS Comput Biol 2018;14:e1005941.

[30] Dignon, G.L., Zheng, W., Best, R.B., Kim, Y.C., Mittal, J.. Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. Proc Natl Acad Sci U S A 2018;115:9929–9934.

[31] Dignon, G.L., Zheng, W., Kim, Y.C., Mittal, J.. Temperature-Controlled Liquid–Liquid Phase Separation of Disordered Proteins. ACS Cent Sci 2019;5:821–830.

[32] Das, S., Lin, Y.H., Vernon, R.M., Forman-Kay, J.D., Chan, H.S.. Comparative roles of charge, π, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. Proc Natl Acad Sci U S A 2020;117:28795–28805.

[33] Regy, R.M., Thompson, J., Kim, Y.C., Mittal, J.. Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. Protein Sci 2021;doi:10.1002/pro.4094.

[34] Dannenhoffer-Lafage, T., Best, R.B.. A Data-driven Hydrophobicity Scale for Predicting Liquid-Liquid Phase Separation of Proteins. J Phys Chem B 2021;125:4046–4056.

[35] Tesei, G., Schulze, T.K., Crehuet, R., Lindorff-larsen, K.. Accurate model of liquid-liquid phase behaviour of intrinsically-disordered proteins from data-driven optimization of single-chain properties. bioRxiv 2021;:1–9doi:10.1101/2021.06.23.449550.

[36] Latham, A.P., Zhang, B.. Maximum Entropy Optimized Force Field for Intrinsically Disordered Proteins. J Chem Theory Comput 2020;16:773–781.

[37] Regmi, R., Srinivasan, S., Latham, A.P., Kukshal, V., Cui, W., Zhang, B., et al. Phosphorylation-Dependent Conformations of the Disordered Carboxyl-Terminus Domain in the Epidermal Growth Factor Receptor. J Phys Chem Lett 2020;11:10037–10044.

[38] Zheng, W., Schafer, N.P., Davtyan, A., Papoian, G.A., Wolynes, P.G.. Predictive energy landscapes for protein-protein association. Proc Natl Acad Sci U S A 2012;109(47):19244–19249.

[39] Latham, A.P., Zhang, B.. Consistent Force Field Captures Homologue-Resolved HP1 Phase Separation. J Chem Theory Comput 2021;17:3134–3144.

[40] Robustelli, P., Piana, S., Shaw, D.E.. Developing a molecular dynamics force field for both folded and disordered protein states. Proc Natl Acad Sci U S A 2018;115:E4758–E4766.

[41] Uversky, V.N.. The alphabet of intrinsic disorder: II. Various roles of glutamic acid in ordered and intrinsically disordered proteins. Intrinsically Disord Proteins 2013;1:e24684.

[42] Van Der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., et al. Classification of intrinsically disordered regions and proteins. Chem Rev 2014;114:6589–6631.

[43] Das, R.K., Pappu, R.V.. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proc Natl Acad Sci U S A 2013;110:13392–13397.

[44] Bryngelson, J.D., Onuchic, J.N., Socci, N.D., Wolynes, P.G.. Funnels, pathways, and the energy landscape of protein folding: A synthesis. Proteins 1995;21:167–195.

[45] Shakhnovich, E.. Protein Folding Thermodynamics and Dynamics : Where Physics , Chemistry , and Biology Meet Fundamental Model of Protein Folding. Chem Rev 2006;106:1559–1588.

[46] Dill, K.A., Chan, H.S.. From levinthal to pathways to funnels. Nat Struct Biol 1997;4:10–19.

[47] Onuchic, J.N., Luthey-Schulten, Z., Wolynes, P.G.. THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective. Annu Rev Phys Chem 1997;48:545–600.

[48] Eastwood, M.P., Hardin, C., Luthey-Schulten, Z., Wolynes, P.G.. Statistical mechanical refinement of protein structure prediction schemes: Cumulant expansion approach. J Chem Phys 2002;117:4602–4615.

[49] Mirny, L.A., Shakhnovich, E.I.. How to derive a protein folding potential? A new approach to an old problem. J Mol Biol 1996;264:1164–1179.

[50] Liwo, A., Arłukowicz, P., Czaplewski, C., Ołdziej, S., Pillardy, J., Scheraga, H.A.. A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: Application to the UNRES force field. Proc Natl Acad Sci U S A 2002;99:1937–1942.

[51] Schommers, W.. Pair potentials in disordered many-particle systems: A study for liquid gallium. Phys Rev A 1983;28:3599–3605.

[52] Lyubartsev, A.P., Laaksonen, A.. Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach. Phys Rev E 1995;52:3730–3737.

[53] Noid, W.G.. Perspective: Coarse-grained models for biomolecular systems. J Chem Phys 2013;139:090901.

[54] Ercolesi, F., Adams, J.B.. Interatomic potentials from first-principles calculations: The force-matching method. EPL 1994;26:583–588.

[55] Izvekov, S., Parrinello, M., Burnham, C.J., Voth, G.A.. Effective force fields for condensed phase systems from ab initio molecular dynamics simulation: A new method for force-matching. J Chem Phys 2004;120:10896–10913.

[56] Izvekov, S., Voth, G.A.. A multiscale coarse-graining method for biomolecular systems. J Phys Chem B 2005;109:2469–2473.

[57] Shell, M.S.. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. J Chem Phys 2008;129:144108.

[58] Noé, F., Tkatchenko, A., Müller, K.R., Clementi, C.. Machine Learning for Molecular Simulation. Annu Rev Phys Chem 2020;71:361–390.

[59] Schneider, E., Dai, L., Topper, R.Q., Drechsel-Grau, C., Tuckerman, M.E.. Stochastic Neural Network Approach for Learning High-Dimensional Free Energy Surfaces. Phys Rev Lett 2017;119:150601.

[60] Ding, X., Lin, X., Zhang, B.. Stability and folding pathways of tetra-nucleosome from six-dimensional free energy surface. Nat Commun 2021;12:1091.

[61] Wang, J., Olsson, S., Wehmeyer, C., Pérez, A., Charron, N.E., De Fabritiis, G., et al. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. ACS Cent Sci 2019;5:755–767.

[62] Husic, B.E., Charron, N.E., Lemm, D., Wang, J., Pérez, A., Majewski, M., et al. Coarse graining molecular dynamics with graph neural networks. J Chem Phys 2020;153:194101.

[63] Noé, F., Olsson, S., Köhler, J., Wu, H.. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. Science 2019;365:eaaw1147.

[64] Ding, X., Zhang, B.. Computing Absolute Free Energy with Deep Generative Models. J Phys Chem B

2020;124:10166–10172.

[65] Wirnsberger, P., Ballard, A.J., Papamakarios, G., Abercrombie, S., Racanière, S., Pritzel, A., et al. Targeted free energy estimation via learned mappings. J Chem Phys 2020;153:144112.

[66] Ding, X., Zhang, B.. DeepBAR: A Fast and Exact Method for Binding Free Energy Computation. J Phys Chem Lett 2021;12:2509–2515.

[67] Vitalis, A., Pappu, R.V.. ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. J Comput Chem 2009;30:673–699.

[68] Choi, J.M., Pappu, R.V.. Improvements to the ABSINTH Force Field for Proteins Based on Experimentally Derived Amino Acid Specific Backbone Conformational Statistics. J Chem Theory Comput 2019;15:1367–1382.

[69] Martin, E.W., Holehouse, A.S., Peran, I., Farag, M., Incicco, J.J., Bremer, A., et al. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. Science 2020;367:694–699.

[70] Ferrie, J.J., Petersson, E.J.. A Unified De Novo Approach for Predicting the Structures of Ordered and Disordered Proteins. J Phys Chem B 2020;124:5538–5548.

[71] Bottaro, S., Lindorff-Larsen, K., Best, R.B.. Variational optimization of an all-atom implicit solvent force field to match explicit solvent simulation data. J Chem Theory Comput 2013;9:5641–5652.

[72] Crehuet, R., Buigues, P.J., Salvatella, X., Lindorff-Larsen, K.. Bayesian-Maximum-Entropy Reweighting of IDP Ensembles Based on NMR Chemical Shifts. Entropy 2019;21:898.

[73] Latham, A.P., Zhang, B.. Improving Coarse-Grained Protein Force Fields with Small-Angle X-ray Scattering Data. J Phys Chem B 2019;123:1026–1034.

[74] Xie, W.J.W., Zhang, B.. Learning the Formation Mechanism of Domain-Level Chromatin States with Epigenomics Data. Biophys J 2019;116:2047–2056.

[75] Qi, Y., Reyes, A., Johnstone, S.E.S., Aryee, M.M.J., Bernstein, B.B.E., Zhang, B.. Data-Driven Polymer Model for Mechanistic Exploration of Diploid Genome Organization. Biophys J 2020;119:1905–1916.

[76] Amirkulova, D.B., White, A.D.. Recent advances in maximum entropy biasing techniques for molecular dynamics. Mol Simul 2019;45:1285–1294.

[77] Rangan, R., Bonomi, M., Heller, G.T., Cesari, A., Bussi, G., Vendruscolo, M.. Determination of Structural Ensembles of Proteins: Restraining vs Reweighting. J Chem Theory Comput 2018;14:6632–6641.

[78] Różycki, B., Kim, Y.C., Hummer, G.. SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. Structure 2011;19:109–116.