Hybrid computational methods combining experimental information with Molecular Dynamics

Arup Mondal [a], Stefan Lenz [b], Justin L. MacCallum [b], Alberto Perez [a,*]

[a] *Quantum Theory Project, Department of Chemistry, University of Florida, Leigh Hall, Gainesville, 32611, Florida, USA*
[b] *Department of Chemistry, University of Calgary, 2500 University Drive NW, Calgary, T2N 1N4, AB, Canada*

Abstract

A goal of structural biology is to understand how macromolecules carry out their biological roles by identifying their metastable states, mechanisms of action, pathways leading to conformational changes, and the thermodynamic and kinetic relationships between those states. Integrative modeling brings structural insights into systems where traditional structure determination approaches cannot help. We focus on the synergies and challenges of integrative modeling combining experimental data with molecular dynamics simulations.

1. Introduction

Techniques like x-ray crystallography[1] provide an exquisite picture of the average structure of stable conformations. But challenges remain for many systems including those unsuitable for crystallization, characterizing ensembles, states with low populations, and the transition pathways between these states. Fortunately, many experimental techniques such as NMR, FRET, DEER, PREs, chemical crosslinking, and others offer indirect structural information that can be obtained even in challenging situations[2, 3].

When enough information is available (data-rich regime), the system is highly constrained, and the data typically defines a narrow uncertainty ensemble[4], which is in some cases deposited in the protein data bank (PDB)[5]. Choices such as the computational method and how the data is modeled can affect this ensemble. Increasingly, however, researchers operate in the data-poor regime, where the system is only weakly constrained so that the data alone is insufficient to fully define the structural ensemble. By *integrating* data from multiple experiments, it is possible to narrow down the possible structures[6]. Even in such scenarios, distinguishing between multiple models requires a *hybrid*

[⋆] Corresponding author
[*] Email address: perez@ufl.edu

approach involving computational software to explore conformational space and identify possible solutions compatible with the data.

Recently, the wwPDB-dev[7] was developed to accept models originating from integrative/hybrid approaches. Whereas the PDB contains over 200,000 structures, the wwPDB-dev holds just 112 entries as of January 2023. The stark difference arises from difficulty in identifying pipelines to model these challenging systems and the need for better assessment tools to validate the quality of the models[8]. This review focuses on integrative/hybrid modeling in which Molecular Dynamics (MD) techniques are used to generate models. Excellent reviews focus on other aspects of integrative/hybrid modeling[9, 6, 10•, 1, 11, 12].

MD is seeing a resurgence[13] thanks to improved force fields [14] and enhanced sampling techniques [15] leading to excellent agreement between experiments and simulation techniques [3]. Efforts to synergize with new sources of experimental data provide optimism for their role in integrative approaches[6, 12, 1]. However, as each experimental technique provides different outputs and uncertainties, there is no unique recipe for hybrid modelling[8, 16]. For instance, density maps can be used to restrain residue/atomic positions[17, 9], many techniques provide distance or orientation information between parts of the molecule[18, 19], and others provide average or ensemble information[20]. MD samples ensembles of conformations that, when processed using statistical mechanics principles, provide insights about the relevant states of the system. Experimental data can be used to bias the ensemble, affecting the probability of sampling different regions in the energy landscape. The resulting ensembles can be processed to learn about states compatible with the data and physics model. Alternatively, data can be used with unbiased MD ensembles to select a subset of structures compatible with the experimental information. The challenges remain similar to other integrative pipelines: determining the number of states represented by the data, how to model uncertainties, or how to validate the ensembles

2. Sampling strategies

The large phase space available to biomacromolecules challenges the identification of important regions (states). Multiple approaches such as quasi-static minimization, random tree forest, or normal mode analysis to name a few have been used to sample phase space. What makes techniques such as MD and Monte Carlo (MC) unique is the foundation of these sampling strategies on physical principles, such as detailed balance, that relates the explored ensembles to the relative importance of different regions of phase space through statistical mechanics. It is this physical foundation that allows us to determine thermodynamic (e.g., free energies) and kinetic (e.g., binding rates) properties. The more accurate the underlying representation of the potential energy landscape (given by the force field), the better agreement MD/MC-based sampling strategies will have with experiments.

In recent years, force field development efforts have increased, leading to a variety of force fields suited for different systems [21]. The type of experimental data as well as the initial structural knowledge determine the type of MD sampling strategy to use[15]. For example, describing the ensemble of conformations available for a known folded structure that is compatible with FRET data might not require access to all configuration space—only appropriate fluctuations around the starting configuration need to be sampled. On the other hand, processes like binding, folding, or describing IDP ensembles will require extensive coverage of configuration space. In *informed sampling* approaches, the data is directly used to bias sampling, leading to ensembles that satisfy the data; whereas in *uninformed sampling*, the data is used in the post-analysis stage to either select structures or reweight the ensemble to obtain agreement with experimental measurements (see Figure 1).

2.1. Informed search

The general premise in informed search is that we know some properties about the end-states and incorporate the knowledge as restraints to guide the system. Often this data is evaluated on the fly on the structure that is being sampled by using a *forward model*[22, 23, 24, 25]. However, experimental data represents an ensemble average, thus, imposing restraints on a single structure could inappropriately bias the results. Alternatively, some groups use time-averaged restraints or ensemble average restraints[26] to represent the ensemble nature of the data better[27]. For instance, XL-MS data might contain an ensemble where a residue might be involved in a particular cross-link in one conformation, and in an alternative cross-link in another conformation [28]. As a second example, the mapping between single structures and NMR observables (e.g. NOESY data) leads to well-known errors. Lindorff-Larsen and his group showed that when multiple states give rise to NOE observables, a consensus structure that agrees with the data might not capture the complexity of the system[29•].

In recent years, this type of informed search has been framed in terms of Bayesian inference to account for the uncertainties arising from the force field, data, and forward model used[8, 30, 31, 32]. Information-driven sampling can thus overcome limitations in the force field, and supplement the experimental data with a physics-based prior to identify structures and states that either method alone could not[23, 33].

2.2. Uninformed search

These approaches yield the prior distribution based on the force field redand work best when starting from a native-like structure. Enhanced sampling methods such as generalized ensemble, weighted ensemble, adaptive sampling, or accelerated MD are also commonly used for higher efficiency than conventional MD[15]. The goal is to promote a broad exploration of the energy landscape. The prior should reflect the Boltzmann distribution based on the force field used – hence, removing any potential bias used in enhanced sampling simulations is important. Failure to do so leads to poor priors.

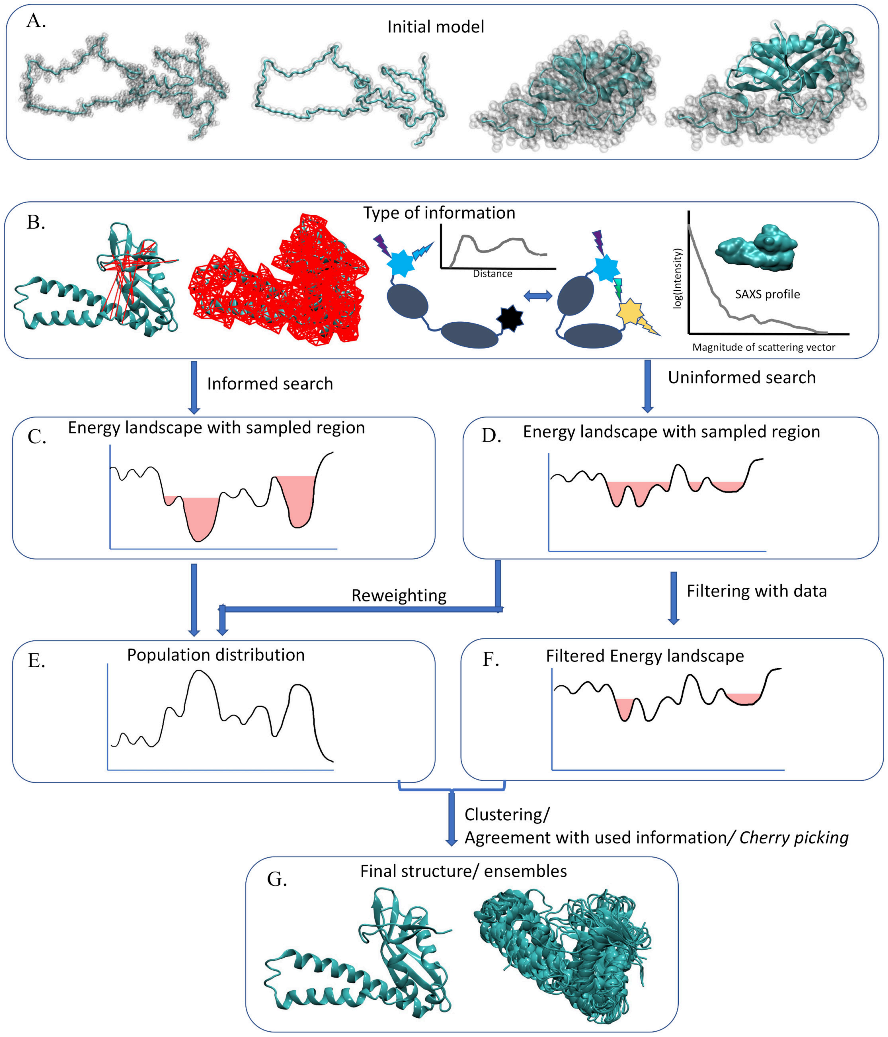
A.
Initial model
B.
Type of information
Distance
log(Intensity)
SAXS profile
Magnitude of scattering vector
Informed search
Uninformed search
C.
Energy landscape with sampled region
D.
Energy landscape with sampled region
Reweighting
Filtering with data
E.
Population distribution
F.
Filtered Energy landscape
Clustering/
Agreement with used information/ Cherry picking
G.
Final structure/ ensembles

Figure 1: **Schematic representation of integrative modelling**. A. Initial models might represent extended or native conformations in an all-atom or coarse representation. B. Schematic of different types of information used to restrain simulations such as pair-wise restraints or density maps. C. *Informed search* introduces energy penalties in regions incompatible with data, focusing exploration on a few energy basins (filled in red). In contrast, *Uninformed search* (D) does not prioritize specific regions of the landscape. E. Analysis of the population distribution of either the prior (*uninformed search)*) or posterior (*informed search*) yields the relevant states. F. Alternatively, the distributions can be filtered in accordance to their agreement with experimental data. G. Final representative structure(s) or ensembles that satisfy the data.

3. Choosing Representative Models from MD ensembles

Traditionally, similarity measures and dimensionality reduction techniques followed by clustering provide information about the relevant states in the system and their relative importance. More recently, Markov State Models (MSM)[34, 35] are becoming increasingly popular to provide kinetic information on the relationships between states. The analysis is straightforward when the ensembles are unbiased and requires prior reweighting otherwise[36].

The availability of data can then help choose the most compatible structures from the ensemble. For instance, data is used to reweight ensembles and choose representative structures or distribution of structures[37, 20]. The most popular approaches use maximum entropy, maximum parsimony, or Bayesian inference principles[38, 16, 39•, 40, 41]. In maxEnt approaches, the relative entropy with respect to the original ensemble (typically measured as a Kullback-Leibler Divergence, see eq. 1) is minimized. Thus, maintaining the original weights from the simulation will have the minimum cross-entropy – while selecting a single structure ($cherry - picking$)[42] will have the maximum relative entropy. The latter was preferred when limited sampling and force field inaccuracies were prevalent. However, with improved force field and sampling, there has been a shift towards ensemble-based methods to represent the system.

$$S(p_1|p_0) = \sum_i^N p_1(x_i) \ln \frac{p_1(x_i)}{p_0(x_i)} \tag{1}$$

Having a good overlap between the model and experimental evidence for the relevant regions of configuration is key to successful reweighting. A recent work describing IDP ensembles finds that the quality of the force field (prior distribution) is more important than the re-weighting method[43]. In other words, reweighting techniques can never recover information about a region that has not been sampled. These approaches are also useful to study conformational transitions through a two-step process which first involves the construction of an MSM, followed by a *refinement* with FRET data through an unsupervised learning model[44].

4. Challenges in hybrid modeling

4.1. Number of states to represent the system

Proteins are dynamic molecules, accessing multiple states that explain structure/function relationships. While the information in the PDB typically yields information about a single state (either as a single structure or narrow ensemble), there are multiple computational tools and experimental approaches that can then use the initial structure to identify other relevant states that explain function. A common challenge in integrative approaches is knowing if the measured experimental observables originating from different experiments represent the same state, or if the observed signal arises from multiple states. When signals between different states overlap, advanced methods such as those based

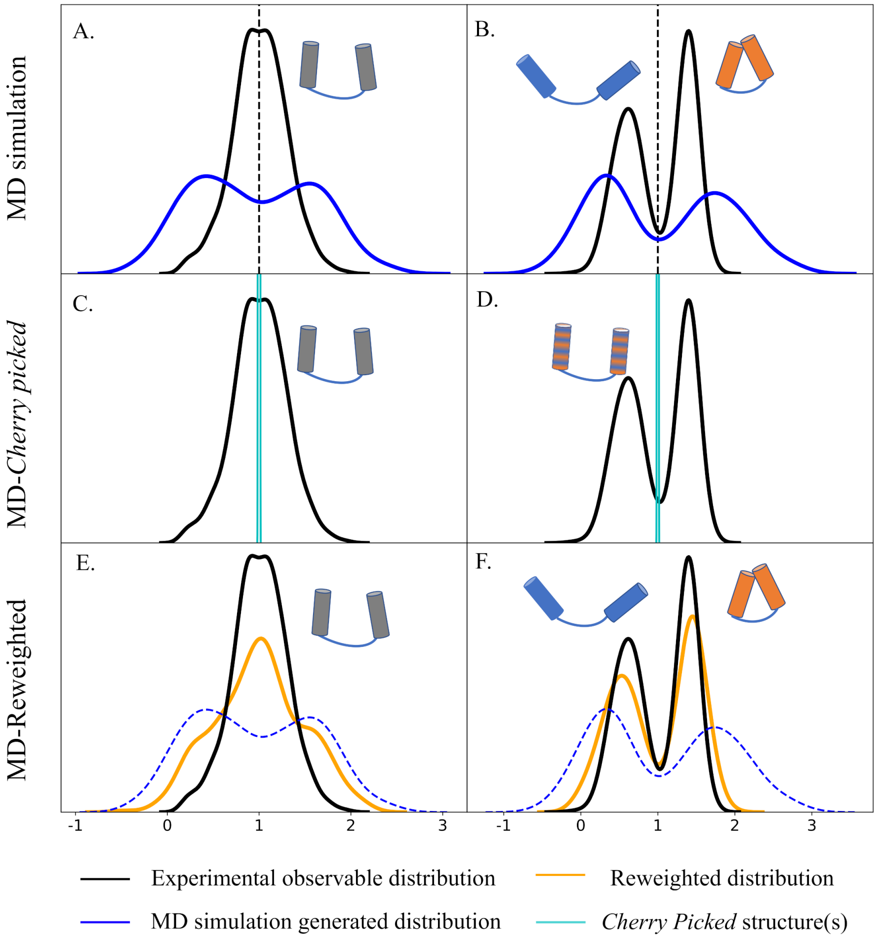

Figure 2: **Scheme for selecting structure(s)** based on distributions of an arbitrary property coming from experiments (black line) or simulations (blue line). The left panel represents an scenario where the experimental data has a unimodal distribution around the average observable (A). In this scenario, choosing a cherry-picked single structure (cyan) (C) or reweighting the MD ensemble (orange) to identify the highest population cluster will provide similar results. On the right panel, the system has two conformations leading to a bimodal distribution of the experimental observable (B). The cherry-picked average structure from the MD distribution is a bad representation of the system (D). However, reweighting the MD ensemble based on the experimental data correctly represents the underlying experimental distribution.”

on maximum entropy are needed to identify the different states present in the distribution[45$^{\bullet\bullet}$, 46].

MD approaches, on the other hand, typically find hundreds or thousands of states through clustering of their ensembles – further classification into metastable states identifies a smaller set of states to compare with experiments[47, 48, 49]. When this process of classification benefits from experimental data, it can further help in identifying multiple biological states[50, 51]. In this context, it is hard to ignore the progress from machine learning methods such as AlphaFold2[52] can also generate thousands of diverse structures of proteins by making clever use of sub-sampled or clustered multiple sequence alignments[53]. As with MD, agreement with experimental data can be used to choose high-accuracy structures representing different states from these ensembles. However, these ensembles lack the physics to relate thermodynamic/kinetic properties from the states, can only be used for systems trained on large databases (e.g, the PDB), include implicit biases based on databases (e.g., holo vs apo structures), and are not yet sensitive to small perturbations such as single point mutations.

4.2. Uncertainty in the data

It remains challenging to separate errors in the physical models (force fields), experiments, and how the data is modeled (e.g., forward models)[16, 8]. Consequently, the error bars and uncertainties reported by different methods and groups do not necessarily reflect the same information [45$^{\bullet\bullet}$] leading to significant discrepancies in the reported uncertainties and lack of robustness in predictions across methods[45$^{\bullet\bullet}$]. A blind study for modeling dynamical systems using smFRET data assessed the performance of 19 participants. Participants arrived at similar results, separating the experimental signal originating from conformational dynamics and other sources that increase ambiguity. The study helped to define standards for error analysis and propagation in the field[54]. A similar community-wide assessment of the reproducibility of SAXS and SANS data[55] for a set of proteins emphasizes current limitations in the field such as accurate solvent subtraction.

4.3. The need for reference standards

Hybrid modeling is an umbrella term for many possible combinations of computational methods and sources of experimental data. Clear protocols and standards exist for determining structures using data-rich regime techniques, such as X-ray crystallography, cryo-EM, or solution NMR. In contrast, the number of tools and, therefore, the number of ways to perform hybrid modeling has increased dramatically. Addressing some of the current limitations will require coordination between scientists with different expertise[56] in order to: 1) produce homogeneous and transferable protocols across research labs (experimental and computational), 2) increase reproducibility across integrative platforms given the same initial data, 3) reduce human intervention in modeling data, and 4) provide the protocols of how the data was modeled along with the original data and output structures. Blind studies have revealed conceptual oversights[45••, 18] due to misinterpretation of what the data represents. These oversights often originate from the lack of standardized protocols, shared expertise, and lack of communication across multiple labs.

The different PDB Task Forces (https://www.wwpdb.org/task/) play a fundamental role in developing and enforcing standards for the community. For example, the NMR exchange format initiative (NEF) [57, 58] strives to create a self-describing format that integrative modeling programs can learn to read to model structures and write to incorporate information on how the data was modeled. The Small Angle Scattering Task Force aims to test and benchmark different methods that use SAXS profiles to model biomolecules. To further these efforts, the experimental data and the associated structural models are deposited in the SAS Biological DataBase (SASBDB, www.sasbdb.org)[59]. More initiatives and widespread adoption of data deposition in software-agnostic formats are needed for other sources of data, including smFRET [60]. Task forces further promote the deposition of centralized databases[59, 61, 62, 63], and models that contain information and unified protocols to promote reproducible and transferable open science. These centralized databases are still limited in some areas like mass spectroscopy[42], where increased awareness will help modelers develop better integrative tools.

As modelers develop software to integrate different types of data, there is a lack of compatibility across tools. Ideally, a method developed by one lab for integrating data from experiment A and a method from another lab for experiment B could be used in a *plug-and-play* style. In practice, different formats and other incompatibilities prevent combining such pipelines into a single workflow[64••].

4.4. How is an ensemble validated?

One major question yet to be resolved is how to validate an ensemble prediction appropriately. Most methods measure ensemble average properties of some type, but many different distributions can produce the same average. The well-known Anscombe quartet[65] illustrates different distributions that have identical descriptive statistics (mean of both variables, variance of both variables, correlation between x and y, linear regression and R^2 between x and y). It is not possible to use the goodness of fit to assess which distribution is correct, as all four distributions agree with the available statistics equally well.

This phenomenon is widespread in integrative and hybrid modeling, where there can be many—typically infinitely many!—different ensembles that are in equally good agreement with the available data. Choosing some ensembles as more correct than others typically boils down to using physical models, and regularization principles like maximum entropy or maximum parsimony [30] as naively forcing a fit to the average quantity can often time lead to incorrect inferences (see Figure 2).

A major challenge facing the field is comparing the outputs of different tools. If two methods predict different ensembles and both ensembles are in equal agreement with the data, how do we say which one is more correct than another? Certain modeling tools may make different assumptions with data, and understanding how that influences an ensemble remains challenging. On a similar note, the noise associated with different data types should be considered when validating an ensemble. All of this must be communicated to users of these tools and consumers of these models in a transparent and accessible way.

Finally, there are few or insufficient databases to view ensembles produced by hybrid or integrative modeling. Generally, the PDB database is not an appropriate venue for these models, and PDBdev lacks sufficient validation tests for the ensembles (like proCheck for X-ray crystal structures) [10•, 66]. In recent years, the number of deposited models and experimental data in the SASBDB has increased rapidly, making it useful for validating models. However, this database is dedicated only to SAXS/SANS data, highlighting the need for similar databases to other types of experimental data. For IDPs even these might not be good enough, requiring databases that consider ensembles[67].

5. Emerging applications of integrative modeling

5.1. Integrative/Hybrid approaches will provide insights into cellular assemblies

As new sources of experimental data and their uncertainties become more prevalent, we see a wide range of spatiotemporal scales that dictates the need for atomistic, Coarse-grained (CG), or ultra-CG approaches[68, 69•, 70]. For instance, Hi-C restraint data is used to model genome assemblies [71, 72, 73]. With growing interest in these areas, new computational methods arise for developing and assessing benchmark sets [74]. Ultimately, the goal is to assemble large teams with overlapping expertise and a unified protocol to tackle the structural biology of whole cells (metamodeling)[75, 76, 77].

5.2. Advances in describing intrinsically disordered systems

Intrinsically disordered proteins (IDP) or proteins featuring intrinsically disordered regions (IDR) make up approximately 20% of eukaryotic proteomes [11]. IDPs are particularly difficult to model accurately, with force fields failing to reproduce the radius of gyration for completely disordered proteins. Most of the force fields fail to reproduce several experimental observables simultaneously[78]. Efforts to improve force field descriptions of IDPs is an ongoing area of research; however, improved descriptions of disordered proteins often come at the expense of modeling ordered proteins[79].

Force fields are often most effective at describing the local structure of ordered proteins. Hybrid approaches utilizing long-distance-based restraints offer an avenue to overcome the limitations of force fields (e.g., compact structures). For example, combining replica exchange discrete molecular dynamics (DMD) with restraints derived from FRET and DEER experiments [80] overcomes the accuracy limitations of DMD and samples helix-coil transitions in SNAP-25, an IDP. Similarly, using NMR-based Bayesian reweighting [81], SAXS-based ensemble optimization methods[82], or combining multiple experiments such as

NMR, SAXS, and single molecule FRET[83•] capture the ensembles of IDP conformations. A recent review discusses approaches to studying IDP conformational ensembles in detail[84].

6. Our current opinion: the ongoing efforts to standardize experimental outputs and computational pipelines will lead to a flourishing integrative structural biology community

Hybrid modeling holds the promise to solve problems in structural biology that other methods alone could not. With increased standardization, we expect better modeling tools and more efficient workflows that will lead to a dramatic increase in wwPDB-dev depositions over the next decade.

7. Conflict of interest statement

The authors declare no conflict of interest.

8. Acknowledgements

The authors thank the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada Foundation for Innovation for their support, and the National Science Foundation. SL thanks the Canada First Research Excellence Fund for his postdoctoral fellowship. JLM thanks the Canada Research Chairs program. AP thanks support from the NSF CAREER award CHE-2235785.

References

[1] Seffernick, J.T. and Lindert, S. (2020). Hybrid methods for combined experimental and computational determination of protein structure. The Journal of Chemical Physics *153*, 240901.

[2] Ziegler, S.J., Mallinson, S.J., John, P.C.S., and Bomble, Y.J. (2021). Advances in integrative structural biology: Towards understanding protein complexes in their cellular context. Computational and Structural Biotechnology Journal *19*, 214–225.

[3] Braitbard, M., Schneidman-Duhovny, D., and Kalisman, N. (2019). Integrative Structure Modeling: Overview and Assessment. Annual Review of Biochemistry , 113–35.

[4] Bonomi, M., Heller, G.T., Camilloni, C., and Vendruscolo, M. (2017). Principles of protein structural ensemble determination. Current Opinion in Structural Biology *42*, 106–116.

[5] Rose, P.W., Prlić, A., Altunkaya, A., Bi, C., Bradley, A.R., Christie, C.H., Costanzo, L.D., Duarte, J.M., Dutta, S., Feng, Z., et al. (2017). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Research *45*, D271–D281.

[6] Ziemianowicz, D.S. and Kosinski, J. (2022). New opportunities in integrative structural modeling. Current Opinion in Structural Biology *77*, 102488.

[7] Burley, S.K., Kurisu, G., Markley, J.L., Nakamura, H., Velankar, S., Berman, H.M., Sali, A., Schwede, T., and Trewhella, J. (2017). PDB-Dev: a Prototype System for Depositing Integrative/Hybrid Structural Models. Structure *25*, 1317–1318.

[8] Schneidman-Duhovny, D., Pellarin, R., and Sali, A. (2014). Uncertainty in integrative structural modeling. Current Opinion in Structural Biology *28*, 96–104.

[9] Bonomi, M. and Vendruscolo, M. (2019). Determination of protein structural ensembles using cryo-electron microscopy. Current Opinion in Structural Biology *56*, 37–45.

[10•] Sali, A. (2021). From integrative structural biology to cell biology. Journal of Biological Chemistry *296*, 100743.

> •Reviews the different aspects of integrative modelling including software and challenges. Shows the progress from the structure determination of single proteins to the aspiration of mapping whole cells through integrative approaches.

[11] Jeschke, G. (2022). Integration of Nanometer-Range Label-to-Label Distances and Their Distributions into Modelling Approaches. Biomolecules *12*, 1369.

[12] Srivastava, A., Tiwari, S.P., Miyashita, O., and Tama, F. (2020). Integrative/Hybrid Modeling Approaches for Studying Biomolecules. Journal of Molecular Biology *432*, 2846–2860.

[13] Schlick, T., Portillo-Ledesma, S., Myers, C.G., Beljak, L., Chen, J., Dakhel, S., Darling, D., Ghosh, S., Hall, J., Jan, M., et al. (2021). Biomolecular Modeling and Simulation: A Prospering Multidisciplinary Field. Annual Review of Biophysics *50*, 1–35.

[14] Nerenberg, P.S. and Head-Gordon, T. (2018). New developments in force fields for biomolecular simulations. Current Opinion in Structural Biology *49*, 129–138.

[15] Henin, J., Lelievre, T., Shirts, M.R., Valsson, O., and Delemotte, L. (2022). Enhanced Sampling Methods for Molecular Dynamics Simulations [Article v1.0]. Living Journal of Computational Molecular Science *4*.

[16] Orioli, S., Larsen, A.H., Bottaro, S., and Lindorff-Larsen, K. (2020). How to learn from inconsistencies: Integrating molecular simulations with experimental data. Progress in Molecular Biology and Translational Science *170*, 123–176.

[17] Shekhar, M., Terashi, G., Gupta, C., Sarkar, D., Debussche, G., Sisco, N.J., Nguyen, J., Mondal, A., Vant, J., Fromme, P., et al. (2021). CryoFold: Determining protein structures and data-guided ensembles from cryo-EM density maps. Matter *4*, 3195–3216.

[18] Robertson, J.C., Nassar, R., Liu, C., Brini, E., Dill, K., and Perez, A. (2019). NMR-assisted protein structure prediction with MELDxMD. Proteins *36*, D402.

[19] Fajardo, J.E., Shrestha, R., Gil, N., Belsom, A., Crivelli, S.N., Czaplewski, C., Fidelis, K., Grudinin, S., Karasikov, M., Karczyńska, A.S., et al. (2019). Assessment of chemical-crosslink-assisted protein structure modeling in CASP13. Proteins: Structure, Function, and Bioinformatics *87*, 1283–1297.

[20] Larsen, A.H., Wang, Y., Bottaro, S., Grudinin, S., Arleth, L., and 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 Computational Biology *16*, e1007870.

[21] Mondal, A., Chang, L., and Perez, A. (2022). Modelling peptide–protein complexes: docking, simulations and machine learning. QRB Discovery *3*.

[22] Lubecka, E.A. and Liwo, A. (2022). A coarse-grained approach to NMR-data-assisted modeling of protein structures. Journal of Computational Chemistry *43*, 2047–2059.

[23] Mondal, A., Swapna, G., Hao, J., Ma, L., Roth, M.J., Montelione, G.T., and Perez, A. (2022). Structure determination of protein-peptide complexes from NMR chemical shift data using MELD. bioRxiv , 2021.12.31.474671.

[24] Sala, D., Huang, Y.J., Cole, C.A., Snyder, D.A., Liu, G., Ishida, Y., Swapna, G., Brock, K.P., Sander, C., Fidelis, K., et al. (2019). Protein structure prediction assisted with sparse NMR data in CASP13. Proteins: Structure, Function, and Bioinformatics *87*, 1315–1332.

[25] Mondal, A. and Perez, A. (2021). Simultaneous Assignment and Structure Determination of Proteins From Sparsely Labeled NMR Datasets. Frontiers in Molecular Biosciences *8*, 774394.

[26] Czaplewski, C., Gong, Z., Lubecka, E.A., Xue, K., Tang, C., and Liwo, A. (2021). Recent Developments in Data-Assisted Modeling of Flexible Proteins. Frontiers in Molecular Biosciences *8*, 765562.

[27] Bonomi, M., Camilloni, C., Cavalli, A., and Vendruscolo, M. (2016). Metainference: A Bayesian inference method for heterogeneous systems. Science Advances *2*, e1501177.

[28] Piersimoni, L., Kastritis, P.L., Arlt, C., and Sinz, A. (2022). Cross-Linking Mass Spectrometry for Investigating Protein Conformations and Protein–Protein InteractionsA Method for All Seasons. Chemical Reviews *122*, 7500–7531.

[29•] Bottaro, S., Nichols, P.J., Vögeli, B., Parrinello, M., and Lindorff-Larsen, K. (2020). Integrating NMR and simulations reveals motions in the UUCG tetraloop. Nucleic Acids Research *48*, gkaa399–.

•Elegant integration of simulations and NMR data ti identify two states that give rise to the experimental observables.

[30] Gaalswyk, K., Muniyat, M.I., and MacCallum, J.L. (2018). The emerging role of physical modeling in the future of structure determination. Current Opinion in Structural Biology *49*, 145–153.

[31] Rieping, W., Habeck, M., and Nilges, M. (2005). Inferential structure determination. Science *309*, 303–306.

[32] Habeck, M., Nilges, M., and Rieping, W. (2005). Bayesian inference applied to macromolecular structure determination. Physical Review E *72*, 031912.

[33] Perez, A., Morrone, J.A., Brini, E., MacCallum, J.L., and Dill, K.A. (2016-11). Blind protein structure prediction using accelerated free-energy simulations. Science Advances *2*, e1601274.

[34] Noé, F., Schütte, C., Vanden-Eijnden, E., Reich, L., and Weikl, T.R. (2009-11). Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. Proceedings of the National Academy of Sciences of the United States of America *106*, 19011–19016.

[35] Pande, V.S., Beauchamp, K., and Bowman, G.R. (2010-9). Everything you wanted to know about Markov State Models but were afraid to ask. Methods *52*, 99–105.

[36] Shirts, M.R. and Chodera, J.D. (2008). Statistically optimal analysis of samples from multiple equilibrium states. Journal of Chemical Physics *129*.

[37] Yagi, K., Re, S., Mori, T., and Sugita, Y. (2022). Weight average approaches for predicting dynamical properties of biomolecules. Current Opinion in Structural Biology *72*, 88–94.

[38] Ge, Y. and Voelz, V.A. (2018). Model Selection Using BICePs: A Bayesian Approach for Force Field Validation and Parameterization. The Journal of Physical Chemistry B *122*, 5610–5622.

[39•] Bottaro, S., Bengtsen, T., and Lindorff-Larsen, K. (2020). Integrating Molecular Simulation and Experimental Data: A Bayesian/Maximum Entropy reweighting approach. Methods Mol Biol *2112*, 219–240.

> •Here the authours describe Bayesian/Maximum Entropy approach to reweight the initial ensemble from simulation with experimental data to generated new ensembles that matches better with experimental observables.

[40] Hummer, G. and Köfinger, J. (2015). Bayesian ensemble refinement by replica simulations and reweighting. The Journal of Chemical Physics *143*, 243150.

[41] Różycki, B., Kim, Y.C., and Hummer, G. (2011). SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. Structure *19*, 109–116.

[42] Biehn, S.E. and Lindert, S. (2022). Protein Structure Prediction with Mass Spectrometry Data. Annual Review of Physical Chemistry .

[43] Gomes, G.N.W., Namini, A., and Gradinaru, C.C. (2022). Integrative Conformational Ensembles of Sic1 Using Different Initial Pools and Optimization Methods. Frontiers in Molecular Biosciences *9*, 910956.

[44] Matsunaga, Y. and Sugita, Y. (2018). Linking time-series of single-molecule experiments with molecular dynamics simulations by machine learning. eLife *7*.

[45••] Götz, M., Barth, A., Bohr, S.S.R., Börner, R., Chen, J., Cordes, T., Erie, D.A., Gebhardt, C., Hadzic, M.C.A.S., Hamilton, G.L., et al. (2022). A blind benchmark of analysis tools to infer kinetic rate constants from single-molecule FRET trajectories. Nature Communications *13*, 5402.

> ••This blind study using smFRET data assesses the performance of 19 groups in modeling the states and kinetic properties for a set of systems. When the data is well separated, groups correctly identify all states and infer correct kinetics. Despite improvements, some failures remain when the data presents overlapping signals between states.

[46] Saurabh, A., Fazel, M., Safar, M., Sgouralis, I., and Pressé, S. (2023). Single-photon smFRET. I: Theory and conceptual basis. Biophysical Reports *3*, 100089.

[47] Sengupta, U., Carballo-Pacheco, M., and Strodel, B. (2019). Automated Markov state models for molecular dynamics simulations of aggregation and self-assembly. The Journal of Chemical Physics *150*, 115101.

[48] Chang, L. and Perez, A. (2022). Deciphering the Folding Mechanism of Proteins G and L and Their Mutants. Journal of the American Chemical Society *144*, 14668–14677.

[49] Copperman, J. and Zuckerman, D.M. (2020). Accelerated Estimation of Long-Timescale Kinetics from Weighted Ensemble Simulation via Non-Markovian “Microbin” Analysis. Journal of Chemical Theory and Computation *16*, 6763–6775.

[50] Hamilton, G.L., Saikia, N., Basak, S., Welcome, F.S., Wu, F., Kubiak, J., Zhang, C., Hao, Y., Seidel, C.A., Ding, F., et al. (2022). Fuzzy supertertiary interactions within PSD-95 enable ligand binding. eLife *11*, e77242.

[51] Dawson, J.E., Smith, I.N., Martin, W., Khan, K., Cheng, F., and Eng, C. (2022). Shape shifting: The multiple conformational substates of the PTEN N-terminal PIP2-binding domain. Protein Science *31*, e4308.

[52] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021-7). Highly accurate protein structure prediction with AlphaFold. Nature .

[53] Stein, R.A. and Mchaourab, H.S. (2022). SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2. PLoS Computational Biology *18*, e1010483.

[54] Agam, G., Gebhardt, C., Popara, M., Mächtel, R., Folz, J., Ambrose, B., Chamachi, N., Chung, S.Y., Craggs, T.D., Boer, M.d., et al. (2022). Reliability and accuracy of single-molecule FRET studies for characterization of structural dynamics and distances in proteins. bioRxiv , 2022.08.03.502619.

[55] Trewhella, J., Vachette, P., Bierma, J., Blanchet, C., Brookes, E., Chakravarthy, S., Chatzimagas, L., Cleveland, T.E., Cowieson, N., Crossett, B., et al. (2022). A round-robin approach provides a detailed assessment of biomolecular small-angle scattering data reproducibility and yields consensus curves for benchmarking. Acta Crystallographica Section D *78*, 1315–1336.

[56] Hamilton, G.L., Alper, J., and Sanabria, H. (2020). Reporting on the future of integrative structural biology ORAU workshop. Front Biosci (Landmark Ed) , 43–68.

[57] Gutmanas, A., Adams, P.D., Bardiaux, B., Berman, H.M., Case, D.A., Fogh, R.H., Güntert, P., Hendrickx, P.M.S., Herrmann, T., Kleywegt, G.J., et al. (2015-06). NMR Exchange Format: a unified and open standard for representation of NMR restraint data. Nature structural & molecular biology *22*, 433 – 434.

[58] Berman, H.M. (2021). Synergies between the Protein Data Bank and the community. Nature Structural & Molecular Biology *28*, 400–401.

[59] Valentini, E., Kikhney, A.G., Previtali, G., Jeffries, C.M., and Svergun, D.I. (2015). SASBDB, a repository for biological small-angle scattering data. Nucleic Acids Research *43*, D357–D363.

[60] Lerner, E., Barth, A., Hendrix, J., Ambrose, B., Birkedal, V., Blanchard, S.C., Börner, R., Chung, H.S., Cordes, T., Craggs, T.D., et al. (2021). FRET-based dynamic structural biology: Challenges, perspectives and an appeal for open-science practices. eLife *10*, e60416.

[61] Burley, S.K., Kurisu, G., Markley, J.L., Nakamura, H., Velankar, S., Berman, H.M., Sali, A., Schwede, T., and Trewhella, J. (2017). PDB-Dev: a Prototype System for Depositing Integrative/Hybrid Structural Models. Structure *25*, 1317–1318.

[62] Vallat, B., Webb, B., Fayazi, M., Voinea, S., Tangmunarunkit, H., Ganesan, S.J., Lawson, C.L., Westbrook, J.D., Kesselman, C., Sali, A., et al. (2021). New system for archiving integrative structures. Acta Crystallographica Section D *77*, 1486–1496.

[63] Sali, A., Berman, H., Schwede, T., Trewhella, J., Kleywegt, G., Burley, S., Markley, J., Nakamura, H., Adams, P., Bonvin, A., et al. (2015). Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. Structure *23*, 1156–1167.

[64••] Hancock, M., Peulen, T.O., Webb, B., Poon, B., Fraser, J.S., Adams, P., and Sali, A. (2022). Integration of software tools for integrative modeling of biomolecular systems. Journal of Structural Biology *214*, 107841.

> ••Eloquent description of the problems facing standardization, cost of software interoperability, and cost collaboration towards more efficient integrative modeling workflows.

[65] Anscombe, F.J. (1973). Graphs in Statistical Analysis. The American Statistician *27*, 17–21.

[66] Vallat, B., Webb, B., Westbrook, J.D., Sali, A., and Berman, H.M. (2018). Development of a Prototype System for Archiving Integrative/Hybrid Structure Models of Biological Macromolecules. Structure *26*, 894–904.e2.

[67] Lazar, T., Martínez-Pérez, E., Quaglia, F., Hatos, A., Chemes, L.B., Iserte, J.A., Méndez, N.A., Garrone, N.A., Saldaño, T.E., Marchetti, J., et al. (2020). PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. Nucleic Acids Research *49*, D404–D411.

[68] Roel-Touris, J., Don, C.G., Honorato, R.V., Rodrigues, J.P.G.L.M., and Bonvin, A.M.J.J. (2019). Less Is More: Coarse-Grained Integrative Modeling of Large Biomolecular Assemblies with HADDOCK. Journal of Chemical Theory and Computation *15*, 6358–6367.

[69•] Roel-Touris, J. and Bonvin, A.M. (2020). Coarse-grained (hybrid) integrative modeling of biomolecular interactions. Computational and Structural Biotechnology Journal *18*, 1182–1190.

•Reviews the software that allow to use coarse grain model and discusses few complxes that are solved by incorporating experimental data

[70] Chen, Y.L. and Habeck, M. (2017). Data-driven coarse graining of large biomolecular structures. PLoS ONE *12*, e0183057.

[71] Buitrago, D., Labrador, M., Arcon, J.P., Lema, R., Flores, O., Esteve-Codina, A., Blanc, J., Villegas, N., Bellido, D., Gut, M., et al. (2021). Impact of DNA methylation on 3D genome structure. Nature Communications *12*, 3243.

[72] Cheng, R.R., Contessoto, V.G., Aiden, E.L., Wolynes, P.G., Pierro, M.D., and Onuchic, J.N. (2020). Exploring chromosomal structural heterogeneity across multiple cell lines. eLife *9*, e60312.

[73] Shinkai, S., Nakagawa, M., Sugawara, T., Togashi, Y., Ochiai, H., Nakato, R., Taniguchi, Y., and Onami, S. (2020). PHi-C: deciphering Hi-C data into polymer dynamics. NAR Genomics and Bioinformatics *2*, lqaa020.

[74] Zheng, Y. and Keleş, S. (2020). FreeHi-C simulates high-fidelity Hi-C data for benchmarking and data augmentation. Nature Methods *17*, 37–40.

[75] Itoh, Y., Woods, E.J., Minami, K., Maeshima, K., and Collepardo-Guevara, R. (2021). Liquid-like chromatin in the cell: What can we learn from imaging and computational modeling? Current Opinion in Structural Biology *71*, 123–135.

[76] Feig, M. and Sugita, Y. (2019). Whole-Cell Models and Simulations in Molecular Detail. Annual Review of Cell and Developmental Biology *35*, 191–211.

[77] Raveh, B., Sun, L., White, K.L., Sanyal, T., Tempkin, J., Zheng, D., Bharath, K., Singla, J., Wang, C., Zhao, J., et al. (2021). Bayesian metamodeling of complex biological systems across varying representations. Proceedings of the National Academy of Sciences *118*, e2104559118.

[78] Mu, J., Liu, H., Zhang, J., Luo, R., and Chen, H.F. (2021). Recent Force Field Strategies for Intrinsically Disordered Proteins. Journal of Chemical Information and Modeling *61*, 1037–1047.

[79] Rahman, M.U., Rehman, A.U., Liu, H., and Chen, H.F. (2020). Comparison and Evaluation of Force Fields for Intrinsically Disordered Proteins. Journal of Chemical Information and Modeling *60*, 4912–4923.

[80] Saikia, N., Yanez-Orozco, I.S., Qiu, R., Hao, P., Milikisiyants, S., Ou, E., Hamilton, G.L., Weninger, K.R., Smirnova, T.I., Sanabria, H., et al. (2021). Integrative structural dynamics probing of the conformational heterogeneity in synaptosomal-associated protein 25. Cell Reports Physical Science *2*, 100616.

[81] Crehuet, R., Buigues, P.J., Salvatella, X., and Lindorff-Larsen, K. (2019). Bayesian-Maximum-Entropy Reweighting of IDP Ensembles Based on NMR Chemical Shifts. Entropy *21*, 898.

[82] Ding, C., Wang, S., and Zhang, Z. (2021). Integrating an Enhanced Sampling Method and Small-Angle X-Ray Scattering to Study Intrinsically Disordered Proteins. Frontiers in Molecular Biosciences *8*, 621128.

[83•] Gomes, G.N.W., Krzeminski, M., Namini, A., Martin, E.W., Mittag, T., Head-Gordon, T., Forman-Kay, J.D., and Gradinaru, C.C. (2020). Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. Journal of the American Chemical Society *142*, 15697–15710.

> •Predicted conformational ensemble of IDPs combining SAXS and NMR data and validated with sm-FRET data.

[84] Thomasen, F.E. and Lindorff-Larsen, K. (2022). Conformational ensembles of intrinsically disordered proteins and flexible multidomain proteins. Biochemical Society Transactions *50*, 541–554.