\Chapter{Summary and perspectives}

\Section{Summary}

The previous Chapters in this thesis have introduced and showcased novel approaches for explorative and integrative modeling in the presence of cryo-EM data and distance restraints. In \inchapter[chapter:powerfit] I presented the PowerFit software, a Python package for fast cross correlation based rigid body fitting of high-resolution structures in low-resolution densities. PowerFit comes with a new more sensitive scoring function, the core-weighted local cross correlation, in addition to an optimized protocol for fast fitting. In \inchapter[chapter:image-pyramids] I reported results of an extensive benchmark of the PowerFit software using 379 subunits of 5 ribosome density maps. The success rate of unambiguously fitting subunits larger than 100 residues reached approximately 90\% up to 12Å resolution, showing that objective fitting methods have matured to usable aids in structural modeling. The limits of rigid body fitting can be leveraged through the use of image pyramids to gain a speedup of a factor of 30 on CPUs and 40 on GPUs, and it allows the identification of possible over-interpreted regions of the density on an objective basis.

\inchapter[chapter:haddock-em] describes the incorporation and benchmarking of cryo-EM data into the data-driven docking program HADDOCK. The approach is flexible and can be fully combined with other available sources of data in HADDOCK, making it a fully integrative modelling approach. It was demonstrated on two ribosome systems, two virus-antibody systems, and a symmetric pentamer. An update of the HADDOCK webserver was presented in \inchapter[chapter:haddock2.2], together with extensive usage statistics of the software all over the world.

\inchapter[chapter:disvis] dealt with explorative modeling using distance restraints in general, and cross-link data specifically. I introduced the concept of the accessible interaction space and presented a method to quantify and visualize it. This directly indicates the information content of distance restraints and shows whether all data are self-consistent and, if not, it gives an indication of which restraint is a false-positive. This was implemented in another Python package, DisVis. The approach is general and can easily be incorporated into FFT-based docking programs allowing the use of distance restraints by combining the 'marriage made in heaven' of sampling and scoring \cite[Vajda2013].

I extended this approach further in \inchapter[chapter:inferring-interface-residues], presenting a method to infer interface residues from distance restraints using the concept of the average-interactions-per-complex (AIC) statistic. The AIC provides an objective probability for a residue to be at the interface based on the available data. Furthermore, I benchmarked the use of cross-link based distance restraints in HADDOCK using four different approaches. My results show that using solely unambiguous distance restraints is suboptimal; instead they should either be complemented with center-of-mass restraints or DisVis-based ambiguous interactions restraints.

\Section{Challenges of integrative modeling}

The field of integrative modeling is still relatively young, with several challenges ahead that the structural biology community will have to face, since integrative approaches are increasingly applied to solve the structure of large macromolecular assemblies. Recently a task force was assigned by the Worldwide PDB (wwPDB) to make recommendations for the field to follow in order to consistently progress and allow a proper assessment of the quality of such integrative models. The results of the First wwPDB Hybrid/Integrative Methods Task Force Workshop were recently published \cite[Sali2015], with 5 main recommendations about data-representation, model validation and data-archives. These were that: 1) the experimental and computational protocols in addition to the structural models should be deposited; 2) multiple model representations should be allowed for multi-scale and multi-temporal models; 3) new procedures should be developed to ascertain model uncertainty and accuracy; 4) a federated system of data archives should be created; and 5) publications standards need to be developed for integrative models as is already the case for X-ray and NMR structures.

Thus point 1, 4 and 5 are mainly about the reproducibility of integrative structural models, point 2 is about what data-structures and format standards to use, so far all more practical matters reflecting the current immature status of the field than real inherent scientific challenges. Point 3 highlights a current challenge in this field with respect to the precision and accuracy together with the validation of integrative models. Even though for several experimental techniques, cross-validation (SAXS \cite[Rambo2013]) and confidence interval (cryo-EM \cite[Volkmann2009]) measures have been developed, they have been infrequently used, except in X-ray crystallography where this has been a standard since years (the concept of the free R-factor \cite[Brunger1992]), and thus far not been combined. For other methods such as cross-links coupled with mass-spectrometry (CXMS) the statistical propensities of derived distance restraints have only been sparsely studied for small benchmark and sample sizes \cite[Kahraman2013, Leitner2014].

Gaining deeper insight into the uncertainty of integrative models and current validation approaches, requires new high-quality and elaborate benchmarks on systems for which high-resolution structures of both the bound and unbound states are available,, of which the protein-protein docking benchmark is a prime example \cite[Vreven2015] (although not really representative of the complexity of systems typically studies by integrative modelling approaches), together with additional experimental data. Especially for upcoming promising techniques as SAXS and CXMS, experimental data on multiple well-investigated systems are missing even for binary protein interactions. Although there are databases for CXMS \cite[Zheng2013, Kahraman2013], they are relatively limited in size, e.g. the XLdb reports 62 intra-chain cross-links of which 34 are coming from a single RNA polymerase II system \cite[Rappsilber2011, Kahraman2013]. The small sample size and questionable reproducibility of the results are major limitations in the development of robust validation and uncertainty assessment tools. Thus, for the integrative structural biology field to properly move forward a \emph{quid pro quo} mentality needs to be established between experimental and computational scientists: additional experiments should be performed for the purpose of further understanding the scope and limitations that the data are providing, so that, in turn, improved computational models can be delivered to answer important biological questions.

\Section{Future guidelines and additional fields of research}

\Subsection{Explorative modeling}

To adequately model the uncertainty of integrative models more emphasis should be put on the data themselves by investigating the amount of information the data carry by searching and quantifying the whole interaction space. I presented in this thesis a methodology in \inchapter[chapter:disvis] to assess the information content of distance restraints, information that can be obtained from a variety of techniques. Note, however, that the approach presented is limited to binary complexes and fully characterizing the interaction space of multi-component systems remains an open challenge. The approach for appreciating the information content of distance restraints can be further extended by using a statistical distance preference function, i.e. a knowledge based potential, inferred from experimental data, to better investigate the probability distribution of the accessible interaction space. Similar approaches can be developed for SAXS (though computationally more expensive as the scattering curve needs to be calculated millions to billions of times, a more CPU-demanding process than a simple distance calculation), and other biochemical and biophysical based potentials, such as surface overlap/van der Waals interactions. Thus, instead of heuristically optimizing the number of acceptable models within the top X best scoring structures using a linear combination of (pseudo-)energies, as is common in the docking field \cite[Vajda2009], the energy distributions can be analyzed to give further indication of the reliability of each measure and from there to define confidence intervals in models. Established probability distributions can afterwards be used as Bayesian priors in an effort to move to Bayesian statistical models.

Furthermore, currently integrative modeling is often used to generate preferably a single or a few models that represents the data, even though the original outset of the approach is to generate all models that can be represented by the data. This requires a different mindset of the structural biologist in general, by regarding a structural model not simply as a single entity, but rather as a whole set of conformations. This approach, however, is gaining ground, e.g. in X-ray crystallography methods are being developed that represent the electron density as a set of conformers \cite[Fraser2011, vandenBedem2013] and ensembles \cite[Burnley2012]. These representations are only the tip of the iceberg within this mindset, as the ensemble space will be significantly bigger in the presence of sparse data, such as CXMS data. Model representations should thus become more diffuse as the accessible interaction space becomes larger, to accurately present the ensembles consistent with the data. Again, explorative modeling techniques can help here by quantifying the information content to provide insight into the magnitude of the ensemble space, while concurrently easing the transition from a single-structure mindset to an elaborate multi-ensemble paradigm.

\Subsection{Formal structural biology}

Further investigating the accuracy of individual experimental methods require scientists that are trained in both computational and experimental techniques, the \emph{hybrid scientist}. This allows the scientist to perform experiments to further guide and validate the computational modeling, ultimately resulting in a \emph{formal structural biology}, where instead of only advancing biological insight, the emphasis is also put on investigating the accuracy and precision of both models and experimental methods \emph{an sich} and the interpretation of the generated results. In the semi-long run, this approach will become a fertile and stabile foundation to build upon for in-depth structural research in challenging and interesting biological systems and networks. This will ultimately result in a more formal approach to structure determination from multiple data sources.

\Section{Conclusion}

The (integrative) structural biology field is a fast moving and exciting field of research, with many experimental and computational advances. The most recent dramatic example is of course the spectacular improvement of the cryo-EM field due to direct electron detectors. Even though atomic resolution can now be achieved for stable complexes, the bulk of the resulting densities need additional data from diverse sources for a structural interpretation, requiring high-end integrative methods. However, there are still significant challenges to overcome for integrative modeling to become a standard tool in the toolbox of structural biologists. These are mainly dealing with the reproducibility of the results and the uncertainty of the models. By showcasing integrative modeling approaches and introducing new methods for quantifying the information content of experimental data this thesis has laid out some new building blocks for the field to build upon and move forward