# Constrained optimization applied to multiscale integrative modeling

THÈSE N$^{O}$ 8630 (2018)

PRÉSENTÉE LE 1$^{ER}$ JUIN 2018

À LA FACULTÉ DES SCIENCES DE LA VIE
UNITÉ DU PROF. DAL PERARO
PROGRAMME DOCTORAL EN BIOTECHNOLOGIE ET GÉNIE BIOLOGIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Giorgio Elikem TAMO

acceptée sur proposition du jury:

Prof. P. De Los Rios, président du jury
Prof. M. Dal Peraro, directeur de thèse
Prof. V. Zoete, rapporteur
Prof. A. Bonvin, rapporteur
Prof. B. Correia, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2018

## Acknowledgements

I wish to sincerely thank and acknowledge the support of my thesis supervisor Prof. Matteo Dal Peraro, whose encouragements, trust, relentless and contagious enthousiasm has motivated me throughout the years undertaken here at EPFL for my thesis research.

Being in a warm and relaxed working environment was a major factor that kept me positive and greatly contributed to keep my motivation level high, especially during times of pressure. Therefore, I also want to heartfully thank the previous and current lab members of the Laboratory Biomolecular modeling for the endless laughter and help.

I wish to thank the people I had the immense pleasure to collaborate with. These particularly include my colleagues Dr. Luciano Abriata, Syvlain Traeger and Deniz Aydin, whose enthusiasm, easy going attitude and sharp mind enabled me to always try improve my critical abilities. A major thank you goes to Dr. Andrea Maesani, who was certainly a source of inspiration for me during my PhD years. working alongside him for a short period, I could learn essential career skills including time, stress and project management. I would also like to thank our external collaborator, Prof. Song at the KAIST university for his enthusiasm and drive witnessed during our collaborative project.

I am also very grateful for all the valuable people working with us here the EPFL, and whith whom I shared countless coffees and light hearted discussions. These include Sebastien Ferrara from the SVIT, Magali Masson, Sonja Bodmer and Julia Prebandier.

Of course, life outside the lab I believe is equally as important as life inside it, as it provides a good mental balance. This is why I wish to thank my family for unconditionally supporting, through giving advice and love. Especially, I would like to dedicate this thesis to my mother, who has pushed me and always encouraged to pursue higher studies, which eventually led me to undertake a PhD here at the EPFL. I would also like to thank my dear fiancée Olga, who helped shape my critical mind by providing an honest and critical look at my scientific work.

## Abstract

Multiscale integrative modeling stands at the intersection between experimental and computational techniques to predict the atomistic structures of important macromolecules. In the integrative modeling process, the experimental information is often integrated with energy potential and macromolecular substructures in order to derive realistic structural models. This heterogeneous information is often combined into a global objective function that quantifies the quality of the structural models and that is minimized through optimization. In order to balance the contribution of the relative terms concurring to the global function, weight constants are assigned to each term through a computationally demanding process. In order to alleviate this common issue, we suggest to switch from the traditional paradigm of using a single unconstrained global objective function to a constrained optimization scheme. The work presented in this thesis describes the different applications and methods associated with the development of a general constrained optimization protocol for multiscale integrative modeling.

The initial implementation concerned the prediction of symmetric macromolecular assemblies throught the incorporation of a recent efficient constrained optimizer nicknamed mViE (memetic Viability Evolution) to our integrative modeling protocol *pow*$^{er}$ (parallel optimization workbench to enhance resolution). We tested this new approach through rigorous comparisons against other state-of-the-art integrative modeling methods on a benchmark set of solved symmetric macromolecular assemblies. In this process, we validated the robustness of the constrained optimization method by obtaining native-like structural models.

This constrained optimization protocol was then applied to predict the structure of the elusive human Huntingtin protein. Due to the fact that little structural information was available when the project was initiated, we integrated information from secondary structure prediction and low-resolution experiments, in the form of cryo-electron microscopy maps and crosslinking mass spectrometry data, in order to derive a structural model of Huntingtin. The structure resulting from such integrative modeling approach was used to derive dynamic information about Huntingtin protein.

At a finer level of resolution, the constrained optimization protocol was then applied to dock small molecules inside the binding site of protein targets. We converted the classical molecular

docking problem from an unconstrained single objective optimization to a constrained one by extracting local and global constraints from pre-computed energy grids. The new approach was tested and validated on standard ligand-receptor benchmark sets widely used by the molecular docking community, and showed comparable results to state-of-the-art molecular docking programs.

Altogether, the work presented in this thesis proposed improvements in the field of multiscale integrative modeling which are reflected both in the quality of the models returned by the new constrained optimization protocol and in the simpler way of treating the uncorrelated terms concurring to the global scoring scheme to estimate the quality of the models.

**Keywords :** Multiscale integrative modeling, molecular modeling, constrained optimization, Huntingtin, structure-based molecular docking.

## Résumé

La modélisation integrative à échelle multiple se situe à l'intersection des techniques experimentales et informatiques pour prédire les structures atomistiques d'importantes macromolecules. Dans le processus de modélisation integrative, l'information expérimentale est souvent intégrée avec des potentielles d'energie et des sous-structures de macromolecules dans le but d'obtenir des models structurels réalistes. Ces termes de nature hétérogène sont souvent combinés en une seule fonction objective globale qui caractérise la qualité de ces modèles structurels et permet leur optimisation par le biais d'une minimisation. Afin d'équilibrer les contributions relatives aux termes contribuant à cette function objective globale, des constantes sont attribuées à chaque terme au moyen d'un processus de calcul coûteux. Dans le but de résoudre ce problème, nous suggérons de passer du paradigme traditional, consistant à utiliser une seule fonction global, à une optimisation sous contraintes. Le travail présenté dans cette thèse décrit les différentes applications et méthodes associées au developpement d'un protocole general d'optimisation sous contraintes pour la modélisation intégrative à échelle multiple.

La mise en œuvre initiale concernait la prédiction d'assemblages macromoléculaires symétriques grâce à l'incorporation d'un optimiseur récent nommé mViE (memetic Viability Evolution) à notre protocole de modélisation intégrative *pow*$^{er}$ (parallel optimization workbench to enhance resolution). Nous avons testé cette nouvelle approche au moyen de comparaisons rigoureuses à d'autres méthodes de modélisation intégrative sur un ensemble d'assemblages macromoléculaires symétriques déjà résolus. Dans ce processus, nous avons validé la robustesse de la méthode d'optimisation sous contrainte en obtenant des modèles structurels ressemblant à ceux résolus par méthode expérimentale.

Ce protocole d'optimisation sous contrainte a ensuite été appliqué pour prédire la structure de la protéine Huntingtin humaine. Pour ce faire, nous avons intégré des informations issues de la prédiction de structure secondaire et des expériences à basse résolution, sous forme d'enveloppes de microscopie cryoélectronique et de données de spectrométrie de masse réticulaires. La structure résultant d'une telle approche de modélisation intégrative a été utilisée pour dériver des informations dynamiques sur la protéine Huntingtin.

À un niveau de résolution plus fin, le protocole d'optimisation sous contrainte a ensuite été appliqué pour amarrer des petites molécules à l'intérieur du site de fixation de protéines. Nous avons entrepris cela en extrayant des contraintes locales et globales à partir de grilles d'énergie pré-calculées. Cette nouvelle approche a été testée et validée sur des complexes ligand-récepteur déjà résolus de façon expérimentale. Ainsi, nous avons obtenus des résultats comparables à des programmes d'amarrage moléculaire de pointe.

En conclusion, les travaux présentés dans cette thèse proposent des améliorations dans le domaine de la modélisation intégrative à échelle multiple qui se traduisent à la fois par la qualité des modèles obtenus par le nouveau protocole d'optimisation sous contrainte, et par le traitement plus simple des termes non corrélés qui caractérisent de la qualité des modèles.

**Mots-clés** : Modelisation intégrative à échelle multiple, modélisation moléculaire, optimisation sous contrainte, Huntingtin, amarrage moléculaire.

# Content

## LIST OF FIGURES

# Chapter 1 Introduction

Adapted from the published papers: "**The importance of dynamics in integrative modeling of supramolecular assemblies**" Giorgio E. Tamò, Luciano A. Abriata and Matteo Dal Peraro, *Current Opinion in Structural Biology*, 2015, and "**Assessment of data-assisted prediction by inclusion of crosslinking/mass-spectrometry and small angle X-ray scattering data in the 12th Critical Assessment of protein Structure Prediction experiment**" Giorgio E. Tamò, Luciano A. Abriata, Guilia Fonti and Matteo Dal Peraro. *Proteins: structure, function and bioinformatics*, 2018

## 1.1 Integrative modeling for molecular assembly

Supramolecular complexes are the cornerstone of cellular architecture and function. Large assemblies of macromolecules are involved in DNA remodeling, translation and transcription, RNA processing, protein synthesis and degradation, import, export and injection of solutes through cell membranes and to different organelles, ATP synthesis, respiration and photosynthesis, packaging of viral nucleic acids, membrane reshaping, just to name some of the systems best characterized to date.

Recent advances in cryo-electron microscopy (cryo-EM) mainly consisting in the use of direct electron detector and better image processing methods [1], are today increasing the discovery and elucidation of near-atomistic resolution of medium to very large assemblies [2]. In fact, the number of electron maps deposited yearly in the EM-databank is growing exponentially with a rate of >1000 maps per year since 2015 (http://www-ebi.emdatabank.org/statistics.html). Noteworthy, the recent elucidation of large macromolecules including the 70s ribosome structure at 3.6Å[3], the 26S proteasome at 4.2 Å [4], the RNA-polymerase I [5], II [6] and III [7] at 3.8 Å, 3.4 Å anf 3.1 Å respectively, and even extremely large macromolecules such as the 150MDa human adenovirus 26 solved at 3.7Å of resolution [8].

In these remarkable examples, the structural elucidation of these dynamic complexes at atomistic resolution was key to shed light on their function. Nevertheless, whereas the atomic structure of single proteins and complexes of relatively low molecular weight can generally be obtained by NMR spectroscopy and/or X-ray crystallography, the atomistic structure of several supramolecular assemblies are not yet routinely accessible using cryo-EM experiments [9]. Despite their tremen-

dous gain in resolution for some macromolecules, the volumetric shapes obtained from cryo-EM techniques might often not have the resolution required to unambiguously define the atomic location of constituent subunits.

Fortunately, the lack of high resolution data for these elusive large macromolecules can be complemented with a broad array of experimental methods that can provide lower resolution information about overall shape, symmetry, composition, contact sites between constituent molecules, angular and distance restraints between domains, as extensively reviewed in ref. [10].

When high resolution data of the subunits composing the macromolecular complex are available, e.g. through X-ray or NMR experiments, these can be combined with experiments such as small-angle X-ray or neutron scattering (SAXS/SANS) experiments, cryo-electron microscopy (cryo-EM) or tomography which would provide coarser details of the assembly such as size, volume and shape. Moreover, contact information such as residue-residue contacts unveiled through mutagenesis, chemical cross-linking and ion mobility mass spectrometry [11], or even from coevolution analysis [12-14], can help to define assembly rules.

In this context, computational approaches commonly called integrative modelling (IM) attempts to consistently combine these heterogeneous, and sometime incomplete, data with the structures of the individual subunits that constitute a complex in order to generate models at near atomistic resolution [15].

To address this challenge, IM method traditionally converts the prediction of macromolecular assemblies into an optimization problem; that is, a problem that can be encoded as a fitness function that describes the quality of the assembly and that can be minimized/maximised by an efficient heuristic algorithm[16-18]. This function usually combines the experimental data, which can be translated to spatial structural constraints, and energy terms, as extracted from common force fields, in order to obtain physically plausible models. The experimental and energy terms are most often linearly combined into a single fitness function. In order to balance the relative contribution of each term concurring to the single fitness function, weighted constants are assigned. The computation of this weights remains challenging because they often require heavy computation and can be biased towards the training set they originate from.

Another issue generally faced by IM approaches consists in protein flexibility. Precisely, the structure of subunits as determined in isolation often differ, with varying degrees, from the conformation they adopt upon supramolecular assembly [19], further complicating the prediction of native

architectures. Such scenario might arise either from artifacts induced by the conditions under which the structure of individual subunits is solved and/or may arise from the native dynamics underlying the molecular recognition pathways that lead to assembly through mechanisms as diverse as conformational selection or induced fit. One way or the other, this calls for the inclusion in integrative modeling protocols of the global and local dynamics of the individual subunits undergoing macromolecular assembly. Notably, the inclusion of dynamic features cannot only lead to more reliable models, but can also reveal fine details of the assembly mechanism and the existence of multiple functional states.

On one hand, some integrative modeling approaches use experimental restraints to drive deformations of the subunits upon assembly, gradually improving the fit to the experimental restraints. This typically involves biasing with the given restraints the algorithms that perform backbone/sidechain refinement, normal modes analysis [20] or molecular dynamics simulations of variable granularity [19,21] (Figure 1.1A). The other main avenue consists in using the experimental restraints to select conformations from existing ensembles, namely compiled from X-ray and NMR ensembles, homology models, and/or MD trajectories (Figure 1.1B). Whereas the first set of strategies resembles and can potentially describe induced-fit mechanisms underlying protein recognition, the second class of approaches seems well positioned to capture structural assembly driven by conformational selection.

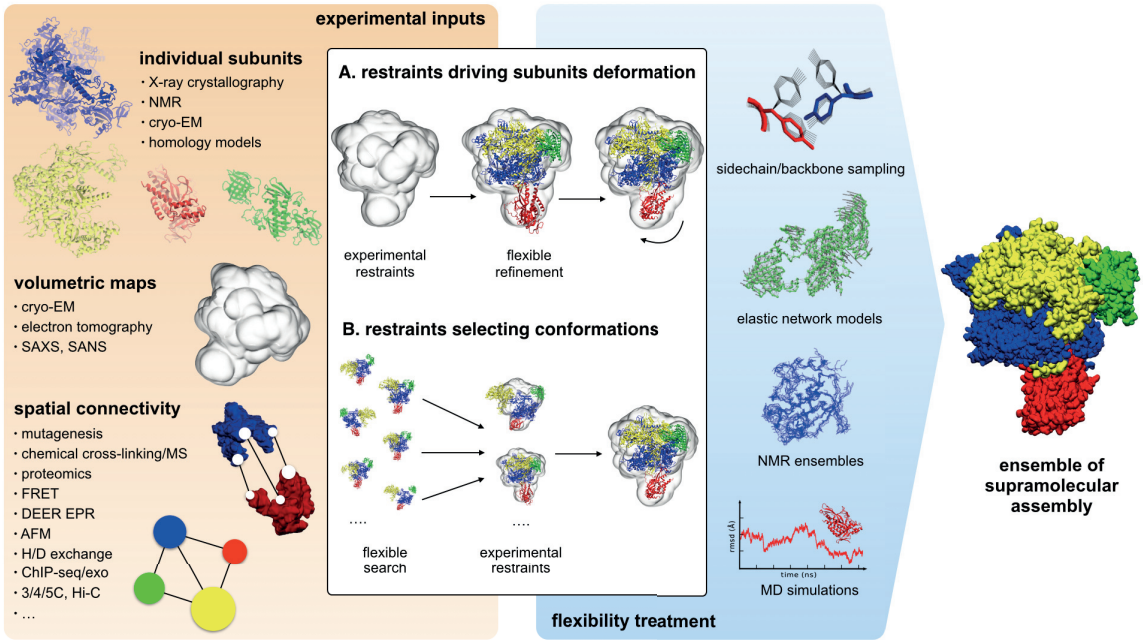



**Figure 1.1 | Integrative modeling strategies accounting for flexibility and dynamics.**

Integrative modeling requires as input the structures of the individual subunits composing the supramolecular complex. Structures should usually have high-resolution, but, when not available, coarser molecular representations can also be used, along with any experimental data (e.g., experimental inputs at decreasing levels of resolution are indicated in the orange box). Experimental restraints (e.g., cryo-EM maps in this example) can be used upfront to guide the prediction of supramolecular complexes (A), or, alternatively, conformations can be sampled without restraint biases and then filtered to satisfy them (B). In both approaches the flexibility of individual subunits within the complex can be accounted for with different accuracy, from local refinement of intermolecular interactions, to elastic network based structure deformation, to MD-based exploration of the conformational space (see the blue box). As a final result, integrative modeling methods aim at producing biologically relevant supramolecular assemblies consistent with the available set of experimental data. The protein structures used for this example correspond to the RNA polymerase II complex (PDB 4A3D [22]).

### 1.1.1 Using experimental restraints to drive deformation of subunits

A number of methods use experimental spatial restraints, as obtained from low-resolution experiments, to directly drive the physical deformation of starting structures into consistent conformations. With a very simple and coarse way to represent protein flexibility, Situs [23] performs very well at fitting assemblies to volumetric maps derived from EM experiments within a broad range of resolution (*e.g.,* up to 30 Å [24]). Within this approach flexibility is considered by converting starting structures into a skeleton that is allowed to sample conformations following distributions observed

in the Protein Data Bank. In this way the fitting protocol uses a knowledge-based force field on which restraints are implemented to penalize shape differences between the assembled model and the experimental volume [23]. Although the dynamics of individual subunits is not completely sampled, this is a simple way to enlarge the conformational space accessible for assembly, which has been successfully applied to several systems [25-28] most notably myosin fibers [24] and full muscle filaments [29].

Similar ways to adjust individual structures into assemblies make use of normal modes computed from a deformable elastic network, as done by DireX [30] and iMODfit [31,32]. Recently, it has been also shown that the combination of diverse flexible fitting protocols of this kind can improve pseudo-atomistic models based on intermediate-resolution EM maps, providing in turn a general way to assess the fits [33].

A possible limitation of this kind of approaches is that conformations generated by normal mode analysis might not always be physically accurate, but further refinement using force field based molecular dynamics (MD) simulations can improve over this drawback. Bock *et al.* [34] used this approach to fit atomic X-ray structures of the *E. coli* ribosome into a set of cryo-EM density maps captured at different stages of the tRNA translocation process. The subsequent MD analysis of the different states was able to reveal their progression along the tRNA translocation pathway, estimating the timescale of transitions and the key driving forces along the process. In the same vein, DireX has been recently used for the refined flexible fitting of cyclic nucleotide-binding potassium channel crystal structure inside EM density maps corresponding to different conformational states. This helped in elucidating the ligand-induced gating mechanism of this protein [35].

At the most accurate end of the dynamic treatment of assembly structures is the direct use of MD simulations restrained by experimental inputs, which profits from an accurate description of the physico-chemical features of the system. Within this group of methods, the molecular dynamics flexible fitting (MDFF) protocol [36] uses the gradient of the electron density distribution to compute forces that are added to those of a classical force field. This method has been widely applied since its introduction [37,38], one of its most recent and striking applications being the assembly of the HIV virus capsid into an 8 Å cryo-EM map [39]. Illustrating the importance of incorporating protein dynamics in integrative modeling, assembly of the HIV capsid revealed that deformation of the planar hexamer-of-hexamer and pentamer-of-hexamer capsid units is required to properly describe protein-protein contacts upon capsid assembly.

### 1.1.2 Using experimental restraints to select functional states from a conformational ensemble

Dealing with a broad variety of experimental inputs as spatial restraints, including cryo-EM, SAXS, NMR and proteomics data, we find the Integrative Modeling Platform (IMP) [40]. In order to satisfy the spatial restraints in an efficient manner, IMP makes use of a large variety of optimization algorithms [40]. This protocol has been successfully used to model the complex assemblies of the 26S proteasome [41], the nuclear pore complex [42-44] and recently the monooxygenase hydroxylase from *M. capsulatus* [45] and the 40S-elF1-elF3 translation initiation complex [46]. Despite this framework does not directly consider dynamic features of the complex subunits during model prediction, functional states of large assemblies can be discriminated from the experimental data using Bayesian inference [47,48]. Recently, two distinct functional dynamic states of the two-component system PhoQ were found to optimally fit a set of disulfide cross-linking data, supporting thus a scissoring mode of signal transduction in sensor His-kinase receptors [49].

Other protocols consider flexibility in a more explicit way, like Rosetta [50], where conformational ensembles can be generated from NMR experiments and Monte Carlo sampling, while further refinement of sidechain and backbone torsional angles is used incrementally to improve the quality of the assembled models [50]. In a recent advancement, dubbed "fold-and-dock" [51], it is also possible to simultaneously sample the internal flexibility of individual subunits under symmetry constraints, which allowed folding of a symmetric homodimer from its sequence and available experimental data. Improvements in comparative modeling (*i.e.,* RosettaCM [52]) and the inclusion of density maps as fitting guides for local refinement or more extensive model rebuilding now allow reaching near atomic resolutions when starting from cryo-EM maps of 4-10 Å resolution [53]. In a recent application, this approach has produced multiple atomistic snapshots based on high-resolution cryo-EM data leading to novel proposals on how GTP hydrolysis leads to microtubule dynamic instability and depolymerization [54].

Following similar premises is the well-established integrative modeling framework HADDOCK [18,55]. Initially designed to employ NMR data, it has also been used with evolutionary data predicting inter-protein contacts [13,56] and adapted to handle SAXS data [57]. Conformational flexibility is here accounted in two main ways; *i.e.* by providing ensembles from NMR structures, collections of X-rays structures or simulation snapshots, rather than static structures to the search algorithm [10], and by performing multidomain docking [58]. This latter scheme, like early programs such as

FlexDock [59] and RNAbuilder [38], splits a flexible binding partner into subdomains based on an elastic network model and treats the parts independently, but under the strong constraints that the two halves of a hinge must be spatially close in the complex [58]. Using this recipe, the dimeric structure of ubiquitin bound to the deubiquitinating enzyme Josephin was determined, helping to pinpoint its cleavage site [60]. In another recent example, this protocol has been used to build a complex of the interleukin 1 receptor bound to an antagonist, unveiling a conformational change as large as 20 Å upon binding [61].

At the most accurate end of this array of methods, one can sample subunit intrinsic dynamics using explicit MD simulations to determine conformational states which are relevant for the supramolecular complex, and have an optimization algorithm select the best set of conformations by filtering them so as to satisfy the restraints. The great advantage of such approach is that native physico-chemical determinants are more thoroughly accounted for and conformations are not biased a priori. Such strategy is at the core of the *pow*$^{er}$ (parallel optimization workbench for enhancing resolution) framework developed in our laboratory [16]. In this scheme, MD trajectories are indexed by principal component analysis and evolutionary based algorithms, like particle swarm optimization, are used to efficiently select optimal conformations that fulfill the set of experimental restraints. This strategy has already proved successful for the prediction of large assemblies like the HIV1 hexameric capsomer complex, the basal body of *Yersinia* type III secretion system and the C4b-binding protein [16,62-64]. Noteworthy, *pow*$^{er}$ was used to elucidate the assembly mechanism of the pore-forming toxin aerolysin [62], for which unbiased MD simulations were able to extensively characterize activated states of the monomeric toxin along the pore formation pathway, which were then used for the prediction of intermediate pre-pore and mature transmembrane pore states. This study revealed the role of the subunit intrinsic flexibility in driving a concerted swirling mechanism that mediates membrane pore insertion.

## 1.2 Integrative modeling for predicting protein tertiary structures

IM techniques have also been extended to elucidate the tertiary structure of proteins. In this case, the low-resolution experiments to be integrated for tertiary structure prediction can take the form of small angle X-ray scattering (SAXS) [57,65,66], crosslinking/mass-spectrometry (CLMS) [67] or even residue-residue contacts predicted from residue co-evolution analysis [68,69] can be integrated for tertiary structure predictions [70]. Residue co-evolution analysis in particular relies on statistical

techniques to detect residues in multiple sequence alignments of homologous proteins which tend to mutate together [71]. From such analysis, probable residue-residue contacts can be inferred and used to assist the modeling of tertiary protein structures. When supplemented with low-resolution experiments, the recent integration of co-evolution analysis has enabled to blindly model protein structures in a successul manner in the Critical Assessment of protein Structure Prediction (CASP) [70,72,73].

Briefly, CASP was created as an effort to objectively assess the state-of-the-art of protein structure prediction methods from amino acid sequences through an international competition in which groups have to predict protein structures secured by the organizers but which have not been released by the Protein Data Bank[72]. The predicted models are then evaluated against the respective experimentally solved structures (targets) in different tracks that look at specific features, such as global fold prediction, refinement of fold and side chain details, oligomer prediction, and of most relevance to this thesis, data-assisted modeling. This category in particular was included in the last 12$^{th}$ edition of CASP and featured the evaluation of IM techniques used to predict protein structures (of less than 50 kDa) by combining CLMS and SAXS data.

The main outcome from this data-assisted category was that co-evolution analysis applied to predict residue-residue contacts was a key factor when determining the tertiary protein structure from a their respective amino acid sequences[70]. In this case, assisting the modeling of protein structures with low-resolution experiments was useful only for few predicting groups but overall did not show meaningful improvements in the quality of the submitted structural models [73]. Nevertheless, the existence of a CASP data-assisted category is very recent and we expect future improvements both in technical advancements in IM methods and in the quality of the low-resolution experimental data to assist the prediction.

## 1.3 Integrative modeling for ligand-protein interactions

Although on a much finer scale, structure-based small molecule docking (SMD) faces similar challenges as IM. SMD is part of the virtual screening (VS) process of the drug discovery pipeline and is used to suggest active compounds from a screening database to be tested experimentally, with the aim to obtain binding affinities in the μM ($10^{-6}$) range. Practically, SMD consists in a computational approach where the structure of the target protein is available, either through experiments or modelling, and commercially available compounds from a large database are iteratively docked to the

target protein, then selected based on their estimated binding affinity. The aim of using VS over traditional high throughput screening experiments (HTS) is to significantly reduce the time and financial costs often associated with obtaining hit compounds. Notably, in a study comparing hit rate enrichment of VS over HTS on the protein phosphatase-1B target, Doman et al. [74] found that in the 400'000 compounds screened using experimental high throughput experiments, 85 compounds inhibited the enzyme with $IC_{50} < 100$ μM, giving an enrichement of 0.021%. In the same study, the authors also virtually screened 235'000 commercially available compounds against the same target. From this procedure, the binding affinity of the suggested 365 best-scoring molecules was experimentally measured and 127 of them were found to have $IC_{50} < 100$ μM, giving a 1700-fold (34.8%) enrichment over the HTS method. In this study, the DOCK[75] software was used. Other popular SMD methods include AutoDock suite[76], Attracting cavity[77], SwissDock [78] and HADDOCK [79]. Noteworthy, the AutoDock Vina docking sofware is freely accessible, open-source and robustly benchmarked [76]. Commercially available software include ICM [80] and Glide[81]/Gold[82]. Similar to IM protocols, SMD methods convert the virtual screening problem as an optimization problem were the SMD algorithm attempts to minimize a global fitness function qualifying and quantifying the quality of the binding between the small molecule and the target protein.

The SMD procedure is traditionally decomposed in two distinct steps, which are referred to as sampling and ranking. Considering the target protein as a rigid receptor fixed in space, the sampling step consists in exhaustively probing the roto-translations and internal flexibility of the small molecule, used as a ligand, in order to obtain a ligand-receptor pose similar to the native one. The second step consists in accurately estimating the binding affinity of the ligand-receptor poses generated from the sampling step.

Unlike IM methods applied to protein structure prediction, SMD methods usually do not rely on experimental data to drive the docking of ligand inside the binding site of a given receptor, but on scoring functions composed of carefully calibrated uncorrelated terms. In principle, if any information susceptible to drive the docking of ligand inside the binding site of receptors were available, greater accuracy could be conferred to the docking procedure. Nevertheless, the tedious and computationally demanding process of re-calibrating an already established scoring function almost forbids the addition of new scoring terms without rebalancing their contribution. In case scoring function terms in the form of geometrical constraints can be computed prior to the docking, IM methods can use such constraints to drive the docking of ligands.

## 1.4 Objective of the thesis

Currently, the atomistic structural determination of macromolecular complex remains a daunting challenge. Despite recent computational and experimental advances to increase structural resolution of these important biological systems, many avenues could be suggested for improvements and the structure of several important macromolecules remains to be elucidated. The work featured in this thesis proposes improvements in the computational methods employed to model important structural complexes consisting both in protein-protein and protein-small molecule complexes.

**Constrained optimization for integrative modeling (Chapter 3)**. Most IM protocols attempt to model the macromolecular assembly problem as an unconstrained problem featuring a single objective, which could lead to non-optimal aggregation of heterogeneous components concurring to the global fitness function. As a possible solution to this problem, we suggest to shift from the traditional unconstrained optimization to a constrained optimization. To face this challenge we incorporated a novel constrained optimization algorithm called memetic viability evolution [83] into our in-house IM protocol *pow*$^{er}$. The aim of this approach was to separate the fitness function components, previously linearly combined into a single unconstrained fitness function, into constraints and objective separately. For validation purposes, we first benchmarked our new IM protocol on symmetrical assembly cases featured in the literature. After making sure our protocol was robust, we then took advantage of the flexibility the new protocol conferred and extended our analysis on symmetrical assembly complexes featuring volumetric maps as obtained from cryo-EM experiments.

**Towards a near-atomistic model of Huntingtin protein: integrative modelling to the rescue (Chapter 4)**. Huntingtin (Httn) is an important protein that is ubiquitously found in the cytoplasm of cells and which function has been associated with membrane vesicles and organelles [84,85], and more generally with membrane trafficking [86]. Importantly, poly-glutamine (poly-Q) expansions at the N-terminus of Httn has been linked to the serious neurodegerative Huntington's disease (HD), for which no treatment currently exists to slow or halt disease progression. HD is an inherited neurodegenerative disorder that is autosomal dominant, and which was found to cause progressive damage to the brain until eventually leading to death. It was observed that HD operates by killing neuron cells located in the basal ganglia through structural aggregation. The extent of poly-Q expansion of Httn has been suggested as the main culprit for protein aggregation and thus of HD

symptoms. Nevertheless, how these factors are related is still unknown to date. Until the 21st of February 2018 (while this thesis was still being written), no atomistic structure of the Httn was available. On this date, the group Guo et al. [87] published the structure of the Q17-Httn solved from a 4 Å cryo-EM map. Importantly in order to obtain a cryo-EM map at such high resolution, the authors stabilized the Q17-Httn through the binding of the HAP40 protein [87], which likely induced conformational differences with this structure and the native monomeric Httn. Recently, we obtained from our collaborators at the Song Lab (KAIST) single particle cryo-EM data solved at 8-11 Å of the full-length monomeric Httn with varying degrees of poly-Qs and CLMS information specifying the intra-protein residue contacts. The work described here is an attempt to model the full-length Httn structure by combining the experimental information available prior to the publication of the Q17-Httn [87]. We also used this new high-resolution structure to assess our modeling and derive the native conformation of Httn in isolation, i.e. without the effect of HAP40, and the impact of poly-Q expansion on its flexibility.

**Constrained optimization for small molecule docking (Chapter 5).** Due to the fact that thousands to millions of compound molecules have to be tested against a given target, SMD protocols need to be fast and computationally efficient in term of resources used. During the sampling steps the best receptor/ligand poses need to be quickly identified from all other generated poses. This difficult task is commonly achieved using a fitness function that linearly combines quantitative features describing the binding pose. The relative weight constants assigned to balance the relative contribution of these features require careful and tedious computations. This almost forbids the addition of new feature terms inside an already established fitness function. Nevertheless, by switching from an unconstrained to a constrained optimization protocol, we envision the addition of new features possible and effortless. To achieve this, a robust fitness function is kept as objective while an unlimited amount of new terms can be added as inequality constraints to guide the optimization towards feasible areas of the search space. Practically, the work described here attempts to solve the SMD problems as a constrained optimization problem with the *pow*$^{er}$*-mViE* framework (Chapter 3). To do so, the well-established Autodock Vina fitness function is used as objective while energy grid-based parameters are added as inequality constraints to guide the assembly towards favorable areas of the optimization search space.

The present thesis is outlined as follows. First, we will briefly describe the computational techniques and optimization concepts employed in this work (Chapter 2). Then we will introduce the topic of constrained optimization applied to macromolecular assembly with symmetrical assembly showcases (Chapter 3). We will then extend this protocol into the integrative modelling of the full Httn protein by combining experimental information derived from cryo-EM and chemical cross-linking (Chapter 4). Finally, the constrained optimization method is applied to flexible small molecule docking with extensive testing and examples (Chapter 5). Eventually we close this thesis with conclusions and perspectives (Chapter 6), together with additional publications and Annexes related to the work presented.

# Chapter 2 Methods

Parts adapted from: "**Disentangling constraints using viability evolution principles in integrative modeling of macromolecular assemblies**" Giorgio Tamò, Andrea Maesani, Sylvain Traeger, Matteo T. Degiacomi, Dario Floreano, Matteo Dal Peraro. *Scientific Report*. 2017.

Integrative modeling techniques aim at finding a reasonable prediction of macromolecular complex structures using the structural information from high-resolution subunits and low-resolution experimental data. One of such established IM protocols is the parallel optimization workbench to enhance resolution (*pow*$^{er}$) framework, which was extensively used in this work. Though it has been implemented to solve virtually any optimization problem, here, it has been applied mainly to solve macromolecular assembly problems. The *in silico* assembly of large macromolecular assemblies is considered difficult due to the large search space the individual subunits have to explore prior to multimerization and the multiple local minima of the fitness function used to quantify the quality of the structural models. To address this challenging task, *pow*$^{er}$ features state-of-the-art optimization algorithms which performance have been extensively tested [16,63,64].

## 2.1 *pow*$^{er}$ general workflow and architecture

The original *pow*$^{er}$ framework has been originally designed and implemented by Matteo T. Degiacomi. Related work can be found in his thesis and in ref. [62]. The *pow*$^{er}$ framework has been conveniently divided into two main effector parts which are (i) the module designed to solve the optimization problem (e.g. symmetric assembly, asymetric assembly) and (ii) the heuristic algorithms performing the optimization. A module is a Python object composed of several classes (blue boxes in Figure 2.1) which manipulate the input data, interact with the optimizer (red box in Figure 2.1) and output a solution as final result.

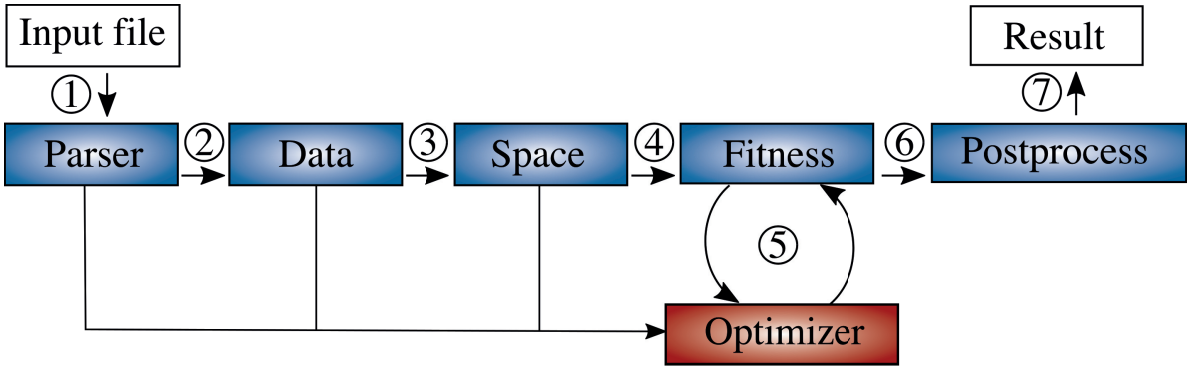



**Figure 2.1 | General workflow of *pow*$^{er}$.**

The blue boxes represent the classes specific for the module designed to solve the optimization problem. In this schematic, data describing the optimization problem proceeds from input (left) to output (right). The optimizer, here represented as a red box, is the engine directly responsible for solving the optimization problem.

1. The protocol leading to the solution that solves the optimization problem starts with an input file containing specific keywords that will be read and incorporated within the module. In this case, the class Parser will verify the integrity of these keywords and translate them into Python variables.

2. From the keywords variables translated into Python variables, the class Data fetches and loads the information necessary to undertake the optimization. For instance, in the case of macromolecular assembly, this information may take the form of protein structure coordinates, crosslinked residue information, cryo-EM volumetric maps, etc.

3. Based on the data and on the keywords entered, the class Space will then define the search space of the optimizer.

4-5. The information defined by the previous classes is then transferred to the class Fitness and Optimizer, which are the core of the optimization process leading to the obtention of the final solution that minimizes the fitness function.

6. Once a termination criteria has been reached (e.g. convergence, budget of function evaluations, etc.) and global minimum is assumed to have been reached, the data is post-processed. The post-processing step includes for instance data sorting, clustering, comparisons, etc.

7. The final solution or clusters of representative solutions are returned.

## 2.2 Data : Experimental information to guide macromolecular assembly

In order to predict the structure of macromolecules, *pow*$^{er}$ integrates information typically extracted from low-resolution experiments. Some of this information can be translated into spatial distances specifying residue-residue contacts or assembly dimensions that must be satisfied during the optimization procedure, and that can take the form of residue mutations or cross-linking-mass-spectrometry (CLMS) data.

Other types of low-resolution information can be used in a more direct manner during the assembly process by optimizing a specific function that quantitatively describes the fit of the computed structural models and the experimental data. For instance, the satisfaction of cryo-EM data can be formulated as the maximization of a cross-correlation coefficient (*ccc*) which essentially quantifies the match between the electron densities present in the experimental density map and those simulated from the protein structure [23]. In a similar fashion, the satisfaction of SAXS data can be formulated as the minimization of a $\chi^2$ value that quantifies the difference between the experimental SAXS profile (*Ln[I(q)] vs* $q^2$) and a simulated SAXS profile from the protein structure [88].

## 2.3 Search space: sampling roto-translations and flexibility of protein subunits

In an aim to derive structural models of macromolecular assemblies, the experimental information described previously is integrated with high-resolution structures of the subunits composing the macromolecular complex. The structures of subunits are typically obtained from high-resolution experiments such as X-ray crystallography or nuclear magnetic resonance (NMR). When high-resolution experiments to characterise the structure of subunits are lacking, structural models can be computed from secondary structure prediction programs [89] or by homology to other know protein structures [90].

### 2.3.1 Rigid assembly

When predicting the assembly of symmetric protein assemblies in a rigid setting, the high-resolution structure of only one subunit is necessary to reconstruct the whole assembly. Thus, taking advantage of the circular symmetry, the coordinates of only one subunit need to be modified since the whole assembly can be reconstructed from circular rotations of that single subunit. The search space parameters consist of [$\alpha, \beta, \gamma, r$], where $\alpha, \beta, \gamma$ are the three Eulerian angles defining the subunit structure orientation according to the following transformation matrix:

$$R = R_z(\gamma) \cdot R_y(\beta) \cdot R_z(\alpha)$$

$$= \begin{pmatrix} cos(\gamma) & -sin(\gamma) & 0 \\ sin(\gamma) & cos(\gamma) & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} cos(\beta) & 0 & sin(\beta) \\ 0 & 1 & 0 \\ -sin(\beta) & 0 & cos(\beta) \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & cos(\alpha) & -sin(\alpha) \\ 0 & sin(\alpha) & cos(\alpha) \end{pmatrix},$$

and *r* is the radius of the symmetric assembly.

For asymmetric assemblies, the dimensions of the search space increase according to the number of protein subunits involved in the assembly. To reduce the search space, usually the biggest subunit is held rigid as a receptor while the other subunits used as ligand(s) are rotated and translated according to the parameters [$\alpha$, $\beta$, $\gamma$, $t_x$, $t_y$, $t_z$] where $\alpha$, $\beta$, $\gamma$ are the three Eulerian angles and $t_x$, $t_y$, $t_z$ are translations along the *x*, *y* and *z* axis.

### 2.3.2 Flexible assembly

The previous section described an ideal assembly case where the conformation of the unbound subunit structure was similar to that of the bound subunit. Such cases are however rare and most often one has to sample the flexibility of the subunits before attempting to model the assembly. In this case subunit flexibility can be sampled before optimization using molecular dynamics (MD) simulations, and used during optimization as a conformational ensemble added to the search space.

MD simulations aim to model the temporal evolution of an atomic ensemble through iteratively computing the forces acting on each atom, based on molecular mechanics force-fields. Briefly this is done through integrating Newton's equation of motion to determine the position and velocity of every atom. The force $\boldsymbol{F}$ acting on the ensemble of atoms at position $\boldsymbol{X(t)}$ and time *t* is computed as the negative gradient of the molecular mechanics potential *U(X(t))* (see below), as extracted from force fields, and is formulated as:

$$F\big(X(t)\big) = -\nabla U(X(\mathrm{t}))$$

This generic scheme is evaluated in an iterative manner and is commonly translated into the Verlet integration algorithm, which predicts the position $X_{n+1}$ and velocity $V_n$ of atoms at timestep *n*, which are defined as:

$X_{n+1} = 2X_n - X_{n-1} + \Delta t^2 M^{-1} \nabla U_n(X(t))$ and

$V_n = \frac{(X_{n+1} - X_{n-1})}{2\Delta t}$,

where $\Delta t$ is the size of the timestep (typically in order of 2fs for atomistic MD) and *M* the mass associated with the atoms of the system.

If sampling a molecular system for an infinite amount of time, the ergodic theorem states that the time average computed over a measurable quantity would be equal to the ensemble average of the phase space (velocity and position) of that system. Given that sampling a system phase space for an infinite amount of time is clearly unfeasible using molecular dynamics techniques, this hypothesis essentially allows to reduce the sampling to timescale long enough to observe a biologically relevant event, with the assumption that the ergodic theorem is satisfied.

## 2.4 Fitness: scoring the quality of the structural models

### 2.4.1 Molecular mechanics

To ensure physical plausibility of the structural models produced by *pow*[er], i.e. preventing steric clashes for instance, potential energy functions are used. The potential energy of such molecular systems can be calculated using classical molecular mechanics techniques from their atomic position. In this case, empirical models called force fields essentially capture the functional form and parameters describing the interaction between the atoms of the molecular system. The basic functional form describing the atomic interactions as found in common force fields can be formulated as follows:

$$U(r^N) = U_{bonded}(R^N) + U_{non-bonded}(R^N),$$

where,

$$U_{bonded}(r^N) = \sum_{i}^{bonds} U_{bond}(r_i) + \sum_{i}^{angles} U_{angle}(\theta_i) + \sum_{i}^{torsions} U_{torsion}(\phi_i) + \sum_{i}^{impropers} U_{improper}(\psi_i)$$

$$U_{non-bonded}(r^N) = \sum_{i}^{N}\sum_{j}^{N} U_{coulomb}(r_{ij}) + U_{vdw}(r_{ij})$$

The potential $U(r^N)$ is computed based on the position *r* of the *N* atoms of the molecular system. The bonded terms $U_{bonded}(r^N)$ is applied to covalently bonded atoms and the non-bonded term $U_{non-bonded}(r^N)$ to all atoms separated by at least 3 covalent bonds.

The bond stretching term $U_{bond}$ representing the displacements between 2 covalently linked atom and can be approximated as:

$$U_{bond}(r) = k_b(r - r_0)^2,$$

where $k_b$ is the force constant constraining 2 covalently linked atoms around their equilibrium distance $r_0$.

The angle potential $U_{angle}$ between 3 convalently linked atoms can be approximated as:

$$U_{angle}(\theta) = k_a(\theta - \theta_0)^2,$$

where $k_a$ is a spring constant constraining the 3 atoms around the equilibrium angle $\theta_0$.

The dihedral potential $U_{torsion}$ between 4 covalently bonded atoms can be approximated as:

$$U_{torsion}(\phi) = \sum_n k_d[1 + \cos(n\phi - \phi_0)],$$

where $k_d$ is a force constant constraining the 4 atoms around the equilibrium angle $\phi_0$ and *n* the dihedral multiplicity term defining the periodicity of the system minima (e.g. *n=1* when the periodicity of rotation is 360°, *n=2* when the periodicity of rotation is 180°, etc)

The improper torsional angle term $U_{improper}$ for an out-of-plane atom is used to enforce correct chirality and planarity and can be approximated as:

$$U_{improper}(\psi) = k_i(\psi - \psi_0)^2,$$

where $k_i$ is the force constant constraining 4 covalently bonded atoms around the equilibrium angle $\psi_0$.

The non-bonded term $U_{vdw}$ describes the van der Waals interaction between 2 non-covalently linked atoms or that are separated by at least 3 covalent atoms and can formulated as a simple 12-6 Lennard-Jones potential:

$$U_{vdw}(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right],$$

where *r* consists in the pairwise distance between two atoms within a distance usually truncated at 12 Å, $\sigma$ is the energy minimum defining the optimal distance separating the two atoms, and $\varepsilon$ the depth of the potential well.

The non-bonded term $U_{coulomb}$ describes the electrostatic interaction between 2 non-covalently linked atoms and is formulated as:

$$U_{coulomb}(r) = \frac{1}{4\pi\varepsilon_0\epsilon_r} \frac{q_1 q_2}{r},$$

where $q_1$ and $q_2$ are the charge of respective interacting atoms, $r$ is the distance separating the atoms and $\varepsilon_0$ the dielectric permittivity of vacuum.

In IM methods, the calculation of potential energy is usually one of the most computationally demanding process since it requires to be re-calculated at each iteration from the position of every atoms for each new structural model. In order to alleviate such computational cost, coarse-grained (CG) representation of macromolecular systems can also be used. Essentially, CG representation reduces and aggregates the information associated with groups of atoms into fewer beads of different chemico-physical properties [91]. The net benefit of using CG representation over all-atom systems includes an acceleration in potential energy computation and a decrease in local energy minima traps associated with a smoother energy landscape [92]. For these reasons, also the Martini force field for coarse grained modelling of proteins (version 2.6) [93] was incorporated into the *pow*$^{er}$ framework.

### 2.4.2 Integrative modeling of macromolecular assembly with *pow*$^{er}$

The aim of the current IM approach is to predict the structure of macromolecular assembly by integrating the information provided from low-resolution experiments describing the spatial connectivity between the subunits, sampling the roto-translational parameters of these subunits and by ensuring their physical plausibility.

Formally, the prediction of macromolecular assembly as stated above can be translated as an optimization problem. Let $f(x)$ be a function where $x \in R^n$. For an optimization problem, the fitness to be minimized is $f$ and $x$ a vector of design variables [$x_1$, $x_2$, …, $x_n$]. The aim of the optimization procedure is to find the optimal solution $x_{min} \in R^n$ that defines $y_{min} = f(x_{min})$, where the fitness function is at global minimum.

For homomultimeric assemblies featuring a circular symmetry, we want to find the optimal combination of rotations and translations set $x_{min}$ that will modify the subunit structures coordinates ($x_1$, $y_1$, $z_1$, … $x_n$, $y_n$, $z_n$) so that the fitness function describing the quality of the assembly $f$ is minimized.

The scoring term concurring to the fitness function $f$ consists in the pseudo-potential energy function that computes the satisfaction of experimental data, which are translated into spatial distances:

$$E_{data} = \sqrt{\sum_i^n (o_i - t_i)^2},$$

that combines the squared sum of the differences between target and observed spatial distances ($n$ is the numbers of geometric constraints, $o_i$ is the Euclidian distance of the constraint $i$ observed in the assembly model and $t_i$ its target distance value, as inferred from experimental data).

The potential energy term used to avoid steric clashes in the structural models is a simple 9-6 Lennard-Jones coarse energy potential [94] acting on the protein Cα-atoms in the form of :

$$E_{phys} = 4\varepsilon \left[ \left(\frac{\sigma}{r}\right)^9 - \left(\frac{\sigma}{r}\right)^6 \right],$$

where $r$ consists in the pairwise distance between the Cα interfacial atoms of the subunits within a distance of 12 Å, $\sigma$ = 4.7 Å and $\varepsilon$ = 1 kcal/mol).

One critical aspect of any optimization procedure is to design a fitness function that accurately describes the system. This implies that the direct minimization of the fitness function should ideally improve model quality. When different terms concurrent to the same fitness function are minimized simultaneously, imbalances may arise. Here, when minimizing $E_{data}$ and $E_{phys}$ simultaneously without balancing their relative contribution, such imbalances can lead to models where either the geometric constraints have been satisfied but include steric clashes, or inversely to physically plausible models not satisfying the experimental data.

## 2.5 Optimizers

### 2.5.1 Unconstrained optimization with particle swarm optimization

To balance the relative contributions of the fitness function components (e.g. $E_{data}$ and $E_{phys}$), these terms can be linearly combined with weight constants as in:

$$f(x) = \mathbf{w} * E_{phys} + (1\text{-}\mathbf{w})\, E_{data},$$

where ***w*** is the weight constant lowering the contribution of the energy component and which has been calibrated at 0.2 using a systematic analysis described in [16].

In $pow^{er}$, $f$ is minimized with the heuristic particle swarm optimization algorithm (PSO). The latter is a robust heuristic optimization algorithm inspired by nature. It was invented in 1995 by James Kennedy and Russel Eberhart, and since then has been applied to several fields including transport [95], geology [96], economics [97], and more particularly to integrative modelling of macromolecular assemblies[16,98].

PSO was originally designed to simulate the social behaviour as depicted by the movement of fish schools or bird flocks. Hence, it evolves a population of candidate solutions that are influenced by the position and fitness of the best individual and by neighboring solutions according to the following pseudocode:

**Algorithm 1**

```
for each timestep t do:
  for each particle p do:

    inertia ← w *v(p,t −1)
    personal ←cp*rand(0,1)*(x(p,t−1)−x_best(p))
        global ←cn*rand(0,1)*(x(p,t−1)−x'_best)

        v(c,t)←inertia+personal +global

        if|v(p,t)|≥size(space)then
                v(p,t)←norm(v(p,t))*size(space)
        end if

        x (p , t ) ← x (p , t − 1) + v (p , t )

        if f(x(p,t))≤f_best(p)then
          f_best(p)← f (x(p,t))
          x_best(p)←x(p,t)
        end if

  end for
end for
```

At initial time $t$, the velocity ***v*** and position ***x*** of each individual candidate solution (particle *p)* are randomised according to the boundaries specified by the search space. The fitness ***f(x)*** of every particle is evaluated based on the particle position ***x***.

Then, at each iteration *t*, for each particle, the velocity $\boldsymbol{v}$ and position $\boldsymbol{x}$ are updated based on the particle's best ever recorded position $\boldsymbol{x_{best}}$, which is the position where *f* was minimal at $f_{best}$, and the position $\boldsymbol{x'_{best}}$ of best neighboring particle. The velocity $\boldsymbol{v}$ is affected by three factors. One of them is the inertia w, which is essentially used to gradually slow down the velocity of the particle through the optimization as it decreases with every timestep *t*. The assumption behind the decrease of inertia is that as the optimization proceeds, interesting areas of the search space have been discovered and thus the velocity of particle should decrease so as to search these areas more carefully. The other two factors cp and cn consist in weighting constants used to balance the influence of the particle best position $\boldsymbol{x_{best}}$ and the best neighbor position $\boldsymbol{x'_{best}}$ on the particle new position at time *t*.

The termination criterion is reached whenever the budget of function evaluation has been used or whenever a maximum number of timesteps have been reached. Other criteria include reaching positional or fitness convergence.

The current PSO algorithm featured in *pow*$^{er}$ contains a modification to avoid local minmima and is called PSO *Kick and Reseed* (PSO-KaR). Briefly, at each iteration, PSO-KaR checks whether the velocity or the fitness of any particle is below a pre-defined threshold value and if true, then randomly reinitialised that particle's position and velocity. This approach was found efficient to increase the PSO sampling procedure. Particularly, when used to find the minimum of standard benchmark functions such as rastrigin or sine, PSO-KaR was able to find lower fitness minima than the original PSO as the dimensionality of the search space increased [16].

Importantly, *pow*$^{er}$ with PSO-KaR was found suitable to solve symmetrical assembly cases with and without the inclusion of protein flexibility. In the assembly cases without flexibility, the subunits were considered as rigid and realistic spatial constraints were extracted a priori and used to guide the assembly during optimization. These symmetric rigid assembly cases consisted in the solved multimeric assembly structures of the chorismate mutase (pdbid 1xho), acyl carrier protein synthase (PDB-id 1fth), SM archeal protein (PDB-id 1i8f) and Escherichia coli Escj (PDB-id 1yj7). Starting from a set of spatial constraints and a structure of one of the subunits, structural models were computed with *pow*$^{er}$ with PSO-KaR and structurally assessed against their respective multimeric assembly structure. The structural similarity metric used to compare the obtained models with solved multimeric assemblies was the root-mean-square-deviation:

$$rmsd = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2},$$

where, given two protein structures having an equivalent number of *N* atoms, δ is the distance between N pairs of equivalent Cα atoms. For all the assembly cases mentioned above, power with PSO-KaR returned solutions with a Cα-RMSD < 2.5 Å, which is considered acceptable.

### 2.5.2 Constrained optimization with memetic Viability Evolution

#### 2.5.2.1 Constrained optimization

Constrained optimization problems are different to unconstrained ones in that they contain one or several inequality constraints to guide the solutions towards desired regions of the search space. Constrained optimization problems can be formulated as follow:

$$\textit{min } f(x), \textit{ s.t. } \begin{cases} l_i \leq x_i \leq u_i, \ i = 1,2,\dots,n \\ g_j(x) \leq 0, \quad j = 1,2,\dots,m \end{cases}$$

where $f(x)$ is the objective function to be minimized, $x \in R^n$ is a vector of design variables [$x_1$, $x_2$, …, $x_n$], $l_i$ and $u_i$ are and respectively the upper and lower admissible boundary ranges defining the search space of each variable $x_i$, $i \in 1,2, \dots, n$, and $g_j(x)$ are the inequality constraints defined on each solution $x$. In this case, the aim of the optimization procedure is to find the optimal solution $x_{min} \in R^n$ that defines $y_{min} = f(x_{min})$, where the fitness function as at global minimum and where the best solution $x_{min}$ satisfies all inequality constraints so that $g_j(x_{min}) \leq 0$.

The macromolecular assembly problem was previously specified as an unconstrained fitness function $f(x) = \boldsymbol{w} * E_{phys} + (1-\boldsymbol{w}) E_{data}$. When redefining this problem in a constrained optimization perspective as for instance, $min\ E_{phys}$, $s.t.\ \{E_{data} \leq 0\}$, the net advantage of the redefinition is the loss of the weights constant $\boldsymbol{w}$ previously used to calibrate the relative contribution of $E_{phys}$ and $E_{data}$.

Such type of constrained optimization problems is considered difficult to solve using traditional evolutionary algorithms, due the typically large number of function evaluation needed to converge. Thus to face this challenge, a novel constrained optimizer called memetic viablity evolution (mViE) [83] was incorporated into the macromolecular assembly prediction protocol of *pow*$^{er}$ (see Chapter 3).

#### 2.5.2.2 Viability evolution theory

In order to optimize a given objective/fitness function, traditional heuristic evolutionary algorithms operate by selecting the candidate solutions with the best phenotypes (i.e. lowest objectives/fitness) and allowing them to reproduce by iteratively modifying their genotypes. Such fitness-based selection process is inspired from nature where competition has been determined as the main factor for driving the selection and reproduction of the fittest individuals. Traditional evolutionary algorithms relying on competition are usually suitable for optimization problems defined by a single objective that uniquely describes the problem. However, in cases where both constraints and objectives are present, a few methods have been implemented [99]. Moreover, inherent to competition-based evolutionary algorithm, by iteratively selecting individuals from a population of candidate solutions, there is gradual loss of genotype diversity which can be associated with early convergence to local minima [100].

In an attempt to alleviate these issue, the work by Maesani et al. [101] suggest "an alternative abstraction to artificial evolution" in which the concept of competition to select the best individuals is replaced by that of viability. Unlike fitness-based evolution, viability evolution (ViE) selects individuals based on a set of criteria defined on physiological or environmental factors [102], called viability boundaries[101]. Figuratively speaking, one can suggest that instead of competing against each other to reproduce, in ViE, individuals are selected for reproduction based on environmental factors that increasingly become harsher by constraining the individuals to increasingly smaller regions, where the minimum of the objective is found.

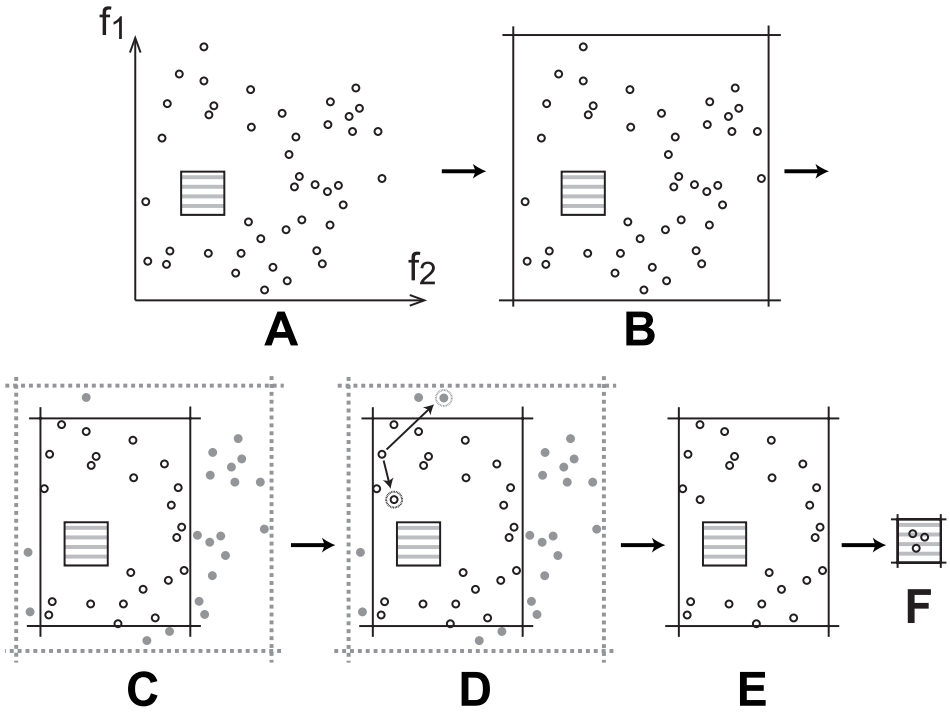



**Figure 2.2 | The viability Evolution (ViE) algorithm (adapted from [101]).**

**A.** Candidate solutions are randomly generated on a two-dimensional search space representing the viability boundaries defined on the inequality constraints f1 and f2 which satisfaction is illustrated by the grey square with horizontal lines. Practically the aim of the optimization is to slowly drive the solutions towards the grey-lined square where the constraints are satisfied. **B.** The viability boundaries are initially set and relaxed so as to encompass all the candidate solutions. **C.** The viability boundaries are tightened in both f1 and f2 toward the area where the constraints are satisfied. **D-E.** Solutions found outside the viability boundaries, generated either from mutations or reproduction, are eliminated. **F.** The process is repeated until the limits of the viability boundaries match those of the inequality constraints and solutions are obtained that satisfy these constraints.

In the context of optimization problems featuring both constraints and objectives, viability boundaries take the form of adaptive inequality constraints that are tightened at each iteration to drive the solutions towards feasible region of the search space where constraints are satisfied (Figure 2.2). Practically, at the beginning of a constrained optimization process, viability boundaries defined on the problem constraints are relaxed beyond the lower and upper limits of the constraints to encompass all the individuals from a population of candidate solutions. At the first iteration of the algorithm, new candidate solutions are generated from reproduction or mutations. The viability

boundaries are then tightened and all individuals not satisfying the new boundary limits are eliminated. Viability boundaries are then continuously tightened around a population of individuals forced to produce offspring or mutate in order to escape the ever constrained boundaries [101].

#### 2.5.2.3 Memetic Viability Evolution

Building from the theory of viability evolution, memetic viability evolution (mViE) [101] is a novel and efficient constrained optimizer recently implemented by Maesani et al. [83]. During optimization with mViE, the viability boundaries are applied to a population of search units that are based on Covariance Matrix Adaptation Evolution Strategy (1+1)-CMA-ES [103], which can be recombined together using a Differential Evolution (DE) [104] operators.

To date, the standard implementation of CMA-ES is considered as one of the most efficient heuristic optimization scheme for unconstrained problems featuring a single objective [105,106]. Briefly in the CMA-ES method, candidate solutions are generated according to a multivariate normal distribution, which variable number equals the number of search space dimensions and which mean is iteratively updated based on the search space position of the fittest solutions [105]. The dependency between the variables in the distribution is recorded through a covariance matrix that is iteratively adapted so as to increase the likelihood to produce better offspring (i.e. with better objective/fitness) [105]. Several flavours of the basic CMA-ES have been suggested to solve single-[105] multi-objective [107] or constrained optimization [108] problems. Of interest to this work, the standard CMA-ES termed ($\mu/\lambda$)-CMA-ES evolves a population of $\lambda$ individuals while the mViE algorithm features a variation termed (1+1)-CMA-ES, which essentially evolves one candidate solution at each iteration.

In this case, the local exploration of the search space is undertaken by the (1+1)-CMA-ES search units while more global search are performed during DE recombinations. Compared to CMA-ES, DE is simpler evolutionary algorithm that was proven to have fast and smooth convergence on multimodal problems, i.e. problems with several local optima [104]. DE operates by maintaining a population of candidate solutions and randomly recombining the parameters of three individuals into a new offspring with the aim to improve on the problem objective [104].

In order to optimally balance the local/global search operations, a specific scheduler was implemented in the mViE algorithms. A global overview of the mViE algorithm can be found in algorithm 2:

## Algorithm 2

```
Psucc_local ← 0.5
Psucc_global ← 0.5
C_α ← 0.1
C_β ← 0.05
initialize termination_criteria
initialize local_search_units()
relax_viability_boundaries(local_search_units)
evaluate_constraints_and_objective(local_search_units)

while termination_criteria not reached:
  decision ← component_scheduler ()

  if decision is excecute_local_search then:
    rank(local_search_units)
    best ← select_best(local_search_units)
    offspring ← generate_offspring(local_search_units)
    evaluate_constraints_and_objective(offspring)

    if offspring is better than best then:
      Psucc_local ← (1 − C_α) * Psucc_local + C_α
      best ← offspring

    else if offspring is worse than best then:
      if viability_boundaries not violated then:
        Psucc_local ← (1 − C_β) * Psucc_local
      else if viability boundaries violated then:
        Psucc_local ← (1 − C_α) * Psucc_local
      end if

    end if

  else if decision is excecute_global_search then:

    best ← select_best(local_search_units)
    parents ← random_selection(local_search_units)
    offspring ← recombine_random(parents)

    evaluate_constraints_and_objective(offspring)

    if offspring is better than parents then:

      if offspring is better than best then:
        Psucc_global ← (1 − C_α) * Psucc_global + C_α
        Psucc_local ← Psucc_global
        best ← offspring
      else if offspring worse than best then:
        Psucc_global ← (1 − C_α) * Psucc_global + C_β
      end if

    else if offspring worse than parents then:
      Psucc_global ← (1 − C_α) * Psucc_global
    end if
  end if

  update viability_boundaries(best)
  update termination_criteria
end while
```

The aim of mViE is to find solutions that satisfy the pre-defined constraints $g_j(\boldsymbol{x})$ and minimize the objective $f(\boldsymbol{x})$. To do this, it either deploys search units that perform local explorations of the search space or recombine the search units to explore the search space more globally. The decision on whether to advance either the local or global search operations is decided by the component scheduler function, which takes as input the parameters *Psucc_local* and *Psucc_global* probabilities as well as other parameters that keep count of the success of the local and global operations respectively.

At the beginning of the mViE algorithm, the population of search units is randomly initialized on the search space and the viability boundaries are relaxed to the maximal boundaries of the $g_j(\boldsymbol{x})$ constraints. At this time the probability of either advancing local or global search is equivalent. As the optimization proceeds, these values are updated based on the carefully calibrated fading parameters $C_\alpha$ and $C_\beta$. These values were calibrated so that the number of function evaluations are optimally spent between local and global search.

In case a local search is decided by the component scheduler through the variable *excecute_local_search*, an offspring is generated from the best ranked search units, in terms of objective and constraint satisfaction, and compared to the global best solution found. The idea is to penalise the offspring which fitness values and constraints satisfaction are worse than the ones of global best solution, whilst making sure the viability constraints are satisfied. Thus a higher penalty on the *Psucc_local* is assigned whenever not only the candidate offspring is worse than the best solution but also when it violates the viability boundaries.

In case of global search specified by the variable *excecute_global_search*, parents are first randomly selected from the local_search_units and an offspring is obtained from recombination of these parents. In this case the highest penalty on the probability of global search success *Psucc_global* is assigned whenever the offspring is worse than both the parents and the global best solution.

For both steps the viability boundaries defined on the constraints are tighthened on the succesful offspring according to the following :

$$b_j = \max\left(0, \min\left(b_j, g_j(y) + \frac{b_j - g_{j_y}(y)}{2}\right)\right),$$

where $b_j$ is the constraint boundary defined on the constraint $g_j$.

The mViE algorithm was shown to outperform state-of-the-art constrained optimizers on a standard set of difficult constrained optimization function [83], and thus was found suitable to be adapted to the prediction of macromolecular assemblies as shown in the next chapter.

# Chapter 3 Disentangling constraints using viability evolution principles in integrative modelling of macromolecular assemblies

Published as the following paper : "**Disentangling constraints using viability evolution principles in integrative modeling of macromolecular assemblies**" Giorgio Tamò, Andrea Maesani, Sylvain Traeger, Matteo T. Degiacomi, Dario Floreano, Matteo Dal Peraro. Scientific. Report. 2017.

## 3.1 Introduction

Macromolecular assemblies are of paramount importance for the functioning of biological cells. Due to their size and complexity, using traditional experimental methods to elucidate their structure and dynamics remains to date a daunting task and a major challenge. Nevertheless, if high-resolution structures of subunits as well as experimental low-resolution information describing their mode of assembly are available (e.g., describing residue contacts between protein subunits), these can be used to assemble the subunits into their native complex. Integrative modeling (IM) [10,43,109] is an *in silico* approach that integrates this experimental information with empirical energy potentials, as found in molecular force fields [110], to generate candidate model assemblies. Along with the recent advances in structural biology [111,112], IM has gained importance by successfully predicting several large molecular assemblies from their isolated subunit components [62,110,113-115].

The energetic and experimental components are usually aggregated into a fitness function that describes the quality of the candidate assemblies and that is minimized by a stochastic search [16,18,110,116]. In order to balance the contribution of the components, relative weights must be assigned [16,113]. The determination of optimal weights, besides being a tedious and computationally expensive process, can heavily influence assembly predictions [45,47]. A further and more general issue consists in the fact that a unique fitness function, notwithstanding its correct component balancing, is often inadequate to select the best candidate assemblies generated by IM protocols [10]. This inaccuracy results from the difficulty to correlate the quality of the candidate solutions to canonical scoring terms [117].

Here we report a widely applicable IM protocol based on an evolutionary method that does not require the individual weighting of the fitness function components. The proposed protocol is based on a novel evolutionary method [101], where candidate solutions can survive and reproduce if

they satisfy a set of viability criteria defined on the problem objectives and constraints. In particular, we adopt a variation of the viability method, named memetic Viability Evolution (mViE) [83], which maintains and recombines multiple sub-population in order to optimize the balance between local and global search and has been shown to outperform several state-of-the-art methods on a standard benchmark suite of constrained optimization problems [83,108] (see Methods). Therefore, rather than mixing objectives and constraints in a single fitness function, this method modifies independently the viability criteria for each objective and constraint during evolution, driving thus the solutions towards desired regions of the search space. Practically, this allows for a natural partitioning of the fitness function components. In the case of macromolecular assembly prediction, this means that fitness function components can be treated independently as objectives and inequalities representing constraints on the search space. The resulting assembly protocol based on mViE is featured as an extension of our protein assembly framework *pow*$^{er}$ (http://lbm.epfl.ch/resources) [16,62], and it is benchmarked here on an extended set of known symmetrical assemblies.

## 3.2 Results and discussion

### 3.2.1 Viability evolution applied to assembly prediction

We initially tested mViE on a benchmark set of symmetric assemblies where inputs are given as (i) high resolution structures of the subunits, (ii) realistic spatial constraints describing connectivity among subunits (as derived from cross-linking mass spectrometry experiments, e.g.), and/or (iii) volumetric density maps as obtained by cryo-Electron Microscopy (EM), e.g. (Figure 3.1a). During the prediction of plausible assemblies (Figure 3.1b), we used mViE to explore the assembly conformational space of subunits. At first, mViE attempts to find non-violating structural models, termed viable, by using the coarse energy potential as well as residue distances as viability boundaries. This is to ensure that structural models without steric clashes which satisfy the imposed distances of contacting interfacial residues are found (Figure 3.1b). Then, mViE attempts to maximize, as objective, the diversity of viable solutions. These solutions are clustered, fitted into their assembly density maps provided as input [24] and finally ranked according to a cross-correlation coefficient (*ccc*) value (Figure 3.1c and 2a**,** see Methods). Low-resolution volumetric maps or, in general, additional experimental inputs are not always available to rank the candidate solutions. Clustering algorithms can alone identify best solutions as they tend to predominantly sample regions of the search space associated to the most native-like assembly [118]. Moreover, multi-resolution energy scoring

functions can be applied to assess assembly predictions on the sole basis of intermolecular contacts [119], as for instance using recent machine learning protocols to discriminate true protein-protein interfaces from incorrect ones [120,121]. Eventually, final clustered solutions can be further refined using more sophisticated and computationally expensive techniques.

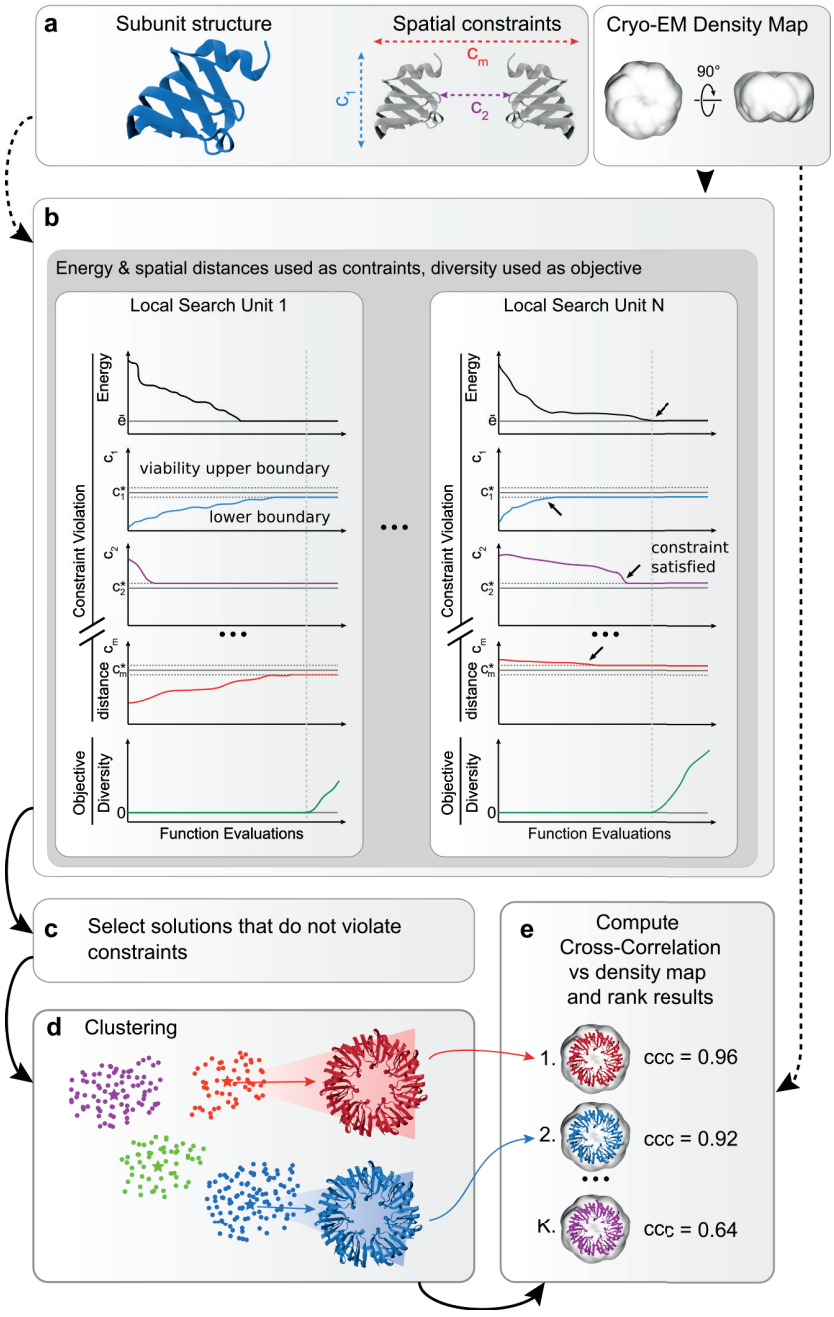


**Figure 3.1 | Assembly prediction using mViE algorithm.**

**a.** The protocol requires as input structures at atomic resolution of the subunits forming the assembly, a set of spatial constrains obtained by experiments that characterize the connectivity of the complex, and/or volumetric density maps providing information about the general complex architecture. **b.** mViE uses a population of multiple search units that try to independently minimize each constraint violation defined either as energy for the coarse energy potential or as $C_{\{1, 2, \ldots m\}}$ for the residue spatial distances. Once a search unit discovers assemblies that satisfy all the constraints, i.e. that are within the viability upper and lower boundaries defined on the constraints (black arrows on left panel), it tries to discover assemblies that maximize the diversity with respect to the assemblies predicted by the other search units. This is performed by favoring solutions that are at a larger Euclidean distance in the search space with respect to other local search units. **c.** Candidate assemblies that do not violate the constraints are (**d)** hierarchically clustered and the clusters centroids are extracted. **e.** The centroids are ranked on their cross-correlation coefficient (*ccc*) computed against the provided density map and the top-ranked predictions are returned to the user. Notice that density maps can be used *a priori* as objective or *a posteriori* for ranking.

### 3.2.2 Performance and versatility of mViE : switching constraints and objectives

In order to assess the feasibility and performance of this new protocol, we collected 18 symmetrical assemblies for which the multimeric conformation is already known, and compared the results with previously proposed methods including a Particle Swarm Optimization (PSO) *pow*$^{er}$ [16], Multifit [122] and SymmDock [123] (see Methods and Supplementary Table S1). For each of the assembly cases, EM maps at 15 Å resolution were synthetically generated and used to rank the best candidate solutions. Ranking by density maps was more efficient in extracting the best models compared to using solely the coarse energy potential, as demonstrated by a better correlation between relative model rmsd and *ccc* value (Figure3.2a and Supplementary Fig. S1 and Table S2). Remarkably, mViE is on par or better (in 13 out 18 protein prediction problems, mViE returned solutions with lower RMSD values than other prediction protocols, Figure3.2b and Supplementary Table S3) than existing IM protocols while it did not require an aggregated fitness function and, most importantly, the *a priori* identification of weights for the fitness components.

In the previous assessment, density map fitting was used as a post-processing step to rank the best assemblies returned by mViE. However, taking advantage of the flexibility of mViE, this kind of input can be used upfront as objective, while the interfacial residue distances and energy potential are still used as constraints on the search space. When applied to a selected set of complexes (i.e., GroEL, GTP-cyclohydrolase, and lymphokine complexes, see Online Methods) we could observe a significant improvement in term of solution quality ($P<0.05$, Wilcoxon-Rank-Sum-Test, Figure3.2c-e and Supplementary Fig. S2), demonstrating the flexibility of mViE in effortless-

ly handling additional and diverse components concurring to global fitness. Along the same lines, mViE was further tested on the problem of blindly docking protein subunits into density maps alone, i.e., without any additional distance constraint. We thus used solely the *ccc* value as objective to be maximized upon satisfaction of the energy potential as constraint. We found that mViE was able to successfully assemble complexes inside their density map for all the three selected assemblies (Cα RMSD < 2.5 Å, Figure3.2f and Supplementary Fig. S2).

To better understand how geometric constraints affect the outcome of mViE, a comprehensive benchmark was performed on the same selected targets (see Methods and Supplementary Fig. S3). We found that increasing the number of distance constraints improved the quality of solutions more directly than changing the accuracy of the constraints, ($P<0.05$, Jonckeere-Terpstra-Test, Figure3.2c-d and Supplementary Fig. S2). Changing density map resolution on the other hand seemed to have little effect on solutions quality. Specifically, the best solutions extracted with density maps at a resolution of 15 Å were not different ($P>0.05$, Wilcoxon-Rank-Sum-Test, Figure3.2e and Supplementary Fig. S2) than those ranked with a resolution of 25 Å.

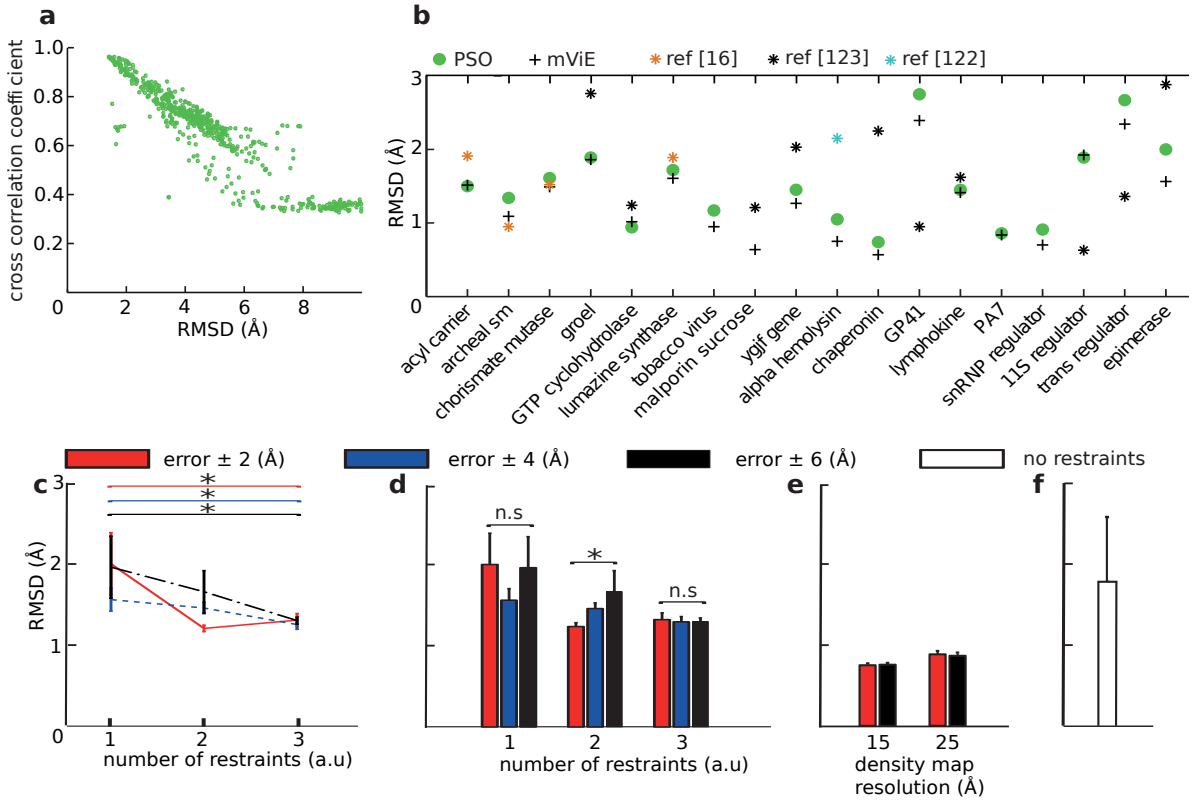


**Figure 3.2 | Performance assessment of mViE.**

**a.** Cross-correlation-coefficient (*ccc*) and backbone RMSD landscape of the lymphokine assembly models is shown as representative result. **b.** Performance comparison between mViE, PSO, and the best candidate assemblies of related protein predictions as reported in the literature[16,122,123] on 18 symmetrical protein assembly predictions. **c.** Effect of number of geometric spatial constraints on the quality of candidate assemblies returned by mViE on the GTP-cyclohydrolase homo 5-mer assembly problem (used as representative case, see **Supplementary Fig. S2** for extended data). Here the spatial constraints and potential energy were used as constraints on the search space, population diversity as objective and density map ranking as a post-processing step. **d**. Effect of the quality of spatial constraints (measured as error ± Å) on the quality of candidate assemblies returned by mViE on the GTP-cyclohydrolase homo 5-mer assembly problem. The assembly conditions were the same as in panel (**c.**) **e.** Results of mViE protocol using density map fitting during assembly of GTP-cyclohydrolase. The optimization was undertaken by using *ccc* as an objective to be maximized once the spatial and energy constraints are satisfied. **f.** Blind docking using mViE on the GTP-cyclohydrolase assembly problem. As input were provided a density map of 15 Å and one of the homo 5-mer subunit. The *ccc* was used as objective to be maximized during the optimization. Only the energy potential was used as constraint during the optimization procedure.

### 3.2.3 Assessing the quality of experimental constraints

While the assembly problems described so far were ideal cases where experimental constraints used as input were in fact correct, in reality ambiguity in the experimental constraints can arise, for example due to intrinsic limitations of techniques or multiple conformational states of the assembly. Unfortunately, in these cases correct and erroneous experimental constraints are difficult to discriminate *a priori* and are both used, possibly leading to incorrect models. A promising approach to solve this challenging problem has been recently proposed within a Bayesian framework [47,113,114]. Within our scheme, mViE search units naturally return solutions that maximize the number of constraints satisfied. Thus, if we realistically assume that the number of incorrect experimentally derived constraints never outnumbers the number of correct constraints, we can expect mViE to return solutions that satisfy a greater majority of correct constraints. To test this hypothesis, we synthetically increased the number of wrong constraints (see Methods and Supplementary Table S4), and could observe that for all the 3 selected cases, mViE was eventually able to return models satisfying a greater majority of correct constraints Figure 3.3.a-b). Therefore, when dealing with a large amount of experimental constraints, mViE represents a promising method to effectively maximize the probability to select native structural features while discarding erroneous or inconsistent experimental inputs during model prediction.

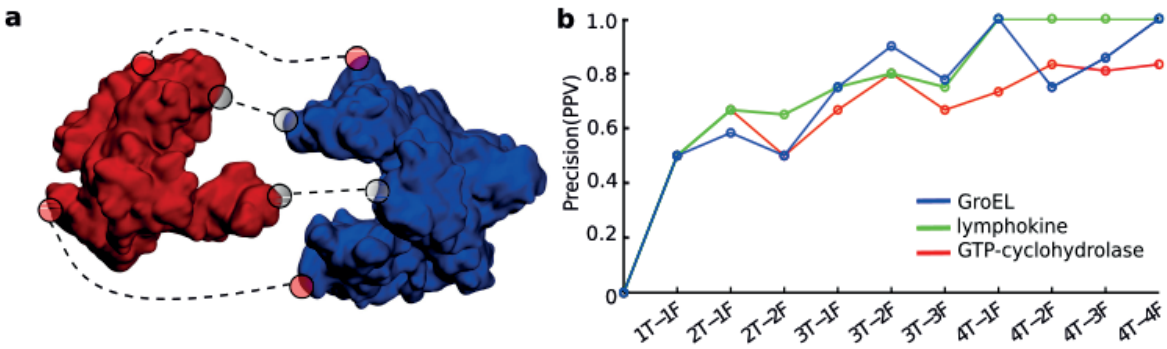



**Figure 3.3 | Detection of wrong constraints using mViE.**

**a.** Protein-protein interaction example where 4 interfacial residue contacts are used as inputs, in this case 2 of these constraints are correct (grey colored circles) and two are erroneous (red colored circles). **b.** Wrong constraints inclusion effect on returned model quality. The quality of the models was assessed by calculating the Precision/Positive Predictive Value [PPV=TP/(TP/FP)] representing how well good constraint (TP) were correctly satisfied in the returned models against wrongly satisfied "bad" constraints (FP). On the x-axis, the 1T-1F label means that 1 correct and 1 wrong spatial constraints were used during the optimization process.

### 3.2.4 Predicting the interface of the PhoQ periplasmic sensor

We finally applied mViE to the assembly of the PhoQ perisplasmic sensor [124]. In order to reconstruct the biological PhoQ homo-dimer interface, we used subunit model structures of PhoQ and a set of crosslinking fractional data as experimental inputs [124]. In this case we were able to directly model the PhoQ complex as a constrained optimization problem; in particular, the energy potential and the distance between Cb of the most highly cross-linked residues were used as constraints on the search space (see Methods). Since the distance between disulfide cross-linked residues is correlated to their degree of cross-linking [124], we used as objective the Pearson correlation coefficient ($R$) between the fractional cross-linking profiles and the Cb distances between cross-linked residues. The predicted PhoQ interface was in agreement with the crosslinking profile obtained experimentally (Pearson $R = 0.91$, $p<0.05$, Figure 3.4a-b). Compared to the original study where 774,165 models were generated [124] to find the optimal solutions, less than 3000 models were generated with mViE.

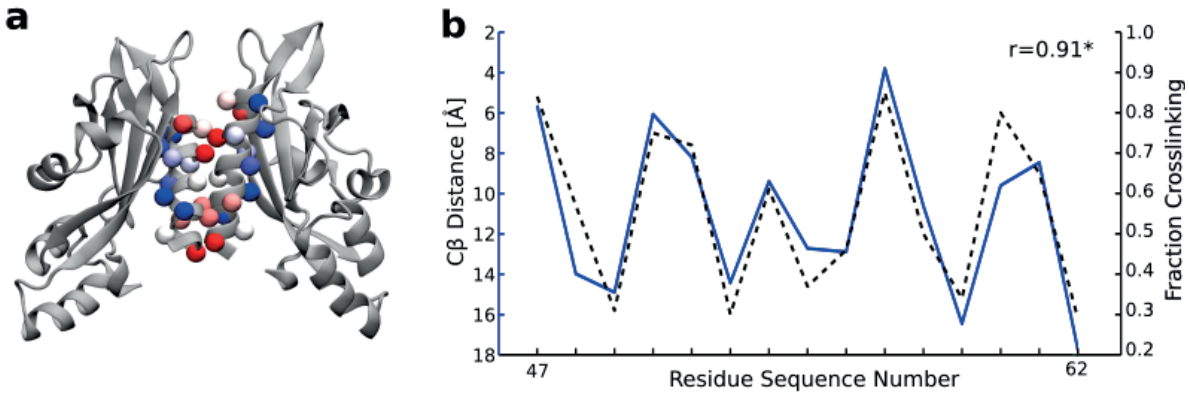



**Figure 3.4 | Assembly of the periplasmic sensor.**

**a.** Best model of the PhoQ periplasmic sensor obtained by mViE. The colors of residue Cb atoms represent the degree of crosslinking from highly (red) to poorly (blue) crosslinked. **b.** Fraction crosslinking vs. Cb distance for the best structural model of the PhoQ periplasmic sensor.

## 3.3 Conclusion

In conclusion, here we described and experimentally validated a new viability evolutionary algorithm that shows great flexibility and efficiency when applied to integrative modeling problems. This new method enables the independent treatment of experimental data without the need of identifying weights and combination of individual constraints into an aggregated fitness function. The method can thus handle a large and heterogeneous variety of experimental inputs related to molecular assemblies, and assess at the same time the quality of the predicted models and experimental inputs used in model prediction. This method could be extended to more challenging and general cases such as the prediction of large non-symmetric heteromultimeric complexes. The absence of symmetry would increase computational demand due to a larger number of dimensions to be explored, likely requiring more spatial constraints to help convergence toward predicted assemblies consistent with the initial inputs.

## 3.4 Methods and materials

***pow$^{er}$ framework for macromolecular assembly***. Please refer to section **Methods 2.4** for the details regarding the symmetric assembly optimization protocol featured in *pow$^{er}$* as well as the unconstrained single objective used to quantify the quality of the structural models.

***Prediction protocol based on viability evolution.*** mViE was incorporated into the *pow*$^{er}$ framework in order to model the prediction of symmetric assemblies as a constrained optimization problem without aggregating geometric distances ($E_{data}$) and energy potential ($E_{phys}$) in the same fitness function. In this case $E_{data}$ was used as a constraint on the search space to guide the assembly of the subunits. During optimization, the energy potential ($E_{phys}$) can be used in two different ways. On one hand, $E_{phys}$ can be used as an objective to be minimized once the $E_{data}$, used as constraints, have been satisfied in the form of:

$$min\ E_{phys},\ s.t.\ \{\ l_i \leq t_i \leq u_i,\ i = 1,2,\dots,n\ \}$$

where $l_i$ and $u_i$ are the lower and upper target boundaries of each of the $n$ contacting residue distances, respectively. On the other hand, $E_{phys}$ can be used together with $E_{data}$ as a constraint on the search space. In this case, maintaining $E_{phys} < 0$ would make sure that the assemblies are not only without steric clashes but also in close proximity.

Population diversity is a measure of how diverse the population of mViE search units are from a candidate solution $x$ generated any time during optimization. Maximizing this term during optimization would increase the diversity of candidate solutions by a better exploration of the feasible search space, i.e. toward regions of the search space where candidate solutions do not violate constraints. Diversity can be measured as:

$$diversity(x_t) = \frac{\sqrt{\sum_{i=0}^{n}(x_t - X(i)_{population})^2}}{n}$$

where $x_t$ is the parameter of a candidate solution generated at step $t$ of the optimization, $X$ the parameter of a mViE local search unit $i$ from a population of size $n$.

In the case where $E_{phys}$ and $E_{data}$ are used as constraints on the search space, population diversity can be used as objective to be maximized once the candidate solutions have satisfied the above constraints in the form:

$$max\ diversity,\ s.t. \begin{cases} l_i \leq t_i \leq u_i,\ i = 1,2,\dots,n \\ E_{phys} < 0 \end{cases}$$

***Ranking of candidate solutions with electron density maps.*** Due to the fact that a large number of solutions were generated during optimization, representative structures of the best models needed to be extracted. To do so, an integrated hierarchical clustering algorithm that selects centroids of clustered solutions with respective backbone root-mean-square-deviation (C$\alpha$-RMSD) lower than 1 Å was used. Input density maps of the reference structures were simulated using the SITUS [23] package command *kercon,* to rank the centroid of each cluster at resolutions 15 and 25 Å.

The centroids were then independently fitted in the simulated maps using the SITUS module *colores* and ranked according to a cross-correlation coefficient (*ccc*), which described the overlap between model and map. SITUS can fit structures into density maps with resolutions as low as 30 Å thanks to an intermediate step of Laplace transformation of the map that improves shape definition [23].

***Assessing the performance of mViE against PSO.*** In order to assess the performance of the mViE protocol against a previously published algorithm Particle Swarm Optimization (*PSO*), we chose 18 symmetrical complexes for which the atomistic structure has already been solved and that are available in the protein databank. As the evolutionary algorithms performing the search were stochastic, we repeated each test for 10 independent runs. Each execution could sample 20,000 candidate predictions before terminating. Information on the stoichiometry of the assemblies and the chosen interfacial spatial restraints can be found in the **Supplementary Table S1**.

As a general rule, a large number of interfacial spatial constraints leads to models of better quality. However, it has been observed that for our benchmarks set no more than three spatial constraints are necessary to efficiently guide the optimization process towards the assembly of realistic models [16]; hence the limitation of the number of spatial constraints generally to three per assembly prediction. In order to account for possible experimental noise on the target measures, an error of 2 Å was chosen per constraint, as previously done in ref [16]. Typical resolutions of macromolecular density maps range from 5 to 25 Å [122,125,126]; thus, an averaged resolution level of 15 Å was chosen for the evaluation mViE and *PSO*.

Evaluation of each method performance was achieved by computing the Cα-RMSD between the already solved reference structures and the best representative structures of the candidate solutions, ranked by ccc. For each prediction problem, the top 5 ranking solutions were returned by each

method as RMSD distances from the true assembly structure, resulting in 50 for each protein prediction problem per assembly protocol.

***Effect of spatial constraints on candidate solution assembly.*** With the aim to cover a broad spectrum of protein assembly cases, we chose 3 proteins complex of different size and stoichiometry. These were the lymphokine, GTP-cyclohydrolase and GroEL homo-multimers (Supplementary Figs. S3a-b**)**. These protein complexes were assembled under different conditions with the mViE protocol to assess how spatial constraints; both in quality and quantity, and density map resolution were affecting the quality of the candidate solutions. To address this task, 5 important contacting residues were identified within the interface of each of the 3 protein complexes and were chosen to serve as spatial restraints during optimization (Supplementary Fig. S3c**)**. For every optimization run the spatial constraints to be satisfied were randomly chosen in order to avoid bias toward preferred configurations and incremented gradually from 1 to 3 random restraints per run. To test for the effect of experimental error, the Euclidian distance for each of the spatial restraints was incremented from 2 to 6 Å. Finally, the density map resolution chosen to rank the best assemblies varied from 15 to 25 Å. In order to increase the significance of the statistics, 5 trials were performed for each condition. A total of 10,000 candidate predictions were generated per trial.

***Density map fitting as objective.*** Taking advantage of the fact that mViE does not require the computation of weights to balance the contribution of scoring function components, we directly used the *ccc* value during optimization instead of using it to rank the best models. Thus, for the three protein cases, the spatial as well as energetic terms of the traditional scoring function were used as inequality constraints on the search space. The *ccc* value was used as an objective to be maximized at each optimization step. Similarly to above, density map resolution and spatial constraint accuracy were varied from 15 to 25 Å and 2 to 6 Å respectively. In this case however, only one spatial constraint was randomly chosen for each optimization run, which was repeated in 5 trials.

In order to reduce computation time and keep the optimization in 4 dimensions [$\alpha, \beta, \gamma, x$], the SITUS package was used every 1000 steps to align the input density map with the representative structures of the candidate models generated by mViE. In the normal optimization steps, the generated models were then aligned to the best SITUS fitted structure and the *ccc* value was computed. Blindly docking proteins inside density maps is a more difficult problem because the only infor-

mation available consists of the density map of the fully assembled complex and the subunits structures. Thus for this problem, we did not use the residue contact information defined as spatial constraints as in the previous assembly cases. Instead, we used solely the information provided from the input density map of the complex at 15 Å and the energy potential as the 9-6 Lennard-Jones. In this case, the *ccc* was used as objective to be maximized and the energetic term as an inequality constraint ($E_{phys} < 0$). Moreover, we did not use the SITUS package to align the best structures to their input density map. Instead, any model generated was first translated to the center of the density map and assembled in 4 dimensions [$\alpha, \beta, \gamma, x$]. In the same optimization step, the fully assembled complex was roto-translated in 6 dimensions in order to fit the complex inside the density map. This amounted to a total of 10 parameters to be optimized by mViE. Due to the inherent difficulty of the problem, a budget of 100,000 functions evaluations was used to blindly dock each of the three protein cases, with 5 trials per run.

***Discerning good from bad experimental spatial constraints.*** During the optimization driven assembly process, the presence of wrong spatial constraints in the constraints dataset may lead to incorrect models. It is therefore important to assess how the inclusion of wrong constraints can affect the quality of the predictions returned by assembly protocols. To this end, we chose three protein structures, which were the lymphokine, GTP-cyclohydrolase and GroEL homo-multimers, in order to test how the inclusion of erroneous spatial constraints into mViE affected the quality of solutions. For each protein case, a total of 5 correct and 5 wrong restraints were selected. The correct constraints corresponded to true contacts and the wrong ones, which corresponded to non-native residue contacts, were each selected randomly across the interacting proteins surface (Supplementary Table S4).

Both types of constraints were used by mViE to assemble protein models. Based on the assumption that wrong constraints should never outnumber corrects ones, they were gradually added from 1 to 4 in a random fashion together with correct ones. At the end of each run, models satisfying the highest numbers of constraints were selected and assessed for how good they were at discriminating good from bad constraints according to:

$$PPV = \frac{TP}{TP+FP}$$

where PPV is the precision score, TP and FP are the number of true positive and false positive satisfied constraints respectively. In this case a TP corresponds to a good constraint being correctly satisfied and FP to a wrong constraint being incorrectly satisfied in a candidate assembly. A total PPV score of 1.0 implies that the candidate model assembly satisfied only correct constraints and inversely for a score of 0.0.

***Assembly of the PhoQ homo-dimer.*** In order to reconstruct the physiological perisplasmic sensor homo-dimer of the PhoQ two-component signaling system (TCS) from a structural model, disulfide scanning mutagenesis experiments were undertaken by [124]. In these experiments, cysteine substitutions were performed on 16 residues spanning the N-terminal helix of the periplasmic region. For each of these residues, the degree of crosslinking was reported in Supplementary Table S5. Given that the degree of crosslinking roughly correlates with Cb distance, this information could be used in the original study [124] to reconstruct the physiological complex from the assembly subunits using a rigid-body grid search where translation and rotations are applied to one of the subunits while maintaining the other fixed. After generating several candidate assemblies, the best models were selected by their absence of steric clashes and by their Pearson correlation (*R)* value characterizing the correlation between the percentage of cross-linked residues curve, and Cb distance from residues spanning the N-terminal helix of the models.

This problem was ideal to test our new protocol on a real biological case since it could be translated into a constrained optimization problem. Thus we used the same subunit model as used as in the study of [124]. Instead of using a grid search to sample the configuration of the homodimer interface, we used mViE. In the aim to get candidate assemblies resembling the physiological dimer, the Pearson value R between Cb distance and fractional crosslinking Cb distances was used as objective to be maximized upon satisfaction of the constraints. These were the Cb distance between the most highly cross-linked residue (Arg50, crosslinking degree 0.95, Supplementary Table S5) and potential energy to avoid steric clashes. The best model was extracted and minimized using CHARMM27 [127]. Our results were then compared with those obtained in [124].

***Statistical analysis.*** Due to the ordinal and non-parametric nature of the RMSD values describing the quality of the best solutions returned by *PSO* and mViE, non-parametric statistical tests were

used to evaluate significance of the results. In this paper, the Wilcoxon Rank Sum and Jonckheere-Terpstra tests were used.

### 3.5 Supplementary information

**Table S1. Protein prediction problems used to compare mViE and *PSO*.**

| Protein Name (PDBid) | Stoichiometry | Spatial restraint type | Target (+/- 2Å) |
|---|---|---|---|
| 11S regulator (1avo) | 7 | resid 196 - 181 | 6 |
| | | resid 203 - 118 | 10 |
| | | resid 221 - 133 | 8 |
| Acyl carrier (1fth) | 3 | width | 60 |
| | | height | 47 |
| | | resid 10 - 105 | 9 |
| Alpha hemolysin (7ahl) | 7 | resid 2 - 56 | 11 |
| | | resid 162 - 35 | 6 |
| | | resid 128 - 131 | 7 |
| Archeal sm (1i8f) | 7 | width | 65 |
| | | height | 37 |
| | | resid 29 - 29 | 4 |
| Chaperonin (1h5x) | 7 | resid 95 - 7 | 4 |
| | | resid 61 - 57 | 12 |
| | | resid 56 - 55 | 10 |
| chorismate mutase (1xho) | 3 | width | 49 |
| | | height | 44 |
| | | resid 74 - 74 | 4.5 |
| Epimerase (1eq2) | 5 | resid 85 - 34 | 5 |
| | | resid 142 - 39 | 7 |
| GP41 (1f23) | 3 | resid 26 - 46 | 8 |
| | | resid 11 - 63 | 15 |
| | | resid 2 - 71 | 9 |
| Groel (1oel) | 7 | resid 518 - 37 | 9 |
| | | resid 257 - 269 | 5 |
| | | resid 283 - 181 | 5 |
| GTP_cyclohydrolase (1fb1) | 5 | resid 241 - 243 | 9 |
| | | resid 224 - 134 | 6 |
| | | resid 183 - 126 | 7 |
| lumazine synthase (1ejb) | 5 | width | 77 |
| | | height | 46 |
| | | resid 103 - 103 | 14 |
| | | resid 103 - 107 | 6 |
| Lymphokine (1tnf) | 3 | resid 124 - 15 | 7 |
| | | resid 103 - 104 | 8 |
| malporin sucrose (1af6) | 3 | resid 197 - 18 | 5 |
| | | resid 81 - 66 | 8 |
| | | resid 58 - 361 | 7 |
| PA7 (1tzo) | 7 | resid 185 - 200 | 8 |

| | | | |
|---|---|---|---|
| | | resid 308 - 669 | 6 |
| | | resid 479 - 470 | 6 |
| snRNP protein (1h64) | 7 | resid 22 - 65 | 10 |
| | | resid 5 - 41 | 6 |
| tobacco virus (3kml) | 17 | height | 32 |
| | | resid 13 - 13 | 33 |
| | | resid 25 - all | -26 |
| trans regulator (1ny6) | 7 | resid 266 - 207 | 9 |
| | | resid 299 - 364 | 10 |
| yjgF gene (1qu9) | 3 | resid 109 - 112 | 7 |
| | | resid 72 - 21 | 6 |

**Table S2. Evaluation of ranking methods using density map cross-correlation coefficient and energy potential.**

The relationship between ccc and energy to RMSD respectively was computed using Pearson's correlation. The r-value indicates the level of correlation between RMSD and ranking method, in this case a value of 1.0 indicates a perfect correlation. The superscripts indicate * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ using a Pearson correlation test, *n.s.* no statistical significance.

| | ccc | | energy | |
|---|---|---|---|---|
| **protein predictions** | **r-val** | **p-val** | **r-val** | **p-val** |
| acyl_carrier | 0.925 | *** | 0.003 | n.s |
| archael_sm | 0.907 | *** | 0.028 | n.s |
| chorismate_mutase | 0.072 | n.s | 0.132 | n.s |
| groel | 0.928 | *** | -0.01 | n.s |
| GTP_cyclohydrolase | 0.915 | *** | 0.122 | n.s |
| lumazine_syntase | 0.869 | *** | 0.003 | n.s |
| malporin_sucrose | 0.875 | *** | 0.035 | n.s |
| tobacco_virus | 0.74 | *** | 0.03 | n.s |
| yjgF_gene | 0.722 | *** | 0.088 | n.s |
| alpha_hemolysin | 0.973 | *** | 0.044 | n.s |
| chaperonin | 0.652 | *** | 0.024 | n.s |
| GP41 | 0.03 | n.s | -0.081 | n.s |
| lymphokine | 0.803 | *** | 0.023 | n.s |
| PA7 | 0.913 | *** | 0.026 | n.s |
| snRNP_protein | 0.764 | *** | 0.042 | * |
| 11S_regulator | 0.868 | *** | 0.108 | * |
| trans_regulator | 0.649 | *** | 0.037 | * |
| epimerase | 0.6812 | *** | 0.0137 | n.s |

## Table S3. Performance comparison between mViE, PSO and other IM protocols.

The best models extracted with mViE were compared to the best models obtained from an earlier implementation of *pow*$^{er}$ [a], SymmDock[b] and Multifit[c] as extracted from the literature. Values in bold format indicate the assembly protocol that obtained the most native resembling models.

| | **Best Models RMSDs (Å-Cα)** | | | |
|---|---|---|---|---|
| **protein predictions** | ***mVie*** | ***PSO*** | **Literature** | **(ref)** |
| acyl carrier | **1.50** | **1.50** | 1.91 | a |
| archeal SM | 1.09 | 1.34 | **0.95** | a |
| chorismate mutase | **1.49** | 1.61 | 1.52 | a |
| groel | **1.86** | 1.89 | 2.76 | b |
| GTP cyclohydrolase | 1.01 | **0.94** | 1.24 | b |
| lumazine synthase | **1.60** | 1.72 | 1.89 | a |
| tobacco virus | **0.95** | 1.17 | - | - |
| malporin sucrose | **0.64** | 7.23 | 1.21 | b |
| ygjf gene | **1.27** | 1.45 | 2.03 | b |
| alpha hemolysin | **0.75** | 1.05 | 2.15 | c |
| chaperonin | **0.57** | 0.74 | 2.25 | b |
| GP41 | 2.40 | 2.75 | **0.95** | b |
| lymphokine | **1.41** | 1.45 | 1.62 | b |
| PA7 | **0.84** | 0.86 | 3.17 | b |
| snRNP protein | **0.70** | 0.91 | 3.44 | b |
| 11S regulator | 1.92 | 1.89 | **0.63** | b |
| trans regulator | 2.34 | 2.67 | **1.36** | b |
| epimerase | **1.56** | 2 | 2.88 | b |

**Table S4. True and False spatial constraints selected for the Lymphokine, GTP-cyclohydrolase and GroEL assembly cases.**

Carbon-$\alpha$ distances between the residues of the subunits are expressed in Å.

| | True restraints | | | False restraints | | |
|---|---|---|---|---|---|---|
| **protein name** | **Subunit 1** | **Subunit 2** | **Distance[Å]** | **Subunit 1** | **Subunit 2** | **Distance[Å]** |
| Lymphokine | Phe124 | His15 | 7 | Arg31 | Ala145 | 36 |
| | Glu103 | Arg104 | 8 | Gln102 | Asn34 | 33 |
| | Glu116 | Lys98 | 13 | Lys98 | His15 | 15 |
| | Lys98 | Glu116 | 6 | Gly54 | Val150 | 22 |
| | Tyr119 | Tyr119 | 8 | Gly122 | Lys112 | 30 |
| GTP Cyclo. | Lys220 | Asp136 | 6 | Ser228 | Ser166 | 18 |
| | Glu183 | His126 | 7 | Pro238 | Phe122 | 20 |
| | Lys224 | Asp134 | 6 | Leu82 | Glu61 | 30 |
| | Leu245 | Leu247 | 7 | Arg216 | Lys93 | 32 |
| | Arg241 | Glu243 | 9 | Als208 | Glu243 | 18 |
| GroEL | Glu518 | Asn37 | 9 | Gly256 | Phe44 | 14 |
| | Glu257 | Lys272 | 11 | Asp5 | Lys34 | 24 |
| | Glu255 | Lys207 | 8 | Asp359 | Gly459 | 46 |
| | Arg197 | Glu386 | 14 | Glu255 | Val387 | 19 |
| | Asp283 | Thr181 | 5 | Glu76 | Glu209 | 22 |

**Table S5. Related to Materials and Methods section; Experimental crosslinking data adapted from residues 47 to 62 spanning the N-terminal helix of the periplasmic sensor domain 3BQ8.**

The theoretical values below are adapted computed using equation [1] found in the work by Goldberg et al. 2008.

| | Degree of Crosslinking | |
|---|---|---|
| **Residue** | **Experimental** | **Theoretical** |
| T47 | 0.69 | 0.84 |
| T48 | 0.38 | 0.57 |
| F49 | 0.32 | 0.31 |
| R50 | 0.95 | 0.75 |
| L51 | 0.92 | 0.72 |
| L52 | 0.81 | 0.3 |
| R53 | 0.5 | 0.61 |
| G54 | 0.85 | 0.82 |
| E55 | 0.09 | 0.37 |
| S56 | 0.34 | 0.46 |
| N57 | 0.81 | 0.85 |
| L58 | 0.71 | 0.5 |
| F59 | 0.28 | 0.34 |
| Y60 | 0.92 | 0.8 |
| T61 | 0.48 | 0.65 |
| L62 | 0.12 | 0.29 |

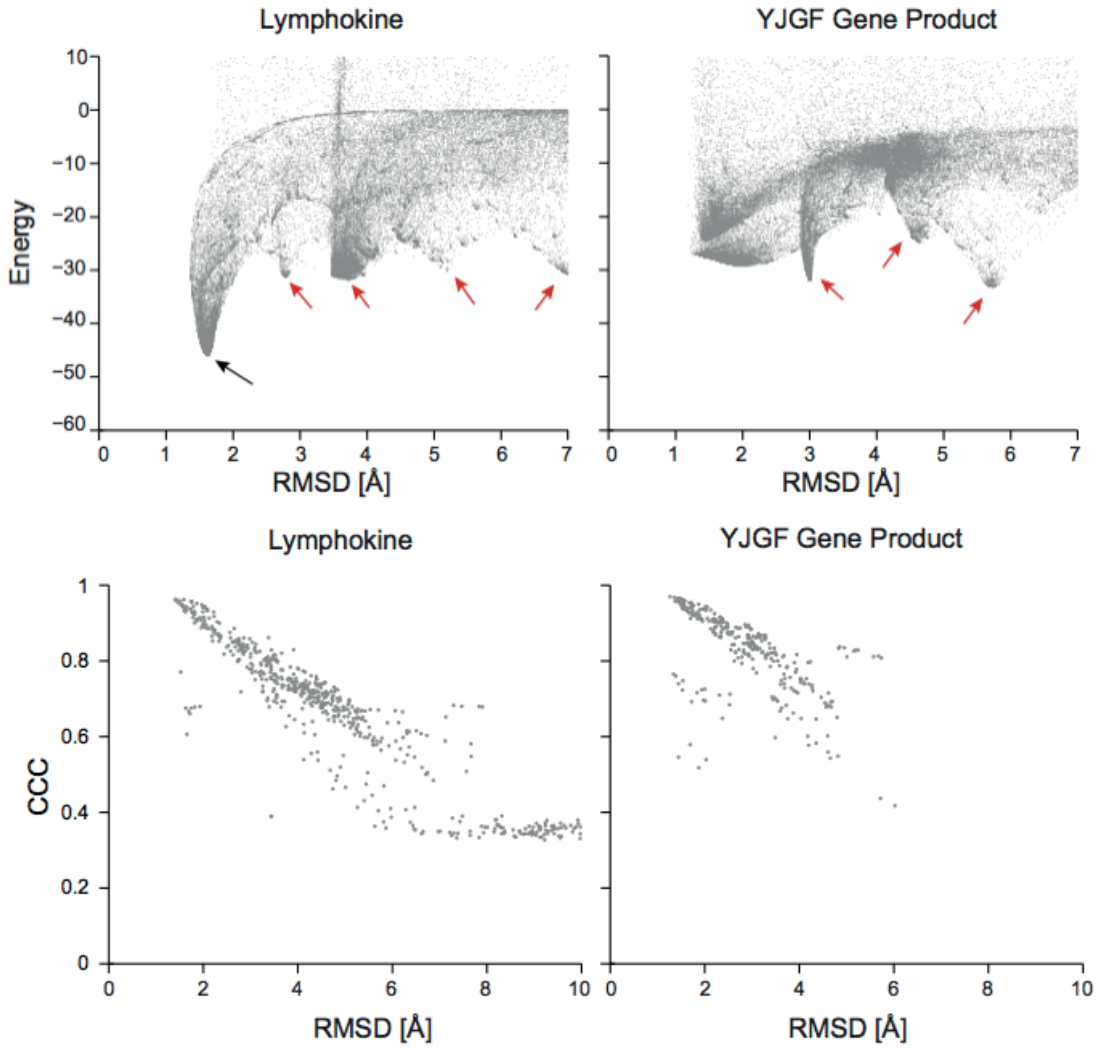



**Figure S1. Assessment of the effect of ccc/energy on backbone RMSD of the model assemblies.** **(Top Panel)** The symmetrical assembly models of the Lymphokine and YJGF gene product were evaluated for a relationship between backbone RMSD and energy as defined by a Lennard-Jones potential. Following the energy gradient may help reaching search areas were assemblies having minimal RMSD are found (black arrow). However, the presence of energy wells can mislead an optimization method (red arrows). In some cases, as in the YJGF gene product, this is even more troublesome as energy wells corresponding to configurations with higher RMSD have the lowest energy. **(Bottom Panel)** In contrast, there was for the same assembly cases a more correlating relationship between RMSD and cross-correlation-coefficient computed against Cryo-EM volumetric maps.

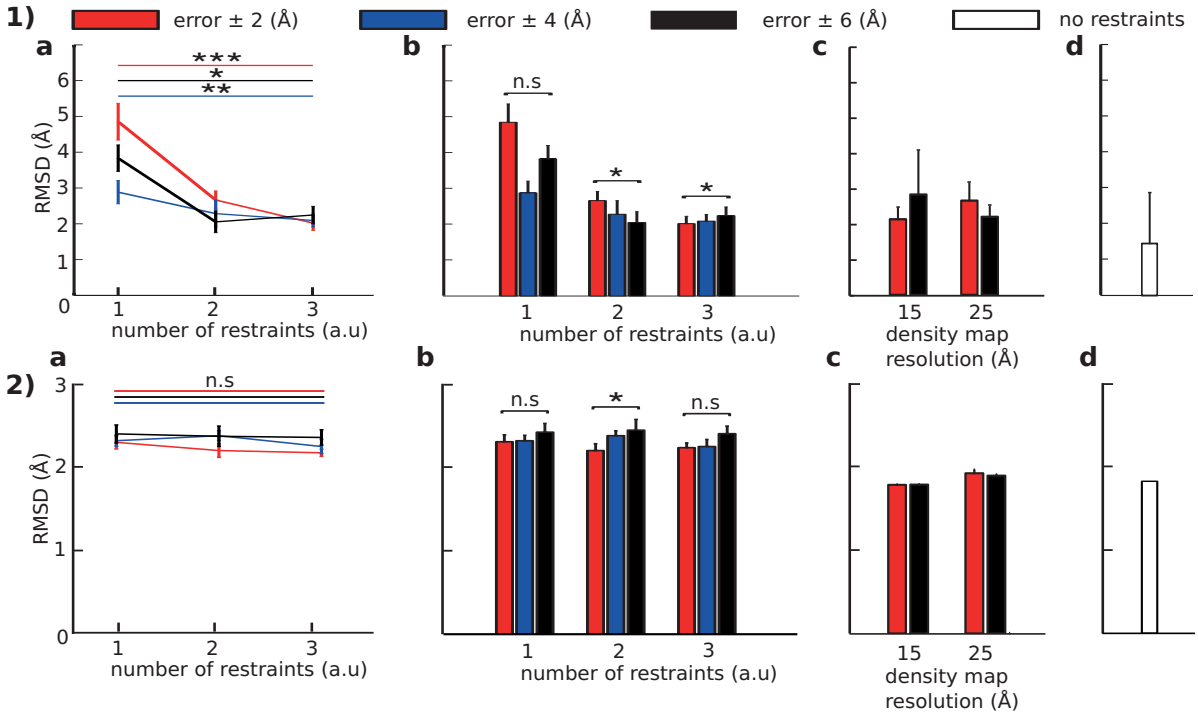
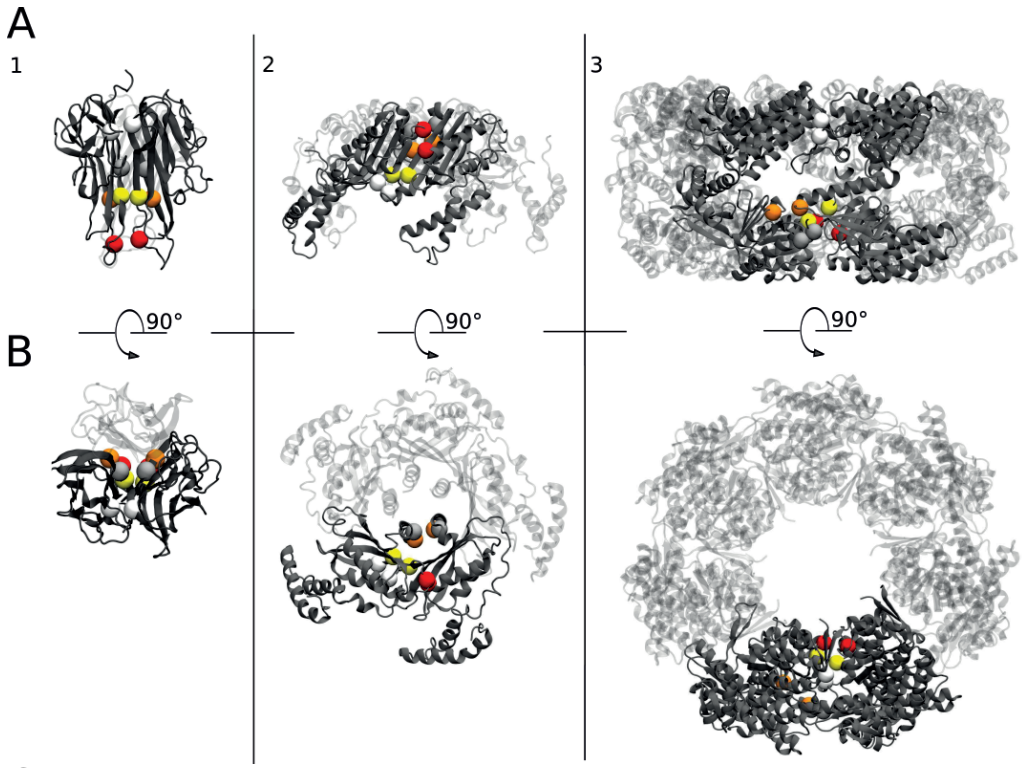



**Figure S2. Performance assessment of mViE on the GroEL (1) and Lymphokine (2) assembly cases.**

**a.** Effect of number of geometric spatial constraints on the quality of candidate assemblies returned by mViE. Here the spatial constraints and potential energy were used as constraints on the search space, population diversity as objective and density map ranking as a postprocessing step. For the assembly cases of GroEL, increasing the number of spatial restraints had a more significant and drastic effect on the quality of constraints than in the case of the Lymphokine case **b**. Effect of the quality of spatial constraints (measured as error ± Å) on the quality of candidate assemblies returned by mViE. The assembly condition were the same as in **a**. In this case of Lymphokine and GroEL, increasing the quality of spatial restraints had little effect on the quality of the assemblies returned by mViE **c.** Results of mViE protocol using density map fitting during assembly of GroEl (1) and Lymphokine (2). The optimization was undertaken by using *ccc* as an objective to be maximized once the geometry and energy constraints are satisfied. **d.** Blind docking using mViE on the GroEL and Lymphokine assembly problems. As input were provided a density map of 15 Å and one of the multimer subunits. The *ccc* was used as objective to be maximized during the optimization. Only the 9-6 Lennard Jones energy potential was used as constraint.




C

| Protein Name | Interaction Color | Interacting residue ids | | Interacting residue name | |
|---|---|---|---|---|---|
| | | subunit 1 | subunit 2 | subunit 1 | subunit 2 |
| Lymphokine [pdb 1tnf] | white | 124 | 15 | Phe | His |
| | red | 103 | 104 | Glu | Arg |
| | yellow | 116 | 98 | Glu | Lys |
| | orange | 98 | 116 | Lys | Glu |
| | gray | 119 | 119 | Tyr | Tyr |
| GTP cyclo. [pdb 1fb1] | white | 220 | 136 | Lys | Asp |
| | red | 183 | 126 | Glu | His |
| | yellow | 224 | 134 | Lys | Asp |
| | orange | 245 | 247 | Leu | Leu |
| | gray | 241 | 243 | Arg | Glu |
| Groel [pdb 1oel] | white | 518 | 37 | Glu | Asn |
| | red | 257 | 272 | Glu | Lys |
| | yellow | 255 | 207 | Glu | Lys |
| | orange | 197 | 386 | Arg | Glu |
| | gray | 283 | 181 | Asp | Thr |

**Figure S3. Circular assembly cases used for the performance assessment of mViE. A-B.**
Assembly cases of the Lymphokine (1), GTP-Cyclohydrolase (2) and GroEL (3). **C.** Important interfacial residues chosen as spatial restraints during the assessment performance of mViE and corresponding color codes (See Methods).

# Chapter 4 Structural analysis of Huntingtin to reveal the link between protein flexibility and disease

## 4.1 Introduction

Huntington's disease (HD) is a neurodegenerative disorder that equally affects approximately 1 in 10'000 men and women of European descent [128], with devastating life consequences. It is an inherited autosomal dominant disease which causes progressive damage to the brain and which resulting symptoms include memory loss, impaired locomotion, and eventually death. Unfortunately, still today, no therapeutic treatment exists to cure or even slow down the progression of the disease. Similar to other notorious neurodegenerative disorder such as Alzheimer's or Parkinson's disease, evidence suggests that HD's symptoms are due to the systematic death of neuron located in the basal ganglia [129]. While the cellular mechanisms underlying the cause of the disease are still unknown, molecular studies suggest differences in a gene encoding for Huntingtin protein (Httn). In particular, compared to healthy individuals, patients suffering from HD typically have a poly-expansion of CAG trinucleotide repeat at the beginning of the gene, which is translated as a large poly-glutamine segment at the N-terminal region of the Httn [67]. Interestingly, the length of the poly-glutamine tract was linked to the severity and onset of the disease, with a general acceptance that a poly-Q expansion above 36 repeats was associated with the disease through a "toxic gain of function"[129] (Figure 4.1).

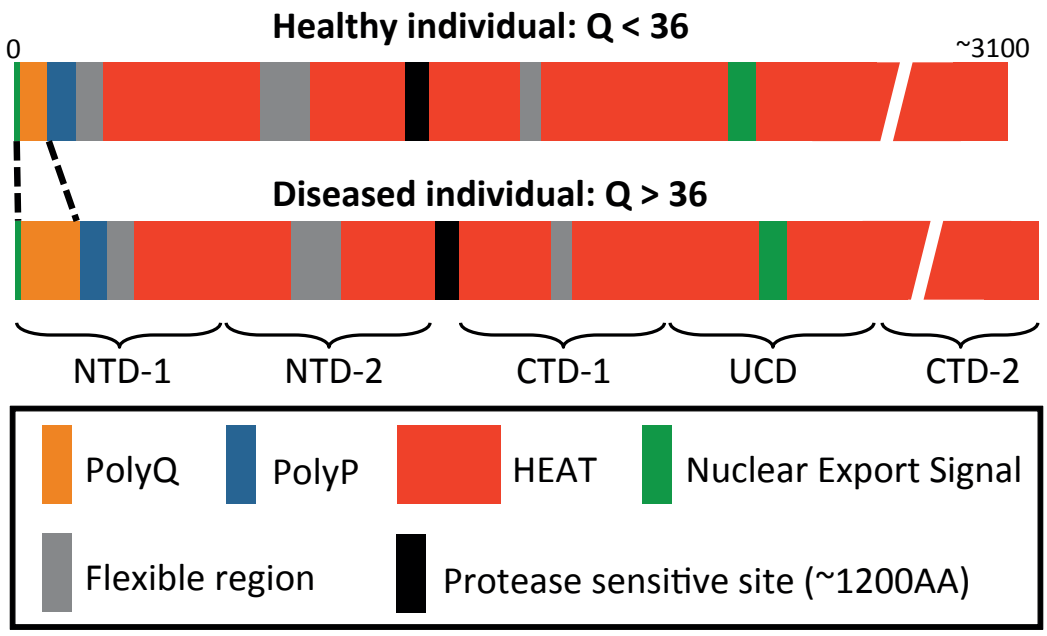



**Figure 4.1 | Schematic representation of Huntingtin main structural elements.**

Huntingtin is divided into five segments (NTD-1, NDT-2, CTD-1, UCD and CTD-2) based on the location of protease sensitivity sites as described in the study by [130]. The Huntingtin structure is composed mainly of HEAT repeats elements which are separated by flexible regions. The first ~100 amino-acids of the N-terminal region of the protein are characterised by the presence of an extended poly-proline (poly-P) and poly-glutamine (poly-Q) region. The extent of poly-Q expansion (poly-Q > 36) has been linked to the apparition of Huntington's disease symptom.

For convenience and ease of reference, the Httn protein sequence was divided in five subdomains consisting in two amino-termini (NTD-1 and NTD-2) and three carboxyl termini (CTD-1, UCD and CTD-2), as previously referred to by [67], and which were defined based on the location of protease sensitive sites [130] (Figure 4.1). The wild type (Q < 36) Httn is a large soluble protein (~350kDa, ~3140 amino acids) that is ubiquitously found in the cytoplasm of cells and which function has been associated with membrane vesicles and organelles [84,85], and membrane trafficking [86]. When mutated (Q > 36), Httn was observed to form aggregation bodies in neurons. Although not determined as the root cause for toxicity, Httn aggregation was found to be associated with the disruption of several important cellular function including the ubiquitination and endosomal degradation pathway, induction of apoptosis and $Ca^{2+}$ signaling [129,131-133]. Despite the fact that several studies have been conducted as an attempt to understand the causal link between poly-Q expansion and toxicity, still very little is known to date.

Also lacking to date is structural information related to the full Httn protein, which is believed to the important for understanding toxicity and subsequently to develop potential therapies. Despite a few studies that solved, via X-ray crystallography, the atomistic structure of first Httn N-terminal amino acids [131,134], most of what is currently known about the structure comes from comparisons between Httn and other proteins sharing sequence similarities [135-137]. Importantly, such structural comparisons have suggested the presence of HEAT (huntingin with elongation factor 3, p65 regulatory A subunit of protein phosphatase 2A and TOR1) domains, which are composed of tandem arrays of Helix-turn-Helix arranged in long solenoid structures. HEAT domains were predicted to be important for cellular function by mediating protein-protein interactions [138].

Although high-resolution experiments revealing the atomistic structure of the Httn are lacking, low-resolution experiments such those undertaken by our collaborators in the Song's lab (KAIST) [67], have recently shed light on the global structure of Httn. Notably in their work, cryo-electron microscopy (cryo-EM) combined with crosslinking mass spectrometry (CLMS) experiments have revealed that Httn folds in a hollow spherical solenoid structure with its C-terminal tail contacting the rest of the protein [67]. As an extension of this remarkable study, within our collaboration they recently obtained single-particle cryo-EM maps of ~7 Å resolution at different poly-Q expansion length (Q21, Q23 and Q78), together with CLMS data indicating the intra-Httn residue contacts. When analysed in isolation, such type of low-resolution experiments can only provide coarse details such the global shape of the protein or the general arrangement of its subunit constituents [110]. However, when integrated with IM approaches greater resolution can be reached.

In the true spirit of integrative modeling, we applied in this work the *pow*$^{er}$ framework to elucidate the structure of the Httn protein. To address this challenging task, we combined the information provided from recently obtained experimental data, which consisted in Httn cryo-EM maps and respective CLMS data, together with atomistic models of the Httn α-helices arranged in HEAT-like super structures. The sequence and position of the α-helices were obtained by applying the secondary structure prediction webservers PSI-pred [89] and J-pred [139] on the amino sequence of full length Httn. Each of these helices was then modeled into atomistic structure using MODELLER [90]. We took advantage of our recently implemented constrained optimization protocol *pow*$^{er}$ to flexibly fit the α-helix models into the provided cryo-EM maps while simultaneously satisfying the spatial constraints described in the CLMS experiments. This allowed us to produce realistic atomic structures of the Httn, which we believe will help understanding the causal association between mutated

Httn and HD. Notably, while writing this thesis a high-resolution cryo-EM structure of Httn was solved [87], and its implication for the present work are discussed in section 4.5.

## 4.2 Experimental data

The experimental information integrated by *pow$^{er}$-mViE* to model the structure of Httn was obtained directly from our collaborators in the song lab (KAIST). These were cryo-EM data of the Q23- and Q21-Httn as well as crosslinking-mass-spectrometry (CLMS) data.

### 4.2.1 Cryo-EM

#### 4.2.1.1 Sample and grid preparation (undertaken by Taeyang Jung in Song lab, KAIST)

FLAG-tag Httn was expressed from pALHD (Q2,23,46,67,78) in the Baculovirus Expression system (Invitrogen). The Sf9 cell lysate, obtained by freezing/thawing in buffer A (50 mM Tris-HCl pH 8.0, 500 mM NaCl, and 5% glycerol) containing complete protease inhibitor cocktail and PhosSTOP phosphatase inhibitor cocktail (Roche Applied Science), was spun at 15,000 rpm (2 hours). The supernatant was incubated with M2 anti-FLAG beads (Sigma) (2 hours, 4 °C). The non-specifically bound proteins were removed by washing extensively with buffer A. To obtain Httn with reduced global phosphorylation, M2-bead bound Httn proteins were subjected to calf intestinal phosphatase (CIP) treatment (in 1X NEB buffer 3 and 1U of enzyme per 10 μg huntingtin) at room temperature for 1-2 hours. CIP was removed by washing extensively with buffer A. FLAG-Httn was eluted with a buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 5% glycerol) containing 0.4 mg/ml FLAG peptide and loaded onto a calibrated Superose 6TM 10/300 column, equilibrated with 50 mM Tris-HCl pH 8.0 and 150 mM NaCl. FLAG-Httn eluted discretely and was estimated to be at least 90% pure by Coomassie staining. Full-length FLAG-tag Httn were subjected to ultracentrifugation at 33,000 rpm for 16 hours with 5-20% sucrose gradient in presence of 0-0.2% glutaraldehyde gradient. A fraction containing only the monomer Httn was collected and the protein was concentrated to above 0.5 mg/ml. For cryoEM, 3.5 ul of sample loaded on the R2/1 300 mesh quantifoil grids covered with Graphene oxide. The grids were vitrified by plunge freezing in ethanol using a vitrobot II with 8 seconds blotting time. The grids were then moved to liquid nitrogen and loaded into the autoloader of Titan Krios microscope with a Gatan K2 Summit direct detector.

#### 4.2.1.2 Data acquisition

The images of Q23-Httn with Volta phase plate (Q23-Httn-VPP) were taken in the Diamond Light Source in UK and Q23-Httn without VPP was taken in Scilife Lab in Sweden. For all data set, the calibrated magnification was 130,000 and it was corresponding to 1.06 Å/pixel. Images were collected at a dose of 3 ~ 4 Electrons per A2 in 6 ~ 8 seconds with 32 to 40 sub-frames. The data was automatically collected with EPU software and focusing was performed next to each hole with drift protection. The target defocus set to -500, and -700 nm for VPP images, and -1.5 to -4.0 for the ones without VPP images. The VPP was replaced to a new position after every 60 images. The gain images were applied while recording movie frames.

#### 4.2.1.3 Image processing

Movie frame stacks were imported into Scipion software to have access to various program suited for further image processing. MotionCor2 used for beam induced motion correction and dose weighted and un-dose weighted micrographs were generated. The CTF estimation was done by using CTFFIND4, and option for detecting extra phase shift was applied on VPP images. Good images were selected based on estimated CTF resolution (<6 Å), defocus range (for normal images, < 4.0, > 1.5 nm, and for VPP images, < 1.0, and > 0.2 nm), and the thickness of graphene oxide. The particle picking was performed with RELION autopicking using small sets of 2D template generated from manually picked 15k particles. The particles were extracted from dose-weighted micrographs with 240 x 240 pixel box size and the contrast was inverted for further image processing. The particle sets were cleaned by 4 rounds of reference free RELION 2D classification, and only particles in good 2D classes were subjected to build 3D initial model using RELION 3D model. Then 3D classification and 3D auto refinement were performed in RELION for final density maps. The overall resolution of Q23-Htnn and Q23-Httn-VPP were, 8 Å and 11 Å, respectively.

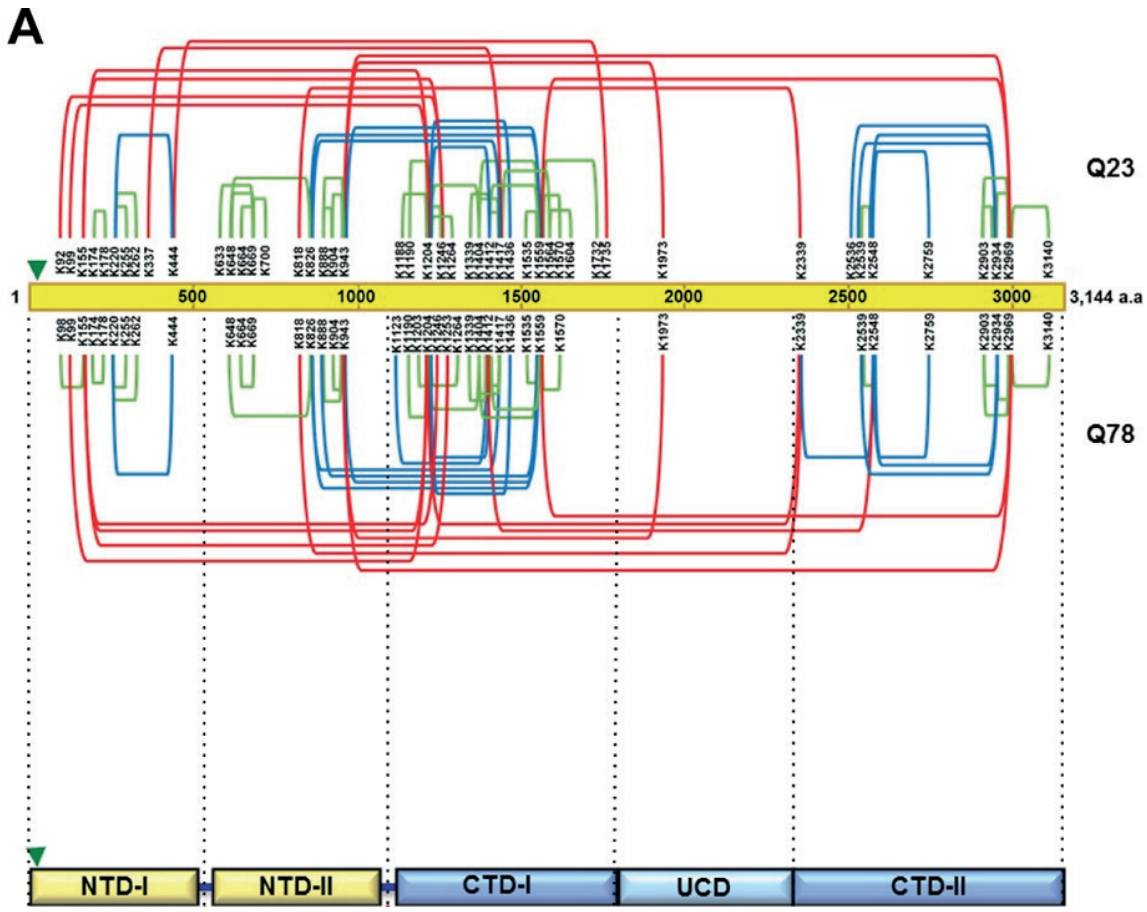



**Figure 4.2 | Q23- and Q78-Httn specific CLMS data, adapted from [67].**

The 3144 amino acid sequence of Q23-Httn is shown as a yellow bar with the poly-Q region indicated by a green arrowhead. The short-, mid- and long-range intra protein contact described by the CLMS data are represented in the green, blue and red colors respectively for Q23-Httn and Q78-Httn. At the bottom of the image is shown the five Httn substructures.

### 4.2.2 Cross-linking-mass-spectrometry

The experimental procedure describing the acquisition of CLMS data is described in detail in the work of our collaborator [67] (Figure 4.2), and thus would not be described further in this thesis. Importantly for the work described here, the CLMS information described 96 intra-protein residue contacts in the Q23-Httn. The crosslinking reagent used was disuccinimidyl suberate (DSS) and the targeted residues were the lysines (Lys) of the Httn proteins.

## 4.3 Predicting Httn substructures

In studies where the resolution is lower than 4 Å, the information provided from the cryo-EM density maps is sufficient to guess atomistic coordinates and hence obtain a comprehensive atomistic models of proteins [3-6]. Being at a maximal resolution of 8 Å, it not possible to apply the same methodology to model the atomistic structure of Httn solely from the EM density maps, hence the decision to use IM approaches. At the core of IM methods, high-resolution structures of the macomolecule subunits, which can be obtained from standard experiments, are combined with experimental data that describe their mode of assembly. In this case, except for a small crystallized fragment of Exon 1 containing the N-terminal 17 polyQs (PDB-id 3ior), no structure solved at atomistic details are available for Httn to be integrated with the experimental data.


**Figure 4.3 | Secondary structure prediction of the full-length Httn sequence.**

The full Q23-Httn amino structure is displayed here together with the secondary structure elements prediction. In this case, the secondary structure assignment was based on the consensus between PSI-pred and J-pred. This means that for a given amino acid residue, both PSI-pred and J-pred should be predict the same secondary structure element for it to be shown.

Thus, the only avenue left was to model the individual subunits composing the Httn protein and attempt to integrate these models with the cryo-EM and CLMS data, in the aim to derive a near-atomistic model. To address this challenging task, we first predicted the secondary structure elements using widely available secondary structure (SS) prediction servers. In this scope, the fasta sequence of the wild-type (Q23) human Httn was retrieved and subjected to SS prediction servers PSI-pred [89] and J-pred [139]. Similar to the study by ref. [140], the decision on whether to assign amino acids to a certain secondary structure element was based on a consensus between PSI-pred and J-pred predictions. This initial analysis led to the observation that 1431/3140 (45%) of amino acids were predicted to be part of 111 α-helices (Figure 4.3).

**Table 1 | Summary of HEAT repeats predicted from the literature.**

| Reference | HEAT-1 | HEAT-2 | HEAT-3 | HEAT-4 |
|---|---|---|---|---|
| Andrade (1995) | 205-329 | 745-942 | 1534-1710 | - |
| Palidwor (2009) | 114-413 | 672-969 | - | 2667-2938 |
| Tartari (2008) | 124-391 | 803-1100 | 1425-1710 | 2798-3107 |
| Modelled segment | 91-401 | 664-944 | 1412-1701 | 2663-2970 |
| Length | 310 | 280 | 289 | 307 |
| % of total seq. | 9.86 | 8.91 | 9.19 | 9.76 |

Interestingly, a significant proportion of the α-helices assigned by the SS were predicted to form HEAT repeat structures (Table 1). In fact, from the different studies undertaken to predict the location and length of the HEAT repeats in Httn, there seem to be a consensus that the Httn amino acid sequence is composed of four large segments forming HEAT repeats, which cover ~40% of the total sequence [135-137] (Table 1).

## 4.4 Modeling the Q23-Httn structure with *pow*$^{er}$

### 4.4.1 Modeling and docking Httn substructures

In an initial unsuccessful strategy to predict the full-length Httn structure, we used *pow*$^{er}$*-mViE* to rigidly dock four major HEAT repeats into the 8 Å Q23-Httn cryo-EM map (Figure 4.4a, top panel) by integrating the available heterogeneous experimental data. As a summary of the approach, the atomistic structures of the four HEAT repeats were modeled according to secondary structure predictions (Figure 4.3) and literature information (Table 1). The comparative cryo-EM analysis suggesting the position of NTD1 and NTD2 (Figure 4.4b) was used to dock the first HEAT structure at the location of NTD1. The CLMS data was used during the rigid docking of the other three HEAT structures as constraints to ensure the inter-HEAT residue contacts defined by crosslinked Lys were satisfied.

Eventually, we obtained a model that was satisfactory in the way that all the HEAT structures fitted well inside the density map (*ccc* = 0.8), no steric clashes were present and all the crosslinked distances were satisfied. Unfortunately, upon further analyzing the suggested structure and interacting with the experimentalist group it became apparent that this model, which contained ~40% of the Httn structure, occupied most of the volume of density map and thus left no other space to model the remaining 60% of the structure. The reason behind this setback was that treating the HEAT structures to be assembled into Httn macromolecules as rigid instead flexible structures forced them to occupy more space inside the density map, subsequently preventing the addition of the remaining Httn structure.

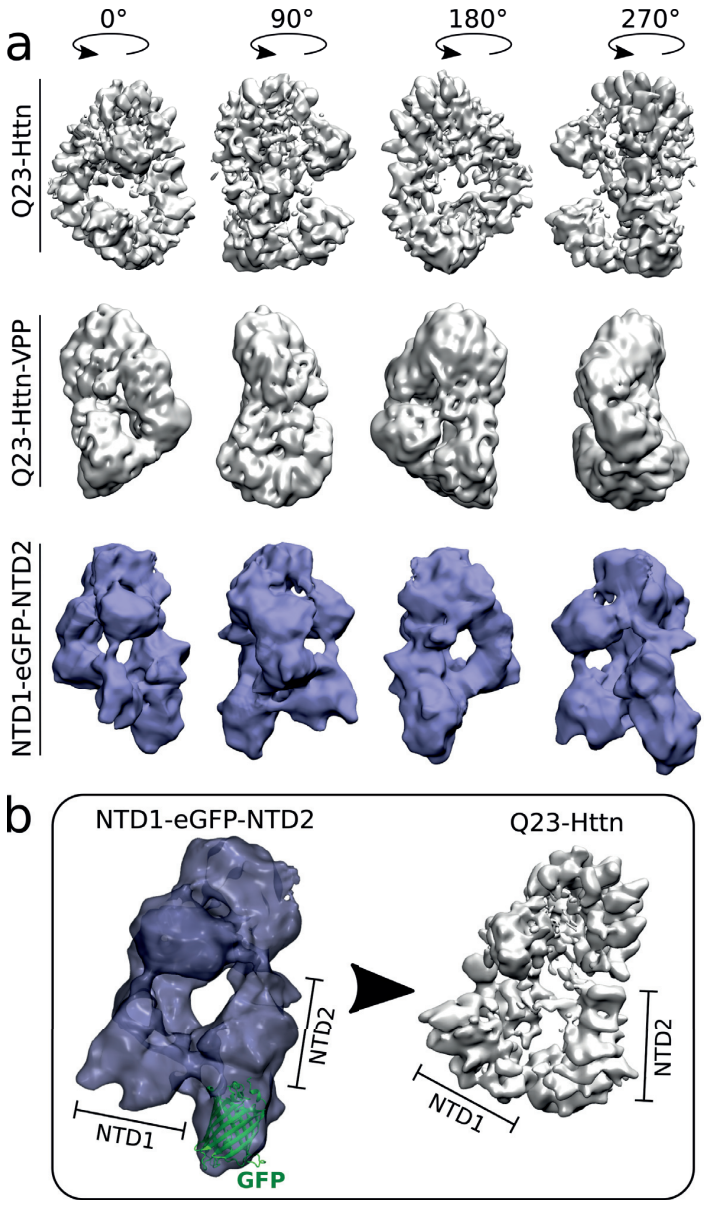



**Figure 4.4 | Cryo-EM maps of the full-length Huntingtin.**

**a.** The two top panels describe cryo-EM maps obtained from monomeric full-length Httn protein featuring poly-Q expansion of 23 with and without VPP with estimated resolution of 8 Å and 11 Å respectively. The bottom panel describes the Huntingtin protein featuring a poly-Q expansion of 21 fused to a eGFP in between NTD-1 and NTD-2 (also shown in Figure 4.6) with an estimated resolution at 15 Å. **b.** Comparison between cryo-EM maps obtained for Q21-Httn with a fused GFP in between NTD1 and NDT2 (NTD1-eGFP-NTD2, left panel) and Q23-Httn (right panel). The eGFP was docked into NTD1-eGFP-NTD2 and was used to determine the location of NTD1 and NTD2 on the Q23-Httn.

In an effort to overcome the setbacks of the previous strategy, we devised a second strategy that consisted in modeling the individual α-helices using the electron density information of the cryo-EM map. Unlike the previous approach, the relative ordering of the α-helices modeled into HEAT repeats and their spatial relationship to one another were not taken into account during modeling. This greatly simplified the modeling problem in which the task consisted solely in finding the location of helices strands inside a density map of medium resolution. A few methods have been implemented to solve such type of problem, with usually a good documentation and ample test cases [141,142].

Thus, starting with as input the 8 Å Q23-Httn density map, we used the volume tracer method implemented within the SITUS package [142] and the *find_helix_strand* command of the PHENIX package [141] to detect the position of α-helices within the Q23-Httn density map. Nevertheless, none of these methods could detect and model α-helices in a satisfactory manner, which could possibly be due either to the low resolution of the density map or to methodological reasons related to the respective detecting methods.

Due to the fact that current methods seemed inadequate to detect the location of α-helices, we envisioned an alternative strategy that consisted in flexibly docking the individual α-helices into the Q23-Httn density map so as to detect the location of helices. In this case, flexibility was taken into account by modifying the length of the α-helices during the assembly process.

Practically, the initial phase of the strategy consisted in producing 60 assembly models, each obtained by iteratively docking 100 α-helices with *pow*$^{er}$*-mViE* (Figure 4.5A) following the optimization protocol formulated as:

$$Max\ ccc(x),\ s.t.\ \begin{cases} l_i \leq x_i \leq u_i,\ i = 1,2,\dots,n \\ E_{phys} < 0 \end{cases}$$

Where the objective function $ccc = \frac{N}{N_{sol}} \frac{\sum_i^N (L \otimes ref)_i\,(L \otimes sol)_i}{\sqrt{\sum_k^N ref_k^2\ sol_k^2}}$ is the cross-correlation-coefficient describing the fit between protein subunits and the provided density map (L corresponds to a 3x3x3 Laplacian filter kernel used to convolute the input density maps with the aim to increase the precision of the density map fitting[126]. *N* is the number of overlapping voxels between the density maps, $N_{sol}$ the total number of voxel of the candidate solution density map, $(L \otimes ref)_i$ is the $i^{th}$ voxel of the reference density map and $(L \otimes sol)_i$ its counterpart from the candidate monomer assembly);

$E_{phys} = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^9 - \left(\frac{\sigma}{r}\right)^6\right]$ is a simple 9-6 Lennard-Jones coarse energy potential [94] to avoid steric clashes (r consists in the pairwise distance between the Cα interfacial atoms of the subunits within a distance of 12 Å, $\sigma$ = 4.7 Å and $\varepsilon$ = 1 kcal/mol);

For each of the 60 assembly models, the aim of the optimization was to individually and successively fit the 100 α-helices into the Q23-Httn density map (through *ccc* maximization), by sampling the α-helix parameters [*x, y, z, α, β, γ, len*], where *x, y*, *z* are translations, *α, β, γ* the orientations and *len* the flexible length of the α-helices (varying from 13 to 25 amino acids), whilst satisfying the constraint specified by $E_{phys} < 0$ to prevent steric clashes. The relative ordering, residue type and the spatial relationship between α-helices were not considered during the docking. Therefore, each α-helix structure was modeled as a simple poly-Ala α-helix where only the backbone Cα-atoms were considered during docking and no inter-helical constraints were imposed by CLMS data. The docking of each of the 100 helix was allocated a budget of 300'000 function evaluations, summing up to a total 1.8 billion function evaluations for the 60 assembly models.

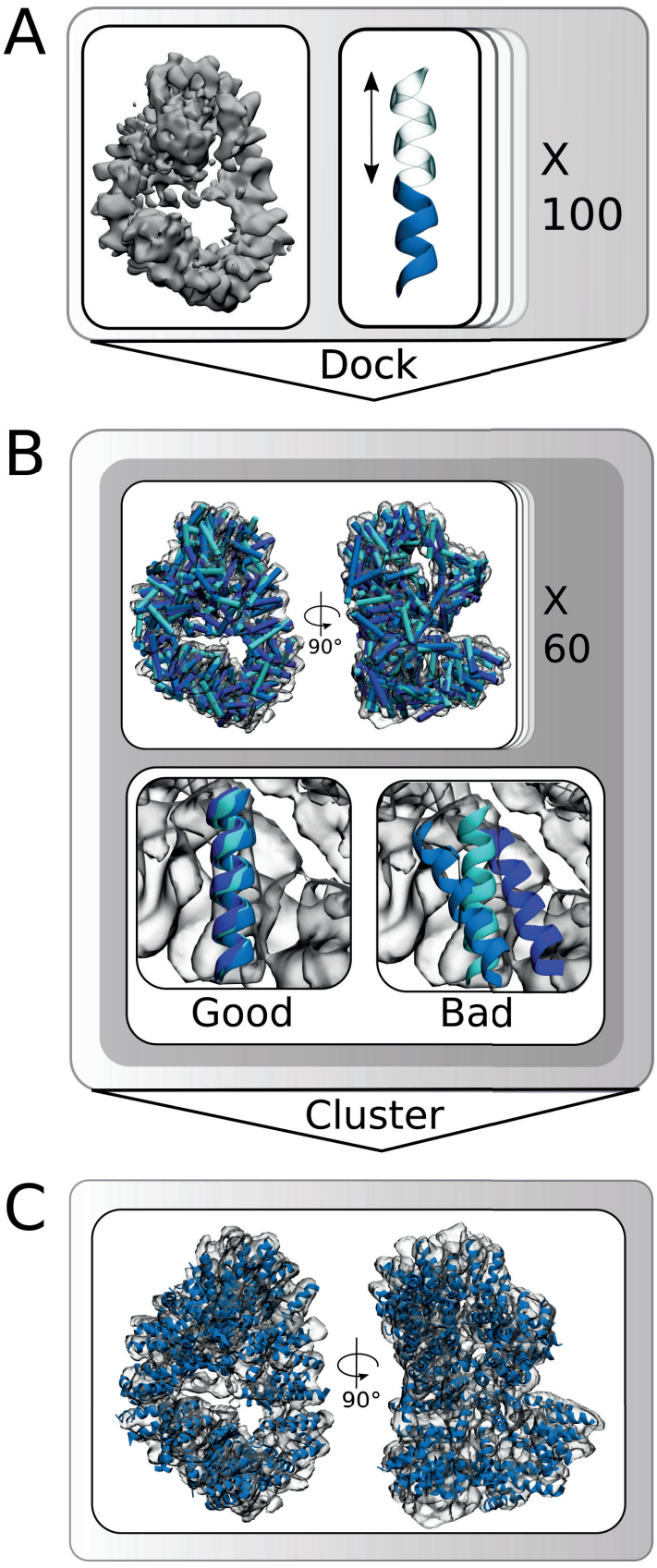

A
X
100
Dock
B
90°
X
60
Good
Bad
Cluster
C
90°

**Figure 4.5 | Modeling and docking process of Httn α-helices into Q23-Httn cryo-EM map.**

**A.** The docking procedure was undertaken with *pow*er-*mViE* and consisted in iteratively docking 100 modeled poly-Ala helices, by simultaneously sampling their roto-translations and modifying their length, into the provided 7 Å cryo-EM map of the Q23-Httn. **B. (top panel)** A total of 60 models, each featuring 100 docked helices, were subjected to a dedicated clustering algorithm. (**bottom panel**) The aim of the clustering algorithm was to extract the Q23-Httn cryo-EM map areas most populated by the poly-Ala of different models. This was done first by assessing the structural overlap between the helices of different Httn models (here, 3 models out of the 60 are represented with different tones of blue). In this case, individual helices of different models found to be in similar density map locations (Cα-RMSD < 4 Å, "good") were clustered together and helical centroids were extracted and added to the final model (**C.**).

Then, the α-helices found at identical locations in most of the 60 models were combined (Figure 4.5B). Given that the helices were fitted into the Q23-Httn map by matching their simulated electron densities to the one of the map, the motivation behind this approach was to detect the density map areas more likely to be populated by the docked α-helices. We then implemented a dedicated clustering algorithm to detect these populated areas and to extract the α-helices overrepresented in these areas. Briefly, the approach of the clustering method was to (i) compute the distance (Cα-RMSD) between helices of different models, (ii) detect clusters of α-helices where the distance between models were minimal (Cα-RMSD < 4 Å) and which location was consistent between models (Figure 4.5B, bottom-panel), (iii) extract, out of each α-helices clusters, a representative α-helix to be included in the final model, and (iv) compute the structural distance between these representative α-helices to make sure their atoms do not overlap and hence obtain a clash free model. For the sake of better understanding, the main idea is illustrated in Figure 4.5B (bottom-panel**),** where α-helices from 3 models out of 60 are shown in different tones of blue. In this idealized visual representation, the clustering algorithm would select and then extract representative α-helix out of the cluster depicted on the left panel ("good") rather than the one the right ("bad") because the α-helices on the left panel have a better positional and structural agreement between different models.

Eventually, a model containing 114 α-helices was obtained, which roughly corresponded to the total number of α-helices predicted by SS prediction software (111). Using the semi-automated *fit_in_map* command of the UCSF Chimera software[143], the fitting of each representative α-helices inside the density map was refined further. During the refinement, α-helices found outside the volume of the density map and in areas of low densities were removed from the model, leaving 96 helices, which according to the SS prediction results, consisted in the number of predicted helices with at least 8 residues.

### 4.4.2 Ordering Httn helices via Monte Carlo optimization

Resulting from the initial phase, we obtained a poly-Ala model scaffold describing the most probable helical location inside the Q23-Httn density map (Figure 4.5C).

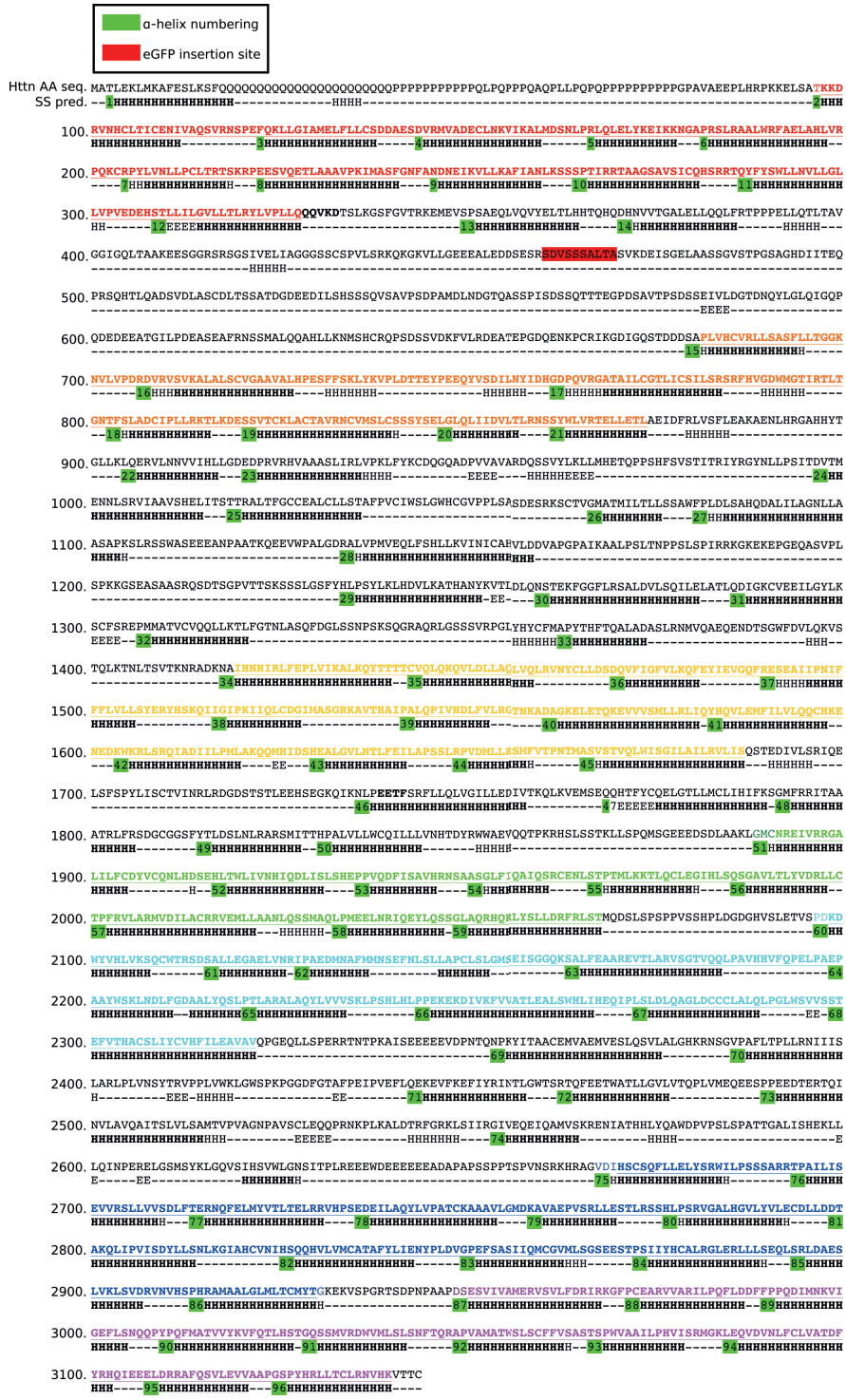


**Figure 4.6 | Amino acid sequence of Q23-Httn and modeled Helices**.

The full-lenght amino acid sequence of Q23-Httn is represented in the top line of each numbered row. At the bottom line of each numbered row are the associated secondary structure elements as predicted from the consensus between J-Pred and PSI-pred. The seven major regions describing the probable HEAT repeats structure within Q23-Httn are represented by the seven rainbow colors. The 96 modeled longest helices (≥8 residues) used to find the real location of Httn helices inside the previously modeled poly-Ala scaffold (Figure 4.5) are numbered and shown with bold format.

Using the helical positions we detected earlier, the next phase was to correctly map the location of the true Httn helices by aligning them to the fixed poly-Ala α-helices, used as scaffold. For the sake of consistency with the number of poly-Ala α-helices, the longest (≥8 residues) 96 Httn helical structures were modeled based on the SS prediction. Details regarding the modeled α-helices amino acid composition and their numbering can be found in sequence found in Figure 4.6.

The task of finding the correct Httn α-helices location and ordering inside the density map was converted into an optimization problem where the discrete search space consisted in all the possible alignments of Httn helices on the poly-Ala α-helices and the objective function combined the satisfaction of residue distance specified by the CLMS and eGFP experiments, and the distance between consecutive α-helices.

Using a grid search to sample all the possible alignment through permutations would have amounted to $96! \approx 1.0e^{150}$ iterations/function evaluations, which would have taken far too long even with modern computational resources. Thus we decided to use the optimization capabilities of *pow*$^{er}$ to solve this optimization problem. However, given that the current *pow*$^{er}$ optimizer (mViE and PSO) work best on a continuous search space rather than a discrete one, which is featured here, we designed a simple Monte Carlo optimization algorithm within *pow*$^{er}$ to the find the optimal Httn helical locations according to the algorithm:

**Algorithm 3**

```
aligned_helices ← Randomize_Helix_alignments(Scaffold_helices, Httn_Helices)
best_score ← compute_spatial_distances(Aligned_helices)

for each timestep t do:

  helix_to_swap_1 ← random_integer(1,96)
  helix_to_swap_2 ← random_integer(1,96)

  if helix_to_swap_1 != helix_to_swap_2 then
    trial_alignment ← swap_helices_location(helix_to_swap_1, helix_to_swap_2)
  end if

  score ← compute_spatial_distances(Aligned_helices)

  if score < best_score then
    best_score ← score
    aligned_helices ← trial_alignment
  end if

end for

return aligned_helices
```

At timestep t=0, each of the 96 helix structures stored in the vector ***Httn_Helices*** is structurally aligned on a randomly chosen poly-Ala helix stored in ***Scaffold_helices*** and their structural coordinates are recorded in the vector ***aligned_helices***. Then an initial score is computed as the addition of satisfaction values (0 if satisfied, 1 if not) on all the measured inter-helical crosslinked Lys distances and the distance between successive helices (i.e. distance between helix 1 and helix 2, … helix 95 and helix 96) and is considered as *best_score*. Then at each timestep *t*, the location of two randomly chosen helices aligned on the scaffold is swapped and their new structural coordinates are recorded as a tentative move in ***trial_alignment***. From the new structural coordinates, a new score is computed. If the trial helical alignment combination is found to satisfy the experimental data better than the previous alignment combination, it is kept for the next iteration and the *best_score* is updated.

In this optimization process, a crosslink distance found below 45 Å and inter-helical distances below 20 Å were considered satisfied. The reason for choosing a crosslink cutoff distance of 45 Å rather than the traditional 35 Å associated with DSS crosslinks [67], is that 98% (94/96) of the Httn crosslinked Lys were located on flexible regions which were not modeled. Due to the fact that only defined helical elements were considered in the current modeling process, the crosslinks locat-

ed on these loops/flexible regions were reported onto the nearest atoms of the modeled helices, so as not to discard precious structural information to guide the optimization process.

In order to further guide the optimization process and reduce the search space complexity, we used the information provided from the eGFP experiments (Figure 4.4) and fixed the position of helices at amino acid position 98 and 374 (respectively numbered as helices 1 and 14 in Figure 4.6) on the poly-Ala helix scaffold so that they are located at the beginning and end of NTD1.

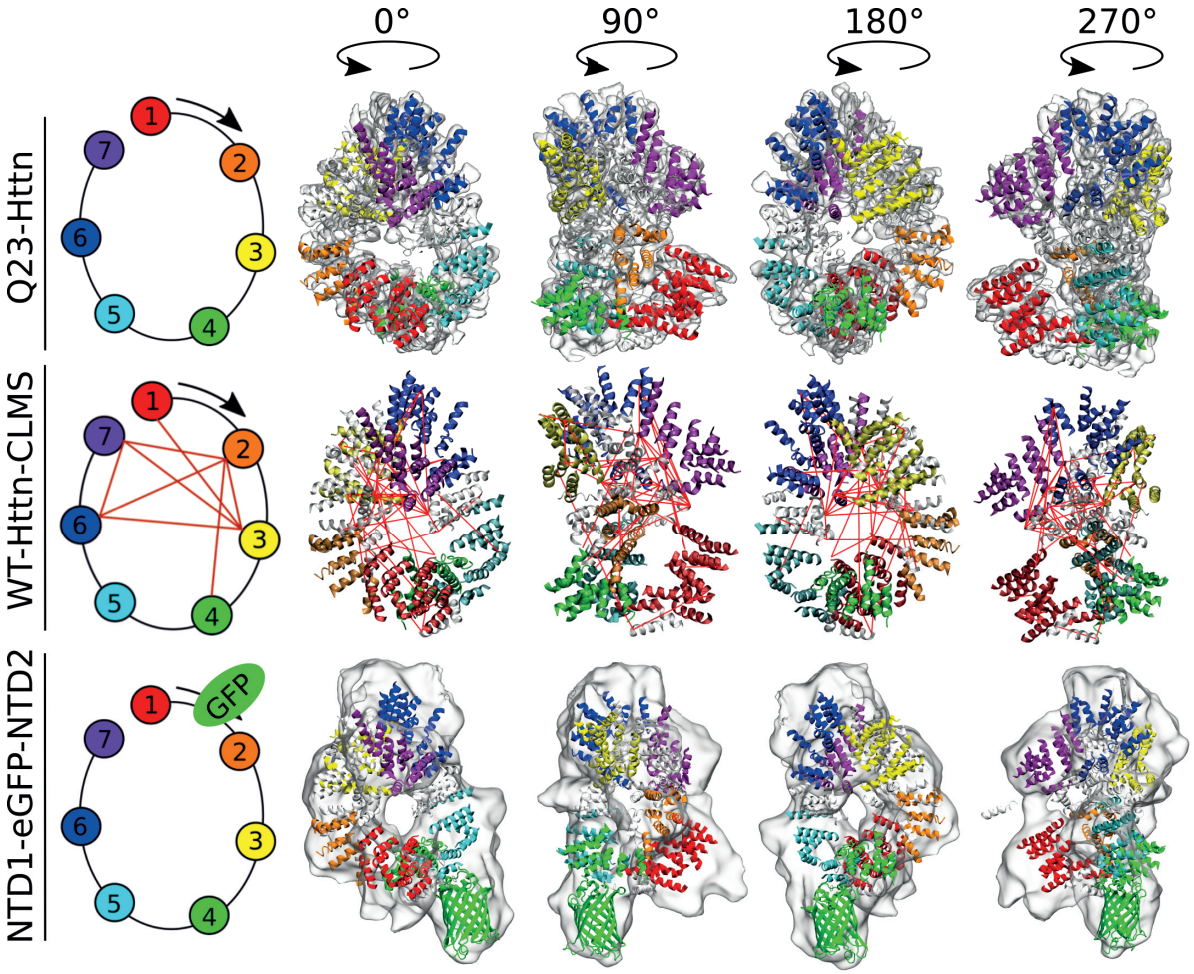



**Figure 4.7 | Final model obtained for Q23-Httn and experimental data satisfaction assessment.**

**Top panel.** The helices assigned to respective HEAT repeat structures were colored in rainbow colors according to Figure 4.6, other helices were colored in white. The modeled α-helices agree well with the densities of the 8 Å Q23-Httn cryo-EM map (ccc = 0.7). **Middle panel.** The CLMS data describing intra-protein contacts, which is predominantly found on residues located on loop/flexible regions, were represented as red lines separating the helices Cα-Cα atoms. The average distance of crosslinked Cα-Cα atoms was 27.9 Å. **Bottom panel.** The Q23-Httn model was rigidly docked into the cryo-EM map featuring a fused GFP between NTD-1 and NTD-2 (see Figure 4.4). In this case a GFP structure

(pdbid: 1gfl) was also docked in the map. In the NTD1-eGFP-NTD2 cryo-EM map (see Figure 4.4), GFP is located in the cleft between the HEAT structures colored in red and orange. In the Q23-Httn amino acid sequence (Figure 4.6), The GFP was fused in the middle of a ~300 amino acid-long flexible region (unresolved in the cryo-EM map) at residue position 466, which was separated by ~130 amino acids from the red HEAT structure and ~200 residues from the orange HEAT structure. Due to this relatively long distance, the GFP might appear instead closer to the green and cyan HEAT structures. Nevertheless in the cryo-EM map, it is located close to helix 14, which is in between the red and orange HEAT structures (closest Cα-Cα atoms < 10 Å, see Figure 4.6).

Eventually, after visually inspecting and carefully analyzing all the best models produced by the Monte Carlo optimization, we narrowed down a model that best described the structure of the Httn protein. The criteria for the selection of this model were that the 96 helices should follow a logical order between each other to be structurally plausible and that the information described by the CLMS and eGFP data should be best satisfied. We further refined that model using the UCSF Chimera software [143] in order to correctly orient the helices with respect to each other and to refine their position within the density map as the density map was not directly used during the Monte Carlo optimization. We then obtained a final structure, which modeled helices position and relationship to one another seemed to agree with the Q23-Httn cryo-EM map (Figure 4.7, top panel**).** The analysis on the satisfaction of the DSS crosslinks mapped to the helices (Figure 4.7, middle panel**)** revealed that the crosslinks tended to be more concentrated in the region spanning the HEAT structures here labeled as 1, 2, 3, 6 and 7 (Figure 4.7, middle panel). The reason for crosslink to aggregate in specific protein location remains unclear as it could be related to several causes including surface accessibility or to the condition in which the crosslinking was performed [144]. The average distance between crosslinked residues (Cα-atoms) was 27.9 Å (± 13.44 Å), which was below the 45 Å cutoff distance used for the optimization process and more importantly, also below the standard 35 Å cutoff value used for DSS crosslinks. Eventually the model was found to satisfy the suggested location of eGFP in the NTD1-eGFP-NTD2 cryo-EM map with the close spatial proximity of helix 14 (located between the HEAT structures 1 and 2 colored in red and orange where the eGFP was fused) to the fused eGFP protein (Cα-Cα atoms < 10 Å, Figure 4.7, bottom panel).

## 4.5 Structural assessment of the Q23-Httn model

While still in the process of refining the model presented in Figure 4.7, in the very last weeks, the structure of the Q17-Httn protein was solved from a 4 Å cryo-EM map [87]. Remarkably in this work, single particle cryo-EM was used to obtain a high-resolution structure by binding the Q17-Httn

structure to HAP40, which is a long HEAT repeat forming protein that is naturally found to interact with Httn. Given the very recent availability of an experimentally solved Httn structure, we performed a structural comparison between our final model and the Q17-Httn structure recently published (PDB-id: 62z8, Figure 4.8).

### 4.5.1 Structural comparison of Q17- against Q23-Httn

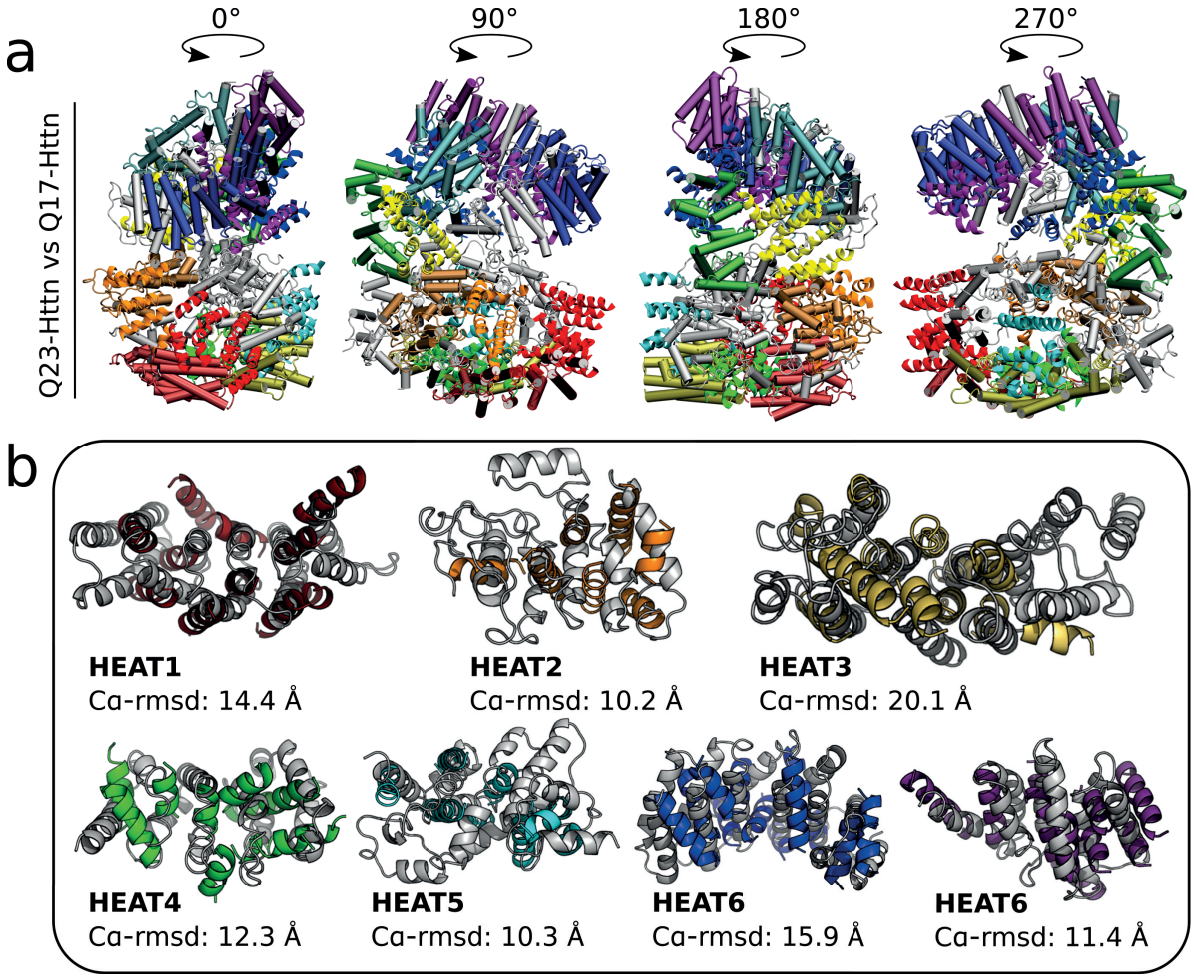



**Figure 4.8 | Structural comparison between HEAT structures of the Q23-Httn model and 4 Å resolution Q17-Httn.**

**a.** Global comparison of the Q23-Httn modelled from the 8 Å cryo-EM map and Q17-Httn solved from a 4 Å cryo-EM map (pdbid 62z8). The helices of the Q23-Httn structure are represented as cartoons [145] atomic models while the helices of Q17-Httn are represented as tubes. The HAP40 protein structure normally bound to Q17-Httn was removed. **b.** Local comparison between HEAT structures of Q23-Httn and Q17-Httn. The Cα–atoms of the Q23-Httn HEAT structures (rainbow colors) were superimposed onto the respective Q17-Httn HEAT structures (grey color) with Pymol [146] and the structural difference were reported as Cα-rmsd values.

Visual inspection of our Q23-Httn model against the published Q17-Httn structure revealed that both structures were mostly composed of defined helical structures which arranged into HEAT repeats. Global structural alignments of both Q23-Httn and Q17-Httn indicated that N-terminal HEAT1 and HEAT2 as well as C-terminal HEAT6 and HEAT7 were in similar geographical locations (Figure 4.8a). Conversely, HEAT3, HEAT4 and HEAT5 had several geographical disagreements between the Q17- and Q23-Httn structures. Local structural alignments between the Cα–atoms of respective HEAT repeats revealed as well that the overall topology of the helices forming the HEAT repeats were not well respected between Q17- and Q23-Httn (Figure 4.8b). More precisely, the structural differences, quantified by Cα-rmsd values, ranged from 10.3 Å (HEAT2, Figure 4.8b) to 20.1 Å (HEAT3, Figure 4.8b). The biological reason for having >10 Å Cα-rmsd values between HEAT structures of respective Httn models could be related to the inherent flexibility of HEAT structures which might naturally vary from one Httn structure to another. There is also the possibility that structural differences observed in the local HEAT structures was induced by the presence of HAP40 bound to Q17-Httn and absent in Q23-Httn, which might have led to more global conformational changes of the entire protein. Obviously, we do not discard the possibility that structural inaccuracies could have arisen from the method employed to model and fit the individual helices that formed the HEAT structures. In fact, we noticed that our method had the tendency to model more compact HEAT structures compared to ones solved in the 4 Å Q17-Httn cryo-EM map(Figure 4.8b). This was especially observed in HEAT1, HEAT2, HEAT3 and HEAT5 structures where the overall Q17-Httn structures appeared more elongated than the respective Q23-Httn structures.

### 4.5.2 Assessment of Q17-Httn experimental data satisfaction

In order to further rationalise the structural differences that occurred between the Q23-Httn and Q17-Httn structures, the satisfaction of the experimental data used to model Q23-Httn was assessed on the new Q17-Httn structure. First, the Q17-Httn without bound HAP40 was docked inside the 8 Å Q23-Httn map used to model our structure. This analysis showed that the Q23-Httn cryo-EM map electron densities that should have described the location of the Q17-Httn N-terminal regions (HEAT1 in red, HEAT2 in orange and flanking in white, Figure 4.9a) were in fact missing. In this respect, we suggest the absence of electron densities in the N-terminal region of the Q23-Httn cryo-EM map could have strongly contributed to the structural discrepancies observed between our model and the cryo-EM Httn structures.

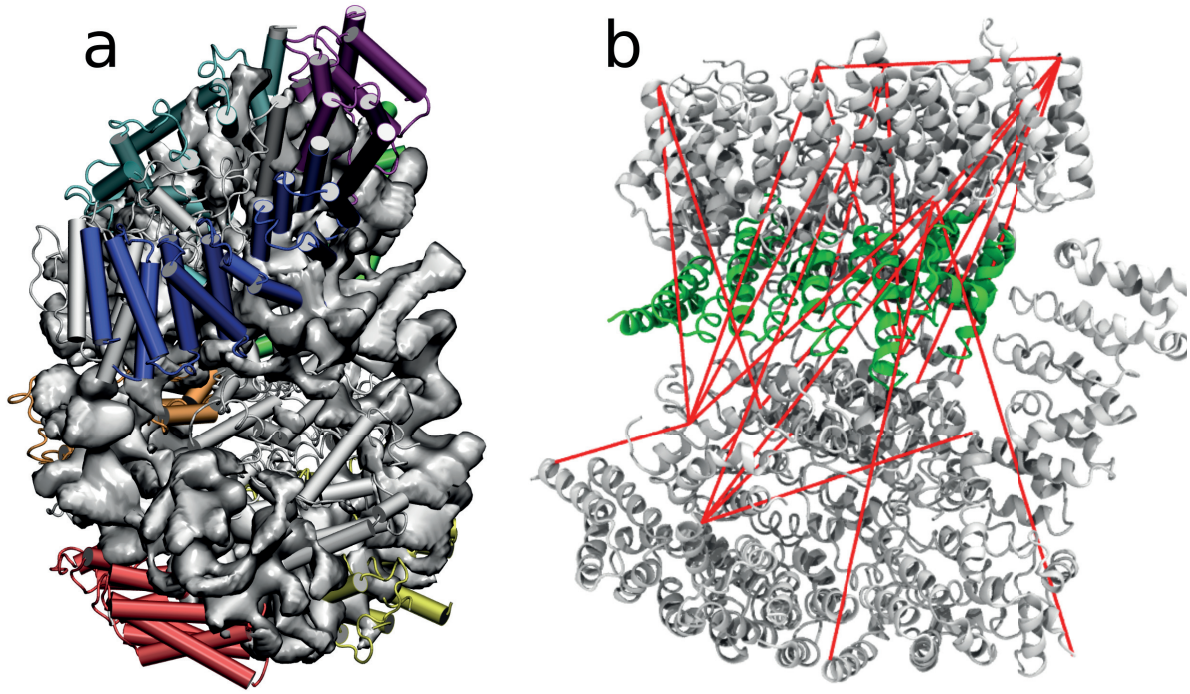



**Figure 4.9 | Structural analysis of Q17-Httn bound to HAP40 (pdbid 62z8) and experimental data satisfaction.**

**a.** The Q17-Httn HEAT structures are colored according to the rainbow colors described in Figure 4.6. The Q17-Httn structure was docked into the 8 Å Q23-Httn cryo-EM map and revealed that important densities were missing the N-terminal regions of the cryo-EM map (illustrated by the red, yellow and white HEAT repeat structures not matching the density location). **b.** The Q17-Httn structure is colored in grey, the bound HAP40 in green and the crosslinked residue distances mapped on the Cα-atoms are colored in red. Here are reported the crosslink distances found above the upper distance limit of 45 Å.

Furthermore, we assessed the level of CLMS data satisfaction by mapping crosslinked residue distances onto the Q17-Httn structure (Figure 4.9b). In this analysis, we observed that the average distance between Cα-atoms of the crosslinked residues was larger (33.9 Å ± 25.44 Å) than our Q23-Httn model (27.9 Å ± 13.44 Å), where the longest violated distances recorded were 115 Å and 63.3 Å for Q17- and Q23-Httn respectively. Further analysis of Q17-Httn structure led to the understanding that these violated distances were in fact mapping the residues separated by the HAP40 protein (colored in green in Figure 4.9b), which, should have been in close proximity in the Q17-Httn monomeric form. The presence of HAP40 could have led to significant conformational changes in Q17-Httn which in turn could represent some of the underlying reasons for the differences observed when comparing our Q23-Httn structure that of Q17-Httn.

## 4.6 Discussion and conclusions

In this work, we described the integrative modeling of Q23-Httn structure with the *pow*$^{er}$ framework by combining cryo-EM and CLMS data. The choice for using *pow*$^{er}$ was based on the difficulty encountered by other state-of-the-art software to model and assemble helical structures inside the available 8 Å cryo-EM map. The obtained Q23-Httn structural model, computed through docking, clustering and finding the correct spatial arrangement of the modeled helices, was found to satisfy the experimental data. Recent availability of a Q17-Httn atomistic structure solved from a 4 Å cryo-EM map enabled to assess the quality of our Q23-Httn model. Similarly to the solved Q17-Httn structure, the Q23-Httn structure also featured a multitude of helices as components of larger HEAT structures. Additionally, the extreme N-terminal and C-terminal HEAT structures were found to have similar geographical locations. Unlike the Q17-Httn structure however, the HEAT structure located more centrally in the amino-acid sequences were found to have different geographical location. Such differences were found to be more pronounced when comparing the individual HEAT structures, where differences were > 10 Å (Cα-rmsd) between respective Q17- and Q23-Httn HEAT structures. Assessing the satisfaction of Q17-Httn on experimental data used to model Q23-Httn led to the understanding that the observed structural differences might have occurred for varying reasons.

A possible reason for the structural discrepancy could have been related to the inherent nature of HEAT repeats to form flexible solenoid rods, which likely created differences from one Httn structure to another. Other reasons could have been related to the poor resolution of the available cryo-EM maps, limited amount of CLMS data and intrinsic methodological errors associated with the current approach used to model the Q23-Httn structure. However, it was not possible to compare the performance of our approach to similar ones since there was no evidence in the literature of methods that have attempted to model such large proteins (~350 kDa) from cryo-EM maps obtained at a resolution of 8 Å or lower, by modeling and docking the individual helical structures. In order to root out the possibility that structural differences in Q23- and Q17-httn might have arisen because of inaccuracies in our suggested method, a thorough evaluation of the robustness of the method would have to be undertaken. Such robustness assessment would be undertaken on other large macromolecules solved from cryo-EM experiments also featuring several HEAT repeats, such as the 3825 amino acid-long transcription co-activator complex SAGA (Tra1, pdbid: 5oej, mass: 437 kDa) solved from 5.7 Å cryo-EM map.

We also suggest a possible reason for Q23- and Q17-Httn structural differences was related to the quality of the experimental data used to assist the modelling of Q23-Httn. Precisely, structural alignments followed by visual inspection of the Q17-Httn structure against the 8 Å Q23-Httn cryo-EM map revealed missing densities in the N-terminal regions which might have led not only to a generally more compact Q23-Httn structure (also reflected in the more compact shape of HEAT1, HEAT2, HEAT3 and HEAT5), but also to mismatches when trying to map the geographical location of the respective HEAT structures. Finally, the violation of several crosslinked residue distances by Q17-Httn indicated that the presence of the bound HAP40 likely conferred a conformation change in the overall Q17-Httn, which was difficult to predict from the monomeric 8 Å Q23-Httn cryo-EM structure, and which was probably also responsible for the structural differences observed in the cryo-EM maps and related models.

Importantly, the presence of the bound HAP40 protein was instrumental to stabilize the Q17-Httn structure and enabled its atomistic resolution at 4 Å [87]. Nevertheless, the structure obtained from this work does not directly allow to understand the role of poly-Q expansion and Huntington's disease. Instead, we suggest this relationship can be elucidated through functional and structural comparisons between the monomeric forms of Httn with different poly-Q expansions (wild-type and disease). In this respect, the Song lab obtained both monomeric Q23- and Q78-Httn cryo-EM maps. The initial aim of this challenging project was to model a reasonable Httn structure from the cryo-EM with the best resolution (here, the Q23-Httn map at 8 Å). Upon model validation, the next phase of the global aim was to extrapolate the Q23-Httn model into the Q78-Httn cryo-EM map obtained at lower resolution of 10 Å, and rationalize the structural similarity/differences that might indicate a reason for the disease.

Instead of a setback to the overall aim, the recent availability of a higher resolution Q17-Httn structure can be used, instead of our model, for the goal specified above. First however, the impact of the bound HAP40 on the overall Httn structure would have to be understood. In this aim, we are now performing molecular dynamics simulations on the Q17-Httn structure without HAP40 in order to relax its structure in isolation, as in our samples, and thus observe the impact of HAP40 on the overall Httn architecture. As shown by preliminary results, we expect that Httn will assume a more compact conformation resembling more closely those observed in our cryo-EM maps. After careful analysis of the results obtained from MD simulations, we plan on filtering using *pow*$^{er}$, the different conformations sampled through MD to find the ones which best satisfy the monomeric Q23-Httn experimental data as well as CLMS data. We would then flexibly fit this monomeric WT

Httn to the 10 Å cryo-EM map of Q78-Httn using available programs such as MDFF [36]. Eventually, we think this procedure would be useful to shed light on the structural differences induced by the extent of poly-Q expansion and help understand the relationship between Httn structure and disease.

# Chapter 5 Using energy grids to constrain the search space of small molecules during molecular docking

## 5.1 Introduction

The field of computer-aided drug discovery has emerged over four decades ago with the ultimate goal of providing the most accurate predictions that would enable the cost- and time-efficient delivery of new drugs. A major part of the drug discovery pipeline involves structure-based drug design (SBDD), which has been no less than instrumental in the development of important drug including the imatinib (Gleevec) [147] against Abl tyrosine kinase, Amprenvir (Agenerase)[148] against HIV protease or zanamivir (Relenza) against neuraminidase [149].

Given the three-dimensional structure of a target protein, the typical goal of SBDD is to suggest potent inhibitors (i.e. with binding affinities in the nM range) in the form of an optimized lead compound, that would undergo phase I clinical trials [150,151]. To address this challenging task, the SBDD process features many iterative cycles synergizing cutting edge experimental and *in silico* methods. In the early cycles of the SBDD, which includes the identification of hit compounds with binding affinities in the μM ($10^{-6}$) range, two types of approaches are commonly used. The first experimental approach consists in using an experimental high throughput screening method (HTC) where a large number of compounds are tested against the target protein, and their associated binding affinities recorded.

In order to alleviate the substantial time and financial cost associated with HTC, the second avenue consists in using virtual screening (VS) methods for the *in silico* hit identification, with the aim of screening commercially available compounds against the target protein, using structure-based small molecule docking methods (SMD) [152]. In the remarkable study of Doman et al. [74] where compounds against protein phosphatase-1B target were screened, VS has been reported to provide a much greater enrichment, in term of number of compounds tested to have $IC_{50}$ < 100 μM over all suggested compounds, when compared to experimental HTC methods.

Given the three-dimensional structures of the unbound protein target and a database of compounds, SMD methods aim to predict the most probable bound conformation of a ligand in the tar-

get binding site first by generating ligand-receptor poses then by estimating their binding energy. To date, more than 60 SMD methods have been implemented, which strength and limitations have been extensively reviewed in the literature [153-157]. Amongst the most popular docking methods one can find GOLD (Genetic Optimization for Ligand Docking) [82], which uses a genetic algorithm to sample the conformational flexibility of the ligand and that of protein residues located in the binding site. Another popular method is ICM (Internal Coordinate Mechanics) [80,158], which uses a Monte Carlo minimization to sample the internal coordinates of the ligand with the aim to rapidly find its correct binding pose. Also featuring a Monte Carlo-based optimization to sample ligand conformation and position, Autodock Vina [76] is currently one of the most cited docking method in the scientific literature [159,160] with over 1000 citations in 2015 [161]. Another recent and efficient docking method called AC (Attracting Cavity) [77] has been implemented which is based on energy minimizations on smooth energy landscapes calculated on a cloud of points surrounding the receptor binding site.

Importantly, GOLD, ICM, AC and Autodock Vina have displayed great accuracies to predict the bounds conformation of ligands and their estimated binding energies. To achieve such performance, they rely on carefully calibrated global scoring functions composed of linearly added energy terms[76] including Lennard Jones potentials, Coulomb potentials, desolvation or hydrophobicity [76,77,80,82]. Coefficients in the form of weight constants are assigned to balance the contribution of these uncorrelated terms, which can be obtained from regression analyses by iteratively fitting the scoring function to experimentally determined binding affinities.

One major limitation of such approach is that, added to the costly calculation of these weights, the regression coefficients can be heavily dependent on the dataset used to compute them[162]. More importantly, the addition of new terms to an already optimized scoring function is nearly impossible without rebalancing their contribution. This limitation comes from the fact that the docking problem is treated as an unconstrained single objective optimization.

In an attempt to optimize several scoring terms without balancing them, Gu et al.[162] converted the SMD problem to a multi-objective optimization where multiple scoring functions were minimized simultaneously by a genetic algorithm. This approach was tested against the GOLD test set and showed remarkable docking accuracies compared to other popular molecular docking methods including GOLD [82], Glide [81], Surflex [163] and Dock [164].

Another way to circumvent this problem would be to use a constrained optimization scheme, which aims at minimizing a single objective function while ensuring the satisfaction of pre-defined constraints [83]. The advantage of this optimization scheme is that an unlimited amount of constraints can be added to guide and accelerate the optimization process without balancing their contribution. Constrained optimization has been applied in various fields including medicine [165], optics [166], engineering [167] and recently to the *in silico* prediction of protein macromolecular assembly [17]. Particularly in this recent work, the protein-protein docking problem was converted into a constrained optimization, where the *pow*$^{er}$ (parallel optimization workbench to enhance resolution) [16,62] framework coupled to a recent constrained optimizer named *mViE* (memetic viability evolution) [83] was used to dock symmetric and hetero-dimeric protein assemblies, by using experimental data to guide the assembly process [17] (See **Chapter 3**).

Unlike the protein-protein docking problem converted to a constrained optimization where experimental data are used as constraints, in SMD, no experimental data is usually available to assist the docking of small compounds into the target binding site. More generally, to our knowledge, no constrained optimization application has yet been reported to predict the docking of small molecules. The reason behind this could be related either to the difficulty to find features to be used as constraints or to the general complexity of the optimization problem.

Thus in this work, we sought to design a constrained optimization method suitable for docking small molecules. To address this task, we modified our previously implemented constrained optimization protocol *pow*$^{er}$*-mViE* by introducing an objective function similar to the one found in AutoDock Vina [76] and inequality constraints in the form of receptor-ligand physicochemical properties, extracted from energy grids. In order to test and validate the method, we used high-resolution receptor-ligand complexes of the PDBbind core set [168] to extract and calibrate the constraints, and the Astex diverse set[169] as the validation set. On a rigid docking setting with the PDBbind core set, our constrained molecular docking approach showed almost a four-fold increase in accuracy compared to an unconstrained docking. On a flexible docking setting with the Astex diverse set, we obtained docking accuracies comparable to that of state-of-the-art software including the AutoDock Vina program. In this study, constraints were extracted from energy grids generated in the binding site of a rigid receptor, and were used to guide the docking of a flexible ligand. While the results obtained using this approach are comparable to other successful SMD methods, we envision that the use of constraints extracted from energy grids would be critical for docking ligands more accurately

when also the native dynamics of the target protein, as for instance accounted from molecular dynamics simulations, is considered during docking.

## 5.2 Methods

The suggested constrained molecular docking protocol requires as input the atomistic structures of the ligand, and the receptor (Figure 5.1). From these structures, atom types are assigned based on the Xscore nomenclature [76,170]. Then, local and global constraints to assist the docking are computed from atom type-based pre-computed energy grid maps, using the procedure outlined in following methods section.

These local and global constraints are simultaneously used during optimization to guide the docking of the ligand inside the receptor binding site. In order to quantify the binding energy of the poses during docking, a scoring function similar to that of AutoDock Vina is minimized. At the end of the optimization, the best receptor-ligand pose, i.e. the pose with the lowest energy which satisfies all the constraints, is returned.

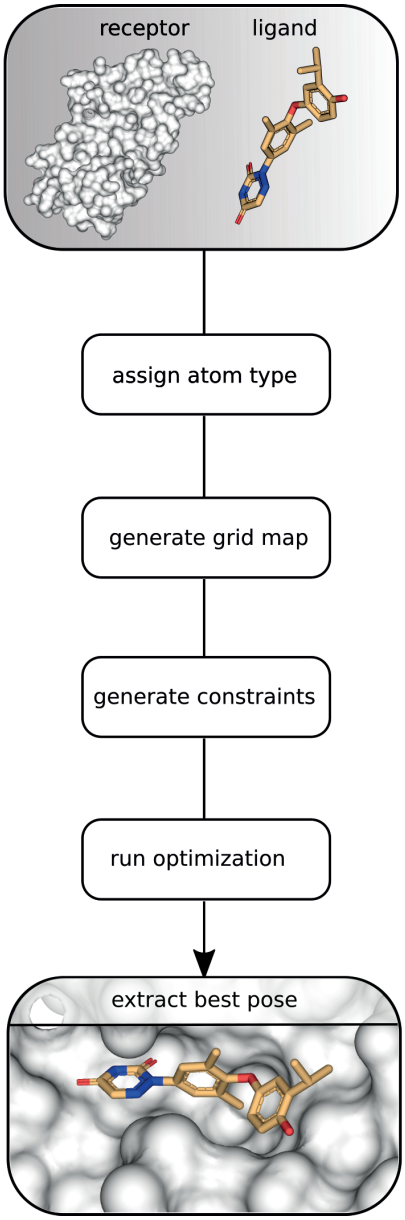



**Figure 5.1 | Constrained ligand docking workflow.**

The *pow*$^{er}$*-mViE* molecular docking method takes as input the isolated structure of the receptor and the ligand. Atom types are assigned to both receptor and ligand structures based on the Xscore nomenclature using MGLTools (http://mgltools.scripps.edu/). Then, grid maps recapitulating the energy of interaction between each ligand atom type and receptor atoms are computed and constraints used to guide the ligand docking into favorable regions of the binding site are extracted. A constrained optimization is then run with the aim of minimizing an objective function estimating binding energy similar to that of AutoDock Vina, while ensuring the satisfaction of the constraints computed in the previous step. At the end of the optimization run, the receptor-ligand associated with the lowest estimated binding energy that satisfies all the constraints is returned.

### 5.2.1 The constrained optimization strategy

The constrained optimization procedure that minimizes the energy function while satisfying the constraints was encoded within the *pow*$^{er}$ framework coupled with the constrained optimizer *mViE*.

Briefly, *mViE* advances a population of solutions based on (1+1)-CMAES to locally explore the search space and recombines these local units using a differential evolution (DE) operator to perform a global search [83]. In this context, *mViE* attempts to find candidate solutions that satisfy pre-defined inequality constraints using the concept of viability evolution [83,101], which is an abstraction of artificial evolution and aims at selecting the promising candidate solution by adapting boundaries, termed viability boundaries, set around the inequality constraints. Initially relaxed around the constraints, these viability boundaries are gradually tightened during optimization. Only the solutions not violating these boundaries are termed viable and are selected for the next iterations. This has the effect of gradually driving the solutions towards "feasible" areas of the search space where the constraints are satisfied and the objective is minimal [101]. The minute workings of *mViE* can be found in the work of Maesani et al. [83] and is also summarized in **Chapter 2**.

### 5.2.2 General formulation of the docking problem

In the context of molecular docking, *pow*$^{er}$*-mViE* was used to sample the position, orientation and flexibility of the ligand encoded in a vector of design variables $X$ = [*x, y, z, α, β, γ, ens*], where *x, y* and *z* correspond to the three translations, *α, β* and *γ* to the three Eulerian angles defining the ligand orientation and *ens* the ligand conformational ensemble generated by sampling torsional angles of the ligand rotatable bonds, in order to solve the docking problem generally formulated as:

$$min\ f(X),\ s.t. \begin{cases} l_i \leq X_i \leq u_i,\ i = 1,2,\dots,n \\ g_j(X) \leq 0, \quad j = 1,2,\dots,m \end{cases} \qquad \textbf{(1)}$$

where $f(X)$ is the **objective function** to be minimized, $l_i$ and $u_i$ the lower and upper boundary ranges defining the search space of the variable $X_i$, and $g_j(X)$ the **inequality constraints** defined on each solution $X$.

### 5.2.3 Objective function

The objective function featured in AutoDock Vina [76] was integrated in the new *pow$^{er}$-mViE* molecular docking method due its proven robustness, and detailed description. Given the atomistic pairwise interaction term $d_{ij} = r_{ij} - R_{ti} - R_{tj}$, where $r_{ij}$ is the measured Euclidian distance between the $i^{th}$ and $j^{th}$ atoms and $R_t$ the van der Waals radius of these atoms of type $t$, the objective function *f(X)* is defined as:

$$AutodockVina\ f(X) = \begin{cases} w_1 * gauss_1 + \\ w_2 * gauss_2 + \\ w_3 * repulsion + \\ w_4 * Hydrophobic + \\ w_5 * Hydrogenbonds \end{cases} \quad (2)$$

where

$$gauss_1(d) = e^{-(\frac{d}{0.5})^2} \quad (3)$$

$$gauss_2(d) = e^{-((d-3)/2)^2} \quad (4)$$

$$repulsion(d) = \begin{cases} d^2, if\ d < 0 \\ 0, if\ d \geq 0 \end{cases} \quad (5)$$

$$Hydrophobic(d) = \begin{cases} 1, if\ d < 0.5 \\ 1.5 - d, if\ 0.5 < d < 1.5, \\ 0, if\ d > 1.5 \end{cases} \quad (6)$$

$$Hydrogenbonds(d) = \begin{cases} 1, if\ d < -0.7 \\ -\frac{d}{-0.7}, if -0.7 < d < 0 \\ 0, if\ d > 0 \end{cases} \quad (7)$$

In the AutoDock Vina scoring function (**eq. 2**), the steric terms from **eq. 3**, **eq. 4** and **eq. 5** essentially reproduce the attractive and repulsive parts found in a standard Lennard-Jones potential and are

applied to every atom irrespective of their type. Instead, the hydrophobic term depicted in **eq. 6** is applicable only when both atoms involved in the interaction are hydrophobic. Similarly to **eq. 7**, the hydrogen bonding term is only applied whenever one of the atoms participating in the interaction is a hydrogen bond acceptor and the other a hydrogen bond donor. The coefficients $w_1$, $w_2$, $w_3$, $w_4$, $w_5$ were respectively calibrated at -0.0356, -0.00516, 0.840, -0.0351 and -0.587 in the original study by Trott and Olson [76] to balance the contribution of the steric terms.

The choice for the Autodock Vina energy function as the objective implemented in *pow$^{er}$-mViE* was based on the fact it was open source, well-document and showed remarkable results when integrated into other molecular docking programs [162,171,172]. Nevertheless, given the flexibility of the *pow$^{er}$-mViE* protocol, in principle any objective function can be integrated or developed.

### 5.2.4 Constraint terms

The information encoded in the physicochemical interactions between the receptor and ligand was used as constraints to guide the docking procedure. Precisely, inequality constraints indicating the ideal ligand position and conformation inside the binding site were extracted as spatial distances from pre-computed grid maps of interaction energies for different atom types (Figure 5.1).

Assuming the knowledge of the receptor binding site location, a cubic lattice box was first generated at that location with evenly spaced voxels (Figure 5.2**,** panel 1). Voxels found at a Euclidian distance too close to any receptor atoms were removed, therefore leaving only non-overlapping voxels, from which two types of constraints were extracted: local and global.

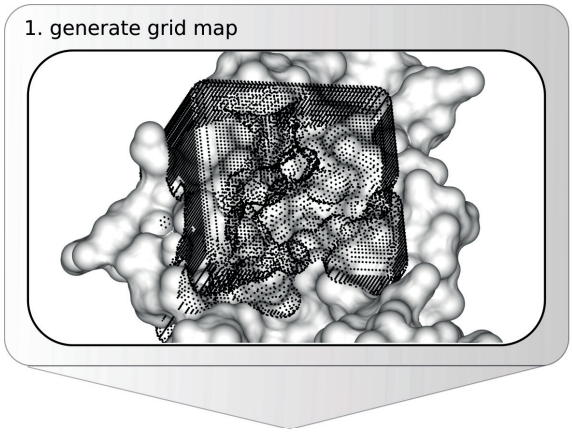

1. generate grid map
2. extract non-overlapping voxels
3. get energy (kcal/mol)
H-bond donors
H-bond acceptors
Hydrophobic beads
combine
a. Local constraints
b. Global constraints
4. dock ligand by satisfying constraints

**Figure 5.2 | Constraint generation from energy grid maps.**

**1.** A grid map represented by a cubic lattice box is generated at the location of the receptor active site. **2.** Non-overlapping voxels from the lattice box are extracted by removing voxels found too close to the receptor atoms **3.** The energy quantifying the interaction between the non-overlapping voxels and receptor atoms is recorded, for each atom type. Voxels are ranked based on their interaction energy calculated by **eq. 2**. The lowest energy voxels are extracted and used in two different ways (a-b). **a.** After clustering, the voxel centroids are used as local constraints during docking by acting as docking "anchors". **b.** Alternatively, the lowest-energy voxels for each atom type are combined based on their coordinates and are converted as global anchors. **4.** The ligand is docked to the binding site by minimizing the objective function estimating the binding energy while simultaneously satisfying the global and local constraints. For the global constraint satisfaction, the ligand is spatially constrained into the cloud of voxels. Instead, for the local constraint, each ligand atom is spatially anchored to the voxel of the same atom type (same color in figure).

Local constraints assist the docking procedure by locally anchoring the ligand atoms to predicted grid points of the same atom type (Figure 5.2, panel a). From the atomistic structure of the ligand, atom types are defined based on the Xscore nomenclature[170] and include five atom types: hydrophobic (H), hydrogen bond acceptor (A), hydrogen bond donor (D), polar (P) and metal (M). Probes, each with the property of the ligand atom types, are iterated over each non-overlapping voxel and the energy of interaction between the atom-type probe and the receptor atoms is evaluated using **eq. 2** (Figure 5.2, panel 3). The non-overlapping voxels associated with the lowest interaction energies are extracted and then subjected to clustering. The resulting cluster centroids are then used as anchors during docking (Figure 5.2**,** panel 4) in the form of local inequality constraint of the type $LocalAnchorDist_j \leq Ldist \ \ (j = 1,2, \dots, m)$ where the $LocalAnchorDist$ is the smallest Euclidian distance measured between one of the ligand atom and a centroid with a similar atom type, for all the atom types *m* of the ligand, and *Ldist* is a cutoff distance expressed in Å. In this case, the constraints are considered satisfied whenever at least one ligand atom of each atom type is found below the cutoff distance *Ldist* of any anchoring centroid with the same atom type (Figure 5.2**,** panel 4).

Global constraints were computed using all the non-overlapping voxels to extract a cloud of grid points (Figure 5.2**,** panel b). To extract local constraints, a clustering was performed on the lowest energy non-overlapping voxels (Figure 5.2, panel a). For global constraints, the lowest energy voxels were combined, based on their 3D coordinates, to form a cloud of grid points located inside the binding site and to be used as a global anchor (Figure 5.2, panel 3 and b). The main idea behind this approach was to derive spatial distances that would constrain the ligand inside the binding site where the energies were minimal. Practically, this was done with the global inequality con-

straint $GlobalAnchorDist \leq Gdist$ Å, where the $GlobalAnchorDist$ is the average Euclidian distance computed from all ligand atoms and their respective nearest voxels located in the cloud of grid points, and *Gdist* the cutoff distance (Å) below which the constraint is considered satisfied. This leads us to the formal definition of the molecular docking problem as:

$$min\ AutodockVinaf(X), s.t. \begin{cases} l_i \leq X_i \leq u_i,\ i = 1,2,\dots,n \\ LocalAnchorDist_j \leq Ldist, j = 1,2,\dots,m \\ GlobalAnchorDist \leq Gdist \end{cases} \quad (\mathbf{8})$$

### 5.2.5 Training the energy grid map parameters to obtain optimal constraints

The cutoff distance values associated with *Ldist* and *Gdist* were determinant for the quality of the local and global constraints as defined in **eq. 8** because cutoff distances that are too high would not be specific enough to constrain the ligand in a desirable area of the binding site, and conversely cutoff values that are too low might be too specific and be satisfied only by a subset of receptor-ligand complexes. In turn, the values of *Ldist* were dependent on the parameter values related to the construction and extraction of grid voxels.

Thus, a comprehensive training benchmark was undertaken to find a set of grid parameter values that minimized *Ldist* and *Gdist* whilst ensuring they were satisfied in most receptor-ligand complexes. A simple combinatorial approach was performed to find the optimal grid parameter values, which were: the distance cutoff that defined an overlap between grid voxels and receptor atom, the number of low-energy voxels used to compute the constraints and the distance separating each voxel of the grid map. Precisely, we combined values ranging from 2.6 to 3.3 Å for the distance cutoff defining an overlap between grid voxels and receptor atom, values ranging from 500 to 2000 voxels for the number of low-energy voxels recorded; and values ranging from 0.375 Å to 1.5 Å defining the voxel spacing of the grid map.

For this benchmark, the receptor-ligand complexes were extracted from the PDBbind core set [168], which was used as the training set and contained 195 high-resolution crystal structures compiled particularly for the evaluation of docking methods. For each combination of these three parameter values, a grid map was generated on bound receptor-ligand complexes, the value of

$LocalAnchorDist$ and $GlobalAnchorDist$ were recorded and a rigid docking was performed to evaluate the quality of the constraints.

### 5.2.6 Step-by-step docking protocol

As a first assessment and in effort to validate the improvement provided by the constraints, docking was performed on the training set (PDBbind core set) considering both the ligand and the receptor as rigid structures. In this case, the location of the binding site was known. Only the binding pocket residues (protein residues within 10 Å from the ligand center of mass) were used as the receptor instead of the whole protein to speed up the calculation. In order to simulate a realistic docking case, the geometric center of the binding site (represented by the bound ligand center of mass) was randomized. The ligand was then removed and its orientation and position were randomized. For both ligand and receptor, the atom types were assigned using the MGLTools (available at http://mgltools.scripps.edu/). The boundaries delimiting the ligand search space were represented as a cubic box of 23.0 Å, which center corresponded to the randomized binding site geometric center. For comparative purposes, the number of function evaluations attributed to each receptor-ligand docking was consistent between docking methods and approximated the number function evaluation attributed to a standard rigid docking performed by AutoDock Vina, which was set to 120,000 function evaluations.

A second, more unbiased and exhaustive, round of assessment was undertaken on the Astex diverse set [169], which contained 85 docking cases of high resolution receptor-ligand complexes. As the grid parameters and constraints were not trained on this set, it served as the basis for an unbiased test. Moreover, unlike the previous assessment where both the ligand and the receptor were treated as rigid bodies, here the flexibility of the ligand was included in the search space in the form of a conformational ensemble. To account for the larger search space, the size of the box defining the search space and the budget of function evaluations were increased to 25.0 Å and 200,000 respectively. Eventually for all optimization methods, the average CPU time required to compute a docking on a single-threaded execution (For AutoDock Vina: AutoDock Vina parameter --cpu 1) was recorded and performed with a Xeon E3-1200 3.6 GHz.

### 5.2.7 Generation of ligand conformation ensemble

Here, the flexibility of the ligand was generated in isolation from the target binding site and added to the ligand search space during optimization. To address this task, we implemented a systematic rotatable bond sampler to generate ligand conformers.

The aim of the suggested implementation was to systematically generate all possible energy favorable conformations of the ligand and eventually add this ensemble of poses, treated as a conformational database, to the optimization search space. Specifically, for a given ligand, the number of active rotatable bonds was extracted using the MGLTools. Then, discrete angle increments of 120° were performed for each rotatable bond. In order to filter out physically implausible ligand conformations, the internal ligand energy was computed on each atom pair separated by at least 3 covalent bonds (1-4 LJ contributions) using a simple 12-6 Lennard-Jones potential, at each step of the permutation. In this case, ligand conformers with energies > 0.0 kcal/mol were not included in the conformational ensemble. The conformers were added as an extra dimension in the optimization space and used for the assessment of the Astex diverse set.

### 5.2.8 Metrics to evaluated accuracy

At end of the docking protocol, the accuracy of the docking was evaluated. For each receptor-ligand complex, the docked ligand pose associated with the lowest interaction energy (rank_1) was extracted. In the case of *pow$^{er}$-mViE*, an additional selection criterion was that the best pose should also satisfy all the pre-defined constraints of equation (8). The position and conformation of the rank_1 ligand pose were compared to that of the original crystal structure using the standard structural similarity metric $RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$, where averaging is done over the $N$ pairs of equivalent heavy atoms and $\delta_i$ is the distance between the two atoms in the $i$-th pair. In this case, a docking was considered successful if the structure of the docked ligand was < 2.0 Å from the original crystal structure. Eventually, a success rate was issued by simply dividing the number of successful docking cases over all docking cases. This standard evaluation protocol has been in used in several studies [76,77,80] in order to benchmark and validate newly developed SMD methods.

## 5.3 Results and discussion

The molecular docking method suggested herein takes inspiration from a recent study in which the prediction of macromolecular assemblies was modeled as an optimization problem[17] (See **Chapter 3**). Given a receptor molecule that is held fixed in space and a ligand that is moved around the receptor, our in-house docking protocol *pow$^{er}$-mViE* attempts to minimize a scoring function describing the quality of candidate poses by sampling the rotation, orientation and conformation of the lig-

and, while simultaneously satisfying pre-defined local and global constraints describing the ideal position of the ligand inside the binding site.

### 5.3.1 Unconstrained rigid docking

As a first exploratory step, we used *pow$^{er}$-mViE* to rigidly dock each of the 195 ligands extracted from the PDBbind dataset into their respective binding sites in an unconstrained setting, that is, by minimizing solely the objective function defined in **eq. 2**.

This unconstrained molecular docking enabled the correct docking of 36/195 receptor-ligand complexes (accuracy: 18.5%, Heavy atom RMSD < 2.0 Å, Figure 5.3a). As an attempt to understand the cause of such a low accuracy, several ligand poses were generated at RMSD distances ranging from 0.0 to > 10.0Å, from the bound ligand, then ranked by an energy score calculated using **eq. 2,** for each of the 195 PDBbind receptor-ligand complexes. In this case, the energy function was able to correctly rank 91% of the receptor-ligand poses for all the PDBbind complex. Thus it was clear the low docking accuracy obtained with the unconstrained *pow$^{er}$-mViE* protocol was due mostly to the optimizer inefficiency rather than the inaccuracy of the energy function. In fact the energy landscape associated with this objective function contained several local minima, which complicated the docking process. The difficulty to find an efficient optimization protocol was also reflected in the original study describing the AutoDock Vina implementation [76] in which several state-of-the-art optimizers were tested to efficiently minimize the scoring function.

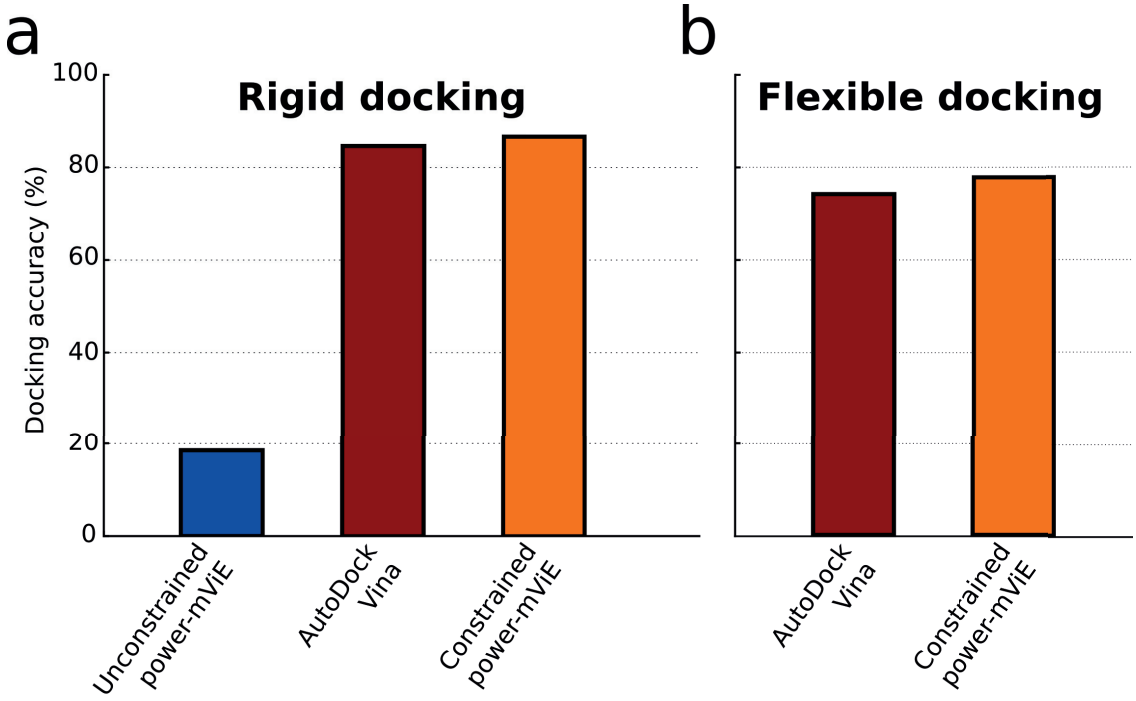



**Figure 5.3 | Comparison of Docking accuracies between *pow$^{er}$-mViE* and AutoDock Vina.**

**a.** Accuracies were computed from the 195 receptor-ligand complexes found in the PDBbind core set, in rigid docking setting. **b**. Reported here are the maximal accuracies obtained by flexibly docking 84 complexes of the Astex diverse set using 5 different sets of random geometric centers (see section **5.3.3**).

### 5.3.2 Constrained rigid docking

In this work, we envisioned to switch from an unconstrained optimization paradigm to a constrained one by supplementing the AutoDock Vina objective function (**eq. 2**) with local and global constraints, where the aim was to "anchor" the ligand to desirable regions of the search space. Briefly, these local and global constraints were extracted from grid maps in the form of a cubic lattice box where atom-type specific energies were computed on each grid voxel (see Methods). The local constraints consisted of voxel centroids extracted by clustering the lowest energy voxels according to their atom types (Figure 5.2**,** panel a) and the global constraint was obtained by combining the coordinates of the lowest energy voxels into one global cloud of grid voxels (Figure 5.2**,** panel b). Local constraints were considered satisfied whenever at least one ligand atom of each atom type was found below the cutoff distance *Ldist* of any anchoring centroid of the same atom type. The satisfaction of global constraints was computed by making sure the average distance between all the ligand atoms and their nearest cloud voxel was smaller than the global cutoff value *Gdist*.

Each constraint type (local and global) required the computation of distinct grids, each associated with different parameter values. The cloud of voxels defining the global constraint required a voxel spacing of 0.375 Å, a cutoff distance defining an overlap between voxel and receptor atoms of 2.8 Å and fixed number 500 low-energy voxels to be combined based on their 3D coordinates, for each atom type. In this respect *Gdist* was calibrated at 1.5 Å. Conversely, the extraction of the centroid voxels defining the local constraints required a voxel spacing of 1.0 Å, a cutoff distance defining an overlap between voxel and receptor atoms of 3.0 Å and fixed number 600 low-energy voxels to be clustered, for each atom type. In this case, the cutoff distance *Ldist,* applied to all atom-type voxels, was calibrated at 2.0 Å.

The optimal set of parameters was obtained through an exhaustive rigid docking benchmark on the PDBbind core set consisting in 195 receptor-ligand complexes where the docking accuracy (86.7%) was on par with the accuracy obtained with AutoDock Vina (84.6%, Figure 5.3, left panel). Noteworthy, the AutoDock Vina docking program was trained on receptor-ligand complexes from the PDBbind refine set [76]. In the study describing the implementation of AutoDock Vina, the authors recorded an accuracy of 78% for the flexible docking of 190 complexes, compared to accuracies >80% for rigid docking, as obtained in this study. More generally, docking accuracies from rigid-body docking methods are higher compared to flexible docking ones [155], since rigid docking problems have a smaller search space with only six degrees of freedom consisting of three translations and three rotations. Several Fast-Fourrier-Transform (FFT)-based programs currently exist which specialize in ligand docking in a rigid setting [155,173] with greatly reduced computation time when compared to molecular docking programs using stochastic optimization such as *pow$^{er}$-mViE*.

Another way to speed up computation is to use pre-computed grid maps that store the interaction energies between ligand and receptor and subsequently simplify the energy calculation during docking[174]. Such approach was used by the AutoDock program with the AutoGrid feature [161] as well as ICM [80]. Grid maps have also been used to detect the location of the binding pocket [175,176].

In this work, the pre-computed grid maps were not used to accelerate the receptor-ligand energy calculation nor to find potential binding sites. Rather they were used to derive a set of local and global constraints to guide the docking process. Importantly, we found that supplementing the objective function with these local and global constraints (**eq. 8**) considerably improved our initial docking accuracy from 18.5% (unconstrained *pow$^{er}$-mViE*) to 86.7% (constrained *pow$^{er}$-mViE*).

### 5.3.3 Constrained flexible docking

The constrained molecular docking approach was further extended to flexibly dock ligands inside their respective binding sites. In this case, the receptor was still kept rigid while the ligand flexibility was sampled. We tested this approach on the Astex diverse set[169], which contained 85 high resolution receptor-ligand complex structures. In the previously described rigid docking benchmark, a single randomly chosen binding site geometric center was used to train the grid parameter values as well the constraints cutoff values *Ldist* and *Gdist*. This was done due to the relatively long computation time required to complete the benchmark and one could argue that the parameters obtained via this benchmark could be biased towards these specific geometric centers. Therefore, in order to evaluate how the location of the geometric center affected the outcome of the $pow^{er}$-*mViE* molecular docking, 5 different sets of the 85 complexes were created, each with different binding site geometric centers. For each of the 5 sets, geometric centers were randomized from the ligand center of mass with continuous values ranging from ± 2.0 Å in the *x*, *y* and *z* plane. For each of the 5 sets, 5 trials were performed to increase the statistics, giving a total of 25 * 85 docking trials. The *Ldist*, *Gdist* and grid parameters were the same as those extracted from the rigid docking benchmark.

Ligand flexibility was encoded as a pre-generated ensemble of ligand conformers obtained through incrementing the torsional angles in a combinatorial approach (see Methods). This conformer ensemble was added as an extra search space dimension and contained the crystal conformation of the ligand. The accuracy computed from this flexible docking method was compared to the one obtained using AutoDock Vina in exactly the same conditions (i.e. geometric center location, starting from bound ligand structure, equal number of function evaluations) and are shown in Table 1 and Figure 5.3**,** right panel.

**Table 2 | Flexible docking of Astex diverse set and reported accuracies for *pow$^{er}$-mViE* and AutoDock Vina respectively.**

| | | ***power-mViE*** | | ***AutoDock Vina*** | |
|---|---|---|---|---|---|
| | trial number | accuracy (%) | mean accuracy(%) | accuracy (%) | mean accuracy (%) |
| random center 1 | 1 | 72.6 | | 73.8 | |
| | 2 | 66.7 | | 71.4 | |
| | 3 | 73.8 | 72.3 ± 3.0 | 70.2 | 71.7 ± 1.6 |
| | 4 | 73.8 | | 72.6 | |
| | 5 | 70.2 | | 70.2 | |
| random center 2 | 1 | 69.0 | | 73.8 | |
| | 2 | 76.2 | | 73.8 | |
| | 3 | 75.0 | 71.2 ± 4.1 | 72.6 | 72.6 ± 1.2 |
| | 4 | 67.9 | | 71.4 | |
| | 5 | 67.9 | | 71.4 | |
| random center 3 | 1 | 69.0 | | 71.4 | |
| | 2 | 66.7 | | 71.4 | |
| | 3 | 71.4 | 68.6 ± 2.0 | 73.8 | 71.9 ± 2.0 |
| | 4 | 66.7 | | 69.0 | |
| | 5 | 69.0 | | 73.8 | |
| random center 4 | 1 | 77.4 | | 70.2 | |
| | 2 | 76.2 | | 72.6 | |
| | 3 | 67.9 | 72.4 ± 4.2 | 73.8 | 72.1 ± 1.4 |
| | 4 | 70.2 | | 71.4 | |
| | 5 | 70.2 | | 72.6 | |
| random center 5 | 1 | 69.0 | | 69.0 | |
| | 2 | 71.4 | | 70.2 | |
| | 3 | 70.2 | 71.4 ± 1.9 | 70.2 | 70.5 ± 1.0 |
| | 4 | 73.8 | | 71.4 | |
| | 5 | 72.6 | | 71.4 | |
| **overall mean accuracy (%)** | | **71.0 ± 3.2** | | **71.8 ± 1.5** | |
| **overall max accuracy (%)** | | **77.4** | | **73.8** | |
| **overall min accuracy (%)** | | **66.7** | | **69.0** | |

In this evaluation, we found that the results obtained using *pow$^{er}$-mViE* protocol matched those obtained with AutoDock Vina (Figure 5.3**,** right panel). Importantly, the location of the geometric center used to compute the constraints did not seem to influence the accuracy of the flexible docking (Table 2) with an average accuracy estimated at 71.0% ± 3.2% for all the 25 docking trials. These results were found comparable with those obtained using AutoDock Vina with respective geometric centers (mean accuracy: 71.8% ± 1.5%). Notably, the maximal accuracies obtained both for AutoDock Vina (73.8%) and *pow$^{er}$-mViE* (77.4%) flexible docking protocols (Table 2) were comparable to the AutoDock Vina accuracies (76.5%) described in the study by Zoete et al. [77] obtained with

similar conditions. From the Astex diverse set, used to validate the method, 5 receptor-ligand complexes which included 1GPK, 1HNN, 1N1M, 1N2V and 1OYT were also present in the PDBbind core set, used to train the grid parameters. Removing these 5 docking cases when computing the overall accuracy did not lead to a different overall mean accuracy (70.2% ± 3.2%). On the same Astex diverse set and starting from the bound ligand conformation, AC obtained docking a maximal accuracy of 88.2%[77], which were higher than those obtained both by the *pow*$^{er}$*-mViE* (77.4%) and AutoDock Vina (73.8%) in this work. When starting from a random conformer, the highest ranking molecular docking programs were ICM with 91%[80], GOLD with 87%[82] and AC with 83.5 %[77].

The results presented here have to be reflected with care since the strategy to sample ligand flexibility was different between AutoDock Vina and *pow*$^{er}$*-mViE*. Precisely, AutoDock Vina samples the angles of each ligand torsion during optimization while *pow*$^{er}$*-mViE* assigns a structure from a pre-computed conformer ensemble [76]. The strategy to sample from a pre-computed set of ligand conformers through a systematic search is a conceptually simpler method compared to sampling each of the ligand torsions [153,154].

Nevertheless when generating conformers of the 1YGC ligand, which contained 14 active torsions, by systematically incrementing each torsional angle by 120°, memory failures occurred. Such inconvenience has already been reported in the literature [154] and is called a "combinatorial explosion" problem where the memory required to store and sample a large number of conformers far exceeds the available computer memory. For this reason, 84 complexes instead of the 85 featured in the Astex diverse set were taken into account and presented in Figure 5.3 and Table 2. When attempting to reduce the angular increments to 60° or 35°, memory errors were encountered that prevented to further test the method. Thus we did not evaluate the robustness of our method when starting from random ligand conformation because it required a finer angular increment.

A possible solution to circumvent this problem could be to incrementally grow the ligand into the binding site[177]. This method in particular was applied by the Dock [178] and FlexX[179] programs and involves iteratively docking ligand fragments into the binding site then covalently linking them. Implementing a similar sampling strategy for the next molecular docking implementation of *pow*$^{er}$*-mViE* would certainly alleviate the combinatorial memory problem and increase the robustness of the approach.

In term of average CPU time required for each of the 84 flexible docking cases, AutoDock Vina was faster (50s ± 40s) than *pow*$^{er}$*-mViE* (5min ± 42s). In the present analysis where a limited

amount of docking rounds were performed, such time differences do not significantly affect the usability of *pow*$^{er}$*-mViE*. However, in docking cases where a large number of compounds need to be virtually screened, *pow*$^{er}$*-mViE* would be less competitive than Autodock Vina. Being the first implementation of *pow*$^{er}$*-mViE* for ligand docking, we expect to reduce the computational time associated with docking in the future implementation of the method, either by a better engineering of the docking algorithm or by incorporating multi-parallel optimization.

### 5.3.4 Evaluation of successes and failures

For each of the 84 flexible docking cases featured in the Astex diverse set, the number of successful trials (i.e. where the best ligand pose is found at RMSD < 2.0 Å of the bound ligand) was computed over the 25 trials. In 44 receptor-ligand docking cases, both AutoDock Vina and *pow*$^{er}$*-mViE* consistently found the native ligand binding mode with success rates ≥ 76% (20/25 trials).

Conversely, in 10 docking cases which comprises complexes 1GM8, 1HVY, 1JD0, 1JJE, 1Q41, 1R58, 1SQ5, 1TZ8, 1W1P and 1UVF, neither AutoDock Vina nor *pow*$^{er}$*-mViE* could consistently find the ligand binding mode as found in the native receptor-ligand complex (success rates ≤ 20%, 5/25 trials). Upon visual inspection of the receptor-ligand complex structures and searching through the literature, it appeared that some of these cases were in fact "usual suspects" previously labeled as difficult docking cases [77,80]. The reason behind such difficulty lies in the presence of explicit water molecules mediating critical receptor-ligand hydrogen bonding interactions in the native complex structure which are removed prior to docking. This was attributed as the root cause for the docking failures of 1GM8, 1HVY, 1JD0, 1SQ5 and 1W1P. Another reason for failure is due to the ligand interaction with metal ions as in complexes 1JD0 or 1JJE. The reason for the failure of other docking cases could not be detected through the visual inspection of the complex crystal structure and thus was attributed to limitations in the scoring function accuracy.

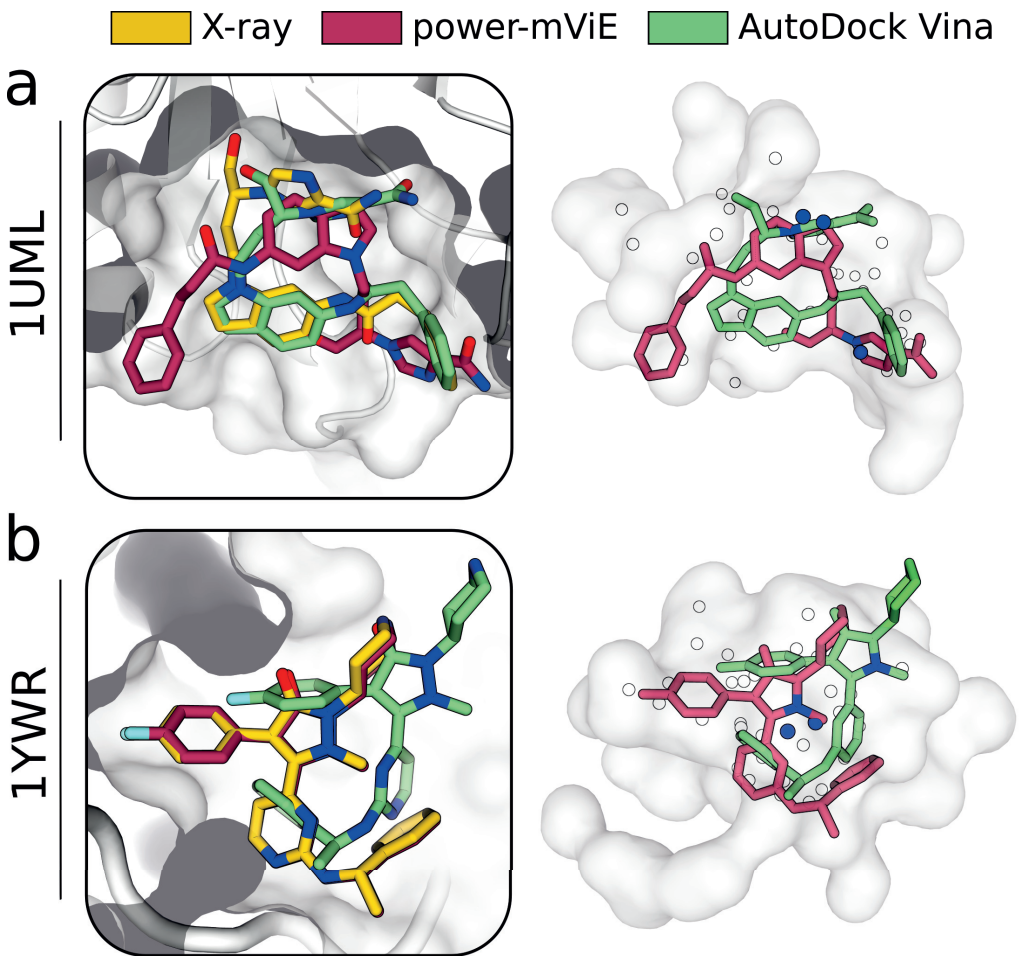



**Figure 5.4 | Docking success and failures of AutoDock Vina and *pow$^{er}$-mViE*.**

**a.** Best ligand poses returned by the *pow$^{er}$-mViE* and AutoDock Vina methods for 1UML (left panel) and associated local and global constraints satisfaction (right panel). **b**. Best ligand poses returned by the *pow$^{er}$-mViE* and AutoDock Vina methods for 1YWR (left panel) and associated local and global constraints satisfaction (right panel). For the representation of constraint satisfaction for both 1UML and 1YWR (right panels), the white cloud surface represents the global constraint, the full blue circles represent the satisfied local polar centroids (Ldist < 2 Å), which are located close to polar ligand atoms also colored in blue, and the empty white circles represent unsatisfied polar centroids.

More interesting docking cases consisted of those where AutoDock Vina and *pow$^{er}$-mViE* showed conflicting success rates. For complexes 1MMV and 1UML (Figure 5.4a) in particular, AutoDock Vina succeeded with a rate of 100% for the 25 trials, while *pow$^{er}$-mViE* was always ≤ 20% (5/25 trials). Visual inspection of the native complex structures did not reveal the presence of neither water molecules nor metal ions mediating the interactions between ligand and receptors. In order to check whether the local and global constraints may have wrongly anchored the ligand into undesir-

able areas of the binding site, we computed the local and global constraint satisfaction as well the energy score, according to **eq. 2,** for the best (lowest estimated binding energy) receptor-ligand poses returned by AutoDock and *pow*$^{er}$*-mViE*. Such analysis showed that both the correctly docked lowest-energy poses computed by AutoDock Vina and the wrongly docked best poses of *pow*$^{er}$*-mViE* were found to satisfy the local and global constraints. However, the lowest energies recorded for best poses of 1MMV (-10.9 kcal/mol) and 1UML (-16.3 kcal/mol) returned by AutoDock Vina were lower than the ones computed on the crystal structures of 1MMV (-9.6 kcal/mol) and 1UML (-14.9 kcal/mol) respectively, which in turn were lower than the best pose computed by *pow*$^{er}$*-mViE* (1MMV: -8.67 kcal/mol, 1UML: -11.7 kcal/mol). This showed in these two cases that while the constraints were satisfied in the best poses returned by both *pow*$^{er}$*-mViE* and AutoDock Vina, failure to compute better poses by *pow*$^{er}$*-mViE* was probably due to the fact that it was unable to overcome a local energy minimum. A possible solution to better find the correct binding pose could be either to increase the number of function evaluations for these two cases or refine the constraints further so as to more precisely anchor the ligand into the binding site.

Conversely, *pow*$^{er}$*-mViE* correctly docked the ligand of complexes 1MEH, 1N2V, 1R55 with success rates of ≥ 84% (21/25 trials) while AutoDock Vina consistently failed to find the native binding mode (0% success rate) for these complexes. Interestingly, visual inspection of these complexes revealed that both 1MEH and 1N2V contained water molecule, and 1R55 a Zn ion bridging the interaction between the ligand and receptor protein. These have been previously noted as being particularly difficult docking cases[77,80]. Comparisons between the estimated binding energy associated with the *pow*$^{er}$*-mViE* best poses (1MEH: -10.2 kcal/mol, 1N2V: -8.4 kcal/mol, 1R55: -9.3 kcal/mol), the energies from the best AutoDock Vina poses (1MEH: -10.4 kcal/mol, 1N2V: -8.6 kcal/mol, 1R55: -10.2 kcal/mol) and crystal structures (1MEH: -9.7 kcal/mol, 1N2V: -8.0 kcal/mol, 1R55: -8.0 kcal/mol) indicated that both *pow*$^{er}$*-mViE* and AutoDock Vina found lower energies than that of the complex structure. Moreover, the local and global constraints were satisfied on the best poses returned by both *pow*$^{er}$*-mViE* and AutoDock Vina and on the three complex crystal structures. Thus, we suggest the reason for the success of *pow*$^{er}$*-mViE* in this case might be due to a combined effect of constraints anchoring the ligand to a suitable region of the binding mode and to the coarse angular step of 120° which might have prevented the ligand to adopt a conformation leading to a non-native binding mode.

The docking cases of 1YWR (*pow*$^{er}$*-mViE*: 72%, AutoDock Vina: 0%, Figure 5.4**b**) and 1G9V (*pow*$^{er}$*-mViE*: 52%, AutoDock Vina: 0%) were found to also be worth mentioning since they

illustrated how the use of constraints helped find the correct binding mode of the ligand. Visual inspection of the 1YWR crystal structure did not reveal the presence of neither water nor metal ion molecules inside the binding site. Evaluating the level of constraint satisfaction on the best poses returned by AutoDock Vina over the 25 trials showed that a local constraint specifying the minimum distance between a polar centroid voxel and a polar atom of the ligand was always > 2.0 Å which is above the *Ldist* cutoff value, and thus violates that local constraint (Figure 5.4b). Unlike the AutoDock Vina best poses, both the crystal structure 1YWR and the best *pow$^{er}$-mViE* poses were found to satisfy all the local and global constraints (Figure 5.4b). Additionally, the estimated binding energy associated with the best pose returned by *pow$^{er}$-mViE* (-13.7 kcal/mol) was lower than the ones of the crystal structure (-12.4 kcal/mol), which was in turn lower than the best AutoDock Vina pose (-12.0 kcal/mol). This remarkable example served as a showcase illustrating how the use of constraints by *pow$^{er}$-mViE* was useful in anchoring the ligand in favorable regions of the binding site where the minimal binding energy was found, leading to a pose resembling that of the native binding mode.

Anecdotic cases like those of 1MEH, 1N2V, 1R55, and particularly 1YWR put forward the usefulness of using pre-computed constraints to assist the docking of ligands inside their binding site. Noteworthy, 1MEH and 1N2V featured water- mediated ligand-receptor interactions and 1R55 contained Zn ions in the binding site. Such cases are considered difficult to solve. Nevertheless, similar water- and metal ion-mediated docking cases such as 1GM8, 1HVY, 1JD0, 1JJE, 1Q41, 1R58, 1SQ5, 1TZ8, 1W1P and 1UVF still remain elusive to most SMD algorithms including *pow$^{er}$-mViE* and AutoDock Vina. If present in the crystallographic structure of the binding site, a suggested strategy to improve the docking performance is to keep and optimize the orientation and position of the water hydrogen atoms, using specific energy minimizations [180,181]. If not present in the receptor crystallographic structure, water molecules can be added into the binding site using grid-based or molecular dynamics simulations and can be selected based on their energetic stability for docking [182]. The presence of metal ions can hinder the performance of molecular docking programs because it may affect the hydration and protonation states of charged residues[183], which are factors still difficult to predict using available methods as they often require more expensive calculation at the quantum level[184].

## 5.4 Conclusion

In the work presented here, we applied the principles of constrained optimization to assist the docking of small molecules inside the binding site of their respective targets. Precisely, local and global constraints were extracted from pre-computed energy grid maps at the location of the binding site. These constraints were essentially used during docking to anchor ligands inside desirable binding site areas by constraining their global position inside low-energy areas, and by locally matching ligand atoms to areas favouring their associated atom types.

Using *pow*$^{er}$*-mViE* on a constrained rigid docking setting, we combined these local and global constraints together with the AutoDock Vina scoring function and witnessed an almost four-fold improvement in docking accuracy compared to a rigid docking setting where an unconstrained optimization was performed. Moreover, in a flexible docking setting, we obtained accuracies comparable to state-of-the art SMD methods.

Despite the encouraging results obtained in this work, we suggest necessary improvements for the future implementation of the method. Namely, we encountered the problem known as ‘’combinatorial explosion’’ when computing the different ligand conformers used as an ensemble to be sampled during the optimization. As a possible solution to this problem and to increase the robustness of our approach, we suggest a sampling method similar to Dock [178] and FlexX[179] programs, which consists in first dividing the ligand into fragments and then individually docking the fragments by respecting their covalent bonding.

Similar to other SMD methods [77,80], we faced difficulties in docking cases where water or metal ions were mediating the interaction between ligand and their receptor. A suggested approach to improve docking accuracies when water mediates ligand-receptor interactions could be to first to refine the hydrogen atom positions of the water molecules found in the complex crystal structure, then to use these waters as part of the receptor structure during docking [180,181]. When not present in the complex crystal structure, water molecules can be added using molecular dynamics-based techniques, which assess their relative positional retention inside the binding site [182]. As for docking cases where protein-ligand interactions are mediated through a metal ion, more refined techniques are suggested which include calculation at the quantum level [184].

Finally, we envision this approach to be advantageous in more difficult, but more realistic docking cases such as those where both the ligand and receptor flexibility are sampled during optimization. Instead of using a single receptor conformation, the inclusion of receptor flexibility in

SMD holds promise for increasing docking accuracy and capturing a better physical representation of the ligand-receptor interactions [160,185]. In the best-case scenario, receptor flexibility can be accounted for by including different experimentally solved structures of the protein target in the SMD process, determined for instance via X-ray crystallography or NMR. However in more realistic docking cases where only a single conformation of the receptor is available, receptor flexibility can be sampled using *in silico* methods. A popular way to treat receptor flexibility using *in silico* methods consists in using amino-acid side-chain rotamer libraries [186], which locally samples the binding site flexibility based on experimentally observed amino-acid side-chain conformations. This approach however has limited use in docking cases where the binding site is expected to undergo substantial backbone conformational change in order to accommodate the ligand [153], e.g. in loop-forming active sites [187] or cryptic pockets [188]. For such flexible cases, more expensive techniques, including molecular dynamics simulations, can be used on the receptor in its apo form to generate different snapshots of binding site conformations. This ensemble of binding site conformations can be used directly during docking, as in ICM (4D docking) [158,189] and is expected to provide not only significant enrichment in VS but also more chemically diverse hits.

# Chapter 6 Conclusion and perspectives

Partly adapted from the published paper: "**Assessment of data-assisted prediction by inclusion of crosslinking/mass-spectrometry and small angle X-ray scattering data in the 12th Critical Assessment of protein Structure Prediction experiment**" Giorgio E. Tamò, Luciano A. Abriata, Guilia Fonti and Matteo Dal Peraro. *Proteins: structure, function and bioinformatics*., 2018

Integrative modeling (IM) techniques have emerged as powerful tools to model the architecture of important proteins [43,49,62,190] by combining computational methods with experimental information. Such experimental inputs can take the form of crosslink/mass-spectrometry (CLMS) [11,67,140] experiments that can be used to pinpoint contacting residues, greatly helping the modeling process of protein structure. Another important experimental input consists in low-resolution structural data describing shape, volume or subunit arrangement such as small-angle X-ray scattering (SAXS) data or cryo-electron microscopy (EM).

In particular, the last decade bears witness to remarkable advancements in the field of cryo-EM, which mainly consisted in the use of direct electron detector and better image processing methods [1]. Such advances have recently allowed the structural elucidation of larger and larger protein macromolecules at atomistic resolution (<4 Å) [3,5,8].

Through the integration of the previously mentioned heterogeneous experimental information together with biophysics-based energy potentials, the ultimate goal of IM techniques is to model native-like atomistic structures of important macromolecular assemblies [18]. Subsequently, the evolution of the IM methods closely follows the progress and advances of the low-resolution experiments they rely on to model their predictions. In this respect, the final accuracy, in term of structural similarity to the native structure, conferred to a macromolecular structure predicted using IM techniques would be strongly dependent on the accuracy depicted in the integrated experimental data.

Therefore, whenever the experimental techniques commonly used to assist the integrative modelling of macromolecular structures gain sufficient accuracy to elucidate their structure at atomistic level, the use of IM approach for such modelling problems should be carefully evaluated. As a striking example to illustrate the previous point, during the present work to model the large Q23-

Httn monomeric structure from a 8 Å cryo-EM map and CLMS data (**Chapter 4**), an independent study [87] described cryo-EM experiments that enabled the resolution at 4 Å of the Q17-Httn structure. Fortunately in this case, we could still proceed from our previous work to predict the monomeric Httn structure since the solved 4 Å Httn structure was stabilised through binding a HAP40 protein, where Httn likely changed conformation in order to bind HAP40.

More generally however, the acceleration of cryo-EM-based macromolecular structure determination suggests a gradual shift of focus from the IM community. Instead of targeting the structural elucidation of large macromolecules that are getting increasingly more reachable by cryo-EM experiments, IM efforts would gradually shift towards larger macromolecular, even cellular, assemblies such as those now being elucidated by cryo-electron tomography with resolution averaging 15 Å [191]. Coming together with an increase in size, the complexity of such large systems would require IM method to be able to integrate more heterogeneous experiments and to better treat protein flexibility.

Still currently, state-of-the art methods applied to predict both large and small molecular complexes usually rely on carefully calibrated scoring functions to quantify the quality of their structural models. Importantly here, the scoring function featured in these methods often contain uncorrelated term which are linearly added and which relative contribution have been balanced by a tedious weight assignment process. Moreover, this weight assignment process often does not permit the addition of other uncorrelated terms without needing to rebalance the contribution of all the other terms.

In this work, we suggested the use of constrained optimization as a way to alleviate the previously mentioned issues, which are omnipresent in multiscale integrative modeling. Precisely, we incorporated within our in-house integrative modeling protocol *pow*$^{er}$ a novel and robust constrained optimizer called memetic Viability Evolution (mViE). The application of this constrained optimization protocol for multiscalte integrative modeling, nicknamed *pow*$^{er}$*-mViE*, was achieved in various projects which are described in the chapters of the thesis. Essentially, our constrained IM approach enabled the optimization of terms related to the global scoring scheme separately without re-balancing their contribution. More importantly, it enabled to seamlessly add any term to the scoring function without rebalancing the relative contributions.

Although the work described here has been applied to model both large (**Chapters 3 and 4**) and fine (**Chapter 5**) molecular systems, we predict its usefulness would extend to model larger

biological systems such as those requiring the integration of several heterogeneous experimental data.

The *in silico* elucidation of such larger assemblies will however require benchmarks and methods to accurately evaluate the robustness of state-of-the-art IM methods. To date, there seems to be a lack of rigorous assessement protocols to validate IM methods used to predict the structure of large asymmetric macromolecules (>50 kDa) by integrating low-reslution experiments such as cryo-EM, SAXS or CLMS. Remarkable efforts have recently been noticed from the protein modeling community with the inclusion of data-assisted categories in the CASP11 and CASP12 experiments, in which NMR, CLMS and SAXS data were used to assist the modeling of protein structures. Although highly promising, it is still in relative infancy and does not include large protein structures which size >300kDa, such as the Httn protein (~350 kDa). We believe the addition of a data-assisted category for modeling protein of this size would be beneficial for the future, which brings the challenge of solving the structure of bigger and bigger proteins.

Furthermore, the work featured in this thesis has triggered the development of other projects related to the modeling of protein structures. One of these projects consists in fitting largely asymetric protein assemblies (more than four asymmetric protomers) inside low-resolution cryo-EM maps (>10 Å). This type of modeling problem is still very difficult to solve today due to the very large degrees of freedom associated with sampling both the roto-translations and flexibility of each of the assembly protomers performed through optimization [192]. In the same spirit as in **Chapter 5**, where voxel centroids were used to constrain the search space of small molecules inside their receptor binding site, the idea for accelerating the fit of asymmetric assemblies would be to pre-generate and detect, from the experimental EM map, areas of high electron densities to be used as anchors by the protomeric subunits during docking.

Another possible extension of the constrained approach is to solve the difficult problem of protein-protein docking. Though also not reported in this thesis, preliminary studies have been undertaken in which we aimed to detect, using an artifial neural network (ANN), the native binding mode of protein-protein interfaces. Practically, this can be achieved by training an ANN on several protein-protein decoys and by assessing whether the ANN correctly classified the protein interfaces as native or non-native. This early analysis led to a classification accuracy of ~80% on the protein-protein benchmark set [193], which showed great promise. In this scope, we envision to integrate the trained ANN together with relevant low-resolution experiment to the scoring scheme of $pow^{er}$-*mViE* in order to further perfect the current modeling procedure. Although applied to protein-protein

docking, the same machine learning principles could be extended to our SMD method to better characterise the ligand-receptor interactions and could serve as more robust estimation of binding affinity, which in turn would lead to higher docking accuracies.

Other developments related to the work featured in this thesis include the creation of a dedicated website, which should be online shortly and which would serve as an IM platform to distribute, maintain and further develop the *pow*$^{er}$ framework. As a future development, we plan on including a dedicated module to evaluate the fit between proteins and SAXS data, the treatment of protein flexibility through normal mode analysis and MD simulations, and finally the seamless addition of novel efficient optimizers.

## 6.1 References and Bibliography


1. Murata K, Wolf M. Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. Biochimica et biophysica acta 2018;1862(2):324-334.
2. Greber BJ, Boehringer D, Leibundgut M, Bieri P, Leitner A, Schmitz N, Aebersold R, Ban N. The complete structure of the large subunit of the mammalian mitochondrial ribosome. Nature 2014.
3. Agirrezabala X, Samatova E, Klimova M, Zamora M, Gil-Carton D, Rodnina MV, Valle M. Ribosome rearrangements at the onset of translational bypassing. Sci Adv 2017;3(6).
4. Ding F, Lu CR, Zhao W, Rajashankar KR, Anderson DL, Jardine PJ, Grimes S, Ke AL. Structure and assembly of the essential RNA ring component of a viral DNA packaging motor. Proceedings of the National Academy of Sciences of the United States of America 2011;108(18):7357-7362.
5. Neyer S, Kunz M, Geiss C, Hantsche M, Hodirnau VV, Seybert A, Engel C, Scheffer MP, Cramer P, Frangakis AS. Structure of RNA polymerase I transcribing ribosomal DNA genes. Nature 2016.
6. Bernecky C, Herzog F, Baumeister W, Plitzko JM, Cramer P. Structure of transcribing mammalian RNA polymerase II. Nature 2016;529(7587):551-554.
7. Abascal-Palacios G, Ramsay EP, Beuron F, Morris E, Vannini A. Structural basis of RNA polymerase III transcription initiation. Nature 2018;553(7688):301-306.
8. Yu XD, Veesler D, Campbell MG, Barry ME, Asturias FJ, Barry MA, Reddy VS. Cryo-EM structure of human adenovirus D26 reveals the conservation of structural organization among human adenoviruses. Sci Adv 2017;3(5).
9. Ward AB, Sali A, Wilson IA. Biochemistry. Integrative structural biology. Science 2013;339(6122):913-915.
10. Rodrigues JP, Bonvin AM. Integrative computational modeling of protein interactions. The FEBS journal 2014;281(8):1988-2003.
11. Baldwin AJ, Lioe H, Hilton GR, Baker LA, Rubinstein JL, Kay LE, Benesch JL. The polydispersity of αB-crystallin is rationalized by an interconverting polyhedral architecture. Structure 2011;19(12):1855-1863.
12. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences of the United States of America 2011;108(49):E1293-E1301.
13. Hopf TA, Scharfe CP, Rodrigues JP, Green AG, Kohlbacher O, Sander C, Bonvin AM, Marks DS. Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife 2014;3.
14. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. eLife 2014;3.
15. Thalassinos K, Pandurangan AP, Xu M, Alber F, Topf M. Conformational States of Macromolecular Assemblies Explored by Integrative Structure Calculation. Structure 2013;21(9):1500-1508.
16. Degiacomi MT, Dal Peraro M. Macromolecular Symmetric Assembly Prediction Using Swarm Intelligence Dynamic Modeling. Structure 2013;21(7):1097-1106.
17. Tamo G, Maesani A, Trager S, Degiacomi MT, Floreano D, Dal Peraro M. Disentangling constraints using viability evolution principles in integrative modeling of macromolecular assemblies. Scientific reports 2017;7(1):235.
18. Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 2003;125(7):1731-1737.

19. Spiga E, Degiacomi MT, Dal Peraro M. New Strategies for Integrative Dynamic Modeling of Macromolecular Assembly. Advances in Protein Chemistry and Structural Biology 2014.
20. Bakan A, Meireles LM, Bahar I. ProDy: protein dynamics inferred from theory and experiments. Bioinformatics 2011;27(11):1575-1577.
21. Spiga E, Alemani D, Degiacomi MT, Cascella M, Dal Peraro M. Electrostatic-Consistent Coarse-Grained Potentials for Molecular Simulations of Proteins. J Chem Theory Comput 2013;9(8):3515-3526.
22. Cheung AC, Sainsbury S, Cramer P. Structural basis of initial RNA polymerase II transcription. The EMBO journal 2011;30(23):4755-4763.
23. Wriggers W, Milligan RA, McCammon JA. Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. J Struct Biol 1999;125(2-3):185-195.
24. Wriggers W. Using Situs for the integration of multi-resolution structures. Biophysical reviews 2010;2(1):21-27.
25. Abe K, Tani K, Friedrich T, Fujiyoshi Y. Cryo-EM structure of gastric H+,K+-ATPase with a single occupied cation-binding site. Proceedings of the National Academy of Sciences of the United States of America 2012;109(45):18401-18406.
26. Tolun G, Makhov AM, Ludtke SJ, Griffith JD. Details of ssDNA annealing revealed by an HSV-1 ICP8-ssDNA binary complex. Nucleic acids research 2013;41(11):5927-5937.
27. Cuervo A, Pulido-Cid M, Chagoyen M, Arranz R, Gonzalez-Garcia VA, Garcia-Doval C, Caston JR, Valpuesta JM, van Raaij MJ, Martin-Benito J, Carrascosa JL. Structural characterization of the bacteriophage T7 tail machinery. J Biol Chem 2013;288(36):26290-26299.
28. Birol M, Enchev RI, Padilla A, Stengel F, Aebersold R, Betzi S, Yang Y, Hoh F, Peter M, Dumas C, Echalier A. Structural and biochemical characterization of the Cop9 signalosome CSN5/CSN6 heterodimer. Plos One 2014;9(8):e105688.
29. Alamo L, Wriggers W, Pinto A, Bartoli F, Salazar L, Zhao FQ, Craig R, Padron R. Three-Dimensional Reconstruction of Tarantula Myosin Filaments Suggests How Phosphorylation May Regulate Myosin Activity. Journal of molecular biology 2008;384(4):780-797.
30. Wang Z, Schroder GF. Real-space refinement with DireX: From global fitting to side-chain improvements. Biopolymers 2012;97(9):687-697.
31. Lopéz-Blanco JR, Chacón P. iMODFIT: Efficient and robust flexible fitting based on vibrational analysis in internal coordinates. J Struct Biol 2013;184(2):261-270.
32. Lopez-Blanco JR, Garzon JI, Chacon P. iMod: multipurpose normal mode analysis in internal coordinates. Bioinformatics 2011;27(20):2843-2850.
33. Pandurangan AP, Shakeel S, Butcher SJ, Topf M. Combined approaches to flexible fitting and assessment in virus capsids undergoing conformational change. J Struct Biol 2014;185(3):427-439.
34. Bock LV, Blau C, Schroder GF, Davydov II, Fischer N, Stark H, Rodnina MV, Vaiana AC, Grubmuler H. Energy barriers and driving forces in tRNA translocation through the ribosome. Nature structural & molecular biology 2013;20(12):1390-1396.
35. Kowal J, Chami M, Baumgartner P, Arheit M, Chiu PL, Rangl M, Scheuring S, Schroder GF, Nimigean CM, Stahlberg H. Ligand-induced structural changes in the cyclic nucleotide-modulated potassium channel MloK1. Nature communications 2014;5:3106.
36. Trabuco LG, Villa E, Schreiner E, Harrison CB, Schulten K. Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography. Methods 2009;49(2):174-180.
37. Li QF, Wanderling S, Paduch M, Medovoy D, Singharoy A, McGreevy R, Villalba-Galea CA, Hulse RE, Roux B, Schulten K, Kossiakoff A, Perozo E. Structural mechanism of

voltage-dependent gating in an isolated voltage-sensing domain. Nature structural & molecular biology 2014;21(3):244-252.
38. Flores SC, Sherman MA, Bruns CM, Eastman P, Altman RB. Fast Flexible Modeling of RNA Structure Using Internal Coordinates. Ieee Acm T Comput Bi 2011;8(5):1247-1257.
39. Zhao G, Perilla JR, Yufenyuy EL, Meng X, Chen B, Ning J, Ahn J, Gronenborn AM, Schulten K, Aiken C, Zhang P. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. Nature 2013;497(7451):643-646.
40. Russel D, Lasker K, Webb B, Velazquez-Muriel J, Tjioe E, Schneidman-Duhovny D, Peterson B, Sali A. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. PLoS biology 2012;10(1):e1001244.
41. Lasker K, Forster F, Bohn S, Walzthoeni T, Villa E, Unverdorben P, Beck F, Aebersold R, Sali A, Baumeister W. Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. Proceedings of the National Academy of Sciences of the United States of America 2012;109(5):1380-1387.
42. Sampathkumar P, Kim SJ, Upla P, Rice WJ, Phillips J, Timney BL, Pieper U, Bonanno JB, Fernandez-Martinez J, Hakhverdyan Z, Ketaren NE, Matsui T, Weiss TM, Stokes DL, Sauder JM, Burley SK, Sali A, Rout MP, Almo SC. Structure, dynamics, evolution, and function of a major scaffold component in the nuclear pore complex. Structure 2013;21(4):560-571.
43. Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprapto A, Karni-Schmidt O, Williams R, Chait BT, Sali A, Rout MP. The molecular architecture of the nuclear pore complex. Nature 2007;450(7170):695-701.
44. Shi Y, Fernandez-Martinez J, Tjioe E, Pellarin R, Kim SJ, Williams R, Schneidman D, Sali A, Rout MP, Chait BT. Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. Molecular & cellular proteomics : MCP 2014.
45. Politis A, Stengel F, Hall Z, Hernandez H, Leitner A, Walzthoeni T, Robinson CV, Aebersold R. A mass spectrometry-based hybrid method for structural modeling of protein complexes. Nature methods 2014;11(4):403-406.
46. Erzberger JP, Stengel F, Pellarin R, Zhang SY, Schaefer T, Aylett CHS, Cimermancic P, Boehringer D, Sali A, Aebersold R, Ban N. Molecular Architecture of the 40S center dot eIF1 center dot eIF3 Translation Initiation Complex. Cell 2014;158(5):1123-1135.
47. Habeck M, Rieping W, Nilges M. Weighting of experimental evidence in macromolecular structure determination. Proceedings of the National Academy of Sciences of the United States of America 2006;103(6):1756-1761.
48. Rieping W, Habeck M, Nilges M. Inferential structure determination. Science 2005;309(5732):303-306.
49. Molnar KS, Bonomi M, Pellarin R, Clinthorne GD, Gonzalez G, Goldberg SD, Goulian M, Sali A, DeGrado WF. Cys-Scanning Disulfide Crosslinking and Bayesian Modeling Probe the Transmembrane Signaling Mechanism of the Histidine Kinase, PhoQ. Structure 2014.
50. Das R, Baker D. Macromolecular modeling with rosetta. Annual review of biochemistry 2008;77:363-382.
51. DiMaio F, Leaver-Fay A, Bradley P, Baker D, Andre I. Modeling symmetric macromolecular structures in Rosetta3. Plos One 2011;6(6):e20450.
52. Song Y, DiMaio F, Wang RY, Kim D, Miles C, Brunette T, Thompson J, Baker D. High-resolution comparative modeling with RosettaCM. Structure 2013;21(10):1735-1742.
53. DiMaio F, Tyka MD, Baker ML, Chiu W, Baker D. Refinement of protein structures into low-resolution density maps using rosetta. Journal of molecular biology 2009;392(1):181-190.

54. Alushin GM, Lander GC, Kellogg EH, Zhang R, Baker D, Nogales E. High-Resolution Microtubule Structures Reveal the Structural Transitions in αβ-Tubulin upon GTP Hydrolysis. Cell 2014;157(5):1117-1129.
55. Kastritis PL, Rodrigues JP, Bonvin AM. HADDOCK(2P2I): a biophysical model for predicting the binding affinity of protein-protein interaction inhibitors. Journal of chemical information and modeling 2014;54(3):826-836.
56. van Dijk AD, Ciofi-Baffoni S, Banci L, Bertini I, Boelens R, Bonvin AM. Modeling protein-protein complexes involved in the cytochrome c oxidase copper-delivery pathway. Journal of proteome research 2007;6(4):1530-1539.
57. Karaca E, Bonvin AMJJ. On the usefulness of ion-mobility mass spectrometry and SAXS data in scoring docking decoys. Acta Crystallogr D 2013;69:683-694.
58. Karaca E, Bonvin AMJJ. A Multidomain Flexible Docking Approach to Deal with Large Conformational Changes in the Modeling of Biomolecular Complexes. Structure 2011;19(4):555-565.
59. Mashiach E, Schneidman-Duhovny D, Peri A, Shavit Y, Nussinov R, Wolfson HJ. An integrated suite of fast docking algorithms. Proteins 2010;78(15):3197-3204.
60. Nicastro G, Todi SV, Karaca E, Bonvin AM, Paulson HL, Pastore A. Understanding the role of the Josephin domain in the PolyUb binding and cleavage properties of ataxin-3. Plos One 2010;5(8):e12430.
61. Karaca E, Bonvin AM. Advances in integrative modeling of biomolecular complexes. Methods 2013;59(3):372-381.
62. Degiacomi MT, Iacovache I, Pernot L, Chami M, Kudryashev M, Stahlberg H, van der Goot FG, Dal Peraro M. Molecular assembly of the aerolysin pore reveals a swirling membrane-insertion mechanism. Nature chemical biology 2013;9(10):623-629.
63. Kudryashev M, Stenta M, Schmelz S, Amstutz M, Wiesand U, Castano-Diez D, Degiacomi MT, Munnich S, Bleck CK, Kowal J, Diepold A, Heinz DW, Dal Peraro M, Cornelis GR, Stahlberg H. In situ structural analysis of the Yersinia enterocolitica injectisome. eLife 2013;2:e00792.
64. Hofmeyer T, Schmelz S, Degiacomi MT, Dal Peraro M, Daneschdar M, Scrima A, van den Heuvel J, Heinz DW, Kolmar H. Arranged Sevenfold: Structural Insights into the C-Terminal Oligomerization Domain of Human C4b-Binding Protein. Journal of molecular biology 2013;425(8):1302-1317.
65. Schneidman-Duhovny D, Kim SJ, Sali A. Integrative structural modeling with small angle X-ray scattering profiles. BMC structural biology 2012;12:17.
66. Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. Nucleic acids research 2016;44(W1):W424-W429.
67. Vijayvargia R, Epand R, Leitner A, Jung TY, Shin B, Jung R, Lloret A, Atwal RS, Lee H, Lee JM, Aebersold R, Hebert H, Song JJ, Seong IS. Huntingtin's spherical solenoid structure enables polyglutamine tract-dependent modulation of its structure and function. eLife 2016;5.
68. Ovchinnikov S, Park H, Varghese N, Huang PS, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D. Protein structure determination using metagenome sequence data. Science 2017;355(6322):294-297.
69. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. Cell 2012;149(7):1607-1621.
70. Abriata LA, Tamo GE, Monastyrskyy B, Kryshtafovych A, Dal Peraro M. Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. Proteins 2018;86 Suppl 1:97-112.

71. van den Bergh T, Tamo G, Nobili A, Tao Y, Tan T, Bornscheuer UT, Kuipers RKP, Vroling B, de Jong RM, Subramanian K, Schaap PJ, Desmet T, Nidetzky B, Vriend G, Joosten HJ. CorNet: Assigning function to networks of co-evolving residues by automated literature mining. Plos One 2017;12(5):e0176427.
72. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. Proteins 2016;84 Suppl 1:4-14.
73. Tamo GE, Abriata LA, Fonti G, Dal Peraro M. Assessment of data-assisted prediction by inclusion of crosslinking/mass-spectrometry and small angle X-ray scattering data in the 12(th) Critical Assessment of protein Structure Prediction experiment. Proteins 2018;86 Suppl 1:215-227.
74. Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, Stallings WC, Connolly DT, Shoichet BK. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. Journal of medicinal chemistry 2002;45(11):2213-2221.
75. Meng EC, Shoichet BK, Kuntz ID. Automated Docking with Grid-Based Energy Evaluation. Journal of computational chemistry 1992;13(4):505-524.
76. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry 2010;31(2):455-461.
77. Zoete V, Schuepbach T, Bovigny C, Chaskar P, Daina A, Rohrig UF, Michielin O. Attracting cavities for docking. Replacing the rough energy landscape of the protein by a smooth attracting landscape. Journal of computational chemistry 2016;37(4):437-447.
78. Grosdidier A, Zoete V, Michielin O. SwissDock, a protein-small molecule docking web service based on EADock DSS. Nucleic acids research 2011;39(Web Server issue):W270-277.
79. Kurkcuoglu Z, Koukos PI, Citro N, Trellet ME, Rodrigues J, Moreira IS, Roel-Touris J, Melquiond ASJ, Geng C, Schaarschmidt J, Xue LC, Vangone A, Bonvin A. Performance of HADDOCK and a simple contact-based protein-ligand binding affinity predictor in the D3R Grand Challenge 2. J Comput Aided Mol Des 2018;32(1):175-185.
80. Neves MAC, Totrov M, Abagyan R. Docking and scoring with ICM: the benchmarking results and strategies for improvement. J Comput Aid Mol Des 2012;26(6):675-686.
81. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. Journal of medicinal chemistry 2004;47(7):1739-1749.
82. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. Journal of molecular biology 1997;267(3):727-748.
83. Maesani A, Iacca G, Floreano D. Memetic Viability Evolution for Constrained Optimization. Ieee T Evolut Comput 2016;20(1):125-144.
84. Difiglia M, Sapp E, Chase K, Schwarz C, Meloni A, Young C, Martin E, Vonsattel JP, Carraway R, Reeves SA, Boyce FM, Aronin N. Huntingtin Is a Cytoplasmic Protein Associated with Vesicles in Human and Rat-Brain Neurons. Neuron 1995;14(5):1075-1081.
85. Sharp AH, Loev SJ, Schilling G, Li SH, Li XJ, Bao J, Wagster MV, Kotzuk JA, Steiner JP, Lo A, et al. Widespread expression of Huntington's disease gene (IT15) protein product. Neuron 1995;14(5):1065-1074.
86. Harjes P, Wanker EE. The hunt for huntingtin function: interaction partners tell many different stories. Trends in biochemical sciences 2003;28(8):425-433.

87. Guo Q, Bin H, Cheng J, Seefelder M, Engler T, Pfeifer G, Oeckl P, Otto M, Moser F, Maurer M, Pautsch A, Baumeister W, Fernandez-Busnadiego R, Kochanek S. The cryo-electron microscopy structure of huntingtin. Nature 2018;555(7694):117-120.
88. Grudinin S, Garkavenko M, Kazennov A. Pepsi-SAXS: an adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles. Acta crystallographica Section D, Structural biology 2017;73(Pt 5):449-464.
89. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics 2000;16(4):404-405.
90. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. Journal of molecular biology 1993;234(3):779-815.
91. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, de Vries AH. The MARTINI force field: coarse grained model for biomolecular simulations. J Phys Chem B 2007;111(27):7812-7824.
92. Kmiecik S, Gront D, Kolinski M, Wieteska L, Dawid AE, Kolinski A. Coarse-Grained Protein Models and Their Applications. Chem Rev 2016;116(14):7898-7936.
93. Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP, Marrink SJ. The MARTINI Coarse-Grained Force Field: Extension to Proteins. J Chem Theory Comput 2008;4(5):819-834.
94. Alemani D, Collu F, Cascella M, Dal Peraro M. A Nonradial Coarse-Grained Potential for Proteins Produces Naturally Stable Secondary Structure Elements. J Chem Theory Comput 2010;6(1):315-324.
95. Wu Q, Cole C, McSweeney T. Applications of particle swarm optimization in the railway domain. Int J Rail Transp 2016;4(3):167-190.
96. Wang X, Qiu X. Application of particle swarm optimization for enhanced cyclic steam stimulation in a offshore heavy oil reservoir. arXiv preprint arXiv:13064092 2013.
97. Maschek MK. Particle Swarm Optimization in Agent-Based Economic Simulations of the Cournot Market Model. Intell Syst Account 2015;22(2):133-152.
98. Moal IH, Bates PA. SwarmDock and the Use of Normal Modes in Protein-Protein Docking. Int J Mol Sci 2010;11(10):3623-3648.
99. Woldesenbet YG, Yen GG, Tessema BG. Constraint Handling in Multiobjective Evolutionary Optimization. Ieee T Evolut Comput 2009;13(3):514-525.
100. Popyack JL. Introduction to evolutionary computing (second edition) (vol 17, pg 197, 2016). Genet Program Evol M 2016;17(2):201-201.
101. Maesani A, Fernando PR, Floreano D. Artificial Evolution by Viability Rather than Competition. Plos One 2014;9(1).
102. Aubin JP, Frankowska H. Set-valued analysis, viability theory and partial differential inclusions. World Congress of Nonlinear Analysis '92, Vols 1-4 1995:1039-1058.
103. Igel C, Suttorp T, Hansen N. A computational efficient covariance matrix update and a (1+ 1)-CMA for evolution strategies. 2006. ACM. p 453-460.
104. Storn R, Price K. Differential evolution - A simple and efficient heuristic for global optimization over continuous spaces. J Global Optim 1997;11(4):341-359.
105. Hansen N, Ostermeier A. Completely derandomized self-adaptation in evolution strategies. Evol Comput 2001;9(2):159-195.
106. Hansen M, Ostermeier A. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. 1996 Ieee International Conference on Evolutionary Computation (Icec '96), Proceedings Of 1996:312-317.
107. Igel C, Hansen N, Roth S. Covariance matrix adaptation for multi-objective optimization. Evol Comput 2007;15(1):1-28.

108. Maesani A, Floreano D. Viability Principles for Constrained Optimization Using a (1+ 1)-CMA-ES. 2014.
109. Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprapto A, Karni-Schmidt O, Williams R, Chait BT, Rout MP, Sali A. Determining the architectures of macromolecular assemblies. Nature 2007;450(7170):683-694.
110. Tamo GE, Abriata LA, Dal Peraro M. The importance of dynamics in integrative modeling of supramolecular assemblies. Current opinion in structural biology 2015;31:28-34.
111. Greber BJ, Bieri P, Leibundgut M, Leitner A, Aebersold R, Boehringer D, Ban N. Ribosome. The complete structure of the 55S mammalian mitochondrial ribosome. Science 2015;348(6232):303-308.
112. Hoffmann NA, Jakobi AJ, Moreno-Morcillo M, Glatt S, Kosinski J, Hagen WJ, Sachse C, Muller CW. Molecular structures of unbound and transcribing RNA polymerase III. Nature 2015;528(7581):231-236.
113. Ferber M, Kosinski J, Ori A, Rashid UJ, Moreno-Morcillo M, Simon B, Bouvier G, Batista PR, Muller CW, Beck M, Nilges M. Automated structure modeling of large protein assemblies using crosslinks as distance restraints. Nature methods 2016.
114. Shi Y, Pellarin R, Fridy PC, Fernandez-Martinez J, Thompson MK, Li Y, Wang QJ, Sali A, Rout MP, Chait BT. A strategy for dissecting the architectures of native macromolecular assemblies. Nature methods 2015;12(12):1135-1138.
115. Sali A, Berman HM, Schwede T, Trewhella J, Kleywegt G, Burley SK, Markley J, Nakamura H, Adams P, Bonvin AM, Chiu W, Peraro MD, Di Maio F, Ferrin TE, Grunewald K, Gutmanas A, Henderson R, Hummer G, Iwasaki K, Johnson G, Lawson CL, Meiler J, Marti-Renom MA, Montelione GT, Nilges M, Nussinov R, Patwardhan A, Rappsilber J, Read RJ, Saibil H, Schroder GF, Schwieters CD, Seidel CA, Svergun D, Topf M, Ulrich EL, Velankar S, Westbrook JD. Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. Structure 2015;23(7):1156-1167.
116. Lasker K, Topf M, Sali A, Wolfson HJ. Inferential Optimization for Simultaneous Fitting of Multiple Components into a CryoEM Map of Their Assembly. Journal of molecular biology 2009;388(1):180-194.
117. Liang S, Wang G, Zhou Y. Refining near-native protein-protein docking decoys by local resampling and energy minimization. Proteins 2009;76(2):309-316.
118. Yueh C, Hall DR, Xia B, Padhorny D, Kozakov D, Vajda S. ClusPro-DC: Dimer Classification by the Cluspro Server for Protein-Protein Docking. Journal of molecular biology 2016.
119. Gromiha MM, Yugandhar K, Jemimah S. Protein-protein interactions: scoring schemes and binding affinity. Current opinion in structural biology 2016;44:31-38.
120. Pfeiffenberger E, Chaleil RA, Moal IH, Bates PA. A machine learning approach for ranking clusters of docked protein-protein complexes by pairwise cluster comparison. Proteins 2016.
121. Fink F, Hochrein J, Wolowski V, Merkl R, Gronwald W. PROCOS: computational analysis of protein-protein complexes. Journal of computational chemistry 2011;32(12):2575-2586.
122. Lasker K, Sali A, Wolfson HJ. Determining macromolecular assembly structures by molecular docking and fitting into an electron density map. Proteins-Structure Function and Bioinformatics 2010;78(15):3205-3211.
123. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. Geometry-based flexible and symmetric protein docking. Proteins 2005;60(2):224-231.
124. Goldberg SD, Soto CS, Waldburger CD, Degrado WF. Determination of the physiological dimer interface of the PhoQ sensor domain. Journal of molecular biology 2008;379(4):656-665.

125. Baker ML, Baker MR, Hryc CF, Dimaio F. Analyses of subnanometer resolution cryo-EM density maps. Methods in enzymology 2010;483:1-29.
126. Jolley CC, Wells SA, Fornme P, Thorpe MF. Fitting low-resolution cryo-EM maps of proteins using constrained geometric simulations. Biophysical journal 2008;94(5):1613-1621.
127. MacKerell AD, Jr., Banavali N, Foloppe N. Development and current status of the CHARMM force field for nucleic acids. Biopolymers 2000;56(4):257-265.
128. Dayalu P, Albin RL. Huntington disease: pathogenesis and treatment. Neurologic clinics 2015;33(1):101-114.
129. Tobin AJ, Signer ER. Huntington's disease: the challenge for cell biologists. Trends in cell biology 2000;10(12):531-536.
130. Seong IS, Woda JM, Song JJ, Lloret A, Abeyrathne PD, Woo CJ, Gregory G, Lee JM, Wheeler VC, Walz T, Kingston RE, Gusella JF, Conlon RA, MacDonald ME. Huntingtin facilitates polycomb repressive complex 2. Human molecular genetics 2010;19(4):573-583.
131. Kim MW, Chelliah Y, Kim SW, Otwinowski Z, Bezprozvanny I. Secondary structure of Huntingtin amino-terminal region. Structure 2009;17(9):1205-1212.
132. Bezprozvanny I. Calcium signaling and neurodegenerative diseases. Trends in molecular medicine 2009;15(3):89-100.
133. Li SH, Li XJ. Huntingtin-protein interactions and the pathogenesis of Huntington's disease|. Trends Genet 2004;20(3):146-154.
134. Kim M. Beta conformation of polyglutamine track revealed by a crystal structure of Huntingtin N-terminal region with insertion of three histidine residues. Prion 2013;7(3):221-228.
135. Andrade MA, Bork P. HEAT repeats in the Huntington's disease protein. Nature genetics 1995;11(2):115-116.
136. Palidwor GA, Shcherbinin S, Huska MR, Rasko T, Stelzl U, Arumughan A, Fouiie R, Porras P, Sanchez-Pulido L, Wanker EE, Andrade-Navarro MA. Detection of Alpha-Rod Protein Repeats Using a Neural Network and Application to Huntingtin. Plos Computational Biology 2009;5(3).
137. Tartari M, Gissi C, Lo Sardo V, Zuccato C, Picardi E, Pesole G, Cattaneo E. Phylogenetic comparison of huntingtin homologues reveals the appearance of a primitive polyQ in sea urchin. Molecular biology and evolution 2008;25(2):330-338.
138. Groves MR, Hanlon N, Turowski P, Hemmings BA, Barford D. The structure of the protein phosphatase 2A PR65/A subunit reveals the conformation of its 15 tandemly repeated HEAT motifs. Cell 1999;96(1):99-110.
139. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. JPred: a consensus secondary structure prediction server. Bioinformatics 1998;14(10):892-893.
140. Guaitoli G, Raimondi F, Gilsbach BK, Gomez-Llorente Y, Deyaert E, Renzi F, Li XT, Schaffner A, Jagtap PKA, Boldt K, von Zweydorf F, Gotthardt K, Lorimer DD, Yue ZY, Burgin A, Janjic N, Sattler M, Versees W, Ueffing M, Ubarretxena-Belandia I, Kortholt A, Gloeckner CJ. Structural model of the dimeric Parkinson's protein LRRK2 reveals a compact architecture involving distant interdomain contacts. Proceedings of the National Academy of Sciences of the United States of America 2016;113(30):E4357-E4366.
141. Terwilliger TC. Rapid model building of alpha-helices in electron-density maps. Acta crystallographica Section D, Biological crystallography 2010;66(Pt 3):268-275.
142. Rusu M, Wriggers W. Evolutionary bidirectional expansion for the tracing of alpha helices in cryo-electron microscopy reconstructions. J Struct Biol 2012;177(2):410-419.

143. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory research and analysis. Journal of computational chemistry 2004;25(13):1605-1612.
144. Schneider M, Belsom A, Rappsilber J, Brock O. Blind testing of cross-linking/mass spectrometry hybrid methods in CASP11. Proteins 2016;84 Suppl 1:152-163.
145. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. Journal of molecular graphics 1996;14(1):33-38, 27-38.
146. DeLano WL. The PyMOL molecular graphics system. http://pymol org 2002.
147. Schindler T, Bornmann W, Pellicena P, Miller WT, Clarkson B, Kuriyan J. Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. Science 2000;289(5486):1938-1942.
148. Kaldor SW, Kalish VJ, Davies JF, 2nd, Shetty BV, Fritz JE, Appelt K, Burgess JA, Campanale KM, Chirgadze NY, Clawson DK, Dressman BA, Hatch SD, Khalil DA, Kosa MB, Lubbehusen PP, Muesing MA, Patick AK, Reich SH, Su KS, Tatlock JH. Viracept (nelfinavir mesylate, AG1343): a potent, orally bioavailable inhibitor of HIV-1 protease. Journal of medicinal chemistry 1997;40(24):3979-3985.
149. Varghese JN. Development of neuraminidase inhibitors as anti-influenza virus drugs. Drug development research 1999;46(3-4):176-196.
150. Anderson AC. The process of structure-based drug design. Chemistry & biology 2003;10(9):787-797.
151. Jorgensen WL. Efficient drug lead discovery and optimization. Accounts of chemical research 2009;42(6):724-733.
152. Gohlke H, Klebe G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angewandte Chemie 2002;41(15):2644-2676.
153. Meng XY, Zhang HX, Mezei M, Cui M. Molecular docking: a powerful approach for structure-based drug discovery. Current computer-aided drug design 2011;7(2):146-157.
154. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nature reviews Drug discovery 2004;3(11):935-949.
155. Pagadala NS, Syed K, Tuszynski J. Software for molecular docking: a review. Biophys Rev 2017;9(2):91-102.
156. Bursulaya BD, Totrov M, Abagyan R, Brooks CL. Comparative study of several algorithms for flexible ligand docking. J Comput Aid Mol Des 2003;17(11):755-763.
157. Ferreira LG, Dos Santos RN, Oliva G, Andricopulo AD. Molecular docking and structure-based drug design strategies. Molecules 2015;20(7):13384-13421.
158. Abagyan R, Totrov M, Kuznetsov D. Icm - a New Method for Protein Modeling and Design - Applications to Docking and Structure Prediction from the Distorted Native Conformation. Journal of computational chemistry 1994;15(5):488-506.
159. Sousa SF, Fernandes PA, Ramos MJ. Protein-ligand docking: Current status and future challenges. Proteins-Structure Function and Bioinformatics 2006;65(1):15-26.
160. Yuriev E, Ramsland PA. Latest developments in molecular docking: 2010-2011 in review. J Mol Recognit 2013;26(5):215-239.
161. Forli S, Huey R, Pique ME, Sanner MF, Goodsell DS, Olson AJ. Computational protein-ligand docking and virtual drug screening with the AutoDock suite. Nature protocols 2016;11(5):905-919.
162. Gu JF, Yang X, Kang L, Wu JY, Wang XC. MoDock: A multi-objective strategy improves the accuracy for molecular docking. Algorithm Mol Biol 2015;10.

163. Spitzer R, Jain AN. Surflex-Dock: Docking benchmarks and real-world application. J Comput Aided Mol Des 2012;26(6):687-699.
164. Venkatachalam CM, Jiang X, Oldfield T, Waldman M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. Journal of molecular graphics & modelling 2003;21(4):289-307.
165. Sauer OA, Shepard DM, Mackie TR. Application of constrained optimization to radiotherapy planning. Medical physics 1999;26(11):2359-2366.
166. Tikhonravov AV, Trubetskov MK, Amotchkina TV. Application of constrained optimization to the design of quasi-rugate optical coatings. Applied optics 2008;47(28):5103-5109.
167. Coello CAC. Use of a self-adaptive penalty approach for engineering optimization problems. Comput Ind 2000;41(2):113-127.
168. Wang R, Fang X, Lu Y, Wang S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. Journal of medicinal chemistry 2004;47(12):2977-2980.
169. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WT, Mortenson PN, Murray CW. Diverse, high-quality test set for the validation of protein-ligand docking performance. Journal of medicinal chemistry 2007;50(4):726-741.
170. Wang RX, Lai LH, Wang SM. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J Comput Aid Mol Des 2002;16(1):11-26.
171. Quiroga R, Villarreal MA. Vinardo: A Scoring Function Based on Autodock Vina Improves Scoring, Docking, and Virtual Screening. Plos One 2016;11(5):e0155183.
172. Tanchuk VY, Tanin VO, Vovk AI, Poda G. A New, Improved Hybrid Scoring Function for Molecular Docking and Scoring Based on AutoDock and AutoDock Vina. Chemical biology & drug design 2016;87(4):618-625.
173. Gabb HA, Jackson RM, Sternberg MJE. Modelling protein docking using shape complementarity, electrostatics and biochemical information. Journal of molecular biology 1997;272(1):106-120.
174. Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. Journal of medicinal chemistry 1985;28(7):849-857.
175. Harris R, Olson AJ, Goodsell DS. Automated prediction of ligand-binding sites in proteins. Proteins-Structure Function and Bioinformatics 2008;70(4):1506-1517.
176. Hendlich M, Rippmann F, Barnickel G. LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. Journal of molecular graphics & modelling 1997;15(6):359-+.
177. Leach AR, Hann MM, Burrows JN, Griffen EJ. Fragment screening: an introduction. Molecular bioSystems 2006;2(9):430-446.
178. Ewing TJ, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des 2001;15(5):411-428.
179. Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking method using an incremental construction algorithm. Journal of molecular biology 1996;261(3):470-489.
180. Roberts BC, Mancera RL. Ligand-protein docking with water molecules. Journal of chemical information and modeling 2008;48(2):397-408.
181. Labbe CM, Pencheva T, Jereva D, Desvillechabrol D, Becot J, Villoutreix BO, Pajeva I, Miteva MA. AMMOS2: a web server for protein-ligand-water complexes refinement via molecular mechanics. Nucleic acids research 2017;45(W1):W350-W355.

182. Bucher D, Stouten P, Triballeau N. Shedding Light on Important Waters for Drug Design: Simulations versus Grid-Based Methods. Journal of chemical information and modeling 2018.
183. Chen DL, Menche G, Power TD, Sower L, Peterson JW, Schein CH. Accounting for ligand-bound metal ions in docking small molecules on adenylyl cyclase toxins. Proteins-Structure Function and Bioinformatics 2007;67(3):593-605.
184. Chaskar P, Zoete V, Rohrig UF. On-the-Fly QM/MM Docking with Attracting Cavities. Journal of chemical information and modeling 2017;57(1):73-84.
185. Lill MA. Efficient Incorporation of Protein Flexibility and Dynamics into Molecular Docking Simulations. Biochemistry-Us 2011;50(28):6157-6169.
186. Leach AR. Ligand docking to proteins with discrete side-chain flexibility. Journal of molecular biology 1994;235(1):345-356.
187. Venkitakrishnan RP, Zaborowski E, McElheny D, Benkovic SJ, Dyson HJ, Wright PE. Conformational changes in the active site loops of dihydrofolate reductase during the catalytic cycle. Biochemistry-Us 2004;43(51):16046-16055.
188. Oleinikovas V, Saladino G, Cossins BP, Gervasio FL. Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations. J Am Chem Soc 2016;138(43):14257-14263.
189. Bottegoni G, Rocchia W, Rueda M, Abagyan R, Cavalli A. Systematic exploitation of multiple receptor conformations for virtual ligand screening. Plos One 2011;6(5):e18845.
190. Grinter R, Josts I, Mosbahi K, Roszak AW, Cogdell RJ, Bonvin AM, Milner JJ, Kelly SM, Byron O, Smith BO, Walker D. Structure of the bacterial plant-ferredoxin receptor FusA. Nature communications 2016;7:13308.
191. Schur FK, Hagen WJ, de Marco A, Briggs JA. Determination of protein structure at 8.5 Å resolution using cryo-electron tomography and sub-tomogram averaging. J Struct Biol 2013;184(3):394-400.
192. Pandurangan AP, Vasishtan D, Alber F, Topf M. γ-tempy: Simultaneous fitting of components in 3d-em maps of their assembly using a genetic algorithm. Structure 2015;23(12):2365-2376.
193. Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, Torchala M, Chaleil R, Jimenez-Garcia B, Bates PA, Fernandez-Recio J, Bonvin AM, Weng Z. Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. Journal of molecular biology 2015;427(19):3031-3041.

## Annexe

Additional publication related to the thesis:

"**Assessment of data-assisted prediction by inclusion of crosslinking/mass-spectrometry and small angle X-ray scattering data in the 12$^{th}$ Critical Assessment of protein Structure Prediction experiment**" Giorgio E. Tamò, Luciano A. Abriata, Guilia Fonti and Matteo Dal Peraro. *Proteins: structure, function and bioinformatics*, 2018

Received: 21 August 2017 | Revised: 16 November 2017 | Accepted: 10 December 2017

DOI: 10.1002/prot.25442

RESEARCH ARTICLE

WILEY PROTEINS

# Assessment of data-assisted prediction by inclusion of crosslinking/mass-spectrometry and small angle X-ray scattering data in the 12th Critical Assessment of protein Structure Prediction experiment

Giorgio E. Tamò[1,2] | Luciano A. Abriata[1,2] | Giulia Fonti[1,2] | Matteo Dal Peraro[1,2]

[1]Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[2]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

**Correspondence**
Matteo Dal Peraro, Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne 1015, Switzerland.
E-mail: matteo.dalperaro@epfl.ch

## Abstract

Integrative modeling approaches attempt to combine experiments and computation to derive structure-function relationships in complex molecular assemblies. Despite their importance for the advancement of life sciences, benchmarking of existing methodologies is rather poor. The 12th round of the Critical Assessment of protein Structure Prediction (CASP) offered a unique niche to benchmark data and methods from two kinds of experiments often used in integrative modeling, namely residue-residue contacts obtained through crosslinking/mass-spectrometry (CLMS), and small-angle X-ray scattering (SAXS) experiments. Upon assessment of the models submitted by predictors for 3 targets assisted by CLMS data and 11 targets by SAXS data, we observed no significant improvement when compared to the best data-blind models, although most predictors did improve relative to their own data-blind predictions. Only for target Tx892 of the CLMS-assisted category and for target Ts947 of the SAXS-assisted category, there was a net, albeit mild, improvement relative to the best data-blind predictions. We discuss here possible reasons for the relatively poor success, which point rather to inconsistencies in the data sources rather than in the methods, to which a few groups were less sensitive. We conclude with suggestions that could improve the potential of data integration in future CASP rounds in terms of experimental data production, methods development, data management and prediction assessment.

**KEYWORDS**

critical assessment of structure prediction, cross-linking mass-spectrometry, integrative modeling, small-angle x-ray scattering, structural biology

## 1 | INTRODUCTION

Integrative modeling (IM) techniques are today emerging as powerful tools to model the architecture of proteins[1–3] by combining computational methods with experimental information. Such experimental inputs usually take the form of low-resolution structural data describing shape, volume, or subunit arrangement such as cryo-electron microscopy (EM) or small-angle X-ray scattering (SAXS) data.[4–6] In contrast, spatial restraints extracted from crosslink/mass-spectrometry experiments[7] can be used to pinpoint contacting residues, greatly helping the modeling process of protein structure.[1] IM techniques are commonly used to model protein quaternary structures[1–3] but can in principle be extended to model protein tertiary structure as well. Moreover, even residue-residue contacts predicted from residue co-evolution analysis can be integrated for tertiary and quaternary structure predictions,[8–10] with associated uncertainty and noise that have to be considered by these methods.[11] Although IM has been successfully applied to reveal the structures of several large important macromolecules,[12–16] there is still no standard benchmark procedure to rigorously assess the performance of current IM approaches on predicting protein tertiary structures.[8,17] This work describes one of such effort, which took place in the context of the Critical Assessment of protein Structure Prediction (CASP).

**Abbreviations:** CASP, critical assessment of protein structure prediction; CLMS, crosslinking/mass-spectrometry; SAXS, small-angle x-ray scattering.

CASP attempts to objectively assess the state of the art of protein structure prediction from amino acid sequences through an international competition in which groups have to predict protein structures secured by the organizers but which have not been released by the Protein Data Bank.[17,18] The predicted models are then evaluated against the respective experimentally solved structures (targets) in different tracks that look at specific features, such as global fold prediction, refinement of fold, and side-chain details, oligomer prediction, and of most relevance to this article, data-assisted modeling. The data-assisted modeling usually takes place on a subset of the targets that undergo data-blind fold prediction, whose predictions we also analyzed in a recent article.[19]

CASP11 was the first edition to incorporate experimental inputs to protein tertiary structure prediction in a data-assisted prediction category[20] featuring simulated NMR data about ambiguous distance restraints,[21] and residue-residue contacts determined experimentally from crosslinking/mass-spectrometry (CLMS) data[20] based on the photo-CLMS protocol established by Brock and Rappsilber.[22] For this 12th edition of CASP (CASP12), organizers managed to include (1) CLMS data produced by the same approach as in CASP11 and obtained also by the Rappsilber group,[22] and (2) SAXS data obtained from the collaboration with the SIBYLS Beamline facility at the Advanced Light Source Synchrotron. Of importance to the practical aspects of implementation during CASP, both the CLMS and SAXS methods used were streamlined to quickly collect data from the same samples used for target crystallization.

In this article, we report the assessment of data-assisted predictions based on SAXS and CLMS data for protein structure prediction in CASP12. We describe results for each target in comparison to the data-blind models, which we also assessed.[19] We discuss the outcome of our assessment in the context of the employed data, the methods used by the predictors and the logistics of the CASP experiment.

## 2 | METHODS

### 2.1 | Analysis of CLMS data

The CLMS-assisted category contained 3 targets that were named Tx892, Tx894, and Tx895. In this category, we considered predicting groups that submitted at least 1 models for the 3 targets. These 11 groups were Unres, Laufer_seed, bugre, Spiders, M2O, Rbo_Human, Kias-Gdansk, Dalton, Font, Lee, and Goal. The predictor names associated with each of the group nicknames can be found on the CASP webpage (http://predictioncenter.org/casp12/docs.cgi?view=groupsbyname).

For each of the 3 Tx targets, we evaluated how the solved crystal structure and predicted models fitted the CLMS data by identifying the pairs of residues predicted to form crosslinks within different Cα-Cα distances, and analyzing the confidence level values associated with each predicted contact. These confidence levels were calculated in ref. 23 and were made available to us through the Prediction Center.[24]

### 2.2 | Analysis of SAXS data

A total of 11 targets named Ts894, Ts895, Ts866, Ts886, Ts896, Ts899, Ts901, Ts909, Ts941, Ts942, and Ts947 were selected for the SAXS-assisted category. Groups who submitted target predictions were Floudas, Spiders, Unres, Rbo_Human, Grudinin_DeepL, Multicom, dimaiolab, Grudinin, Kias_Gdansk, and Lee. More detail about each group participant can be found on the CASP webpage (http://predictioncenter.org/casp12/docs.cgi?view=groupsbyname).

We used SCÅTTER[25] to generate Guinier plots (Supporting Information Figure S1 and Table S1), and, based on the linearity of the $ln[I(q)]$ versus $q^2$ plot, classified the samples as monodispersed (if linear) or aggregated (if not linear). For the current SAXS-assisted targets, the information regarding the molecular mass and oligomeric states was obtained from ref.[23]

The agreement between the experimental data and the structure of both models and targets was assessed using the software *Crysol*[26] by calculating the $\chi$ value, which quantifies the difference between the experimental SAXS profile and simulated SAXS profile generated from the model. We used the SITUS[27] package to dock the structures inside the computed SAXS envelopes using the *kercon* module and quantified the fit between the structure and envelop with the cross-correlation (*cc*) value.

### 2.3 | Improvement assessment of data-assisted predictions

The relative improvement of data-assisted models over their equivalent models predicted in the data-blind category was evaluated. This was done by visual inspection and structural comparison between the models and respective crystal structures using structural similarity metrics. These included the Global Distance Test Total Score (GDTTS),[28] which is a standard measure of model quality used in CASP. This metric computes the maximum number of model residues that can be superimposed on the target structure under the cutoffs values of 1, 2, 4, and 8 Å, normalized by the number of residues. It ranges from 0 to 100 with the highest value describing models whose Cα atoms can all be aligned within 1 Å from their positions in the reference structure. The GDTTS is computed using an automated procedure and is normally available for each of the models from the Prediction Center.[24] For some comparisons that we did, where GDTTS was not available, it was computed using the maxCluster package.[29]

To complement the scoring obtained from GDTTS upon model evaluation relative to the targets, other metrics including QCS,[30] Handedness,[31] DFM,[31] CoDM,[31] and TMalign scores[32] were used. The QCS[30] score compares structural similarities between secondary structure elements. The Handedness and DFM scores[31] use the relative orientations and distortions of atom tetrads to estimate conformation differences. CoDM[31] measures residue-residue distance matrices correlations to quantify the structural differences between protein structures. The TMalign score[32] is a sequence-independent version of the TM score for measuring structural similarity by weighting and combining distance residue-residue distances in model and target. All these scores were available from the Prediction Center.[24] Root-mean-square-deviations (rmsd) between Cα atoms were computed using PyMol.[33]

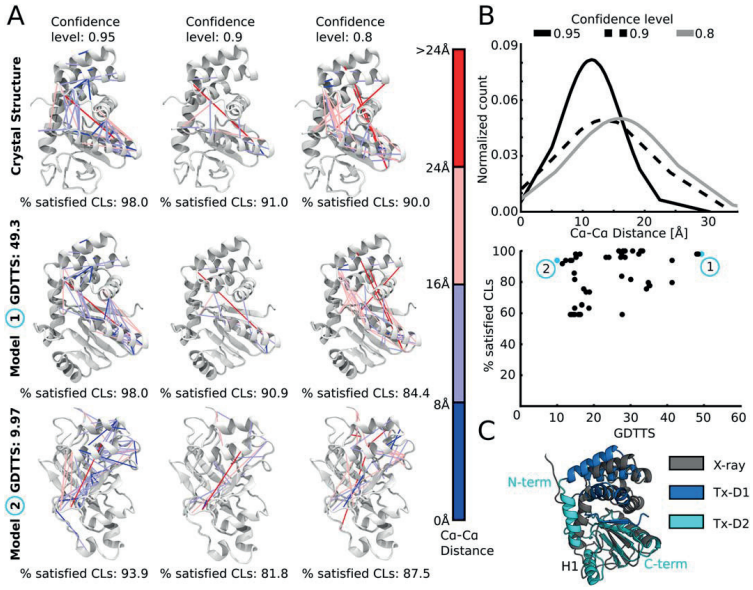



**FIGURE 1** Assessment of CLMS-based modeling for target Tx892. **A**, Target crystal structure (top), best model 1 of Goal (middle), and worst model (bottom) in terms of GDTTS. Pairs of residues expected to be close in space according to CLMS data at 3 confidence levels are connected by lines color-coded by the actual distance as measured in the target crystal structure. **B**, Top: Distribution of Cα-Cα distances, as measured in the target crystal structure, for pairs of residues crosslinked at confidence levels of 0.8, 0.9, and 0.95. The un-normalized and un-smoothened distribution can be found in the Supporting Information Figure S2. Bottom: Fraction of crosslinks satisfied by each model against their GDTTS scores, for all submitted CLMS-assisted models. **C**, Structural alignment of the best Tx892 model on the target crystal structure aligned considering residues 84–193 (i.e., domain 2: Tx-D2) to highlight the different arrangements of the 2 domains in the model compared to the target structure. Helix H1 illustrates a normally disordered region being modeled as a defined secondary structure

## 3 | RESULTS

### 3.1 | CLMS driven data-assisted modeling

In the CLMS-assisted category of CASP12, predictors were asked to model 3 target structures, namely Tx892, Tx894, and Tx895. For each target, the information available to the predictors was the amino acid sequence, the pairs of residues found crosslinked by mass spectrometry (MS), and the confidence level associated with each crosslinked pair. The confidence level, which range from 0.8 to 0.95, is estimated directly by the group providing these data.[23] We here analyze the models produced by the 11 predictors for these 3 targets employing the cross-linking data, in comparison to the target structures; and interpret results in the context of domain difficulty, the available data-blind models, and the quality of the CLMS data as defined based on the target structure.

Target Tx892 was expected to be rather hard for modeling because sequence- and structure-level searches to the Protein Data Bank do not retrieve any clear templates, especially for the sequence segment 84–193.[34] Ranking of models by GDTTS, Handedness, DFM, CoDM, and TM scores all invariably point at models 1 and 2 by the Goal group and model 1 by Lee as best models. Importantly, regarding the assessment of data-driven modeling, these 3 models score slightly better than the best data-blind models, according to all 6 scores (e.g., reaching GDTTS in the range 48.1–49.4 versus 42.6 for the best data-blind model). These models were found to be structurally similar to each other (<1.5 Å Cα-rmsd). Model 1 submitted by Goal was the best model by GDTTS, and is displayed in Figure 1.

Upon visual comparison of the top data-assisted models to the target structures, it becomes clear that the region spanning residues 84–193 is modeled less accurately than the remaining N-terminal region (residues 15–83). In the crystal structure the residue segment 125–138 lacked any regular structure, but was modeled as an α-helix in the 3 best models (Figure 1C). The topology of both domains is captured reasonably well, but not their overall arrangement, which decreases the overall score of the full model. In detail, the region 15–83 of the best model superimposes very well with the corresponding region in the target structure (Cα-rmsd = 1.7 Å), whereas the region 84–193 aligns with a Cα-rmsd of 5.8 Å.

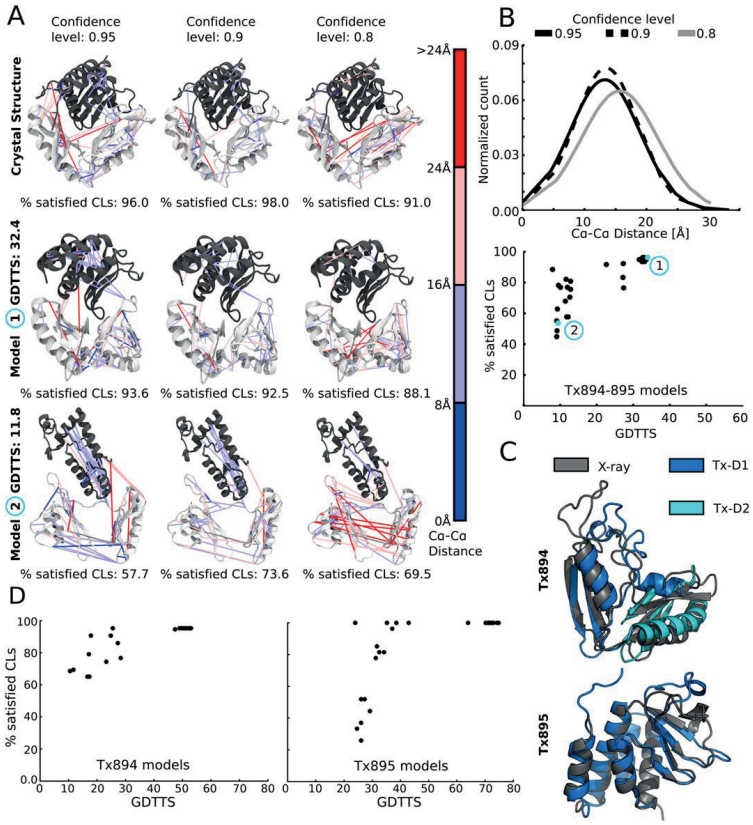



**FIGURE 2** Assessment of CLMS-based modeling for Tx894–895. **A**, Target crystal structure (top), best model 2 of Lee (middle) and worst model (bottom) in terms of GDTT for target Tx894 (in white) and Tx895 (in black). Pairs of residues expected to be close in space according to CLMS data at 3 confidence levels are connected by lines color-coded by the actual distance as measured in the target crystal structure. **B**, Top: Distribution of Cα-Cα distances, as measured in the target crystal structure, for pairs of residues crosslinked at confidence levels of 0.8, 0.9, and 0.95. The un-normalized and un-smoothened distributions can be found in the Supporting Information Figure S2. Bottom: Fraction of crosslinks satisfied by each Tx894–895 complex model against their GDTTS scores, for all submitted CLMS-assisted models. **C**, Structural alignment of the best Tx894 and Tx895 models against the target crystal structure. **D**, Fraction of crosslinks satisfied by each Tx894 and Tx895 model, separately, against their GDTTS scores, for all submitted CLMS-assisted models and based on distance cutoff value of 25 Å

Modeling based on CLMS data resulted in a modest improvement on the best predictions for this target, relative to the data-blind predictions; but did not contribute to a major improvement of the fold. Although part of this could be due to methodological problems or to problems with CLMS data quality, we investigated other potential explanations. CLMS data as provided by the experimentalist group consists of pairs of crosslinked residues accompanied by confidence levels. For Tx892 we observe that at the highest confidence level (0.95), 98% of the residue pairs actually have Cα-Cα distances lower than 25 Å, which is the upper bound suggested by Belsom et al.[22] (Figure 1A,B). At lower confidence levels the distributions of Cα-Cα distances shift slightly to higher values, but still most contacts remain within 25 Å (Figure 1A,B). One could propose that either (1) the density of residue-residue crosslinks across the whole protein is not high enough, or too heterogeneous, to effectively help modeling; or (2) the 25 Å cutoff distance is just too large to lead to concrete improvements in model quality. Visually from Figure 1A the CL density looks well distributed, with a slight tendency for crosslinked residues spanning α-helices to be around the central core of the protein. The effect of the rather long 25 Å upper bound distance of the CLMS-derived contacts would then be more important, especially considering that the protein's main axes measure around 49, 36, and 29 Å Indeed, the worse model, with a

GDTTS of 10.0, still satisfies >80% of the CL distances, even 94% at 0.95 confidence level (Figure 1A, bottom panel). As we show also in Figure 1A bottom panel, the worst model still satisfies most crosslinked pairs, giving support to our idea that the 25 Å upper bound distance for the CLMS experiment is way too large, at least for a protein of this size. Last, the observation that this (worst) model even satisfies CLMS constraints between a pair of residues located actually far from each other in the target structure (red line) shows (although it does not prove) that incorporation of some constraints might actually have been deleterious.

Targets Tx894 and Tx895 constitute a heterodimer, crystallized as such and now released in the PDB as 5hkg. Predictors were informed that these 2 targets form a dimer but were instructed to submit them as separate protomers that would represent their model of the dimer when put together. Tx895 is expected to be an easy target because sequence- and structure-level searches on the PDB returned sufficient templates.[34] Based on the same criteria, residue region 271–324 of Tx894 is expected to be easy to model, whereas residue region 182–270 is expected to be difficult.

GDTTS, Handedness, DFM, CoDM, and TM unequivocally indicate that the best models submitted for Tx894 were models 2, 4, and 5 from Goal with GDTTS ranging from 52.0 to 53.0, versus 59.1 for the best data-blind model. For Tx895, the best models were model 1 and 2 of Lee and model 4 of Goal with GDTTS from 72.9 to 74.8, versus 75.4 for the best data-blind model. For Tx894 and Tx895, respectively, the top 3 best models were structurally very similar to each other (<1.0 Å Cα-rmsd). Clearly in this case, incorporation of the CLMS data did not produce models better than the best models submitted in the data-blind category; actually, they are slightly worse (see below).

From the alignment of Tx894 and Tx895 best models (model 2 of Goal for Tx894 and model 2 of Lee for Tx895, Figure 2C) against their respective crystal structure, we observed that the target structure topology was well respected, especially in regions where secondary structure elements were defined. Structural disagreements were found mostly in flexible/disordered regions. This trend is best illustrated in the modeling of Tx895 (Figure 2C, bottom panel), in which the first 86 out of 120 residues found in secondary structure regions aligned very well with the respective crystal structure (<1.0 Å Cα-rmsd), but the inclusion of the following region (residues 87–124) resulted in a higher Cα-rmsd of 4.4 Å.

Both the Tx894 and Tx895 best models were modeled worse than in the best models of the data-blind category. In order to rationalize this lack of improvement, we first assessed the agreement between the CLMS experimental data and the target crystal structure (pdb id 5hkg). In this case, at the highest confidence interval (0.95), 96% of the Cα-Cα crosslinked residue distances were found below 25 Å (Figure 2A, top panel). Even, at lower confidence intervals, still > 90% of Cα-Cα crosslinked residue distances were found below 25 Å (Figure 2A,B, top panel). Then we assessed the agreement between CLMS data and models (Figure 2B, bottom panel and 2D). In this case, the best models were found to have a high fraction of satisfied CL distances, but were of low GDTTS (<35.0). This again highlights the fact that a threshold of 25 Å might be too permissive and allow residues far in the structure to be considered as contacts.

Moreover, if the predictors used all data with confidence level above 0.8, it might have happened that this 10% of incorrect CLMS-derived contacts led to deformations, and thus worsening, of the models.

## 3.2 | SAXS-driven data-assisted modeling

CASP12 featured for the first time the inclusion of SAXS data to assist target modeling. In this SAXS-assisted category, 10 predictors submitted models for 11 targets with varying degree of oligomerization and templates availability. The information available to the predictors was the amino acid sequence of the targets and their experimentally derived SAXS profiles. Here, we analyze the models produced by the predicting groups which used SAXS data to assist their modeling, in comparison to the target structures. We then interpret the results in context of modeling difficulty, available data-blind models and quality of the SAXS data.

On a general note, it is important to mention that the stoichiometry of models predicted by SAXS was not always consistent with the crystal structure of the target, like in the cases of Ts866 and Ts886, which consisted of 6 and 12 subunits, respectively, in their crystal structure (PDB id 5uw2 and 5fhy), but were observed to be different in SAXS[23] (Table 1). Moreover, the fit between SAXS experiments and target structures should be interpreted with care since a significant proportion of flexible residues (2–56%, Table 1),[23] which are present in the target sequence, were unresolved in the crystal structures. Ts886 in particular is a very elongated protein and was found to be very flexible from the non-linearity observed in the Guinier plots (Supporting Information Table S1 and Figure S1), and missed 50% if its residues in the crystal structure (Table 1). Therefore, given the large discrepancy the modeling of Ts886 will not be further discussed during this assessment, where we report a detailed analysis of the most notable predictions assisted by SAXS data.

Target Ts909 is present as a trimer in its crystal structure (pdb id 5q5n) and is expected to be easy to model having a good structural template readily available (pdb id 3suc: chain A).[34] All the scoring schemes used point to model 1 and 2 of Lee, model 2 from Grudinin and model 5 of Kias-Gdansk as highest quality models with GDTTS ranging from 56.2 to 61.9 versus 62.0 for the data-blind predictions. Therefore, inclusion of the SAXS data did not lead to significant global improvements.

Structural alignments between the best model of Lee and 1 protomer of the target structure showed good superimposition of the β-strand regions with respect to the same atoms in the crystal structure (1.6 Å Cα-rmsd with 227 out of 333 residues, Figure 3A, top panel), but the addition of the flexible loops was responsible for the decrease in model quality (6.5 Å Cα-rmsd). It would have been interesting to understand how the SAXS data, which captured structural information of the whole assembly, was used during the modeling of Ts909. Unfortunately, since predictors were requested

**TABLE 1** Summary information of SAXS-assisted target prediction[a]

| | Target | Stoichiometry from SAXS | Mass (kDa) from SAXS | Mass (kDa) from seq. | χ Target (X-ray) | % seq missing in x-ray |
|---|---|---|---|---|---|---|
| Monomers | Ts896 | monomer | 24 | 54 | 2.77 | 10 |
| | Ts899 | monomer | 57 | 47 | 6.94 | 15 |
| | Ts941 | monomer | 45 | 51 | 17.42 | 56 |
| | Ts942 | monomer | 65 | 54 | 11.16 | 21 |
| | Ts947 | monomer | 18 | 25 | 24.65 | 20 |
| Multimers | Ts894–895 | heterodimer | 30 | 51 | 26.83 | 2 |
| | Ts886 | dimer | ... | 39 | 1394.38 | 50 |
| | Ts901 | Filament | 216 | 36 | ... | 11 |
| | Ts909 | Trimer | 85 | 37 | 21.45 | 10 |
| | Ts866 | Large Heteromer | 142 | 13 | ... | 36 |

[a]The data relative to target stoichiometry, molecular mass, and percentage of residue unresolved in X-ray were obtained from ref.[23]

to submit an isolated protomer, this analysis could not be performed.

Targets Ts894 and Ts895, presented in the CLMS section as Tx894–895, are the only targets in this CASP edition for which 2 different sources of data were provided for assisting predictions. Unfortunately, CLMS and SAXS data were not simultaneously used, thus, while we described above how CLMS affected prediction, we separately report here the role of SAXS data on modeling. The best models for Ts894 according to our 6 scoring metrics were model 2 of Grudinin and models 1 and 3 by the Lee group, with GDTTS of 51.4 and 50.5, respectively, versus 59.1 for the best data-blind model. For Ts895, the 5 best models were submitted by the Lee group and were completely

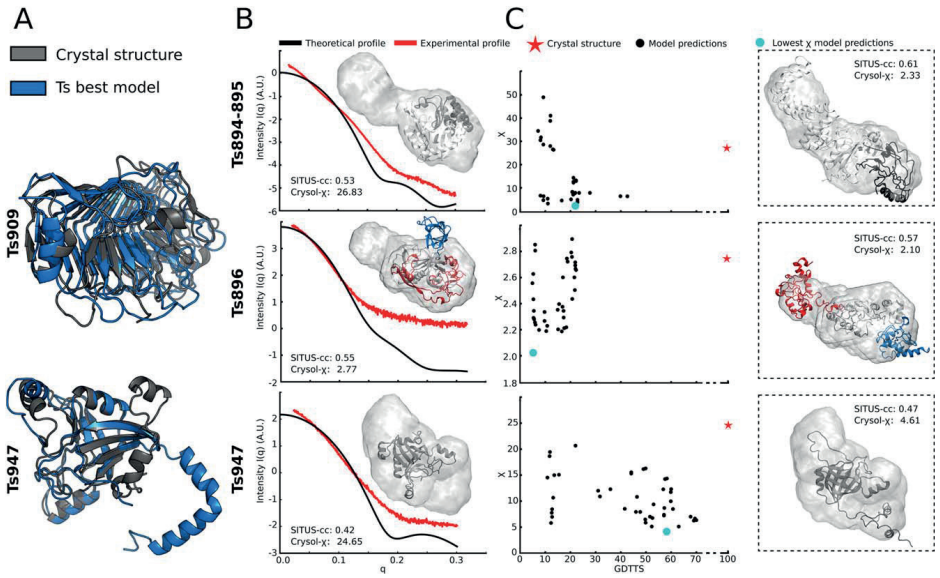



**FIGURE 3** Assessment of SAXS-based modeling for targets Ts894–895, Ts896, Ts909, and Ts947. **A**, Structural alignments of best Ts909 and Ts947 models against their respective crystal structure. **B**, Evaluation of fit between Ts894–895, Ts896, and Ts947 crystal structure and SAXS data. Residue regions 39–124, 125–325, and 326–486 of Ts896 were colored in blue, red, and gray, respectively. **C**, χ value associated with each Ts894–895, Ts896, and Ts947 model, separately, against their GDTTS scores, for all submitted models. Models with lowest χ value were extracted, fitted into their SAXS envelope and color-coded similarly to the crystal structure

identical to each other with GDTTS of 72.5 versus 75.4 for the best data-blind model. Also in this case, the inclusion of the SAXS data did not lead to any clear improvement.

The main structural differences between the crystal structure of Ts894 against the model 2 of Grudinin, and the crystal structure of Ts895 against model 5 of Lee, came from the misalignment of the flexible regions. Indeed, when aligning the best Ts895 model to its crystal structure, a Cα-rmsd of <1.0 Å was obtained using 83 out of 120 residues which were associated with secondary structure elements, whereas the addition of the remaining 37 residues, which were part of flexible regions (residues 1–15, 54–68, and 74–81), increased the Cα-rmsd to 4.5 Å. From this analysis, it can be inferred that the lack of improvement in the SAXS-assisted models for Ts894–895 was due to structure differences in the flexible regions.

The differences we observe do not exclude though that they could be due to the integration of SAXS data pointing to protein conformations different than observed in the crystal structure. In fact, the Ts894–895 crystal structure did not fit very well the SAXS data ($\chi = 26.8$, Figure 3B, top panel), and could point to the presence of different dimeric structure conformations in the sample solution. Analyzing using EPPIC[35] whether any of the dimer configurations present in the crystal unit cell of Ts894–895 (PDB id 5hkq) could alternatively and better fit the SAXS data led to no suitable dimeric conformation. Models with the lowest $\chi$ values did not resemble the crystal structure (Figure 3C, top panel), showing that the predictors' efforts in using the SAXS data actually led them to make models worse. Overall, we conclude that the SAXS data was not overall useful to model target Ts894–895.

Target Ts896 had several templates for the region 39–325, which itself could be visually separated into 2 regions (residues 39–124 and 125–325) linked by a long loop (residues 121–127), but no clear templates for region 326–486,[34] making the later hard to model. GDTTS- and QCS-based ranking points at model 1 and 3 of Lee as best models, while TM ranking puts model 2 of Multicom as being the best models, even though their GDTTS were low (21.9–22.2 versus 24.5 for the best data-blind model). Such low GDTTS values, however, are close to what "random spaghetti" models would have, so the values are hard to compare. The QCS metric[30] compares secondary structures at low GDTTS and is designed to match the results of visual inspection. For model 3 of Lee and the best data-blind models, QCS reaches a value of 55.1, meaning that there is some resemblance of these models to the target.

Visually, the 2 models by the Lee group were structurally similar to each other (2.7 Å Cα-rmsd). These 2 models and Multicom's model appeared different from the crystal structure when considering Cα-rmsd (>20 Å), despite the high QCS score.

We then aligned separately the regions and found that the secondary structure region of region 39–124 was well modeled (1.1 Å Cα-rmsd when considering 61/86 residues and 5.6 Å when all residues considered), but the other regions were not so well modeled (>10 Å Cα-rmsd). It is possible that also that their spatial arrangement was not captured. Since SAXS information can help infer spatial arrangement between structural domains,[4] we evaluated its agreement against the crystal structure and submitted models (Figure 3B, middle panel).

Though the crystal structure had a low $\chi$ (2.8) with respect to the SAXS data, computing the $\chi$ of all Ts896 models showed a tendency to decrease GDTTS upon minimizing $\chi$. This means that incorporating SAXS data during modeling effectively decreased model resemblance to the crystal structure, explaining why Ts896 models failed to improve over the data-blind models.

Target Ts947 is expected to be very easy to model, as there are several templates available in the Protein Data Bank.[34] Ideally, incorporation of experimental data should help to refine details about loop conformations. In this case, the 5 best models submitted were all produced by Lee, with GDTTS scores ranging from 67.1 to 70.0 compared to 65.9 for the best model obtained in the data-blind category. These 5 models were structurally very similar in the C-terminal region spanning residues 97–216 (<0.4 Å Cα-rmsd) and differed in the N-terminal region spanning residues 42–96 (>3.0 Å Cα-rmsd). The best model 1 of Lee was also found to be very similar to the crystal structure, especially in regions were secondary structure elements were defined (Cα-rmsd < 1.0 Å when considering 119 out of 175 residues, Figure 3A bottom panel). Adding the remaining 56 residues, associated with more flexible parts, decreased the quality of the alignment (8.9 Å Cα-atoms).

For this target, although the crystal structure did not seem to fit the SAXS data well ($\chi = 24.7$, Figure 3B, bottom panel), the models do show a tendency to increase their GDTTS upon minimizing $\chi$. This suggests, somewhat paradoxically, that using SAXS data, which was seemingly inconsistent with the crystal structure, did help in modeling of Ts947.

## 4 | DISCUSSION

### 4.1 | CLMS-assisted category

In the CLMS-assisted category the 2 best performing groups were Goal and Lee (Figure 4B, top and middle panel), both of which came from the same group. Similar to the Kias-Gdansk predicting group, Lee/Goal[36] adds CLMS data as restraints to the energy function describing the quality of their models.

Compared to the global best models submitted in the data-blind category, Lee/Goal showed marginal to no improvement in ΔGDTTS ranging from −5.4 to 6.8 for the 3 targets, and was the only group to improve relative to its own data-blind models with improvement for target Tx892 as high as 21.5 ΔGDTTS (Figure 4B, bottom panel). This means that unlike the other groups, Lee/Goal managed not to decrease the quality of their models by adding CLMS data. In fact, the other 6 groups that submitted models in both categories worsened the quality of their models (mean ΔGDTTS < 0.0, Figure 4A,B, middle panels).

Generally, the quality of the best models submitted for the 3 targets in the CLMS category did not improve relative to the best models submitted in the data-blind category (Figure 4A, top panel), even though they were found to satisfy most of the CL distances under the distance cutoff of 25 Å at the highest confidence level (0.95). In particular, this trend was observed when evaluating the fraction of satisfied CLs both for the CLMS-assisted (Figure 1) and data-blind Tx892 models (Supporting Information Figure S3).

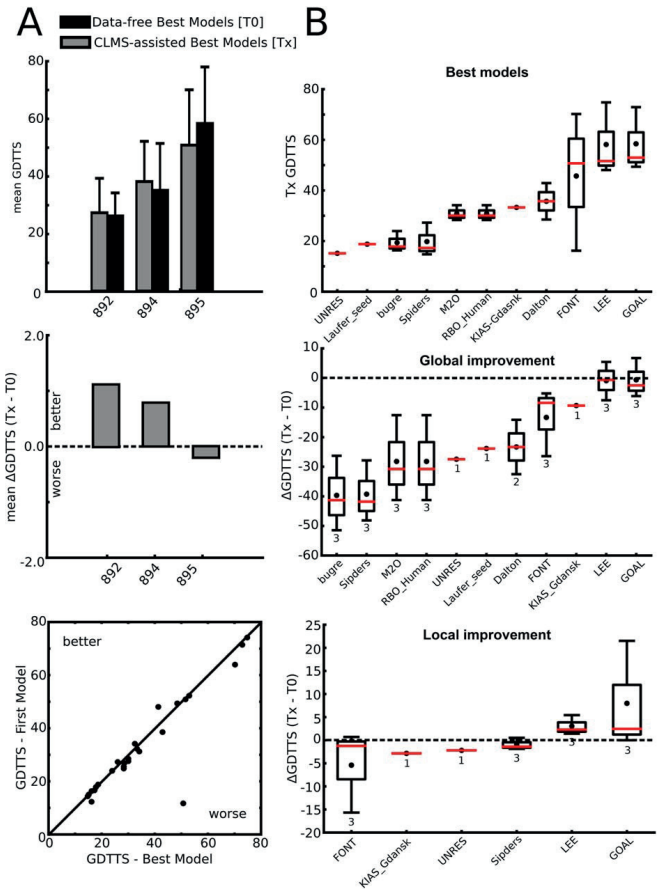



**FIGURE 4** Evaluation of model quality in the CLMS-assisted category. **A**, Top: Evaluation of best models submitted for each target and comparison to best data-blind models. Middle: Improvement analysis of CLMS-assisted over data-blind models. Only groups who submitted in both categories were considered. Mean $\Delta$GDTTS $> 0.0$ implied that groups improved on their data-blind model while using CLMS data and conversely for values $<0.0$. Bottom: Evaluation of groups ability to correctly estimate the quality of their model. The GDTTS of what groups considered as their best models (model 1) was plotted against the GDTTS of the best model of the remaining 4. Points above the diagonal separation line represent models correctly estimated as best by predictors. **B**, Top: Evaluation of best models submitted by each group for all targets. The dot inside the boxplots represents the mean GDTTS (vertical axis) of the best models. Middle: Evaluation of global improvement of CLMS-assisted over data-blind best models for each group. The numbers below the boxplots indicate the number of targets each group submitted models for. Bottom: Evaluation of local improvement of CLMS-assisted over data-blind best models for each group. Only groups that submitted models in both the CLMS assisted and data-blind were considered

On top of possible problems in the methodologies used by the predictors, we envision 2 main reasons as to why integrating CLMS data to the models of targets did no increase their quality. First, this could be related to the density distribution of crosslinks, which could be in turn related to a number of factors including surface accessibility, environmental reactivity or digestion cleavage site distribution.[20] Second, the cutoff distance of 25 Å to discriminate contacting crosslinked residues in models was too permissive to accurately determine true residue contacts. When assessing the relationship between cutoff distance and ability to separate good models from bad (in term of GDTTS, Supporting Information Figure S4), we found that lowering the cutoff value did not result in a better estimation of model quality. This leaves us with the suggestion that the chemical spacer used to crosslink residue was probably too long and could crosslink distant residues in the target structure.

Moreover, conflicting contacts were observed in the target crystal structures and appeared as residues expected to form crosslinks but which are found separated by long distance in the crystal structure. Conflicting contacts could be either due to wrongly assigned cross-linked residues from the MS analysis or could point to the existence of extensive protein flexibility (either fast dynamics, slow conformational fluctuations, or actually having multiple static conformers) that results in artefactual pairs of residues appearing in contact despite being far in the target. Moreover, conflicting contacts can also be caused by the presence of other protein particles, which would form inter-protein crosslinks. Unfortunately, the presence of such conflicting data increases the challenges for prediction as one must find a way to filter out the conflicting contacts during the modeling process[37–39] or include only the residue pairs at high confidence intervals as distance constraints during modeling.

Finally, all the 11 groups submitting CLMS-based models had difficulty estimating the quality of their own models with only 6 out of the 23 correct estimations of their first model as their highest quality model (Figure 4A, bottom panel).

### 4.2 | SAXS-assisted category

The models predicted in the SAXS-assisted category did not show significant improvement when compared to the data-blind ones. Nevertheless, the top performers in this category (Lee, Kias-Gdansk, and Grudinin, Figure 5B, top-right panel) performed relatively well on the homo-multimeric targets Ts866 and Ts909. This means that unlike other groups, the top performers did not decrease the quality of their models when adding SAXS data.

This result was found interesting since the task of modeling structures in a multimeric state is often more complicated than that of modeling monomers in isolation. This is because the proximity of 1 protomer can influence the conformations of others. Predicting this type of conformational rearrangement is still a longstanding problem in structural biology[40] and constitutes one of the main challenges to be faced in the Critical Assessment of PRediction of Interaction (CAPRI[41]), as further described in ref.[42]

We also observed that the best Ts866 and Ts909 models submitted by these groups shared close structural similarities to the best models submitted in the data-blind category, which could be explained by the fact that some predictors used pre-computed models extracted from widely available webservers as a starting point for their own prediction. For instance, Kias-Gdansk extracted restraints from models computed from webservers, such as Goal,[36] Rosetta-Baker,[43] Quark,[44] and iTasser[45] in order to drive the modeling of their predictions. For homo-oligomeric targets, Kias-Gdansk initially assembled the protomers into their multimeric state and refined their best multimeric models with SAXS data.

The modeling strategy consisting in refining pre-computed models with SAXS data was also employed by the Grudinin group. In the case of homo-oligomeric targets, the Grudinin group started by modeling the complex by first assembling the multimeric structure using SAM,[46] a symmetric assembly protocol based on fast Fourier transform. Their best multimeric models were then relaxed by minimizing the $\chi$ values with respect to pre-computed normal modes.[47]

Other positive results for the top performers include monomeric targets Ts942 and Ts947. For these targets, the Grudinin group improved the quality of their own Ts942 and Ts947 models over their T0 equivalents with $\Delta$GDTTS of 9.9 and 17.7, respectively. The predictors Grudinin, Lee, and Multicom, used amongst other metrics $\chi$ to quantify the difference between the experimental and theoretical SAXS profiles.

The group Multicom[48] used not only the $\chi$ value but also other metrics, including radius of gyration ($R_g$), to quantify the fit between models and SAXS data. Unlike most groups, Kias-Gdansk did not use the $\chi$ value but the SAXS distance-distribution profile converted to a maximum-likelihood penalty and used as a restraint term inside their pseudo-energy function. During our analysis, we could not see any meaningful improvement in terms of model quality between these groups compared to others that minimized only the $\chi$ value (Figure 5B).

Despite a few remarkable results achieved by the individual top performing groups of this category, generally no meaningful improvement was observed amongst groups in term of mean GDTTS over the data-blind category (Figure 5B, middle panel). Although for some targets using the SAXS data tended to increase model GDTTS, as for Ts942 and Ts947, (Figure 3B, bottom panel), they overall failed to globally improve over the data-blind ones (Figure 5A). For these models, secondary structure elements were in general well modeled, unlike the disordered/flexible regions, which contributed to decrease their GDTTS (Figure 3A, bottom panel). Though structural information related to the flexible/disordered protein regions is described by the SAXS data, it is unlikely that including it would improve the modeling of these regions to resemble one of the crystal structure. Briefly, this is due to the fact that SAXS data captures the contribution from all conformations a protein can adopt in solution, while the X-ray structure represents only one of those conformations trapped in the crystal lattice.

On the other hand, there were models that were globally badly modeled (mean GDTTS $< 25.0$) and where using SAXS data had the opposite effect of decreasing model quality. For Ts896, Ts899, and Ts941 in particular, minimizing $\chi$ tended to lower their GDTTS (Figure 3B, Middle panel). This implies that predicting groups relying on minimization of the $\chi$ such as Grudinin or Lee, are likely to be led astray. One way to solve this issue for future CASP rounds would be to enforce the structural compatibility between SAXS data and target crystal structures, prior to distribution of data for predictions.

Similarly, to the CLMS-assisted category, groups had difficulties estimating the relative quality of their own models. This was illustrated in Figure 5A, bottom panel, with only 28 of the 82 models correctly assigned as first best model.

### 4.3 | Open challenges in data-assisted protein structure prediction

The current installment of CASP12 featured proportionally more difficult targets compared to the ones presented in CASP11.[34] Despite this difficulty, the outcome of this edition of data-assisted prediction can

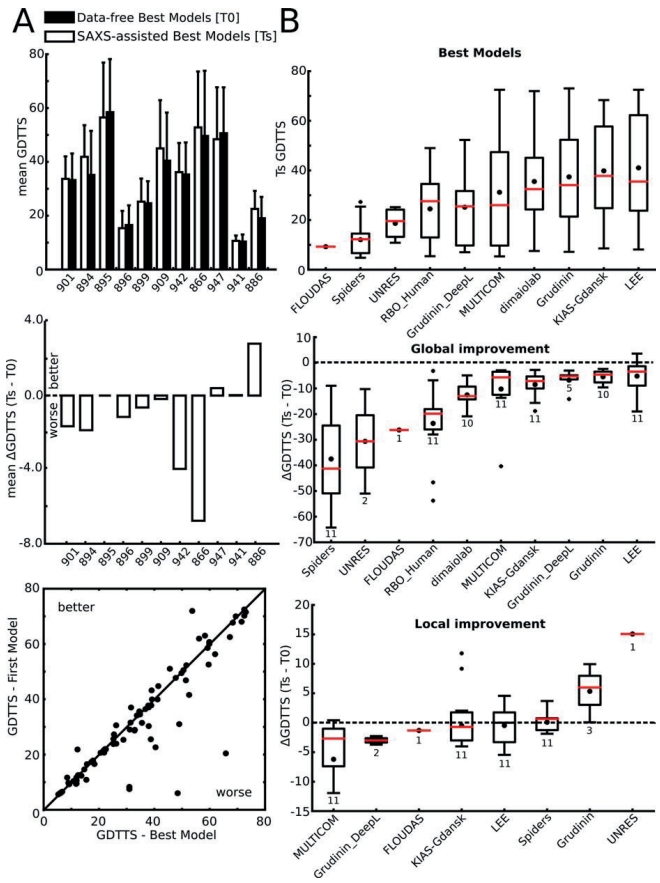



**FIGURE 5** Evaluation of model quality in the SAXS-assisted category. **A**, Top: Evaluation of best models submitted for each target and comparison to best data-blind models. Middle: Improvement analysis of SAXS-assisted over data-blind models. Only groups who submitted in both categories were considered. Mean $\Delta$GDTTS $>$ 0.0 implied that groups improved on their data-blind model while using SAXS data and conversely for values $<$0.0. Bottom: Evaluation of groups ability to correctly estimate the quality of their model. The GDTTS of what groups considered as their best models (model 1) was plotted against the GDTTS of the best model of the remaining 4. Points above the diagonal separation line represent models correctly estimated as best by predictors. **B**, Top: Evaluation of best models submitted by each group for all targets. The dot inside the boxplots represents the mean GDTTS (vertical axis) of the best models. Middle: Evaluation of global improvement of SAXS-assisted over data-blind best models for each group. The numbers below the boxplots indicate the number of targets each group submitted models for. Bottom: Evaluation of local improvement of SAXS-assisted over data-blind best models for each group. Only groups that submitted models in both the SAXS-assisted and data-blind were considered

be considered promising, providing the fact that this category is still in its infancy and a completely new data source was introduced for the first time (i.e., SAXS). Generally, for models submitted in the data-assisted category, no significant improvement was observed with respect to the data-blind ones. Therefore, there still is a number of challenges that need to be faced in order to truly benefit from the potential of integrating additional experimental for assisting CASP experiments.

Similar to previous CASP editions the time allocated to submit models was short ($\sim$2 weeks for regular prediction and another $\sim$2 weeks for data-assisted prediction), which constitutes one of the main challenges faced by the predictors. One way some predictors bypassed this issue was to start from models pre-computed from modeling webservers. For instance, this strategy was adopted by the predicting groups Kias-gdanks, Grudinin, and Spiders, who optimized pre-computed structures modeled from well-established webservers.

The other time-related challenge is linked to CLMS and SAXS data collection. This task has to follow the deposition of targets and requires an additional level of management, which will likely need to be further automatized with the experimental groups and/or facilities in order to provide data in a time-effective manner. A period of only 2 weeks between sample arrival and the CASP deadline for prediction was available to perform data collection and analysis. For the SAXS experimentalists, such tight timeframe possibly hindered the task of performing thorough testing and quality assessment of the data. Surely, with time the workflow of this pipeline will be improved, but extending the time needed for prediction within this category could be anyway beneficial for both supporting experimentalists and predictors.

Another challenge is related to the small amount of targets and low participation of groups in the current data-assisted category, which produced a small statistics. With growing advances both in the experimental techniques and in methods development, we expect not only more targets to be featured, but also more groups to actively participate in the data-assisted category. This will eventually increase the statistical relevance of the assessment and enable to draw clearer conclusions on the significance of data-assisted model prediction. As for now, despite an overall quite poor improvement in data-assisted modeling, anecdotic cases showed a promising potential for integrative modeling applied to protein structure prediction.

Eventually, the poor improvement in data-assisted models might be related to the fact that both CLMS and SAXS data could have captured different protein conformations than that of the target crystal structure. In the case where crystal or NMR structures are the only reference to structurally validate a model, full data consistency between the experimental data and target structures would have to be ensured before the prediction period begins. Though this approach would certainly be more demanding and time consuming, it would provide true complementarity between the structural information featured in the CLMS, SAXS and crystallography experiments, and ensure a more stringent assessment of predictions.

In particular for the 3 targets featured in the CLMS-assisted category, the models were found to satisfy most of the crosslinked residue distances under 25 Å but globally failed to improve over the data-blind models. We suggest this problem is related to the length of the crosslink spacer used for the CASP12 experiment, which might be too long and could erroneously indicate contacts between residues that are far apart in the structure. A possible suggestion for the next CASP rounds could be to choose a shorter crosslink spacer. However, compatibility with the photo-CLMS protocol established by Brock and Rappsilber[22] would have to be ensured. For small proteins, as the ones featured in the CLMS-assisted category, such long spacer would not enable to accurately pinpoint residue contact, but would certainly be helpful for bigger proteins.

Moreover, depending on their size and complexity, crosslink spacer as well as the residues they crosslink could be very flexible and wrap around protein surfaces. Subsequently, 2 crosslinked residues could in principle be separated by a distance different than the one specified by the crosslink spacer length.[49,50] In this case Euclidian distances used as constraints, as was done by Lee/Gaol, Kias-Gdansk, and other studies,[37,38,48] either during modeling or as a selective filter on the computed models, are not adequate to capture this flexibility. Using computational methods such as Xwalk,[51] which uses non-linear distances to describe crosslink interaction between residues could help.

Furthermore, false contacts predicted by MS can be mixed with correct ones that are usually difficult to distinguish. In order to deal with such complication, methods using a Bayesian framework[52,53] or which output models maximizing the number of satisfied constraints have been implemented.[38] Although such approaches are more computationally demanding, we think they could be useful in better treating CLMS data and could lead to improvement in model quality by more efficiently filtering out false positives.

As for the SAXS-assisted category, generally no meaningful improvement was observed. In cases where the GDTTS of models was high but not higher than the best data-blind models, the overall fold of the proteins was well modeled, especially in the well-defined secondary structure regions. In these cases, the difference between the crystal structure and the SAXS-assisted model came from the flexible/disordered regions, which information is captured by the SAXS experiments but could not be used to model structures to resemble the X-ray structures. This point to a more general issue related to the fact that crystallography and SAXS experiments might be capturing different conformations of the same target and such would prove difficult to combine structurally.

Moreover, it also happens that the interface between protomers described in the crystal structures could rather be artifacts of the technique rather than a real physiological interface. In such cases, combining the SAXS information describing the solution assembly to the crystal interfaces would be unsuccessful. Thus, in order to reliably score models obtained by adding SAXS information, a different metric to evaluate model quality by including the flexibility that SAXS describes would be preferable. As suggested in Ref. [20] this new metric could consist in inferring for instance the function of the target from the modeled structure or in answering specific biological questions in order to guide life-scientists to propose new experiments. In fact, we do not discard, especially for oligomeric complexes, the possibility that good models originating from SAXS data integration (i.e., having low $\chi$) would prove to be more biologically relevant, that is, having loops of domain conformations more suitable to perform their biological function than their respective models more closely resembling the X-ray conformer (i.e., having low GDTTS).

In this CASP, we noticed a case of possible sample aggregation for Ts886 and a different oligomerization state of Ts866 when compared to their reference crystal structures. Together with ensuring complementarity between experiments and target crystal structure, supplementing the SAXS experiments with size-exclusion-chromatography (SEC-SAXS) would likely improve the separation of aggregates from sample and ensure a better sample quality for future CASP rounds. Additionally, more precise methods such as multi-angle laser light scattering (MALLS) could be used to estimate molecular weight and eventually get a better estimation of the target complex stoichiometry, even though this will certainly increase the complexity and timeframe needed for the supporting experimental pipeline.

## 5 | CONCLUSIONS

This edition of CASP featured CLMS-assisted and SAXS-assisted categories, made possible by the efforts of the CASP organizing committee that managed to coordinate work from multiple supporting groups. Although for a few models some groups performed slightly better than in the top data-blind category, generally no significant improvements were noticed upon inclusion of CLMS and SAXS data. Nevertheless, due to its early age, added to the small number of targets and of participating groups when compared to other well-established CASP categories, we feel that it is still premature to draw any definitive conclusion on the impact of experimental data on structure prediction. We have suggested some general recommendations regarding both the improvement of the assessment criteria for data-assisted prediction and the experimental data acquisition and quality control. We hope that the future CASP rounds will increasingly feature more targets with additional experimental information and more participation from the predicting groups. With the recent advances of the cryo-EM field, both in terms of resolution and data acquisition, we think that the inclusion of medium-to-low resolution cryo-EM data for future editions of this CASP category would be an interesting extension. Eventually, in the true spirit of integrative modeling, we think that having a category which simultaneously combines all sorts of experimental data would have to be the final and overarching goal for data-assisted protein structure and assembly prediction in CASP. Crosslinking data for instance has already been shown to be compatible with several low resolution data including SAXS.[4] All of the above would enable the collection of relevant data, which in turn will contribute to give a clearer picture on the impact of integrative modeling methods for tertiary and quaternary protein structure prediction.

### ACKNOWLEDGMENTS

We would like to dedicate this work to Anna Tramontano, we thank her for the unstoppable dedication to the CASP initiative and for her continuous support and motivation during our assessment, and to Guido Capitani with whom we shared this experience as CASP assessors. We would like to thank the experimentalists from the SIBYLS Beamline facility at the Advanced Light Source Synchrotron for useful discussion and suggestions. Finally, we thank the CASP12 organizers for the invitation to participate in target domain definition and classification, and as assessors for the topology and data-assisted modeling evaluations; and to the groups that contributed structures and experimental data for CASP12.

### ORCID

*Luciano A. Abriata* http://orcid.org/0000-0003-3087-8677
*Giorgio E. Tamò* http://orcid.org/0000-0003-1634-9357

### REFERENCES

[1] Tamo GE, Abriata LA, Dal Peraro M. The importance of dynamics in integrative modeling of supramolecular assemblies. *Curr Opin Struct Biol.* 2015;31:28–34.

[2] Rodrigues JP, Bonvin AM. Integrative computational modeling of protein interactions. *FEBS J.* 2014;281(8):1988–2003.

[3] Ward AB, Sali A, Wilson IA. Biochemistry. Integrative structural biology. *Science.* 2013;339(6122):913–915.

[4] Schneidman-Duhovny D, Kim SJ, Sali A. Integrative structural modeling with small angle X-ray scattering profiles. *BMC Struct Biol.* 2012;12(1):17

[5] Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res.* 2016;44(W1):W424–W429.

[6] Karaca E, Bonvin AM. On the usefulness of ion-mobility mass spectrometry and SAXS data in scoring docking decoys. *Acta Crystallograph D Biol Crystallogr.* 2013;69(Pt 5):683–694.

[7] Vijayvargia R, Epand R, Leitner A, et al. Huntingtin's spherical solenoid structure enables polyglutamine tract-dependent modulation of its structure and function. *eLife.* 2016;5:e11184

[8] Ovchinnikov S, Park H, Varghese N, et al. Protein structure determination using metagenome sequence data. *Science.* 2017;355(6322):294–297.

[9] Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell.* 2012;149(7):1607–1621.

[10] Malinverni D, Marsili S, Barducci A, De Los Rios P, Punta M. Large-scale conformational transitions and dimerization are encoded in the amino-acid sequences of Hsp70 chaperones. *PLoS Comput Biol.* 2015;11(6):e1004262

[11] Morcos F, Pagnani A, Lunt B, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA.* 2011;108(49):E1293–E1301.

[12] Sali A, Berman HM, Schwede T, et al. Outcome of the first wwPDB hybrid/integrative methods task force workshop. *Structure.* 2015;23(7):1156–1167.

[13] Degiacomi MT, Iacovache I, Pernot L, et al. Molecular assembly of the aerolysin pore reveals a swirling membrane-insertion mechanism. *Nat Chem Biol.* 2013;9(10):623–629.

[14] Alber F, Dokudovskaya S, Veenhoff LM, et al. The molecular architecture of the nuclear pore complex. *Nature.* 2007;450(7170):695–701.

[15] Upla P, Kim SJ, Sampathkumar P, et al. Molecular architecture of the major membrane ring component of the nuclear pore complex. *Structure.* 2017;25(3):434–445.

[16] Kudryashev M, Stenta M, Schmelz S, et al. In situ structural analysis of the *Yersinia enterocolitica* injectisome. *eLife.* 2013;2:e00792

[17] Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction: progress and new directions in round XI. *Proteins.* 2016;84:4–14.

[18] Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins.* 2003;53(Suppl6):334–339.

[19] Abriata LA, Tamò GE, Monastyrskyy B, Kryshtafovych A, Dal Peraro M. Assessment of hard target modeling in CASP12 reveals an emerging role of alignment based contact prediction methods. *Proteins.* 2017. https://doi.org/10.1002/prot.25423.

[20] Schneider M, Belsom A, Rappsilber J, Brock O. Blind testing of cross-linking/mass spectrometry hybrid methods in CASP11. *Proteins.* 2016;84:152–163.

[21] Kinch LN, Li WL, Monastyrskyy B, Kryshtafovych A, Grishin NV. Assessment of CASP11 contact-assisted predictions. *Proteins.* 2016;84:164–180.

[22] Belsom A, Schneider M, Brock O, Rappsilber J. Blind evaluation of hybrid protein structure analysis methods based on cross-linking. *Trends Biochem Sci.* 2016;41(7):564–567.

[23] Ogorzalek LT, Hura GL, Belsom A, et al. Small angle X-ray scattering and cross-linking for data assisted protein structure prediction in CASP 12 with prospects for improved accuracy. *Proteins.*

[24] Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP11 statistics and the prediction center evaluation system. *Proteins.* 2016;84(Suppl1): 15–19.

[25] Hura GL, Menon AL, Hammel M, Rambo RP, Poole FL, Tsutakawa SE, Jenney FE, Jr., Classen S, Frankel KA, Hopkins RC, Yang SJ, Scott JW, Dillard BD, Adams MW, Tainer JA. Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). *Nat Methods.* 2009;6:606–612.

[26] Svergun D, Barberato C, Koch MHJ. CRYSOL: a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J Appl Crystallogr.* 1995;28(6):768–773.

[27] Wriggers W, Milligan RA, McCammon JA. Situs: a package for docking crystal structures into low-resolution maps from electron microscopy. *J Struct Biol.* 1999;125(2–3):185–195.

[28] Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. *Proteins.* 1999;(Suppl3):22–29.

[29] Herbert A, Sternberg M. MaxCluster: a tool for protein structure comparison and clustering. 2014. http://www.sbg.bio.ic.ac.uk/~maxcluster/

[30] Kinch L, Yong Shi S, Cong Q, Cheng H, Liao Y, Grishin NV. CASP9 assessment of free modeling target predictions. *Proteins.* 2011;79 (S10):59–73.

[31] Tai CH, Bai H, Taylor TJ, Lee B. Assessment of template-free modeling in CASP10 and ROLL. *Proteins.* 2014;82:57–83.

[32] Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins.* 2004;57(4):702–710.

[33] DeLano WL. The PyMOL molecular graphics system. http://pymol.org 2002.

[34] Abriata LA, Kinch LN, Tamò GE, Monastyrskyy B, Kryshtafovych A, Dal Peraro M. Definition and classification of evaluation units for tertiary structure prediction in CASP12 facilitated through semi-automated metrics. *Proteins.* https://doi.org/10.1002/prot.25403.

[35] Duarte JM, Srebniak A, Scharer MA, Capitani G. Protein interface classification by evolutionary analysis. *BMC Bioinformatics.* 2012;13:334

[36] Joo K, Joung I, Lee SY, et al. Template based protein structure modeling by global optimization in CASP11. *Proteins.* 2016;84:221–232.

[37] Habeck M, Rieping W, Nilges M. Weighting of experimental evidence in macromolecular structure determination. *Proc Natl Acad Sci USA.* 2006;103(6):1756–1761.

[38] Tamo G, Maesani A, Trager S, Degiacomi MT, Floreano D, Dal Peraro M. Disentangling constraints using viability evolution principles in integrative modeling of macromolecular assemblies. *Sci Rep.* 2017;7(1):235

[39] Molnar KS, Bonomi M, Pellarin R, et al. Cys-scanning disulfide crosslinking and bayesian modeling probe the transmembrane signaling mechanism of the histidine kinase, PhoQ. *Structure.* 2014;22 (9):1239–1251.

[40] Kuroda D, Gray JJ. Pushing the backbone in protein-protein docking. *Structure.* 2016;24(10):1821–1829.

[41] Lensink MF, Wodak SJ. Docking, scoring, and affinity prediction in CAPRI. *Proteins.* 2013;81(12):2082–2095.

[42] Lafita A, Bliven S, Kryshtafovych A, Bertoni M, Monastyrskyy B, Duarte JM, Schwede T, Capitani G. Assessment of protein assembly prediction in CASP12. *Proteins.* 2017. https://doi.org/10.1002/prot.25408.

[43] Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins.* 1999; 171–176.

[44] Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins.* 2012;80(7):1715–1735.

[45] Zhang Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics.* 2008;9:40

[46] Ritchie DW, Grudinin S. Spherical polar Fourier assembly of protein complexes with arbitrary point group symmetry. *J Appl Crystallogr.* 2016;49(1):158–167.

[47] Grudinin S, Garkavenko M, Kazennov A. Pepsi-SAXS: an adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles. *Acta Crystallogr D Struct Biol.* 2017;73(Pt 5):449–464.

[48] Li J, Cao R, Cheng J. A large-scale conformation sampling and evaluation server for protein tertiary structure prediction and its assessment in CASP11. *BMC Bioinformatics.* 2015;16(1):337

[49] Kahraman A, Herzog F, Leitner A, Rosenberger G, Aebersold R, Malmström L. Cross-link guided molecular modeling with ROSETTA. *PLoS One.* 2013;8(9):e73411

[50] Herzog F, Kahraman A, Boehringer D, et al. Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. *Science.* 2012;337(6100):1348–1352.

[51] Kahraman A, Malmstrom L, Aebersold R. Xwalk: computing and visualizing distances in cross-linking experiments. *Bioinformatics.* 2011; 27(15):2163–2164.

[52] Ferber M, Kosinski J, Ori A, et al. Automated structure modeling of large protein assemblies using crosslinks as distance restraints. *Nat Methods.* 2016;13(6):515.

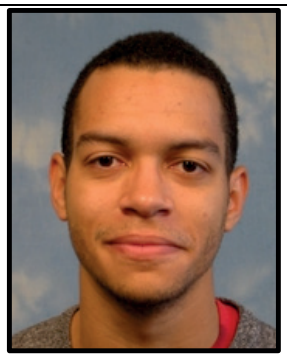
[53] Shi Y, Pellarin R, Fridy PC, et al. A strategy for dissecting the architectures of native macromolecular assemblies. *Nat Methods.* 2015; 12(12):1135–1138.

## SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

**How to cite this article:** Tamò GE, Abriata LA, Fonti G, Dal Peraro M. Assessment of data-assisted prediction by inclusion of crosslinking/mass-spectrometry and small angle X-ray scattering data in the 12th Critical Assessment of protein Structure Prediction experiment. *Proteins.* 2018;86:215–227. https://doi.org/10.1002/prot.25442

# Giorgio Elikem Tamò



Nationality: Swiss – DOB: 20.06.1989
Route de Chavannes 5, 1007, Lausanne, Switzerland
Phone: +41(0)76 735 13 79, E-mail: giorgiotamo@gmail.com,
LinkedIn: https://www.linkedin.com/in/giorgio-tam%C3%B2-94220a82

**Summary:**

- 11.2013-04.2018: PhD position in the laboratory of Prof. Matteo Dal Peraro (EPFL), with a main interest in applying heuristic optimization, statistical and structural methods to solve biological problems. These resulted in several publications and collaborations.

- 12.2012-11.2013: Internship at (EPFL) where daily duties were to create and optimize software linked to the in-silico assembly of lipid membranes (http://lipidbuilder.epfl.ch/home). This resulted in a publication and usage of our tool by the community.

- 02.2012-11.2012: Internship at Bio-Prodict (Medium-size company) where daily duties were to extract proteomics related data for protein engineering (https://www.bio-prodict.nl/), which still generates revenue, lead to collaborations with pharmaceutical companies and resulted in a publication.

**Relevant Education and qualification:**

- **Ecole Polytechnique Federale de Lausanne (EPFL)** **Lausanne, CH**
  Ph.D., Bio-Engineering, Thesis: 'In-silico Macromolecular Assembly' 2013-2018
- **Wageningen University** **Wageningen, NL**
  M.Sc., Bioinformatics 2011-2013
- **Kingston University** **London, GB**
  Bs.c., Marine and Freshwater biology 2007-2010

**Relevant programming skills:**

- Python, Javascript, MySQL, C++, PHP, SAS, Perl, Graphlab, Github

**Volunteering experience/extracurricular activities:**

- 06.2008 to 07.2008: Volunteered in Grenada (Caribbean islands) as a research assistant and worked on 'the leatherback sea turtle conservation project'
- 02.2015-02.2017: Treasurer at the ISCB-RSG (Bioinformatics) EPFL student association

**Languages:**

- French: mother tongue
- Italian: fluent
- English: fluent
- German: basic

**Prizes and hobbies:**

- Awarded overall best student (Honors) during Bachelor studies
- Amateur triathlete and travelling passionate

**Publications in peer-reviewed scientific journals:**

- Bovigny*, C., **Tamò*, G.**, Lemmin, T., Maïno, N. & Dal Peraro, M. 2015. LipidBuilder: A Framework To Build Realistic Models for Biological Membranes. *J. Chem. Info. Modeling*. 55(12), 2491-9. DOI: 10.1021/acs.jcim.5b00501
- **Tamò*, G.**, Abriata*, L. & Dal Peraro, M. 2015. The importance of dynamics in integrative modeling of supramolecular assemblies. *Curr. Opin. Struct. Biol*. 31, 28-34. DOI: 10.1016/j.sbi.2015.02.018
- **Tamò*, G.**, Maesani*, A., Träger, S., Degiacomi, M.T., Floreano, D. & Dal Peraro, M. 2017, Disentangling constraints using viability evolution principles in integrative modeling of macromolecular assemblies. *Sci. Rep.*, 7 (1), 235.
- Bergh*, T., **Tamò*, G.**, Nobili, A., Tao, Y., Tan, T., Bornscheuer, U.T., Kuipers, R.K.P., Vroling, B., de Jong, R.M., Subramanian, K., Schaap, P.J., Desmet, T., Nidetzky, B., Vriend, G. & Joosten H.J. 2017, CorNet: Assigning Function to Networks of Co-Evolving Residues by Automated Literature Mining. *PloS One*, 12(5), 1-19.
- **Tamò, G.**, Abriata, L., Fonti G. & Dal Peraro, M. 2017, Assessment of data-assisted prediction by inclusion of crosslinking/mass-spectrometry and small angle X -ray scattering data in CASP12. *Proteins: struct. Funct. and Bioinfo.* DOI: 10.1002/prot.25442
- Abriata, L.A, Kinch, L.N., **Tamò, G.**, Monastyrskyy, B., Kryshtafovych, A. & Dal Peraro M. 2017, Definition and classification of evaluation unit s for tertiary structure prediction in CASP12 facilitated through semi-automated metrics. *Proteins: struct. Funct. and Bioinfo.* DOI: 10.1002/prot.25403
- Abriata, L.A, **Tamò, G.**, Monastyrskyy, B., Kryshtafovych, A. and Dal Peraro M. 2017, Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins: struct. Funct. and Bioinfo.* DOI: 10.1002/prot.25423