Chapter 11

Protein–Protein Modeling Using Cryo-EM Restraints

Mikael Trellet, Gydo van Zundert, and Alexandre M. J. J. Bonvin

Abstract

Recent improvements in cryo-electron microscopy (cryo-EM) in the past few years are now allowing to observe molecular complexes at atomic resolution. As a consequence, numerous structures derived from cryo-EM are now available in the Protein Data Bank. However, if for some complexes atomic resolution is reached, this is not true for all. This is also the case in cryo-electron tomography where the achievable resolution is still limited. Furthermore the resolution in a cryo-EM map is not a constant, with often outer regions being of lower resolution, possibly linked to conformational variability. Although those low- to medium-resolution EM maps (or regions thereof) cannot directly provide atomic structure of large molecular complexes, they provide valuable information to model the individual components and their assembly into them. Most approaches for this kind of modeling are performing rigid fitting of the individual components into the EM density map. While this would appear an obvious option, they ignore key aspects of molecular recognition, the energetics and flexibility of the interfaces. Moreover, this often restricts the modeling to a unique source of data, the EM density map.

In this chapter, we describe a protocol where an EM map is used as restraint in HADDOCK to guide the modeling process. In the first step, rigid-body fitting is performed with PowerFit in order to identify the most likely locations of the molecules into the map. These are then used as centroids to which distance restraints are defined from the center of mass of the components of the complex for the initial rigid-body docking. The EM density is then directly used as an additional restraint energy term, which can be combined with all the other types of data supported by HADDOCK. This protocol relies on the new version 2.4 of both the HADDOCK webserver and software. Preparation steps consisting of cropping the EM map and rigid-body fitting of the atomic structure are explained. Then, the EM-driven docking protocol using HADDOCK is illustrated.

Key words Biomolecular interactions, Information-driven docking, Cryo-EM data, Flexibility, HADDOCK, Molecular modeling

1 Introduction

To drive all essential functions of the cells, biomolecules interact with each other forming complexes of different scales and stabilities. Deciphering the three-dimensional (3D) structure of such

Electronic supplementary material: The online version of this chapter (https://doi.org/10.1007/978-1-0716-0270-6_11) contains supplementary material, which is available to authorized users.

Zoltán Gáspári (ed.), *Structural Bioinformatics: Methods and Protocols*, Methods in Molecular Biology, vol. 2112, https://doi.org/10.1007/978-1-0716-0270-6_11, © Springer Science+Business Media, LLC, part of Springer Nature 2020

molecular complexes provides insights into the molecular determinants of these interactions and opens the route to tuning them in order to prevent or promote functions linked, for example, to diseases. Several experimental techniques exist to solve the 3D structure of molecules. Depending on the flexibility, mobility, and environment of those proteins, some techniques will be more efficient than others. They might also picture the system at different resolutions. X-ray crystallography and NMR have been for a long time the sole providers of high-resolution atomic structures stored in the Protein Data Bank (PDB). However, the past few years have seen the rise in the number of high-resolution structures solved by cryo-electron microscopy (cryo-EM). Cryo-EM has undergone a revolution in terms of the achievable resolution, thanks to both technical (e.g., the direct electron detectors) and software advances [1, 2].

Despite those advances, there will still be plenty of cases where cryo-EM will not achieve atomistic resolution (also typically difficult to reach in cryo-electron tomography). The resolution within one large macromolecular complex is also not a constant, meaning that parts of the complexes, often on the periphery or the more flexible parts, might only be seen at lower resolution. In those cases, one has to rely to fitting structures or models of the components of a complex into the density. This can be done via different ways: Manual fitting using specialized tools [3, 4], exhaustive search and rigid-body fitting [5], or flexible fitting, using different strategies to account for the atomic structures flexibility [6]. Often this modeling does not take into account flexibility (or only to a limited degree) and usually ignores the energetics at the interface of the fitted components, with the result that the interfaces in those complexes often have a poor quality with many clashes.

We have previously published a protocol that makes use of cryo-EM densities in flexible docking based on our information-driven, integrative modeling platform HADDOCK [7]. In this chapter, we illustrate the use of cryo-EM data as restraints to drive the modeling of a protein–protein complex using the new HADDOCK2.4 web portal, which now supports such kind of data. The protocol illustrates various steps, from the preparation/cropping of the original cryo-EM map to rigid-body fitting into the cryo-EM density to extract centroids position and finally to the setup of HADDOCK-EM run using its web portal version.

2 Overview

This section describes the different steps and their background in order to perform a protein–protein docking run in HADDOCK using an EM density map as restraint.

HADDOCK makes use of a variety of restraints (often expressed in terms of ambiguous or unambiguous distance restraints) throughout the entire docking process to drive and score the complex formation. These restraints can be derived from various experimental information sources such as NMR chemical shift perturbations, hydrogen/deuterium exchange, chemical cross-linking detected by mass spectrometry, mutagenesis, etc. [8–11].

When using cryo-EM data, however, HADDOCK needs to first convert the information provided by the EM map into distance restraints in order to drive the molecules to their potential location. This can be done by extracting centroids from the EM map as described in [12]. The centroids are provided as 3D coordinates to HADDOCK and are automatically converted to unambiguous (or ambiguous in cases where circular symmetry is present or the identity between subunits is uncertain) distance restraints between the centroids and the center of mass of the subunits, as illustrated in Fig. 1. These restraints draw, during the initial rigid-body step of HADDOCK, the molecules toward their location within the EM map. Once the rigid complex is formed and oriented correctly in the density, the cryo-EM density-based restraint energy term in HADDOCK is applied, and the refinement protocol proceeds through the various steps of HADDOCK. For details, *see* Subheading 2.3 and the original HADDOCK-EM publication [7].

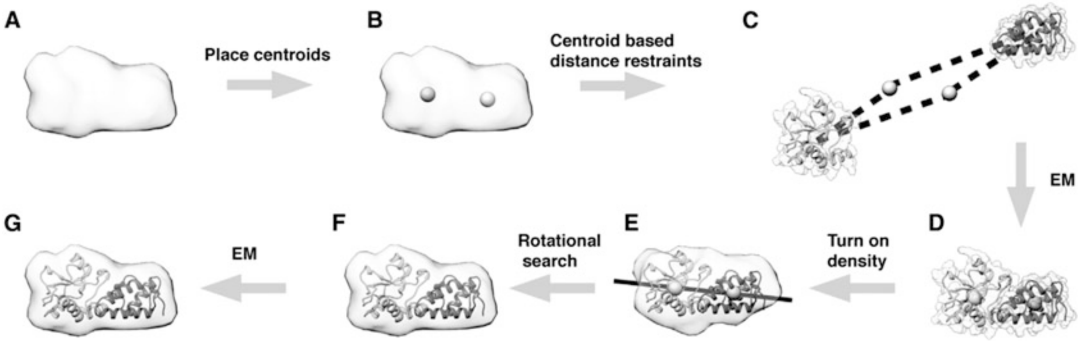


Fig. 1 Representation of the Rigid-Body Docking Protocol in HADDOCK-EM as illustrated in [7]. (**a**) Simulated cryo-EM data of colicin E7/IM7 complex (PDBid 7CEI). (**b**) Centers of mass of each subunit represented with gray spheres within the EM map. (**c**) Distance restraints in HADDOCK it0 step are defined between the COM of chain A (light gray) and B (dark gray) and their corresponding centroids. (**d**) Example of a complex obtained after the first rigid-body minimization (it0). (**e**) After the position, the relative orientation of each subunit should be determined. (**f**) A line drawn between the two centroids is used as axis to perform a rotational search. The complex with the highest cross-correlation value is chosen. (**g**) Excluding the centroid-based restraints, a final rigid-body minimization is performed against the cryo-EM data and assessed, thanks to a cross-correlation-based potential

2.1 High-Resolution Atomic Structure Rigid-Body Fitting into Cryo-EM Densities

The rigid-body fitting into the cryo-EM map will be performed using PowerFit [12], making use of our web server [13]. PowerFit fits atomic structures into density maps by performing a full-exhaustive six-dimensional cross-correlation search between the atomic structure and the density. It takes as input an atomic structure in PDB or mmCIF format and a cryo-EM density with its resolution, and outputs positions and rotations of the atomic structure corresponding to high correlation values and the top ten best scoring rigid poses. PowerFit uses the local cross-correlation function as its base score. The score is by default enhanced with an optional Laplace prefilter and a core-weighted version that minimizes the effect overlapping densities from neighboring subunits.

From the fitted structure, one can extract the 3D coordinates of the centroids (their center of mass position into the map), an information required by HADDOCK-EM.

2.2 Cryo-EM Density Map Cropping

In order to reduce data noise and save computational time, we strongly advise to crop the cryo-EM map to the region of interest. Cropping can be straightforwardly performed using UCSF Chimera [3]. A step-by-step protocol to extract a subregion of a density map is available at https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/midas/mask.html. In this protocol, we will use fitting results from PowerFit to crop the map with respect to the predicted molecular subunits' location.

2.3 Protein–Protein HADDOCKing with EM Restraints

2.3.1 Docking Protocol

The docking protocol in HADDOCK consists of three successive steps:

– *it0*: Rigid-body energy minimization (RBEM),
– *it1*: Semiflexible simulated annealing (SA) in torsion angle space (TAD/SA),
– *water*: Final restrained molecular dynamics in explicit solvent.

Pre- and post-processing steps are performed: (1) to build missing atoms in the preliminary step and (2) to launch a variety of analyses and clustering of solutions in the final step. For further details, please refer to [14, 15].

The HADDOCK-EM protocol requires as input an EM density map and its resolution together with the centroid coordinates of each of the subunits to be docked. Some changes have been made to the default HADDOCK docking protocol to account for the cryo-EM data parameters, mainly in it0, where centroids, approximate location of the subunits' COMs in the density map obtained during the fitting step (*see* Subheading 2.2), are used to place the subunits. As for the center of mass docking protocol of HADDOCK [16], additional distance restraints are generated between the COMs of the subunits. The main difference here lies in the fact that distance restraints are not created between the subunits themselves but between each subunit and one or several (in case of ambiguity) centroid coordinates.

Other cryo-EM-related required parameters for HADDOCK are either directly extracted from the map or have optimized default values. Some of these can be controlled through the web portal interface, for expert tuning of results.

Rigid-Body Energy Minimization (RBEM, it0)

In the initial docking stage, the interacting partners are considered rigid and separated in space and placed on a sphere centered on the midpoint of the centroids. For each docking trial, each subunit is randomly rotated around its center of mass and translated within a 10 Å box of to ensure unbiased starting configurations. In the case of unambiguous centroid-based restraints, HADDOCK will fit the subunits' COMs on the centroids to which they are associated. In the case of ambiguous restraints, each subunit would be ambiguously linked to any of the centroid given as input. Then, selection of the best conformation will solely rely on the HADDOCK score.

The centroid-based distance restraint is described by a soft square potential between two pseudo-atoms, one of which corresponds to the centroid and the other to the COM of the subunit.

Optimization steps have been performed to derive the best values for (1) the force constant of the centroid-based distance restraints that drives the COMs to the centroids, (2) the weight for the cross-correlation energy term, and (3) the weight of the LCC term in the HADDOCK score for *it0*. The default values in our protocol stand, respectively, at 50, 15,000, and 100. Those three values can be changed in the submission interface of HADDOCK2.4.

Binary systems will undergo a supplementary optimization step that aims at optimizing their orientation within the EM map. For this, an exhaustive 4° rotation search along the axis joining the centroids is performed, and at each step, the cross-correlation value is calculated to assess the pose. The orientation with the maximal cross-correlation value is kept. Finally, a rigid-body minimization is performed against the map using a combination of the cross-correlation, van der Waals, and electrostatic energy terms. Models are then scored by the traditional HADDOCK score plus a LCC term that reports on the overall quality of the fitting within the EM map. Typically, 2000 models are generated and scored from which typically the 400 models with the best HADDOCK score (*see* Subheading "Scoring") will go to the semiflexible simulated annealing stage of HADDOCK.

Semiflexible Simulated Annealing in Torsion Angle Space (TAD/SA, *it1*)

After a first rigid-body simulated annealing stage, the semiflexible simulated annealing stage, which starts with a short rigid-body molecular dynamics phase, optimizes the side chain conformations at the interface and then both backbone and side chains. The flexible regions are automatically defined for each docking model as the residues within 5 Å from a partner molecule. The parameters for *it1* are the same as in a typical docking run with HADDOCK, with the exception of adding the cross-correlation energy term used both during the simulated annealing protocol and in the scoring.

Restrained Molecular Dynamics in Explicit Solvent (Water)

The structures obtained after simulated annealing are finally refined in an explicit solvent layer to further improve the scoring. This is done by a short molecular dynamics simulation in water, solvating the complex in an 8 Å shell of TIP3P water molecules [17].

Scoring

The EM protocol introduces a new term to the HADDOCK score, namely, the local cross-correlation value (LCC) computed for a given model which is added to the equation defining the score, with an optimal weight for the three stages:

$$\mathrm{HS}_{\mathrm{EM}-it0} = 0.01 * \boldsymbol{E}_{\mathrm{vdw}} + 1.0 * \boldsymbol{E}_{\mathrm{elec}} + 0.01 * \boldsymbol{E}_{\mathrm{AIR}} + 1.0 * \boldsymbol{E}_{\mathrm{desolv}} - 0.01 * \mathrm{BSA} - 400 * \mathrm{LCC}.$$

$$\mathrm{HS}_{\mathrm{EM}-it1} = 1.0 * \boldsymbol{E}_{\mathrm{vdw}} + 1.0 * \boldsymbol{E}_{\mathrm{elec}} + 0.1 * \boldsymbol{E}_{\mathrm{AIR}} + 1.0 * \boldsymbol{E}_{\mathrm{desolv}} - 0.01 * \mathrm{BSA} - 10{,}000 * \mathrm{LCC}$$

$$\mathrm{HS}_{\mathrm{EM}-itw} = 1.0 * \boldsymbol{E}_{\mathrm{vdw}} + 0.2 * \boldsymbol{E}_{\mathrm{elec}} + 0.1 * \boldsymbol{E}_{\mathrm{AIR}} + 1.0 * \boldsymbol{E}_{\mathrm{desolv}} - 10{,}000 * \mathrm{LCC}$$

The other terms of the scoring function are the intermolecular van der Waals (E_{vdw}) and electrostatic (E_{elec}) energies calculated with the OPLS force field and an 8.5 Å nonbonded cutoff [18], an empirical desolvation potential (E_{desolv}) [19], the ambiguous interaction restraint energy (E_{AIR}), and the buried surface area (BSA).

2.3.2 Clustering of Final Solutions

All models generated by HADDOCK are clustered either based on their fraction of common contacts [20] (FCC, default) or on their interface-ligand-RMSD (i-l-RMSD) depending on the user's choice.

3 Methods

The HADDOCK-EM protocol requires some preliminary steps outside the traditional HADDOCK pipeline and independent from the web server. As explained in the previous sections, atomic structures will first be fit into an EM map region, then the EM map will be cropped, followed by a final fitting step.

To follow our protocol in its entirety, the 3D viewer program UCSF Chimera (https://www.cgl.ucsf.edu/chimera/) is needed. The protocol described here is based on version 1.12.0. Python2 or Python3 should also be installed (we recommend the latest stable versions Python 2.7.15 or Python 3.6). All other steps will simply make use of a standard web browser with JavaScript enabled. A registration to the CSB portal is required to use both PowerFit and HADDOCK (*see* **Note 1**). Complementary to the HADDOCK registration, users must request GURU access via their profile page to get access to the EM restraint parameters.

In the following sections, we illustrate our protocol on a test case taken from the use cases illustrated in [7]. The complex studied describes the interaction between two proteins of the 30S subunit of the ribosome (chains F and R). An atomic model of the entire complex is available (PDBid: 2YKR) as well as the 9.8 Å resolution cryo-EM map from which it was derived (EMDBid: 1884). The necessary files are provided in a tar archive available in the Supplementary Material. This protocol describes a two-body docking example. The same recipe can be extended to more components by repeating the PowerFit steps as many times as there are components and providing all the independent structures and centroid positions to the HADDOCK submission portal. The protocol should be able to run on any operating system since it mainly relies on a web browser, Chimera, and some Python scripts.

3.1 Preprocessing of the Cryo-EM Map

In this section, we will crop the cryo-EM map to only keep the part that is relevant for our docking. This step is optional and significantly depends on (1) the size of the map and (2) the preliminary information we have about the structure localization within this map. In our example, we already know the location of the subunits we want to dock in the EM map. Without this information, a very first step would have been to perform a fitting of the subunits within the EM map we plan to use to identify their possible location (as described in Subheading 3.2, for instance). Such fitting should always start from the largest components since these are easier to identify in the EM map.

To crop the map, we will use UCSF Chimera. Chimera has a very complete support for density maps and allows to quickly observe, analyze, and manipulate such maps via a customised user interface. We will follow the instructions given in the Chimera documentation [21] with little modifications accounting for most recent versions of Chimera. For an online version of the documentation, please refer to https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/midas/mask.html.

1. Open Chimera.
2. Load the cryo-EM map (`emd_1884.map`).
3. Load the crystal structure of the 30S subunit bound to RsgA (`2ykr_FR.pdb`).

   Warning: At this stage, be careful to not move the complex independently from the EM map during the session. This could lead to erroneous results during the next steps.
4. If the *Model Panel* window is not displayed, go to *Tools* > - *General Controls* > *Model Panel.*
5. Select "2ykr_FR.pdb."
6. Click on *Action* > *Surface* > *Show* to generate the surface of the protein.
7. A new line should appear in the *Model Panel* window with the name "MSMS main surface of 2ykr_FR.pdb."
8. Click on *Tools* > *General Controls* > *Command Line.*

9. In the new dialog window that opened at the bottom of the viewer main window type:

   mask #0 #1

   where #0 represents the identifier of your EM map and #1 the identifier of your protein–protein surface.

10. A new volume representation should appear (*see* Fig. 2) together with a new line in the *Model Panel* window named "emd_1884.map masked."

11. Save the new masked map, in the *Volume Viewer* window: *File* > *Save map as*... (`1884_masked.map`).

3.2 Getting Centroid Coordinates by Fitting the Atomic Structure into the New Cryo-EM Map

In this section, we will use the PowerFit web server to obtain the centroid coordinates of the two subunits of the complex. PowerFit performs an exhaustive search to identify the best fit of our crystal structure within the new masked cryo-EM map obtained in Subheading 3.1. The best solutions are ranked according to a cross-correlation score (similar to the one used in the HADDOCK protocol).

Note that access to the PowerFit web server requires registration (https://wenmr.science.uu.nl/auth/register/—select **PowerFit** as registered service).

1. Go to http://milou.science.uu.nl/services/POWERFIT.
2. Add the cryo-EM map file (`1884_masked.map`) to the "Cryo-EM map" field.
3. Add the atomic structure of the complex (`2ykr_F.pdb`) to the "Atomic structure" field.
4. Put `9.8` as "Map resolution" (in Angstroms).

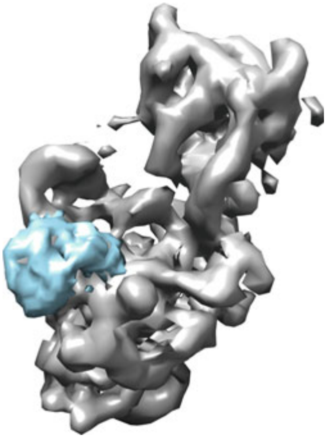


Fig. 2 Chimera snapshot illustrating the EM map of 30S ribosomal subunit with RsgA bound in the presence of GMPPNP, EMDBid 1884 [24], in white and, in blue, subpart of the EMDBid 1884 EM map masked by the subunits F and R of atomic structure 2ykr

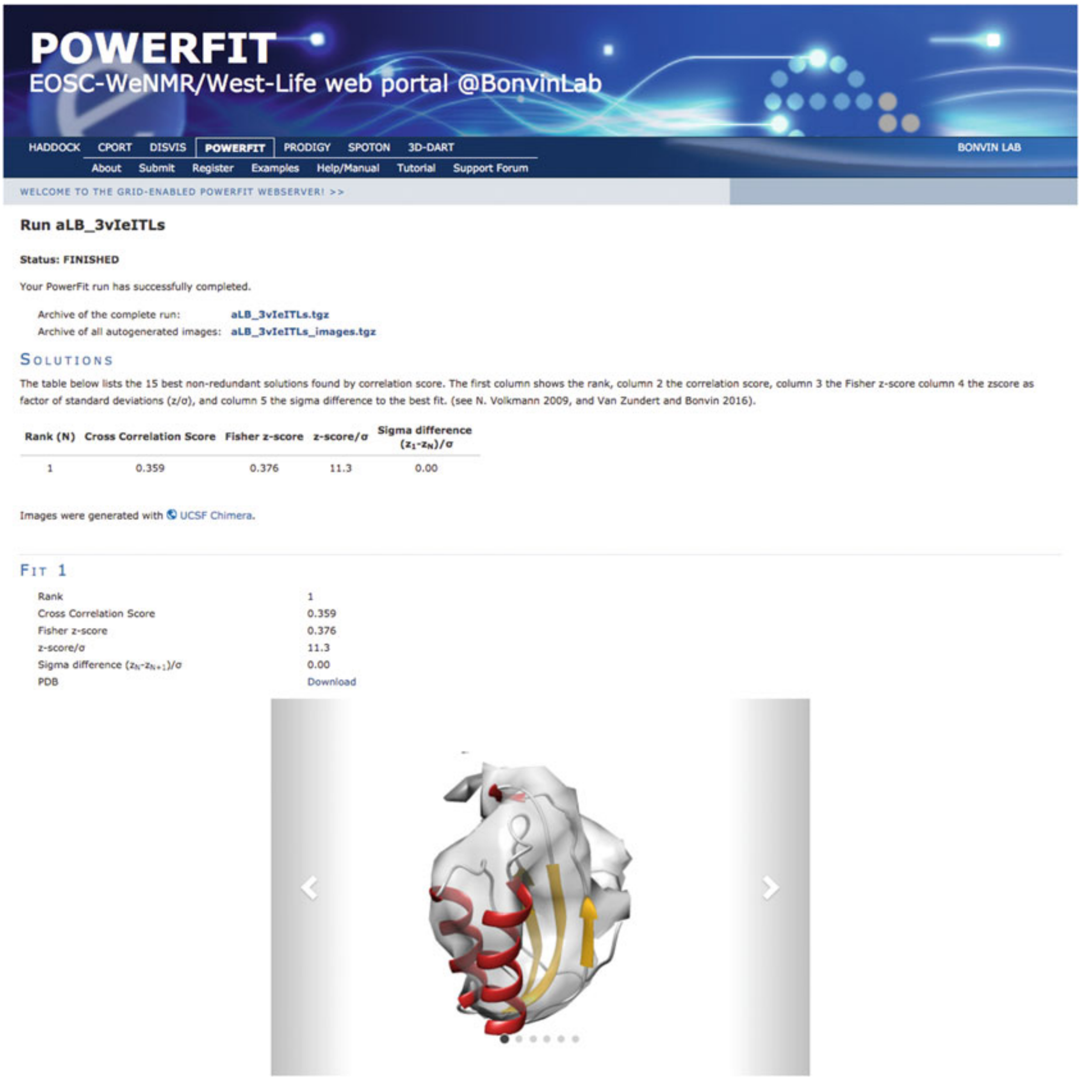
5. By default the server will redirect the computation to GPGPU grid resources provided by the federated sites of EGI. To run locally on our server, you might choose to uncheck "Redirect submission to grid (GPU) resources."
6. Enter your credentials (email + password) and click on "Submit."
7. The run should take about 5 min (*see* **Note 2**). The status of your job will be updated every 30 s. Once the job is finished, you will get an email, and, if you have left the page open, you will be redirected to the results page, similar to the one shown in Fig. 3.
8. On this page, reach the "Solutions" section. The table presented here reports the 15 best nonredundant solutions ranked by correlation score. We will focus on the best solution.
9. Click on the first link of the page corresponding to "Archive of the complete run." This will download the output of PowerFit under an archive file.
10. Untar the archive.
11. Redo **steps 1–10** by only changing the PDB file provided in **step 3**. But give this time the other protein, `2ykr_R.pdb`, as atomic structure input.
12. *Optional*: For each run (and then each set of output files extracted from the archives), open files `lcc.mrc` and `fit_1.pdb` with Chimera. Check that the atomic structure is well fitted within the density map file.
13. Using a terminal (or the windows command-prompt), run the python script `em_tools/centroid-from-structure.py` providing the best fit chains PDB files (`fit_1.pdb`) in each run archive as unique arguments.

```
> python centroid-from-structure.py fit_1.pdb
Parsed file: fit_1.pdb
Corresponding centroid (x, y, z):
11.90 -2.48 75.54
> python centroid-from-structure.py fit_1.pdb
Parsed file: fit_1.pdb
Corresponding centroid (x, y, z):
17.40 5.80 58.00
```

14. Save the centroid coordinates for later.

3.3 Preparation of Input Files

Each PDB provided to HADDOCK has to respect the PDB format with proper syntax and clear chain identifiers (*see* **Note 3**). The two input chains for the docking run are the chains F and R of 2YKR and are, respectively, provided in files `2ykr_F.pdb` and `2ykr_R.pdb`.



Run aLB_3vIeITLs

Status: FINISHED

Your PowerFit run has successfully completed.

Archive of the complete run:	aLB_3vIeITLs.tgz
Archive of all autogenerated images:	aLB_3vIeITLs_images.tgz

SOLUTIONS

The table below lists the 15 best non-redundant solutions found by correlation score. The first column shows the rank, column 2 the correlation score, column 3 the Fisher z-score column 4 the zscore as factor of standard deviations (z/σ), and column 5 the sigma difference to the best fit. (see N. Volkmann 2009, and Van Zundert and Bonvin 2016).

Rank (N)	Cross Correlation Score	Fisher z-score	z-score/σ	Sigma difference $(z_1-z_N)/\sigma$
1	0.359	0.376	11.3	0.00

Images were generated with UCSF Chimera.

FIT 1

Rank	1
Cross Correlation Score	0.359
Fisher z-score	0.376
z-score/σ	11.3
Sigma difference $(z_N-z_{N+1})/\sigma$	0.00
PDB	Download

Fig. 3 Screen capture of PowerFit results page after fitting of chain F of 2ykr in the masked map obtained from EMDBid 1884

The PDB file of the protein must be checked to avoid any double occupancies or residue insertion codes. If present, these can be removed by manual editing of the file or automatically by using the `pdb_delocc.py` script provided as part of the PDB-tools repository maintained by the HADDOCK team (https://github.com/haddocking/pdb-tools).

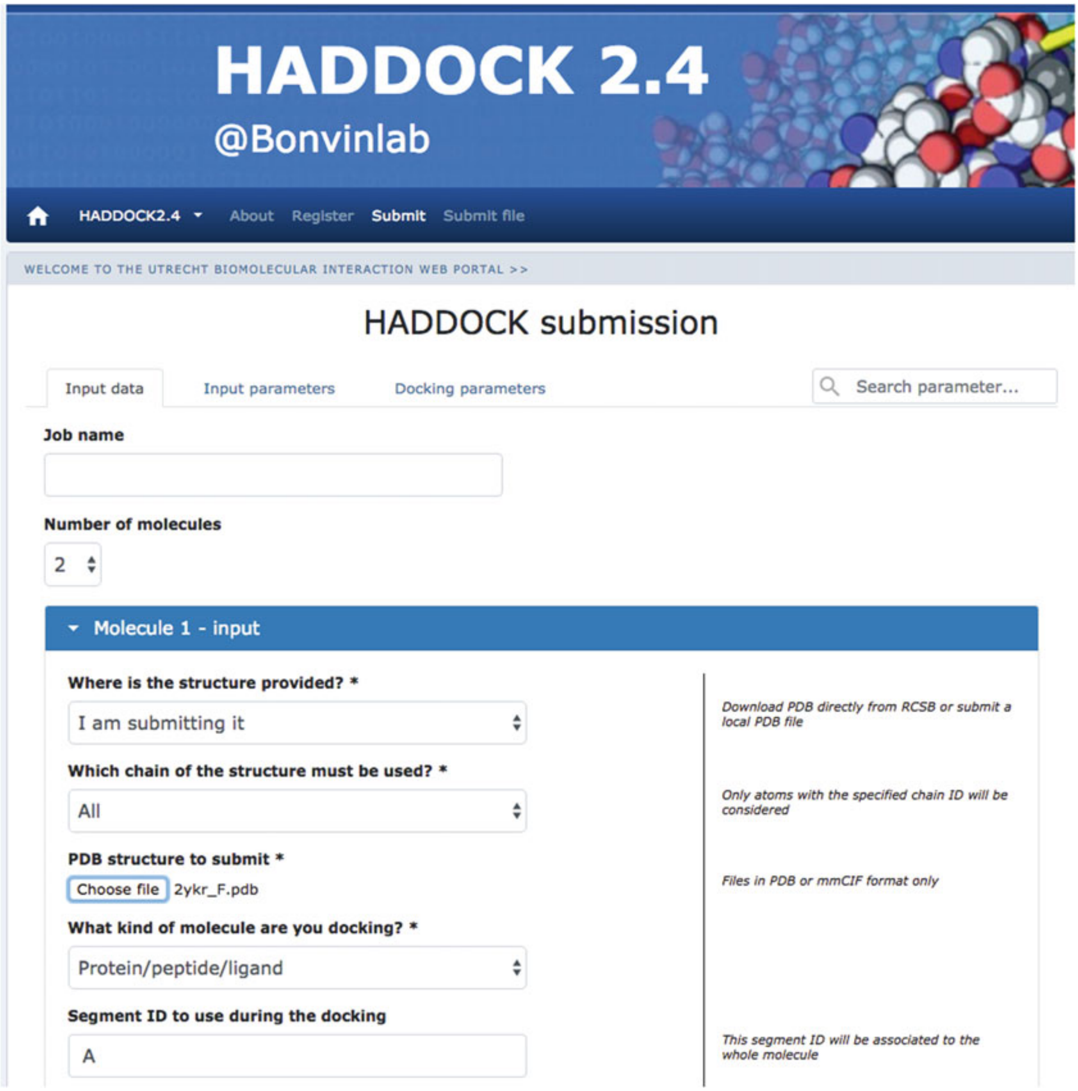
The EM map obtained after the previous step of cropping can be submitted as it is. The HADDOCK2.4 new web server processes and converts automatically any map under MRC or CCP4 format to XPLOR format, the latter being the only one read by CNS

(Crystallography and NMR System) [22], the computational engine used by HADDOCK.

3.4 Docking Two Subunits of the 30S Ribosome with the HADDOCK2.4 Web Server

For this docking, we will make use of the new HADDOCK web server available in its beta version at (https://wenmr.science.uu.nl/haddock2.4/). Registration is required to make use of the new interface and can be accessed through the corresponding submenu in the portal. Following the activation by the HADDOCK support team, users must request GURU access to be able to use EM restraints. This can be done in their own user profile page.

1. Open an Internet browser and go to https://wenmr.science.uu.nl/haddock2.4/. Click on the Submit subsection. You will find the page illustrated in Fig. 4.
2. We advise to give a name to your docking run. Be aware that no space or special characters other than "-" or "_" are allowed. We propose here to name the run "2ykr_em_modelling."
3. Define the number of molecules to dock (in this case, the default value of **2**).
4. There is no precise order for the molecules, and either of the PDB files can be provided first, but we do advise as a general rule to provide the largest component as first molecule (*see* **Note 4**). By default, we will use chain F as first molecule. In the section "First molecule," at the entry *"Where is the structure provided?"* Leave option `I am submitting it`. Leave "Which chain of the structure must be used?" to `All` (*see* **Note 3**). Next to "PDB structure to submit," press the `Choose file` button and move to the location where the tutorial data were unpacked. Go to the `pdbs/`directory and select the *2ykr_F.pdb* file. Keep both Nter and Cter to `False`.
5. In the section "Second molecule," at the entry "Where is the structure provided?" Leave option `I am submitting it`. Leave "Which chain of the structure must be used?" to `All` (*see* **Note 3**). Next to "PDB structure to submit," press the "Choose file" button and move to the location where the tutorial data were unpacked. Go to the `pdbs/`directory and select the `2ykr_R.pdb` file. Keep both Nter and Cter to False.
6. Click "Next" and wait for the second step interface to load (should not take more than a few seconds).
7. Leave the Molecule 1 and 2 parameters empty. Go to section "EM restraints (optional)" and unfold it as illustrated in Fig. 5.
8. Check `Use density/XREF restraints?` (set to True)
9. Next to "EM map," press the "Choose file" button and move to the location where the tutorial data were unpacked. Go to the `em_maps/` directory and select the `1884_masked.map` file (or select the one you generated at Subheading 3.2).



HADDOCK submission

Job name

Number of molecules: 2

Molecule 1 - input

Where is the structure provided? *: I am submitting it — Download PDB directly from RCSB or submit a local PDB file

Which chain of the structure must be used? *: All — Only atoms with the specified chain ID will be considered

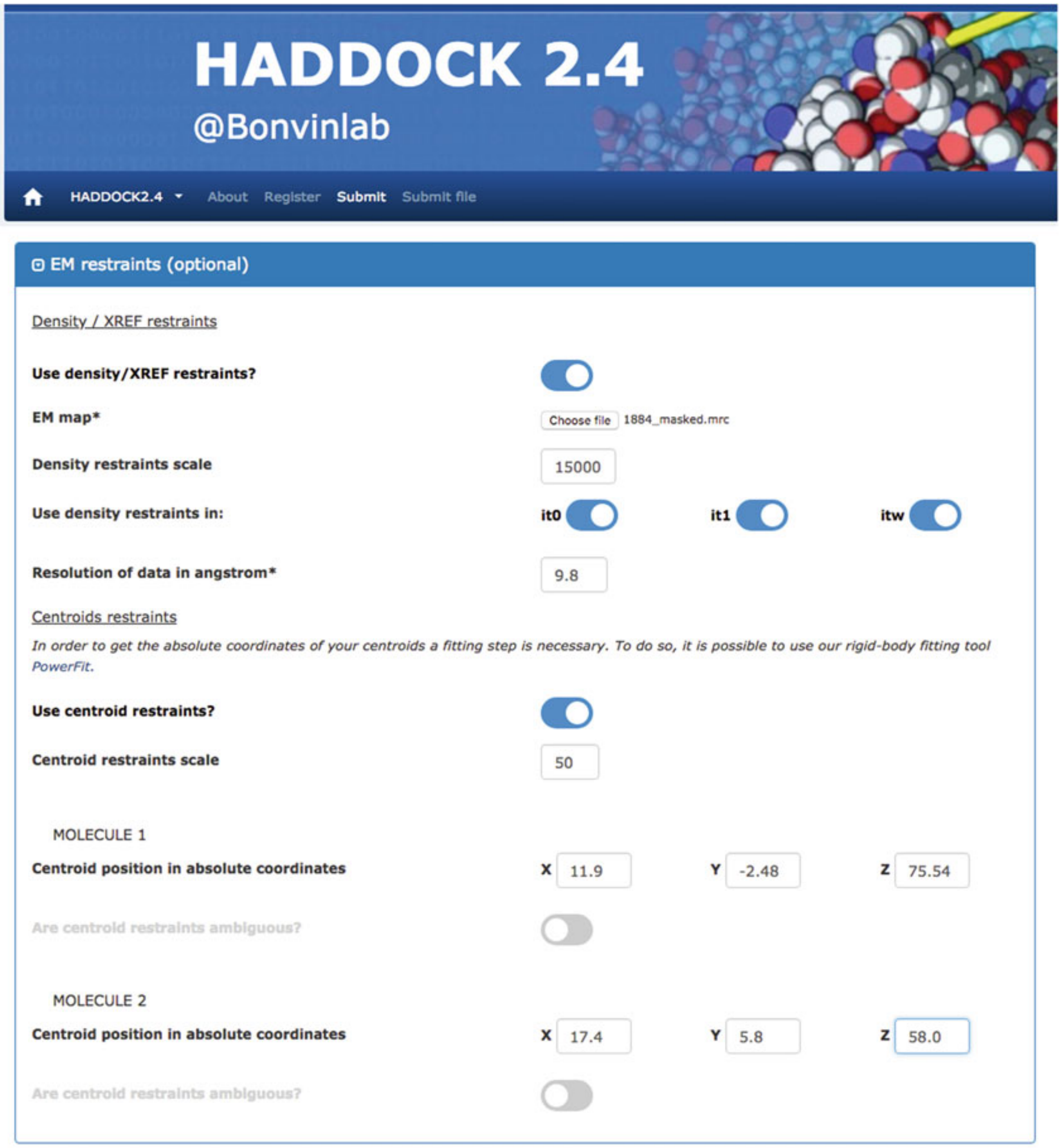
PDB structure to submit *: 2ykr_F.pdb — Files in PDB or mmCIF format only

What kind of molecule are you docking? *: Protein/peptide/ligand

Segment ID to use during the docking: A — This segment ID will be associated to the whole molecule

Fig. 4 Illustration of HADDOCK 2.4 submission page at the *Input data* first step

10. Set **9.8** in “Resolution of data in angstrom” field.
11. If this is not the case, check `Use centroid restraints?` (set to True).
12. In “MOLECULE 1 > Centroid position in absolute coordinates”, enter the coordinates you saved from Subheading 3.3 for chain A.
13. In “MOLECULE 2 > Centroid position in absolute coordinates”, enter the coordinates you saved from Subheading 3.3 for chain B.
14. Click `Next` and wait for the third step interface to load (should not take more than a few seconds).



⊟ EM restraints (optional)

Density / XREF restraints

Use density/XREF restraints?	●
EM map*	Choose file 1884_masked.mrc
Density restraints scale	15000
Use density restraints in:	it0 ● it1 ● itw ●
Resolution of data in angstrom*	9.8

Centroids restraints

In order to get the absolute coordinates of your centroids a fitting step is necessary. To do so, it is possible to use our rigid-body fitting tool PowerFit.

Use centroid restraints?	●
Centroid restraints scale	50

MOLECULE 1

	X	Y	Z
Centroid position in absolute coordinates	11.9	-2.48	75.54
Are centroid restraints ambiguous?	○		

MOLECULE 2

	X	Y	Z
Centroid position in absolute coordinates	17.4	5.8	58.0
Are centroid restraints ambiguous?	○		

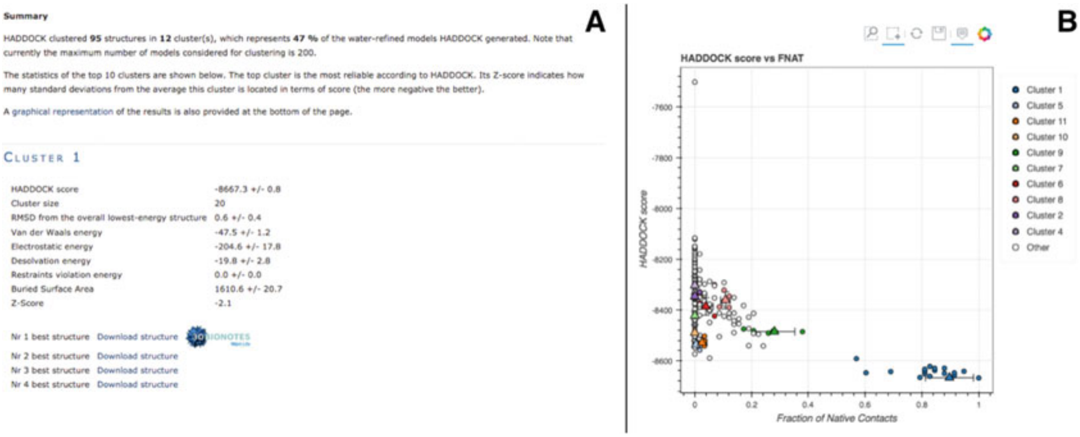
Fig. 5 Illustration of HADDOCK 2.4 submission page at the *Input parameters* second step

15. Leave default parameters and click `Submit` at the bottom of the page.
16. After a few seconds, you will be redirected to a page reporting the status of your job, a short summary of the docking input, and a progression report. This page will be updated every 30 s to report the progression of your job.
17. Within typically a few hours, depending on the web server load, you will receive another email reporting the final status of your job. If successful, a result page will be available at the link given

in the email, or, if you left the status page open, the page will be automatically loaded with a results summary. On this page, you will find the name of your docking run as well as a link to download it as a gzipped tar file. A link to the unique file containing input data and parameters is again provided.

18. The results page also indicates the number of clusters created by HADDOCK and how many structures coming from the *water* steps have been clustered. In our example, **12** clusters are created, gathering **47%** of the top 200 models. For an easier visualization of the results, only the ten best clusters based on the average HADDOCK score of its top four models are displayed in the summary page. You can find information and analyses of the last cluster in the gzipped tar file. For each cluster, information relative to the HADDOCK score of the top four models, the cluster size, and different statistics and energy values are reported as we can see for cluster 1 in Fig. 6a (*see* **Note 5**).

19. At last, an interactive representation of different CAPRI assessment criteria with respect to the HADDOCK score is provided for the ten best clusters in the "Results analysis" section. The first three plots show the HADDOCK score versus the fraction of common contacts (FCC—*see* **Note 6**), the i-RMSD, and the l-RMSD calculated using the top-ranked model as reference, respectively (*see* **Note 7**). The last three plots show the van der Waals, electrostatics, and AIR energy versus i-RMSD. An example of one of the plots is shown in Fig. 6b. One can note that the Eair values are all equal to 0 because no other



Summary

HADDOCK clustered **95** structures in **12** cluster(s), which represents **47 %** of the water-refined models HADDOCK generated. Note that currently the maximum number of models considered for clustering is 200.

The statistics of the top 10 clusters are shown below. The top cluster is the most reliable according to HADDOCK. Its Z-score indicates how many standard deviations from the average this cluster is located in terms of score (the more negative the better).

A graphical representation of the results is also provided at the bottom of the page.

CLUSTER 1

HADDOCK score	-8667.3 +/- 0.8
Cluster size	20
RMSD from the overall lowest-energy structure	0.6 +/- 0.4
Van der Waals energy	-47.5 +/- 1.2
Electrostatic energy	-204.6 +/- 17.8
Desolvation energy	-19.8 +/- 2.8
Restraints violation energy	0.0 +/- 0.0
Buried Surface Area	1610.6 +/- 20.7
Z-Score	-2.1

Nr 1 best structure Download structure
Nr 2 best structure Download structure
Nr 3 best structure Download structure
Nr 4 best structure Download structure

Fig. 6 Illustration of HADDOCK 2.4 results page after docking subunits F and R from 2ykr using as sole restraints the EM map information. (**a**) Extract of the cluster analysis for cluster 1. (**b**) Snapshot of one of the interactive plots provided in the Model Analysis section. In this plot, the HADDOCK score is plotted against the Fraction of Native Contacts

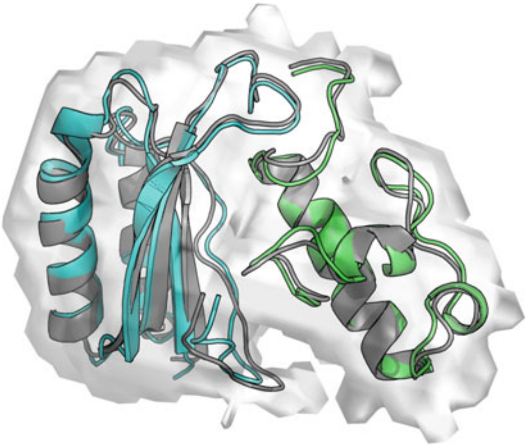


Fig. 7 Comparison of the best scoring models generated by HADDOCK, in blue (chain F) and green (chain R), and the reference structure (PDBid 2ykr) in dark grey. The EM map used to fit the two subunits and drive the docking run is shown as a transparent surface

restraints than the EM map-derived ones have been used to drive this docking.

20. It is possible to manually compare a reference structure with the best models of each cluster generated by HADDOCK. The 3D structures of these models can be directly downloaded from the results page. They are also located in the root of the docking run you downloaded as a gzipped tar file. Their name follows the following syntax: `cluster2_1.pdb`. This file is, for instance, the best model according to its HADDOCK score in the second cluster given by HADDOCK. The clusters are reported on the result page in the order of their HADDOCK score (from best to worst) (*see* **Note 8**).

    You can use fitting software such as ProFit [23] to get precise values of RMSD. PyMol is also useful since it has its own fitting algorithm and will give you a RMSD value as well as a visual feedback of the differences between the clustered models and the reference structure. Keep in mind that your reference structure has to be formatted in the same way that the PDB models generated by HADDOCK. ProFit considers only structures with an identical number of atoms. A superposition between the best HADDOCK model and the reference structure is reported in Fig. 7.

4 Notes

1. Registration to the CSB portal is mandatory to make use of both PowerFit and HADDOCK and can be done following

this link: https://wenmr.science.uu.nl/register. Once the registration has been done, check your mailbox for a confirmation link and click on this link (or copy/paste it in your web browser) to give us the possibility to activate your account. PowerFit and HADDOCK will only work for users logged in with a validated account.

2. The computational time of PowerFit scales almost linearly with the size of the system. However, using GPU resources allows to keep a PowerFit run with default parameters under 30 min for the largest systems. A hard limit of 200 MB for the size of the files that can be uploaded on the server prevents too large systems to be considered without previous trimming and/or cleaning.
3. Defining the largest molecule as first molecule for docking can be important for the final clustering because, in case of RMSD clustering, the structures are first fitted on the interface residues of the first molecule and then the RMSD is calculated on the interface residues of the second molecule. The interface residues are defined from an analysis of contacts in the generated models (at it1 and water, respectively). Defining the largest molecule first should thus result in a better fitting and clustering. However, one should note that the default clustering method is FCC and the order of the molecules does not impact the FCC calculation algorithm.
4. The PDB files provided to HADDOCK have to be correctly formatted to avoid any issues during the simulation process. There should be no overlap in residue numbering between different chains of a PDB. One can check the proper format of its PDB file using the pdb_format.py script provided as part of the PDB-tools repository maintained by the HADDOCK team (https://github.com/haddocking/pdb-tools). Missing atoms in the PDB files are not problematic since HADDOCK will rebuild them automatically.
5. The *Z*-score indicates how many standard deviations from the average a cluster is located in terms of its HADDOCK score. So the more negative, the better.
6. The FCC stands for fraction of common contacts and is calculated by comparing the lists of contacts at the interface between the components of a complex for two different structures. A contact is defined when two residues from different chains of the complex are closer than 5 Å from each other. The FCC is calculated as the fraction of common residue pairs shared between the two structures.
7. All reported RMSDs are calculated with respect to the lowest scoring model (the best model according to the HADDOCK score). The i-l-RMSD, which is used for clustering, is calculated on the interface backbone atoms of all chains except the

first one after fitting on the backbone atom of the interface of the first molecule. The i-RMSD is calculated by fitting on the backbone atoms of all the residues involved in intermolecular contacts within a cutoff of 10 Å. The l-RMSD is obtained by first fitting on the backbone atoms of the first molecule and then calculating the RMSD on the backbone atoms of the remaining chains.

8. The naming of clusters in HADDOCK is linked to their size and not their score. This originates from the clustering software. By definition, the largest cluster is always called cluster1, followed by cluster2 and so on. The cluster size however does not correlate per se with the HADDOCK score. Refer to the result page (or open in a web browser the `index.html` file provided in the tar archive) to see the cluster order based on the HADDOCK score.

Acknowledgments

This work is supported by the European H2020 e-Infrastructure grants (West-Life grant no. 675858 and BioExcel grants no. 675728 and 823830).

References

1. Bai X, McMullan G, Scheres SH (2015) How cryo-EM is revolutionizing structural biology. Trends Biochem Sci 40:49–57
2. Kimanius D, Forsberg BO, Scheres SH et al (2016) Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. Elife 5. https://doi.org/10.7554/eLife.18722
3. Pettersen EF, Goddard TD, Huang CC et al (2004) UCSF Chimera – a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612
4. Baker TS, Johnson JE (1996) Low resolution meets high: towards a resolution continuum from cells to atoms. Curr Opin Struct Biol 6:585–594
5. Esquivel-Rodríguez J, Kihara D (2013) Computational methods for constructing protein structure models from 3D electron microscopy maps. J Struct Biol 184:93–102
6. McGreevy R, Teo I, Singharoy A et al (2016) Advances in the molecular dynamics flexible fitting method for cryo-EM modeling. Methods 100:50–60
7. van Zundert GCP, Melquiond ASJ, Bonvin AMJJ (2015) Integrative modeling of biomolecular complexes: HADDOCKing with cryo-electron microscopy data. Structure 23:949–960
8. van Dijk AD, Boelens R, Bonvin AMJJ (2005) Data-driven docking for the study of biomolecular complexes. FEBS J 272:293–312
9. Melquiond ASJ, Bonvin AMJJ (2010) Data-driven docking: using external information to spark the biomolecular rendez-vous. In: Zacharias M (ed) Protein-protein complexes. Imperial College Press, London, pp 182–208
10. Karaca E, Bonvin AMJJ (2013) Advances in integrative modeling of biomolecular complexes. Methods 59:372–381
11. Rodrigues JPGLM, Bonvin AMJJ (2014) Integrative computational modelling of protein interactions. FEBS J 281:1988–2003
12. van Zundert GCP, Bonvin AMJJ (2015) Fast and sensitive rigid-body fitting into cryo-EM density maps with PowerFit. AIMS Biophys 2:73–87
13. van Zundert GCP, Trellet M, Schaarschmidt J et al (2017) The DisVis and PowerFit web servers: explorative and integrative modeling of biomolecular complexes. J Mol Biol 429:399–407

14. Dominguez C, Boelens R, Bonvin AM (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 125:1731–1737
15. de Vries SJ, van Dijk AD, Krzeminski M et al (2007) HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. Proteins 69:726–733
16. de Vries SJ, Melquiond ASJ, Kastritis PL et al (2010) Strengths and weaknesses of data-driven docking in critical assessment of prediction of interactions. Proteins 78:3242–3249
17. Jorgensen WL, Chandrasekhar J, Madura JD et al (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935
18. Jorgensen WL, Tirado-Rives J (1988) The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. J Am Chem Soc 110:1657–1666
19. Fernández-Recio J, Totrov M, Abagyan R (2004) Identification of protein–protein interaction sites from docking energy landscapes. J Mol Biol 335:843–865
20. Rodrigues JPGLM, Trellet M, Schmitz C et al (2012) Clustering biomolecular complexes by residue contacts similarity. Proteins 80:1810–1817
21. Goddard TD, Huang CC, Ferrin TE (2005) Chimera documentation – subregions. https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/midas/mask.html
22. Brünger AT, Adams PD, Clore GM et al (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54:905–921
23. Martin ACR, Porter C (2010) ProFit. http://www.bioinf.org.uk/software/profit/
24. Guo Q, Yuan Y, Xu Y et al (2011) Structural basis for the function of a small GTPase RsgA on the 30S ribosomal subunit maturation revealed by cryoelectron microscopy. Proc Natl Acad Sci U S A 108:13100–13105