# M3: an integrative framework for structure determination of molecular machines

Ezgi Karaca[1,5], João P G L M Rodrigues[2,5], Andrea Graziadei[1], Alexandre M J J Bonvin[2] & Teresa Carlomagno[1,3,4]



**We present a broadly applicable, user-friendly protocol that incorporates sparse and hybrid experimental data to calculate quasi-atomic-resolution structures of molecular machines. The protocol uses the HADDOCK framework, accounts for extensive structural rearrangements both at the domain and atomic levels and accepts input from all structural and biochemical experiments whose data can be translated into interatomic distances and/or molecular shapes.**

Cellular functions rely on the concerted action of biomolecules that form so-called molecular machines. Classical structural biology approaches often prove inadequate to elucidate the structure–function relationships of these large and dynamic complexes. Integrative structural biology[1,2] approaches combine data from multiple techniques to compensate for the shortcomings of each individual technique. However, while the experiments provide structural information at atomic resolution, the data is usually incomplete and nonhomogenously distributed. Purpose-tailored structure calculation protocols have been developed to incorporate specific combinations of hybrid data[3–10]. Most of these protocols are based on the concept of data-driven docking[1,10], whereby the complex is built from the structures of its individual monomers or subunits, and these structures are determined using classical methods[11].

Here, we present a general structure determination protocol, Model Molecular Machines (M3), which employs the user-friendly HADDOCK framework[10,12] to assemble molecular machine structures from their building blocks under the guidance of hybrid data (**Fig. 1**). In this work, the term 'molecular machines' refers to high-molecular-weight assemblies, irrespective of their function. The M3 protocol is built on three pillars. First, it uses an all-atom representation of the building blocks, which allows structural rearrangements both at the domain and atomic levels, as well as a physics-based force field. Second, it handles all combinations of shape and distance restraints, irrespective of their nature. Third, M3 uses a statistical analysis for structure selection, which, at the same time, probes the adequateness of the data to drive the structure calculation toward specific regions of the conformational space. If the data are found to be inadequate, the protocol indicates the need for additional experimental information. Here, we evaluate the performance of M3 with five molecular machines of different composition and size using either simulated (three complexes) or experimental (two complexes) data.

## RESULTS

### The M3 framework

**Figure 1** shows the M3 workflow. The building blocks, from which M3 starts the docking, are domains, monomers or subcomplexes; and these blocks largely preserve their 3D structure upon complex formation. M3 uses complementary and orthogonal experimental information—i.e. interatomic distances and molecular shapes. Interatomic distances are typically measured by nuclear magnetic resonance (NMR) spectroscopy, cross-linking mass spectrometry (XL-MS), electron paramagnetic resonance (EPR) spectroscopy or Foerster resonance energy transfer (FRET)[1,11]. Molecular shape information is obtained from electron microscopy (EM) or small angle scattering (SAS)[13,14]. The M3 protocol as well as the benchmark cases that demonstrate the performance of the method are publicly available at https://github.com/ezgikaraca/ISD-files and as **Supplementary Software** and **Supplementary Data**.

The sampling consists of two steps. At the start, the building-block structures are randomly placed and rotated on the surface of a sphere to avoid any initial configuration bias. To this end, we developed a module that, unlike the standard HADDOCK protocol, can handle an unlimited number of individual components (see Online Methods). During the first global search step (**Fig. 1b**), the building blocks are pulled together as rigid bodies under the effect of the total energy function, $E_{total} = E_{ff} + w_{exp} \times E_{exp,}$ where $E_{ff}$ is the force-field term accounting for nonbonded interactions, $E_{exp}$ measures the agreement between experimental and back-calculated distance restraints and $w_{exp}$ is the weight of the restraint energy terms, as defined in HADDOCK[12]. Distance restraints obtained with different experimental techniques concern different parts of the complex; i.e., protein–protein NOEs define interprotein interfaces, protein–RNA cross-links define protein–RNA interfaces, etc. During the conformational search, to eliminate any bias toward the interfaces defined by the largest

[1]European Molecular Biology Laboratory, Structural and Computational Biology Unit, Heidelberg, Germany. [2]Bijvoet Center for Biomolecular Research, Faculty of Science, Utrecht University, Utrecht, the Netherlands. [3]Leibniz University Hannover, Centre for Biomolecular Drug Research, Hannover, Germany. [4]Helmholtz Centre for Infection Research, Group of Structural Chemistry, Braunschweig, Germany. [5]Present addresses: Izmir International Biomedicine and Genome Institute (iBG-izmir), Dokuz Eylül University Saglik Yerleskesi, Izmir, Turkey (E.K.) and Department of Structural Biology, Stanford University School of Medicine, Stanford, California, USA (J.P.G.L.M.R.). Correspondence should be addressed to T.C. (teresa.carlomagno@oci.uni-hannover.de).

RECEIVED 21 APRIL; ACCEPTED 5 JULY; PUBLISHED ONLINE 14 AUGUST 2017; DOI:10.1038/NMETH.4392

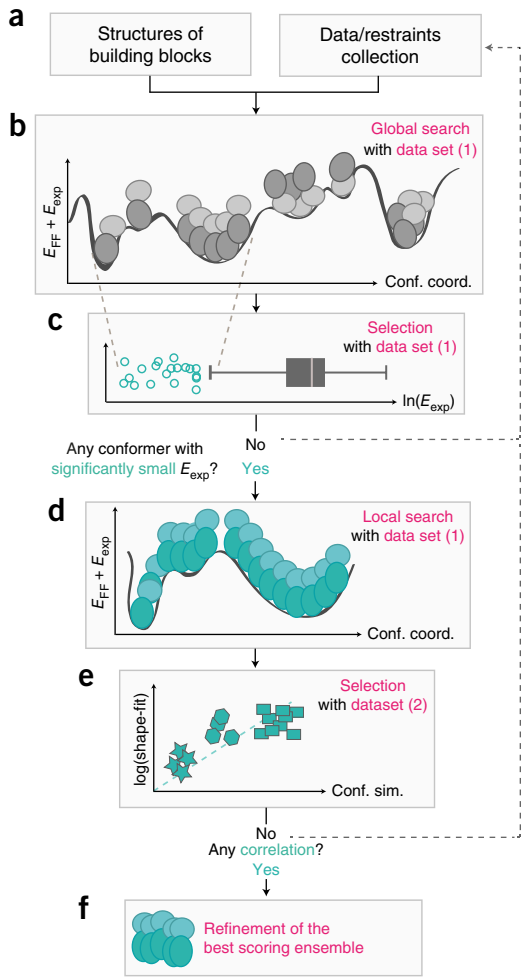


**Figure 1** | Workflow of the integrative structure determination protocol M3. (**a**) The protocol starts with preparation of the building blocks and conversion of the experimental data into structural restraints. (**b**) Experimental distance restraints (data set (1)) guide complex formation through the rigid-body docking step (global search). Conf. coord., conformation coordinates. (**c**) Complexes with significantly small restraint violation energy (green circles), ln($E_{exp}$), are isolated and passed on to the local search step (the gray vertical line corresponds to the median of the distribution). (**d**) A high-temperature SA (local search) protocol is used to explore the conformational space around the selected complexes. (**e**) The local search structures are clustered according to their conformational similarity (conf. sim.). The cluster with the best fitness to the shape data (data set (2)) is selected and (**f**) refined in explicit solvent.

set of restraints, we match the number of restraints in each set. Failure to balance the size of the restraints sets results in random conformational sampling of the interface defined by the set with the fewest restraints. The global search ends when no new structures are generated with significantly smaller $E_{exp}$ and different geometry (for all tested cases, ~1,000–5,000 structures, corresponding to an effective sampling of 10,000–50,000 conformers; see Online Methods).

The low-energy regions of the conformational space, to be explored during the second step, are selected differently from the standard HADDOCK protocol. For sparse experimental data, the energy function $E_{exp}$ follows a non-normal right-skewed distribution, where structures with significantly lower $E_{exp}$ are indistinguishable from the rest of the structures (**Supplementary Fig. 1**). Transformation of $E_{exp}$ into ln($E_{exp}$) generates a left-skewed distribution[15], whose tail contains structures with significantly low $E_{exp}$ that differ from the pool of random conformations. To identify these structures, M3 uses nonparametric box-and-whisker statistics. Here, 50% of the data around the median is defined as one interquartile range (IQR) and represented as a box. Whiskers can be extended from opposite sides of the box by multiple IQRs to cover the data spread. Any data falling outside the whisker expansion is classified as an outlier. We observe that, when the whiskers are extended by two IQRs, a small number of structures with significantly low $E_{exp}$ values emerge as outliers corresponding to the left tail of the ln($E_{exp}$) distribution. Conversely, absence of outliers indicates that the experimental information is insufficient to generate nonrandom structures.

After the global search selection, the outlier structures are subjected to the second sampling step. Here, a local search protocol generates ten structures per selected conformer using high-temperature simulated annealing in torsion-angle space applying the same form of $E_{total}$ as during global search (**Fig. 1d**). This step uses a modified version of the HADDOCK's simulated annealing protocol to allow extensive search of the conformational space around the selected conformers (see Online Methods). The structures resulting from the local search are separated into clusters according to their structural similarity and scored with respect to their agreement with molecular shape data measured by SAS or EM (**Fig. 1e**). While clustering, the structural similarity is measured by the orientational-r.m.s. deviation (o-r.m.s. deviation), which calculates the root mean square deviation between the translation and rotation vectors of each building block (see Online Methods). o-r.m.s. deviation ensures that the structure similarity is not dominated by the largest subunit, as it is the case for the coordinate r.m.s. deviation. We use the parameters $\chi$ and ccor (**Supplementary Note**) to evaluate the fitness (fit) of the conformers to SAS and EM data, respectively. If the structures of one cluster (or a subset thereof) distinguish themselves with low ln($\chi$) or high ln(ccor) from the rest of the population, these structures are returned as the final ensemble, after a short molecular dynamics simulation in explicit solvent (**Fig. 1f**). If the conformers with the best fitness belong to different clusters, their heterogeneity is analyzed to guide the acquisition of additional experimental data and resolve the ambiguity. Thus, throughout M3, structures are selected exclusively by experimental data; this is different from HADDOCK, whose score also includes force field and empirical energy terms.

Accounting for the conformational changes upon complex formation is a major challenge when modeling molecular machines. M3 addresses conformational changes at three levels. First, interdomain reorientations are addressed in the first and second sampling steps by 'breaking' the interdomain linkers and treating the domains as individual building blocks. The interdomain linkers are kept fully flexible during the local search, and their integrity is restored after the final structure selection. Second, small-to-medium structural rearrangements of side chains and loops at the interfaces between building blocks are addressed in the local search through an extensive high-temperature simulated annealing protocol. Third, large-scale conformational changes of long loops and linkers are addressed by representing the affected building block with an ensemble of conformations rather than with a single conformer.

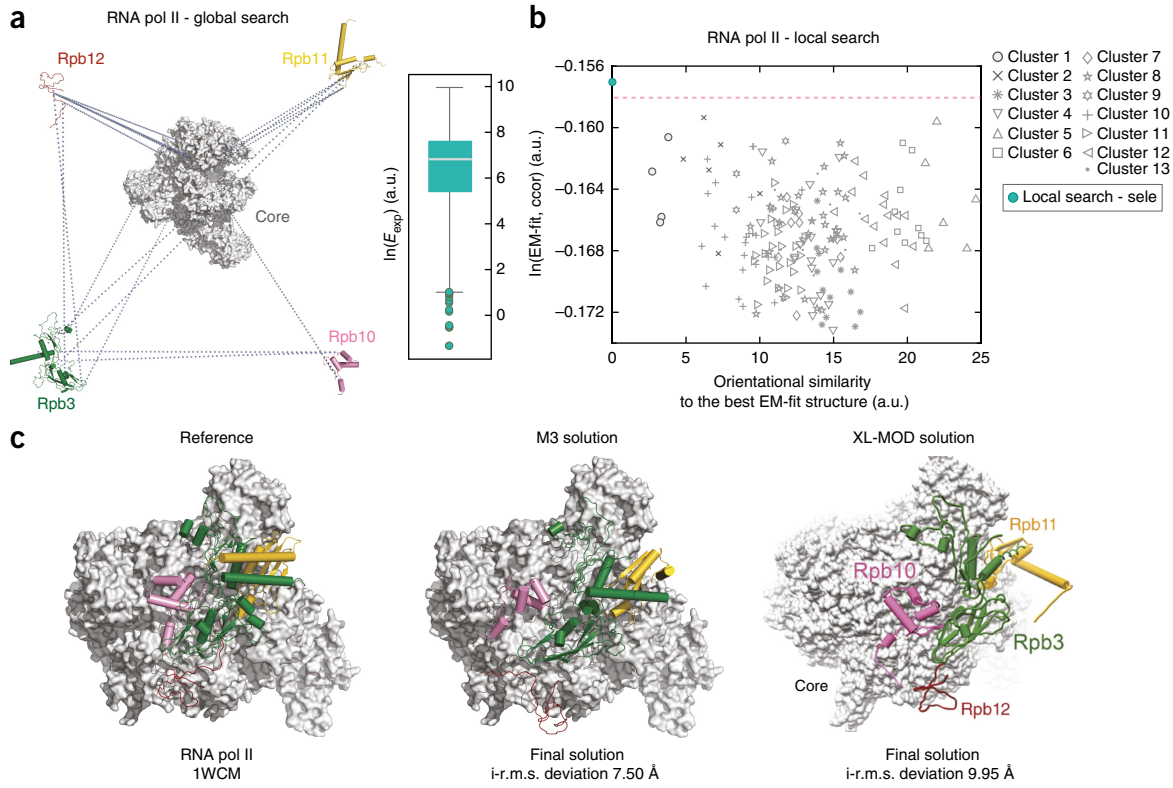


**Figure 2** | Application to the yeast RNA polymerase (pol) II demonstrates M3's ability to translate sparse data into a structural model. (**a**) Left, graphical representation of the separated building blocks (monomers Rpb3, Rpb10, Rpb11 and Rpb12 and the polymerase core) before structure calculations together with the 19 XL-MS-derived distance restraints used during the global search. Right, box-and-whiskers statistics identifies 18 low-energy structures among the 500 generated conformers. Center line, median; box limits, interquartile range; whiskers length, 2× interquartile range; points, outliers. (**b**) The 180 structures generated by the local search step were scored with respect to their agreement with the EM map EMDB ID 2784 (ref. 22) and grouped in 13 clusters. The structure with the highest ccor belongs to cluster 1 and is chosen as the final solution (sele, selected; a.u., arbitrary units). (**c**) The M3 solution (middle) is compared to the reference crystal structure (PDB ID: 1wcm; left) and a previously published model calculated by XL-MOD (reprinted from ref. 5). For comparison purposes, interface r.m.s. (i-r.m.s.) deviation was used to measure the accuracy of the M3 solution (see Online Methods).

We applied the M3 framework to four published high-resolution structures (PDB ID: 1k8k[16], 4wzj[17], 1i6h[18], 1wcm[19]) using either simulated (1k8k, 4wzj, 1i6h) or experimental data (1wcm). In addition, we calculated the structure of a molecular machine (PDB ID: 4by9 (ref. 7)) for which experimental data were collected in our laboratory.

### M3 validation with simulated data

We validated the M3 protocol using different kinds of simulated experimental data, including methyl-detected nuclear Overhauser effects (NOEs) and paramagnetic relaxation enhancement effects; UV zero cross-links between proteins and RNA; Lys–Lys interprotein cross-links; electron microscopy map and small angle scattering curves (**Supplementary Table 1** and **Supplementary Note**).

Using the example of the heptameric Arp2/3 protein complex (1k8k), we demonstrate the ability of M3 to assess whether the information content of the experimental data is sufficient to drive the structure calculation toward a well-defined minimum (**Supplementary Fig. 2** and **Supplementary Note**). Here, we performed parallel global search runs using random sets of inter-monomer NOEs comprising 50, 30 and 10 distances. The absence of outliers in the 10 NOEs run indicated that the input data were insufficient to define the 3D geometry of the complex.

Using the example of the U4 Sm proteins–RNA complex (4wzj), we tested the performance of NMR- and XL-MS-derived data to describe protein–protein interfaces (**Supplementary Fig. 3** and **Supplementary Note**). We found that the distance set derived by methyl-detected-paramagnetic relaxation enhancement effects (mPRE) measured by NMR performs superiorly to Lys–Lys distances measured by XL-MS on account of the higher completeness of the data. For both complexes (1k8k and 4wzj), M3 was able to find the native structure with an accuracy of <3 Å using either 30 precise (1k8k) or 136 loose (4wzj) distances.

With the yeast RNA polymerase II (1i6h), we probed the performance of M3 on a large and heterogeneous molecular machine composed of ten monomers (one RNA–DNA hybrid and nine proteins; **Supplementary Fig. 4**). The calculation was driven by 55 simulated distance restraints representing protein–protein (50) and protein–nucleic acid (5) XL-MS data. In addition, we simulated a 10-Å-resolution EM map. In the final M3 ensemble, all monomers, except for Rpb11, displayed the native orientation. Rpb11 was incorrectly placed on account of the uneven distribution of the XL-MS-derived distances that define its position (**Supplementary Figs. 4** and **5**). When excluding Rpb11, the final ensemble is only 4.8 ± 0.6 Å Cα/P-r.m.s. deviation away from the native structure 1i6h (total Cα/P-r.m.s. deviation, 7.7 ± 1.2 Å).

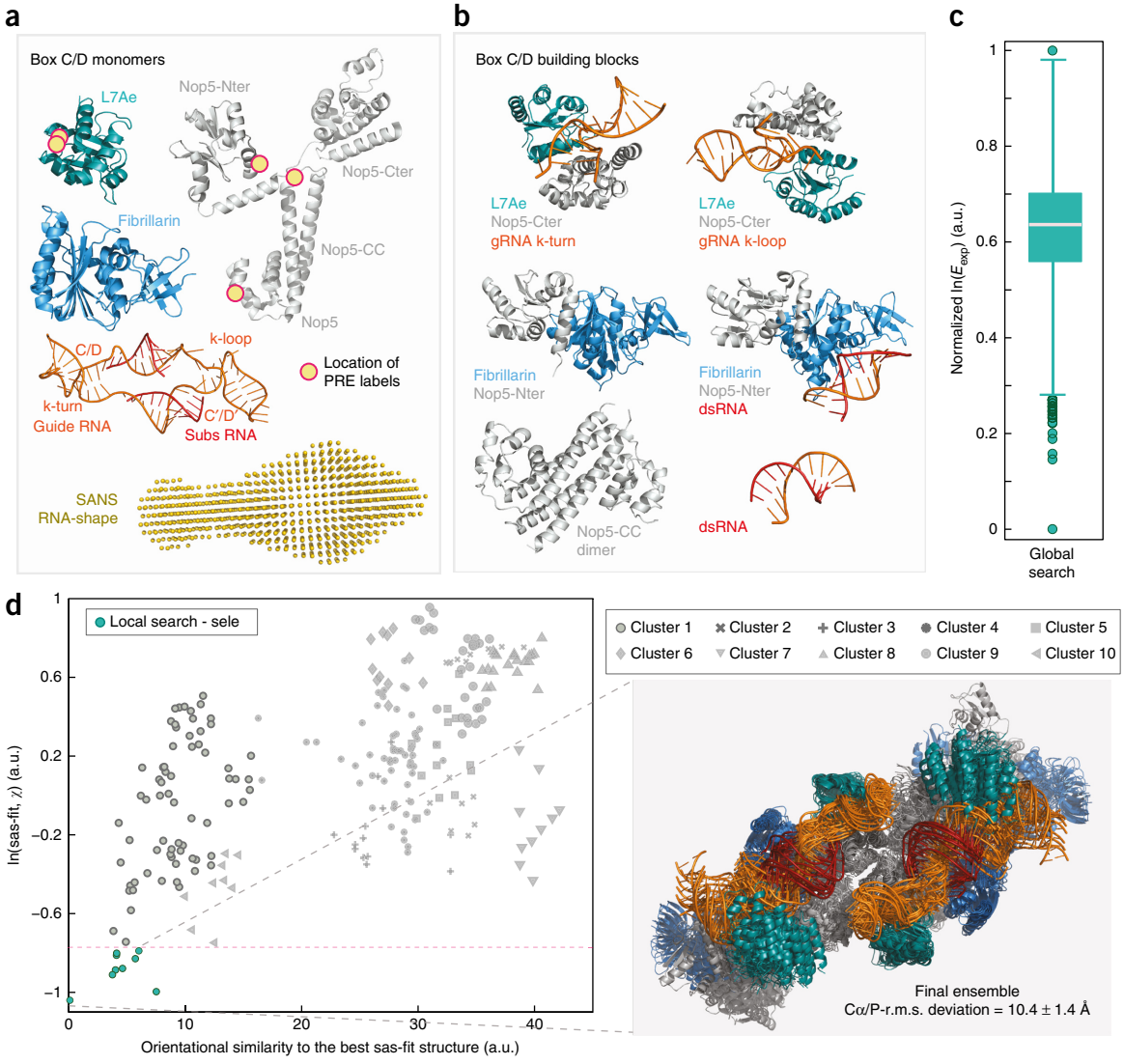


**Figure 3** | Structure determination of the Box C/D RNP underpins the robustness of the M3 protocol. (**a**) The Box C/D machinery is composed of four copies of three core proteins—L7Ae, fibrillarin and Nop5—and two copies of gRNA, which base pair with four substrate RNA molecules. The structural properties of Nop5 and gRNA, the location of the five PRE labels and the dummy-atom model of the RNA shape are shown (as in 4by9). (**b**) Substructures that enter the complex as rigid bodies are defined as individual building blocks (CC, coiled-coil; N/C-ter, N/C-terminal). (**c**) At the end of the global search, the ln($E_{exp}$) distribution, resulting from a combination of PRE, connectivity and SANS shape restraint violations, generated 27 outliers. (**d**) Cluster 1 contains nine structures with significantly better fitness to the SAS data (mean SAXS-$\chi$ = 1.9 ± 0.1; mean fibrillarin-SANS-$\chi$ = 2.6 ± 0.3; mean Nop5-SANS-$\chi$ = 6.8 ± 0.4)[7]. The precision of the ensemble is given in the figure; the similarity to 4by9 is 10.4 ± 1.4 Å (Cα/P-r.m.s. deviation).



**XL-MS- and EM-driven modeling of a yeast RNA polymerase II**

Next, the 12-subunit yeast RNA polymerase II (1wcm) was used to test the performance of M3 with 19 published experimental distances derived by XL-MS (**Supplementary Table 1** and **Supplementary Note**). This case was recently used to validate XL-MOD[5], thus it offered the opportunity to compare the performance of M3 with an existing state-of-the-art modeling approach. The available XL-MS data describe the position of four RNA polymerase II subunits—Rpb3, Rpb10, Rpb11 and Rpb12—with respect to the core complex[20] (**Fig. 2a**). Because of the limited number of degrees of freedom and experimental restraints, the energy surface could be sampled with only 500 structures (effectively 5000 conformers) during global search, and this resulted in 18 low-energy outliers. Extension of the conformational search to 1,000 structures did not generate any structure with better fit to the experimental data or significantly different geometry (**Supplementary Fig. 6a**,**b**). The 18 low-energy outliers were subjected to local search. In contrast to the U4 Sm proteins–RNA complex, during local search $E_{exp}$ increased upon refinement of the interaction interfaces (**Supplementary Fig. 6c**,**d**). This indicates that the physical forces and the distance restraints do not have a common minimum in the conformational space explored by the local search (**Supplementary Fig. 6e**,**f**) and suggests that the structures are not close to the native state of the complex[21]. Notwithstanding this warning, we proceeded to rank the 180 conformations obtained from local search using the shape information from the EM map of the yeast RNA polymerase II (EMDB ID: 2784)[22]. Despite all showing a good fit to the

EM data (ccorr between 0.84 and 0.85), the structures displayed substantially different orientations of the monomers with respect to the core, as detected by o-r.m.s. deviation (**Fig. 2b**). In a standard M3 workflow, the user would be warned that the data are not sufficient to determine the orientation of the four monomers unambiguously. The structural heterogeneity is caused by the ill-defined orientations of Rpb10 and Rpb11 as a result of either too few (Rpb10) or unevenly distributed (Rpb11) cross-link data[5] (**Fig. 2a**). As readily indicated by the comparison of $\mathbf{E}_{exp}$ in the global and local search steps (**Supplementary Fig. 4d**), the experimental distance restraints are not sufficient to drive the global search toward the native conformation. M3's best solution (ccor = 0.855, **Fig. 2b**) predicts the interfaces between Rpb3, Rpb10, Rpb11 and Rpb12 and the RNA polymerase II core with 7.50-Å accuracy. In particular, we found that the orientation of Rbp3 is much closer to the native structure than in the solution offered by XL-MOD (**Fig. 2c**).

### The structure of the Box C/D enzyme is determined by a combination of NMR-PRE and SAS data

Finally, we tested M3 with an experimental case from our laboratory describing the substrate-bound state of the Box C/D complex, an RNA methylation enzyme comprising of three proteins and a guide RNA (gRNA, **Fig. 3a**). One copy of gRNA binds two substrate RNAs and two copies of each protein. The architecture and the functionality of the complex suggest conformational dynamics hinging on the flexible regions of the Nop5 protein, separating this protein's three domains, and of the gRNA (**Fig. 3a**). The structure of the Box C/D enzyme in its substrate-bound form (4by9) was determined previously by a custom-tailored protocol in the ARIA framework using a combination of NMR and SAS data[7] (**Supplementary Table 1**).

The active Box C/D enzyme comprises two copies of gRNA and four copies of the core proteins (i.e., 16 proteins, two gRNAs and four substrate RNAs). We decreased the system complexity by grouping the monomers as subcomplexes that enter the Box C/D machinery as preassembled pseudorigid units. Both NMR and biochemical analysis identified these building blocks as (**Fig. 3b**) (i,ii) the gRNA k-turn/k-loop elements in complex with L7Ae and Nop5-C-terminal domain; (iii) the dimer of Nop5-coiled-coil domains; (iv) fibrillarin in complex with Nop5-N-terminal domain; (v) the substrate–gRNA duplex and the (vi) the substrate–gRNA duplex bound to the fibrillarin–Nop5-N-terminal domain dimer. The distinction between building blocks (v) and (vi) followed our previous NMR data indicating that only two of the four fibrillarin copies can bind the substrate–gRNA duplexes simultaneously[7]. Each building block is present twice in the complex, for a total of 12 subunits. Connectivity between the separated domains of either Nop5 or the gRNA was enforced through distance restraints[23]. 205 interprotein distances were measured by NMR-PRE experiments, (**Fig. 3a** and **Supplementary Table 1**). The RNA shape in the assembled complex was measured by contrast-matching small angle neutron scattering (SANS) and represented as molecular envelope, which was used during the conformational search to restrain the space explored by the RNA. To this end, highly ambiguous distance restraints were defined between RNA heavy atoms and the pseudoatoms representing the envelope (**Fig. 3b**; see Online Methods).

For the substrate-bound form of the Box C/D enzyme, 27 low-energy structures were selected from the initial global search (**Fig. 3c**). The 270 conformers resulting from local search were scored with respect to the consensus-$\chi$ score of three SAS curves—one SAXS curve, describing the shape of the entire complex, and two SANS curves acquired with contrast matching to report on the shape of fibrillarin and Nop5 in the assembled complex. The cluster in best agreement with the shape data (dark green circles, **Fig. 3d**) comprises nine structures and is consistent with the complex conformation in 4by9 (Cα/P-r.m.s. deviation, 10.4 ± 1.4 Å). This result confirms that M3 performs comparably well to the system-tailored protocol used to obtain the 4by9 structure[7].

## DISCUSSION

M3 is a versatile method for structure calculation of macromolecular complexes from hybrid data. It accepts both distance and shape information and can handle both proteins and proteins–nucleic acids complexes. With respect to other state-of-the-art methodologies[3–9], M3 has the advantage of permitting the incorporation of virtually any type of distance restraints in a straightforward manner, without requiring restraint-type-specific potentials. Shape information is typically applied after the local search step to select the native structures; however, when required, molecular shapes can be used directly during global search by defining ambiguous distances between the backbone atoms and the position of dummy atoms representing the shape (as described for the RNA shape in the Box C/D enzyme). Other protocols, such as XL-MOD[5], which address structural rearrangements using Bayesian analysis and restraints reweighting, have not been demonstrated with combinations of different types of distance or shape restraints.

In addition, M3 preserves the description of physical forces at the atomic level. This is critical when experimental information is scarce, as illustrated for the case of RNA polymerase II; here M3 performs better than XL-MOD using the same restraints set and is computationally less expensive due to its simpler algorithm.

Compared to other generic protocols, such as IMP[9], M3 is unique in the way it addresses structural rearrangements at the atomic level and reports on the adequacy of the input data. However, the whole-atom representation used in M3 impedes its application to very large assemblies such as the nuclear pore[24].

Owing to its ease of use, versatility and general applicability, we believe that M3 will be a useful tool for structure calculations with hybrid data.

## METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

**ACKNOWLEDGMENTS**

This work was supported by the EMBL, the EU FP7 ITN project RNPnet (contract number 289007) and the DFG grant CA294/3-2. E.K. acknowledges support from the Alexander von Humboldt Foundation through a Humboldt Research Fellowship for Postdoctoral Researchers. We thank J. Kirkpatrick for critical reading of the manuscript and B. Simon for discussion and support with CNS. A.M.J.J.B. acknowledges funding from the European H2020 e-Infrastructure grants West-Life (grant no. 675858) and BioExcel (grant no. 675728).

**AUTHOR CONTRIBUTIONS**

E.K. designed the studies, developed software, performed structure calculations, analyzed and interpreted data and wrote the manuscript, J.P.G.L.M.R. developed software; A.G. analyzed experiments; A.M.J.J.B. provided software and assisted in software development; T.C. designed the studies, assisted in data interpretation, wrote the manuscript and supervised the project.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.



1. Karaca, E. & Bonvin, A.M. Advances in integrative modeling of biomolecular complexes. *Methods* **59**, 372–381 (2013).
2. Ward, A.B., Sali, A. & Wilson, I.A. Biochemistry. Integrative structural biology. *Science* **339**, 913–915 (2013).
3. Morag, O., Sgourakis, N.G., Baker, D. & Goldbourt, A. The NMR-Rosetta capsid model of M13 bacteriophage reveals a quadrupled hydrophobic packing epitope. *Proc. Natl. Acad. Sci. USA* **112**, 971–976 (2015).
4. Duss, O., Yulikov, M., Jeschke, G. & Allain, F.H. EPR-aided approach for solution structure determination of large RNAs or protein–RNA complexes. *Nat. Commun.* **5**, 3669 (2014).
5. Ferber, M. *et al.* Automated structure modeling of large protein assemblies using crosslinks as distance restraints. *Nat. Methods* **13**, 515–520 (2016).
6. Kalinin, S. *et al.* A toolkit and benchmark study for FRET-restrained high-precision structural modeling. *Nat. Methods* **9**, 1218–1225 (2012).
7. Lapinaite, A. *et al.* The structure of the box C/D enzyme reveals regulation of RNA methylation. *Nature* **502**, 519–523 (2013).
8. Politis, A. *et al.* A mass spectrometry-based hybrid method for structural modeling of protein complexes. *Nat. Methods* **11**, 403–406 (2014).
9. Russel, D. *et al.* Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* **10**, e1001244 (2012).
10. van Zundert, G.C. *et al.* The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. *J. Mol. Biol.* **428**, 720–725 (2016).
11. Carlomagno, T. Present and future of NMR for RNA–protein complexes: a perspective of integrated structural biology. *J. Magn. Reson.* **241**, 126–136 (2014).
12. Dominguez, C., Boelens, R. & Bonvin, A.M. HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737 (2003).
13. Gabel, F. Small-angle neutron scattering for structural biology of protein-RNA complexes. *Methods Enzymol.* **558**, 391–415 (2015).
14. Madl, T., Gabel, F. & Sattler, M. NMR and small-angle scattering-based structural analysis of protein complexes in solution. *J. Struct. Biol.* **173**, 472–482 (2011).
15. Feng, C. *et al.* Log-transformation and its implications for data analysis. *Shanghai Arch. Psychiatry* **26**, 105–109 (2014).
16. Robinson, R.C. *et al.* Crystal structure of Arp2/3 complex. *Science* **294**, 1679–1684 (2001).
17. Leung, A.K., Nagai, K. & Li, J. Structure of the spliceosomal U4 snRNP core domain and its implication for snRNP biogenesis. *Nature* **473**, 536–539 (2011).
18. Gnatt, A.L., Cramer, P., Fu, J., Bushnell, D.A. & Kornberg, R.D. Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. *Science* **292**, 1876–1882 (2001).
19. Armache, K.J., Mitterweger, S., Meinhart, A. & Cramer, P. Structures of complete RNA polymerase II and its subcomplex, Rpb4/7. *J. Biol. Chem.* **280**, 7131–7134 (2005).
20. Chen, Z.A. *et al.* Architecture of the RNA polymerase II–TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J.* **29**, 717–726 (2010).
21. Raman, S. *et al.* NMR structure determination for larger proteins using backbone-only data. *Science* **327**, 1014–1018 (2010).
22. Plaschka, C. *et al.* Architecture of the RNA polymerase II–Mediator core initiation complex. *Nature* **518**, 376–380 (2015).
23. Karaca, E. & Bonvin, A.M. A multidomain flexible docking approach to deal with large conformational changes in the modeling of biomolecular complexes. *Structure* **19**, 555–565 (2011).
24. Alber, F. *et al.* The molecular architecture of the nuclear pore complex. *Nature* **450**, 695–701 (2007).

## ONLINE METHODS

**Complex structures used for benchmarking M3.** Asymmetric protein–protein heptamer; (PDB ID 1k8k[16]) crystal structure of Arp2/3 complex (225 kDa) from *Bos taurus*.

Asymmetric RNP octamer; (PDB ID 4wzj[17]) crystal structure of the spliceosomal U4 snRNP core domain (102 kDa) from *Homo sapiens*.

Asymmetric RNP decamer; (PDB ID 1i6h[18]) crystal structure of RNA polymerase II (477 kDa) from *Saccharomyces cerevisiae*.

Asymmetric RNP pentamer; (PDB ID 1wcm[19]) crystal structure of RNA polymerase II (509 kDa) from *Saccharomyces cerevisiae* (docking of four subunits to the polymerase core).

C2-symmetric RNP tetradecamer; (PDB ID 4by9 (ref. 7)) solution NMR structure of the Box C/D enzyme (386 kDa) from *Pyrococcus furiosus* in the holo form.

*Building blocks.* For the Arp2/3 protein complex, the building blocks were defined as the individual proteins; for the spliceosomal U4 snRNP, as individual proteins and the RNA monomer; for the RNA polymerase II, using synthetic data, as individual proteins and the central RNA–DNA hybrid; for the RNA polymerase II, using experimental data, as individual proteins and the RNA polymerase II core. In the absence of their free forms, the conformations of the building blocks were extracted from the PDB coordinates of the complexes. For the Box C/D complex, the identity of the building blocks is described in the main text; the PDB coordinates of 3 nmu[25] served as template structure. The interdomain flexibility was addressed by dividing both Nop5 and the gRNA into three and four separated, well-structured domains, respectively.

**Integrative structure determination protocol.** *Initial placement of the molecules.* Before global search, the building blocks were uniformly distributed on the surface of a sphere using the following protocol. First, each starting molecule was treated as a particle of unitary negative charge. Second, all particles were distributed on the surface of a sphere using the golden section spiral algorithm[26]. Finally, the placement of the particles was optimized by minimizing the Coulomb energy of the system via a steepest descent algorithm[27]. The radius of the final sphere was scaled so that the minimum distance between any particle is at least equal to the [maximum dimension of the largest molecule + 25 Å]. 25 Å corresponds to a value that is larger than 2.5 times the cutoff for nonbonded interactions—i.e., 8.5 Å[12]. After the initial placement, we followed the HADDOCK protocol, where each building block is randomly rotated and translated within a 10 Å sided cube. The U4 RNA from 4wzj and the molecular envelope of Box C/D gRNA were not translated.

*Global search (it0) parameters.* CNS 1.3 was used as the structure calculation engine[28]. During the global search step, the number of rotational/translational rigid-body minimization steps was increased from its original HADDOCK value of 250 to 1,000 to account for the high number of building blocks. Sampling of 180°-rotated molecules was disabled. After each repetition of ten rigid-body minimization steps ($n_{trials} = 10$), the lowest energy structure was written to the disk. 5,000 global search structures were saved, which effectively corresponds to sampling 50,000 conformers. For Box C/D, the number of rotational/translational rigid-body minimization steps was increased to 10,000; while the number of initial rotational minimization steps (before the rotational/translational minimization) was increased to 125. These numbers correspond to the smallest number of steps generating structures with a good fit to the connectivity restraints. The rest of the parameters were kept at their default HADDOCK values.

*Restraint energy-based scoring after global search.* At the end of the global search, the right-skewed distribution of the restraint violation energy $E_{exp}$ was transformed into a left-skewed distribution by taking $\ln(E_{exp})$[15]. Here, the structures with statistically significant low $\ln(E_{exp})$ values were identified as outliers in box-and-whisker plots using a whisker length of two times the IQR ($2 \times$ IQR). Box-and-whisker statistics were calculated with Matlab7.6 (ref. 29). The absence of outliers was interpreted as inadequacy of the experimental data to drive the structure calculation protocol toward nonrandom regions of the conformational space (**Supplementary Fig. 2**). We note that the $\ln(E_{exp})$ distribution of a structure-calculation run using a large, redundant number of distance restraints could also fail to show outliers in this statistical test when, for example, all conformers converge to the same structure. Evaluation of the structural similarity of the complex structures generated by the global search distinguishes the two scenarios—a high (low) level of structure similarity is the result of a converged (nonconverged) structure-calculation run that uses a very high (low) number of distance restraints. For large complexes, high numbers of redundant experimental restraints are usually not accessible.

To build the $\ln(E_{exp})$ distribution of the Box C/D enzyme, the violation energies of different restraint classes (PRE distances, shape, connectivity were individually normalized using a standard rank-preserving normalization function:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

The restraint energies were summed after normalization, so that each class makes a similar contribution to the total energy.

*Local search (it1) parameters.* The torsion-angle-dynamics simulated-annealing (TAD-SA) step of HADDOCK consists of three consecutive SA protocols[12]—first, the orientations of the rigid bodies are optimized by rigid-body molecular dynamics at high temperature, followed by cooling; second, the side chains of the interfacial regions are refined while slowly cooling the system; third, larger conformational rearrangements are allowed during refinement of both interfacial side chains and backbone in another round of slow cooling. To address the intricate conformational space of molecular machines and to allow for additional flexibility, for the TAD-SA steps of HADDOCK we used modified parameters similar to those employed in standard NMR structure-calculation protocols[7]. The temperature of the rigid-body TAD search was increased from 2,000 to 5,000 K; while the number of steps was increased from 500 to 20,000. Correspondingly, the rigid-body cooling was performed over 20,000 steps rather than 500. The factor time step of annealing was decreased from 8 to 4 to ensure a robust sampling. During the local search step for the Box C/D complex, the building blocks consisting of subcomplexes were kept together via contact restraints that preserved the intermolecular interfaces; in addition, nucleic acid conformations were restrained by applying HADDOCK nucleic-acid restraints (also applied for 4wzj)[30], and the positions of the dummy atoms representing the RNA SANS shape were fixed. Linkers between domains were treated as fully flexible.

*Selection of the final ensemble.* In this work, the orientational r.m.s. deviation (o-r.m.s. deviation) was used as the similarity measure

within clustering. The o-r.m.s. deviation measure was developed to adequately describe the similarity of subunit orientations in a complex even when the subunits significantly differ in size. This measure is based on a coarse-grained representation, where a vector defines each building block. The vector encodes the orientation and the translation of a given building block with respect to its imaginary copy placed at the origin and aligned along the principal axes system of the anisotropy tensor of the most anisotropic building block. The translation component of this vector has three dimensions (in $x$, $y$, $z$), while the orientation component has four dimensions (the rotation axis defined in $x$, $y$, $z$ and the degree of rotation around this axis). This leads to a seven-dimensional complex vector representing each building block, and a $7 \times N$ vector representing the assembled complex ($N$ being the number of building blocks). o-r.m.s. deviation is given in arbitrary units (a.u.).

Before o-r.m.s. deviation calculations, all solutions were fitted onto the most anisotropic building block of the system. The complex vectors were built via the coor orient function of CNS 1.3, while the similarity between each complex vector was calculated by the pdist function of Matlab7.6. Subsequently, the similarities were hierarchically linked with the ward method (Matlab7.6) and the linkages clustered with the k-means clustering (Matlab7.6). For the Box C/D complex, flexible monomers (two copies of fibrillarin and two copies of Nop5-N-terminal domain) were not included in the complex vector representation.

The fitness of each model ($\chi$ for SAXS/SANS and ccor for EM) with respect to shape data was calculated with Crysol (for SAXS), Cryson (for SANS) and Chimera (for EM)[31,32]. The scatter plot of ln(fit) versus o-r.m.s. deviation (from the structure with the best fitness) was used to determine the clustering cutoff, $k$—i.e., the number of clusters; we chose the minimum number of clusters that avoided overlap of members of different clusters. Finally, if the structures of one cluster (or a subset thereof) distinguished themselves with low ln($\chi$) or high ln(ccor) from the rest of the structures, then these structures were subjected to water refinement in HADDOCK and returned as the final solution.

In the final ensemble, we define precision as the average Cα- and P-r.m.s. deviation (Cα/P-r.m.s. deviation) to the structure with best fitness (or smallest restraint violation energy for the Arp2/3 complex) and accuracy as the average Cα/P-r.m.s. deviation to the reference, experimentally determined structure. To be able to compare the performance of M3 with that of XL-MOD, the accuracy for the RNA pol II structure, obtained from experimental data, is calculated as i-r.m.s. deviation. i-r.m.s. deviation defines the positional r.m.s. deviation of all interface residues (calculated for the Cα, N, C, and O atoms); a residue is defined to be at the interface if any of its heavy atoms is within 10 Å of any other atom of the interacting partners[33].

The M3 protocol is available at https://github.com/ezgikaraca/ISD-files, before integration in the HADDOCK server, together with all starting structures, restraint files and analysis scripts for the validation and test cases. The standard release of Haddock is available at http://www.bonvinlab.org/software/haddock2.2/installation/.

**Structural restraints.** *Synthetic restraints.* HADDOCK uses a soft-square flat-bottom $E_{exp}$ potential[34]. The target distance $d$ together with its lower and upper error bounds ($d_-$, $d_+$) defines the width of the flat bottom. In the calculations, synthetic NOE restraints, which describe interatomic distances of less than 6 Å, were imposed as $d$ with $d_- = d$ and $d_+ = 0.5$ Å. For synthetic PRE restraints, hypothetical labels were positioned on solvent-exposed residues. If necessary, *in silico* cysteine mutations were incorporated at the position of the label. ILV-PRE restraints were defined between the SG atom of the label and the CD1/CG1 atoms of Ile, Leu and Val residues. They were classified in three groups—short ($d < 15$ Å), medium ($15 \le d < 25$ Å), long ($d \ge 25$ Å) and imposed as $d = 15$ Å ($d_-$, $d_+ = 15$ Å, 0.5 Å), $d$ = PRE-measured-interatomic-distance ($d_-$, $d_+ = 0.5$ Å, 0.5 Å) and $d = 25$ Å ($d_-$, $d_+ = 0.5$ Å, 75 Å), respectively[35]. When randomly discarding a percentage of restraints, we retained at least one contact for each pair of building blocks. Protein cross-link restraints were generated for the disuccinimidylsuberate (DSS) cross-linker. Linked lysine-Cαs were allowed to be a maximum of 26 Å apart (11.4 Å for the extended DSS linker + the combined length of the two Lys side chains + 1 Å error) during the global and local search steps[36]. Zero-cross-link restraints defined the proximity between lysine NZ and pyrimidine P atoms with $d = 5$ Å ($d_-$, $d_+ = 5$ Å, 0.5 Å)[37]. Small-angle scattering curves were simulated with Crysol and Cryson from the ATSAS package[31,38]; while the EM map was simulated with Chimera[32]. Further details on the restraints are provided in **Supplementary Table 1**.

*Restraints for the spliceosomal U4 snRNP 4wzj.* Using 4wzj as a template, we generated 247 or 132 synthetic intermonomer PREs among ILV (isoleucine, leucine, valine) side chains of the Sm proteins (methyl groups are the only NMR-detectable moieties in a 100 kDa complex[39], **Supplementary Note**). 14 paramagnetic tags were introduced at the following residue positions: 61,78 (SmA); 52,69 (SmB); 77,99 (SmC); 46,70 (SmD); 48,70 (SmE); 50,71 (SmF); 46, 69 (SmG). *In silico* cysteine mutations were applied where necessary.

A second set of protein–protein distances was generated *in silico* from XL-MS. 39 intermonomer distance pairs were calculated by the xWalk program using standard settings[36]. Four distances were eliminated, as the putative cross linker crossed the RNA in the center of the complex. The final set of 35 contained distance information for the following monomer pairs: SmA–SmB, SmA–SmD, SmB–SmC, SmD–SmG, SmE–SmF.

The protein–RNA distances were defined by UV zero cross-links between RNA-uridines and histidine/phenylalanine residues of SmA and SmG, as measured by Urlaub *et al.*[37]. The number of protein–RNA XL-MS distances was multiplied to match the number of protein–protein distances, which would otherwise dominate the structure calculation. For molecular shape data, we used a synthetic SAXS curve predicted from 4wzj. Restraint statistics are given in **Supplementary Table 1**.

*Synthetic restraints for the RNA polymerase II 1i6h.* The synthetic data consisted of 15% of all possible distances between cross-linkable lysines[36,40], leading to 50 sparsely distributed protein–protein distances. For six of the protein pairs, the data set contained only one distance (**Suppplementary Fig. 4** and **Supplementary Table 1**). Protein–nucleic acid interactions were described by five zero cross-links between lysine residues of the two largest protein subunits (Rbp1/2) and pyrimidine bases of the RNA–DNA hybrid, which was treated as a single building block during the structure calculation. The number of nucleic acid–protein XL-MS distances was multiplied to match the number of protein–protein distances, which would otherwise dominate the structure calculation. For molecular shape data, we simulated an EM map at 10 Å resolution[32].





doi:10.1038/nmeth.4392

*Experimental restraints for the RNA polymerase II 1wcm.* 19 published experimental restraints derived from XL-MS defined the positions of the Rpb3, Rpb10, Rpb11 and Rpb12 monomers on the core of the yeast RNA pol II[20]. Crystallographic restraints were used to keep the core intact during the local search. The EM map, with EMDB id 2784, is of 6.6 Å resolution[22].

*Experimental restraints for the Box C/D enzyme 4by9.* PRE restraints had been previously measured[7,11]. In our structure calculations, we did not explicitly include the paramagnetic label 3-(2-iodoacetamido)-2,2,5,5,tetramethyl-1-pyrrolidinyloxy. Instead, the distance was defined between the SG atom of the corresponding cysteine and the CD1/CG1 atoms of the ILV residues (**Supplementary Table 1**). The presence and the flexibility of the paramagnetic tag were reflected in values of $d_-$, $d_+ = 6$ Å, which, when added to the experimental error of 2 Å, gave $d_-$, $d_+ = 8$ Å. The SANS curves of the holo Box C/D complex with [$^2$H]Nop5 and [$^2$H]Fibrillarin were measured in a 42%/58% $D_2O/H_2O$ buffer. DAMMIN was used to obtain a low-resolution molecular envelope of the gRNA within the Box C/D RNP from the [$^2$H]gRNA SANS curve[41]. The dummy-atom representation of this envelope was used to restrict the conformational space of the gRNA by defining ambiguous distance restraints (up to 4 Å) between each dummy atom and all P, C1 and C4 atoms of the RNA and vice versa. C2 symmetry restraints were applied between equivalent RNA molecules and Nop5 coiled-coil dimers, as suggested by the NMR data. The molecular integrity of Nop5 and the gRNA, whose domains were defined as independent building blocks, was enforced via connectivity restraints imposed between the artificially detached N-C and P-O3′ atoms, respectively. During the global and local search steps, we allowed detached N-C and P-O3′ atoms to be separated by a maximum distance of 5.0 and 1.3 Å, respectively. The number of connectivity restraints was scaled to match the total number of shape (ambiguous), PRE (unambiguous) and C2 symmetry restraints. For structure selection after the global search, the energy contribution of each distance restraint class (ILV-PRE-derived interprotein distances, SANS-derived RNA shape and interdomain connectivities) to $E_{exp}$ was normalized and summed (**Supplementary Fig. 7**).

**Code availability.** The M3 protocol is publicly available at https://github.com/ezgikaraca/ISD-files and as **Supplementary Software**. A user guide is available as a **Supplementary Protocol** and at the *Protocol Exchange*[42].

**Data availability statement.** The structural coordinates used in this study have the following accession codes: Protein Data Bank accessions 1k8k, 4wzj, 1i6h, 1wcm, 4by9; and The Electron Microscopy Data Bank accession 2784. Restraint files, starting structures and final models are available in **Supplementary Data**. A **Life Sciences Reporting Summary** is available.


25. Xue, S. *et al.* Structural basis for substrate placement by an archaeal box C/D ribonucleoprotein particle. *Mol. Cell* **39**, 939–949 (2010).
26. Saff, E.B. & Kuijlaars, A.B.J. Distributing many points on a sphere. *Math. Intell.* **19**, 5–11 (1997).
27. Rodrigues, J.P. *Computational Structural Biology of Macromolecular Interactions* (Ridderprint BV, 2014).
28. Brunger, A.T. Version 1.2 of the crystallography and NMR system. *Nat. Protoc.* **2**, 2728–2733 (2007).
29. MATLAB and Statistics Toolbox Release v. *R2008a* (Version 7.6) (Natick, 2008).
30. van Dijk, M. & Bonvin, A.M. Pushing the limits of what is achievable in protein–DNA docking: benchmarking HADDOCK's performance. *Nucleic Acids Res.* **38**, 5634–5647 (2010).
31. Petoukhov, M.V. *et al.* New developments in the ATSAS program package for small-angle scattering data analysis. *J. Appl. Cryst.* **45**, 342–350 (2012).
32. Pettersen, E.F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
33. Méndez, R., Leplae, R., De Maria, L. & Wodak, S.J. Assessment of blind predictions of protein–protein interactions: current status of docking methods. *Proteins* **52**, 51–67 (2003).
34. Nilges, M., Gronenborn, A.M., Brünger, A.T. & Clore, G.M. Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints. Application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. *Protein Eng.* **2**, 27–38 (1988).
35. Rosenzweig, R., Moradi, S., Zarrine-Afsar, A., Glover, J.R. & Kay, L.E. Unraveling the mechanism of protein disaggregation through a ClpB-DnaK interaction. *Science* **339**, 1080–1083 (2013).
36. Kahraman, A., Malmström, L. & Aebersold, R. Xwalk: computing and visualizing distances in cross-linking experiments. *Bioinformatics* **27**, 2163–2164 (2011).
37. Urlaub, H., Kühn-Hölsken, E. & Lührmann, R. Analyzing RNA-protein crosslinking sites in unlabeled ribonucleoprotein complexes by mass spectrometry. *Methods Mol. Biol.* **488**, 221–245 (2008).
38. Karaca, E. & Bonvin, A.M. On the usefulness of ion-mobility mass spectrometry and SAXS data in scoring docking decoys. *Acta Crystallogr. D Biol. Crystallogr.* **69**, 683–694 (2013).
39. Mund, M., Overbeck, J.H., Ullmann, J. & Sprangers, R. LEGO-NMR spectroscopy: a method to visualize individual subunits in large heteromeric complexes. *Angew. Chem. Int. Edn Engl.* **52**, 11401–11405 (2013).
40. Mühlbacher, W. *et al.* Conserved architecture of the core RNA polymerase II initiation complex. *Nat. Commun.* **5**, 4310 (2014).
41. Petoukhov, S.V. The system-resonance approach in modeling genetic structures. *Biosystems* **139**, 1–11 (2016).
42. Karaca, E. *et al.* M3: an integrative framework for structure determination of molecular machines. *Protocol Exchange* http://dx.doi.org/10.1038/protex.2017.093 (2017).