Macromolecular symmetric assembly prediction using swarm intelligence dynamic modeling

Matteo T. Degiacomi [1#] and Matteo Dal Peraro [1*]

[1] Institute of Bioengineering, School of Life Sciences,

Ecole Polytechnique Fédérale de Lausanne - EPFL, CH-1025 Lausanne

[#] Current address: Physical and Theoretical Chemistry Laboratory, South Parks Road, Oxford, OX1 3QZ

*Correspondence should be addressed to Dr. Matteo Dal Peraro

Institute of Bioengineering, School of Life Sciences,

Ecole Polytechnique Fédérale de Lausanne,

AAB 010 Station 19, CH-1015 Lausanne

E-mail: matteo.dalperaro@epfl.ch

SUMMARY

Proteins often assemble in multimeric complexes to perform a specific biological function. Trapping these high-order conformations is however difficult experimentally. Therefore, predicting how proteins assemble using *in silico* techniques can be of great help. The size of the associated conformational space and the fact that proteins are intrinsically flexible structures make, however, this optimization problem extremely challenging. Nonetheless, known experimental spatial restraints can guide the search process, contributing to model biologically relevant states. We present here a swarm intelligence optimization protocol able to predict the arrangement of protein symmetric assemblies by exploiting a limited amount of experimental restraints and steric interactions. Importantly, within this scheme the native flexibility of each protein subunit is taken into account as extracted from molecular dynamics simulations. We show that this is a key ingredient for the prediction of biologically functional assemblies when, upon oligomerization, subunits explore activated states undergoing significant conformational changes.

INTRODUCTION

Proteins often assemble in multimeric states to perform specific biological functions (Gavin et al., 2002). As the native folded state of single proteins, the most stable conformation of these macromolecular assemblies corresponds to an arrangement that is unique. Unfortunately, it is usually difficult to characterize the structure of multimeric assemblies at atomistic resolution. This is due to both their size and complexity, which make the production of sufficiently pure crystal for X-ray crystallography challenging. Moreover, if the assembly is amphipathic, such as in the case of transmembrane structures, crystallization is even more difficult. For this reason, *in silico* methodologies aiming at predicting how multiple protein copies arrange to form a multimeric complex would be desirable. The prediction of protein assembly can be interpreted as a minimization problem having an extremely large and complex search space. In principle, knowing the structure of an individual subunit should be sufficient to reconstruct the structure of the whole assembly. However, both the complexity of the search space and the fact that proteins are intrinsically dynamic objects, often undergoing conformational changes when multimerizing (Bahadur and Zacharias, 2008), contribute to making the prediction of a macromolecular assembly structure an extremely challenging task.

To tackle this problem several solutions have been proposed to date. Some approaches, such as SymmDock (Schneidman-Duhovny et al.), M-ZDOCK (Pierce et al., 2005) or MolFit (Berchanski et al., 2005), exploit the fact that often multimers respect a specific symmetry (Plaxco and Gross, 2009), (Goodsell and Olson, 2000). These methods reduce the search space by imposing a specific symmetry and subsequently rigidly dock the binding partners, so that a predefined energy function is minimized. The aforementioned schemes are *ab initio*, i.e. they do not directly exploit any previous knowledge about the system being studied. Still, it is important to point out that, while producing an X-ray structure of a macromolecular assembly is challenging, low-resolution data is usually more accessible. While this structural information can be used to filter or re-rank the models produced by *ab initio* methods, it can also be used to directly guide the assembly process by providing geometric restraints that the final assembly should respect (Lensink and Wodak, 2010),(Alber et al., 2008). The advantage of such approach is that the search space can be greatly reduced, as a consequence reducing the computational effort needed to explore the assembly's conformational space. Shih *et al.*, presenting a thorough analysis on the effect of quantity and precision of distance constraints, showed

that their prediction scheme based on distance restraints, DPPD, performs equally or better than *ab initio* methods (Shih and Hwang, 2012). In the context of experiment-driven protein assembly prediction one of the major efforts is represented by IMP (Integrative Modeling Platform) (Russel et al., 2012). IMP is able to deal with a variety of experimental restraints and predicts the rigid body arrangement of very large and heterogeneous macromolecular assemblies (Alber et al., 2007), (Lasker et al., 2012) via Monte Carlo and conjugate gradient search. On a similar side, HADDOCK (de Vries et al., 2007) can not only deal with a large variety of experimental restraints, but also with protein flexibility by first rigidly docking up to six subunits according to a predefined symmetry, and subsequently refining the result via simulated annealing. Another approach, Rosetta, has been shown to precisely predict multimeric arrangements according to several symmetries, also keeping into account both backbone and side-chain flexibility via a Monte Carlo-based multistep refinement procedure (Andre et al., 2007).

Upon assembly, constituent subunits often explore regions of their conformational space that can be markedly different from those captured by high-resolution X-ray crystallography or NMR structures of the single subunit. If additional activation processes are required to trigger assembly, such as proteolytic cleavage, post-translational modifications, etc., the conformation of subunits in the assembled (bound) state is expected to significantly differ from that observed in the free, unbound state. Importantly, there is increasing evidence of conformational selection playing an important role in the recognition process of protein binding (Boehr et al., 2009), (Lange et al., 2008), (Peters and de Groot, 2012). Protein bound conformations are thus expected to already populate the accessible conformational space explored by the unbound protein. Therefore, this opens the intriguing possibility to characterize the ensemble of states that are relevant for assembly by a thorough exploration of the conformational space of single subunits. Nowadays this task is possible thanks to the ever-increasing sampling power of large-scale molecular dynamics (MD) simulations of proteins in their native environment (Dror et al., 2012). Importantly, improved performances have been demonstrated by methods keeping into account protein's conformational space by docking ensembles of structures generated via MD simulations (Grunberg et al., 2004) (Smith et al., 2005), linear combinations of eigenvectors extracted by essential dynamics analysis (Mustard and Ritchie, 2005) or NMR experiments (Chaudhury and Gray, 2008). A hurdle in ensemble docking is however represented by its larger computational weight.

Our goal is to predict multimeric symmetric arrangements using both the structural and dynamic information of the monomeric state, guided by low-resolution spatial restraints characterizing the final assembly. In this context, minimization algorithms based on classical mathematical approaches (such as steepest descent or conjugate gradient) are unsuitable, since they tend to fail to converge to the global minimum. We tackled this problem by adopting a Particle Swarm Optimization (PSO) algorithm. PSO is a distributed heuristic optimization technique which has been shown to be both highly robust to local minima and usually converging as fast, and in some cases even faster, than other heuristic approaches such as genetic algorithm or simulated annealing (Besozzi et al., 2009), (Elbeltagi et al., 2005), (Namasivayam and Günther, 2007), (Angeline, 1998), (Abraham and Liu, 2009). PSO has been used successfully for docking small molecules in protein active sites (Chen et al., 2007), (Morris et al., 1999), (Meier et al., 2010) but, in our knowledge, our implementation is the first example of PSO applied to the broad domain of protein multimeric assembly.

In order to keep large portions of protein conformational space into account and understand the assembly mechanism, we exploit the increasing sampling capabilities of MD simulations as an additional resource. Importantly, in our method this conformational space is directly added to PSO search space. We show that our approach can quickly predict a reasonable protein arrangement in a symmetrical multimeric conformation by selecting, if necessary, the most suitable structure within the available conformations. Assembly prediction is performed in a variety of assembly situations, and guided by a limited amount of geometric restraints.

RESULTS

We aim at finding quickly a reasonable prediction for a multimeric structure arrangement on the basis of structural information about its subunits and experimental measures acting as search restraints (Figure 1). In a first step, an ensemble of monomer conformations is generated from molecular dynamics simulations or, alternatively, from structural biology experiments; this will be treated as a conformational database (see Experimental Procedure). The advantage of such an approach is that assembly prediction is performed using physically plausible structures. Subsequently, upon definition of a list of geometric restraints and a specific circular symmetry, a new kind of Particle Swarm Optimization

(PSO) search dynamically explores the conformational database. PSO searches for a multimeric assembly respecting all the given restraints and presenting no steric clashes. Geometric restraints can be typically provided by low-resolution electron density maps or experiments such as cross-linking disulfide scanning, labeling or FRET. If necessary, multimers can be assembled on a given substrate. At PSO search completion, a large set of solutions having a good score is usually generated. A smaller set of representative solutions, typically less than ten (Table 1), is returned by clustering the accepted solutions according to their respective Root Mean Square Deviation (RMSD). At present, the structure of hetero-dimers (thus addressing general protein-protein interactions) or homo-multimers with or without a target substrate (if a circular symmetry is imposed) can be predicted. This process is usually very fast, and can produce small ensemble of solutions being sufficiently good to generate biologically sound working hypotheses, and act as seeds for further optimization steps using more computationally expensive techniques.

We tested our protocol on a variety of systems having been already solved in their multimeric conformation. These cases ranged from a dimer having no imposed symmetry up to a 24-mer having an imposed circular symmetry, where we use the structure of subunits in the bound conformation to perform prediction. In all tests, a variety of geometric restraints plausibly derived from experiments of different nature were exploited and their influence on final predictions tested. In a second phase, we applied our method to three cases in which both the protein's bound and unbound conformations are known. Here, the conformational spaces of the unbound states were characterized by means of molecular dynamics simulations. Finally, we tackled a real case: the prediction of the heptameric conformation of pore-forming toxin aerolysin on the basis of known cryo-EM maps. In this case, we demonstrate that accounting for protein flexibility plays a key role in the prediction of biologically relevant states.

Rigid Assembly using Symmetry and Stoichiometry

We applied our method to protein complexes exploiting information usually accessible, like their stoichiometry and relative produced symmetry, to test its correctness. We selected a set of bound complexes (for which we use the bound monomeric conformation to rigidly reconstruct the assembly) spanning several symmetry classes, namely Chorismate Mutase (C3 symmetry) Acyl Carrier Protein Synthase (C3), Lumazine Synthase (C5), SM

Archeal Protein (C7) and EscJ (C24). We also included one case, PhoQ (C2), where no symmetry was imposed *a priori* and disulfide cross-linking data were used to guide assembly ((Goldberg et al., 2008)(Lemmin et al., 2013), see SI for further details). It should be noted that some of these proteins have been also adopted as test cases in the context of *de novo* prediction (Andre et al., 2007). In all the cases we attempted to reconstruct the original multimer on the basis of a known stoichiometry by providing various combinations of experimentally plausible geometric restraints (see details in Supplemental Information, SI). In SI further tests aimed at assessing the influence of restraint choice (i.e., quantity, quality and nature) on the final prediction are also presented. Results with relative computation timing are reported in Table 1, and best structures superimposed to the known crystal structure are shown in Figure 2A-D.

With the increase of geometric restraints stringency, the amount of obtained models decreases and their quality increases (Figure 3A, and Figure S3 for a complete benchmark). Importantly, in all our test cases at least one good model (RMSD < 2 Å, Table 1, Figure 2A-D) was generated and, when using at least 4 distance restraints, most produced models were consistent with the original crystal. Since docking was rigid at this stage, i.e. proteins were not deformed during optimization, part of RMSD difference is explained by the small differences within assembled subunits. Despite producing clash-free models consistent with imposed restraints, the fitness function in its current simple implementation (see Experimental Procedure) is not able to optimally rank obtained structures. On the other side, this limitation is balanced by the very limited set of solutions produced by our method (Table 1), which allows a direct (e.g. *in silico* and/or *in vitro*) assessment of the biological significance of the ensemble. Execution time is affected by the protein size and the complexity of desired geometric restraints. The largest contributor to execution time is however the post-processing phase, which is affected by the amount of final solutions that need to be generated. By imposing more severe filtering and clustering thresholds, a smaller number of solutions can be obtained. In these examples, a small number of possible assemblies were produced, in most cases in less than five minutes.

Assembly Considering Protein Native Flexibility

Whereas the previous tests are to be considered as ideal cases for exploring the general correctness and performance of our method, keeping into account the intrinsic flexibility of the unbound monomeric subunit can be crucial for sampling conformations more favorable to form the final assembly. To address this point, we selected three cases where both the unbound and bound structures have been solved at atomistic resolution. The first was phospholipase A2, having an RMSD within bound and unbound conformation equal to 0.8 Å. The second and third, harder cases, were *Flavivirus* trimeric envelope glycoprotein and *HIV-1* hexameric capsomer, having an RMSD of 4.4 Å and 10.5 Å, respectively. These proteins have been also used by (Mashiach-Farkash et al., 2011) for benchmarking their refinement method SymmRef against the *de novo* modeling program SymmDock. As before, tests have been run for assessing the effect and nature of the restraints for prediction (Figure S4-5). In this context, our strategy was to explore the conformational space of the unbound state using MD simulations and to include this conformational ensemble during the optimization step (see Experimental Procedure).

When the difference between bound and unbound states is mild, such in the case of phospholipase A2, no major difference between the results obtained with a rigid and flexible approach can be observed (Table 1, Figure 2E). In fact, rigidly docking the unbound structure was sufficient to obtain very good results. By imposing more than one geometric restraint, less than 3 models could always be produced, all of them having a RMSD smaller than 2 Å with respect to the known assembly (Figure 3B, S4 and S5). This result is better than the first acceptable solution produced by SymmDock (9.2 Å) (Mashiach-Farkash et al., 2011).

Unlike phospholipase A2, the case of *Flavivirus* envelope glycoprotein could not be considered as simple. A ~250 ns simulation was run and exploited as conformational database. While the RMSD between crystallographic bound and unbound states is 4.4 Å, during MD simulation RMSD values as low as 3.5 Å were explored. A more thorough observation of the trajectory revealed that the protein, very elongated, flexes along its main axis. Seven eigenvectors, that represented 80% of protein motion, were automatically selected as additional dimensions in the search space (see Experimental Procedure). Exploring this conformational space, 13 models were finally produced, the best one having a RMSD of 4.2 Å with respect to the known assembled crystal, 3.8 Å excluding most flexible and solvent exposed loops (Figure 2F, Table 1). Interestingly, the final RMSD is

lower than that between unbound and bound state and is, in fact, built by automatically selecting one of the best MD frames available (i.e., RMSD of 3.6 Å with respect to the bound state). By comparison with the unbound structure the selected frame has the important advantage of better capturing the protein-wide motion involved in the binding process, making it therefore a more suitable candidate for assembly prediction. This result is comparable to the first acceptable result produced by SymmDock (3.5 Å) (Mashiach-Farkash et al., 2011).

A more striking case is constituted by the HIV-1 hexameric capsomer, which is composed of two domains connected by a flexible linker. We performed a ~500 ns long MD simulation of the unbound conformation and tracked its RMSD with respect to the known bound conformation. While a large majority of structures were very different from the bound state (RMSD greater than 10 Å), during a rearrangement of the two domains the RMSD was as low as 2.9 Å (Figure 4A). This indicates that states very close to the bound conformation, although less populated, are present in the conformational space of the unbound state. We ran three optimizations one for the bound state, one for the unbound state, and one keeping into account the MD ensemble, using the same geometric restraints. Not surprisingly, no acceptable model was obtained when using the unbound conformation. Conversely, when using the bound state, 9 solutions were produced, the best one having a RMSD equal to 0.6 Å with respect to the known crystal. As comparison, this result is better than the first acceptable solution produced by SymmDock (1.7 Å) and, remarkably, comparable to its refinement *via* SymmRef (0.9 Å) (Mashiach-Farkash et al., 2011). Finally, we attempted to produce a model by exploiting the whole MD trajectory of the protein's unbound state. Two eigenvectors were sufficient to describe 84% of the protein motion (Figure 4B), and PSO finally produced 16 models, the best one having a RMSD equal to 3.7 Å. Importantly, our method assembled this model by automatically selecting as building block the MD frame with the lowest RMSD with respect to the bound state (Figure 4C,D). Moreover, all the produced models were built using monomers having RMSD values lower than 3.5 Å with respect to the bound state.

These latter results clearly highlight how our method is able to pinpoint the most suitable structure for assembly within a conformational ensemble, outperforming a purely rigid docking approach and producing states that can be further refined by minimization or molecular simulation techniques. If the complete conformational space could be sampled by MD, one might expect that the PSO-based search would eventually find the optimal

state to assemble the multimeric complex. In practice, in the case of *Flavivirus* envelope glycoprotein MD was not able to explore states very close to the bound conformation in a 10^2 ns timescale. This is likely because the rearrangements involved upon assembly would have required a longer sampling time. On the other hand, in the same timescales MD explored states close to the bound conformation for the case of the HIV-1 hexameric capsomer, due to the quaternary architecture of the protein.

Assembly Considering Activation-Induced Protein Flexibility

In order to study a real assembly prediction situation, we applied our method to the prediction of the heptameric conformation of *Aeromonas hydrophila* pore-forming toxin aerolysin (monomeric pdb: 1PRE (Parker et al., 1994)). For this protein, flexibility plays an extremely important role: in fact, pore-forming toxins often undergo large conformational changes in order to assemble into a transmembrane complex. Interestingly, point mutation Y221G has been found to be able to impair the aerolysin assembly membrane insertion, leading to the creation of a hydrophilic multimer (i.e., prepore state) for which a cryo-EM map was obtained at a resolution of 16 Å (Tsitrin et al., 2002). We have already shown that the aerolysin C-terminal peptide (CTP, Figure 5A) acts as an intramolecular chaperone to preserve the correct folding of the monomer, but its cleavage and removal from aerolysin main lobe is then necessary for pore formation (Iacovache et al., 2011). Therefore, we characterized the dynamic features of both the inactive Y221G aerolysin (i.e. complexed with CTP as in the X-ray structure, Figure 5A) and the activated form, where the CTP was removed from the main lobe. We observed that the presence of the CTP largely affects the internal motion of aerolysin: while the inactive form explores conformations very close to the X-ray structure, the active form shows a larger interdomain flexibility and additional internal rotation of the domain originally holding the CTP (Figure 5A,B).

We used the knowledge about the activated conformational space of aerolysin in order to model its heptameric prepore assembly within our optimization framework. As restraints, the measures of Y221G cryo-EM map were adopted (i.e. height and width equal to 85±5 Å and 150±5 Å respectively, and pore radius equal to 5±2 Å, EM maps has been filtered at 0.5 and width and height were measured on the resulting volume, Figure 5C,D). In particular, two independent runs were performed, one providing as input an ensemble of

conformations generated from a ~200 ns MD simulation of Y221G aerolysin without the CTP, and one providing the X-ray structure of Y221G aerolysin pruned *a posteriori* of the CTP coordinates; the latter being representative of the relatively compacted conformational ensemble explored by MD (Figure 5B). In the first case, by keeping into account monomer flexibility as described by the 3 main PCA eigenvectors (94.5% of protein's movement), we obtained 6 models. All the models when docked into the Y221G EM using Situs (Wriggers, 2010) presented a good cross-correlation coefficient (CCC) with respect to it. The highest ranked structure had a CCC of 0.72, a single subunit of this model having an RMSD equal to 4.3 Å with respect to the original aerolysin crystal (Figure 5C). Importantly, this assembly appears to recapitulate all known biophysical and biochemical data for wild-type and Y221G aerolysin (whose impact is investigated in a forthcoming work). Conversely, when only the X-ray based CTP-bound conformation was used to assemble a multimer, none of the solutions explored met the solutions filtering criteria (Figure 5B,D). This indicates that no structure satisfying the given geometric restraints could be found. We docked the best solution (fitness equal to 4.75) into the Y221G cryo-EM map obtaining a CCC of 0.57, value significantly lower than what obtained using the dynamic modeling protocol and presenting several inconsistencies with respect to known experimental data.

Altogether, this blind result, along with tests on unbound cases (Figure 4), shows that using an ensemble of structures representative of monomer native flexibility, which also can take into account activated states, leads eventually to improved performances in assembly prediction and to the generation of more biologically sound structural models.

Effect of Quantity and Quality of Restraints

The quantity and accuracy of used restraints may greatly influence the performance of any protein assembly prediction methodology. To better validate our method, we performed a set of benchmarks involving a variable number of restraints, as well as a different error on the given experimental target measures (Figure 3, see also SI and Figures S3-5 for extended benchmarks). We selected the trimeric phospholipase A2 (both using its unbound structure and a ~500 ns MD ensemble) and acyl carrier protein as test structures. For both proteins, an ensemble of protein-protein contacts was first identified for the available multimeric structure. In order to test how our pipeline is affected by the amount of

provided restraints, we selected combinations from one to five contacts as restraint (Figure 3). To test the effect of noise on the target measure, all these tests were repeated by imposing errors from 1 to 6 Å on every target measure. Results are summarized in Figures 3 and S3-S5.

In all tests we observe that, not surprisingly, the amount of obtained models decreases rapidly the larger the amount of used restraints. The average solutions quality also improves with a higher number of restraints. Results show that, when using at least 4 restraints, a large majority of produced structures is consistent with the original complex (Figure 3), even when imposing a fairly large error on the measures, such as 6 Å (Figure S3-S5). Importantly, in any performed test at least one good solution (bRMSD < 2 Å) was present within the produced models. Using a MD trajectory as input (Figures 3B and S5) instead of a single structure (Figure S4) results in a larger amount of solutions to be produced when using a small amount of constraints. Phospholipase A2 features a flexible loop, which is the major responsible of a different RMSD within the produced models. If no restraints involving this loop are set, a large amount of solutions is produced. However, using a larger amount of restraints (more than 3) leads to performances comparable to the ones obtained using a single structure as input. We observe that the location of restraints on the assembly surface can affect the optimization process. Contacts involving residues located in a protein cavity, for instance, would impose more severe geometric restraints than contacts involving completely exposed protein regions. Relative location of restraints also plays an important role. Better results are usually obtained when identified contacts are spread on the whole protein-protein interface, especially when the error on individual measures becomes large.

DISCUSSION

We presented a new method based on swarm intelligence optimization that is able to predict, exploiting a limited set of low-resolution experimental spatial restraints, the conformation of homo-multimers according to a predefined circular symmetry or general protein-protein interactions. A first major strength of our algorithm is that, thanks to a novel, robust and efficient PSO implementation, it can quickly return a small set of possible structures respecting the imposed restraints and presenting no severe clashes. These models can be subsequently refined and re-ranked with techniques such as SymmRef

(Mashiach-Farkash et al., 2011) or MDFF if a density map is available (Trabuco et al., 2008).

Furthermore, our method is designed to take native protein flexibility into account by automatically extracting relevant conformations from a provided structural ensemble. In this context we use the sampling power of molecular dynamics that currently allow exploring relevant portions of the conformational space accessible to protein subunits. The advantage of such approach is that significant conformational changes, involving for instance inter-domain rearrangements, can be kept into account. Recent evidence indicates conformational selection as a major mechanism implicated in recognition processes as protein binding and assembly. Namely, bound conformations are already expected to populate the accessible conformational space explored by unbound proteins. Thus, in principle, the investigation of the conformational ensemble of single subunits alone can already provide useful information about their state upon assembly. Using aerolysin as a real case, we demonstrated indeed that exploiting a structural ensemble produced by MD simulations greatly improved the prediction outcome. We expect that, with the ever-growing sampling capabilities of current molecular dynamics simulations, the hybrid strategy implemented in our method will offer in the future an unprecedented, robust and efficient way to address molecular assembly through dynamic modeling.

As a matter of fact, such conformational ensemble could be generated via other techniques (e.g. Monte Carlo simulation, NMR experiments) allowing the exploration of a macromolecule conformational space. In this context, one of the possible extensions is the possibility of automatically performing a normal modes analysis on a single provided structure. Exploration of the conformational space constituted by a linear combination of the main discovered modes would subsequently take place. The immediate advantage of this approach is clearly its extremely affordable computational cost. On the other side, flexibility extracted from a harmonic approximation of protein molecular interactions is limited to a specific equilibrium state and cannot access further conformational arrangements as seen in MD simulations, especially for further activated states (Figure 5).

Assembly predictions can be guided by virtually any possible experimental results providing insights about proteins structure. Macroscopic measures such as height and width extracted from cryo-EM experiments, atomic distances obtained for instance from FRET, disulfide cross-linking, chemical cross-linking coupled to mass spectrometry experiments or identification of location of specific amino acids in the assembly structure

(for instance from gold labeling or alanine scanning) can all effectively lead to a correct assembly prediction. At present, cryo-EM maps only provide our protocol with information about the assembly general shapes (height, width, concavity, pore radius, etc.). PSO can, however, deal with any kind of fitness function. Thus, a natural extension of our protocol will be the direct assembly of homo- and hetero-multimers into a provided electron density map.

It has to be pointed out that our approach is based on a purely geometric optimization, side chain arrangement is therefore not refined at the moment. In fact, our aim at this stage was to quickly generate a small ensemble of reasonable protein arrangements. Even though we do not produce a refined *de novo* prediction, our results can provide important insights into multimeric arrangement and guide the design of new experiments. The current energetic contribution to the fitness function is simply constituted by a coarse potential on the protein scaffold used to avoid steric clashes. Energy scoring based on molecular mechanics contribution of the monomer-to-monomer binding would likely help to rank the best solutions and address completely experimentally blind assembly searches. The goal is to obtain a lightweight but more accurate energetic contribution, better describing van der Waals interactions and accounting for electrostatic contributions. In this respect, taking care of side chains refinement could also lead to significant improvements.

In conclusion, we believe we just scratched the surface of the capabilities of this novel approach. In fact, since the PSO engine is not sensitive to the kind of imposed symmetry, implementing other common symmetries (such as helical or icosahedral) is trivial, and is part of our future development plans. Moreover, improvements on the fitness function, by including a broader set of geometric restraints, and the energy scoring, with the use of more accurate molecular mechanics potentials, will certainly enhance the quality of the predicted assemblies, and eventually address the prediction of protein-protein interactions. Finally, the increasing capability of MD coupled to this optimization protocol has the great advantage to account for native protein flexibility, activation processes and conformational selection of relevant states upon assembly. Given the number of multi-protein complexes, crucial for key cellular functions, that are amenable to low-resolution analyses but not X-ray crystallography, we believe that this dynamic modeling approach will be widely used in the future.

EXPERIMENTAL PROCEDURES

Kick-and-Reseed Particle Swarm Optimization

Let a function f(**x**), where $\mathbf{x} \in F^n \subset R^n$. We call f the *fitness function*, and *search space* the multidimensional real space F^n in which the function is defined. We want to find a point $\mathbf{x}_{min} \in F^n$ such that $y_{min}=f(\mathbf{x}_{min})$ is the function's global minimum. PSO aims at finding the point $\mathbf{x}_{min}$ having the lowest fitness. This technique represents the search process as a model of birds' social behavior when flocking (Kennedy and Eberhart, 1995). To do so, an ensemble of solutions (also called particles) has its position and velocity randomly initialized in the multidimensional search space. Along the whole optimization process, every particle will keep track of the value f_{best} and position $\mathbf{x}_{best}$ of its best-found solution. At the beginning of every discrete timestep, particles *p* are updated about the swarm status, i.e. the current position of all particles, as well as their respective best found solution value and location. Subsequently, they will independently update their own velocity **v**, which in turn will be used to update their position **x**. This is repeated until a maximal number of timesteps is reached. Since, after having been updated about the swarm status, particles act as independent agents, PSO can be considered as embarrassingly parallel, easily benefiting from current multi-core architectures. Velocity update is affected by 3 factors. The first, *inertia*, determines how a particle's trajectory is preserved along time. The second, *personal best*, attracts particles towards their own best solution. The third, *global best*, attracts particles towards the best solution found by neighboring particles (Figure 6). In order to increase PSO robustness and convergence rate, we implemented a new kind of PSO called PSO *kick and reseed* (PSO-KaR, see full details in SI). In PSO-KaR, particles being too slow (i.e. stagnating in a minimum) have their position and velocity randomly reinitialized, and their memory erased. This PSO-based method is implemented in a code called *Parallel Optimization Workbench – pow*^{er}, available as open source at http://lbm.epfl.ch. All the tests were performed running 3 PSO-KaR repetitions of 200 steps with 80 particles. PSO behavioral constants w_{start}, w_{end} , *cp* and *cn* were set to 0.9, 0.4, 1.2 and 1.4 (see SI). 4 processors on an Intel AMD64 dual-quad core machine were used. All reported execution times (Table 1) also include post-processing, i.e. solution filtering, clustering and generation of corresponding PDB files.

Search Space Definition and Data Manipulation

Our scheme aims at predicting the assembly of two proteins or, by imposing a predefined symmetry, of multimeric assemblies. In particular, when docking two proteins, the search space is six-dimensional (protein translation and rotation around the x, y and z axis). If an ensemble of protein structures is available, namely obtained from a MD trajectory (or alternatively NMR or X-ray experiments), flexibility (or multiple conformations) can be introduced as set of further dimensions in the search space. To do so, a principal component analysis (PCA) is initially performed on the ensemble (Figure 1). The projection value of every trajectory frame along the most relevant eigenvectors, also called fluctuations, is computed. These are used as a way to index the trajectory frames, which we can consider as a protein conformation database. The search space is eventually characterized by three rotations, three translations, and *n* fluctuations. In order to produce the assembly corresponding to a specific position in the search space, the subunit in the database having its eigenvector projection being the closest to the desired fluctuation values is first extracted. Subsequently, the selected conformation is roto-translated. The main advantage of using an ensemble of conformations is that the protein conformations used to assemble the multimer will more closely respect protein native flexibility.

This scheme is extended to assemble circular symmetric structures given a known stoichiometry. In this context, the conformational space is defined by the three rotation angles of a single monomer with respect to a center of symmetry aligned along the *z* axis, and a displacement with respect to it, which represents the radius of the assembly in its narrowest point. This methodology has been already successfully applied to model several isoforms of human C4b-binding protein (Hofmeyer et al., 2013). Our method has been also extended to flexibly assemble a multimeric complex on a rigid receptor, as recently done for the prediction of the basal body YscDJ complex of the *Yersinia* type III secretion system (in press). In this case additional degrees of freedom, i.e. the translation of the whole assembly along the z axis and the rotation around it, are also kept into account.

Fitness Function and Clustering

The fitness function scoring the quality of an assembly depends on two factors, geometry and energy. As geometric contribution, specific measures of the current multimer *m* are compared to target values **t** being experimentally known. Let **c**(m) an ensemble of

measures performed on a multimer. The geometric score G(m) of a multimer is determined by the Euclidean distance within obtained and target measures:

$$G(m)= \{[\mathbf{t}-\mathbf{c}(m)]\cdot[\mathbf{t}-\mathbf{c}(m)]\}^{1/2} \quad (1)$$

In order to avoid steric clashes during assembly, a coarse energy potential is also taken into account. This "minimalistic" contribution is constituted by a 9-6 Lennard-Jones-type potential describing all the C_α and C_β atoms of two neighboring monomers extracted from the assembly:

$$E(m)= 4\varepsilon [(\sigma / r)^9 – (\sigma / r)^6] \quad (2)$$

where *r* are all the distances within couples of atoms being at a distance smaller than 12 Å. ε=1 kcal/mol and σ=4.7 Å correspond instead to a coarse-grained model for C_α atoms (Alemani et al., 2010). The final fitness function *f* mixes geometric and energetic contributions by means of the following weighted sum:

$$f(m)=c*E(m)+(1-c)*G(m) \quad (3)$$

where *c* is a real value within 0 and 1. In preliminary systematic tests, we found that when *c* has a value smaller than 0.05 no suitable solution is usually found, whereas with values greater than 0.2 a large amount of solutions is returned (not shown). In our tests we set *c*=0.2. It should be noted that the rough energy function in eq. 1 only avoids clashes, and is therefore not sufficiently precise to allow a blind docking (i.e. without geometric restraint) and accurate ranking of the solutions. During execution, solutions having a fitness lower than 0 (i.e. most likely clash free and respecting the given geometric restraints) are retained. Solutions having their respective RMSD smaller than 1 Å are subsequently clustered; cluster representatives (centers) are finally selected. More details can be found in Supplemental Information.

ACKNOWLEDGEMENTS

We gratefully thank Thomas Lemmin for helping with PhoQ calculations; Marco Stenta, Davide Alemani and Enrico Spiga for having tested and helped debugging *pow*^{er} first version; Gisou van der Goot and Ioan Iacovache for the inspiration for developing this method. This work is supported by the Swiss National Science Foundation (SNF grant numbers: 200021_122120 and 200020_138013).

REFERENCES

Tsitrin, Y., Morton, C. J., el-Bez, C., Paumard, P., Velluz, M. C, Adrian, M., Dubochet, J., Parker, M. W., Lanzavecchia, S.,van der Goot, F. G., (2002). Conversion of a transmembrane to a water-soluble protein complex by a single point mutation. Nat Struct Biol *9*, 729-733.

Abraham, A., and Liu, H. (2009). Turbulent Particle Swarm Optimization Using Fuzzy Parameter Tuning. Foundations of Computational Intelligence Volume 3, 291-312.

Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.T., and others (2007). Determining the architectures of macromolecular assemblies. Nature *450*, 683-694.

Alber, F., Forster, F., Korkin, D., Topf, M., and Sali, A. (2008). Integrating diverse data for structure determination of macromolecular assemblies. Annual review of biochemistry *77*, 443-477.

Alemani, D., Collu, F., Cascella, M., and Dal Peraro, M. (2010). A Nonradial Coarse-Grained Potential for Proteins Produces Naturally Stable Secondary Structure Elements. J Chem Theory Comput *6*, 315-324.

Andre, I., Bradley, P., Wang, C., and Baker, D. (2007). Prediction of the structure of symmetrical protein assemblies. Proceedings of the National Academy of Sciences.

Angeline, P. (1998). Evolutionary optimization versus particle swarm optimization: Philosophy and performance differences. In Evolutionary Programming VII, pp. 601-610.

Bahadur, R.P., and Zacharias, M. (2008). The interface of protein-protein complexes: analysis of contacts and prediction of interactions. Cellular and molecular life sciences : CMLS *65*, 1059-1072.

Berchanski, A., Segal, D., and Eisenstein, M. (2005). Modeling oligomers with Cn or Dn symmetry: application to CAPRI target 10. Proteins: Structure, Function, and Bioinformatics *60*, 202-206.

Besozzi, D., Cazzaniga, P., Mauri, G., Pescini, D., and Vanneschi, L. (2009). A comparison of genetic algorithms and particle swarm optimization for parameter estimation in stochastic biochemical systems. Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, 116-127.

Boehr, D.D., Nussinov, R., and Wright, P.E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. Nature chemical biology *5*, 789-796.

Chaudhury, S., and Gray, J.J. (2008). Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles. Journal of Molecular Biology *381*, 1068-1087.

Chen, H.M., Liu, B.F., Huang, H.L., Hwang, S.F., and Ho, S.Y. (2007). SODOCK: swarm optimization for highly flexible protein-ligand docking. J Comput Chem *28*, 612-623.

de Vries, S.J., van Dijk, A.D.J., Krzeminski, M., van Dijk, M., Thureau, A., Hsu, V., Wassenaar, T., and Bonvin, A.M.J.J. (2007). HADDOCK versus HADDOCK: new features and performance of HADDOCK2. 0 on the CAPRI targets. Proteins: structure, function, and bioinformatics *69*, 726-733.

Dror, R.O., Dirks, R.M., Grossman, J.P., Xu, H., and Shaw, D.E. (2012). Biomolecular simulation: a computational microscope for molecular biology. Annual review of biophysics *41*, 429-452.

Elbeltagi, E., Hegazy, T., and Grierson, D. (2005). Comparison among five evolutionary-based optimization algorithms. Advanced Engineering Informatics *19*, 43-53.

Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M.*, et al.* (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141-147.

Goldberg, S.D., Soto, C.S., Waldburger, C.D., and DeGrado, W.F. (2008). Determination of the physiological dimer interface of the PhoQ sensor domain. Journal of molecular biology *379*, 656-665.

Goodsell, D.S., and Olson, A.J. (2000). Structural symmetry and protein function. Annual review of biophysics and biomolecular structure *29*, 105-153.

Grunberg, R., Leckner, J., and Nilges, M. (2004). Complementarity of structure ensembles in protein-protein binding. Structure *12*, 2125-2136.

Hofmeyer, T., Schmelz, S., Degiacomi, M.T., Dal Peraro, M., Daneschdar, M., Scrima, A., van den Heuvel, J., Heinz, D.W., and Kolmar, H. (2013). Arranged sevenfold: structural insights into the C-terminal oligomerization domain of human C4b-binding protein. J Mol Biol *425*, 1302-1317.

Iacovache, I., Degiacomi, M.T., Pernot, L., Ho, S., Schiltz, M., Dal Peraro, M., and van der Goot, F.G. (2011). Dual Chaperone Role of the C-Terminal Propeptide in Folding and Oligomerization of the Pore-Forming Toxin Aerolysin. PLoS pathogens *7*, e1002135.

Kennedy, J., and Eberhart, R. (1995). Particle swarm optimization. In Neural Networks, 1995. Proceedings, IEEE Internat.Conf., pp. 1942-1948.

Lange, O.F., Lakomek, N.A., Fares, C., Schroder, G.F., Walter, K.F., Becker, S., Meiler, J., Grubmuller, H., Griesinger, C., and de Groot, B.L. (2008). Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. Science *320*, 1471-1475.

Lasker, K., Förster, F., Bohn, S., Walzthoeni, T., Villa, E., Unverdorben, P., Beck, F., Aebersold, R., Sali, A., and Baumeister, W. (2012). Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. Proceedings of the National Academy of Sciences *109*, 1380-1387.

Lemmin, T., Soto, C.S., Clinthorne, G., DeGrado, W.F., and Dal Peraro, M. (2013). Assembly of the transmembrane domain of E. coli PhoQ histidine kinase: implications for signal transduction from molecular simulations. PLoS computational biology *9*, e1002878.

Lensink, M.F., and Wodak, S.J. (2010). Docking and scoring protein interactions: CAPRI 2009. Proteins *78*, 3073-3084.

Mashiach-Farkash, E., Nussinov, R., and Wolfson, H.J. (2011). SymmRef: a flexible refinement method for symmetric multimers. Proteins *79*, 2607-2623.

Meier, R., Pippel, M., Brandt, F., Sippl, W., and Baldauf, C. (2010). ParaDockS: A Framework for Molecular Docking with Population-Based Metaheuristics. J. Chem. Inf. Model *50*, 879-889.

Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K., and Olson, A.J. (1999). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. Journal of Computational Chemistry *19*, 1639-1662.

Mustard, D., and Ritchie, D.W. (2005). Docking essential dynamics eigenstructures. Proteins-Structure Function and Bioinformatics *60*, 269-274.

Namasivayam, V., and Günther, R. (2007). PSO@ Autodock: A fast flexible molecular docking program based on swarm intelligence. Chemical Biology & Drug Design *70*, 475-484.

Parker, M.W., Buckley, J.T., Postma, J.P.M., Tucker, A.D., Leonard, K., Pattus, F., and Tsernoglou, D. (1994). Structure of The Aeromonas toxin proaerolysin in its water-soluble and membrane-channel states. Nature *367*, 292-295.

Peters, J.H., and de Groot, B.L. (2012). Ubiquitin dynamics in complexes reveal molecular recognition mechanisms beyond induced fit and conformational selection. PLoS computational biology *8*, e1002704.

Pierce, B., Tong, W., and Weng, Z. (2005). M-ZDOCK: a grid-based approach for Cn symmetric multimer docking. Bioinformatics *21*, 1472-1478.

Plaxco, K.W., and Gross, M. (2009). Protein complexes: the evolution of symmetry. Current biology : CB *19*, R25-26.

Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B., and Sali, A. (2012). Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. PLoS Biology *10*, e1001244.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H.J. (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic acids research *33*, W363-W367.

Shih, E.S., and Hwang, M.J. (2012). On the use of distance constraints in protein-protein docking computations. Proteins *80*, 194-205.

Smith, G.R., Sternberg, M.J.E., and Bates, P.A. (2005). The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes

with an application to protein-protein docking. Journal of Molecular Biology *347*, 1077-1101.

Trabuco, L.G., Villa, E., Mitra, K., Frank, J., and Schulten, K. (2008). Flexible Fitting of Atomic Structures into Electron Microscopy Maps Using Molecular Dynamics. Structure *16*, 673-683.

Tsitrin, Y., Morton, C.J., el-Bez, C., Paumard, P., Velluz, M.C., Adrian, M., Dubochet, J., Parker, M.W., Lanzavecchia, S., and van der Goot, F.G. (2002). Conversion of a transmembrane to a water-soluble protein complex by a single point mutation. Nat Struct Biol *9*, 729-733.

Wriggers, W. (2010). Using Situs for the integration of multi-resolution structures. Biophysical reviews *2*, 21-27.

TABLES

Table 1. Summary of assembly prediction using various experimental restraints

Symmetry, execution time (including post-processing), final number of independent solutions and best protein backbone RMSD (bRMSD) with respect to the known assembly are indicated. In brackets, the rank of the best solution is also indicated. Bound, unbound and aerolysin cases are reported in this order. Relative structural superimpositions are reported in Figures 2, 4 and 5.

Protein	symmetry	time	solutions	bRMSD
PhoQ	C2	3m00s	6 (1)	1.85 Å
Chorismate Mutase	C3	2m16s	20 (3)	1.52 Å
Acyl Carrier	C3	2m59s	3 (1)	1.91 Å
Lumazine Syntaze	C5	3m53s	6 (5)	1.89 Å
SM Archeal Protein	C7	2m30s	6 (1)	0.95 Å
EscJ	C24	8m52s	9 (3)	2.04 Å
Phospholipase A2 *	C3	1m36s	2 (1)	1.58 Å
Envelope Glycoprotein *	C3	7m03s	14 (4)	3.81 Å
Hexameric Capsomer *	C6	6m50s	16 (14)	3.74 Å
Aerolysin *	C7	6m35s	6 (4)	0.72 **

* unbound cases
** cross-correlation coefficient with respect to the density map

FIGURES LEGENDS

Figure 1. Macromolecular assembly prediction workflow

When a structural ensemble is provided (e.g. from a MD trajectory), principal component analysis (PCA) is first performed. Fluctuations of main eigenvectors are added to the search space, which also includes protein roto-translation dimensions. Every point in the search space is a protein assembly, which is evaluated by a fitness function taking into account both assembly energy and geometric restraints typically based on experimental evidence. Fitness function minima in the search space are sampled via a novel kind of Particle Swarm Optimization (PSO) algorithm (see Figure 6). The best solutions are filtered and clustered, and their corresponding multimeric structures produced. These can be subsequently refined via more expensive computational techniques. This method for macromolecular assembly is implemented in an open source code called *Parallel Optimization Workbench – pow*^{er}, which is available at http://lbm.epfl.ch.

Figure 2. Molecular assembly predictions

Best symmetrical assembly predictions (in yellow) superimposed to known X-ray crystal structures (in blue) of bound (**B, C, D**) and unbound (**E, F**) cases. In panel **A**, no symmetry is imposed *a priori* for assembling the PhoQ dimer, and red spheres highlight the position of residues displaying high disulfide cross-linking efficiency used as distance restraints. See Table 1 for the relative RMSD numerical values.

Figure 3. Assessing the effect of quality and quantity of imposed restraints

An ensemble of six native contacts (distance 4±2 Å) has been selected for acyl carrier (**A**, bound case) and phospholipase A2 (**B**, unbound case). For every combination of these, an assembly prediction has been run. The amount of produced solutions (left column) and the distribution of model RMSD with respect to the known structure (right column, where red lines represent median, blue boxes the intervals between lower and upper quartiles, whiskers the lowest and highest observations, and red crosses the outliers) have been finally extracted. When increasing the number of imposed restraints, a smaller amount of

solutions is returned, most having an RMSD below 2 Å with respect to the known assembly. See Figures S3-S5 for a complete benchmark considering a broader noise-range contribution on restraints.

Figure 4. Molecular assembly prediction accounting for native dynamics

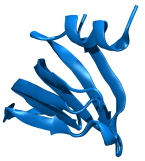
(A) Time evolution of monomeric *HIV-1* hexameric capsomer RMSD with respect to initial unbound reference X-ray state. At around 300 ns, the protein visits conformations with low RMSD (less than 3 Å, indicated by a red dot) with respect to the bound state. **(B)** Projection of the first two eigenvectors calculated by PCA reveals that, in the MD eigenspace of the unbound state, some structures appear close to the bound state. **(C)** Superposition of the unbound structure and the best frame extracted from MD simulation that is close to the bound state. **(D)** Best model produced by our method (3.7 Å RMSD, in yellow) is superimposed to known X-ray crystal structure of the complex (in blue). A frame having a low RMSD in the conformational database (2.9 Å, indicated by the red dot in panels A and B) was automatically selected as building block.

Figure 5. Molecular assembly prediction accounting for native flexibility upon activation

(A) Aerolysin crystal structure and schematic representation of its two main modes of motion (black arrows). Removal of C-terminal peptide (CTP, in grey) activates the toxin, while mutation Y221G (in yellow) locks aerolysin assembly in a prepore conformation. **(B)** Projection of the first two eigenvectors calculated from PCA of the MD trajectories with CTP (in black) and without (in blue). Upon activation, a much larger conformational space with respect to the X-ray static structure (in red) is explored. **(C)** By keeping into account protein native flexibility upon activation (see **B**), a very good fit (CCC=0.72) is obtained with respect to the available EM map, and a structure fully compatible with available biological data is predicted. **(D)** By only using aerolysin crystal structure to predict the heptameric conformation of the prepore state, a model having a CCC equal to 0.57 is instead assembled.

Figure 6. PSO-KaR algorithm for molecular assembly prediction

Pseudocode of the new Kick-and-Reseed Particle Swarm Optimization. At every timestep *t*, the position ***x*** and velocity ***v*** in the search space of every particle *p* is updated according to three factors: *inertia*, *personal best* and *global best*. These contributions are weighted by *w, cp* and *cn* coefficients, respectively. Every particle keeps track of the position $\boldsymbol{x}_{best}$ and value f_{best} of its best-found solution. When a particle's velocity drops below a predefined threshold $\boldsymbol{v}_{min}$ (code in the rectangular box) its velocity is randomly reinitialized ("kick"), whereas if the current fitness value is lower than a threshold f_{min}, its position is restarted as well, and its memory erased ("reseed"). See Figures S1-S2 for performance on classic benchmark functions and Supplemental Information for details about the algorithm.



input: monomeric structure

flexibility analysis

either PCA on structures ensemble, or rigid body on single structure

generate fitness function

1 energetic (Lennard-Jones) and n experiment-based geometric contributions

docking

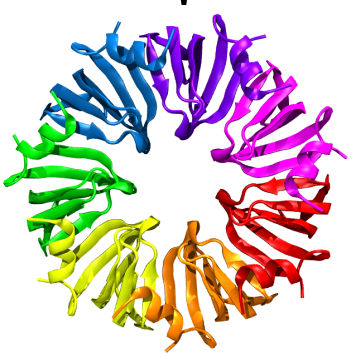
Particle Swarm Optimization

cluster solutions

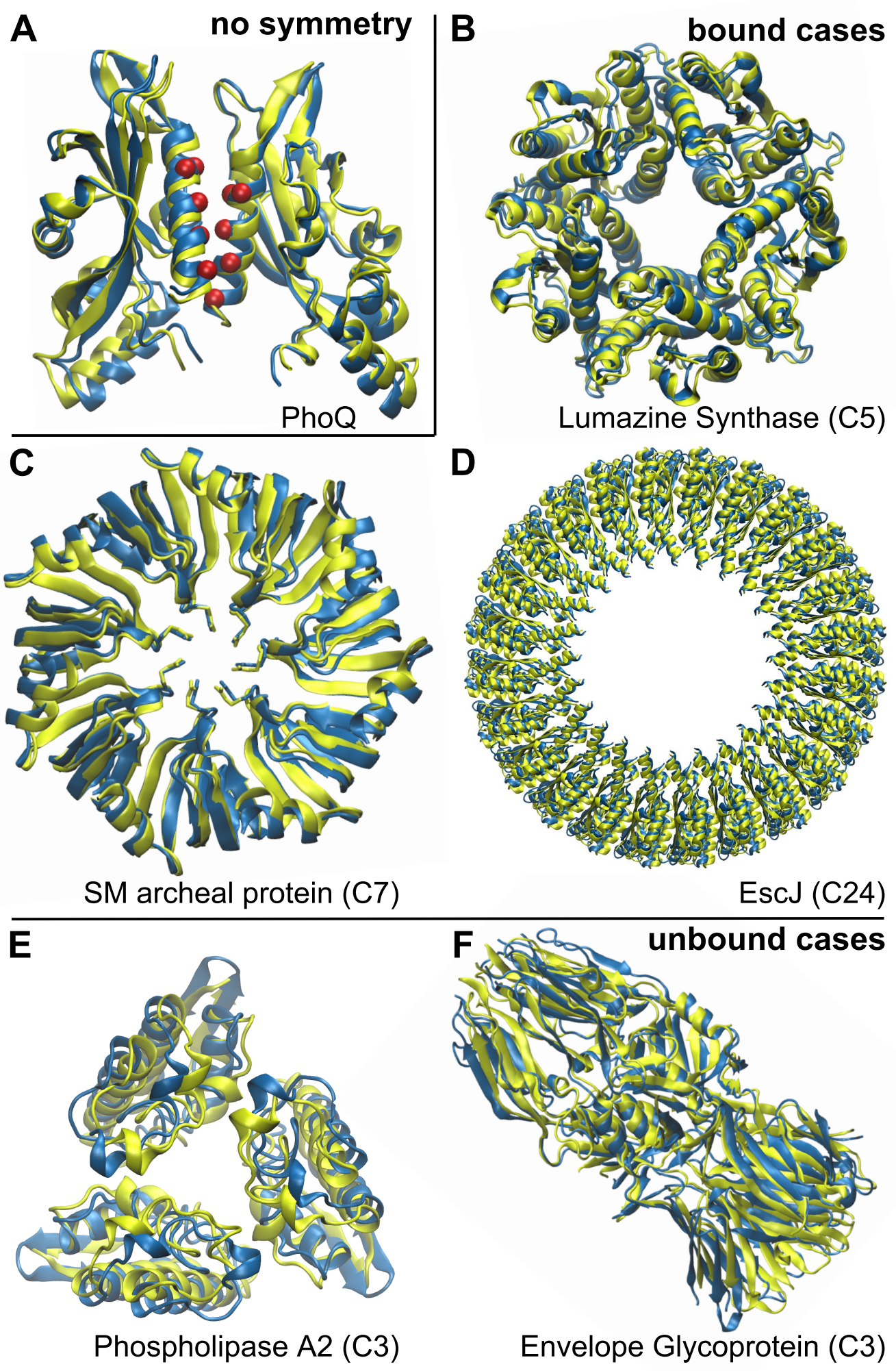
best solutions RMSD-based clustering

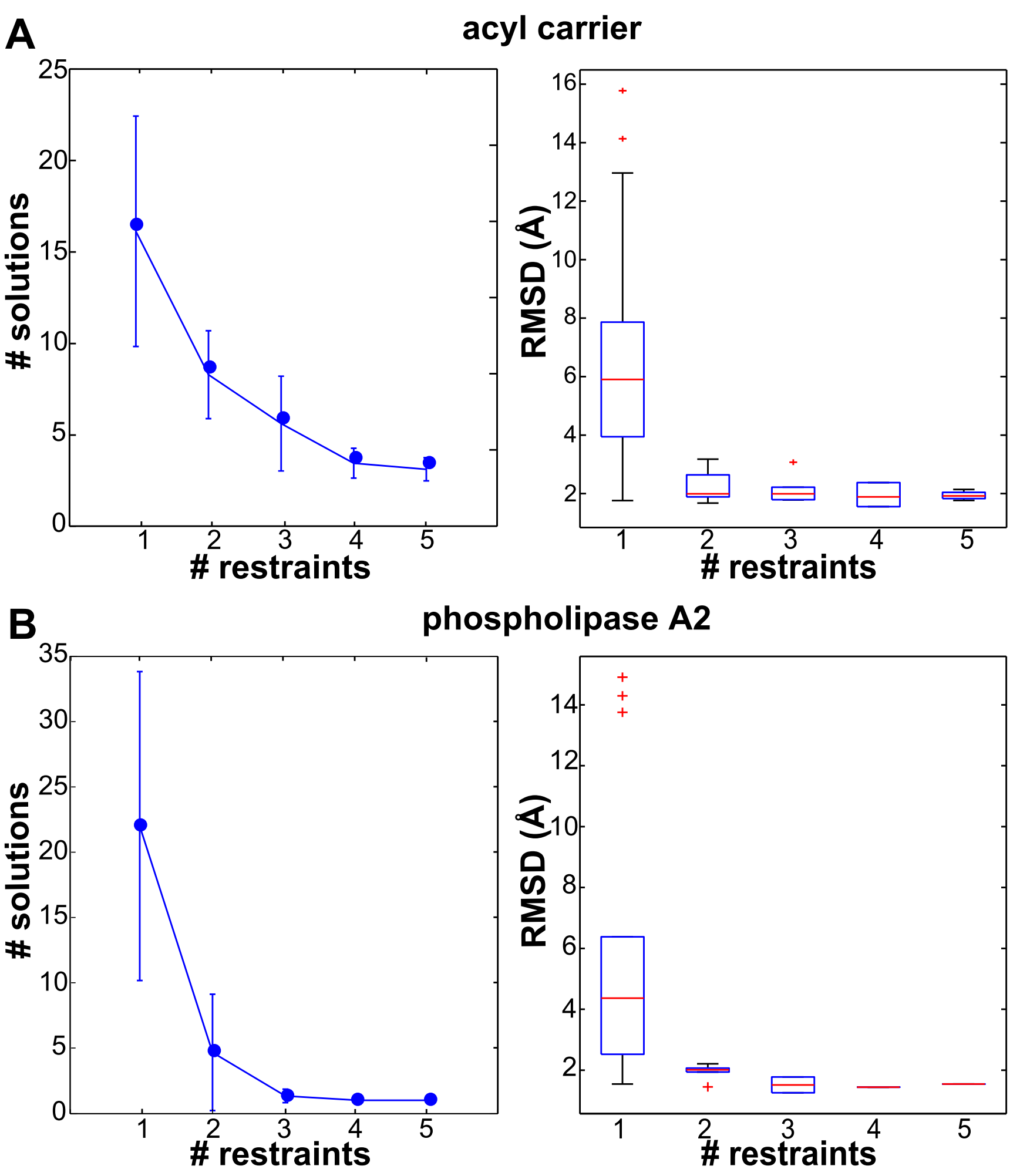
refine

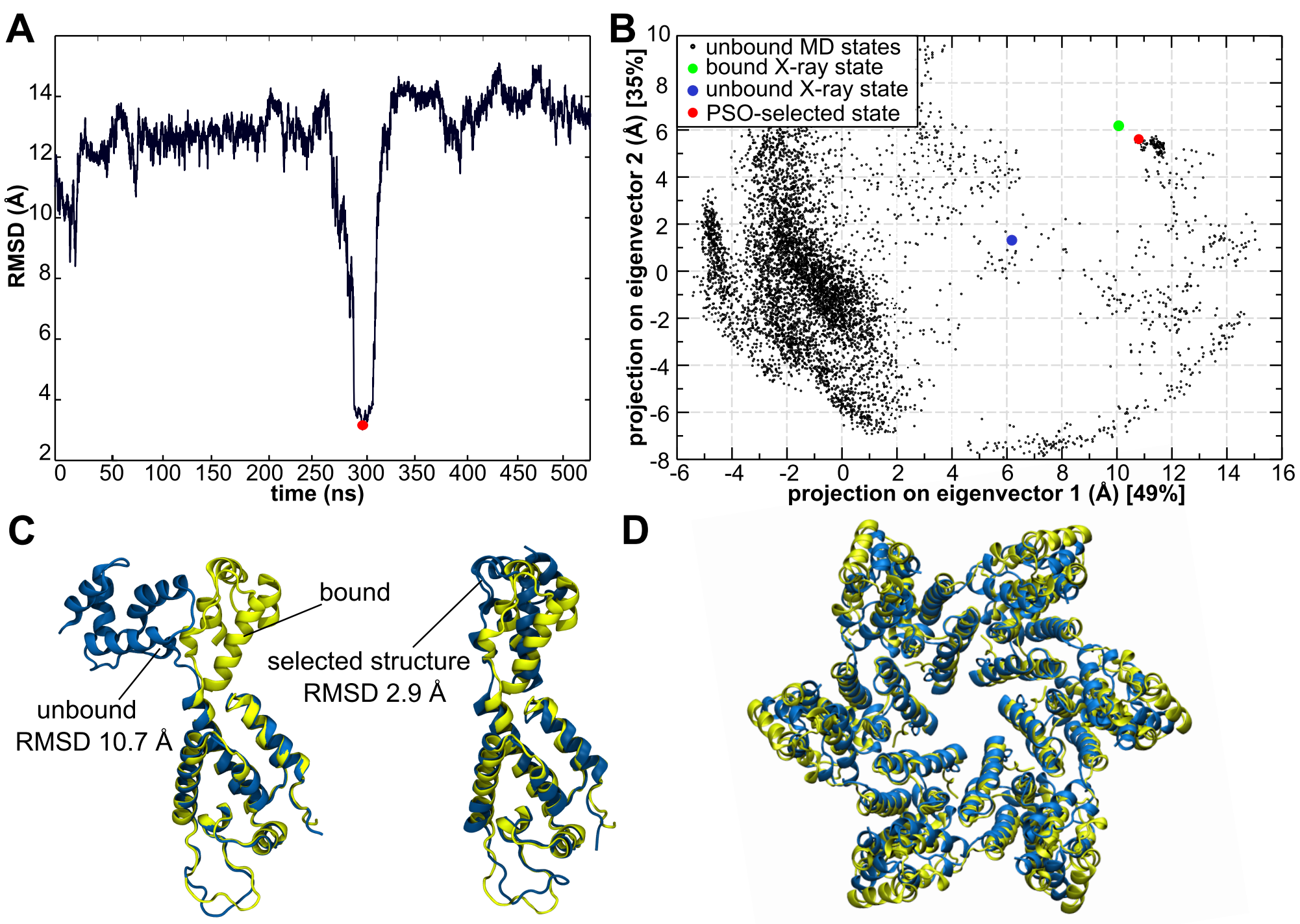
minimization, molecular dynamics...

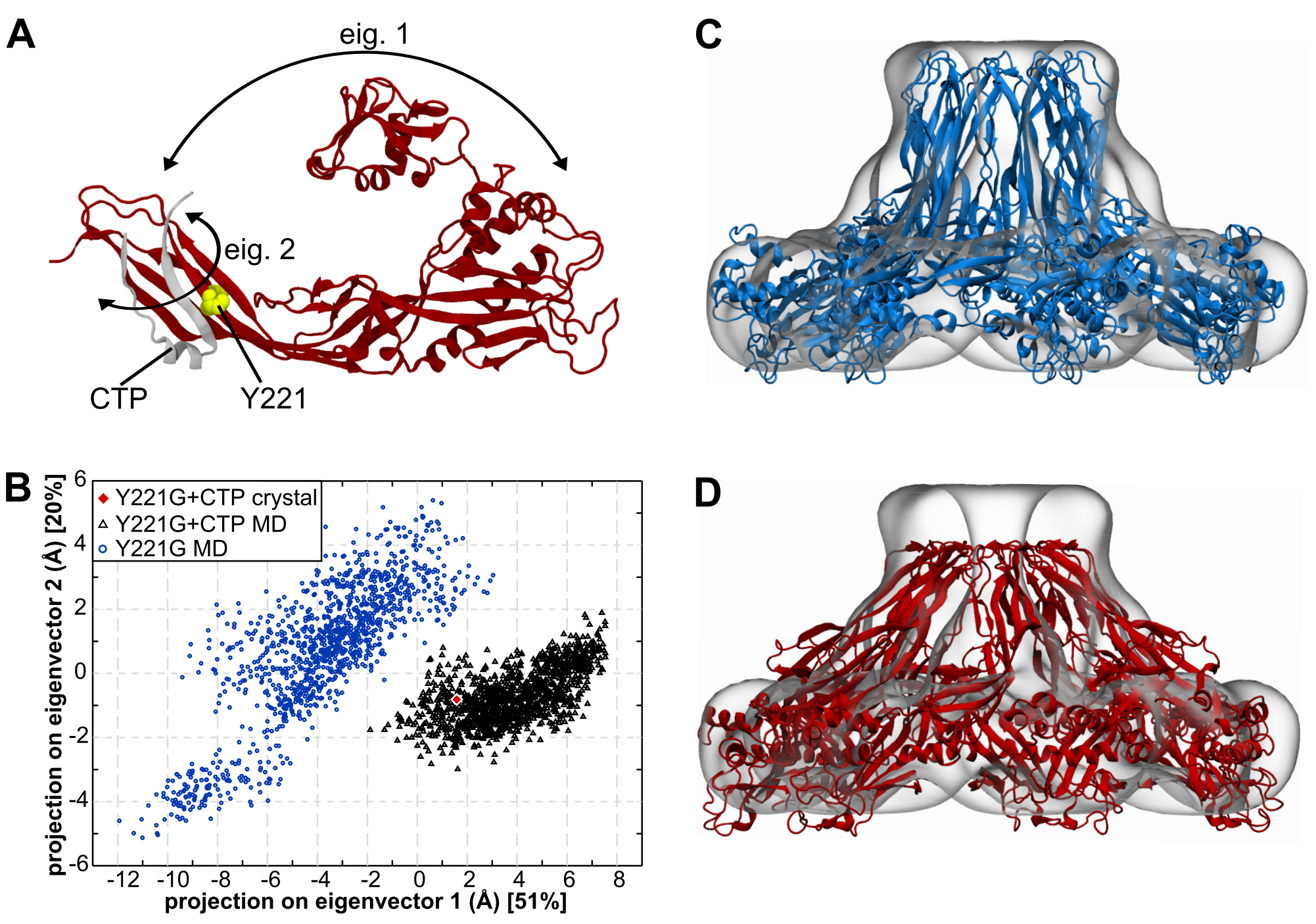


output: multimeric assembly


A
no symmetry
PhoQ
B
bound cases
Lumazine Synthase (C5)
C
SM archeal protein (C7)
D
EscJ (C24)
E
Phospholipase A2 (C3)
F
unbound cases
Envelope Glycoprotein (C3)


A
acyl carrier
solutions
restraints
RMSD (Å)
restraints
B
phospholipase A2
solutions
restraints
RMSD (Å)
restraints


A
RMSD (Å)
time (ns)
B
unbound MD states
bound X-ray state
unbound X-ray state
PSO-selected state
projection on eigenvector 2 (Å) [35%]
projection on eigenvector 1 (Å) [49%]
C
bound
selected structure
RMSD 2.9 Å
unbound
RMSD 10.7 Å
D


A
eig. 1
eig. 2
CTP
Y221
C
B
Y221G+CTP crystal
Y221G+CTP MD
Y221G MD
projection on eigenvector 2 (Å) [20%]
projection on eigenvector 1 (Å) [51%]
6
4
2
0
-2
-4
-6
-12
-10
-8
-6
-4
-2
0
2
4
6
8
D

```
for every timestep t do:
    for every particle p do:
        inertia ← w*v(p,t-1)
        personal ← cp*rand(0,1)*(x(p,t-1) - x_best(p))
        global ← cn*rand(0,1)*(x(p,t-1) – x'_best)
        v(c,t) ← inertia + personal + global
        if |v(p,t)| ≥ size(space) then
            v(p,t) ← norm(v(p,t))*size(space)
            x(p,t) ← x(p,t-1) + v(p,t)
        else if |v(p,t)| ≤ v_min and f(x(p,t)) ≥ f_min then
            v(p,t) ← rand(0,1)*v_min
            x(p,t) ← x(p,t-1) + v(t)
        else if |v(p,t)| ≤ v_min and f(x(p,t)) ≤ f_min then
            v(p,t) ← rand(0,1)*v_min
            x(p,t) ← rand(0,1)*space
        else
            x(p,t) ← x(p,t-1) + v(p,t)
        end if
        if f(x(p,t)) ≤ f_best(p) then
            f_best(p) ← f(x(p,t))
            x_best(p) ← x(p,t)
        end if
    end for
end for
```

SUPPLEMENTAL INFORMATION FOR:

Macromolecular symmetric assembly prediction using swarm intelligence dynamic modeling

Matteo T. Degiacomi [1#] and Matteo Dal Peraro [1*]

[1] Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne - EPFL, 1022 Lausanne, Switzerland

[#] Current address: Physical and Theoretical Chemistry Laboratory, South Parks Road, Oxford, OX1 3QZ

SUPPLEMENTAL FIGURES

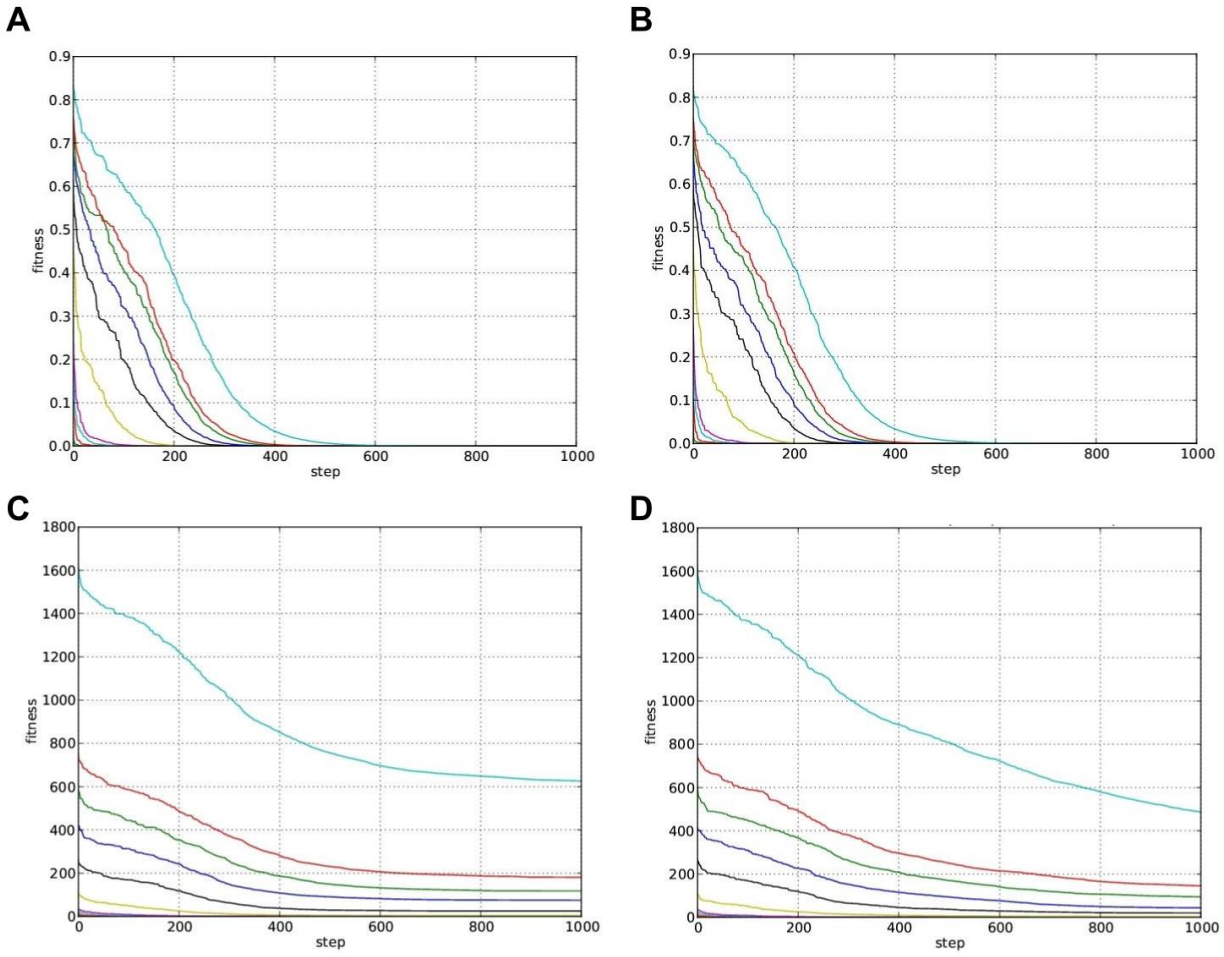


Figure S1, related to Figure 6: Results on multidimensional Sine (A, B) and Rastrigin (C, D) using PSO (A, C) and PSO-KaR (B, D). Every result is an average of 10 independent optimization runs. The more the function dimensionality increases, the more the fitness convergence is slow (cyan = 100D, red = 50D, green = 40D, blue = 30D). See Figure 6 of main text for the PSO-KaR pseudo-code.

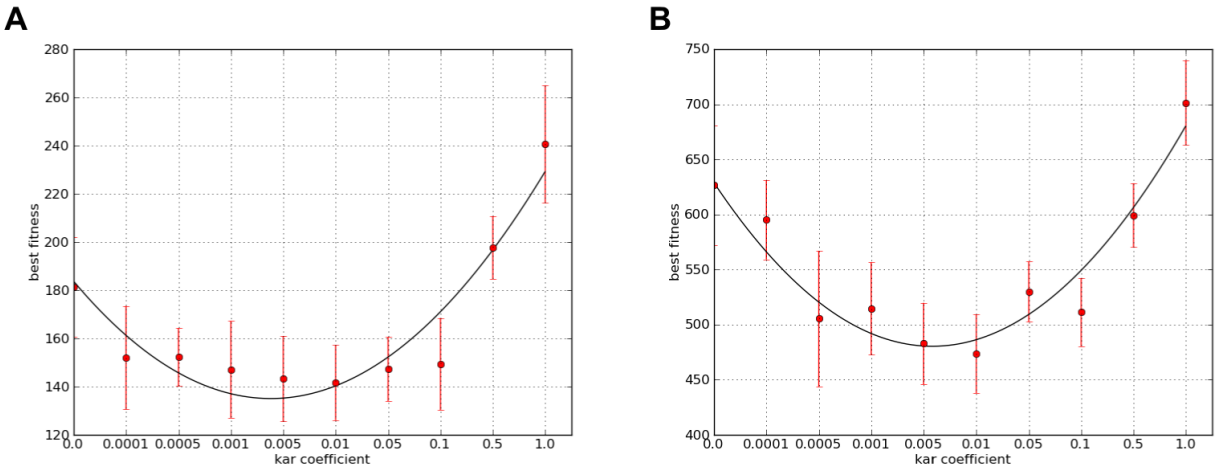


Figure S2, related to Figure 6: Results on multidimensional Rastrigin function using different velocity thresholds for PSO-KaR optimization. The final fitness achieved with different threshold values is shown for runs on the 100D Rastringin (A) and 50D Rastrigin (B). Error bars indicate the standard deviation over 10 runs. In both cases, a too high threshold introduces too much noise, and a too low one that does not perturb the swarm sufficiently. Best results are obtained, in both cases, with a velocity threshold equal to 0.01. See Figure 6 of main text for the PSO-KaR pseudo-code.

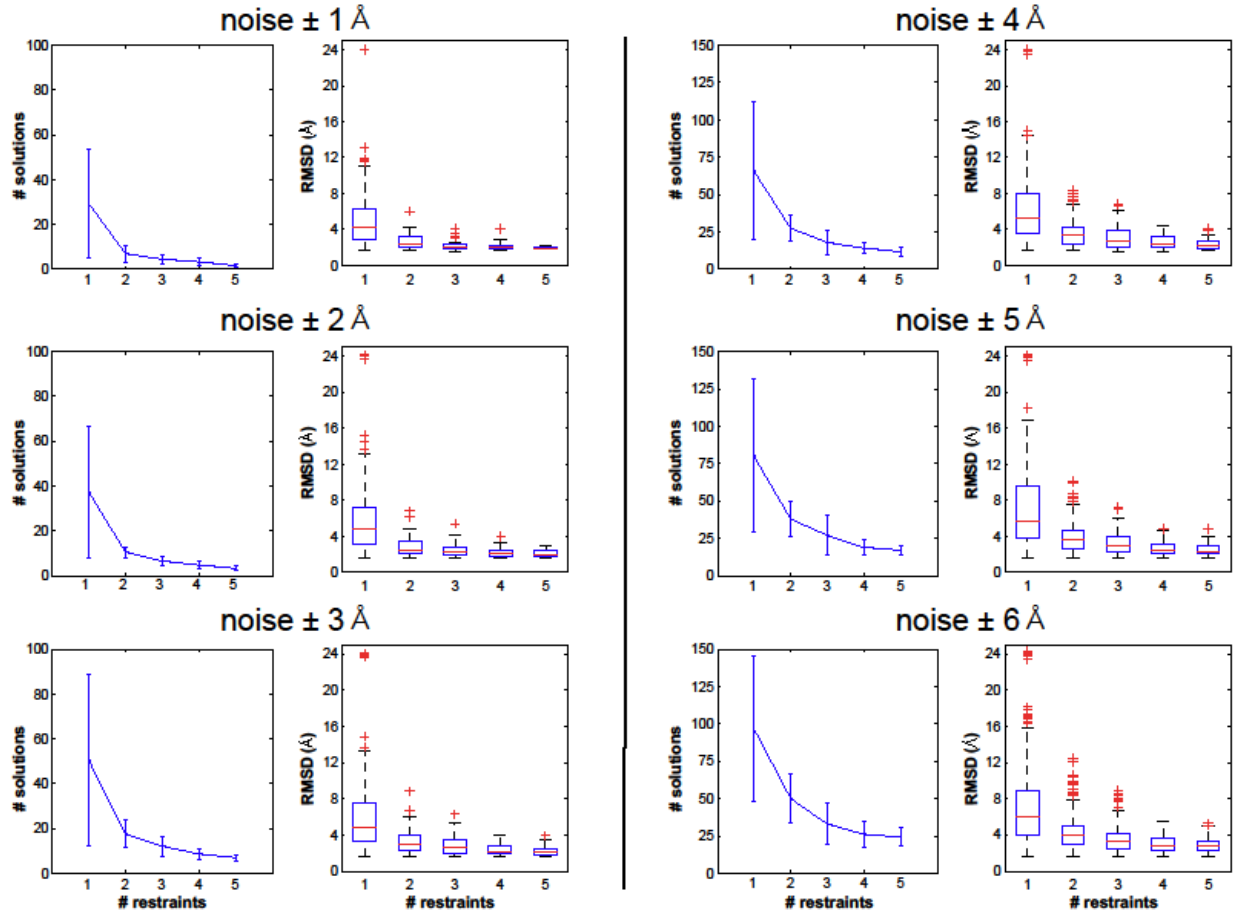


Figure S3, related to Figure 3: Performance on the prediction of multimeric conformation of Acyl Carrier using different quantities of distance restraints and imposing different errors on measurements. Left plots display the amount of final solutions; right plots the distribution of bRMSD [Å] of obtained solutions with respect of the known crystal structure. See Figure 3 of main text for a specific extract of these complete results.

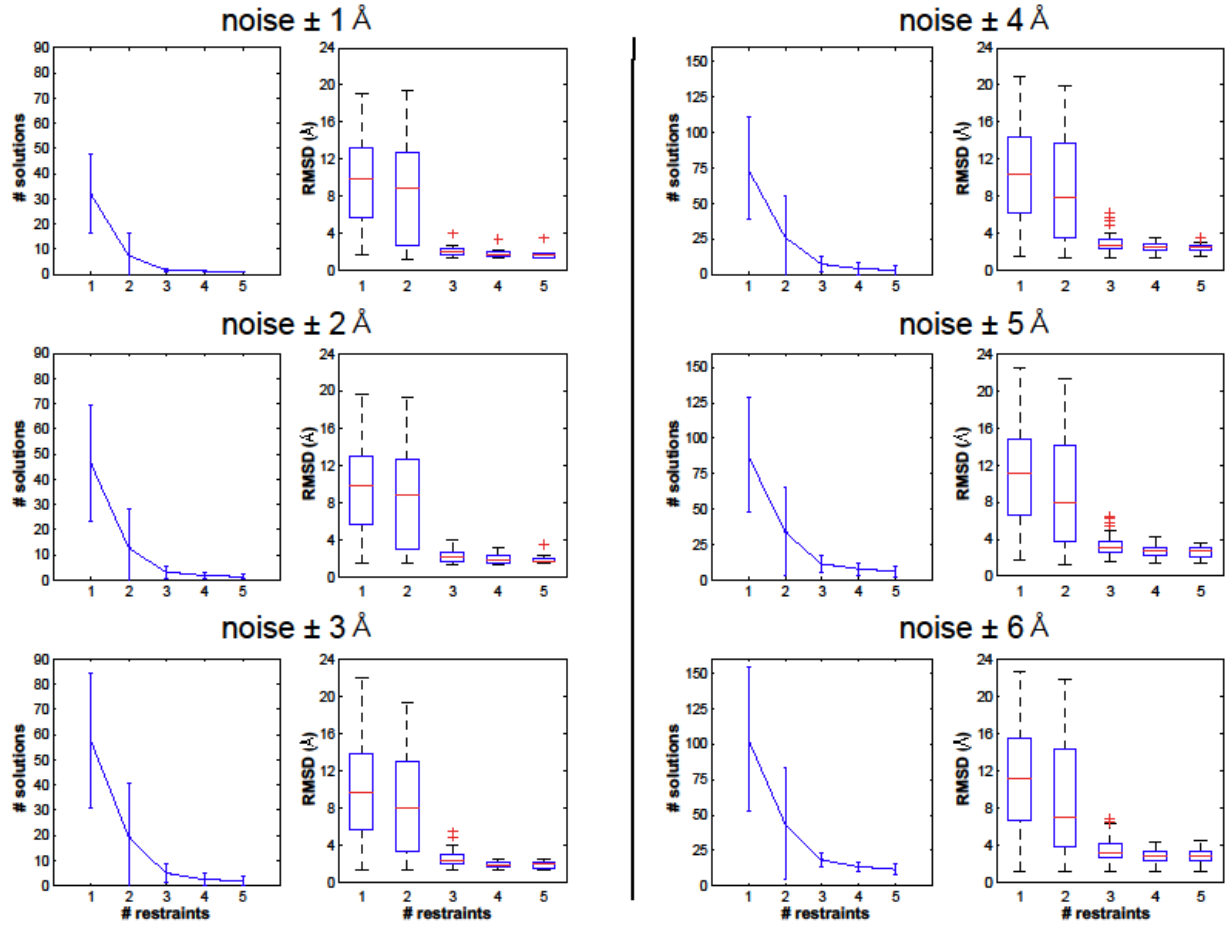


Figure S4, related to Figure 3: Performance on the prediction of multimeric conformation of phospholipase A2 using different quantities of distance restraints and imposing different errors. Left plots display the amount of final solutions, right plots the distribution of bRMSD [Å] of obtained solutions with respect of the known crystal structure. See Figure 3 of main text for a specific extract of these complete results.

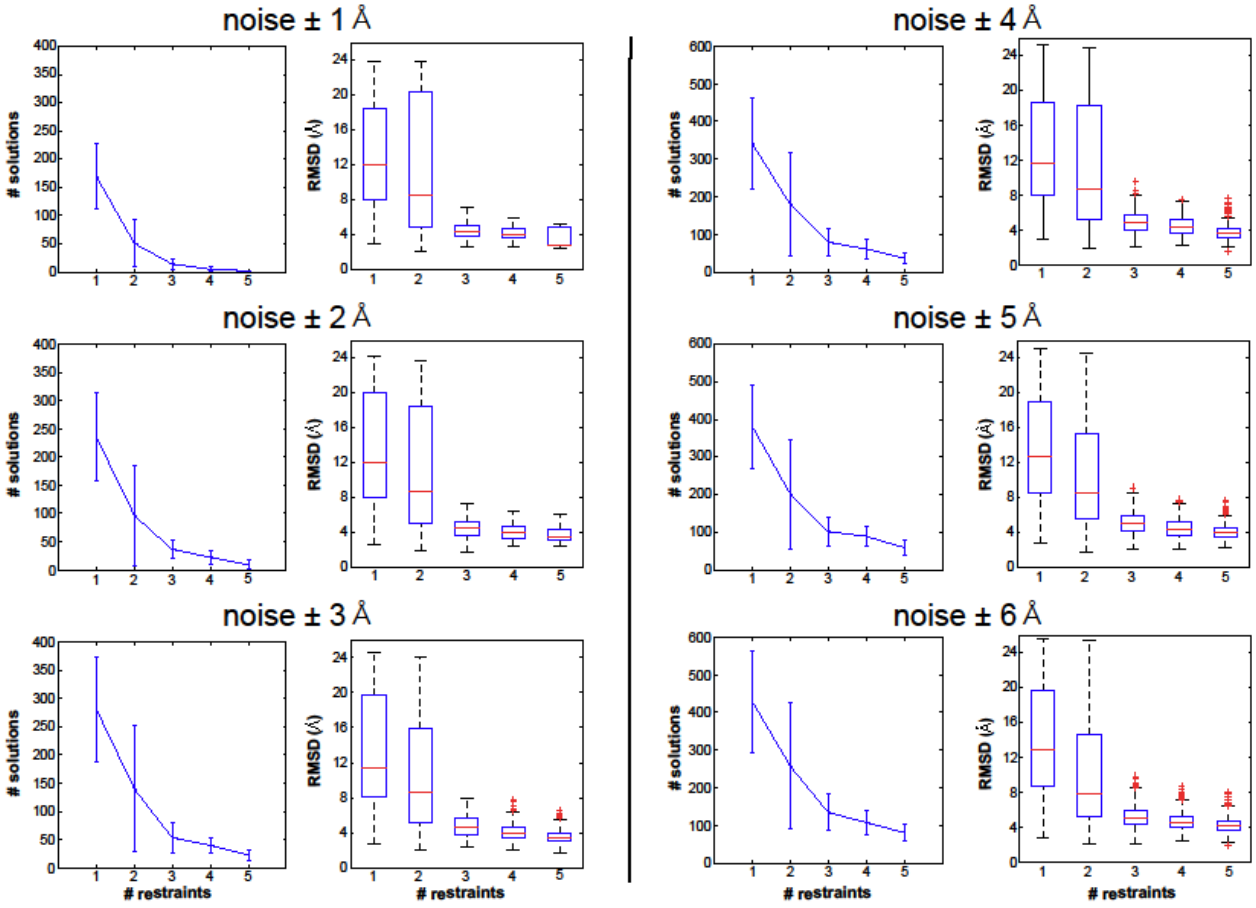


Figure S5, related to Figure 3: Performance on the prediction of multimeric conformation of phospholipase A2 using a ~500 ns MD simulation as input. Different quantities of distance restraints as well as errors on the measures are imposed. Left plots display the amount of final solutions, right plots the distribution of bRMSD [Å] of obtained solutions with respect of the known crystal structure. See Figure 3 of main text for a specific extract of these complete results.

SUPPLEMENTAL DATA

1. PSO and PSO-KaR benchmark

In order to assess the effect of the *Kick and Reseed* procedure (KaR) on PSO performance, and identify the most suitable value for the Kick threshold, we ran a set of function optimization tests. Two classic benchmark multidimentional functions were used: sine (used as a trivial test) and Rastrigin (used as an hard test). The sine function is defined as follows:

$f(\mathbf{x}\}=1+\sum_n \sin(x_n) / n$

In the interval [0,2π] one unique minima exists, f(3/2 π)=0. The Rastrigin function is defined as follows:

$f(\mathbf{x})=10+ \sum_n [x_n^2-10*\cos(2\pi x_n)]$

This function is particularly hard to optimize. In the interval [-5.12,5.12] it contains a large number of local minima, that get increasingly deep the closer they are to the unique global minima located in f(0)=0, 1, 2, 3, 5, 5, 10, 20, 30, 40, 50 and 100 dimensional Rastrigin and Sine function were submitted to PSO and PSO-KaR optimization. For PSO-KaR, the minimal velocity threshold was set to 0.01. For both PSO and PSO-KaR, every of these functions was optimized in 10 independent runs with 1000 steps and 80 particules. In both cases particles neighbors were defined as the particles having preceding and following indexes. The average fitness at every optimization step is shown in Figure S1.

The KaR procedure improves fitness convergence when hard fitness functions are optimized, whereas it has no effect when optimizing easy functions. We notice that, in every case, the more a function dimensionality increases, the more fitness convergence rate decreases. Increasing the size of the search space has the effect of reducing the particles density in it. Therefore, finding the funnel leading to the global minima, and the right trajectory to find its lowest point, becomes increasingly difficult. On Rastrigin function, early convergence is immediately observable when using PSO. This phenomena becomes more dramatic when dimensionality increases. This is due both to the increased number of almost equivalent local minima, and by the already mentioned reduced particles density in the search space.

For simple fitness functions (such as sine or low dimensional Rastrigin) no relevant difference can be observed within the PSO and PSO-KaR procedures. However, when increasing the fitness function complexity, the effect the KaR procedure become more and more relevant. This shows that, while being comparable to PSO for easy fitness functions, PSO-KaR prevents early convergence in hard optimization problems. To assess the effect of the *kick and reseed* procedure and define an ideal value for the kicking threshold, we tested different velocity thresholds while running PSO-KaR optimizations on the two hardest fitness function used above, i.e. the 50 and 100 dimensional Rastrigin functions. Values equal to 1, 0.5, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001 and 0 (equivalent to the standard PSO) were tested. Averages of 10 independent runs with 1000 optimization steps and 80 particles are visible in Figure S2.

For both the 50 and the 100 dimensional Rastringin, a threshold value equal to 1 turns out to be too high. High thresholds introduce excessive noise in the swarm, which in turn converges with more difficulty. Conversely, thresholds being too low (such as 0.0001), do not perturb the swarm sufficiently to have a real positive effect in its search process. The best result, leading to improved performances with respect of the standard PSO, was obtained in both tests by a velocity threshold equal to 0.01. This value was therefore set as default in PSO implementation. Importantly, the KaR procedure does not affect negatively the precision of search process. Indeed, despite the addition of noise in the search process, the standard deviation in PSO-KaR does not increase with respect of the standard PSO. By comparing the results of the 100 and 50 dimensional Rastrigin, we observe that in the second case a broader range of threshold values leads to similar performances. Previously, we also observed that for very simple functions (sine) the KaR procedure has no effect. This indicates that the more the fitness function becomes complex, the more a good choice of velocity threshold becomes important.

This algorithm is implemented in an open source code called *Parallel Optimization Workbench – pow*^{er}, which is available at http://lbm.epfl.ch.

2. Molecular Assembly Prediction Details

2.1 Structural Models Used in Test Cases

We applied our scheme for assembly prediction to a set of model systems, on which we tested the robustness and accuracy of our approach on bound and unbound relevant cases. For each system a heterogeneous set of restraints were used and led to the results shown in Table 1 and Figures 2 and 4. Moreover, extended tests were performed in order to assess the influence of the number and nature of restraints on the overall result (Figures 3, and S3-S5)

Tests on bound cases

Trimeric *Streptococcus pneumoniae* Acyl carrier protein synthase (AcpS) was crystallized at a resolution of 1.9 Å with 3'5'-ADP bound in two of its three active sites, located at protein's interfaces (pdb: 1FTH (Chirgadze et al., 2000)). Chain A was extracted as representative structure. We imposed the assembly radius to be 0±2 Å, a negative radius leading to multimers crossing the center of symmetry. We generated a restraint by identifying the residues constituting the protein's binding site. In AcpS, Asp10 and His105 bind the same phosphate group. For this reason, we set their distance to be 9±2 Å. Additionally, in order to assess the performance of our pipeline as a function of number of used constraints (Figures 3A and S5), combinations of six specific contacts (Asp10-His105, Ile4-Glu117, Lys64-Ile103, Ile11-Ser104, Ile11-His105, His7-Ser113) were adopted. Backbone RMSD was computed taking into account all residues common to all three chains, i.e. residues 3 to 68, 75 to 99 and 101 to 118.

Heptameric *Pyrobaculum aerophilum* SM Archeal Protein cationic pore was crystallized at a resolution of 1.75 Å (pdb: 1I8F (Mura et al., 2001)). We extracted chain A as a representative monomeric structure. We imposed the assembly's narrowest radius to be 4±2 Å, which is a reasonable size for a cationic channel. a single restraint was imposed: we imposed residue Asp29 to be exposed to the pore lumen, therefore imposing its distance from the pore central axis to be smaller than 6 Å (i.e. the largest possible pore radius). Lumen-exposed residues can be typically identified by detecting channel conductivity alteration upon cysteine mutation and addition of MTS probes in solution. Residues 15 to 80 of all chains (i.e. the residues common to all the seven subunits in the crystal) were selected to compute the backbone RMSD within the PSO-generated solutions and the known crystal.

Clostridium thermocellum Chorismate Mutase was crystallized in a trimeric conformation (pdb: 1XHO). Assembly radius was bound to be within 0 and 3 and the height (42±4 Å) and width (49±4 Å) of the final assembly were adopted as unique geometric restraints. Such a restraint that can be typically obtained by measuring low-resolution electron density maps obtained, for instance, via cryo-electron microscopy.

Pentameric Lumazine Synthase from *Saccharomyces cerevisiae* was crystallized at a resolution of 1.85 Å (pdb: 1EJB (Meining et al., 2000)). Since the first 16 N-Terminal residues show a different coil conformation, we excluded them from our test. After removal of the first 16 N-Terminal residues, we supposed that all the five monomers could be considered as identical. Chain A was selected as representative structure. In this test we combined four different restraints. Height and width were set to be equal to 75±2 and 49±2 Å respectively. Furthermore, we restrained the distance with two residues couples. The distance within two Asp103 facing each other was set to 3.8±2 Å, and the distance within Asp103 and neighboring His107 to 9.4±2 Å. Such measures are typically obtainable via cross-linking experiments. Assembly radius was set to 5±1 Å. Backbone RMSD was computed taking into account all the residues.

Escherichia coli 24mer EscJ, part of type III secretion system's basal body, had a 4-mer basic unit crystallized (pdb: 1YJ7 (Yip et al., 2005)). Chain A was extracted as representative structure. We set the assembly radius to be 38±2 Å and imposed three different restraints. We set assembly's width to be 176±2 Å, its height to 55±2 Å and residue Pro99 to face inwards (distance from the origin smaller than 40, the largest acceptable radius). From the obtained 24-mer predicted models we extracted a 4-mer basic unit which was compared to the available crystal. Backbone RMSD was computed with respect of the tetrameric basic unit taking into account all residues common to all three chains, i.e. residues 21 to 91, 98 to 133 and 141 to 186.

PhoQ is a two-component system histidine kinase responsible of detection of divalent cations at the inner bacterial membrane. Two X-ray structures of its periplasmic sensor domain (a.a. 45-186) are available: one from *Salmonella enterica* (pdb: 1YAX), and one from *Escherichia coli* (pdb: 3BQ8). The crystals being different, Goldberg et al. (Goldberg et al., 2008) determined via cross-linking experiments on the dimer's interface (α-helix 45-62) that the structure from *E.coli* shows the most physiologically relevant arrangement. From the dimeric crystal 3BQ8, two individual monomers were extracted and defined as ligand (chain A) and receptor (chain B) respectively, while five cross-linking measures

displaying high efficiency (Thr47, Leu51, Gly54, Asn57, Leu58, see Figure 2A in main text) were used as geometric restraints. Since cross-linking efficiency does not provide a high-resolution information about residues distance, we considered that the five most efficient cross-links simply indicate that residues' Cα atoms are 9±3 Å apart, consistently with disulfide bond distances. All structures were compared to the known crystal by computing their backbone RMSD on residues Phe44 to Val183. C-Terminal residues Glu184 to Ser188, being random coils, were excluded. 6 different solutions were finally produced, the best one having a backbone RMSD equal to 1.85 Å. Its interfacing residues, located on α-helix 46-62, had a backbone RMSD of 0.76 Å. Importantly, all the solutions respected the given geometric restraints and presented no steric clashes, possibly being representative conformations of alternative signaling states accessible to this sensor domain.

Tests on unbound cases

Phospholipase A2 from *Naja naja* was crystallized both in its monomeric (pdb: 1POA (Scott et al., 1990)) and trimeric (pdb: 1A3F (Segelke et al., 1998)) form. The RMSD within the unbound and bound conformation (residues 1 to 117) is 0.7 Å, and is mainly due to a slightly different arrangement of a loop. In order to perform a molecular dynamics (MD) simulation of the unbound state, we first solvated the structure in a rectangular box of pre-equilibrated TIP3P water, and neutralized the system's total charge by the addition of Na^+ and Cl^- ions. Molecular dynamics has been performed using the Amber parm99sb force field (Case et al., 2010) on NAMD molecular dynamics engine (Phillips et al., 2005), with SHAKE algorithm on all the bonds, and Particle-mesh Ewald treating the electrostatic interactions in periodic boundary conditions. We chose an integration step of 2 fs. The systems had their energy initially minimized by means of 1000 conjugate gradient steps, and have subsequently been gradually heated from 0 to 300 K in 1 ns at 1 Atm. Simulations were run in the nPT ensemble at 1 atm and 300K. Temperature has been controlled by mean of Langevin forces, using a damping constant of 1 ps^{-1}. The protein was simulated for 480 ns. From this simulation, one frame every 100 ps, i.e. 4800 frames, were extracted and used as conformational database for PSO optimization. In all tests, we set the assembly radius to be within 1 and 3 Å. In order to assess the performance of our pipeline as a function of number of used constraints, we selected six specific contacts (Lys6-Glu55, Tyr3-Tyr63, Tyr110-Lys115, Trp19-Gly31, Asn111-Lys115, Asn1-Trp61). These restraints were used for tests performed using the single unbound structure as

starting monomer, and the MD trajectory of unbound state described above (Figures 3E, S4 and S5).

Flavivirus envelope glycoprotein was crystallized both in its monomeric (pdb: 1SVB, (Rey et al., 1995)) and trimeric state (1URZ, (Bressanelli et al., 2004)). The protein is composed of a large flexible domain (residues 1 to 300) connected to a small globular domain. We simulated 1SVB with MD for 230 ns using the same protocol detailed above for Phospholipase A2. In our tests, we subsequently used the larger domain only (having an RMSD of 4.4 Å with respect of the bound state), and used as restraints two contacts: Asp10-His105 and His7-Ser113. For both restraints we set a target distance equal to 4±2 Å (Figure 2F). 2300 frames were then used as conformational database for our optimization protocol. In order to compute the obtained model's RMSD with respect of the known bound state, we considered residues 8 to 97, 110 to 146, 160 to 202, resid 210 to 247 and resid 252 to 296, i.e. all residues common to both structures, excluded long solvent exposed unstructured regions.

HIV-1 Hexameric Capsomer was crystallized both in its monomeric (pdb: 1E6J, complexed with FAB13B5 protein (Monaco-Malbet et al., 2000)) and trimeric (pdb: 3MGE (Pornillos et al., 2010)) form. The RMSD between the two states (residues 11 to 140, 144 to 176 and 189 to 219) is equal to 10.5 Å. The unbound protein, consisting of two rigid domains connected by a flexible linker, was extracted from1E6J crystal simulated for 500 ns using the same protocol detailed above for Phospholipase A2. One frame every 100 ps, i.e. 5000 structures, was extracted and used as conformational database for our optimization protocol. We set the assembly radius to be within 0 and 3 Å, and set one geometric constraint per domain, i.e. Glu212-Lys140 and the disulfide bridge Cys42-Cys51.

A crystal structure of dimeric proaerolysin mutant Y221G has been obtained with a resolution of 2.2 Å (pdb: 3C0N). A single monomer was extracted from the crystal. After 100 ns production run, we manually removed the 40 residues long C-Terminal peptide (CTP). We assumed that this would mimic the effect of proteolytic cleavage leading to protein activation. The newly obtained protein, active Y221G aerolysin, was equilibrated and simulated during 200 ns using the same simulation protocol detailed above. 2000 structures, one every 100 ps, were extracted from this latter trajectory, and submitted to PSO optimization as detailed in the main text.

3. Molecular assembly predictions

3.1. Effect of quantity and quality of restraints

Contacts selected in phospholipase A2 were Lys6-Glu55, Tyr3-Tyr63, Tyr110-Lys115, Trp19-Gly31, Asn111-Lys115, Asn1-Trp61. Contacts selected in acyl carrier were Asp10-His105, Ile4-Glu117, Lys64-Ile103, Ile11-Ser104, Ile11-His105, His7-Ser113.

We selected combinations of 1,2,3,4 and 5 contacts as restraint. Every test was repeated using six different combinations, and results collected in a unique dataset. To test the effect of error size on the target measure, all these tests were repeated by imposing errors of 1,2,3,4,5 and 6 Å on every target measure. Therefore, in this benchmark we ran a total of (3 proteins)*(5 contacts)*(6 combinations)*(6 errors) = 540 optimizations. Notice that benchmark data presented in main text (Figure 3A and 3B) are extracted from this same dataset (results obtained using an error equal to 2 A).

RMSD distributions in all the experiments have been represented using boxplots computed using Matlab and default parameters (Figures 3, S3, S4 and S5). In these plots the red lines represent medians, blue boxes the intervals between lower and upper quartiles, whiskers the lowest and highest observations, and red crosses the outliers.

3.2. Mixing different restraints

In this section we report results obtained, for representative proteins, by adopting different geometric restraints and using the same experimental setup stated in main text (see Material and Methods). Not surprisingly, the amount of obtained valid models usually increases the less stringent the used geometric restraints are, and *vice versa*. In general, the more precise (and reasonable) restraints are provided, the better the result (i.e. a smaller amount of possible models is produced). In the following, *w* stands for “width”, *h* stands for “height” and *d(r1,r2)* the Euclidean distance within two residues, *r1* and *r2*.

SM Archeal Protein. This data is related to Table 1 and Figure 2C in main text. Table S1 reports results using a variety of different geometric restraints. We tested the effect of making the same geometric condition less stringent, using a global measures (width and height), imposing a residue to face outside the complex (Pro80) or two residues in the pore center (Arg29 and Asp30) to be in contact.

Restraint	time	solutions	bRMSD
w=66±2 Å, h=40±2 Å	2m16s	2	1.52 Å
d(Arg29,(0,0,0))<10 Å, w=66±2 Å, h=40±2 Å	2m53s	3	1.63 Å
d(Arg29,(0,0,0))<10 Å	2m53s	15	1.28 Å
d(Arg29,Asp30)<4 Å	2m56s	27	1.52 Å

Table S1, related to Figure 3 and Table 1: Summary of PSO prediction results on SM Archeal Protein heptamer using several geometric restraints

Chorismate Mutase. This data is related to Table 1 in main text. We tested the effect of making the same global geometric condition more stringent, and also applied geometric restraints on specific atoms. On this aspect, we constrained the N-Terminal residues (Val2) as well as residue Met74 (in protein's core) to be close. We notice that, as soon as restraints become more stringent (for instance by restraining the distance within two atoms), the amount of produced multimers drops.

Restraint	time	solutions	bRMSD
w=49±2 Å, h=42±2 Å	2m20s	20	1.89 Å
w=49±2 Å, h=42±2 Å, d(Met74,Met74)=3.5±1 Å	2m20s	3	1.72 Å
w=49±4 Å, h=42±4 Å, d(Met74,Met74)=3.5±1 Å	2m26s	6	1.94 Å
d(Met74,Met74)=3.5±1 Å, d(Val2,Val2)=8±1 Å	1m50s	2	1.59 Å
w=49±3, h=42±3, d(Val2,Val2)=8±1 Å	1m55s	7	2.49 Å

Table S2, related to Figure 3 and Table 1: Summary of PSO prediction results on Chorismate Mutase trimer using several geometric restraints

Acyl Carrier. This data is related to Table 1 in main text Table S3 reports results using a variety of different geometric restraints. The influence of using global measures (width and height), other residues part of the binding site (i.e. Lys64 and Asp55) and various mixes of these quantities were tested.

Restraint	**time**	**solutions**	**bRMSD**
w=60±2 Å, h=47±2 Å	2m12s	9	1.80 Å
w=60±2 Å, h=47±2 Å, d(Asp10,His105)=9±2 Å	1m59s	3	1.82 Å
w=60±4 Å, h=47±4 Å, d(Asp10,His105)=9±4 Å	1m59s	8	1.98 Å
d(Lys64,Asp55)=10±4 Å	2m14s	8	1.86 Å
d(Lys64,Asp55)=10±4 Å, d(Asp55,His105)=9±4 Å	2m20s	6	1.95 Å

Table S3, related to Figure 3 and Table 1: Summary of PSO prediction results on Acyl Carrier trimer using several geometric restraints

EscJ. This data is related to Table 1 and Figure 2D in main text. Table S4 reports results using a variety of different geometric restraints. On this test case, we tested the effect of using a unique, stringent set of global measures (height and width). We also coupled such measures with residues specific restraints. In particular, we tested the effect of imposing residue Lys178 to face upwards, and to combine this restraint with others.

Restraint	**time**	**solutions**	**bRMSD**
w=176±2 Å, h=180±2 Å	6m52s	7	2.58 Å
Lys178 on top, d(Pro99,(0,0,0))<40	8m40s	10	2.06 Å

Table S4, related to Figure 3 and Table 1: Summary of PSO prediction results on EscJ using several geometric restraints

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

1. Kick-and-Reseed Particle Swarm Optimization (PSO-KaR)

Even though PSO usually displays a fast convergence and is robust to local minima, it has also been shown that when the fitness function's profile becomes extremely rough, the search might still terminate with a premature convergence (that is, with all particles stagnating in local minima (Angeline, 1998)). In order to increase PSO robustness and convergence rate, several approaches have been proposed (Abraham and Liu, 2009) (Pasupuleti and Battiti, 2006) (Binkley and Hagiwara, 2008) (Clerc, 1999) (Ratnaweera et al., 2004) (Clerc and Kennedy, 2002). We implemented a new PSO flavor that we call *kick and reseed* (PSO-KaR, box in Figure 6). This method is meant to avoid early convergence and leads to an increased sampling. To do so, particles velocities are constantly monitored. When a particle slows below a predefined threshold velocity $\mathbf{v}_{min}$, two possible actions are taken. If the current fitness value is above a predefined threshold f_{min}, the particle velocity is randomly reinitialized. Conversely, if the current fitness is below said threshold, the particle is also randomly reseeded in a new position of the search space, and its memory about its personal best erased. PSO-KaR has been tested on classic benchmark function, and proved to be more effective than standard PSO when dealing with rough fitness functions (see Figure S1-S2).

In PSO-KaR, *w*, *cp* and *cn* scale the influence of inertia, personal best and global best, respectively. *w* varies within 0 and 1. It has been shown that improved performance can be achieved by starting inertia at high values and gradually reducing it while optimization proceeds (Shi and Eberhart, 1998). A high inertia value produces a more "turbulent" swarm, which is ideal for an initial exploratory phase. Reducing the inertia has a "cooling" effect, which is more suitable when areas of interest have been discovered. While in literature values for initial and final *w* to 0.9 and 0.4 are usually accepted, values of *cn* and *cp* are subject to debate. In this context, meta-optimization approaches (optimization of parameters to improve PSO performance for a specific problem) have been proposed (Pedersen and Chipperfield, 2010) (Meissner et al., 2006).

In a typical PSO run, within 20 and 80 particles are initialized. A particle's neighborhood can be defined either as indexed, either as geographic. In the first case, an index is assigned to every particle, and particles having consecutive indices are

considered as neighbors. In geographic neighborhood, a particle will select as neighbors only the first *n* particles being close in the search space. Interestingly, it has been observed that a fully connected swarm would converge faster, but perform poorly than a partially connected one, being more sensitive to local minima (Kennedy, 1999) (Kennedy and Mendes, 2002). Search space boundary conditions can be enforced either as periodic or reflexive. In periodic boundary conditions, the unit cell containing the search space is replicated periodically in every direction. A particle leaving the unit cell will reappear in the neighboring cell. In reflexive boundary conditions, cell borders are considered as hard walls. Particles would therefore bounce elastically against them. An upper threshold for particles' velocity corresponding to the search space size is usually set.

All the tests were performed running 3 PSO repetitions of 200 steps with 80 particles. PSO behavioral constants w_{start} and w_{end} were set to 0.9 and 0.4, respectively. On the base of preliminary systematic tests on *cp* and *cn* constants, we identified *cp*=1.2 and *cn*=1.4 as the combination leading to the most effective performance in terms of space exploration and exploitation of areas having low fitness. Neighborhood was indexed and had size equal to 1, while and boundary conditions were periodic. 4 processors on an Intel AMD64 dual-quad core machine were used.

2. Fitness Function and Clustering

In preliminary systematic tests on Archeal SM and Acyl Carrier, we found that best results are obtained when setting as weighting factor c (see eq. 3 in main text) a value within 0.05 and 0.2 (not shown). When using smaller values, no good solution is usually found whereas, when using larger values, a large amount of solutions is returned. Importantly, for problems having a small amount of constraints, the contribution of the energy term turned out to be critical for good predictions.

All fitness evaluations obtained during PSO are collected, and solutions having a fitness lower than a predefined threshold are retained. In most applications the filtering criteria is set to 0. Such a value indicates that, most likely, the system's energy is negative (i.e. clash free) and geometric restraints are well respected. Since several solutions usually represent similar conformations, clustering is performed. Two *ad hoc* clustering approaches able to determine automatically the number of required clusters are available: the first groups solutions being close enough in the search space, whereas the second

clusters solutions generating assemblies having a small RMSD within themselves. We used here the latter approach. Cluster representatives are selected (cluster centers), ranked according to their fitness, and their corresponding assemblies returned as an ensemble of PDB files.

SUPPLEMENTAL REFERENCES

Bressanelli, S., Stiasny, K., Allison, S.L., Stura, E.A., Duquerroy, S., Lescar, J., Heinz, F.X., and Rey, F.A. (2004). Structure of a flavivirus envelope glycoprotein in its low-pH-induced membrane fusion conformation. EMBO J *23*, 728-738.

Case, D., Darden, T., Cheatham III, T., Simmerling, C., Wang, J., Duke, R., Luo, R., Walker, R., Zhang, W., Merz, K., and others (2010). AMBER 11. University of California, San Francisco.

Chirgadze, N.Y., Briggs, S.L., McAllister, K.A., Fischl, A.S., and Zhao, G. (2000). Crystal structure of Streptococcus pneumoniae acyl carrier protein synthase: an essential enzyme in bacterial fatty acid biosynthesis. The EMBO journal *19*, 5281-5287.

Goldberg, S.D., Soto, C.S., Waldburger, C.D., and DeGrado, W.F. (2008). Determination of the physiological dimer interface of the PhoQ sensor domain. Journal of molecular biology *379*, 656-665.

Meining, W., Mörtl, S., Fischer, M., Cushman, M., Bacher, A., and Ladenstein, R. (2000). The atomic structure of pentameric lumazine synthase from Saccharomyces cerevisiae at 1.85 \AA resolution reveals the binding mode of a phosphonate intermediate analogue1. Journal of molecular biology *299*, 181-197.

Monaco-Malbet, S., Berthet-Colominas, C., Novelli, A., Battai, N., Piga, N., Cheynet, V., Mallet, F., and Cusack, S. (2000). Mutual conformational adaptations in antigen and antibody upon complex formation between an Fab and HIV-1 capsid protein p24. Structure *8*, 1069-1077.

Mura, C., Cascio, D., Sawaya, M.R., and Eisenberg, D.S. (2001). The crystal structure of a heptameric archaeal Sm protein: Implications for the eukaryotic snRNP core. Proceedings of the National Academy of Sciences *98*, 5532.

Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kalé, L., and Schulten, K. (2005). Scalable molecular dynamics with NAMD. J Comput Chem *26*, 1781-1802.

Pornillos, O., Ganser-Pornillos, B.K., Banumathi, S., Hua, Y., and Yeager, M. (2010). Disulfide bond stabilization of the hexameric capsomer of human immunodeficiency virus. J Mol Biol *401*, 985-995.

Rey, F.A., Heinz, F.X., Mandl, C., Kunz, C., and Harrison, S.C. (1995). The envelope glycoprotein from tick-borne encephalitis virus at 2 A resolution. Nature *375*, 291-298.

Scott, D.L., White, S.P., Otwinowski, Z., Yuan, W., Gelb, M.H., and Sigler, P.B. (1990). Interfacial catalysis: the mechanism of phospholipase A2. Science *250*, 1541-1546.

Segelke, B.W., Nguyen, D., Chee, R., Xuong, N.H., and Dennis, E.A. (1998). Structures of two novel crystal forms of Naja naja naja phospholipase A2 lacking Ca2+ reveal trimeric packing. J Mol Biol *279*, 223-232.

Yip, C.K., Kimbrough, T.G., Felise, H.B., Vuckovic, M., Thomas, N.A., Pfuetzner, R.A., Frey, E.A., Finlay, B.B., Miller, S.I., and Strynadka, N.C.J. (2005). Structural characterization of the molecular platform for type III secretion system assembly. Nature *435*, 702-707.