

# Principles of assembly reveal a periodic table of protein complexes

Sebastian E. Ahnert\*<sup>1</sup>, Joseph A. Marsh\*<sup>2,3</sup>, Helena Hernández<sup>4</sup>, Carol V. Robinson<sup>4</sup>,  
and Sarah A. Teichmann<sup>1,3,5</sup>

<sup>1</sup>*Theory of Condensed Matter, Cavendish Laboratory, University of Cambridge, JJ Thomson Avenue, Cambridge CB3 0HE, United Kingdom*

<sup>2</sup>*MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, United Kingdom*

<sup>3</sup>*European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom*

<sup>4</sup>*Physical and Theoretical Chemistry Laboratory, Department of Chemistry, University of Oxford, South Parks Road, Oxford OX1 3QZ, United Kingdom*

<sup>5</sup>*Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom*

*\*These authors contributed equally to this work.*

## Abstract

Structural insights into protein complexes have had a broad impact on our understanding of biological function and evolution. Here we seek a comprehensive understanding of the general principles underlying quaternary structure organisation in protein complexes. To do this, we first examine the fundamental steps by which protein complexes can assemble using experimental and structure-based characterisation of assembly pathways. Most assembly transitions can be classified into three basic types, which can then be used to exhaustively enumerate a large set of possible quaternary structure topologies. These topologies, which include the vast majority of observed protein complex structures, give rise to a natural organisation into a periodic table. Based upon this, we are then able to accurately predict the expected frequencies of quaternary structure topologies, including those not yet observed. Overall, these results have important implications for quaternary structure prediction, modelling and engineering.

## Introduction

Evolution has given rise to an enormous variety of protein complexes (1–3). The organising principles that underlie this diversity remain poorly understood, particularly in comparison with protein folds, which have been classified extensively in terms of their architecture (4–6) and evolution (7, 8). However, network models have shown considerable promise for characterising and comparing protein complexes in recent years. For example, complexes are often represented as networks of associations between proteins, with little consideration for structure or stoichiometry. Alternatively, one can use a graph representation, which we introduced several years ago, that captures the main features of quaternary structure topology (9). In this model, the nodes are the polypeptide chains, defined by their amino acid sequence and often referred to as *subunits*, while the edges are the interfaces between physically interacting chains, weighted according to size.

Many protein complexes assemble spontaneously *via* ordered pathways *in vitro*, and we have shown that these assembly pathways have a strong tendency to be evolutionarily conserved (10, 11). Furthermore, there are striking similarities between protein complex assembly and evolutionary pathways, with assembly pathways often being reflective of evolutionary histories, and *vice versa* (12). In other words, quaternary structure evolution can essentially be thought of as an assembly process occurring on an evolutionary time scale. This suggests it may be useful to consider the types of protein complexes that have been formed in evolution from the perspective of assembly pathways.

Here we attempt to understand and explain the organisation of protein complexes in quaternary structure space using the principles of assembly. First, by characterising the assembly pathways of a large number of protein complexes, we find that assembly can be generally explained by three basic steps: dimerisation, cyclisation and subunit addition. Combinations of these steps allow us to exhaustively enumerate possible quaternary structure topologies within a given region of quaternary structure space.

To achieve this, we consider each polypeptide chain as a distinct self-assembly building block and consider all the ways in which interfaces can be distributed across the chains that are

present in the complex. The large variety of possible topologies that are generated by this procedure can then be compared to those structures that are actually observed. We find that ~92% of known protein complex structures are compatible with this model.

A major benefit of this assembly-centric view of protein complexes is that it gives rise to a natural organisation of complexes into a “periodic table”, by numbers of subunit repeats and numbers of subunit types. Exceptions are primarily the result of quaternary structure assignment errors or cases where sequence-identical subunits can form different interactions and thus introduce asymmetry. Many of these asymmetric complexes also fit the paradigm of a periodic table when their assembly role (rather than their subunit identity) is considered.

Finally, by combining the periodic table with our enumeration, we introduce a model to predict the expected frequencies of different quaternary structure topologies. Not only does this show an excellent correspondence with known protein complex structures, it also predicts new topologies most likely to be observed in the future.

## **A survey of transitions in the assembly pathways of protein complexes**

If we wish to understand the principles that underlie quaternary structure organisation, it is useful to begin by considering the different ways in which protein complexes can assemble. We therefore first seek to determine the assembly and disassembly, *i.e.* (dis)assembly, pathways for as many protein complexes as possible. Previously, we used electrospray mass spectrometry to characterise the (dis)assembly of 8 homomers (10) and 8 heteromers (11, 13). While the homomers followed quite simple pathways, more diversity was observed for the heteromeric complexes. For this reason, here we experimentally characterise the (dis)assembly pathways of 9 additional heteromers with widely varying quaternary structures (Fig. 1). In all of these cases, well-defined intermediate subcomplexes could be identified under at least one set of experimental conditions. All of the 8 homomers and 15/17 heteromers characterised by electrospray mass spectrometry to date have stoichiometries under native conditions consistent with the published biological units in the Protein Data Bank (PDB).

We also searched the literature for protein complexes of known structure where experimental (dis)assembly data is available, as we have done previously (10, 11). Often, these are cases where at least two different oligomeric states have been observed under equilibrium conditions. In total, we identified 11 homomers and 13 heteromers for which some (dis)assembly information is available in the literature.

Finally, we attempted to obtain further information on protein assembly by considering the large number of protein complexes of known structure. We searched for pairs of protein complexes where the quaternary structure of one complex could be described as a subset of the other. This includes, for example, a homodimer and a homotetramer with highly similar or identical sequences, suggesting that the tetramer assembles *via* a dimeric intermediate. This also includes homomer:heteromer pairs, where the heteromer has acquired a subunit with respect to the homomers. In total, this provided 154 homomers and 263 heteromers with putative structure-based assembly information.

Importantly, we recognise that the structure-based pathways do not represent direct characterisation of assembly. Instead, they indicate that two or more different quaternary structure states have been observed, and we assume that assembly transitions can occur between them. It is worth noting that even for biophysically characterised assembly pathways, we do not always have evidence that they are physiologically relevant. However, the fact that the biophysical and structure-based pathways have a strong tendency to reflect evolutionary history (10) and be evolutionarily conserved (11) does suggest that they have a functional relevance.

Given this large set of assembly data, we next asked what quaternary structure transitions (*i.e.* assembly steps) tend to be observed. For homomeric complexes, we can classify all possible transitions into three different types (Fig. 2A, left). First, there is *dimerisation*, where a doubling of the complex occurs and a twofold axis of rotational symmetry is formed (*e.g.* monomer-to-dimer or dimer-to-tetramer). Second, there is *cyclisation*, which involves the assembly of a ring-like quaternary structure with higher-order rotational symmetry (*e.g.* monomer-to-trimer or monomer-to-tetramer). Finally, there is *fractional transition*, an inherently asymmetric step in which the quaternary structure changes by a non-integer ratio (*e.g.* dimer-to-trimer or trimer-to-tetramer).

For each homomer with assembly data, we identified all the assembly steps that could account for the transitions between the free monomers, the observed subcomplexes and the full complex (see Methods). The distributions of these three different assembly steps are shown in Fig. 2B. All three datasets show a similar trend, with dimerisation being the most common step, cyclisation being the next most common, and fractional transitions being quite rare. This is consistent with previous observations of the favourable assembly and evolutionary transitions between homomers with different symmetries (10).

In heteromers, there are two further assembly steps that are possible, in addition to the three steps observed for homomers. These are illustrated in Fig. 2A (right): *subunit addition*, in which a new subunit is acquired (*e.g.* monomer-to-heterodimer), and *non-stoichiometric transition*, in which, the types of subunits within the heteromer remain the same, but the relative ratios between them change (*e.g.* assembly from 1:1 to 2:1 stoichiometry).

The distributions of all five possible assembly steps for heteromers are shown in Fig. 2C. The same basic trend is observed for the homomeric steps, with dimerisation being the most common and few fractional transitions. However, the most common observed step for heteromers from all three datasets is heteromeric subunit addition.

Within the heteromers, there is an interesting difference between the transitions observed in the mass spectrometry data and those observed in the other datasets. Specifically, non-stoichiometric transitions are much more common with mass spectrometry, evident from the considerable number of subcomplex intermediates with uneven stoichiometry (*i.e.* different numbers of each subunit type) observed in Fig. 1. This can be attributed due to two factors: the sensitivity of the mass spectrometry measurements to low-populated assembly intermediates, and the way in which the mass spectrometry experiments are performed, over a range of destabilising solution conditions designed to progressively disrupt the quaternary structure of the complex. We know that such non-stoichiometric transitions must occur in many cases where

they are not observed. For example, consider the transition from an  $\alpha\alpha$  homodimer to a  $\beta\alpha\alpha\beta$  heterotetramer, where there is no interaction between the two  $\beta$  subunits. In this case, an  $\alpha\alpha\beta$  assembly intermediate should form, as it is highly improbable that two separate  $\beta$  subunits would bind simultaneously. However, this asymmetric subcomplex is unlikely to be observed under non-destabilising conditions and without highly sensitive mass spectrometry measurements.

## Enumeration of the topological space of protein complexes

Next, we seek to explore quaternary structure space by combining different assembly steps to determine which protein complex topologies are possible. Given that the protein complex assembly pathways from above are dominated by dimerisation, cyclisation and subunit addition, we have focused on these three steps.

An important consideration here is interface symmetry. Dimerisation results in a twofold axis of rotational symmetry, and therefore the interface formed by dimerisation will be isologous (*i.e.* symmetric or head-to-head) and involve two identical surfaces on subunits of the same type (14). In contrast, cyclisation results in higher-order rotational symmetry and is associated with interfaces that are heterologous (*i.e.* asymmetric or head-to-tail) and involve two different surfaces on the same type of subunit. Finally, there are heteromeric interfaces, formed between two distinct polypeptide chains, and hence by definition also heterologous.

Proteins are inherently asymmetric at the level of individual polypeptide chains, so we can make the assumption that the same interface surface cannot appear twice on the same protein, or on two structurally different proteins. Together with this fundamental assumption, the three transitions (dimerisation, cyclisation, and subunit addition) all lead to symmetric protein complexes with even subunit stoichiometry. This is because we can view subunit addition as the formation of a larger multi-protein 'subunit', or protomer, which means that we can extend the homomeric definitions of dimerisation and cyclisation to homomers formed of these multi-protein subunits, leading to equal multiples of each type of protein (see Fig. 3).

Every homomeric complex (of single-protein or multi-protein subunits) can have at most two isologous or heterologous interfaces, as each new homomeric interface imposes a new axis of rotational symmetry. In other words, symmetry constrains the number of homomeric interface types to a maximum of two. One or two interfaces of two possible types give us five scenarios: a) one isologous, b) one heterologous, c) two isologous, d) two heterologous, and e) one isologous and one heterologous.

In order to elucidate all possible heteromeric topologies that can arise under these constraints, we start by enumerating all topologies of  $s$  distinct subunits (see Fig. 3). For this enumeration, we represent heteromeric topologies as trees rather than all possible graphs, because we wish to distinguish between essential and circumstantial interfaces in the complex. We do this by removing all interfaces that do not break apart the complex in order of size (see next section and Methods for details). For each of the five scenarios described as a) to e) above, we then consider all topologically distinct ways (that is, distinct under symmetry operations on the tree) in which the interfaces can be distributed across the set of subunits and pairs of subunits of the tree. The final step is to construct the topologies of the complexes from these distributions of

interfaces across the tree. Some of these are isomorphic (taking into account interface types and subunit identities), which reduces the overall number of topologies.

An important difference between our idealised model and real protein complexes is that real complexes can have more interfaces. However, we can directly relate real protein topologies to the above idealised forms, if we consider some of the weaker intersubunit contacts as circumstantial. In other words, stronger interfaces - one could call them 'essential' - exist that bind the complex together by themselves. We can distinguish between the essential and circumstantial interfaces by successively cutting away as many interfaces as possible in increasing order of size without giving rise to disconnected components of the complex (see Methods for details), leading to the simplest possible graph representation of a quaternary structure topology. This contrasts with the previous approach used in 3D Complex (9), where all intersubunit interfaces are considered. Thus this representation effectively sits above the more detailed classification of 3D Complex: a single, simplified topology used here can correspond to multiple 3D Complex topologies.

The vast majority of real protein complexes are compatible with our model: 92.5% of homomers and 91.7% of heteromers have topologies identified in our exhaustive enumeration (Fig. 4). In these complexes, structurally identical proteins inhabit the same *topological environment*, meaning the same local environment in terms of the interfaces they form with other subunits in a complex. We therefore define as *bijjective* complexes those having a one-to-one (*i.e.* bijjective) correspondence between their polypeptide sequence and their topological environment.

By contrast, all of the real protein complexes not compatible with our enumeration are *non-bijjective*, *i.e.* sequence-identical subunits exist in non-equivalent topological environments (Fig. 4). The difference between bijjective and non-bijjective complexes is further illustrated in Fig. S1.

In contrast to our simple enumeration model that requires only three types of assembly steps, non-bijjective complexes would require other asymmetric fractional transition and non-stoichiometric transition assembly steps. To explore this, we performed an exhaustive enumeration of all possible bijjective and non-bijjective topologies for complexes with specific stoichiometries. We find that for complexes with 2:2 stoichiometry, there are two possible bijjective topologies compared to seven possible non-bijjective topologies (Fig. S2). For complexes with 3:3 stoichiometry, there are also two possible bijjective topologies, while the number of possible non-bijjective topologies explodes to 250 (Fig. S3). This illustrates a major benefit of our approach: by limiting our model to only three simple assembly steps, we are able to cover the vast majority of observed protein complexes with a much smaller set of possible quaternary structure topologies.

To further justify our classification into bijjective and non-bijjective complexes, we utilised the fact that the quaternary structure assigned to a protein complex is often incorrect and does not represent the quaternary structure in solution or within the cell (15, 16). Using a database of manually confirmed quaternary structure assignments (17), complemented by additional manual assignments of our own, we compared error rates for bijjective and non-bijjective homomers and heteromers (Fig. 4). Strikingly, we found that, while bijjective complexes have a low rate of quaternary structure error (~10%), more than half of the non-bijjective structures are the result of errors. Thus, most non-bijjective protein complex structures do not represent genuine examples

of biological asymmetry, but instead are due to artefacts or errors in the structure determination process. This also suggests that the fact that a protein complex is non-bijective could be very useful for identifying likely quaternary structure assignment errors.

An exception to the above is the non-bijective complexes with uneven stoichiometry, which have only a 20% quaternary structure error rate. We recently studied these in detail and elucidated several different structural mechanisms by which they can form, which include varying degrees of pseudosymmetry, steric occlusion, and subunit flexibility leading to conformational differences between identical subunits (18). Notably, we found that those for which a structural mechanism for uneven stoichiometry could not be ascertained were mostly the result of quaternary structure assignment errors (18).

## A periodic table of protein complexes

Analysis of the real and enumerated quaternary structure topologies above shows that all bijective heteromers can be related to simpler homomeric topologies. Specifically, if the different subunit types are grouped together as a protomer, then the interactions between protomers will be equivalent to a bijective homomer topology, or to a monomer for cases with no subunit repeats. This suggests a natural classification for protein complexes: by their equivalent homomeric topologies in one dimension, and by the number of unique subunits in the other. Fig. 5 illustrates this “Periodic Table of Protein Complexes” for all topologies with  $\leq 12$  repeats and  $\leq 4$  unique subunits. In this classification, complexes related to the equivalent homomers are contained in the same column of the table, thus allowing the similarities between different heteromeric complexes to be easily recognized.

Most symmetry groups are associated with a single homomeric topology, except dihedral and cubic groups with  $\geq 6$  subunits, which have two topologies each. While this graph representation incorporates similarities between binding surfaces based upon the identities of interacting residues, it does not require any non-local geometric information. Our graph model therefore inherently includes the symmetry group information of a complex, and is to our knowledge the first network representation of complexes that does this.

Fig. S4 shows the frequencies of each equivalent homomer symmetry group for complexes with varying numbers of unique subunits. Interestingly, complexes with different numbers of unique subunits show very similar distributions. Thus homomers and heteromers populate the horizontal axis of the periodic table in a very similar manner, although complexes with more unique subunits do tend to have fewer repeats.

The regions of the periodic table that correspond to higher numbers of repeats and subunits are sparsely populated. This can be attributed to two factors. First, there is a significant bias amongst structurally characterised protein complexes towards those with fewer numbers of unique subunits, whereas evidence suggests that protein complexes *in vivo* will tend to have more distinct components (19, 20). Second, as can be seen in Fig. S4, there is a clear tendency for topologies towards the right of the periodic table to be less common, which suggests that cyclic or dihedral complexes with more repeated subunits may be less stable or more difficult to evolve. These regions can also be expected to be filled in coming years, at least to a certain extent. Fig. 5A shows the rate at which new topologies have been discovered: roughly four per

year for the last 20 years with no signs of slowing. In order to illustrate the space of possible topologies, the number of discovered topologies versus the number possible determined through exhaustive enumeration is shown in each cell of the periodic table (see example in Fig. 5B).

We note that this table is not “periodic” in the same sense as the periodic table of the elements, as it is in principle open-ended, as opposed to periodic with respect to atomic number. There are no theoretical limitations to quaternary structure topology space in either dimension, although the vast majority of known structures can be placed on the table in Fig. 5. In Fig. S5, we have provided an expanded version of the periodic table, where complexes with up to 14 unique subunits and 48 subunit repeats can be visualised. Overall, we believe that the analogy to the periodic table of the elements is useful, as it provides a means of organising quaternary structure topologies and visualising similarities. Furthermore, just as the periodic table of the elements has successfully predicted many new chemical elements, our periodic table of proteins has considerable predictive power by elucidating the regions of quaternary structure space that remain to be populated.

We showed above that the majority of non-bijective complexes are the result of quaternary structure assignment errors. The exception to this is complexes with uneven stoichiometry, most of which do represent genuine cases of biological asymmetry. Therefore, we sought to reconcile uneven stoichiometry with our model of the periodic table. Interestingly, we find that if we consider the periodic table at the level of local topological environments, rather than at the level of subunits, then two sequence-identical subunits can play different roles within the graph representing the complex. Examination of the topologies of non-bijective complexes revealed that many of them were equivalent to the same symmetric homomer topologies observed for the bijective periodic table. Fig. S6 illustrates this with a periodic table made for non-bijective heteromers with 2:1 subunit stoichiometry. For these cases, the 2:1 protomer can be considered analogous to a heterotrimer with three unique subunits. The only difference between 2:1 heteromers here and 1:1:1 heteromers from the main periodic table (*i.e.* the third row in Fig. 5) is that sequence-identical subunits can still sometimes form isologous interfaces, despite existing in different local environments. Thus, the results of our quaternary structure enumeration can be easily applied to complexes with uneven stoichiometry if the repeated subunits from the protomer are simply considered to be different subunit types in our enumeration model.

## **Predicting likely yet unobserved quaternary structure topologies**

The exhaustive enumeration allows us to determine what quaternary structure topologies are possible, but it does not tell us which are most likely or should be most abundant in nature. To address this, we adapted our enumeration procedure in order to produce topologies according to the observed distribution in the periodic table. Essentially, we know that each cell on the periodic table can be defined by a specific set of assembly steps needed to build the topologies within that cell. Combining the steps in different ways can produce all the topologies compatible with a given cell. Therefore, we sampled cells of the periodic table according to the observed distribution in real complexes, each time randomly combining the assembly steps associated with each cell. This was repeated  $3 \times 10^7$  times, with full details provided in the Methods.



All of the quaternary structure topologies present on the periodic table were observed at least once in our calculations. Furthermore, in addition to the previously observed quaternary structure topologies, our model also predicted 579 topologies that were not seen in any of the complexes in our dataset. To independently validate this result, we compiled an extended set of heteromeric complexes not present in our original dataset because they were published more recently, determined with electron microscopy (which we did not initially include), or were originally excluded due to structural criteria (see Methods).

The extended set of heteromers contained 53 different quaternary structure topologies, 14 of which were not present in the main dataset. These 14 have a striking tendency to be among the most highly predicted topologies in our model. For example, six of them were observed amongst the top 20 most likely predicted topologies (Fig. 6), out of a total of 579 predicted ( $P$ -value:  $2 \times 10^{-6}$ , Fisher's exact test). Fig. S7 illustrates how the observed topologies cluster within the most highly predicted rankings, thus supporting the predictive utility of our model.

We also employed a complementary approach for the prediction of the relative abundances of topologies within a given cell, which makes fewer assumptions, but also yields less specific predictions. We consider the number of distributed interfaces (that is, single interfaces that are spread across two subunits), and the number of topological equivalents (marked by red crosses in Fig. S8) of a given interface distribution. If we compare topologies pairwise within the cells of the periodic table with  $\leq 4$  unique subunits and  $\leq 12$  subunit repeats, and count the instances in which topology A has fewer distributed interfaces and more or equal topological equivalents than topology B, or fewer or equal distributed interfaces and more topological equivalents, then we observe that out of the 30 such instances, 21 (70%) times topology A is more abundant than topology B. This is because distributed interfaces restrict the order in which evolutionary steps can happen, making topologies with more such interfaces rarer. Larger numbers of topological equivalents on the other hand make topologies more common as there are more ways in which such complexes can evolve.

Finally, to further validate our predictive model, we compared the predicted frequencies of heteromeric quaternary structure topologies to those observed in both the main and extended datasets (Fig. S9). Overall, the correlations are very good, with the predictions nicely recapitulating the observed frequencies. Although the predictions are partially fit to the frequencies of topologies observed in the main dataset, the high correlation with the extended dataset provides strong independent validation of our model.

## Conclusion

In this study we have shown that the assembly of protein complexes is dominated by three main transition types, which in combination can explain most observed quaternary structure topologies. This also leads to a natural organisation of protein complexes in the form of a periodic table, in which heteromeric protein complexes are grouped according to their equivalent homomeric quaternary structure topologies. The periodic table illustrates both the variety of observed protein complexes as well as the space of possible topologies through exhaustive enumeration, analogous to previous strategies investigating network topologies (21, 22). Given that new topologies have been discovered at a fairly constant rate of four per year over the past

two decades, we can expect new additions to the unfilled or partially filled cells of the periodic table in the near future. These unfilled or partially filled cells constrain the total space of expected protein complexes, similar to how an upper bound of 10,000 total types of interacting domain pairs has been proposed (23).

A major practical application of the periodic table framework will be in predicting and modelling the quaternary structure of protein complexes. Specifically, our results show that bijective quaternary structure topologies are far more likely to occur than non-bijective topologies, despite the fact that there are far more possible non-bijective topologies. We also provide predictions for the relative likelihoods of different bijective topologies. This knowledge can inform the interpretation of high-throughput interaction experiments (24), or structure-based interaction predictions (25) by highlighting the quaternary structure topologies that are possible and most likely to occur. Homology information can aid these quaternary structure predictions, and give further insight into the evolution and assembly of complexes, as subcomplexes often arise as evolutionary precursors and assembly intermediates (12). Similarly, the periodic table can tell us which evolutionary precursor topologies are likely to have given rise to a specific complex. The periodic table can also provide constraints for multi-subunit docking or modelling, both on the relative arrangements of subunits, and the overall complex symmetry (26–29). This could be further integrated into hybrid methods combining different experimental measurements (30), such as electrospray (31) or cross-linking (32) mass spectrometry.

This work could be of significant utility in the bioengineering of protein complexes. The self-assembly formalism introduced here allows one to specify exactly what essential interfaces would need to be engineered to form a protein complex of a given topology. This could facilitate *de novo* engineering of oligomeric assemblies (33–35) and allow for directed modulation of existing quaternary structures, stepping either across or down the “periodic table” in an incremental manner.

Despite its strong predictive power, it is important to reiterate that basic periodic table model does not account for ~8% of known protein complex structures. More than half of these exceptions arise due to quaternary structure assignment errors. Indeed, a benefit of this approach is that it highlights likely quaternary structure misassignments, particularly by identifying non-bijective complexes with even subunit stoichiometry. However, this still leaves us with ~4% of correct structures that are not compatible with the periodic table.

Although non-bijective complexes are possible, they are rare and so this should be given consideration in any protein modelling or engineering attempts. Related to this, it would be particularly interesting to see whether chaperones are more frequently involved in the assembly of non-bijective complexes, to stabilise the required asymmetric transitions. To model the non-bijective protein complexes, additional assembly steps involving fractional and non-stoichiometric transitions are needed. However, as we showed, this would also greatly expand the number of possible quaternary structure topologies. Therefore, we consider the periodic table in its current implementation to be a reasonable compromise, allowing the vast majority of existing quaternary structure topologies to be explained and the most likely unobserved topologies to be predicted.

## Methods

### ***nESI-MS experiments***

The complexes were kindly donated as follows: *Saccharomyces cerevisiae* SAGA deubiquitinating module (Protein Data Bank [PDB] ID: 3MHH; C. Wolberger, Johns Hopkins University School of Medicine, Baltimore); *Thermus thermophilus* MglA/MglB complex (PDB ID: 3T1Q; A. Wittinghofer, Max Planck Institute for Molecular Physiology, Dortmund); *Geobacillus stearothermophilus* PDH E1 subunit (PDB ID: 3DVA; B. Luisi, University of Cambridge); *Streptococcus pyogenes* toxin-antitoxin complex (PDB ID: 3Q8X; A. Meinhart, Max Planck Institute for Medical Research, Heidelberg); *Saccharomyces cerevisiae* phosphofructokinase (PDB ID: 3O8O; T. Schöneberg, University of Leipzig); *Ruegeria pomeroyi* propionyl-CoA carboxylase (PDB ID: 3N6R; L. Tong, Columbia University, New York); *Homo sapiens* MSL1-MSL2 complex (PDB ID: 4B7Y; J. Kadlec, European Molecular Biology Laboratory, Grenoble); *Synechococcus elongatus* acetylglutamate kinase / PII complex (PDB ID: 2V5H; V. Rubio, Instituto de Biomedicina de Valencia, Valencia); *Pseudomonas aeruginosa* 3-methylcrotonyl-CoA carboxylase (PDB ID: 3U9T, 3U9S; L. Tong, Columbia University, New York).

The nESI-MS experiments on complex disassembly and reassembly were performed as previously described using QToF2 mass spectrometers modified for high m/z operation (11). A list of all (sub)complexes observed under various experimental conditions is provided in Table S1.

### ***Protein complex datasets***

The full set of protein X-ray crystal structures was taken from the Protein Data Bank (PDB) on 2012-08-08. Only polypeptide chains with at least 30 residues were considered. Backbone only models were ignored, as well as structures containing nucleic acids, or > 10% non-water heteroatoms. Heteromeric protein complexes formed by polypeptide cleavage were also ignored. This procedure also has the effect of removing complexes with protein chains that lack unique *db\_id* sequence identifiers. Complexes with > 59 subunits or split over multiple PDB entries were excluded. In total, the final dataset contained 30469 monomers, 28935 homomers and 5543 heteromers. Manual quaternary structure assignments came from PiQSi, and include those with errors assigned as “probably yes/no” (17) and from additional manual searching of the literature.

The size of the interface between each pair of subunits from all protein complexes were calculated with AREAIMOL (36). For each complex, all interfaces of the same type were identified by calculating the correlation between atom-specific buried surface areas for each pair of interfaces. Only interfaces > 200 Å<sup>2</sup> were considered. Two interfaces were considered to be of the same type if the Pearson correlation between the buried surface areas in terms of equivalent amino acids is > 0.7. Interface sizes were averaged over all interfaces of the same type. Similarly, homomeric interfaces were classified as isologous if the correlation between the residue-specific buried surface area for each subunit is > 0.7.

The extended set of quaternary structure topologies, used to validate our predictive model, was taken from a more recent set of protein complex structures from the PDB on 2014-12-16. In

addition to the crystal structures used in the main set, electron microscopy structures were also considered here. Furthermore, the constraints on the main dataset were loosened, so that complexes formed *via* cleavage, as well as complexes containing nucleic acids or other heteroatoms were also included (although only protein chains were considered). In total, this extended set possessed 4214 bijective heteromers (with  $\leq 4$  unique subunits and  $\leq 12$  subunit repeats) not present in the main dataset. Within this set, there were 53 different quaternary structure topologies, 14 of which are new.

For certain analyses, we used non-redundant subsets of the full and extended datasets. Proteins were filtered for redundancy at the 50% sequence identity level, essentially as was done previously (18, 20). The non-redundant sets were used for the histogram in Fig. S4, the comparison between observed and predicted in Fig. S9, and for fitting the expected periodic table quaternary structure distribution in our model.

Both the main set and the extended set of quaternary structure topologies used in this study are provided in Table S2, in the form of pairwise interfaces formed between subunits.

### ***Determination of assembly pathways***

All of the mass spectrometry and literature-identified (dis)assembly pathways involved protein complexes of known structure where the subunit composition of at least one subcomplex intermediate could be identified. To complement this, we also performed a structural analysis where we identified similar protein complexes with different quaternary structures. Starting from the full set of homomeric and heteromeric complexes in our main dataset, we searched for complexes where another complex could be considered as a subset of the full complex. For example, a homodimer was considered to be a subset of a homotetramer if the subunits shared  $> 90\%$  sequence identity. All these subset complexes were considered to be similar to the experimentally identified subcomplexes for the purpose of defining assembly transitions. All of the experimentally identified subcomplexes and structural subsets are provided in Table S3.

Although we can observe the subcomplexes formed during (dis)assembly, we do not directly observe the assembly or disassembly steps that occur in solution. However, we can infer these from the subcomplexes identified. For every subcomplex and full complex, we identify the largest subcomplex that can be considered to be a subset of that (sub)complex. If no subcomplex is observed, then (dis)assembly is assumed to occur *via* free monomers. We then assign the transition between the two states into one of the five categories from Fig. 2. All (dis)assembly transitions are provided in Table S4.

### ***Interface cutting procedure***

In order to distinguish subunit interfaces that are circumstantial from those that are essential to the self-assembly process, we remove interfaces from the weighted subunit contact graph of a given complex, in increasing order of interface size. We skip interfaces if cutting them would result in the complex becoming disconnected, and stop when no further interfaces can be cut.

This procedure will not necessarily result in a tree-like graph, as the same interface may appear several times in the same complex. For example, a cyclic ring will not be cut further, as one would have to cut all interfaces, which are of the same size, at once.

The advantages of this approach are: a) it vastly simplifies the number of possible topologies, and unifies clearly similar topologies that would be treated differently if all interfaces were taken into account; (b) it is less arbitrary than conventional thresholds of interface size, and therefore is a more fundamental description of a complex; and (c) symmetry emerges directly from the graph topology through the number of distinct interfaces of a subunit.

### ***Topological enumeration procedure***

First let us consider what combinations of different interface types can be present in a homomeric structure. These are:

- one symmetric interface ( $C_2$  symmetry)
- one asymmetric interface (cyclic symmetry with more than two subunits)
- one symmetric and one asymmetric interface (dihedral or tetrahedral symmetry)
- two symmetric interfaces (dihedral symmetry)
- two asymmetric interfaces (only in tetrahedral symmetry)

Larger numbers of interface types are not possible in homomers, due to the constraints placed on the symmetry of the complex by these interface types.

As explained above, any heteromer that is formed using a combination of cyclisation, dimerisation and subunit addition can be represented as a homomer of multiple copies of the same heteromeric module, in which each subunit type appears exactly once. We can therefore enumerate all possible heteromeric topologies that can arise by carrying out the following procedure (illustrated in Fig. S8):

1. **Enumerate all trees of  $s$  distinct subunits.** These are the heteromeric modules, multiple copies of which will be joined together into ‘homomers’. The reason for restricting ourselves to trees instead of all possible graphs is because we aim to only consider the most important interfaces in the original complexes, by using the interface cutting procedure outlined in the previous section. This will always lead to tree-like structures as the repeated heteromeric modules.
2. **For each of the five possible combinations of interface types in homomers discussed at the beginning of this section, consider all topologically distinct ways in which the interfaces can be distributed across the set of subunits and pairs of subunits of the tree.** Topologically distinct here means that we cannot convert two distributions of interfaces into each other by only using symmetry operations of the tree.
3. **Construct the topologies of the complexes from these distributions of interfaces across the tree.** For this we also need to consider the possible number of repetitions for each cyclic interface (see for example the ring-like complexes of varying size in the periodic table). In complexes with 12 or more subunits we can have more than one pair of divisors with at least one even divisor (e.g. in the case of 12, three & four, and two & six), which leads to several possible topologies for the same total number of subunits (see the  $D_6$  and  $T$  examples in the  $r=12$  column of the  $s=1$  row of the periodic table).
4. **Distinguish isomorphic topologies. In some cases different distributions of interfaces in the enumeration procedure will lead to isomorphic topologies.** These can be easily identified using an isomorphism check, giving us the final enumeration of

topologies.

The numbers in the bottom right of each cell in the periodic table (Fig. 5) give the observed and total numbers of different topologies for that particular symmetry group and number of subunits  $s$ .

### ***Enumeration of all possible bijective and non-bijective topologies***

We enumerate all possible bijective and non-bijective topologies with 2:2 stoichiometry, and all possible bijective and non-bijective topologies with 3:3 stoichiometry and up to six interfaces. We do this by considering all possible four-node graphs with three to six edges (2:2), shown in Fig. S2, and all six-node graphs with five or six edges (3:3), shown in Fig. S3. On these we consider all distributions of equal numbers of two node colours (corresponding to the two protein species). Furthermore, for all edges between nodes of the same colour we consider two possibilities, corresponding to isologous and heterologous interfaces. Edges can thus have three colourings (heteromeric, homomeric isologous, homomeric heterologous). Finally we also consider different interface sizes by considering all possible relative size ranks (including equal ranks) of the different edges and subsequent interface cutting according to the same procedure followed in the periodic table. An isomorphism check (which includes the colourings of nodes and edges) is then used to identify unique topologies in this enumeration. Two additional constraints are that an equal interface size can appear more than once only for one type of subunit pair, and that the same interface size can only appear once on each subunit for homomeric isologous and heteromeric interfaces, and once or twice for homomeric heterologous interfaces.

### ***Prediction of expected frequencies of quaternary structure topologies***

To predict the expected frequencies of quaternary structure topologies, we employed an approach whereby we attempt to recapitulate the observed distribution of complexes within cells on the periodic table, considering complexes with up to 12 subunit repeats and 4 unique subunits. This prediction procedure can be divided into three parts:

**1. Selecting a cell from the periodic table.** To do this, we first randomly select a structure from the non-redundant set of complexes and monomers. The row of the periodic table (*i.e.* the number of unique subunits) is directly taken from this randomly selected structure. However, the column of the periodic table is not taken directly from this structure. This is because the sampling of cells on the periodic table is sparse for the lower rows, and thus we would miss cells with no current structures. Instead, each structure can be classified into one of three groups: first column (*i.e.* monomeric or no repeated subunits), cyclic (including  $C_2$ ), or dihedral/tetrahedral. Then, another structure is randomly selected out of all of those from the first row of the table (*i.e.* a monomer or homomer) that belong to the same group. This structure is used to define the column of the periodic table. Thus the distribution of homomers defines the distribution of predicted heteromers in the horizontal axis, with a correction for the fact that complexes with more subunit repeats tend to be less likely to have cyclic or dihedral subunit repeats.

**2. Defining the assembly steps.** Each cell of the periodic table is associated with a specific set of subunit addition, dimerisation and/or cyclisation assembly steps required to get there from a monomer. Therefore, to generate a random quaternary structure topology compatible with a given cell, we first randomise the order of the assembly steps. There are two exceptions to this: i) for  $D_n$  ( $n > 2$ ) topologies where the homomer has all isologous interfaces (e.g. trimer of dimers rather than dimer of trimers), the dimerisation step must occur before the cyclisation step; and ii) for tetrahedral complexes, the cyclic trimerisation must occur before the tetramerisation.

**3. Constructing a quaternary structure topology.** Given a defined set of assembly steps, we can construct a quaternary structure topology. A new interface is added to the quaternary structure topology for each assembly step. When the subcomplex is heteromeric, there are multiple ways an interface could be formed. In these cases, the subunit(s) to be involved are randomly selected. For example, if a dimerisation step was applied to an A-B subcomplex, then a new isologous interface could be formed between two A subunits or two B subunits, or a pair of identical heteromeric interfaces could be formed between each pair of A and B subunits. In the latter case, the isologous interface is “distributed” across two subunits. In Fig. S10, we illustrate the random quaternary structure topology construction process for a single cell of the periodic table.

All predicted quaternary structure topologies are provided in Table S5.

### ***Enumeration-based shorthand notation of protein complex topologies***

The representation of interface distributions across the subunits, as shown in the figures describing the enumeration process (Figs. 3 and S9), allows for a natural shorthand notation of topologies: The subunits are labelled *A*, *B*, *C*, etc., and cyclic, dihedral, and heteromeric interfaces are denoted *c*, *d* and *h*. The location of an interface follows the type, and in the case of distributed interfaces both subunits are given. A cyclic homomer therefore is *cA*, the first structure with two subunits in Fig. 3 is *cA hAB*, and the last structure with three subunits in the same figure is *dAC hAB hBC*.

## **Acknowledgements**

We thank C. Wolberger, A. Wittinghofer, B. Luisi, A. Meinhart, T. Schöneberg, L. Tong, J. Kadlec and V. Rubio for providing protein complex samples. We thank H. Rees for assistance with manual quaternary structure assignments and identification of literature-derived assembly pathways. We thank P. Beltrao, S. Edelstein, T. Flock, D. Gfeller, M. Hein, F. Krueger, R. Laskowski, E. Levy, S. MacKinnon, I. Moal, E. Natan, T. Perica, B. Stauch, S. Velankar, S. Wodak and X. Zhang for helpful discussions and comments on the manuscript. This work was supported by the Royal Society (S.E.A. and C.V.R.), the Human Frontier Science Program (J.A.M.), the Medical Research Council grant G1000819 (H.H. and C.V.R.) and the Lister Institute for Preventative Medicine (S.A.T.).

## **References**

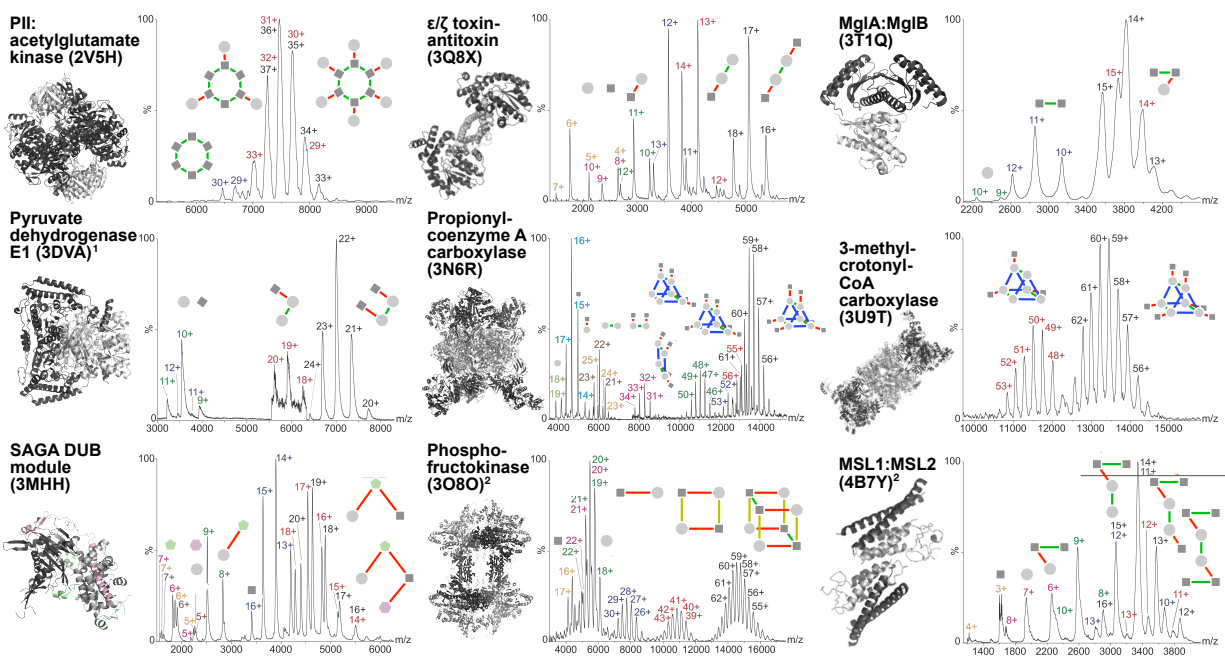
1. D. S. Goodsell, A. J. Olson, Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 105–153 (2000).
2. J. Janin, R. P. Bahadur, P. Chakrabarti, Protein-protein interaction and quaternary structure.

- Q. Rev. Biophys.* **41**, 133–180 (2008).
3. J. A. Marsh, S. A. Teichmann, Structure, dynamics, assembly and evolution of protein complexes. *Annu. Rev. Biochem.* **84** (2015), doi:10.1146/annurev-biochem-060614-034142.
  4. M. Levitt, C. Chothia, Structural patterns in globular proteins. *Nature*. **261**, 552–558 (1976).
  5. C. A. Orengo, D. T. Jones, J. M. Thornton, Protein superfamilies and domain superfolds. *Nature*. **372**, 631–634 (1994).
  6. W. R. Taylor, A “periodic table” for protein structures. *Nature*. **416**, 657–660 (2002).
  7. A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
  8. C. A. Orengo *et al.*, CATH—a hierarchic classification of protein domain structures. *Struct. Lond. Engl.* **1993**, **5**, 1093–1108 (1997).
  9. E. D. Levy, J. B. Pereira-Leal, C. Chothia, S. A. Teichmann, 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.* **2**, e155 (2006).
  10. E. D. Levy, E. B. Erba, C. V. Robinson, S. A. Teichmann, Assembly reflects evolution of protein complexes. *Nature*. **453**, 1262–1265 (2008).
  11. J. A. Marsh *et al.*, Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell*. **153**, 461–470 (2013).
  12. J. A. Marsh, S. A. Teichmann, Parallel dynamics and evolution: protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *BioEssays*. **36**, 209–218 (2014).
  13. Z. Hall, A. Politis, C. V. Robinson, Structural Modeling of Heteromeric Protein Complexes from Disassembly Pathways and Ion Mobility-Mass Spectrometry. *Structure*. **20**, 1596–1609 (2012).
  14. J. Monod, J. Wyman, J.-P. Changeux, On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* **12**, 88–118 (1965).
  15. K. Henrick, J. M. Thornton, PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **23**, 358–361 (1998).
  16. E. Krissinel, K. Henrick, Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).
  17. E. D. Levy, PiQSi: protein quaternary structure investigation. *Struct. Lond. Engl.* **1993**, **15**, 1364–1367 (2007).
  18. J. A. Marsh, Rees, H A, Ahnert, S E, Teichmann, S A, Structural and evolutionary versatility in protein complexes with uneven stoichiometry. *Nat. Commun.* **6**, 6394 (2015).
  19. T. Perica *et al.*, The emergence of protein complexes: quaternary structure, dynamics and allostery. *Biochem Soc Trans.* **40**, 475–491 (2012).
  20. J. A. Marsh, S. A. Teichmann, Protein flexibility facilitates quaternary structure assembly and evolution. *PLOS Biol.* **12**, e1001870 (2014).
  21. S. S. Shen-Orr, R. Milo, S. Mangan, U. Alon, Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31**, 64–68 (2002).
  22. W. Ma, A. Trusina, H. El-Samad, W. A. Lim, C. Tang, Defining network topologies that can achieve biochemical adaptation. *Cell*. **138**, 760–773 (2009).
  23. P. Aloy, R. B. Russell, Ten thousand interactions for the molecular biologist. *Nat. Biotechnol.* **22**, 1317–1321 (2004).
  24. P. C. Havugimana *et al.*, A census of human soluble protein complexes. *Cell*. **150**, 1068–1081 (2012).
  25. Q. C. Zhang *et al.*, Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*. **490**, 556–560 (2012).
  26. Y. Inbar, H. Benyamini, R. Nussinov, H. J. Wolfson, Prediction of multimolecular assemblies by multiple docking. *J. Mol. Biol.* **349**, 435–447 (2005).



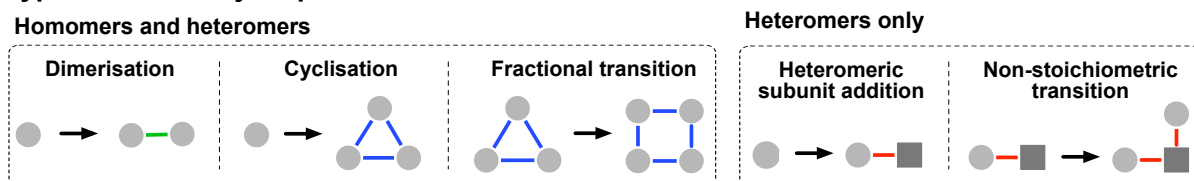
27. F. DiMaio, A. Leaver-Fay, P. Bradley, D. Baker, I. André, Modeling symmetric macromolecular structures in Rosetta3. *PLoS One*. **6**, e20450 (2011).
28. J. Esquivel-Rodríguez, D. Kihara, Evaluation of multiple protein docking structures using correctly predicted pairwise subunits. *BMC Bioinformatics*. **13 Suppl 2**, S6 (2012).
29. B. G. Pierce *et al.*, ZDOCK server: interactive docking prediction of protein–protein complexes and symmetric multimers. *Bioinformatics*. **30**, 1771–1773 (2014).
30. F. Alber, F. Förster, D. Korkin, M. Topf, A. Sali, Integrating Diverse Data for Structure Determination of Macromolecular Assemblies. *Annu. Rev. Biochem.* **77**, 443–477 (2008).
31. H. Hernández, C. V. Robinson, Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. *Nat. Protoc.* **2**, 715–726 (2007).
32. J. Rappsilber, The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *J. Struct. Biol.* **173**, 530–540 (2011).
33. N. P. King *et al.*, Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy. *Science*. **336**, 1171–1174 (2012).
34. Y.-T. Lai, D. Cascio, T. O. Yeates, Structure of a 16-nm cage designed by using protein oligomers. *Science*. **336**, 1129 (2012).
35. J. Zhang, F. Zheng, G. Grigoryan, Design and designability of protein-based assemblies. *Curr. Opin. Struct. Biol.* **27**, 79–86 (2014).
36. M. D. Winn *et al.*, Overview of the CCP 4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 235–242 (2011).

# Figures

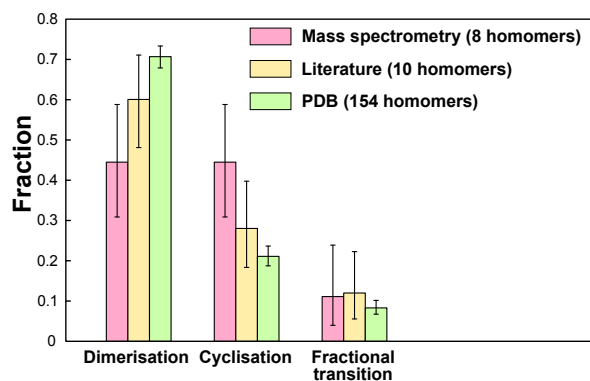


**Figure 1: Mass spectrometry characterisation of heteromer (dis)assembly pathways.** For each characterised complex, the known three-dimensional structure is shown, as well as representative mass spectrum, showing the full complex and subcomplexes formed in a graph representation. In all cases the full complex is represented by the rightmost graph representation in each mass spectrum. A full list of subcomplexes is provided in Table S1. The structures of 3DVA, 3O8O and 4B7Y shown here differ from those in the PDB: 3DVA is missing the  $\gamma$  subunit, as it was not present in our sample, and the 4:4 model of 3O8O and 4:2 model of 4B7Y were built from the unit cell to match the solution data.

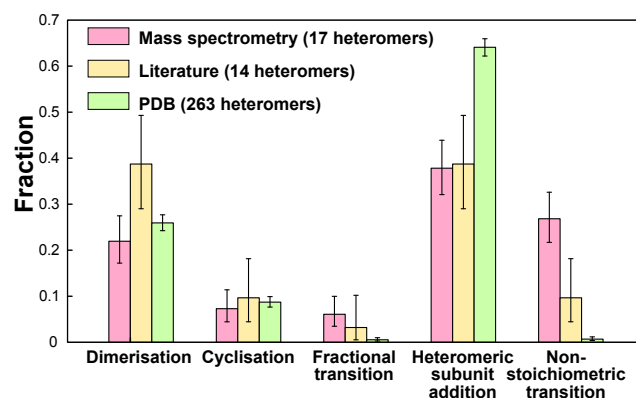
### A) Types of assembly steps



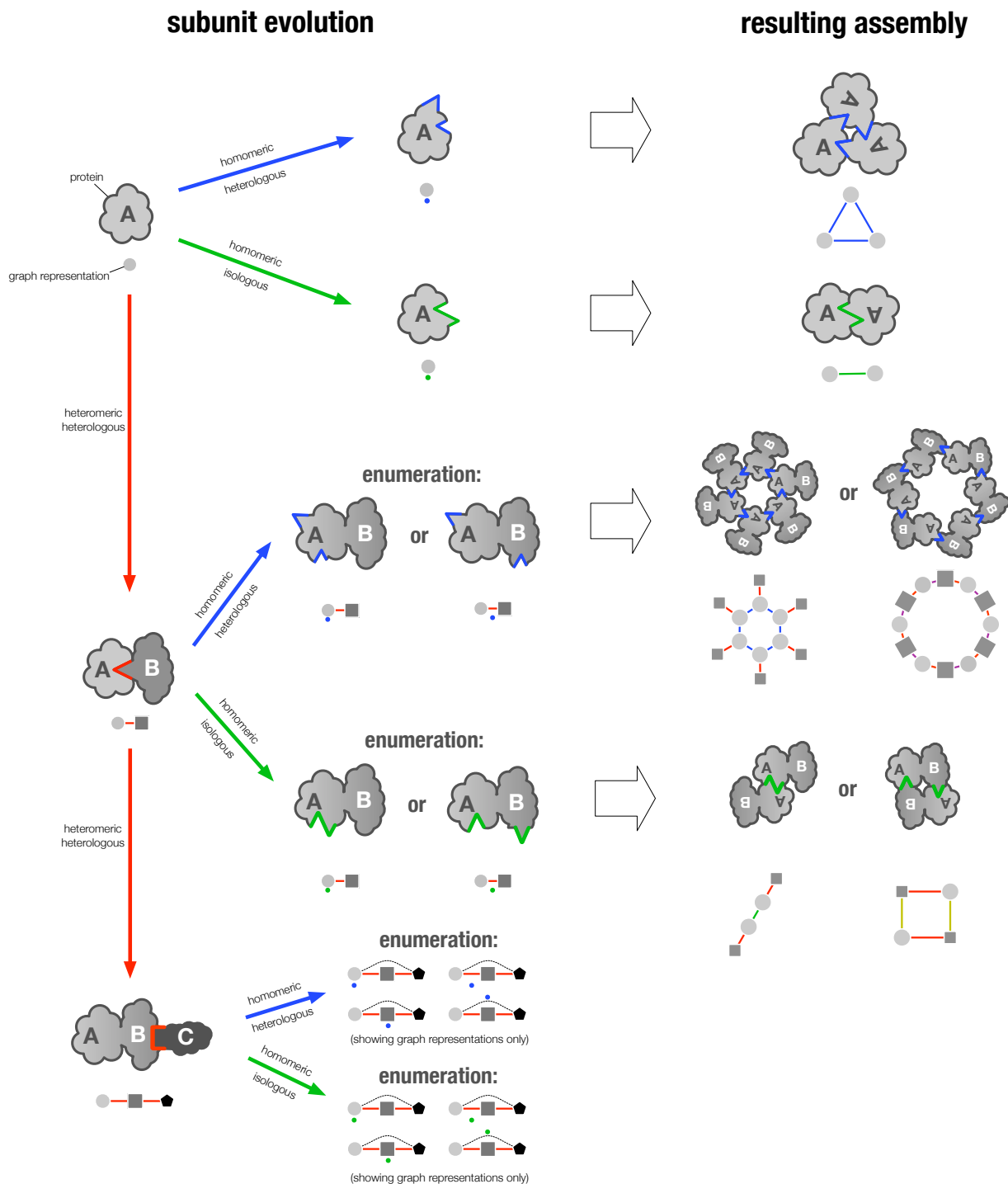
### B) Homomers



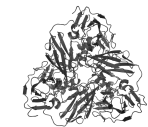

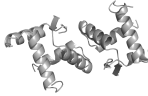
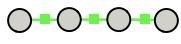

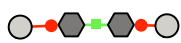



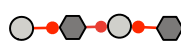
### C) Heteromers



**Figure 2: Types of assembly steps observed in homomeric and heteromeric complexes.** (A) Illustration of the five possible types of assembly steps. (B-C) Distribution of observed assembly steps for homomers and heteromers from mass spectrometry experiments, assembly pathways identified in the literature and from complexes with varying quaternary structure in the PDB. Error bars represent 68% Clopper-Pearson confidence intervals.



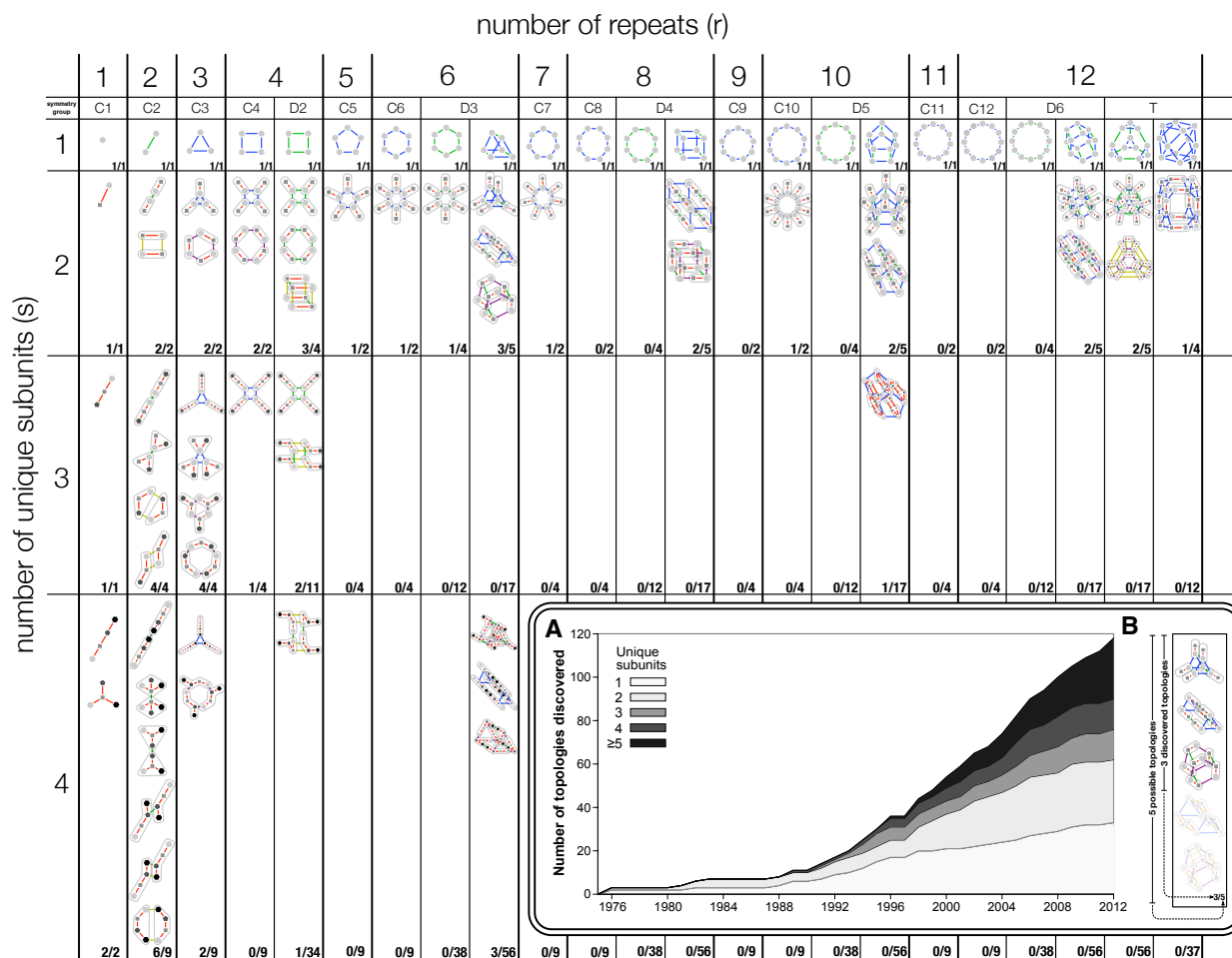
**Figure 3: Three assembly transitions give rise to the topological space of protein complexes.** These transitions are cyclisation (blue), dimerisation (green), and subunit addition (red). We can enumerate all possible topologies arising from these steps by calculating all ways in which a cyclic or dihedral interface can be distributed across a heteromer with 1:1 stoichiometry. For heterodimers, there are two such ways for both the cyclisation and the dimerisation steps. For heterotrimers, there are four such ways for each step.

|                   | Example   | Graph representation  | Bijectionity  | Occurrence | Error rate**  |
|-------------------|---|---|---|------------|---------------|
| <b>Homomers</b>   | <br>1as6 |  | <b>bijection homomer</b><br>one sequence, one topological environment   | 92.5%      | 10.6% (3.9%)  |
|                   | <br>1baz |  | <b>non-bijection homomer</b><br>one sequence, more than one topological environment   | 7.5%       | 60.1% (48.9%) |
| <b>Heteromers</b> | <br>1wbj |  | <b>bijection heteromer</b><br>multiple sequences, which map bijectionly (i.e. one-to-one) to topological environments   | 91.7%      | 9.0% (9.3%)   |
|                   | <br>1xd2 |  | <b>non-bijection heteromer with uneven stoichiometry</b><br>multiple sequences, which <b>do not</b> map bijectionly (i.e. one-to-one) to topological environments, and <b>do not</b> all appear an equal number of times. | 6.4%       | 20.4% (20.1%) |
|                   | <br>3a33 |  | <b>non-bijection heteromer with even stoichiometry</b><br>multiple sequences, which <b>do not</b> map bijectionly (i.e. one-to-one) to topological environments, but <b>do</b> all appear an equal number of times.       | 1.9%       | 58.6% (58.6%) |

\* In this example the central protein forms two different interfaces with the two outer proteins due to the inherent asymmetry of proteins. The topological environments of the outer proteins therefore differ.

\*\* values in brackets exclude 'probably yes' and 'probably no' error assignments in PiQSi from the analysis

**Figure 4:** Frequencies of protein complex types and their quaternary structure assignment error rates. Among non-bijection heteromers we can further distinguish between those that exhibit even stoichiometry and those with uneven stoichiometry. The former are much more likely to be the result of quaternary structure assignment errors. The latter are more likely to represent biologically relevant quaternary structure. In the last column we give alternative error rates in brackets that exclude the PiQSi (17) error assignments 'probably yes' and 'probably no' from the analysis. These error rates follow the same pattern for non-bijection heteromers of even versus uneven stoichiometries.

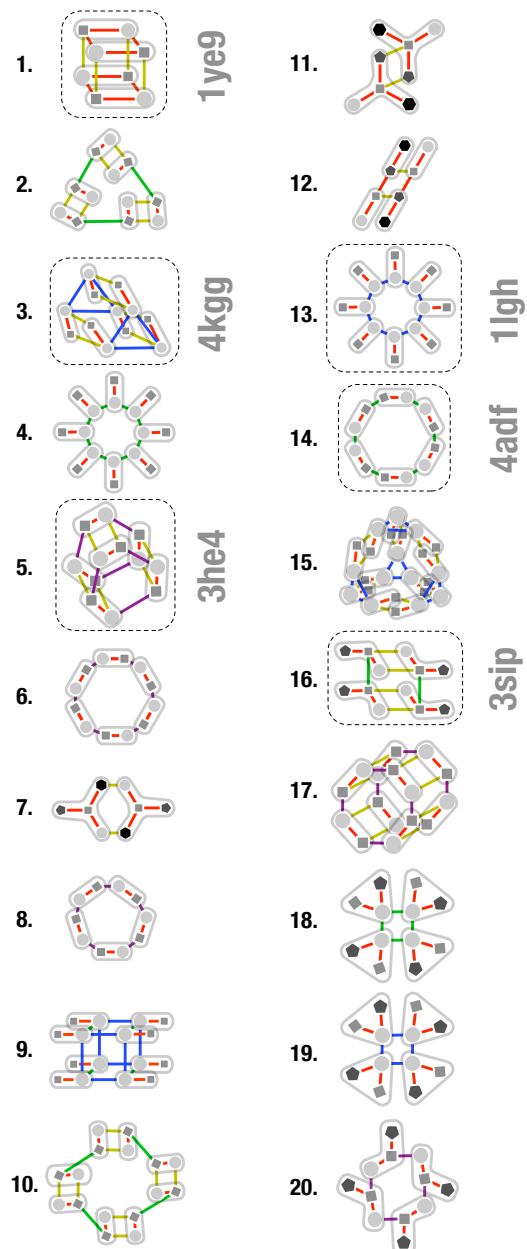


**Figure 5: The Periodic Table of Protein Complexes.** All bijective protein complex topologies can be arranged according to the number of different subunit types  $s$  and the number of times these subunits are repeated  $r$ . Isologous interfaces between the same subunits (*i.e.* dihedral interfaces), are shown in green, and heterologous interfaces between subunits of the same types (cyclic interfaces), are shown in blue. Heteromeric interfaces are shown in red, apart from those that correspond to a symmetric dimerisation (yellow), or to higher order cyclisation (purple). The topologies in the  $s=1$  row are the equivalent homomers of the heteromeric structures in the  $s>1$  rows. To clarify this equivalence, subunits in the heteromers are grouped according to the repeated subcomplexes. In addition the yellow and purple interfaces of the heteromeric complexes highlight interfaces that are dihedral (green) and cyclic (blue) in the equivalent homomers. The ratio in the bottom right of each cell indicates the number of topologies that have been observed and the total number of possible topologies of this type. The table shown here is an excerpt ( $s < 5$ ,  $r < 13$ ) of the full table. An interactive version of this table with information on the structures represented by each topology can be found at: <http://sea31.user.srcf.net/periodictable/> INSET: A) Number of discovered topologies as a function of time, which has been steadily increasing at a rate of about four topologies per year for the last two decades. B) An illustration of observed topologies versus all possible topologies with six repeats and two subunits ( $r=6$ ,  $s=2$ ). Three of the possible five topologies have been observed thus far.

## Top 20 predicted topologies

Out of 579 predicted topologies, a total of 14 are observed in the extended data.

Six of these observed topologies are among the top 20 predicted.



**Figure 6: The top 20 most likely quaternary structure topologies from our model not observed in the main dataset.** Of these, 6 are observed in the extended dataset, validating the power of the model ( $P$ -value:  $2 \times 10^{-6}$ ). The remaining 14 topologies are also expected to occur relatively frequently in nature, and thus to be observed in experimentally determined structures soon. The distribution of all the new topologies observed in the extended dataset compared to the expected frequencies of all predicted topologies is shown in Fig. S7.