

# Determining the architectures of macromolecular assemblies

Frank Alber<sup>1\*</sup>, Svetlana Dokudovskaya<sup>2\*†</sup>, Liesbeth M. Veenhoff<sup>2\*†</sup>, Wenzhu Zhang<sup>3</sup>, Julia Kipper<sup>2†</sup>, Damien Devos<sup>1†</sup>, Adisetyantari Suprpto<sup>2†</sup>, Orit Karni-Schmidt<sup>2†</sup>, Rosemary Williams<sup>2</sup>, Brian T. Chait<sup>3</sup>, Michael P. Rout<sup>2</sup> & Andrej Sali<sup>1</sup>

**To understand the workings of a living cell, we need to know the architectures of its macromolecular assemblies. Here we show how proteomic data can be used to determine such structures. The process involves the collection of sufficient and diverse high-quality data, translation of these data into spatial restraints, and an optimization that uses the restraints to generate an ensemble of structures consistent with the data. Analysis of the ensemble produces a detailed architectural map of the assembly. We developed our approach on a challenging model system, the nuclear pore complex (NPC). The NPC acts as a dynamic barrier, controlling access to and from the nucleus, and in yeast is a 50 MDa assembly of 456 proteins. The resulting structure, presented in an accompanying paper, reveals the configuration of the proteins in the NPC, providing insights into its evolution and architectural principles. The present approach should be applicable to many other macromolecular assemblies.**

A mechanistic understanding of the cell requires the structural characterization of the thousands of its constituent biological assemblies<sup>1</sup>. So far, conventional approaches have provided a valuable but limited window into the structures of these assemblies. For example, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy can resolve the atomic details of individual proteins and small complexes, whereas electron microscopy produces morphological maps but can lack the ability to identify and detail specific components in the map of the whole assembly. As a result, we do not yet have atomic-resolution structures, or even low-resolution representations, for the vast majority of complexes in the cell. How, then, are we to resolve the molecular architectures of these assemblies?

In an attempt to address this problem, we have taken the yeast (*Saccharomyces cerevisiae*) nuclear pore complex (NPC) as a case in point. The NPC is among the largest macromolecular assemblies in the cell, mediating the exchange of molecules that pass between the nuclear and cytoplasmic compartments. Yeast NPCs are ~50 MDa structures built of multiple copies of some 30 different proteins (nucleoporins), totalling at least 456 protein molecules<sup>2</sup>. Each NPC is a plastic structure embedded in the nuclear envelope and is composed of eight morphologically similar 'spokes' surrounding a central tube<sup>3–6</sup>. Filling this tube and projecting into both the cytoplasmic and nuclear sides are flexible filamentous domains from proteins termed FG (phenylalanine-glycine) repeat nucleoporins; these domains form the docking sites for transport factors that carry macromolecular cargoes through the NPC.

The NPC represents a significant challenge for conventional structure determination approaches owing to its large size and the high degree of flexibility of the complex and its components. Thus, although electron microscopy has provided valuable insights into

the overall shape of the NPC, its molecular architecture (that is, the spatial configuration of its component proteins) has yet to be revealed, and atomic structures have only been solved for domains covering ~5% of its component protein sequences<sup>7</sup>. The NPC therefore encapsulates many of the obstacles that will be encountered in the detailed structural examination of other macromolecular assemblies.

We describe here a set of proteomics experiments and a computational platform for converting the resulting data into the structures of macromolecular assemblies. Central to this approach is the realization that many kinds of biophysical and proteomic data contain valuable structural information about assemblies.

## Overview of integrative structure determination

Our approach to structure determination can be seen as an iterative series of four steps: data generation by experiment, translation of the data into spatial restraints, calculation of an ensemble of structures by satisfaction of these restraints, and an analysis of the ensemble to produce the final structure (Fig. 1). The structure calculation part of this process is expressed as an optimization problem, a solution of which requires three main components: (1) a representation of the assembly in terms of its constituent parts; (2) a scoring function, consisting of individual spatial restraints that encode all the data; and (3) an optimization of the scoring function, which aims to yield structures that satisfy the restraints.

Formally, our approach is similar to the determination of protein structures by NMR spectroscopy, in which the folding of the polypeptide chain is determined by satisfying distance restraints between pairs of atoms<sup>8</sup>. As with NMR spectroscopy, a structure is computationally determined from experimental data. Here, atoms

<sup>1</sup>Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, Byers Hall, Suite 503B, 1700 4th Street, University of California at San Francisco, San Francisco, California 94158-2330, USA. <sup>2</sup>Laboratory of Cellular and Structural Biology, and <sup>3</sup>Laboratory of Mass Spectrometry and Gaseous Ion Chemistry, The Rockefeller University, 1230 York Avenue, New York, New York 10065, USA. <sup>†</sup>Present addresses: Laboratory of Nucleocytoplasmic Transport, Institut Jacques Monod, 2 place Jussieu, Tour 43, Paris 75251, France (S.D.); Department of Biochemistry, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands (L.M.V.); German Aerospace Center (PT-DLR), Heinrich-Konen-Strasse 1, D-53227 Bonn, Germany (J.K.); Structural Bioinformatics, EMBL, Meyerhofstrasse 1, D-69117 Heidelberg, Germany (D.D.); Office of Technology Transfer, The Rockefeller University, 1230 York Avenue, New York, New York 10065, USA (A.S.); Herbert Irving Comprehensive Cancer Center, Columbia University, 1130 St Nicholas Avenue, New York, New York 10032, USA (O.K.-S.).

\*These authors contributed equally to this work.

are replaced by proteins, and their positions and relative proximities are restrained on the basis of data from a variety of proteomics and other experiments, including affinity purification, ultracentrifugation, electron microscopy and immuno-electron microscopy (immuno-EM).

**Data generation.** The most important aspect of our approach is its potential to use simultaneously almost any conceivable type of information to determine assembly structures. For example, sedimentation analysis of the isolated proteins can be used to infer their shapes; immuno-EM can give an approximate localization of each protein in the assembly; and affinity purification of tagged proteins and protein complexes can yield information about the arrangement and interactions of proteins within the assembly. These data can be of a kind not normally used for structure determination (for example, complexes identified by affinity purification, can refer to different levels in the structural hierarchy (for example, a protein domain, a whole protein, or a protein complex), and can be ambiguous in terms of their structural interpretation (for example, the uncertainty as to which copy of the protein is involved in an interaction, when multiple copies exist).

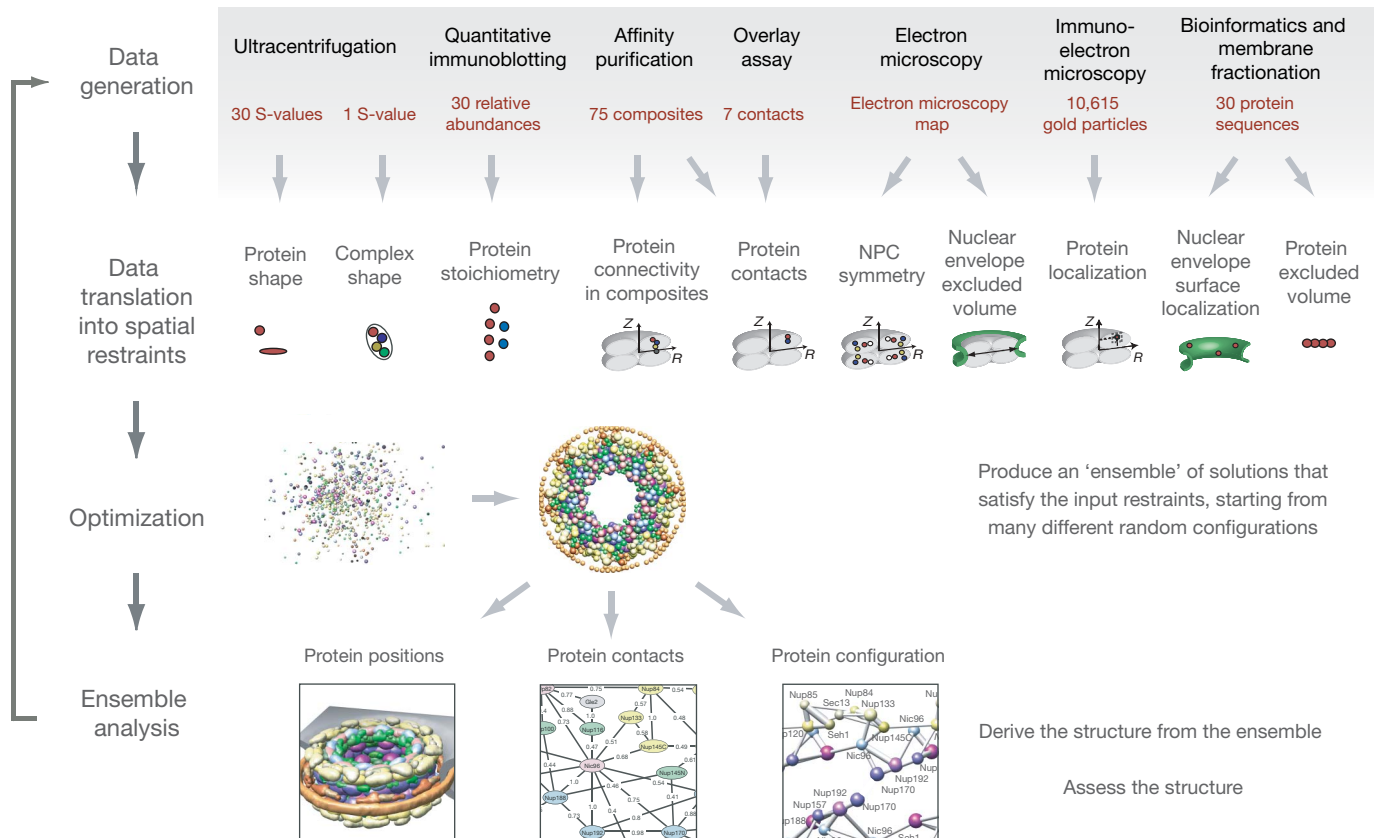
The use of such data for structure determination presented us with four major challenges. First, large amounts of suitable data must be collected to give sufficient spatial information to define structures; fortunately, the proteomic revolution has provided methodologies that allow us to garner enough information. Second, much of the data can be of relatively low precision; thus, to avoid over-interpretation, appropriate tolerances must be used in its structural interpretation. Third, the possibility of false-positive data must be minimized and taken into consideration. Fourth, ambiguity of the data in terms of its structural interpretation must be treated when multiple copies of the same protein are present in an assembly and the experiment does not

determine which specific instance of a protein is detected. All of these challenges can be addressed by an integrative approach that incorporates information varying greatly in terms of its accuracy and precision; limitations of any particular type of data can be overcome by the use of large and diverse data sets derived from synergistic experimental methods<sup>1,9</sup>.

**Data translation into spatial restraints.** The data can be used to restrain many different features of the assembly, such as the positions of proteins, protein contacts, proximity between proteins, and the shape and symmetry of the whole assembly. A 'restraint' specifies values of the restrained feature that are consistent with the experimental information about it; a perfectly satisfied restraint is indicated here by 0, whereas values larger than 0 correspond to a violated restraint. Thus, a restraint encodes our uncertainty in the restrained feature. In essence, restraints can be thought of as generating a 'force' on each component in the assembly, to mould them into a configuration that satisfies the data used to define the restraints.

**Optimization.** All the restraints are summed to obtain a scoring function, which determines the degree of consistency between the restrained spatial features in a structure and the experimental information; a perfect structure is indicated by 0, reflecting the summed values of perfectly satisfied restraints, whereas values larger than 0 correspond to a structure that increasingly violates restraints. The scoring function is then optimized to calculate a structure that minimizes violations of the restraints. It is necessary to generate many such structures to provide a good sampling of structures that are consistent with the data (that is, the 'ensemble').

**Ensemble analysis.** All of the structures that satisfy the input restraints are clustered into distinct sets, on the basis of their similarities. There are three possible outcomes of such clustering. First, if only a single cluster of structures satisfies all the input information,



**Figure 1 | Determining the architecture of the NPC by integrating spatial restraints from proteomic data.** First, structural data (red) are generated by various experiments (black). Second, the data are translated into spatial restraints. Third, an ensemble of structural solutions that satisfy the data are

obtained by minimizing the violations of the spatial restraints, starting from many different random configurations. Fourth, the ensemble is clustered into distinct sets of structures on the basis of their similarities, and analysed in terms of protein positions, contacts and configuration.

there is probably sufficient data for determining the unique native state. Second, if different clusters are consistent with the input information, either the data are insufficient to define the single native state or there are multiple native structures. If the number of clusters is small, the structural differences between them may suggest additional experiments so as to narrow down the possible solutions. Third, if no structures satisfy all input information, either the data or their interpretation in terms of the restraints is incorrect. Given the first two outcomes, the ensemble can be analysed to determine different aspects of the native state, such as protein positions, contacts and configuration. The variability of the ensemble provides an estimate of the precision of the structure determination.

We illustrate our approach by determining the configuration of the protein components in the NPC from the yeast *S. cerevisiae* (Fig. 1).

### Data generation

As no single experimental technique has been sufficient to solve the molecular architecture of the NPC, we used a variety of techniques, each of which gave different and synergistic information about the structure; the techniques were chosen to generate the needed structural information with a defined level of accuracy.

**An NPC component list.** To determine any structure, we must first define its parts (Fig. 2). In the case of the NPC, we have already determined that some 30 nucleoporins constitute the assembly<sup>2</sup>. Although the exact composition is still uncertain because some proteins interact relatively transiently with the NPC, potential omission of a small fraction of such transient components is unlikely to interfere with structure determination.

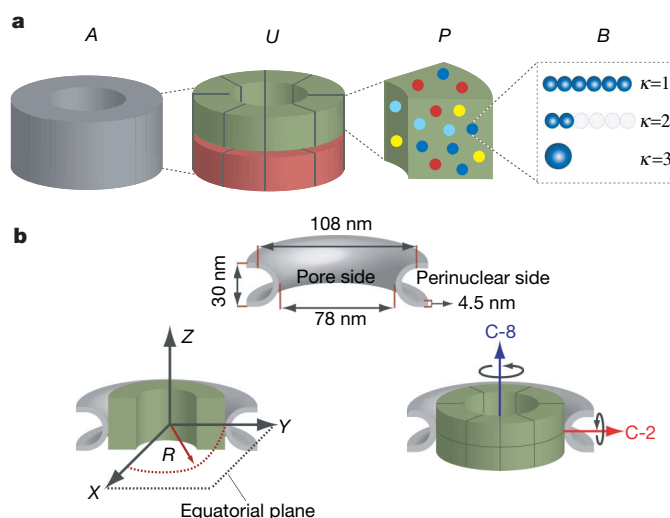
**The stoichiometry of each component in the NPC.** The stoichiometry of each nucleoporin in each half-spoke has been previously established<sup>2</sup>. However, having found the stoichiometry of Nup82 to be ambiguous, we re-examined it with new strains and found that Nup82 is present in two copies per spoke (Fig. 3 and Supplementary Fig. 7).

**The shape and size of each component.** Next, we must represent the structures of the constituent nucleoporins. Because atomic structures have not yet been solved for most nucleoporins, we estimated their shapes based primarily on their sedimentation coefficients determined by ultracentrifugation of the purified proteins (Fig. 3 and Supplementary Information). The sedimentation behaviour of most FG nucleoporins agrees with their predicted filamentous, native disordered structure<sup>10,11</sup>. Pom152, an integral membrane component, appeared to be a highly elongated structure, consistent with its multiple domains modelled as  $\beta$ -cadherin-like folds<sup>7</sup>. Most of the other nucleoporins appear to have a relatively compact tertiary structure that is again in agreement with their predicted fold assignments<sup>7,12</sup>. The seven-protein Nup84 complex<sup>13</sup> could be separated into two smaller complexes on sedimentation: an elongated tetramer (composite 30, see below) and an elongated hexamer (composite 45, see below), consistent with their elongated appearance when visualized by electron microscopy<sup>14</sup>.

**The size, shape and symmetry of the NPC.** It is also helpful to have some information on the overall shape and symmetry of the NPC. The position of the nuclear envelope membrane relative to the NPC and the NPC's symmetry are based on our electron microscopy and cryo-electron microscopy (cryo-EM) data<sup>5</sup>. These studies have revealed an eight-fold rotational symmetry of the yeast NPC and an approximate two-fold rotational symmetry between the nucleoplasmic and cytosolic halves of the NPC, defining the 'half-spoke' as a 16-fold pseudo-symmetry unit of the NPC (Fig. 2). We have also previously shown that heparin treatment of isolated NPCs produced a ring-like substructure ('Pom rings'), which is associated with the pore membrane and perinuclear space in the intact NPC<sup>15</sup>. We isolated and examined these rings (Supplementary Information), and found that they had a maximum diameter of  $\sim 106$  nm, consistent with the measured maximum NPC diameter of  $\sim 97$  nm<sup>5</sup>.

**The localization of each component in the NPC.** We have previously obtained the coarse localization of most nucleoporins within the NPC by immuno-EM, relying on a gold-labelled antibody that specifically interacted with the localized protein through its carboxy-terminal PrA tag (Fig. 4a)<sup>2</sup>. We have now generated a more accurate and complete immunolocalization map of the NPC, in which its constituent proteins, except Sec13, have been localized using a larger data set and improved analysis (Fig. 4b and Supplementary Information).

Inherent limitations in the immuno-EM method allow it to provide only a broad range of allowed axial and radial values for each nucleoporin. Nevertheless, these ranges are smaller than the dimensions of the half-spoke and so are still informative. Notably, most nucleoporins are found on both the nuclear and cytoplasmic sides of the NPC and are tightly packed within a region adjacent to the nuclear membrane (Fig. 4). Most of the FG nucleoporins are found on both sides of the NPC, with a small number found exclusively on the cytoplasmic or nuclear side; for simplicity, we consider Nup116 and Nup100 to be cytoplasmically disposed and Nup145N to be nucleoplasmically disposed, although  $\sim 20\%$  of the signal of each is found on the opposite side. Most of the non-FG nucleoporins are also found on both sides. The membrane proteins are found close to the nuclear envelope membrane, and Pom152–PrA is localized to the lumen of the nuclear envelope. Our immuno-EM map agrees almost entirely with independent localizations performed by other groups. For example, Nup159 and Nup82 have previously been shown to be restricted to the peripheral cytoplasmic face<sup>16</sup>; Nup1 was found on the peripheral nuclear face<sup>17</sup>; and Nup157, Nup170, Nup53 and Nup59 were shown to localize proximally to both sides of the

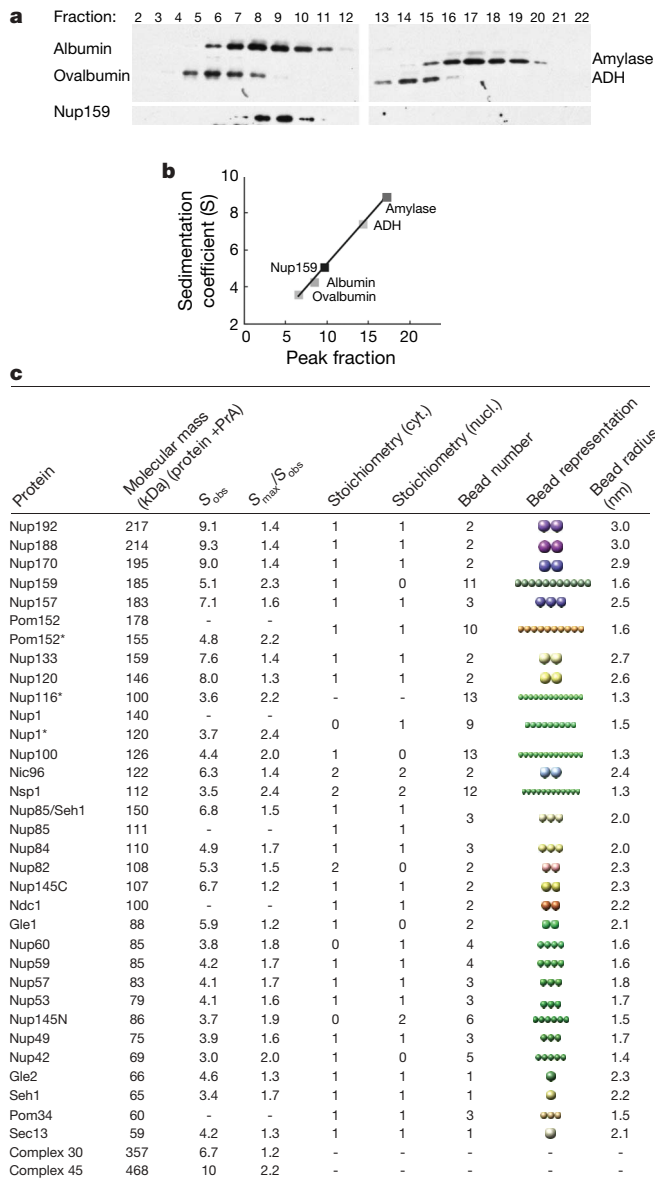


**Figure 2 | Structural representation of the NPC.** **a**, Hierarchical representation of the NPC that facilitates the expression of the experimental data in terms of spatial restraints. Formally, we define the whole NPC assembly *A* as a set of symmetry units *U* of two different types with eight instances each, referred to as half-spokes. Half-spokes of the first type (green) reside at the cytoplasmic side and half-spokes of the second type (red) reside at the nucleoplasmic side of the nuclear envelope. Two adjacent half-spokes, one of each type, form a spoke. Each of the 16 NPC half-spokes consists of a set of proteins *P* that are described by their type and index. Each protein is represented by a flexible string of beads *B* in the root representation  $\kappa = 1$ . Additional representations  $\kappa > 1$  can be derived from the root representation (for example, by omitting some beads as in  $\kappa = 2$  or by combining beads as in  $\kappa = 3$ ). For the NPC, each protein is described with up to nine different representations. **b**, Top panel: the dimensions of the nuclear envelope, as taken from cryo-EM images (ref. 5). Bottom-left panel: the coordinate system we use has the origin at the centre of the nuclear envelope pore. The nuclear envelope is indicated in grey. Bottom-right panel: the eight-fold (C-8) and two-fold (C-2) symmetry axes of the NPC, as revealed primarily by cryo-EM<sup>5</sup>. We apply the two-fold symmetry only to proteins that appear with identical stoichiometry in both the nucleoplasmic and cytoplasmic half-spokes.



NPC<sup>18</sup> (other independent localizations are listed in Supplementary Table 10).

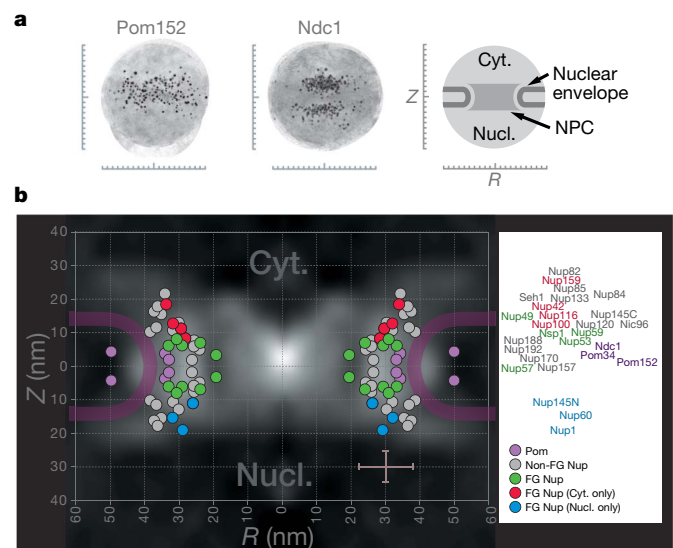
**How the NPC components fit together.** The coarse shape, approximate position and stoichiometry of each nucleoporin are not



**Figure 3 | Protein shape and stoichiometry information.** **a**, Protein shape from hydrodynamic experiments. Purified native PrA-tagged nucleoporins were sedimented on sucrose gradients, together with a set of biotin-labelled marker proteins. Fractions were collected and analysed by immunoblotting of the biotin and PrA tags. An immunoblot of fractions from a typical sedimentation analysis is shown, indicating the position of the tagged protein (Nup159–PrA) together with the markers ovalbumin (3.6 S), bovine serum albumin (4.3 S), alcohol dehydrogenase (ADH, 7.4 S) and  $\beta$ -amylase (8.9 S). **b**, Peak positions for the sedimenting proteins were determined and linear regression was used to calibrate the sedimentation coefficients of the PrA-tagged nucleoporin. **c**, Bead representations  $\kappa = 1$  of the NPC proteins and their stoichiometries per half-spoke. The stoichiometry of a protein in the cytoplasmic (cyt.) and nucleoplasmic (nucl.) half-spoke, as measured by quantitative immunoblotting<sup>2</sup>, is shown.  $S_{\text{max}}$  values were calculated based on the molecular mass (kDa) of each protein;  $S_{\text{max}}/S_{\text{obs}} < 1.4$  indicates a globular protein; 1.6–1.9, moderately elongated;  $> 2$ , highly elongated<sup>45</sup>. An asterisk indicates that C-terminal fragments were measured. Also shown is a visualization of the protein as a flexible bead chain (shown here in its most extended configuration), which is based on sedimentation analysis, identification of domains by sequence comparison and secondary structure prediction.

enough to build an accurate picture of the NPC: rather like the pieces in a jigsaw puzzle, we also need information on the interactions between nucleoporins. We obtained this information from a large number of overlay assays and affinity purification experiments, as well as from the composition of the Pom rings (consisting of Pom34 and Pom152). An overlay assay identifies a pair of proteins that interact with each other, whereas an affinity purification identifies one or more proteins that interact directly or indirectly with the bait protein (Figs 5 and 6 and Supplementary Information). An affinity purification produces a distinctive set of co-isolating proteins, which we term a composite. A composite may represent a single complex of physically interacting proteins or a mixture of such complexes overlapping at least at the tagged protein. We only used overlay and affinity purification data with a signal-to-noise ratio above a demanding threshold (Supplementary Information).

We designed several affinity purification methods to obtain a large and diverse set of composites (Supplementary Information). PrA was used as a high-affinity C-terminal purification tag on each nucleoporin. Different cell fractions from the tagged strains served as starting materials, although most fractions were produced by whole-cell cryolysis, which proved to be rapid and convenient, yielding high amounts of each complex with minimal losses and proteolytic damage. We generated  $\sim 20$  variants of extraction buffers with diverse properties to release different kinds of complexes from the fractions. Complexes were isolated via the tagged nucleoporins using antibody conjugated to either Sepharose or magnetic beads, although we preferred magnetic beads as it permitted rapid, high-yield isolations, and eliminated an upper size limit on the purified complexes (Supplementary Information). We also performed affinity purifications from diploid compared with haploid strains to detect a potential second, untagged copy of a given nucleoporin in the complex—a strong indication of a homotypic interaction for that nucleoporin; Pom152–PrA and Nup82–PrA were the only two nucleoporins giving composites containing a second untagged copy. Although



**Figure 4 | Localization of proteins by immuno-EM.** **a**, Immuno-EM montages for Pom152–PrA nuclei and Ndc1–PrA nuclear envelopes. Scale bars are graduated in 10-nm intervals using the coordinate system defined in Fig. 2b. The major features in each montage are shown schematically at the right, showing how the position of every gold particle in each montage was measured from both the central Z-axis of the NPC ( $R$ ) and from the equatorial plane of the nuclear envelope ( $Z$ ). **b**, Estimated position of the C terminus of each protein in the NPC relative to the central Z-axis of the NPC ( $R$ ) and the equatorial plane ( $Z$ ) superimposed on the protein density map of a cross-section of the yeast NPC obtained by cryo-EM<sup>2</sup>. The average allowed ranges along the  $R$  and  $Z$  coordinates ( $\pm 8$  nm and  $\pm 4.5$  nm, respectively) are indicated by the brown bars in the bottom right corner.



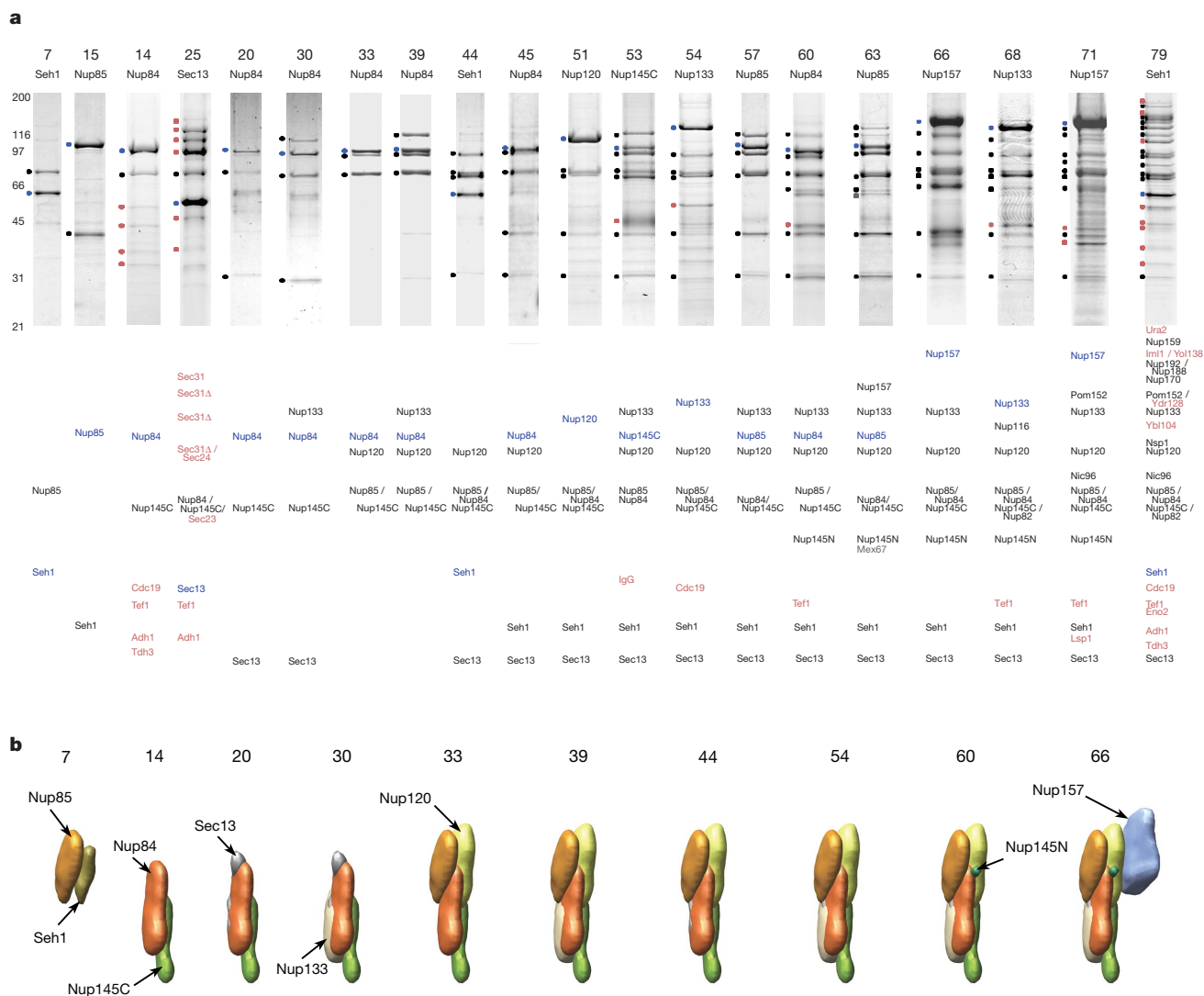
we originally designed our approaches for the purification of NPC complexes, they have proved to be useful for the isolation of many types of complexes from different cells<sup>7,12,19–24</sup>.

Identification of proteins was performed by mass spectrometry<sup>25,26</sup>. Generally, the most vicinal associates of the tagged protein should be approaching stoichiometric amounts in the purified complexes; conversely, distally associating proteins may be less abundant. By concentrating on only Coomassie-stainable SDS–polyacrylamide gel electrophoresis (PAGE) bands, we ensured that we identified only the more abundant proteins in any given affinity purification and avoided trace residuals (Fig. 5a). Polypeptides below ~20 kDa were excluded from this analysis for technical reasons<sup>27</sup>; however, due to their small volume, their exclusion is not likely to significantly affect structure determination.

Affinity purifications of tagged versions of all yeast nucleoporins, as well as the NPC-associated messenger RNA transport factors Gle1 and Gle2 (refs 28, 29), yielded 73 distinct composites; together with overlay

assays and Pom ring data, we have defined a total of 82 composites (Fig. 6a and Supplementary Information). The composites varied in complexity from dimers to those containing 20 proteins (composite 82) and, importantly, shared significant overlap in composition (Fig. 6b). Therefore, we expect considerable synergy among the composites when used to map the architecture of the whole assembly.

A good example of the compositional overlap is the Nup84 complex (Fig. 5a, b)<sup>13,14,30</sup>. The smallest building blocks of this complex are heterodimers (Fig. 5, composites 7, 14, 15). Under different isolation conditions, these dimers can be purified with an increasing number of additional proteins, such as trimers (25, 20), a tetramer (33), a pentamer (39), hexamers (44, 45, 51), and the full septameric Nup84 complex (53, 54, 57). This full complex interacts with Nup157 (63, 66) and Nup145N (60). Finally, the entire Nup84 complex co-precipitates together with the Nup170 complex and an Nsp1-containing complex (79). Our data also agree with composites generated by other groups. For example, the Nup84 composites<sup>13,14,30</sup>, a



**Figure 5 | Protein interactions of the Nup84 complex.** **a**, A sample of affinity purifications containing Nup84 complex proteins. Affinity-purified PrA-tagged proteins and interacting proteins were resolved by SDS–PAGE and visualized with Coomassie blue. The name of the PrA-tagged protein together with a corresponding identification number for the composite is indicated above each lane (Supplementary Information). Molecular mass standards (kDa) are indicated to the left of the panel. The bands marked by filled circles at the left of the gel lanes were identified by mass spectrometry (either of the example shown here or of a parallel version; Supplementary Information). The identity of the co-purifying proteins is indicated in order

below each lane; PrA-tagged proteins are indicated in blue, co-purifying nucleoporins in black, NPC-associated proteins in grey, and other proteins (including contaminants) in red. Each individual gel image was differentially scaled along its length so that its molecular mass standards aligned to a single reference set of molecular mass standards, and contrast-adjusted to improve visibility. **b**, The mutual arrangement of the Nup84-complex-associated proteins as visualized by their localization volumes. The localization volumes, obtained from the final NPC structure (Fig. 9), allow a visual interpretation of the relative proximities of the proteins.

Nup116 composite<sup>31</sup>, a Nup170 composite<sup>18</sup>, a Nup42–Gle1 dimer<sup>29</sup>, a Nic96 composite<sup>32</sup> and others (Supplementary Table 9) have been previously described, and are completely consistent with the composites identified here.

### Data translation into spatial restraints

The next step is to translate the experimental data about the NPC structure into spatial restraints (Fig. 1). These restraints were numerous, overlapping and varied in type, and thus were expected to be sufficient for defining the architecture of the NPC.

**Restraints and the scoring function.** Structure determination is enabled by expressing information as a scoring function, the global optimum of which corresponds to the structure of the native

assembly<sup>33</sup>. One such function is a joint probability density function (PDF) of protein positions, given the available information  $I$  about the system,  $p(C/I)$ , where  $C = (c_1, c_2, \dots, c_n)$  is the list of the cartesian coordinates ( $c_i$ ) of the  $n$  component proteins in the assembly (that is, the configuration of the proteins). This joint PDF gives the probability density that a component  $i$  of the native configuration is positioned very close to  $c_i$  given the information  $I$  we wish to consider in the calculation. In general,  $I$  may include any structural information from experiments, physical theories, or statistical preferences. The complete joint PDF is generally unknown, but can be approximated as a product of PDFs  $p_f$  that describe individual assembly features (for example, distances or relative orientations of proteins):

$$p(C/I) = \prod_f p_f(C/I_f)$$

The scoring function  $F(C)$  is then defined as the logarithm of the joint PDF:

$$F(C) = -\ln \prod_f p_f(C/I_f) = \sum_f r_f(C)$$

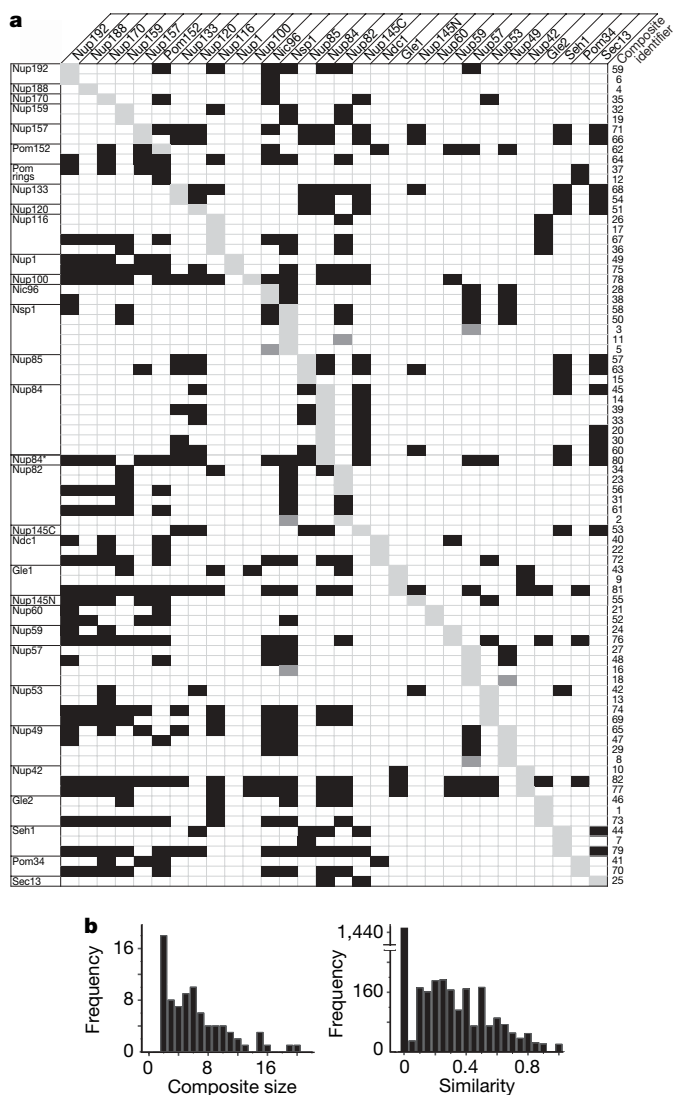
For convenience, we refer to the logarithm of a feature PDF as a restraint  $r_f$  and the scoring function is therefore the sum of the individual restraints.

**Setting up the representation of the NPC.** To define restraints on the components of an assembly, we must first specify the symmetry unit of the assembly (that is, the half-spoke in the case of the NPC) (Fig. 2a) and the stoichiometry of its components (Fig. 3). In addition, we must define the representations of the components. Each nucleoporin was represented by a flexible chain consisting of a small number of connected beads (Figs 2a and 3). The number and radii of the beads were chosen to reproduce the protein masses and the sedimentation coefficients<sup>34</sup>. The flexibility of the representation and the low granularity of the NPC structure are sufficient to accommodate uncertainties in the measured S-values and their interpretation. For the FG nucleoporins, no restraints other than the chain connectivity and excluded volume were imposed on the beads representing the FG-repeat regions.

Given the symmetry unit and the protein representations, we can formally represent the NPC with a four-level hierarchy corresponding to the whole NPC, the half-spokes, proteins and beads representing each protein (Fig. 2a). In addition, the nuclear envelope was represented as a rigid surface of many small beads, providing a mould in which the NPC forms (Fig. 2b).

**Symmetry of the NPC.** The eight-fold and approximate two-fold rotational symmetries of the NPC (Fig. 2b) were imposed by requiring essentially identical configurations of the proteins in common within each half-spoke; the corresponding restraint is formally the root-mean-square of the differences between equivalent intra-half-spoke distances. Although any individual NPC assembly may be perturbed from this perfect symmetry at any given point in time, restraints on the symmetry are nevertheless justified by the relatively low-resolution structure reported here, our intent to characterize the average structure, and exclusion of the FG-repeat regions from the symmetry restraints.

**Protein positions from immuno-EM.** To reflect the uncertainty in the immuno-EM data, we do not restrain a protein to a specific position. Instead, the C-terminal bead of each protein, corresponding to the tag position, was restrained by imposing lower and upper harmonic bounds on its  $Z$  and  $R$  coordinates (Fig. 2b), corresponding to the ranges allowed by the immuno-EM data. On average, the allowed area spans 16 and 9 nm along the  $R$  and  $Z$  coordinate, respectively (Fig. 4 Supplementary Tables 2 and 7, and Supplementary Fig. 8). With such large allowed ranges, the immuno-EM data provide little more information to the structure calculation than which side of the nuclear envelope each nucleoporin is on, and whether it is close to or distal from the NPC equatorial plane and the NPC axis.



**Figure 6 | Protein proximity by affinity purification.** **a**, Composites determined by affinity purification. The affinity-purified nucleoporin–PrA is indicated on the vertical axis, and the corresponding nucleoporins in each composite are shown on the horizontal axis. Composite identifiers are indicated to the right. Presence of a nucleoporin in a composite is indicated by a black box, and the tagged nucleoporin is indicated by a light grey box. In composite 64 (Pom152) and in composites 31 and 61 (Nup82), a second untagged copy of a corresponding protein is present, indicated by a black box. A direct interaction determined by overlay assay is indicated by a dark grey box. The asterisk for Nup84 indicates that the data were obtained with GFP-tagged Nup84. **b**, Distributions of composite size (left) and composite similarity (right). The similarity between two composites is defined by  $2a/(2a + b + c)$ , where  $a$  is the number of proteins that occur in both composites,  $b$  is the number of proteins present only in the first composite, and  $c$  is the number of proteins present only in the second composite.

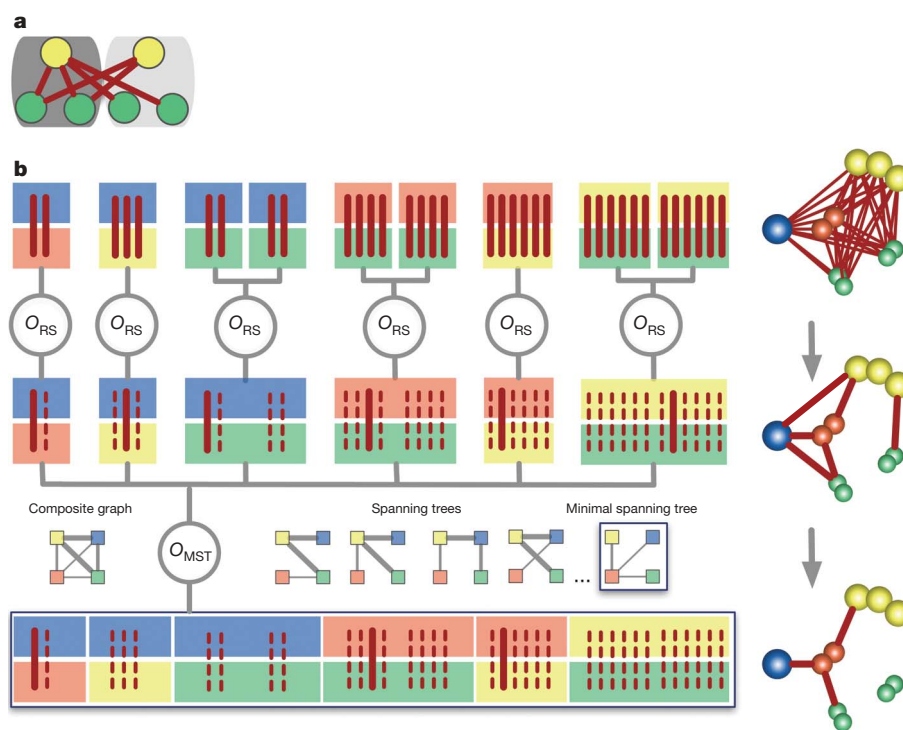
**Protein positions using the nuclear envelope as a mould.** The transmembrane-spanning helices of the three membrane proteins Pom152, Ndc1 and Pom34 were predicted by the program TMHH<sup>35</sup>. The corresponding beads were then restrained to the surface of the nuclear envelope by harmonic positional restraints. In addition, the terminal regions of each protein were restrained either to the pore or perinuclear sides of the nuclear envelope, on the basis of the immuno-EM data and the number of predicted transmembrane helices<sup>2</sup>.

**Protein proximities from overlay assays and affinity purifications.**

The overlay assays and affinity purifications carry information about protein proximities, and so are encoded by the same type of spatial restraint. These data provide the richest set of restraints for our NPC structure.

To interpret each composite in terms of a spatial restraint, we must consider three ambiguities. First, there is an ambiguity as to what contacts are present in a composite when it contains more than two proteins. A composite implies only that a copy of each protein in the composite must directly interact with at least one copy of another

protein in the composite; any structure that satisfies this condition is consistent with the observed composite. In other words, a composite of  $n$  proteins implies at least  $n-1$  such interactions between proteins of all types in the composite. Thus, each allowed combination of protein interactions corresponds to a 'spanning tree' of a 'composite graph' (as explained in Fig. 7b). Second, when there are multiple copies of the same protein in the assembly, there is an ambiguity as to which copy is involved in a given type of interaction (Fig. 7a). A measured interaction implies only that at least one copy of the protein is involved in that interaction. Third, when multiple beads are used to represent a protein, there is an ambiguity as to which bead is involved in the interaction (Fig. 2a). A measured interaction implies only that at least one bead of the protein is involved in that interaction. As a result of these three ambiguities, we need to encode a composite by a 'conditional restraint', ensuring that all allowed combinations of alternative assignments of interacting bead pairs are considered (Fig. 7b). Finding the assignment of interactions to specific beads that satisfies the data becomes part of the optimization process (see below). Other minor restraints were also derived from



**Figure 7 | Ambiguity in data interpretation and conditional restraints.**

**a**, The ambiguity for a protein interaction between proteins of green and yellow types is illustrated. The ambiguity results from the presence of multiple copies of the same protein in the same or neighbouring symmetry unit. In our NPC calculations, both neighbouring half-spokes on the cytoplasmic and nucleoplasmic sides are considered, for a total of four neighbouring half-spokes (not shown). **b**, The conditional restraint is illustrated by an example of a composite of four protein types (yellow, blue, red, green), derived from an assembly containing a single copy of the yellow, blue, and red protein and two copies of the green protein; proteins are represented by a single bead (blue protein), a pair of beads (green and red proteins), and a string of three beads (yellow protein) (right panel). This composite implies that at least three of the following six possible types of interaction must occur: blue–red, blue–yellow, blue–green, red–green, red–yellow and yellow–green. In addition, (1) the three selected interactions must form a 'spanning tree' of the 'composite graph' (defined below); (2) each type of interaction can involve either copy of the green protein (in general, all alternatives must be considered as illustrated in **a**); and (3) each protein can interact through any of its beads. These considerations can be encoded through a tree-like evaluation of the conditional restraint. At the top level, all optional bead–bead interactions between all protein copies are clustered by protein types. Each alternative bead interaction is restrained by

a harmonic upper bound on the distance between the beads; these are 'optional restraints', because only a subset is selected for contribution to the final value of the conditional restraint. Next, a 'rank-and-select' operator ( $O_{RS}$ ) selects only the least violated optional restraint from each interaction type, resulting in six restraints (thick red line) at the middle level of the tree. Finally, the minimal spanning tree operator ( $O_{MST}$ ) finds the combination of three restraints that are most consistent with the composite data (thick red line); here the edge weights in the minimal spanning tree (defined below) correspond to the restraint values given the current assembly structure. The column on the right shows a structural interpretation of the composite with proteins represented by their coloured beads and alternative interactions indicated by edges between them. The composite graph (shown on the left) is a fully connected graph that consists of nodes for all identified protein types and edges for all pairwise interactions between protein types; in the context of the conditional restraint, the edge weights correspond to the restraint values. Five of the sixteen possible spanning trees are also shown. A spanning tree is a graph with the smallest possible number of edges that connect all nodes. The minimal spanning tree is the spanning tree with the minimal sum of edge weights. This restraint evaluation process is executed at each optimization step based on the current configuration, thus resulting in possibly different subsets of selected optional restraints at each step.



the overlay assay and affinity purification data (Supplementary Information).

### Optimization

With the scoring function in hand, the positions of the proteins are determined by optimization of the scoring function (Supplementary Information), resulting in structures that are consistent with the data (Fig. 1). The optimization starts with a random configuration of the constituent proteins' beads, and then iteratively moves them so as to minimize violations of the restraints (Fig. 8). In essence, the restraints cooperate to slowly 'pull together' the proteins into a good-scoring configuration. We use standard methods of conjugate gradients and molecular dynamics with simulated annealing (Supplementary Information). These methods allow the evolving structure some 'breathing room' to explore the scoring function landscape, minimizing the likelihood of getting caught in local scoring function minima (Fig. 8a). To comprehensively sample structures consistent with the data, independent optimizations of randomly generated initial configurations were performed until an ensemble of 1,000

structures satisfying the input restraints was obtained (approximately 200,000 trials were required, running for approximately 30 days on 200 CPUs) (Fig. 8b).

### Ensemble interpretation

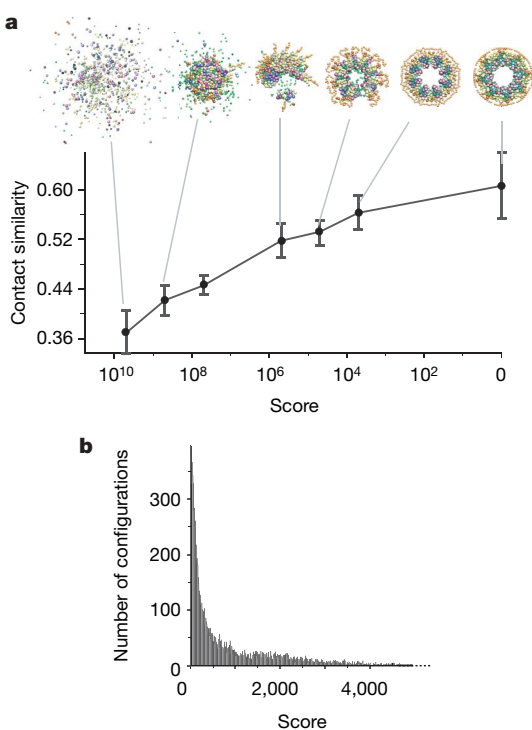
We analysed the ensemble of 1,000 structures that satisfy the input data (Fig. 8b) in terms of protein positions, contacts and configuration (Figs 9 and 10).

**Protein positions.** These 1,000 structures were first superposed (Fig. 9a) (Supplementary Information). Next, the superposed structures were converted into the probability of any volume element being occupied by a given protein (that is, the 'localization probability') (Fig. 9b). The spread around the maximum localization probability of each protein describes how precisely its position was defined by the input data. The positions that have a single narrow maximum in their probability distribution in the ensemble are determined most precisely. When multiple maxima are present in the distribution at the precision of interest, the input restraints are insufficient to define the single native state of that protein (or there are multiple native states).

The actual localization probabilities yielded single pronounced maxima for almost all proteins, demonstrating that the input restraints define one predominant structure. The average standard deviation for the distance between neighbouring protein centroids is 5 nm; the precision of the larger, centrally positioned proteins seems to be higher than that of the anchor domains of some FG nucleoporins. This level of precision defines a region smaller than the diameters of many nucleoporins. Thus, our map is sufficient to determine the relative positions of proteins in the NPC; we do not interpret features smaller than this precision. On the basis of the localization probabilities (Fig. 9b), we also define the volume most likely occupied by each protein, termed the 'localization volume' (Figs 9c and 10a). The localization volumes of the proteins overlap only to a small degree, such that only 10% of the NPC volume is assigned to two or more proteins, again underscoring how well the position of each nucleoporin is resolved. On the basis of our current data, we are not able to distinguish between the two possible mirror-symmetric structures; here, we present one of them.

**Protein contacts.** The proximities of any two proteins in the structure can be measured by their relative 'contact frequency', which is defined by how often the two proteins contact each other in the ensemble (Fig. 10b). Contacts are highly conserved among the ensemble structures, despite some variability; 32 protein pairs have a contact frequency higher than 65%. Of all the 435 contact frequencies, 7% are high (65–100%) and 73% are low (0–25%); this again demonstrates that the structure is well defined, as an ensemble of varied structures would yield mainly medium contact frequencies. Notably, few high-contact frequencies are seen between proteins of the same type, indicating that the NPC is held together primarily by heterotypic interactions.

We can improve our determination of contacts by considering not only the contact frequencies but also the composite data (Fig. 10c). More specifically, we define two proteins to be 'adjacent' if their relative contact frequency is larger than 65% or if they appear in the maximal spanning tree of any composite graph whose edge weights correspond to contact frequencies (as explained in Fig. 10c). If two proteins are adjacent, they are more likely to interact with each other in the native NPC structure than when they are not adjacent<sup>36</sup>. In total, 51 types of adjacencies were found (Fig. 10d). A particularly large number of adjacencies are observed for Nic96 and Nup82, which both appear in two copies per symmetry unit, as well as for the core proteins Nup192 and Nup188. Whereas the latter two proteins bridge the bulk of the NPC to the membrane proteins and also provide anchor sites for FG nucleoporins, Nic96 bridges major ring structures of the NPC and also serves as an anchor site for FG nucleoporins<sup>37</sup>. Most FG nucleoporins are peripherally located and therefore show only a few adjacencies.



**Figure 8 | Calculation of the NPC bead structure by satisfaction of spatial restraints.** **a**, Representation of the optimization process as it progresses from an initial random configuration to an optimal structure. The graph shows the relationship between the score (a measure of the consistency between the configuration and the input data) and the average contact similarity. The contact similarity quantifies how similar two configurations are in terms of the number and types of their protein contacts; a contact between two proteins occurs if the distance between their closest beads is less than 1.4 times the sum of the bead radii (Supplementary Information). The average contact similarity at a given score is determined from the contact similarities between the lowest scoring configuration and a sample of 100 configurations with the given score. Error bars indicate standard deviation. Representative configurations at various stages of the optimization process from left (very large scores) to right (with a score of 0) are shown above the graph; a score of 0 indicates that all input restraints have been satisfied. As the score approaches zero, the contact similarity increases, showing that there is only a single cluster of closely related configurations that satisfy the input data. **b**, Distribution of configuration scores. The presence of configurations with the score close to 0 demonstrates that our sampling procedure finds configurations consistent with the input data. These configurations satisfy all the input restraints within the experimental error.

**Protein configuration.** We can now combine the protein positions and adjacencies into a configuration of the NPC proteins (Fig. 10e, f). This representation allows us to deconvolute the composites into their constituent complexes (for example, see Figs 5b and 10g).

**Synergy among restraints.** How our data act synergistically is best demonstrated by the progressive increase in the certainty about the protein positions, as a result of an incremental addition of information (Fig. 11a). Hence, the variability among the 1,000 structures is significantly smaller than the uncertainties in any of the original data. For example, the allowed ranges for protein localization by immuno-EM are reduced from  $\pm 4.5$  and  $\pm 8$  nm along the Z-axis and the radial coordinate, respectively, to  $\pm 2$  and  $\pm 3$  nm in the ensemble, as a direct result of data integration. Similarly, data integration also improves the prediction of protein interactions (Fig. 11b).

### Assessment of precision and accuracy

The accuracy of a model is defined as the difference between the model and the native structure. Therefore, it is currently impossible to know with certainty the accuracy of the determined NPC structure. Nevertheless, five lines of evidence indicate that the accuracy of our structure is similar to its precision, and thus representative of the true configuration of the NPC.

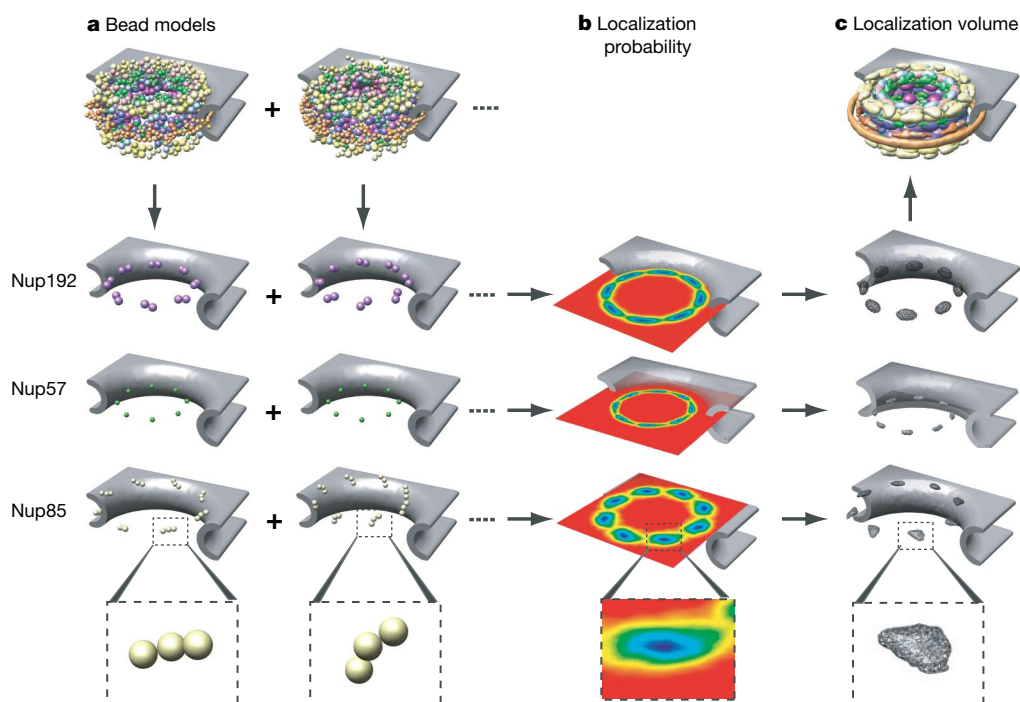
**Self-consistency of the experimental data.** Inconsistencies in the experimental data or its interpretation can be identified when the optimization generates only frustrated structures that do not satisfy the input restraints. This is not the case for our NPC calculations; we find only a single cluster of NPC structures that satisfy all the input restraints. To show that it is not trivial to find structures satisfying all restraints, we repeated the calculations with a comparable, but partly

incorrect set of restraints (Supplementary Information). Specifically, all untagged proteins were randomly swapped between composites, leaving the number of composites, the number of proteins in each composite, and all other restraints unchanged. An optimization using this modified restraint set failed to produce any structures that satisfied all restraints.

**Variability in the ensemble.** We have confirmed that the ensemble of 1,000 structures is sufficiently large for the precision of the NPC architecture to be determined reliably: the reproducibility of contact frequencies calculated from random subsets of the ensemble was plotted as a function of the subset size (Supplementary Information). The similarity between two sets of contact frequencies converges for random subsets of  $\sim 100$  structures.

**The ability of a restraint set to define a native state.** We have previously described an approach to test whether or not a given restraint set is sufficient to reconstruct a known native state<sup>36</sup>. In this approach, a native structure is assumed, the restraints to be tested are simulated from this structure, the structure is then reconstructed based only on these restraints, and finally the reconstruction is compared to the original assumed structure. Using this approach, we have simulated composite restraints based on our NPC structure, reproducing the number of composites and the distribution of their size in the original data set; all other restraints were kept the same as in the real application. The accuracy of the reconstructed model was comparable to the precision of the current NPC model.

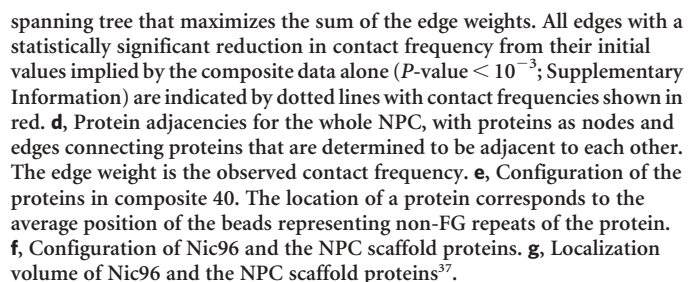
**Patterns unlikely to occur by chance.** The distribution of nucleoporins in our structure is expected to reflect their functionality and evolution, and so should be decidedly nonrandom. Indeed, as discussed at length in the accompanying paper<sup>37</sup>, there is a striking co-segregation of proteins by fold type to particular locations in the



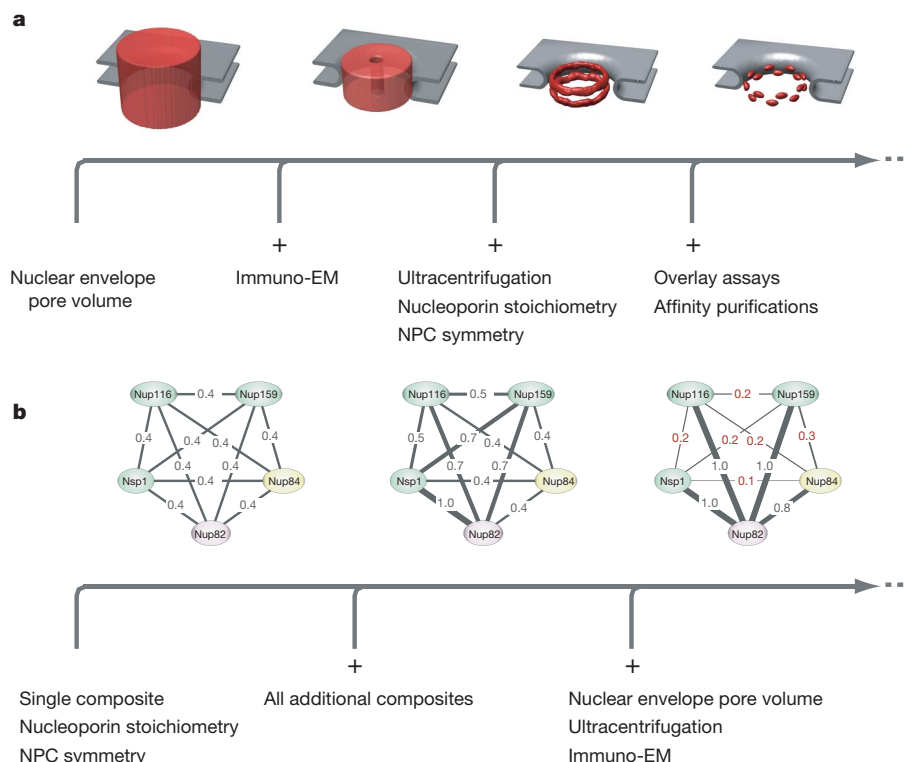
**Figure 9 | Bead model, ensemble, localization probability and localization volume.** **a**, Top: two representative bead models of the NPC (excluding the FG-repeat regions) from the ensemble of 1,000 superposed structures satisfying all restraints (Fig. 8b). The eight positions of three sample proteins (Nup192, Nup57 and Nup85) on the cytoplasmic side are shown, with a detailed view of the bead representation of one copy of Nup85 at the bottom. **b**, The localization probability for each protein type is obtained by converting the ensemble into the probability of any volume element being occupied by the protein. Shown are contour maps of the cross-sections in the plane parallel

to the equatorial plane that contains the maximum value of the protein localization probability. **c**, The localization volume of the sample proteins, derived from the localization probability. The volume elements are first sorted by their localization probability values. The localization volume then corresponds to the top-ranked elements, the volume of which sums to the protein volume, estimated from its molecular mass. The localization volume of a protein reveals its most probable localization. Because of the limited precision of the information used here, the localization volume of a protein should not be mistaken for its density map, such as that derived by cryo-EM.

regions), in good agreement with the experimentally reported maximal diameter of transported particles<sup>43</sup>. Fourth, Nup133, which has been experimentally shown to interact with highly curved membranes via its ALPS-like motif, is adjacent to the nuclear envelope in our structure<sup>44</sup>. Moreover, three of the four additional scaffold nucleoporins that are predicted to contain the ALPS-like motif are also close to the nuclear envelope. Finally, perhaps the best example is that of the Nup84 complex. Our configuration for this complex (Fig. 5b)<sup>37</sup> is completely consistent with previous results<sup>13,14,30</sup>. Specifically, Nup85 and Seh1 form a dimer that together with Nup120 forms the trimeric ‘head’ of the complex, consistent with the top two arms of the ‘Y’-shaped Nup84 complex (Fig. 5b)<sup>14</sup>. Similarly, Nup145C, Nup84, Sec13 and Nup133 form the ‘tail’ in







**Figure 11 | The structure is increasingly specified by the addition of different types of synergistic experimental information.** **a**, Protein positions. As an example, each panel illustrates the localization of 16 copies of Nup192 in the ensemble of NPC structures, generated using the data sets indicated below. The localization probability is contoured at 65% of its maximal value (red). The smaller the volume, the better localized are the proteins. The NPC structure is therefore essentially moulded into shape by the large amount of diverse experimental data. **b**, Protein contacts. Prediction of protein interactions from contact frequencies improves as more data are used. As an example, each panel illustrates the contact frequencies between proteins found in composite 34. Contact frequencies are shown as edge weights and indicated by the thickness of the lines

connecting the proteins. Left: when only a single composite is used (together with stoichiometry and symmetry information), all interactions are equally likely (initial contact frequency, Supplementary Information). Middle: when the highest likelihood of interaction between a particular protein pair from all composites is used, the uncertainty about the interactions is reduced. Right: when all data are used, the contact frequencies are either very high ( $>0.65$ ) or very low ( $<0.25$ ), thus allowing a strong prediction of protein interactions. Contact frequencies reflect the likelihood that a protein interaction is formed given the data considered and are calculated from the ensemble of optimized structures. Numbers in red indicate final contact frequencies that significantly decreased (at a  $P$ -value  $<10^{-3}$ ) from their initial values (Supplementary Information).

both our structure and the Y-shaped complex (Fig. 5b)<sup>14</sup>. Here, we resolve the relative positions of the proteins in this complex and show how the complex is integrated into the architecture of the entire NPC.

Together these assessments indicate that our data are sufficient to determine the configuration of the proteins comprising the NPC. Indeed, it is hard to conceive of any combination of errors that could have biased our structure towards a single solution that resembles known NPC features in so many ways.

## Conclusions

We have devised an integrative approach to solve the structure of the NPC using diverse biophysical and proteomic data. This approach has several advantages. First, it benefits from the synergy among the input data. Data integration is in fact necessary for structure determination, because none of the individual data sets contains sufficient spatial information on its own. Despite the little structural information in each individual restraint, the concurrent satisfaction of all restraints derived from independent experiments markedly reduces the degeneracy of the final structures. Second, the integrative approach can potentially survey all the structures that are consistent with the data. Alternatively, if no structure is consistent with the data, then some experiments or their interpretations are incorrect. Third, this approach can make the process of structure determination more efficient, by indicating which measurements would be most informative. Fourth, the approach can, in principle, incorporate essentially any structural information about a given assembly. Thus, it is straightforward to adapt it for calculating higher resolution

structures by including additional spatial restraints from higher resolution data sets, such as atomic structures of proteins, chemical cross-linking, footprinting, small angle X-ray scattering (SAXS) and cryo-EM. It is conceivable that these additional data sets might allow us to determine pseudo-atomic structures of assemblies as complex as the NPC. Furthermore, by obtaining detailed structural information concerning different stages of a dynamic process, our approach may animate the NPC's assembly and transport mechanisms<sup>6</sup>.

The molecular architecture of many macromolecular complexes could, in principle, be resolved using a similar integrative approach. With regards to the NPC, the resulting structure has already provided abundant insights into the function and evolution of the cell<sup>37</sup>.

## METHODS SUMMARY

See Supplementary Information for a detailed description of our Methods. The experimental data, the Integrative Modelling Platform software and the NPC structural model are available at <http://ncdir.org/npc>.

Received 30 August; accepted 22 October 2007.

1. Sali, A., Glaeser, R., Earnest, T. & Baumeister, W. From words to literature in structural proteomics. *Nature* **422**, 216–225 (2003).
2. Rout, M. P. *et al.* The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* **148**, 635–651 (2000).
3. Macara, I. G. Transport into and out of the nucleus. *Microbiol. Mol. Biol. Rev.* **65**, 570–594 (2001).
4. Weis, K. Nucleocytoplasmic transport: cargo trafficking across the border. *Curr. Opin. Cell Biol.* **14**, 328–335 (2002).

5. Yang, Q., Rout, M. P. & Akey, C. W. Three-dimensional architecture of the isolated yeast nuclear pore complex: functional and evolutionary implications. *Mol. Cell* **1**, 223–234 (1998).
6. Beck, M., Lucic, V., Förster, F., Baumeister, E. & Medalia, O. Snapshots of nuclear pore complexes in action captured by cryo-electron tomography. *Nature* **449**, 611–615 (2007).
7. Devos, D. *et al.* Simple fold composition and modular architecture of the nuclear pore complex. *Proc. Natl Acad. Sci. USA* **103**, 2172–2177 (2006).
8. Havel, T. F. & Wüthrich, K. A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of intramolecular <sup>1</sup>H–<sup>1</sup>H proximities in solution. *Bull. Math. Biol.* **46**, 673–698 (1984).
9. Malhotra, A., Tan, R. K. & Harvey, S. C. Prediction of the three-dimensional structure of *Escherichia coli* 30S ribosomal subunit: a molecular mechanics approach. *Proc. Natl Acad. Sci. USA* **87**, 1950–1954 (1990).
10. Denning, D. P., Patel, S. S., Uversky, V., Fink, A. L. & Rexach, M. Disorder in the nuclear pore complex: the FG repeat regions of nucleoporins are natively unfolded. *Proc. Natl Acad. Sci. USA* **100**, 2450–2455 (2003).
11. Lim, R. Y. *et al.* Flexible phenylalanine-glycine nucleoporins as entropic barriers to nucleocytoplasmic transport. *Proc. Natl Acad. Sci. USA* **103**, 9512–9517 (2006).
12. Devos, D. *et al.* Components of coated vesicles and nuclear pore complexes share a common molecular architecture. *PLoS Biol.* **2**, e380 (2004).
13. Siniossoglou, S. *et al.* Structure and assembly of the Nup84p complex. *J. Cell Biol.* **149**, 41–54 (2000).
14. Lutzmann, M., Kunze, R., Buerer, A., Aebi, U. & Hurt, E. Modular self-assembly of a Y-shaped multiprotein complex from seven nucleoporins. *EMBO J.* **21**, 387–397 (2002).
15. Strambio-de-Castillia, C., Blobel, G. & Rout, M. P. Isolation and characterization of nuclear envelopes from the Yeast *Saccharomyces*. *J. Cell Biol.* **131**, 19–31 (1995).
16. Miller, A. L. *et al.* Cytoplasmic inositol hexakisphosphate production is sufficient for mediating the Gle1-mRNA export pathway. *J. Biol. Chem.* **279**, 51022–51032 (2004).
17. Solsbacher, J., Maurer, P., Vogel, F. & Schlenstedt, G. Nup2p, a yeast nucleoporin, functions in bidirectional transport of importin alpha. *Mol. Cell Biol.* **20**, 8468–8479 (2000).
18. Marelli, M., Aitchison, J. D. & Wozniak, R. W. Specific binding of the karyopherin Kap121p to a subunit of the nuclear pore complex containing Nup53p, Nup59p, and Nup170p. *J. Cell Biol.* **143**, 1813–1830 (1998).
19. Archambault, V. *et al.* Genetic and biochemical evaluation of the importance of Cdc6 in regulating mitotic exit. *Mol. Biol. Cell* **14**, 4592–4604 (2003).
20. Archambault, V. *et al.* Targeted proteomic study of the cyclin-Cdk module. *Mol. Cell* **14**, 699–711 (2004).
21. Tackett, A. J. *et al.* I-DIRT, a general method for distinguishing between specific and nonspecific protein interactions. *J. Proteome Res.* **4**, 1752–1756 (2005).
22. Cristea, I. M., Williams, R., Chait, B. T. & Rout, M. P. Fluorescent proteins as proteomic probes. *Mol. Cell. Proteomics* **4**, 1933–1941 (2005).
23. Niepel, M., Strambio-de-Castillia, C., Fasolo, J., Chait, B. T. & Rout, M. P. The nuclear pore complex-associated protein, Mlp2p, binds to the yeast spindle pole body and promotes its efficient assembly. *J. Cell Biol.* **170**, 225–235 (2005).
24. Cristea, I. M. *et al.* Tracking and elucidating alphavirus-host protein interactions. *J. Biol. Chem.* **281**, 30269–30278 (2006).
25. Zhang, W. & Chait, B. T. ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.* **72**, 2482–2489 (2000).
26. Krutchinsky, A. N., Kalkum, M. & Chait, B. T. Automatic identification of proteins with a MALDI-quadrupole ion trap mass spectrometer. *Anal. Chem.* **73**, 5066–5077 (2001).
27. Stelter, P. *et al.* Molecular basis for the functional interaction of dynein light chain with the nuclear-pore complex. *Nature Cell Biol.* **9**, 788–796 (2007).
28. Murphy, R., Watkins, J. L. & Wente, S. R. GLE2, a *Saccharomyces cerevisiae* homologue of the *Schizosaccharomyces pombe* export factor RAE1, is required for nuclear pore complex structure and function. *Mol. Biol. Cell* **7**, 1921–1937 (1996).
29. Murphy, R. & Wente, S. R. An RNA-export mediator with an essential nuclear export signal. *Nature* **383**, 357–360 (1996).
30. Lutzmann, M. *et al.* Reconstitution of Nup157 and Nup145N into the Nup84 complex. *J. Biol. Chem.* **280**, 18442–18451 (2005).
31. Bailer, S. M. *et al.* Nup116p associates with the Nup82p-Nsp1p-Nup159p nucleoporin complex. *J. Biol. Chem.* **275**, 2354–23548 (2000).
32. Grandi, P., Doye, V. & Hurt, E. C. Purification of NSP1 reveals complex formation with 'GLFG' nucleoporins and a novel nuclear pore protein NIC96. *EMBO J.* **12**, 3061–3071 (1993).
33. Shen, M. Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507–2524 (2006).
34. Harding, S. E. Determination of macromolecular homogeneity, shape, and interactions using sedimentation velocity analytical ultracentrifugation. *Methods Mol. Biol.* **22**, 61–73 (1994).
35. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
36. Alber, F., Kim, M. F. & Sali, A. Structural characterization of assemblies from overall shape and subcomplex compositions. *Structure* **13**, 435–445 (2005).
37. Alber, F. *et al.* The molecular architecture of the nuclear pore complex. *Nature* doi:10.1038/nature06405 (this issue).
38. Akey, C. W. & Radermacher, M. Architecture of the *Xenopus* nuclear pore complex revealed by three-dimensional cryo-electron microscopy. *J. Cell Biol.* **122**, 1–19 (1993).
39. Stoffer, D. *et al.* Cryo-electron tomography provides novel insights into nuclear pore architecture: implications for nucleocytoplasmic transport. *J. Mol. Biol.* **328**, 119–130 (2003).
40. Kiseleva, E. *et al.* Yeast nuclear pore complexes have a cytoplasmic ring and internal filaments. *J. Struct. Biol.* **145**, 272–288 (2004).
41. Hinshaw, J. E., Carragher, B. O. & Milligan, R. A. Architecture and design of the nuclear pore complex. *Cell* **69**, 1133–1141 (1992).
42. Beck, M. *et al.* Nuclear pore complex structure and dynamics revealed by cryoelectron tomography. *Science* **306**, 1387–1390 (2004).
43. Pante, N. & Kann, M. Nuclear pore complex is able to transport macromolecules with diameters of about 39 nm. *Mol. Biol. Cell* **13**, 425–434 (2002).
44. Drin, G. *et al.* A general amphipathic  $\alpha$ -helical motif for sensing membrane curvature. *Nature Struct. Mol. Biol.* **14**, 138–146 (2007).
45. Schurmann, G., Haspel, J., Grumet, M. & Erickson, H. P. Cell adhesion molecule L1 in folded (horseshoe) and extended conformations. *Mol. Biol. Cell* **12**, 1765–1773 (2001).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We thank H. Shio for performing the electron microscopic studies; J. Fanghänel, M. Niepel and C. Strambio-de-Castillia for help in developing the affinity purification techniques; M. Magnasco for discussions and advice; A. Kruchinsky for assistance with mass spectrometry; M. Topf, D. Korkin, F. Davis, M.-Y. Shen, F. Foerster, N. Eswar, M. Kim, D. Russel, B. Peterson and B. Webb for many discussions about structure characterization by satisfaction of spatial restraints; C. Johnson, S. G. Parker and C. Silva, T. Ferrin and T. Goddard for preparation of some figures; and S. Pulapura and X. J. Zhou for their help with the design of the conditional diameter restraint. We are grateful to J. Aitchison for discussion and insightful suggestions. We also thank all other members of the Chait, Rout and Sali laboratories for their assistance. We acknowledge support from an Irma T. Hirsch Career Scientist Award (M.P.R.), a Sinsheimer Scholar Award (M.P.R.), a grant from the Rita Allen Foundation (M.P.R.), a grant from the American Cancer Society (M.P.R.), the Sandler Family Supporting Foundation (A.S.), the Human Frontier Science Program (A.S., L.M.V.), NSF (A.S.), and grants from the National Institutes of Health (B.T.C., M.P.R., A.S.), as well as computer hardware gifts from R. Conway, M. Homer, Intel, Hewlett-Packard, IBM and Netapp (A.S.).

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). Correspondence and requests for materials should be addressed to A.S. ([sali@salilab.org](mailto:sali@salilab.org)), M.P.R. ([rout@rockefeller.edu](mailto:rout@rockefeller.edu)), or B.T.C. ([chait@rockefeller.edu](mailto:chait@rockefeller.edu)).

## **Determining the architectures of macromolecular assemblies by integrating spatial restraints from proteomic data**

**Frank Alber\*, Svetlana Dokudovskaya\*, Liesbeth M. Veenhoff\*, Wenzhu Zhang, Julia Kipper, Damien Devos, Adisetyantari Suprpto, Orit Karni-Schmidt, Rosemary Williams, Brian T. Chait, Michael P. Rout, and Andrej Sali**

## **The molecular architecture of the nuclear pore complex**

**Frank Alber\*, Svetlana Dokudovskaya\*, Liesbeth M. Veenhoff\*, Wenzhu Zhang, Julia Kipper, Damien Devos, Adisetyantari Suprpto, Orit Karni-Schmidt, Rosemary Williams, Brian T. Chait, Andrej Sali, and Michael P. Rout**



# 1 Biochemical Materials and Methods

## 1.1 Immuno-Electron Microscopy of the Tagged Nups

### 1.1.1 Immuno-Electron Microscopy

A diagram of the NPC showing a nomenclature for its main features is shown in Supplementary Figure 1. Yeast strains expressing C-terminal PrA-tagged nups from genomic loci were constructed as described<sup>1</sup>. The C-terminus of Pom152-PrA (and therefore the tag) is in the NE lumen and thus inaccessible to antibodies<sup>2</sup> (data not shown). As permeabilization of isolated NEs risked compromising their fragile architecture, the more robust Pom152-PrA nuclei were instead permeabilized and successfully immunolabeled<sup>3</sup> (Supplementary Figure 2b). All other PrA-tagged nups were localized by immuno-electron microscopy (immunoEM) on pre-embedding labeled nuclear envelopes isolated from tagged strains<sup>1</sup>. The isolation of nuclei and nuclear envelopes from those tagged strains was performed as described<sup>4</sup>. NEs were extracted with low concentrations of heparin prior to immunostaining<sup>1</sup>. A 50  $\mu$ l aliquot of tagged NE was diluted with 200  $\mu$ l of 10 mM Bis-Tris (pH 6.5), 0.1 mM  $MgCl_2$ , 20% DMSO (BT-DMSO buffer), 1:500 dilution of solution P (2% PMSF, 0.04% pepstatin A in absolute ethanol) and either 0.01, 0.03, 0.1 or 0.3 mg/ml heparin, then incubated on ice for ~1 hour. The extracted NEs were pelleted by centrifugation at 100,000  $g_{max}$  for 20 min, 4 °C. The supernatant and pellet were processed for immunoblot analysis to detect the PrA tag<sup>1</sup>. The correct titer of heparin was determined to be the maximum concentration where the PrA-tagged protein was almost entirely retained in the pellet fraction (Supplementary Table 1)<sup>1</sup>; if heparin extraction was used, the immunolabeling pattern of at least two different concentrations of heparin were compared to ensure the localizations were not significantly altered.

Immunostaining and fixation of the NEs was performed as previously, with minor modifications<sup>5</sup>. Yeast NEs were diluted with 3 volumes of BT-DMSO buffer containing the appropriate concentration of heparin (above) and incubated on ice for 1 hour. The extracted NEs were centrifuged onto the bottom of a well of a microtiter plate coated beforehand with 2.5% glutaraldehyde and polylysine. The NE pellet was fixed for 5 min at 25 °C with 4% formaldehyde in 1.25 M sucrose-BT-DMSO solution, quenched with 50 mM  $NH_4Cl$  in BT-DMSO buffer, and blocked with immunoEM buffer (0.5% BSA, 0.5X PBS, 10  $\mu$ M  $CaCl_2$ , 10  $\mu$ M  $ZnCl_2$ , 10 mM  $MgCl_2$ , 0.05%  $NaN_3$ ). Whole nuclei from the Pom152-PrA strain were permeabilized with 0.03% Triton X-100,

0.03% sodium N-lauroyl-sarcosine in BT-DMSO buffer before pelleting onto the microtiter wells<sup>3</sup>. For immunostaining, a primary antibody (1:20 dilution in immunoEM buffer of rabbit anti-mouse IgG (ICN/Cappel #55480) pre-adsorbed onto fixed, permeabilized yeast cells to reduce background<sup>6</sup>, was added and incubated overnight at 4°C. After washing three times in immunoEM buffer, the bound primary antibody was visualized with an overnight incubation of a 1:10 dilution of 5 nm gold anti-rabbit IgG (Janssen Life Sciences) in immunoEM buffer followed by further washing with the same buffer. Subsequent processing of immunolabeled NEs and nuclei for electron microscopy was performed as described<sup>5</sup>. Sections were viewed and photographed unstained using a JEOL 100 CV Electron Microscope operated at 80 kV with 33,000X magnification. To avoid experimenter bias, the immunolocalization of the nups was conducted double-blind; the identity of the nup was not known during imaging.

### 1.1.2 Image Processing and Extracting the Estimated Position of the Nups in the NPC

Negatives containing labeled NEs were scanned into a Macintosh computer using an AGFA Arcus II flatbed scanner at 300% magnification, 300 lpi resolution with Adobe Photoshop v. 4.0.1 software. Regions of the NE containing gold-labeled NPCs were selected with circles corresponding to 200 nm in diameter (diagrammed in Supplementary Figure 2). The center of the circle was manually aligned with the intersection point of the central Z-axis and equatorial plane of the NPC. Only NPC images meeting the following criteria were chosen for further analysis: they had to be specifically labeled, sectioned parallel to the central Z-axis, display no distortion or gross morphological alteration of the surrounding NE, have clear NPC morphology, and have unambiguous nucleocytoplasmic orientation. Batches of 20 such NPC images were manually aligned (using their NPC central axes and equatorial planes) to generate a montage; the position of every gold particle in each montage was measured from both the central Z-axis of the NPC (x) and from the equatorial plane of the NE (z) using NIH Image v.1.6.2 (Supplementary Figure 3a). This process was repeated until approximately 300 particle positions were collected for each nup (Supplementary Table 1). Our micrographs do not have sufficient resolution to define the spoke positions; thus, we cannot establish a gold particle's azimuthal angle. Moreover, the scarcity of *en face* views necessitates the use of side views, which determine only the apparent radial values. Each montage also demonstrates a high degree of gold particle scatter, due to the rotation of the antibodies around the tags, the inherently flexible nature of NPCs and NEs<sup>7</sup>, distortion or damage of NPCs during preparation, inaccuracies in the alignment of the individual NPC images, and rotation of the NPCs around their central axes in the case of the x values.

To account for this scatter, we first created a scenario where the position of the PrA antigen was known, by coating a flat plastic surface with PrA (Supplementary Figure 9). To a first approximation, each face of an NPC (embedded in the NE) presents a topology that is similar to the control plastic surface; like this surface, antibodies attached to a PrA-tagged nup cannot penetrate the NE membranes or the dense structure of the NPC, but can freely rotate above them. The position of ~5,000 gold particles above the labeled surface was mapped, and the resulting distribution was modeled by a Gaussian function with a standard deviation of 11 nm, similar to previous observations<sup>8</sup>.

Next, the mean axial (*Z*) and radial nup positions (*R*) of a nup were estimated by sliding model Gaussian distributions along the *Z*- and *R*-axes, so as to maximize an overlap between the calculated and empirically determined distributions:

$$\text{Overlap} = 1 - \frac{1}{2} \sum_{i=1}^{N_b} |E_i - C_i|$$

where *E* and *C* are respectively the experimental and calculated distributions normalized to sum to 1, and *N<sub>b</sub>* is the number of the 2.5 nm bins spanning the distributions from 0 to a large positive value. The calculated distribution is simulated by 50,000 points with a Gaussian distribution and standard deviation of 11 nm (using only positive final values). A perfect overlap between the experimental and calculated distributions is 1 while no overlap between them is 0 (Supplementary Figure 3B). To determine the mean radial position of the tagged nup from its projection on the *R*-axis, it was necessary to account for the random rotation of the NPC around its central *Z*-axis before the maximization of the overlap. This was achieved by multiplying the calculated radial position for each of the 50,000 gold particles by the cosine of a random angle between 0° and 90° (Supplementary Figure 3C).

Finally, it was necessary to correct for localization errors presumed due to the steric hindrance of the NE and NPC (see above). With the exception of Nup1 and Nup60, the immunolabeled PrA tags were attached to structural (i.e., non-FG repeat) domains of all the nups. These structural portions must be embedded in the structure of the NPC as visualized in the cryoEM map. Hence, the total immunoEM map (excepting Nup60 and Nup1) was manually aligned to maximize its overlap with the cryoEM map<sup>9</sup>. Any mass density cues below 17 nm radius in the cryoEM map were excluded, because the central pore diameter of the NPC is ~35 nm (such that no structural nup domains could be anchored at a smaller radius)<sup>9,10</sup>. The maximum overlap between the immunoEM and cryoEM maps was obtained by adjusting the *R* and *Z* positions for each by adding 5 nm and subtracting 23 nm, respectively. Such a large *Z* adjustment is supported by the control PrA-coated

surface data (above), whose gold distribution in the labeled surfaces peaked at ~17 nm from the protein coat.

The resulting estimated *R* and *Z* positions are listed in Supplementary Table 2 (with their estimated allowed ranges), and diagrammed in Supplementary Figure. 8.

## 1.2 Affinity Purification of Nups and Their Neighbors

### 1.2.1 General Remarks

In this study, we have modified or developed several affinity purification methods to obtain a large variety of nup composites. The key steps of our methodology are as follows:

*PrA as an affinity purification tag.* PrA has a high affinity for IgG, but can be readily eluted using high pH, low pH, or denaturants, making it an excellent and widely used affinity purification tag<sup>11,12</sup>. In addition, we have also developed a strategy to release PrA from IgG using a competing peptide, which yields intact composites or pure native PrA-tagged nups in high yield<sup>13</sup>.

*Cell lysis and fractionation.* Different cell fractions were used as starting materials for the affinity purification procedure, increasing the likelihood of producing distinct composites. Thus, we isolated PrA-tagged nups from solubilized NE or highly enriched NPC fractions<sup>4,14</sup>. We also used lysates produced by whole cell cryolysis. This latter approach proved to be rapid and convenient, yielding high amounts of each composite because there are no subfractionation losses, and because proteolytic damage to the isolated proteins is minimized. As such, it became our major technique for generating composites.

*Extraction buffers.* Our base extraction buffer originates from conditions we used previously to affinity purify karyopherin-cargo composites, which are held together by relatively weak interactions<sup>15</sup>. We generated ~20 variants of this buffer with diverse extraction properties, allowing us to obtain nup composites of distinct compositions (Supplementary Table 3).

*Sepharose resin and magnetic bead isolation.* Initially, we used IgG Sepharose resins for isolation of the tagged composites, but we now use IgG-magnetic beads due to their superior speed, higher yield, and absence of an upper size limit for the isolated composite<sup>16</sup>. This optimized method allows for minimal processing of the fraction and short incubation times, which in turn (i) permits the isolation of labile composites that might otherwise dissociate during long incubation times, (ii)



further reduces proteolytic damage to the composites, and (iii) allows much larger composites to be isolated.

*Affinity purification from haploid and diploid strains.* The main reason for performing affinity purification in diploid *versus* haploid strains is the potential for detecting a potential second, untagged copy of a given nup in the composite. The presence of such an additional copy may indicate that the nup can form a dimer. Interestingly, we also found that some composites (e.g., Nup42-PrA (#82 and #77), Supplementary Figure 4) can be purified only from diploid strains.

### 1.2.2 Cell Disruption

Yeast strains, carrying PrA-tagged versions of nups<sup>1</sup> were grown in Wickerham media (0.3% Bacto Malt Extract, 0.3% Yeast Extract, 0.5% Bacto Peptone, 1% glucose) until early log phase, i.e.  $\sim 1 \times 10^7$  for diploid strains and  $\sim 2 \times 10^7$  for haploid strains<sup>4</sup>. Cells were harvested, washed twice with water and once with buffer containing 20 mM K-HEPES (pH 7.4), 1.2% PVP, 1mM DTT, 1:200 of protease inhibitor cocktail (PIC solution) (#P-8340; Sigma, St. Louis, Missouri, United States) and 1:200 of solution P. Cells were centrifuged at  $\sim 2,000$  g, 20 min at 4°C and the pellet was pushed through a plastic syringe into a 50 ml Falcon tube filled with liquid nitrogen, and the resulting frozen “noodles” were stored at -80°C. Cell lysis was achieved by cryogenic grinding of  $\sim 5$ g of “noodles” in a 25-ml stainless steel grinding jars and ball mill (#MM301; Retsch, Haan, Germany) in five sessions of 3 min at 30 Hz. The jars were cooled in liquid nitrogen between each grinding session. Typically >90% of the yeast cells are disrupted. Frozen ground cells were stored at -80°C.

### 1.2.3 Conjugation of Sepharose with IgG

3.75 g of CNBr-activated Sepharose 4B (Pharmacia #17-0430-01) were rehydrated in 30 ml of 1 mM HCl. The suspension was then filtered, washed once while on the filter with 0.1M NaHCO<sub>3</sub> (pH 8.5), transferred into 50 ml Falcon tube and centrifuged at  $\sim 500$  g<sub>max</sub>, 5 min. 50 mg of rabbit IgG (Cappel #55944) and 7.38 ml of 0.1 M NaHCO<sub>3</sub>, pH 8.5 were added to the resin and the suspension was incubated with rotation for 2h at room temperature or overnight at 4°C. The slurry was divided equally into four 50 ml Falcon tubes, centrifuged at  $\sim 500$  g<sub>max</sub>, 5 min, washed once with 40 ml of 1M ethanolamine pH 9, and incubated with rotation in this buffer for 2h at room temperature or overnight at 4 °C. The suspension was centrifuged and subsequently washed as follows: twice with 0.1 M NaHCO<sub>3</sub>, pH 8.5, 1 M NaCl; once with 0.1 M glycine, pH 2.8; once with 0.2 M glycine, pH 2.8; twice with water and twice with PBS. The resin was finally resuspended in an equal volume of PBS (10-12 ml) and NaN<sub>3</sub> was added to the final concentration 0.05%.

### **1.2.4 Affinity purification of Nup Composites from Nuclear Envelopes Using IgG-Sepharose**

Nuclear envelopes were prepared as described previously<sup>4</sup>. 0.6 –1.2 mg of NEs were used for a small-scale affinity purification, and 6-12 mg for a large scale preparation. NEs were diluted with ~0.4 volumes of cold 10 mM Bis-Tris (pH 6.5), 0.1 mM MgCl<sub>2</sub> (BT buffer) until the sucrose molarity has reached 1.2M (refractive index (RI) = 1.392). The NEs were pelleted by centrifugation for 1h at ~100,000 g at 4 °C. The NE-pellet was resuspended and diluted until reaching a final total protein concentration of 1 mg/ml in the appropriate cold extraction buffer (Supplementary Table 3) and vortexed 3 times for 1 min. The suspension was cooled on ice for at least 1 min between each mixing step. NEs in extraction buffer were incubated for 1h on ice, diluted 10 times with cold 20 mM K/HEPES (pH 7.4), 110 mM KOAc, 2 mM MgCl<sub>2</sub>, 0.1% Tween 20 (TBT buffer) and centrifuged at ~100,000 g<sub>max</sub> for 1 h. IgG Sepharose, pre-equilibrated with TBT, buffer was added to the supernatant (1 bed volume of Sepharose per 2.5 volumes of extract) and incubated overnight at 4°C on a rotating wheel. The resin was recovered by centrifugation at ~2,000 g<sub>max</sub> for 20 min at 4 °C, transferred to siliconized eppendorf tubes (Fisher, # 02-681-331) and washed 6 times with 1ml of TBT buffer. The resin was transferred to Bio-Spin columns (Bio-Rad #732-6008) pre-equilibrated with TBT buffer, and washed 2 times with 150 µl TBT buffer and 2 times with 150µl of 100 mM NH<sub>4</sub>OAc (pH 7.4). Proteins were eluted at room temperature with 300 µl of 0.5 M acetic acid (pH 3.4). The eluant was lyophilized, resuspended in buffer A (0.5 M Tris-base, 5% SDS) followed by addition of an equal volume of buffer B (75% glycerol, 125 mM DTT, 0.05% bromphenol blue), separated on a 4-20% Tris-glycine gel (Invitrogen #EC60255), and visualized with Coomassie blue (R-250) staining.

### **1.2.5 Affinity Purification of Nup Composites from Nuclear Pore Complex**

#### **Preparations**

Enriched fractions of nuclear pore complexes were prepared as previously described<sup>17,18</sup>. 100-300 µl of NPC fraction was used for the small-scale affinity purifications, and 4 ml for a large-scale preparation. The NPC fraction was diluted 1.5 fold in the appropriate 1.6x extraction buffer (Supplementary Table 3) and sonicated 3 times for 5-8 seconds. The sample was cooled on ice between each round of sonication. The extraction mixture was further incubated on ice for 1h, then diluted 10 times with TBT buffer, containing 20% DMSO and centrifuged at ~100,000 g<sub>max</sub> for 2h. IgG Sepharose pre-equilibrated with TBT buffer was added to the supernatant (10 µl of bed-volume per 15 ml of extract) and incubated overnight at 4°C on a rotating wheel. The resin was

recovered by centrifugation at  $\sim 2,000\text{ g}_{\text{max}}$ , for 20 min at 4°C, and further processed as described above.

### 1.2.6 Conjugation of Magnetic Beads with IgG

Magnetic beads (2.8  $\mu\text{m}$  Dynabeads M-270 Epoxy (Invitrogen #143-02-D)) were conjugated to rabbit IgG (ICN #55944 or Sigma #I5006) according to the manufacturer's instructions with the following modifications. The magnetic beads were resuspended in 0.1 M sodium phosphate buffer, pH 7.4 to give a concentration of  $\sim 19\text{ mg/ml}$  (or  $\sim 10^9$  beads per ml) and incubated for 10 min at room temperature while rotating. The wash with 0.1 M sodium phosphate buffer was repeated and beads were collected on the magnet. 1 mg of IgG per 15 mg of beads were used for the conjugation reaction. Prior to the conjugation, the necessary amount of lyophilized IgG was resuspended in water to a concentration of 17 mg/ml and centrifuged for 10 min at  $\sim 16,000\text{ g}_{\text{max}}$  4°C to remove aggregates. The supernatant was subsequently diluted with 10.3 volumes of 0.1 M sodium phosphate buffer pH 7.4 and 5.7 volumes of 3M ammonium sulfate to final concentrations of 60 mM phosphate buffer, 1M ammonium sulfate and 1 mg/ml IgG. 1 ml of the resulting IgG solution was added per 15 mg of washed beads. The reaction mixture was incubated overnight at 30°C while rotating. Coated magnetic beads were quickly washed once each with 100 mM glycine-HCl (pH 2.5), 10 mM Tris-HCl (pH 8.8), 100 mM freshly prepared triethylamine, four times with PBS and once with PBS plus 0.5% Triton X-100. Finally the beads were incubated for 15 min in PBS, 0.5% Triton X-100 at room temperature while rotating and then washed once with PBS. The beads were resuspended in PBS containing 0.02%  $\text{NaN}_3$  to a final concentration of 150  $\mu\text{g}/\mu\text{l}$  (or  $10^7$  beads per 1  $\mu\text{l}$ ) and stored at 4°C.

### 1.2.7 Affinity Purification of Nup Composites from Whole Cell Lysates Using

#### Magnetic Beads

0.5g of frozen ground cells were resuspended in 4.5 ml of the appropriate extraction buffer (Supplementary Table 3) with 1 mM DTT, and 1:200 dilutions of solution P and PIC solution. The suspension was homogenized with a Polytron PT-K (Kinematica, Switzerland) for 30 sec at 5.5 setting and the soluble fraction was isolated by centrifugation at  $\sim 2,000\text{ g}_{\text{max}}$ , 4°C for 15 minutes. This soluble extract was incubated with gentle agitation on a rocking platform for 1 hour at 4°C with  $\sim 8 \times 10^8$  of IgG-conjugated magnetic beads per gram of cells. The magnetic beads were then collected with a magnet, washed five times with 1 ml of the specified ice-cold wash buffer (Supplementary Table 3) and once with 0.1 M  $\text{NH}_4\text{OAc}$ , 0.1 mM  $\text{MgCl}_2$ , 0.02% Tween-20. The protein composites were eluted with 0.5 M  $\text{NH}_4\text{OH}$ , 0.5 mM EDTA by incubation for 20 min at room

temperature. The eluant was lyophilized, resuspended in SDS-PAGE sample buffer (as above), separated on a 4-20% Tris-glycine gel, and visualized with Coomassie blue (R-250) staining.

### 1.2.8 Mass Spectrometric Identification of Co-purifying Proteins

Identification of proteins was performed as described<sup>19</sup>. Examples of SDS-PAGE separations of each of our composite isolations, together with their protein identifications, are shown in Supplementary Figure 4 and listed in Supplementary Table 3. Proteins that directly interact with the tagged proteins should be approaching stoichiometric amounts in the purified composites. Conversely, distally associating proteins will be less abundant than the intermediate proteins they require for association. Thus, we have concentrated on identifying only the more abundant proteins in any given affinity purification (i.e., the observable Coomassie-stained bands), as these will generally be the most vicinal associates of the tagged protein. Small molecules and polypeptides below ~20 kDa were excluded from this analysis for technical reasons; however, due to their small size, their exclusion will not significantly affect the calculation of our structure.

### 1.2.9 Analysis of Nup Composites

A subjective quality criterion was assigned to each affinity purification based on the strength of the associating proteins compared with both the tagged nup and the “background” (i.e., obvious non-nup proteins) (Supplementary Table 3). Only those affinity purifications that met the high quality criteria were considered, and their relative quality was taken into account in our final structure calculations.

Composites containing FG nups also sometimes copurified with kaps and other transport factors, as expected<sup>20</sup>. The composites were also occasionally contaminated with common impurities such as heat shock proteins, metabolic enzymes, chaperones, translational factors, and ribosomal proteins. These proteins are abundant in the cell, and often found in other studies of protein-protein interactions<sup>16,21</sup> and, therefore, cannot be considered as specific.

The immunoisolated material varied in complexity from heterodimers (e.g., #1, #7, #10) to composites containing 20 proteins (e.g., #82) (Supplementary Figure 4, Supplementary Table 3). Other than being isolated from different PrA-tagged proteins, some of these composites are essentially identical; they are certain dimers (#1 and #17, Gle2/Nup116; #9 and #10, Gle1/Nup42; #7 and #15, Seh1/Nup85), composites containing nups within the Nup84 complex (#20 and #25; #45 and #51; #53, #54 and #57; #63 and #66), and composites containing nups in the Nsp1 complexes (#19 and #23; #27, #28 and #29).



## 1.3 Direct Contacts from Overlay Assays

### 1.3.1 General Remarks

Overlay assays were used for the identification of direct protein-protein contacts within the NPC. In these assays, the PrA-tagged nups were immobilized on nitrocellulose ('baits', Supplementary Figure 5A) and then probed with purified, native, biotin-labeled PrA-tagged nups ('probe') (Supplementary Figure 5B). Alternatively, overlay assays were performed on protein preparations of purified NPCs that were fractionated by reverse phase chromatography in the first dimension and by SDS-PAGE in the second, and probed with purified PrA-tagged Nsp1.

### 1.3.2 Affinity Purification of Individual Native Nups

2.5 grams of frozen ground cells were thawed into 25 ml of the cold extraction buffer 20 mM K-HEPES (pH 7.4), 0.1 mM MgCl<sub>2</sub>, 1.0% Triton X-100, 0.5% sodium deoxycholate, 0.1% sodium N-lauroyl-sarcosine, 1 mM DTT, 1:500 dilution of solution P and a 1:500 dilution of PIC solution. Extraction buffer for Nup60 and Nup100 was the following: 20 mM Tris-HCl pH 8.5, 1.0% Triton X-100, 0.5% sodium deoxycholate, 0.3% sodium N-lauroyl-sarcosine, 1 mM EDTA, 1 mM DTT, 1:200 dilution of solution P and a 1:200 dilution of PIC solution. The suspension was homogenized with a Polytron and incubated for 1 hour at 4°C on a rotating wheel. The soluble fraction was isolated by centrifugation at ~2,000 g<sub>max</sub>, 4°C for 15 min followed by centrifugation for 1 hour at 280,000 g<sub>max</sub>. The extract was incubated overnight at 4°C with 25 µl of IgG-Sepharose (10 µl of bed volume resin per gram of cells). The resin was transferred to BioSpin columns (BioRad #732-6008), washed 6 times with 1 ml 20 mM K-HEPES (pH 7.4), 1 mM EDTA, 0.1% Triton X-100, 0.05% sodium deoxycholate, 0.01% sodium N-lauroyl-sarcosine and placed into eppendorf tubes. Nup60 and Nup100 were washed with the buffer 20 mM Tris-HCl, pH 8.5, 0.1% Triton X-100, 0.05% sodium deoxycholate, 0.03% sodium N-lauroyl-sarcosine, 1 mM EDTA, 1:200 solution P. The PrA-tagged proteins were eluted by incubation of the resin with 62.5 µl (2.5 bed volumes) of 440 µM Bio-Ox peptide, dissolved in 20 mM K-HEPES (pH 7.4), 1 mM EDTA, 0.1% Tween 20, for 2 hours at 4 °C and 2 hours at room temperature while rotating. The suspension was transferred to BioSpin columns (Bio-Rad #732-6008), which were place into eppendorf tubes, centrifuged at ~16,000 g<sub>max</sub> for 1 min. Glycerol was added to 10% and the proteins were stored at -20 °C.

### 1.3.3 Overlay Assays

Overlay assays were performed essentially as described<sup>15</sup>. 100 ng of each purified nup was resolved by 4-20% Tris-glycine SDS-PAGE and electrophoretically transferred to nitrocellulose membranes. After blocking for 1 hour with TBT, 5% milk, 2% BSA and a quick wash in TBT, 0.5% milk, 0.2% BSA, 0.5 mM DTT, 1:100 solution P the membranes were incubated overnight at 4 °C in 0.7 ml of the same buffer but with 2 µg of purified biotinylated nup (see below). The membranes were washed two times 10 min in TBT and once 10 min in TBT, 2% milk, 1% BSA, and incubated for 1 hour with a 1:3000 dilution of Streptavidin Horseradish Peroxidase conjugate in the same buffer. After washing once for 10 minutes and 3 times 5 minutes with TBT, the blots were developed using Lumi-Light Western Blotting Substrate (Roche, #12 015 2000 001). Alternatively, overlay assays were performed on protein preparations of purified nuclear pore complexes isolated as described<sup>14</sup> and fractionated by RP-HPLC in the first dimension and by SDS-PAGE in the second<sup>1</sup>. The proteins were transferred to a nitrocellulose membrane and probed with purified native PrA-tagged Nsp1 (see above). The signal on the nitrocellulose was matched with a stained SDS-PAGE gel and the corresponding band was analyzed by mass spectrometric analysis<sup>19</sup>.

### 1.3.4 Data Analysis

There were large differences in the level of nonspecific binding for each protein, baits as well as probes. As for each probe we expect that only a few baits potentially interact we in principle have some 25 baits as negative control for background binding of this particular probe. Similarly, from comparison between the 23 overlay assays we could determine whether some baits had a higher tendency to bind any probe. We have therefore quantified each probe-bait interaction relative to the average signal this probe gives with all baits (on one overlay blot) and relative to the average signal that a particular bait gives with all probes (on the different overlays).

The signal ( $S_{i,j}$ ) from each bait ( $i$ ) with a given probe ( $j$ ) was normalized to the general background signal across all  $n$  baits for that probe. A similar correction was applied for the general background signal for each bait across all  $k$  probes, to generate a normalized signal for each bait with each probe ( $Y_{i,j}$ ):

$$Y_{i,j} = \frac{S_{i,j}}{\left( \sum_{i=1}^{n_i} \sum_{j=1}^{k_j} S_{i,j} \right)}$$

Only a probe-bait interaction giving a signal more than 10 times above its average was considered significant (Supplementary Figure 5C, Supplementary Table 3).

## 1.4 Hydrodynamic Estimation of Nup Shapes

### 1.4.1 Sucrose Gradient Velocity Centrifugation

The shapes of the individual nups and the Nup84 complex were examined using velocity gradient sedimentation. Each nup was affinity purified from whole cell extracts (see section 3.2, above). The sedimentation coefficient of each nup was determined by comparison against a set of marker proteins in the same gradient.

5.1 ml Linear 5-20% (w/w) sucrose gradients in buffers as listed in Supplementary Table 4 were made using the BioComp Gradient Master. Layered on top of each gradient were, in a total volume of 100  $\mu$ l, 0.5  $\mu$ g of the different pure proteins isolated as described above together with marker proteins (2  $\mu$ g each) that were labeled with biotin following instructions of the manufacturer (Biotin-XX, sulfosuccinimidyl ester sodium salt, Molecular Probes #F-6347). The markers used were chicken egg white ovalbumin (3.6 S, Amersham Biosciences #17-0442-01), bovine serum albumin (4.3 S, Pierce #23209), *Saccharomyces cerevisiae* alcohol dehydrogenase (7.4 S, Sigma #A8656) and  $\beta$ -amylase from sweet potato (8.9 S, Sigma #A8781). The samples were centrifuged at  $\sim 300,000 g_{max}$  for the indicated time in an SW55 Ti rotor at 4  $^{\circ}$ C (Supplementary Table 4). 200  $\mu$ l fractions were collected from the top of the gradient and analyzed by SDS-PAGE and western blot analysis. The biotin tagged marker proteins and the PrA-tagged nups were detected with Streptavidin Horseradish Peroxidase conjugate (Molecular Probes #S-911) and rabbit anti-mouse IgG (ICN/Cappel #55480) / anti-rabbit Horseradish Peroxidase conjugate (Amersham #NA934V), respectively. Band intensities were quantified, and the peak fractions of the marker proteins were plotted as function of its sedimentation coefficient and fitted with a standard curve by linear regression ( $r$ -squared > 0.99 in all cases). Thus, for each gradient a calibration curve describing the sedimentation behavior of the marker proteins was generated and used to determine the sedimentation coefficient of the nup. The error in the measurement estimated from the width of the nup protein peak at half height is  $\sim 0.8$  S. The sedimentation behavior of Nup145N-PrA and Nup57-PrA could not be assayed using purified protein and in these cases cleared whole cell extracts of ground yeast cells were analyzed instead. Neither full length Pom152-PrA nor Nup116-PrA were analyzed; a C-terminal proteolytic fragment missing the N-terminal transmembrane spanning helix

from Pom152-PrA and a ~100 kDa C-terminal proteolytic fragment of Nup116-PrA were analyzed instead. For each protein, the S-value was used to generate a bead model of each nup (Supplementary Table 5).

Nup84 complex was purified from 10g of frozen cell powder using TBT, 0.5M NaCl as an extraction buffer and eluted from the IgG-Sepharose with the appropriate amount of Bio-Ox peptide (see above). The complex was loaded on 6.25-25% w/w sucrose gradient in 50 mM Tris-HCl, pH 8, 0.5 mM EDTA, 300 mM NaCl, 1 mM DTT, 1:100 solution P, 1:100 PIC and centrifuged for 7.5h at ~300,000 g. 200  $\mu$ l fractions were collected from the top of the gradient and the refractive index measured in each fraction. Fractions were precipitated with TCA, loaded on SDS-PAGE and visualized with Coomassie blue (R-250) staining.

## 1.5 Structure and Composition of the Pom Ring

### 1.5.1 Gradient Isolation of Pom Rings

The preparation of Pom rings follows a modification of a previously described protocol<sup>22</sup>. We examined a negatively stained preparation of pom rings by electron microscopy (Supplementary Figure 6). The size difference between the pom rings and other heparin extracted material allowed them to be segregated from each other on a velocity sedimentation gradient.

For each gradient, 1.0 ml of highly enriched NPCs<sup>14</sup> was first diluted with 1.7 ml of BT-DMSO buffer, then 0.3 ml of a stock 100 mg/ml heparin solution was added and the mixture vortexed for 30 sec before being incubated for 30 min at 25 °C, followed by a further incubation for 30 min at 0°C. The sample is then overlayed onto a centrifuge gradient in a Beckman SW55 tube, precooled to 4°C and containing a bottom cushion of 100  $\mu$ l of 1.75 M sucrose-BT-DMSO<sup>14,17</sup> beneath a continuous 4 ml gradient of 1.48 M sucrose-BT-DMSO to 0.5 M sucrose-BT-DMSO. The gradient was centrifuged at 4°C for 5 h at ~300,000  $g_{max}$ . The gradient was then sequentially fractionated into 14 x ~650  $\mu$ l aliquots. Proteins from these aliquots were recovered by TCA precipitation and resolved by SDS-PAGE; after staining with Coomassie blue, protein bands were manually excised and identified by mass spectrometry<sup>19</sup> (data not shown).



## 1.6 Quantitation of nups

Nup stoichiometry was determined as described previously<sup>1</sup>; a slightly modified version of this procedure was used to re-check the stoichiometry of Nup82 (Supplementary Figure 7).

## 2 Computational Methods

### 2.1 Theory

The complete process of structure determination can be seen as an iterative series of four steps: data generation by experiment, translation of the data into spatial restraints, calculation of an ensemble of structures by satisfaction of these spatial restraints, and an analysis of the ensemble to produce the final structure (Supplementary Figure 10). The structure calculation part of this process is conveniently expressed as an optimization problem, a solution of which requires three main components: (i) the representation of the assembly in terms of its constituent parts; (ii) the scoring function, consisting of individual spatial restraints that encode all the data; and (iii) the optimization of the scoring function, which aims to yield structures that satisfy the restraints.

### 2.2 Representation

#### 2.2.1 Four-level hierarchical representation of the NPC

In the NPC, we consider 30 protein types (nups) and their relative stoichiometries, leading to a total of 456 protein molecules<sup>1</sup> (Supplementary Table 5). CryoEM shows the NPC as a ring with an eight-fold rotational axis perpendicular to the NE plane (Supplementary Figures 1, 12)<sup>9</sup>. This symmetry indicates that the NPC is composed of 8 identical building blocks, termed spokes. Our immunoEM experiments have localized each nup to the nucleoplasmic, cytoplasmic, or both sides of the equatorial plane (Supplementary Figure 12b, Supplementary Table 5)<sup>1</sup>. Based on these observations, we formally represent the NPC composition and protein stoichiometry with a 4-level hierarchy, consisting of the whole NPC (*assembly*,  $A$ ), the half spoke (*unit*,  $U$ ), the nup (*protein*,  $P$ ), and bead (*particle*,  $B$ ) levels (Supplementary Figure 12).

Each of the eight half-spoke units  $U$  at the cytosolic side is composed of 27 different types of nups  $\tau$ , of which two are present in two copies each, totaling 29 protein instances. Similarly, each of the eight half-spoke units  $U$  at the nucleoplasmic side contains 28 protein instances of 25 different types. The cytosolic and nucleoplasmic half-spoke compositions and the corresponding protein stoichiometries are defined in Supplementary Table 5.

### 2.2.2 Protein representation

Every protein  $P$  is represented as a set of beads  $B$ , each with associated attributes (e.g., radius, mass) (Supplementary Table 5). The number of beads and their attributes determine the resolution (granularity) of the protein representation. The most detailed data about the shape of most nups come from hydrodynamic experiments. Accordingly, we approximate the coarse shape and volume of each protein with a linear chain of equally-sized beads that best reproduce the observed sedimentation coefficients (below)<sup>23</sup> and are also consistent with our fold assignments<sup>24</sup>. It is possible that protein conformations in the NPC differ from their conformations in solution. Partly for this reason, we represent each protein as a flexible chain, to allow for any conformation of the flexible bead chain, from maximally extended to maximally compact (“Protein chain restraint”); no specific conformation is explicitly enforced. The bead chain describes a protein at the highest resolution in our representation (the “root” representation  $\kappa = 1$ ).

As a convenient way of further representing their structure, each protein can be described by several additional representations  $\kappa$  that are derived from the “root” representation, but capture different aspects about the structural and biological properties of the protein (Supplementary Table 5). For instance, a representation may contain only a subset of beads from the root representation, as is the case for representation  $\kappa = 2$  (Supplementary Figure 12) that contains only beads corresponding to protein regions with defined native structures, while representation  $\kappa = 3$  represents the same regions with a single bead per protein. Here, we used up to 9 representations per protein (Supplementary Table 5).

Each protein is described simultaneously by several structural representations  $\kappa$  (Supplementary Table 5, Supplementary Figure 12). Each representation consists of a set of particles  $B_j^\kappa$  and their attributes, such as the particle radii, partial protein mass, and the Cartesian coordinates; the subscript and superscript indicate an instance and type, respectively. Except for the “root” representation ( $\kappa = 1$ ), the attributes of a particle are fully or partly derived from particle attributes of another representation of the same protein. For instance, the Cartesian coordinates of all particles in representations  $\kappa$  from 2 to 9 are calculated from the particle coordinates in  $\kappa = 1$ , either by inheriting the coordinates from one of the particles in the root representation or by averaging the positions of some or all particles in the root representation. A configuration of the assembly is defined by the specific values of the particle attributes of all particles in  $B$ .

The root representation  $\kappa = 1$  describes a protein at the highest resolution, as a chain of equally sized beads in contact with each other. The number of beads and their radii were determined to

best reproduce the data from hydrodynamics experiments. The Cartesian coordinates of all particles in representations  $\kappa$  from 2 to 9 are coupled to the particle coordinates in representation  $\kappa = 1$ , either by enforcing the particle coordinates to be identical to one of the particles in the root representation or by calculating the particle coordinate as the weighted average coordinate of a group of particles from the root representation.

*Globular protein domains* are represented at  $\kappa = 2$  as a subset of particles from  $\kappa = 1$  whose corresponding protein regions are predicted to fold into globular domains. These particles are identified by a manual coarse mapping of the predicted domains in the protein sequence onto protein particles at the root representation.

*Globular protein domains* are represented at  $\kappa = 3$  as a single particle corresponding to the average positions of all particles in  $\kappa = 2$ . The particle radius is calculated from the total associated mass for all particles in  $\kappa = 2$ <sup>25</sup>.

*Unstructured protein regions* are represented at  $\kappa = 4$  as a subset of particles from  $\kappa = 1$  whose corresponding sequence segments are predicted to be unstructured in their native state<sup>24</sup>. These unstructured regions include so-called FG-repeat motifs. These particles are identified by a manual coarse mapping of the unstructured segments onto protein particles at the root representation.

*Non membrane-spanning protein regions* are represented at  $\kappa = 5$  as a subset of particles from  $\kappa = 1$  that excludes all membrane-spanning regions of the membrane proteins Pom152, Pom34, and Ndc1. The trans-membrane helical regions are predicted by the program TMHMM<sup>26</sup>.

*Membrane-anchor particles* are represented at  $\kappa = 6$  by a single particle per protein (from  $\kappa = 1$ ), selected as the central particle in the membrane-spanning regions found only in Pom152, Pom34, and Ndc1.

*Representations  $\kappa = 7$  and  $\kappa = 8$*  are subsets of particles from  $\kappa = 1$  describing regions of the membrane-spanning proteins that are localized either at the perinuclear side ( $\kappa = 7$ ) or the pore side of the NE ( $\kappa = 8$ ) (Supplementary Figure 12b, Supplementary Table 5). The assignment of these particles is based on the experimentally determined localization of each protein's C-terminus relative to the NE (above; Supplementary Tables 2,7), sequence positions of the predicted transmembrane-spanning helices, and the parity of the number of transmembrane-spanning helices.

Homotypic interactions between two copies of Pom152 are restrained through representation  $\kappa = 9$ , thereby allowing only interactions between the C-terminal regions.



The NE pore provides a mould for the NPC. We represent the NE by a large number of partially overlapping, fixed, and hard spheres spanning the NE bilayer (located on a half-torus surface with torus radii of 54 and 12.75 nm, respectively) (Supplementary Figure 12). The spheres have a diameter of 4.5 nm, corresponding to the average thickness of the bilayer membrane. Together, these spheres form a continuous membrane layer effectively representing the overall dimensions of the NE pore as defined by EM<sup>9</sup> (Supplementary Figure 12).

### 2.2.3 Bead number and radius of the root representation from ultracentrifugation experiments

The root representation of a nup corresponds to a flexible chain of beads, whose number and diameter are chosen to be consistent with the measured sedimentation coefficient, as follows.

First, the frictional ratio of each nup was calculated from its sedimentation coefficient  $S_0$  (Supplementary I), the molar mass, and the partial specific volume (Eq. 1 in <sup>23</sup>). The mass and the partial specific volume were estimated from the amino acid sequence <sup>27</sup> and corrected for the mass and partial specific volume of the Protein-A tag.

Second, the molecular shape function  $P$  of a nup was calculated from its frictional ratio and hydration (Eq. 2 in <sup>23</sup>); the hydration is defined as the ratio between the mass of physically bound water molecules and the mass of the protein. We estimated hydration from the frictional ratio, relying on an empirical function derived from a benchmark dataset of 20 protein structures, each of known frictional ratio and hydration:

The estimated hydration values range from 0.2 (for a compact shape with a low frictional ratio) to 1.4 (for an elongated shape with a high frictional ratio; corresponding to 5 water molecules per amino acid residue).

The axial ratio of a prolate ellipsoid of revolution representing each nup was calculated by the inversion formula (Tables 1 and 3 in <sup>23</sup>). The number of beads per protein was then set to the nearest integer value of the axial ratio. The beads corresponding to the Protein-A tag were removed from the representation. Finally, the radii of the beads were scaled to reproduce the volume of the nup without the Protein-A tag. The distance between the two consecutive beads in the flexible chain of beads is constrained to the diameter of the bead.

## 2.2.4 Nuclear Envelope from cryoEM

The NE pore provides a mould for the NPC, described by EM<sup>9</sup> (Supplementary Figure 12). The shape and dimension of the NE are taken into account by specific restraints that allow membrane spanning protein beads to freely penetrate the NE layer while all other protein beads are prevented from doing so (“Excluded volume restraints” below).

## 2.3 Spatial restraints

### 2.3.1 Protein chain restraint

To represent each protein as a flexible chain of particles, distance restraints are imposed between two neighboring chain particles at protein representation  $\kappa = 1$  with the restrained distance being the sum of the two particle radii (Supplementary Table 5).

### 2.3.2 Excluded volume restraint

A harmonic penalty is imposed if the distance between two particles is smaller than the sum of their radii (Supplementary Tables 5 and 6). Excluded volume restraints between protein particles are imposed for representation  $\kappa=1$ ; excluded volume restraints between particles of proteins and the NE are imposed for representation  $\kappa=5$ , so that membrane spanning protein regions are able to freely penetrate the NE layer, whereas all other protein particles are prevented from doing so (Supplementary Tables 5 and 6).

### 2.3.3 Pore side restraints, perinuclear volume restraints, and membrane surface restraints

To localize membrane-spanning proteins relative to the NE half-torus surface, we imposed three types of restraints. Particles defined in representation  $\kappa=6$  are restrained to the NE surface. Particles defined in representation  $\kappa=7$  are restrained to be in the perinuclear volume by imposing an upper bound on the length of the normal vector between it and the NE surface (Supplementary Tables 5 and 6). Particles defined in representation  $\kappa=8$  are restrained to be on the pore side of the NE by imposing a lower bound on the length of the normal vector between it and the NE surface (Supplementary Tables 5 and 6).

### 2.3.4 Protein localization restraint

The coarse localization of the C-terminal region of each protein (except Sec13) was determined by immunoEM, relying on a gold-labeled antibody that specifically interacted with the localized protein through its C-terminal PrA tag (above) <sup>1</sup>. For each protein, the C-terminal particle (Supplementary Tables 5 and 6) was restrained by imposing lower and upper bounds on its Z- and radial-coordinates (Supplementary Figure 12, Supplementary Table 6). The upper and lower bounds along both coordinates were calculated from the location and spread of the gold distribution and corrected for the dimension of the PrA tag (above and Supplementary Table 7). These bounds restrict the protein C-termini to an interval of ~16 and 8 nm on the Z- and radial coordinate, respectively (Supplementary Table 7). This relatively low precision reflects the large variability of gold particle positions, typically due to errors in EM image alignment, linker protein flexibility, and possibly distortions in the NPC structure.

Trans-membrane-spanning regions of the three membrane proteins Pom152, Ndc1, and Pom34 were restrained to the surface of the NE (Supplementary Figure 12 and Supplementary Table 6). ImmunoEM localizes the C-terminal regions of these proteins to the pore side of the NE (Supplementary Figure 12, Supplementary Table 6) <sup>1</sup>. The localizations of the C-terminal regions can then be derived from the parity of the number of predicted trans-membrane spanning helices <sup>24</sup>. Accordingly, only the C-terminal region of Pom152 is restrained to the perinuclear volume between the inner and outer NE, whereas the C-terminal region of Pom34 as well as the N-terminal regions of Pom34, Ndc1, and Pom152 are restrained to the pore-side of the NE (pore side restraint) (Supplementary Table 6).

### 2.3.5 Restraints on protein-protein interactions

#### 2.3.5.1 Conditional restraints

An interpretation of the data in terms of a spatial restraint generally involves identifying the specific proteins that are restrained (i.e., structural interpretation) and the limitations on the possible values of the feature implied by the data. While the data above were straightforward to apply to specific proteins, we focus here on the data that cannot be uniquely assigned to specific proteins (i.e., their structural interpretation is ambiguous). For example, the knowledge that two proteins of specified type(s) are forming an interaction cannot be assigned to a single pair of proteins, if there are multiple copies of the same protein in the assembly. An interaction between any one of all possible pairs of protein instances would be consistent with the data. In general, any information about spatial features is ambiguous if it is derived from experiments that identify only types rather than

instances of proteins, multiple copies of the same protein are present in the assembly, and the spatial feature does not apply to all protein copies.

To use ambiguous information, we introduce “conditional restraints”. A conditional restraint considers all alternative structural interpretations of the data, each one of which corresponds to a set of one or more independent “optional restraints”. The selection of the best alternative interpretation (i.e., the optimal set of optional restraints) is a result of the optimization process. At each optimization step, only the most likely interpretation is chosen; as a result, only a subset of all optional restraints are activated in the scoring function, while all others are ignored. Formally, the activation of optional restraints is achieved through “operator functions”  $O$  that evaluate all optional restraints  $R$  based on the current NPC configuration  $C$  and return the subset of restraints  $O(R, C) = R^*$  that lead to the smallest total restraint violation (Optional restraints and Operator functions are defined below).

### 2.3.5.2 Potential interactions $I_{\alpha\beta}(\theta, s)$

All alternative interactions  $I_{\alpha\beta}(\theta, s)$  between two nup instances of types  $\tau=\alpha$  and  $\tau=\beta$  are defined per half-spoke  $U_s^\theta$  and include all pairwise combinations of all nup instances of type  $\alpha$  and  $\beta$  in  $U_s^\theta$  with all potential interaction partners; a nup is a potential interaction partner if it is assigned to the same or any of the four adjacent half-spokes (Supplementary Figure 13). This definition is based on the expected dimensions of the proteins compared to that of a half-spoke, preventing direct physical interaction between proteins from non-neighboring half-spokes (except potentially for the FG-repeat containing regions, whose interactions are not restrained).

Specifically, the list of all potential interactions is the union of two groups of possible interactions: The first group includes all pairwise combinations of all proteins of type  $\alpha$  in  $U_s^\theta$  with all potential interaction partners of type  $\beta$ . The second group contains all combinations of all proteins of type  $\beta$  in  $U_s^\theta$  with its potential interaction partners of type  $\alpha$  in half-spokes  $U_{s'}^{\theta'}$  where  $(\theta', s') \in N(\theta, s)$  (Supplementary Figure 13). Finally, the list of all possible interactions  $I_{\alpha\beta}(\theta, s)$  between proteins of type  $\alpha$  and  $\beta$  defined for half-spoke unit  $U_s^\theta$  is the union of both groups.



### 2.3.5.3 Optional restraints $R_j$

An interaction between a pair of nup instances is expressed by optional distance restraints  $r_i$  that encode what is structurally known about the interaction ( $R_j = \{r_i \mid i \in (1, 2, \dots, N_j)\}$ ). Because nup interactions have not been structurally characterized, we make no assumptions about their binding interfaces (*i.e.*, about the individual beads forming the contacts).  $R_j$  is therefore a set of all possible distance restraints  $r_i$  between any two particles, one from each of the two interacting nups (Supplementary Figure 14a). Each restraint  $r_i$  is an upper distance bound with the bound corresponding to the sum of the two interacting particle radii multiplied by a tolerance factor of 1.3. This tolerance factor was chosen in an *ad hoc* manner to allow for the uncertainty in the protein radii prediction and the possibility of indirect interactions through small absent proteins that may have not been detected.

In general, optional restraints are imposed on particles from protein representation  $\kappa = 2$ , assuming that nup interactions are only mediated through the structured protein regions and not by the unstructured FG-repeat containing regions. This assumption is reasonable because the unstructured FG-repeat containing regions are likely to mediate interactions with transport cargo and not with other proteins in the NPC<sup>28</sup>. The only exceptions are the interactions Gle1-Nup42 and Gle2-Nup116. Gle1 and Gle2 are RNA transport factors. Therefore, interactions of Nup42 and Nup116 with these proteins are presumably mediated through their FG-repeat containing regions<sup>29</sup> defined at representations  $\kappa = 4$ , whereas interactions with any other protein are imposed at protein representation  $\kappa = 2$ . We also assume that the homotypic interactions of Pom152 occur through their C-terminal regions and are therefore imposed at protein representation  $\kappa = 9$ ; all other interactions of Pom152 with any other protein are imposed at representation  $\kappa = 2$ .

### 2.3.5.4 Operator functions

Operator functions  $O_n^{type} : R \times C \rightarrow R^*$  return an activated subset of  $n$  restraints, depending on optional restraints  $R$  and the current NPC configuration  $C$ . Three different operator functions are used:

The *rank-and-select operator*  $O_n^{RS}$  evaluates optional restraints based on the current NPC configuration and activates the  $n$  restraints with the smallest restraint value.

The *add-and-select operator*  $O_n^{AS}$  evaluates optional restraints and sums up the restraint values for subsets of restraints, ranks these subsets based on the total restraint value, and activates the restraints in the subset with the smallest total restraint value.

The *minimal-spanning tree (MST) operator*  $O_n^{MST}$  defines a fully connected undirected graph  $G = (C, E)$  with protein types as nodes  $\mu$  and edges  $(\mu, \nu) \in E$  connecting all nodes  $\mu$  and  $\nu$ . Each edge is assigned a weight  $\omega : E \rightarrow W$  with  $\omega(\mu, \nu)$  corresponding to the value of the optional restraint associated with the edge. The Kruskal algorithm<sup>30</sup> is used to determine the minimal-spanning tree of  $G$ . The MST defines the acyclic subset of all edges that connect all of the nodes in  $G$  and whose total weight  $w(MST) = \sum_{(u,v) \in MST} w(\mu, \nu)$  is minimized. All optional restraints that are represented by edges in the *MST* are activated.

### 2.3.5.5 Protein contact restraint

Protein contact restraints (Supplementary Figure 14a, Supplementary Table 6) are imposed to ensure the observed binary interactions between 13 pairs of nups. First, for each nup pair, the list of all alternative interactions  $I_{\alpha\beta}(\theta, s)$  (Supplementary Figure 13) between all nup instances of the two types  $(\tau=\alpha, \tau=\beta)$  per half-spoke  $U_s^\theta$  is constructed (Supplementary Figure 14a). Each alternative interaction  $J \in I_{\alpha\beta}(\theta, s)$  is then translated into a set of optional distance restraints  $R_J = \{r_i \mid i \in (1, 2, \dots, N_J)\}$  (above) that encode what is known structurally about this particular interaction  $J$ .

At each step during the course of the optimization, the rank-and-select operator function  $O_n^{RS}$  is applied to the set of all optional restraints  $R_I = \{R_J\}$  using the current NPC configuration  $C$  as input. The restraint with the smallest value is activated for contribution to the scoring function, while all other restraints in  $R_I$  are ignored (Supplementary Figure 14a).

### 2.3.5.6 Competitive binding restraint

A competitive binding restraint enforces the binding of one copy of Nsp1 to either Nup82 or Nup49 as well as Nup57, but not to all three at the same time. Because Nup82 occurs only at the cytoplasmic side, the restraint is imposed only on proteins in the half-spoke unit  $U_s^\theta$  with  $\theta=1$  and  $s=1$ . A three-level restraint hierarchy is needed, as follows (Supplementary Figure 14b). All possible interactions  $I$  (Nup82-Nsp1, Nsp1-Nup49, and Nsp1-Nup57) are translated into sets of

optional restraints  $R_j$  (above). At each optimization step, the rank-and-select operator  $O_1^{RS}$  activates only one of the optional restraints per restraint set at the first restraint level. These restraints are grouped into subsets that describe the alternative solutions for the competitive binding. The add-and-select operator  $O_n^{AS}$  then activates all optional restraints in the group with the smallest total restraint value. The activated restraints are consistent with the conditional dependencies of competitive binding (Supplementary Figure 14b).

### 2.3.5.7 Restraint on protein connectivity in composites

Affinity purification experiments produce a list of proteins  $C_m = \{\tau_1, \tau_2, \dots, \tau_n\}$  that are co-purified with a tagged bait protein. This list is a subset of the assembly composition  $C_m \in T$  and is often referred to as a complex. However, we have termed this set of proteins a “composite” (Supplementary Table 3 and 8) to emphasize that the immunoprecipitation does not necessarily provide the composition of a single complex (*i.e.*, a set of proteins with defined composition, stoichiometry, and configuration). In fact, a composite could correspond to a mixture of different complexes, which all contain a tagged protein as a common member. A composite with more than two proteins does not explicitly contain information on direct protein interactions. For instance, composite data cannot provide information on how many proteins of each type are in a complex, how many direct protein interactions are formed between those proteins and which particular proteins (types and instances) are interacting in a complex. However, we can infer the minimal necessary number of interactions between proteins listed in a composite: to be able to form at least one complex, at least one protein of each type in a composite  $C_m$  must be in contact with at least one protein of another type in  $C_m$ . A composite with  $n$  protein entries implies that there are at least  $n-1$  direct interactions connecting proteins of all types listed in the composite. We refer to this requirement as the connectivity condition of a composite; this condition applies to each composite  $C_m$  in each half-spoke  $U_s^\theta$ . The connectivity condition holds true even if the composite represents a mixture of complexes and does not depend on the copy numbers of each protein type in these complexes.

As for all described conditional restraints, the optimal connectivity between proteins in the composite is also determined as part of the structure calculation. To this end, all alternative protein connectivities are considered as optional solutions, and only the one with the smallest restraint violation is activated in the scoring function. To encode the conditional dependencies between the optional restraints, we constructed a three-layer restraints hierarchy (Supplementary Figure 15) that activates the set of  $n-1$  optional restraints that satisfy the composite connectivity condition and

lead to the smallest restraint violation, while all other optional restraints are ignored. The restraint activation is re-evaluated at each optimization step for each individual composite connectivity restraint.

The complex connectivity restraint is imposed for all 63 composites with more than 2 components (above and Supplementary Table 8), for proteins at the cytoplasmic half-spoke unit  $U_s^\theta$  (with  $\theta=1$  and  $s=1$ ) as well as the nucleoplasmic half-spoke unit  $U_s^\theta$  (with  $\theta=2$  and  $s=1$ ). For a composite  $C$  and half-spoke  $U_s^\theta$ , the interaction set  $I^{tot}$  includes all pairwise combinations of protein types in  $C$ :

$$I^{tot}(\theta, s) = \{I_{\mu\nu}(\theta, s) \mid (\mu, \nu) \in C \times C, \mu \neq \nu\}$$

where  $I_{\mu\nu}(\theta, s)$  is the list of all potential interactions of nup instances of types  $\nu$  and  $\mu$  (above and Supplementary Figure 13). Each individual potential interaction  $J \in I_{\mu\nu}(\theta, s)$  is then translated into optional distance restraints  $R_J$  (above).

The complex connectivity restraint relies on a three-level restraint hierarchy (Supplementary Figure 15). At each optimization step, the rank-and-select operator  $O_1^{RS}$  activates only one of the optional restraints per restraint group  $R_I$  at the first restraint level. The resulting activated restraints are subjected to the MST operator  $O_n^{MST}$  at the next hierarchical level. The MST operator activates the  $n-1$  restraints (where  $n$  is the size of the composite) that satisfy the connectivity condition with the lowest total restraint value.

### 2.3.5.8 Complex diameter restraint

The maximal diameter of the complex  $C_{45}$  (corresponding to composite #45) (Supplementary Table 8) is enforced by the following conditional restraint (Supplementary Figure 16). At each optimization step, a lower bound distance restraint is activated between two protein particles that span the maximal diameter of the largest of all connected complexes that contain exactly one copy of each nup in  $C_{45}$ . The largest complex is identified as follows. Based on the current NPC configuration, a protein contact graph  $G=(C, E)$  is calculated, where each protein is a node  $u$  of type  $\tau$ , with  $\tau \in C_{51}$  and edges  $(u, v) \in E$  that connect only nodes  $u$  and  $v$  whose corresponding proteins are in contact with each other. A contact is defined if at least one distance between any two protein particles is smaller than the sum of their radii multiplied by a tolerance factor of 1.4. All

connected subgraphs  $\{g_i\}$  of  $G$  with exactly one node of each type  $\tau$  are identified using a modified breadth-depth search algorithm, similar to the algorithm used to solve the modified graph-coloring problem<sup>30</sup>. Optional lower bound restraints are imposed between any two particles of two different nups that are part of the same subgraph  $g_i$ . All optional restraints are evaluated and only the restraint with the smallest restraint value is activated in the scoring function. The lower bound  $LB$  is defined as the  $LB = D - rad_1 - rad_2$ , where  $D = 19.2$  nm is the diameter of the complex derived by ultracentrifugation (including a 14% tolerance) (above), and  $rad_1$  and  $rad_2$  are the radii of the restrained protein beads (Supplementary Table 5). The contact graphs and the diameters of the complexes are calculated with protein representation  $\kappa = 2$ . For computational efficiency the complex diameter restraint was applied as a filter on pre-optimized structures.

### 2.3.5.9 Protein proximity restraint

We can approximate the maximal diameter of a hypothetical composite complex from its maximal possible mass, derived in turn from its maximal possible protein stoichiometry. We can then infer that for each pair of protein types identified in the composite at least one instance is within the upper distance bounds defined by the maximal composite diameter (Supplementary Table 6). For computational efficiency, we kept only the smallest upper bounds on pairwise distances when the same pair of protein types occurs in different composites. Conditional proximity restraints were only imposed for composites with more than four proteins. Although they are partially redundant relative to the composite connectivity restraints, they were used during the early stage of the optimization, when the composite connectivity restraints are not yet imposed, to facilitate finding good scoring configurations.

### 2.3.5.10 Maximal diameter of a complex

Using the PIBASE database of all structurally defined protein complexes<sup>31</sup>, we found empirically that for 95% of complexes the maximal dimensions along their longest principal axes are smaller than  $0.495 n^{1/3}$  nm, where  $n$  is the total number of residues in a complex. Due to the uncertainty in the determination of the complex mass, resulting from the uncertainty in the experimental determination of the complex composition, we scaled  $n$  with a “composite quality” factor  $q > 1$  (Supplementary Table 8). Lastly, to insure against over-interpretation of the maximal diameter data, the final estimate of the maximal diameter was scaled by a “tolerance factor” of 1.35. Thus, the maximal estimated diameter  $D_{\max}$  is  $1.35 \cdot 0.495 \cdot (q \cdot n)^{\frac{1}{3}}$ .



### 2.3.5.11 Symmetry restraint

The eight-fold and two-fold symmetries were effectively imposed through several independent restraints forcing two configurations of protein particles to be identical. Protein particles are selected from representation  $\kappa = 2$ , thereby excluding all unstructured FG-repeat containing regions. The symmetry is imposed by restraining the distance RMS (DRMS) between the two indexed sets of protein particles to 0<sup>32</sup>. The following pairs of protein sets were restrained:

(i) Eight pairs of protein sets from neighboring half-spokes within each half-spoke ring (Supplementary Figure 12), for a total of 16 such pairs, imposed for all proteins  $\tau \in T^\theta$ :

$$G = U_s^\theta \text{ and } G' = U_{(s \bmod 8)+1}^\theta \text{ where } \theta \in (1, 2) \text{ and } s \in (1, 2, \dots, 8)$$

(ii) Eight pairs of protein sets from half-spokes in opposite rings. Although the nup stoichiometries of the cytosolic and nucleoplasmic half-spokes differ from each other, most of the nups (22 out of 29 nup types) appear in both half-spoke types with identical stoichiometry. In particular, all scaffold nups are distributed equally, while some FG nups and Nup82 appear only on one of the two sides. We impose this symmetry restraint only between nups that are present with the same copy number in both half-spoke types:

$$G = U_s^\theta \text{ and } G' = U_s^{\theta \bmod 2 + 1} \text{ where } \theta \in (1, 2), s \in (1, 2, \dots, 8)$$

(iii) Eight pairs of protein sets from pairs of neighboring half-spokes located within the same ring, between all nups in a half-spoke:

$$G = \{U_s^\theta \cup U_{(s \bmod 8)+1}^\theta\} \text{ and } G' = \{U_s^{\theta \bmod 2 + 1} \cup U_{(s \bmod 8)+7}^{\theta \bmod 2 + 1}\} \text{ where } \theta \in (1, 2), s \in (1, 2, \dots, 8)$$

(iv) Eight pairs of protein sets from pairs of neighboring half-spokes located in the opposite rings, only between nups that are present with the same copy number in both half-spoke types:

$$G = \{U_s^\theta \cup U_s^{\theta \bmod 2 + 1}\} \text{ and } G' = \{U_s^\theta \cup U_{(s \bmod 8)+1}^{\theta \bmod 2 + 1}\} \text{ where } \theta \in (1, 2), s \in (1, 2, \dots, 8)$$

(v) One pair of protein sets  $G$  and  $G'$  that each contain proteins from both cytoplasmic and nucleoplasmic rings:

$G = \{^1U_s^\theta \cup ^2U_s^\theta\}$ , where  $^1U_s^\theta$  is the set of proteins with  $(\theta, s) = (1, 1)$ ,  $\tau \in T^1$ , and  $^2U_s^\theta$  is the set of proteins with  $(\theta, s) \in \{(2, 1), (1, 5), (2, 5)\}$  and  $\tau \in T^2$ .

$G' = \{^1U_s^\theta \cup ^2U_s^\theta\}$ , where  $^1U_s^\theta$  is the set of proteins with  $(\theta, s) = (2, 2)$ ,  $\tau \in T^1$ , and  $^2U_s^\theta$  is the set of proteins with  $(\theta, s) \in \{(1, 1), (2, 5), (1, 5)\}$  and  $\tau \in T^2$ .

$T^1$  and  $T^2$  are two equally large unique sets of proteins that appear with the same copy number on the cytoplasmic and nucleoplasmic sides, respectively. Moreover  $T^1 \not\subset T^2$ .

### 2.3.5.12 Chirality restraint

Identical chirality between two half-spokes in the same ring was achieved by restraining to the same value two dihedral angles defined by two arbitrarily chosen sets of 4 protein particles taken from 4 different proteins in representation  $\kappa = 2$  and located in two different half-spokes in the same ring. The same chirality of the protein configurations in the two opposing rings is achieved by restraining to the same value the dihedral angles defined by the same type of 4 protein particles located in opposing half-spokes. Based on our current data, we are not able to distinguish between the two possible mirror-symmetric solutions; here, we present one of these two solutions.

### 2.3.5.13 Phase angle restraint: restraints on the relative order of protein indices

Restraints were also imposed to ensure a relative order of the phase angles in the equatorial plane of the symmetry related proteins. Importantly, these restraints do not add additional spatial information about protein localizations or interactions; they merely facilitate the analysis of the optimized NPC configurations.

For convenience of the analysis of the resulting NPC configurations, we imposed restraints that ensure a defined relative order of the phase angles in the X-Y membrane equatorial plane of the symmetry related nups. Importantly, these restraints do not add additional spatial information about protein localization or interactions.

Upper bound distance restraints were imposed between the C-terminal protein particles  $B_j^{\kappa=1}(\theta, s, \tau, i)$  and  $B_j^{\kappa=1}(\theta, s \bmod 8 + 1, \tau, i)$  with identical values of  $\tau$ ,  $\theta$ , and  $i$ . The restraint was imposed for all proteins with  $\theta \in \{1, 2\}$  and  $s \in \{1, 2, \dots, 8\}$ . The upper bound was derived from the maximal upper bound along the radial coordinate determined by immunoEM for each protein type (Supplementary Table 7).

Moreover, restraints were imposed enforcing the localization of the N-terminal protein particles  $B_j^{\kappa}(\theta, s, \tau, i)$  and  $B_j^{\kappa}(\theta, (s+3) \bmod 8 + 1, \tau, i)$ , where  $\kappa=1$ ,  $\theta \in \{1, 2\}$ ,  $s \in \{1, 2, \dots, 8\}$ , to the diametrically opposite sides of the half-spoke ring. An angle restraint was imposed between three particles  $B_j^{\kappa}(\theta, s, \tau, i)$ ,  $l_{(s-2) \bmod 8 + 1}$ , and  $B_j^{\kappa}(\theta, (s+3) \bmod 8 + 1, \tau, i)$ , where  $\theta \in \{1, 2\}$  and  $s \in \{1, 2, \dots, 8\}$ . The particle coordinates of  $l_{(s-2) \bmod 8 + 1}$  are determined as the average particle coordinates

of the two particles  $B_j^{\kappa}(\theta, (s-2) \bmod 8 + 1, \tau, i)$  and  $B_j^{\kappa}(\theta, (s-2) \bmod 8 + 1, \tau, i)$ . The mean and the standard deviation of the restraint are  $180^\circ$  and  $0.1^\circ$ .

Angle restraints are imposed on three particles  $B_j^{\kappa}(\theta, s, \tau, i)$ ,  $l_{(s-2) \bmod 8 + 1}$ , and  $c$ , where  $\theta \in \{1, 2\}$ ,  $s \in \{1, 2, \dots, 8\}$ .  $c$  is the coordinate origin positioned in the center of the nuclear pore. The mean and the standard deviation of the restraint are  $90^\circ$  and  $0.1^\circ$ .

## 2.4 Modeling by optimization of the scoring function

We generate NPC structures by simultaneously minimizing the violations of all restraints, resulting in configurations that minimize the scoring function. In general, a number of different configurations may be consistent with the input restraints. The aim is to obtain as many structures as possible that satisfy all input restraints. The entire optimization is split into two stages (Supplementary Figure 11). First, 200,000 independently optimized configurations, starting each time from a different random bead configuration, are generated by a coarse sampling protocol. Each of the independent optimizations consists of an iteration of approximately ten thousand small shifts of protein beads guided by either conjugate gradient or molecular dynamics methods. Second, the best 10% configurations from the sampling stage are further optimized by a refinement protocol. Both sampling and refinement protocols have been constructed by trial-and-error, with the aim to obtain the maximal number of well-scoring models at a minimal cost in CPU time. The entire calculation takes 30 days on a 200 CPU cluster.

At each conjugate gradient or molecular dynamics step, the activation of optional restraints in all conditional restraints is reevaluated based on the current configuration, leading to the activation of only a subset of all optional restraints. For instance, out of the 120,393 optional restraints in the scoring function, maximally 1,130 restraints are dynamically activated at each optimization step (Supplementary Table 6).

## 2.5 Analysis

A *protein contact* at representation  $\kappa$  is present if the distance between any two protein particles is smaller than the sum of their radii multiplied by a tolerance factor of 1.4, which is slightly larger than the factor of 1.3 used in imposing interaction and connectivity restraints. The value of 1.4 is also justified by the highest accuracy of predicting native protein contacts in model systems similar to the NPC<sup>25</sup>.

The *contact frequency* of a pair of protein types ( $f_{obs}$ ) is the fraction of configurations in the ensemble that contain at least one protein contact between any protein instances of the two types.

The *contact instance frequency* of a pair of protein instances is defined as the fraction of configurations that contain a protein contact between the two protein instances.

The *initial contact frequency* of a pair of protein types ( $f_{ini}$ ) is the probability that the pair is part of the composite minimal connectivity. In a single composite, all possible protein contacts are *a priori* equally likely. Then,  $f_{ini}$  is equal to the number of times this protein pair is found in any of the composite's connectivities divided by the total number of all protein connectivities per composite:

$f_{ini} = \frac{n-1}{n(n-1)/2} = \frac{2}{n}$ , where  $n$  is the number of protein types in the composite. The same protein

pair may be present in multiple composites and only the largest resulting  $f_{ini}$  is considered.

We estimate the *statistical significance* of the difference between the observed contact frequency  $f_{obs}$  and the initial contact frequency  $f_{ini}$  for a given pair of proteins ( $\Delta f = f_{obs} - f_{ini}$ ). Based on the binomial distribution, the cumulative *P*-value for the null hypothesis that the initial frequency  $f_{ini}$  of a given protein pair is consistent with the observed contact frequency  $f_{obs}$  is:

$$P = \sum_{k=n+1}^N \binom{N}{N \cdot f_{obs}} f^{(N \cdot f_{obs})} (1-f)^{N(1-f_{obs})} \text{ with } f = \begin{cases} f_{ini} & \text{for } \Delta f > 0 \\ (1-f_{ini}) & \text{for } \Delta f < 0 \end{cases}$$

where  $N=1000$  is the number of models in the ensemble.  $\Delta f$  is judged to be statistically significant if the null hypothesis is rejected with a *P*-value  $> 0.001$ .

The most probable *composite minimal connectivity* is the set of all protein pairs defined in the minimal spanning tree of the fully connected composite graph with protein types as nodes and edge weights of  $1-f_{obs}$ .

A pair of proteins of type(s)  $\tau$  is defined as being *adjacent* if its observed contact frequency  $f_{obs}$  is larger than 65% or is part of a composite minimal connectivity.

In a *configuration* (ie, *3D adjacency graph*), each protein instance is represented as a node that is located at the protein's average position at representation  $\kappa=3$ . Edges are drawn between nodes if the corresponding pair of protein types is *adjacent*.

The *2D localization probability* (*2D-lp*) gives the probability of localizing proteins of type  $\tau$  at the radial (*R*) and axial (*Z*) coordinates (Supplementary Figure 12b). It is calculated on a discrete 50 x

50 grid with 1 nm spacing and normalized to sum to 1. The  $2D\text{-}lp$  is determined for proteins in half-spoke  $U_s^\theta$  (where  $\theta \in (1,2)$  and  $s = 1$ ) from representation  $\kappa = 3$  (one particle per structured domain) (Supplementary Table 5).

*Chirality.* The imposed distance restraints cannot differentiate between the two possible mirror images of the NPC. Thus, we transformed (by reflection) all good scoring configurations into one of the two randomly selected mirror images.

*Structural superposition.* To facilitate a structural analysis of the ensemble, we superposed all configurations on top of each other, using a previously validated procedure<sup>25</sup>. We first selected an optimal *template* configuration and then superposed each one of the other configurations on it using a modified pairwise rigid-body superposition that minimizes the RMSD between the superposed configurations<sup>33</sup>. The superposition template is the configuration whose protein contact map is most similar to the adjacency map of the NPC; the similarity between two contact maps is  $2a / (2a + b + c)$ , where  $a$  is the number of contacts that occur in both contact maps,  $b$  is the number of contacts present only in the first map, and  $c$  is the number of contacts present only in the second map. The rigid body superposition had to take into account that the NPC is structurally invariant to permutations of symmetry-related proteins  $P_i^\tau(\theta, s)$  (i.e., proteins that share the same set of indices  $\theta$  and  $\tau$ , but differ in their half-spoke index  $s$  and stoichiometry index  $i$ ). As a result, independent optimizations may lead to protein configurations that differ only in the permutation of their  $s$  and  $i$  indices but are otherwise identical. Therefore, we attempted to identify the permutation with the smallest RMSD by a deterministic non-exhaustive search. To facilitate this search, we imposed during structure optimization spatial restraints on the phase angles in the equatorial plane (Supplementary Figure 12, Supplementary Table 6) for the nups that differ only in indices  $s$ , such that the ordering of the nups by the phase angle is the same as that by  $s$ . As a result, the relevant permutations are restricted to cyclic permutations between protein particles from  $S = \{1, 2, \dots, 8\}$ . Moreover, only protein types that form the core structure of the NPC are superposed (i.e., Nup120, Nup192, Nup188, Nup170, Nup133, Nup84, Nup85, and Nup145C). Each protein is represented by a single particle corresponding to its structured domains (representation  $\kappa = 3$ ). The superposition is performed twice, independently for proteins in the cytoplasmic and nucleoplasmic half-spokes. The variability among the superposed configurations provides an upper bound on the precision of our structure determination.

The protein *localization probability* ( $3D\text{-}lp$ ) is the probability that a given volume element is occupied by the excluded volume of a particular protein; correspondingly, a given protein occupies multiple volume elements according to its size. It is calculated from all superposed configurations:



The volume of a protein is projected onto a discrete 100 x 100 x 100 grid with the spacing of 1 nm, with the coordinate system origin at (50, 50, 50). The 3D- $\rho$  is calculated with the aid of “pseudo-Gaussian” smoothing:

$$\rho_{\tau}(\mathbf{r}) = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^B (f + (1-f)e^{-\frac{1}{2}\left(\frac{dis}{\sigma(rad_j)}\right)^2})$$

$$\text{with } f = \begin{cases} 1 & \text{for } |\mathbf{r}_j - \mathbf{r}| \leq \frac{1}{2} rad_j \\ 0 & \text{for } |\mathbf{r}_j - \mathbf{r}| > \frac{1}{2} rad_j \end{cases}, \quad dis = \begin{cases} |\mathbf{r}_j - \mathbf{r}| & \text{for } |\mathbf{r}_j - \mathbf{r}| \leq \frac{1}{2} rad_j \\ |\mathbf{r}_j - \mathbf{r}| - \frac{rad_j}{2} & \text{for } |\mathbf{r}_j - \mathbf{r}| > \frac{1}{2} rad_j \end{cases}$$

where  $\sigma(rad_j) = 0.8493 \cdot \frac{rad_j}{2}$ ,  $\mathbf{r}$  and  $\mathbf{r}_j$  describe the positions of the grid voxel and the particle  $j$ , respectively,  $rad_j$  is the radii of particle  $j$  at representation  $\kappa$ ,  $B$  is the total number of protein particles of all proteins of type  $\tau$ , and  $N = 1,000$  is the total number of configurations in the ensemble.

The *localization volume* of a protein is calculated from its 3D- $\rho$ . The 3D- $\rho$  grid points are first sorted by their values. The localization volume then corresponds to the top grid points whose volume sums to the protein volume. The protein's volume was estimated from its sequence and the partial specific volumes of the 20 standard residue types<sup>27</sup>. The localization volume of a protein reveals its most probable localization, given the input spatial restraints. For visualization, the localization volume is blurred by averaging the values of direct voxel neighbors.

The *protein localization point* is defined for each protein as the location with the highest 3D- $\rho$ .

*Implementation.* To enable the computations described in this paper, we developed the RESTRAINER module that will be available as part of the *Integrative Modeling Platform* (IMP) software (<http://salilab.org/imp/>).

## 2.6 Assessment of precision and accuracy

### 2.6.1 Self-consistency of the experimental data

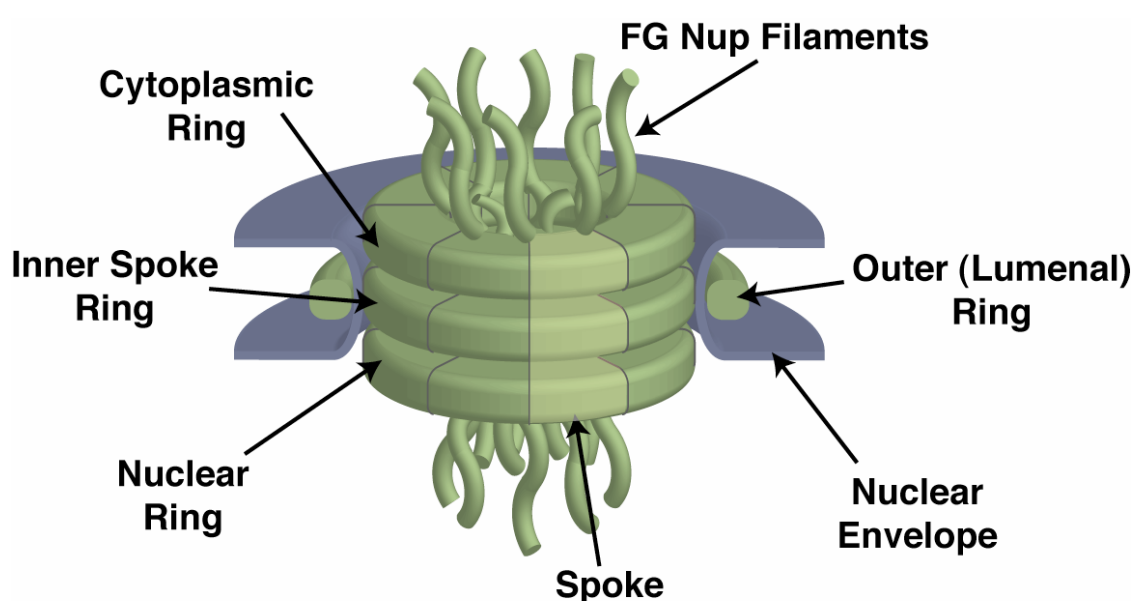
In our approach, inconsistencies in experimental data can be identified when only frustrated models that do not satisfy the input restraints are generated. This is not the case for our NPC

calculations. We find NPC models that satisfy all input restraints, demonstrating the self-consistency among our diverse experimental data and their interpretations in terms of spatial restraints (Supplementary Figure 17a). This observation increases our confidence in the data.

To demonstrate that it is not trivial to find models satisfying all restraints, we repeated the calculations with a comparable, but partly incorrect set of restraints. Specifically, all untagged proteins were randomly swapped between composites, leaving the number of composites and the number of proteins in each composite unchanged. All other restraints also remained unchanged. An optimization using this modified restraint set did not lead to any structures that satisfied all spatial restraints.

In another calculation, instead of using randomized composite data, we used protein-protein interactions from public depositories<sup>34</sup>. The public interaction data, sometimes generated on the large-scale, may contain false positive interactions<sup>35</sup>. As a consequence, we were unable to generate any models that satisfy all of the restraints (the best score was 12,100 in comparison to 0.5 for the ensemble) (Supplementary Figure 17b).

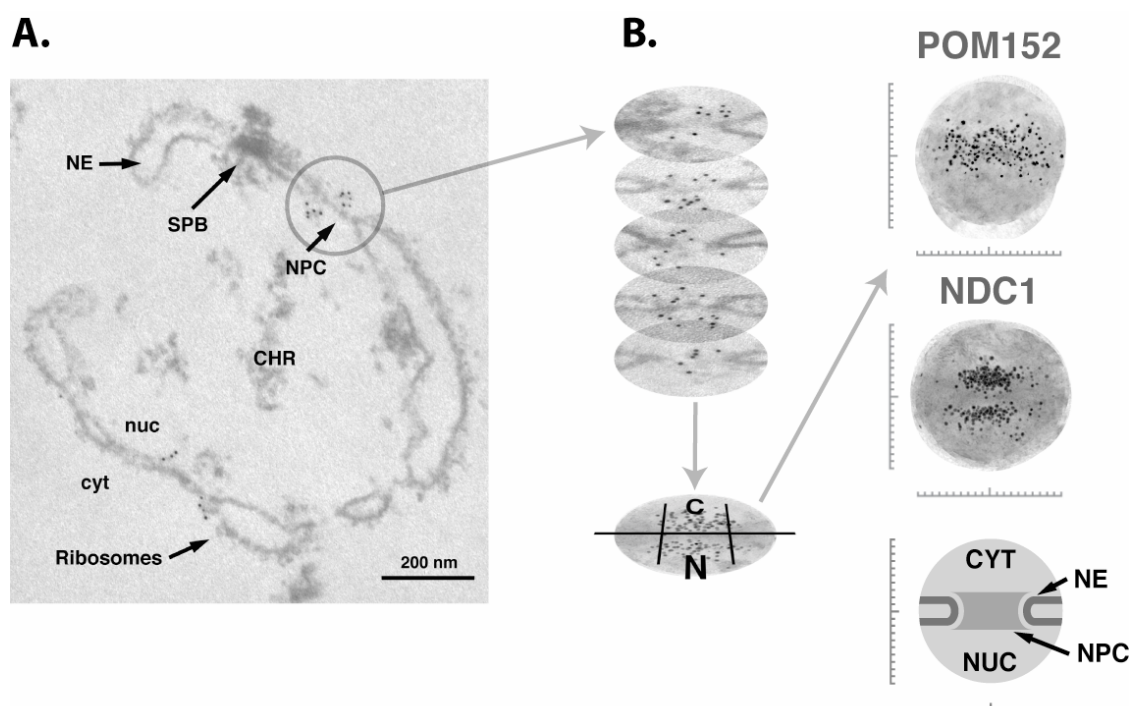
### 3 Supplementary Figures



### Supplementary Figure 1

**Supplementary Figure 1. Diagram of the main structural features of the NPC.**

Diagram of the main structural features of the NPC, showing the commonly-used published nomenclature. The nuclear basket has been omitted for clarity.



**Supplementary Figure 2.**

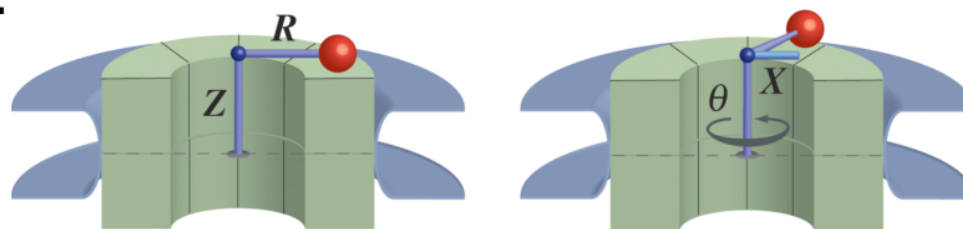
**Supplementary Figure 2. Electron micrographs of immunolabeled heparin-extracted NEs.**

(A). Single example of an immunolabeled NE from Nup116-PrA-tagged cells. NE preparation and immunolabeling did not compromise the integrity of the NE, which can be oriented by virtue of ribosomes on its cytoplasmic side (cyt), chromatin remnants (CHR) on its nucleoplasmic side (nuc), and embedded spindle pole bodies (SPB). The NPCs are specifically immunolabeled with 5 nm gold-conjugated antibodies.

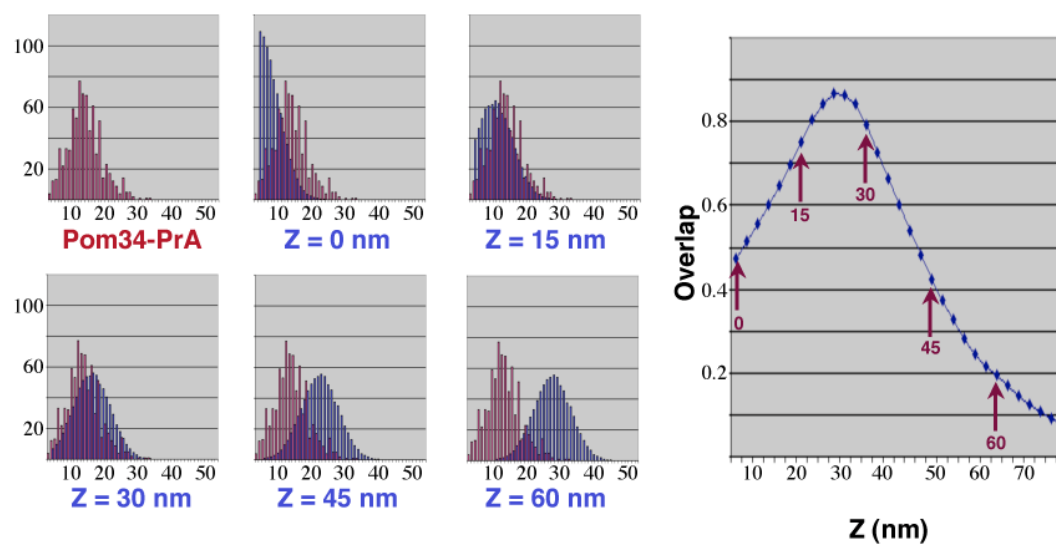
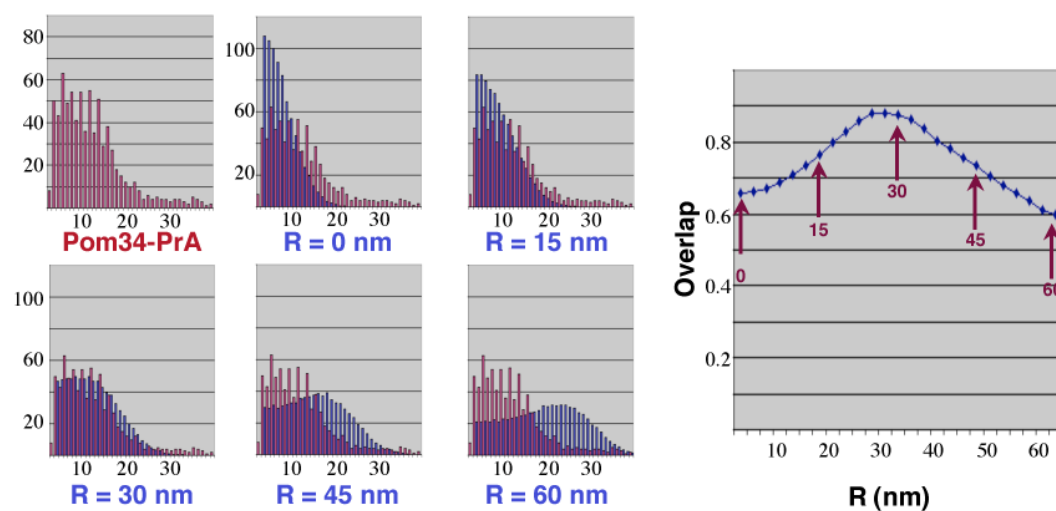
(B). Left panel: Regions of the NE containing immunolabeled NPCs were selected with circles 200 nm in diameter around the center of each NPC. These selections were superimposed upon one another and aligned to generate a montage. Right panel: The resulting montages for Pom152-PrA NEs and Ndc1-PrA NEs. Scale bars are graduated in 10 nm intervals, with the horizontal bar centered on the cylindrical axis of the NPC and the

vertical bar centered on the plane of NPC pseudo-mirror symmetry. The major features in each montage are diagrammed schematically at the bottom.



**A.**

$$X = R \cos \theta$$

**B.****C.**

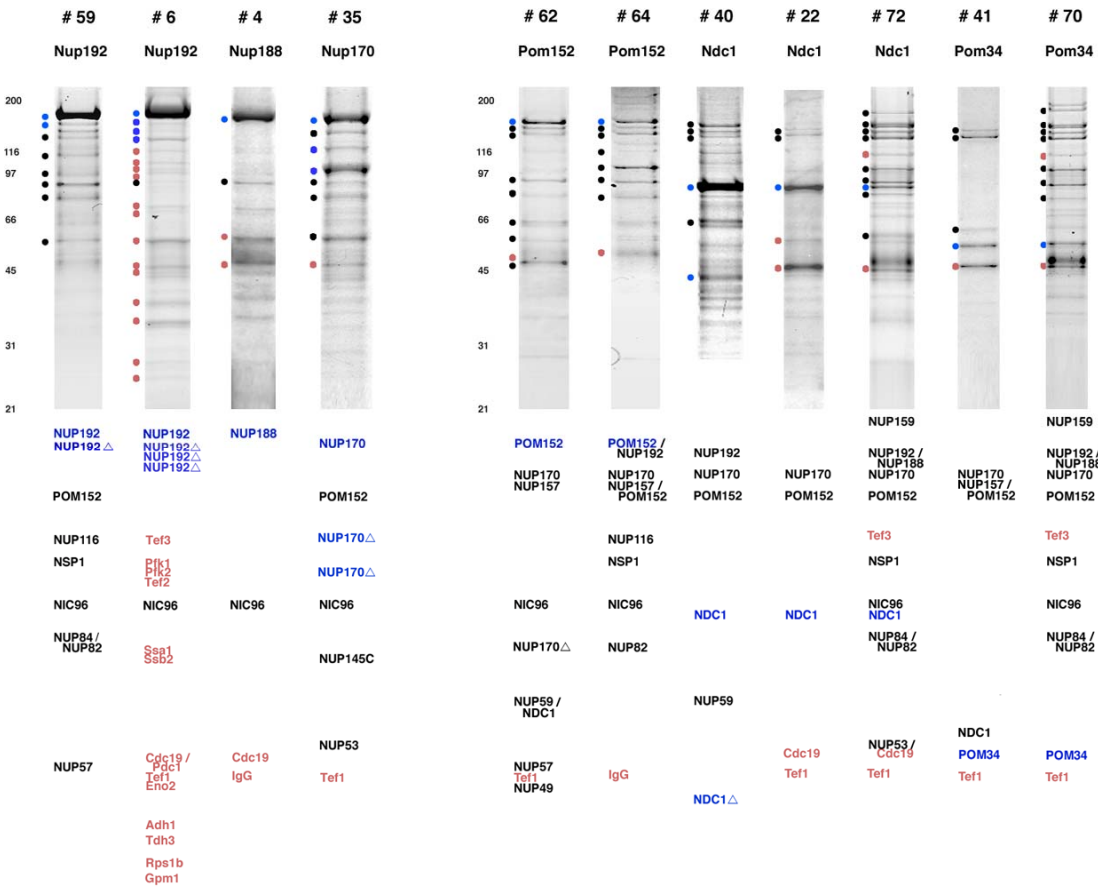
Supplementary Figure 3.

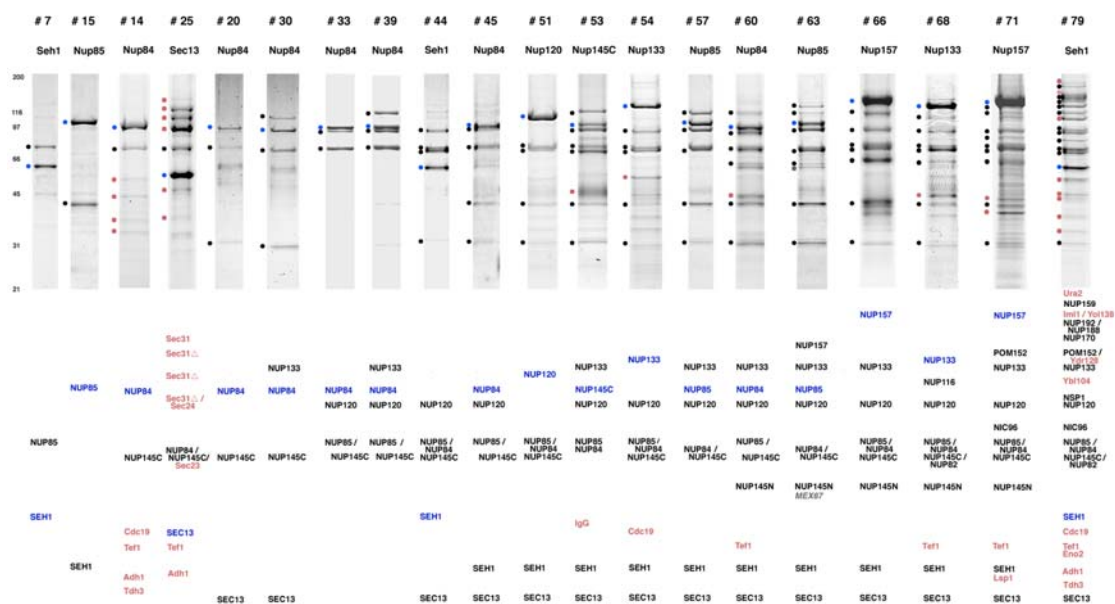
**Supplementary Figure 3. Estimation of the actual position of a tagged nup from the distribution of its immunoEM labeling.**

(A). Diagram illustrating a gold particle (red sphere) labeling an NPC. Left panel: Cross-section through an NPC (green) in which the position of a gold particle along the cylindrical axis ( $Z$ ) and radial position ( $R$ ) is indicated. Right panel: The position of the gold particle relative to the central  $Z$ -axis as viewed from a direction perpendicular to the section plane ( $X$ ) is dependent on the radial distance of the gold from the central  $Z$ -axis ( $R$ ) and the angle of rotation of the NPC ( $\theta$ ).

(B). Extracting the  $Z$  position of Pom34-PrA from the distribution of gold particles in the montage. The red histograms show the actual distribution of gold particles in the montage of Pom34-PrA, which is compared with the modeled distributions (blue histograms) of the gold at different distances from the equatorial plane of the NPC ( $Z$ ). The right panel shows a graph illustrating the degree of overlap between the actual and modeled distributions at different distances (Materials and Methods).

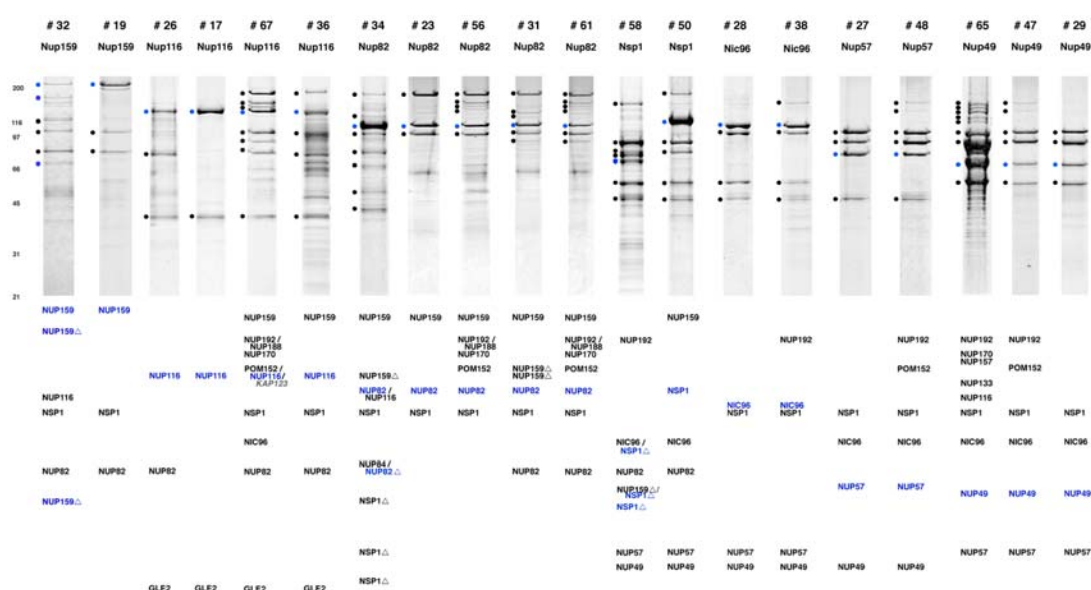
(C). Extracting the  $R$  position of Pom34-PrA from the distribution of gold particles in the montage. The red histograms show the actual distribution of gold particles relative to the central  $Z$ -axis in the montage, which is compared with the modeled distributions (blue histograms) of the gold at different distances from the central  $Z$ -axis of the NPC ( $R$ ). The right panel shows a graph illustrating the degree of overlap between the actual and modeled distributions at different distances.

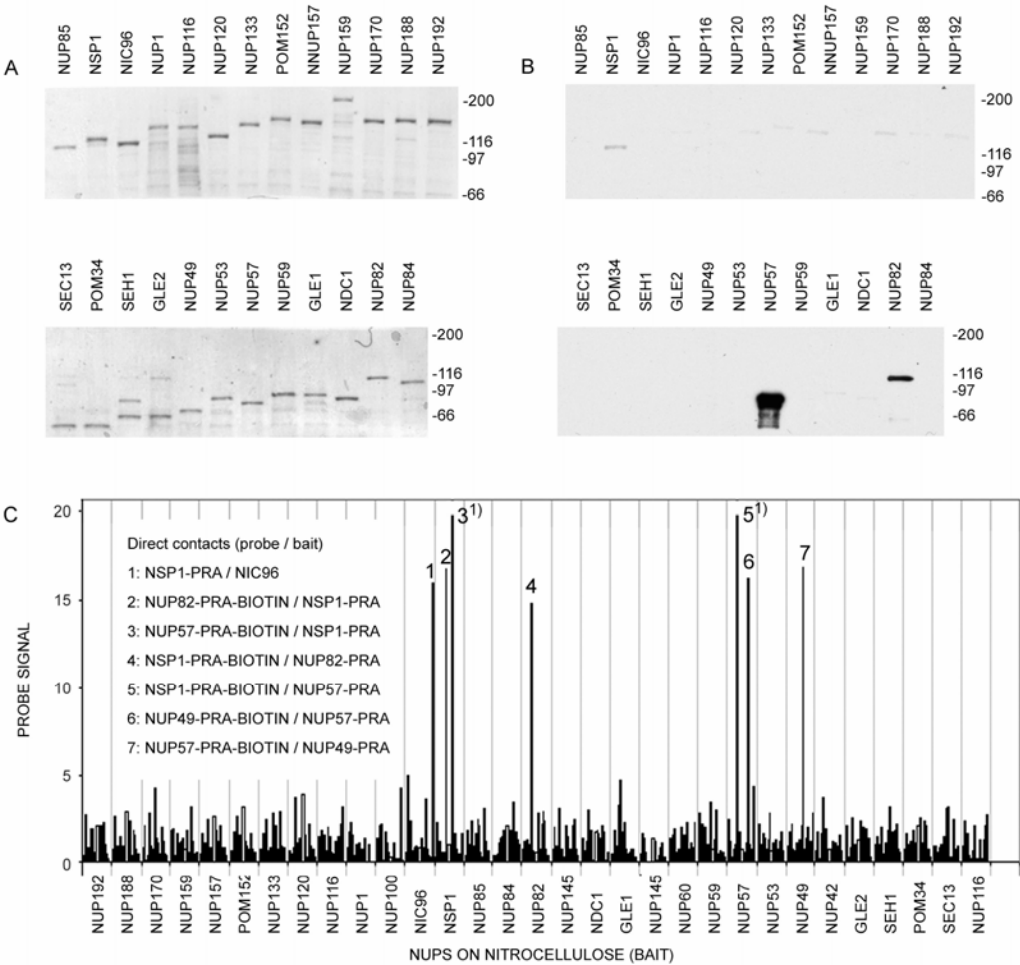












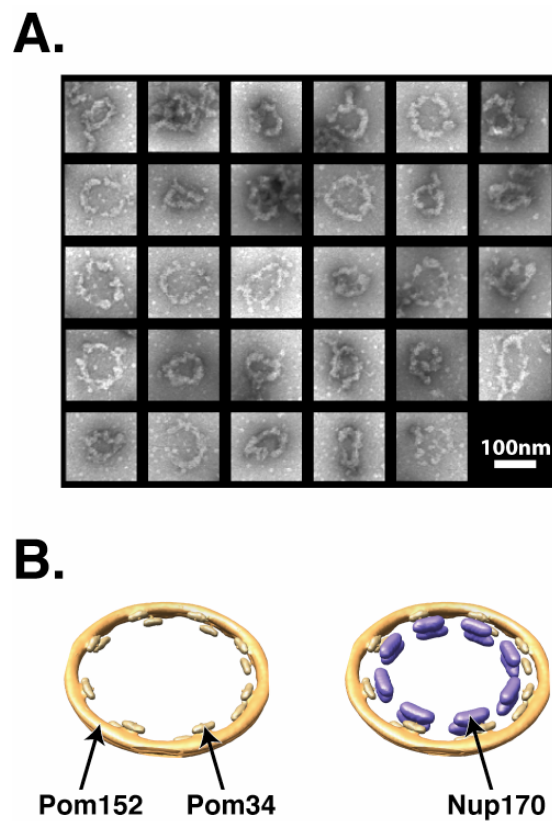
Supplementary Figure 5

Supplementary Figure 5. Overlay assays detecting direct nup-nup interactions.

(A). Nitrocellulose filter with Amido black-stained protein bands of purified PrA-tagged nups (bait). In the case of Seh1 and Gle2, two bands are visible as Seh1-PrA was purified together with Nup85 and Gle2-PrA with Nup116.

(B). Signal from specific binding of biotinylated Nsp1-PrA to Nup82-PrA and Nup57-PrA immobilized on nitrocellulose, detected by binding to streptavidin-peroxidase conjugate and conversion of a chemiluminescent substrate.

(C). Overview of all interactions probed in overlay assays. Plotted are the probe-bait interactions of 23 biotinylated PrA-tagged nups as probe on all nups as bait, and of Protein A tagged Nsp1 as probe on all proteins from purified NPCs. Each probe-bait interaction was quantified relative to the average signal with all baits on the overlay blot and relative to the signal this bait gives with all the probes on the other 22 overlays. Only a probe-bait interaction giving a signal more than 10 times above its average was considered significant. The signals 3 and 5 are larger than 20 but are not quantified as they are outside the linear range of detection.

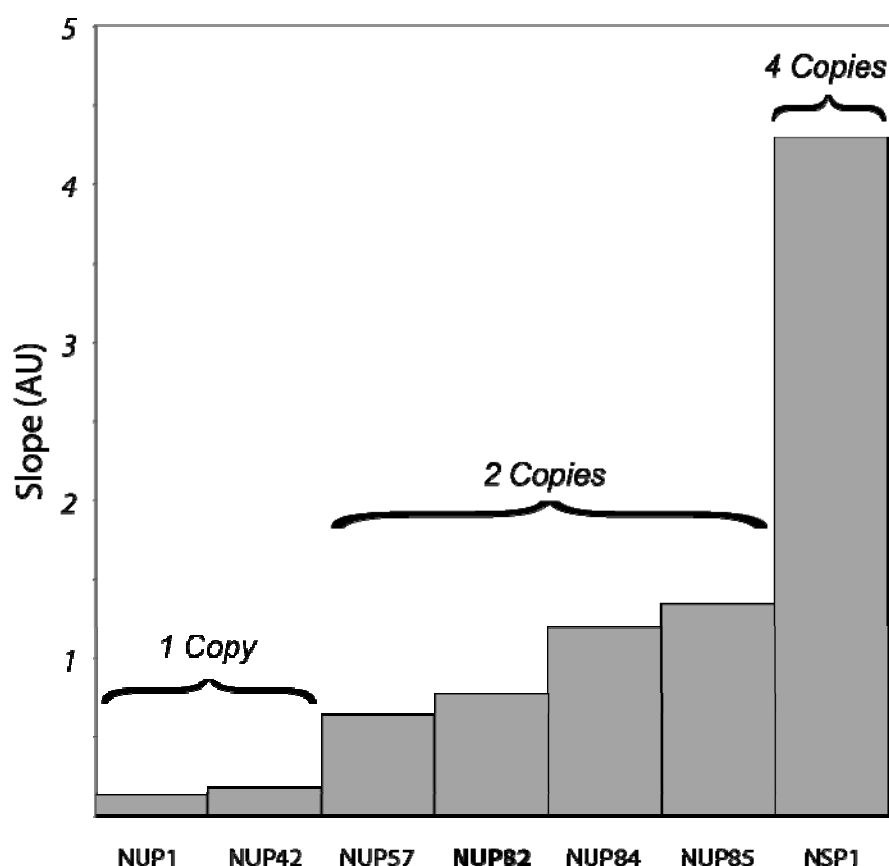


**Supplementary Figure 6.**

**Supplementary Figure 6. Analysis of the Pom rings.**

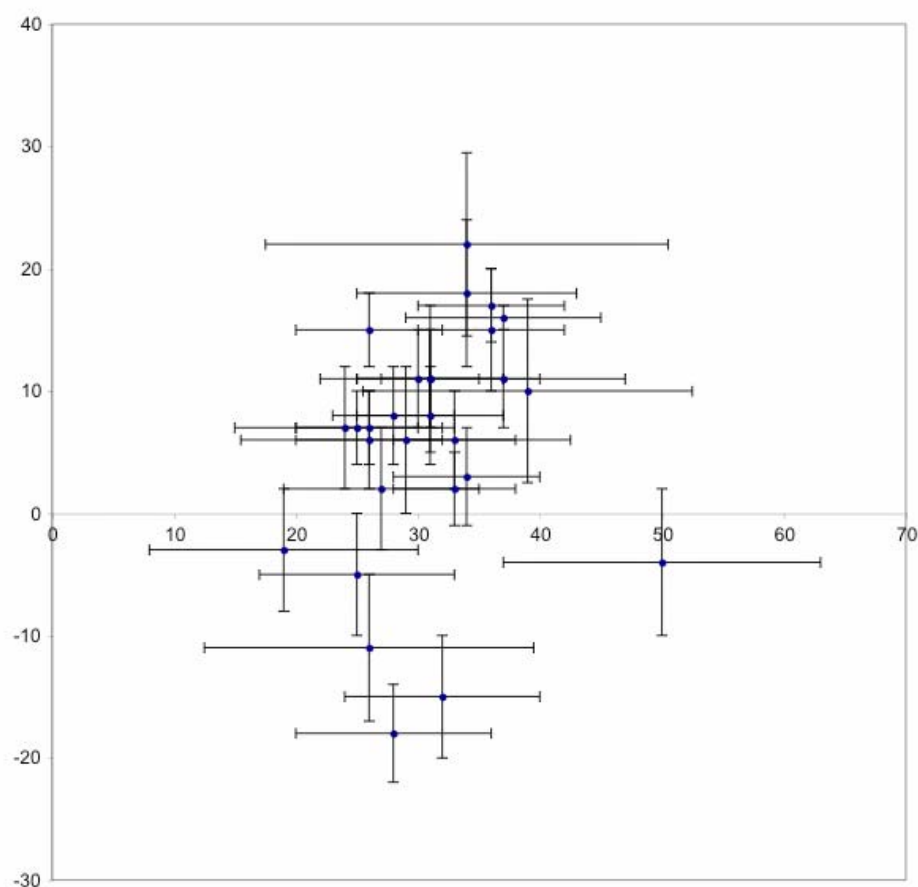
(A). Montage of negatively stained samples of individual Pom rings, visualized by transmission electron microscopy (scale bar, 100 nm).

(B). Localization volume representation of proteins present in the Pom rings (left) and Pom rings connected to Nup170 (right).



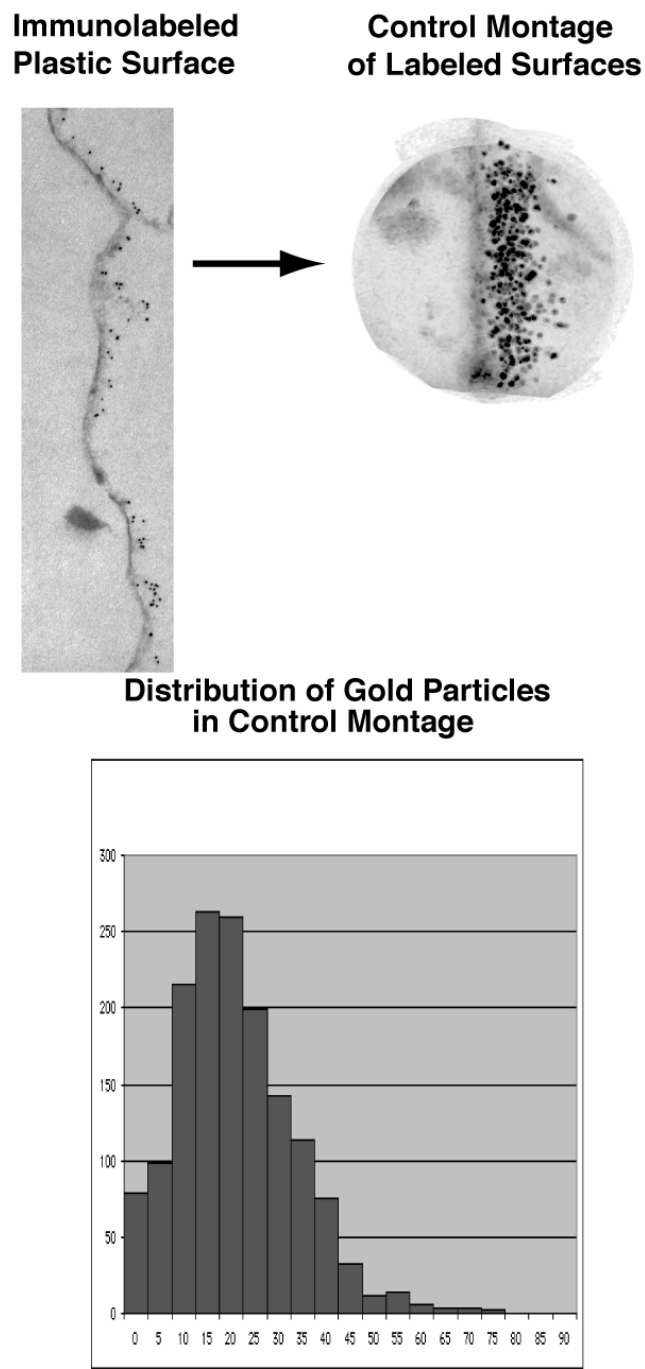
### Supplementary Figure 7. Identification of Nup82 copy number.

Aliquots of NE preparations from PrA tagged strains equivalent to 3.6, 6, 10 and 15  $\mu\text{g}$  were processed for immunoblot analysis. The strains with known copy number – Nup42, Nup1 (1 copy per spoke), Nup57, Nup84, Nup85 (2 copies per spoke) and Nsp1 (4 copies per spoke) were used as a control <sup>36</sup>. The membranes were probed first with MAb118C3 to detect Pom152 (the internal standard) and then with HRP conjugated IgG to detect both MAb118C3 and the PrA tag. The signal intensities of both the standard and tagged protein were quantified with ImageQuant. For each Nup sample, the signal intensity of PrA was plotted against that of the internal standard in Microsoft Excel. Regression analysis showed a linear plot. The resulting slope is a measure of the abundance of the Nup82-PrA relative to internal standard <sup>36</sup>. Shown here are the slope values for each Nup, with value for Nup82 falling into the same range as the values of the 2 copy per spoke Nups.

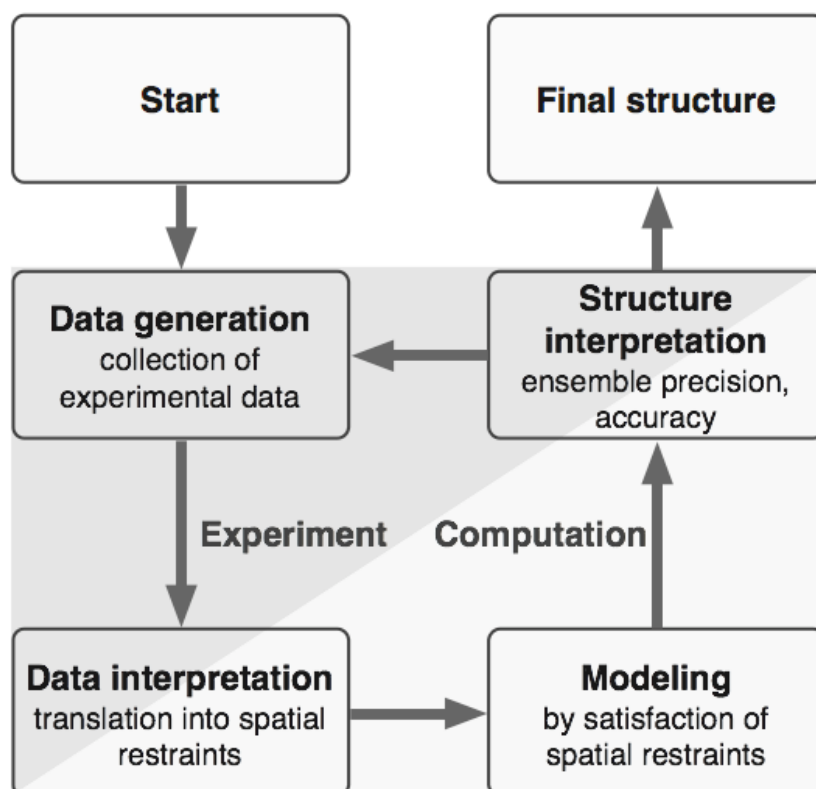


**Supplementary Figure 8.** Immunoelectron Microscopy Ranges. The allowed ranges of R and Z for each Nup, listed in Supplementary Table 2, are illustrated here. The scale is in nm, with the position of each Nup indicated with a point and its ranges with error bars. The identity of each Nup can be determined by comparing with Figure 4 in <sup>37</sup>; only the cytoplasmic copy of each two-sided Nup is shown, for clarity.



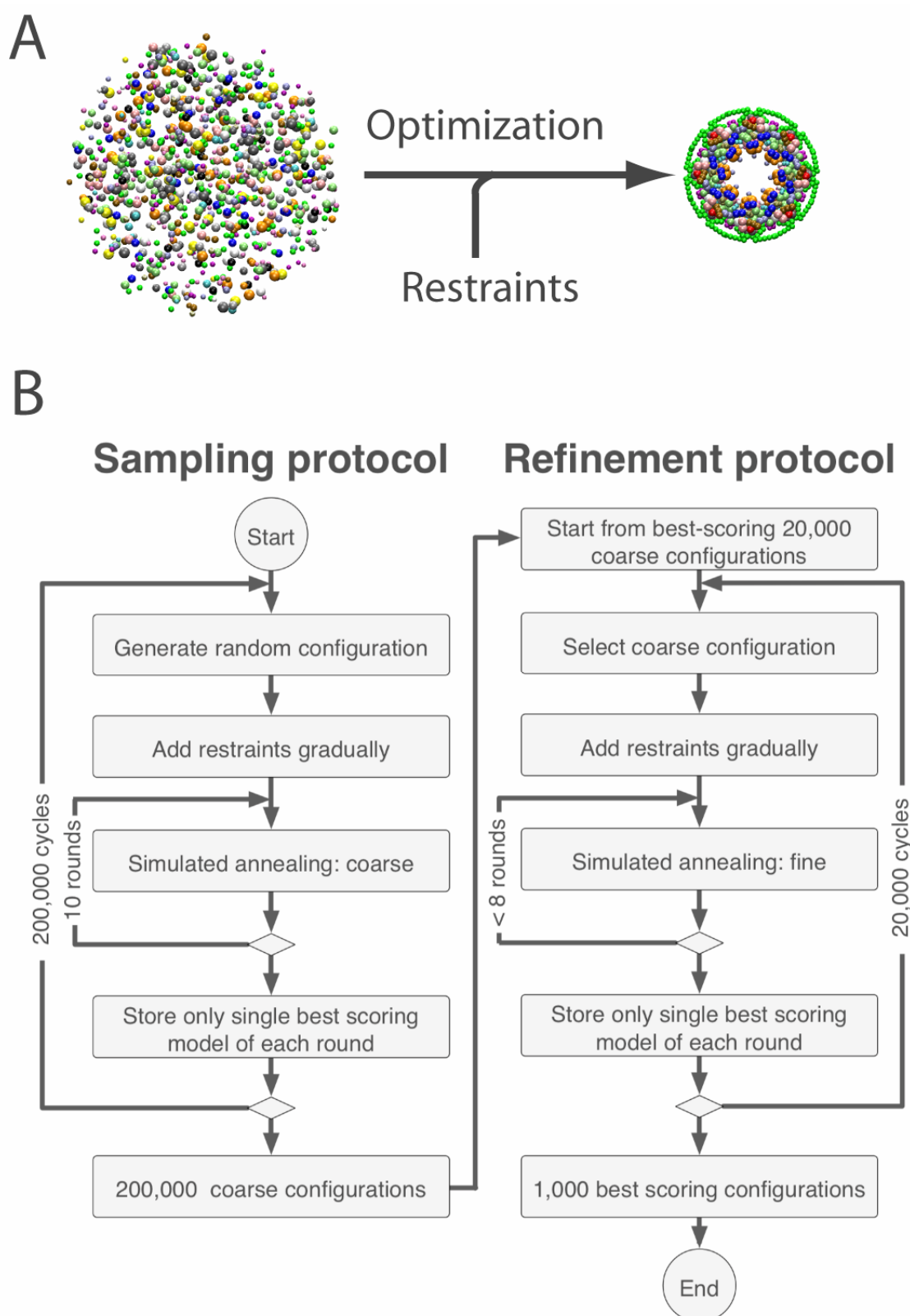


**Supplementary Figure 9.** Distributions of gold particles localized through immunoEM on a plastic surface.



**Supplementary Figure 10: Structure determination process.**

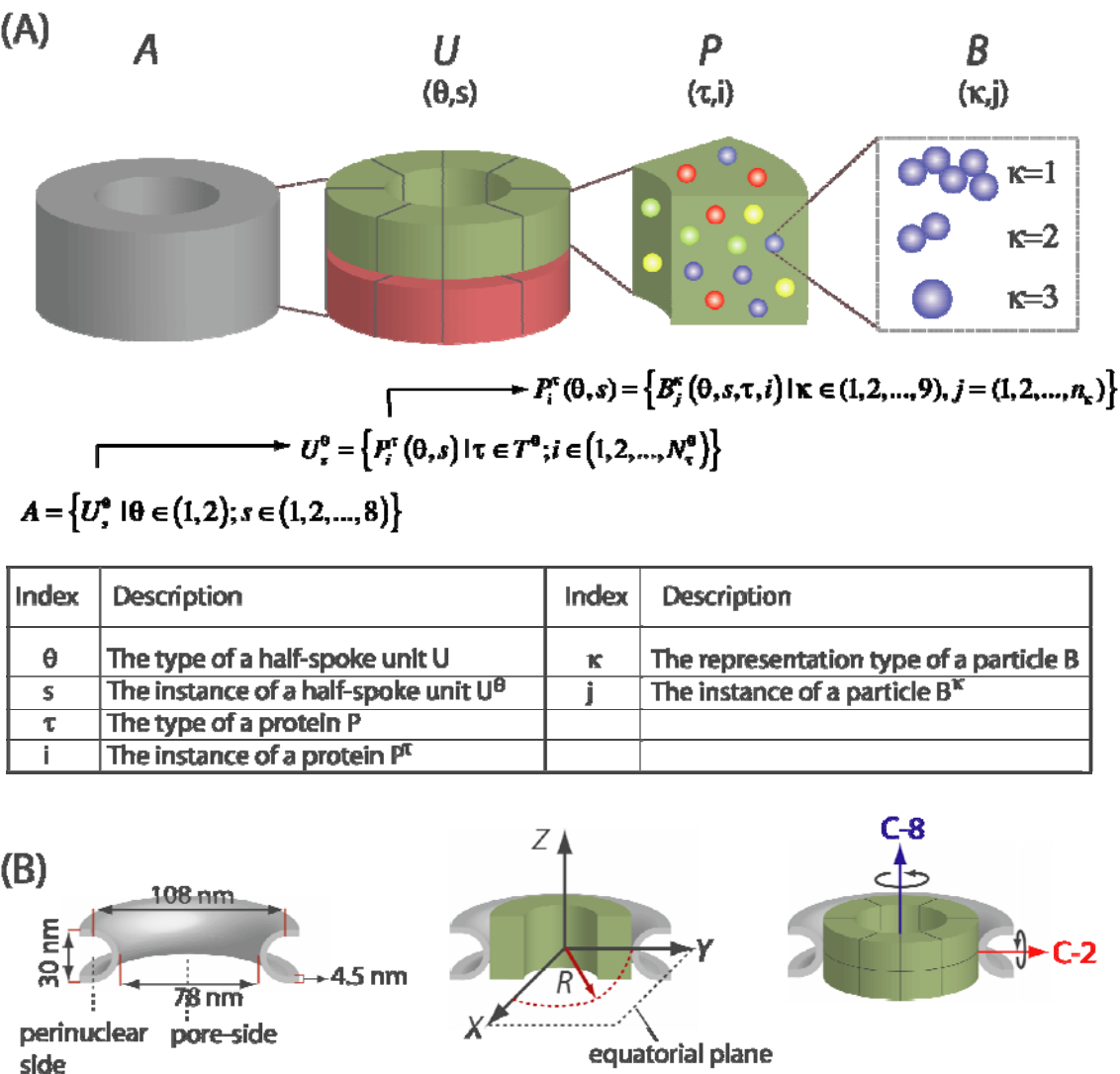
Iterative process of assembly structure determination by satisfaction of spatial restraints.



**Supplementary Figure 11: Determination of the structure of the NPC by satisfaction of spatial restraints.**

**a**, Schematic view of the optimization process. (Left) Randomized configuration of all beads at  $\kappa=2$ . (Right) Resulting configuration after the optimization of the scoring function that incorporates all restraints.

**b**, The optimization is split into two stages. First, a coarse sampling protocol (left column) generates 200,000 coarse configurations, starting each time from a different random configuration. This protocol relies on a variable target function method<sup>38</sup> that consists of gradually increasing the number of restraints that are included in the scoring function, finally culminating in the full scoring function  $F$ . At each stage of the variable target function method, a combination of the conjugate gradient (CG) minimization and a molecular dynamics (MD) simulation with simulated annealing is applied. In total, a single optimization of an initial random configuration consists of an iteration of approximately ten thousand small shifts of protein particles (guided by either CG or MD). Second, a refinement protocol (right column) further refines the best 10% configurations from the sampling stage.



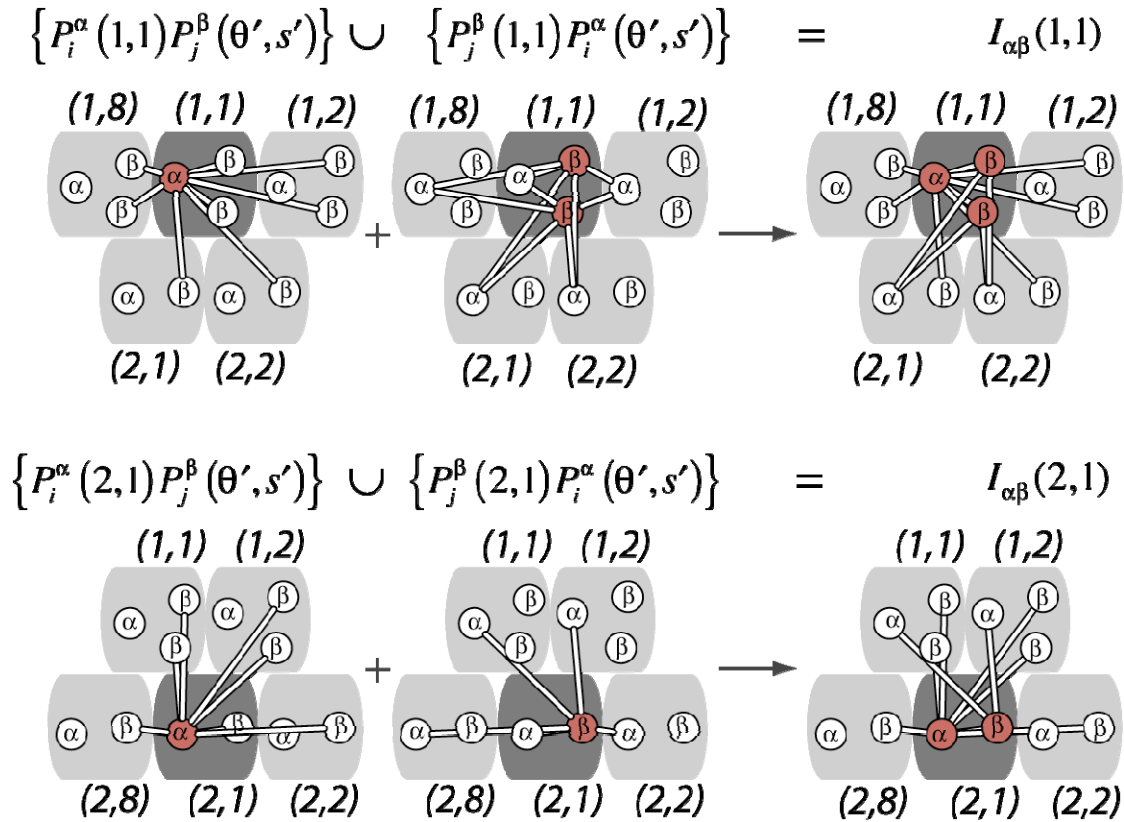
Supplementary Figure 12: Hierarchical organization of the composition and structural representation of the NPC.

a, Formally, we define the NPC assembly  $A$  as a set of symmetry units  $U$  of two different types  $\theta \in \{1, 2\}$  and eight instances  $s$ , each of which is referred to as a ‘half-spoke’. Units of type  $\theta = 1$  reside at the cytosolic side and units of type  $\theta = 2$  reside at the nucleoplasmic side of the NE. Two half-spokes, one of each type, form a spoke. This larger symmetry unit of the NPC is repeated 8 times to form a ring. Each of the 16 NPC half-spokes  $U_s^\theta$  consists of a set of nups  $P_i^\tau$  that are defined by their type  $\tau$  (i.e., nup identity) and copy number  $i$  (i.e., nup instance number). Each protein is described simultaneously by several structural representations. Each protein

representation consists of a set of particles defined by their representation type  $\kappa$  and particle instance  $j$ . Formally, a protein  $P_i^{\tau}$  is represented by the set of all particles in all of its representations. For each component in the hierarchy, we use Greek letters (in superscript) to refer to variables of component types and Latin letters (in subscript) for variables of a particular instance of the component. Component variables that are defined at a higher hierarchical level are listed in brackets.

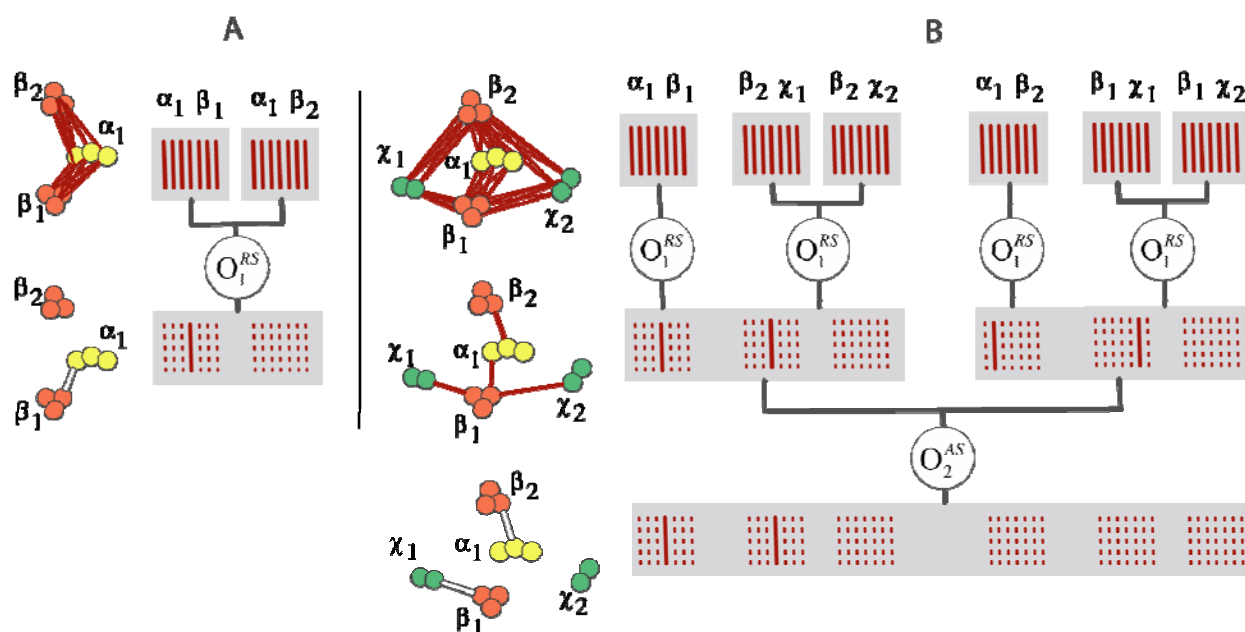
**b**, Left column: Effective NE dimensions taken from cryoEM images <sup>9</sup>. Middle column: Coordinate system with the origin at the center of the NE pore. The NE is indicated in grey, a hypothetical NPC mass is indicated in green. Right column: Eight (blue) and two-fold symmetry axis (red) of the NPC revealed by cryoEM <sup>9</sup>.





**Supplementary Figure 13: Definition of potential protein interactions.**

The list of all alternative interactions (thick lines) between protein instances of type  $\tau = \alpha$  and  $\beta$  defined for proteins in half-spokes  $U_{s=1}^{\theta=1}$  (dark unit, upper row) and  $U_{s=1}^{\theta=2}$  (dark unit, lower row). Left column: All pairwise combinations of proteins of type  $\alpha$  in  $U_s^\theta$  (red circles) with all potential interaction partners of type  $\beta$  in half-spokes  $U_{s'}^{\theta'}$  with  $(\theta', s') \in N(\theta, s)$  (below). Middle column: all combinations of proteins of type  $\beta$  in  $U_s^\theta$  (red circles) with all potential interaction partners of type  $\alpha$  in half-spokes  $U_{s'}^{\theta'}$ , where  $(\theta', s') \in N(\theta, s)$ . Right column: all possible interactions  $I_{\alpha\beta}(\theta, s)$  between proteins of type  $\alpha$  and  $\beta$  for half-spoke  $U_s^\theta$  defined as the union of both groups.  $N(\theta, s) \in \{(\theta, s), (\theta, (s-2) \bmod 8 + 1), (\theta, s \bmod 8 + 1), (\theta \bmod 2 + 1, s), (\theta \bmod 2 + 1, s \bmod 8 + 1)\}$ .

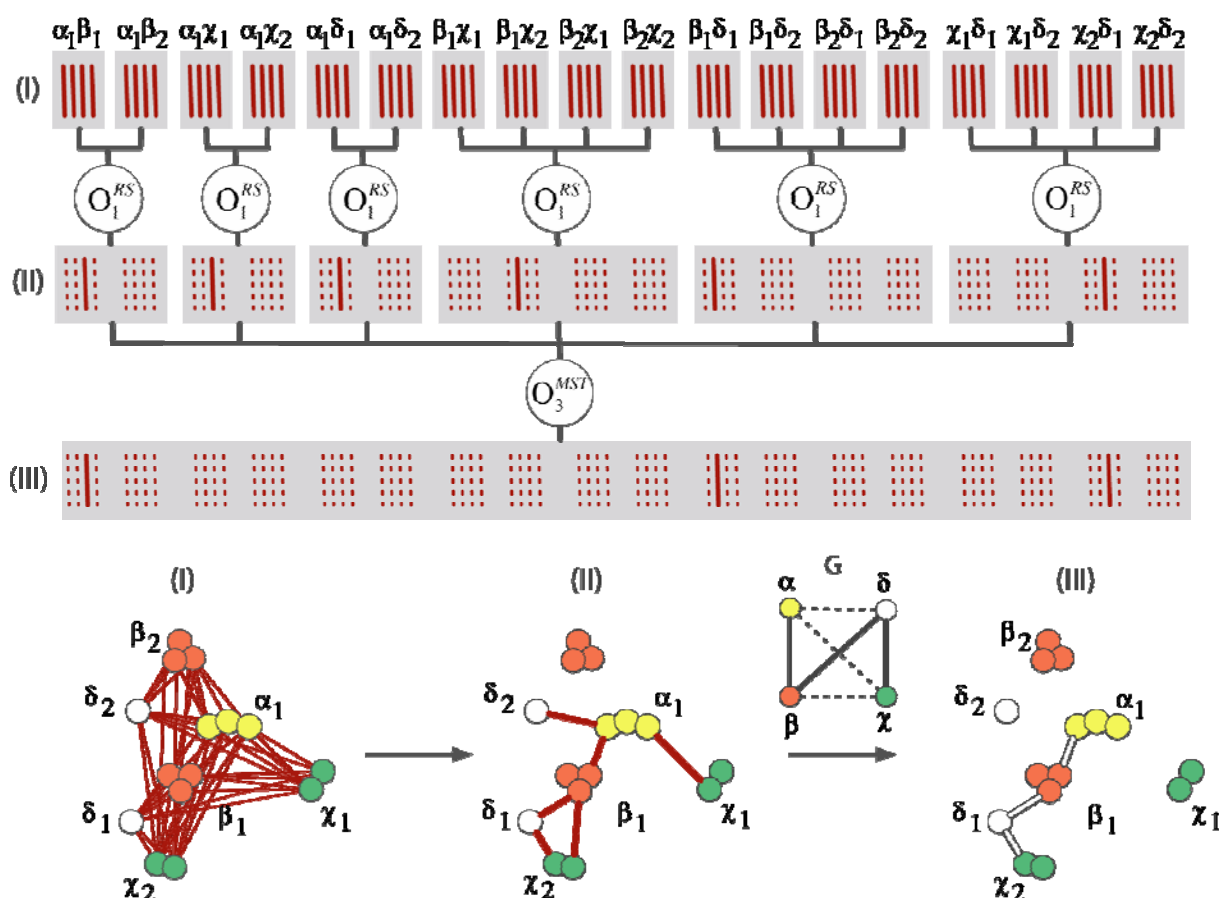


**Supplementary Figure 14: Conditional protein interaction and competitive binding restraint.**

**a**, Protein interaction restraint imposing an interaction between proteins of type  $\alpha$  and  $\beta$  in an assembly with one and two instances, respectively. Left: bead representation of the proteins. Right: schematic view of the restraint activation process. All optional distance restraints are indicated as red lines. The optional restraints describing the two possible interactions  $I_{\alpha\beta}(1,1) = \{(P_1^\alpha P_1^\beta), (P_1^\alpha P_2^\beta)\}$  are grouped in two separate sets (grey boxes). At each optimization step, a rank-and-select operator function ( $O_n^{RS}$ ) activates the optional restraints with the smallest restraint value, based on the current assembly configuration.

**b**, Conditional competitive binding restraint. Left: bead representation of an assembly of 5 proteins. Right: schematic view of the restraint activation process. The restraint enforces two interactions between proteins of type  $\beta$ - $\alpha$  and  $\beta$ - $\chi$  under the condition that proteins of types  $\alpha$  and  $\chi$  cannot interact to the same instance of protein  $\beta$  ( $\alpha_1 \beta_n$  and  $\beta_m \chi_l \mid n \neq m$ , for any  $l$ ). Conditional dependence is encoded in a three-level restraint hierarchy. Upper row: The optional restraints describing the two possible interactions of  $I_{\alpha\beta}(1,1)$  and four possible interactions of  $I_{\beta\chi}(1,1)$  are grouped into six separate sets (grey boxes indicate the groups and red thick lines the optional restraints). At the first restraint level, a rank-and-select operator  $O_1^{RS}$  activates one out of the optional restraints per restraint group  $R_i$ . The resulting activated restraints become optional restraints at the next hierarchical level and are organized into two groups that each are consistent

with the conditional dependencies of competitive binding. At the third restraint level, an operator  $O_n^{AS}$  (add-and-select operator) activates the optional restraints in one of the two groups.



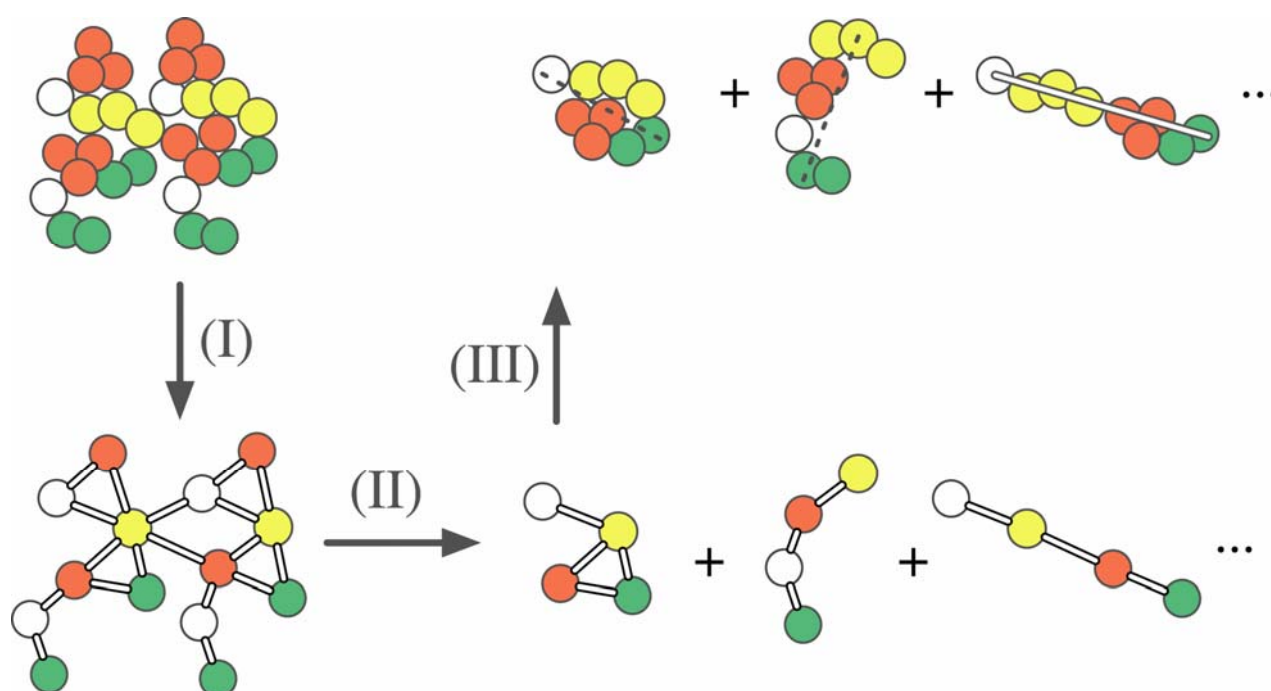
**Supplementary Figure 15: Complex connectivity restraint.**

**a**, Schematic view of the restraint activation process for the complex connectivity restraint of a composite containing 4 protein types present in an assembly with 7 proteins (1  $\alpha$ , 2  $\beta$ , 2  $\chi$ , and 2  $\delta$ ).

**b**, Bead representation of the same assembly of 7 proteins.

Each potential interaction of the 6 different types of interactions ( $I^{tot}(1,1) = \{I_{\alpha\beta}, I_{\alpha\chi}, I_{\alpha\delta}, I_{\beta\chi}, I_{\beta\delta}, I_{\chi\delta}\}$ ) is translated into a set of optional distance restraints (red thick lines). At the first restraint level, a rank-and-select operator  $O_1^{RS}$  activates only one out of the optional restraints per restraint group. The resulting 6 activated restraints (red full lines) are subjected to the minimal spanning tree (MST) operator  $O_n^{MST}$  at the next hierarchical level. This

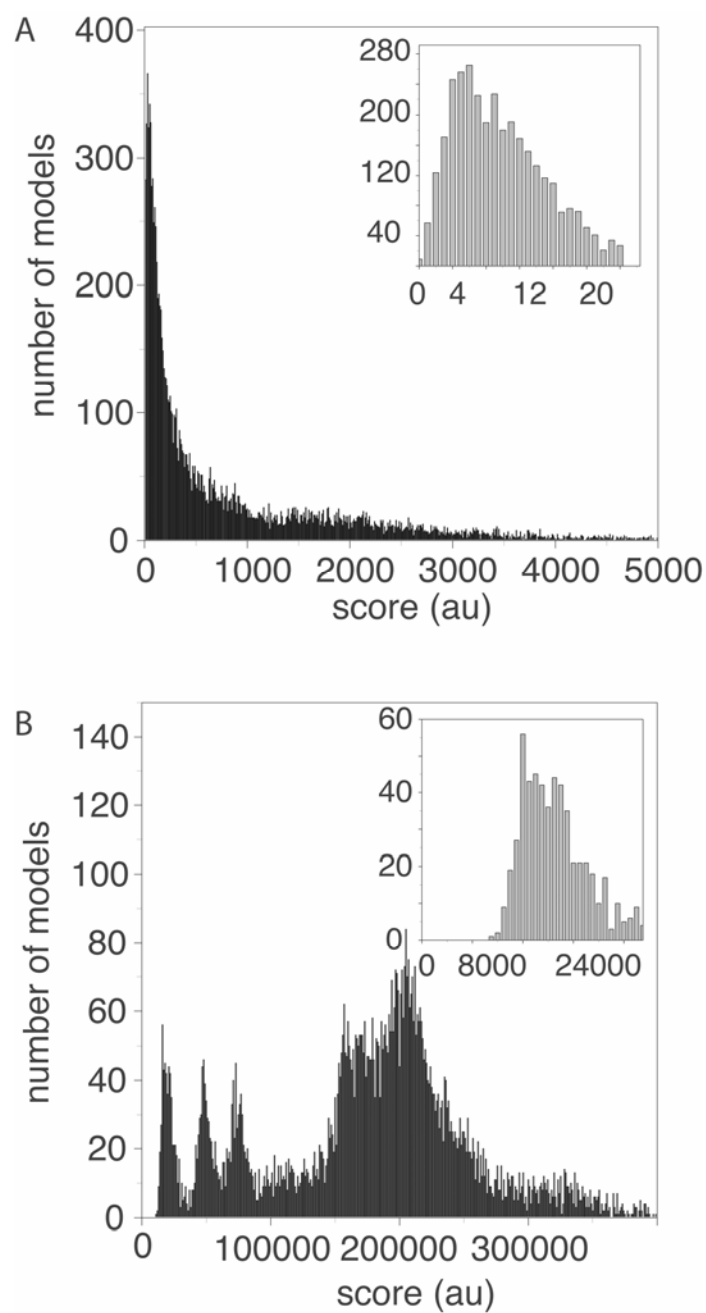
operator activates only the 3 restraints that satisfy the connectivity condition with the lowest total restraint values.



**Supplementary Figure 16: Maximal dimension of a complex defined by protein types.**

Schematic description for the identification of the largest dimension of a complex in an assembly containing two identical units each with 7 proteins of four different types:

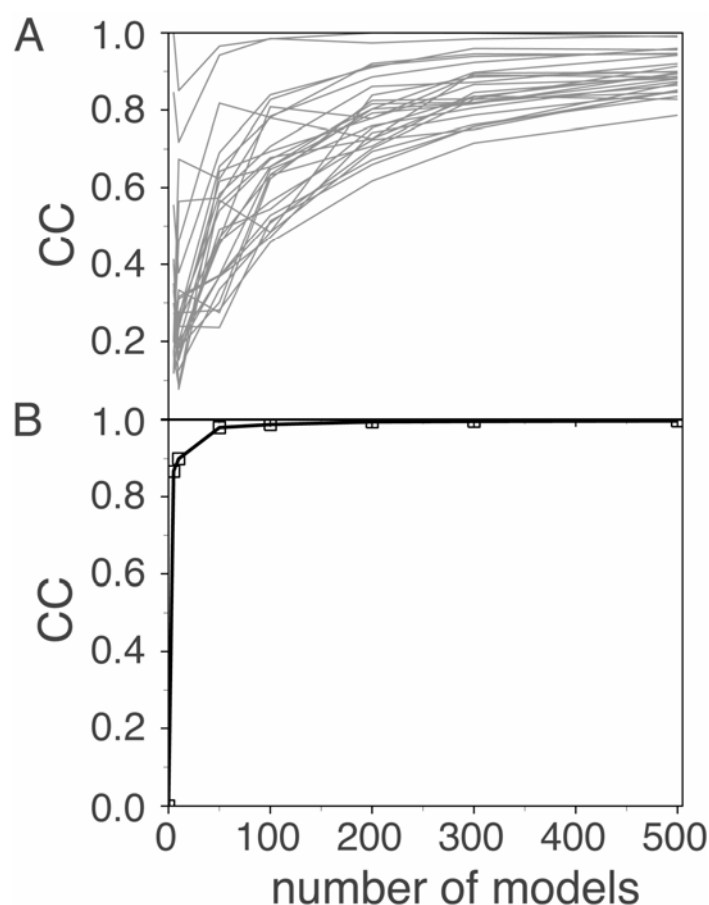
$U_1^1 = \{P_1^\alpha, P_1^\beta, P_2^\beta, P_1^\gamma, P_2^\gamma, P_1^\delta, P_2^\delta\}$ . (I) generation of a contact graph, (II) identification of all connected subgraphs with a defined composition, and (III) identification of the largest distance in complexes whose composition is defined by the identified subgraphs.



**Supplementary Figure 17. Distributions of scores in the ensemble of configurations.**

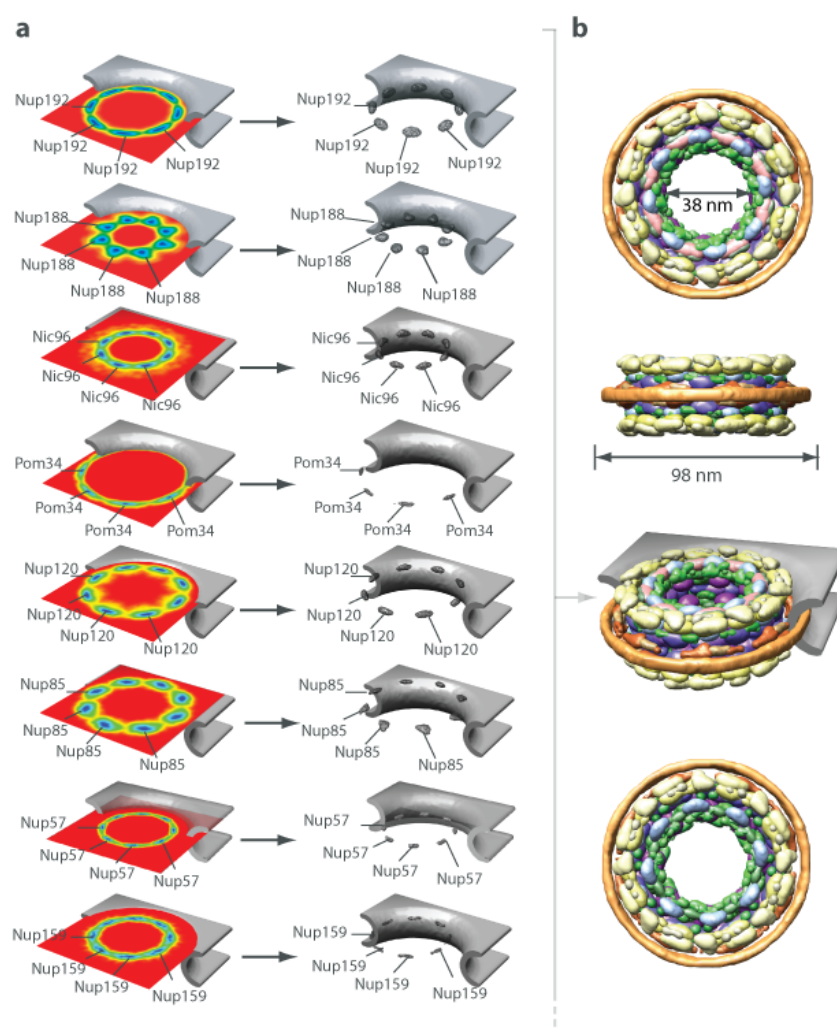
(a) generated with a scoring function based on our data and (b) generated with a scoring function based on publically available data.





**Supplementary Figure 18. Testing how representative is the ensemble.**

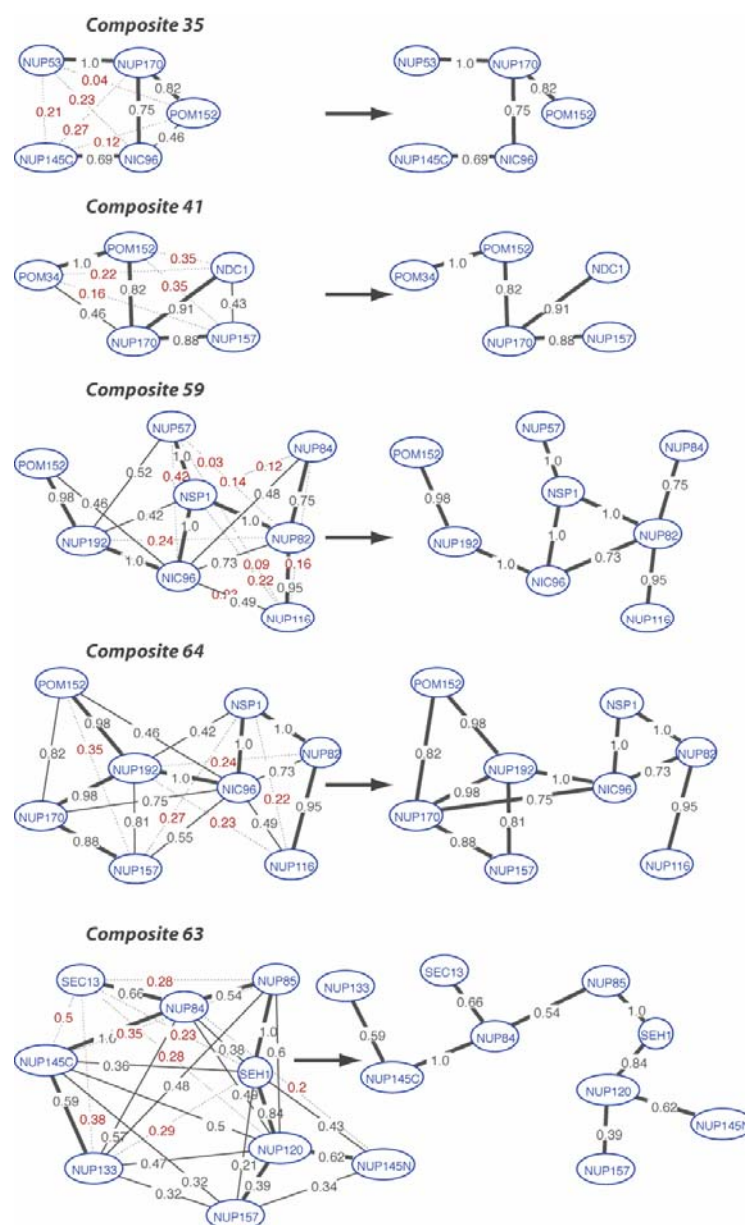
Cross-correlation coefficient (CC) of 2D localization probabilities and contact frequency maps calculated for independent ensembles with 5, 10, 50, 100, 200, 300, and 500 configurations. (A) Structural similarity of two independent ensembles calculated as the cross-correlation between their corresponding 2D localization probabilities. Each line corresponds to the cross-correlation values at a given ensemble size for a single protein type. (B) Similarity of two independent ensembles calculated as the cross-correlation between their corresponding contact frequency maps.



### Supplementary Figure 19: Protein localization.

**a**, The localization probability of nups. The contour level diagram is shown in the plane parallel to the equatorial plane that contains the maximum value of the protein localization probability. Only protein instances in the eight cytosolic half-spoke units are shown; if a protein type is not present at the cytosolic side, the contour level diagram is shown for proteins at the nucleoplasmic side (at  $Z < 0$ ).

**b**, Localization volumes of the structured domains of all 456 Nups in the NPC in front (left) and side views (right) derived from protein representation  $\kappa=2$ . The diameter of the transport channel and the NPC are also indicated. The proteins are color-coded according to their assignment to the 6 NPC modules.



**Supplementary Figure 20: Protein-protein contact analysis.**

Protein connectivity graphs of selected composites. Nups are nodes connected by edges with the observed contact frequency as the edge weight. Edges that are part of the most probable composite minimal connectivity are shown by thick edges. All edges with a statistical significant reduction in contact frequency from their initial values ( $P$ -value  $< 10^{-3}$ ) are indicated by dotted lines. For clarity, in composites #63 and #64, all edges with observed contact frequencies  $< 0.25$  are omitted. Also shown are the corresponding composite adjacency graphs, with edges shown either when they are part of the minimal composite connectivity or have a contact frequency larger than 65%.

**Composite 34**

Inner rings + FG nups + linker nups

**Composite 63**

Inner rings + FG nups + linker nups

**Composite 64**

Inner rings + FG nups + linker nups

**Inner rings + membrane ring**

30°

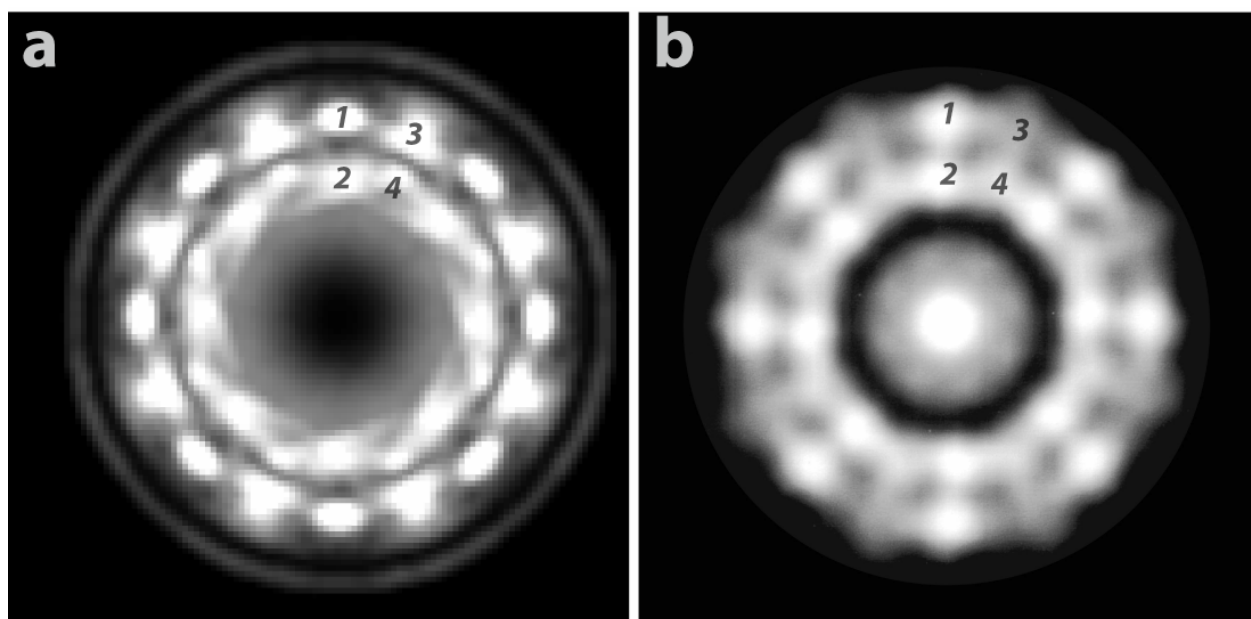
1 - Nup170  
2 - Nup192  
3 - Nup157  
4 - Nup188  
5 - Ndc1  
6 - Pom34  
7 - Pom152

Nup170 Nup159  
Nup192 Nup116  
Nup157 Nup100  
Nup188 Nup1  
Ndc1 Nsp1  
Pom34 Nup60  
Pom152 Nup59  
Nic96 Nup57  
Nup82 Nup49  
Gle1 Nup42  
Gle2 Nup53  
Nup59

**a**, Adjacency graphs of selected composites in a single spoke (left) and adjacency graphs of selected composites for nups in all half-spokes (right). The adjacency graph representation allows

resolving composites and nups in terms of individual complexes and provides information about how these complexes are integrated into the complete NPC.

**b**, Localization volume of selected proteins.



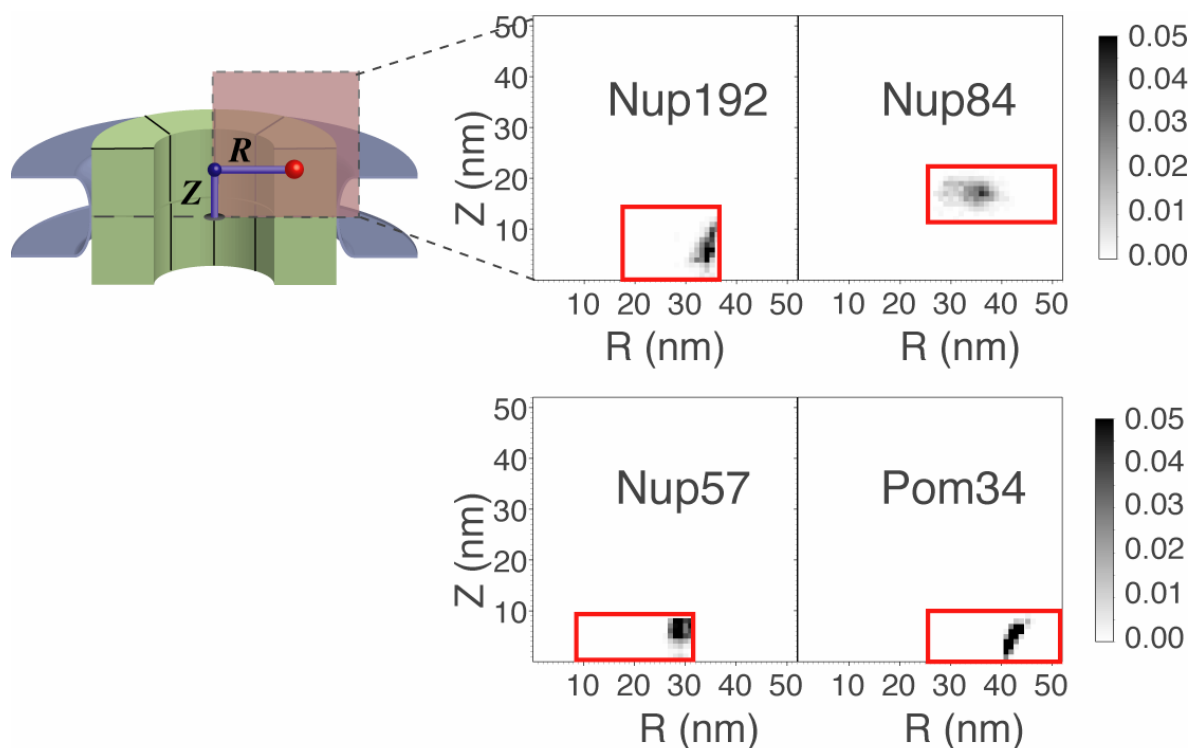
**Supplementary Figure 22: Aspects of the NPC molecular architecture.**

**a**, Nup mass density (white) in *en face* view showing a density projection along the Z-axis from Z of -50 nm to Z of +50 nm. The mass density includes the localization volumes of all structured domains and the normalized localization probability of all unstructured FG regions (see Figure 1 in ref.<sup>39</sup>). Labels indicate mass density peaks present in a single spoke that correspond to two equivalent groups of proteins (1-2 and 3-4).

**b**, Slice of the mass density distribution from a cryoEM map of the Yeast NPC<sup>9</sup>. As in panel **a**, two equivalent peaks per spoke unit are seen for the outer and inner rings.

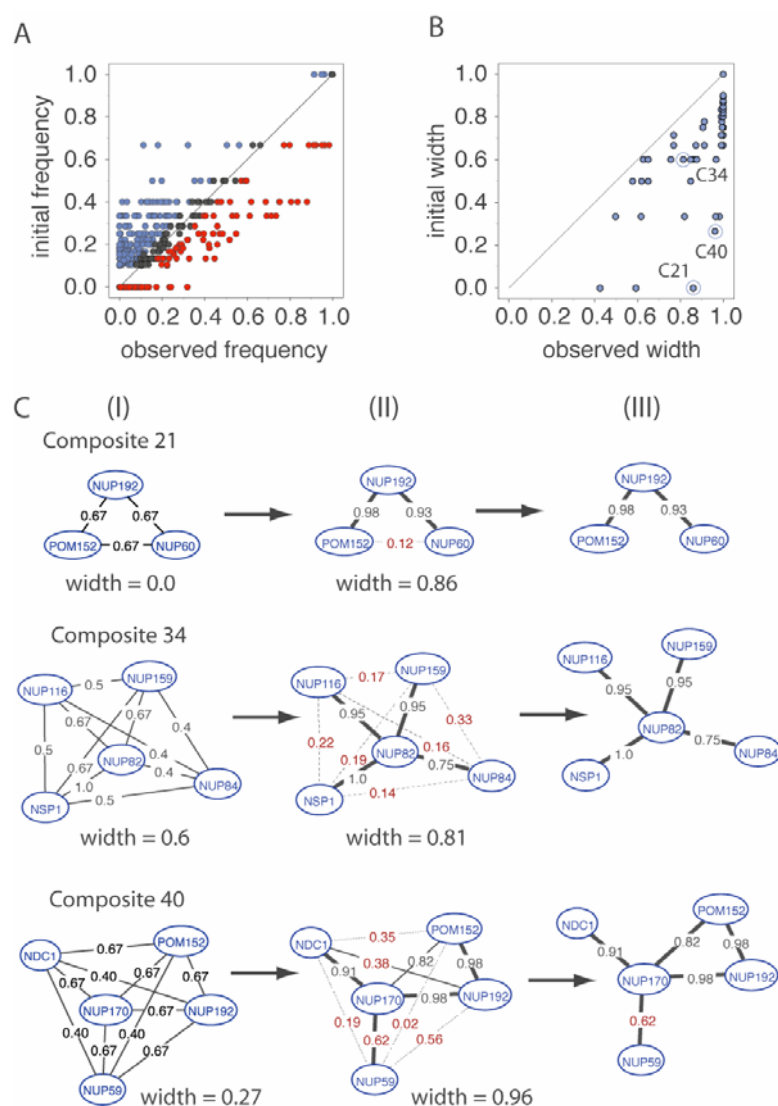
We would like to thank Dr. Chris Akey for providing Supplementary Figure 22 as well as for countless inspiring discussions about the NPC.





### Supplementary Figure 23: 2D localization probability.

The 2D-localization probability ( $2D-lp$ ) of selected nups determined at representation  $\kappa=3$ . The initial  $2D-lp$  based only on the immuno-EM data of each nup is uniformly distributed between the upper and lower bounds along the radial and axial coordinate (red dotted box). The observed  $2D-lp$  upon integration of all additional data is shown as a density plot in grey scale, showing the final, significantly more precise localizations.

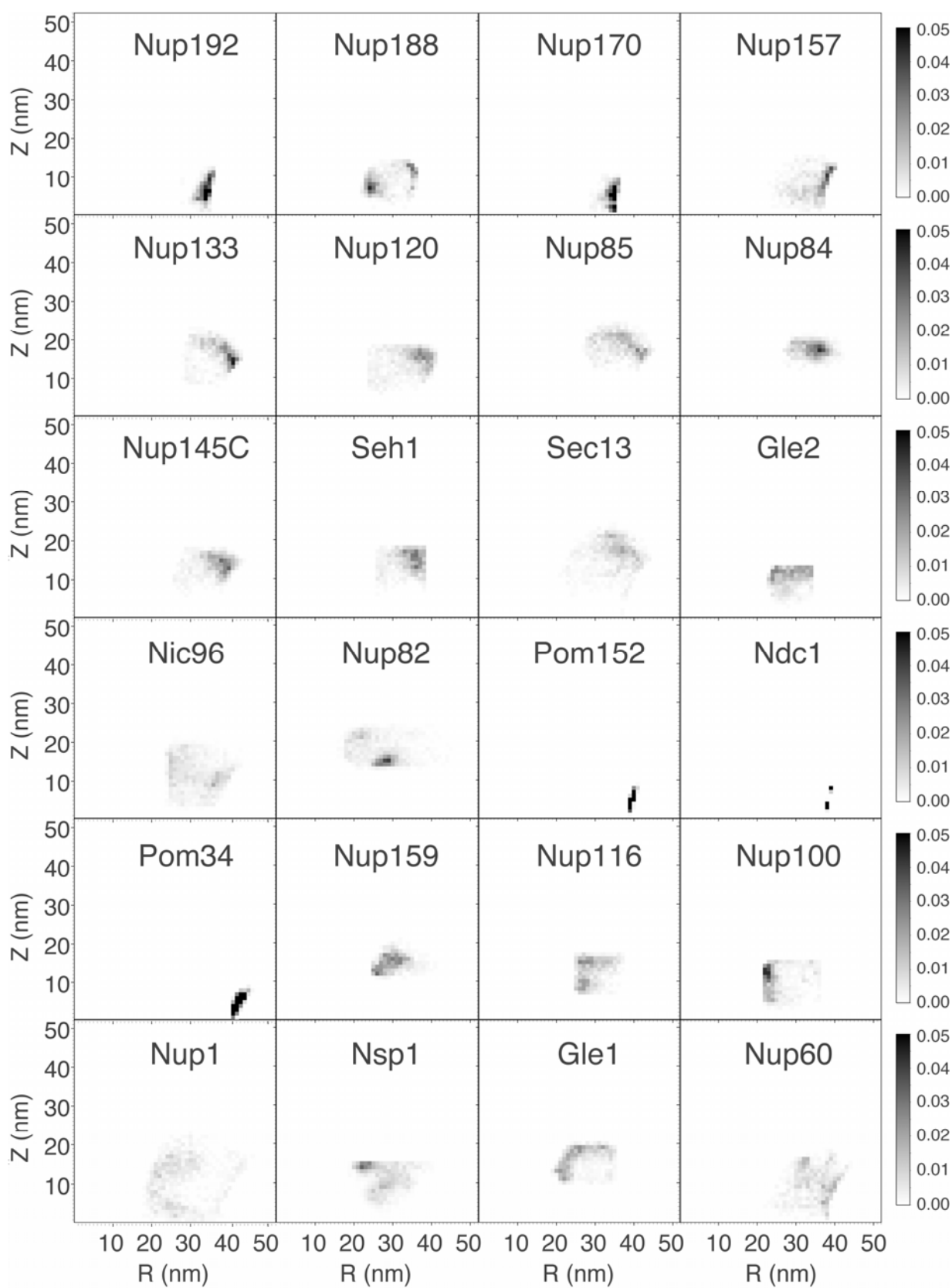


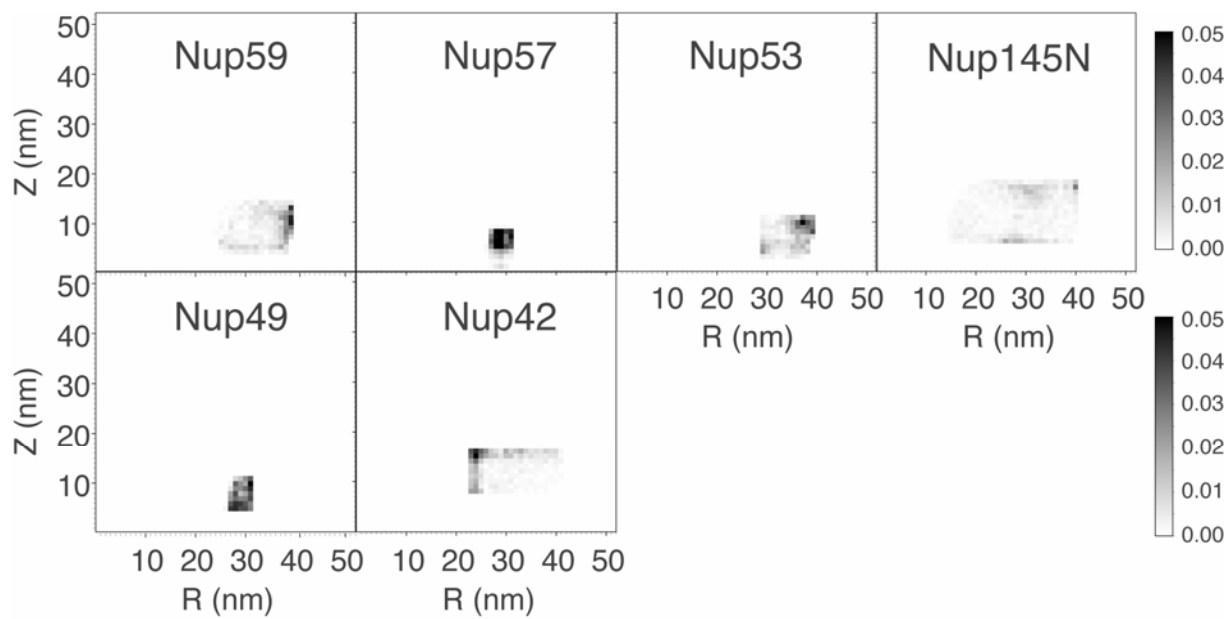
### Supplementary Figure 24. Benefits of data integration.

**a**, Statistical significance of the change in contact frequencies upon data integration: blue and red dots represent contact frequencies with significantly decreased and increased observed contact frequencies ( $P\text{-value} < 10^{-10}$ ), respectively, compared to their initial values.

**b**, Scatter plot of the widths of the initial and observed contact frequency distributions for each composite.

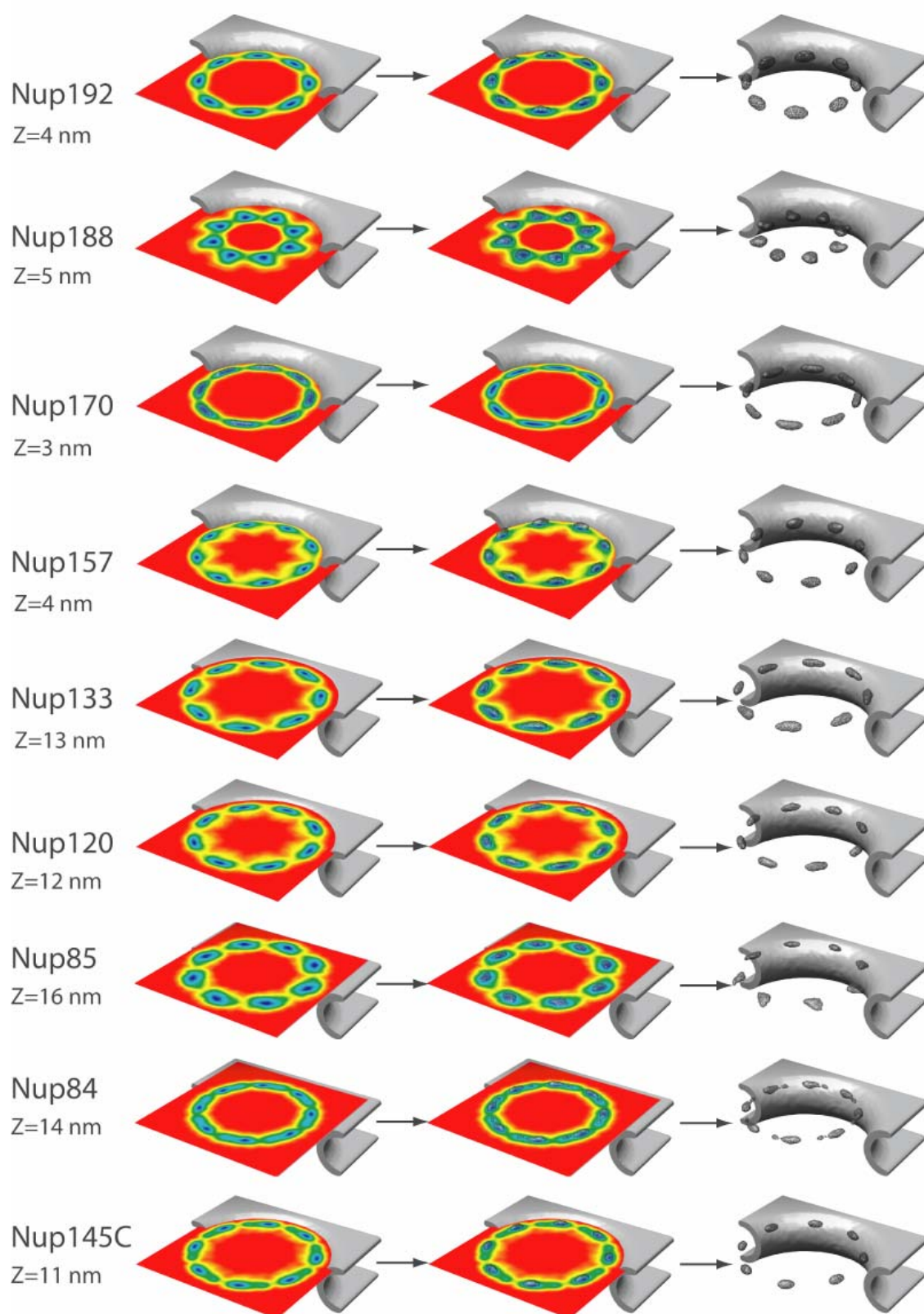
**c**, Composite connectivity graphs (left panel) with initial frequencies as edge labels and observed contact frequencies (right panel). Thick lines indicate the most probable protein connectivity for the composite. All edges with a statistical significant reduction in contact frequency from their initial values ( $P\text{-value} < 10^{-3}$ ) are indicated by dotted lines.



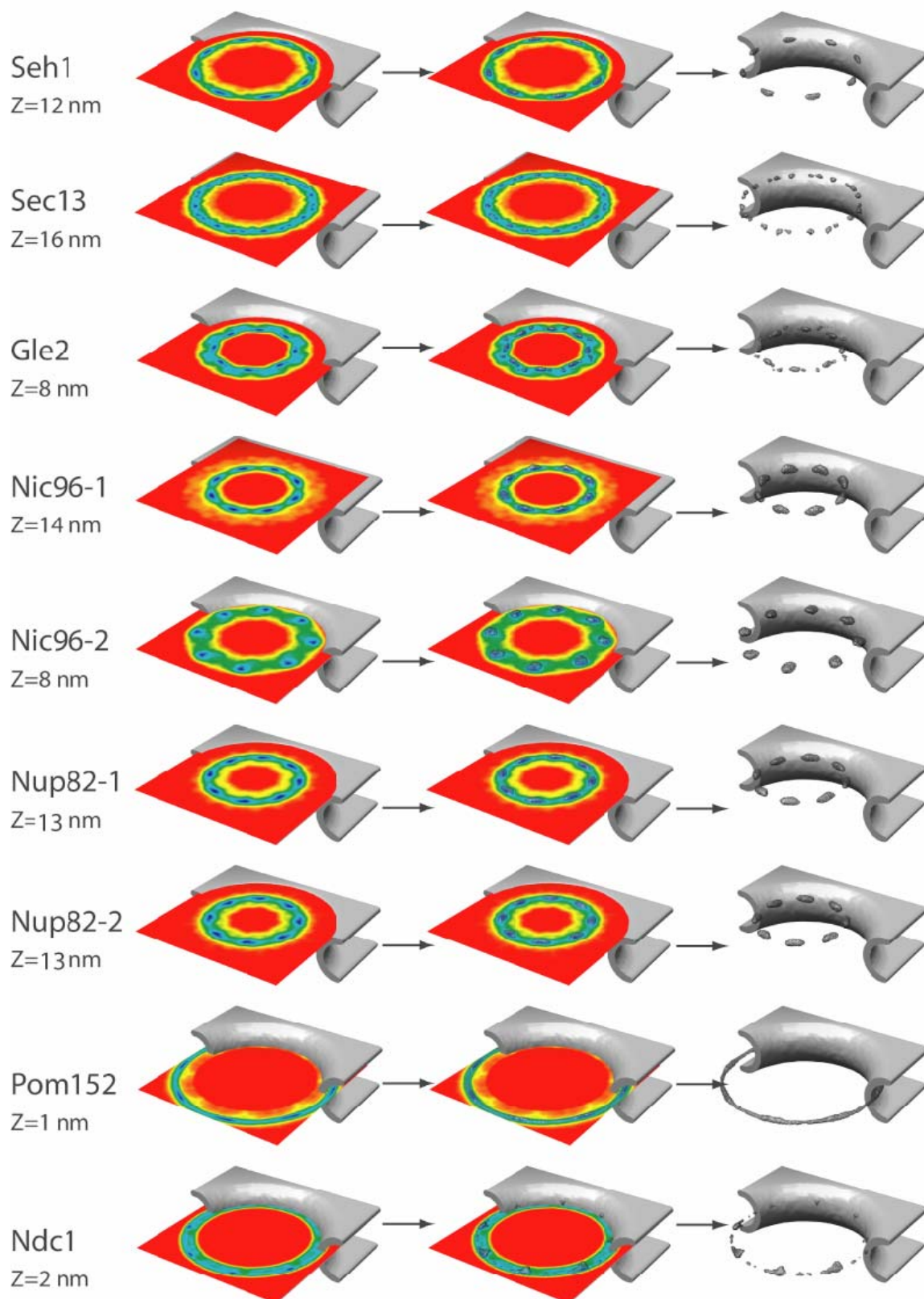


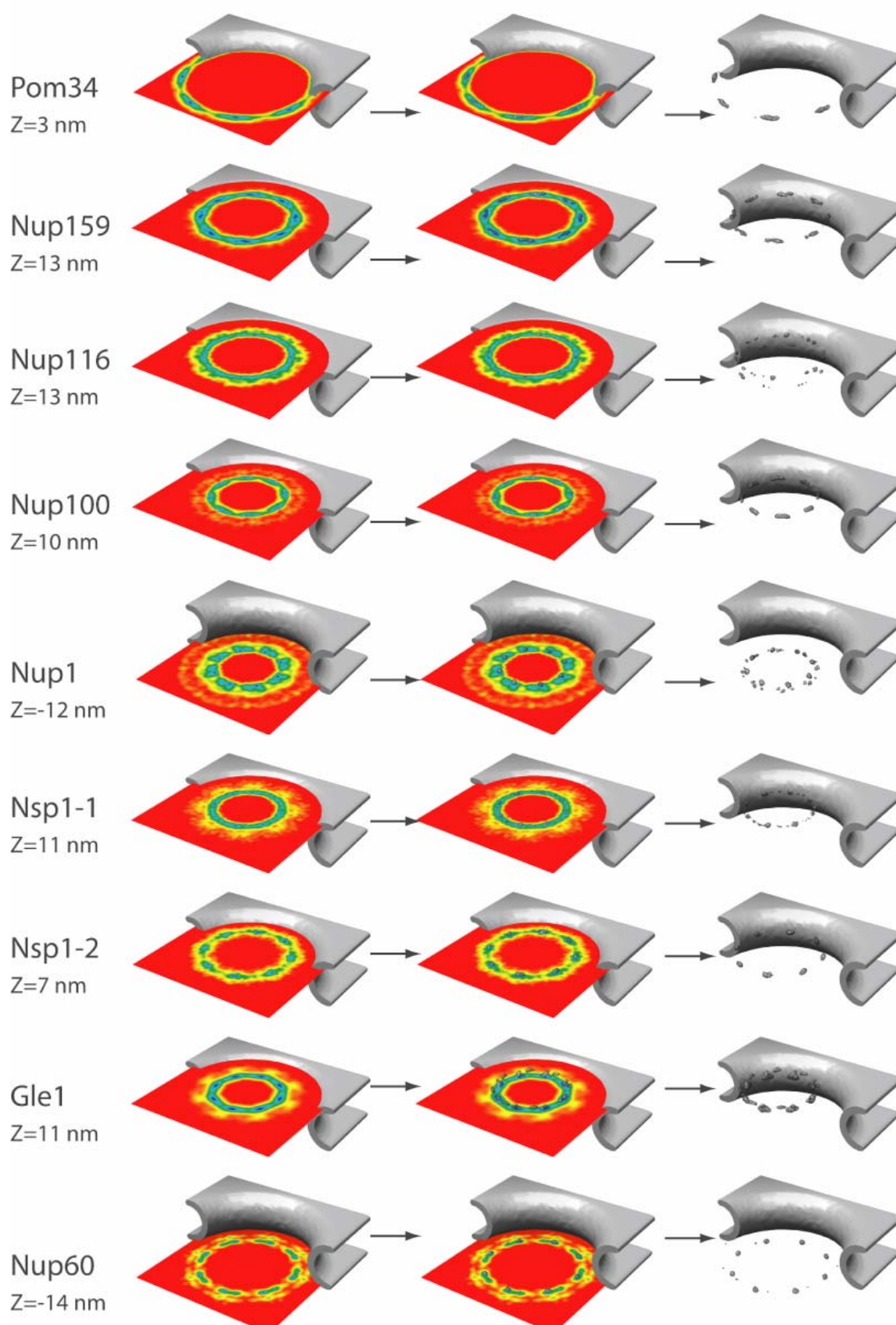
**Supplementary Figure 25. 2D localization probabilities of each nup type.**

The 2D-localization probability ( $2D-lp$ ) of nups determined at representation  $\kappa=3$ .

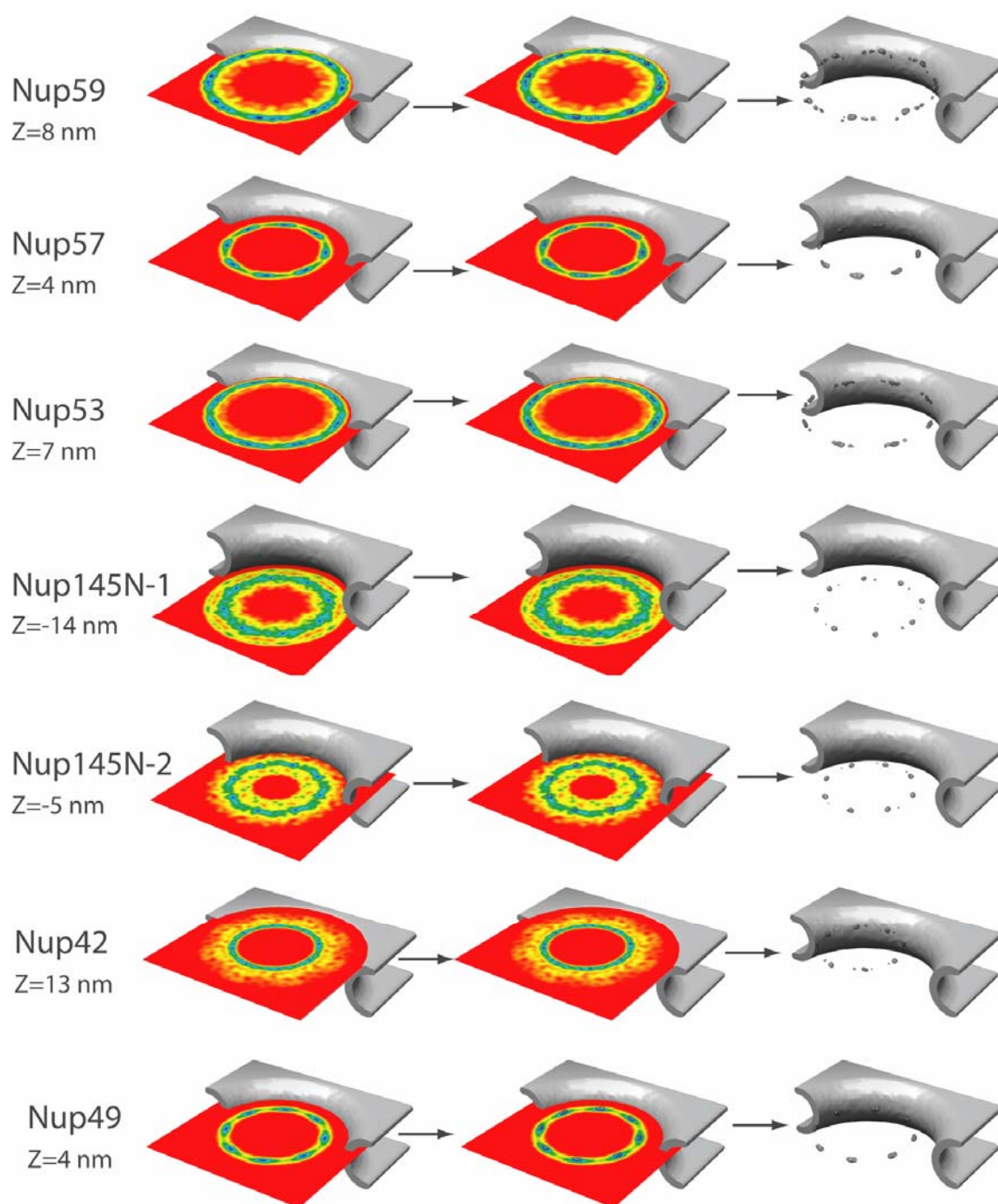










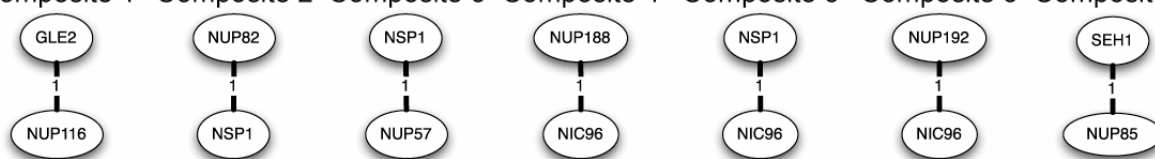


**Supplementary Figure 26. Localization probabilities of eight copies of each nup type at the cytoplasmic side.**

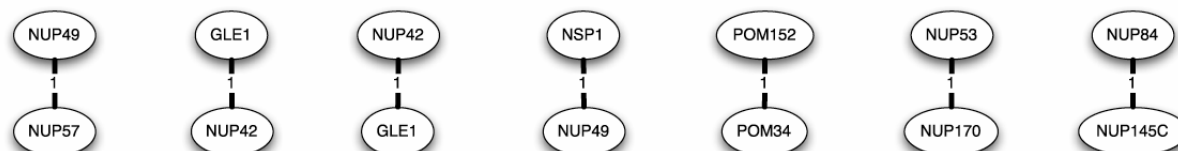
If the nup is present only at the nucleoplasmic side, the localization probabilities is shown for the eight nup copies at the nucleoplasmic side. Each set of nups is shown in three representations:

(left column) the contour level diagram is shown in a plane parallel to the equatorial plane that contains the maximum value of the protein localization probability; (middle column) the contour level diagram is shown together with the localization volume of each protein. The localization volume reveals its most probable localization; (right column) only the localization volumes are shown for each protein.

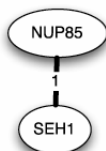
Composite 1 Composite 2 Composite 3 Composite 4 Composite 5 Composite 6 Composite 7



Composite 8 Composite 9 Composite 10 Composite 11 Composite 12 Composite 13 Composite 14



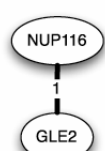
Composite 15



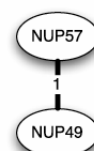
Composite 16



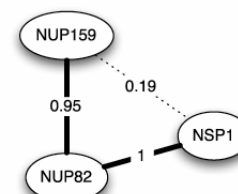
Composite 17



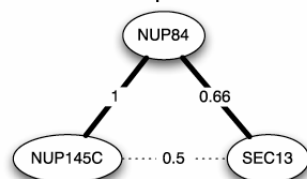
Composite 18



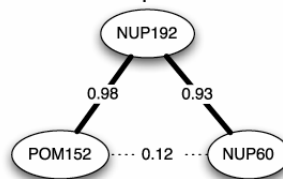
Composite 19



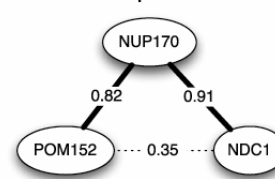
Composite 20



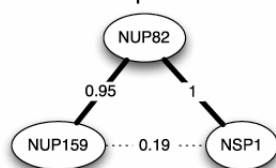
Composite 21



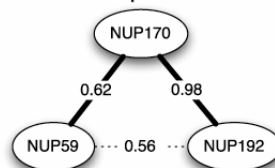
Composite 22



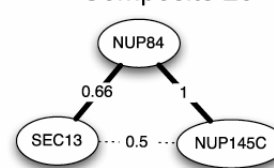
Composite 24



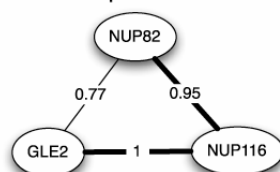
Composite 25



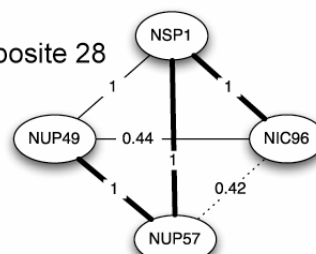
Composite 26

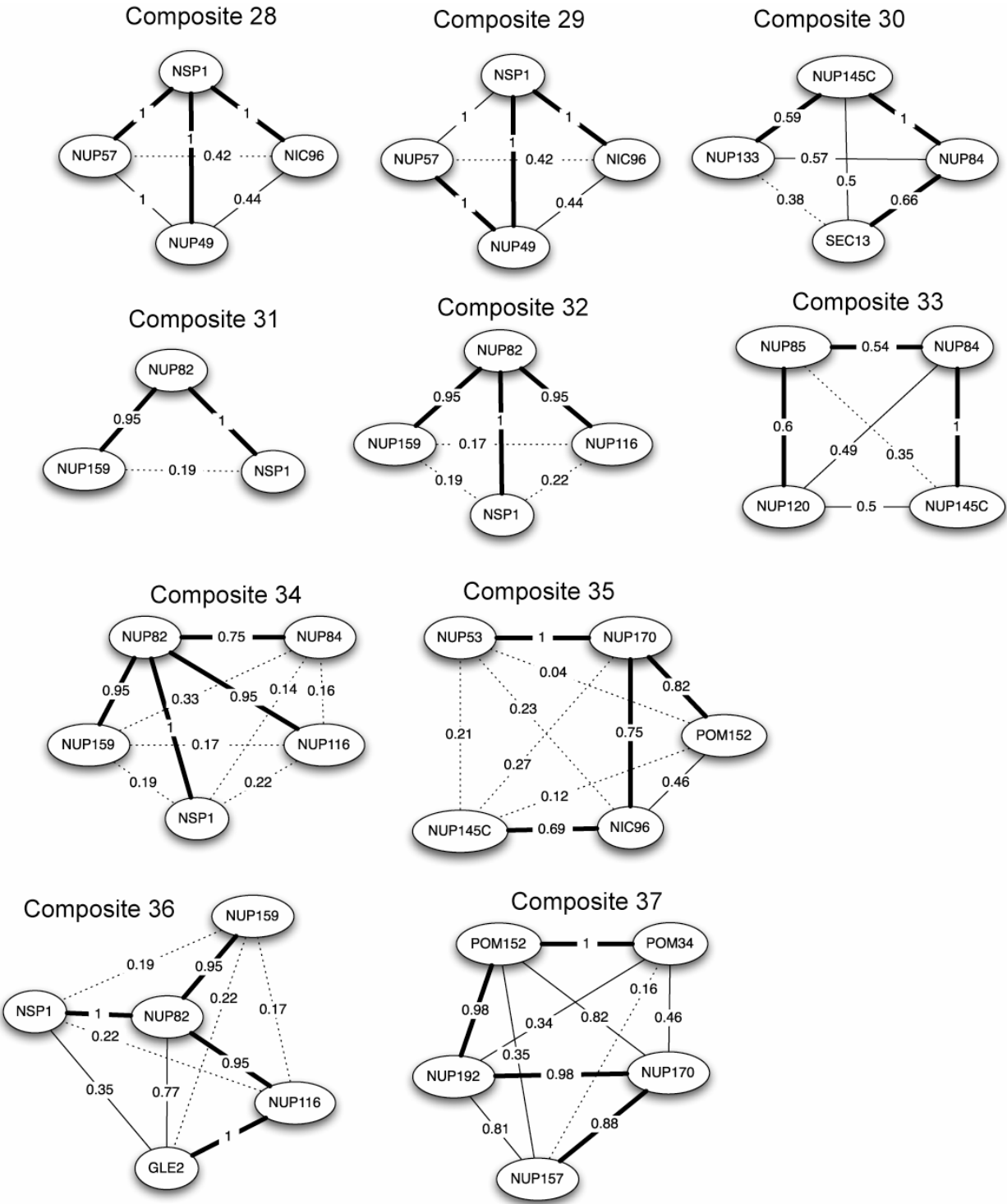


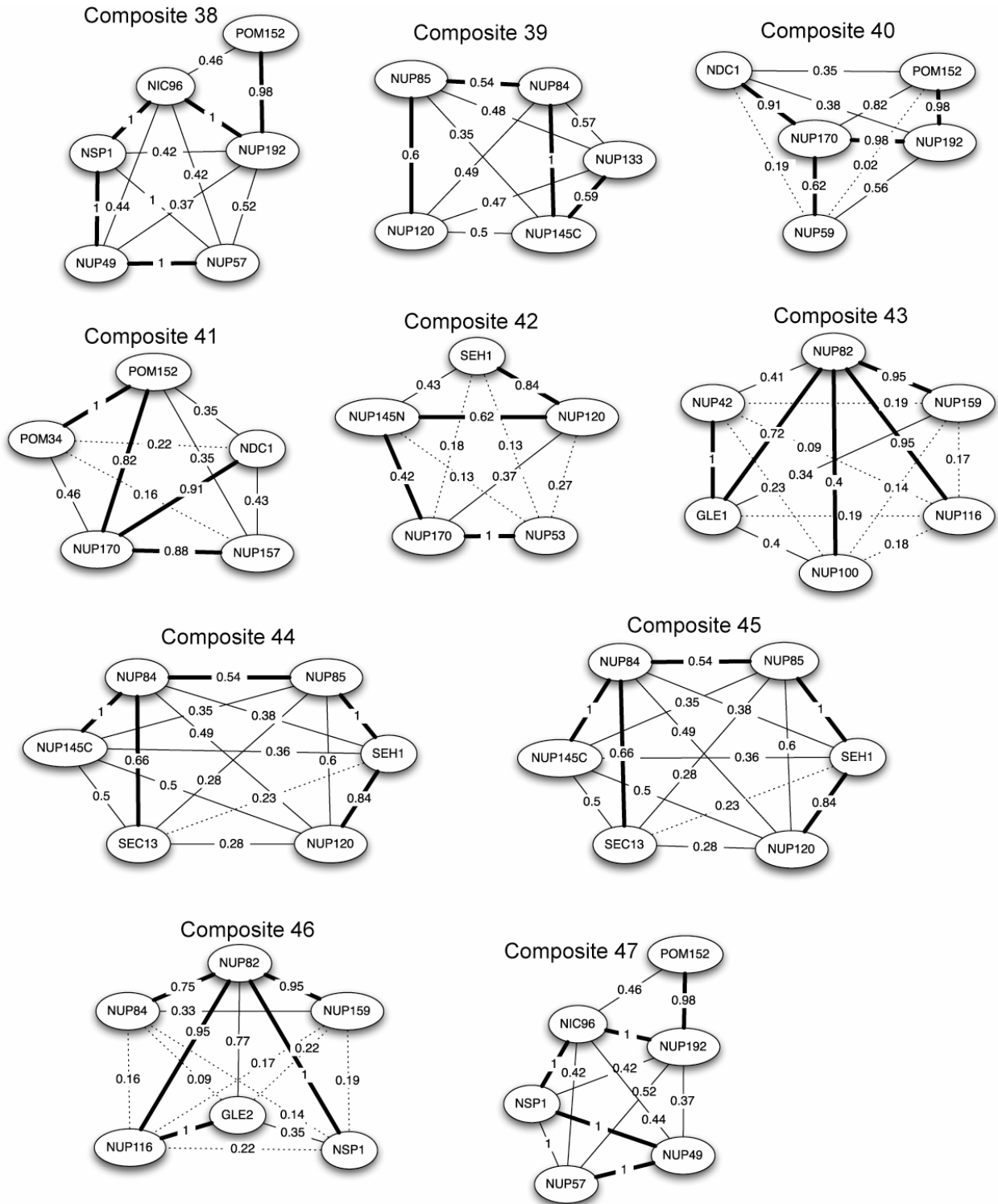
Composite 27



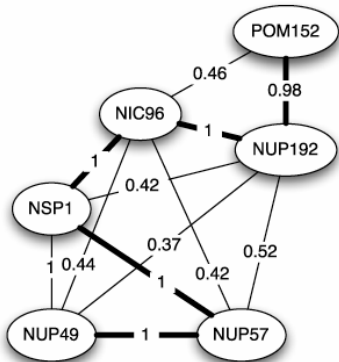
Composite 28



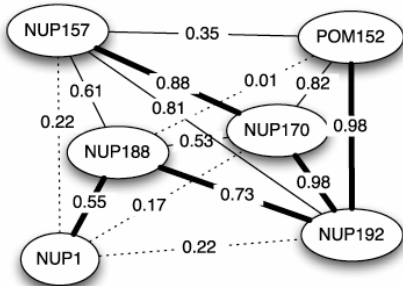




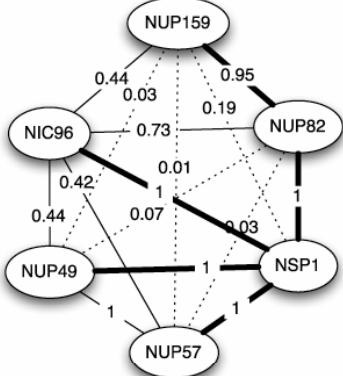
Composite 48



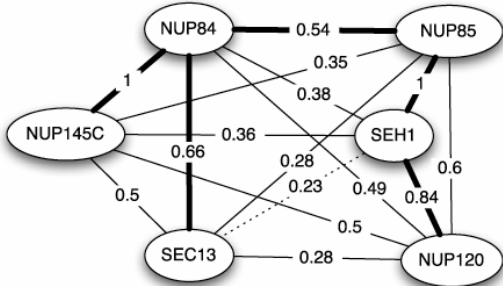
Composite 49



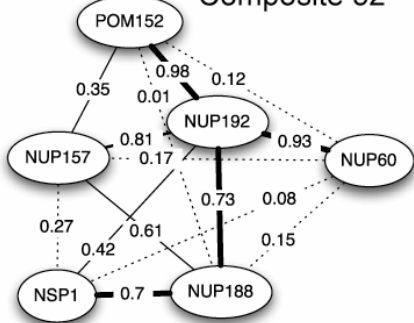
Composite 50



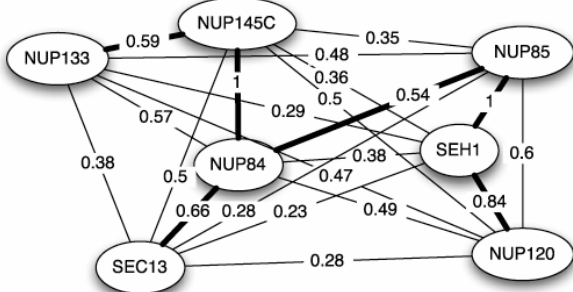
Composite 51



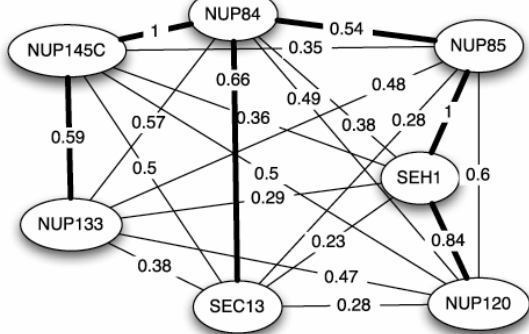
Composite 52



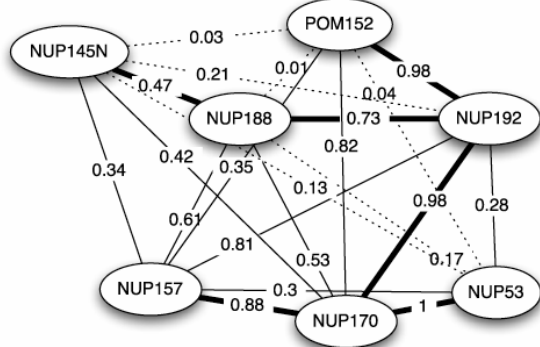
Composite 53



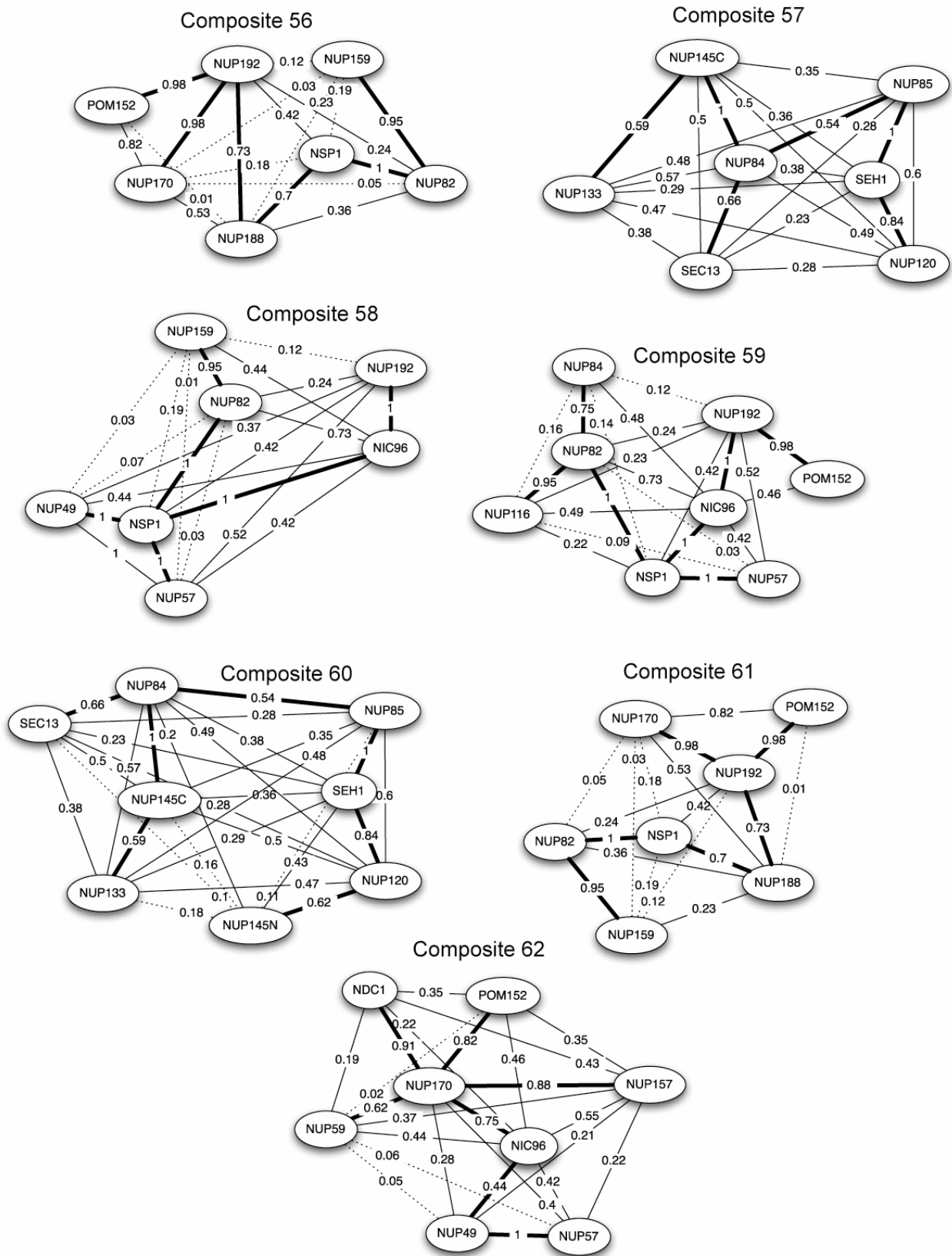
Composite 54



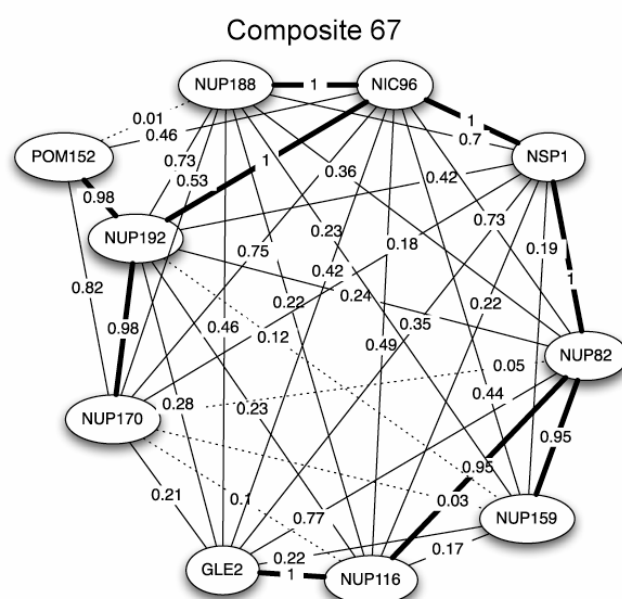
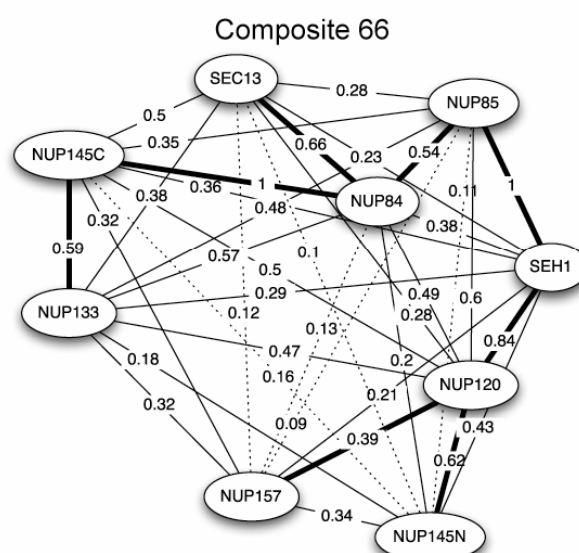
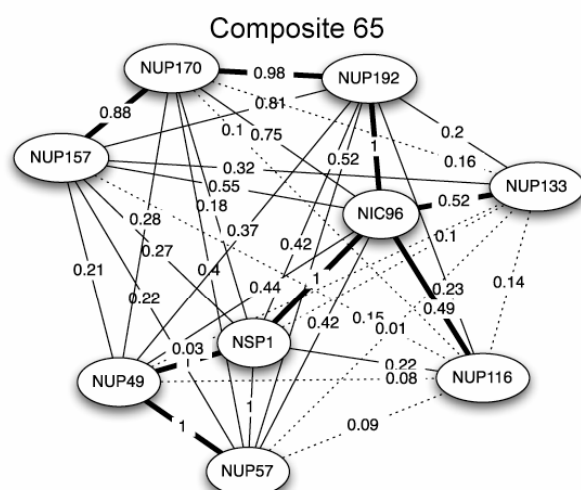
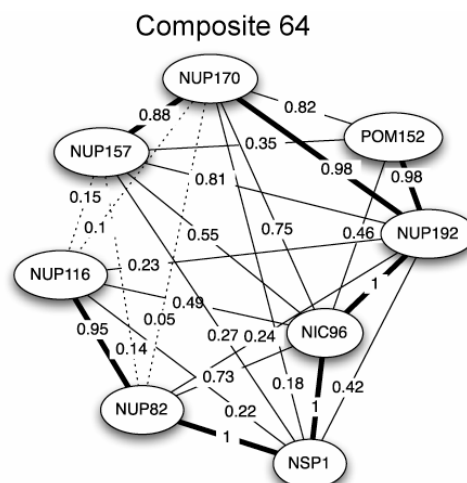
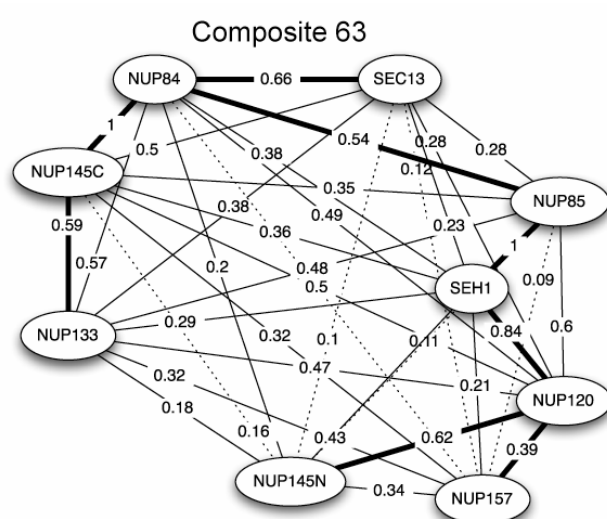
Composite 55



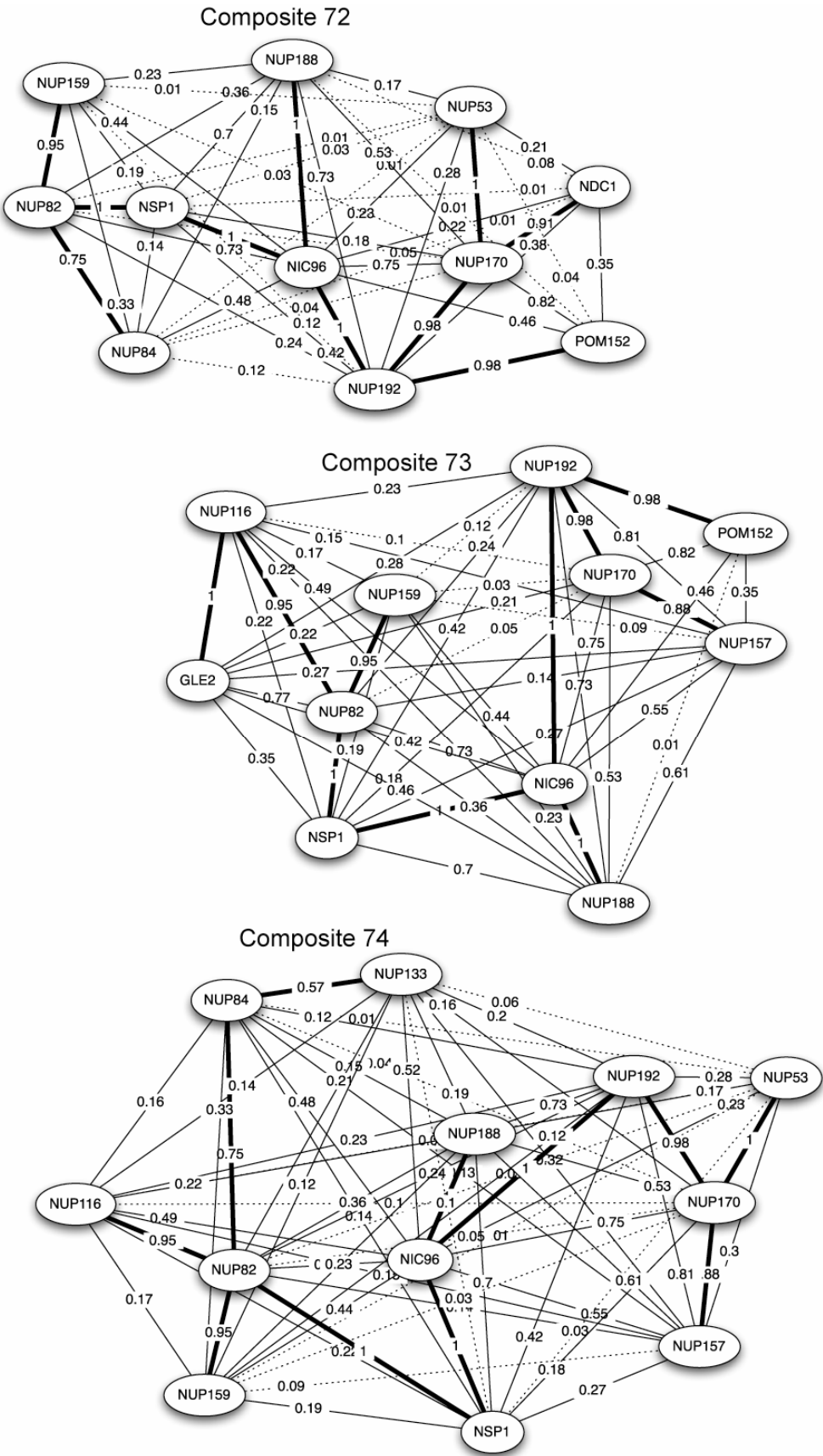






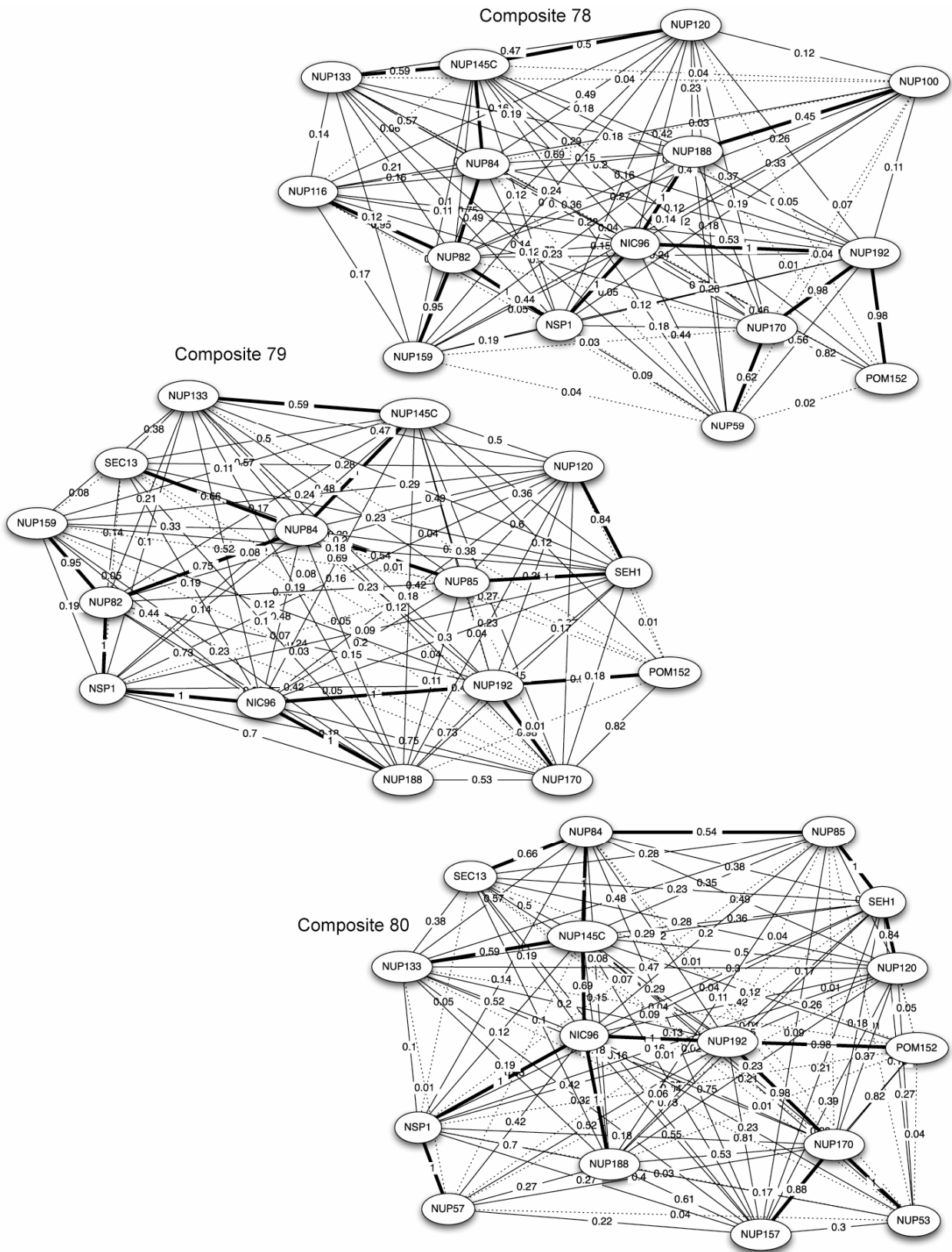


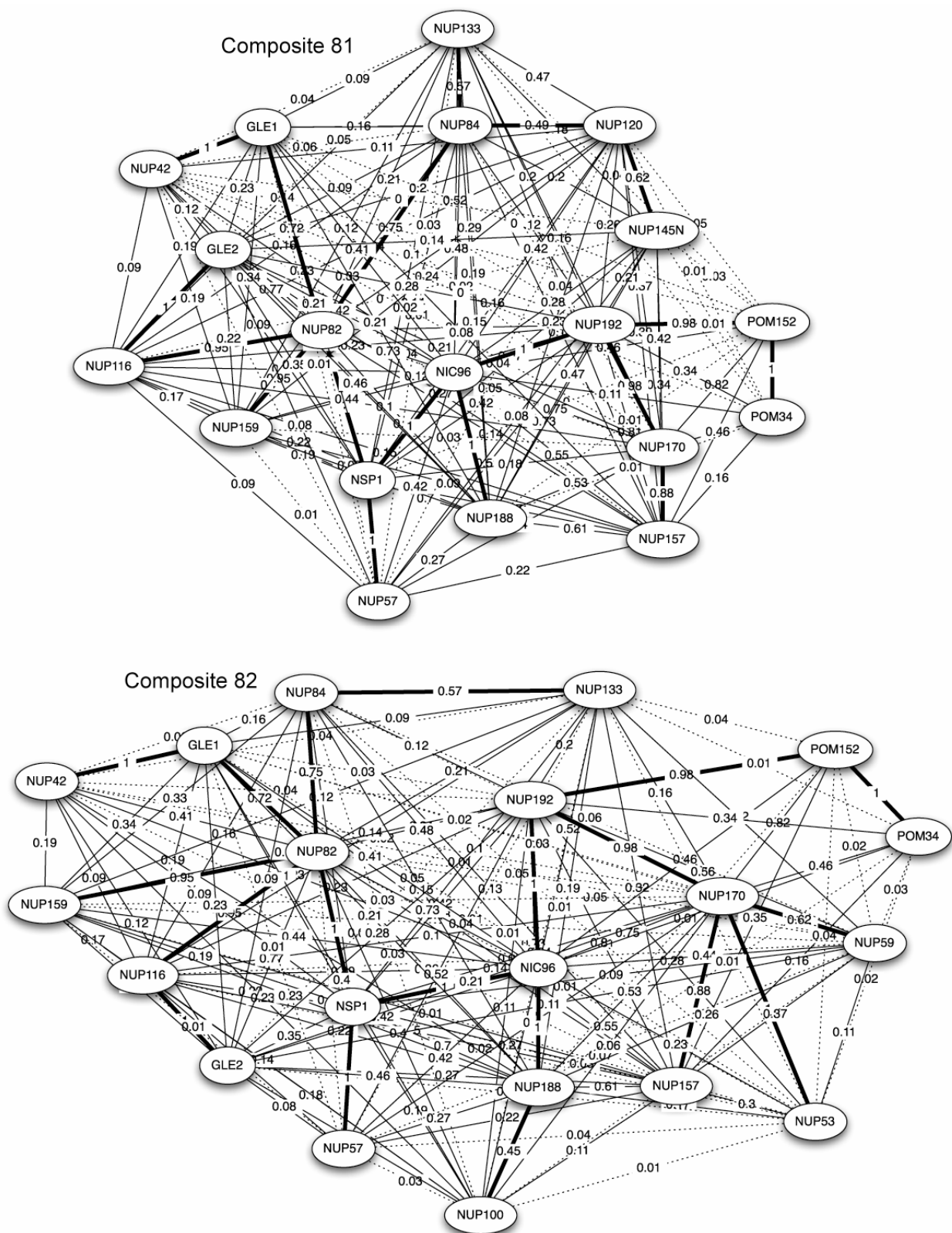












**Figure 27. Protein connectivity graphs of all composites.**

Nups are nodes connected by edges with the observed contact frequency as the edge weight. Edges that are part of the most probable composite minimal connectivity are shown in thick lines. All edges with a statistical significant reduction in contact frequencies (given the size of the corresponding composite) are indicated with dotted lines.



## 4 Supplementary Tables

**Supplementary Table 1. Number of gold particles in each montage and concentrations of heparin used to extract the NEs.**

Supplementary Table 1. Number of gold particles in each montage and concentrations of heparin used to extract the NEs		
Gene name	Number of particles	Heparin Concentration (mg/ml)
GLE1	699	0.01, 0.03
GLE2	190	0.01, 0.03
NDC1	343	0, 0.01
NIC96	298	0, 0.01, 0.03
NSP1	304	0.01, 0.03
NUP1	208	0, 0.01
NUP100	358	0.01, 0.03
NUP116	649	0.01, 0.03
NUP120	411	0.03, 0.1
NUP133	293	0.01, 0.03
NUP145C	274	0.01, 0.03
NUP145N	221	0.01, 0.03, 0.1
NUP157	325	0, 0.3
NUP159	229	0, 0.01, 0.03
NUP170	309	0.03, 0.1
NUP188	392	0.03, 0.1
NUP192	300	0.01, 0.03
NUP42	315	0.03, 0.1
NUP49	522	0.01, 0.03
NUP53	565	0.01, 0.03, 0.1
NUP57	186	0.03, 0.1
NUP59	436	0.01, 0.03
NUP60	326	0, 0.01
NUP82	229	0, 0.01
NUP84	559	0
NUP85	379	0, 0.01
POM34	674	0.1, 0.3
SEH1	483	0.01, 0.03
POM152	138	0
$\Sigma$	10615	

**Supplementary Table 2. Immunolocalization of the nups in the NPC relative to the NE.**

The estimated position of the nups relative to the central Z-axis (*R*) and the equatorial plane (*Z*) of the NPC using the best-fit analysis of the gold particle distribution in the nup montages is shown, as are the estimated errors.

Supplementary Table 2. Immunolocalization of the nups in the NPC relative to the NE				
Gene name	R (nm)	R error (nm)	Z (nm)	Z error (nm)
GLE1	26	6	15	3
GLE2	24	9	7	5
NDC1	34	6	3	4
NIC96	39	13.5	10	7.5
NSP1	29	13.5	6	6
NUP1	28	8	-18	4
NUP100	28	5	8	4
NUP116	30	5	11	4
NUP120	31	6	11	4
NUP133	36	6	15	5
NUP145C	37	10	11	4
NUP145N	26	13.5	-11	6
NUP157	27	8	2	5
NUP159	34	9	18	6
NUP170	25	8	-5	5
NUP188	26	6	7	3
NUP192	26	6	6	4
NUP42	31	9	11	4
NUP49	25	5	7	3
NUP53	33	5	6	4
NUP57	19	11	-3	5
NUP59	31	6	8	4
NUP60	32	8	-15	5
NUP82	34	16.5	22	7.5
NUP84	37	8	16	1
NUP85	36	6	17	3
POM34	33	5	2	3
SEH1	31	6	11	6
POM152	50	13	-4	6

Supplementary Table 3. Affinity purification and overlay assays of the PrA-tagged nups.

Composite Number - fold	Nup-PrA	Affinity Purified Nups <sup>(1)</sup>	Purified out non-nups (by MS)	1n/2n <sup>(2)</sup>	Source <sup>(3)</sup>	EB <sup>(4)</sup>	WB <sup>(5)</sup>	Resin <sup>(6)</sup>	Quality <sup>(7)</sup>
59 6	Nup192	POM132 (MS/MS), NUP116 (4E-3), NIC96 (6E-18), NUP84 (6E-3), NUP82 (6E-3), NSP1, NUP57 (MS/MS) NIC96 (MS/MS)	Tef3, Plk1, Plk2, Tef2, Ssa1, Ssb2, Cdc19, Pdc1, Tef1, Eno2, Adh1, Tdh3, Rps1b, Gpm1	2n 1n	NPC WCE	4 12	n/a 12	MB MB	3 3
4	Nup188	NIC96 (3E-7)	Cdc19, IgG	1n	WCE	14	20	S	3
35	Nup170	POM152(2E-13), NIC96(L), NUP145C(4E-5), NUP53(3E-15)	Tef1	1n	NE	5	n/a	S	4
32	Nup159	NUP116(MS/MS), NSP1(2E-13), NUP82(1E-27)		1n	NE	2	n/a	S	4
19		NSP1(L), NUP82 (L)		1n	WCE	9	18	MB	4
71	Nup157	POM152(1E-8), NUP133(1E-15), NUP120(1E-21), NIC96(3E-31), NUP85(5E-6), NUP84(8E-13), NUP145C(2E-6), NUP145N(8E-12), SEH1(5E-8), SEC13(2E-7)	Tef1, Lsp1	1n	NE	4	n/a	S	3
66		NUP133(MS/MS), NUP120 (L), NUP84(L), NUP85(L), NUP145C(L), NUP145N(L), SEH1(L), SEC13(L)		2n	NE	3	n/a	S	4
62	Pom152	NUP170(1E-8 and MS/MS), NUP157(MS/MS), NIC96(5E-5), NUP59(1E-15), NDC1(MS/MS), NUP57(4E-4), NUP49(7E-4)	Tef1	1n	NPC	16	n/a	S	3
64		NUP192(2E-36), NUP170(3E-13), NUP157(2E-2), NUP116(MS/MS), NSP1(1E-43), NIC96(3E-21), NUP82(2E-2 & MS/MS), POM152 (no PrA)(2E-22)	IgG	2n	WCE	8	7	MB	3
37		NUP192(MS/MS), NUP170(5E-20), NUP157(MS/MS), POM34(MS/MS)		2n	PR	n/a	n/a	n/a	3
12		POM34(5E-11)		2n	PR	n/a	n/a	n/a	3
68	Nup133	NUP116(3E-3), NUP120(4E-18), NUP85(2E-18), NUP84(2E-37), NUP82(5E-4), NUP145C(1E-7), NUP145N(1E-13), SEH1(L), SEC13(L)	Tef1	1n	NE	3	n/a	S	3
54		NUP120(8E-40), NUP85(9E-10), NUP84(3E-26), NUP145C(2E-18), SEH1(L, 3E-7), SEC13(L, 5E-19)	Cdc19	1n	WCE	11	18	MB	4
51	Nup120	NUP85(L), NUP84(3E-7), NUP145C(2E-27), SEH1(1E-10), SEC13 (L)		1n	WCE	9	18	MB	5
26	Nup116	NUP82(3E-4), GLE2 (L)		1n	NE	3	n/a	S	4
17		GLE2(4E-10)		1n	NE	14	19	S	5
67		NUP192(8E-45), NUP188(MS/MS), NUP170(2E-65), NUP159(1E-69), POM152(MS/MS), NSP1(5E-37), NIC96(4E-51), NUP82(3E-33), GLE2 (L)	Kap95, Kap123	1n	WCE	9	18	MB	5
36		NUP159(2E-15), NSP1(2E-39), NUP82(4E-16), GLE2(MS/MS)	Hsp70	2n	WCE	10	n/a	S	4
49	Nup1	NUP192 (L), NUP188 (L), NUP170(L), NUP157(1E-3), POM152(3E-15)	Kap95, Kap60	1n	WCE	7	6	MB	3
75		NUP192(2E-38), NUP188(5E-34), NUP170(1E-34), NUP159(8E-10), NUP157(1E-4), POM152(7E-26), NSP1(1E-51), NUP120(5E-7), NUP82(1E-33), NUP84(7E-39), NUP145C(MS/MS)	Mlp1, Kap95, Kap60	1n	WCE	7	6	MB	3
78	Nup100	NUP192(1E-34), NUP188(3E-18), NUP170(8E-81), NUP159(2E-81), POM152(L), NUP133(8E-72), NUP120(1E-8), NUP116(2E-23), NSP1(1E-41), NIC96(8E-53), NUP84(4E-42), NUP82(5E-35), NUP145C(5E-9), NUP59(4E-12)	Mex67	1n	WCE	10	6	MB	4
28	Nic96	NSP1(5E-18), NUP57(2E-48), NUP49(7E-43)		1n	NE	5	n/a	S	5
38		NUP192(4E-23), NSP1(L), NUP57(L), NUP49(L)		2n	WCE	9	18	MB	5
58	Nsp1	NUP192(7E-67), NUP159(6E-4), NIC96(7E-53), NUP82(9E-3), NUP57(5E-12), NUP49(4E-13)		1n	NPC	2	n/a	S	4
50		NUP159(4E-14), NIC96(6E-9), NUP82(4E-5), NUP57(L), NUP49(4E-4),		1n	WCE	2	n/a	S	5
3		NUP57		1n	OL	MM	MM	S	5
11		NUP82		1n	OL	MM	MM	S	5
5		NIC96		1n	OL	MM	MM	S	5
57	Nup85	NUP133(1E-22), NUP120(L), NUP84(2E-14), NUP145C(1E-6), SEH1(9E-6), SEC13(5E-12)		1n	NE	4	n/a	S	5
63		NUP157 (7E-4), NUP133(L), NUP120 (1E-51), NUP84(L), NUP145C(L), NUP145N(3E-5), SEH1(L), SEC13(L)	Mex67	1n	NE	2	n/a	S	5
15		SEH1 (L)		1n	WCE	14	19	S	5
45	Nup84	NUP120(L), NUP85(L), NUP145C(8E-14), SEH1(1E-13), SEC13(4E-10)		1n	NE	4	n/a	S	4
14		NUP145C(6E-40)	Cdc19, Tef1, Adh1, Tdh3	1n	WCE	14	19	S	5
39		NUP133(L), NUP120(L), NUP85(L), NUP145C(L)		1n	WCE	13	13	MB	5
33		NUP120(L), NUP85(L), NUP145C(L)		1n	WCE	11	18	MB	5
20		NUP145C(L), SEC13(L)		1n	SG	12	21	S	4
30		NUP133(L), NUP145C(L), SEC13 (L)		1n	SG	12	21	S	4
60		NUP133(MS/MS), NUP120(L), NUP85(MS/MS), NUP145C(MS/MS), NUP145N(L), SEH1(L), SEC13(L)	Tef1	2n	WCE	11	n/a	S	3
80	Nup84-GFP	NUP192, NUP188, NUP170, NUP157, POM152, NUP133, NUP120, NSP1, NIC96, NUP145C, NUP85, NUP53, NUP57, SEH1, SEC13	Ssa1, Ssb1, Tef1, Adh1, G3PD, Rpl15a, Rps24b	1n	WCE	10	10	MB	5
34	Nup82	NUP159(2E-54), NUP116(4E-5), NSP1(2E-52), NUP84(4E-7)		1n	NE	3	n/a	S	3
23		NUP159(L), NSP1(L)		1n	WCE	13	18	MB	5
56		NUP159 (MS/MS), NUP192 (MS/MS), NUP188 (MS/MS), NUP170 (MS/MS), POM152 (MS/MS), NSP1(L)		1n	WCE	9	18	MB	5
31		NUP159 (MS/MS), NSP1 (MS/MS), NUP82 (no PrA) (MS/MS)		2n	WCE	13	18	MB	5
61		NUP159(L), NUP192(L), NUP188(L), NUP170(L), POM152(L), NSP1(L), NUP82 no PrA (L)		2n	WCE	9	18	MB	5
2		NSP1		1n	OL	MM	MM	S	5
53	Nup145C	NUP133(2E-80), NUP120(8E-25), NUP85(1E-38), NUP84(3E-57), SEH1(1E-18), SEC13(3E-11)	IgG	2n	WCE	13	13	MB	5
40	Ndc1	NUP192(8E-9), NUP170(3E-4), POM152(MS/MS), NUP59(1E-22)		1n	NE	3	n/a	S	3
22		NUP170 (MS/MS), POM152(MS/MS)	Tef1, Cdc19	1n	WCE	15	n/a	S	3
72		NUP192(9E-44), NUP188(7E-19), NUP170(3E-54), NUP159(2E-61), POM152(3E-53), NSP1(2E-82), NIC96(4E-35), NUP82(2E-10), NUP84(3E-3), NUP53(3E-5)	Tef3, Tef1	2n	WCE	7	18	MB	4
43	Gle1	NUP159(1E-20), NUP116(6E-10), NUP100(4E-13), NUP82(5E-9), NUP42(1E-12)	Tef1	1n	NE	3	n/a	S	3
9		NUP42(MS/MS)		1n	WCE	12	n/a	S	5
81		NUP192(L), NUP188(L), NUP170(2E-46), NUP159(L), POM152(4E-68), NUP157(4E-41), NUP133(4E-59), NUP116(2E-15), NSP1(1E-27), NUP120(MS/MS), NIC96(1E-59), NUP82(4E-23), NUP84(1E-10), NUP145N(MS/MS), NUP57(3E-11), GLE2(MS/MS), POM34(MS/MS), NUP42(MS/MS)	Mex67, Cdc19, EF1a, Eno2, Adh1, Tdh3	1n	WCE	9	18	MB	5
55	Nup145N	NUP192(L), NUP188(L), NUP170(5E-5), NUP157(L), POM152(L), NUP53 (4E-12)		2n	NE	5	n/a	S	3
21	Nup60	NUP192(3E-8), POM152(5E-4)	Kap123, Nup2, Kap95, Kap60	2n	WCE	7	18	MB	3
52		NUP192(5E-19), NUP180(5E-10), POM152(2E-20), NUP157(MS/MS), NSP1(L)	Ura2, Kap123, Nup2, Kap95, Kap60	2n	WCC	7	7	MD	3
24	Nup59	NUP192(6E-19), NUP170(4E-4)		1n	NE	2	n/a	S	3
76		NUP192(3E-45), NUP188(6E-42), NUP170(1E-69), NUP159(9E-3), NUP157(7E-30), NSP1(5E-57), POM152(1E-79), NIC96(1E-49), NUP82(1E-8), NUP53(7E-12), POM34(7E-13), GLE2(4E-5 and MS/MS)		2n	WCE	7	18	MB	4

27	Nup57	NSP1(L,1E-37), NIC96(L,2E-60), NUP49(L,3E-19)		1n	NE	2	n/a	S	4
48		NUP192 (L), POM152(L), NSP1(L), NIC96(L), NUP49(L)		1n	WCE	9	18	MB	5
16		NSP1		1n	OL	MM	MM	S	5
18		NUP49		1n	OL	MM	MM	S	5
42	Nup53	NUP170(2E-40), NUP120(2E-7 and MS/MS), NUP145N(3E-29), SEH1(3E-14)	Kap121, Lsp1	1n	NE	2	n/a	S	3
13		NUP170(3E-35)		1n	WCE	17	n/a	S	4
74		NUP159(1E-51), NUP192(3E-14), NUP188(8E-19), NUP170(4E-39), NUP157(3E-4 and MS/MS), NUP133(2E-37), NUP116(1E-7), NIC96(6E-26), NUP84(5E-26), NUP82(3E-27), NSP1(1E-61)	Kap121	2n	WCE	10	10	MB	4
69		NUP159(L), NUP192(L), NUP188(L), NUP170(L), NUP116(L), NIC96(L), NUP84(L), NUP82(L), NSP1(L)	Kap121	2n	WCE	10	10	MB	3
65	Nup49	NUP192(1E-45), NUP170(2E-29), NUP157(8E-41), NUP133(2E-75), NUP116(4E-20), NIC96(5E-43), NSP1(8E-47), NUP57(3E-29)		1n	NE	2	n/a	S	4
47		NUP192(2E-34), POM152(5E-12), NSP1(L), NIC96(L), NUP57(L)		1n	WCE	9	18	MB	5
29		NSP1(L), NIC96(L), NUP57(L)		1n	WCE	13	18	MB	5
8		NUP57		1n	OL	MM	MM	S	5
10	Nup42	GLE1(4E-19)	IgG	1n	NE	3	n/a	S	5
82		NUP192 (L), NUP188 (L), NUP170 (L), NUP159 (L), NUP157(2E-33), POM152(2E-67), NUP133(2E-52), NUP116(2E-26), NUP100 (MS/MS), NSP1(4E-82), NIC96(3E-48), NUP84(3E-27), NUP82(1E-22), GLE1(2E-5 and MS/MS), NUP59(1E-11), NUP53(5E-12), NUP57(6E-11), GLE2(3E-6 and MS/MS), POM34(2E-7 and MS/MS)	Cdc19, Tdh3	2n	WCE	7	18	MB	5
77		NUP159(L), NUP192(L), NUP188(L), NUP170(L), NUP116(L), NUP100(L), NSP1(L), NIC96(L), NUP84(L), NUP82(L), GLE1(L), NUP59(L), NUP53(L), NUP57(L)		2n	WCE	10	18	MB	3
46		Gle2	NUP159(3E-20), NSP1(3E-88), NUP116(8E-21), NUP82(3E-36), NUP84,(2E-6)		1n	NE	3	n/a	S
1	NUP116(1E-11)			1n	NE	4	n/a	S	5
73	NUP192(L), NUP188(L), NUP170(L), NUP157 (L), POM152(L), NUP159(L), NUP116(L), NSP1(L), NIC96(L), NUP82(3E-19)		Hsp71	1n	WCE	9	18	MB	3
44	Seh1	NUP120(8E-19), NUP84(1E-38), NUP85(3E-7), NUP145C(1E-21), SEC13(2E-26)		1n	NE	10	n/a	S	5
7		NUP85 (L)		1n	WCE	14	19	S	5
79		NUP192(1E-66), NUP188(2E-29), NUP170(6E-60), NUP159(2E-14), POM152(1E-33), NUP133(5E-89), NSP1(2E-56), NUP120(4E-44), NIC96(2E-56), NUP85(2E-17), NUP84(4E-49), NUP145C(2E-36), NUP82(8E-11), SEC13(4E-36)	Ura2, Iml1p, Yol138, Ydr128, Ybl104c, Cdc19, Tef1, Eno2, Adh1, Tdh3	1n	WCE	9	18	MB	5
41	Pom34	NUP170(MS/MS), NUP157(2E-5), POM152(4E-67), NDC1(2E-5 and MS/MS)	Tef1	1n	WCE	15	n/a	S	4
70		NUP192(L, 9E-44), NUP188(L, 5E-18), NUP170(L,3E-54), NUP159(L,2E-61), POM152(L, 3E-53), NSP1(L, 2E-82), NIC96(L, 4E-35), NUP84(3E-8), NUP82(5E-5)	Tef3, Tef1	2n	WCE	7	18	MB	4
25	Sec13	NUP84(5E-9), NUP145C(4E-7)	Sec31, Sec23, Sec24, Tef1, Adh1	2n	WCE	19	n/a	S	4

- (1)

L - a protein was identified by "lining up" the band in that lane with the band in the adjacent lane of a similar or identical preparation (data not shown) where the protein in that band was identified by MS.
- (2)

1n - haploid yeast; 2n - diploid yeast
- (3)

NPC - enriched nuclear pore complex preparations; NE - nuclear envelope preparations; WCE - whole cell extracts; PR - Pom rings; OL - data from overlay assay; SG - sucrose gradient velocity centrifugation of Nup84 subcomplex
- (4)

Extraction Buffers (EB)

1TB ( 20 mM K/HEPES pH 7.4, 110 mM KOAc, 2 mM MgCl<sub>2</sub>)

2TB, 1% Triton X-100, 150 mM NaCl

3TB, 1% Triton X-100, 250 mM NaCl

4TB, 1% Triton X-100, 500 mM NaCl

5TB, 1% Triton X-100, 1M NaCl

6TBT ( 20 mM K/HEPES pH 7.4, 110 mM KOAc, 2 mM MgCl<sub>2</sub>, 0.1% Tween 20)

7TBT, 1% Triton X-100

8TBT, 1% Triton X-100, 50 mM NaCl

9TBT, 1% Triton X-100, 75 mM NaCl

10TBT, 1% Triton X-100, 150 mM NaCl

11TBT, 1% Triton X-100, 250 mM NaCl

12TBT, 1% Triton X-100, 500 mM NaCl

13TBT, 1% Triton X-100, 1M NaCl

1420 mM K/HEPES pH 7.4, 0.1 mM MgCl<sub>2</sub>, 1% Triton X-100, 0.5% sodium deoxycholate, 0.1% sodium N-lauroyl-sarcosine

1520 mM K/HEPES pH 7.4, 0.1 mM MgCl<sub>2</sub>, 0.2% sodium deoxycholate, 0.01% sodium N-lauroyl-sarcosine

1610 mM Bis-Tris pH 6.5, 0.1 mM MgCl<sub>2</sub>, 20% DMSO, 1% Triton X-100, 20 mM PIPES pH 6.8, 0.3 mg/ml

1750 mM Tris-HCl pH 8, 1% Triton X-100, 150 mM NaCl

MMSee "Materials and Methods" for details

n/anot applicable

(5)

Washing Buffers (WB)

18TBT, 1mg/ml heparin

1920 mM K/HEPES, pH 7.4, 1 mM EDTA, 0.1% Triton X-100, 0.05% sodium deoxycholate, 0.01% sodium N-lauroyl-sarcosine

20150 mM MgCl<sub>2</sub>

21TBT, 0.5% Triton X-100

(6)

S - Sepharose; MB - magnetic beads; n/a - not applicable

(7)

All the immunoprecipitations were given a "quality note" based on a five note scale: 5-excellent, 4-good, 3-average. Immunoprecipitation with a quality mark below 3 were omitted.

(8)

Cristea et al., Mol Cell Proteomics, 2005, 4(12), 1933-41
- www.nature.com/nature
- 90

**Supplementary Table 4. Conditions used for sedimentation analysis of nups and their complexes.**

Gene name	Buffer <sup>(1)</sup>	Time (hours) at ~300.000 g <sub>max</sub>
NUP192	A	8
NUP188	B	8
NUP170	A	8
NUP159	C	12
NUP157	A	8
POM152	A	11
NUP133	A	11
NUP120	A	11
NUP116	C	13
NUP1	C	13
NUP100 <sup>(2)</sup>	D	13
NIC96	A	13
NSP1	A	14
NUP85	A	14
NUP84	D	14
NUP82	A	14
NUP145C	A	14
NUP145N <sup>(2)</sup>	D	19
NDC1	A	15
GLE1 <sup>(2)</sup>	D	15
NUP60	C	17
NUP59	C	17
NUP57 <sup>(2)</sup>	D	17
NUP53	A	17
NUP49	C	17
NUP42	C	17
GLE2	A	20
SEH1	A	20
POM34	A	20
SEC13	A	20
PrA	D	28
Nup84 subcomplex	E	8

**(1) Buffer composition**

A: sucrose in TB, 0.1% Tween 20, 1 mM DTT, 1:200 solution P, 1:200 PIC

B: sucrose in TB, 0.1% Tween 20, 0.1% Triton X100, 0.03% sodium N-lauroyl-sarcosine, 0.05% sodium deoxycholate, 1 mM DTT, 1:200 solution P, 1:200 PIC.

C: sucrose in 20 mM HEPES pH 8.2, 110 mM KOAc, 1 mM EDTA, 0.1% Tween 20, 1 mM DTT, 1:100 solution P, 1:100 PIC

D: sucrose in 20 mM HEPES pH 7.4, 1 mM EDTA, 0.1% Triton X-100, 0.03% sodium N-lauroyl-sarcosine, 0.05% sodium deoxycholate, 1 mM DTT, 1:100 solution P, 1:100 PIC
































E: sucrose in 50 mM Tris-HCl, pH 8, 0.5 mM EDTA, 300 mM NaCl, 1 mM DTT, 1:100 solution P, 1:100 PIC

**(2) Cleared whole cell extracts of ground yeast cells were analyzed**

Supplementary Table 5: Bead representations of each nup and their stoichiometries.

$\tau$  is the nup type.  $N_\tau^\theta$  is the number of nup instances of type  $\tau$  in each cytoplasmic ( $\theta = 1$ ) and nucleoplasmic half-spoke ( $\theta = 2$ ).  $\{B_j^\kappa\}$  is the set of beads for each nup at representation  $\kappa$ .  $n_\kappa$  is the total number of particles (beads) per nup representation  $\kappa$ .  $r$  is the radius of each bead. Each nup is described with up to 9 representations  $\kappa$ . The Cartesian coordinates of beads in representations at  $\kappa > 1$  are inherited from particles in the root representation; these beads are shown opaque whereas all other beads in the root representation are translucent.





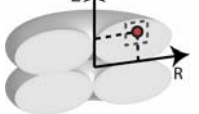
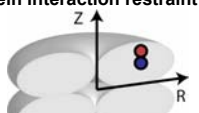
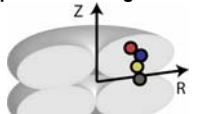
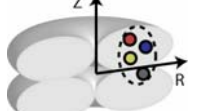
$\tau$	$N_\tau^1$	$N_\tau^2$	$\kappa$	$\{B_j^\kappa\}$	$n_\kappa$	$r$	$\tau$	$N_\tau^1$	$N_\tau^2$	$\kappa$	$\{B_j^\kappa\}$	$n_\kappa$	$r$
Nup192	1	1	1,2,5		2	3.0	Nup1	0	1	1,5		9	1.5
			3	-	1	-				2		2	1.5
Nup188	1	1	1,2,5		2	3.0				3	-	1	-
			3	-	1	-				4		7	1.5
Nup170	1	1	1,2,5		2	2.9	Nsp1	2	2	1,5		12	1.3
			3	-	1	-				2		3	1.3
Nup157	1	1	1,2,5		3	2.5				3	-	1	-
			3	-	1	-				4		9	1.3
Nup133	1	1	1,2,5		2	2.7	Gle1	1	0	1,2,5		2	2.1
			3	-	1	-				3	-	1	-
Nup120	1	1	1,2,5		2	2.6	Nup60	0	1	1,5		4	1.6
			3	-	1	-				2,3		1	1.6
Nup85	1	1	1,2,5		3	2.0	Nup59	1	1	4		3	1.6
			3	-	1	-				1,5		4	1.6
Nup84	1	1	1,2,5		3	2.0				2		2	1.6
			3	-	1	-				3	-	1	-
Nup145C	1	1	1,2,5		2	2.3	Nup57	1	1	4		2	1.6
			3	-	1	-				1,5		3	1.8
Seh1	1	1	1,2,3,5		1	2.2				2,3		1	1.8
Sec13	1	1	1,2,3,5		1	2.1				4		2	1.8
Gle2	1	1	1,2,3,5		1	2.3	Nup53	1	1	1,5		3	1.7
Nic96	2	2	1,2,5		2	2.4				2,3		1	1.7
			3	-	1	-				4		2	1.7
Nup82	1	1	1,2,5		2	2.3	Nup145N	0	2	1,5		6	1.5
			3	-	1	-				2,3		1	1.5

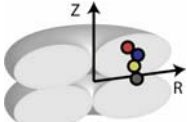
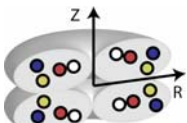
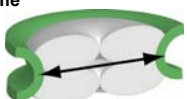
Pom152	1	1	1		10	1.6	Nup49	1	1	4		5	1.5
			2,3,8		1	1.6				1,5		3	1.7
			4	-	0	1.6				2,3		1	1.7
			5		7	1.6				4		2	
			6		1	1.6	Nup42	1	0	1,5		5	1.4
			7		6	1.6				2,3		1	1.4
			9		2	1.6				4		4	1.4
Ndc1	1	1	1		2	2.2	Nup159	1	0	1,5		11	1.6
			2,3,8		1	2.2				2		2	1.6
			4,5,7		0	2.2				3	-	1	-
			6		1					4		9	1.6
							Nup116	1	0	1,5		13	1.4
Pom34	1	1	1		3	1.5				2		1	1.4
			2		2	1.5				3	-		
			3	-	1	-				4		11	1.4
			4,5,7		0		Nup100	1	0	1,5		13	1.4
			6		1	1.5				2		3	1.4
			8		2	1.5				3	-	1	-
										4		10	1.4



**Supplementary Table 6: Form of the restraints.**

Each restraint is a harmonic function  $(f - f_0)^2 / \sigma^2$ , where  $f$  is the restrained feature and  $\sigma$  is the parameter that determines the strength of the restraint. For upper bounds, the score is 0 for  $f < f_0$ ; for lower bounds, the score is 0 for  $f > f_0$ . The restraints are imposed on particles  $B = \{B_j^\kappa(\theta, s, \tau, i)\}$  where  $\theta \in (1, 2)$ ,  $s \in (1, 2, \dots, 8)$ ,  $\tau \in T^\theta$ ,  $i \in (1, 2, \dots, N_\tau^\theta)$ ,  $\kappa \in (1, 2, \dots, N_\kappa^\tau)$ ,  $j \in (1, 2, \dots, n_\kappa)$ , unless explicitly stated otherwise.  $R_C$  is the total number of conditional restraints,  $R_O$  is the number of all optional restraints, and  $R_A$  is the total number of activated restraints.  $N_\tau^\theta$  is the number of nup instances of type  $\tau$  in a half-spoke of type  $\theta$ .  $n_\kappa$  is the total number of beads for a nup of type  $\tau$  and representation  $\kappa$ .

Data generation		Data interpretation				
Method	Experiments	Restraint	$R_c$	$R_o$	$R_A$	Functional form of activated feature restraint
Bioinformatics and Membrane fractionation	30 nup sequences	Protein excluded volume restraint 	-	-	1,864 1,863/2	<b>Protein-protein:</b> Violated for $f < f_0$ , $f$ is the distance between two beads, $f_0$ is the sum of the bead radii, and $\sigma$ is 0.01 nm. Applied to all pairs of particles in representation $\kappa=1$ : $B^m = \{B_j^{\kappa=1}(\theta, s, \tau, i)\}$
	30 nup sequences	Surface localization restraint 	-	-	48	<b>Membrane-surface location:</b> Violated if $f \neq f_0$ , $f$ is the distance between a protein particle and the closest point on the NE surface (half-torus), $f_0 = 0$ nm, and $\sigma$ is 0.2 nm. Applied to particles: $B^m = \{B_j^{\kappa=6}(\theta, s, \tau, i)   \tau \in (\text{Ndc1}, \text{Pom152}, \text{Pom34})\}$
	30 Nup sequences and immuno-EM (see below)		-	-	64	<b>Pore-side volume location:</b> Violated if $f < f_0$ , $f$ is the distance between a protein particle and the closest point on the NE surface (half-torus), $f_0 = 0$ nm, and $\sigma$ is 0.2 nm. Applied to particles: $B^m = \{B_j^{\kappa=8}(\theta, s, \tau, i)   \tau \in (\text{Ndc1}, \text{Pom152}, \text{Pom34})\}$
			-	-	80	<b>Perinuclear volume location:</b> Violated if $f > f_0$ , $f$ is the distance between a protein particle and the closest point on the NE surface (half-torus), $f_0 = 0$ nm, and $\sigma$ is 0.2 nm. Applied to particles: $B^m = \{B_j^{\kappa=7}(\theta, s, \tau, i)   \tau \in (\text{Pom152})\}$
Hydrodynamics experiments	1 S-value	Complex shape restraint 	1	164	1	<b>Complex diameter</b> Violated if $f < f_0$ , $f$ is the distance between two protein particles representing the largest diameter of the largest complex, $f_0$ is the complex maximal diameter $D=19.2R$ , where $R$ is the sum of both particle radii, and $\sigma$ is 0.01 nm. Applied to particles of proteins in composite $C_{45}$ : $B^m = \{B_j^{\kappa=1}(\theta, s, \tau, i)   \tau \in C_{51}\}$
	30 S-values	Protein chain restraint 	-	-	1,680	<b>Protein chain</b> Violated if $f \neq f_0$ , $f$ is the distance between two consecutive particles in a protein, $f_0$ is the sum of the particle radii, and $\sigma$ is 0.01 nm. Applied to particles: $B = \{B_j^{\kappa}(\theta, s, \tau, i)   \kappa = 1\}$
Immuno-Electron microscopy	10,940 gold particles	Protein localization restraint 	-	-	456	<b>Z-axis position</b> Violated for $f < f_0$ , $f$ is the absolute Cartesian Z-coordinate of a protein particle, $f_0$ is the lower bound defined for protein type $\tau$ , and $\sigma$ is 0.1 nm. Applied to particles: $B = \{B_j^{\kappa}(\theta, s, \tau, i)   \kappa = 1, j = 1\}$
			456	Violated for $f > f_0$ , $f$ is the absolute Cartesian Z-coordinate of a protein particle, $f_0$ is the upper bound defined for protein type $\tau$ , and $\sigma$ is 0.1 nm. Applied to particles: $B = \{B_j^{\kappa}(\theta, s, \tau, i)   \kappa = 1, j = 1\}$		
			-	-	456	<b>Radial position</b> Violated for $f < f_0$ , $f$ is the radial distance between a protein particle and the Z-axis in a plane parallel to the X and Y axes, $f_0$ is its lower bound defined for protein type $\tau$ , and $\sigma$ is 0.1 nm. Applied to particles: $B = \{B_j^{\kappa}(\theta, s, \tau, i)   \kappa = 1, j = 1\}$
			456	Violated for $f > f_0$ , $f$ is the radial distance between a protein particle and the Z-axis in a plane parallel to the X and Y axes, $f_0$ is its upper bound defined for protein type $\tau$ , and $\sigma$ is 0.1 nm. Applied to particles: $B = \{B_j^{\kappa}(\theta, s, \tau, i)   \kappa = 1, j = 1\}$		
Overlay assays	13 contacts	Protein interaction restraint 	20	112	20	<b>Protein contact</b> Violated for $f > f_0$ , $f$ is the distance between two protein particles, $f_0$ is the sum of the particle radii multiplied by a tolerance factor of 1.3, and $\sigma$ is 0.01 nm. Applied to particle: $B = \{B_j^{\kappa}(\theta, s, \tau, i)   \kappa \in (2, 4, 9), \theta \in (1, 2, 3)\}$
Affinity purification	4 complexes	Competitive binding restraint 	1	132	4	<b>Protein contact</b> Violated for $f > f_0$ , $f$ is the distance between two protein particles, $f_0$ is the sum of the particle radii multiplied by a tolerance factor of 1.3, and $\sigma$ is 0.01 nm. Applied to : $B = \{B_j^{\kappa}(\theta, s, \tau, i)   \theta \in (1, 2, 3), \kappa \in (2, 4, 6), \tau = (\text{Nup82}, \text{Nic96}, \text{Nup49}, \text{Nup57})\}$
	64 complexes	Protein proximity restraint 	692	25,348	692	<b>Protein proximity</b> Violated for $f > f_0$ , $f$ is the distance between two protein particles, $f_0$ is the maximal diameter of a composite complex, and $\sigma$ is 0.01 nm. Applied to particles: $B = \{B_j^{\kappa}(\theta, s, \tau, i)   \theta \in (1, 2, 3), \kappa \in (2, 4, 9)\}$

		<div>Composite connectivity restraint</div> 	64	94,637	413	<div>Protein contact</div> <p>Violated for <math>f &gt; f_o</math>. <math>f</math> is the distance between two protein particles, <math>f_o</math> is the sum of the particle radii multiplied by a tolerance factor of 1.3, <math>\sigma</math> is set to 0.01 nm (Supplementary Table 8). Applied to particles:</p> $B = \{B_j^\kappa(\theta, s, \tau, i)   \theta \in (1, 2, 3), \kappa \in (2, 4, 9)\}$
Electron microscopy	EM map	<div>Assembly symmetry restraint</div> 	-	-	2,097,441	<div>Protein group configuration</div> <p>Violated if <math>f \neq f_o</math>. <math>f</math> is the difference of two distances between two pairs of protein particles, <math>f_o</math> is 0, and weight <math>\sigma</math> is set to 0.4-2.4 nm. Applied to particles:</p> $B = \{B_j^\kappa(\theta, s, \tau, i)   \kappa = 2\}$
					6	<p>Violated if <math>f \neq f_o</math>. <math>f</math> is the difference between a pair of dihedral angles, each defined by four protein particles, <math>f_o</math> is 0, and <math>\sigma</math> is 0.01 nm. Applied to particles:</p> $B = \{B_j^\kappa(\theta, s, \tau, i)   \kappa = 2\}$
		<div>Nuclear volume Envelope excluded</div> 	-	-	876*1864/2	<div>NE-protein:</div> <p>Violated for <math>f &lt; f_o</math>. <math>f</math> is the distance between a protein and a NE bead, <math>f_o</math> is the sum of the bead radii, and <math>\sigma</math> is 0.01 nm. Applied to all pairs of particles of the NE and proteins:</p> $B = \{B_j^\kappa(\theta, s, \tau, i)   \kappa = 5\}$

**Supplementary Table 7. Upper and lower bounds on radial and Z-coordinate of the C-terminal bead of nups at the cytoplasmic and nucleoplasmic sides.**

\*No experimental data was available for Sec13, instead the maximal possible extension of the NPC was chosen as the boundary limits. \*\*If lower bounds lie on the opposite side of the equatorial plane, their values were set to zero and the corresponding error range was increased by 50 %.

Restraint type	Z-axial restraints				R-radial restraints	
	For all proteins in $U_s^\theta$ with $\theta = 1, s = 1 - 8$		For all proteins in $U_s^\theta$ with $\theta = 2, s = 1 - 8$		For all proteins $U_s^\theta$ with $\theta = 1 - 2, s = 1 - 8$	
Protein type ( $\tau$ )	Min Z	Max Z	Max Z	Min Z	Min R	Max R
GLE1	11.0	17.0	-	-	20.0	32.0
GLE2	2.0	12.0	-2.0	-12.0	15.0	33.0
NDC1**	0.0	9.0	0.0	-9.0	28.0	140.0
NIC96	2.5	17.5	-2.5	-17.5	25.5	52.5
NSP1	0.0	12.0	0.0	-12.0	15.5	42.5
NUP1	-	-	-14.0	-22.0	20.0	36.0
NUP100	4.0	12.0	-	-	23.0	33.0
NUP116	7.0	15.0	-7.0	-15.0	25.0	35.0
NUP120	7.0	15.0	-7.0	-15.0	25.0	38.0
NUP133	10.0	20.0	-10.0	-20.0	30.0	42.0
NUP145C	7.0	15.0	-7.0	-15.0	27.0	47.0
NUP145N	-	-	-5.0	-17.0	12.5	39.5
NUP157**	0.0	9.5	0.0	-9.5	19.0	35.0
NUP159	12.0	24.0	-	-	25.0	43.0
NUP170**	0.0	7.5	0.0	-7.5	17.0	33.0
NUP188	4.0	10.0	-4.0	-10.0	20.0	33.0
NUP192	2.0	10.0	-2.0	-10.0	20.0	32.0
NUP42	7.0	15.0	-	-	22.0	40.0
NUP49	4.0	10.0	-4.0	-10.0	20.0	30.0
NUP53	2.0	10.0	-2.0	-10.0	28.0	38.0
NUP57**	0.0	7.5	0.0	-7.5	8.0	30.0
NUP59	4.0	12.0	-4.0	12.0	25.0	37.0
NUP60	-	-	-10.0	-20.0	24.0	40.0
NUP82	14.5	29.5	-14.5	-29.5	17.5	50.5
NUP84	15.0	17.0	-15.0	-17.0	29.0	45.0
NUP85	14.0	20.0	-14.0	-20.0	30.0	42.0
POM152**	0.0	9.5	0.0	-9.5	37.0	163.0
POM34**	0.0	6.5	0.0	-6.5	28.0	138.0
SEH1	5.0	17.0	-5.0	-17.0	25.0	37.0
SEC13*	0.0	50.0	0.0	-50.0	5.0	55.0

**Supplementary Table 8. Binary protein interactions and composite compositions from overlay assays (OL) and affinity purification experiments (AP).**

For affinity purification experiments, the tagged proteins are indicated in bold fonts. The order of proteins in each row is arbitrary. We also show a quality score indicating the relative reliability of the experimental data (from top to medium quality ranging from 1 to 3). A homotypic interaction of Pom152 was imposed based on the Pom ring data (supplementary Figure 6) supported by the diploid pull outs.

	Binary protein interactions	Quality	Method
1	Gle2-Nup116	1	AP
2	Nup82-Nsp1	1	OL
3	Nsp1-Nup57	1	OL
4	Nup188-Nic96	1	AP
5	Nsp1-Nic96	1	OL
6	Nup192-Nic96	1	AP
7	Seh1-Nup85	1	AP
8	Nup49-Nup57	1	OL
9	Gle1-Nup42	1	AP
10	Nup42-Gle1	1	AP
11	Nsp1-Nup49	1	OL
12	Pom152-Pom34	1	AP
13	Nup53-Nup170	1	AP
14	Nup84-Nup145C	1	AP
15	Nup85-Seh1	1	AP
16	Nup57-Nsp1	1	OL
17	Nup116-Gle2	1	AP
18	Nup57-Nup49	1	OL
	<b>Composites</b>	<b>Quality</b>	
19	<b>Nup159</b> -Nsp1-Nup82	2	AP
20	<b>Nup84</b> -Nup145C-Sec13	2	AP
21	<b>Nup60</b> -Nup192-Pom152	3	AP
22	<b>Ndc1</b> -Nup170-Pom152	3	AP
23	<b>Nup82</b> -Nup159-Nsp1	1	AP
24	<b>Nup59</b> -Nup192-Nup170	3	AP
25	<b>Sec13</b> -Nup84-Nup145C	2	AP
26	<b>Nup116</b> -Nup82-Gle2	1	AP
27	<b>Nup57</b> -Nsp1-Nic96-Nup49	1	AP
28	<b>Nic96</b> -Nsp1-Nup57-Nup49	1	AP
29	<b>Nup49</b> -Nsp1-Nic96-Nup57	1	AP
30	<b>Nup84</b> -Nup133-Nup145C-Sec13	2	AP
31	<b>Nup82</b> -Nup159-Nsp1-Nup82	1	AP
32	<b>Nup159</b> -Nup116-Nsp1-Nup82	2	AP
33	<b>Nup84</b> -Nup120-Nup85-Nup145C	1	AP
34	<b>Nup82</b> -Nup159-Nup116-Nsp1-Nup84	3	AP
35	<b>Nup170</b> -Pom152-Nic96-Nup145C-Nup53	2	AP

36	<b>Nup116</b> -Nup159-Nsp1-Nup82-Gle2	2	AP
37	<b>Pom152</b> -Nup192-Nup170-Nup157-Pom34	2	AP
38	<b>Nic96</b> -Nup192-Nsp1-Nup57-Nup49	1	AP
39	<b>Nup84</b> -Nup133-Nup120-Nup85-Nup145C	1	AP
40	<b>Ndc1</b> -Nup192-Nup170-Pom152-Nup59	3	AP
41	<b>Pom34</b> -Ndc1-Pom152-Nup157-Nup170	2	AP
42	<b>Nup53</b> -Nup170-Nup120-Nup145N-Seh1	3	AP
43	<b>Gle1</b> -Nup159-Nup116-Nup100-Nup82-Nup42	3	AP
44	<b>Seh1</b> -Nup120-Nup84-Nup85-Nup145C-Sec13	1	AP
45	<b>Nup84</b> -Nup145C-Seh1-Sec13-Nup85-Nup120	3	AP
46	<b>Gle2</b> -Nup159-Nsp1-Nup116-Nup82-Nup84	2	AP
47	<b>Nup49</b> -Nup192-Pom152-Nsp1-Nic96-Nup57	1	AP
48	<b>Nup57</b> -Nup192-Pom152-Nsp1-Nic96-Nup49	2	AP
49	<b>Nup1</b> -Pom152-Nup157-Nup192-Nup188-Nup170	3	AP
50	<b>Nsp1</b> -Nic96-Nup82-Nup57-Nup49-Nup159	1	AP
51	<b>Nup120</b> -Nup84-Nup145C-Seh1-Nup85-Sec13	1	AP
52	<b>Nup60</b> -Nup192-Pom152-Nup157-Nup188-Nsp1	3	AP
53	<b>Nup145C</b> -Nup133-Nup120-Nup85-Nup84-Seh1-Sec13	1	AP
54	<b>Nup133</b> -Nup120-Nup85-Nup84-Nup145C-Seh1-Sec13	2	AP
55	<b>Nup145N</b> -Nup192-Nup188-Nup170-Nup157-Pom152-Nup53	3	AP
56	<b>Nup82</b> -Nup159-Nup192-Nup188-Nup170-Pom152-Nsp1	1	AP
57	<b>Nup85</b> -Nup145C-Nup84-Seh1-Sec13-Nup133-Nup120	1	AP
58	<b>Nsp1</b> -Nup192-Nic96-Nup82-Nup57-Nup49-Nup159	2	AP
59	<b>Nup192</b> -Pom152-Nup116-Nic96-Nup84-Nup82-Nsp1-Nup57	3	AP
60	<b>Nup84</b> -Nup133-Nup145C-Nup85-Seh1-Sec13-Nup120-Nup145N	3	AP
61	<b>Nup82</b> -Nup159-Nup192-Nup188-Nup170-Pom152-Nsp1-Nup82	1	AP
62	<b>Pom152</b> -Nup170-Nup157-Nic96-Nup59-Nup57-Nup49-Ndc1	3	AP
63	<b>Nup85</b> -Nup145C-Nup84-Seh1-Sec13-Nup133-Nup120-Nup145N-Nup157	1	AP
64	<b>Pom152</b> -Nup192-Nup170-Nup157-Nup116-Nsp1-Nic96-Nup82-Pom152	3	AP
65	<b>Nup49</b> -Nup192-Nup170-Nup157-Nic96-Nsp1-Nup57-Nup133-Nup116	2	AP
66	<b>Nup157</b> -Nup133-Nup120-Nup84-Nup85-Nup145C-Nup145N-Seh1-Sec13	2	AP
67	<b>Nup116</b> -Nup192-Nup188-Nup170-Nup159-Pom152-Nsp1-Nic96-Nup82-Gle2	1	AP
68	<b>Nup133</b> -Nup116-Nup120-Nup85-Nup84-Nup82-Nup145C-Nup145N-Seh1-Sec13	2	AP
69	<b>Nup53</b> -Nup159-Nup192-Nup188-Nup170-Nup116-Nic96-Nup84-Nup82-Nsp1	3	AP
70	<b>Pom34</b> -Nup82-Nup84-Nup159-Nup170-Pom152-Nup192-Nup188-Nsp1-Nic96	2	AP
71	<b>Nup157</b> -Pom152-Nup133-Nup120-Nic96-Nup85-Nup84-Nup145C-Nup145N-Seh1-Sec13	3	AP
72	<b>Ndc1</b> -Nup159-Nup170-Pom152-Nup192-Nup188-Nsp1-Nic96-Nup82-Nup84-Nup53	2	AP
73	<b>Gle2</b> -Nup159-Nsp1-Nup116-Nup82-Nic96-Nup192-Nup188-Nup170-Pom152-Nup157	3	AP
74	<b>Nup53</b> -Nup159-Nup192-Nup188-Nup170-Nup157-Nup133-Nup116-Nic96-Nup84-Nup82-Nsp1	2	AP
75	<b>Nup1</b> -Nup192-Nup188-Pom152-Nup159-Nup170-Nup157-Nsp1-	3	AP

	Nup120-Nup82-Nup84-Nup145C		
76	<b>Nup59</b> -Nup159-Nup188-Nup192-Nup170-Pom152-Nup157-Nsp1-Nic96-Nup82-Nup53-Pom34-Gle2	2	AP
77	<b>Nup42</b> -Nup159-Nup170-Nup192-Nup188-Nup116-Nup100-Nsp1-Nic96-Nup84-Nup82-Gle1-Nup59-Nup53-Nup57	3	AP
78	<b>Nup100</b> -Nup159-Nup192-Nup188-Nup170-Pom152-Nup133-Nup116-Nsp1-Nup120-Nic96-Nup84-Nup82-Nup145C-Nup59	2	AP
79	<b>Seh1</b> -Nup159-Nup188-Nup192-Nup170-Pom152-Nup133-Nsp1-Nup120-Nic96-Nup85-Nup84-Nup145C-Nup82-Sec13	1	AP
80	<b>Nup84</b> -Nup192-Nup188-Nup170-Nup157-Pom152-Nup133-Nup120-Nsp1-Nic96-Nup145C-Nup85-Nup53-Nup57-Seh1-Sec13	1	AP
81	<b>Gle1</b> -Nup170-Pom152-Nup157-Nup133-Nup116-Nsp1-Nup120-Nic96-Nup82-Nup84-Nup145N-Nup57-Gle2-Pom34-Nup159-Nup188-Nup192-Nup42	1	AP
82	<b>Nup42</b> -Pom152-Nup157-Nup133-Nup116-Nup100-Nsp1-Nic96-Nup84-Nup82-Gle1-Nup59-Nup53-Nup57-Gle2-Pom34-Nup159-Nup170-Nup192-Nup188	1	AP



**Supplementary Table 9.** Table listing published yeast nup composites other than those described in ref.<sup>36</sup> or the current work. The list is comprehensive but likely not exhaustive. Notes: “Nup84 complex” contains Nup133, Nup120, Nup145C, Nup85, Nup84, Seh1, and Sec13.

<i>Nup</i>	<i>Composite</i>	<i>References</i>
<b>Nup192</b>	Nic96	40
<b>Nup188</b>	Nup100, Nup170, Nup188, Nup84 complex	41
<b>Nup170</b>	Nup53, Nup59, Nup157	42,43
<b>Nup159</b>	Nup116, Nup82, Nsp1, Nup159	44
	Nup82	45
	Nsp1	46
<b>Nup157</b>	Nup53, Nup59, Nup170	42
	Nup84 complex	47
<b>Nup133</b>	Nup84 complex	47-49
<b>Nup120</b>	Nup84 complex	47-49
<b>Nup116</b>	Nup82, Nsp1, Nup159	44
	Gle2	29
<b>Nup100</b>	Nup188, Nup170, Nup188, Nup84 complex	41
	Gle1	50
<b>Nic96</b>	Nup57, Nup49, Nsp1	12
	Nup192	40
<b>Nsp1</b>	Nup57, Nup49, Nsp1	12
	Nup116, Nup82, Nup159	44

<b>Nup85</b>	Nup84 complex	20,47,49
<b>Nup84</b>	Nup84 complex	20,47,49
<b>Nup82</b>	Nup116, Nup159, Nsp1, Nup159	44
	Nup159	45
<b>Nup145C</b>	Nup84 complex	20,47,49
<b>Ndc1</b>		
<b>Gle1</b>	Nup42, Nup100	50
<b>Nup145N</b>	Nup84 complex	47
<b>Nup59</b>	Nup53, Nup170, Nup157	42,43
<b>Nup57</b>	Nup49, Nic96, Nsp1	12
<b>Nup53</b>	Nup59, Nup170, + Nup157	42,43
<b>Nup49</b>	Nup57, Nic96, Nsp1	12
<b>Nup42</b>	Gle1	50
	Nup84 complex	41
<b>Gle2</b>	Nup116	29
<b>Seh1</b>	Nup84 complex	20,47,49
<b>Sec13</b>	Nup84 complex	20,47,49

The large scale studies<sup>51</sup> also reproduced many of the listed interactions, but they were not specifically listed.

**Supplementary Table 10.** Table listing published nup sublocalizations within the NPC other than those described in ref.<sup>36</sup> or the current work. In cases where no yeast nup immunolocalization exists, any available vertebrate homolog localizations have been listed (as “vertebrate”).

<b>Nup</b>	<b>ImmunoEM Localization</b>	<b>References</b>
<b>Nup192</b>	Nuclear	40
	Nuclear & Cytoplasmic, proximal (vertebrate)	52
<b>Nup188</b>	Nuclear & Cytoplasmic	53
<b>Nup170</b>	Nuclear & Cytoplasmic, proximal	43
<b>Nup159</b>	Cytoplasmic peripheral	5,54
<b>Nup157</b>	Nuclear & Cytoplasmic, proximal	43
<b>Pom152</b>	Nuclear Envelope Lumen	22
<b>Nup133</b>	Nuclear & Cytoplasmic (vertebrate)	55
<b>Nup116</b>	Cytoplasmic (mainly), Nuclear (minor), peripheral	56
<b>Nup1</b>	Nuclear peripheral	57
<b>Nic96</b>	Nuclear & Cytoplasmic, proximal (vertebrate)	52
	Nuclear & Cytoplasmic, proximal; basket	58
<b>Nsp1</b>	Nuclear & Cytoplasmic, proximal; basket	58
	Nuclear & Cytoplasmic	58
<b>Nup84</b>	Nuclear & Cytoplasmic, peripheral (vertebrate)	52,55
<b>Nup82</b>	Cytoplasmic peripheral	45
<b>Nup145C</b>	Nuclear & Cytoplasmic, peripheral (vertebrate)	52
<b>Ndc1</b>	Nuclear Envelope Membrane	59
<b>Gle1</b>	Cytoplasmic (mainly), Nuclear (minor), peripheral	54

<b>Nup145N</b>	Nuclear & Cytoplasmic, proximal (vertebrate)	52,60
<b>Nup60</b>	Nuclear peripheral (vertebrate)	52
<b>Nup59</b>	Nuclear & Cytoplasmic, proximal	43
<b>Nup57</b>	Nuclear & Cytoplasmic	61
<b>Nup53</b>	Nuclear & Cytoplasmic, proximal	43
<b>Nup49</b>	Nuclear & Cytoplasmic	61
<b>Nup42</b>	Cytoplasmic (mainly), Nuclear (minor), for overexpressed protein	62
<b>Pom34</b>	Nuclear Envelope Membrane	63

## 5 References

- 1 M. P. Rout, J. D. Aitchison, A. Suprpto et al., *J. Cell Biol.* **148** (4), 635 (2000).
- 2 S. E. Tcheperegine, M. Marelli, and R. W. Wozniak, *J Biol Chem* **274** (8), 5252 (1999).
- 3 C. Strambio-de-Castillia, G. Blobel, and M. P. Rout, *J Cell Biol* **144** (5), 839 (1999).
- 4 J. Kipper, C. Strambio-de-Castillia, A. Suprpto et al., *Methods Enzymol* **351**, 394 (2002).
- 5 D. M. Kraemer, C. Strambio-de-Castillia, G. Blobel et al., *J Biol Chem* **270** (32), 19017 (1995).
- 6 M. P. Rout and J. V. Kilmartin, *J. Cell Biol.* **111**, 1913 (1990).
- 7 C. W. Akey, *J Mol Biol* **248** (2), 273 (1995).
- 8 A. J. North, W. G. Bardsley, J. Hyam et al., *J Cell Sci* **112** ( Pt 23), 4325 (1999).
- 9 Q. Yang, M. P. Rout, and C. W. Akey, *Mol. Cell* **1** (2), 223 (1998).
- 10 C. M. Feldherr, D. Akin, and R. J. Cohen, *J Cell Sci* **114** (Pt 24), 4621 (2001).
- 11 D. A. Stirling, A. Petrie, D. J. Pulford et al., *Mol Microbiol* **6** (6), 703 (1992); J. D. Aitchison, M. P. Rout, M. Marelli et al., *J Cell Biol* **131** (5), 1133 (1995).
- 12 P. Grandi, V. Doye, and E. C. Hurt, *EMBO J.* **12** (8), 3061 (1993).
- 13 Tetenbaum-Novatt J. Strambio-de-Castillia C., Imai B.S., Chait B.T., Rout M.P., *Journal of Proteome Research* (2005).
- 14 M. P. Rout and G. Blobel, *J Cell Biol* **123** (4), 771 (1993).
- 15 J. D. Aitchison, G. Blobel, and M. P. Rout, *Science* **274** (5287), 624 (1996); M. P. Rout, G. Blobel, and J. D. Aitchison, *Cell* **89** (5), 715 (1997).
- 16 M. Niepel, C. Strambio-de-Castillia, J. Fasolo et al., *J Cell Biol* **170** (2), 225 (2005).
- 17 M. P. Rout and C. Strambio-de-Castillia, in *Cell Biology: A Laboratory Handbook*, edited by J. E. Celis (Academic Press, London, 1998), Vol. 2, pp. 143.
- 18 M.P. Rout and J.V. Kilmartin, in *Cell Biology: A Laboratory Handbook*, edited by J. E. Celis (Academic Press, London, 1998), Vol. 2, pp. 120.
- 19 W. Zhang and B. T. Chait, *Anal Chem* **72** (11), 2482 (2000); A. N. Krutchinsky, M. Kalkum, and B. T. Chait, *Anal Chem* **73** (21), 5066 (2001).
- 20 N. P. Allen, S. S. Patel, L. Huang et al., *Mol Cell Proteomics* **1** (12), 930 (2002).
- 21 V. Archambault, E. J. Chang, B. J. Drapkin et al., *Mol Cell* **14** (6), 699 (2004); A. J. Tackett, D. J. Dilworth, M. J. Davey et al., *J Cell Biol* **169** (1), 35 (2005); I. M. Cristea, R. Williams, B. T. Chait et al., *Mol Cell Proteomics* (2005).
- 22 C. Strambio-de-Castillia, G. Blobel, and M. P. Rout, *J. Cell. Biol.* **131**, 19 (1995).
- 23 S. E. Harding and H. Colfen, *Anal Biochem* **228** (1), 131 (1995).

- 24 D. Devos, S. Dokudovskaya, R. Williams et al., *Proc Natl Acad Sci U S A* **103** (7), 2172 (2006).
- 25 F. Alber, M. F. Kim, and A. Sali, *Structure* **13** (3), 435 (2005).
- 26 A. Krogh, B. Larsson, G. von Heijne et al., *J Mol Biol* **305** (3), 567 (2001).
- 27 Y. Harpaz, M. Gerstein, and C. Chothia, *Structure* **2** (7), 641 (1994).
- 28 L. A. Strawn, T. Shen, N. Shulga et al., *Nat Cell Biol* **6** (3), 197 (2004); K. Weis, *Curr Opin Cell Biol* **14** (3), 328 (2002).
- 29 S. M. Bailer, S. Siniosoglou, A. Podtelejnikov et al., *Embo J* **17** (4), 1107 (1998).
- 30 TH. Corman, CE. Leiserson, RL. Rivest et al., *Introduction to Algorithms*, Second ed. (Cambridge, Massachusetts, 2001).
- 31 F. P. Davis and A. Sali, *Bioinformatics* **21** (9), 1901 (2005).
- 32 M. Levitt, *J Mol Biol* **168** (3), 621 (1983).
- 33 W. Kabsch, *Act Crystallographic* **32 A**, 922 (1976).
- 34 G. D. Bader, D. Betel, and C. W. Hogue, *Nucleic Acids Res* **31** (1), 248 (2003); H. W. Mewes, D. Frishman, U. Guldener et al., *Nucleic Acids Res* **30** (1), 31 (2002).
- 35 C. von Mering, R. Krause, B. Snel et al., *Nature* **417** (6887), 399 (2002).
- 36 M. P. Rout, J. D. Aitchison, A. Suprapto et al., *J Cell Biol* **148** (4), 635 (2000).
- 37 F. Alber, S. Dokudovskaya, L. Veenhoff et al., (*Submitted*) (2007).
- 38 W. Braun and N. Go, *J Mol Biol* **186** (3), 611 (1985).
- 39 F. Alber, S. Dokudovskaya, L. Veenhoff et al., (*Submitted*) (2007).
- 40 B. Kosova, N. Pante, C. Rollenhagen et al., *J Biol Chem* **274** (32), 22646 (1999).
- 41 N. P. Allen, L. Huang, A. Burlingame et al., *J Biol Chem* **276** (31), 29268 (2001).
- 42 T. Iouk, O. Kerscher, R. J. Scott et al., *J Cell Biol* **159** (5), 807 (2002).
- 43 M. Marelli, J. D. Aitchison, and R. W. Wozniak, *J Cell Biol* **143** (7), 1813 (1998).
- 44 S. M. Bailer, C. Balduf, J. Katahira et al., *J Biol Chem* (2000).
- 45 M. E. Hurwitz, C. Strambio-de-Castillia, and G. Blobel, *Proc Natl Acad Sci U S A* **95** (19), 11241 (1998).
- 46 N. Belgareh, C. Snay-Hodge, F. Pasteau et al., *Mol Biol Cell* **9** (12), 3475 (1998).
- 47 M. Lutzmann, R. Kunze, K. Stangl et al., *J Biol Chem* **280** (18), 18442 (2005).
- 48 M. Lutzmann, R. Kunze, A. Buerer et al., *Embo J* **21** (3), 387 (2002).
- 49 S. Siniosoglou, M. Lutzmann, H. Santos-Rosa et al., *J Cell Biol* **149** (1), 41 (2000).
- 50 R. Murphy and S. R. Wentz, *Nature (London)* **383** (6598), 357 (1996).
- 51 A. C. Gavin, P. Aloy, P. Grandi et al., *Nature* **440** (7084), 631 (2006); A. C. Gavin, M. Bosche, R. Krause et al., *Nature* **415** (6868), 141 (2002); Y. Ho, A. Gruhler, A. Heilbut et al., *Nature* **415** (6868), 180 (2002); N. J. Krogan, G. Cagney, H. Yu et al., *Nature* **440** (7084), 637 (2006).
- 52 S. Krull, J. Thyberg, B. Bjorkroth et al., *Mol Biol Cell* **15** (9), 4261 (2004).
- 53 U. Nehrbass, M. P. Rout, S. Maguire et al., *J Cell Biol* **133** (6), 1153 (1996).

- 54 A. L. Miller, M. Suntharalingam, S. L. Johnson et al., *J Biol Chem* **279** (49), 51022 (2004).
- 55 N. Belgareh, G. Rabut, S. W. Bai et al., *J Cell Biol* **154** (6), 1147 (2001).
- 56 A. K. Ho, T. X. Shen, K. J. Ryan et al., *Mol Cell Biol* **20** (15), 5736 (2000).
- 57 J. Solsbacher, P. Maurer, F. Vogel et al., *Mol Cell Biol* **20** (22), 8468 (2000).
- 58 B. Fahrenkrog, J. P. Aris, E. C. Hurt et al., *J Struct Biol* **129** (2-3), 295 (2000).
- 59 H. J. Chial, M. P. Rout, T. H. Giddings et al., *J Cell Biol* **143** (7), 1789 (1998).
- 60 E. R. Griffis, S. Xu, and M. A. Powers, *Mol Biol Cell* **14** (2), 600 (2003).
- 61 B. Fahrenkrog, E. C. Hurt, U. Aebi et al., *J Cell Biol* **143** (3), 577 (1998).
- 62 Y. Strahm, B. Fahrenkrog, D. Zenklusen et al., *Embo J* **18** (20), 5761 (1999).
- 63 M. Miao, K. J. Ryan, and S. R. Wente, *Genetics* **172** (3), 1441 (2006).