Title
Inferring structures, free energy differences, and kinetic rates of biological macromolecular assemblies by integrative modeling

Author
Chemmama, Ilan E.

Publication Date
2020

Peer reviewed|Thesis/dissertation

Inferring structures, free energy differences, and kinetic rates of biological macromolecular assemblies by integrative modeling

by
Ilan Edmond Chemmama

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Andrej Sali
8F7A6AB94F2C4F4...

Andrej Sali
Chair

DocuSigned by:
David A. Agard
[redacted]

David A. Agard

DocuSigned by:
Yifan Cheng
D69B6947B6AC442...

Yifan Cheng

Committee Members

Acknowledgements

During my time at UCSF, I have had the opportunity to benefit from the knowledge and advice of many mentors. I first want to thank Andrej Sali, my thesis advisor, for his mentorship and support over the last few years. Andrej has been a great mentor, always helpful, supportive, and encouraging of my work. Andrej's approach to science has been a standard for me to try to match: his vision, outlook, and rigor to doing science right.

I want to thank the members of the Sali lab. They provided such a unique, supportive, and scientifically exciting environment that made learning and doing science over the past few years that much more enjoyable, and I thank all of them for that opportunity. I want to especially thank my mentors I had the chance to work with closely, Dina Schneidman, Seung Joong ("SJ") Kim, Shruthi Viswanath, Charles Greenberg, and Ignacia Echeverria. I also want to thank Seth Axen, Sara Calhoun, and Sai Ganesan for the countless scientific and non-scientific discussions we have had over the many years. None of my work would have been done without the smooth running of our cluster and the maintenance and improvements to our lab software, IMP; thus, I want to thank Ben Webb and the IMP developers for building the tools and for always providing support when I needed it. I would also like to thank all the brilliant people I overlapped with in the Sali lab, in random order, Daniel, Peter, Barak, Rakesh, Thomas, Aji, Jeremy, Leah, Nikita, Adrian, Kate, Tanmoy, and Kala.

I want to acknowledge the tremendous support and mentorship of faculty I had the opportunity to benefit from over the years. Bernard Gerstman and Prem Chapagain at FIU, my undergraduate research advisors, gave me the opportunity to learn and start my journey to become a scientist. They gave me opportunities to work independently on

projects, and their mentorship, passion, and trust made me want to pursue a career in science. Yifan Cheng and Matt Jacobson, my rotation advisors, allowed me to grow and expand my interests. David Agard and Yifan Cheng, my committee members, challenged me by asking the right questions and were always happy to provide guidance when needed.

I want to thank my many collaborators. I want to acknowledge all the brilliant people that were involved with NPC projects and the ABC Transporter P01 group at UCSF. The level of science rigor, expertise, and talent truly made this experience unique. I want to thank Evan Green, with whom I work on the determination of kinetics and thermodynamics parameters from negative stain EM. I want to acknowledge Lan Huang and her group at UC Irvine, who have been amazing collaborators. I would like to acknowledge Craig Gutierrez, a graduate student in Lan's lab, for the work he did on the COP9 signalosome project. I want to particularly thank Lan Huang for her trust, mentorship, and support when working together. I finally want to acknowledge my funding sources: the NSF Graduate Student Fellowship, NIH training grant, and the various NIH grants to Andrej.

I want to acknowledge the many friends and classmates that made this journey as enjoyable as it was. I want to thank Seth, Sara, JP, Reed, Kevin, and Diego for their friendship. I want to acknowledge Ben and Evan for their friendship and their constant readiness to escape the lab to play squash. Leo and Brad for the fun, the poker games, and their friendship. I also want to thank Kevin, Vanessa, Cindy, Jessica, and Analucia for their friendship, for their support, for constantly reminding me that there is a world outside of science, and for everything we shared along the way.

Last but not least, I want to thank my family. Their love, support, and encouragements made the pursuit of this dissertation even more enjoyable. I want to thank my amazing parents for everything they did, and no word can even begin describing how grateful I am to them. There is not enough space for me to express how grateful and lucky I am to have such parents. I cannot even start enumerating how many sacrifices they made so that my siblings and I can have the best life we can have. I want to express how grateful I am to my grandmothers, who played such critical roles when I was growing up; they will always remain in my thoughts and heart. I want to thank my sister Sarah, my brother Jeremy, my sister-in-law Leah, and the new addition to our family, my niece Ella; I am so lucky and grateful for them. I cannot even start expressing how grateful I am to my family and how much I love them.

Contributions

Several chapters in this dissertation contain previously published materials. Supplementary information, data, and figures have not been included but can be found online in the corresponding publications. A section describing author contributions is included in each chapter.

Chapter I

Ilan E. Chemmama.

Introduction.

Chapter II

Gutierrez, Craig*; Chemmama, Ilan E*; Mao, Haibin; Yu, Clinton; Echeverria, Ignacia; Block, Sarah A; Rychnovsky, Scott D; Zheng, Ning; Sali, Andrej; Huang, Lan. Structural dynamics of the human COP9 signalosome revealed by cross-linking mass spectrometry and integrative modeling. **Proceedings of the National Academy of Sciences** Feb 2020, 117 (8) 4088-4098; DOI: 10.1073/pnas.1915542117 *Contributed equally to this work.

Chapter III

Chemmama, Ilan E*; Green, Evan M*; Agard, David A; Cheng, Yifan; Sali, Andrej. Thermodynamic and kinetic estimates from electron microscopy particle images. *Contributed equally to this work.

Chapter IV

Viswanath, Shruthi*; Chemmama, Ilan E*; Cimermancic, Peter; Sali, Andrej. Assessing exhaustiveness of stochastic sampling for integrative modeling of macromolecular structures. **Biophysical journal** 113 11 2344, 2353 2017 *Contributed equally to this work.

Chapter V

Wang, Xiaorong; Chemmama, Ilan E; Yu, Clinton; Huszagh, Alexander; Xu, Yue; Viner, Rosa; Block, Sarah Ashley; Cimermancic, Peter; Rychnovsky, Scott D; Ye, Yihong; *et al*. The proteasome, interacting Ecm29 protein disassembles the 26S proteasome in response to oxidative stress. **Journal of Biological Chemistry**. jbc. M117. 803619 2017

Chapter VI

Fernandez, Martinez, Javier; Kim, Seung Joong; Shi, Yi; Upla, Paula; Pellarin, Riccardo; Gagnon, Michael; Chemmama, Ilan E; Wang, Junjie; Nudelman, Ilona; Zhang, Wenzhu; *et al*. Structure and function of the nuclear pore complex cytoplasmic mRNA export platform **Cell** 167 5 1215, 1228. e25 2016

Chapter VII

Upla, Paula; Kim, Seung Joong; Sampathkumar, Parthasarathy; Dutta, Kaushik; Cahill, Sean M; Chemmama, Ilan E; Williams, Rosemary; Bonanno, Jeffrey B; Rice, William J; Stokes, David L; *et al*. Molecular architecture of the major membrane ring component of the nuclear pore complex **Structure** 25 3 434, 445 2017

Chapter VIII

Kim, Seung Joong; Fernandez, Martinez, Javier; Nudelman, Ilona; Shi, Yi; Zhang, Wenzhu; Raveh, Barak; Herricks, Thurston; Slaughter, Brian D; Hogan, Joanna A; Upla,

Paula; Chemmama, Ilan E; *et al*. Integrative structure and functional anatomy of a nuclear pore complex. **Nature** 555 7697475 2-18

Chapter IX

Chen, Qi; Vieth, Michal; Timm, David E; Humblet, Christine; Schneidman-Duhovny, Dina; Chemmama, Ilan E; Sali, Andrej; Zeng, Wei; Lu, Jirong; Liu, Ling. Reconstruction of 3D structures of MET antibodies from electron microscopy 2D class averages. **PloS one** 12 4 e0175758 2017

Inferring structures, free energy differences, and kinetic rates of biological macromolecular assemblies by integrative modeling

Ilan Edmond Chemmama

Abstract

Biological macromolecular assemblies play crucial roles in most cellular processes. The determination of their structures, thermodynamics, and kinetics is essential to understand their function, evolution, modulation, and design. Determining such models, however, remains challenging. One particularly powerful approach to constructing models in general is integrative modeling. Integrative modeling aims to maximize the accuracy, precision, and completeness of models, by simultaneously utilizing all available information, including experimental data, physical principles, statistical analyses, and other prior models. The goal of this thesis is to expand the scope of integrative modeling to the inference of spatial, thermodynamic, and kinetic aspects of macromolecular assemblies.

In **Chapter I**, I introduce the integrative modeling framework for spatiotemporal modeling of biological macromolecular assemblies. In **Chapter II**, I demonstrate how the synergy between multi-chemistry cross-linking mass spectrometry and integrative modeling can map the structural dynamics of macromolecular assemblies, by application to the human Cop9 signalosome complex. In **Chapter III**, I present a method for determining structures, free energy differences, and kinetic rates of macromolecular assemblies along their functional cycle, mainly from negative stain electron microscopy (EM). We apply the method to the yeast Hsp90 to estimate the free energy differences

and kinetic parameters along its nucleotide hydrolysis cycle, which includes open and closed states of Hsp90. In **Chapter IV**, I describe a validation of stochastic sampling in integrative modeling. The remaining chapters describe applications of integrative modeling to assemblies of various sizes and scales, using various sources of information, thus illustrating the flexibility of the integrative modeling approach. Specifically, I apply integrative modeling to the human ECM29-Proteasome assembly under oxidative stress (**Chapter V**), the yeast nuclear pore complex (NPC) cytoplasmic mRNA export platform (**Chapter VI**), the major membrane ring component of the yeast NPC (**Chapter VII**), the entire yeast NPC (**Chapter VIII**), and the reconstruction of 3D structures of MET antibodies (**Chapter IX**).

Table of Contents

List of Figures

Chapter II

Chapter III

Chapter IV

Chapter V

Chapter VI

Chapter VII

Chapter VIII

Chapter IX

List of Tables

Chapter IV

Chapter VI

Chapter VII

Chapter IX

Chapter I – Introduction

Introduction

Biological macromolecular assemblies play crucial roles in most cellular processes. These assemblies vary in size and composition, consisting of proteins and sometimes nucleic acids, other macromolecules, and small molecules. The determination of their structures, thermodynamics, and kinetics is thus essential to understand their function, evolution, modulation, and design (1). However, inferring structural, thermodynamic, and kinetics aspects of these assemblies remains challenging. One particularly powerful approach to constructing models in general is integrative modeling. Integrative modeling aims to maximize the accuracy, precision, and completeness of models, by simultaneously utilizing all available information, including experimental data, physical principles, statistical analyses, and other prior models. The goal of this thesis is to expand the scope of integrative modeling to the inference of spatial, thermodynamic, and kinetic aspects of macromolecular assemblies.

Integrative modeling as an optimization problem

Integrative structure modeling in principle relies on all available information about the modeled system (1, 2, 3). Thus, it can maximize the accuracy, precision, completeness, and efficiency of modeling (1, 2). Integrative modeling framework is best formulated as an optimization problem, thus requiring model representation, a scoring function, and a sampling scheme.

Representation

Representation of a model specifies the variables whose values are modeled based on input information. The choice of representation must facilitate (1) answering biological questions of interest, (2) constructing an accurate and efficiently computed scoring function to quantify the consistency of a model with the input information, and (3) efficient sampling of alternative models. Examples of variables in integrative structure determination are atomic positions, the number of different states, and the population proportion of each state.

Scoring

The scoring function quantifies the degree of a match between a tested model and the input information. The most objective scoring function is a Bayesian posterior model probability density, based on information from one or more different experiments, physical theories, statistical analyses, and/or prior models. Bayes' law provides a formula for the posterior probability density of a model *M* given the data *D* and prior information *I*,

$$p(M|D,I) \propto p(D|M,I) \cdot p(M|I).$$

The term $p(D|M,I)$, called the likelihood function, is the probability of observing *D* given *M* and *I*. To define a likelihood for integrative modeling, we define a forward model and a noise model based on our belief about the family of processes that generated *D* given *M* and *I*. The forward model specifies a mapping of the model parameters to simulated measurements given the knowledge of the process of data generation in the absence of experimental noise. Thus, the forward model returns a simulated noiseless measurement given a realization of model *M*. The noise model specifies the distribution of the deviation

between the experimental measurement and the simulated measurement computed from the forward model.

The term $p(M|I)$ is the density of the prior probability distribution, that is the distribution of model *M* given *I*. The prior probability distribution expresses beliefs about model variables before any data is taken into consideration. Priors in integrative modeling are derived from physical principles, prior statistical analyses, and/or other models. Examples of such priors in integrative modeling are excluded volume restraints, sequence connectivity restraints, a molecular mechanics force field, statistical potentials, and previously determined structures (*e.g.*, by using X-ray crystallography, electron microscopy, NMR spectroscopy, and comparative modeling).

Finally, a Bayesian scoring function in integrative modeling is defined as the negative logarithm of the posterior probability density

$$S(M) = -\log\left[p(M|D,I)\right].$$

Advantages of using the Bayesian formulation are as follows. First, the model *M* is a density of models described by the posterior probability density, not a single representative model, with possibly overfitted variables. Indeed, the distribution of models reflects the uncertainty in the experimental data as well as in the prior information. Second, the distribution of models derived by Bayesian modeling is in principle more accurate than the models derived by using traditional least-squares scoring functions, because model *M* (i.e., the model ensemble) was determined by objectively mixing different sources of information with their associated uncertainties (3). Third, we can objectively estimate the uncertainty associated with model *M*. Finally, multiple choices

about model representation and scoring functions can be quantified and compared *via* model selection criteria (4, 5).

Sampling

The sampling methods must efficiently and accurately sample the posterior probability density. Variables are often sampled stochastically (6, 7). For example, many Monte Carlo schemes have been developed to efficiently and accurately sample the posterior probability density, including simulated annealing, replica exchange, Gibbs sampling, and Hamiltonian Monte Carlo methods (8).

The integrative modeling framework

Integrative modeling iterates through four stages to transform input information into a model (1, 2, 9–14): (1) gathering all available experimental data and prior information; (2) translating information into representations of assembly components and a scoring function for ranking alternative assembly structures; (3) sampling structural models; and (4) validating the model.

Stage 1: Gather information

In the first stage of modeling, we gather all information available about a system needed to solve the spatial and temporal arrangements of assemblies at the highest precision (i.e., smallest uncertainty). First, information can be experimental data, such as observations by mass spectrometry (e.g., stoichiometry (13), list of cross-links (13–17)), spectroscopy (e.g., Förster resonance energy transfer [FRET] (18) and nuclear magnetic resonance [NMR] (19)), and microscopy (e.g., electron microscopy images (13, 16, 20, 21) and electron microscopy density maps (13, 22, 23)). Second, information can be

physical principles and statistical analyses, such as a molecular mechanics force field, statistical potentials, excluded volume, and connectivity between contiguous beads. Finally, information can be prior structural models determined by integrative modeling, X-ray crystallography, NMR spectroscopy, electron microscopy, and comparative protein structure modeling.

Stage 2: Translate information into representations of assembly components and a scoring function

Information gathered in the first stage can then be used to inform the priors and likelihoods, effectively resulting in defining the representation of the system, the scoring function, and the sampling space. Information gathered in the first stage can also be used for filtering and validation whether or not it is used for representation, scoring, and sampling.

First, information can be used to define the representation by specifying the variables whose values will be determined by modeling. These variables determine the components of the system (identity, copy numbers, granularity, and number of states), the coordinates of these components (position, orientation, and conformations). They also include auxiliary variables that are fit to input information. For example, previously obtained component structures can be used to specify rigid bodies, thus reducing the number of sampled variables from three per atom to six per rigid body.

Second, information can be used to construct and compute the scoring function. As described before, the most objective scoring function to assess the match between a tested model and the input information is the Bayesian posterior model probability density. Information can specify the likelihood function (i.e., the forward model and the noise

model), and the prior probability density. For example, experiments (e.g., chemical cross-linking), physical principles and statistical analyses (e.g., excluded volume, connectivity restraints) can be used to inform part of the model's spatial arrangements (e.g., positions, orientations, and conformations) and their associated uncertainty. Consequently, these prior probability densities can constrain and limits the model search space. For example, physical principles dictate that a kinetic rate is a real non-negative scalar, and thus we can constrain the sampling of a kinetic rate value to positive real numbers.

Third, information can be used to filter the sample models after they are produced by the sampling scheme. Such filtering is especially useful for information that is computationally expensive to evaluate many times as part of the scoring function used for sampling. For example, experimental data (e.g., a set of electron microscopy 2D images) can be used only to filter a small subset of models that already satisfy other information.

Finally, any subset of information can be set aside to be used only for validation of the good-scoring models, without changing or filtering them.

Stage 3: Sample structural models

The purpose of the third stage is to find a sample of all models that are sufficiently consistent with input information, as quantified by the scoring function; the goal of sampling methods is to accurately and efficiently sample the posterior model density. This search is often achieved by a stochastic sampling of alternative structures, avoiding the biases and limitations intrinsic to searches by humans. The sampling must be done at a precision that is higher than needed for interpreting the models and higher than the precision of the scoring function landscape.

Stage 4: Validate the model

Validation is essential for avoiding overinterpretation of any model of any type. For integrative structures, we currently perform validation in these five steps: (1) select the models for validation; (2) estimate sampling precision; (3) estimate model precision; (4) quantify the degree to which a model satisfies the information used to compute it; and (5) quantify the degree to which a model satisfies relevant information not used to compute it.

First, we need to select all models that represent the posterior model probability density (i.e., the ensemble). Due to the nature of stochastic sampling, the model is often sampled independently multiple times, each time starting from a different initial random configuration. Models generated in the early equilibration burn-in phase of each sampling are discarded, resulting in an equilibrated sample that maximizes the number of uncorrelated samples (24).

Second, the sampling precision can be estimated for stochastic sampling methods by splitting the ensemble of models into two independent sets, followed by quantifying the difference between the two sets. It is important to properly estimate the sampling precision (uncertainty) because only the features of the model larger than the sampling precision are well estimated (3, 25). When using stochastic sampling methods, sampling precision can be increased simply by increasing the number of independently computed models. High sampling precision is necessary but not sufficient for exhaustive sampling (25, 26).

In the third step, model uncertainty (precision) is estimated. The most explicit description of model uncertainty is provided by the set of all models that are sufficiently

consistent with the input information (i.e., the ensemble). Model precision can be quantified by the variability among the models in the ensemble; in the end, the ensemble can be described by one or more representative models and their uncertainties. The uncertainty is often not distributed evenly across the ensemble; thus, computing local estimates of precision can help improving the modeling and the inference of the biology. For example, model representation can be optimized such that sampling is exhaustive at a precision commensurate with the precision of the representation (4).

It is often convenient for model interpretation if the ensemble structures are first clustered on the basis of their structural similarity. As a result, only a structure representative of each major cluster can potentially be used for interpretation. Many clustering methods exist. They vary in terms of the criterion used to quantify a similarity between two structures, such as the distance root-mean-square deviation between structure coordinates that avoids the need for structure superposition (27). They also vary in the method for converting pairwise similarities into clusters (28). Clustering generally depends on an arbitrary threshold parameter that determines how many clusters are produced. Minor clusters containing few structures might be ignored, especially if the scoring function approximates a Bayesian posterior model density where minor clusters represent unlikely solutions. However, discarding these minor clusters may result in an underestimation of uncertainty and overestimation of the confidence in the models. The clustering threshold is selected such that the following three criteria are satisfied: first, the number of major clusters is minimized for parsimony; second, the precision of these clusters is worse than the sampling precision; and finally, the cluster precision is high enough for interpreting the models.

Model uncertainty results from insufficient input information and sample heterogeneity (3). It is difficult to deconvolve the impact of these two sources of model uncertainty. In general, only the total model uncertainty is reported. In addition, model uncertainty can also result from uncertain model representation, uncertainty in the scoring function, and insufficient sampling. This uncertainty in particular is often not considered but can be large. For example, a mistake in representation is not recoverable in the current modeling schemes, because they assume that the representation is correct and so do not even attempt to generate the correct representation (e.g., when a protein subunit structure is incorrect, incorrectly assumed to be rigid, or incorrect stoichiometry is enforced during modeling a structure of a complex).

Fourth, the model is assessed by quantifying the degree to which it satisfies the information used to compute it. This data satisfaction is considered on the scale of the uncertainty of the corresponding information. When data used to construct the model are insufficiently satisfied, the model is not validated. Such violations can occur when the data are more uncertain than assumed, the representation is incorrect, and/or the sampling is not sufficient.

Finally, a model is tested against information that was not used to compute it. For example, one can perform a jackknifing test consisting of repetitively omitting a random subset of chemical cross-links, recomputing the model, and comparing these models against the omitted cross-links, to validate both the model and the cross-links, similarly to R_{free} in X-ray crystallography (29).

Thesis summary

In **Chapter II**, I demonstrate how the synergy between multi-chemistry cross-linking mass spectrometry and integrative modeling can map the structural dynamics of macromolecular assemblies, by application to the human Cop9 signalosome complex. In **Chapter III**, I present a method for determining structures, free energy differences, and kinetic rates of macromolecular assemblies along their functional cycle, mainly from negative stain electron microscopy (EM). We apply the method to the yeast Hsp90 to estimate the free energy differences and kinetic parameters along its nucleotide hydrolysis cycle, which includes open and closed states of Hsp90. In **Chapter IV**, I describe a validation of stochastic sampling in integrative modeling. The remaining chapters describe applications of integrative modeling to assemblies of various sizes and scales, using various sources of information, thus illustrating the flexibility of the integrative modeling approach. Specifically, I apply integrative modeling to the human ECM29-Proteasome assembly under oxidative stress (**Chapter V**), the yeast nuclear pore complex (NPC) cytoplasmic mRNA export platform (**Chapter VI**), the major membrane ring component of the yeast NPC (**Chapter VII**), the entire yeast NPC (**Chapter VIII**), and the reconstruction of 3D structures of MET antibodies (**Chapter IX**).

References

1. M. P. Rout, A. Sali, Principles for Integrative Structural Biology Studies. *Cell* **177**, 1384–1403 (2019).
2. F. Alber, *et al.*, Determining the architectures of macromolecular assemblies. *Nature* **450**, 683–694 (2007).
3. D. Schneidman-Duhovny, R. Pellarin, A. Sali, Uncertainty in integrative structural modeling. *Current Opinion in Structural Biology* **28**, 96–104 (2014).
4. S. Viswanath, A. Sali, Optimizing model representation for integrative structure determination of macromolecular assemblies. *PNAS* **116**, 540–545 (2019).
5. K. H. Knuth, M. Habeck, N. K. Malakar, A. M. Mubeen, B. Placek, Bayesian evidence and model selection. *Digital Signal Processing* **47**, 50–67 (2015).
6. M. P. Allen, D. J. Tildesley, *Computer Simulation of Liquids* (Clarendon Press, 1989).
7. N. Metropolis, S. Ulam, The monte carlo method. *Journal of the American statistical association* **44**, 335–341 (1949).
8. M. Betancourt, A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434 [stat]* (2018) (February 27, 2020).
9. F. Alber, B. T. Chait, M. P. Rout, A. Sali, “Integrative Structure Determination of Protein Assemblies by Satisfaction of Spatial Restraints” in *Protein-Protein Interactions and Networks: Identification, Characterization and Prediction.*, A. Panchenko, T. Przytycka, Eds. (Springer-Verlag, 2008), pp. 99–114.
10. F. Alber, *et al.*, The molecular architecture of the nuclear pore complex. *Nature* **450**, 695–701 (2007).

11. A. Sali, *et al.*, Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure* **23**, 1156–67 (2015).
12. D. Russel, *et al.*, Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. *PLOS Biology* **10**, e1001244 (2012).
13. S. J. Kim, *et al.*, Integrative structure and functional anatomy of a nuclear pore complex. *Nature* **555**, 475–482 (2018).
14. C. Gutierrez, *et al.*, Structural dynamics of the human COP9 signalosome revealed by cross-linking mass spectrometry and integrative modeling. *PNAS* **117**, 4088–4098 (2020).
15. Y. Shi, *et al.*, Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. *Mol Cell Proteomics* **13**, 2927–2943 (2014).
16. J. Fernandez-Martinez, *et al.*, Structure and Function of the Nuclear Pore Complex Cytoplasmic mRNA Export Platform. *Cell* **167**, 1215-1228 e25 (2016).
17. X. Wang, *et al.*, The proteasome-interacting Ecm29 protein disassembles the 26S proteasome in response to oxidative stress. *J Biol Chem* **292**, 16310–16320 (2017).
18. M. Bonomi, *et al.*, Determining protein complex structures based on a Bayesian model of in vivo Förster resonance energy transfer (FRET) data. *Mol. Cell Proteomics* **13**, 2812–2823 (2014).
19. W. Rieping, M. Habeck, M. Nilges, Inferential structure determination. *Science* **309**, 303–6 (2005).

20. P. Upla, *et al.*, Molecular Architecture of the Major Membrane Ring Component of the Nuclear Pore Complex. *Structure* **25**, 434–445 (2017).
21. Q. Chen, *et al.*, Reconstruction of 3D structures of MET antibodies from electron microscopy 2D class averages. *PLoS One* **12** (2017).
22. P. Robinson, *et al.*, Molecular architecture of the yeast Mediator complex. *eLife* **4**, e08719 (2015).
23. S. Hanot, *et al.*, Multi-scale Bayesian modeling of cryo-electron microscopy density maps. *https://doi.org/10.1101/113951* **Preprint** (2017).
24. J. D. Chodera, A Simple Method for Automated Equilibration Detection in Molecular Simulations. *J. Chem. Theory Comput.* **12**, 1799–1805 (2016).
25. S. Viswanath, I. E. Chemmama, P. Cimermancic, A. Sali, Assessing Exhaustiveness of Stochastic Sampling for Integrative Modeling of Macromolecular Structures. *Biophysical Journal* **113**, 2344–2353 (2017).
26. A. Gelman, D. B. Rubin, Inference from Iterative Simulation Using Multiple Sequences. *Statist. Sci.* **7**, 457–472 (1992).
27. P. Koehl, Protein structure similarities. *Current Opinion in Structural Biology* **11**, 348–353 (2001).
28. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Science & Business Media, 2009).
29. A. T. Brunger, Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–5 (1992).

Chapter II - Structural dynamics of the human COP9 signalosome revealed by cross-linking mass spectrometry and integrative modeling

Contributing authors

Craig Gutierrez[1,*], Ilan E. Chemmama[2,*], Habin Mao[3], Clinton Yu[1], Ignacia Echeverria[2], Sarah A. Block[4], Scott D. Rychnovsky[4], Ning Zheng[3,5], Andrej Sali[2,6], and Lan Huang[1,7]

[1]Department of Physiology and Biophysics, University of California, Irvine, CA 92697, USA

[2]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA 94158, USA

[3]Department of Pharmacology, University of Washington, Seattle, WA 98195, USA

[4]Department of Chemistry, University of California, Irvine, CA 94697, USA

[5]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA

[6]Department of Pharmaceutical Chemistry, Institute of Quantitative Biosciences, University of California, San Francisco, San Francisco, CA 94158, USA

[7]Contact: lanhuang@uci.edu

[*]Craig Gutierrez and Ilan E. Chemmama contributed equally to this work.

Abstract

The COP9 signalosome (CSN) is an evolutionarily conserved eight subunit (CSN1–8) protein complex that controls protein ubiquitination by deneddylating Cullin-RING E3 ligases (CRLs). The activation and function of CSN hinges on its structural dynamics, which has been challenging to decipher by conventional tools. Here, we have developed

a multichemistry cross-linking mass spectrometry approach enabled by three mass spectrometry-cleavable crosslinkers to generate highly reliable cross-link data. We applied this approach with integrative structure modeling to determine the interaction and structural dynamics of CSN with the recently discovered ninth subunit, CSN9, in solution. Our results determined the localization of CSN9 binding sites and revealed CSN9-dependent structural changes of CSN. Together with biochemical analysis, we propose a structural model in which CSN9 binding triggers CSN to adopt a configuration that facilitates CSN–CRL interactions, thereby augmenting CSN deneddylase activity. Our integrative structure analysis workflow can be generalized to define in-solution architectures of dynamic protein complexes that remain inaccessible to other approaches.

Significance

Structural plasticity is a critical property of many protein complexes that has been challenging to study using conventional structural biology tools. Cross-linking mass spectrometry (XL-MS) has become an emergent technology for elucidating architectures of large protein complexes. While effective, current XL-MS methods mostly rely on lysine reactive cross-linking chemistry and have limited capacity in fully defining dynamic structures of protein complexes. Here, we have developed an integrated structural approach based on three MS-cleavable cross-linkers with distinct chemistries. This approach enabled us to obtain highly reliable and comprehensive cross-link data that significantly facilitate integrative structural modeling of dynamic protein complexes. In addition, it has been successfully applied to the COP9 signalosome to determine its structural dynamics associated with its function.

Introduction

The COP9 signalosome (CSN) is an evolutionarily conserved and essential multisubunit protein complex involved in diverse cellular and developmental processes in animals and plants (1–3). The CSN functions as a deneddylase, specific for cleaving Nedd8 modification from cullin proteins, the key components of Cullin–RING ubiquitin E3 ligases (CRLs) (4–8). CRLs represent the largest evolutionarily conserved superfamily of multisubunit E3s (5, 6), which embody ~30% of all human E3 proteins and coordinate degradation of ~20% of the proteins processed by the proteasome. The dynamic cycle of neddylation and deneddylation of cullins is a critical step in regulating the assembly and activity of CRLs (6, 9, 10). In addition to enzymatic regulation of CRLs, the CSN can inactivate CRLs noncatalytically by direct binding, preventing their association with E2 enzymes and ubiquitination substrates (11–14). While abnormal CRL activity is frequently associated with various human diseases, multiple studies have also identified the CSN as a positive regulator of oncogenes and negative regulator of tumor suppressors (15–19). Moreover, elevated expression of CSN subunits has been found in a number of human tumors, often with poor prognosis (20, 21). Therefore, better understanding of the CSN structure would provide new insights on their function and the regulation of CRLs associated with human pathology.

The canonical CSN complex (hereafter referred to as CSN) typically consists of eight subunits (CSN1–8) (1, 3), including six different PCI (proteasome lid-CSN-initiation factor 3) domain-containing subunits (CSN1 to CSN4, CSN7, and CSN8) and two MPN (MPR1/PAD1 amino-terminal) domain-containing proteins (CSN5 and CSN6). Among them, CSN5 is the catalytic subunit directly responsible for CSN deneddylase activity (4).

The CSN complex shares sequence similarities to the 19S proteasome lid subcomplex and the eukaryotic translation initiation complex eIF3, which also contain PCI and MPN domains (1, 3). The X-ray structure of recombinant human CSN has revealed that CSN5 and CSN6 MPN domains form a heterodimer, while the six remaining PCI subunits assemble into a horseshoe-shaped ring from which their arm-like α-helical domains project (22). The PCI subunits provide a scaffold, primarily through CSN2 and CSN4, which facilitates the recruitment of neddylated CRLs. Meanwhile, the two MPN subunits are slightly juxtaposed, exposing the active MPN catalytic core in CSN5 (12, 23–25). All eight subunits are united in a helical bundle formed by their C-terminal carboxyl α-helices, which are stacked between the CSN5–CSN6 dimer and PCI ring. Interestingly, substrate-free CSN exists in an inactive, autoinhibited state (23). Structural and biochemical characterization of CSN–CRL complexes have revealed substrate-induced structural dynamics associated with CSN activation (12, 23–26). Binding of neddylated CRLs to CSN triggers substantial remodeling and extensive conformational changes of the complex, activating the isopeptidase activity of CSN5. Although the structural plasticity of the CSN is important for CSN activation and function in regulating CRL activities in cells, it has not been well characterized due to limitations in existing technologies.

Recently, the ninth CSN subunit, CSN9 (also known as CSNAP [CSN acidic protein]), has been discovered to complex with CSN1–8 stoichiometrically to form a nine-membered noncanonical CSN complex (also known as CSN9-bound CSN, hereafter referred to as CSNn) (27). As canonical CSN subunits (CSN1–8) have a one-to-one correspondence to the subunits of the 19S proteasome lid subcomplex (3, 28), CSN9 is homologous to DSS1, the smallest component of the 19S lid. While CSN9 is not essential

for the assembly and catalytic activity of CSN (27), a recent study has suggested that CSN9 reduces the affinity of CSN–CRL interactions, contributing to steric regulation of CRLs (14). The depletion of CSN9 appears to have a global impact on CRL-associated activities, leading to altered reproductive capacity, suppressed DNA damage response, decreased viability, and delayed cell cycle progression (14). It has also been suggested that the C terminus of CSN9 is important in its incorporation within the CSN complex, likely through interactions with CSN3, CSN5, and CSN6 (27). However, due to its small size and highly disordered structure, it remains challenging to accurately determine interaction interfaces between CSN9 and CSN. As a result, no high-resolution structures are available for the CSN9-bound CSN complex. Thus, alternative strategies to dissect the architecture of the noncanonical CSN complex and determine how CSN9 interacts with CSN1–8 are needed to help us uncover structural details underlying the functional importance of CSN9 in cells.

In recent years, cross-linking mass spectrometry (XL-MS) has become a powerful strategy for probing protein–protein interactions (PPIs) (29–31). While effective, XL-MS possesses several inherent challenges, including unambiguous identification of cross-linked peptides due to their complex fragmentation when conventional (i.e., non-cleavable) cross-linkers are used. To facilitate MS identification, we have developed a suite of sulfoxide-containing MS-cleavable cross-linkers (e.g., disuccinimidyl sulfoxide [DSSO]) (32–36). These MS-cleavable reagents contain symmetric MS-labile C-S bonds (adjacent to the sulfoxide group) that are selectively and preferentially fragmented prior to peptide backbone cleavage during collision-induced dissociation (CID) (31–36). Such fragmentation has proven robust, thus enabling simplified and accurate identification of

cross-linked peptides by MS^n analysis. Among them, DSSO is an amine-reactive sulfoxide-containing MS-cleavable cross-linker that has been successfully applied for in vitro studies of purified protein complexes (32, 37, 38) and cell lysates (39, 40). Although lysine-reactive reagents are most popular, they alone cannot provide a full PPI mapping as some interaction interfaces do not contain proximal lysines for cross-linking (31). Therefore, we have developed dihydrazide sulfoxide (DHSO) for acidic residues (35) and bismaleimide sulfoxide (BMSO) for cysteine cross-linking (36), complementing the lysine-reactive DSSO and expanding PPI coverage on residue-specific protein interconnectivity. In addition to PPI mapping, XL-MS data has been successfully used for integrative structure modeling of protein complexes as observed cross-links impose upper distance bounds on pairs of cross-linked residues (41–44). Coupling cross-link data with other biophysical data (43, 44) and utilizing cross-linkers with different reactive chemistries (43) can significantly increase the accuracy of the resulting structures by integrative modeling. In comparison to conventional structural tools, XL-MS approaches can uniquely characterize large, heterogeneous, and dynamic protein assemblies in solution (31).

In this work, we developed and employed a multichemistry XL-MS approach enabled by three MS-cleavable cross-linkers to obtain comprehensive PPI maps of the CSN (CSN9-free) and CSNn (CSN9-bound) complexes to significantly improve precision and accuracy of their models. Based on our cross-link data, X-ray structures, and comparative models of CSN subunits, we computed the complete integrative structures of CSN and CSNn at 16- and 22-Å precisions, respectively. The integrative structures have maintained the core architecture of the known X-ray structure of CSN (PDB ID code 4D10), but importantly revealed additional conformations and configurations of CSN in

solution that were absent in the static structure. The integrative structure of CSNn has defined the CSN9 binding site in a cleft formed among CSN1, CSN3, and CSN8, resulting in local subunit reorientations that more likely contribute to CSN9-dependent increase of CSN deneddylase activity in vitro. Collectively, this work not only provides molecular features for us to better determine the structure dynamics of the CSN complex, but also reveals the structural basis underlying the role of CSN9 in CSN-mediated activities. Moreover, the integrated structural approach presented here is effective and can be generalized to define in-solution structures of dynamic protein complexes that remain inaccessible to other approaches.

Results

Multichemistry XL-MS Strategy for CSN Complexes

To define the architectures of CSN and CSNn complexes, we aimed to perform a comprehensive XL-MS analysis to maximize PPI mapping and to facilitate integrative structure modeling. To this end, we developed a combinatory XL-MS approach based on multiple MS-cleavable cross-linkers that carry specific but complementary cross-linking chemistries. Specifically, we selected three sulfoxide-containing MS-cleavable cross-linkers that target lysines (DSSO) (32), acidic residues (DHSO) (35), and cysteines (BMSO) (36). This combination is based on the critical roles of the selected reactive residues in protein structures, and the complementarity of the resulting cross-links for mapping PPIs. Both lysines and acidic residues are highly prevalent and often enriched at protein interaction interfaces, whereas cysteines are less abundant but can be more selective for targeting specific regions. In addition, no disulfide bonds have been reported

for CSN subunits, indicating that cysteine cross-linking would be suited for structural analysis of CSN. Importantly, the usage of these reagents has shown to significantly improve the coverage of PPI mapping even for simple proteins (35, 36). The general workflow of our multichemistry XL-MS strategy is illustrated in SI Appendix, Fig. S1. CSN complexes were purified under reducing condition after coexpression in *Escherichia coli* (Datasets S1 and S2), which were catalytically active and used for all XL-MS experiments. It is noted that CSN7 exists as two functionally redundant homologs in mammalian cells, CSN7a and CSN7b (45). Here, CSN7b was expressed and incorporated into CSN complexes for structural analysis. Each purified complex was first subjected to DSSO, DHSO, and BMSO cross-linking separately (SI Appendix, Fig. S1). The resulting cross-linked complexes were then enzymatically digested and separated to enrich cross-linked peptides by peptide size-exclusion chromatography (SEC) (46). The cross-links identified by liquid chromatography (LC)-MS^n analysis were then used for generating 2D cross-link maps to describe intersubunit interactions and for integrative structure modeling.

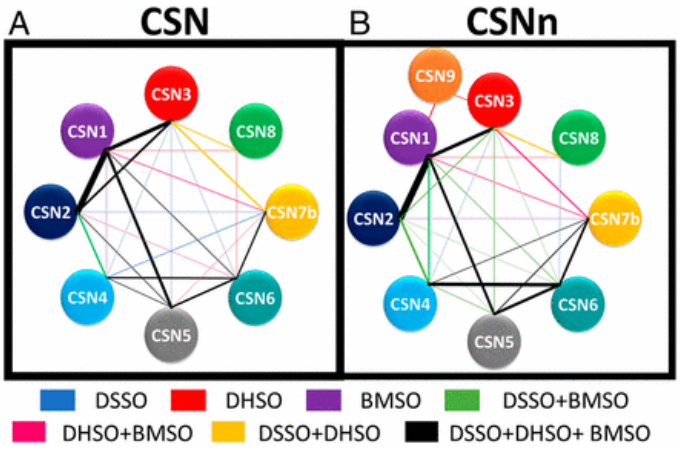

Figure 2.1 | PPI maps of the CSN complexes based on cross-link data from all three linkers (DSSO, DHSO, BMSO)
(A) CSN (CSN1–8). (B) CSNn (CSN1–9). Each CSN subunit is represented by colored nodes. The edges between two connected nodes are color-coded to describe PPIs resulted from individual or combinations of cross-linkers: That is, blue, DSSO; red, DHSO; purple, BMSO; lime, DSSO+BMSO; magenta, DHSO+BMSO; gold, DSSO+DHSO; black, DSSO+DHSO+BMSO. Edge thickness was determined by the total number of unique cross-links identified between the interactors.

Identification of CSN Cross-Linked Peptides

To illustrate cross-link identification, representative MS^n spectra of DSSO, DHSO, and BMSO cross-linked peptides of CSN are displayed in SI Appendix, Fig. S2. As DSSO, DHSO, and BMSO cross-linked peptides all carry two symmetric MS-cleavable bonds adjacent to the central sulfoxide in linker regions, cleavage of either one during MS^2 analysis physically separates cross-linked peptide constituents (α and β), resulting in the detection of two characteristic fragment ion pairs modified with complementary cross-linker remnants (α_A/β_T and α_T/β_A), regardless of linker chemistries (SI Appendix, Fig. S2 A–C). MS^3 analyses of these characteristic MS^2 fragment ion pairs enabled accurate identification of their sequences (SI Appendix, Fig. S2 D–I). In combination with MS^1 and MS^2 data, DSSO, DHSO, and BMSO cross-linked peptides were identified unambiguously. In this work, we have performed at least four biological replicates for each

XL-MS experiment. As a result, from all of XL-MS experiments, we identified a total of 682 DSSO, 275 DHSO, and 456 BMSO unique cross-linked peptides of CSN (Datasets S3–S5), and a total of 856 DSSO, 723 DHSO, and 576 BMSO unique cross-linked peptides of CSNn (Datasets S6–S9). Based on the identified cross-linked peptides, residue-to-residue linkages were determined (SI Appendix, Fig. S3). To ensure the validity of subsequent analyses, we decided to only use highly reproducible residue-to-residue linkages that have ≥ 60% occurrence among all biological replicates of each experiment. Thus, we obtained a total of 452 highly reproducible cross-links for CSN, including 214 K-K, 169 D/E-D/E, and 69 C-C linkages, describing 205 intersubunit (74 DSSO, 91 DHSO, and 40 BMSO) and 247 intrasubunit interactions (140 DSSO, 78 DHSO, 29 BMSO) (Datasets S10–S12). For CSNn, a total of 544 highly reproducible cross-links were acquired with 269 K-K, 167 D/E-D/E, and 108 C-C linkages, representing 244 intersubunit (86 DSSO, 83 DHSO, and 75 BMSO) and 300 intrasubunit interactions (183 DSSO, 84 DHSO, 33 BMSO) (Datasets S13– S15). These high-confidence cross-links were used for subsequent analyses (SI Appendix, Fig. S3).

The CSN Interaction Topology

To define intersubunit physical contacts, we generated experimentally derived interaction topology maps of CSN complexes based on the highly reproducible cross-link data (**Fig. 2.1** and Datasets S10–S15). As a result, extensive interaction networks were formulated comprising a total of 26 and 24 unique pairwise interactions for CSN and CSNn, respectively (Datasets S16 and S17). Among the three linkers, DSSO yielded the most connectivity within CSN, indicating lysine reactive reagents best-suited for general assessment of PPIs within CSN. While DHSO and BMSO identified less overall, they did

yield additional subunit contacts. Specifically, DSSO alone identified five unique PPIs; in comparison, DHSO and BMSO yielded a total of seven unique PPIs (Dataset S16). To better assess linker-dependent interactions, we constructed DSSO, DHSO, and BMSO PPI maps separately for each CSN complex (SI Appendix, Fig. S4 A–F). Since the majority of CSN subunits possess similar percentages of K, D/E, and C residues in their primary sequences, the number of cross-links representing each intersubunit interaction is more likely dependent on the number of cross-linkable residues at their interaction interfaces, as well as the detectability of resulting cross-linked peptides. This is further illustrated by 2D cross-link maps (SI Appendix, Fig. S4 G–L). For example, for the two smallest subunits of CSN, CSN7b has a relatively high percentage of acidic residues and its interactions were mostly revealed by DHSO, whereas CSN8 interactions were better described by DSSO due to its relatively high percentage of lysines (**Fig. 2.1A** and SI Appendix, Fig. S4 A–C).

Similar to the CSN complex, all three linkers have yielded extensive and complementary cross-links to represent subunit interconnectivities of CSNn (Dataset S17). Importantly, 16 CSN9-containing cross-links have been identified (Dataset S14), demonstrating its physical interactions within CSN at the residue level. Specifically, the C-terminal tail of CSN1 and several regions across CSN3 have been found to interact with CSN9. Since CSN9 is highly acidic with few lysine and no cysteine residues, only DHSO was able to capture CSN9 interactions within the CSNn complex. With the addition of CSN9, it appears that CSNn presented unique characteristics in its cross-link maps in comparison to those of CSN (**Fig. 2.1B** and SI Appendix, Fig. S4 D–F). This suggests that CSN9 may induce local changes in the CSNn complex that impact cross-link

formation. Collectively, our results have demonstrated the effectiveness and complementarity of our combinatory XL-MS strategy in mapping PPIs within CSN complexes. Integration of multichemistry cross-linking not only enabled cross-validation of intersubunit interactions, but also expanded interaction coverage due to the distinct capabilities of uncovering interactions at specific protein regions.

Mapping of CSN Cross-Links to the X-Ray Structure

To assess whether the cross-links agreed with the X-ray structure, we first mapped the identified K-K, D/E-D/E, and C-C linkages of CSN complexes to the existing CSN X-ray structure (3.8 Å, PDB ID code 4D10) by determining their Cα-Cα spatial distances (Datasets S10–S15). Considering linker spacer arm lengths (i.e., DSSO [10.1 Å], DHSO [12.4 Å], and BMSO [24.2 Å]), side-chain lengths of targeted amino acids (i.e., K [5.4 Å], D/E [2.5/3.7 Å], and C [2.8 Å]), as well as side-chain flexibility and dynamics, we have estimated the maximum Cα-Cα distances spanned by each linker: DSSO at 30 Å, DHSO at 30 Å, and BMSO at 45 Å. Thus, cross-links with distances above these thresholds were considered nonsatisfying or violating. For intersubunit interactions, 60% of DSSO cross-links of CSN were considered violating (SI Appendix, Fig. S5A). This is surprising as usually less than 20% of lysine-reactive cross-links are violated when mapped onto existing high-resolution structures (30, 38, 40). Similar discrepancies with the X-ray structure were observed for DHSO and BMSO data as 55% DHSO and 87% of BMSO intersubunit cross-links were beyond the expected thresholds (SI Appendix, Fig. S5 B and C). In contrast, most intrasubunit cross-links of CSN were satisfied in the Xray structure, with only 12% of DSSO, 15% of DHSO, and 21% BMSO violating intrasubunit cross-links (SI Appendix, Fig. S5 A–C). Since the high-resolution structure of CSNn has

not been resolved, we also mapped CSNn cross-links onto the same CSN structure. Similarly, a significant portion of intersubunit cross-links of CSNn from all three linkers (i.e., 57% of DSSO, 52% of DHSO, and 84% BMSO) were nonsatisfying (SI Appendix, Fig. S5 D–F), whereas for the intrasubunit cross-links, only 10% of DSSO, 10% of DHSO, and 23% BMSO were nonsatisfying (SI Appendix, Fig. S5 D–F). The high proportion of violating intersubunit crosslinks is more likely due to the additional conformations that CSN complexes may adopt in solution beyond the one observed in the X-ray structure.

Integrative Structure Modeling of the CSN Complex

To determine CSN structure in solution, we performed integrative structure modeling using the previously described four-stage workflow (SI Appendix, Supplemental Method and Fig. S6 and Dataset S18) (38, 43, 44, 47–50). The input information included the highly reproducible cross-link datasets (Datasets S10–S15), the X-ray structure of CSN (PDB ID code 4D10), and two comparative models of CSN7b subunit domains based on the structure of the CSN7a subunit in the X-ray structure of CSN. The representation of the system used for modeling of CSN was chosen as follows. First, the helical bundle comprising segments from each of the eight subunits was constrained based on the X-ray structure. Second, the remaining structures of subunits CSN1–8 were represented by 15 rigid bodies, corresponding to different domains of the proteins (SI Appendix, Supplemental Method and Fig. S7H and Dataset S18). Finally, short (4 to 13 residues) segments linking rigid bodies and regions missing in the X-ray structure (2 to 52 residues long) were modeled as flexible strings of 2 to 10 residue beads each. Next, we exhaustively sampled configurations of the 16 rigid bodies (i.e., the helical bundle and the 15 rigid bodies) that satisfy the cross-links as well as sequence connectivity and excluded

volume restraints, using a Monte Carlo method that started with a random initial structure. The modeling did not rely on any knowledge of the X-ray structure except for the shapes of the 16 rigid bodies. The sampling yielded 71,350 representative models that sufficiently satisfied the input restraints. The clustering of the ensemble identified a single distinct cluster containing the majority (76%) of the individual models (SI Appendix, Fig. S7 A–D), corresponding to the complete integrative structure of CSN in solution. The precision of the cluster corresponds to the variability among the clustered ensemble and defines the overall precision (uncertainty) of the integrative CSN structure (**Fig. 2.2A** and SI Appendix, Fig. S7), which was quantified by the average RMSD with respect to the centroid of 16 Å (SI Appendix, Supplemental Method). The centroid structure is the most similar structure to all of the other structures in the cluster.

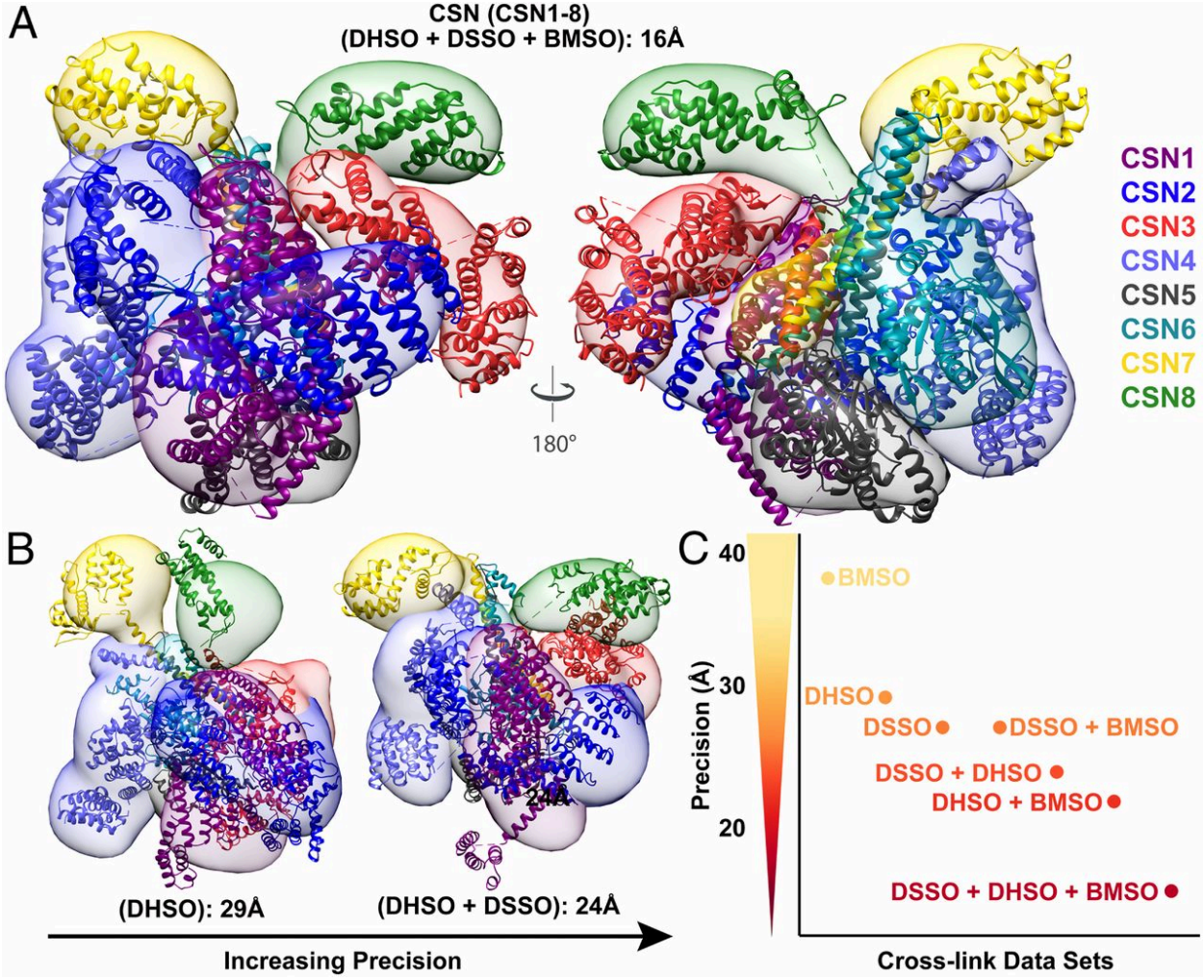

Figure 2.2 | Integrative structures of CSN
(A) The integrative structure of CSN determined at 16-Å precision when all three cross-link datasets (DSSO+DHSO+BMSO) were used for modeling. For each subunit, the localization probability density of the ensemble of models is shown with a representative structure (the centroid) from the ensemble embedded within it. (B) Integrative modeling of CSN determined using DHSO or DHSO+DSSO datasets yielded models determined at 29- and 24-Å precision, respectively. (C) Graphical representation of determined model precisions with seven combinations of our three cross-link datasets, illustrating that increasing the number of cross-linking chemistries (abscissa axis) for integrative structure modeling leads to increased model precision (ordinate axis). CSN subunit was color-coded as illustrated.

Validation of the Integrative Structure of the CSN Complex

To validate the integrative structure of CSN, we first assessed how well it satisfied the input cross-links used to compute it. The integrative structure of CSN satisfied 98% of the cross-links. The remaining 2% of the cross-links would be satisfied if the threshold was increased by 10 Å (SI Appendix, Fig. S7F). These violations can be rationalized by experimental uncertainty, coarse-grained representation of the complex, and finite structural sampling. Next, we evaluated the integrative structure of CSN by cross-validation against different input cross-link datasets. Namely, we independently repeated integrative modeling described above with six different subsets of CSN cross-links (Datasets S10–S12), including: 1) DSSO only, 2) DHSO only, 3) BMSO only, 4) DSSO and DHSO, 5) DSSO and BMSO, and 6) DHSO and BMSO. The results were examined in three ways as follows. First, we gauged how well each of the six CSN model ensembles satisfied different subsets of the cross-links. All six models satisfied more than 95% of all cross-links, whether or not they were used for modeling, thus increasing our confidence in modeling. Second, we showed that increasing the amount of input information improved the precision of the output model when sampling was exhaustive. This result is expected when the choice of model representation (here, the 16 rigid bodies) is appropriate for input information (here, mainly the cross-links) as encoded in the scoring function. In addition to validating the model and the data, the improved precision of the model resulting from increasing the number of cross-linking chemistries demonstrate the complementarity of the three cross-linking datasets (**Fig. 2.2B** and **C**). Specifically, the model precision increased from 37 Å for BMSO cross-links only to 16 Å for all three types of cross-links (i.e., DSSO+DHSO+BMSO). Third, we calculated the overlaps between the

integrative structure ensemble using all cross-links and each of the six model ensembles based on a subset of crosslinks. The overlap was quantified by the ratio of the distance between ensemble centroids to three times the sum of the ensemble precisions (SI Appendix, Supplemental Method). The distance between two ensemble centroids is defined by their RMSD. The ensemble precision is defined by the RMSD from the centroid averaged over all models in the ensemble. In particular, two structural aspects were evaluated, including the tertiary structure of each individual subunit (a total of 8 subunits) as described by the intramolecular distances as well as the relative positions and orientations of all pairs of subunits (a total of 28 pairs) in the complex as described by the intermolecular distances. For each of the 8 subunits and each of the 28 pairs of subunits, the integrative structure based on all cross-links overlapped with the integrative structures based on each of the 6 cross-link subsets (SI Appendix, Fig. S8). Therefore, these cross-validations further increased our confidence in the integrative structure of CSN.

Comparison of Integrative and X-Ray Structures of CSN

To compare the integrative and X-ray structures of CSN, we first examined how well both structures satisfied our cross-link datasets and determined that the integrative structure did much better than the X-ray structure, for both intrasubunit (98% vs. 85%) and intersubunit (99% vs. 39%) cross-links (SI Appendix, Fig. S7F and Datasets S19–S21). These results indicate that the integrative structure ensemble is a better representation of CSN conformations in solution than the X-ray structure.

Next, we inspected whether or not the integrative model preserved the core of the previously determined CSN structures, which contains three main features: 1) The PCI ring (in the order of CSN7-CSN4-CSN2-CSN1-CSN3-CSN8), 2) the CSN5–CSN6 dimer,

and 3) a helical bundle consisting of a helix from each of the eight subunits (23, 45, 51). During our modeling, while the helical bundle was constrained as a rigid body (Figs. 2A and 3 A and D), the order of the PCI ring and CSN5–CSN6 dimer were not enforced. However, the latter two features emerged from our simulation and resemble those in the X-ray structure (Figs. 2A and 3B and SI Appendix, Fig. S7G). This preservation is important especially for the CSN5–CSN6 dimer, as it is crucial for keeping CSN5 inactive in the absence of a substrate, and releasing CSN5 for activation upon substrate binding (12, 23, 24, 52). The CSN5–CSN6 dimer was well-represented by our crosslink data, resulting in the highest precisions among the 28 pairs of subunits in the integrative structure of CSN (16 Å) (Figs. 2A and 3 B and D and SI Appendix, Fig. S7). Moreover, subunits CSN3 and CSN8 also adopted similar positions and orientations relative to other subunits in both the integrative and X-ray structures (Figs. 2A and 3A and SI Appendix, Fig. S7G), albeit the precision of the CSN3–CSN8 pair in the integrative structure was relatively low (25 Å). In summary, the core of CSN integrative structure in solution is similar to previous X-ray and electron microscopy (EM) structures (23, 45, 51).

Finally, we computed the RMSD between the CSN X-ray and integrative structure centroids to assess whether the RMSD was larger than three times the precision of the integrative structure, as the resolution of the X-ray structure is much higher than that of the integrative structure. The crystallographic structures of three subunits (i.e., CSN2, CSN4, and CSN5) and four pairs of subunits (i.e., CSN2–CSN4, CSN2–CSN5, CSN4–CSN5, and CSN4–CSN6) were found to lie further than three times the integrative structure precision from the ensemble centroid (**Fig. 2.3D**), indicating significant differences in these regions between the two compared structures. The observed

differences were further supported by the largest RMSDs measured in these regions between the X-ray and integrative structure centroid of CSN (SI Appendix, Fig. S9A). The detected discrepancies are unlikely the result of integrative modeling uncertainty; instead, they likely reflect different functional states in solution or differences between the solution and X-ray structures. Specifically, the C terminus of CSN4 interacts tightly with the C terminus of CSN6 (precision of 20 Å) (**Fig. 2.3D** and SI Appendix, Fig. S7G), opposite from CSN5 in the integrative structure (**Figs. 2.2A** and **2.3A**). In contrast, CSN4 does not interact with CSN6 in the X-ray structure (**Fig. 2.3D**). The relative positions and orientations of CSN2, CSN4, CSN5, and CSN6 in the integrative structure were determined by satisfying all but 1 of the 47 intersubunit crosslinks. In contrast, the X-ray structure only satisfied 30 of these cross-links.

Although the arrangement order of CSN1, CSN2, and CSN3 remained unchanged, the N terminus of CSN2 was found to wrap around CSN1 toward CSN3 in the integrative structure (**Figs. 2.2A** and **2.3 A** and **C**), whereas it projected outwards without contacting either CSN1 or CSN3 in the X-ray structure. The relative positions and orientations of CSN1, CSN2, CSN3, and CSN4 in the integrative structure were determined by satisfying all but 1 of the 98 intersubunit cross-links. In contrast, the X-ray structure only satisfied 28 of these cross-links and none of the 16 crosslinks between CSN2 and CSN3. Taken together, the results demonstrate that integrative structure modeling of CSN based on our comprehensive cross-link data were able to not only recapitulate the core architecture common to all known CSN structures, but also uncover significant quaternary differences relative to the X-ray structure.

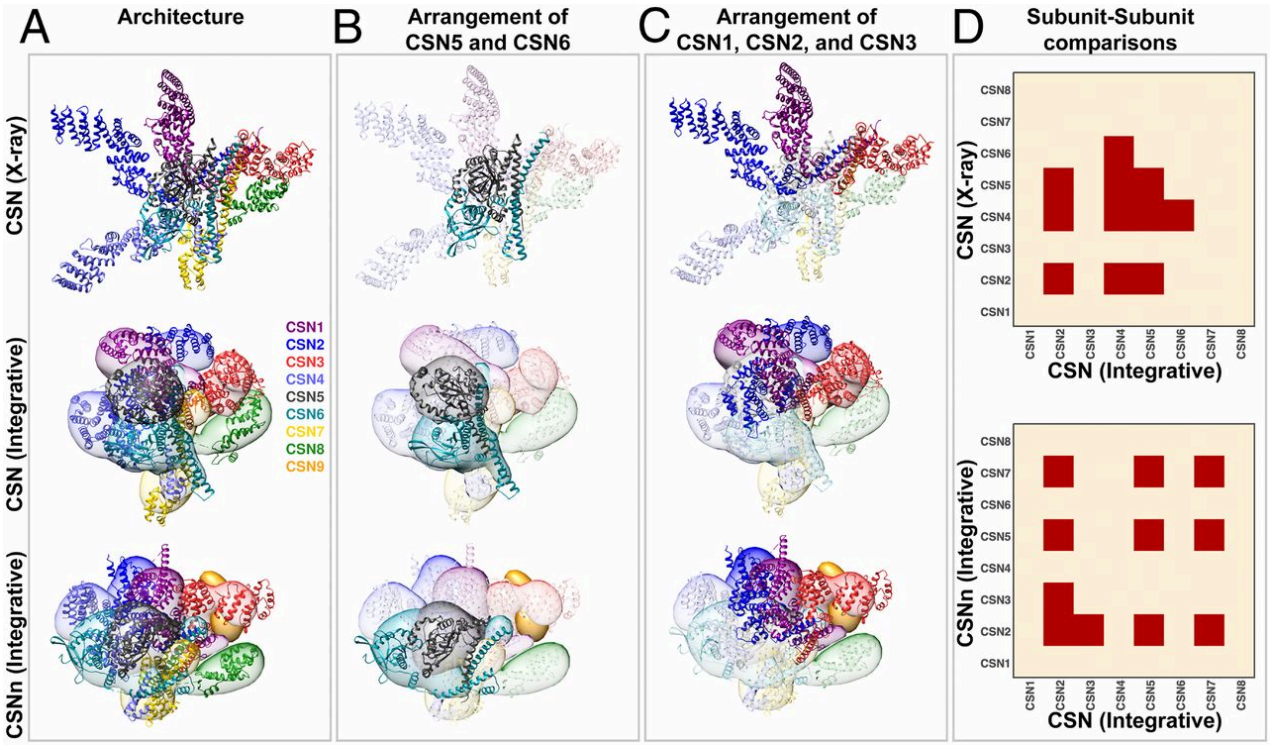

Figure 2.3 | Comparison of integrative and X-ray structures of the CSN complexes.

(A) Overall architectures of CSN: X-ray structure (PBD ID code 4D10) (Top), CSN integrative structure (Middle), and CSNn integrative structure (Bottom). For each subunit in the integrative structures, the localization probability density of the ensemble of models is shown with a representative structure (the centroid) from the ensemble embedded within it. The CSN and CSNn structures show that the models adopt a more condensed state as compared to the X-ray structure, but they generally retain the overall architecture with only the helical bundle being constrained during modeling. (B) The arrangement of the CSN5–CSN6 (MPN domain containing subunits) dimer was an emerging feature in integrative structures; however, a slight shift in the interface was observed in the CSNn model. (C) Models indicate that the arrangement of CSN1, CSN2, and CSN3 was altered in the presence of CSN9; CSN2 moved from a state interacting with CSN3 in CSN to an opened state in the CSNn model, resembling the overall architecture of the CSN X-ray structure. (D) Respective binary subunit–subunit comparison of the CSN integrative structure with the CSN X-ray structure (Upper) and the CSNn integrative structure (Lower), respectively. The structures were compared by calculating their ensemble overlap; the overlap was quantified by the ratio of the distance between ensemble centroids to three times the sum of the ensemble precisions. Differences are shown in red. The CSN subunit was color-coded as illustrated.

Integrative Structure Modeling of the CSNn Complex

To localize the CSN9 subunit and map its interactions with the CSN complex, we also performed integrative structure modeling of CSNn (CSN9-bound CSN), based primarily on 619 highly reproducible cross-links for CSNn from all three cross-linkers (SI Appendix, Supplementary Method and Fig. S10 and Dataset S22). Integrative structure modeling of CSNn was performed the same way as described above for CSN. The structure of CSN9, a 57 amino acid-long acidic protein, is unknown and cannot be modeled. Therefore, it was represented as a string of flexible beads corresponding to two residues each. The sampling of the CSNn complex yielded 125,750 representative models that sufficiently satisfied the input restraints. The clustering of the ensemble identified a single distinct cluster containing the majority (79%) of the individual models (SI Appendix, Fig. S10), corresponding to the complete integrative structure of CSNn in solution. The precision of the cluster is 22 Å (SI Appendix, Fig. S10 A–D), which is sufficient to map all positions and relative orientations of CSN1–9 subunits (Figs. 3A and 4A and SI Appendix, Fig. S10E). Moreover, the integrative structure of CSNn satisfied 99% of the input cross-links (intersubunit and intrasubunit) (Datasets S23–S25). Importantly, the resulting structure of CSNn has precisely localized CSN9 at a cavity formed by the C terminus of CSN1, all of CSN3, and CSN8 (**Fig. 2.4A**). The position of amino acid residues 20 to 57 of CSN9 was specified by satisfying all of the 16 CSN9-containing intersubunit cross-links (**Fig. 2.4 A** and **B**). It is noted that the exact position of the first 19 amino acid residues of CSN9 could not be accurately determined since cross-linked peptides involving this region were not identified. Regardless, we were able to determine the interactions of CSN9 with CSN1–8 in the integrative structure. We consider a contact between CSN9 and any of the CSN1–

8 subunit if the two subunits are within 12 Å from each other; a contact is defined as an interaction if the contact frequency across the ensemble is at least 75% (SI Appendix, Fig. S10G). As a result, CSN1 and CSN3 were found in the closest proximity to CSN9 across the ensemble and thus were identified as CSN9 interactors, corroborating well with our cross-link data. Therefore, the CSN9–CSN interactions have been precisely determined by integrative structure modeling (**Fig. 2.4** and SI Appendix, Fig. S10G), providing CSN9's binding cavity and its interactors.

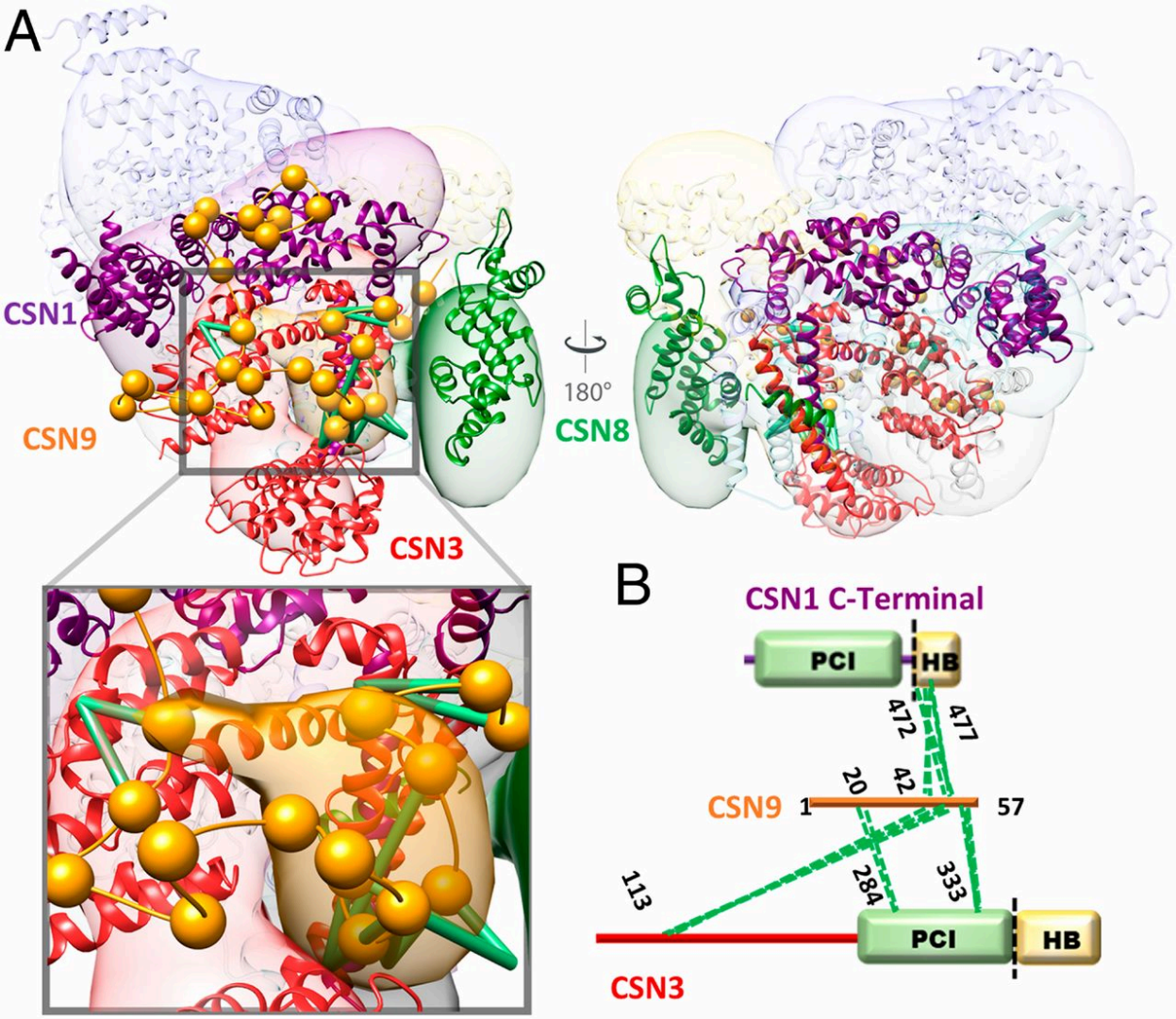

Figure 2.4 | Binding of CSN9 in the CSNn integrative structure.
(A) The integrative structure of CSNn determined at 22-Å precision using all three cross-link datasets (DSSO+DHSO+BMSO). For each subunit, the localization probability density of the ensemble of models is shown with a representative structure (the centroid) from the ensemble embedded within it. The higher probable localization of CSN9, corresponding to its C terminal, on the CSNn model is represented by the orange localization probability density, and a representative structure from the ensemble is shown with spheres corresponding to two residues per beads connected by an extrapolated trace of the backbone. CSN9 primarily interacts with the main body of CSN3 (red) while its C-terminal tail also falls into the cavity between CSN1 (purple), CSN3 (red), and CSN8 (green). The Inset displays a closer view of CSN9 interaction. Green lines represent CSN9-containing DHSO cross-links. (B) Two-dimensional DHSO cross-link map linking CSN9 to CSN1 and CSN3 at specific residues.

Comparison of Integrative Structures of the Canonical and Noncanonical CSNs

To compare the two CSN complexes in light of their precisions, we then examined their structural differences among the conformations of single subunits and configurations of pairs of subunits by assessing whether the differences are larger than the sum of their precisions (**Fig. 2.3D**) and by computing the RMSD between their respective centroid (SI Appendix, Fig. S9B). While a large portion of the two compared structures was similar, the conformation of 3 of the 8 subunits (i.e., CSN2, CSN5, and CSN7) and 3 of the 28 pairs of subunits (i.e., CSN2–CSN3, CSN2–CSN5, and CSN2–CSN7) had notable differences in these regions (**Fig. 2.3D** and SI Appendix, Fig. S9B). Both the integrative structures of CSN and CSNn maintained similar core structures (i.e., ordering of the PCI ring, the CSN5–CSN6 dimer, and the helical bundle) (**Fig. 2.3B**). However, CSN2 changed its conformation and position relative to its neighbors (i.e., CSN3, CSN5, and CSN7) (**Fig. 2.3 A**, **C**, and **D** and SI Appendix, Fig. S9B). Specifically, in the integrative structure of CSNn, CSN2, and CSN4 localize adjacent to one another, allowing the formation of the CSN9-binding cavity (**Figs. 2.3D** and **2.4**). The conformation and relative position of the CSN2 subunit in the integrative structure of CSNn were determined by satisfying all 74 intersubunit cross-links obtained for CSNn. Therefore, our results suggest that CSN2 possesses structural plasticity, enabling its interaction with CSN1 and CSN3 to yield a more open configuration in CSN9-bound CSN than in CSN9-free CSN.

To explore the potential role of CSN9-mediated structural changes, we compared the integrative structures of CSN and CSNn to the cryo-EM structure of the CRL4A-bound CSN complex (at resolution of 6.4 Å) (24). Specifically, we assessed whether the structure of the CSN complexed with neddylated CRL4A overlapped with the two integrative

structures. The structure of CRL4A-bound CSN differs from the integrative structure of CSN for one subunit (i.e., CSN2) and two pairs of subunits (i.e., CSN2–CSN4 and CSN2–CSN5) (SI Appendix, Fig. S9C). In contrast, the structure of CRL4A-bound CSN has no significant differences with the integrative structure of CSNn (SI Appendix, Fig. S9D). Similar comparisons were performed with the structure of CRL1-bound CSN (at resolution of 7.2 Å) (25). While the structure of CSN bound to neddylated CRL1 differs from the integrative structure of CSN for two subunits (i.e., CSN2 and CSN5) and three pairs of subunits (i.e., CSN2–CSN4, CSN2–CSN5, and CSN2–CSN6) (SI Appendix, Fig. S9E), it has no significant differences with the integrative structure of CSNn (SI Appendix, Fig. S9D). Collectively, these assessments suggest that CSN9-bound CSN is structurally similar to CRL-bound CSN (24, 25). Upon CSN9 binding, the integrative structure of CSNn displays local structural changes, mainly on the conformation and position of CSN2. Specifically, CSN2 moves closer to CSN4, causing CSN9-bound CSN to adopt a configuration resembling CRL-bound CSN (24, 25).

Biochemical Validation of CSN9 Binding

In order to validate the interactions of CSN9 with the CSN complex revealed by XL-MS and structural modeling, we performed in vitro binding assays using purified CSN subunits. CSN9 only interacts with CSN1-2-3 and CSN1-2-3-8 subcomplexes, whereas no binding was detected with CSN4-6-7, CSN4-6-7-5, or CSN4-6-7-5-8 subcomplexes (SI Appendix, Fig. S11). These results confirm that CSN1 and CSN3 are present in the subcomplex required for CSN9 binding onto CSN. To understand the importance of CSN9, we have compared in vitro deneddylase activities of CSN and CSNn with neddylated Cullin 1 as the substrate. Similar results were obtained for the same assay

performed at different time scales (SI Appendix, Fig. S12), demonstrating that CSNn displayed markedly increased activity over CSN and CSN9 can enhance CSN activity in vitro.

Quantitative Validation of the Structural Dynamics of the CSN Complexes

To validate the observed structural differences between CSN models with and without CSN9, parallel reaction monitoring (PRM)-based targeted quantitation of CSN crosslinks was utilized (53). Since DHSO cross-linking yielded the most intersubunit linkages best describing CSN9-induced structural changes, we individually cross-linked CSN and CSNn with DHSO for PRM experiments. To perform unbiased quantitative analysis, we generated a total of 341 PRM targets based on highly reproducible DHSO cross-linked peptides previously obtained from CSN and CSNn complexes (Datasets S11 and S14). Peptide quantitation was derived from the summation of peak areas of all transitions through Skyline software. As exemplified in **Fig. 2.5A**, an intra-CSN4 cross-link (E306–E345) from both CSN and CSNn samples displayed similar abundance, indicating that this interaction is independent of CSN9. In contrast, a CSN2–CSN3 cross-link (CSN2:E63–CSN3:E333) was only observed in CSNn and not in CSN, suggesting a CSN9-induced conformational change. In total, 229 DHSO cross-linked peptides were quantified, which represent 18 intersubunit interactions (Dataset S26). As shown in **Fig. 2.5B**, the vast majority of quantified cross-links remained unchanged between CSN and CSNn, confirming that CSN9 does not trigger major organizational changes within the CSN complex during its binding. This corroborates well with the modeling results as both of our CSN models satisfied 99% of DHSO cross-links from both complexes. Apart from

unchanged interactions, a total of 22 cross-linked peptides were found with significant changes (>2.5-fold, greater than 3σ) between the two compared complexes (**Fig. 2.5B**).

Besides cross-links involving CSN9, two additional crosslinked peptides corresponding to two intersubunit interactions (i.e., CSN4–CSN6 and CSN6–CSN7) have decreased CSN/CSNn ratios, suggesting that these cross-links are favored in CSNn. In contrast, 18 cross-linked peptides describing 7 intersubunit interactions (CSN1–CSN2, CSN1–CSN3, CSN1–CSN5, CSN2– CSN3, CSN2–CSN7, CSN4–CSN5, and CSN6–CSN7) and 1 intra-CSN1 interaction have increased CSN/CSNn ratios, implying that these cross-links are preferably formed in CSN. Apart from CSN9-containing interactions, five quantifiable CSN2– CSN3 cross-links exhibited the most significant changes between the two compared complexes with CSN/CSNn ratios all greater than 10.2 (**Fig. 2.5C**), indicating that CSN2–CSN3 interactions were severely disrupted upon CSN9 binding. This is consistent with the structural differences between CSN and CSNn revealed by integrative modeling as these linkages were only satisfied by the CSN models (**Fig. 2.5D**). Since CSN1 closely interacts with CSN2, CSN3, and CSN9, the decreased abundance of CSN1– CSN2 and CSN1–CSN3 cross-links in CSNn supports the CSNn model, suggesting that the main body of CSN2 swings away from CSN1 and CSN3 into a more open state. Collectively, PRM based targeted quantitation of CSN cross-links strongly supports structural similarities and differences between the integrative models of the two CSN complexes.

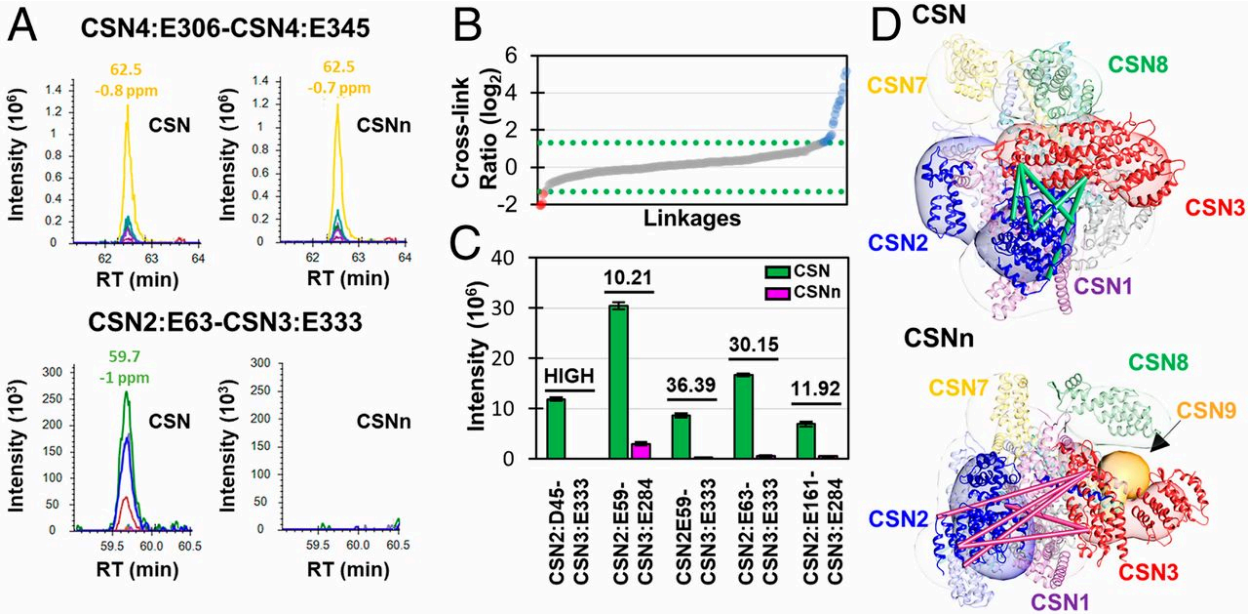

Figure 2.5 | PRM-based targeted quantitation of DHSO cross-linked peptides to validate CSN9-induced structural changes in CSN.

(**A**) Skyline outputs for PRM quantitation of a representative DHSO intrasubunit (CSN4:E306–CSN4:E345) (Upper) and an intersubunit (CSN2:E63–CSN3:E333) (Lower) cross-linked peptides to compare their relative abundance in the CSN and CSNn complexes. Based on peak areas, the relative abundance ratio (CSN/CSNn) of the intrasubunit cross-link was determined as 1.11 (Upper), indicating no significant change. In contrast, the relative abundance of the intersubunit cross-link (CSN/CSNn) was determined as 30.15 (Lower), suggesting a significant change. (B) The distribution of cross-link ratios (CSN/CSNn) of 229 DHSO cross-linked peptides (represented as log2 values) determined by PRM quantitation, in which only 22 cross-linked peptides displayed significant changes (>2.5-fold, greater than 3σ), including 4 with decreased ratios (red dots) and 18 with increased ratios (blue dots). The cross-link ratios (CSN/CSNn) describe the relative abundance of cross-linked peptides in the two compared complexes. (C) Abundance of five quantifiable CSN2–CSN3 cross-links (CSN2:D45–CSN3:E333, CSN2:E59–CSN3:E284, CSN2E59–CSN3:E333, CSN2:E63–CSN3:E333, and CSN2:E161–CSN3:E284) detected in the CSN and CSNn complexes. The underlined numbers shown represent relative abundance ratios (CSN/CSNn) of the selected cross-linked peptides between the two complexes, indicating that these interactions are favored in CSN. (D) The five cross-links shown in (C) were mapped on CSN and CSNn integrative structures. The linkages in the CSN model (green) are satisfied within the expected distance (<30 Å), which are not satisfied in the CSNn model (magenta). Details on PRM quantitation of the cross-linked peptides are listed in Dataset S26.

Discussion

In this work, we have developed a multichemistry XL-MS approach based on three distinct MS-cleavable cross-linkers (i.e., DSSO, DHSO, and BMSO) to comprehensively map PPIs and facilitate integrative structure modeling of CSN complexes. The large number of cross-links identified in this work is highly complementary, allowing expanding PPI coverage and cross-validating results. This approach enables us to obtain the most extensive intrasubunit and intersubunit interaction maps of CSN (CSN9-free) and CSNn (CSN9-bound) complexes. It is noted that CSN9-containing interactions were only identified through DHSO cross-linking, not by DSSO and BMSO, signifying the need of multichemistry XL-MS to fully characterize PPIs of CSN complexes. Importantly, the combinatory XL-MS data enabled structural characterization of the CSN complexes with complete sequences and significantly enhanced the precision of integrative structure modeling, resulting in the precisions of 16 and 22 Å for CSN and CSNn, respectively. These are considerably higher than the precision of models from single and dual cross-linking chemistries (24 to 37 Å). While lysine-to-lysine and acidic-to-acidic residue cross-links have been successfully applied for structural mapping and modeling (31, 35, 38, 46, 54, 55), we demonstrate here that cysteine-to-cysteine cross-links are as effective for structure determination of protein complexes. This is illustrated by the fact that a single integrative structure (i.e., a single cluster of models) satisfies most of the BMSO cross-links, similarly to DSSO and DHSO cross-links (**Fig. 2.2** and SI Appendix, Fig. S7). In addition, we obtained highly overlapping model ensembles based on seven different combinations of the three types of cross-link data (i.e., DSSO, DHSO, and BMSO crosslinks) (**Fig. 2.2C** and SI Appendix, Fig. S8), confirming the validity and coherence of

our cross-link data. Therefore, coupling combinatory XL-MS based on multiple cross-linking chemistries with integrative structure modeling facilitates the determination of the interaction and structure dynamics of CSN complexes. The same strategy can be directly adopted for characterizing architectures of other dynamic protein complexes in solution.

During XL-MS analyses, we have found that although the majority of intrasubunit cross-links of CSN from all three linkers were satisfied by the known X-ray structure (PDB ID code 4D10), most of intersubunit cross-links were classified as violating. This implies that CSN has much more flexible intersubunit than intrasubunit interactions. Since X-ray crystallography only reveals static structures with a single conformation, distance violation of cross-links suggests the presence of multiple conformations and configurations of CSN in solution. Similar results have been obtained for the CSNn complex, further confirming the interaction and structural plasticity of CSN complexes. While CSN is known to carry structural flexibility to allow its interaction with a diverse array of CRLs to regulate their activities (12, 23–25), our XL-MS results provide additional evidence to support CSN structural heterogeneity in solution. Because of this, our cross-link dataset generated here is comprised of a wide range of possible conformations of CSN complexes. Therefore, to minimize complexity, only highly reproducible cross-link data were used to derive structural ensembles that represent major conformations of CSN complexes in solution. The integrative structures of CSN complexes have satisfied 98% of all of the cross-links obtained in this work, considerably better than the X-ray structure. This result further indicates that CSN contains additional accessible states other than the one determined by X-ray crystallography. In contrast to the observed conformational and configurational differences in intersubunit interactions, the core structure of CSN is preserved. Indeed,

we have found that the CSN model maintains overall configuration with the presence of the PCI ring and the positioning of CSN5–CSN6 dimer, apart from a rearrangement of CSN2 with respect to CSN1, CSN3, and CSN4 positioning in the complex. The core structure of CSN has also been detected in the CSNn model, which was derived from a completely different set of cross-link data used for CSN modeling. As these core modules are crucial for the CSN assembly, structure, and function (12, 23–25, 45, 51), their determination by integrative modeling based primarily on cross-links further demonstrates the effectiveness of our approach and the validity of the determined integrative structures.

Here, we have determined that CSN9 predominantly interacts with CSN3 and CSN1, and is localized in a cavity formed by CSN1-3-8 in the CSNn structure. Although CSN3–CSN9 interaction has previously been shown biochemically (27), our results have identified interaction contacts between the two interactors. Importantly, we have identified CSN1 as an additional CSN9 interactor and determined CSN9 binding sites within the CSN complex. While it has been suggested that CSN9 may bind to CSN5 and CSN6 (27), no cross-links between CSN9 and CSN5 or CSN6 were identified and the integrative structure of CSNn shows that both subunits are much farther away from CSN9 than CSN1 and CSN3. Interestingly, CSN1, CSN3, and CSN8 form a connected submodule in the integrative and X-ray structures of CSN (23), and the assembled CSN1–3-8 subcomplex can be isolated in mammalian cells (56). It is known that each CSN subunit has a corresponding homolog in the nine-subunit 19S lid complex (27, 57). Recently, the proteasome subunit DSS1/ Rpn15, the homolog of CSN9, has been determined to interact with Rpn3 (homolog of CSN3) and Rpn7 (homolog of CSN1), which forms a subcomplex prior to the 19S lid assembly (58), corroborating well with the close

interactions of CSN9 with CSN3 and CSN1. These results further indicate interaction similarities between the CSN and the 19S lid complexes.

Apart from similarities in organizational architectures in the CSN integrative and X-ray structures, we have observed structural differences between the integrative structures of CSN and CSNn that may contribute to CSN dynamics. One notable difference is the CSN2–CSN3 interaction and its relative location to CSN1 subunit. Specifically, in the CSN integrative structure, the CSN2 N terminus wraps around CSN1 toward CSN3 and away from CSN4, whereby CSN2 is not readily available to interact with Cullin and Rbx1. This is of importance because CSN2 plays a major role along with CSN4 in stabilizing the CSN–CRL interaction when CSN binds to CRLs (12, 23–25). CSN1 has been shown to bind to the CRL4A adaptor DDB1, which is important in stabilizing Cul4A and required for efficient deneddylation (24). However, CSN1 involvement appears to be specific for CRL4 and not CRL2 and CRL3 complexes (24, 26). While CSN3 has not been shown to directly contact CRL components, overexpression of CSN3 leads to increased amounts of CSN in cells and downregulation of CSN3 causes the destruction of CSN and cell death (59). Thus, we speculate that the observed changes of interactions among CSN1, CSN2, and CSN3 may represent one of the major conformations of CSN that is needed to interact with specific subsets of CRLs in cells.

While the integrative structures of CSN and CSNn have both maintained the core structure of CSN, CSN9 binding causes a major shift in CSN2 and its interactions with neighboring subunits that have been confirmed by quantitative XL-MS analysis. Given the critical importance of CSN2 in CSN–CRL interactions (12, 24, 25), we suspect that CSN9-induced structural changes may be associated with the augmented CSN in vitro

deneddylase activity observed in this work. Comparative analysis has revealed that the major differences between canonical CSN (CSN9-free) and CRL-bound CSN lie in the relative position of CSN2 and its interaction with CSN5 (SI Appendix, Fig. S9 C and E), indicating that CSN2 has to undergo conformational changes to fulfill its role in facilitating CSN binding to CRLs (24, 25). Therefore, the observed structural alterations at CSN2 would be important for the formation of the CSN–CRL complex, the prerequisite for subsequent deneddylation. The structure similarity between CSN9-bound CSN and CRL-bound CSN (SI Appendix, Fig. S9D) strongly supports the biological relevance of CSN9-induced structural changes. Thus, these results prompt us to propose a structural model in which CSN9 causes the canonical CSN to adopt a configuration favorable for interacting with CRLs (**Fig. 2.6**). In the absence of CSN9, binding of neddylated CRL to CSN results in a series of conformational changes, among which the initial important steps involve the movement of N-terminal domains of CSN2 and CSN4 toward cullin (12, 24, 25). These rearrangements occur prior to the release and activation of CSN5.

In contrast, the addition of CSN9 triggers CSN to undergo conformational changes by repositioning the N terminus of CSN2 away from CSN3 but closer to CSN4 (**Fig. 2.6**). As the resulting conformation and configuration of CSN9-bound CSN are highly similar to those of CRL-bound CSN, we suspect that CSN9 may enhance the affinity (or recognition) between CSN and its substrate, neddylated CRLs, thus facilitating the assembly of CSNneddylated CRL complex to enhance CSN activation and deneddylation of CRLs. In addition, the conformation of CSNn may also enable its faster release from deneddylated CRLs as reported (14). In the absence of CSN9, the assembly/disassembly of the CSN–CRL complex would more likely be much slower due to substantial

conformational changes required for the activation of CSN upon binding to CRLs, thus leading to slower deneddylation rate. Therefore, the differences in the assembly/disassembly of CSN–CRL complexes more likely contribute to their interaction affinity, and slower disassembly of the CSN–CRL complex could imply tighter interaction.

In summary, CSN9-induced conformational changes related to CSN2, are biologically relevant, especially in preparing CSN for associating with neddylated CRLs, thereby contributing to augmenting deneddylation activity of CSN. The integrative structures of CSN complexes determined in this work have established a structural basis for us to further dissect condition-induced structural dynamics of CSN in the future, unraveling molecular insights into its activation, function, and regulation under different physiological and pathological conditions.

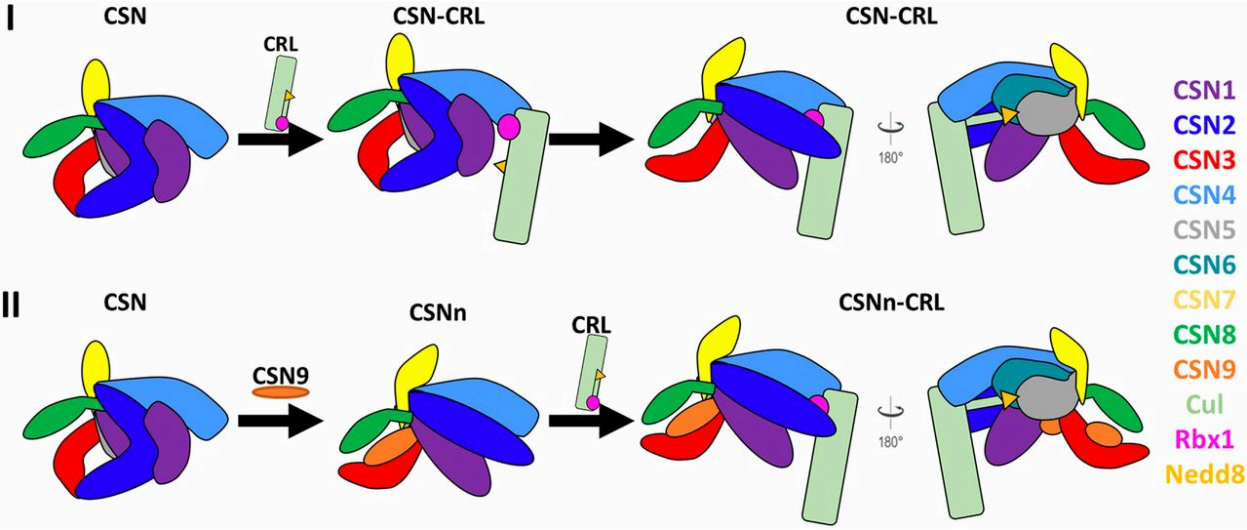

Figure 2.6 | The proposed structural model of CSN9 binding to facilitate CSN interaction with neddylated CRLs.
CSN and neddylated CRL subunits were color-coded as illustrated. (I) CSN9-free CSN needs to undergo substantial conformational changes upon binding to a neddylated CRL. In comparison, (II) CSN9-bound CSN adopts a configuration better suited for CRL binding.

Methods

Expression and Purification of CSN Complexes

Eight of total nine subunits of the human CSN complex, except CSN5, were overexpressed and purified from *E. coli*. Two three-subunit subcomplexes, CSN1-2-3 and CSN4-6-7, were prepared through coexpression. Briefly, CSN2 was subcloned into a modified pGEX4T1 (Amersham Biosciences) vector containing a GST tag followed by a tobacco etch virus (TEV) protease cleavage site, while both CSN1 and CSN3 were subcloned into a modified pET15b (Novagen) vector containing a chloramphenicol resistance cassette. After coexpression in BL21(DE3) (Novagen), the CSN1-2-3 formed a complex and was purified by glutathione-affinity chromatography. Following TEV cleavage, the CSN1-2-3 subcomplex was further purified by anion exchange and gel-filtration chromatography. CSN4-6-7 was prepared in the same way. CSN8 and CSN9 were subcloned into the pGEX4T1 vector individually and subjected to the same purification procedure. Recombinant full-length CSN5 inserted into a modified GTE vector (Invitrogen). It has a GST tag that was removed during purification and was prepared from insect cells using a baculovirus expression system. Two CSN complexes, with or without CSN9, were reconstituted by incubating the purified subcomplexes and individual subunits in equimolar ratio and polished by SEC. Neddylated Cul1–Rbx1 complex was prepared as described previously (60).

XL-MS Analysis of CSN Complexes

Affinity-purified human CSN complex with or without CSN9 were cross-linked with DSSO, DHSO, or BMSO, respectively. Each CSN complex was reacted with a selected cross-linker at their optimized molar ratios (protein to linker) respectively: DSSO (1:250), BMSO

(1:400), and DHSO (1:30) (32, 35, 36). DMTMM was used to activate acidic residues for DHSO cross-linking (35). All reactions were performed for 1 h at room temperature. The resulting cross-linked proteins were digested by lys-C and trypsin. Cross-linked peptides were enriched by peptide SEC, analyzed by LC MSn , and identified through database searching, as previously described (SI Appendix, Supplemental Method) (35, 36).

PRM Targeted Quantitation of Cross-Linked Peptides

The 341 PRM targets were obtained based on highly reproducible DHSO cross-linked peptides of CSN and CSNn complexes, as summarized in Datasets S11 and S14. For targeted analysis, the mass spectrometer was operated with the following settings: No survey scan collected, tMS2 resolving power 30,000, AGC target 5e4, maximum injection time 54 ms, isolation window 1 m/z, and CID normalized collision energy of 23%. A total of 341 cross-links were monitored over 3 separate targeted analyses for each sample, along with a set of 16 heavy-labeled AQUA peptides. Targeted analysis of AQUA peptides used the same settings as cross-link ions except were subjected to higher-energy collision dissociation with normalized collision energy of 30%. Transition lists based on expected cross-link fragmentation ions were generated and quantified using Skyline v.4.2.0.19072. Once exported, extracted intensities were normalized within sample sets using relative intensities of AQUA peptides based on quantified b and y ions.

In Vitro Deneddylation Assay

A mixture containing 5 μM Nedd8-Cul1-Rbx1 and 20 nM CSN was incubated in reaction buffer of 50 mM Hepes (pH 7.5), 150 mM NaCl, and 1 mM TCEP. The reactions were carried out at room temperature and stopped by adding SDS/PAGE sample buffer at indicated time points, then analyzed by 9% SDS/PAGE and stained with Coomassie blue.

Biochemical Validation of the CSN9 Interactors

Purified components of CSN, including CSN5, CSN8, and subcomplex CSN1-2-3 and CSN4-6-7, were used for pull-down assay. His-GB1 fused CSN9 served as the bait protein. The prey samples (different combinations of CSN subunits) were mixed with His-GB1- CSN9 at molar ratio 2:1. After 10-min incubation, His Mag Sepharose Ni beads (GE Healthcare) were added into the samples and suspended by gently tapping the sample tubes for 5 min to immobilize His-GB1-CSN9 and its binding partners. Then the beads were washed with 20 mM Tris·HCl (pH 8.0), 300 mM NaCl, 20 mM imidazole five times. The beads were further eluted with 200 mM imidazole and the elution was analyzed on a 4 to 15% MiniPROTEAN TGX Gel (Bio-Rad). To identify the binding partners of CSN9, all of the purified CSN components were loaded on the same gel.

Integrative Structure Modeling

Integrative structure modeling was carried out to determine the structures of the human canonical and noncanonical CSN complexes (SI Appendix, Supplemental Method and Datasets S18 and S22). Mass spectrometry raw data have been deposited at the PRIDE Archive proteomics data repository site (dataset identifier PXD014673). All of the relevant scripts, data, and results are available at GitHub, https://salilab.org/ CSN2019. The integrative structures of CSN and CSNn are deposited at PDB-Dev (https://pdb-dev.wwpdb.org/), with ID codes PDBDEV_00000037 and PDBDEV_00000038, respectively.

Data deposition

Mass spectrometry raw data have been deposited at the PRIDE Archive proteomics data repository site (dataset identifier PXD014673). All the relevant scripts, data, and results

are available at GitHub, https://salilab.org/CSN2019. The integrative structures of CSN and CSNn are deposited at PDB-Dev, https://pdb-dev.wwpdb.org/ (PDB ID codes PDBDEV_00000037 and PDBDEV_00000038).

Supplemental Information

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1915542117/-/DCSupplemental.

Acknowledgments

We thank Prof. A. L. Burlingame and Robert Chalkley for their support of the development version of Protein Prospector. This work was supported by National Institutes of Health Grants R01GM074830 and R01GM130144 (to L.H.), and P41GM109824, R01GM083960, and S10OD021596 (to A.S.); National Science Foundation graduate research Fellowship 1650113 (to I.E.C.); and National Science Foundation Grant CHE 1807612 (to S.D.R.). N.Z. is a Howard Hughes Medical Institute Investigator.

Author contributions

C.G., I.E.C., and L.H. designed experiments; C.G. performed all XL-MS experiments and data analyses; I.E.C. performed integrative structure modeling and analysis; H.M. and N.Z. purified CSN complexes and performed biochemical validation; C.Y. performed quantitative XL-MS experiments and assisted all MS analyses; I.E. assisted on structure modeling; S.A.B. and S.D.R. synthesized cross-linking reagents; A.S. supervised structure modeling; L.H. conceived the study and directed the research; C.G., I.E.C., A.S., and L.H. wrote the manuscript with input from other authors.

References

1. N. Wei, X. W. Deng, The COP9 signalosome. *Annu Rev Cell Dev Biol* **19**, 261–86 (2003).
2. D. A. Wolf, C. Zhou, S. Wee, The COP9 signalosome: an assembly and maintenance platform for cullin ubiquitin ligases? *Nat Cell Biol* **5**, 1029–33 (2003).
3. N. Wei, G. Serino, X. W. Deng, The COP9 signalosome: more than a protease. *Trends Biochem Sci* **33**, 592–600 (2008).
4. G. A. Cope, *et al.*, Role of predicted metalloprotease motif of Jab1/Csn5 in cleavage of Nedd8 from Cul1. *Science* **298**, 608–11 (2002).
5. R. J. Deshaies, C. A. Joazeiro, RING domain E3 ubiquitin ligases. *Annu Rev Biochem* **78**, 399–434 (2009).
6. M. D. Petroski, R. J. Deshaies, Function and regulation of cullin-RING ubiquitin ligases. *Nat Rev Mol Cell Biol* **6**, 9–20 (2005).
7. G. A. Cope, R. J. Deshaies, COP9 signalosome: a multifunctional regulator of SCF and other cullin-based ubiquitin ligases. *Cell* **114**, 663–71 (2003).
8. J. E. Lee, *et al.*, The steady-state repertoire of human SCF ubiquitin ligase complexes does not require ongoing Nedd8 conjugation. *Mol Cell Proteomics* **10**, M110 006460 (2011).
9. J. R. Skaar, J. K. Pagan, M. Pagano, Mechanisms and function of substrate recruitment by F-box proteins. *Nat Rev Mol Cell Biol* **14**, 369–81 (2013).
10. L. Jia, Y. Sun, SCF E3 ubiquitin ligases as anticancer targets. *Curr Cancer Drug Targets* **11**, 347–56 (2011).

11. E. D. Emberley, R. Mosadeghi, R. J. Deshaies, Deconjugation of Nedd8 from Cul1 is directly regulated by Skp1-F-box and substrate, and the COP9 signalosome inhibits deneddylated SCF by a noncatalytic mechanism. *J Biol Chem* **287**, 29679–89 (2012).
12. R. I. Enchev, *et al.*, Structural basis for a reciprocal regulation between SCF and CSN. *Cell Rep* **2**, 616–27 (2012).
13. E. S. Fischer, *et al.*, The molecular basis of CRL4DDB2/CSA ubiquitin ligase architecture, targeting, and activation. *Cell* **147**, 1024–39 (2011).
14. M. G. Füzesi-Levi, *et al.*, CSNAP, the smallest CSN subunit, modulates proteostasis through cullin-RING ubiquitin ligases. *Cell Death & Differentiation* (2019).
15. G. A. Cope, R. J. Deshaies, Targeted silencing of Jab1/Csn5 in human cells downregulates SCF activity through reduction of F-box protein levels. *BMC Biochem* **7**, 1 (2006).
16. Y. H. Lee, *et al.*, Molecular targeting of CSN5 in human hepatocellular carcinoma: a mechanism of therapeutic response. *Oncogene* **30**, 4175–84 (2011).
17. Y. Pan, F. X. Claret, Targeting Jab1/CSN5 in nasopharyngeal carcinoma. *Cancer Lett* **326**, 155–60 (2012).
18. G. Zhong, H. Li, T. Shan, N. Zhang, CSN5 silencing inhibits invasion and arrests cell cycle progression in human colorectal cancer SW480 and LS174T cells in vitro. *Int J Clin Exp Pathol* **8**, 2809–15 (2015).
19. H. Zhang, *et al.*, COPS5 inhibition arrests the proliferation and growth of serous ovarian cancer cells via the elevation of p27 level. *Biochem Biophys Res Commun* **493**, 85–93 (2017).

20. M. H. Lee, R. Zhao, L. Phan, S. C. Yeung, Roles of COP9 signalosome in cancer. *Cell Cycle* **10**, 3057–66 (2011).
21. K. S. Richardson, W. Zundel, The emerging role of the COP9 signalosome in cancer. *Mol Cancer Res* **3**, 645–53 (2005).
22. E. S. Fischer, *et al.*, Structure of the DDB1-CRBN E3 ubiquitin ligase in complex with thalidomide. *Nature* **512**, 49–53 (2014).
23. G. M. Lingaraju, *et al.*, Crystal structure of the human COP9 signalosome. *Nature* **512**, 161–5 (2014).
24. S. Cavadini, *et al.*, Cullin-RING ubiquitin E3 ligase regulation by the COP9 signalosome. *Nature* **531**, 598–603 (2016).
25. R. Mosadeghi, *et al.*, Structural and kinetic analysis of the COP9-Signalosome activation and the cullin-RING ubiquitin ligase deneddylation cycle. *Elife* **5** (2016).
26. S. V. Faull, *et al.*, Structural basis of Cullin-2 RING E3 ligase regulation by the COP9 signalosome. *bioRxiv*, 483024 (2018).
27. S. Rozen, *et al.*, CSNAP Is a Stoichiometric Subunit of the COP9 Signalosome. *Cell Rep* **13**, 585–598 (2015).
28. M. Sharon, *et al.*, Symmetrical modularity of the COP9 signalosome complex suggests its multifunctionality. *Structure* **17**, 31–40 (2009).
29. A. Sinz, C. Arlt, D. Chorev, M. Sharon, Chemical cross-linking and native mass spectrometry: A fruitful combination for structural biology. *Protein Science* **24**, 1193–1209 (2015).

30. A. Leitner, M. Faini, F. Stengel, R. Aebersold, Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. *Trends Biochem Sci* **41**, 20–32 (2016).
31. C. Yu, L. Huang, Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology. *Anal Chem* **90**, 144–165 (2018).
32. A. Kao, *et al.*, Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. *Mol Cell Proteomics* **10**, M110.002212 (2011).
33. C. Yu, W. Kandur, A. Kao, S. Rychnovsky, L. Huang, Developing new isotope-coded mass spectrometry-cleavable cross-linkers for elucidating protein structures. *Anal Chem* **86**, 2099–106 (2014).
34. R. M. Kaake, *et al.*, A new in vivo cross-linking mass spectrometry platform to define protein-protein interactions in living cells. *Mol Cell Proteomics* **13**, 3533–43 (2014).
35. C. B. Gutierrez, *et al.*, Developing an Acidic Residue Reactive and Sulfoxide-Containing MS-Cleavable Homobifunctional Cross-Linker for Probing Protein-Protein Interactions. *Anal Chem* (2016).
36. C. B. Gutierrez, *et al.*, Development of a Novel Sulfoxide-Containing MS-Cleavable Homobifunctional Cysteine-Reactive Cross-Linker for Studying Protein-Protein Interactions. *Anal Chem* **90**, 7600–7607 (2018).
37. A. Kao, *et al.*, Mapping the structural topology of the yeast 19S proteasomal regulatory particle using chemical cross-linking and probabilistic modeling. *Mol Cell Proteomics* **11**, 1566–77 (2012).

38. X. Wang, *et al.*, Molecular Details Underlying Dynamic Structures and Regulation of the Human 26S Proteasome. *Mol Cell Proteomics* **16**, 840–854 (2017).
39. F. Liu, D. T. Rijkers, H. Post, A. J. Heck, Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nat Methods* **12**, 1179–84 (2015).
40. F. Liu, P. Lossl, B. M. Rabbitts, R. S. Balaban, A. J. R. Heck, The interactome of intact mitochondria by cross-linking mass spectrometry provides evidence for coexisting respiratory supercomplexes. *Mol Cell Proteomics* **17**, 216–232 (2018).
41. F. Herzog, *et al.*, Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. *Science* **337**, 1348–52 (2012).
42. J. P. Erzberger, *et al.*, Molecular architecture of the 40SeIF1eIF3 translation initiation complex. *Cell* **158**, 1123–35 (2014).
43. S. J. Kim, *et al.*, Integrative structure and functional anatomy of a nuclear pore complex. *Nature* **555**, 475–482 (2018).
44. M. P. Rout, A. Sali, Principles for Integrative Structural Biology Studies. *Cell* **177**, 1384–1403 (2019).
45. D. Dubiel, B. Rockel, M. Naumann, W. Dubiel, Diversity of COP9 signalosome structures and functional consequences. *FEBS Lett* **589**, 2507–13 (2015).
46. A. Leitner, T. Walzthoeni, R. Aebersold, Lysine-specific chemical cross-linking of protein complexes and identification of cross-linking sites using LC-MS/MS and the xQuest/xProphet software pipeline. *Nat Protoc* **9**, 120–37 (2014).
47. F. Alber, *et al.*, Determining the architectures of macromolecular assemblies. *Nature* **450**, 683–694 (2007).

48. D. Russel, *et al.*, Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. *PLOS Biology* **10**, e1001244 (2012).

49. A. Sali, *et al.*, Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure* **23**, 1156–67 (2015).

50. D. Schneidman-Duhovny, R. Pellarin, A. Sali, Uncertainty in Integrative Structural Modeling. *Curr Opin Struct Biol* **28**, 96–104 (2014).

51. B. Rockel, T. Schmaler, X. Huang, W. Dubiel, Electron microscopy and in vitro deneddylation reveal similar architectures and biochemistry of isolated human and Flag-mouse COP9 signalosome complexes. *Biochem Biophys Res Commun* **450**, 991–7 (2014).

52. M. Birol, *et al.*, Structural and biochemical characterization of the Cop9 signalosome CSN5/CSN6 heterodimer. *PLoS One* **9**, e105688 (2014).

53. J. D. Chavez, *et al.*, A General Method for Targeted Quantitative Cross-Linking Mass Spectrometry. *PLoS One* **11**, e0167547 (2016).

54. X. Zhang, *et al.*, Carboxylate-Selective Chemical Cross-Linkers for Mass Spectrometric Analysis of Protein Structures. *Anal Chem* **90**, 1195–1201 (2018).

55. A. Leitner, *et al.*, Chemical cross-linking/mass spectrometry targeting acidic residues in proteins and protein complexes. *Proc Natl Acad Sci U S A* **111**, 9455–60 (2014).

56. M. Heusel, *et al.*, Complex-centric proteome profiling by SEC-SWATH-MS. *Molecular Systems Biology* **15**, e8438 (2019).

57. M. Sharon, T. Taverner, X. I. Ambroggio, R. J. Deshaies, C. V. Robinson, Structural organization of the 19S proteasome lid: insights from MS of intact complexes. *PLoS Biol* **4**, e267 (2006).

58. M. Bai, *et al.*, In-depth Analysis of the Lid Subunits Assembly Mechanism in Mammals. *Biomolecules* **9**, 213 (2019).

59. A. Peth, C. Berndt, W. Henke, W. Dubiel, Downregulation of COP9 signalosome subunits differentially affects the CSN complex and target protein stability. *BMC Biochem* **8**, 27 (2007).

60. C. Yu, *et al.*, Characterization of Dynamic UbR-Proteasome Subcomplexes by In vivo Cross-linking (X) Assisted Bimolecular Tandem Affinity Purification (XBAP) and Label-free Quantitation. *Mol Cell Proteomics* (2016).

Chapter III - Thermodynamic and kinetic estimates from electron microscopy particle images

Contributing authors

Ilan E. Chemmama[1,*], Evan M. Green[2,*], Yifan Cheng[2,3], David A. Agard[2,3], and Andrej Sali[1,4]

[1]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94158, USA

[2]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA 94158, USA

[3]Howard Hughes Medical Institute, University of California, San Francisco, San Francisco, CA 94158, USA

[4]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94158, USA; California Institute of Quantitative Biosciences, University of California, San Francisco, San Francisco, CA 94158, USA.

[*]Ilan E. Chemmama and Evan M. Green contributed equally to this work.

Abstract

Mapping the functional cycle of a biomolecular machine requires determining the structures of key states, their stabilities and the kinetics of their interconversions under native conditions. Here, we describe a method that outputs these free energy differences and kinetic rates based on time-dependent negative-stain electron microscopy particle images for a given set of structural states and transitions between them. First, proportions

of these states in micrographs prepared at different lag times after triggering a process of interest are estimated by assigning each particle image to one of the states. Second, the free energy difference between each pair of connected states is estimated from the corresponding proportion and the rates of conversion between all pairs of connected states are estimated by fitting the time-dependent proportions to the system of first order reaction rate equations. In principle, different sets of transitions and reaction rate equations can be explored using model selection criteria. The accuracy of the estimates is limited by the uniformity of sampling structural states during particle picking and by the accuracy of assigning particle images to the structural states. The uncertainty of the estimates is determined by the numbers of particle images and is estimated by bootstrapping. We illustrate our method by applying it to the yeast Hsp90 chaperone engaged in its ATPase cycle in the presence of two ATP analogs, the non-hydrolysable AMP·PNP or the slowly hydrolysable ATPγS. The resulting estimates of the thermodynamic and kinetic parameters are in agreement with previous experiments. The proposed method is applicable to a large number of systems, including to those mapped by cryo-electron microscopy.

Introduction

Biological macromolecular assemblies play crucial roles in most cellular processes. Spatiotemporal models of these assemblies facilitate the understanding of their function, evolution, and modulation (1–3). Traditional structure determination methods, such as X-ray crystallography, provide mostly static structural models of these assemblies, but they are largely uninformative about thermodynamics and kinetics of transitions between states. Other methods can provide some dynamic information. For example, Förster resonance energy transfer (FRET) spectroscopy can probe the distance between dyes, with a spatial resolution of ~5 Å and temporal resolution of ~5 ms. However, numerous FRET measurements with dyes engineered at different locations on an assembly are required to infer global structural features along a functional cycle. To fully describe these assemblies, it is thus necessary to have a method for accurately, precisely, and completely determining the key states in the functional cycle, their structures and stabilities, as well as the kinetics of their interconversions under native conditions.

Single particle electron microscopy (EM) allows for imaging of non-crystalline specimens (single particles) of assemblies. EM has recently undergone a “resolution revolution” (4). The number of high-resolution EM structures has greatly increased due to improvements in sample preparation, electron microscopes, image sensors, image processing software, and molecular modeling software (5, 6). For example, direct electron detectors have replaced film and charge-coupled device (CCD) cameras due to their improved quantum efficiency (i.e., a better conversion of incident radiation to photons) (5, 7, 8). Image processing improvements are mostly due to the introduction of statistical treatment of the sample heterogeneity and experimental uncertainty (9–13). Thus, it is

now possible to image assemblies in solution at high resolution (14–18). As a result, EM has a unique opportunity to image assemblies along their functional cycle. An important role for modeling will include interpreting the single particle images in terms of time-dependent populations of multiple states along the functional cycle in solution.

Multiple structures of assemblies can be observed by EM if the differences between the structures are larger than the resolution of the images. Different multiple structures can be imaged in a sample by coupling structural changes to a process of interest, e.g. by modulation of a physical state variables (e.g., temperature) (19) or chemical composition (e.g., addition of a ligand) (20, 21). Furthermore, time-resolved EM aims at determining structures along the functional cycle (22). Time-resolved EM has been successfully used to describe (i) slow-processes such as the biogenesis of ribosomes (23) and the coupling between ribosome dynamics and tRNA movement (19) as well as (ii) sub-second processes such as the molecular mechanism of the bacterial translation initiation (24). In these experiments, both structures and population sizes agreed with prior experiments (21, 24).

Here, we show that using time-resolved EM enables determining structures of states, free energy differences between pairs of states, and kinetic rate parameters for interconversions between them. We illustrate our method by applying it to the yeast Hsp90 chaperone engaged in its ATPase cycle in the presence of two ATP analogs, the non-hydrolysable AMP·PNP or the slowly hydrolysable ATPγS. We determined the structures of the open and closed states by negative stain EM, both in agreement with previously determined structures. Then, we estimated population fractions of these two states in micrographs prepared at different lag times, from which we estimated free

energy differences and kinetic rates, which agreed with previous estimates of these parameters. We discuss the limitations of the approach as well as potential future developments, including the possibility to determine even more state variables, such as entropy differences. The approach may contribute towards accurate, precise, and complete descriptions of macromolecular free energy landscapes.

Methods

Negative stain EM sample preparation, data acquisition, and analysis

Purified *S. cerevisae* Hsp90 was mixed with either AMP·PNP or ATPγS, resulting in the final concentration of 110 nm for the Hsp90 dimer 5 mM for the nucleotide, in a buffer consisting of 20 mM HEPES pH 7.5, 150 mM KCl, and 5 mM $MgCl_2$. Samples were incubated at room temperature (~20 °C) and aliquots were taken from a common stock at the specified time points for negative staining. For the 0 second time point, no nucleotide was added. Negative staining was performed following published protocols (25). Briefly, 2.5 µL of sample was applied to glow-discharged carbon-coated copper grids and stained with 0.75% (w/v) uranyl formate. Negatively stained samples were imaged on a Tecnai T20 electron microscope (FEI Company) equipped with a LaB6 filament operating at a 200kV acceleration voltage. Images were recorded at a nominal magnification of 50,000× using a TemF816 8K × 8K CMOS camera (TVIPS GmbH) with a calibrated pixel size of 1.57 Å. Images were 2 × 2 binned yielding a final pixel size of 3.14 Å. Particles were picked and processed by cryoSPARC (26).

Determination of the estimate of free energy and kinetic parameters.

To estimate the free energy differences and kinetic parameters, we formulate a model based on rate theory. For an N-species reversible reaction:

$$S_i \rightleftharpoons S_j \text{ (for i,j = 1, …, } N) \wedge (i \neq j).$$

We describe the chemical process by a master equation:

$$\frac{d[S_i]}{dt} = \sum_{j \neq i} k_{ji}[S_j] - \sum_{i \neq j} k_{ij}[S_i]$$

with its matrix representation:

$$\begin{pmatrix} [\dot{S}_1] \\ \vdots \\ [\dot{S}_N] \end{pmatrix} = \begin{pmatrix} -\sum_{j \neq 1} k_{1j} & \cdots & k_{N1} \\ \vdots & \ddots & \vdots \\ k_{1N} & \cdots & -\sum_{j \neq N} k_{Nj} \end{pmatrix} \begin{pmatrix} [S_1] \\ \vdots \\ [S_N] \end{pmatrix}.$$

The system is thus described by a system of ordinary differential equations (ODE), with the solution of

$$[\vec{S}] = \sum_i c_i e^{-\lambda_i t} \overrightarrow{u_i}$$

where λ_i, $\overrightarrow{u_i}$, and c_i are the *i*th eigenvalue, eigenvector, and integration constant, respectively. We simultaneously fit these parameters to the population fractions for each state *i* normalized for each time point, as determined from EM images. In addition, from the obvious constraint $\sum_i N_i = 1$:

$$\frac{[S_i]}{\sum_j [S_j]} = \frac{\frac{N_1'}{\mathcal{N}_A V}}{\frac{1}{\mathcal{N}_A V}\sum_j N_j'} = \frac{N_i'}{\sum_j N_j'} = N_i,$$

where $[S_i]$ is the concentration of state *i* at some time *t*, N_i' is the particle count observed by EM, *V* is the volume of the solution, and $\mathcal{N}_A$ is the Avogadro's number.

Results

The open and closed conformations of the yeast Hsp90

We selected single particle images for the open conformation of the yeast Hsp90 and refined a single class using cryoSPARC (26) without enforcing the C2 symmetry. The resulting negative stain EM density map was determined at a nominal resolution of ~15Å (**Fig. 3.1C**). We also refined these particle images enforcing a C2 symmetry. The resulting negative stain EM density map was determined to a nominal resolution of ~15Å. We then assessed the overlap between the two maps by computing the cross-correlation coefficient between them. The overlap between the two maps had a cross-correlation coefficient of 0.99. The negative stain EM density maps were of sufficient quality to confidently dock the rigid extended conformation of Hsp90 previously determined by fitting a molecular model to a SAXS profile (27). We assessed the overlap between the model and the EM density map of the open conformation by computing the cross-correlation between them. The model-map overlap had a cross-correlation coefficient of 0.90.

Similarly, we reconstructed a closed state of the yeast Hsp90 determined by negative stain EM. The resulting negative stain EM density map was determined at a nominal resolution of ~15Å (**Fig. 3.1D**). We assessed the model-map overlap between the yeast Hsp90 closed conformation (PDB code: 2CG9 (28)) by computing a cross-correlation between them. The model-map overlap had a cross-correlation coefficient of 0.82. We conclude that both conformations observed by negative stain EM agree with previously observed structures.

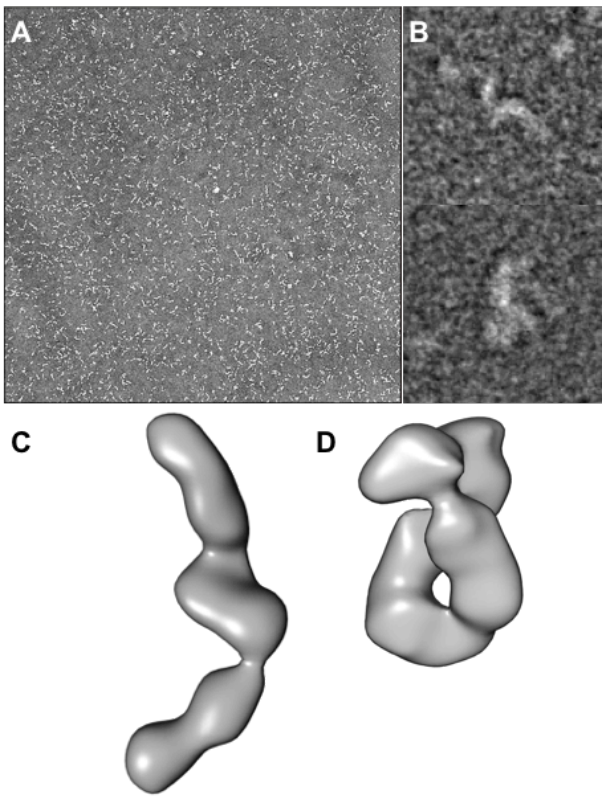

Figure 3.1 | Structure of the open and closed conformations of the yeast Hsp90
(A) Representative negative stain image of the yeast HSP90. (B) Example of particle images in an open (top) and closed (bottom) conformation used for analysis. (C) 3D reconstruction of the open conformation of the yeast Hsp90 determined by negative stain EM at a nominal resolution of ~15Å. (D) 3D reconstruction of the closed conformation of the yeast Hsp90 determined by negative stain EM at a nominal resolution of ~15Å.

The rate of conversion from open to closed state of yeast Hsp90 with AMP·PNP

In the presence of the non-hydrolysable ATP analog AMP·PNP, Hsp90 closure is an irreversible two-state process from the open to closed conformation, because negligible hydrolysis occurs on the time scale of the experiment (29). The solution of the two-state system of ODEs for an irreversible process is:

$$\begin{cases} N_o(t) = & +c_2 e^{-kt} \\ N_c(t) = c_1 & -c_2 e^{-kt} \end{cases}, \quad (1)$$

where N_o and N_c are the fractions of observed populations for the open and closed conformation, respectively, c_1 and c_2 are integration constants and k is the rate of closure of Hsp90 in the presence of AMP·PNP. Without loss of generality, we apply boundary conditions (limit as t goes to 0 and ∞) and the normalization constraint (Methods) to determine that $c_1 = 1$. We note that c_1 corresponds to the fraction of particles in a closed state at $t \to \infty$ and c_2 corresponds to the fraction of particles in the open state at time $t = 0$.

We imaged seven time points for a total of 297 micrographs and automatically picked 793,319 particles using a Gaussian sphere as a reference. We then used multiple rounds of classification (2D and 3D) to determine population sizes in the open and closed conformations (**Fig. 3.2**). We fitted simultaneously the parameters in Equation 1 to the evolution of the population fractions over time. We determined that c_2 = (0.79 ± 0.02) and k = (3.5 ± 0.3)·10^{-4} s^{-1}. Free energy differences at temperature T = 20 °C between the open and closed conformations of Hsp90 in the absence of the nucleotide can be calculated using $\Delta G_{o\to c} = -RT \ln\left[\frac{N_c}{N_o}\right] = -RT \ln\left[\frac{1-c_2}{c_2}\right]$, propagating uncertainty using $d\Delta G_{o\to c} = \sqrt{\sum_q \left|\frac{\partial \Delta G_{o\to c}}{\partial q}\right|^2 dq^2}$, where q is one of T or c_2. We measure the free energy difference to be (0.77 ± 0.05) kcal·mol^{-1}, in agreement with the fraction of populations previously observed (21).

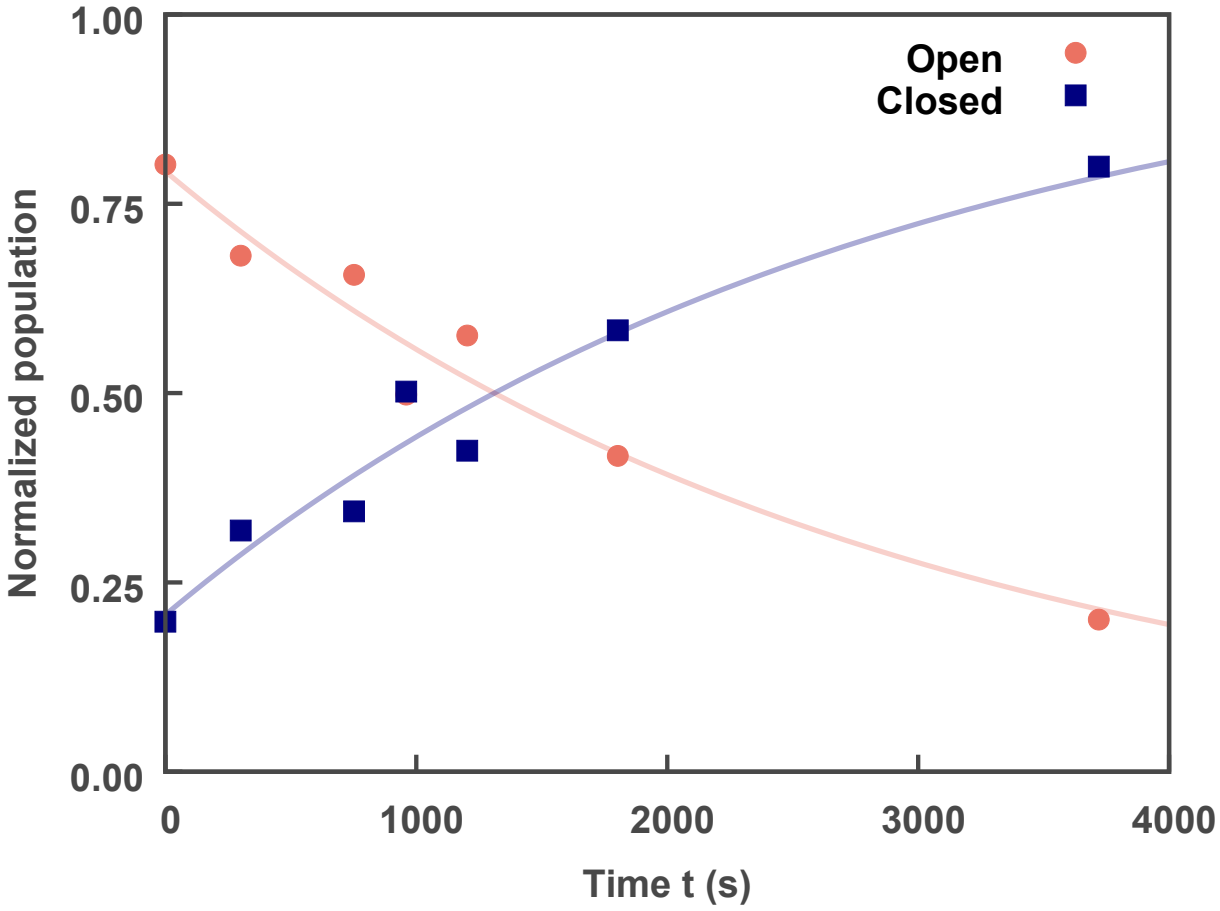

Figure 3.2 | Two-state model for the kinetics of closure of the yeast Hsp90 in the presence of AMP·PNP.
Experimentally determined kinetic traces of the irreversible process between the open and closed states. Population fractions of the open state (red filled circles) and closed state (blue filled squares) are shown as a function of time. The lines show the least-squares fit of the first order kinetics of closing to the data (opaque lines; the fitting parameters are c_2 = (0.79 ± 0.02) and k_c = (3.5 ± 0.3)·$10^{-4}$$s^{-1}$).

The kinetic rate parameter of closure of the yeast Hsp90 with ATPγS

In the presence of the slowly hydrolysable ATP analog ATPγS, Hsp90 closure and opening is a reversible two-state process. The solution for the two-state system of ODEs for a reversible process is

$$\begin{cases} N_o(t) = c_1 \dfrac{k_o}{k_c} & + c_2 e^{-(k_o+k_c)t} \\ N_c(t) = c_1 & - c_2 e^{-(k_0+k_c)t} \end{cases},$$

where N_o and N_c is the fraction of observed population for the open and closed conformations, respectively, c_1 and c_2 are integration constants, and k_o and k_c are the rates of opening and closure of Hsp90 in the presence of ATPγS, respectively. Without loss of generality, we apply boundary conditions and the normalization constraint (Methods) and determine that $c_1 = \frac{k_c}{k_o+k_c}$. The solution is

$$\begin{cases} N_o(t) = \frac{k_o}{k_o+k_c} & + c_2 e^{-(k_o+k_c)t} \\ N_c(t) = \frac{k_c}{k_o+k_c} & - c_2 e^{-(k_0+k_c)t} \end{cases}. \quad (2)$$

We imaged 13 time points, resulting in the total of 967 micrographs. We present here preliminary results for 8 of these time points. We picked 715,475 particles using a Gaussian sphere as a reference. We then used multiple rounds of classification (2D and 3D) to determine population sizes in the open and closed conformations (**Fig. 3.3**). We fitted simultaneously the parameters in Equation 2 to the evolution of the population fractions over time. We determined that c_2 = (0.34 ± 0.05), k_o = (6 ± 2)·10^{-4} s^{-1}, and k_c = (9 ± 3)·10^{-4} s^{-1}. The rate of closure is in agreement with the rate of closure observed by FRET spectroscopy (30). Similarly to above, the free energy difference at temperature T = 20 °C between the open and closed conformations of Hsp90 in the absence of ATPγS

is (0.6 ± 0.4) kcal·mol^{-1}, in agreement with a previous estimate (21). At equilibrium in the presence of ATPγS, the free energy difference between the two states is (-0.2 ± 0.3) kcal·mol^{-1}. As expected, the closed conformation is favored in the presence of the nucleotide, as previously observed (21, 29, 31, 32).

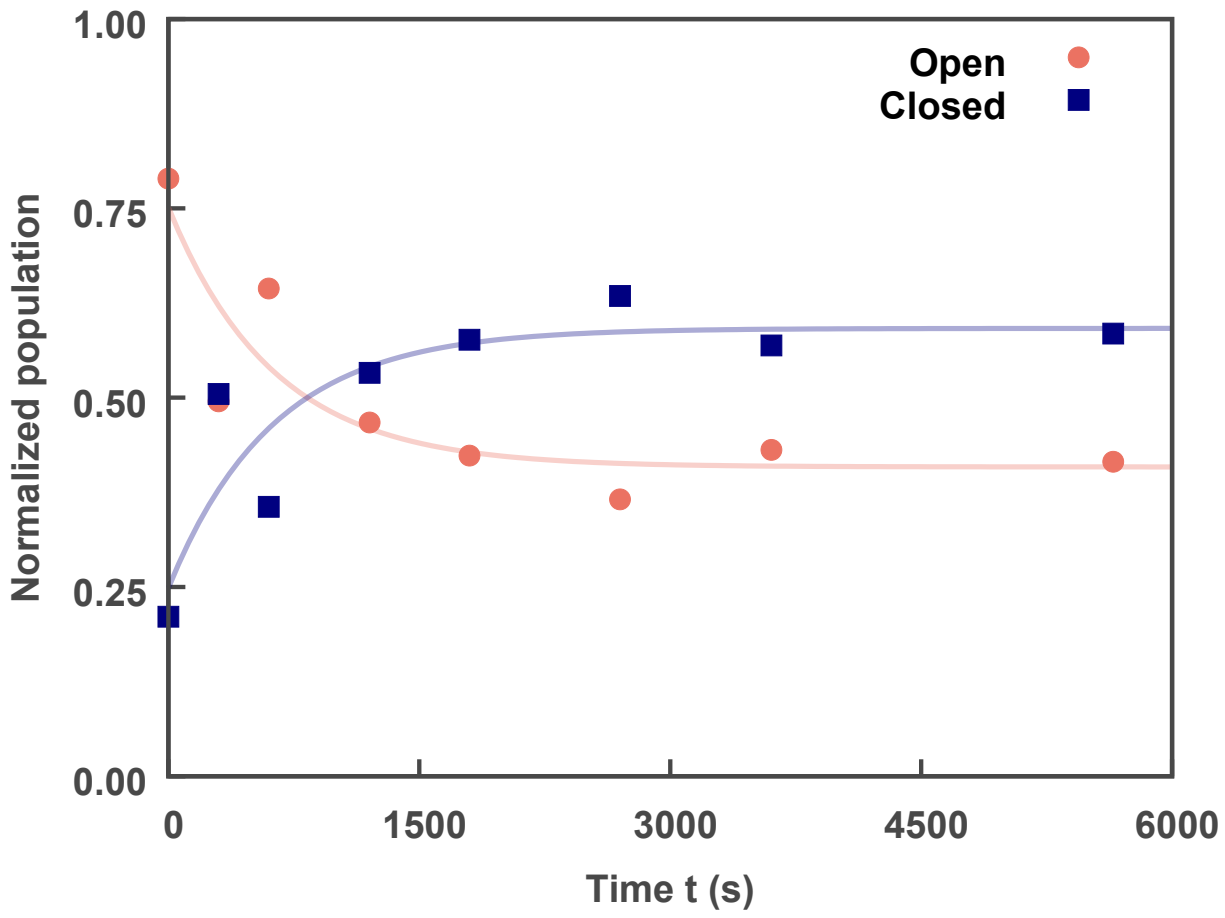

Figure 3.3 | Two-state model for the kinetics of closure of the yeast Hsp90 in the presence of ATPγS.
Experimentally determined kinetic traces of the reversible process between the open and closed states. Population fractions of the open state (red filled circles) and closed state (blue filled squares) are shown as a function of time. The lines show the least-squares fit of the first order kinetics of closing and opening to the data (opaque lines; the fitting parameters are c_2 = (0.34 ± 0.05) and k_c = (9 ± 3)·10^{-4} s^{-1}, k_o = (7 ± 2)·10^{-4} s^{-1}).

Discussion

Electron microscopy is able to determine the structures of biological macromolecular assemblies. In addition, because EM is a true single particle method, it is also possible to use it to map the kinetics and thermodynamics of a system that exists in multiple states. As an example, we used here a slow process of closing the open state of yeast Hsp90. In particular, we determined the structures of the two key states, the open and closed states, as well as the evolution of their populations as a function of time. The population sizes enabled us to compute free energy differences and kinetic parameters for the yeast Hsp90 along its functional cycle, in agreement with previous estimates (21, 29, 31, 32). The modeling of kinetic and thermodynamic parameters can also be applied to cryo-EM images, which could allow us to characterize processes involving more similar structures than possible for negative stain EM (20, 33–35). Moreover, temperature dependent experiments (19) as well as ligand-dependent experimental conditions (20, 21) can modulate the population sizes and enable, in principle, the determination of entropy differences and structures of on-pathway intermediates.

The method suffers from experimental, computational, and statistical limitations. First, the kinetics of interconversion between different states needs to be slower than the experiment setup time (e.g., pipetting, staining, or the time of plunge freezing (~10^{-7} sec)). Second, the approach is limited by the ability to classify and count particle images into their correct states; the classification can be improved by using cryo-EM for leveraging higher-resolution information. Third, the least stable state still needs to be sufficiently frequent for accurate counting; in principle, increasing the number of images will ameliorate this shortcoming, although the population frequency is an exponentially

decreasing function of thermodynamic stability. Finally, the particles in the different states need to be counted with equal bias relative to the uncounted particles. In principle, the sensitivity of the population estimates as a function of the accuracy of assigning single particle images to different states can be tested by cross-validation (e.g., bootstrapping).

The results presented here are preliminary. We are currently investigating robustness and further validation of the process model. First, we need to estimate the uncertainty of the population fractions. This estimation requires to perform bootstrapping on each of the datasets and compute expectations and deviations for each time point of each sample. Lastly, we need to test whether the representation of our process model is appropriate. For example, we need to assess whether or not the kinetic model of Hsp90 in the presence of ATPγS is a two-state system; it is possible that the data suggest additional on-pathway intermediates that have not yet been accounted for. The on-pathway intermediates may be classifiable using negative-stain EM or may require higher resolution information available by imaging Hsp90 by cryo-EM.

Acknowledgments

We thank Dr. Daniel Elnatan for kindly providing purified sample of the yeast Hsp90.

Author contributions

I.E.C., E.M.G, D.A.A, and A.S. designed experiments; E.M.G. performed the negative stain EM experiments; I.E.C. performed modeling and analysis; D.A.A. and Y.C. supervised the EM experiments; A.S. supervised modeling and analysis.

References

1. M. P. Rout, A. Sali, Principles for Integrative Structural Biology Studies. *Cell* **177**, 1384–1403 (2019).
2. F. Alber, *et al.*, Determining the architectures of macromolecular assemblies. *Nature* **450**, 683–694 (2007).
3. A. Sali, *et al.*, Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure* **23**, 1156–67 (2015).
4. W. Kühlbrandt, The Resolution Revolution. *Science* **343**, 1443–1444 (2014).
5. Y. Cheng, N. Grigorieff, P. A. Penczek, T. Walz, A primer to single-particle cryo-electron microscopy. *Cell* **161**, 438–449 (2015).
6. X. C. Bai, G. McMullan, S. H. Scheres, How cryo-EM is revolutionizing structural biology. *Trends Biochem Sci* **40**, 49–57 (2015).
7. X. Li, S. Q. Zheng, K. Egami, D. A. Agard, Y. Cheng, Influence of electron dose rate on electron counting images recorded with the K2 camera. *J Struct Biol* **184**, 251–60 (2013).
8. R. S. Ruskin, Z. Yu, N. Grigorieff, Quantitative characterization of electron detectors for transmission electron microscopy. *J. Struct. Biol.* **184**, 385–393 (2013).
9. F. J. Sigworth, A Maximum-Likelihood Approach to Single-Particle Image Refinement. *Journal of Structural Biology* **122**, 328–339 (1998).
10. S. H. W. Scheres, A Bayesian View on Cryo-EM Structure Determination. *J Mol Biol* **415**, 406–418 (2012).
11. S. H. Scheres, RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* **180**, 519–30 (2012).

12. S. H. Scheres, Processing of Structurally Heterogeneous Cryo-EM Data in RELION. *Methods Enzymol* **579**, 125–57 (2016).
13. D. Lyumkis, A. F. Brilot, D. L. Theobald, N. Grigorieff, Likelihood-based classification of cryo-EM images using FREALIGN. *J Struct Biol* **183** (2013).
14. M. Liao, E. Cao, D. Julius, Y. Cheng, Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504**, 107–112 (2013).
15. Y. Gao, E. Cao, D. Julius, Y. Cheng, TRPV1 structures in nanodiscs reveal mechanisms of ligand and lipid action. *Nature* **534**, 347–351 (2016).
16. M. G. Campbell, D. Veesler, A. Cheng, C. S. Potter, B. Carragher, 2.8 Å resolution reconstruction of the Thermoplasma acidophilum 20S proteasome using cryo-electron microscopy. *eLife* **4**.
17. A. Bartesaghi, D. Matthies, S. Banerjee, A. Merk, S. Subramaniam, Structure of β-galactosidase at 3.2-Å resolution obtained by cryo-electron microscopy. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 11709–11714 (2014).
18. A. Bartesaghi, *et al.*, 2.2 Å resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor. *Science* **348**, 1147–1151 (2015).
19. N. Fischer, A. L. Konevega, W. Wintermeyer, M. V. Rodnina, H. Stark, Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy. *Nature* **466**, 329–333 (2010).
20. S. Hofmann, *et al.*, Conformation space of a heterodimeric ABC exporter under turnover conditions. *Nature* **571**, 580–583 (2019).

21. D. R. Southworth, D. A. Agard, Species-Dependent Ensembles of Conserved Conformational States Define the Hsp90 Chaperone ATPase Cycle. *Molecular Cell* **32**, 631–640 (2008).

22. J. Frank, Time-resolved Cryo-Electron Microscopy: Recent Progress. *J Struct Biol* **200**, 303–306 (2017).

23. A. M. Mulder, *et al.*, Visualizing ribosome biogenesis: parallel assembly pathways for the 30S subunit. *Science* **330**, 673–677 (2010).

24. S. Kaledhonkar, *et al.*, Late steps in bacterial translation initiation visualized using time-resolved cryo-EM. *Nature* **570**, 400–404 (2019).

25. D. S. Booth, A. Avila-Sakar, Y. Cheng, Visualizing proteins and macromolecular complexes by negative stain EM: from grid preparation to image acquisition. *JoVE (Journal of Visualized Experiments)*, e3227 (2011).

26. A. Punjani, J. L. Rubinstein, D. J. Fleet, M. A. Brubaker, cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods* **14**, 290–296 (2017).

27. K. A. Krukenberg, F. Förster, L. M. Rice, A. Sali, D. A. Agard, Multiple Conformations of E. coli Hsp90 in Solution: Insights into the Conformational Dynamics of Hsp90. *Structure* **16**, 755–765 (2008).

28. M. M. U. Ali, *et al.*, Crystal structure of an Hsp90-nucleotide-p23/Sba1 closed chaperone complex. *Nature* **440**, 1013–1017 (2006).

29. M. Hessling, K. Richter, J. Buchner, Dissection of the ATP-induced conformational cycle of the molecular chaperone Hsp90. *Nature Structural & Molecular Biology* **16**, 287–293 (2009).

30. H. Girstmair, *et al.*, The Hsp90 isoforms from S. cerevisiae differ in structure, function and client range. *Nat Commun* **10**, 3626 (2019).
31. S. Sattin, *et al.*, Activation of Hsp90 Enzymatic Activity and Conformational Dynamics through Rationally Designed Allosteric Ligands. *Chemistry – A European Journal* **21**, 13598–13608 (2015).
32. T. Weikl, *et al.*, C-terminal regions of Hsp90 are important for trapping the nucleotide during the ATPase cycle11Edited by R. Huber. *Journal of Molecular Biology* **303**, 583–592 (2000).
33. N. Elad, *et al.*, The dynamic conformational landscape of gamma-secretase. *J Cell Sci* **128**, 589–98 (2015).
34. X. C. Bai, E. Rajendra, G. Yang, Y. Shi, S. H. Scheres, Sampling the conformational space of the catalytic subunit of human gamma-secretase. *Elife* **4** (2015).
35. A. N. Rizo, *et al.*, Structural basis for substrate gripping and translocation by the ClpB AAA+ disaggregase. *Nature Communications* **10**, 1–12 (2019).

Chapter IV - Assessing Exhaustiveness of Stochastic Sampling for Integrative Modeling of Macromolecular Structures

Contributing authors

Shruthi Viswanath[1,*], Ilan E. Chemmama[1,2,*], Peter Cimermancic[1,#], and Andrej Sali[1,2,3]

[1]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94158, USA

[2]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94158, USA

[3]Institute of Quantitative Biosciences, University of California, San Francisco, San Francisco, California, CA 94158, USA

[*]Shruthi Viswanath and Ilan E. Chemmama contributed equally to this work.

Contact: shruthi@salilab.org and sali@salilab.org

[#]Peter Cimermancic's present address is Verily Inc., South San Francisco, California

Abstract

Modeling of macromolecular structures involves structural sampling guided by a scoring function, resulting in an ensemble of good-scoring models. By necessity, the sampling is often stochastic, and must be exhaustive at a precision sufficient for accurate modeling and assessment of model uncertainty. Therefore, the very first step in analyzing the ensemble is an estimation of the highest precision at which the sampling is exhaustive. Here, we present an objective and automated method for this task. As a proxy for sampling exhaustiveness, we evaluate whether two independently and stochastically

generated sets of models are sufficiently similar. The protocol includes testing 1) convergence of the model score, 2) whether model scores for the two samples were drawn from the same parent distribution, 3) whether each structural cluster includes models from each sample proportionally to its size, and 4) whether there is sufficient structural similarity between the two model samples in each cluster. The evaluation also provides the sampling precision, defined as the smallest clustering threshold that satisfies the third, most stringent test. We validate the protocol with the aid of enumerated good-scoring models for five illustrative cases of binary protein complexes. Passing the proposed four tests is necessary, but not sufficient for thorough sampling. The protocol is general in nature and can be applied to the stochastic sampling of any set of models, not just structural models. In addition, the tests can be used to stop stochastic sampling as soon as exhaustiveness at desired precision is reached, thereby improving sampling efficiency; they may also help in selecting a model representation that is sufficiently detailed to be informative, yet also sufficiently coarse for sampling to be exhaustive.

Introduction

Integrative structure determination is an approach for characterizing the structures of large macromolecular assemblies that relies on multiple types of input information, including from varied experiments, physical theories, and statistical analysis (1–4). Therefore, it maximizes the accuracy, precision, completeness, and efficiency of structure determination. Moreover, it can often produce a structure for systems that are refractive to traditional structure determination methods (5–11), such as x-ray crystallography, electron microscopy, and NMR spectroscopy. Integrative structure determination proceeds in four stages. First, all information that describes the system of interest, including data from wet lab experiments, statistical tendencies such as atomic statistical potentials (12–14), and physical laws such as molecular mechanics force fields (15, 16), is collected. Second, a suitable representation for the system is chosen depending on the quantity and resolution of the available information. The available information is then translated into a set of spatial restraints on the components of the system. The spatial restraints are combined into a single scoring function that ranks alternative models based on their agreement with input information. Third, the alternative models are sampled using a variety of techniques, such as conjugate gradients, molecular dynamics, Monte Carlo (17), and divide-and-conquer message passing methods (18). The sampling generates an ensemble of models that are as consistent with the input information as possible. Finally, input information and output structures need to be analyzed to estimate structure precision and accuracy, detect inconsistent and missing information, and suggest more informative future experiments. Assessment begins with structural clustering of the modeled structures produced by sampling, followed by assessment of the thoroughness

of structural sampling, estimating structure precision based on variability in the ensemble of good-scoring structures, quantification of the structure fit to the input in-formation, structure assessment by cross-validation, and structure assessment by data not used to compute it. Integrative modeling can iterate through these four stages until a satisfactory model is built.

A key challenge in integrative modeling of biomolecular structures is to map the complete ensemble of models consistent with the input information (good-scoring models) (1, 2, 19, 20). The variation among the models in this ensemble quantifies the uncertainty of modeling (model precision). Because sampling large macromolecular systems is often necessarily stochastic, we can only aim to find representative good-scoring models. These representative models sample all good-scoring models at some precision, which we define as the sampling precision. Clearly, the sampling precision imposes a lower limit on the model precision. Therefore, exhaustive sampling of good-scoring models is a prerequisite for accurate modeling and assessment of model precision. Sampling is exhaustive at a certain precision when it generates all sufficiently good-scoring models at this precision. Importantly, sampling exhaustive-ness and sampling precision are invariably intertwined. There is always a precision at which any sampling is exhaustive; for example, even a single structure provides an exhaustive sample at a precision worse than the scale of the system.

Accurate estimation of model precision is key in assessing an integrative structure. It is perhaps more important to assess the precision of a model than to compute a model in the first place. The reason is that the utility of a model is determined significantly by its precision. First, model precision provides an estimate of the aggregate uncertainty in the

input information; second, it likely provides the lower bound on model accuracy; finally, applications of models strongly depend on their accuracy, with different applications having varied requirements for accuracy and precision (20–22). Further, only when model precision is estimated accurately, can the model be used to inform future choices, such as whether to gather more data, change the system representation, scoring functions, or sampling algorithms. Commonly used structural features for estimating model precision include the particle positions, distances, and contacts (5, 6, 23, 24), although specific systems may benefit from the use of derived features, such as the distance to a membrane in a transmembrane assembly. Of particular interest are the features that have a single maximum in their probability distribution. The spread around the maximum describes how precisely the feature was determined by the input information.

Sampling convergence in Monte Carlo simulations for protein and RNA structure prediction has been assessed by checking for abundance of structures close to the lowest energy structure(s) (25–32). Convergence in molecular dynamics (MD) simulations has been measured by counting the number of structural clusters (33–35) and their relative populations (36–40), cosine of the principal components (41), distance between the free energy surfaces of different parts of the simulation (42), and drift in the free energies (43). Some methods assess convergence in MD simulations by comparing different trajectories via a difference in populations for each cluster (36–40). For example, models from a ‘‘reference’’ simulation are first clustered based on a predetermined cutoff (38), followed by assigning models from additional simulation to the nearest cluster in the reference simulation; thus, each simulation produces a histogram of populations of clusters that enables comparison of any two simulations.

As mentioned above, testing for sampling exhaustiveness is the first step of the analysis and validation stage of our four-stage integrative modeling process, immediately following the sampling stage (2, 4, 7–9, 44). Here, we present an objective and automated protocol that aims to estimate the precision at which sampling is exhaustive, thereby assessing sampling exhaustiveness for integrative structural modeling. As a proxy for assessing sampling exhaustive-ness, we evaluate whether or not two independently and stochastically generated sets of models (model samples) are sufficiently similar. Model samples for assessment can be obtained, for example, from two independent simulations using random starting models or different random number generator seeds. The protocol for evaluating exhaustiveness includes two tests that consider the model scores, followed by two tests that consider the model structures.

There are at least two major limitations of our approach. First, sampling convergence is at best an approximation of sampling exhaustiveness. Although similarity between independent model samples does indicate sampling convergence, we can only hypothesize that the convergence of stochastic sampling at some precision also indicates sampling exhaustiveness at that precision, for scoring function landscapes like those used in integrative structure modeling (many dimensions, rugged, few major minima). This hypothesis is supported by all five examined cases of binary docking solutions enumerated at a specified precision. Accordingly, passing the proposed tests is a necessary, but not sufficient condition for exhaustive sampling; a positive outcome of the test may be misleading if, for example, the landscape contains only a narrow, and thus difficult to find, pathway to the pronounced minimum corresponding to the native state. Second, our tests are also not applicable to methods whose sampling is not stochastic

(e.g., a conjugate gradients minimization from a fixed unique starting point) or so expensive that they cannot generate a large enough sample of independent models. The rest of the article is organized as follows. In Methods, we describe the four-part protocol for estimating sampling precision and assessing sampling exhaustiveness, including its application to five illustrative cases of binary protein complexes. In Results, we demonstrate the protocol on the illustrative cases and validate it by comparing stochastic model samples with models from exhaustive enumeration using rigid docking (45, 46). Parameters of the protocol, its applicability and uses, its shortcomings, the relationship between various kinds of precision in integrative modeling, relation to prior work, and future work are addressed in Discussion.

Methods

The protocol for estimating sampling precision and assessing sampling exhaustiveness (**Fig. 4.1**) consists of four tests that are increasingly stringent; each test needs to be passed before it makes sense to proceed to the next test. Given two model samples and their scores as input, the tests check 1) convergence of the model score, 2) whether model scores for the two model samples were drawn from the same parent distribution, 3) whether each structural cluster includes models from each sample proportionally to its size, and 4) whether there is sufficient structural similarity between the two model samples in each cluster. Next, each step in the flowchart (**Fig. 4.1**) is described in turn.

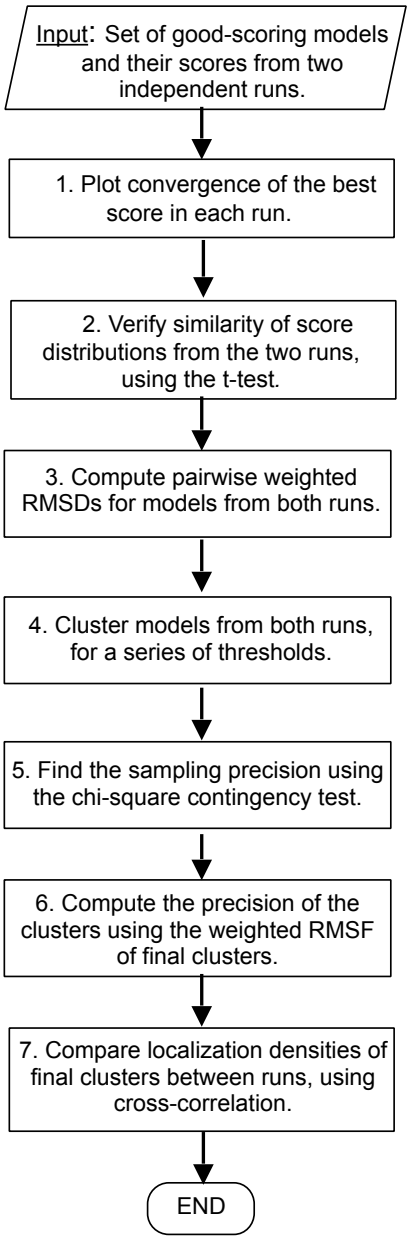

Figure 4.1 | Flowchart of the protocol for estimating sampling precision and assessing sampling exhaustiveness

Generating inputs for the protocol

The necessary input for the protocol is two model samples of approximately equal size and their scores. Each model sample consists of random, independently generated models. Both model samples must be generated using the same sampling method. In

integrative structure modeling, we are generally not interested in all sampled models, but only in models that are good-scoring (i.e., those that are sufficiently consistent with input information) (2, 4, 7–9, 44).

The precise definition of good-scoring models is left to the user and can be application-dependent. Example choices include all models scoring better than a threshold on the total score (2, 4, 7–9, 44), or all models satisfying all types of input information within acceptable thresholds. For example, if a protein complex is modeled by fitting its components into an electron microscopy density map subject to cross-linking, excluded volume, and sequence connectivity, the corresponding scoring function can be a sum of the correlation coefficient between the EM map and a model as well as harmonic (Gaussian) restraints for chemical cross-links, pairs of overlapping atoms, and sequence connectivity; a good-scoring model can then be defined as a model that fits the EM density with a cross-correlation > 0.80 and violates (e.g., a restraint value > 2 SD from the mean) a smaller number of harmonic restraints than expected for the corresponding Gaussian distributions (e.g., 5%).

The samples are usually generated by a stochastic sampling algorithm. One such algorithm is the Metropolis Monte Carlo scheme (47) that starts from a different configuration and/or random number seed for each run. For our purposes, a larger number of shorter runs is preferable over a smaller number of longer runs for two reasons. First, a larger number of runs benefits more easily from parallel execution than a smaller number of runs. Second, independent runs are guaranteed to result in uncorrelated models, whereas, additional care is needed to ensure the lack of correlation for models from a single run.

Convergence of the best score

The first test assesses whether the best model score continues to improve as more models are sampled. This test operates on random subsets of the model scores of the two samples combined. Model score subsets of several sizes (e.g., 20, 40, 60, 80%, and the complete set) are each created several times (replicates). The best score in each subset is averaged across the replicates. Plotting the average best score for each model subset size shows whether the best score converges as the number of models is increased.

Similarity of scores

The second test confirms that good-scoring models in the two model samples have similar score distributions (i.e., satisfy the data equally well). Specifically, the nonparametric Kolmogorov-Smirnov two-sample test (48, 49) tests the null hypothesis that the distributions of model scores in the two model samples were drawn from the same parent distribution. The *p* value from the Kolmogorov-Smirnov two-sample test is a measure of the statistical significance of the difference between the two distributions. A *p* value lower than the cutoff of significance (usually 0.05) indicates that the difference in the two score distributions is statistically significant.

Even a tiny difference between two distributions can be significant if the samples are large (50, 51). Therefore, we additionally use an effect size measure for the Kolmogorov-Smirnov two-sample test. Conveniently, the Kolmogorov-Smirnov two-sample test statistic, *D*, is itself a proportion (48, 49). The proportion ranges from 0 to 1, where 0 represents no difference between the two samples and 1 no overlap between

the two samples. A value of 0.30 (medium effect size) or higher suggests that the two score distributions are different (48, 49).

Finally, we conclude that the score distributions are similar if the difference is not statistically significant (p value > 0.05) or if the difference is significant (p value < 0.05) but its magnitude is small (D < 0.30).

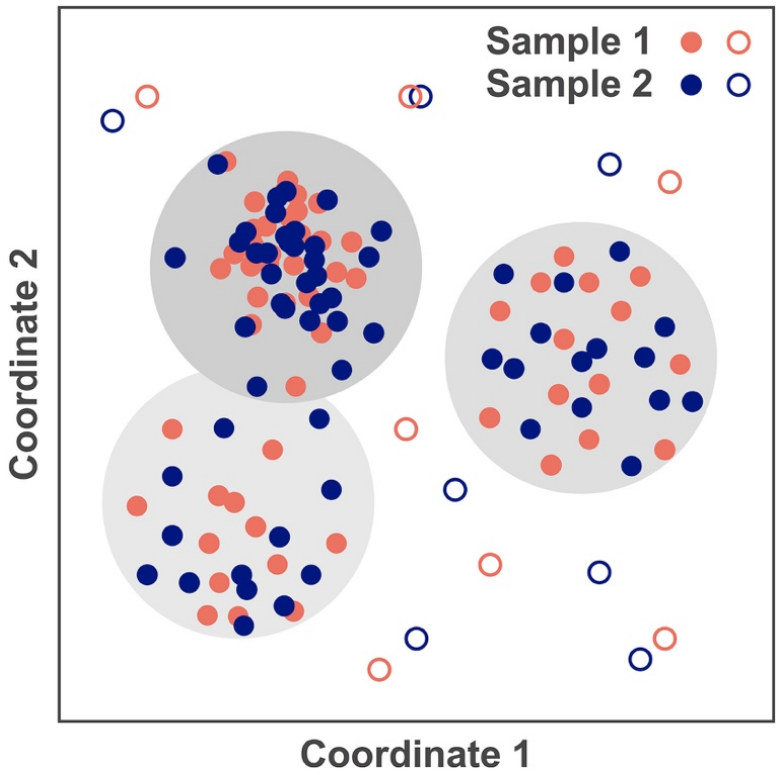

Figure 4.2 | Conceptual representation of the χ2 test for sampling exhaustiveness, showing models in a 2D coordinate space.
Two independent equal-sized random samples of good-scoring models are shown in red and blue. Models in the two samples are clustered together. The gray circles indicate cluster boundaries and the gray-scale indicates the density of models in the cluster. The size of the circles indicates the clustering threshold. The test assesses whether the proportion of models from the two samples (red and blue) is similar in each significant cluster. Note that some models are shown as open circles, indicating that these models belong to insignificant clusters (i.e., small clusters containing few models).

Computing pairwise root-mean-square deviations

The third test assesses whether models from each sample are present in each structural cluster proportionally to the sample size; when the sample sizes are equal, each cluster should contain approximately the same number of models from each sample. The test requires clustering models from both samples combined. It may be necessary to select sufficiently small random subsets of the two model samples, to make clustering computationally feasible.

The first step of clustering is to compute root-mean-square deviation (RMSD) values between all pairs of models from both samples combined (8):

$$RMSD_{i,j} = \left(\sum_{k=1}^{b} n_k \left(\vec{x}_{i,k} - \vec{x}_{j,k} \right)^2 / \sum_{k=1}^{b} n_k \right)^{\frac{1}{2}}$$

where $\vec{x}_{i,k}$ is the Cartesian coordinate of the *k*th of *b* beads in model *i*, n_k is the number of residues in bead k, and n is the total number of models; other structural dissimilarity or similarity measure may be used.

Finding the sampling precision

A stochastic sampling method does not enumerate all good-scoring models, but generates only a sample of them. Here, the sampling precision is defined as the radius of the clusters in the finest clustering for which each sample contributes models proportionally to its size (considering both significance and magnitude of the difference) and for which a sufficient proportion of all models occur in sufficiently large clusters (**Fig. 4.2**).

Clustering models using several thresholds

To find the sampling precision, we evaluate increasingly coarser clusterings, obtained using the following threshold-based clustering method (33). For each model, we first find all neighboring models, defined as models whose RMSD distance (above) from the model is less than the input threshold. Initially all models are unclustered. The unclustered model with the maximum number of neighbors and its neighbors are added to form a new cluster, and the list of unclustered models is updated. The last step is repeated until no unclustered models remain. This clustering is performed for all thresholds sampling the interval between the minimum and maximum RMSDs in steps of 2.5 Å. The next three paragraphs describe the three criteria evaluated for each clustering. *Significance*. To assess the significance of the difference between the proportions of each sample in the clusters, we use the χ^2 test for homogeneity of proportions (52). This test evaluates the null hypothesis that the two model samples are distributed nearly equally (for equal-sized samples) or approximately in proportion to their sizes (for unequal sized samples) in all major clusters. The *p* value from the test is a measure of the statistical significance of proportionate contributions to clusters from both samples. A *p* value lower than the cutoff of significance (usually 0.05) indicates that the difference in the two distributions is statistically significant.

Magnitude. To assess the magnitude of the difference between the proportions of each sample in the clusters, we use an effect size measure for the χ^2 test, Cramer's *V* (53). This test measures the magnitude of the difference between the distributions of the two samples across clusters. Cramer's *V* is defined as $\sqrt{\chi^2/n}$, where χ^2 is the χ^2 test metric and *n* is the total number of models in both samples. A value of *V* of at least 0.1 suggests

that the difference between the two distributions is large; it corresponds approximately to a *p* value of 0.05 for the case of two clusters and 500 models per sample.

Population. The calculation of the p value and Cramer's V requires that each sample has at least 10 expected models per cluster (54). Therefore, we remove all clusters containing < 10 models from either sample. To allow us to proceed with the assessment, we also require that at least 80% of the models remain after this removal.

Computing precision of clusters

For defining clusters and visualization, any threshold equal to or worse than the sampling precision can be chosen. The sampling precision is the smallest clustering threshold at which sampling is exhaustive; choosing a larger threshold will result in fewer, larger clusters, and may be preferable for analysis and/or visualization.

Although the sampling precision limits the maximum radius of a cluster (**Fig. 4.2**), models could be more tightly distributed inside a cluster. To quantify the actual spread of models in clusters, we define the cluster precision as the weighted root-mean-square fluctuation (RMSF) of all models in the cluster. Weighted RMSF accounts for differing sizes of beads often used to represent integrative structures (8, 55). It is computed using

$$\langle RMSF^2 \rangle^{1/2} = \left[\left(\sum_{n=1}^{b} n_k \sum_{i=1}^{n} \left(\vec{x}_{i,k} - \langle \vec{x}_{i,k} \rangle \right)^2 \right) / n \left(\sum_{k=1}^{b} n_k \right) \right]^{1/2}.$$

The cluster precision is ~1.4 times the sampling precision, reflecting the general relationship between RMSD and RMSF (8, 55).

Computing localization densities and their cross correlation

The final test involves computing the cross correlation between the model densities from each sample, for each cluster. The density maps are created at a resolution equal to the

threshold used for defining clusters (above). The cross-correlation coefficients between the maps are calculated using the software UCSF Chimera (56).

Validation of the protocol

We illustrate our protocol by relying on five binary protein complexes of known structure from the ZDOCK Benchmark 4.0 (57), spanning a range of docking difficulty, and 5–7 simulated distance restraints per complex. We modeled the structure of each complex by stochastic sampling as implemented in an integrative modeling platform (IMP; Supporting Material). We assessed the sampling exhaustiveness protocol based on a comparison of stochastic sampling with exhaustive enumeration, as follows.

The quality of the sampling exhaustiveness protocol is quantified by the fraction of good-scoring models from exhaustive enumeration (below) that are located within any sufficiently large cluster of the good-scoring models from the tested sampling, for the clustering threshold equal to the tested sampling precision; an enumerated model is located in a cluster, if its distance to the cluster center is within the tested sampling precision.

Fast-Fourier transform-based protein docking algorithms (45, 58–60) efficiently construct models of binary protein complexes by enumerating all possible rigid rotations and translations on a uniform 3D grid. The set of all models (57) produced by 1.2 Å and 6º uniform sampling on an FFT grid was used. Good-scoring models from enumeration were identified as in stochastic sampling (models for which at least 90% of cross-links span a Cα-Cα distance of < 12Å). For each good-scoring ZDOCK model, its distance to the nearest major cluster center from IMP was calculated.

The distribution of models from stochastic sampling in IMP cannot be compared directly to enumerated models computed by ZDOCK. The ZDOCK models are enumerated on a uniform grid, whereas IMP samples the posterior probability of models and therefore produces a nonuniform model distribution. In addition, ZDOCK and IMP use different representations (atomic and coarse-grained, respectively).

Results

We demonstrate the sampling exhaustiveness protocol on an example from the Protein Data Bank (PDB), 1AVX. The remaining four examples are described in Figs. S1–S5. There are 3369 good-scoring models for PDB: 1AVX (1896 in sample 1 and 1473 in sample 2). The score convergence test shows that the best score does not continue to improve significantly with an increase in the number of models sampled (**Fig. 4.3A**; to visualize the relatively rapid convergence in model scores, see Fig. S6). The two score distributions are similar to each other, as shown by the overlap in the score histograms and the insignificant *p* value and small *D* value from the Kolmogorov-Smirnov two-sample test (**Fig. 4.3B**). Next, exhaustiveness is examined at varying thresholds between the minimum and maximum RMSDs of 0.43 and 42.93 Å (**Fig. 4.3C**; **Table 4.1**). Based on the *p* value, Cramer's *V*, and the population of models in the contingency table, the χ^2 test is satisfied from the threshold of 12.93 Å onwards (**Table 1**). Hence, the sampling precision is 12.93 Å. In general, stricter (smaller) clustering thresholds result in many small clusters, which are ignored (**Table 1**, last column; **Fig. 4.3C**). In contrast, more lenient (larger) clustering thresholds result in fewer, larger clusters that are more likely to be retained in the analysis. For example, for the lowest clustering threshold of 0.43 Å,

each model is in its own cluster and hence all clusters are small and eliminated from the contingency table. In contrast, for thresholds > 25.43 Å (**Fig. 4.3C**), only one cluster containing all models remains. Finally, we chose the sampling precision as the clustering threshold for visualizing clusters. Inspection of the cluster populations (**Fig. 4.3D**) shows that they are similar for the two samples. The sampling precision is ~1.4 times the cluster precision, as expected from the general relationship between RMSD and RMSF (Methods; (55)). The agreement between the localization densities for samples 1 and 2 (**Fig. 4.3**, **E** and **F**) is demonstrated by the high cross-correlation coefficient of 0.99 for each cluster.

Validation by comparison to exhaustive enumeration

The sampling exhaustiveness protocol was validated by showing that 99.2% of the good-scoring ZDOCK models were within an IMP cluster for PDB: 1AVX (**Fig. 4.4**); the corresponding fraction was 100% for the other four examples (Fig. S5). For PDB: 1AVX, out of 510 good-scoring ZDOCK models, 506 were within the sampling precision of the center of a significant cluster and the distances for the other four models were less than one grid spacing further away (**Fig. 4.4**). Similarly, the largest distances between good-scoring ZDOCK and IMP models were 1.52, 3.96, 1.56, and 0.96 Å short of their sampling precisions for PDB: 1I2M, 1SYX, 2IDO, and 7CEI, respectively (Fig. S5). In conclusion, the sampling exhaustiveness protocol neither overestimates nor underestimates the sampling precision, for the five examined cases.

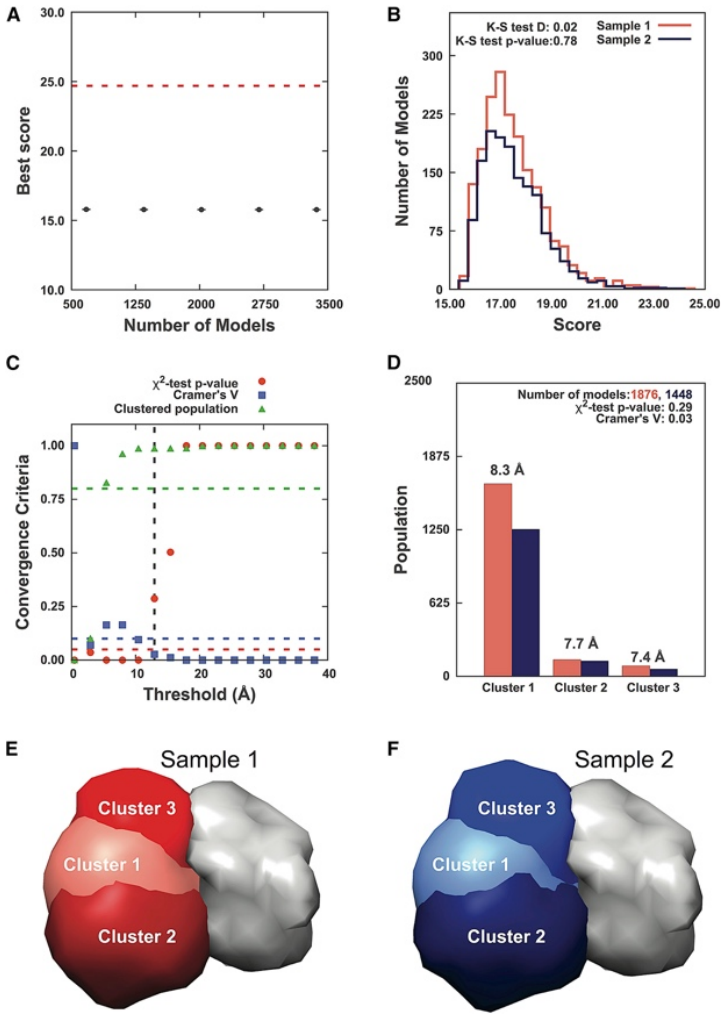

Figure 4.3 | Results for sampling exhaustiveness protocol for PDB: 1AVX
(A) Shown here are results of test 1, convergence of the model score, for the 3369 good-scoring models; the scores do not continue to improve as more models are computed essentially independently. The error bar represents the SD of the best scores, estimated by repeating sampling of models 10 times. The red dotted line indicates a lower bound on the total score. (B) Shown here are results of test 2, testing similarity of model score distributions between samples 1 (red) and 2 (blue); the difference in distribution of scores is not significant (Kolmogorov-Smirnov two-sample test p value > 0.05)and the magnitude of the difference is small (the Kolmogorov-Smirnov two-sample test statistic D is 0.02); thus, the two score distributions are effectively equal. (C) Shown here are results of test 3, containing three criteria for determining the sampling precision (y axis), evaluated as a function of the RMSD clustering threshold (33) (x axis). First, the p value is computed using thec2testfor homogeneity of proportions (52) (red dots). Second, an effect size for the $\chi 2$ test is quantified by the Cramer's V value (blue squares). Third, the population of models in sufficiently large clusters (containing at least 10 models from each sample) is shown as green triangles. The vertical dotted gray line indicates the RMSD clustering threshold at which three conditions are satisfied (p value > 0.05 (dotted red line), Cramer's V < 0.10 (dotted blue line), and the population of clustered models > 0.80 (dotted green

line)),thus defining the sampling precision of 12.93 Å. (D) Populations of sample 1 and 2 models in the clusters are obtained by threshold-based clustering using the RMSD threshold of 12.93 Å. Cluster precision is shown for each cluster. (E and F) Shown here are results of test 4: comparison of localization probability densities of models from sample 1 (red) and sample 2 (blue) in each cluster. The density map of the receptor, which is kept fixed through the simulation, is shown in gray. All densities were visualized at a threshold equal to one-third the maximum. The cross-correlation of the density maps of the two samples is 0.99 for each of the three clusters.

Table 4.1 | Three criteria for determining the sampling precision for PDB: 1AVX, evaluated as a function of the clustering threshold

The three criteria are 1) *p* values and 2) Cramer's *V*, both from the χ^2 test; and 3) the population of models in the contingency table after eliminating small clusters.

Threshold (in Å)	*p* value	Cramer's *V*	Population of models in contingency table (%)
0.4	0.0	1.0	0.0
2.9	0.0	0.1	10.1
5.4	0.0	0.2	82.7
7.9	0.0	0.2	86.2
10.4	0.0	0.1	98.7
12.9	0.3	0.0	98.7
15.4	0.5	0.0	98.7
17.9	1.0	0.0	98.8
20.4	1.0	0.0	99.8
22.9	1.0	0.0	100.0
25.4	1.0	0.0	100.0
27.9	1.0	0.0	100.0

Threshold (in Å)	*p* value	Cramer's *V*	Population of models in contingency table (%)
30.4	1.0	0.0	100.0
32.9	1.0	0.0	100.0
35.4	1.0	0.0	100.0
37.9	1.0	0.0	100.0
40.4	1.0	0.0	100.0
42.9	1.0	0.0	100.0

Discussion

Summary of the protocol

Accurate assessment of model uncertainty in integrative modeling necessitates that sampling is exhaustive at a precision sufficient for assessing model uncertainty. In this article, we introduce a protocol for determining the sampling precision of integrative structural models computed by a stochastic sampling algorithm. The protocol requires two samples of independently and stochastically generated sets of models and their scores. It includes two tests for convergence of the score and two tests for convergence of the structures. The tests for score convergence assess whether the scores in the two samples are from similar distributions. The tests for structural convergence rely on structural clustering of the models, followed by assessing whether the models in the two samples are distributed similarly across the clusters. The five illustrative cases

demonstrate the relative accuracy of the sampling exhaustiveness protocol (Figs. 3 and 4; Figs. S1–S5). Below, we discuss the parameters used in the protocol, and its applicability, shortcomings, and relationship among various kinds of precision in integrative modeling; we then address overfitting in integrative modeling, relation to prior work, and future work.

Parameters

All parameters used in the protocol are listed next; their values are chosen based on rules-of-thumb in statistics literature. First, the significance cutoff for the KS test is 0.05 and the magnitude cutoff for the KS statistic, *D*, is 0.3, the latter corresponding to medium effect size . Second, due to the inability of the χ^2 test to handle small expected cell counts, we eliminate clusters with 80%. Finally, the significance cutoff for the χ^2 test is 0.05 and the magnitude cutoff on Cramer's *V* is 0.1, the latter corresponding approximately to a *p* value of 0.05 for the case of two clusters and 500 models per sample.

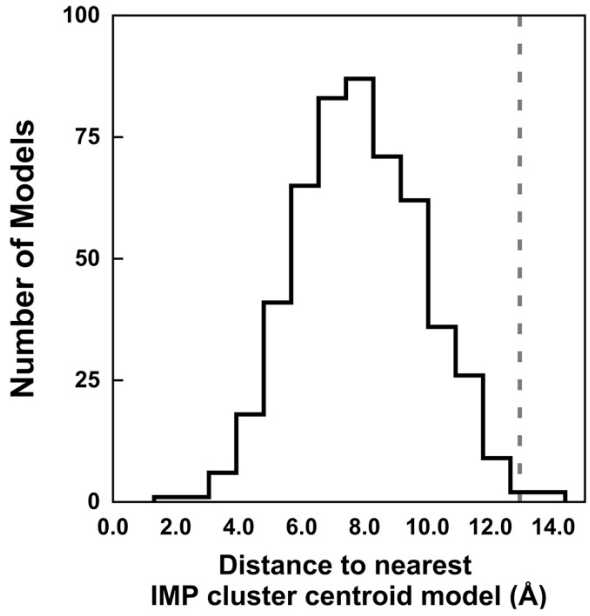

Figure 4.4 | Histogram showing the distribution of distance (measured by weighted ligand RMSD) of a good-scoring PDB: 1AVX model from enumeration (ZDOCK) to the nearest cluster centroid model from stochastic sampling (IMP)
The dotted line indicates the sampling precision for the IMP model sample determined by the sampling exhaustiveness protocol.

Applicability and uses

The sampling exhaustiveness protocol is broadly applicable to a range of sampling methods, a range of clustering or binning methods, features of models other than model scores, and models other than macromolecular structures, and it can be used dynamically during sampling to stop as soon as desired sampling precision is reached, as follows.

First, any stochastic sampling method that generates a large number of independent model samples is appropriate. Metropolis Monte Carlo sampling can satisfy this requirement of independence by 1) sampling models from multiple independent trajectories (e.g., starting from different random initial configurations) and 2) sampling models at sufficiently distant intervals on a single trajectory, such that samples are effectively uncorrelated with each other. The sampling exhaustiveness protocol can only

compare model samples produced by the same sampling algorithm (e.g., samples from uniform sampling and importance sampling are clearly not directly comparable).

Second, any one of the variety of clustering or binning methods for grouping models based on their similarity could be used instead of the distance threshold-based clustering. In principle, even a uniform grid could be applied to bin the models. This clustering is used as a relatively rapid method to assign most models to a relatively small number of groups of similar precision. As a result, we can easily quantify the sampling precision across the entire space of models and convey the results in terms of a small number of model clusters. In contrast, for example, k-means clustering generally results in clusters of varying precision, thus obfuscating the relationship between the cluster precision and sampling precision.

Third, any quantity of interest, such as radius of gyration and distance to a membrane, can be tested in the same manner as the model scores here.

Fourth, the protocol is applicable to stochastic sampling of any kind of a model, not just a structural model.

Fifth, and finally, the protocol can be applied to estimate sampling precision dynamically during a simulation, so that sampling is stopped as soon as desired sampling precision is reached, maximizing sampling efficiency. Assessment of exhaustiveness is particularly important for modeling large systems with many degrees of freedom, where exhaustive sampling of representative good-scoring solutions is particularly difficult.

Critique

In the absence of enumeration, exhaustiveness of stochastic sampling cannot be proved with complete certainty. Therefore, we suggest that even a statistical test such as the one

proposed here is better than no test. As a proxy for assessing exhaustiveness, our protocol evaluates whether two independent random model samples are similar to each other (Introduction). Our tests are not applicable to methods that do not generate independent random samples (e.g., a conjugate gradients minimization from a fixed unique starting point), or are so expensive that they cannot generate a large enough sample of independent models. Further, passing the proposed tests is a necessary, but not sufficient, condition for exhaustive sampling; a positive outcome of the test may be misleading if, for example, the landscape contains only a narrow, and thus difficult to find, pathway to the pronounced minimum corresponding to the native state. Nevertheless, based on the five examples, we argue that convergence of stochastic sampling at some precision often also indicates sampling exhaustiveness at that precision.

Precision in integrative modeling

In this article, we used the model (ensemble) precision, sampling precision, and cluster precision. In addition, the data precision (uncertainty) reflects the experimental noise (systematic and random error) (4); and the representation precision can be defined, for example, by the diameter of the largest primitive (Gaussian, bead) used to represent the system. We now discuss these five precisions in the context of each other.

First, the sampling precision imposes a lower limit on the model precision. The shape of the scoring function landscape at precisions better than the sampling precision is not sampled accurately by definition; thus, any features of the model landscape more precise than the sampling precision are unlikely to be estimated accurately.

Second, because the model ensemble is divided into one or more clusters, the model precision is always equal to or worse than any cluster precision.

Third, for the final description of the model ensemble, it only makes sense to cluster the models using a clustering threshold that is equal to or larger than the sampling precision (due to the first point above; see **Fig. 4.4**).

Fourth, and lastly, the sampling precision is in turn limited by the representation precision and data precisions. Although the model, sampling, and cluster precisions, as defined here, are directly comparable to each other, the representation and data precisions are defined on different scales. Nevertheless, qualitatively speaking, the sampling precision cannot be significantly higher than the representation and data precisions; moreover, it is likely not beneficial to use a representation with a precision that is significantly higher than the data precision.

Addressing overfitting in integrative structure modeling

Overinterpretation of the data (overfitting) is a frequent concern in any modeling. For example, a single high-resolution atomic model may fit an EM density map at intermediate resolution well; proposing such a model as the solution is often a case of overfitting because there are likely many other atomic models that also fit the data equally well. Our sampling exhaustiveness test provides a potential insurance against overfitting. When a test is passed, overfitting is not a problem (at the sampling precision) because all models (at this precision) that are consistent with the data are provided in the output model ensemble.

Relation to prior work

The methods most related to that in this article, applied in the context of MD simulations, are those in (36–39) (also used in (40)). In (36, 37), models from multiple MD simulations are combined and compared in terms of their relative populations. In (38), a new

simulation is compared against a reference simulation, by clustering models from the reference simulation based on a predetermined cutoff. The models of the new simulation are then assigned to the nearest cluster from the reference simulation. Thus, each simulation produces a histogram of populations across clusters and any two simulations can be compared by the difference in their populations for each cluster. In (39), this method is expanded by computing the number of independent samples in an MD trajectory as a way of assessing the sampling quality. The number of independent samples in an MD simulation is determined by comparing the observed variance in the population of a cluster to the expected analytical variance from an independent and identically distributed sample, for various subsample sizes.

Our protocol additionally determines the significance and magnitude of the difference in population distributions across clusters, using the χ^2 test. More importantly, our protocol also determines the sampling precision objectively, by applying the χ^2 test for a range of clustering thresholds (Figs. 3 and 4). Moreover, we test both score convergence and convergence of structural coordinates (**Fig. 4.1**). A few minor differences exist in our respective clustering methods as well: 1) similarly to (36, 37), we cluster models from all simulations, potentially producing a more comprehensive set of clusters, in contrast to clustering only models from the reference simulation (38, 39); and 2) our cluster centers are chosen based on the density of models close to the cluster center, in contrast to choosing cluster centers randomly (38), choosing clusters of uniform probability (39), or choosing cluster centers based on average linkage with a similarity cutoff (36, 37). Finally, our statistical test applies to independent samples from a stochastic algorithm such as Monte Carlo sampling, whereas some other methods do not

require the samples to be independent (36–40). Preliminary versions of our sampling exhaustiveness protocol have been already used in several integrative modeling applications (9–11, 61–63). Earlier, sampling exhaustiveness for integrative modeling was estimated, at best, by manual visual inspection of localization densities of clusters (7, 8, 44).

Future directions

Future directions include expanding this protocol to establish more detailed tests for exhaustiveness. For instance, it will be useful to determine not just the sampling precision for the entire macromolecular system, but also the sampling precision for different components of the system (e.g., proteins, domains) separately. Such more detailed information would be useful in the analysis stage of the iterative four-stage integrative modeling process (2, 4) to determine, for instance, what representations to change and what input data to reexamine to improve the sampling precision for the entire system.

Structures of macromolecular systems are increasingly computed by integrative modeling that relies on various types of experimental data and theoretical information (20). However, validation of integrative models and data is a major open research challenge. It is particularly timely because of the Worldwide Protein Data Bank effort to expand the scope of its archive to integrative structures (20). We suggest that a sampling exhaustiveness protocol, such as the one described here, is the first assessment applied to all integrative models.

Availability

Benchmark data and code used in this article are freely available at http://salilab.org/sampcon. The code relies on our open source IMP package (https://integrativemodeling.org).

Author contributions

S.V., I.E.C., and A.S. designed research. S.V. and I.E.C. performed research. P.C. contributed computational tools. S.V., I.E.C., and A.S. analyzed data. S.V., I.E.C., and A.S. wrote the manuscript.

Acknowledgments

The authors thank the members of their research group for useful suggestions and Dr. Benjamin Webb for helping to implement the method in IMP. This work was supported by the National Institutes of Health (NIH) (P01 GM105537, R01 GM083960, and P41 GM109824 to A.S.) and the National Science Foundation (NSF) (graduate research fellowship 1650113 to I.E.C.). Molecular graphics images were produced using the UCSF Chimera package from the Computer Graphics Laboratory, University of California, San Francisco, California (supported by NIH P41 RR-01081).

References

1. A. Ward, A. Sali, I. Wilson, Integrative structural biology. *Science* **339**, 913–915 (2013).
2. D. Russel, *et al.*, Putting the pieces together: integrative structure determination of macromolecular assemblies. *PLoS Biol* **10**, e1001244 (2012).
3. B. Webb, *et al.*, “Modeling of proteins and their assemblies with the Integrative Modeling Platform” in *Methods Mol Biol*, Y. Chen, Ed. (Humana Press, 2014), pp. 277–295.
4. D. Schneidman-Duhovny, R. Pellarin, A. Sali, Uncertainty in Integrative Structural Modeling. *Curr Opin Struct Biol* **28**, 96–104 (2014).
5. F. Alber, *et al.*, The molecular architecture of the nuclear pore complex. *Nature* **450**, 695–701 (2007).
6. K. Lasker, *et al.*, Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. *Proc Natl Acad Sci USA* **109**, 1380–1387 (2012).
7. Y. Shi, *et al.*, Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. *Mol Cell Proteomics* **13**, 2927–2943 (2014).
8. P. Robinson, *et al.*, Molecular architecture of the yeast Mediator complex. *eLife* **4**, e08719 (2015).
9. J. Fernandez-Martinez, *et al.*, Structure and Function of the Nuclear Pore Complex Cytoplasmic mRNA Export Platform. *Cell* **167**, 1215–1228 (2016).

10. P. Upla, *et al.*, Molecular Architecture of the Major Membrane Ring Component of the Nuclear Pore Complex. *Structure* **25**, 434–445 (2017).
11. S. Viswanath, *et al.*, The molecular architecture of the yeast spindle pole body core determined by Bayesian integrative modeling. *Mol Biol Cell* (2017).
12. M. J. Sippl, Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* **213**, 859–83 (1990).
13. M. Y. Shen, A. Sali, Statistical potential for assessment and prediction of protein structures. *Protein Sci* **15**, 2507–24 (2006).
14. S. Viswanath, D. V. Ravikant, R. Elber, Improving ranking of models for protein complexes with side chain modeling and atomic potentials. *Proteins* **81**, 592–606 (2013).
15. W. L. Jorgensen, D. S. Maxwell, J. Tirado-Rives, Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society* **118**, 11225–11236 (1996).
16. B. R. Brooks, *et al.*, CHARMM: the biomolecular simulation program. *J Comput Chem* **30**, 1545–614 (2009).
17. N. Metropolis, S. Ulam, The monte carlo method. *Journal of the American statistical association* **44**, 335–341 (1949).
18. K. Lasker, M. Topf, A. Sali, H. J. Wolfson, Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly. *Journal of molecular biology* **388**, 180–194 (2009).

19. F. Alber, B. T. Chait, M. P. Rout, A. Sali, “Integrative Structure Determination of Protein Assemblies by Satisfaction of Spatial Restraints” in *Protein-Protein Interactions and Networks: Identification, Characterization and Prediction.*, A. Panchenko, T. Przytycka, Eds. (Springer-Verlag, 2008), pp. 99–114.
20. A. Sali, *et al.*, Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure* **23**, 1156–67 (2015).
21. T. Schwede, *et al.*, Outcome of a workshop on applications of protein models in biomedical research. *Structure* **17**, 151–159 (2009).
22. D. Baker, A. Sali, Protein structure prediction and structural genomics. *Science* **294**, 93–96 (2001).
23. H. Tjong, K. Gong, L. Chen, F. Alber, Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome Res* **22**, 1295–305 (2012).
24. A. Loquet, *et al.*, Atomic model of the type III secretion system needle. *Nature* **486**, 276–9 (2012).
25. R. Abagyan, M. Totrov, Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J Mol Biol* **235**, 983–1002 (1994).
26. Y. Zhang, D. Kihara, J. Skolnick, Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins* **48**, 192–201 (2002).
27. Y. Shen, *et al.*, Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci U S A* **105**, 4685–90 (2008).
28. A. Roy, A. Kucukural, Y. Zhang, I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **5**, 725–38 (2010).

29. Y. Song, *et al.*, High-resolution comparative modeling with RosettaCM. *Structure* **21**, 1735–42 (2013).
30. D. Bhattacharya, J. Cheng, De novo protein conformational sampling using a probabilistic graphical model. *Sci Rep* **5**, 16332 (2015).
31. Z. Zhang, C. E. Schindler, O. F. Lange, M. Zacharias, Application of Enhanced Sampling Monte Carlo Methods for High-Resolution Protein-Protein Docking in Rosetta. *PLoS One* **10**, e0125941 (2015).
32. J. D. Yesselman, R. Das, Modeling Small Noncanonical RNA Motifs with the Rosetta FARFAR Server. *Methods Mol Biol* **1490**, 187–98 (2016).
33. X. Daura, *et al.*, Peptide folding: When simulation meets experiment. *Angewandte Chemie-International Edition* **38**, 236–240 (1999).
34. X. Daura, W. F. van Gunsteren, A. E. Mark, Folding-unfolding thermodynamics of a beta-heptapeptide from equilibrium simulations. *Proteins* **34**, 269–80 (1999).
35. L. J. Smith, X. Daura, W. F. van Gunsteren, Assessing equilibration and convergence in biomolecular simulations. *Proteins-Structure Function and Genetics* **48**, 487–496 (2002).
36. A. Okur, *et al.*, Improved efficiency of replica exchange simulations through use of a hybrid explicit/implicit solvation model. *Journal of Chemical Theory and Computation* **2**, 420–433 (2006).
37. A. Okur, D. R. Roe, G. Cui, V. Hornak, C. Simmerling, Improving Convergence of Replica-Exchange Simulations through Coupling to a High-Temperature Structure Reservoir. *J Chem Theory Comput* **3**, 557–68 (2007).

38. E. Lyman, D. M. Zuckerman, Ensemble-based convergence analysis of biomolecular trajectories. *Biophys J* **91**, 164–72 (2006).

39. E. Lyman, D. M. Zuckerman, On the structural convergence of biomolecular simulations by determination of the effective sample size. *J Phys Chem B* **111**, 12876–82 (2007).

40. A. Grossfield, S. E. Feller, M. C. Pitman, Convergence of molecular dynamics simulations of membrane proteins. *Proteins* **67**, 31–40 (2007).

41. B. Hess, Convergence of sampling in protein simulations. *Phys Rev E Stat Nonlin Soft Matter Phys* **65**, 031910 (2002).

42. W. J. Son, S. Jang, S. Shin, A simple method of estimating sampling consistency based on free energy map distance. *J Mol Graph Model* **27**, 321–5 (2008).

43. C. Neale, W. F. Bennett, D. P. Tieleman, R. Pomes, Statistical Convergence of Equilibrium Properties in Simulations of Molecular Solutes Embedded in Lipid Bilayers. *J Chem Theory Comput* **7**, 4175–88 (2011).

44. J. Luo, *et al.*, Architecture of the human and yeast general transcription and DNA repair factor TFIIH. *Mol Cell* **59**, 794–806 (2015).

45. R. Chen, Z. Weng, Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins* **47**, 281–94 (2002).

46. R. Chen, L. Li, Z. Weng, ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **52**, 80–7 (2003).

47. W. K. Hastings, Monte-Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **57**, 97–109 (1970).

48. S. Siegal, Nonparametric statistics for the behavioral sciences (McGraw-hill, 1956).

49. D. McCarroll, *Simple Statistical Tests for Geography* (CRC Press, 2016).

50. S. Nakagawa, I. C. Cuthill, Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews* **82**, 591–605 (2007).

51. S. Greenland, *et al.*, Statistical tests, P. *European journal of epidemiology* **31**, 337–350 (2016).

52. J. H. McDonald, *Handbook of biological statistics* (Sparky House Publishing Baltimore, MD, 2009).

53. H. Cramer, *Mathematical Methods of Statistics* (Princeton University Press, 1946).

54. W. G. Cochran, Some methods for strengthening the common χ 2 tests. *Biometrics* **10**, 417–451 (1954).

55. A. Kuzmanic, B. Zagrovic, Determination of ensemble-average pairwise root mean-square deviation from experimental B-factors. *Biophysical journal* **98**, 861–871 (2010).

56. E. F. Pettersen, *et al.*, UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry* **25**, 1605–1612 (2004).

57. H. Hwang, T. Vreven, J. Janin, Z. Weng, Protein-protein docking benchmark version 4.0. *Proteins* **78**, 3111–4 (2010).

58. S. R. Comeau, D. W. Gatchell, S. Vajda, C. J. Camacho, ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics* **20**, 45–50 (2004).

59. A. Tovchigrechko, I. A. Vakser, GRAMM-X public web server for protein–protein docking. *Nucleic acids research* **34**, W310–W314 (2006).

60. S. Viswanath, D. Ravikant, R. Elber, DOCK/PIERR: web server for structure prediction of protein–protein complexes. *Protein Structure Prediction*, 199–207 (2014).

61. D. J. Saltzberg, *et al.*, A Residue-Resolved Bayesian Approach to Quantitative Interpretation of Hydrogen–Deuterium Exchange from Mass Spectrometry: Application to Characterizing Protein–Ligand Interactions. *The Journal of Physical Chemistry B* **121**, 3493–3501 (2016).

62. X. Wang, *et al.*, The Proteasome-Interacting Ecm29 Protein Disassembles the 26S Proteasome in Response to Oxidative Stress. *Journal of Biological Chemistry* (2017).

63. C. Y. Zhou, *et al.*, Regulation of Rvb1/Rvb2 by a Domain within the INO80 Chromatin Remodeling Complex Implicates the Yeast Rvbs as Protein Assembly Chaperones. *Cell Reports* **19**, 2033–2044 (2017).

Chapter V - The proteasome-interacting Ecm29 protein disassembles the 26S proteasome in response to oxidative stress

Contributing authors

Xiaorong Wang[1], Ilan E. Chemmama[2], Clinton Yu[1], Alexander Huszagh[1], Yue Xu[3], Rosa Viner[4], Sarah A. Block[5], Peter Cimermancic[2], Scott D. Rychnovsky[5], Yihong Ye[3], Andrej Sali[2], Lan Huang[1]

[1]Department of Physiology and Biophysics, University of California, Irvine, California 92697

[2]Department of Bioengineering and Therapeutic Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biosciences, University of California, San Francisco, California 94158

[3]Laboratory of Molecular Biology, NIDDK, National Institutes of Health, Bethesda, Maryland 20892

[4]Thermo Fisher Scientific, San Jose, California 94134

[5]Department of Chemistry, University of California, Irvine, California 92697

Contact: lanhuang@uci.edu

Abstract

Oxidative stress has been implicated in multiple human neurological and other disorders. Proteasomes are multi-subunit proteases critical for the removal of oxidatively damaged proteins. To understand stress-associated human pathologies, it is important to uncover

the molecular events underlying the regulation of proteasomes upon oxidative stress. To this end, we investigated H_2O_2 stress–induced molecular changes of the human 26S proteasome and determined that stress-induced 26S proteasome disassembly is conserved from yeast to human. Moreover, we developed and employed a new proteomic approach, XAP (in vivo cross-linking–assisted affinity purification), coupled with stable isotope labeling with amino acids in cell culture (SILAC)– based quantitative MS, to capture and quantify several weakly bound proteasome-interacting proteins and examine their roles in stress-mediated proteasomal remodeling. Our results indicate that the adapter protein Ecm29 is the main proteasome-interacting protein responsible for stresstriggered remodeling of the 26S proteasome in human cells. Importantly, using a disuccinimidyl sulfoxide– based crosslinking MS platform, we mapped the interactions of Ecm29 within itself and with proteasome subunits and determined the architecture of the Ecm29–proteasome complex with integrative structure modeling. These results enabled us to propose a structural model in which Ecm29 intrudes on the interaction between the 20S core particle and the 19S regulatory particle in the 26S proteasome, disrupting the proteasome structure in response to oxidative stress.

Introduction

Oxidative stress has been associated with the aging process and implicated in many human diseases, particularly neurodegenerative disorders (1). Protein oxidation can lead to unwanted changes in protein structure and function, resulting in the accumulation of severely oxidized proteins, subsequent cytotoxicity, and, ultimately, cell death. Therefore, oxidatively damaged proteins must be repaired or removed in a timely fashion to maintain cell homeostasis. Most oxidized proteins undergo selective proteolysis, and abundant evidence has indicated that proteasomes play a critically important role in the removal of oxidized proteins to preserve cell viability in response to oxidative stress (2–4). In addition, proteasomes are highly regulated during cellular responses to oxidative stresses. However, the molecular details underlying such modulation remain largely unexplored, particularly in human cells.

The 26S proteasome is a macromolecular machine responsible for ubiquitin/ATP-dependent protein degradation and comprises two subcomplexes: a 20S core particle (CP) and a 19S regulatory particle (RP) (5, 6). The 20S CP harbors various catalytic activities, including chymotrypsin-like, trypsin-like, and caspase-like peptidase activities. It is composed of seven and seven subunits in eukaryotes that form a conserved cylindrical structure of four heptameric stacked rings assembled in the order of $\alpha\beta\beta\alpha$. Activation of the 20S CP requires binding to proteasome activator proteins (5). The 19S RP is one of the major proteasome activators and consists of at least 19 distinct subunits that constitute the base and lid subcomplexes. The base is composed of six ATPases (Rpt1– 6) and four non ATPase subunits (Rpn1, 2, 10, and 13), whereas the remaining nine subunits (Rpn3, 5–9, 11, and 12 and Rpn15/Sem1) comprise the lid structure. The

19S RP carries multiple functions to facilitate substrate degradation, including substrate recognition, deubiquitination, protein unfolding, substrate translocation, and gating of the 20S CP. In contrast to the highly ordered and stable structure of the 20S CP, the 19S RP appears to be much more flexible and dynamic. Nevertheless, the overall architectures of the 19S RP and the 26S holocomplex are highly conserved from yeast to human (7–11).

During oxidative stress, the proteasome system is highly regulated to fulfill its function in maintaining cell homeostasis (3, 4, 12). To facilitate the removal of oxidatively damaged proteins, ubiquitin/ATP-independent degradation by the 20S CP is significantly enhanced because of the increased amount of free 20S CP in cells, which is not the result of transcriptional control but, rather, of oxidative stress–triggered disassembly of the 26S proteasome (13–15). In yeast, such proteasome dissociation is dependent on the proteasome-interacting protein Ecm29 (13). However, it remains unclear whether human Ecm29 possesses a similar function because of low sequence similarity to its yeast ortholog (20%) and the relative complexity of human systems. In addition, Ecm29-dependent regulation of the proteasome system can have a multifaceted effect on cell physiology (11, 13, 16–19) by inhibiting ubiquitin-dependent protein degradation (11, 19), stabilizing proteasomes (16, 17), and assisting membrane-associated localization of proteasomes (20, 21) and TLR3-dependent signaling (22). However, how Ecm29 regulates the activity of the 26S proteasome in human cells, particularly in response to oxidative stress, is largely unknown. Therefore, further studies are needed to fully describe the molecular details underlying stress-mediated regulation of the 26S proteasome.

Here we quantitatively examined oxidative stress–mediated changes in the human 26S proteasome by developing a new affinity purification-MS strategy. In addition, we investigated the recruitment of Ecm29 to the proteasome and its associated biological implications. Moreover, we employed cross-linking mass spectrometry to define interactions within the Ecm29– proteasome complex, which were used for integrative structure modeling. Together, the results allow us to propose a structural model in which Ecm29 intrudes on the interaction between the 20S CP and 19S RP, thus modulating the function of the proteasome in response to oxidative stress.

Results

H_2O_2-mediated molecular changes in the human 26S proteasome

To evaluate the compositional changes of the human 26S proteasome under oxidative stress, we applied a single-step affinity purification procedure using an HTBH-tagged proteasome subunit (i.e. Rpn11-HTBH) as bait (23). This method has proven to be fast and effective to obtain functional human proteasome complexes and to identify proteasome-interacting proteins (23, 24). However, one of the key proteasome regulators, Ecm29, was not co-purified with human proteasomes. This is not surprising, as dynamic or transient interactions are often lost during conventional affinity purification (24, 25). To circumvent this problem, we developed a new strategy named *in vivo* cross-linking–assisted affinity purification MS (XAPMS) (**Fig. 5.1A**). XAP-MS integrates mild *in vivo* formaldehyde (FA) cross-linking (0.1%) prior to cell lysis, which enables better preservation of 26S proteasome intactness and proteolytic activities during native lysis (26). To quantify H2O2-induced changes, we coupled XAP-MS with SILAC-based

quantitation, in which one population of $293^{Rpn11\text{-}HTBH}$ cells was grown in light (L) medium and treated with H_2O_2, whereas the other population of the same cells was grown in heavy (H) medium as a control without treatment. The relative abundances of proteasome subunits between treated and untreated cells are represented by average peptide SILAC ratios (i.e. L/H) of the two compared samples (**Fig. 5.1B** and supplemental Table S2). As shown, all identified 19S RP subunits have SILAC ratios close to 1, indicating that the abundances of these subunits in the purified samples were unaffected by H_2O_2 stress (**Fig. 5.1B**). In contrast, the SILAC ratios of all 20S CP subunits decreased substantially with SILAC ratios of < 0.4, demonstrating that oxidative stress resulted in dissociation of the 20S CP from the 19S RP. These results were validated using quantitative immunoblot analysis (supplemental Fig. S1). Similarly, we carried out XAP-SILAC MS experiments using $293^{\alpha7/Pre10\text{-}HTBH}$ cells. As expected, 20S CP subunits remained unchanged, whereas copurified 19S RP subunits decreased substantially upon H_2O_2 stress (supplemental Fig. S2). Taken together, we confirmed that H_2O_2 stress–induced disassembly of the 26S proteasome is conserved in mammalian cells.

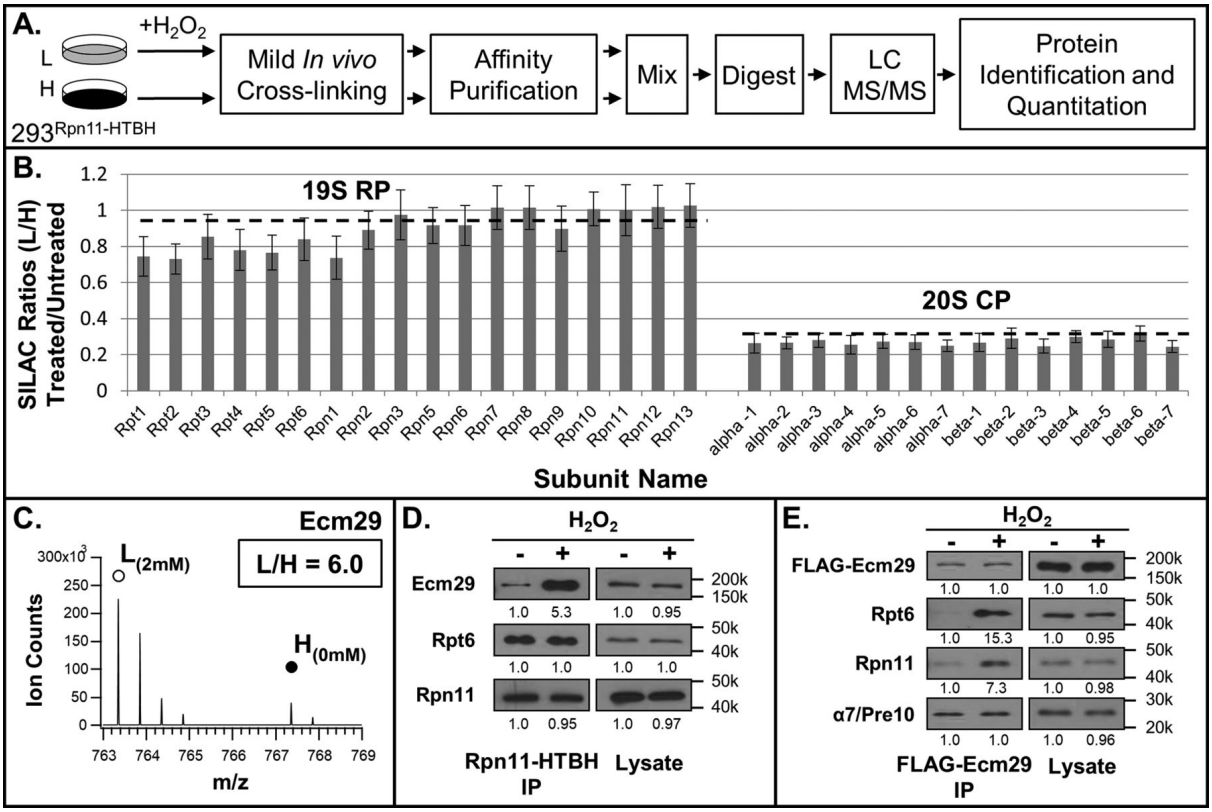

Figure 5.1 | Determination of 26S proteasome disassembly and enrichment of Ecm29 upon oxidative stress in human cells.

A, the general workflow of the SILAC-based quantitative XAP-MS strategy. **B**, relative abundance changes (i.e. SILAC (L/H) ratios) of the human 26S subunits purified from $293^{Rpn11\text{-}HTBH}$ cells in the presence and absence of H_2O_2 treatment. L ($[^{12}C^{14}N]$Arg/Lys), treated cells; H ($[_{13}C_{15}N]$Arg/Lys), untreated control cells. **C**, MS spectrum of a representative Ecm29 tryptic peptide pair (m/z 763.352^{2+} versus 767.352^{2+}) with the SILAC ratio (L/H) as 6.0 when comparing treated with untreated samples. ○, treated (L); ●, untreated (H). **D**, immunoblot analysis of Ecm29 abundance in proteasomes purified from $293^{Rpn11\text{-}HTBH}$ cells before and after H_2O_2 treatment. IP, immunoprecipitation. **E**, immunoblot analysis of protein complexes purified using FLAG–Ecm29 that was transiently transfected in $293^{Rpn11\text{-}HTBH}$ cells. Specific antibodies against Rpt6 and α7/Pre10 were used to probe their abundance. Streptavidin–HRP was used to probe HTBH-tagged Rpn11. All of the treated cells were incubated with 2 mM H_2O_2 for 30 min; untreated cells served as controls. The numbers under the immunoblot bands represent quantitative measurements using a Fuji LAS4000 scanning system.

Ecm29-dependent regulation of the 26S proteasome upon oxidative stress

Although it is known that Ecm29 can be recruited to the 19S proteasome during oxidative stress to regulate the proteasome composition in yeast (13), it remains unclear whether Ecm29 is a general player in modulating proteasome structure upon oxidative stress in eukaryotic systems. To this end, we first determined H_2O_2-triggered enrichment of human Ecm29 in the 19S RP using quantitative XAP-MS and immunoblotting analyses (**Fig. 5.1**, **C** and **D**). This was further validated by reciprocal affinity purification using FLAG-Ecm29 and quantitative immunoblotting (**Fig. 5.1E**). As shown, increased amounts of the two selected 19S subunits (i.e. Rpn11 and Rpt6) were co-purified with Ecm29 upon H_2O_2 treatment, whereas the amount of a 20S subunit, α7/Pre10, did not change in the purified Ecm29 complex. This result suggests that, although there is a dramatic increase in Ecm29 binding to the 19S upon oxidative stress, there is no change in Ecm29 binding to the 26S, thus implying two populations of Ecm29 detected here.

To understand how human Ecm29 regulates the 26S proteasome, we generated two Ecm29 knockdown (KD) cells (i.e. $293^{\text{Rpn11-TB\_Ecm29KD\#53}}$ and $293^{\text{Rpn11-TB\_Ecm29KD\#55}}$) (**Fig. 5.2A**). Quantitative immunoblotting analysis of purified proteasomes revealed that the abundances of the selected proteasome subunits (i.e. Rpt6 and α7/Pre10) are similar in Ecm29 KD and control KD cells under unstressed conditions (**Fig. 5.2B**). In addition, the 26S proteasomal activities were comparable in the two knockdown cells (**Fig. 5.2C**). Together, these results show that human Ecm29 is not essential for the assembly and function of the 26S proteasome. However, when cells were treated with H_2O_2 to induce oxidative stress, the 26S proteasome was rapidly disassembled in control KD cells, but such dissociation was considerably reduced in $293^{\text{Rpn11-TB\_Ecm29KD\#53}}$ and $293^{\text{Rpn11-}}$

$^{TB\_Ecm29KD\#55}$ cells (**Fig. 5.2B**). As expected, H_2O_2 stress also significantly reduced 26S proteasomal activities in control KD cells but not in Ecm29 KD cells (**Fig. 5.2C**). The observed changes in proteasomal activities correlated well with the levels of ubiquitinated proteins detected in these cells in the presence and absence of H_2O_2 stress, respectively (**Fig. 5.2D**). Collectively, our data strongly suggests that Ecm29 regulates H_2O_2-induced 26S proteasome disassembly in human cells.

To further evaluate the function of Ecm29, we overexpressed human FLAG-Ecm29 in $293^{Rpn11\text{-}HTBH}$ cells. Quantitative immunoblotting analyses of the purified proteasomes revealed that overexpression of Ecm29 resulted in more proteasome bound Ecm29 and, concurrently, less 19S RP-associated 20S CP under normal conditions (**Fig. 5.3A** and supplemental Fig. S3). This suggests that increased abundance of Ecm29 under nonstress conditions can disrupt normal 26S proteasome integrity, albeit to a lesser extent compared with the impact of oxidative stress. Interestingly, upon H_2O_2 stress, an increased amount of Ecm29 was also detected at the 19S RP even in the presence of overexpressed Ecm29, similar to wild-type cells, resulting in increased separation of the 20S CP from the 19S RP. These results were subsequently confirmed by the measurements of 26S proteasomal activities, as illustrated in **Fig. 5.3B**. Our data demonstrate that Ecm29 plays an evolutionarily conserved role in regulating the 26S proteasome, especially upon oxidative stress.

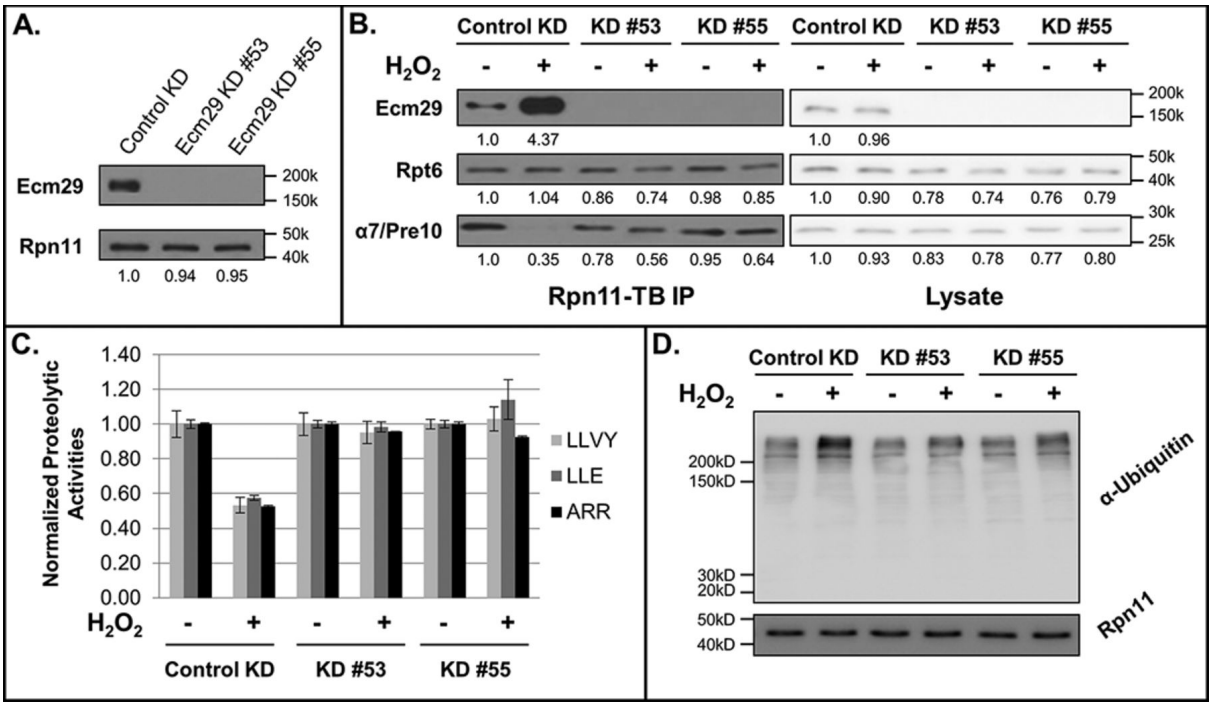

Figure 5.2 | The effect of human Ecm29 on H2O2-induced 26S proteasome disassembly.

A, evaluation of Ecm29 KD efficiency. **B**, the 26S proteasome was affinity-purified with Rpn11-TB from $293^{\text{Rpn11-TB\_controlKD}}$ (control KD), $293^{\text{Rpn11-TB\_Ecm29KD\#53}}$ (KD #53), and $293^{\text{Rpn11-TB\_ Ecm29KD\#55}}$ (KD #55) cells and analyzed by quantitative Western blotting with antibodies against Rpt6 (19S) and α7/Pre10 (20S). **C**, effect of Ecm29 knockdown on the activity of the 26S proteasome. The proteasomal proteolytic activities in $293^{\text{Rpn11-TB\_controlKD}}$ (control KD), $293^{\text{Rpn11-TB\_ Ecm29KD\#53}}$ (KD #53), and $293^{\text{Rpn11-TB\_ Ecm29KD\#55}}$ (KD #55) cells (treated or untreated) were determined by in-solution peptidase activity assays. Three fluorogenic peptide substrates were used: SUC-LLVYAMC for chymotrypsin-like activity, SUC-LLE-AMC for peptide hydrolase activity, and SUC-ARR-AMC for trypsin-like activity. The activities were normalized to Rpn11 (a 19S subunit) in each sample. Data were obtained from three experiments. **D**, detection of total ubiquitin conjugates after H_2O_2-induced stress by Western blot analysis with an antibody against ubiquitin. Equivalent loading was determined by analysis of Western blots with an antibody against Rpn11 and by staining the membrane with Amido Black. The numbers under the immunoblot bands represent quantitative measurements using a Fuji LAS4000 scanning system.

Ecm29 is the main PIP responsible for stress-induced proteasome disassembly

In addition to Ecm29, we identified 9 proteasome-interacting proteins (PIPs) that displayed H_2O_2-induced abundance changes alongside purified 19S RP (i.e. SILAC ratios ≥ 2) using quantitative XAP-MS analysis of $293^{Rpn11\text{-}HTBH}$ cells (supplemental Table S2). Among them, five have known functions in the proteasome system: 19S assembly chaperones (p27/Nas2 and Rpn14/Gankyrin), deubiquitinase (Usp15), ubiquitin ligase (Ube3A), and Hsp70. Hsp70 has been shown to be important for proteasome reassembly after H_2O_2 stress (14). Three of the four remaining PIPs (i.e. Bag6, Ubl4A, and Trc35) are components of a ubiquitin ligase–associated multiprotein transmembrane recognition complex (TRC), which is unique to mammalian systems. With the exception of Bag6 (27), the other factors have not been linked to regulation of proteasome function. To confirm the MS results, respective immunoblotting analyses of purified Rpn11-containing proteasomes and FLAG–Ubl4A complexes were performed (supplemental Fig. S4). Our results demonstrate that the TRC complex interacts with the 26S proteasome via the 19S complex and that their interactions can be modulated by H_2O_2 stress. Given the similarity in H_2O_2-induced enrichment of the TRC complex to that of Ecm29 at the 19S RP, we initially hypothesized that the TRC complex may regulate the 26S proteasome complex in the same manner as human Ecm29 during oxidative stress. However, CRISPR-mediated knockout or siRNA-mediated silencing of Bag6 did not affect the interaction between 20S and 19S proteasomes in either normal or H_2O_2-treated cells (supplemental Fig. S5), suggesting that Bag6 is not required for H_2O_2-induced 26S proteasome disassembly. In addition, knockout of Bag6 did not interfere with H_2O_2-induced recruitment of Ecm29 to the 19S proteasome and vice versa. These results imply that

Bag6 is not associated with Ecm29-dependent regulation of the 26S proteasome complex despite its increased association with the 19S RP upon H_2O_2 treatment. Because the TRC complex is thought to chaperone polypeptides en route to the proteasome to facilitate the degradation of folding-defective proteins, which include retrotranslocation products from the endoplasmic reticulum and mislocalized membrane proteins (28, 29), the increased association of Bag6 with the 19S complex may result from the loss of communication between the 19S and the 20S proteasome, which presumably traps ubiquitinated proteins together with their chaperones on the 19S complex. From these results, we concluded that, although oxidative stress-induced changes in the proteasome interactome are not limited to Ecm29, proteasome disassembly is specifically regulated by the 19S-associated Ecm29.

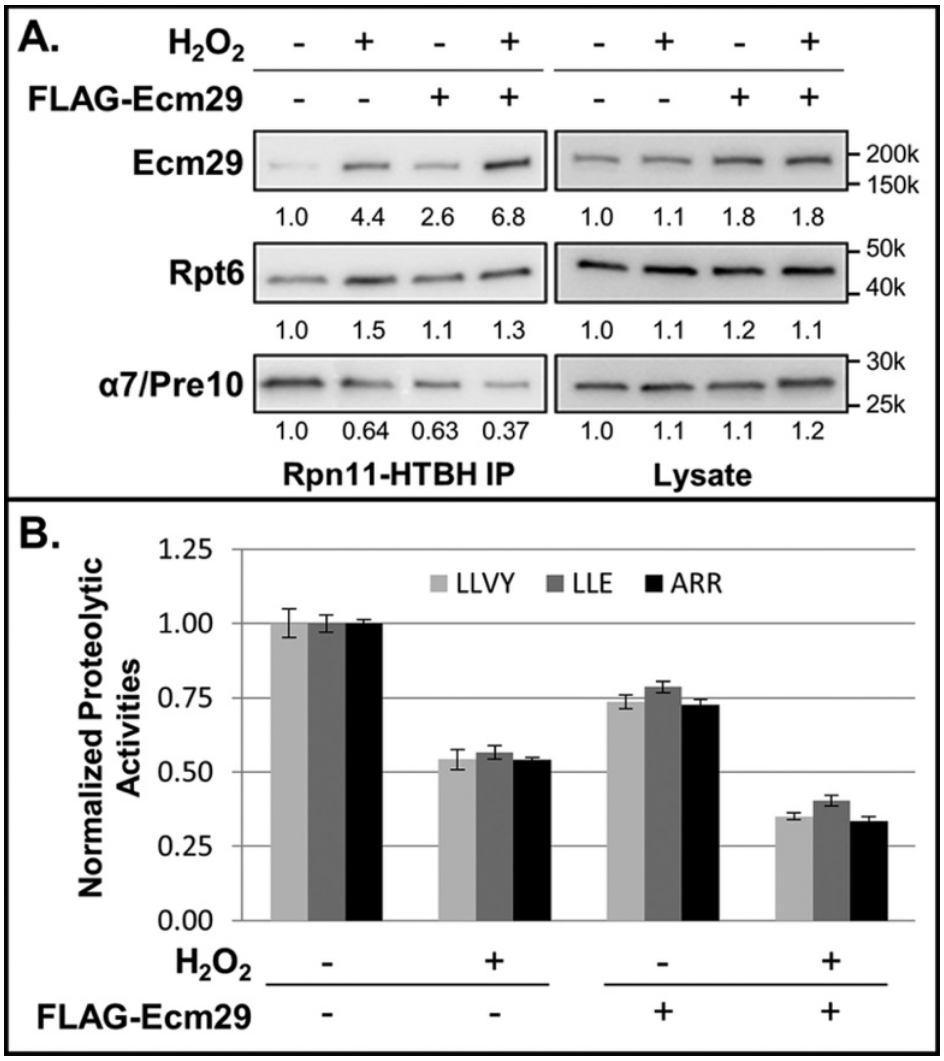

Figure 5.3 | Modulation of the human 26S proteasome by Ecm29 overexpression. **A**, effect of Ecm29 overexpression on 26S proteasome integrity. Shown are quantitative immunoblot analyses of purified proteasomes (left) and cell lysates (right) in the presence or absence of overexpressed Ecm29. $293^{Rpn11\text{-}HTBH}$ cells were transiently transfected with FLAG–Ecm29 either treated with H_2O_2 or untreated as a control. Quantitative Western blotting was performed with antibodies against Ecm29, Rpt6 (19S), and α7/Pre10 (20S). **B**, effect of Ecm29 overexpression on 26S proteasome activities. The proteasomal proteolytic activities were determined by in-solution peptidase activity assays with three fluorogenic peptide substrates: SUC-LLVY-AMC (LLVY); SUC-LLE-AMC (LLE), and SUC-ARR-AMC (ARR). The activities were normalized to Rpn11 (a 19S subunit) in each sample. Data were from three experiments. The numbers under the immunoblot bands represent quantitative measurements using a Fuji LAS4000 scanning system.

Physical interactions of Ecm29 with the proteasome

Because of the lack of structural details on Ecm29 alone and its complex with proteasomes, how Ecm29 is recruited to the 19S particle during oxidative stress is largely unknown. To understand how Ecm29 regulates the proteasome upon oxidative stress, we employed our previously developed XL-MS strategy to determine protein interaction contacts at specific residues (30). This XL-MS strategy enables simplified and accurate identification of cross-linked peptides by integrating an MS-cleavable homobifunctional amine-reactive NHS ester, disuccinimidyl sulfoxide (DSSO), with multistage tandem mass spectrometry (MS^n) (30). To ensure the capture of a sufficient amount of Ecm29–proteasome complexes, we co-expressed HTBH-Ecm29 in $293^{HTBH\text{-}Rpt6}$ cells and treated the cells with H_2O_2 stress prior to cell lysis. Single-step affinity purification by binding to streptavidin beads was carried out to isolate Ecm29–proteasome complexes for *in vitro* DSSO cross-linking, similar as described previously (11). The resulting DSSO cross-linked peptides were analyzed by LC/MS^n for identification (30). As an example, a representative MS^n analysis of a DSSO interlinked peptide is illustrated in **Fig. 5.4**. As shown, the cleavage of either of the two symmetric MS-cleavable C–S bonds in the linker region of the DSSO interlinked peptide α-β (m/z 848.66974^{4+}) resulted in detection of two characteristic fragment pairs: α_A/β_T (m/z 762.412^{2+} and 925.922^{2+}) and α_T/β_A (m/z 778.392^{2+} and 909.942^{2+}) during MS^2 analysis (**Fig. 5.4A**). MS^3 sequencing of the α_T and β_A fragment ion pair yielded series of b and y ions that unambiguously identified them as $^{284}HSVATAADLELK_TSK^{287}$ of Ecm29 and $^{372}K_AMNVSPDVNYEELAR^{386}$ of Rpt5, respectively (**Fig. 5.4B**, left and right panels). Together, the MS^n analysis determined a cross-link between Lys-285 of Ecm29 and Lys-372 of Rpt5 (**Fig. 5.4**). In total, LC/MS^n

analysis identified 69 unique Lys–Lys linkages involving Ecm29, seven of which were interprotein and 62 intraprotein interactions (supplemental Tables S3 and S4). Ecm29 was determined to interact with five 19S base subunits, Rpt1, Rpt4, Rpt5, Rpn1, and Rpn10, as illustrated in **Fig. 5.5**. Among them, Ecm29 has the most contacts with Rpt5, as their interactions are supported by multiple Lys–Lys linkages with the highest number of redundant counts. Based on the cross-link map (**Fig. 5.5**), we concluded that Ecm29 interacts with the 19S base subunits through its multiple HEAT repeat domains. Although the N termini of Rpt4 and Rpt5 are both proximal to the N terminus of Ecm29, the C termini of Rpt5, Rpt1, and Rpn10 also have close contacts with Ecm29. In addition, Lys-397 in Rpn1 near its T2 site was cross-linked to Lys-694 of Ecm29. Our results suggest that Ecm29 most likely interacts with the proteasome through multiple contact sites, with Rpt5 as the major docking point.

To evaluate the cross-links identified between Ecm29 and 19S RP subunits, we first performed phylogenetic alignment analysis of Ecm29 derived from five selected organisms: *Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila*, mouse, and human. Despite a relatively low degree of sequence homology, six of the seven lysine residues of Ecm29 that crosslinked with 19S base subunits were located proximally to conserved sequence regions (i.e. within five amino acid residues); only Lys-852 of Ecm29 was found to be relatively far from the nearest conserved site, 18 residues away (supplemental Table S5). When a similar phylogenetic alignment was performed for Rpt1, Rpt4, Rpt5, Rpn1, and Rpn10, the lysine residues on each respective subunit found to be cross-linked to Ecm29 were mostly located within or directly adjacent to highly conserved regions as well, with the exception of Lys-16 of Rpt5 (supplemental Table S5). Therefore,

almost all of the residues cross-linked between Ecm29 and 19S base subunits correspond to evolutionarily conserved regions, suggesting that these cross-links more likely represent the functional protein interaction interfaces between Ecm29 and the proteasome complex.

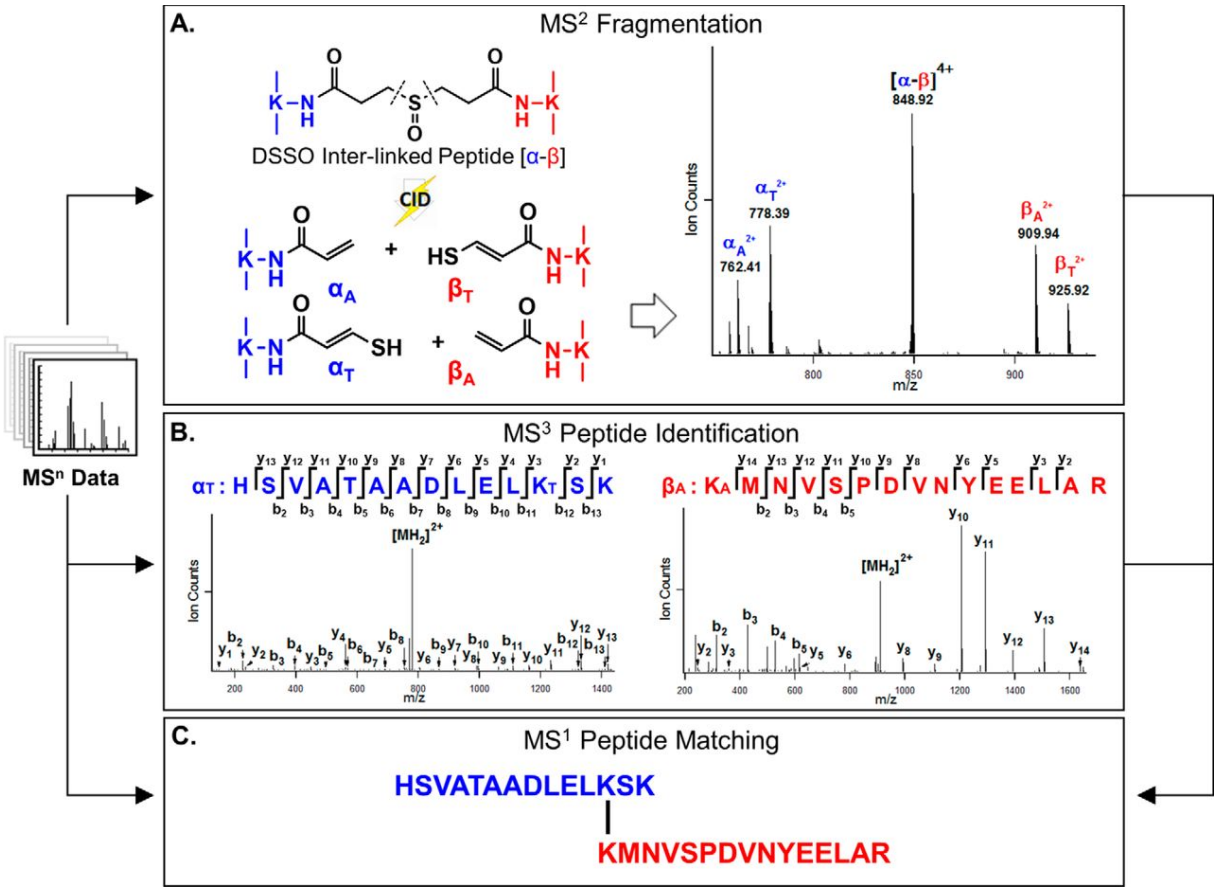

Figure 5.4 | Representative MSn analysis of a selected DSSO cross-link between Ecm29 and Rpt5.

A, collision-induced dissociation (CID) cleavage of the DSSO cross-linked peptide α-β (m/z 848.6697^{4+}) in MS^2 resulted in the formation of two predicted fragment ion pairs: $\alpha_A^{2+}/\beta_T^{2+}$ and $\alpha_T^{2+}/\beta_A^{2+}$. **B**, respective MS^3 spectra of the two selected MS^2 fragment ions α_T^{2+} (m/z 778.39^{2+}) and β_A^{2+} (909.94^{2+}). A series of b and y ions unambiguously identified the peptides as ^{284}HSVATAADLELKTSK^{287} of Ecm29 and ^{372}KAMNVSPDVNYEELAR^{386} of Rpt5, respectively. **C**, along with peptide mass matching at the MS^1 level, three lines of evidence (including MS^2 cross-linker fragmentation and individual MS^3 peptide sequencing) confirm the identification of a DSSO cross-link between Lys-285 of Ecm29 and Lys-372 of Rpt5.

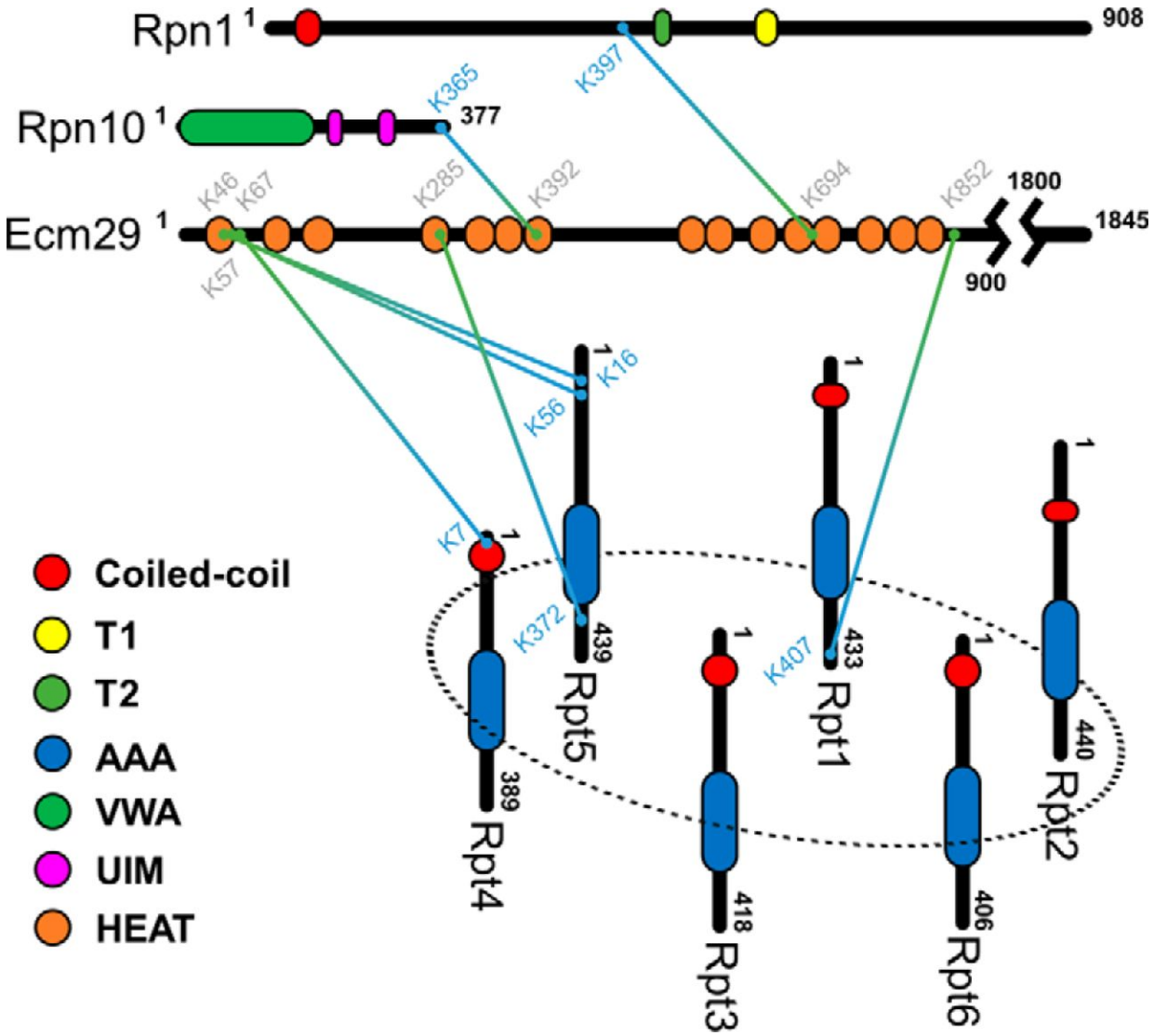

Figure 5.5 | Cross-link map of Ecm29-proteasome interactions.
Diagram of unique Lys–Lys linkages between Ecm29 and the 19S base subunits Rpn1, Rpn10, Rpt1, Rpt4, and Rpt5. The remaining members of the AAA-ATPase ring are also included for spatial context. Cross-linked Ecm29 residues are labeled in green, whereas the residues of the respective 19S RP subunits are shown in blue. Various functional domains of individual proteins are labeled according to SMART: coiled-coil domain, *orange*; AAA-ATPase, *blue*; von Willebrand factor (*VWA*), *green*; ubiquitin-interacting motif (*UIM*), *magenta*; HEAT repeats, *brown*. The T1 and T2 domains of Rpn1 are shown in *yellow* and *orange*, respectively.

Integrative modeling of the Ecm29–proteasome complex

To understand how Ecm29 docks onto the proteasome, the architecture of the Ecm29–proteasome complex was determined using the integrative structure modeling approach described previously (31–35). All available structural information on the Ecm29–proteasome complex was used for computational analyses (supplemental Fig. S6 and Methods). The proteasome was represented by the high-resolution structure of the human 26S proteasome (PDB code 5GJR) (10), whereas Ecm29 was represented by two comparative models built with MODELLER 9.17 (36) based on known template structures detected by HHPred (37) (supplemental Methods). Regions with unknown structures were modeled as flexible strings of beads. Finally, the proximity between specific residue pairs was determined by DSSO XL-MS experiments, which identified a total of 69 unique Lys–Lys linkages (supplemental Tables S3 and S4) describing seven Ecm29-containing interprotein interactions and 62 Ecm29 intraprotein interactions. The maximum C–C distance between any two lysine residues cross-linked by DSSO was estimated to be 30 Å, based on the spacer length of DSSO (10.1 Å), flexibility of lysine side chains, and backbone dynamics. Next, 3,750,000 Ecm29–proteasome models were computed by optimizing spatial proximities, as informed by cross-linking data, excluded volume, and sequence connectivity from 500 random initial models. This process yielded 109,951 good-scoring models (i.e. the ensemble) that satisfy the cross-linking data, the excluded volume, and sequence connectivity restraints used in computing the models. The clustering of the good-scoring models identified two distinct clusters (**Fig. 5.6A**), including 60% (89% of intersubunit cross-links satisfied; **Fig. 5.6**, **B**, **left panel**, and **C**) and 31% (100% of intersubunit crosslinks satisfied; **Fig. 5.6, B**, **right panel**, and **C**) of the models;

the precision of both clusters is 60 Å root mean square deviation for all Ecm29 Cs (supplemental Fig. S7). In general, an ensemble of good-scoring models can be visualized as a localization probability density map. The map specifies the probability of any volume element being occupied by a given bead in superposed good-scoring models. The probability localization density for the structured regions of Ecm29 is sufficiently precise to define the position, but not the orientation, of Ecm29 relative to the proteasome for each of the two clusters (**Fig. 5.6A**). The binding sites on the proteasome are different between the two clusters, indicating that Ecm29 may interact with the proteasome in two different states (although it is also conceivable that we simply do not have enough data to define a single state). In the first state (cluster 1), Ecm29 interacts with the 19S within 10 Å of Rpn1, Rpt2, Rpt4, Rpt5, and Rpn10 (**Fig. 5.6A**, left panel). In this binding mode, Ecm29 does not overlap with the 20S particle and is proximal to the 1 subunit. In the second state (cluster 2), Ecm29 is similarly vicinal to the same partners in the 19S RP (**Fig. 5.6A**, right panel), although its relative position is different and inconsistent with the presence of 20S (**Fig. 5.6A**).

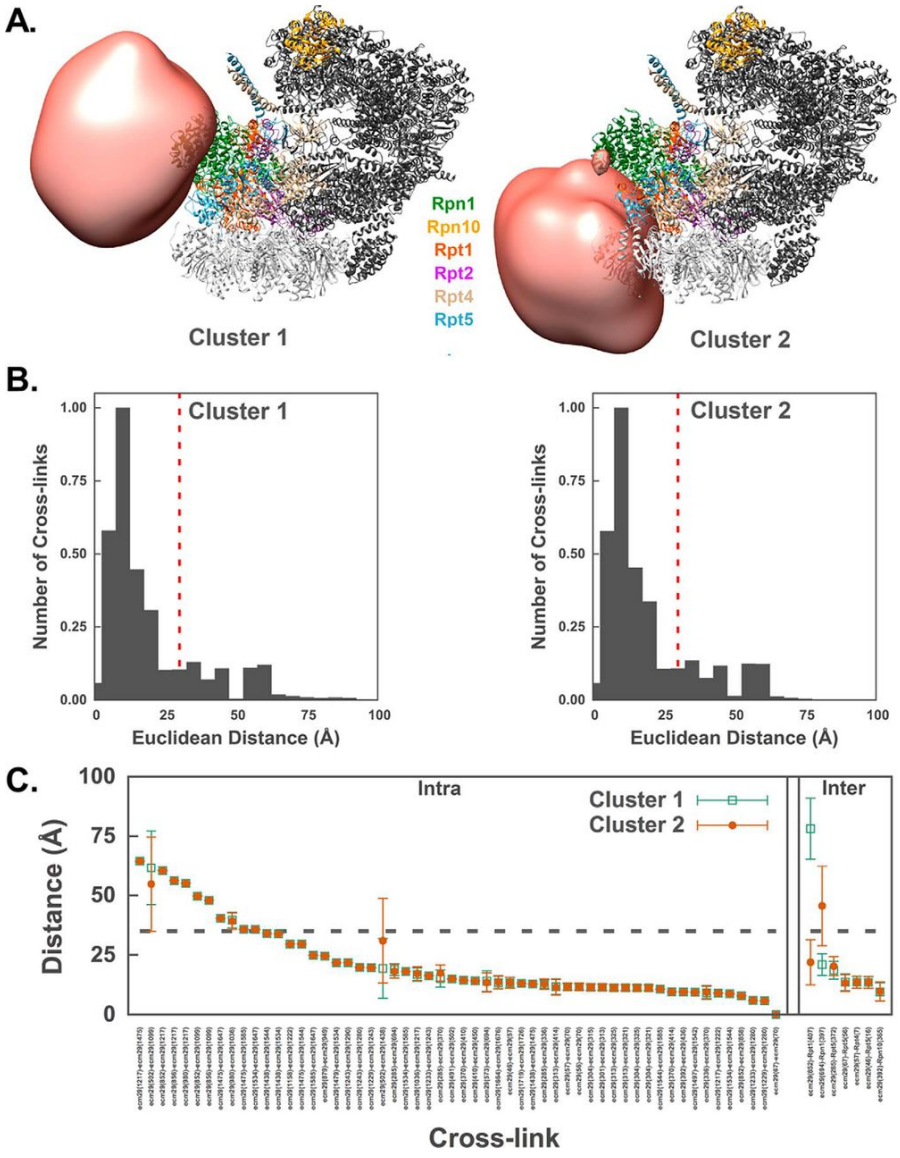

Figure 5.6 | Integrative structure modeling of the Ecm29–proteasome complex. ***A***, localization probability density of the structured part of Ecm29 from cluster 1 (*left panel*) and 2 (*right panel*). The 19S structure is shown in *dark gray*, and the interacting partners of 19S with Ecm29 are *colored*. The 20S α ring is shown in *light gray* for reference. ***B***, Euclidean Cα–Cα distance distributions of all measured cross-links in the ensemble of solutions for each cluster. The *y* axis provides the normalized number of cross-links that were mapped onto the model. The *dashed red* line denotes the expected maximum reach of a cross-link. ***C***, Euclidean Cα–Cα distance statistics for each cross-link in both clusters (cluster 1 in *green empty squares* and cluster 2 in *orange disks*). The cross-links are sorted by average distance (ordinate axis); intra- and interprotein cross-links are separated (*left* and *right*, respectively). The *error bars* represent the standard deviation of the distance across all models in the clusters.

Discussion

Here we developed and employed the XAP-MS strategy to dissect the H_2O_2-dependent compositional dynamics of the human 26S proteasome complex. The results confirmed that the 26S proteasome is highly regulated during oxidative stress and that its disassembly yields more free 20S CP for the removal of oxidatively damaged proteins, corroborating previous reports (13, 14). In addition, with the XAP-MS strategy, we were able to co-purify human Ecm29 with proteasomes reliably with and without H_2O_2 treatment for the first time. This is significant, as the interaction of human Ecm29 with proteasomes appears to be much weaker and/or more transient than yeast Ecm29–proteasome interaction, thus preventing its capture using conventional affinity purification–MS approaches. The reliability of co-purification of human Ecm29 with proteasomes enabled us to confirm H_2O_2-induced enrichment of Ecm29 onto the 19S RP in human cells, similar to its yeast ortholog (13). Our results further indicate that the mild *in vivo* FA cross-linking implemented in the XAP-MS strategy is indeed beneficial for preserving weak, transient, and/or dynamic interactors of protein complexes under native conditions, as described previously (26, 38). Thus, it helps to maintain the integrity of protein complexes as well as to prevent reorganization and loss of protein–protein interactions. Therefore, the XAP-MS method can be applied to study the dynamic interactors of other protein complexes and identify their regulators through protein–protein interactions.

Ecm29 has been shown to be critical in modulating proteasome structure and function (11, 13, 16–19). Interestingly, knockdown of human Ecm29 did not seem to have much impact on the structure and function of proteasomes, as the 26S holocomplex remained intact, and its proteolytic activities were not impaired. These observations are

consistent with those in yeast *ECM29Δ* cells (13) and Ecm29 (i.e. KIAA0368)-deficient mice (39). Collectively, our results suggest that Ecm29 is nonessential for the assembly, integrity, and function of proteasomes under unstressed conditions. However, the impact of Ecm29 on proteasomes becomes noticeably more apparent upon H_2O_2-induced stress. Here we demonstrated that Ecm29 has a conserved function in regulating oxidative stress-triggered 26S proteasome disassembly in eukaryotes, which has been shown to be important for cell survival, particularly for recovery from oxidative stress (13). In addition, we have shown that increased interaction between Ecm29 and the 19S RP is directly associated with remodeling of the 26S proteasome and that Ecm29–proteasome interaction is regulated by oxidative stress. These results indicate that the modes of Ecm29 function are diverse and depend on cellular conditions. Moreover, we found that overexpression of Ecm29 alone could not induce the same level of effects on the 26S proteasome as oxidative stress, suggesting that additional signal(s) would be needed for the recruitment of Ecm29 to dissociate the 20S CP from the 19S RP. It has been shown that oxidation of cysteine residues within proteasome subunits can activate 20S proteolytic activities by inducing gate opening (40) or modulate proteasome disassembly in yeast upon mitochondrial stress (15). Future studies are needed to determine whether protein oxidation and Ecm29 work hand in hand to reshape the structure of the 26S proteasome during oxidative stress.

We combined XL-MS studies and integrative structural modeling to explore the action mechanisms of Ecm29. Our XL-MS studies provide the first physical evidence at peptide resolution to model the positions of Ecm29 on the proteasome structure. It is not surprising that Ecm29 can have multiple contact sites on the proteasome, as it contains

1845 amino acids in the form of HEAT-like repeats throughout its sequence (41). In addition, EM imaging at low resolution has indicated that Ecm29 has an elongated and curved structure (16). Among the five 19S base subunits, Rpt5 seems to be the one that interacts most intimately with Ecm29, based on the number of detected Ecm29–Rpt5 cross-links. This is supported by previous studies showing that Ecm29 localizes in close proximity to Rpt5 in yeast (11, 25). We computed a structural model of the Ecm29–proteasome complex and show that either one of two similar states are consistent with the input cross-link data as well as atomic models of Ecm29 and the proteasome. The Ecm29–19S model shows an elongated structure for Ecm29, forming contacts with Rpt1, Rpt4, Rpt5, Rpn1, and Rpn10, with its C terminus reaching 20S, in agreement with an earlier EM study (16).

Apart from Rpt5, the 20S subunit α7 has been shown to interact with Ecm29 in yeast, depending on the phosphorylation of the α7 tail (15). Interestingly, α7 phosphorylation is constitutive and has been shown to be important for modulating the stability of the CP–RP interactions in a human system (42) but has not been associated with Ecm29 function. We found that phosphorylation of the α7 tail at Ser-250 did not change in response to oxidative stress (data not shown), suggesting that it might not be important for Ecm29 interaction during oxidative stress. In addition, integrative structure modeling has indicated that oxidative stress-mediated proteasome-bound Ecm29 is not in close proximity to α7. Indeed, either localization of Ecm29 on the 19S proteasome suggests that the closest distance between Ecm29 and the α7 subunit of 20S is more than 30 Å. However, this finding is not completely unexpected, as our XL-MS experiments were designed to localize Ecm29 on the 19S and not the 20S or 26S proteasomes.

XL-MS analysis of affinity-purified complexes often reveals multiple conformational states (11). The purified proteasome sample for DSSO cross-linking in this study is expected to contain two major populations of proteasomes: free 19S RP and the 26S proteasome. Each of the two good-scoring models of Ecm29–19S satisfies the input data equally well (**Fig. 5.6**), including six distinct interprotein cross-links (**Fig. 5.6C**, right panel). The major difference between the two clusters is that cluster 1 suggests a closer interaction between Ecm29 and Rpn1, whereas cluster 2 suggests a closer interaction with Rpt1. Although it is possible that we simply did not collected sufficient information to determine the Ecm29–19S structure precisely, it is also conceivable that Ecm29 interacts with the proteasome in multiple (at least two) conformations to fulfill its role in modulating the disassembly of the 26S proteasome upon oxidative stress. The two localizations of Ecm29 on the 19S RP suggest a possible mechanism for the dissociative role of Ecm29 on the proteasome under oxidative stress: recognition of the 19S (cluster 1, state 1) and its inhibition of the 20S–19S interaction (cluster 2, state 2) (**Fig. 5.7**). Thus, these binding modes imply two sequential events: the recruitment of Ecm29 to trigger the 26S proteasome disassembly and relocalization of Ecm29 on the 19S proteasome to block 26S proteasome reassembly. As shown, an increased amount of 19S-bound Ecm29 would lead to elevated competition between Ecm29 and 20S for the binding site on the 19S, thus keeping the 20S and the 19S separated after Ecm29-mediated dissociation upon oxidative stress.

In summary, we examined oxidative stress-triggered molecular changes in the human 26S proteasome using quantitative XAP-MS, biochemical methods, XL-MS, and integrative modeling. In addition, we were able to capture proteasome-bound Ecm29 and determined that Ecm29 binds to the 19S in response to H_2O_2 stress. Importantly, we

demonstrated the biological role of human Ecm29 in modulating 26S proteasome disassembly and mapped specific residue–residue interactions between Ecm29 and multiple 19S RP subunits. The molecular architecture of the Ecm29–proteasome complex allows us to propose a model of Ecm29-dependent regulation of the 26S proteasome during oxidative stress. This model provides a basis for further exploring the diverse roles of Ecm29 in the proteasome system.

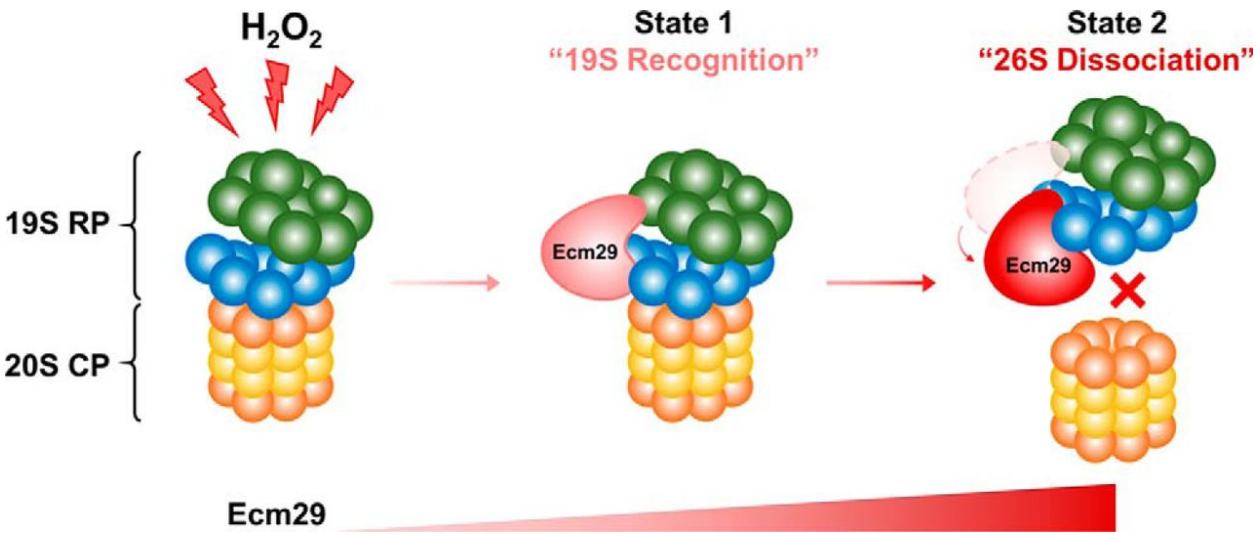

Figure 5.7 | The proposed model of Ecm29-mediated disassembly of the 26S proteasome upon H2O2 stress.
The amount of Ecm29 at the 19S proteasome increases with oxidative stress, which is illustrated with increased intensity in red.

Experimental procedures

Materials

Regular DMEM, SILAC DMEM (deficient in lysine and arginine), ImmunoPure streptavidin, horseradish peroxidase– conjugated antibody, Super Signal West Pico chemiluminescent substrate, and TurboFect transfection reagent were obtained from

Thermo Fisher Scientific. [$^{13}C_6$ $^{15}N_4$]arginine and [$^{13}C_6$ $^{15}N_2$]lysine were purchased from Cambridge Isotope Laboratories. [$^{12}C_6$ $^{14}N_4$]arginine, [$^{12}C_6$ $^{14}N_2$]lysine, anti-FLAG M2 affinity gel, and Ecm29 (KIAA0368) MISSION® shRNA bacterial glycerol stocks (catalog nos. TRCN0000263353 and TRCN0000263355) were obtained from Sigma. MISSION® pLKO.1-puro non-target shRNA bacterial glycerol stocks (catalog no. SHC016-1EA) were a kind gift from Dr. Anand Ganesan at the University of California, Irvine. Antibodies against human Rpt6 and Pre10 were obtained from Biomol International. Initially, human Ecm29 antibody was a kind gift from Dr. Carlos Gorbea (University of Utah, School of Medicine); later on it was purchased from Thermo Fisher Scientific. Ubiquitin antibody was from Santa Cruz Biotechnology. Endoproteinase Lys-C was from Wako Chemicals. Sequencing-grade trypsin was purchased from Promega. The proteasome substrates SUC-LLVY-AMC, SUC-LLE-AMC, and SUC-ARR-AMC were purchased from Boston Biochem. All other general chemicals for buffers and culture media were purchased from Thermo Fisher Scientific or VWR International.

Generation of Ecm29 knockdown cells and BAG6 knockdown cells

Lentiviruses were produced and knockdown cells were generated as described previously (43). Briefly, lentiviruses were generated by transfecting HEK293 cells with the pLKO.1-Ecm29shRNA vectors together with the packaging vectors pMDG and pCMVΔR8.91. Lentiviruses were collected 24 and 48 h post-transfection for target cell infection. 293Rpn11-TB (Hygro) (26) cells were transduced with recombinant lentivirus and selected with 2.5 μg/ml puromycin 48 h after viral infection to produce the stable cell line expressing Ecm29shRNA ($293^{Rpn11\text{-}TB\_Ecm29KD}$). Bag6 knockout cells were generated by

using CRISPR technology (44) from $293^{Rpn11\text{-}TB(Hygro)}$ cells to get $293^{Rpn11\text{-}TB\_Bag6KO}$ cell lines.

Cloning of pQCXIP-HBTH-Ecm29

Ecm29 was PCR-amplified using FLAG-Ecm29 as the template with the following primers: forward, TTAATTAACGCTGGAAAGGCCGGTGAAGGTG; reverse, GAATTCTCACATCCCTAACTCTCCTT-GAAAG. The CSN5 fragment in pQCXIP-HBTH-CSN5 (45) was removed, and Ecm29 PCR fragment was inserted. Cell culture and purification of human 26S proteasomes Nine cell lines ($293^{Rpn11\text{-}HTBH}$, $293^{Rpn11\text{-}TB}$, $^{293HBTH\text{-}Rpt6}$, $293^{\alpha7/Pre10\text{-}HTBH}$, $293^{HBTH\text{-}Ecm29}$, $293^{Rpn11\text{-}TB\_controlKD}$, $293^{Rpn11\text{-}TB\_Ecm29KD\#53}$, $293^{Rpn11\text{-}TB\_Ecm29KD\#53}$, and $293^{Rpn11\text{-}TB\_Bag6KO}$) were used in this work as listed in supplemental Table S1. Cells were grown to (90%) confluence in DMEM and either treated with 2 mM H_2O_2 for 30 min or left untreated as a control. Prior to harvesting, cells were incubated with 0.05% FA for 10 min at 37 °C. The human 26S proteasome was purified by binding to streptavidin–agarose resin (23), which was on-bead digested for MS analysis or eluted with SDS loading buffer for Western blotting. For SILAC experiments, stable cell lines were grown in SILAC DMEM as described previously (24). A Mix-After-Purification SILAC strategy was used to compare proteasome compositions before and after treatment (24).

Transient transfection and affinity purification of FLAG–Ecm29 and FLAG–Ubl4A

$293^{Rpn11\text{-}HTBH}$ cells were transiently transfected with FLAG– Ecm29 or FLAG–Ubl4A using TurboFect transfection reagent as described in the protocol of the manufacturer (Thermo Fisher Scientific). After 24 h, cells were treated with 2 mM H_2O_2 at 37 °C for 30 min or left untreated as a control, followed by 0.05% FA incubation for 10 min at 37 °C in PBS before harvesting. The respective Ecm29 and Ubl4A complexes were affinitypurified by anti-

FLAG M2 affinity gel and eluted with 0.1 M glycine following the protocol of the manufacturer (Sigma).

Proteasome proteolytic activity assay

In-solution proteolytic activity assays for human proteasomes in cell lysates were performed with the fluorogenic pep tide substrates SUC-LLVY-AMC, SUC-LLE-AMC, and SUCARR-AMC as described previously (23).

Quantitative immunoblot analysis

The purified proteasome complexes were analyzed by Western blotting as described previously (13). Primary antibodies against Rpt6, α6/MCP20, α7/Pre10, Ecm29, and Bag6 were utilized, followed by an HRP-conjugated mouse or rabbit secondary antibody against mouse IgG. Protein bands were detected and quantified using a Fuji LAS4000 scanning system (Fujifilm Life Sciences).

Protein identification and quantification by MS

Purified proteasome complexes were digested in-solution with Lys-C/trypsin and analyzed by LC/MS-MS using an EasynLC 1000 coupled with a linear ion trap (LTQ) Orbitrap XL mass spectrometer (Thermo Fisher, San Jose, CA) as described previously (13). The LC/MS-MS data were searched using Batch-Tag within a developmental version (v. 5.17.0) of Protein Prospector at the University of California, San Francisco against a decoy database consisting of a normal SwissProt database concatenated with its randomized version (SwissProt.2013.06.17.random.concat with a total of 455,294 protein entries) (13). Proteins were identified by at least two peptides with an FDR of ≤ 0.5%.

For SILAC experiments, the Search Compare program within Protein Prospector was used to calculate the relative abundance ratios of Arg/Lys-containing peptides based on ion intensities of monoisotopic peaks observed in the LC/MS spectra when the peptides were sequenced and subsequently identified during database searching as described previously (13, 24).

DSSO cross-linking of Ecm29–proteasome complexes

$293^{HBTH\text{-}Rpt6}$ cells were transiently transfected with HBTH–Ecm29. After 24 h, the cells were treated with 5 mM H_2O_2 at 37 °C for 30 min to maximize the interaction between Ecm29 and 19S RP, followed by 0.025% FA at 37 °C in PBS for 10 min. Single-step affinity purification of the human Ecm29–proteasome complexes was achieved by binding to streptavidin–agarose resin. The bound protein complexes were cross-linked on-bead in PBS buffer (pH 7.5) with 0.5 mM DSSO for 1 h at 37 °C and then quenched, reduced/alkylated, and digested as reported previously (11, 30). The resulting peptide mixture was extracted and desalted prior to LC/MSn analysis.

LC/MS^n analysis of DSSO cross-linked peptides

LC MS^n analysis was performed using a Thermo Scientific™ EASY-nLC™ 1200 ultrahigh pressure liquid chromatography (UPLC) system coupled with an Orbitrap Fusion Lumos™ MS (46). Briefly, a 25 cm x 75 μm PepMap EASY-Spray column was used to separate peptides over acetonitrile gradients of 6% to 35% at a flow rate of 300 nl/min. Two different types of acquisition methods were utilized to maximize the identification of DSSO cross-linked peptides: top four data-dependent MS^3 and targeted MS^3 acquisition (46). Two biological replicates were analyzed, and each of them was analyzed with at least two technical replicates.

Data analysis to identify DSSO cross-linked peptides

MS^n data were extracted, searched, and analyzed as described previously (11). Briefly, MS^3 data were subjected to Batch-Tag against a decoy database consisting of a normal SwissProt database concatenated with its randomized version (SwissProt.2014.12.4.random.concat with a total of 20,194 protein entries). Peptides were identified from MS^3 data with a FDR of 1.7%. MS^n data and MS^3 database search results were integrated in xl-Discoverer (an in-house script) to automatically generate, summarize, and validate identified cross-linked peptide pairs. The final FDR of identified interlinked peptides was determined to be 0.1%. The reduction in FDR for the identification of cross-linked peptides occurs as a result of MSn data integration, which improves identification accuracy.

Integrative modeling of the Ecm29–proteasome complex

Comparative and integrative modeling was carried out to elucidate the architecture of the human Ecm29–26S proteasome complex (11) (supplemental Methods).

Author contributions

L. H. conceived the study and directed the research. X. W., I. E. C., A. S., Y. Y., and L. H. designed the experiments. X. W. generated constructs and stable cell lines and carried out quantitative XAP-MS, biochemical, and XL-MS experiments and data analyses. C. Y. and R. V. performed the LC/MS^n analysis. C. Y. analyzed the XL-MS data. I. E. C. performed integrative structure modeling. P. C. did the initial modeling analysis. X. W., C. Y., and I. E. C. prepared figures and tables. A. H. contributed to XL-MS data analysis. S. A. B. and S. D. R. synthesized the cross-linking reagent. Y. X. and Y. Y. provided

biochemical reagents and generated CRISP-based Bag6 knockout cells. X. W., I. E. C., C. Y., Y. Y., A. S., and L. H. contributed to the writing of the manuscript.

Acknowledgments

We thank Prof. A. L. Burlingame and Dr. Robert Chalkley (University of California, San Francisco) for support of the development version of Protein Prospector.

References

1. K. J. Barnham, C. L. Masters, A. I. Bush, Neurodegenerative diseases and oxidative stress. *Nat Rev Drug Discov* **3**, 205–214 (2004).
2. V. Byvaltsev, *et al.*, Nanostructural changes of intervertebral disc after diode laser ablation. *World Neurosurg* **77**, 6–7 (2012).
3. C. T. Aiken, R. M. Kaake, X. Wang, L. Huang, Oxidative stress-mediated regulation of proteasome complexes. *Mol Cell Proteomics* **10**, R110 006924 (2011).
4. M. G. Goebl, *et al.*, The yeast cell cycle gene CDC34 encodes a ubiquitin-conjugating enzyme. *Science* **241**, 1331–5 (1988).
5. D. Voges, P. Zwickl, W. Baumeister, The 26S proteasome: a molecular machine designed for controlled proteolysis. *Annu Rev Biochem* **68**, 1015–68 (1999).
6. D. Finley, Recognition and processing of ubiquitin-protein conjugates by the proteasome. *Annu Rev Biochem* **78**, 477–513 (2009).
7. G. C. Lander, *et al.*, Complete subunit architecture of the proteasome regulatory particle. *Nature* **482**, 186–91 (2012).
8. K. Lasker, *et al.*, Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. *Proc Natl Acad Sci USA* **109**, 1380–1387 (2012).
9. A. Schweitzer, *et al.*, Structure of the human 26S proteasome at a resolution of 3.9 A. *Proc Natl Acad Sci U S A* (2016).
10. X. Huang, B. Luan, J. Wu, Y. Shi, An atomic structure of the human 26S proteasome. *Nat Struct Mol Biol* (2016).

11. X. Wang, *et al.*, Molecular Details Underlying Dynamic Structures and Regulation of the Human 26S Proteasome. *Mol Cell Proteomics* **16**, 840–854 (2017).
12. G. Ben-Nissan, M. Sharon, Regulating the 20S proteasome ubiquitin-independent degradation pathway. *Biomolecules* **4**, 862–84 (2014).
13. X. Wang, J. Yen, P. Kaiser, L. Huang, Regulation of the 26S proteasome complex during oxidative stress. *Sci Signal* **3**, ra88 (2010).
14. T. Grune, *et al.*, HSP70 mediates dissociation and reassociation of the 26S proteasome during adaptation to oxidative stress. *Free Radic Biol Med* **51**, 1355–64 (2011).
15. N. Livnat-Levanon, *et al.*, Reversible 26S proteasome disassembly upon mitochondrial stress. *Cell Rep* **7**, 1371–80 (2014).
16. D. S. Leggett, *et al.*, Multiple associated proteins regulate proteasome structure and function. *Mol Cell.* **10**, 495–507 (2002).
17. M. F. Kleijnen, *et al.*, Stability of the proteasome can be regulated allosterically through engagement of its proteolytic active sites. *Nat Struct Mol Biol* **14**, 1180–8 (2007).
18. S. Park, W. Kim, G. Tian, S. P. Gygi, D. Finley, Structural defects in the regulatory particle-core particle interface of the proteasome induce a novel proteasome stress response. *J Biol Chem* **286**, 36652–66 (2011).
19. S. Y. Lee, A. De la Mota-Peynado, J. Roelofs, Loss of Rpt5 protein interactions with the core particle and Nas2 protein causes the formation of faulty proteasomes that are inhibited by Ecm29 protein. *J Biol Chem* **286**, 36641–51 (2011).

20. C. Gorbea, G. M. Goellner, K. Teter, R. K. Holmes, M. Rechsteiner, Characterization of mammalian Ecm29, a 26 S proteasome-associated protein that localizes to the nucleus and membrane vesicles. *J Biol Chem* **279**, 54849–61 (2004).
21. C. Gorbea, *et al.*, A protein interaction network for Ecm29 links the 26 S proteasome to molecular motors and endosomal components. *J Biol Chem* **285**, 31616–33 (2010).
22. C. Gorbea, M. Rechsteiner, J. G. Vallejo, N. E. Bowles, Depletion of the 26S proteasome adaptor Ecm29 increases Toll-like receptor 3 signaling. *Sci Signal* **6**, ra86 (2013).
23. X. Wang, *et al.*, Mass spectrometric characterization of the affinity-purified human 26S proteasome complex. *Biochemistry* **46**, 3553–65 (2007).
24. X. Wang, L. Huang, Identifying dynamic interactors of protein complexes by quantitative mass spectrometry. *Mol Cell Proteomics* **7**, 46–57 (2008).
25. C. Guerrero, T. Milenkovic, N. Przulj, P. Kaiser, L. Huang, Characterization of the proteasome interaction network using a QTAX-based tag-team strategy and protein interaction network analysis. *Proc Natl Acad Sci U S A* **105**, 13333–8 (2008).
26. C. Yu, *et al.*, Characterization of Dynamic UbR-Proteasome Subcomplexes by In vivo Cross-linking (X) Assisted Bimolecular Tandem Affinity Purification (XBAP) and Label-free Quantitation. *Mol Cell Proteomics* (2016).
27. T. Akahane, K. Sahara, H. Yashiroda, K. Tanaka, S. Murata, Involvement of Bag6 and the TRC pathway in proteasome assembly. *Nat Commun* **4**, 2234–2234 (2013).
28. J. Binici, J. Koch, BAG-6, a jack of all trades in health and disease. *Cell Mol Life Sci* **71**, 1829–1837 (2014).

29. Q. Wang, *et al.*, A ubiquitin ligase-associated chaperone holdase maintains polypeptides in soluble states for proteasome degradation. *Mol Cell* **42**, 758–770 (2011).
30. A. Kao, *et al.*, Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. *Mol Cell Proteomics* **10**, M110.002212 (2011).
31. P. Upla, *et al.*, Molecular Architecture of the Major Membrane Ring Component of the Nuclear Pore Complex. *Structure* **25**, 434–445 (2017).
32. J. Fernandez-Martinez, *et al.*, Structure and Function of the Nuclear Pore Complex Cytoplasmic mRNA Export Platform. *Cell* **167**, 1215–1228 (2016).
33. A. Sali, *et al.*, Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure* **23**, 1156–67 (2015).
34. Y. Shi, *et al.*, Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. *Mol Cell Proteomics* **13**, 2927–43 (2014).
35. Y. Shi, *et al.*, A strategy for dissecting the architectures of native macromolecular assemblies. *Nat Methods* **12**, 1135–8 (2015).
36. A. Sali, T. L. Blundell, Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779–815 (1993).
37. J. Soding, A. Biegert, A. N. Lupas, The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* **33**, W244-8 (2005).
38. R. I. Subbotin, B. T. Chait, A pipeline for determining protein-protein interactions and proximities in the cellular milieu. *Mol Cell Proteomics* **13**, 2824–35 (2014).

39. M. A. Collart, O. O. Panasenko, The Ccr4–not complex. *Gene* **492**, 42–53 (2012).

40. G. M. Silva, *et al.*, Redox control of 20S proteasome gating. *Antioxid Redox Signal* **16**, 1183–94 (2012).

41. A. V. Kajava, C. Gorbea, J. Ortega, M. Rechsteiner, A. C. Steven, New HEAT-like repeat motifs in proteins regulating proteasome structure and function. *J Struct Biol* **146**, 425–430 (2004).

42. S. Bose, F. L. Stratford, K. I. Broadfoot, G. G. Mason, A. J. Rivett, Phosphorylation of 20S proteasome alpha subunit C8 (alpha7) stabilizes the 26S proteasome and plays a role in the regulation of proteasome complexes by gamma-interferon. *Biochem J* **378**, 177–84 (2004).

43. K. J. Shin, *et al.*, A single lentiviral vector platform for microRNA-based conditional RNA interference and coordinated transgene expression. *Proc Natl Acad Sci U S A* **103**, 13759–64 (2006).

44. F. A. Ran, *et al.*, Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380–1389 (2013).

45. L. Fang, *et al.*, Characterization of the human COP9 signalosome complex using affinity purification and mass spectrometry. *J Proteome Res* **7**, 4914–25 (2008).

46. C. Yu, *et al.*, Developing a Multiplexed Quantitative Cross-Linking Mass Spectrometry Platform for Comparative Structural Analysis of Protein Complexes. *Anal Chem* **88**, 10301–10308 (2016).

Chapter VI - Structure and Function of the Nuclear Pore Complex Cytoplasmic mRNA Export Platform

Contributing authors

Javier Fernandez-Martinez[1,8], Seung Joong Kim[2,8], Yi Shi[3,8], Paula Upla[4,8], Riccardo Pellarin[2,7,8], Michael Gagnon[5], Ilan E. Chemmama[2], Junjie Wang[3], Ilona Nuldelman[1], Wenzhu Zhang[3], Rosemary Williams[1], William J. Rice[6], David L. Stokes[4], Daniel Zenklusen[5], Brian T. Chait[3,*], Andrej Sali[2,*], and Michael P. Rout[1,9,*]

[1]Laboratory of Cellular and Structural Biology, The Rockefeller University, New York, NY 10065, USA

[2]Department of Bioengineering and Therapeutic Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biosciences, University of California, San Francisco, California 94158

[3]Laboratory of Mass Spectrometry and Gaseous Ion Chemistry, The Rockefeller University, New York, NY 10065, USA

[4]Skirball Institute of Biomolecular Medicine, Department of Cell Biology, New York University School of Medicine, New York, NY 10016, USA

[5]Département de Biochimie et Médecine Moléculaire, University of Montréal, Montréal, QC H3C3J7, Canada

[6]Simons Electron Microscopy Center at New York Structural Biology Center, New York, NY 10027, USA

[7]Structural Bioinformatics Unit, Institut Pasteur, CNRS UMR 3528, 75015 Paris, France

[8]Co-first author

[9]Lead contact

[*]Contacts: chait@rockefeller.edu (B.T.C.), sali@salilab.org (A.S.), and rout@rockefeller.edu (M.P.R.)

Abstract

The last steps in mRNA export and remodeling are performed by the Nup82 complex, a large conserved assembly at the cytoplasmic face of the nuclear pore complex (NPC). By integrating diverse structural data, we have determined the molecular architecture of the native Nup82 complex at subnanometer precision. The complex consists of two compositionally identical multiprotein subunits that adopt different configurations. The Nup82 complex fits into the NPC through the outer ring Nup84 complex. Our map shows that this entire 14-MDa Nup82- Nup84 complex assembly positions the cytoplasmic mRNA export factor docking sites and messenger ribonucleoprotein (mRNP) remodeling machinery right over the NPC's central channel rather than on distal cytoplasmic filaments, as previously supposed. We suggest that this configuration efficiently captures and remodels exporting mRNP particles immediately upon reaching the cytoplasmic side of the NPC.

Introduction

The nuclear pore complex (NPC) is a large cylindrical structure with eight symmetrically arranged spokes embedded in the nuclear envelope (NE) and is composed of multiple copies of ~30 different nucleoporins (Nups). Discrete Nup subcomplexes associate to form the different substructures of the NPC, consisting of coaxial outer, inner, and membrane rings surrounding a central channel and linked to peripheral components such as the nuclear basket. Approximately one-third of all Nups, termed FG Nups, contain intrinsically disordered domains comprising multiple Phe-Gly (FG) repeats between hydrophilic spacers. These FG repeat regions populate the NPC central channel and, through their specific interaction with cargo-carrying transport factors, mediate transport (1).

Although much of transport across the NPC is mediated by the karyopherin family of transport factors, the export of mRNAs follows a different mechanism that requires a special platform located at the cytoplasmic face of the NPC, called the Nup82 complex in budding yeast (2), which in turn associates with Dyn2, Nup116, Gle2, and Gle1 (3). The central role of this complex is underscored by the fact that its mammalian homolog, the Nup88 complex, is a nexus for disease-associated mutations (4, 5). The Nup82 complex and its associated proteins have proven challenging for structural analyses due to their flexibility and the presence of intrinsically disordered domains. The core of the Nup82 complex is composed of the proteins Nup82, Nup159, and Nsp1. Fragments of each have been solved crystallographically (6–8), and negative stain electron microscopy (EM) revealed this complex to have an overall ‘‘P’’-shaped morphology (9), but no structures

exist for either the whole complex or how it interacts with its associated proteins and the NPC.

mRNA export is achieved in several stages. First, mRNAs, packaged into export-competent messenger ribonucleoprotein (mRNP) particles, are docked into the nuclear basket; the mRNP particle then travels across the NPC through interaction of the non-karyopherin transport factors Mex67-Mtr2 with FG repeats that fill the NPC's central channel (2). Once the mRNP particle reaches the cytoplasmic face of the NPC, the coordinated action of the DEAD-box RNA helicase Dbp5, the nucleoporin Gle1, and the N-terminal b-propeller of Nup159 leads to active remodeling of the mRNP (3, 10). Mex67-Mtr2 and other transport factors are removed during remodeling (11), preventing the mRNA from traveling back to the nucleus. In the final stage, the remodeled mRNA is released into the cytoplasm for translation.

Unfortunately, the precise coordination of these processes at the molecular scale has not been elucidated, in large part due to the lack of sufficiently detailed information on the spatial arrangement of transport and remodeling components relative to each other and the NPC. Localization studies have led to the proposal that the Nup82 complex forms filaments that project orthogonally from the cytoplasmic face of the NPC; such a location would imply that exporting mRNPs must first transit the central channel of the NPC before being transferred out to these peripheral cytoplasmic filaments, where the final stages of mRNP remodeling and export would occur distally from the central channel of the NPC (reviewed in (1–3). However, exactly how this transfer would be accomplished, and how central channel transit and mRNP processing could be coordinated, remained unclear.

To understand these processes, we solved the structure of the endogenous Nup82 complex by using an integrative approach that relies on multiple structural and proteomic data sources (12, 13). We also determined how the Nup82 complex is anchored to the cytoplasmic face of the NPC via the Nup84 complex, a seven-member assembly forming the outer rings. In addition, we used a combined structural and functional mapping analysis to elucidate the major mechanism responsible for mRNA export defects affecting Nup84 complex components. Finally, we integrate our data into a detailed map of the whole cytoplasmic mRNA export and remodeling machinery. We show that, surprisingly, the Nup82 complex positions the cytoplasmic FG repeats and mRNP remodeling machinery right over the NPC's central channel rather than on distal cytoplasmic filaments, as previously supposed.

Results

Solving the Structure of the Endogenous Nup82 Holocomplex

We solved the structure of the endogenous native Nup82 holocomplex (**Fig. 6.1**) using an integrative modeling approach that has previously allowed us and others to successfully determine the molecular architecture of numerous other large native assemblies (14). Such integrative strategies have proven to be suited for the structural analysis of large endogenous complexes that are by nature flexible, contain unstructured regions, and are conformationally heterogeneous (13, 15).

We measured the native stoichiometry of the purified Nup82 holo-complex by a combination of QConCAT-MS (16) and classical Siegel and Monte biophysical measurements (Fig. S1; Methods). The consensus of our analyses results in a

stoichiometry of 2:2:2:2 (Nup159:Nup82:Nsp1:Dyn2), consistent with that previously measured (9) for a truncated overexpressed version of the complex, with the exception of the Dyn2 dimer, a labile component that, unless overexpressed (Fig. S1E), is present as a single dimer in the average native complex. The morphology and dimensions of the complex were determined by negative stain EM, where 4,266 particles were classified into 23 class averages (Fig. S2C); a majority of these (21) showed what appears to be a single dimer of Dyn2, in agreement with a previous study (9) and with our stoichiometry (see above), and were thus included in the calculation. Interestingly, two of the class averages seemingly presented two consecutive dimers of Dyn2 (Fig. S2C, arrowheads), underscoring the previously observed heterogeneity of the complex *in vivo* (9). Instead of using a highly uncertain 3D map computed via single-particle reconstruction based on a heterogenous set of images, we relied on much more robustly computed 2D class averages, following a previously demonstrated procedure (13). Only the structured portions of the complex were constrained by the EM data, because we showed that the unstructured FG repeats are not revealed by negative stain EM (Fig. S2D).

All components of the complex were used in the final calculation, including FG repeats to account for their excluded volume and emanating points. Protein representations were derived from the atomic structures in the Protein Data Bank, where available, or comparative models were built with MODELER 9.13 (17) based on the closest homolog with a known structure detected by HHPred (18) (Fig. S3; Table S1); disordered FG-repeat-containing regions were modeled as flexible strings of beads, guided by our recent nuclear magnetic resonance (NMR) data (19). Finally, the residue-specific spatial proximity and orientation of the different subunits were determined by a

comprehensive chemical crosslinking with mass spectrometry readout (CX-MS) method, using two complementary cross-linkers (Fig. 2A and S2A) (13). To reduce the intrinsic ambiguity of cross-link data arising from the presence of two copies of each protein, we also analyzed a strain expressing an exogenous homolog of Nup82 (skNup82) from the yeast *Saccharomyces kudriavzevii* (20) (Fig. S2A; Methods), whose distinct protein sequence allows crosslinks to it to be distinguished from the endogenous Nup82. We identified a total of 1,131 cross-links (Table S2) that include 662 unique disuccinimidyl suberate (DSS) and 126 unique 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride (EDC) cross-links from the wild-type yeast strain and 343 unique DSS cross-links from the skNup82-containing complex (Fig. S2A). The majority of the identified inter-molecular cross-links mapped to the coiled coil, C-terminal regions of Nup159 and Nsp1 and the whole Nup82 and Dyn2 proteins. Few inter-molecular cross-links were found to connect to the FG regions of Nup159 or Nsp1 and none connected to the b-propeller domain of Nup159, strongly indicating that those domains are dynamic, peripheral, and not located in proximity to the core of the complex (9).

We computed the structure of the Nup82 complex (**Fig. 6.1**) through our integrative modeling approach as implemented in the Integrative Modeling Platform (IMP) program (21) using the data described above. A detailed assessment of the input data and the resulting model are shown in Table 1 and Methods. In summary, the 463 best-scoring solutions satisfy within stringent tolerances the data used to compute them. The clustering analysis of the best-scoring solutions identified a single dominant cluster of 370 similar structures. The corresponding localization probability density map represents the probability of any volume element being occupied by a given protein (**Fig. 6.1**). The 9.0

Å precision of the core structured region is sufficiently high to pinpoint the locations and orientations of the constituent proteins and domains, demonstrating the quality of the input data, including the cross-links and EM 2D class averages (Fig. S4; **Table 6.1**).

Our structure is validated by seven considerations as follows. First, the EDC and DSS cross-links are highly consistent with each other, despite different chemistries, and there is significant highly non-random clustering of both EDC and DSS cross-links into equivalent ‘‘cliques’’ (**Fig. 6.2A**). These represent immediately adjacent regions in the complex, as validated by those cliques that coincide with known crystallographic interface regions, such as Nup159:Dyn2 (PDB: 4DS1) (22) and Nup159:Nup82 (PDB: 3PBP) (8) (**Fig. 6.2B**); indeed, in our final calculated structure these cliques represent immediately adjacent regions in the complex. Second, those few cross-links in violation of strict distance limits in our structure are nevertheless right next to one of the cliques; they are thus consistent with the structure when locally limited flexibility is taken into account (**Fig. 6.2A** and S4D). Third, mass tagging of our structure is consistent with the localization of GFP tags on both the Nup82 and Nup159 C termini (**Fig. 6.2C**). Fourth, our structure is consistent with the previously published data, including an independent negative stain 3D density map (Fig. S5A) (9). Fifth, the trimeric coiled-coil structure is recapitulated even when computed using the chemical cross-linking data alone (Fig. S5C). Sixth, our structure is in agreement with small angle X-ray scattering (SAXS) profiles and ab initio shapes of Nup82 constructs spanning residues 4–220, 4–452, and 572–690 (**Fig. 6.2D** and S5D–S5F; Table S4). Notably, the Nup82 coiled-coil (572–690) forms a kinked structure, and the corresponding SAXS profile shows a monotonous increase in the Kratky plot (**Fig. 6.2D** and S5F), indicating a high degree of flexibility between coiled-coil

segments in solution, as would be expected for coiled-coils that form two different conformers seen in the final structure. Finally, our structure is also validated by the non-random and clustered distribution of cross-links connecting the Nup82 holo-complex to other parts of the NPC, revealing interaction sites, as described below.

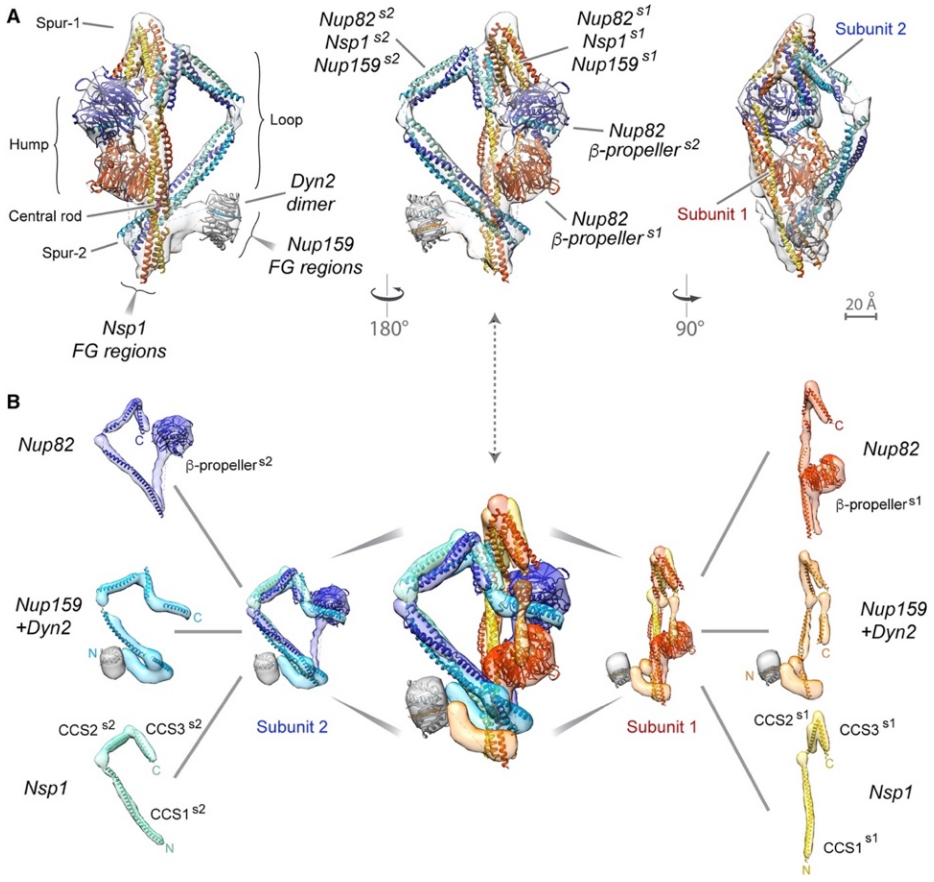

Figure 6.1 | Structure of the Core Nup82 Holo-Complex.
(**A**) Three views of the localization probability density map corresponding to the Nup82 holo-complex ensemble are shown (light gray), with a single representative ribbon structure embedded; the proteins, subunits, and different structural features of the complex are indicated. Subunit assignment is indicated with a superscript “s1” (subunit 1) or “s2” (subunit 2). In all views, the components of each subunit are colored in tones of red (subunit 1) or blue (subunit 2) (see also **B**). (**B**) Exploded view of the Nup82 holo-complex subunits and protein components, with the whole complex shown in the center and the two subunits and the different components shown on the right (subunit 1, colored in red tones) or the left side (subunit 2, colored in blue tones). CCS, coiled-coil segment (as described in the main text).

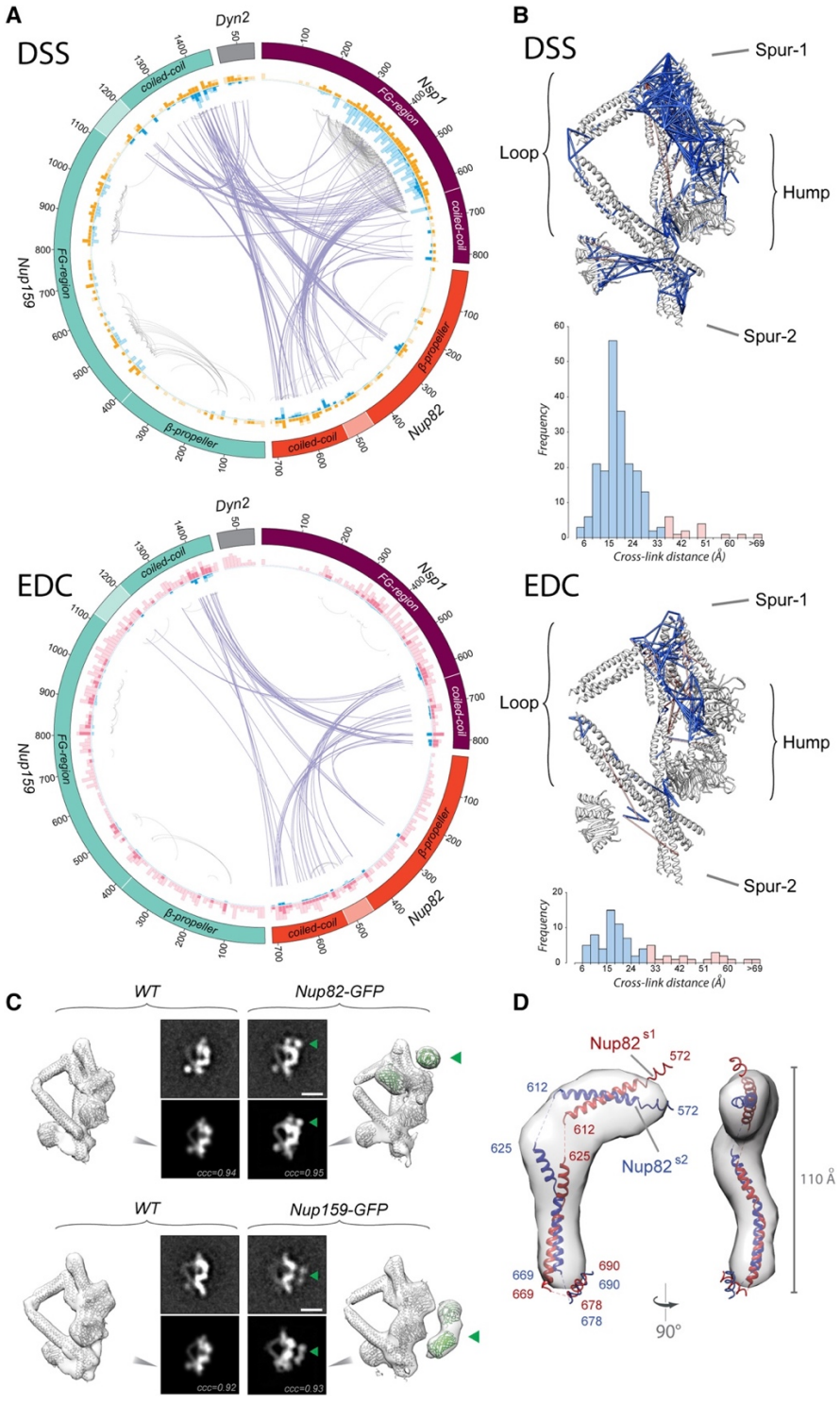

Figure 6.2 | Nup82 Holo-complex Structure Validation.
(**A**) Circos-XL plots showing the distribution of all DSS (top plot) or EDC (bottom plot) cross-links mapping within the core of the Nup82 holo-complex. Each protein is represented as a colored segment, with the amino acid residue indicated on the outside of the plot and relevant domains indicated inside each segment; regions without reliable

fold assignment are identified by lighter shading. Inter-molecular cross-links are depicted as purple lines and intra-molecular cross-links as gray lines. The internal circles include bars representing the density of cross-links per ten residues in DSS and EDC (blue and light blue color for inter-molecular cross-links and intra-molecular cross-links, respectively) and the density of lysines in DSS (orange and light orange bars for cross-linked and uncross-linked residues, respectively) or the density of lysine/carboxylic acid in EDC (pink and light pink bars for cross-linked and uncross-linked residues, respectively). (**B**) Structure of the Nup82 holo-complex showing the cross-links falling within the expected C_α-C_α maximum distance threshold (blue) or outside of that threshold (orange). Below the structure, a bar graph shows the C_α-C_α distance distribution of all DSS or EDC cross-links in the structure. DSS threshold = 35 Å; EDC threshold = 30 Å.
(**C**) GFP mass-tagging analysis of the Nup82 holo-complex. Analyses of a Nup82-GFP tagged version (top diagram) or a Nup159-GFP tagged version (bottom diagram) of the holo-complex are shown. For each diagram, a view of the native Nup82 holo-complex structure is shown (wild-type [WT]), and the tagged version of the structure shown on the right side. The top panels show a representative negative stain 2D class average of the native complex (left) and the tagged version (right; green arrowhead, GFP). The bottom panels show 2D projections of the native structure (left) and the calculated GFP-tagged version (right; green arrowhead, GFP). ccc, cross correlation coefficient. Scale bar, 10 nm. (**D**) SAXS analysis of the Nup82 (572–690) fragment, showing two views of the computed ab initio shape (gray envelope), with ribbon representations of the equivalent Nup82 fragments in the conformation they adopt within the Nup82 holo-complex; subunits 1 (red) and 2 (blue) are indicated.

Features of the Nup82 Holo-complex

The C termini of Nup82, Nup159, and Nsp1 share a common domain arrangement, formed by consecutive helical coiled-coil regions of different length, connected by flexible linkers. They assemble (together with Dyn2) to form the Nup82 holo-complex, a roughly ‘‘D’’-shaped particle, which is formed by the asymmetric assembly of two compositionally identical subunits (termed subunit 1 [s1] and subunit 2 [s2] in **Fig. 6.1**). Each subunit consists mainly of parallel, three-stranded, hetero-trimeric coiled-coils connected by flexible linkers, consisting of a single copy of the C termini of Nup82, Nup159, and Nsp1. However, the two subunits adopt different configurations, mainly due to the different degree of flexion of the hinges between heterotrimeric coiled-coil segments (termed CCSs) and the relative position of the Nup82 b-propellers. Subunit 1 mainly forms the

‘‘rod,’’ while subunit 2 forms the ‘‘loop’’ of the holo-complex, with both subunits contributing to the spurs (**Fig. 6.1**). The CCS1^{s2} and CCS2^{s2} trimers constitute the extended loop that can be observed in certain orientations of the particle (**Fig. 6.1A**, left and center). The denser region of the complex is formed by trimeric parallel CCS domains that form the slightly bent, elongated central rod. Both Nup82 b-propellers are located side by side on top of the rod formed by subunit s1, with Nup82 b-propellers2 located in trans in a distal position from the CCS1-2^{s2} loop. The two ends of the central rod are each formed by the C-terminal (spur-1) and the N-terminal (spur-2) bundles of the CCS domains. Two copies of Dyn2 form a dimer that is perpendicular with spur-2 and seems to help lock the two subunits into their asymmetric arrangement. Dyn2 also helps to orient the two Nup159 copies, so that their FG regions emanate in parallel from that end of the complex. Interestingly, the FG regions of Nsp1 also project from spur-2, forming, together with the Nup159 FGs, an intrinsically disordered plume. In agreement with prior work, the hump formed by the Nup82 b-propellers helps to lock down the C termini of Nup159 and form the attachment site for two Nup116 copies (8) (see below).

Structure of the Nup82-Nup84 Complex Assembly and the Cytoplasmic mRNA Export Platform

To understand how the Nup82 holo-complex is associated with the whole NPC, we isolated it under conditions that preserved its interaction with other Nups (23). CX-MS was used to analyze those proteins proximally associated with each of the Nup82 holo-complex’s components (Table S3). Notably, most of the identified cross-links connected the spur-1 region of the Nup82 holo-complex to components of the Nup84 complex hub (**Fig. 6.3**; Table S3) (13); indeed, a direct physical connection between the Nup82 and

Nup84 complexes was recently demonstrated in *Chaetomium thermophilum* (24). Our data, together with our prior map of the Nup84 complex (13), crystallographic data on the Nup84 complex (25, 26), and the previous map of the entire NPC (12), were sufficient to allow us to dock the two complexes together to generate a map of the entire ~1.3-MDa, 15-protein, Nup82-Nup84 complex assembly (**Fig. 6.3A**). All our solutions were similar, differing only in the degree of rotation along the Nup82 complex long axis relative to the Nup84 complex (Fig. S6). The Nup82 holo-complex body associates through its spur-1 region with the Nup85/Seh1 arm on the Y-complex hub and the N-terminal side of Nup145C (**Fig. 6.3A**), with the two complexes oriented orthogonally with respect to their long axis (**Fig. 6.3A**). Our arrangement is supported by the tight clustering of cross-links between the Nup82 and Nup84 complexes mainly to two discrete locations, one on spur-1 and the other on a single region of the Nup85-Seh1 arm, respectively.

It has been previously shown that the Nup84 complex long axis orientation is approximately parallel to the plane of the NE in the NPC's outer ring (27, 28). Consequently, our structure reveals that the Nup82 holo-complex long axis is orthogonal to that of the Nup84 complex, forming a potential linker between the outer and inner ring. The coiled-coil bundles of the Nup82 holo-complex body form a scaffold, and their downward orientation makes it so that the FG plume in spur-2 projects from the bottom of the complex. The FG regions of Nsp1 and Nup159 would thus face the central transport channel and be adjacent to the Nsp1 FG regions emanating from the inner-ring Nic96 complex (**Fig. 6.3**).

Our CX-MS analysis of the higher-order assembly also identified cross-links connecting other known components of the mRNA export machinery (Gle1, Nup42, and

Nup116) to the Nup82 holo-complex (**Fig. 6.3**; Table S3). The identified crosslinks are fully consistent with previous work showing physical connections between some of these components, such as the C-termini of Gle1 and Nup42 (29) and the C terminus of Nup116 to Nup82 (8), indicating that our CX-MS analysis is targeting bona fide physical connections within the mRNA export machinery. In combination with published crystal structures of labile components of this machinery (10, 30), our data allowed us to assemble a physical map of the whole cytoplasmic mRNA export platform comprising 16 different proteins (some in multiple copies, so comprising 24 subunits) with a mass of ~1.8 MDa (**Fig. 6.3B**). The organization of the assembly reveals that the components actively involved in the mRNP remodeling process (Dbp5, Gle1, and Nup159 N terminus) and associated FG regions (Nup42, Nup116, Nup159, and Nsp1) are localized around the Nup82 holo-complex and the short arms of the Nup84 complex. Remarkably, we identified ten cross-links connecting Gle1 to the Nup82 holo-complex, delineating for the first time the position and orientation of Gle1 in the NPC, adjacent to the Nup82 holo-complex and oriented with its N terminus toward the holo-complex hump, its middle region running parallel to spur-2, and its C-terminal and the Dbp5-interacting domain facing downward toward the NPC central channel (**Fig. 6.3**). Through its interaction with Gle1, Nup42 is also seemingly localized toward the central channel, in agreement with a recent report that showed how the FG region of Nup42 is fully functional if fused to the Gle1 C terminus (31). Thus, in our map, both the core of the Nup82 holo-complex and the Nup84 complex form a flexible scaffold, which organizes and properly orients the two functional ends (FG regions and enzymatic activities) of the cytoplasmic mRNA export machinery.

Functional Relationship between the Nup82 Holocomplex and the Nup84 Complex

To functionally annotate our Nup82-Nup84 complex assembly structure, we sought to investigate its relationship to mRNA export. Mutations affecting both Nup84 and Nup82 complex components have previously been shown to display characteristic mRNA export defects (33). Although the direct involvement of components of the Nup82 holo-complex in mRNA export has been long established (33), until now, the association of mRNA defects with the Nup84 complex has remained unclear. Thus, to identify regions of the Nup84 complex that are most relevant for mRNA export, we analyzed a collection of truncation mutants (23). The mRNA export defect of each mutant was quantified and heat-mapped into the Nup84 complex structure (**Fig. 6.4** and S7). We detected a clear hotspot mapping to the Nup85/Seh1 arm (**Fig. 6.4**), different from those determined for other Nup84 complex phenotypes (23). Notably, this hotspot maps to where the Nup85-Seh1 arm connects to the Nup82 holo-complex (**Fig. 6.3**). This significant structure-function correlation supports the idea that the mRNA export phenotype, focused to this part of the Nup84 complex, is largely associated with a defective incorporation of the Nup82 complex into the NPC. To test this idea, we analyzed the in vivo localization of Nup82- GFP in several Nup84 complex truncation mutants affecting different parts of the Y-shaped complex. As shown in **Fig. 6.4**, the Nup82-GFP construct is indeed significantly mislocalized to the cytoplasm only in mutations affecting the Nup85/Seh1 arm, while a control Nup49-GFP reporter did not show similar behavior (23). Thus, we conclude that the mRNA export phenotype found in Nup84 complex mutants is mainly the consequence of a defective or weakened incorporation of the Nup82 holo-complex into the NPC.

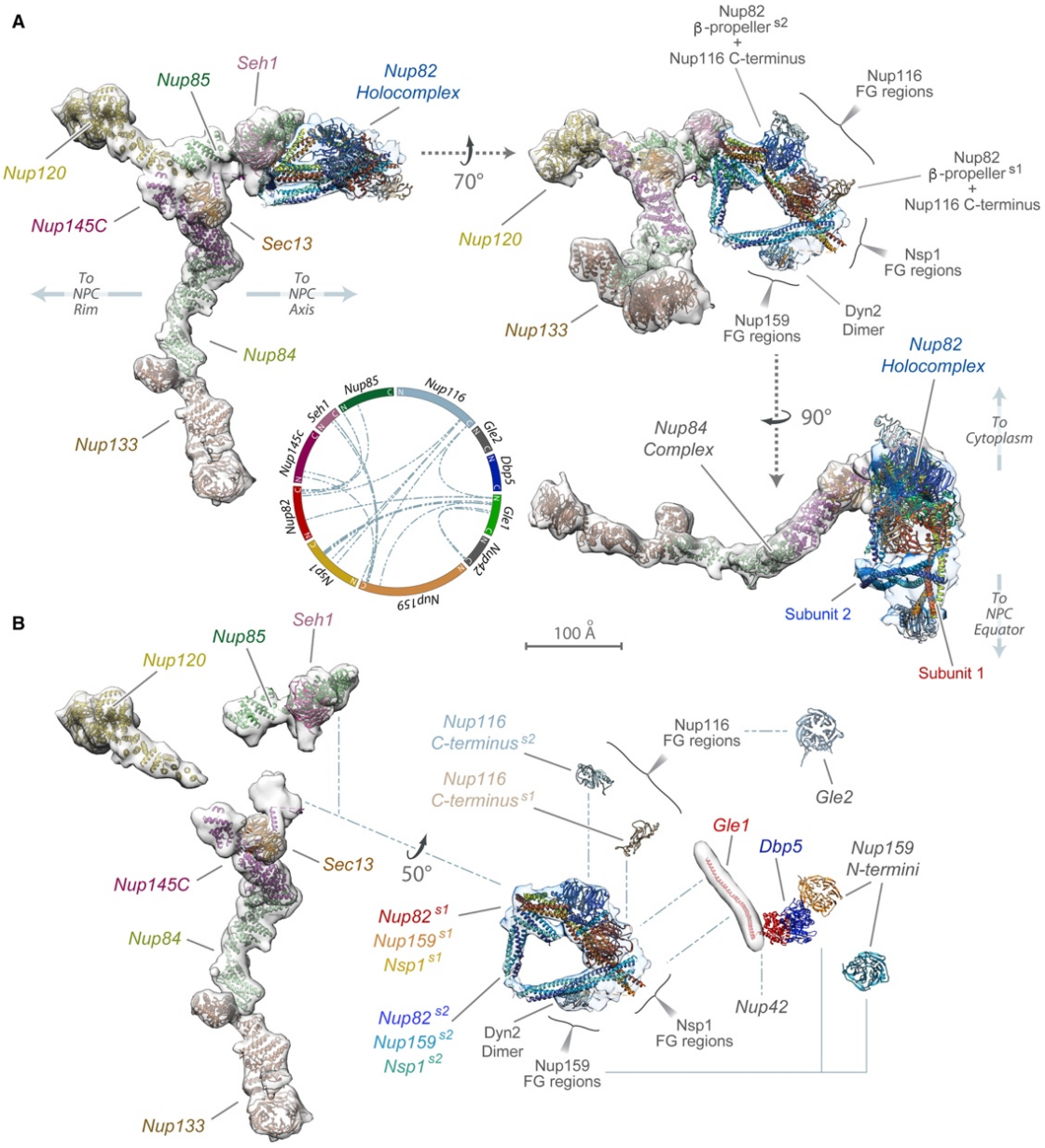

Figure 6.3 | Molecular Architecture of the Cytoplasmic mRNA Export and Remodeling Platform.

(**A**) Structure of the Nup82-Nup84 complex assembly. Three views of the structural arrangement formed by the Y-shaped Nup84 complex (light gray density) and the Nup82 holo-complex (light blue density) calculated using CX-MS data. Each component and structural feature of the different complexes are labeled and shown as a density with fitted ribbon representations of their component Nups. A Circos plot shows the distribution of cross-links (dashed, light blue lines) identified between components of the Nup82 complex, Nup84 complex, and mRNP export/remodeling machinery, used for the calculation of the assembly and the map described in (**B**). (**B**) Molecular architecture of the cytoplasmic mRNA export and remodeling platform. An exploded view of the different platform components is presented (solid blue lines, covalent attachment; dashed blue

lines, CX-MS-identified associations). When available, components are represented as crystal structures (Dbp5, Gle1, and Nup159 N termini; PDB: 3RRM (10); Gle2/RAE1; PDB: 3MMY (30); Nup116 C termini; PDB: 3PBP (8); and 3NF5 (32)). The Gle1 N terminus is represented with a homology model of its predicted coiled-coil region as a red ribbon inside a light gray density of the approximate expected size for the domain.

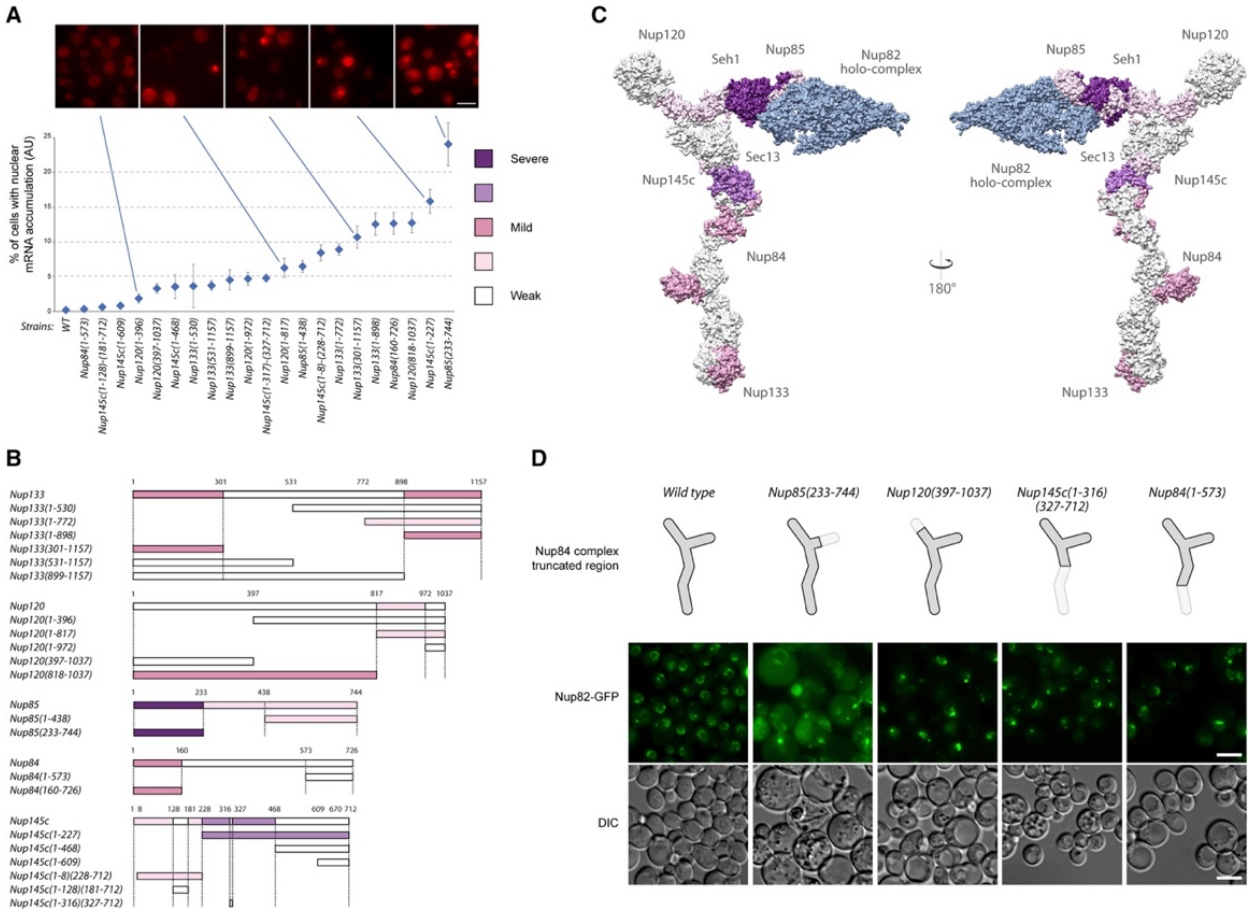

Figure 6.4 | mRNA Export Phenotype in Nup84 Complex Mutants Is Associated with Defective Incorporation of the Nup82 Holo-complex into the NPC.

(**A**) The mRNA export defect phenotype was quantified and plotted (mean value; n = 4) for each Nup84 complex component mutant in order of increasing level of nuclear poly(A) mRNA accumulation as observed by fluorescence in situ hybridization (FISH) (see Methods and Fig. S7 for details) and assigned to five divisions of increasing level of accumulation (white to dark purple) (23). Representative examples of strains included in each division are shown on the top. AU, arbitrary units. Error bars represent SEM. Scale bar, 5 µm. (**B**) Mapping of the color code described in (**A**) into the Nup84 complex components. Horizontal lines represent the amino acid residue length of each protein and truncated version; amino acid residue positions are shown on top of the lines. (**C**) The

severity of nuclear mRNA accumulation phenotype (detailed in **A** and **B**) for specific truncations of the Nup84 complex components are shown mapped into the Nup82-Nup84 complex assembly. The color code is the same as the one described in (**A**). The Nup82 holo-complex density is shown in light blue. (**D**) Subcellular localization of Nup82-GFP in Nup84 complex truncation mutants. Top: diagrams representing the Nup84 complex, with the corresponding truncated region of the complex shown. Middle: localization of the genomically tagged Nup82-GFP reporter as determined by fluorescence microscopy. Bottom: differential interference contrast (DIC) image of the same cells. Scale bar, 5 μm.

Conservation of the Cytoplasmic mRNA Export Platform in Opisthokonts

We tested whether our current structure was consistent with previous maps of the whole NPC. When the Nup82-Nup84 complex assembly is docked into our yeast NPC map (12), the arrangement of their common components is fully consistent, as shown in **Fig. 6.5A**. The Nup82 holo-complex overlaps with the localization density of Nup82, facing down into the central channel, and is in close proximity to the Nup85 arm of the Nup84 complex. Previous attempts to align a single EM envelope for the yeast Nup82 complex to a human cryo-EM NPC map (28) led to divergent and ambiguous results (9). However, we were able to unambiguously dock the yeast Nup82-Nup84 complex assembly into the available human cryo-EM maps (28, 34). When the Nup84 complex was aligned to the corresponding inner copy of its homolog (the Nup107-160 complex), the Nup82 holo-complex aligned with a density projecting only from the cytoplasmic ring, pointing toward the central channel (**Fig. 6.5B** and **6.5C**). It has been suggested that this protrusion might indeed represent some aspect of the Nup88-Nup214 complex, the vertebrate counterpart to the Nup82 holo-complex (28). The yeast and human alignments both support an overall conservation for certain major features of NPC architecture between fungi and metazoa and provide further independent validation of our Nup82-Nup84 complex assembly structure. Importantly, the position of the Nup82 holo-complex FG repeat regions with

respect to the whole NPC is suggestive of an organized arrangement of transport factor docking sites (see Discussion).

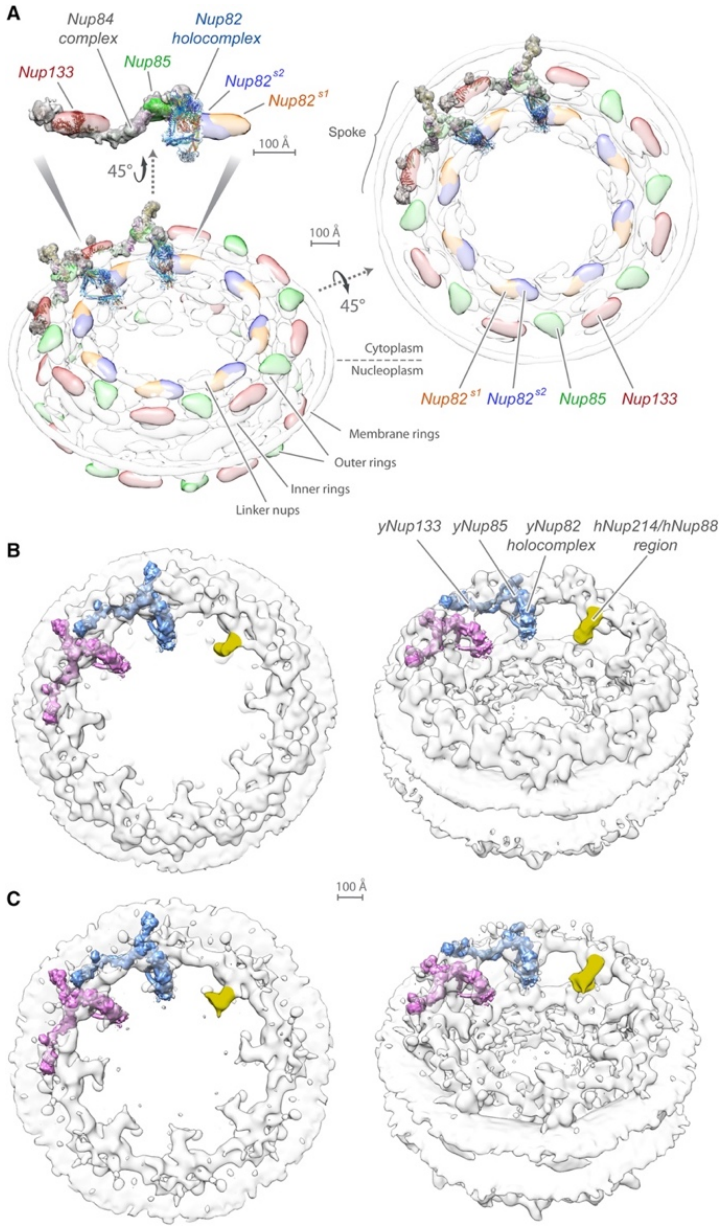

Figure 6.5 | Position of the Nup82-Nup84 Complex Assembly within the NPC. (A) Fitting to the yeast NPC map. Two views of the optimized alignment of two *S. cerevisiae* Nup82-Nup84 complex assemblies into the *S. cerevisiae* NPC localization probability density map (transparent gray), together with a side view of the detailed alignment (12); Nup85 (green), Nup133 (red), and two Nup82 units (blue and orange) are indicated. Scale bars, 100 Å. (B) Comparison with the human NPC tomographic cryo-EM map (EMDB: 2444) (28). Two views of the optimized alignment of two *S. cerevisiae* Nup82-Nup84 complex assemblies (pink and blue) into the human NPC

map (CCC = 0.72). One suggested localization for the human Nup214/Nup88 complex is colored in yellow. (C) Comparison with the mutant human NPC tomographic cryo-EM map (EMDB: 3104) (34), lacking an outer cytoplasmic Y-complex ring (CCC = 0.81).

Discussion

Structure and Evolution of the Nup82 Holo-complex

We present the structure of the Nup82 holo-complex and show how it assembles with the Nup84 complex and other proteins to form the 24-subunit, ~1.8- MDa cytoplasmic mRNA export platform in the NPC. Our structural analysis therefore covers close to one-third of the yeast NPC mass (12), which is now mapped in molecular detail. Unexpectedly, the Nup82 holo-complex and its associated machinery do not form any kind of cytoplasmic filament, in contrast to how it has been pictured in the literature. On the contrary, it forms a strut that faces the central channel. The Nup82 holo-complex exhibits an unusual architecture, with two compositionally identical trimers forming an asymmetric structure. Hinges in coiled-coils allow flexibility to convert two otherwise identically arranged subunits into two similar but morphologically distinct subunits. This structural arrangement, with flexibility in the subunits permitting alternate assemblies, is reminiscent of how vesicle-coating proteins form variable architectures within the same coat complex, such as the hexagonal versus pentagonal architectures observed in clathrin coated vesicles (35). Perhaps this variability is another echo of the evolutionary origin of the NPC in an ancient coating complex (36), and it may also contribute to the observed flexibility of the NPC as a whole. Another feature shared by the NPC and its related coating complexes is the presence of compositionally distinct but structurally and evolutionarily related modules within the entire assembly that arose from ancient duplication events (12, 23, 36). Indeed, there is another NPC subcomplex that also uses a trimeric bundle and

appears to be homologous and evolutionarily related to the Nup82 holo-complex. We discovered this relationship through a homolog detection search using HHPred (18), aiming to find structures comparable to the coiled-coil regions of the three core Nup82 complex components. Remarkably, the top and highly significant hit (HHpred *p* = 4.5E-60, 3.3E-9, and 0.0053 for Nup82, Nsp1, and Nup159, respectively) was another complex from the NPC also containing a heterotrimer of coiled-coils: the *Xenopus laevis* Nup93:Nup62:Nup58:Nup54 complex (6) and its *Chaetomium thermophilum* Nic96:Nsp1:Nup57:Nup49 complex homolog (7) (Fig. S3). This similarity aided in generating high-confidence comparative models for our calculations (Methods). The C termini of both complexes share a common domain arrangement, formed by three consecutive helical coiled-coil regions of different lengths, connected by flexible linkers (**Fig. 6.1**), and both complexes share a common component, Nsp1. Collectively, these observations further support the idea that both complexes evolved from a single common precursor structure, providing yet another example of an ancient duplication now generating diverse modules within the NPC, as postulated by our original protocoatomer hypothesis (36).

Spatial Organization of the FG Repeats

A common architecture and evolutionary origin might also imply a degree of shared functionality. In the case of the Nup82 holocomplex, the coiled-coil region serves as a strut to position various transport factor docking sites out from the core scaffold and toward the central channel of the NPC, where nucleocytoplasmic exchange is mediated (**Fig. 6.6**). We therefore suggest that the coiled-coil trimeric region of the homologous Nic96:Nsp1:Nup57:Nup49 complex and that of the Nup82 holo-complex perform

analogous functions, namely to serve as struts for the correct positioning of transport factor docking sites along the nucleocytoplasmic axis of the central transport channel (**Fig. 6.6**). Being intrinsically disordered, the FG repeat regions themselves cannot form ordered structures to span the central channel. However, by providing a semi-rigid support, the coiled-coil regions of the two complexes may act as flexible struts, placing the FG Nup docking sites so that they efficiently occupy the central channel to form an effective selective barrier, perhaps such that the struts plus FG repeats together comprise the observed ‘‘central transporter’’ (37). Indeed, space-filling models based on size data for FG repeats (38) (**Fig. 6.6**) show that the FG regions would project from the Nup82 holo-complex in such a manner as to essentially span the NPC’s central channel and essentially form the top, cytoplasmic part of the central transporter.

The Nup82 Complex Projects into the NPC’s Central Channel to Coordinate Efficient Export and Remodeling of mRNPs

The FG repeats associated with the Nup82 holo-complex project from the end of the complex adjacent to the Nic96 complex, toward the midplane of the central channel (**Fig. 6.6**); there, they would neighbor the Nsp1, Nup57, and Nup49 FG repeats at the equator of the NPC (7, 39, 40). It is known that the relative position of FG repeats in the Nup82 holo-complex are crucial (31) and that the Mex67/Mtr2 dimer mediating mRNA export directly engages the FG repeats associated with the Nup82 holo-complex (41, 42) (**Fig. 6.3**). Collectively, these results suggest that the type and position of FG repeats in the Nup82 holo-complex are key for an efficient mRNA export mechanism.

Surprisingly, we show that the Nup82 holo-complex does not project outward from the cytoplasmic face of the NPC, as previously assumed. Instead, it projects inward, both

radially and vertically. This arrangement has several important functional consequences. First, based on the organization of the Nup82 holo-complex, this places the associated cytoplasmically disposed FG repeat regions in intimate contact with the symmetrically positioned FG repeat regions in the central channel, forming a continuous conduit of transport factor docking sites from the nuclear to cytoplasmic sides of the NPC. Second, this arrangement also places the mRNP remodeling machineries at the immediate cytoplasmic end of this channel (**Fig. 6.5** and **6.6**). We suggest that the Nup82 holo-complex and Nup84 complex position these cytoplasmic docking and remodeling sites right over the central channel to efficiently capture exporting mRNP particles immediately upon reaching the cytoplasmic end of the central channel; once captured, they can be directly processed by the proximally tethered Gle1/Dbp5/Nup159N remodeling machinery rather than requiring a transfer mechanism to previously supposed distal processing sites on cytoplasmic filaments. Third, the transport factors released during remodeling are also potentially well positioned to be recycled back into the nucleus, while the now translationally primed mRNP exits to the cytoplasm (**Fig. 6.6**). Our molecular architecture is fully consistent with proposed mRNP remodeling models (3, 10), as well as with the observation that cytoplasmic release, but not translocation, is a ratelimiting step during mRNA export (2). When translated into the overall NPC architecture, the presence of eight remodeling hubs surrounding the central channel ensures a highly efficient system consistent with the fast mRNA export rates observed *in vivo* (43–45). Other types of ribonucleoproteins are also actively exported through the NPC, using pathways and components that largely overlap with those of mRNA export (46). It is thus reasonable to

expect that our structural analysis would also serve as a framework for revealing the mechanisms governing their transit and maturation through the NPC.

While the Nup82 holo-complex is a major nexus for RNA export and remodeling processes, its human homolog when altered is also a major nexus for numerous diseases, as underscored by the fact that the mammalian orthologs of Nup82 (Nup88), Nup159 (Nup214), and Nup116 (Nup98) represent the Nups most prevalent in cancer and developmental diseases (47). Hence, our structure may also help rationalize the modifications in this machinery that lead to severe human diseases. For example, mutations in the human homolog of Gle1 are associated with lethal congenital contracture syndrome 1, lethal arthrogryposis with anterior horn cell disease (5), and amyotrophic lateral sclerosis (4). We have been able to localize and orient Gle1 at the hump and spur-2 region of the Nup82 holo-complex, facing the NPC central channel (**Fig. 6.3**). Our data (Table S3) also indicate that the C terminus of Nup42 is associated with the C terminus of Gle1 (29), where the Nup159 N-terminal b-propellers are dynamically associated, with the Nup159 FG regions oriented toward the arms of the Nup84 complex (**Fig. 6.6**) and Dbp5 physically associated with its ATPase cycle modulators Gle1 and the Nup159 b-propeller (10, 48). Strikingly, the residues equivalent to those causing disease states in human Gle1 (49) all map to sites that anchor the yeast protein to either the Nup82 holo-complex or to Nup42 and Dbp5-Nup159N (**Fig. 6.6**). These results, taken together with our structural and functional analyses, underscore the importance of the Nup82 complex as a hub for anchoring the mRNA transport and processing machineries into the heart of the NPC itself and help explain why this complex is a focus for so many developmental, oncogenic, and viral diseases.

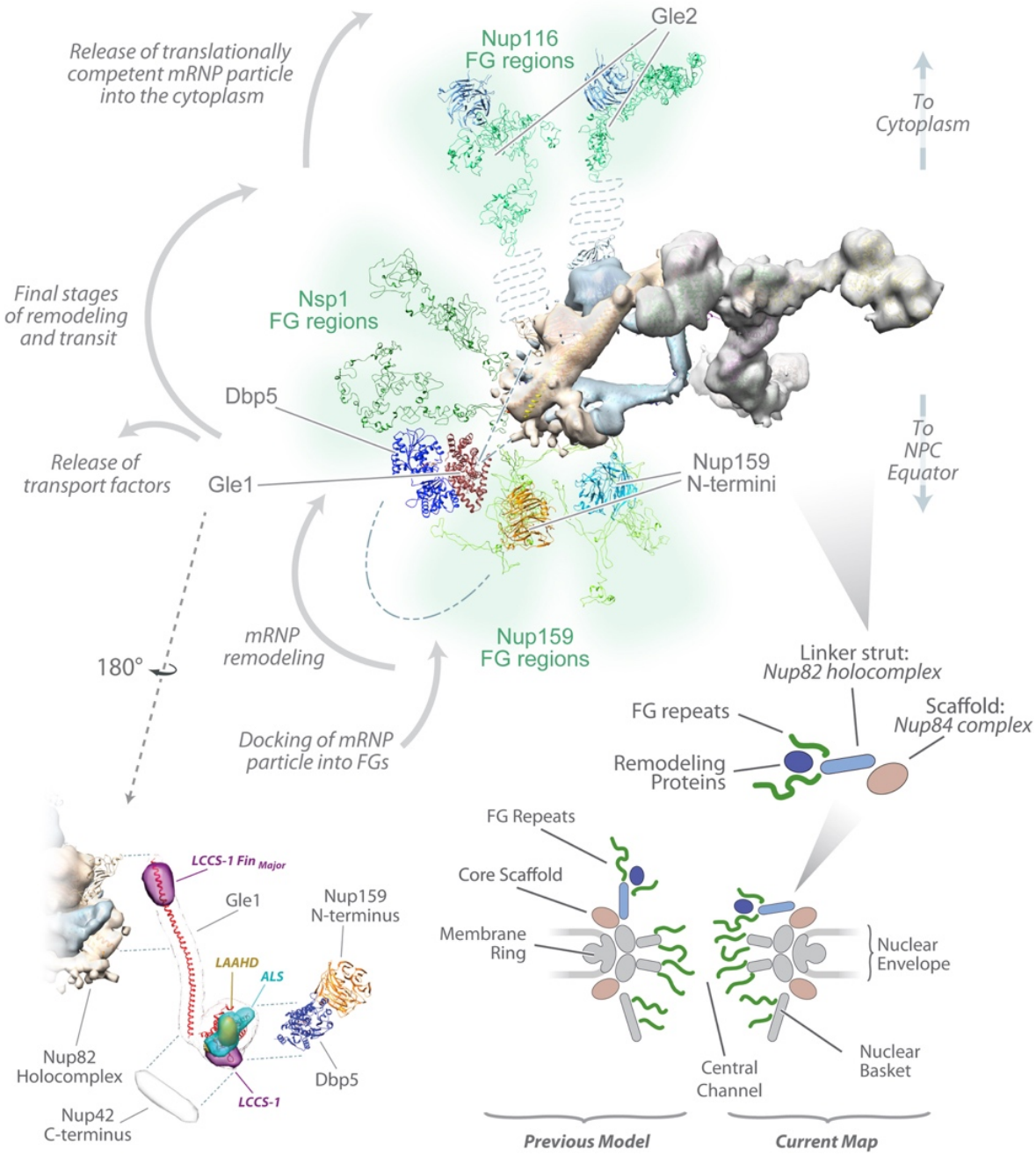

Figure 6.6 | The Nup82-Nup84 Complex Assembly Acts as a Scaffold to Organize the FG Region and mRNP Remodeling Sites in the NPC.
Top: model for the arrangement of the FG regions associated to the Nup82 holo-complex. FG regions were modeled using molecular dynamics. The position of the Nup116 FG regions is based on the position of their C termini (PDB: 3PBP (8)) but could vary significantly, depending on the orientation of the unstructured region connecting the FG domains (dotted blue line). N termini of Nup159 can interact with Dbp5 during mRNP remodeling, as indicated by the dashed blue line. Sequential mRNP export and remodeling steps associated with each region of the complex are shown on the left.

Bottom left: mapping of disease-associated Gle1 mutations into our model for the mRNA export platform. The yeast Gle1 region equivalent to where disease-related mutations have been found in human Gle1 were colored in purple (lethal congenital contracture syndrome 1 [LCCS-1]), gold (lethal arthrogryposis with anterior horn cell disease [LAAHD]), and cyan (amyotrophic lateral sclerosis [ALS]), based on data described previously (4, 49, 50). Proteins are represented as described in **Fig. 6.3B**. Dashed blue lines indicate identified protein-protein associations. **Bottom right**: schematic representation comparing the previous view (left) of the Nup82 complex as components of cytoplasmically oriented filaments, with the new view (right) of how it instead forms struts projecting toward the NPC central channel, positions the FG regions to fill the channel, and forms the top part of the central transporter region.

Methods

Yeast Strains

All Saccharomyces cerevisiae strains used in this study are listed in the Table 2, with the exception of the Nup84 complex truncation mutants that were described in detail in (23). The Nup82 complex tagged strains were constructed in a W303 (Mata/alpha ade2-1 ura3-1 his3-11,15 trp 1-1 leu2- 3,112 can1-100) background. Otherwise stated, strains were grown at 30ºC in YPD media (1% yeast extract, 2% bactopeptone, and 2% glucose). The *Saccharomyces kudriavzevii* strain was obtained from the American Type Culture Collection (ATCC 2601) and grown in the same conditions as referred above for *S. cerevisiae*.

Affinity Purification of Protein Complexes

To purify the native Nup82 complex, that we will call from now on Nup82 holo-complex (as it includes all its intact, full-length endogenous components), we constructed strains in which the NUP encoding gene was genomically tagged with a variant of the Staphylococcus aureus Protein-A, preceded by the human rhinovirus 3C protease (ppx) target sequence (GLEVLFQGPS). The sequence was introduced by PCR amplification of the transformation cassette from the plasmid pProtA/HIS5. Harvested yeast cells,

grown in YPD at 30ºC to mid-log phase were frozen in liquid nitrogen and cryogenically lysed in a Retsch PM100 planetary ball mill (http://lab. rockefeller.edu/rout/protocols). A total of 10-20 g of frozen cell powder were resuspended in 9 volumes of IP buffer (20mM HEPES pH 7.4, 300mM NaCl, 2mM $MgCl_2$, 0.1% Tween 20, 1mM DTT). Cell lysate was clarified by centrifugation at 20,000 g for 10 min. IgG Ab conjugated magnetic beads (Invitrogen) at a concentration of 50 µL slurry/g of frozen powder were added to the clarified cell lysate and incubated for 30 min at 4ºC. Beads were washed three times with 1 mL of IP buffer without protease inhibitors. The native complex was released from the affinity matrix by PreScission protease digestion in the same buffer. The recovered sample was then centrifuged at 20.000 g for 10 min. The supernatant (50-100 µl) was loaded on top of a 5%–20% sucrose gradient made in IP buffer without Tween 20 plus 1/1000 of protease inhibitors. Gradients were ultracentrifuged on a SW55 Ti rotor (Beckman) at 42.000 rpm and 5ºC for 17 hr. Gradients were manually unloaded from the top in 12 fractions of 410 µl. Fractions were analyzed by SDS-PAGE and R250 Coomassie or Sypro Ruby staining.

A higher order complex, containing the Nup84 complex plus several other nups, including the Nup82 holo-complex components, was identified previously (23). The complex was affinity purified from a Nup84-ppx-PrA strain (Methods; Table 2) as described above using as IP buffer 20mM HEPES pH 7.4, 20mM NaCl, 150mM potassium acetate, 2mM $MgCl_2$, 0.5% Triton X-100, 0.1% Tween 20, 1mM DTT, and processed for cross-linking and mass spectrometry analysis.

Stoichiometry of the Nup82 Holo-complex

Diploid strains, carrying one wild-type and one Protein-A-tagged version of each of the major Nup82 holo-complex components were analyzed by affinity purification as described above and the identity of the bands verified by mass spectrometry (Fig. S1). To determine the Stokes radius (R_s) of the Nup82 holo-complex, the natively eluted complex was run through a calibrated Superose 6 GL 30/100 column in 20mM HEPES pH 7.4, 150mM NaCl, 0.1% Tween 20 buffer, and the results plotted against reference protein standards (Ovalbumin, R_s: 3.05; Aldolase, R_s: 4.81; Ferritin, R_s: 6.1; Thyroglobulin, R_s: 8.5). The sedimentation coefficient ($S_{20,w}$) of the Nup82 holo-complex was estimated from the peak of the complex banded in sucrose gradients, run as described above, using the formula $S_{20,w} = \Delta I/(\omega^2 \cdot t)$, where ΔI is the time integral, ω the angular velocity (seconds^{-1}), and t is time (seconds) (see also (51)). The mass of the holo-complex was then calculated using the Siegel-Monte equation (Fig. S1A and S1B) (52). Quantification of the relative amounts of each protein in the purified complex was performed using a synthetic concatamer of tryptic peptides or QconCAT (16) based on the Nup82 complex components (Fig. S1D). Quantotypic peptides for each of the four nucleoporins of the Nup82 complex were selected based on their mass spectrometric behavior (Nup82: 7-LSALPIFQASLSASQSPR-24, 636-NQILQFNSFVHSQK-649; Nup159: 301-TNAFDFGSSSFGSGFSK-717, 948-TSESAFDTTANEEIPK-963; Nsp1: 779-TTNIDINNEDENIQLIK-795, 806-SLDDNSTSLEK-816; Dyn2: 64-NFGSYVTHEK-73, 53- YGNTWHVIVGK-63). A synthetic gene (called Nup82 QconCAT) was designed by concatenation of the sequences encoding the referred peptides and addition of a 6xHis c-terminal tag: (MKEIRNQILQFNSFVHSQKTNAFDFGSSSFGSGFSKNFGSYV

THEKTTNIDINNEDENIQLIKLSALPIFQASLSASQSPRTSESAFDTTANEEIPKYGNTWH
VIVGKSLDDNSTSLEKQINSIKHHHHHH).

The E. coli codon optimized sequence was cloned into plasmid pGEX6p-1, resulting in the expression of a protein with an n-terminal GST tag that was used both as a purification tag and sacrificial peptide (53). The Nup82-QconCAT protein was expressed by growing 300ml of BL21 E. coli cells at 37ºC to OD600 = 0.6 in minimal M9 media (16) supplemented with heavy arginine and lysine (L-arginine:HCl $^{13}C_6$; L-lysine:2HCl $^{13}C_6$, Cambridge Isotope Laboratories Inc.). IPTG (1mM) was used to induce expression of the construct for 3 hr at 37ºC. Harvested cells were processed using BugBuster Extraction Reagent (Novagen) as indicated by the manufacturer. The full-length Nup82 QconCAT was then purified using a two-step method that ensures a final fulllength product by consecutive purification from the n and c-terminal tags: i) Clarified soluble material was incubated with 500 µL of gluthatione Sepharose 4b (GE Healthcare) at room temperature for 1 hr at 4ºC, and the retained proteins eluted using 2x 1ml of elution buffer (20mM HEPES pH 7.4, 150mM NaCl, 45mM imidazole, 6M guanidinium hydrochloride, 1mM TCEP, 1/500 protease inhibitor cocktail (PIC) (Sigma)). ii) The elution volume was then passed through an equilibrated His-Trap HP (GE Healthcare) at room temperature. The retained Nup82 QconCAT was then eluted in 20mM HEPES pH 7.4, 500mM imidazole, 150mM NaCl, 6M guanidinium hychloride, 1mM TCEP, 1/500 PIC. The resulting elution was analyzed by SDS-PAGE to ensure the presence of a full-length, pure protein.

For the MS analysis, the Nup82 holo-complex was purified as described above. The gradient fractions containing the complex were collected and concentrated by

centrifugation at 355,000 g for 6 hr in a TLA 120.1 rotor at 4ºC. The concentrated complex was then resuspended in a final 1x Nupage LDS Sample buffer (Thermo Fisher Scientific), 10mM TCEP (Thermo Fisher Scientific). The Nup82-QconCAT was ethanol precipitated and washed to eliminate the guanidinium chloride and resuspended in 1x Nupage LDS Sample buffer, 10mM TCEP. Approximately equimolar amounts of complex and Nup82-Qconcat were combined to give a final protein amount of 1 µg. The combined sample was heated at 72ºC for 10 min and then alkylated using a final concentration of 30mM iodoacetamide (Sigma). The sample was then loaded into a 4% (37.5:1) in-house prepared stacking acrylamide SDS-PAGE gel. The resulting band, containing a mixture of Nup82 complex and stable-isotopically labeled Nup82 QconCAT proteins, was excised and sequentially digested by endoproteinase LysC (Roche) and trypsin (Roche) inside gel matrix, followed by LC-MS analyses to determine L/H ratio of standard peptides. LC-MS analyses were performed on an Orbitrap Fusion Mass Spectrometer (Thermo Scientific), with an Easy-nLC 1000 HPLC (Thermo Scientific) and an Easy-Spray electrospray source (Thermo Scientific). L/H ratios of standard peptides were determined using the MaxQuant software (version 1.2.2.5) (54).

Overexpression of Dyn2 was performed mimicking the conditions described in (9): the *S. cerevisiae* Dyn2 coding sequence was cloned into the 2-micron plasmid p424-Gal1, under the control of the Gal-1 promoter. Overexpression was achieved by growing the transformed yeast cells in yeast synthetic minimal media supplemented with 2% glucose, 1% raffinose, harvesting the cells in mid-log phase, washing them with ddH2O and then transferring them to yeast synthetic minimal media supplemented with 2% galactose, 1% raffinose for 3 hr at 30ºC. Cells were then harvested and cryo-milled and

the endogenous Nup82 holo-complex was purified as described above using Nup82-PrA as the handle. Purified complexes were run in SDS-PAGE gels, stained with SyproRuby (Thermo Fisher Scientific) and the relative intensity of the different bands were quantified using ImageJ (http://imagej.net).

Chemical Cross-linking and Mass Spectrometry

The natively eluted complex (250 µl, in buffer 1- 20mM HEPES pH 7.4, 300mM NaCl, 0.1% Tween, 2mM $MgCl_2$, 1mM DTT) was crosslinked via the addition of DSS-H12/D12 (DiSuccinimidylSuberate) cross-linker (Creative Molecules) to yield a final concentration of 0.25 mM and incubated for 45 min at 25ºC with gentle agitation in a shaker (900 rpm). The reaction was then quenched by 50 mM ammonium bicarbonate. In the case of cross-linking using EDC reagent (Pierce), the sample was equilibrated and natively eluted in EDC cross-linking buffer (10mM BisTris pH 6.5, 100mM NaCl, 2mM $MgCl_2$, 0.1% Tween, 1mM DTT). EDC (20 mM) and N-hydroxysulfosuccinimide (0.4 mM) (i.e., 2% molar ratio with respect to EDC) were then added to cross-link the sample. The sample was incubated for 45 min at 25ºC with gentle agitation. After cross-linking, Tris-HCl pH 8.0 (50 mM) and β-mercaptoethanol (20 mM) were added to the cross-linked sample to quench the reaction. After Cysteine reduction and alkylation, cross-linked samples were separated in a 4%–12% NuPage SDS-PAGE (Invitrogen). Gels were briefly stained by GelCode Blue Stain Reagent (Thermo Fisher Scientific) to enable the visualization of the cross-linked protein complexes. The cross-linked complexes were then digested in-gel with trypsin or chymotrypsin to generate cross-linked peptides as previously described (Shi et al., 2014). After in-gel digestion, the cross-linked peptide mixtures were fractionated by peptide SEC (Superdex Peptide PC 3.2/30, GE Healthcare) by an offline

HPLC (Agilent Technologies). Two or three SEC fractions covering the molecular mass range of ~2.5 kD to ~10 kD were subsequently collected and analyzed by LC/MS. For cross-link identifications, the purified peptides were dissolved in the sample loading buffer (5% MeOH, 0.2% FA) and analyzed by a LTQ Velos Orbitrap Pro mass spectrometer or an Orbitrap Q Exactive (QE) Plus mass spectrometer (Thermo Fisher). For the analysis by the Velos Orbitrap mass spectrometer, briefly, the dissolved peptides were pressure loaded onto a self-packed PicoFrit column with integrated electrospray ionization emitter tip (360 O.D, 75 I.D with 15 µm tip, New Objective). The column was packed with 10 cm reverse-phase C18 material (3 µm porous silica, 200 Å pore size, Dr. Maisch GmbH). Mobile phase A consisted of 0.5% acetic acid and mobile phase B of 70% ACN with 0.5% acetic acid. The peptides were eluted in a 120 or a 140 min LC gradient (8% B to 50% B, 0-93 min, followed by 50% B to 100% B, 93-110 min and equilibrated with 100% A until 120 or 140 min) using a HPLC system (Agilent), and analyzed with a LTQ Velos Orbitrap Pro mass spectrometer. The flow rate was ~200-250 nL/min. The spray voltage was set at 1.9-2.3 kV. The instrument was operated in the data-dependent mode, where the top eight-most abundant ions were fragmented by higher energy collisional dissociation (HCD) (normalized collisional energy 27-29) and analyzed in the Orbitrap mass analyzer. The target resolution for MS^1 was 60,000 and 7,500 for MS^2. The QE instrument was directly coupled to an EASY-nLC 1200 System (Thermo Fisher) and experimental parameters were similar to those of the Velos Orbitrap. The cross-linked peptides were loaded onto an Easy-Spray column heated at 35ºC (C18, 3mm particle size, 200 Å pore size, and 50 µm X 15cm, Thermo fisher). The top 8 or 10 most abundant ions (with charge stage of 3-7) were selected for fragmentation by HCD. The raw data were searched by

pLink (Yang et al., 2012a) using a FASTA database containing protein sequences of the complexes. An initial MS^1 search window of 5 Da was allowed to cover all isotopic peaks of the cross-linked peptides. The data were automatically filtered using a mass accuracy of $MS^1 \leq 10$ ppm (parts per million) and $MS^2 \leq 20$ ppm of the theoretical monoisotopic (A0) and other isotopic masses (A+1, A+2, A+3, and A+4) as specified in the software. Other search parameters include cysteine carbamidomethyl as a fixed modification, and methionine oxidation as a variable modification. A maximum of two trypsin missed-cleavage sites was allowed. The initial search results were obtained using a default 5% false discovery rate (FDR) – expected by target-decoy search strategy. All spectra were manually verified. ~94% of the cross-link identifications have a MS^1 mass accuracy within 6 ppm. The cross-link data was visualized and analyzed by the CX-Circos software (manuscript in preparation).

Chemical Cross-linking and Mass Spectrometry Analysis of the *S. cerevisiae/S. kudriavzevii* Nup82 Holo-complex

To define the relative orientation of the two copies of Nup82 present in the Nup82 holo-complex we expressed an exogenous copy of Nup82 from the yeast *Saccharomyces kudriavzevii* (called from now on skNup82). We selected *S. kudriavzevii* because it is a closely related species that forms natural hybrids with *S. cerevisiae*, some of them used for wine fabrication (Borneman et al., 2012), and the level of conservation at the amino acid level between both species is particularly high, ensuring functionality of the skNup82 version and enough sequence variation to identify the specific peptides from each species protein version. *S. kudriavzevii* strain was obtained from ATCC (ATCC 2601) and genomic DNA was prepared using standard methods. The 30 UTR and open reading

frame for skNup82 was amplified and sequenced to account for potential mutations detected in the sequence available in the public database (GenBank: EHN01740.1). The wild-type verified skNup82 sequence was found to encode a 716 amino acid protein with 75% identity to the scNup82 primary sequence (alignment available upon request). The upstream 190 nucleotides (promoter) region and the gene sequence were amplified using primers skN82Prom-F(5´-CACCGAAAGTTTATAGATTCAT-3´) and skN82GTW_R2 (5´-GCTGGGCCCCTGGAACAGAACTTCCAGGCCGTTTTTTGGCTGAGTATTAGTG-3´) that introduces an in-frame prescission protease cleavage site at the end of the skNup82 coding sequence. The PCR product was cloned using the pENTR/D-TOPO Cloning Kit (Thermo Fisher Scientific) and then transferred to a modified pAG305GPD-ccdb-EGFP plasmid (Addgene), where the GPD promoter had been eliminated through a SacI-XbaI (New England Biolabs) cleavage and refill. The resulting integrative plasmid, pAG305-skNup82ppx-EGFP, was linearized using ClaI (New England Biolabs) and transformed into a diploid w303 *S. cerevisiae* strain. Successful integrations were assessed by PCR; correct expression and localization of the skNup82-EGFP construct were confirmed by western-blot and fluorescence microscopy, that showed the characteristic nuclear rim staining of a properly localized nucleoporin. Affinity purification of the Nup82 complex using skNup82-EGFP as a handle showed all the components of the native Nup82 complex, including a substoichiometric amount of scNup82, showing correct incorporation of the construct into the native Nup82 complex. The isolated, purified complex (see above for details on purification) was analyzed by CX-MS (see above).

Negative Stain Electron Microscopy

Purified endogenous Nup82 complex samples were applied to glow-discharged carbon-coated copper grids and stained with 1% uranyl formate. Images were collected on a Tecnai F20 (FEI Inc., USA) transmission electron microscope operating at an acceleration voltage of 80 kV at 50,000x magnification and underfocus ~1.5 µm. Images were recorded on a Tietz F224 4096x4096 CCD camera (15 µm pixels) at 2x binning. The pixel size at the specimen level was 3.23 Å. Particles were selected using Boxer from EMAN (Ludtke et al., 1999). The contrast transfer function (CTF) of the normalized images was determined using ctfit from EMAN and the phases were flipped accordingly. After that, the particles were subjected to Iterative Stable Alignment and Clustering (ISAC; (55)) technique. A pixel error of $2\sqrt{3}$ was used for the stability threshold. For comparison, the Nup82 holo-complex class averages were aligned and paired with Nsp1-FGD class averages or with GFP-tagged Nup82 complex class averages using the modified Spider 'AP SH' operation. Then the Nsp1-FGD class averages were subtracted from the Nup82 holo-complex class averages and the Nup82 holo-complex class averages were subtracted from the GFP-tagged Nup82 complex class averages and difference maps generated.

Fluorescence *In Situ* Hybridization

FISH on wild-type and Nup84 complex truncation mutant strains was performed in 96-well plates. A 35 nucleotide long oligo dT probe (synthetized by Exiqon) and labeled post-synthesis with cy5 was used to detect poly A+ RNA [TT+TTT+TTTT+TTT+TTT+TT. TT+TTT+TTT+TTT+TTT+TTTT, T+ represents locked nucleic acids (LNA). Cells were grown in SD complete at 25ºC to OD 600 = 0.5-0.6 and fixed by the addition of para-

formaldehyde at a final concentration of 4% for 45min at room temperature. Cells were washed 3x with buffer B (1.2M Sorbitol, 100mM KHPO4 pH7.5), suspended in spheroplast buffer [1.2M Sorbitol, 100mM KHPO4 pH7.5, 20mM Ribonucleoside-vanadyl complex (NEB #S1402S), 20mM β-mercaptoethanol, 25U lyticase / 1OD600 of cells (Sigma cat # L2524)] and incubated at 37ºC until cell walls were digested. Digested cells were washed 2x with cold buffer B, attached to polyA lysine (0.01%) treated 96 glass bottom MicroWell plate (MGB096-1-2-LG-L #0325289L2L) and stored in 70% ethanol at -20ºC. For hybridization cells were washed twice with 2 3 saline sodium citrate (SSC) and 1x 35% formamide/2 3 SSC. 20ng of labeled dT LNA probe was resuspended in 35% (v/v) formamide, 2 3 SSC, 1 mg·ml^{-1} BSA, 10 mM ribonucleoside vanadyl complex (NEB #S1402S), 5 mM NaHPO4, pH 7.5, 0.5 mg·ml^{-1} Escherichia coli tRNA and 0.5 mg·ml^{-1} single-stranded DNA and denatured at 95ºC for 3 min and cells hybridized overnight in the dark at 37ºC. Cells were then washed in 35% formamide/2 3 SSC at 37ºC 2x 30 min, followed by a 1 min wash in 1 3 PBS at room temperature followed by the addition of DAPI containing mounting medium to each well (Prolong Gold - Invitrogen #P36935). Images were acquired using a Zeiss Z1 inverted microscope, a 100x 1.43 NA oil objective and a AxioCam mRm CCD camera and the following filter sets: Zeiss 488050-9901-000 (Cy5), Zeiss 488049-9901-000 (DAPI). Three-dimensional datasets were generated by acquiring multiple 200 nm z stacks spanning the entire volume of cells, 3D datasets reduced to 2D datasets by applying a maximum projection function in FiJi. The polyA accumulation phenotype was quantified by determining the fraction of cells showing strong nuclear polyA accumulation. For each strain, at least 200 cells from at least 3 different fields were quantified.

Fluorescence Microscopy

Nup82 was genomically tagged with GFP on selected Nup84 complex truncation yeast mutant strains using standard techniques. Cells were grown in YPD media at 30C and visualized with a 63x 1.4 numerical aperture plan-apochromat objective using a Carl Zeiss Axioplan 2 microscope equipped with a Hamamatsu Orca ER-cooled CCD camera. The system was controlled with Openlab imaging software (Perkin Elmer). Images were treated with ImageJ (http://imagej.net/Welcome) and Adobe Photoshop (Adobe) softwares.

Integrative Structure Determination

Our integrative structure determination of the Nup82 holo-complex proceeded through four stages (Fig. S3D) (12, 27): (1) gathering of data, (2) representation of subunits and translation of the data into spatial restraints, (3) configurational sampling to produce an ensemble of structures that satisfies the restraints, and (4) analysis and validation of the ensemble structures. The modeling protocol (i.e., stages 2, 3, and 4) was scripted using the Python Modeling Interface (PMI), version c7411c3, a library for modeling macromolecular complexes based on our open-source Integrative Modeling Platform (IMP) package, version 2.5 (https://integrativemodeling.org) (21). Further details of the integrative modeling procedures are provided in Table 1, as well as previous publications (13). Files containing the input data, scripts, and output structures are available online (https://salilab.org/nup82; https://github.com/salilab/nup82).

Table 6.1. Summary of Integrative Structure Determination of the Nup82 Complex

Modeling Programs	Python Modeling Interface (PMI), version c7411c3; Integrative Modeling Platform (IMP), version 2.5; MODELER 9.13
Homology Detection and Structure Prediction	HHPred, PSIPRED, DISOPRED, DomPred, COILS/PCOILS, and Multicoil2 (see also Figure S3 and Table S1)
Spatial Restraints	Chemical cross-links, electron microscopy 2D, excluded volume, sequence connectivity, and five homo-dimer cross-links restraints (see also Methods)
Sampling Method	Replica exchange Gibbs sampling, based on the Metropolis Monte Carlo algorithm; 8–16 replicas were used through 270 (initial step) and 80 (refinement step) independent runs, at the temperature range of 1.0–2.5
Monte Carlo Moves	Random translation and rotation of rigid bodies (up to 2 Å and 0.04 radians, respectively)
	Random translation of individual beads in the flexible segments (up to 3 Å)
Number of Structures Generated	1,350,000 (initial step) and 10,000 (refinement step) structures
	463 top-scoring structures were subjected to the clustering analysis
Clustering Analysis	2 clusters of 370 (80%) and 93 (20%) structures (see also Figures S4 and S5)
Sampling Exhaustiveness	p = 0.972
Precision of the Clusters	9.0 Å (cluster 1: 370 structures) / 16.3 Å (cluster 2: 93 structures)
Stoichiometry	2:2:2:2 (Nup82:Nup159:Nsp1:Dyn2; see also Figure 6.1)
Chemical Cross-links Satisfied in the Cluster	88.5% combined (93.3% DSS and 74.1% EDC within 35 and 30 Å distances, respectively; see also Figures 6.2B and S4D)
EM 2D Class Averages	Average ccc for 21 class averages is 0.931. See also Figures 6.2C and S2C.

GFP Mass-Tagging EM 2D Class Averages	ccc = 0.932 (GFP mass-tagging at the Nup159 C termini); ccc = 0.953 (GFP mass-tagging at the Nup82 C termini) (see also Figure 6.2C)
Small Angle X-Ray Scattering (SAXS)	χ = 1.66 (Nup824–220), 2.55 (Nup824–452), and 6.47 (Nup82572–690) (see also Figures 6.2D and S5D–S5F and Table S4)
Human NPC cryo-EM Map	ccc = 0.72 (wild-type) and 0.81 (mutant) (see also Figure 6.5 and S6)
Visualization and Plotting	UCSF Chimera 1.10, CX-Circos, matplotLib, and GNUPLOT

Stage 1: Gathering of Data

The stoichiometry was determined via biochemical quantitation of the density-gradient purified Nup82 complex (Fig. S1). 1,131 cross-links were identified via mass spectrometry (**Fig. 6.2A**; Table S2). The atomic structures for some of the yeast Nup82 complex components had been previously determined via X-ray crystallography (Table S1) (8, 22, 32, 56). Their close homologs were identified by HHPred (Table S1) (18). Secondary structure and disordered regions were predicted by PSIPRED (57) and DISOPRED (58), respectively (Table S1). Coiled-coil regions of Nup82, Nsp1, and Nup159 were predicted by COILS/PCOILS (59) and Multicoil2 (60) (Table S1). 21 EM class averages (Fig. S2C) and 3 SAXS profiles (Fig. S5D–S5F) were obtained as described in Methods and Table S4.

Stage 2: Representation of Subunits and Translation of the Data into Spatial Restraints

The domains of the Nup82 complex subunits were coarse-grained using beads of varying sizes representing either a rigid body or a flexible string, based on the available crystallographic structures and comparative models (Table S1). In a rigid body, the beads have their relative distances constrained during configurational sampling, whereas in a flexible string the beads are restrained by the sequence connectivity (13). The residues in the rigid bodies and flexible strings corresponded to 37.3% and 62.7% of the Nup82 complex, respectively. To maximize computational efficiency while avoiding using too coarse a representation, we represented the Nup82 complex in a multi-scale fashion, as follows.

First, the crystallographic structures of each Nup82 complex domain were coarse-grained using two categories of resolution, where beads represented either individual

residues or segments of up to 10 residues. For the one-residue bead representation, the coordinates of a bead were those of the corresponding Ca atoms. For the 10-residue bead representation, the coordinates of a bead were the center of mass of all atoms in the corresponding consecutive residues (each residue was in one bead only). The crystallographic structures covered 25.6% of the residues in the Nup82 complex.

Second, for predicted non-disordered domains of the remaining sequences, comparative models were built with MODELER 9.13 (17) based on the closest known structure detected by HHPred (18) and the literature (Table S1) (6, 7). Notably, structurally defined remote homologs (PDB: 5C3L and 5CWS) (6, 7) were detected for the C-terminal coiled-coil regions of Nup82, Nup159, and Nsp1 (Fig. S3; Table S1). Similarly to the X-ray structures, the modeled regions were also coarse-grained using two categories of resolution, resulting in the 1-residue and 10-residue bead representations. The comparative models covered 11.7% of the residues in the Nup82 complex.

Finally, the remaining regions without a crystallographic structure or a comparative model (i.e., regions predicted to be disordered without a known homolog) were represented by a flexible string of beads corresponding to up to 100 residues each. We used the low-resolution representation (100 residues per bead) only for the unstructured FG repeats, whose structure is ‘‘decoupled’’ from the configurations of the core of the Nup82 holo-complex (27). The residues in these beads corresponded to 62.7% of the Nup82 complex.

To improve the accuracy and precision of the structure ensemble obtained through the satisfaction of spatial restraints (below), we also imposed constraints based on crystallographically defined interfaces: $Dyn2^{7-92}$-$Nup159^{1117-1126}$ (PDB: 4DS1) (22) and

Nup82$^{7-452}$-Nup159$^{1429-1456}$-Nup116$^{966-1111}$ (PDB: 3PBP) (8). The latter interface of ScNup116$^{966-1111}$ was compared with the structure of CgNup116$^{882–1034}$ (PDB: 3NF5) (32), leading to the conclusion that the Nup116 interfaces are consistent among different species. Subcomplexes including these interfaces were simply represented as rigid bodies.

With this representation in hand, we next encoded the spatial restraints into a Bayesian scoring function (13) based on the information gathered in Stage 1, as follows. First, the collected DSS and EDC cross-links were used to construct the Bayesian scoring function that restrained the distances spanned by the cross-linked residues (13), taking into account the ambiguity due to multiple copies of identical subunits; the ambiguous cross-link restraint considers all possible pairwise assignments in multiple copies of identical subunits, weighting more the least violated distance(s).

Second, the excluded volume restraints were applied to each bead in 10-residue (or the closest) bead representations, using the statistical relationship between the volume and the number of residues that it covered (27).

Third, we applied the sequence connectivity restraint, using a harmonic upper bound on the distance between consecutive beads in a subunit, with a threshold distance equal to four times the sum of the radii of the two connected beads. The bead radius was calculated from the excluded volume of the corresponding bead, assuming standard protein density (13, 27).

Fourth, 5 homo-dimer DSS cross-links between Nup159 residues of 1384-1384, 1387-1387, 1414-1414, 1417-1417, and 1432- 1432 as well as one homo-dimer DSS

cross-link between Nup82 residues of 517-517 were transformed to upper-harmonic distance restraints (up to 30 Å), enforcing the homo-dimer formation of the helices. Finally, the EM 2D restraint (13) was imposed on the highest resolution representation of each subunit, using a negative logarithm of the cross-correlation coefficient between the EM class average density and the best-matching density projection of the structure as the em2D score (Stage 3). For sufficient precision, 100 projections were generated by uniform sampling of the unit sphere (13). The pixel size of the resulting projection image was equal to the pixel size of the class average (3.23Å). The relative weight of the final EM 2D restraint in the total score of a structure was set to 10^4, so that the scale of the em2D score matched those of the other restraint types.

Most of the remaining information (stoichiometry, crystallographic structures of the subunits, their homologs, and the two crystallographic interfaces) is included in the representation, whereas the SAXS profiles, immuno-EM class averages, and the density map from single-particle EM reconstruction (9) were used only for validating our final structures. See the IMP scripts for details (https://salilab.org/nup82; https://github.com/salilab/nup82).

Stage 3: Conformational Sampling

Structural models of the Nup82 complex were computed using Replica Exchange Gibbs sampling, based on the Metropolis Monte Carlo algorithm (13). The Monte Carlo moves included random translation and rotation of rigid bodies (up to 2 Å and 0.04 radians, respectively) and random translation of individual beads in the flexible segments (up to 3 Å). 8 to 16 replicas were used for each run, with temperatures ranging between 1.0 and 2.5 (Table 1). A structure model was saved every 10 Gibbs sampling steps, each

consisting of a cycle of Monte Carlo steps that moved every rigid body and flexible bead once. The entire sampling procedure (Steps 1 to 3) took ~4 weeks on a cluster of ~5,000 cores.

Step 1—Initial modeling against each corresponding EM 2D class

21 subsets of independent sampling runs were performed, each sampling run starting with a random initial configuration and sampled against the EM 2D restraint of the corresponding class. The calculations were repeated 10 to 20 times per subset, producing a total of ~1,350,000 structures through the 270 independent runs.

Step 2—Application of the EM 2D filter

From the ~1,350,000 structures from Step 1, we selected 650 structures whose em2D cross-correlation coefficient was at least 0.89 for at least 10 of the 21 class averages (Fig. S4B).

Step 3—Refinement against all 21 EM 2D class averages

80 independent refinement runs were performed, each one starting with one of the 650 structures from Step 2. The scoring function included em2D scores for all 21 class averages as well as other restraints listed above. The sampling produced a total of ~10,000 structures. 463 top-scoring structures from Step 3 were subjected to the subsequent analysis in Stage 4.

Stage 4: Analysis and Validation of the Ensemble Structures

Input information and output structures need to be analyzed to estimate structure precision and accuracy, detect inconsistent and missing information, and to suggest more informative future experiments. Assessment begins with structural clustering of the modeled structures produced by sampling, followed by assessment of the thoroughness

of structural sampling, estimating structure precision based on variability in the ensemble of good-scoring structures, quantification of the structure fit to the input information, structure assessment by cross-validation, and structure assessment by data not used to compute it. These validations are based on the nascent wwPDB effort on archival, validation, and dissemination of integrative structure models, which we lead (14). We now discuss each one of these points in turn.

Clustering

A prerequisite for structure analysis is the clustering of the structures generated by satisfying the input data (12, 13). We used Cα root-mean-square deviation (RMSD) quality-threshold clustering (13). In general, there are three possible modeling outcomes, based on the number of clusters of models and consistency between the models and information (13). First, if only a single model (or a cluster of similar models) satisfies all restraints and all input information, there is likely sufficient information for determining the structure (with the precision corresponding to the variability within the cluster). Second, if two or more different models are consistent with the input restraints, the information is insufficient to define the single state or there are multiple significantly populated states. If the number of distinct models is small, structural differences between models may suggest additional experiments to narrow down the number of possible solutions. Third, if no model satisfies all input information, the information or its interpretation in terms of the inferred spatial restraints is incorrect, in which case the representation needs to be modified to include additional degrees of freedom, and/or sampling needs to be improved. In the case of the Nup82 complex, the clustering analysis identified a single dominant cluster of 370 similar structures (Fig. S4A and S5B), corresponding to the most favorable

outcome of the three possibilities described above. The average RMSD between the major (370 structures) and minor clusters (93 structures) is relatively low at approximately 20 Å, considering the resolution of the data, the resolution of the coarse-grained molecular representation, and the variation within each cluster (13) (Fig. S4A). As a result, localization of all components is effectively identical between the major and minor clusters, differing only in the orientation of the Nup82 β-propeller (Fig. S5B). Most importantly, our functional interpretation of the structure is completely robust with regard to the differences between the means of the two clusters.

Convergence of Sampling

Any structure determination or computational modeling exercise can be described as a structural sampling process, guided by a scoring function (27). Generally, good-scoring structures need to be found by a sampling, optimization, or enumeration scheme. Unless structures are enumerated, the very first test needs to estimate the thoroughness of structural sampling or optimization (13), which is often stochastic (e.g., Monte Carlo and Molecular Dynamics simulations). For stochastic methods, thoroughness of sampling can be assessed by showing that two independent runs (e.g., using random starting configurations or different random number generator seeds) do not result in significantly different solutions (13, 23, 27). Given two or more sets of structures from independent runs, we first cluster structures from all sets together, followed by assessing whether or not the runs contribute evenly to the population of each cluster, using the p value from the χ^2 contingency test for homogeneity of proportions (61).

For the Nup82 complex, the highly significant *p value* of 0.972 (Table 1) indicated that our Monte Carlo algorithm sampled all top scoring solutions at the resolution better

than the precision of the dominant cluster. The caveat is that passing this sampling test is not absolute evidence of thorough sampling; a positive outcome of the test may be misleading if, for example, the landscape contains only a narrow, and thus difficult to find, pathway to the pronounced minimum corresponding to the correct structure.

Estimating Structure Precision Based on Variability in the Ensemble of Good-Scoring Structures

The ensemble of the top-scoring structures is analyzed in terms of the precision of its structural features (12, 27). In general, commonly-used features include particle positions, distances, and contacts. Precision is defined by the feature variability in the ensemble with a measure similar to the crystallographic isotropic temperature factor (B_{iso}) (Fig. S4C), and likely provides the lower bound on its accuracy. Of particular interest are features present in most configurations in the ensemble that have a single maximum in their probability distribution. The spread around the maximum describes how precisely the feature is determined from the input information. The precision of component position is quantified as the average root-mean-square fluctuation (RMSF) across all pairs of structures in the cluster, after least-squares superposition onto the centroid structure (13). For the Nup82 complex, the 9.0 Å precision of the core structured region in the dominant cluster was sufficiently high to pinpoint the locations and orientations of the constituent proteins and domains (**Fig. 6.1** and S4C; **Table 1**), demonstrating the quality of the data including the cross-links and EM 2D class averages. The localization probability density maps of every Nup82 subunit as well as the whole complex were computed from the dominant cluster of the 370 solutions (**Fig. 6.1** and S4A).

Fit to Input Information

An accurate structure needs to satisfy the input information used to compute it. The ensemble of solutions was assessed in terms of how well they satisfied information from which they were computed, including the cross-links, the excluded volume, sequence connectivity, and the EM two-dimensional restraints.

First, the dominant cluster satisfied 88.5% of all combined cross-links (93.3% and 74.1% of the DSS and EDC cross-links, respectively) (**Fig. 6.2B** and S4D; **Table 1**); a cross-link restraint was satisfied by the cluster ensemble if the median Cα-Cα distance of the corresponding residue pairs (considering restraint ambiguity) was < 35 Å and 30 Å for the DSS and the EDC cross-links, respectively. Our cross-link data (10 DSS and 1 EDC cross-links) is in complete agreement with the crystal structure of Nup82$^{7-452}$-Nup159$^{1429-1456}$-Nup116$^{966-1111}$ (PDB: 3PBP) (**Fig. 6.3**).

Second, the EDC and DSS cross-links are highly consistent with each other, despite different chemistries, and there is significant highly non-random clustering of both EDC and DSS cross-links into equivalent ‘‘cliques’’ (**Fig. 6.2A**). These represent adjacencies, as validated by those cliques that coincide with known crystallographic interface regions, such as Nup159:Dyn2 (PDB: 4DS1) (22) and Nup159:Nup82 (PDB: 3PBP) (8); indeed, in our final calculated structure these cliques represent immediately adjacent regions in the complex (**Fig. 6.2B**).

Third, considering the more abundant DSS cross-links, as can be seen from Fig. S4D (left), relatively few cross-links (< 7%) remain unsatisfied by our structures. Of those that are not satisfied, most involve relatively modest distance violations that can clearly be rationalized by locally limited flexibility of the proteins, as shown in **Fig. 6.2B** (cross-

link distance distributions). Moreover, those few cross-links in violation of strict distance limits in our structure are nevertheless right next to one of the cliques; they are thus consistent with the structure when locally limited flexibility is taken into account (**Fig. 6.2A** and S4D) (13).

Fourth, the solutions also fit the EM class averages, with an average cross-correlation coefficient of 0.931 (**Fig. 6.2C**; **Table 1**). Finally, 99% of the top 463 solutions satisfied the excluded volume and sequence connectivity restraints under the combined score threshold of 500.

Satisfaction of Data that Were Not Used to Compute Structures

In principle, our Bayesian modeling already effectively includes cross-validation via its Bayesian scoring function and sampling (13). However, the most direct test of a modeled structure is by comparing it to the data that were not used to compute it (a generalization of cross-validation). A structure can be validated directly against experimental data deliberately omitted from the structural model calculation (62). This goal is achieved by excluding a subset of the experimental data from structure calculation, followed by evaluation of the resulting structures against the omitted subset of data. This procedure is analogous to the one used for calculating the crystallographic R_{free} parameter and can be used to assess both the structure and the input data.

First, mass tagging of our structure is consistent with the localization of GFP tags on both the Nup82 and Nup159 C-termini (See ‘‘GFP mass-tagging analysis of the Nup82 holo-complex by immuno-EM’’ below and **Fig. 6.2C**).

Second, our structure is consistent with the previously published data, including an independent negative stain 3D density map (Fig. S5A) (9). Our asymmetric ~19 nm

long structure bears a general resemblance to the Nup82 complex class averages by Gaik *et al.*, except for having mostly one Dyn2 dimer at its end instead of five dimers (9).

Third, the trimeric coiled-coil structure between the helical Nup82-Nup159-Nsp1 regions is recapitulated even when computed using the chemical cross-linking data alone, without using the EM class averages (Fig. S5C). We modeled the trimer using the available crystallographic structures, the helical regions predicted by PSIPRED (57), and the cross-links. All crystallographic structures and predicted helical regions were kept rigid. We used an ideal helix template to construct the coordinates of the predicted helical regions. We adopted the same multi-scale approach used to represent the entire Nup82 complex described above. The 500 best-scoring solutions satisfied all cross-links. The structural clustering of the 500 best-scoring solutions revealed that regions Nup82$^{522\text{-}612}$-Nup159$^{1211\text{-}1321}$-Nsp1$^{637\text{-}727}$ were consistently arranged into a trimeric helical bundle.

Fourth, our structure is in agreement with SAXS profiles and *ab initio* shapes of Nup82 constructs spanning residues 4-220, 4-452, and 572-690 (**Fig. 6.2D** and S5D–S5F; Table S4). Notably, the Nup82 coiled-coil (572-690) forms a kinked structure and the corresponding SAXS profile shows a tendency of monotonous increase in the Kratky plot (Fig. S5F), indicating a high degree of flexibility between coiled-coil segments in solution, as would be expected for coiled-coils that form two different conformers as seen in the final structure.

Finally, our structure is also validated by the non-random and clustered distribution of cross-links connecting the Nup82 holo-complex to other parts of the NPC, revealing interaction sites, as described in ‘‘Docking of the Nup82 holo-complex and the Y-shape Nup84 complex’’ below.

GFP Mass-Tagging Electron Microscopy

Two different types of GFP-tagged structures of the Nup82 holo-complex were generated by attaching a rigid-body GFP structure (PDB: 1GFL) to either the Nup82 or Nup159 C-termini via the 14 linker residues of DPLALPVATPGIPM. For the Nup82 complex, the best-scoring structure was used. The configuration of the GFP tags was optimized using the replica exchange Gibbs sampling as described above using IMP. In summary, 10 independent sampling runs were performed, each run starting with a random initial configuration of the GFP tags. 4 replicas were used for each run, with temperatures ranging between 1.0 and 2.5. We produced a total of 50,000 structures each for the Nup82 and Nup159 GFP tags, using the EM 2D restraint of the corresponding immuno-EM class average. As a result, the best-scoring model structures are consistent with the localization of the GFP tags on both the Nup82 (ccc = 0.953) and Nup159 (ccc = 0.932) C-termini (**Fig. 6.2C**).

Docking of the Nup82 Holo-complex and the Y-Shape Nup84 Complex

A structure of the Nup82 holo-complex interacting with the Y-shape Nup84 complex was obtained by rigid-body docking restrained by 9 chemical cross-links identified at the interface (Table S3), using the replica exchange Gibbs sampling using IMP, as described above. For the Nup82 complex, the best-scoring structure was used. For the Nup84 complex, our previous structure (Shi et al., 2014) was refined by using new crystallographic structures of the complex subunits (PDB: 4XMM and 4YCZ) (25, 26). Next, 20 independent sampling runs were performed, each run starting with a random initial configuration. 6 replicas were used for each run, with temperatures ranging between 1.0 and 2.5. We produced a total of 100,000 structures using the crosslink restraints

spanning the interface between the Nup82 holo-complex and the Nup84 complex. Subsequently, 200 top-scoring structures were subjected to the clustering analysis, identifying 3 clusters (clusters A, B, and C; 86, 70, and 44 structures, respectively) of solution structures (Fig. S6A). At least 7 out of the 9 chemical cross-links were satisfied by the 200 top-scoring structures, within the distance threshold of 35 Å. All our solutions were similar, differing only in the degree of the Nup82 complex rotation along its long axis, relative to the Nup84 complex (Fig. S6B). Precisions of the Nup82 holo-complex in the 3 clusters were 30.2, 11.0, and 39.0 Å, respectively. Among the three clusters, only cluster C satisfied the cross-links used to compute them (Table S3) and the *S. cerevisiae* NPC localization probability density map (fit score by overlapping volume = 0.46, Fig. 5A and S6C) (12). Notably, this cluster of solutions is also the only one that aligns with the wild-type human NPC tomographic cryo-EM map (**Fig. 6.5B** and S6D, EMDB 2444) (28) and the mutant one lacking an outer cytoplasmic Y-complex ring (**Fig. 6.5C**, EMDB 3104) (34). The cross-correlation coefficients between the Nup82 holo-complex structure and the human NPC tomographic cryo-EM maps are 0.72 (wild-type, **Fig. 6.5B** and S6D) and 0.81 (mutant, **Fig. 6.5C**) in cluster C (**Table 1**). The cross-correlation coefficients were calculated using the measure correlation command in the UCSF Chimera software (https://www.cgl.ucsf.edu/chimera/).

Data and software availability

Software

The modeling protocol (i.e., stages 2, 3, and 4) was scripted using the Python Modeling Interface (PMI), version c7411c3, a library for modeling macromolecular complexes

based on our open-source Integrative Modeling Platform (IMP) package, version 2.5 (https:// integrativemodeling.org) (21).

To display the CX-MS data we used the software CX-Circos (http://cx-circos.net).

Data Resources

The chemical cross-linking with mass spectrometric readout data used in this study was deposited in the Chorus database (https:// chorusproject.org/pages/index.html).

Files containing the input data, modeling scripts, and output structures are available online (https://salilab.org/nup82; https:// github.com/salilab/nup82).

Supplemental Information

Supplemental Information includes seven figures, four tables, and two movies and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2016.10.028.

Author Contributions

Conceptualization, J.F.-M., S.J.K., Y.S., R.P., B.T.C., A.S., and M.P.R.;

Investigation, J.F.-M., S.J.K., P.U., Y.S., R.W., I.N., W. Z., W.J.R., M.G. and D.Z.;

Formal Analysis, S.J.K., R.P., J.W., I.E.C. and A.S.;

Writing, J.F.-M., S.J.K., A.S., B.T.C., and M.P.R.;

Funding Acquisition, D.L.S., D.Z., A.S., B.T.C., and M.P.R.;

Supervision, D.L.S., D.Z., A.S., B.T.C., and M.P.R.

Acknowledgments

We would like to thank S.R. Wente, R. Sadeh, and A. Krutchinsky for sharing yeast strains and plasmids; K. Uryu and the EMRC Resource Center at The Rockefeller University for assistance with negative stain EM; NYSGRC for providing samples for SAXS; T. Matsui and T.M. Weiss at SSRL, SLAC National Accelerator Laboratory for assistance with

collecting SAXS data; and B. Raveh at UCSF for computing FG Nup models. Support was provided by the Simons Foundation grant 349247 (Simons Electron Microscopy Center, NYSBC), the NSERC, Canadian Institutes of Health Research grant MOP232642, and the Canadian Foundation for Innovation (D.Z.), as well as NSF graduate research fellowship 1650113 (I.E.C.) and NIH grants U54 GM103511 (B.T.C., A.S., and M.P.R.), R01 GM112108 (M.P.R.), P41 GM109824 (M.P.R., A.S., and B.T.C.), P41 GM103314 (B.T.C.), and R01 GM083960 (A.S.).

References

1. K. E. Knockenhauer, T. U. Schwartz, The Nuclear Pore Complex as a Flexible and Dynamic Gate. *Cell* **164**, 1162–71 (2016).
2. M. Oeffinger, D. Zenklusen, To the pore and through the pore: a story of mRNA export kinetics. *Biochim Biophys Acta* **1819**, 494–506 (2012).
3. A. W. Folkmann, K. N. Noble, C. N. Cole, S. R. Wente, Dbp5, Gle1-IP6 and Nup159: a working model for mRNP export. *Nucleus* **2**, 540–8 (2011).
4. H. M. Kaneb, *et al.*, Deleterious mutations in the essential mRNA metabolism factor, hGle1, in amyotrophic lateral sclerosis. *Hum Mol Genet* **24**, 1363–73 (2015).
5. H. O. Nousiainen, *et al.*, Mutations in mRNA export mediator GLE1 result in a fetal motoneuron disease. *Nat Genet* **40**, 155–7 (2008).
6. H. Chug, S. Trakhanov, B. B. Hulsmann, T. Pleiner, D. Gorlich, Crystal structure of the metazoan Nup62*Nup58*Nup54 nucleoporin complex. *Science* **350**, 106–10 (2015).
7. T. Stuwe, *et al.*, Architecture of the fungal nuclear pore inner ring complex. *Science* **350**, 56–64 (2015).
8. K. Yoshida, H. S. Seo, E. W. Debler, G. Blobel, A. Hoelz, Structural and functional analysis of an essential nucleoporin heterotrimer on the cytoplasmic face of the nuclear pore complex. *Proc Natl Acad Sci U S A* **108**, 16571–6 (2011).
9. M. Gaik, *et al.*, Structural basis for assembly and function of the Nup82 complex in the nuclear pore scaffold. *J Cell Biol* **208**, 283–97 (2015).
10. B. Montpetit, *et al.*, A conserved mechanism of DEAD-box ATPase activation by nucleoporins and InsP6 in mRNA export. *Nature* **472**, 238–42 (2011).

11. M. K. Lund, C. Guthrie, The DEAD-box protein Dbp5p is required to dissociate Mex67p from exported mRNPs at the nuclear rim. *Mol Cell* **20**, 645–51 (2005).
12. F. Alber, *et al.*, The molecular architecture of the nuclear pore complex. *Nature* **450**, 695–701 (2007).
13. Y. Shi, *et al.*, Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. *Mol Cell Proteomics* **13**, 2927–2943 (2014).
14. A. Sali, *et al.*, Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure* **23**, 1156–67 (2015).
15. Y. Shi, *et al.*, A strategy for dissecting the architectures of native macromolecular assemblies. *Nat Methods* **12**, 1135–8 (2015).
16. J. M. Pratt, *et al.*, Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nat Protoc* **1**, 1029–43 (2006).
17. A. Sali, T. L. Blundell, Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779–815 (1993).
18. J. Soding, Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–60 (2005).
19. L. E. Hough, *et al.*, The molecular mechanism of nuclear transport revealed by atomic-scale measurements. *Elife* **4** (2015).
20. A. R. Borneman, *et al.*, The genome sequence of the wine yeast VIN7 reveals an allotriploid hybrid genome with Saccharomyces cerevisiae and Saccharomyces kudriavzevii origins. *FEMS Yeast Res* **12**, 88–96 (2012).

21. D. Russel, *et al.*, Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. *PLOS Biology* **10**, e1001244 (2012).
22. E. M. Romes, A. Tripathy, K. C. Slep, Structure of a yeast Dyn2-Nup159 complex and molecular basis for dynein light chain-nuclear pore interaction. *J Biol Chem* **287**, 15862–73 (2012).
23. J. Fernandez-Martinez, *et al.*, Structure-function mapping of a heptameric module in the nuclear pore complex. *J Cell Biol* **196**, 419–34 (2012).
24. N. Kellner, *et al.*, Developing genetic tools to exploit Chaetomium thermophilum for biochemical analyses of eukaryotic macromolecular assemblies. *Sci Rep* **6**, 20937 (2016).
25. K. Kelley, K. E. Knockenhauer, G. Kabachinski, T. U. Schwartz, Atomic structure of the Y complex of the nuclear pore. *Nat Struct Mol Biol* **22**, 425–31 (2015).
26. T. Stuwe, *et al.*, Nuclear pores. Architecture of the nuclear pore complex coat. *Science* **347**, 1148–52 (2015).
27. F. Alber, *et al.*, Determining the architectures of macromolecular assemblies. *Nature* **450**, 683–694 (2007).
28. K. H. Bui, *et al.*, Integrated structural analysis of the human nuclear pore complex scaffold. *Cell* **155**, 1233–43 (2013).
29. Y. Strahm, *et al.*, The RNA export factor Gle1p is located on the cytoplasmic fibrils of the NPC and physically interacts with the FG-nucleoporin Rip1p, the DEAD-box protein Rat8p/Dbp5p and a new protein Ymr 255p. *EMBO J* **18**, 5761–77 (1999).

30. Y. Ren, H. S. Seo, G. Blobel, A. Hoelz, Structural and functional analysis of the interaction between the nucleoporin Nup98 and the mRNA export factor Rae1. *Proc Natl Acad Sci U S A* **107**, 10406–11 (2010).
31. R. L. Adams, L. J. Terry, S. R. Wente, Nucleoporin FG domains facilitate mRNP remodeling at the cytoplasmic face of the nuclear pore complex. *Genetics* **197**, 1213–24 (2014).
32. P. Sampathkumar, *et al.*, Atomic structure of the nuclear pore complex targeting domain of a Nup116 homologue from the yeast, Candida glabrata. *Proteins* **80**, 2110–6 (2012).
33. E. Fabre, E. Hurt, Yeast genetics to dissect the nuclear pore complex and nucleocytoplasmic trafficking. *Annu Rev Genet* **31**, 277–313 (1997).
34. A. von Appen, *et al.*, In situ structural analysis of the human nuclear pore complex. *Nature* **526**, 140–3 (2015).
35. Y. Cheng, W. Boll, T. Kirchhausen, S. C. Harrison, T. Walz, Cryo-electron tomography of clathrin-coated vesicles: structural implications for coat assembly. *J Mol Biol* **365**, 892–9 (2007).
36. D. Devos, *et al.*, Components of coated vesicles and nuclear pore complexes share a common molecular architecture. *PLoS Biol* **2**, e380 (2004).
37. Q. Yang, M. P. Rout, C. W. Akey, Three-dimensional architecture of the isolated yeast nuclear pore complex: functional and evolutionary implications. *Mol Cell* **1**, 223–34 (1998).

38. J. Yamada, *et al.*, A bimodal distribution of two distinct categories of intrinsically disordered structures with separate functions in FG nucleoporins. *Mol Cell Proteomics* **9**, 2205–24 (2010).
39. J. Kosinski, *et al.*, Molecular architecture of the inner ring scaffold of the human nuclear pore complex. *Science* **352**, 363–5 (2016).
40. D. H. Lin, *et al.*, Architecture of the symmetric core of the nuclear pore. *Science* **352**, aaf1015 (2016).
41. K. Strasser, J. Bassler, E. Hurt, Binding of the Mex67p/Mtr2p heterodimer to FXFG, GLFG, and FG repeat nucleoporins is essential for nuclear mRNA export. *J Cell Biol* **150**, 695–706 (2000).
42. C. Trahan, M. Oeffinger, Targeted cross-linking-mass spectrometry determines vicinal interactomes within heterogeneous RNP complexes. *Nucleic Acids Res* **44**, 1354–69 (2016).
43. D. Grunwald, R. H. Singer, In vivo imaging of labelled endogenous beta-actin mRNA during nucleocytoplasmic transport. *Nature* **467**, 604–7 (2010).
44. A. Mor, *et al.*, Dynamics of single mRNP nucleocytoplasmic transport and export through the nuclear pore in living cells. *Nat Cell Biol* **12**, 543–52 (2010).
45. C. Smith, *et al.*, In vivo single-particle imaging of nuclear mRNA export in budding yeast demonstrates an essential role for Mex67p. *J Cell Biol* **211**, 1121–30 (2015).
46. P. Nerurkar, *et al.*, Eukaryotic Ribosome Assembly and Nuclear Export. *Int Rev Cell Mol Biol* **319**, 107–40 (2015).
47. D. N. Simon, M. P. Rout, Cancer and the nuclear pore complex. *Adv Exp Med Biol* **773**, 285–307 (2014).

48. K. N. Noble, *et al.*, The Dbp5 cycle at the nuclear pore complex during mRNA export II: nucleotide cycling and mRNP remodeling by Dbp5 are controlled by Nup159 and Gle1. *Genes Dev* **25**, 1065–77 (2011).
49. A. W. Folkmann, T. R. Dawson, S. R. Wente, Insights into mRNA export-linked molecular mechanisms of human disease through a Gle1 structure-function analysis. *Adv Biol Regul* **54**, 74–91 (2014).
50. F. Kendirgi, D. M. Barry, E. R. Griffis, M. A. Powers, S. R. Wente, An essential role for hGle1 nucleocytoplasmic shuttling in mRNA export. *J Cell Biol* **160**, 1029–40 (2003).
51. O. M. Griffith, Techniques of Preparative, Zonal, and Continuous Flow Ultracentrifugation (Beckman Instruments, Inc., 1994).
52. H. P. Erickson, Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy. *Biol Proced Online* **11**, 32–51 (2009).
53. C. Ding, *et al.*, Quantitative analysis of cohesin complex stoichiometry and SMC3 modification-dependent protein interactions. *J Proteome Res* **10**, 3652–9 (2011).
54. J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367–72 (2008).
55. Z. Yang, J. Fang, J. Chittuluru, F. J. Asturias, P. A. Penczek, Iterative stable alignment and clustering of 2D transmission electron microscope images. *Structure* **20**, 237–47 (2012).

56. C. S. Weirich, J. P. Erzberger, J. M. Berger, K. Weis, The N-terminal domain of Nup159 forms a beta-propeller that functions in mRNA export by tethering the helicase Dbp5 to the nuclear pore. *Mol Cell* **16**, 749–60 (2004).

57. D. T. Jones, Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**, 195–202 (1999).

58. J. J. Ward, L. J. McGuffin, K. Bryson, B. F. Buxton, D. T. Jones, The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**, 2138–9 (2004).

59. A. Lupas, M. Van Dyke, J. Stock, Predicting coiled coils from protein sequences. *Science* **252**, 1162–4 (1991).

60. J. Trigg, K. Gutwin, A. E. Keating, B. Berger, Multicoil2: predicting coiled coils and their oligomerization states from sequence in the twilight zone. *PLoS One* **6**, e23519 (2011).

61. J. H. McDonald, *Handbook of biological statistics* (Sparky House Publishing Baltimore, MD, 2009).

62. M. T. Degiacomi, *et al.*, Molecular assembly of the aerolysin pore reveals a swirling membrane-insertion mechanism. *Nat Chem Biol* **9**, 623–9 (2013).

Chapter VII - Molecular Architecture of the Major Membrane Ring Component of the Nuclear Pore Complex

Contributing authors

Xiaorong Wang[1], Ilan E. Chemmama[2], Clinton Yu[1], Alexander Huszagh[1], Yue Xu[3], Rosa Viner[4], Sarah A. Block[5], Peter Cimermancic[2], Scott D. Rychnovsky[5], Yihong Ye[3], Andrej Sali[2], Lan Huang[1]

[1]Department of Physiology and Biophysics, University of California, Irvine, California 92697

[2]Department of Bioengineering and Therapeutic Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biosciences, University of California, San Francisco, California 94158

[3]Laboratory of Molecular Biology, NIDDK, National Institutes of Health, Bethesda, Maryland 20892

[4]Thermo Fisher Scientific, San Jose, California 94134

[5]Department of Chemistry, University of California, Irvine, California 92697

Contact: lanhuang@uci.edu

Abstract

The membrane ring that equatorially circumscribes the nuclear pore complex (NPC) in the perinuclear lumen of the nuclear envelope is composed largely of Pom152 in yeast and its ortholog Nup210 (or Gp210) in vertebrates. Here, we have used a combination of

negative-stain electron microscopy, nuclear magnetic resonance, and small-angle X-ray scattering methods to determine an integrative structure of the ~120 kDa luminal domain of Pom152. Our structural analysis reveals that the luminal domain is formed by a flexible string-of-pearls arrangement of nine repetitive cadherin-like Ig-like domains, indicating an evolutionary connection between NPCs and the cell adhesion machinery. The 16 copies of Pom152 known to be present in the yeast NPC are long enough to form the observed membrane ring, suggesting how interactions between Pom152 molecules help establish and maintain the NPC architecture.

Introduction

The eukaryotic nucleus is delimited by the nuclear envelope (NE), composed of two distinct membranes, the inner and the outer nuclear membranes, that enclose the perinuclear lumen. The outer and inner nuclear membranes join to form specialized circular apertures (nuclear pores), containing large proteinaceous assemblies termed nuclear pore complexes (NPCs) (1). The yeast NPC is a large (~50 MDa) cylindrical assembly composed of multiple copies of ~30 different proteins, termed nucleoporins or Nups, arranged to form eight symmetrically arranged spokes linked by coaxial outer, inner, and membrane rings (2, 3). NPCs facilitate the active transport of macromolecules between the nucleoplasm and cytoplasm and are involved in other multiple essential roles, including controlling genome organization and expression (4). As a consequence, disruptions of the NPC can lead to human disease (5).

It has been shown that the NPC has at its heart a cage-like core scaffold consisting of Nups composed entirely of either a β-propeller fold, an α-solenoid fold, or a distinctive arrangement of both folds, a combination otherwise unique to vesicle-coating complexes (6). These similarities suggest a common evolutionary origin for NPCs and coated vesicles in an early membrane-curving module or ‘‘protocoatomer’’ that led to the formation of the internal membrane systems defining the features of modern eukaryotes (6, 7). This coatomer-like core scaffold is anchored to the pore membrane through two different mechanisms. First, ALPS (amphipathic lipid packing sensor) motifs, membrane-binding α-helical ‘‘fingers,’’ are found on the membrane-facing surface of the NPC core scaffold (8–10). Second, several Nups, termed pore membrane proteins or Poms, carry trans-membranous α-helices (11–13). Curiously, none of these transmembrane domains

or ALPS motifs seem individually essential for NPC assembly or membrane anchoring, suggesting functional redundancy (14). One particular Pom stands out by virtue of its size and its apparent homo-oligomerization (15, 16) to form the membrane ring that equatorially circumscribes the NPC in the perinuclear lumen of the NE (1). In yeast, this protein is termed Pom152 (17), a type II integral membrane protein (15) that has an N-terminal NPC-associating region followed by a single transmembrane domain, whereas its presumed vertebrate homolog Nup210 (also known as Gp210) has its transmembrane domain near the C terminus followed by the NPC-associating region (13, 18). Both homologs have a large luminal domain that was previously suggested to be formed by repeated domains (17) of an Ig or cadherin-like fold (19). However, there is no available experimental evidence defining the precise number and structure of these domains or for the protein as a whole. Although Pom152 is non-essential in yeast, its overexpression significantly inhibits cell growth (17), and it has been implicated in helping to form an early intermediate structure during NPC assembly (20). In vertebrates, Nup210 is a key regulator of cell-fate adaptation (21, 22). Mutation and mis-regulation of Nup210 have been related to severe human diseases, including numerous cancers (23, 24). Here, we have used a combination of negative-stain electron microscopy (EM), nuclear magnetic resonance (NMR), and small-angle X-ray scattering (SAXS) methods to determine an integrative structure of the ~120 kDa luminal domain of Pom152.

Results

Negative-Stain Electron Microscopy Analysis of Pom152

To determine the overall shape and dimensions of the native full-length Pom152 (Pom152^{FL}; **Fig. 7.1A**), we purified the endogenous protein as a monomer using affinity purification, native elution, and sucrose density gradients (25) (**Fig. 7.1** and S1A). The samples were visualized by negative-stain EM. The individual particle images display varying degrees of curvature, although the dominant form was elongated with a pronounced curvature, convex on the side adjacent to the head (**Fig. 7.1**). The particle images were analyzed using an iterative stable alignment and clustering (ISAC) method to generate class averages that were reproducible in multiple classification trials (26). All 35 resulting class averages showed a thin elongated shape for the isolated monomeric protein molecule (**Fig. 7.1B**). The average end-to-end distance of Pom152^{FL} class averages is 38.7 nm (±1.5 nm; 30 class averages) (**Fig. 7.1G**).

To further assess the structural features of Pom152^{FL}, we relied on the random conical tilt method (27) to compute an initial 3D map (**Fig. 7.1C**) and then used the Relion program to compute the final 3D map at 25 Å resolution (28) (**Fig. 7.1D**). The resolution is limited by incomplete angular coverage of particle views and conformational heterogeneity (Fig. S1D and S1E). Nevertheless, the map faithfully recapitulates the major features seen in the 2D class averages. Pom152^{FL} has a prominent head that is attached to a long tail resembling a string-of-pearls (**Fig. 7.1B**). This tail is apparently formed by nine consecutive globular domains (**Fig. 7.1D**) and exhibits heterogeneity in the observed conformational states, probably due to changes in the relative orientations of the globular domains with respect to each other (**Fig. 7.1B** and **7.1C**). The arrangement

of the first three domains (the neck) is relatively linear, while domains 3–9 form a curved shape in all observed classes. The estimated inter-domain angle between the latter globular domains (3–9) in most particle images ranges from -10º to +5º, with an average of -4.1º (**Fig. 7.1E**, **7.1B**, and S1C). Some curvature is retained even as the domains are sequentially removed, supporting the idea that the curvature is an intrinsic property of the tail (**Fig. 7.1F**; below).

To define which of the morphological features correspond to the N- and C-terminal domains of Pom152, we used negative-stain EM to analyze two C-terminally truncated versions of the protein. All resulting class averages for the C-terminal truncations $Pom152^{1\text{-}1,135}$ and $Pom152^{1\text{-}936}$ showed an intact head, but were missing a number of globular domains proportional to the size of the deletion, as reflected by their average end-to-end distance of 28.9 nm (±1.6 nm; 28 class averages) for $Pom152^{1\text{-}1,135}$ and 24.4 nm (±1.0 nm; 33 class averages) for $Pom152^{1\text{-}936}$ (**Fig. 7.1A**, **7.1F**, and **7.1G**). This finding indicated that the tail corresponds to the C-terminal luminal domain of Pom152 ($Pom152^{LD}$) and that the head contains the N-terminal NPC-associating region and the transmembrane domain. The truncated Pom152 particles showed a degree of heterogeneity similar to that of $Pom152^{FL}$. The difference between the end-to-end distances of $Pom152^{FL}$ and the truncated forms ($Pom152^{1\text{-}1,135}$ and $Pom152^{1\text{-}936}$) suggested an average size of 4.0 nm (±2.4 nm; 20 class averages) for each globular domain. The average width of the domains is 2.9 nm (±1.6 nm; 37 class averages).

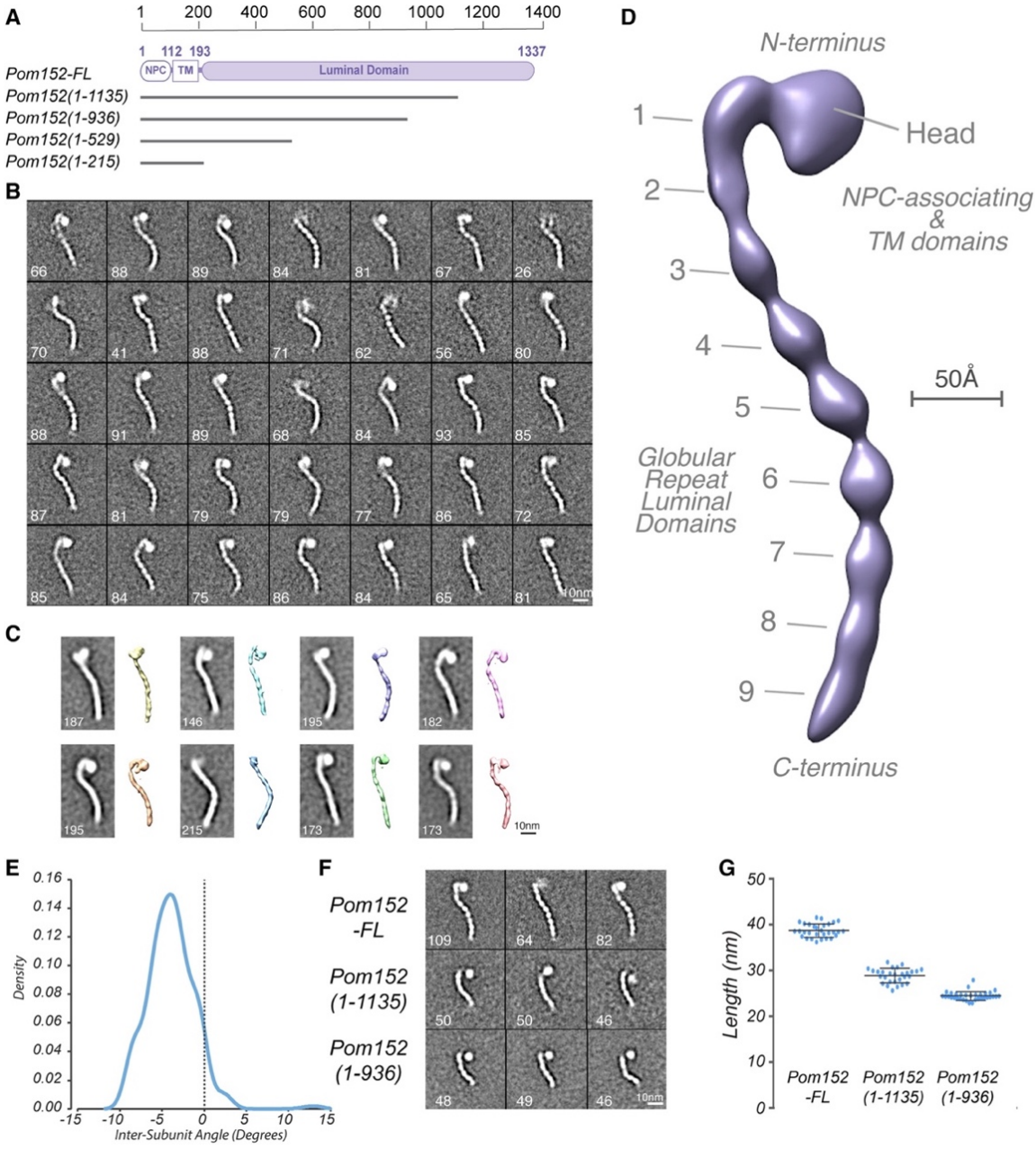

Figure 7.1 | Negative-Stain EM Analysis Shows that Pom152 has an Extended, String-of-Pearls-Shaped Luminal Domain.

(A) Domain organization of Pom152FL and four truncations drawn to scale. Pom152FL exhibits a three-domain organization with the NPC-associating domain (NPC), the transmembrane segment (TM), and the domain inside the perinuclear lumen of the nuclear envelope (luminal domain). The numbers indicate amino acid residue positions. Horizontal gray lines for the truncations represent the number of amino acid residues in each segment. (B) Thirty-five representative negative-stain EM class averages of Pom152FL. The number of particles in each class is shown. Bar, 10 nm. (C) Representative random conical tilt 3D maps (right) are aligned to each of the corresponding class averages (left). The number of particles in each class is shown. Bar,

10 nm. (D) Negative-stain EM density map of Pom152FL. The nine globular domains in the C-terminal lumen (1–9) and the N-terminal head region containing the NPC-associating and TM domains are indicated. Bar, 50 Å. (E) The average inter-domain angle for the last seven repetitive regions was estimated in 37 representative Pom152FL class averages using the ImageJ angle tool. One angle was measured between domains 3 and 9, its difference from 180° was determined, and the resulting value divided between the seven globular domains involved in the estimation. Distribution of the resulting inter-domain angles is shown in a Kernel density plot with a peak of −4.1°, indicating a small negative curvature for the particles. (F) Representative negative-stain EM class averages of three assigned views of Pom152FL, Pom1521−1,135, and Pom1521−936. The number of particles in each class is shown. Bar, 10 nm. (G) End-to-end distances of Pom152FL, Pom1521−1,135, and Pom1521−936 in samples of 30, 28, and 33 class averages, respectively. The lines on the data points indicate the mean and SD: 38.7 ± 1.5 nm, 28.9 ± 1.6 nm and 24.4 ± 1.0 nm for the samples, respectively.

The Luminal Domain of Pom152 Stabilizes the NPC's Association with the Pore Membrane

Complete deletions of Pom152 have not been observed to change the fitness phenotype (17) but have been shown to cause synthetic defects when combined with mutations affecting inner-ring nucleoporins (15) or other integral membrane proteins involved in NPC biogenesis, such as Heh1 (16) or Apq12 (29). Pom152 is a part of the NPC membrane ring and has been suggested to be involved in shaping and stabilizing the NPC's pore membrane (1, 16, 30). Indeed, Nups with ALPS-motif associated with the pore membrane also appear to help stabilize the membrane (8, 20, 31), as is exemplified by sensitivity to growth in the membrane-destabilizing reagent benzyl-alcohol (25). We thus decided to test the phenotype of $Pom152^{LD}$ truncation mutants in the presence of benzyl-alcohol.

Addition of benzyl-alcohol caused especially clear growth defects in the luminal domain mutants (**Fig. 7.2A**). Complete deletion of Pom152 does not show such obvious

defects, suggesting that the observed phenotype is mainly related to the functional role of the luminal domain. When the same mutants were tested for thermosensitivity, no growth defect was observed (**Fig. 7.2A**), showing that the benzyl-alcohol phenotype is a specific and not a general response to stress; also, addition of the chemical chaperone tauroursodeoxycholic acid (TUDCA) did not compensate for the benzyl-alcohol growth defect (not shown), suggesting that the phenotype is not associated with the ER stress response pathway, as shown for hNup210 (22). To control for the possibility of subcellular mislocalization causing the Pom152 truncation mutant phenotypes, we used fluorescently tagged reporters to observe the localization of the Pom152 truncations in relation to other NPC markers. As shown in **Fig. 7.2B**, the truncated form of Pom152 shows the nuclear rim staining characteristic of a nucleoporin; the localization and distribution of NPCs, revealed by co-localization of the inner-ring component Nup188, appears normal. However, upon treatment with benzyl-alcohol, both Nup188 and the truncated Pom152 start to co-accumulate at cytoplasmic foci. Our results thus show that the luminal domain of Pom152 is functionally relevant and suggest that it plays key roles in shaping and stabilizing the NPC structure.

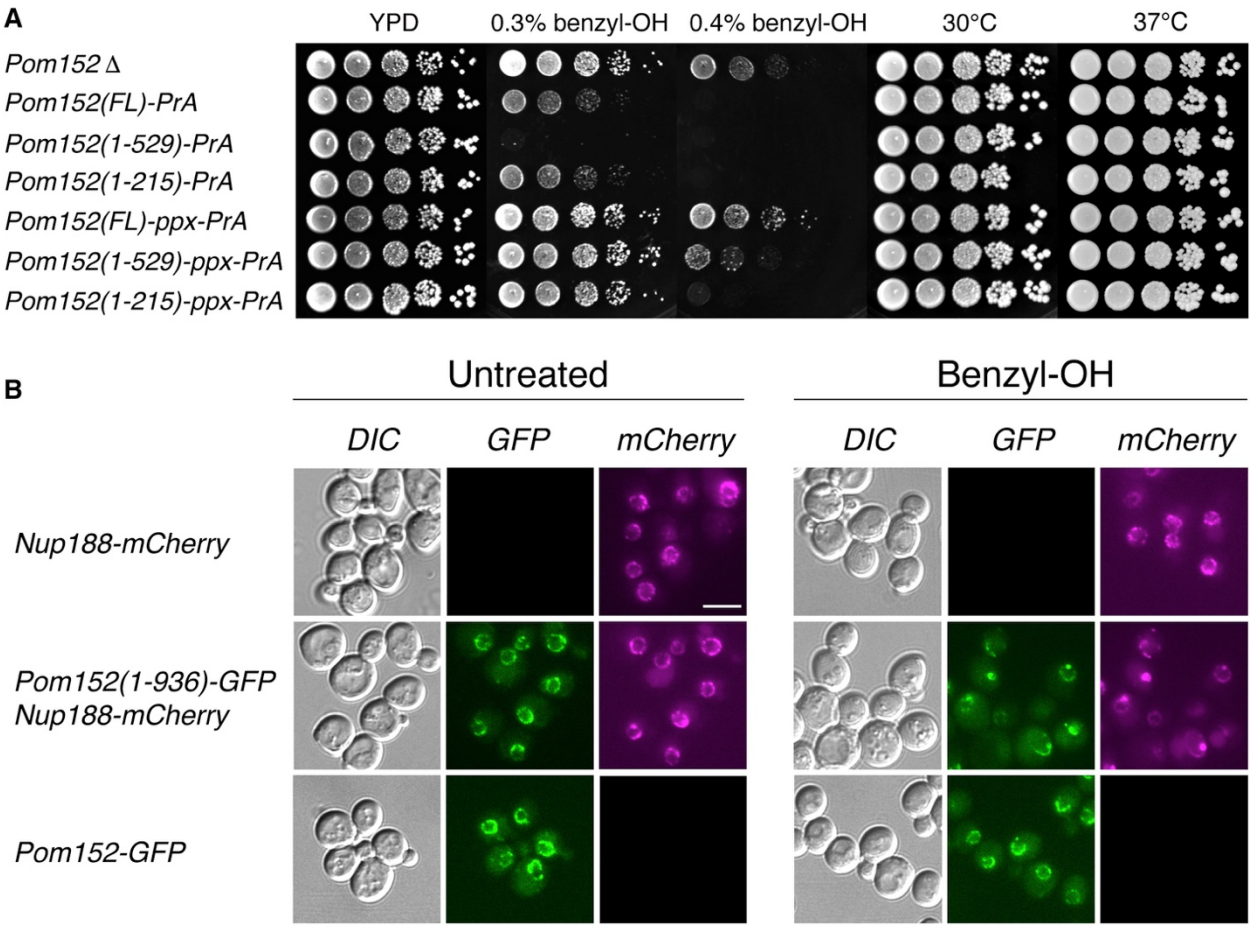

Figure 7.2 | Functional Analysis of Truncations Affecting the Pom152 Luminal Domain.

(**A**) Growth phenotypes of $Pom152^{FL}$ tagged with Protein A or Prescission protease cleavage site (PPX)-Protein A and related truncation mutants (see Fig. 1A). Serial 10-fold dilutions of cells were spotted on YPD plates in the absence or presence of 0.3% and 0.4% benzyl-OH at 30°C or on YPD plates and grown at the indicated temperatures for 1–3 days. (**B**) Subcellular localization of benzyl-OH treated Pom152-GFP constructs. Panels show the localization of the indicated genomically tagged Pom152-GFP or Nup188-mCherry (used as NPC localization control) constructs as determined by fluorescence microscopy. Cells were grown on liquid yeast minimal medium supplemented with 2% glucose at 30°C (untreated) or with an additional 0.1% benzyl-OH for 3 hr at 30°C (benzyl-OH). Differential interference contrast (DIC). Bar, 5 μm.

Structure Determination of Pom152$^{718–820}$, the Luminal Ig-like Domain, Using NMR Spectroscopy

Negative-stain EM revealed that Pom152^{LD} is formed by nine globular domains whose arrangement displays a significant but limited degree of heterogeneity and/or flexibility (**Fig. 7.1**), making this domain challenging for structure determination. However, their small size made individual domains suitable for structure determination by solution NMR spectroscopy. Two segments (Pom152$^{603-820}$ corresponding to domains 3–4 and Pom152$^{718-820}$ corresponding to domain 4) were evaluated by recording their 2D ^{1}H-^{15}N HSQC (heteronuclear single-quantum coherence) spectra (Fig. S2). The line shapes and intensities of the peaks seen for Pom152$^{603-820}$ were non-uniform, making this segment intractable by NMR spectroscopy, whereas Pom152$^{718-820}$ HSQC spectra showed well-dispersed peaks with uniform line shapes and intensities. Pom152$^{718-820}$ was thus chosen for full structure characterization due to its apparently higher conformational homogeneity.

The backbone and side-chain resonances as well as distance restraints were determined for a [U-^{13}C, ^{15}N] Pom152$^{718-820}$ sample. The 20 best-scoring structures in the ensemble superimpose well on each other, with root-mean-square deviation (RMSD) of 0.60 ± 0.11 Å over N, Cα, and C backbone atoms; they also have good stereochemistry (**Fig. 7.3A** and **Table 1**). The solution structure of Pom152$^{718-820}$ revealed an Ig-like fold (**Fig. 7.3C** and **7.3D**) containing nine β strands that form two β sheets with a typical β-sandwich topology (**Fig. 7.3B**). The two β sheets are made of the ABE and C´CFGG´A´ β strands (using the standard Ig-like domain annotation) (**Fig. 7.3B**). The fold does not contain any inter-sheet disulfide bonds, resulting in a less compact β sandwich with a distance between the two sheets of 11.4 Å (between B3 [Ser739] and F3 [Ile792] Cα

atoms) that is larger than that in related Ig-like fold structures, such as cadherins (e.g., 9.4 Å between I33 and V79 in the neural cell adhesion molecule [NCAM]; PDB: 1EPF; (32)). Accordingly, we consider the Ig-like domain of Pom152$^{718\text{-}820}$ to be a variant of the C3-subtype Ig-like fold family (Table 2 in (33)). Structural mapping of sequence conservation of Pom152$^{718\text{-}820}$ among 37 fungal species reveals a conserved surface region lined by residues from the F, G, and G´ β strands and the FG loop, as well as highly conserved cysteine residues at the termini of the Ig domains (**Fig. 7.3E**, **7.3F**, and S6). The former region has a net negative electrostatic potential in its center, partly bounded by patches with a net positive potential (**Fig. 7.3G** and **7.3H**). We tested the possibility that this region is a binding site for calcium ions, as is the case for many cadherin-like proteins. However, addition of EDTA or calcium chloride did not significantly alter chemical shifts of Pom152$^{718\text{-}820}$ or the negative-stain EM images (not shown), suggesting that Pom152 does not bind calcium ions, unlike many other cadherin-like proteins.

Pom152 Luminal Domain Is Formed by an Array of Ig-like Domains

We predicted that the remaining eight domains in the luminal domain of Pom152 are also Ig-like domains, as was already revealed by NMR spectroscopy for Pom152$^{718\text{-}820}$. This prediction was based on the following four considerations. First, sequence-based predictions of secondary structure, disordered regions, and domain boundaries are consistent with an Ig-like fold (Fig. S3A–S3C). Second, the eight domains share statistically significant sequence alignments to Pom152$^{718\text{-}820}$ (E-values ranging from 3.3 x 10^{-49} to 6.8 x 10^{-39}) (**Fig. 7.4B**, **7.4C**, and S3D). Third, comparative models of these domains, constructed with MODELLER (34) using the NMR structure of Pom152$^{718\text{-}820}$ as

a template (**Fig. 7.4B** and **7.4C**), have reasonable stereochemistry, indicating all nine domains might assume the same fold. Finally, the cysteine residues in Pom152$^{718\text{-}820}$ that we identified as conserved across different species (above) are also conserved in the remaining eight domains (**Fig. 7.3E**, **7.3F**, and S6). As an aside, it is conceivable that putative disulfide bridges involving these cysteine residues help stabilize interfaces between the Ig-like domains of Pom152, intra- and/or inter-molecularly.

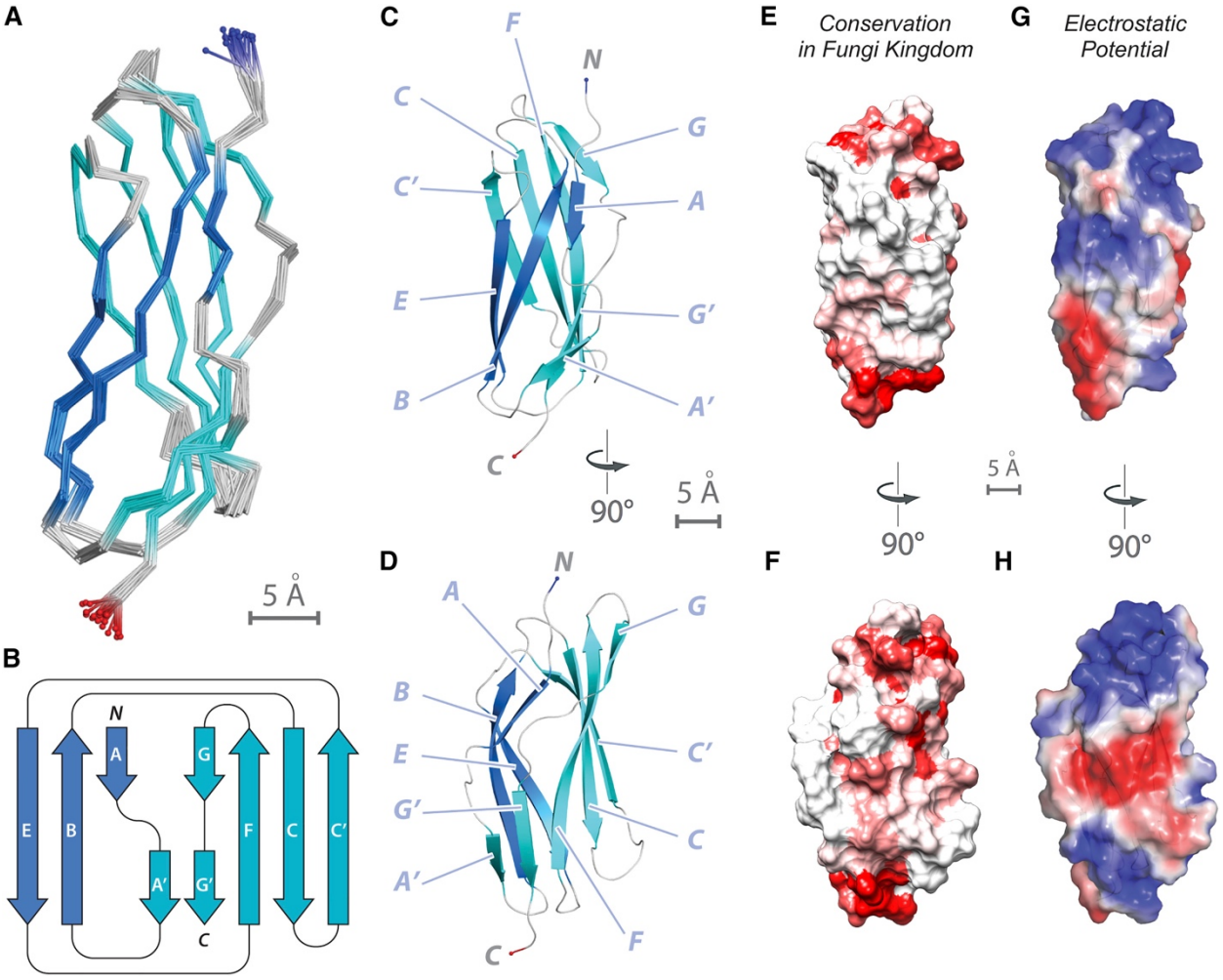

Figure 7.3 | NMR Structure Determination of Pom152718-820 Reveals a Conserved Ig-like Fold Domain

(**A**) Superimposition of the 20 lowest-energy structures of Pom152$^{718\text{-}820}$ that satisfy the NMR restraints best. The two β sheets are made of the ABE (blue) and C´CFGG´A β strands (cyan). Scale bar, 5 Å. (**B**) Topology of the Pom152 fold. β strands are represented as thick arrows and the linkers connecting them as lines. Each strand is

named following the standard Ig-like domain annotation (Halaby et al., 1999). (**C** and **D**) Two views of the best-scoring structure, showing the arrangement of β strands and features labeled as in (**B**). (**E** and **F**) Sequence conservation of Pom152$^{718\text{-}820}$ among 37 fungal homologs (Figure S6), plotted on the surface of the Pom152$^{718\text{-}820}$ structure; low conservation in white, high conservation in red. (**G** and **H**) Electrostatic potential on the surface of the Pom152$^{718\text{-}820}$ structure, calculated using the PyMOL tool APBS; negative (2 kT/e) and positive (+2 kT/e) potentials are shown in red and blue, respectively. See also Figure S2 and Table 1.

Table 7.1 | NMR Restraints and Structural Statistics for the 20 Lowest-Energy Structures of Pom152$^{718-820}$

aHydrogen-bond restraints were an HN-O distance of 1.8–2.3 Å and an N-O distance of 2.8–3.3 Å. bStructural characteristics for the final ensemble of 20 water-refined structures. cRMSD of the mean structure from individual structures in the ensemble. dRMSD for residues 718–820 shown. eRamachandran plot data shown for residues 718–820 (in bracket structured region: 722–725; 731–733; 737–746; 751–759; 766–773; 778–785; 789–800; 804–807; 813–817). See also Fig. 3 and S2.

Restraints and Statistics	Wild-Type
Total number of restraints	2,590
NOE restraints	2,346
Unambiguous	2,086
Intra-residue	863
Sequential	478
Short-range	115
Medium-range	27
Long-range	603
Ambiguous	260
Inter-molecular	
Dihedral angle restraints	176
Hydrogen-bond restraints[a]	68
Structure statistics[b]	

Restraints and Statistics	Wild-Type
NOE violations >0.5 Å	0%
Dihedral violations >5°	0%
RMSD from average structure[c,d]	
All residues (236–298)	
Backbone (N, Cα, C) (Å)	0.60 ± 0.11
Heavy atoms (Å)	1.20 ± 0.10
Ramachandran statistics[e]	
Most favored region (%)	78.9 (89.7)
Additionally allowed (%)	18.4 (8.9)
Generously allowed (%)	2.6 (1.4)
Disallowed (%)	0.0 (0.0)

A Structural Model of a Luminal Ig-like Domain in Human Nup210

Nup210 is the human ortholog of Pom152. Its luminal domain is a key regulator of cell differentiation (21, 22). The two orthologs share domain composition, although the order of the domains is swapped, with the luminal domain located at the N terminus of Nup210, while the transmembrane and NPC-associating domains are at its C terminus (**Fig. 7.4D**). The low sequence identity between the yeast and human luminal domains (~20%) prevented us from unambiguously defining the number and boundaries of the Nup210 Ig-like domains. Nevertheless, we were able to construct a comparative model of $Nup210^{1,079\text{-}1,152}$ by relying on our structure of $Pom152^{718\text{-}820}$ as a template (**Fig. 7.4E** and Supplemental Experimental Procedures). The NCBI Conserved Domain (35) server predicts that $Nup210^{1,079\text{-}1,152}$ is an Ig-like fold domain, based on an alignment (E-value 1.02×10^{-16}) to the Pfam family PF02368 (a bacterial Ig-like domain family), suggesting that the luminal domain of Nup210 is organized similarly to that of Pom152.

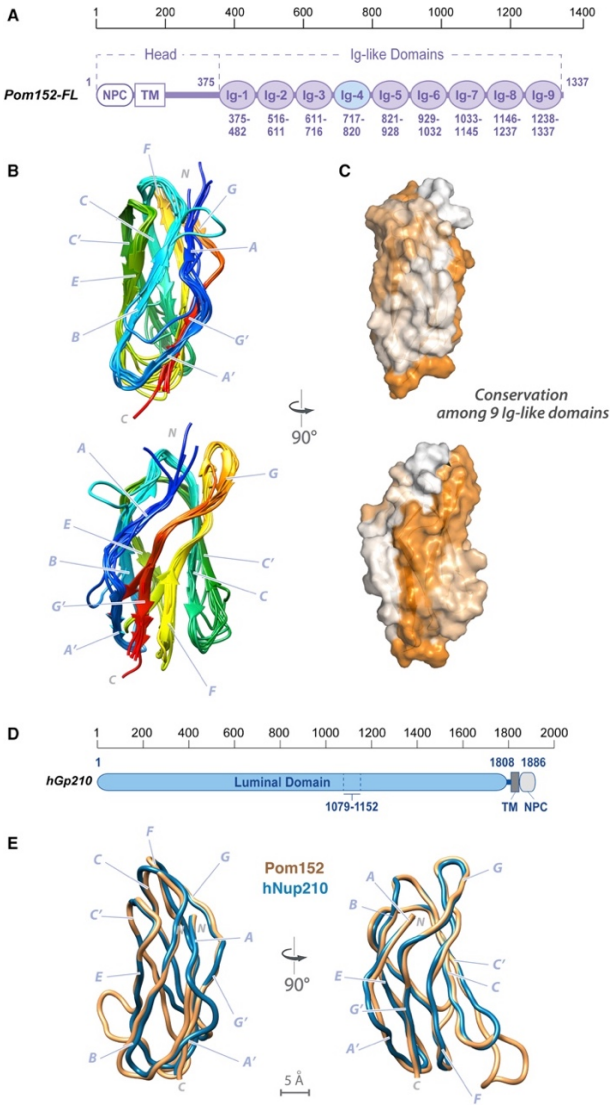

Figure 7.4 | Comparative Models of Eight Ig-like Domains and Comparison with an Ig-like Domain in Human Nup210

(A) Domain organization of Pom152^{FL} and its domains drawn to scale. The head region contains the domain that faces the NPC inner ring (NPC) and the transmembrane segment (TM), followed by the luminal domain composed of nine Ig-like repeats. The amino acid boundaries for each Ig-like repeat are indicated below them. The repeat analyzed by NMR (Ig-4, Pom152$^{718-820}$) is highlighted in light blue. (B) Comparative models of the remaining eight Ig-like domains were built using the Pom152$^{718-820}$ NMR structure as the template, followed by superposing them on the Pom152$^{718-820}$ structure. The eight domains share statistically significant sequence alignments to Pom152$^{718-820}$ (E-values ranging from 3.3 x 10^{-49} to 6.8 x 10^{-39}). (C) Sequence conservation among the nine Ig-like domains was mapped on the structure of Pom152$^{718-820}$ using ConSurf (36). Low conservation, white; high conservation, orange. (D) Domain organization of human Nup210. See legend of Fig. 1A. (E) Superposition of comparative models for yeast Pom152$^{718-820}$ (orange) and human Nup210$^{1,079-1,152}$ (blue). See also Fig. S3.

Integrative Structure Determination of Pom152^{FL}

We determined the structure of Pom152^{FL} through an integrative modeling approach (**Fig. 7.5**) that has proven useful for structural analysis of flexible and thus conformationally heterogeneous proteins, such as Pom152 (9, 37). We represented Pom152^{FL} in a coarse-grained fashion to reflect the relatively low resolution of structural information about it, as follows. Each of the nine Ig-like domains was represented as a rigid body at the resolution of 1 residue per bead, computed from the atomic NMR structure (**Fig. 7.3C**) or a comparative model (**Fig. 7.4B** and **7.4C**). In addition to these nine domains connected by flexible linkers (from 4 to 10 residues per bead), a Pom152^{FL} model also included a flexible string of beads (from 20 to 100 residues per bead) corresponding to the N-terminal NPC-associating domain (residues 1–100), the transmembrane domain (residues 101–200), and the linker domain (residues 201–374). Next, 100,000 Pom152^{FL} models were computed by flexibly fitting random initial models into the negative-stain EM density map while avoiding steric clashes and retaining sequence connectivity.

The 500 best-scoring models (i.e., the ensemble) fit the EM map as well as satisfy the excluded volume and sequence connectivity restraints used to compute the models. The structures also fit the EM class averages, with cross-correlation coefficients ranging from 0.84 to 0.87 for representative class averages (Fig. S4C).

In addition to satisfying information used to compute them, the models are also similar to each other. The clustering of the best-scoring models identified a single dominant cluster of 364 similar structures, with a precision (Supplemental Experimental Procedures) of 7.0 Å for the luminal domain (Fig. S4A).

In general, an ensemble of good-scoring models can be visualized as a localization probability density map. The map gives the probability of any volume element being occupied by a certain bead in superposed good-scoring models. **Fig. 7.6A** shows the localization density for each of the nine Ig-like domains, as sampled by the 364 good-scoring models in the dominant cluster. The 7.0 Å precision of this cluster is sufficiently high to pinpoint the locations, but not the orientations, of the constituent Ig-like domains (Fig. S4B). The integrative structure further supports the similarity between the molecular architectures of Pom152^{LD} and cadherins (**Fig. 7.6A**) (19).

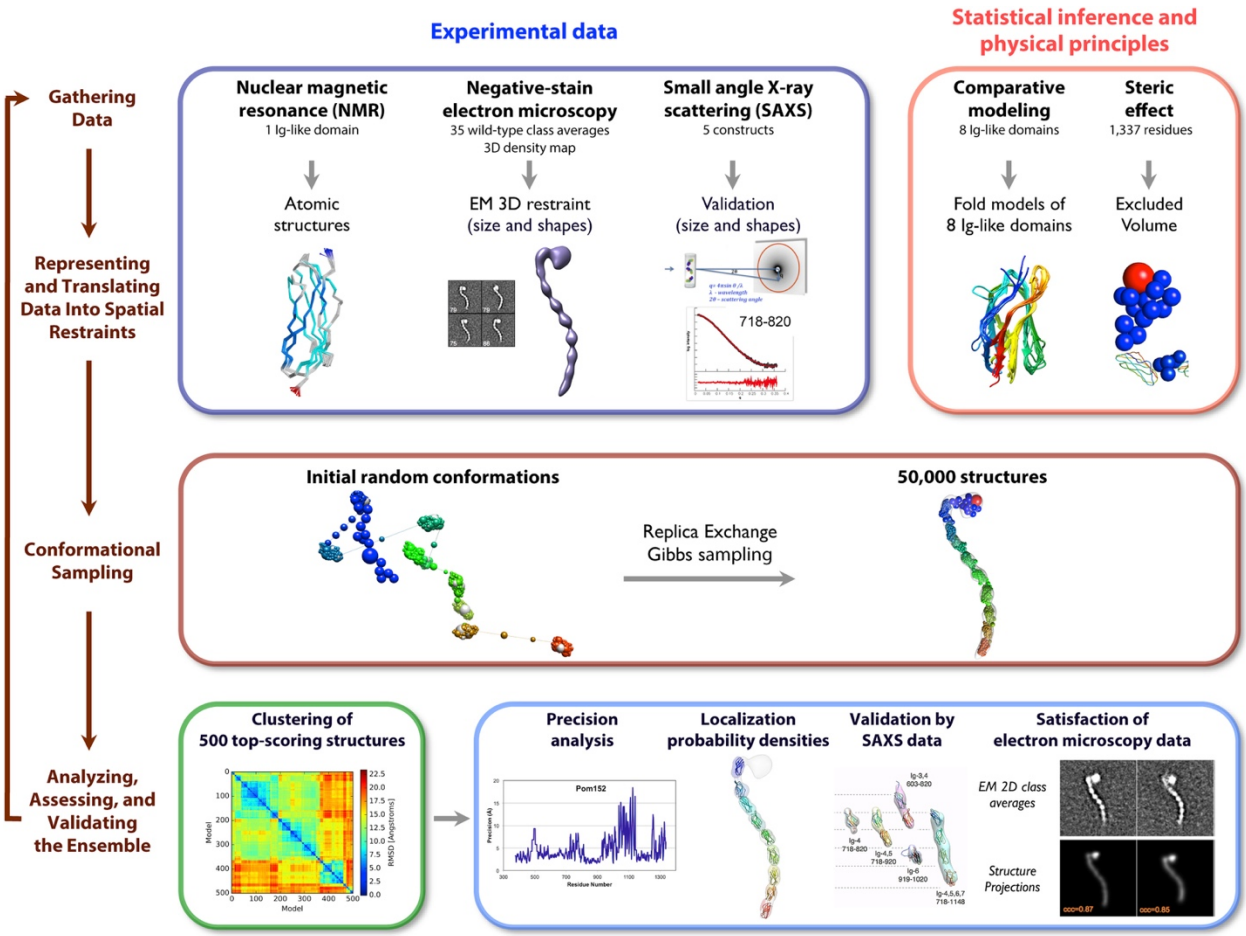

Figure 7.5 | Four-Stage Scheme for Integrative Structure Determination of Pom152FL

The integrative structure determination of Pom152FL proceeds through four stages: (1) gathering data, (2) representing and translating data into spatial restraints, (3) conformational sampling to produce an ensemble of structures that satisfies the restraints, and (4) analyzing, assessing, and validating the ensemble structures. The modeling protocol (i.e., stages 2, 3, and 4) was scripted using the Python Modeling Interface (PMI), version 4d97507, a library for modeling macromolecular complexes based on our open-source Integrative Modeling Platform (IMP) package, version 2.6 (http://integrativemodeling.org) (38).

Validation of the Pom152^{LD} Structure Using SAXS Data

We validated our integrative structure of Pom152^{LD} by SAXS data. Specifically, the computed SAXS profiles match measured SAXS profiles for all five experimentally characterized segments, spanning residues 718–820 (Ig-4), 718–920 (Ig-4,5), 603–820 (Ig-3,4), 919–1020 (Ig-6), and 718–1148 (Ig-4,5,6,7) (Fig. S5 and Table S1). In addition, the shapes of these segments in our mode also match the envelopes (*ab initio* shapes) computed from the corresponding SAXS profiles (**Fig. 7.6** and S5). The linearity of the Guinier plots confirms a high degree of homogeneity for each of the five Pom152 SAXS samples. Each radius of gyration (R_g) and maximum particle size (*Dmax*) are consistent with those of the corresponding Pom152 segments (Table S1). Notably, all five SAXS profiles show well-defined bell-shaped curves in Kratky plots (Fig. S5, middle), indicating relatively rigid conformations of the individual Ig-like domains. In addition, plateaus at the high q region (0.2–0.3 Å^{-1}) in the Kratky plots indicate some flexibility between Ig-like domains in solution, consistent with the heterogeneity inferred from the negative-stain EM class averages (**Fig. 7.1B**).

Position of Pom152^{LD} within the NPC

Pom152 is the main component of the membrane ring of the NPC (1, 15, 17). To determine whether or not the shape and length of our integrative structure of Pom152 account for the formation of the membrane ring, we fitted the luminal domains of 16 copies of Pom152 into our previously published yeast NPC map (1) (Fig. 6B and S7). A good fit positions two copies of the extended Pom152^{LD} molecule in an antiparallel fashion on top of each other, forming a homodimer; it is ambiguous whether the tail-to-head direction is clockwise or counter-clockwise. The length of the elongated Pom152^{LD} (~38.7 nm) is

sufficient to span the spoke width. The arrangement suggests that the membrane ring is made possible by tight interactions between the Ig-like domains in the two antiparallel luminal domains, in agreement with the observed homo-dimerization of Pom152 *in vivo* (1, 15).

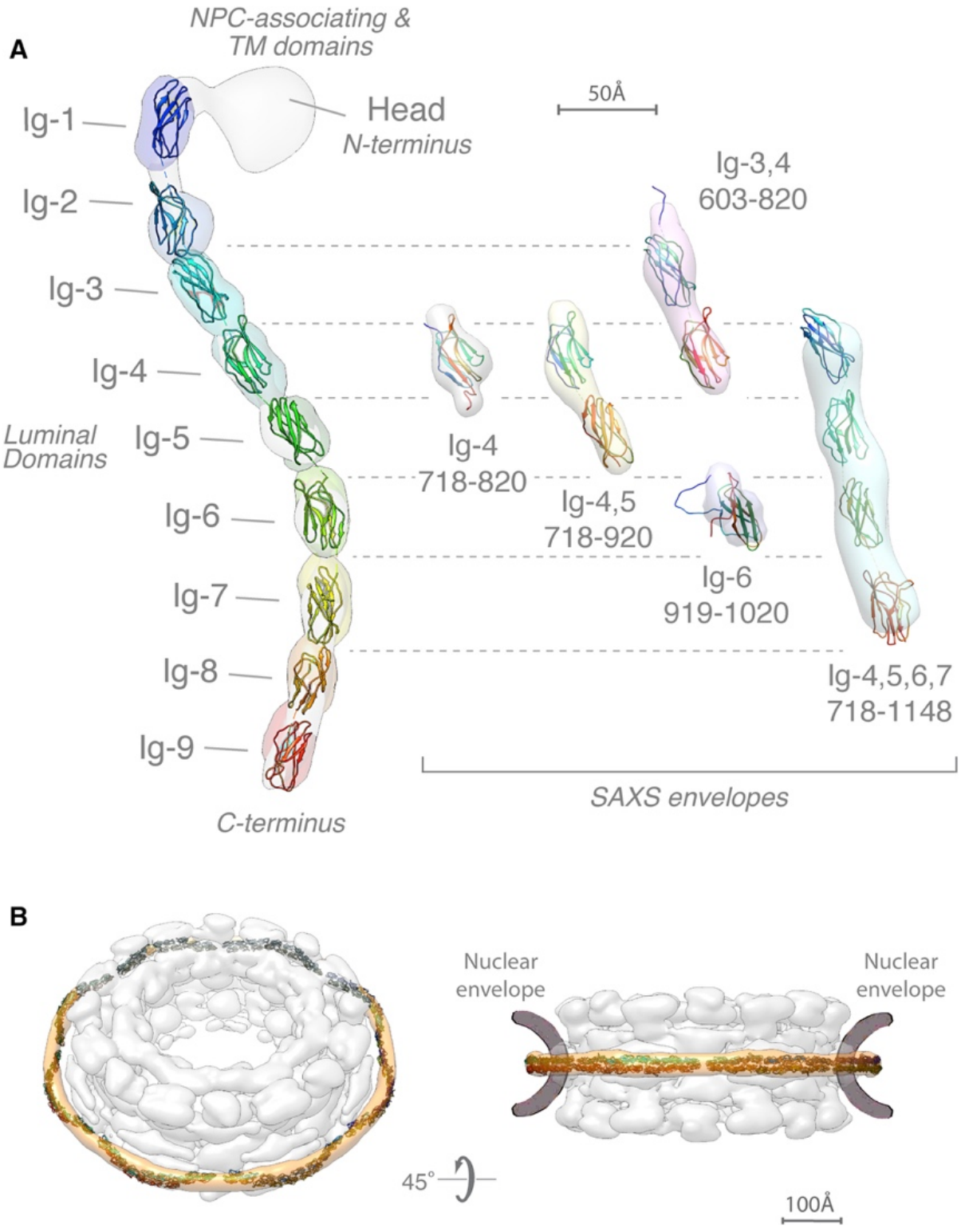
A
NPC-associating &
TM domains
Head
N-terminus
50Å
Ig-1
Ig-2
Ig-3
Ig-4
Luminal
Domains
Ig-5
Ig-6
Ig-7
Ig-8
Ig-9
C-terminus
Ig-4
718-820
Ig-4,5
718-920
Ig-3,4
603-820
Ig-6
919-1020
Ig-4,5,6,7
718-1148
SAXS envelopes
B
Nuclear
envelope
Nuclear
envelope
45°
100Å

Figure 7.6 | Integrative Structure of Pom152FL Based on NMR Spectroscopy, Negative-Stain EM, and SAXS

(**A**) (*Left*) The localization probability density map computed from 364 superposed structures that satisfy the input spatial restraints shows the location of each of the nine Ig-like domains (ranging from blue to red). The negative-stain EM density map is superposed in gray. A representative molecular model of $Pom152^{LD}$ (ribbon plot) was obtained by adjusting the relative orientations of adjacent Ig-like domains to resemble those in known cadherin structures (PDB: 1L3W, 1EPF, 1NCI, 4ZI9, 5K8R) (32, 39–42). (*Right*) Validation of the integrative structure of $Pom152^{LD}$ by SAXS data for five Pom152 segments spanning residues 718-820 (Ig-4), 718-920 (Ig-4,5), 603-820 (Ig-3,4), 919-1020 (Ig-6), and 718-1148 (Ig-4,5,6,7). The shapes of these segments in our integrative structure match the envelopes (*ab initio* shapes) computed from the corresponding SAXS profiles. (**B**) Fit of 16 copies of $Pom152^{LD}$ into the yeast NPC map (1). A good fit positions two copies of the extended $Pom152^{LD}$ molecular in an anti-parallel fashion on top of each other, forming a homodimer; we only show the potential arrangement of the anti-parallel homodimer, which is implied by the C2 symmetry of the NPC (1, 43, 44). See also Fig. 5, S4-S7, and Table S1.

Discussion

Since the first high-resolution structure of a nucleoporin domain was solved in 2002 (45), the combined efforts of many groups have provided a detailed structural picture for most of the major components of the NPC (46). However, conspicuously absent was any such information for the NPC membrane ring. Our structure of $Pom152^{FL}$ is thus the first detailed view of the molecular architecture of the NPC's transmembrane components. The structure confirmed our previous suggestion (19) that the domain arrangement in Pom152, with repeated Ig-like β-sandwich folds forming an extended luminal module, most strongly resembles the organization of another conserved eukaryotic family of proteins, classical (type I) cadherins (47). The similarities between these two types of proteins extend beyond each repetitive unit, including: (1) they share an overall arch shape (**Fig. 7.1** and (48)); (2) they are both able to dimerize (15, 16, 48); and (3) they share a similar domain organization, with a segment of repetitive Ig-like domains connected through a single-pass transmembrane region to a short protein-protein interaction domain (although in cadherins the orientation of the domains is equivalent to that in Nup210; see **Fig. 7.1** and **7.4** and(48)). In the case of cadherins, their cytoplasmic short region is extended and largely unstructured (49) and has been shown to interact with the arm-repeat α-solenoid proteins p120 catenin (50) and β-catenin (49). The NPC-associating domains of Pom152 and Nup210 are of similar size (~150 amino acid residues) and both are predicted to be largely disordered, although the orientation of this domain in Nup210 is at the C-terminus while that of Pom152 is at the N-terminus (13, 17). Variability in the position and type of the membrane domain between potential Nup homologs has previously been seen in the Trypanosome NPC (51); interestingly, Pom152

localizes to the NPC when expressed in mammalian cells (17), suggesting conservation in assembly mechanism and even function between vertebrates and fungi.

The N-terminal Pom152 NPC-associating domain is positioned close to the NPC inner ring (1). Indeed, genetic interactions with inner-ring components (Nic96, Nup59, Nup170, and Nup188) that require the presence of the Pom152 N terminus have been described (15); moreover, Pom152 appears to physically interact with several inner-ring proteins (1). It is thus reasonable to suggest that, similar to the cadherin-catenin interactions, the disordered N-terminus of Pom152 connects to the α-solenoids of the NPC inner-ring components (1, 43, 44), just as cadherins use a largely unstructured linker to interact with the α-solenoid catenin proteins.

Despite their common organization, some clear differences are also observed between Pom152 and cadherins. First, Pom152 lacks the conserved tryptophans that mediate the ‘‘strand-swap’’ mechanism of dimerization in cadherins (48). Not a single Trp residue is present at the C-terminus of Pom152 (Fig. S6), and our fitting into the whole NPC suggests that Pom152 molecules likely dimerize through more extensive contacts (**Fig. 7.6B** and S7). The distribution of electrostatic potential and sequence conservation on the structure of Pom152$^{718\text{-}820}$ (**Fig. 7.3E–7.3H** and S6) indicates a mechanism where facing Ig-like domains from opposite strands generate complementary interaction surfaces. Second, unlike cadherins, Pom152 does not seem to be able to bind calcium. Cadherin Ig-like domains coordinate calcium through highly conserved residues, stiffening the connections between successive domains and imparting a strong curvature to the full-length ectodomain (39, 52). Removal of calcium leads to a disordering of inter-domain orientations that can be observed by negative-stain EM (53). However, we were

not able to detect calcium binding or any obvious calcium-dependent Pom152 shape changes (data not shown). Perhaps this observation is unsurprising in proteins that diverged more than a billion years ago.

Pom152 and Nup210 homologs exist beyond the opisthokonts (Fungi and Metazoa). Clear homologs are found in Amoebozoa and in plants, although all these are more similar to Nup210, the closer homologs of Pom152 being restricted to the Fungi. Although clearly structurally and functionally similar, the precise nature of the evolutionary relationship between the Pom152-like and Nup210-like homologs therefore remains somewhat unclear. Even so, cadherin-like proteins predate the eukaryota, being found in bacteria where they may mediate cell-cell contact (54), consistent with an ancient evolutionary origin for Pom152 and Nup210, and the entire NPC (1, 6, 19, 55).

Experimental Procedures

Affinity Purification of Endogenous Pom152 and Truncation Mutants

Native Pom152 and the truncation mutants Pom152$^{1\text{-}1,135}$ and Pom152$^{1\text{-}936}$ were affinity purified, natively eluted, and further purified in 5%–20% sucrose gradients as previously described (7).

Yeast Strains

Yeast strains (Table S2) were constructed in a W303 (MATa/alpha ade2-1 ura3-1 his3-11, 15 trp 1-1 leu2-3,112 can1-100) background using standard techniques (25). Unless otherwise stated, strains were grown at 30ºC in YPD medium (1% yeast extract, 2% bactopeptone, and 2% glucose).

Electron Microscopy Analyses and 3D Reconstruction of Pom152

Purified Pom152^{FL} and the truncated versions Pom152$^{1\text{-}1,135}$ and Pom152$^{1\text{-}936}$ were applied to glow-discharged carbon-coated copper grids and stained with 1% uranyl formate. Images were collected on a JEOL JEM-2100F transmission electron microscope (JEOL USA) or a Philips CM200 transmission electron microscope (FEI) and analyzed using ISAC (26). The 3D density map of Pom152^{FL} was generated through the random conical tilt reconstruction method and Relion (28). Angles were measured with the ImageJ angle tool. Further details are provided in the Supplemental Experimental Procedures.

Phenotypic Assays

To analyze the growth phenotype, 10-fold dilutions of yeast cultures were spotted on YEPD plates in the presence or absence of 0.3% or 0.4% benzyl-alcohol and incubated at the indicated temperatures (25). Indicated fluorescently tagged proteins were visualized, in the presence of absence of 0.1% benzyl-alcohol, using a 63x 1.4 NA Plan-Apochromat objective using a microscope (Axioplan 2, Zeiss) equipped with a cooled charge-coupled device camera (ORCA-ER, Hamamatsu) (25).

Expression and Purification of Pom152 Constructs

Pom152 constructs used for SAXS studies were expressed as SeMET labeled proteins and purified following a standard procedure (9, 56). Pom152 constructs used for the NMR studies were expressed in minimal medium following a standard protocol (57) with minor modifications. See Supplemental Experimental Procedures for details.

NMR Resonances Assignments and Structure Calculation of Pom152$^{718–820}$

The [U-^{13}C,^{15}N] Pom152$^{718\text{-}820}$ samples were used for backbone and side-chain resonance assignments using multidimensional NMR experiments (58). All NMR data

were acquired at 25ºC using either Varian 600 MHz or Bruker 600 and 900 MHz spectrometers equipped with cryogenic probes capable of applying pulse-field gradients along the z axis. Structure calculations were carried out using distance, dihedral, and hydrogen-bond restraints using the ARIA/CNS program (59). A total of 2,095 restraints were used to solve the structure of Pom152$^{718\text{-}820}$ (**Fig. 7.3A**). Twenty best-scoring structures with no distance restraint violations larger than 0.5 Å and no dihedral restraint violations larger than 5º were chosen to represent the structural ensemble consistent with the NMR data. Further details are provided in the Supplemental Experimental Procedures.

Small-Angle X-Ray Scattering Experiments

SAXS measurements for five Pom152 segments were carried out at the SSRL Beamline 4-2 in the SLAC National Accelerator Laboratory (Menlo Park, CA) (**Fig. 7.6A** and S5; Table S1). Further details of SAXS analysis are provided in Supplemental Experimental Procedures and our previous publications (9, 56).

Comparative Modeling of Luminal Ig-like Domains in Pom152 and Human Nup210

Comparative models of eight luminal Ig-like domains of Pom152 (**Fig. 7.4B**) and a single luminal Ig-like domain of human Nup210$^{1,079\text{-}1,152}$ (**Fig. 7.4E**) were computed with MODELLER (34), using the NMR structure of Pom152$^{718\text{-}820}$ as the template. Further details are provided in Supplemental Experimental Procedures.

Integrative Structure Determination of Pom152^{FL}

The integrative structure determination of Pom152^{FL} proceeded through four stages (**Fig. 7.5**) (1, 2, 25, 60, 61): (1) gathering data, (2) representing and translating data into spatial restraints, (3) conformational sampling to produce an ensemble of structures that satisfies

the restraints, and (4) analyzing, assessing, and validating the ensemble structures. The modeling protocol (i.e., stages 2, 3, and 4) was scripted using the Python Modeling Interface (PMI), version 4d97507, a library for modeling macromolecular complexes based on our open-source Integrative Modeling Platform (IMP) package, version 2.6 (http://integrativemodeling.org) (38). Further details are provided in Supplemental Experimental Procedures and our previous publications (9, 37, 62).

Accession Numbers

The coordinates and spatial restraints for NMR structure determination have been deposited in the PDB (PDB: 5TVZ). NMR resonance assignments have been deposited in the Biological Magnetic Resonance Bank (BMRB: 30201). 3D negative-stain EM reconstruction of Pom152 has been deposited in the Electron Microscopy Data Bank (EMD: EMD-8543). The accession numbers for the SAXS profiles for five Pom152 segments have been deposited in the Small Angle Scattering Biological Data Bank (SASDB: SASDBV9, SASDBW9, SASDBX9, SASDBY9, SASDBZ9). Files containing the input data, scripts, and output structures are available online (https://salilab.org/pom152; https:// github.com/salilab/pom152).

Author Contributions

Conceptualization, P.U., S.J.K., P.S., A.S., J.F.-M., and M.P.R.; Investigation, P.U., S.J.K., P.S., K.D., I.E.C., S.M.C., R.W., and J.F-M.; Formal Analysis, P.U., S.J.K., P.S., K.D., I.E.C., S.M.C., J.B.B., W.J.R., D.L.S., D.C., A.S., M.P.R., and J.F.M.; Writing, J.F.-M., S.J.K., I.E.C., A.S., P.S., K.D., D.C., and M.P.R.; Funding Acquisition, D.L.S., D.C., S.C.A., A.S., and M.P.R.; Supervision, J.F.M., D.L.S., D.C., S.C.A., A.S., and M.P.R.

Acknowledgments

We thank NYULMC OCS Microscopy Core for assistance with negative-stain EM, NYSGRC for providing samples for SAXS and NMR analyses; and T. Matsui and T.M. Weiss at SSRL, SLAC National Accelerator Laboratory, for assistance with collecting SAXS data. The instrumentation in the Einstein NMR Resource is supported by the Albert Einstein College of Medicine, and the Bruker 600 NMR instrument was purchased using funds from NIH award 1S10OD016305. Some of the NMR experiments were performed at the New York Structural Biology Center, which is a STAR center supported by the New York Office of Science, Technology and Academic Research. The New York Structural Biology Center is supported by grant number 349247 from the Simons Foundation. Support was provided by NSF GRF (no. 1650113) (I.E.C.); and NIH grants U54 RR022220 to B.C., A.S., and M.R.; R01 GM112108 to M.R.; U54GM094662 to S.C.A; GM 117212 to D.C.; P41 GM109824 to M.R., A.S., and B.C.; U01GM098256 to M.P.R.; P41 GM103314 to B.C.; and R01 GM083960 to A.S.

References

1. F. Alber, *et al.*, The molecular architecture of the nuclear pore complex. *Nature* **450**, 695–701 (2007).
2. F. Alber, *et al.*, Determining the architectures of macromolecular assemblies. *Nature* **450**, 683–694 (2007).
3. M. P. Rout, *et al.*, The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J Cell Biol.* **148**, 635–51 (2000).
4. A. Ibarra, M. W. Hetzer, Nuclear pore proteins and the control of genome functions. *Genes Dev* **29**, 337–49 (2015).
5. D. N. Simon, M. P. Rout, Cancer and the nuclear pore complex. *Adv Exp Med Biol* **773**, 285–307 (2014).
6. D. Devos, *et al.*, Components of coated vesicles and nuclear pore complexes share a common molecular architecture. *PLoS Biol* **2**, e380 (2004).
7. P. Sampathkumar, *et al.*, Structure, dynamics, evolution, and function of a major scaffold component in the nuclear pore complex. *Structure* **21**, 560–71 (2013).
8. G. Drin, *et al.*, A general amphipathic alpha-helical motif for sensing membrane curvature. *Nat Struct Mol Biol* **14**, 138–46 (2007).
9. S. J. Kim, *et al.*, Integrative structure-function mapping of the nucleoporin Nup133 suggests a conserved mechanism for membrane anchoring of the nuclear pore complex. *Mol Cell Proteomics* (2014).
10. A. von Appen, *et al.*, In situ structural analysis of the human nuclear pore complex. *Nature* **526**, 140–3 (2015).

11. H. J. Chial, M. P. Rout, T. H. Giddings, M. Winey, Saccharomyces cerevisiae Ndc1p is a shared component of nuclear pore complexes and spindle pole bodies. *J Cell Biol* **143**, 1789–800 (1998).
12. M. Miao, K. J. Ryan, S. R. Wente, The integral membrane protein Pom34p functionally links nucleoporin subcomplexes. *Genetics* **172**, 1441–57 (2006).
13. R. W. Wozniak, E. Bartnik, G. Blobel, Primary structure analysis of an integral membrane glycoprotein of the nuclear pore. *J Cell Biol* **108**, 2083–92 (1989).
14. H. L. Liu, C. P. De Souza, A. H. Osmani, S. A. Osmani, The three fungal transmembrane nuclear pore complex proteins of Aspergillus nidulans are dispensable in the presence of an intact An-Nup84-120 complex. *Mol Biol Cell* **20**, 616–30 (2009).
15. S. E. Tcheperegine, M. Marelli, R. W. Wozniak, Topology and functional domains of the yeast pore membrane protein Pom152p. *J Biol Chem* **274**, 5252–8 (1999).
16. W. T. Yewdell, P. Colombi, T. Makhnevych, C. P. Lusk, Lumenal interactions in nuclear pore complex assembly and stability. *Mol Biol Cell* **22**, 1375–88 (2011).
17. R. W. Wozniak, G. Blobel, M. P. Rout, POM152 is an integral protein of the pore membrane domain of the yeast nuclear envelope. *J Cell Biol* **125**, 31–42 (1994).
18. U. F. Greber, A. Senior, L. Gerace, A major glycoprotein of the nuclear pore complex is a membrane-spanning polypeptide with a large lumenal domain and a small cytoplasmic tail. *EMBO J* **9**, 1495–502 (1990).
19. D. Devos, *et al.*, Simple fold composition and modular architecture of the nuclear pore complex. *Proc Natl Acad Sci U S A* **103**, 2172–7 (2006).

20. M. Marelli, C. P. Lusk, H. Chan, J. D. Aitchison, R. W. Wozniak, A link between the synthesis of nucleoporins and the biogenesis of the nuclear envelope. *J Cell Biol* **153**, 709–24 (2001).
21. M. A. D'Angelo, J. S. Gomez-Cavazos, A. Mei, D. H. Lackner, M. W. Hetzer, A change in nuclear pore complex composition regulates cell differentiation. *Dev Cell* **22**, 446–58 (2012).
22. J. S. Gomez-Cavazos, M. W. Hetzer, The nucleoporin gp210/Nup210 controls muscle differentiation by regulating nuclear envelope/ER homeostasis. *J Cell Biol* **208**, 671–81 (2015).
23. M. A. Chapman, *et al.*, Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467–72 (2011).
24. T. Rajkumar, *et al.*, Identification and validation of genes involved in cervical tumourigenesis. *BMC Cancer* **11**, 80 (2011).
25. J. Fernandez-Martinez, *et al.*, Structure-function mapping of a heptameric module in the nuclear pore complex. *J Cell Biol* **196**, 419–34 (2012).
26. Z. Yang, J. Fang, J. Chittuluru, F. J. Asturias, P. A. Penczek, Iterative stable alignment and clustering of 2D transmission electron microscope images. *Structure* **20**, 237–47 (2012).
27. M. Radermacher, Three-dimensional reconstruction of single particles from random and nonrandom tilt series. *J Electron Microsc Tech* **9**, 359–94 (1988).
28. S. H. Scheres, RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* **180**, 519–30 (2012).

29. J. J. Scarcelli, C. A. Hodge, C. N. Cole, The yeast integral membrane protein Apq12 potentially links membrane dynamics to assembly of nuclear pore complexes. *J Cell Biol* **178**, 799–812 (2007).

30. E. Onischenko, L. H. Stanton, A. S. Madrid, T. Kieselbach, K. Weis, Role of the Ndc1 interaction network in yeast nuclear pore complex assembly and maintenance. *J Cell Biol* **185**, 475–91 (2009).

31. N. Meszaros, *et al.*, Nuclear pore basket proteins are tethered to the nuclear envelope and can regulate membrane curvature. *Dev Cell* **33**, 285–98 (2015).

32. C. Kasper, *et al.*, Structural basis of cell-cell adhesion by NCAM. *Nat Struct Biol* **7**, 389–93 (2000).

33. D. M. Halaby, A. Poupon, J. Mornon, The immunoglobulin fold family: sequence analysis and 3D structure comparisons. *Protein Eng* **12**, 563–71 (1999).

34. A. Sali, T. L. Blundell, Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779–815 (1993).

35. A. Marchler-Bauer, *et al.*, CDD: NCBI's conserved domain database. *Nucleic Acids Res* **43**, D222-6 (2015).

36. H. Ashkenazy, *et al.*, ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res* **44**, W344-50 (2016).

37. J. Fernandez-Martinez, *et al.*, Structure and Function of the Nuclear Pore Complex Cytoplasmic mRNA Export Platform. *Cell* **167**, 1215–1228 (2016).

38. D. Russel, *et al.*, Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. *PLOS Biology* **10**, e1001244 (2012).
39. T. J. Boggon, *et al.*, C-cadherin ectodomain structure and implications for cell adhesion mechanisms. *Science* **296**, 1308–13 (2002).
40. J. M. Nicoludis, *et al.*, Structure and Sequence Analyses of Clustered Protocadherins Reveal Antiparallel Interactions that Mediate Homophilic Specificity. *Structure* **23**, 2087–98 (2015).
41. J. M. Nicoludis, *et al.*, Antiparallel protocadherin homodimers use distinct affinity- and specificity-mediating regions in cadherin repeats 1-4. *Elife* **5** (2016).
42. L. Shapiro, *et al.*, Structural basis of cell-cell adhesion by cadherins. *Nature* **374**, 327–37 (1995).
43. J. Kosinski, *et al.*, Molecular architecture of the inner ring scaffold of the human nuclear pore complex. *Science* **352**, 363–5 (2016).
44. D. H. Lin, *et al.*, Architecture of the symmetric core of the nuclear pore. *Science* **352**, aaf1015 (2016).
45. A. E. Hodel, *et al.*, The three-dimensional structure of the autoproteolytic, nuclear pore-targeting domain of the human nucleoporin Nup98. *Mol Cell* **10**, 347–58 (2002).
46. T. U. Schwartz, The Structure Inventory of the Nuclear Pore Complex. *J Mol Biol* **428**, 1986–2000 (2016).
47. N. Ishiyama, M. Ikura, The three-dimensional structure of the cadherin-catenin complex. *Subcell Biochem* **60**, 39–62 (2012).

48. L. Shapiro, W. I. Weis, Structure and biochemistry of cadherins and catenins. *Cold Spring Harb Perspect Biol* **1**, a003053 (2009).

49. A. H. Huber, D. B. Stewart, D. V. Laurents, W. J. Nelson, W. I. Weis, The cadherin cytoplasmic domain is unstructured in the absence of beta-catenin. A possible mechanism for regulating cadherin turnover. *J Biol Chem* **276**, 12301–9 (2001).

50. N. Ishiyama, *et al.*, Dynamic and static interactions between p120 catenin and E-cadherin regulate the stability of cell-cell adhesion. *Cell* **141**, 117–28 (2010).

51. S. O. Obado, *et al.*, Interactome Mapping Reveals the Evolutionary History of the Nuclear Pore Complex. *Plos Biology* **14** (2016).

52. B. Nagar, M. Overduin, M. Ikura, J. M. Rini, Structural basis of calcium-induced E-cadherin rigidification and dimerization. *Nature* **380**, 360–4 (1996).

53. S. Pokutta, K. Herrenknecht, R. Kemler, J. Engel, Conformational changes of the recombinant extracellular domain of E-cadherin upon calcium binding. *Eur J Biochem* **223**, 1019–26 (1994).

54. M. Fraiberg, I. Borovok, R. M. Weiner, R. Lamed, Discovery and characterization of cadherin domains in Saccharophagus degradans 2-40. *J Bacteriol* **192**, 1066–74 (2010).

55. M. C. Field, J. B. Dacks, First and last ancestors: reconstructing evolution of the endomembrane system with ESCRTs, vesicle coat proteins, and nuclear pore complexes. *Curr Opin Cell Biol* **21**, 4–13 (2009).

56. P. Sampathkumar, *et al.*, Atomic structure of the nuclear pore complex targeting domain of a Nup116 homologue from the yeast, Candida glabrata. *Proteins* **80**, 2110–6 (2012).

57. D. J. Weber, *et al.*, NMR docking of a substrate into the X-ray structure of staphylococcal nuclease. *Proteins* **13**, 275–87 (1992).

58. M. Sattler, J. Schleucher, C. Griesinger, Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Progress in Nuclear Magnetic Resonance Spectroscopy* **34**, 93–158 (1999).

59. J. P. Linge, M. Habeck, W. Rieping, M. Nilges, ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* **19**, 315–6 (2003).

60. K. Lasker, *et al.*, Integrative structure modeling of macromolecular assemblies from proteomics data. *Mol Cell Proteomics* **9**, 1689–702 (2010).

61. K. Lasker, *et al.*, Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. *Proc Natl Acad Sci USA* **109**, 1380–1387 (2012).

62. Y. Shi, *et al.*, Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. *Mol Cell Proteomics* **13**, 2927–2943 (2014).

Chapter VIII - Integrative structure and functional anatomy of a nuclear pore complex

Contributing authors

Seung Joong Kim[1,*], Javier Fernandez-Martinez[2,*], Ilona Nudelman[2,*], Yi Shi[3,†,*], Wenzhu Zhang[3,*], Barak Raveh[1], Thurston Herricks[4], Brian D. Slaughter[5], Joanna A. Hogan[6], Paula Upla[7], Ilan E. Chemmama[1], Riccardo Pellarin[1,†], Ignacia Echeverria[1], Manjunatha Shivaraju[5], Azraa S. Chaudhury[2], Junjie Wang[3], Rosemary Williams[2], Jay R. Unruh[5], Charles H. Greenberg[1], Erica Y. Jacobs[3], Zhiheng Yu[8], M. Jason de la Cruz[8,†], Roxana Mironska[2], David L. Stokes[7], John D. Aitchison[4,9], Martin F. Jarrold[6], Jennifer L. Gerton[5], Steven J. Ludtke[10], Christopher W. Akey[11], Brian T. Chait[3], Andrej Sali[1], Michael P. Rout[2]

[1]Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, University of California, San Francisco, San Francisco, California 94158, USA.

[2]Laboratory of Cellular and Structural Biology, The Rockefeller University, New York, New York 10065, USA.

[3]Laboratory of Mass Spectrometry and Gaseous Ion Chemistry, The Rockefeller University, New York, New York 10065, USA.

[4]Institute for Systems Biology, 401 Terry Ave. N., Seattle, Washington 98109, USA.

[5]Stowers Institute for Medical Research, Kansas City, Missouri 64110, USA.

[6]Department of Chemistry, Indiana University, Bloomington, Indiana 47405, USA.

[7]Skirball Institute and Department of Cell Biology, New York University School of Medicine, New York, New York 10016, USA.

[8]Janelia Research Campus, Howard Hughes Medical Institute, 19700 Helix Drive, Ashburn, Virginia 20148, USA.

[9]Center for Infectious Disease Research, Seattle, Washington 98109, USA.

[10]Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, 1 Baylor Plaza, Houston, Texas 77030, USA.

[11]Department of Physiology and Biophysics, Boston University School of Medicine, 700 Albany Street, Boston, Massachusetts 02118, USA.

[†]Present addresses: Structural Bioinformatics Unit, Institut Pasteur, CNRS UMR 3528, Paris, France (R.P.); Department of Cell Biology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15260, USA (Y.S.); Structural Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA (M.J.d.l.C.).

[*]These authors contributed equally to this work.

Contacts: rout@rockefeller.edu (M.P.R), chait@rockefeller.edu (B.T.C), sali@salilab.org (A.S), cakey@bu.edu (C.W.A.), and sludtke@bcm.edu (S.J.L.)

Abstract

Nuclear pore complexes play central roles as gatekeepers of RNA and protein transport between the cytoplasm and nucleoplasm. However, their large size and dynamic nature have impeded a full structural and functional elucidation. Here we determined the structure of the entire 552-protein nuclear pore complex of the yeast *Saccharomyces*

cerevisiae at sub-nanometre precision by satisfying a wide range of data relating to the molecular arrangement of its constituents. The nuclear pore complex incorporates sturdy diagonal columns and connector cables attached to these columns, imbuing the structure with strength and flexibility. These cables also tie together all other elements of the nuclear pore complex, including membrane-interacting regions, outer rings and RNA-processing platforms. Inwardly directed anchors create a high density of transport factor-docking Phe-Gly repeats in the central channel, organized into distinct functional units. This integrative structure enables us to rationalize the architecture, transport mechanism and evolutionary origins of the nuclear pore complex.

Introduction

Nuclear pore complexes (NPCs) are large proteinaceous assemblies studded through the nuclear envelope, the double-membraned barrier that surrounds the nucleus; NPCs are the sole mediators of macromolecular transport between the nucleus and the cytoplasm, and carry key regulatory platforms for numerous nuclear processes (1). NPCs are also major targets for viral manipulation and defects in this transport machine are directly linked to human diseases, including cancers (2). Each NPC is an eight-fold symmetric, cylindrical assembly consisting of approximately 550 copies of about 30 different proteins of the nucleoporin family (Nups). These Nups assemble into sub-complexes that form higher-order structures called spokes. Eight spokes assemble into even larger modules: coaxial outer and inner rings form a symmetric core scaffold, which is connected to a membrane ring, a nuclear basket and cytoplasmic RNA export complexes (3). The scaffold surrounds a central channel that is formed in part by multiple intrinsically disordered Phe-Gly (FG) repeat motifs that extend from nucleoporins termed FG Nups. These FG motifs mediate selective nucleocytoplasmic transport through specific interactions with nuclear transport factors (NTFs), which carry their cognate macromolecular cargoes (4). It has also previously been suggested that the central channel contains a feature called the central transporter (5). Although partial structures have previously been described (3, 6, 7), a complete, high-resolution structure for the entire NPC in any organism has hitherto been lacking, leaving open key questions as to how the NPC is organized and functions, and how it evolved. To address these questions, we have determined an integrative structure of the yeast NPC at sub-nanometre precision.

Results

Solving the structure of the S. cerevisiae NPC

We developed a method to rapidly and gently isolate native yeast NPCs, enabling us to determine the type and amount of each Nup in the NPC, the proximities between Nups resolved to the amino acid residue level, and the mass and detailed morphology of the entire NPC. These data were then used to solve the structure using an integrative modelling approach (8, 9) (Extended Data Fig. 1, Supplementary Results and Discussion, and Methods).

We determined the mass of the entire NPC and a definitive stoichiometry for every Nup and associated molecules using mass spectrometric and in vivo imaging methods. The native NPC has a mass of 52 MDa, or about 87 MDa when including the membrane, cargo and NTFs (**Fig. 8.1** and Extended Data Figs 2, 3c). To inform the proximities, orientations and conformations of the Nups, isolated NPCs were subjected to cross-linking with mass spectrometric readout (9, 10). This approach identified 3,077 unique cross-linked pairs of residues, and provided the distance restraints between them, both within and between Nups (**Fig. 8.2**, Supplementary Table 1 and Methods). The morphology of the NPC was determined using cryo-electron tomography (cryo-ET) and sub-tomogram averaging (11) (Methods). This approach provided a final 3D map at approximately 28Å resolution, with a local resolution of 20–25Å for the inner ring, which has approximate C2 symmetry (**Fig. 3** and Extended Data Figs 4–6). The NPCs retained a considerable amount of nuclear envelope membrane, which forms a continuous belt around the midline of the structure (**Fig. 8.3a**, **b**, **d**, **e**). We found that a membrane protein ring interconnects adjacent spokes within the nuclear envelope lumen (**Fig. 8.3a**, **e**), a

feature largely absent from recent electron microscopy maps. A cylindrically averaged, bi-lobed density fills the central channel (the 'central transporter', **Fig. 8.3a**, **b**). Individual Nups and their domains, as well as the subcomplexes of the NPC, were represented on the basis of published crystallographic structures, integrative structures and comparative models (9, 10, 12) (Supplementary Table 2 and Methods), and validated by small-angle X-ray scattering profiles for 18 Nups (147 constructs; Supplementary Table 6 and Methods). An ensemble of structural solutions for the NPC that sufficiently satisfied all experimental data was calculated by extensive configurational sampling (8, 9) (Supplementary Table 3 and Methods). Variability among these solutions defines the precision of our structure, as quantified by the average root-mean square deviation between solutions in the final ensemble (9). Our final structure defines the positions of 552 Nups (**Fig. 8.4** and Supplementary Videos 1–3), with an overall precision of about 9Å (Extended Data Fig. 1e, f). The centroid solution is used as the representative structure. The structure was validated by numerous independent tests (Extended Data Figs 1, 7, 8, Supplementary Tables 3, 4 and **Methods**).

The multiple functionalities and enormous size of the NPC present unique and substantial structural challenges: it must form a stable passageway with a fixed inner diameter; it must be anchored to the nuclear envelope and stabilize the pore membrane within which it resides, with a height appropriate for the thickness of the nuclear envelope; it must correctly position the transport machinery; and it must resist stresses that might lead to disassembly or malfunction. Our structure suggests how each of these challenges is met and — by comparison with the vertebrate scaffold (6) — how different organisms

may meet these challenges (see the section ‘Evolutionary origin and diversity of the NPC’).

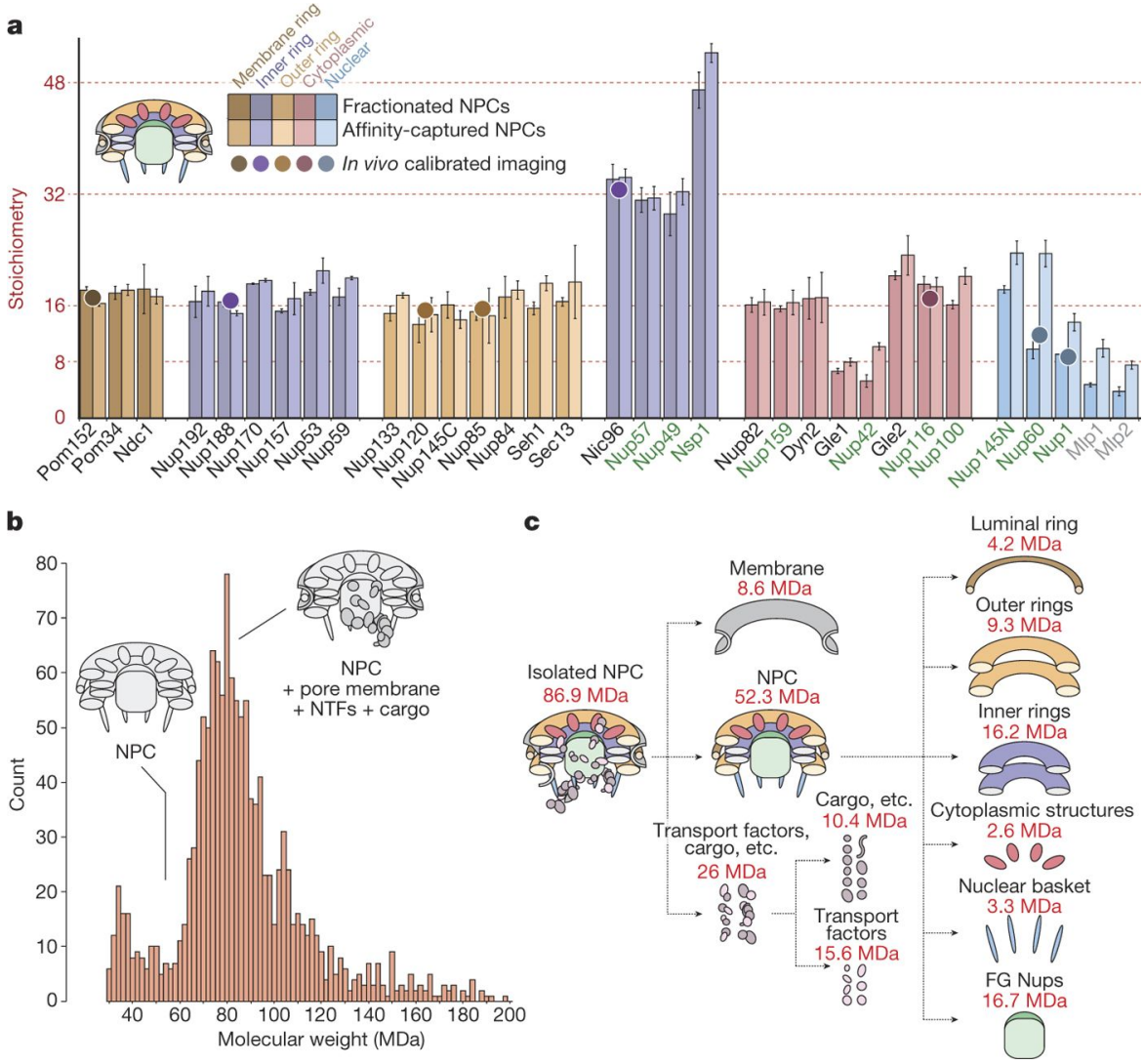

Figure 8.1 | Defining the mass, composition and stoichiometry of the native NPC. **a**, Stoichiometry of the entire complement of NPC components determined by quantitative mass spectrometry (bar plot) and by in vivo calibrated imaging of Nup–GFP reporters (dots) (Extended Data Fig. 3a, b). Darker and lighter color bars (average ± s.d.) represent measurements from a diploid non-tagged *S. uvarum* strain (n=2 or 3 technical and 2 biological replicas) and haploid tagged *S. cerevisiae* strains (n=1–3 technical and 4 biological replicas), respectively. Each Nup is colored on the basis of its localization, as depicted in the cartoon. FG-repeat-containing Nups are labelled in green. **b**, Affinity captured whole NPCs were analysed intact by charge detection mass spectrometry, and a representative mass spectrum is shown. n=2 biological replicas; more than 3 runs with over 1,500 individual NPCs per run. **c**, Dissection of the mass and composition of an NPC.

Forming a stable and defined passageway

The fitness defects of strains containing Nup truncations provide an estimate of the structural importance of the truncated regions (9, 13). We quantified the fitness defect of strains containing systematic truncations of every major symmetric Nup using ODELAY (14) (an automated phenotypic analysis platform; Extended Data Fig. 9). Results were heat mapped onto the NPC structure to reveal critical elements of NPC stability (**Fig. 8.5a**). The inner ring of the NPC contains crucial stabilizing elements, including Nic96, which forms the heart of a diagonally oriented column within each spoke (**Fig. 8.5b**) and interacts with every other protein in the inner ring (**Fig. 8.4d–f**). This high connectivity explains why Nic96 is an essential keystone, holding in place much of the scaffold of the NPC. The remainder of each diagonal column is made of Nup157 and Nup170, which flank Nic96 (**Fig. 8.5b**); Nup157 and Nup170 are functionally redundant but are synthetically lethal (15) and together form another essential element of the diagonal column. Inter-spoke connections represent a second crucial stabilizing element. Nup192 probably serves as a cross-brace between adjacent spokes (**Fig. 8.5a**, **c**). The N termini of Nup170 and disordered regions of Nup53 and Nup59 also form key connections between adjacent spokes (16) (**Figs. 8.4d**, 8.**5c**). The inter-spoke connections are established largely through small, hinge-like contacts that may confer flexibility to the interface between adjacent spokes; the diagonal arrangement of the central columns may also enable rotation or local flexing (**Fig. 8.5ba**), by accommodating compression and expansion forces from nuclear envelope distortions and from the central transporter and the transit of cargoes. Nup188 and Nup192 act as radial separators between the Nic96 column and the triple coiled-coil domains of Nsp1, Nup57 and Nup49, which form a

discontinuous ring that defines the narrowest part of the passageway and may allow some dilation of the NPC (**Fig. 8.4d–f**). This architecture sets a soft upper limit of about 40nm for the size of cargoes that can transit the NPC (4).

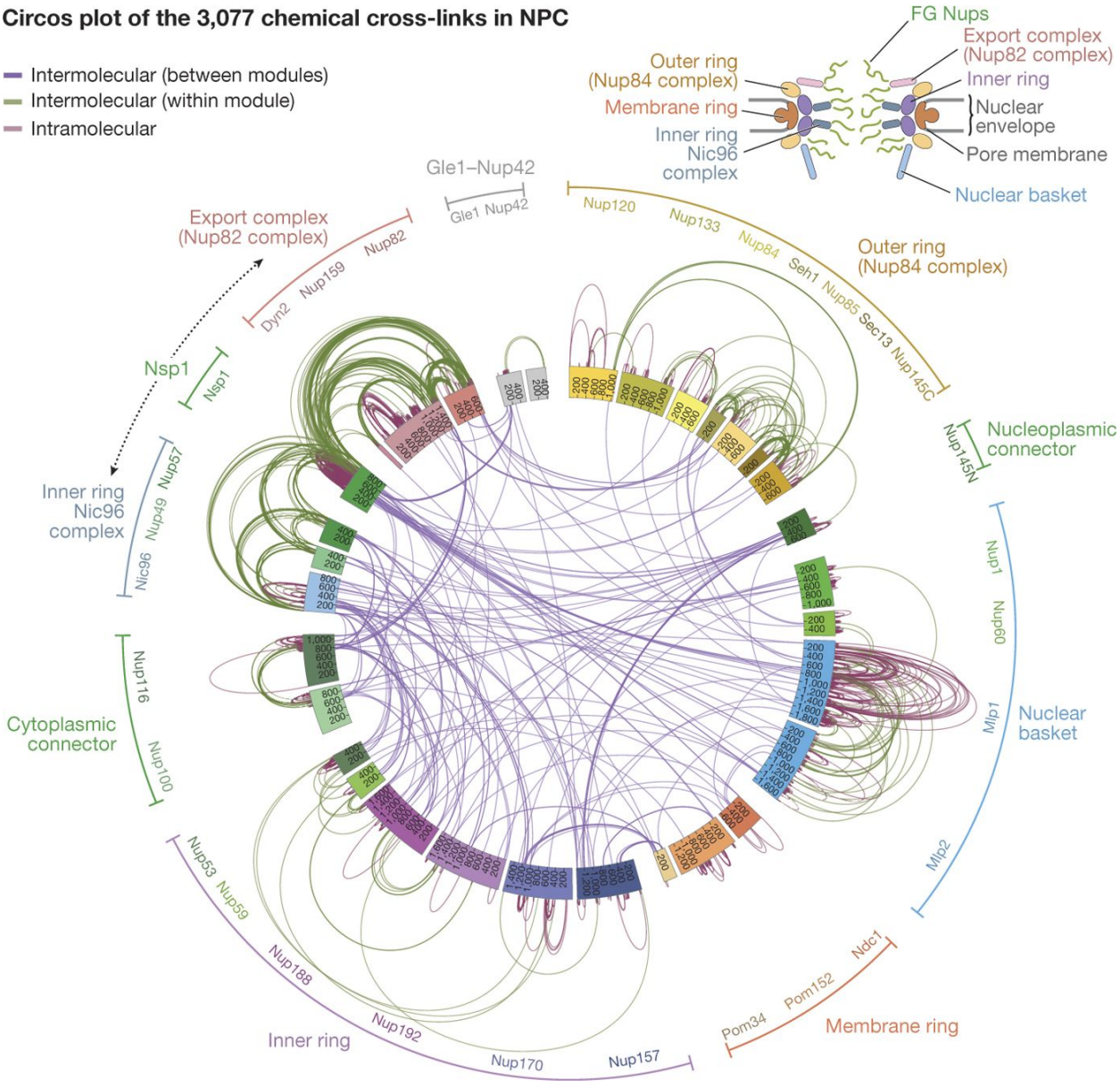

Figure 8.2 | Chemical cross-linking and mass spectrometry reveals nucleoporin connectivity in the NPC.

Circular plot showing the distribution of chemical crosslinks (Supplementary Table 1), mapped to each nucleoporin represented as a coloured segment and with the amino acid residues indicated. The identity of each module and Nup is shown in the periphery of the plot. Key in top left gives the types of cross-links; diagram in top right illustrates the relative positions of modules in the NPC.

How the NPC shapes the nuclear envelope

The pore membrane, where the inner and outer membranes of the nuclear envelope join, defines the inner surface of a torus and therefore has both concave and convex curvatures (**Fig. 8.5d**). The inner ring is anchored to the pore membrane through membrane-binding motifs (MBMs) on the β-propellers at the N termini of Nup157 and Nup170, and on the C termini of Nup53 and Nup59 (refs. (17–19)). These proteins also interact with the scaffold-facing regions of the transmembrane domain (TMD) carrying Nups, such as Pom152, Ndc1 and Pom34 (**Fig. 8.4d**). Together, these MBMs and TMDs form an NPC-anchoring girdle of membrane-associated motifs around the scaffold equator, defining the concave curve of the pore membrane. The convex curvature is defined by both outer and inner rings (**Figs 8.4a, e, 8.5d**). Each outer ring is formed by eight Y-shaped Nup84 complexes arranged head-totail and joined by an interaction between the N termini of Nup120 and Nup133 (20), creating another hinged spoke-to-spoke interface and a minor fitness hotspot (**Fig. 8.5a**, **c**). The outer rings also help define the overall height of the NPC such that it is appropriate for the width of the nuclear envelope. Each Nup84 complex is anchored to the pore membrane by MBMs situated within the N-terminal β-propellers of Nup133 and Nup120 (10, 12, 21) (**Fig. 8.5d**). The convex curvature of the pore membrane is thus defined and stabilized by a ring of MBMs underneath the outer rings and by the thick girdle of MBMs and TMDs around the NPC equator. On the nuclear side, MBMs from Nup1 and Nup60 help anchor the basket to the nuclear envelope (22) (**Fig. 8.5d**, **e**).

In the membrane ring, the luminal domain of Pom152 is composed of nine immunoglobulin-like fold repeats (23) that oligomerize in an anti-parallel fashion to form

eight circumferential arches within the nuclear envelope lumen, producing additional connections between adjacent spokes. Pom152 appears to be pre-stressed by assembly into these arches (**Fig. 8.4e**); the resulting tension may minimize elliptical distortion of the NPC (23). Each arch also delimits a channel (300×120Å wide) between itself and the underlying pore membrane (**Figs 8.3e**, **8.4e**). The outer rings form a series of circumferential arches that align with the Pom152 luminal arches (**Fig. 8.4b**, **e**). These arches align with hinges in the inner ring (**Fig. 8.5c**) that could flex to form lateral openings between spokes. This juxtaposition of arches and transient openings may delineate conduits for nucleocytoplasmic transport of transmembrane proteins (24), potentially resolving the issue of how membrane proteins transit the NPC (25).

Positioning the RNA processing platforms

Whereas the core scaffold is symmetric about the plane of the nuclear envelope, two machineries associated with RNA processing and transport—the basket and export platform— are located at the nuclear and cytoplasmic faces of the NPC, respectively (**Fig. 8.4d–f**). At the core of the export platform is the Nup82 complex, which has a coiled-coil bundle that is attached to the Nup85–Seh1 arm and hub region of the Nup84 complex in the cytoplasmic outer ring (**Fig. 8.4f**). Together, they form a lateral platform that faces the central channel. An α-helical rod that extends from the Nup82 complex holds Gle1, the RNA helicase Dbp5 and the FG-repeat-carrying Nup42 over the middle of the central channel (9, 25, 26). As a result, numerous transport-factor-docking sites and ATP-dependent RNA remodeling proteins are aligned above the cytoplasmic exit of the NPC to efficiently receive exporting RNAs, to remodel and then release them into the cytoplasm. Likewise, Mlp1 and Mlp2 in the nuclear basket are anchored to the core

scaffold mainly by the Nup85–Seh1 arm, similar to the Nup82 complex (**Figs 8.4d, e, 8.5e**). The nuclear basket serves as a platform for the first stages of RNA processing and export (27), and the export platform organizes the last stages of export (26). Similarities between the export platform and basket suggest that these structures are ancient homologues (Extended Data Fig. 10); their asymmetric localization directs unidirectional export of transcripts out of the nucleus.

Flexible connectors tie the NPC together

Certain disordered connectors have recently been shown to be important for holding parts of the scaffold together (6, 7, 16); the complete structure presented here highlights the extent to which such connectors are critical to NPC integrity. Remarkably, flexible connectors run the entire length of each spoke, tying together every major element in the NPC (**Fig. 8.5c**, **e**). They link the periphery and outer rings to the inner rings, both inner rings to the pore membrane and adjacent spokes to one another. We identified two types of connectors (Supplementary Results and Discussion). First, there are vertical connections, aligned parallel to the cylindrical axis of the NPC and constituting the main anchor points between the export platform and the inner ring. On the nuclear side, similar connections are present between the nuclear basket and the inner ring, with an additional connection between the basket and outer ring (**Figs 8.4d**, **8.5e**). Second, there are horizontal flexible connectors that link the central channel to the pore membrane between adjacent spokes (**Fig. 8.5e**). Collectively, these flexible connectors may serve to allow limited movement of the more rigid modules with respect to one another, thereby providing the NPC with another degree of flexibility in response to deformation (28).

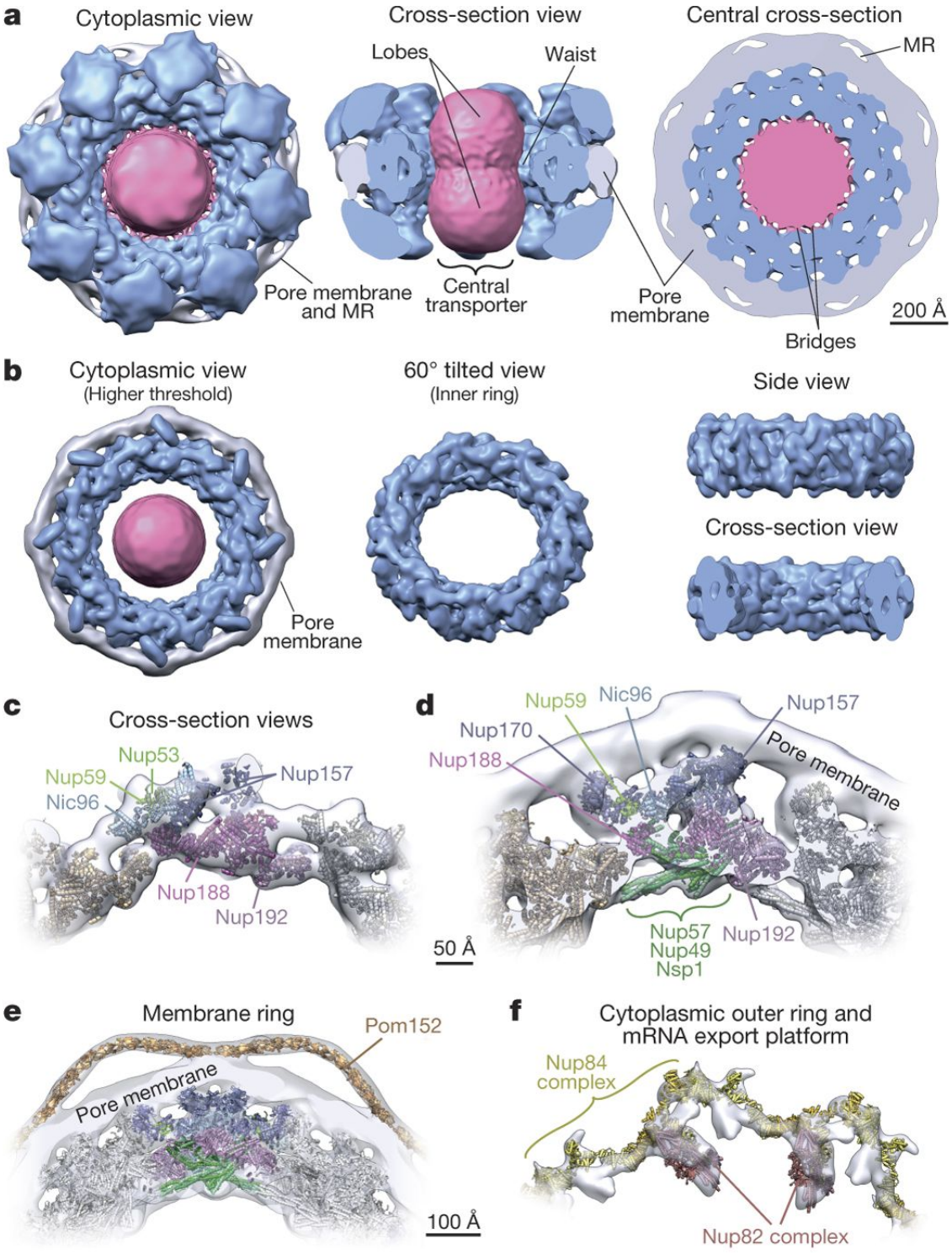

Figure 8.3 | Morphology of the NPC.
a, b. Cryo-ET map of the NPC: core scaffold, blue; membrane region, grey; central transporter, pink. MR: membrane ring. In **a**, cytoplasmic top view (left); cross-section side view (middle); and central cross-section top view (right). **b**, Cryo-ET map from a presented at a higher threshold. Top view (left); 60°-tilted view of the inner ring (middle); and side

(right, top) and cross-section views (right, bottom) of the inner ring. Scale bar, 200Å. **c–f**, Cross-section views show a representative structure embedded within the cryo-ET density (grey), presented with different filtering and thresholding to show the good fit to the cryo-ET map in the inner ring (**c**, **d**), the membrane ring (**e**) and the cytoplasmic outer ring and mRNA export platform (**f**). Nups indicated as in **Fig. 8.4**. Scale bars, 50Å (**c**, **d**) and 100Å (**e**, **f**).

Organization of the transport machinery

Despite its critical function, the central gating machinery has largely been excluded from recent NPC maps (6, 7, 29) and its properties have remained controversial (4). Here we confirm the existence of a large central transporter with two high-density 'lobes' connected by a narrower 'waist' of lower density (5) (**Figs 8.3**, **8.6** and Extended Data Figs 4–6). This central transporter comprises multiple FG repeats that account for about 9 MDa, together with approximately 26 MDa of NTFs and their cargoes caught in transit (though they may be somewhat averaged out in our map) (**Fig. 8.1b**, **c** and Extended Data Fig. 3c). Indeed, even after isolation, each NPC carried 10–80 copies of each of the observed NTFs (30), reflecting the huge and varied transport flux through NPCs.

The localization of FG-repeat anchor points reveals three patterns. First, a vertical path is formed along each spoke by a continuous array of FG repeats (**Fig. 8.6a**, **c** and Extended Data Fig. 11b–d). By binding to these repeats, NTFs may follow these paths across the entire NPC. Second, the FG anchor points of Nsp1–Nup57–Nup49 form a central ring on the equator of the NPC (**Fig. 8.6b**). Thin bridges in our cryo-ET map coincide with the location of these FG anchor points, which indicates that these bridges comprise the FG repeats themselves, emanating from their anchor points (**Fig. 8.6b**). Third, the structured regions of the NPC largely direct the FG-repeat regions inwards toward the axis of the central channel (**Fig. 8.6a**), instead of projecting from the NPC

towards the cytoplasm and nucleoplasm as they are often represented (9). This geometry generates a highly concentrated (25–150 mM) and dynamic FG-repeat phase through which cargo-carrying NTFs readily pass, facilitated by their specific FG interactions, whereas nonspecific macromolecular diffusion is hindered by this same dense phase (31).

It has previously been suggested that the two main types of FG repeat ('Phe-X-Phe-Gly/Phe-Gly' (FXFG/FG) and 'Gly-Leu-Phe-Gly' (GLFG)) are segregated in the NPC to define functionally distinct zones of the gating machinery (32). Consistent with this, and with the known role of FXFG/FG-type repeats in docking RNAs during export (33), we find that FXFG/FG-type repeats are enriched in the nuclear and cytoplasmic peripheries of the NPC, where the RNA-associating export platform and basket reside (**Fig. 8.6d** and Extended Data Fig. 11c). By contrast, the GLFG-type repeats are enriched in regions adjacent to the inner ring and near the cytoplasmic entrance to the central channel. This cytoplasmic localization coincides with the position of FG repeats that are most important for limiting the passage of nonspecific macromolecules (**Fig. 8.6e** and Extended Data Fig. 11d), and is consistent with the known role of GLFG-type repeats in maintaining the passive permeability barrier (34, 35).

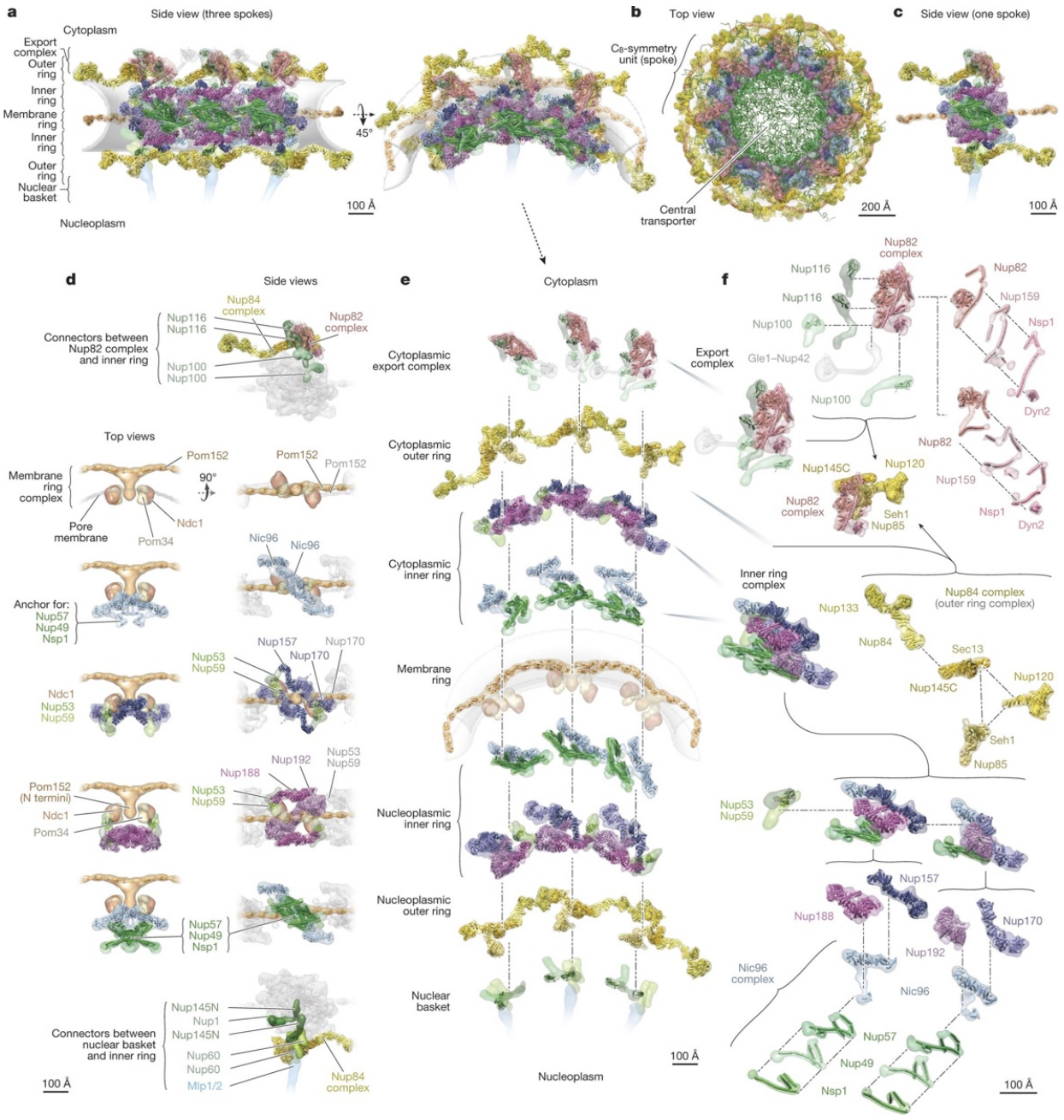

Figure 8.4 | Structural dissection of the NPC.
Complete structure of the NPC and its components shown in different orientations, with a model of the pore membrane region shown in grey (Supplementary Videos 1, 2, 3). For each Nup, the localization probability density of the ensemble of structures is shown with a representative structure from the ensemble embedded within it (Supplementary Table 2). **a**, Two views of three consecutive NPC spokes (C_8-symmetry units), showing how the coaxial outer, inner and membrane rings run continuously between spokes. **b**, Cytoplasmic top view of the complete NPC structure with modelled FG-repeat regions (green). **c**, Side view of a single NPC spoke. **d**, Relative position of major NPC components and connections both within and between spokes. Top (left) and side (right)

views are shown. The membrane ring (beige) is included for reference. Flexible connectors between outer and inner rings are shown in the top and bottom panels, with the inner and membrane rings shown as faded grey densities. **e**, Exploded view of three consecutive spokes, spanning from the cytoplasmic face (top) to the nuclear face (bottom), with dashed lines connecting neighboring rings. **f**, Cytoplasmic mRNA export complex (top), the Nup84 complex (center) and the inner ring complex, including the Nic96 complex (bottom), from a single spoke. The complexes are shown as an exploded diagram, with dashed lines connecting neighboring components.

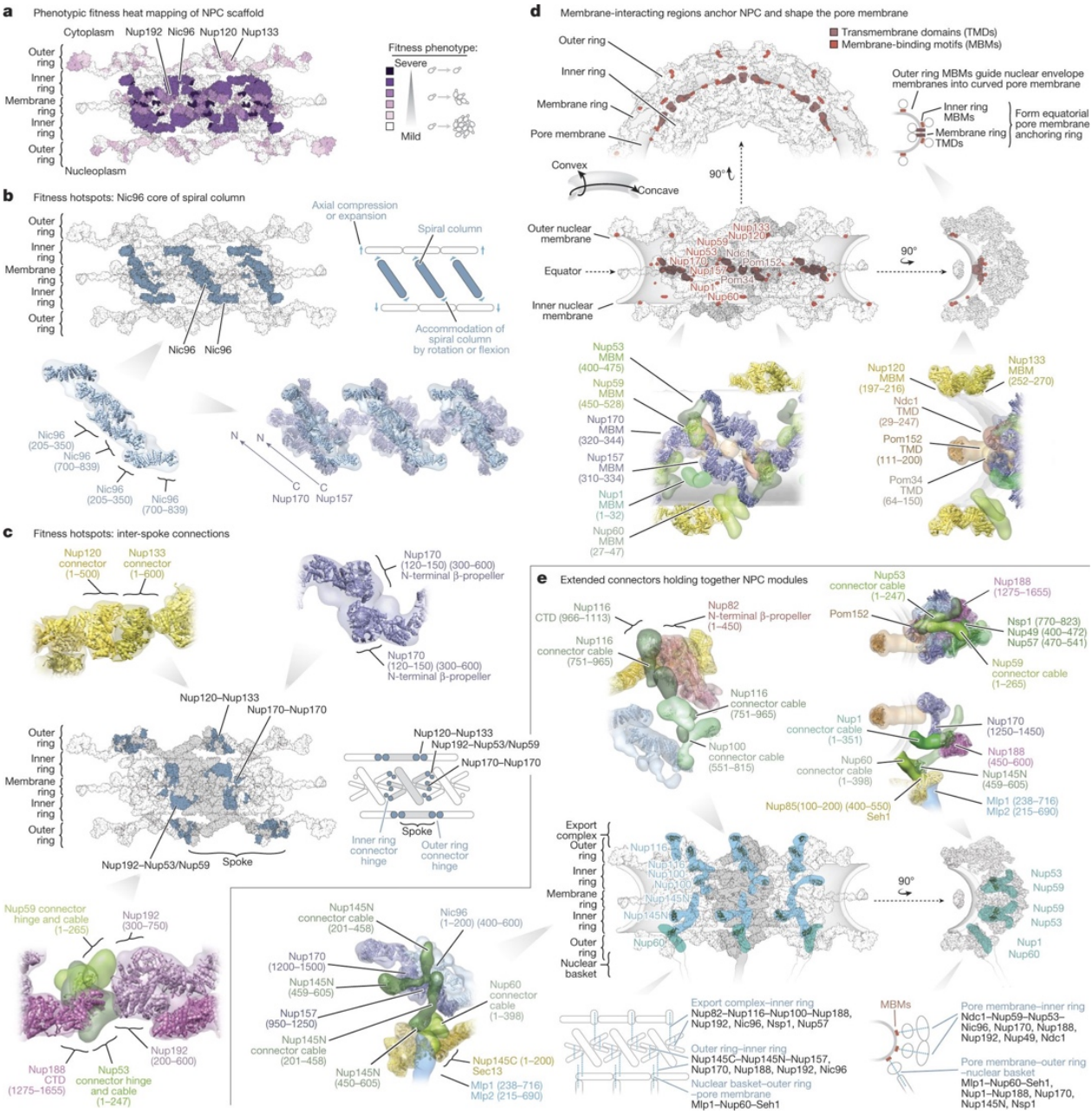

Figure 8.5 | Key NPC architectural features and principles.
a, Severity of fitness defects, indicated in increasing intensity of purple shades for specific truncations of nucleoporins (Extended Data Fig. 9), mapped onto three spokes of the NPC. **b**, **c**, Structures corresponding to the position of the most severe defects (dark blue). In top panels in **b**, diagonally oriented columns reinforcing the core scaffold may accommodate NPC compression and expansion (diagram to right). Bottom, molecular details of Nup arrangement (relevant residue numbers indicated). In middle panels in **c**, the position of hotspots coincides with spoke-to-spoke connections. Central spoke, grey; flanking spokes, white (schematic, middle right). Top and bottom, molecular details of

spoke-to-spoke connector hinges. **d**, Top left and centre left, three spokes shown as top and front views; centre right, one spoke in side view. Schematic indicates convex and concave pore membrane curvatures. Positions of TMDs and MBMs are depicted and their proteins are labelled in brown and orange, respectively. Top right, diagrammatic side view showing how the MBMs and TMDs curve the pore membrane. Bottom, molecular details of the Nups containing the TMDs and MBMs. **e**, Second row left, three spokes in front view, showing how vertical connector Nups (cyan) spanning from the cytoplasmic to nuclear sides of the NPC connect the rings. Second row right, one spoke in side view, showing how horizontal connector Nups (aquamarine) connect modules spanning from the pore membrane to the central channel. First row and bottom row left show molecular details of the connectors within the NPC. Bottom row centre and right, diagrammatic views of the connectors depicted as blue dotted lines; modules connected labelled in blue; major Nups being contacted by connectors listed in grey.

Evolutionary origin and diversity of the NPC

NPCs share architectural features with vesicle coating complexes (Extended Data Fig. 10), which led us to hypothesize that they share a common evolutionary ancestor, the 'protocoatomer' (36). Two major families of coating complexes exist: COPI/clathrin and COPII, each of which have discrete vesicle recognition and trafficking roles (37, 38). We find both COPI/clathrin-like and COPII-like features in the NPC, suggesting that ancestral COPI and COPII coating families evolved first and were followed by the NPC, which may have evolved through a partnership of COPI and COPII coats. This hypothesis implies that the nucleus was a later addition in the evolutionary path of the first eukaryotes (Supplementary Results and Discussion).

Despite substantial conservation of some elements of NPC architecture, other elements can vary widely between species. Generally, the inner ring appears most conserved (39, 40), as is seen in a comparison of our yeast structure with that of the human scaffold (6), although the latter is more expanded (Extended Data Fig. 12). By contrast, peripheral elements exhibit considerable lineage-specific losses and duplications (40, 41). In yeast, each outer ring is formed by 8 copies of the Nup84 complex (**Figs 8.3f, 8.4**), whereas in

vertebrates each outer ring contains 16 copies of the equivalent Y-shaped complex arranged in two interlocked rings (17, 42). Moreover, we see neither an additional copy of Nup157 or Nup170 connecting the outer and inner rings nor Nup188 or Nup192 in the outer rings, as indicated in humans (6) (**Fig. 8.4d–f** and Extended Data Fig. 12a). Another previous model assumed that fungal and human core scaffolds have essentially identical structures (7). Our data invalidate this assumption (**Fig. 8.1**, Supplementary Table 2a and Supplementary Results and Discussion), as well as an earlier ‘fencepost’ model (43). In summary, there is no single universal NPC structure; instead, similar structural elements are used in somewhat different arrangements to generate many lineage-specific adaptations.

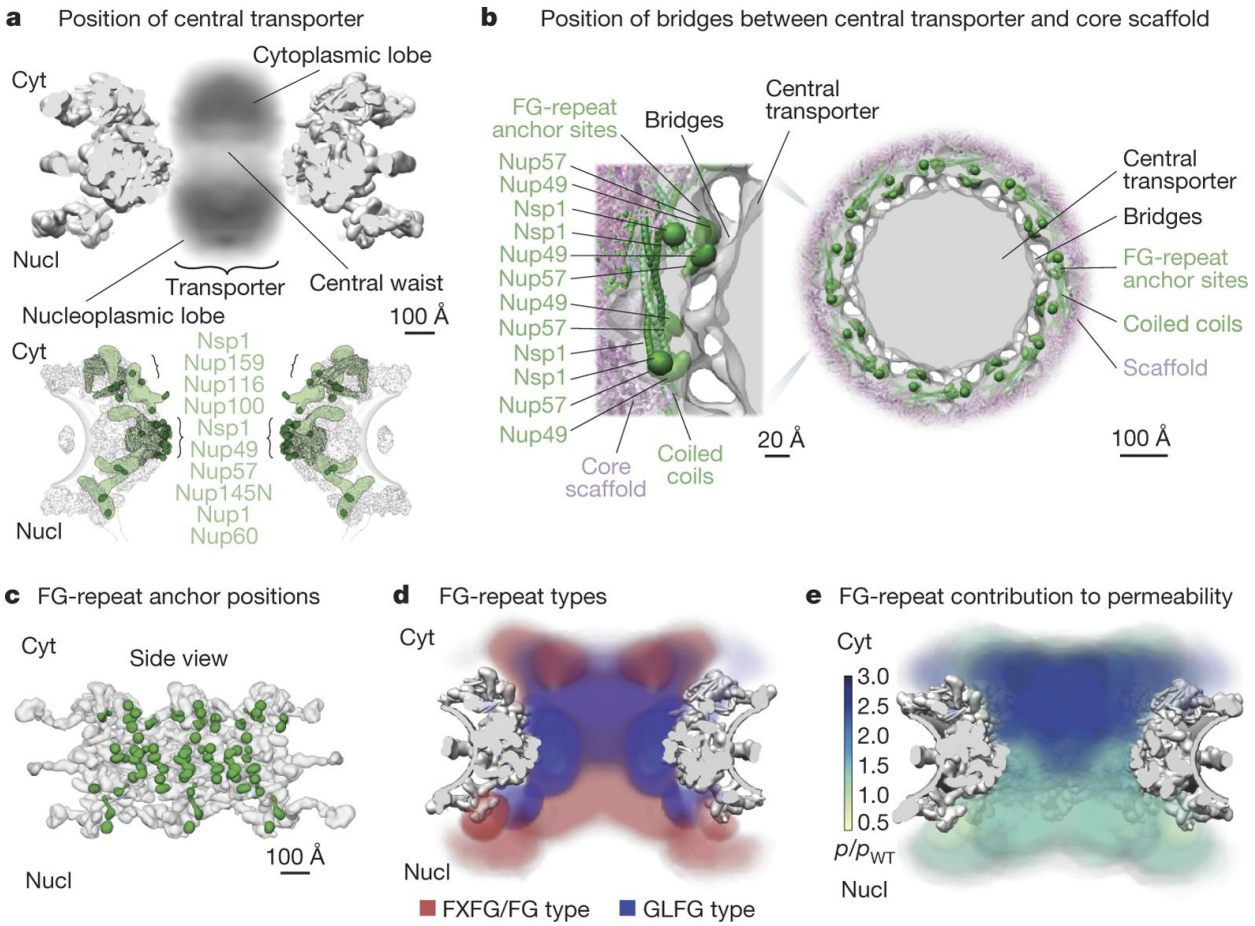

Figure 8.6 | The distributions of FG repeats informs the NPC transport mechanism.

a, Central transporter density from the cryo-ET map (**Fig. 8.3**) is shown within the structure of the NPC scaffold (grey) (top). Features of the central transporter are indicated. Anchors (light green) in FG Nups largely direct the FG-repeat emanating points (dark green) towards the central channel (bottom). Scale bar, 100Å. **b**, Central cross-section of the cryo-ET map (grey) with embedded representative NPC structure (**Fig. 8.4**), showing the central transporter and the bridges connecting it to the core scaffold in top view (scale bar, 100Å), with a magnified view of one spoke on the left (scale bar, 20Å). The anchor points for the FG repeats of Nup49, Nup57 and Nsp1 are depicted as green densities. **c**, Position of FG-repeat anchor points (green) within a side view of three spokes of the scaffold (grey). Scale bar, 100Å. **d**, Heat mapping of repeats of FXFG/FG type (red) and GLFG type (blue), from Brownian dynamics simulations (Methods), showing partitioning to different regions of the central channel. Scale bar, 100Å. **e**, Heat mapping of the effect of FG-repeat region truncations on NPC permeability; the severity of the permeability defect (34) (measured as permeability relative to permeability in wild type, (p/pWT)) is indicated in increasing shades from minor (light green) to severe (dark blue). Cyt, cytoplasm; nucl, nucleoplasm. Scale bar, 100Å.

Conclusions

We have described the structure of the entire yeast NPC at sub-nanometre precision. At the heart of the inner ring, rigid diagonal columns reinforce the structural integrity of the NPC. Membrane-binding and transmembrane Nups are strategically placed throughout the core scaffold to stabilize pore membrane curvature and clamp the NPC to the nuclear envelope. Connectors run the length of each spoke, flexibly tying together all the major modules in the NPC. The architecture of the NPC is reminiscent of a suspension bridge, in which rigid supporting columns are firmly anchored to a substrate and flexible suspension cables connect the columns and roadway to provide a strong and resilient structure. We show that most FG Nup anchor points face inwards toward the NPC central channel to generate a highly concentrated milieu of FG repeats: FXFG/FG repeats form mRNA docking ‘traps’ at the entrance and exit of the channel, and GLFG repeats help form a cytoplasmically biased permeability barrier.

Despite differences, yeast and human NPCs retain a notable degree of structural conservation (Extended Data Fig. 12a, b). As a result, many of the conclusions drawn here should be applicable to the human NPC. To illustrate this point, we mapped the positions of yeast homologues of the oncogenic hotspot human Nup214, Nup98 and Tpr (2) (Extended Data Fig. 12c). Rather than being randomly scattered, these positions coincide with RNA-binding platforms on the cytoplasmic and nucleoplasmic faces of the NPC as well as with several critical connectors and associated FG regions. This conservation suggests that alterations in RNA export, and changes in NPC architecture induced by defective connectors, may underlie the altered behaviour of NPCs in cancer

cells. Thus, our yeast structure provides a roadmap with the potential to advance our understanding of NPC physiology and nuclear transport in general.

Methods

Yeast strains and materials

All S. cerevisiae strains used in this study are listed in Supplementary Table 5, with the exception of the Nup84 complex truncation mutants (13) and the Pom152 truncation mutants (23). Unless otherwise stated, strains were grown at 30°C in YPD medium (1% yeast extract, 2% bactopeptone and 2% glucose). The diploid *S. uvarum* strain (ATCC 9080) was grown and processed for nuclear envelope purification as previously described (44). The following materials were used in this study: Dynabeads M-270 Epoxy (143.02D; Invitrogen); rabbit IgG (55944; MP Biomedicals); protease inhibitor cocktail (P-8340; Sigma-Aldrich); and Solution P (2mg Pepstatin A, 90mg PMSF, and 5ml of absolute ethanol).

Immuno-purification of the endogenous *S. cerevisiae* NPC.

An immunopurification protocol for the isolation of endogenous whole NPCs from *S. cerevisiae* was developed using previously published methodology (3, 45–49). S. cerevisiae Mlp1-, Nup84- or Nup82-encoding genes were genomically tagged with PrA preceded by the human rhinovirus 3C protease (PPX) target sequence (GLEVLFQGPS). Cells were grown in YPD medium at 30°C until early log phase (~2×107 cells/ml), collected, frozen in liquid nitrogen and cryogenically lysed in a planetary ball mill PM 100 (Retsch) (http://lab.rockefeller.edu/rout/protocols). Frozen cell powder was resuspended in 9 volumes of resuspension buffer (20mM HEPES–KOH pH 7.4, 50 mM potassium

acetate, 20 mM NaCl, 2 mM MgCl2, 0.5% (w/v) Triton X-100, 0.1% (w/v) Tween-20, 1mM DTT, 10% (v/v) glycerol, 1/500 (v/v) protease inhibitor cocktail (Sigma)). Cell lysate was clarified by centrifugation at 2,500 RCF for 5min followed by filtration through 1.6-µm filters (Whatman glass microfibre syringe filters). Magnetic beads (Invitrogen) conjugated to rabbit IgG antibodies (http://lab.rockefeller.edu/rout/protocols) were added to the clarified cell lysate at a concentration of 50µl slurry per 1g of frozen cell powder and incubated for 30min at 4°C. Beads were washed once with 1ml of elution buffer without protease inhibitors (20mM HEPES–KOH pH 7.4, 50mM potassium acetate, 20mM NaCl, 2mM MgCl2, 0.1% (w/v) Tween-20, 1mM DTT, 10% (v/v) glycerol). For native elution of the complex, the desired volume of elution buffer with PreScission protease (GE Healthcare) (1/15 (v/v)) was added to the beads and incubated for 45min at 4°C. A magnet was used to remove the beads and collect the supernatant. Beads were subsequently washed with the desired volume of elution buffer containing 1/500 (v/v) protease inhibitor cocktail (Sigma). The total elution volume was centrifuged at 20,000g for 5min to remove the residual magnetic beads. Typical yield of the immuno-purification is ~4µg of isolated NPCs per 1g frozen cell powder (see Extended Data Fig. 2b for SDS-PAGE analysis; for gel source data, see Supplementary Fig. 1).

Mass and stoichiometry of the native *S. cerevisiae* NPC.

Quantification of the absolute stoichiometry of each nucleoporin in the native NPCs was performed using a strategy that combined several orthologous methods (Extended Data Fig. 2a): (1) use of synthetic concatemers of tryptic peptides or QconCATs (50) to define the relative stoichiometry of each component by quantitative mass spectrometry in affinity-captured NPCs; (2) in vivo calibrated imaging analysis of GFP-tagged Nups (51),

to quantify the absolute copy number per NPC of Nups selected to represent each major module of the NPC; and (3) charge detection mass spectrometry to measure the total mass of affinity-captured NPCs (52). For the calculation of the integrative NPC structure, the final copy numbers were rounded to fit the known NPC C8-symmetry and these values are indicated in Supplementary Table 2a.

NPC QconCAT design and purification

Mass spectrometry quantification of the relative amounts of each nucleoprotein in the purified NPC complex was performed using two specifically designed, heavy-labelled synthetic internal standards or QconCATs (50, 53) (Extended Data Fig. 2d, e) formed by concatenated quantotypic nucleoporin peptides. To minimize the potential effect of having different residues flanking the trypsin cleavage site on the cleavage efficiency, we included the native three-residue flanking sequences framing the trypsin cleavage site for each peptide (54). For QconCAT-A (Extended Data Fig. 2d), two peptides for each of the nucleoporins and one peptide for Staphylococcus aureus protein A and Aequorea victoria GFP proteins were selected (Supplementary Table 7) on the basis of their favourable signal responses in liquid chromatography–mass spectrometry analyses of NPC samples and by fitting to the following criteria (when possible): (1) the native three-residue flanking sequences at both sides of the trypsin cleavage sequence do not contain Lys or Arg; (2) avoid the presence of Cys or Met residues within the peptide; (3) avoid the presence of potential internal trypsin cleavage sites (Lys or Arg residues); (4) peptides should be less than 3,000 Da (small size); and (5) avoid peptides showing obvious interferences from co-eluting peptides during the liquid chromatography separation for mass spectrometry analyses. QconCAT-B included two quantotypic peptides for Nup159, Mlp2, Nup192,

Nup84, Nup85, Nup120, Nup49, Nup57, Pom152 and Nic96, and the same GFP peptide as in QconCAT-A (Supplementary Table 7). As an internal control, both QconCAT-A and -B included the same peptides for Nic96, Pom152 and GFP. Each synthetic gene was designed by concatenation of the sequences encoding the selected peptides and addition of a 6× His C-terminal tag (Extended Data Fig. 2d). A 3×FLAG peptide was also included at the N terminus of QconCAT-A, resulting in a protein of 148.2kDa. The Escherichia coli codon optimized sequences were cloned into: (1) plasmid pET15-b (as a NcoI–XhoI fragment) in the case of QconCAT-A; and (2) pGEX6p-1 (as a BamHI–XhoI fragment) in the case of QconCAT-B, resulting in the expression of a 68.1kDa protein with a N-terminal GST tag that was mainly used as a sacrificial peptide (55). The QconCAT proteins were expressed by growing 300ml of BL21 E. coli cells at 37°C to OD600=0.6 in minimal M9 medium without ammonium chloride (50, 53) supplemented with light amino acids and 0.5mg/ml of heavy arginine and lysine (l-arginine:HCl 13C6; l-lysine:2HCl 13C6, Cambridge Isotope Laboratories). IPTG (1mM) was used to induce expression of the constructs for 3h at 37°C. Collected cells were processed using BugBuster Extraction Reagent (Novagen), as indicated by the manufacturer, to isolate the inclusion bodies where the QconCAT protein is accumulated. The full-length QconCAT-A was then purified by resuspending the inclusion bodies pellet in binding buffer (20mM sodium phosphate pH 7.4, 45mM imidazole, 500mM NaCl, 6M guanidinium chloride, 10mM TCEP (0.5M Bond-Breaker TCEP solution, Thermo Fisher Scientific), 1/500 protease inhibitor cocktail (Roche)) and passed through an equilibrated His-Trap HP (GE Healthcare) at room temperature. The retained NPC QconCAT-A was then eluted in 20mM sodium phosphate pH 7.4, 500mM imidazole, 500mM NaCl, 6M guanidinium hydrochloride, 1mM TCEP,

1/500 protease inhibitor cocktail. One-hundred-microlitre aliquots of the resulting elution were precipitated to eliminate the guanidinium hydrochloride by adding ice-cold ethanol to a final concentration of 90% and incubating the samples at −20°C for 2h. Samples were then centrifuged for 10min at 14,000r.p.m. and 4°C to pellet the precipitated protein. The resulting pellet was washed with ice-cold 90% ethanol and allowed to air-dry until most of the liquid was evaporated, leaving a wet pellet. These pellets were solubilized with 5% SDS, 500mM Tris–HCl pH 8.0, 5mM TCEP buffer, by incubating for 5min at room temperature and 5min at 72°C and centrifuged for 10min at 14,000 r.p.m. at room temperature. The supernatants were recovered and two of them combined and injected into a TSKgel G4000SWxl size-exclusion column (TOSOH Bioscience) coupled to a TSKgel SWxl guard column (TOSOH Bioscience), pre-equilibrated in running buffer (40 mM HEPES–KOH pH 7.0, 150 mM NaCl, 0.1% SDS, 5 mM TCEP, 1mM EDTA). Two-hundred-microlitre fractions were collected and analysed by SDS-PAGE to detect the presence of the QconCAT-A peak. Fractions containing the full-length pure protein were supplemented with a final 20% glycerol (v/v), aliquoted, flash-frozen in liquid nitrogen and stored at −80 °C for further use. In the case of QconCAT-B, the protein was purified using His-Trap HP and the elution precipitated and prepared as described for QconCAT-A. The resulting sample was injected into a TSKgel Super SW3000 size-exclusion column (TOSOH Bioscience) pre-equilibrated in running buffer (40mM HEPES–KOH pH 7.0, 150mM NaCl, 0.1% SDS, 5mM TCEP, 1mM EDTA). One-hundred-microlitre fractions were collected and analysed by SDS-PAGE to detect the presence of the QconCAT-B peak. Fractions containing the full-length pure protein were stored as indicated for QconCAT-A.

For the quantitative mass spectrometry analysis, the native NPCs from *S. cerevisiae* PPX–PrA-tagged haploid strains were affinity captured as described above, or purified as enriched NPCs from a diploid *S. uvarum* strain using a subfractionation method previously described in detail (44, 56–58) (http://lab.rockefeller.edu/rout/protocols), using 0.035mg heparin per mg of fraction protein. For affinitycaptured NPCs, the natively eluted NPCs (5µg) were concentrated by pelleting at 40,000r.p.m. for 20min at 4°C in a TLA 55 rotor (Beckman). In the case of subfractionation-enriched NPCs, a volume of the 1.45M:1.85M sucrose gradient fraction that contained an estimated 5µg of NPCs was diluted 1/5 (v/v) in bt-DMSO buffer (10mM bis-Tris-HCl pH 6.5, 0.1mM MgCl2, 20% DMSO) and pelleted at 15,000r.p.m. for 450min at 4°C in a TLA 55 rotor (Beckman). For in-solution mass spectrometry analysis of subfractionation-enriched NPCs, 0.1µg of QconCAT-A were immobilized on Dynabeads His-Tag Isolation and Pulldown resin (Thermo Fisher Scientific) pre-equilibrated in binding buffer (20mM HEPES, 150mM NaCl, 8M urea, 5mM TCEP). The purified protein sample was incubated with the resin for 20min at room temperature, and washed with binding buffer 5×200µl to eliminate residual SDS; in-solution and in-gel analyses showed consistent results (not shown), so most of the further analyses were performed in-gel to improve consistency, speed and throughput. For the solid-state in-gel mass spectrometry analyses, pelleted NPCs were solubilized in 10µl of 0.5M Tris-HCl pH 8.0, 5% SDS by incubating at 72 °C for 5min and then diluted 1:1 with 20% glycerol, 50mM TCEP, 0.5mM EDTA, 0.05% (w/v) bromophenol blue. Approximately equimolar amounts of 0.1µg of purified QconCAT-A or 0.045µg of purified QconCAT-B were added to each 5-µg NPC sample. Samples were then incubated at 72°C for 10min, cooled to room temperature and treated with a final

30mM of iodoacetamide (Sigma), at room temperature in the dark for 30min. Samples was then loaded into a 4% (37.5:1) stacking acrylamide SDS-PAGE gel prepared in-house. The resulting bands, containing a mixture of whole NPCs and QconCAT protein (labelled with a stable isotope), were excised and processed for quantitative mass spectrometry analyses.

Mass spectrometry characterization of QconCAT labelled with a stable isotope

The mass of purified intact QconCAT-A protein, labelled with a stable isotope, was analysed by matrix-assisted laser desorption/ionization (MALDI) (Extended Data Fig. 2e) on a JEOL JMS-S3000 SpiralTOF mass spectrometer using the ultra-thin-layer sample preparation method (59, 60) in which α-cyano-4- hydroxycinnamic acid (Sigma) was used as the matrix. The mass of QconCAT-A was internally calibrated with horse myoglobin. Mass calibration and background subtraction were carried out with the JEOL msTornado control software, and additional analyses were carried out with the MoverZ software (61). The QconCAT-A protein was also characterized by peptide mapping, in which tryptic peptides from in-gel digestion were loaded onto a PicoFrit column (New Objective) with an integrated emitter tip (360-mm O.D., 50-mm I.D., 10-mm tip) self-packed with 6cm of reverse-phase C18 material (ReproSil-Pur C18-AQ, 3-mm beads, Dr. Maisch GmbH), and analysed with a LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific), with a Agilent 1200 series HPLC system (Agilent) and a micro electrospray source built in-house. The purified QconCAT-B was characterized by peptide mapping on a Thermo Orbitrap Fusion mass spectrometer, with a Thermo Easy-nLC 1000 HPLC and a Thermo Easy-Spray electrospray source.

Stoichiometry quantification of NPC using QconCAT and by mass spectrometry

Mixtures of yeast NPC proteins and QconCAT, labelled with a stable isotope, were enzymatically digested either in solution in the presence of urea or inside a SDS-PAGE gel matrix. For in-solution digestion, a mixture of the NPCs and immobilized QconCATs on His-Dynabeads were sequentially digested at room temperature by Endoproteinase Lys-C in 8M urea for 66h and by trypsin in 2M urea for 3h. For in-gel digestion, proteins in the gel matrix were digested in 100mM TrisHCl at room temperature either sequentially by 0.25–2µg Endoproteinase Lys-C for 66h and by 3–25µg trypsin for 3h, or—in later experiments—by 25µg trypsin alone for 3h. The resulting peptides were analysed in duplicate by liquid chromatography–mass spectrometry using a Thermo Fusion or a Thermo Q Exactive Plus mass spectrometer, with a Thermo Easy-nLC 1000 HPLC and a Thermo EasySpray electrospray source. The ratios of light nucleoporin (L) to heavy QconCAT proteins (H) for standard peptides were obtained using MaxQuant (62), complemented with manual determination.

We incorporated two standard peptides from each nucleoporin into the QconCAT standard to enable us to check for internal consistency of the measured L/H ratios for each nucleoporin. Our check required that the relative standard deviations of L/H ratios for two standard peptides from two duplicate liquid chromatography–mass spectrometry runs—that is, for a total of four measurements per nucleoporin—be≤25%. When deriving relative stoichiometry for any given preparation of NPCs analysed in different replication experiments, we corrected for variations in the mixing ratio of light nucleoporins and heavy QconCAT proteins by scaling the measured L/H ratios to minimize the sum of the relative standard deviations of the L/H ratios over all nucleoporins. The resulting scaled L/H ratios from different experiments were used to derive the average L/H ratios and

standard deviations. To assay for potential nucleoporin stoichiometry bias arising from capture through particular affinity handles, we used stable isotope labelling with amino acids in cell culture followed by mass spectrometry (SILAC–MS) analysis of these preparations versus the nuclear envelope preparation. We performed n=2 or 3 technical and 2 biological replicas for NPCs purified by subfractionation procedures from a diploid, non-tagged *S. uvarum* strain, and n=1–3 technical and 4 biological replicas for the nuclear-envelope-corrected affinity-captured NPCs from haploid, tagged *S. cerevisiae* strains (**Fig. 8.1a**).

The absolute stoichiometry (**Fig. 8.1** and Supplementary Table 2a) was then determined by normalizing the summed copies of Nup188, Nup120 and Nic96 per NPC to 64 copies (that is, 16 for Nup188 and Nup120, and 32 for Nic96).

SILAC-MS analyses of the NPC stoichiometry

A preparation of yeast nuclear envelopes obtained by a previously established subfractionation method (44) does not involve disruption of the nuclear envelope membrane by detergents and generates sheets of nuclear envelope studded with intact NPCs. To assess the degree to which the affinity-captured NPCs were intact, we used SILAC-MS to compare the levels of each Nup in the affinity-captured preparation relative to those in the nuclear envelope preparation. To do this, the nuclear envelope sample labelled with light isotopes was mixed with Mlp1-PPX-PrA affinity-captured NPC sample labelled with a heavy isotope (l-lysine:2HCL $^{13}C_6$) in a SILAC experiment. Mixtures of nuclear envelope proteins and NPCs labelled with stable isotopes, purified using the Mlp1–PPX-PrA handle, were digested sequentially in gel matrix by Endoproteinase LysC and by trypsin. Resulting peptides were analysed by liquid chromatography–mass

spectrometry on a Thermo Q Exactive Plus mass spectrometer, with a Thermo Easy-nLC 1000 HPLC and a Thermo Easy-Spray electrospray source. H/L ratios for all peptides were obtained using MaxQuant (62), complemented with manual examinations. The H/L ratios of peptides were used to derive the H/L ratios of nucleoporins and standard deviations (data not shown). The result showed that the affinity-capture process does not affect the overall ratios of the major Nups and NPC modules relative to the nuclear envelope samples (data not shown), indicating that the affinity-capture procedure generates intact NPCs. We also used this comparison relative to nuclear envelopes to correct for the slight increases observed in the ratios of Nups closely associated with the Mlp1 handle in the affinity-captured NPCs (**Fig. 8.1a**).

In vivo calibrated imaging analysis of GFP-tagged Nups

Calibrated imaging data were acquired as previously described (51). Using the avalanche photodiode imaging module of a Zeiss confocor 3, confocal z-stacks of live yeast were acquired with a 40× 1.2 NA Plan-Apochromat water objective. The 488-nm laser line was used to excite GFP, with a 405/488/561 dichroic. Emission was reflected with a LP580 emission dichroic and collected through a BP 505–540-nm emission filter. The pinhole was set to 1 Airy unit. The zoom was set to maintain a pixel size of 55nm, and a z-step size of 400nm was used. After acquisition, images were binned in XY by 2, resulting in an effective pixel size of 110nm, and anaphase cells were analysed for diffraction-limited Nup spots along the anaphase bridge. These spots, when present, were fit to a 2D Gaussian curve to obtain the amplitude of the signal. The z-slice with the maximum signal intensity of the spot was analysed. Fluorescence correlation spectroscopy was used to convert the amplitude of the Gaussian fit of the Nup spot number of molecules of GFP. In

brief, using a strain expressing only cytosolic GFP, fluorescence correlation spectroscopy determined the average number of molecules in the focal volume, as previously described (51). Then, the amplitude of the signal of the Nup spot was compared to the intensity of cytosolic GFP, taken with the same imaging setup. For all measurements, number 1.5 coverslips were measured for uniformity, and the correction collar of the water objective was optimized for this thickness using signal intensity of Alexa Fluor 488 in solution. For each day data were acquired, the calibration using cytosolic monomer GFP was obtained.

Phospholipid analysis

These analyses were performed by Avanti Polar Lipids using their standard protocols.

Label-free mass spectrometry quantification of the NPC and associated proteins

Raw mass spectrometry files from QconCAT Mlp1-PPX-PrA immuno-isolation experiments were analysed using the MaxQuant iBAQ method (63). Only peptides that were not isotopically labelled (that is, not QconCAT) were considered. Proteins were filtered to require more than three unique peptides per protein, and stoichiometries normalized to the absolute minimum value of the difference between label-free and the QconCAT stoichiometry for all the Nups (Extended Data Fig. 3c and Supplementary Table 8). Stoichiometries were multiplied by molecular weight to obtain mass per NPC complex and the results summed to obtain total mass of the NPC (**Fig. 8.1c** and Extended Data Fig. 3c).

Living mass of the NPC with charge detection mass spectrometry

The charge detection mass spectrometry instrument has previously been described (52, 64). In brief, the measurements are made by trapping single ions in a linear electrostatic ion trap. As the ions oscillate back and forth in the trap, they pass through a cylindrical

electrode. The charge induced on the electrode is detected by a charge sensitive preamplifier. The resulting signal is amplified and digitized, and then analysed using fast Fourier transforms. The fundamental frequency provides the m/z and the magnitude is proportional to the charge. The mass of each ion is then obtained by multiplying the charge and m/z. Each NPC sample was characterized by measuring the masses of several thousand ions individually and then binning the masses to yield a true mass spectrum (**Fig. 8.1b, c**).

Chemical cross-linking and mass spectrometry analysis of the cross-linked NPC

NPCs were immuno-purified from Mlp1-PPX-PrA, Nup82-PPX-PrA and Nup84-PPX-PrA S. cerevisiae strains. After native elution, 0.5 or 1.0mM disuccinimidyl suberate (DSS) was added and sample incubated at room temperature for 30min with gentle shaking (~1,000 r.p.m.). The reaction was quenched with 50mM ammonium bicarbonate or SDS-PAGE buffer containing 100mM Tris-HCl.

The sample was then precipitated using 90% methanol at −80°C or concentrated in a speed vacuum before separation by SDS electrophoresis. The sample was reduced by 10 mM tris-(2-carboxyethyl)-phosphine (Invitrogen) at 80°C for 15–20min, cooled to room temperature and alkylated by 50mM iodoacetamide for 20min in the dark to block the formation of disulfide bonds. After reduction and alkylation, the cross-linked complexes were separated by 3–8% SDS-PAGE (NuPAGE Tris-Acetate Fisher) to reduce the complexity of the sample. For in-gel digestion, the high-molecular-weight-region gel bands (>460kDa, estimated by the high-molecular-weight protein markers, Invitrogen) corresponding to the cross-linked NPC proteins were sliced and proteolysed by trypsin as previously described (10, 65). In brief, gel plugs were crushed into small

pieces, ~5–10μg of sequencing-grade trypsin (Promega) per ~100μg protein was added with subsequent 6–8h incubation. This proteolysis was repeated once more to ensure optimal results. Peptides were extracted by formic acid and acetonitrile, desalted on C18 cartridges (Waters) and snap-frozen before fractionation.

To reduce the complexity of the sample, proteolysed mixtures were separated by an orthogonal two-step fractionation strategy. First, size exclusion chromatography (66) was used for size-based separation of peptides into 2–4 fractions (~2–10 kDa). Then, a secondary fractionation using a self-packed basic (at pH 10) C18 resins (Dr. Maisch GmbH) resulted in 10–12 peptide fractions, which were subsequently analysed by liquid chromatography–mass spectrometry.

Each peptide fraction was dissolved in the sample loading buffer (5% MeOH and 0.2% FA) and analysed either by an Orbitrap Q Exactive Plus mass spectrometer or a LTQ Velos Orbitrap Pro mass spectrometer (Thermo Fisher Scientific). The Q Exactive Plus instrument was directly coupled to an easy-nLC system (Thermo Fisher Scientific) for electrospray. The cross-linked peptides were loaded onto the Easy-Spray columns (15-cm prepacked columns that are filled with C18 reversephase material of 2 or 3μm particle size, 200Å pore size and 50μm inner diameter, Thermo Fisher Scientific) that were heated to 35 °C. Mobile phase A consisted of 0.1% formic acid and mobile phase B of 100% ACN with 0.1% formic acid. Peptides were eluted in liquid chromatography gradients of 120min (for example, a liquid chromatography gradient of 3–7% B, 0-6min; 7–28% B, 6-101min; 28–100%B, 101–113min; followed by equilibration with 100% A until 120min). Flow rates were set at ~250–275nl/min. Other instrumental parameters for chemical cross-linking and mass spectrometry analyses include: capillary temperature:

250–275 °C; target mass resolutions (at 200 Th): 70,000 for mass spectrometry and 17,500 for tandem mass spectrometry; AGC targets: 1–3×106 (full mass) and 2×105 (tandem mass spectrometry); mass spectrometry range of 300–1,700 Th; isolation window: 1.3–1.7 Th; higher-energy collisional dissociation normalized energy: 24–29; dynamic exclusion allowed once per 75–90s. The top 8 most abundant ions (with charge stage of 3–7 and intensity thresholds of 3,000–7,500 ions) were selected for fragmentation by higher-energy collisional dissociation. The maximum injection times were set at 200ms (for mass spectrometry) and 500–800ms (for tandem mass spectrometry). For samples that were analysed by Orbitrap Velos, the cross-linked peptide mixtures were pressure-loaded onto a self-packed PicoFrit column with integrated electrospray ionization emitter tip (360 O.D, 75 I.D with 15μm tip, New Objective). The column was packed with 10–15 cm reverse-phase C18 material (3μm porous silica, 200Å pore size, Dr. Maisch GmbH). Mobile phase A consisted of 0.5% acetic acid and mobile phase B of 70% ACN with 0.5% acetic acid. The peptides were eluted in a 120- or a 140-min liquid chromatography gradient (8% B to 50% B, 0–93min, followed by 50% B to 100% B, 93–110min and equilibrated with 100% A until 120 or 150min) using a HPLC system (Agilent), and analysed with a LTQ Velos Orbitrap Pro mass spectrometer using similar parameters to the Q Exactive Plus instrument.

The raw data were searched by pLink (67) using a FASTA database containing 34 NPC protein sequences. An initial MS^1 search window of 5 Da was allowed to cover all isotopic peaks of the cross-linked peptides. The data were automatically filtered using a mass accuracy of $MS^1 \leq 10$ p.p.m. (parts per million) and $MS^2 \leq 20$ p.p.m. of the theoretical monoisotopic (A0) and other isotopic masses (A + 1, A + 2, A + 3 and A + 4) as specified

in the software. Other search parameters included cysteine carbamidomethyl as a fixed modification and methionine oxidation as a variable modification. A maximum of two trypsin missed-cleavage sites was allowed. The initial search results were obtained using a default 5% false discovery rate expected by target–decoy search strategy. All spectra were manually verified as previously described (9, 10, 65, 68, 69). The cross-linking data were analysed and plotted by an online software tool, CX-Circos (http://cx-circos.net; W.J. et al., manuscript in preparation) (**Fig. 8.2**).

Cryo-electron tomography of whole NPCs

We used cryo-ET and sub-tomogram averaging (Extended Data Fig. 4a) to obtain a final map with a global resolution of 28Å; the inner ring was solved at 20–25Å (Extended Data Fig. 5 and Supplementary Table 9). To create this map, NPCs were immuno-purified from Mlp1-PPX-PrA *S. cerevisiae* strain, in a final buffer of 20mM HEPES (at pH 7.5), 50mM Potassium Acetate, 20mM NaCl, 2mM MgCl2, 0.1% Tween 20 and 1mM DTT (see 'Immuno-purification of the endogenous S. cerevisiae NPC' for details). The concentration was estimated by SDS-PAGE to be ~0.3–0.4mg/ ml. Freshly cleaned Quantafoil 300 mesh copper grids with 2-µm holes in the support film were prepared with a continuous carbon support film that spanned the holes. Before use, the grids were glow discharged in air, floated on 5µl sample drops for 45min and then washed by serial transfer on 4×20µl drops of elution buffer without glycerol. Each grid was mounted on forceps in a Mark III Vitrobot (FEI) at room temperature and 100% relative humidity. Buffer on the grid was removed by blotting from the bottom with a tool that held a filter paper wedge, using access through the left-hand port. Then, 2µl of freezing buffer was added to the grid from the right-hand port and the grid was plunge-frozen in liquid ethane after blotting.

Cryo-ET data collection was done on a Titan-Krios electron microscope operating at 300kV, equipped with an X-FEG, a post-column energy filter set to 20 eV and a spherical aberration (Cs) corrector (Supplementary Table 9). Images were recorded with a Gatan K2 Summit direct electron detector in integration mode, with single frames taken at each tilt with UCSF Tomo (70), at a nominal spacing of 5.6Å per pixel. A total of 253 tilt series were collected in steps between −60°, 0° and 60° in increments of 2.5–4° for different tilt series. Although the full tilt range was used for tomogram reconstruction, in the final sub-tomogram averaging step only data up to±45° tilt from each sub-tomogram were included in the final average. The dose target for each tilt series was 90–100 electrons per $Å^2$ and followed a cosine α dose curve with a flux of 20 electrons per pixel per second, and a dose of 3.5 electrons per $Å^2$ for the zero tilt image. Extended Data Fig. 4a presents the strategy we used to reconstruct the 3D map of the whole yeast NPC. The 3dmod viewer in IMOD71 was used to visually screen tilts for defects with the Fourier transform, to gauge image motion. In total, 120 tilt series with a defocus range between −4.6 and −7.5 μm were kept for further processing (Supplementary Power Point Presentation slides 1, 2). After interactive test runs with etomo (71), we processed the tilt series in an automated fashion with batchtomo using 7×7 patch tracking to create aligned tilt series and calculated tomograms with backprojection and the simultaneous iterative reconstruction technique (SIRT), which were contrast transfer function (CTF)-corrected by phase-flipping each image in the tilt series (Supplementary Power Point Presentation slides 3–6). The final SIRT tomograms were binned 3× and used for interactive sub-tomogram ‘particle’ picking with e2spt_boxer.py in the EMAN2 single particle tomography package (72, 73) with a low pass filter of 100Å. In total, 6,416 fully sampled unfiltered

sub-tomograms were extracted from the back-projection tomograms in 300×300×300 voxel volumes.

In the alignment and averaging process, new algorithms for high-speed 3D alignment with automatic missing wedge compensation and averaging (73) were used throughout, and were critical for processing such large sub-tomograms. An initial reference was prepared by averaging a small subset of sub-tomograms to produce a low-resolution reference using the C8-symmetry of the complex (74). Owing to the large size and distinct shape of the particles, alignments were unambiguous. The alignment and averaging strategy for the final map was adapted from previously described procedures (72), and applied iteratively. The observed flexibility of the NPC ring initially limited the overall resolution to ~38Å with 5,245 sub-tomograms (data not shown); the sub-tomograms discarded at this stage were those with the worst quality when compared to the average, generally owing to higher noise levels but in some cases due to particle damage or false positives during sub-tomogram picking. We realized that observed flexibility of the NPCs limited our resolution to ~38Å and therefore used a tactic to locally align all individual spokes (C8- symmetry units) to the reference rather than aligning whole rings, which could contain long-range deviations from a perfect toroid. It is important to note that these deviations are not large; across the entire NPC, they are on the order of the 38Å resolution achieved without local alignment. All previous NPC cryo-ET maps have been produced using the approach of dividing the NPC into subunits (6, 29, 75, 76). In brief, two reference volumes were prepared; (1) the entire NPC and (2) a masked volume in which one-fourth of the ring had been retained, centred roughly on the mass of a single subunit. Each NPC was rotationally and translationally aligned to the reference ring. Using

this initial alignment, each NPC was replicated into its eight pseudo-symmetric orientations, then a translation-only alignment against the masked reference was performed. This had the effect of bringing one asymmetric unit per replicated ring into register with the reference at a consistent radius. Although small per-subunit rotations might have occurred, this possibility was not included in the local alignment. The average from the eight subunits was then used to construct a symmetric ring by applying an azimuthal linear ramp mask centred on the mask used for alignment, which fell to zero at an angle of 45° in both directions, and then imposing the C8-symmetry. This interpolates smoothly from one side of the subunit to the other symmetry-related side, to produce a complete symmetrized ring. This processing dramatically reduces the blurring caused by local fluctuations in subunit position and the resulting 3D volume was used as a reference in the next cycle of iterative refinement, which was repeated until no further improvement was observed.

At this stage we realized that the preferred orientation of the particles within our tomograms was leading to anisotropic resolution in the final structure, with 2/3 of the NPC rings oriented within 30° of the C8-symmetry axis (as clearly observed in our raw data). Producing an isotropic average required balancing the various ring orientations by discarding the lowest-contrast rings in the over-populated orientations to more evenly balance the orientation distribution (77–79). This normalizing of orientations, and not 3D classification, was what led to discarding a fraction of our data (Extended Data Fig. 4a). When doing this, we elected to discard noisier sub-tomograms. This was achieved by comparing the agreement of each sub-tomogram with the overall average. In each angular range, we then retained roughly the same number of sub-tomograms, keeping

those with the best quality. The discarded sub-tomograms were nearly as good, meaning we could equally well have used the next best subset of the sub-tomograms with virtually no effect on the final structure. Indeed, in the less common orientations, we were forced to use virtually all of the sub-tomograms irrespective of quality. Thus, in our work, a large fraction of the sub-tomograms were discarded not owing to their poor quality or conformational variability, but rather because of the preferred orientation of the particles within the tomogram. The final reconstruction used 1,864 (of the 6,416 initial) sub-tomograms. These sub-tomograms were further divided randomly into two groups for resolution assessment. ‘Gold standard’ refinement was used for resolution testing and to ensure self-consistency (Supplementary Power Point Presentation slides 7, 8). Both global and local resolution assessments were done using a set of tiled Gaussian masks to estimate the local resolution and reproducibility of the structure, which is one of the standard methods for local resolution assessment. In brief, a 3D Gaussian shape is generated with a ‘full width at half maximum’ parameter of at least 2× the anticipated resolution, and generally even larger. This Gaussian shape is then applied as a mask to both maps and a Fourier shell correlation (FSC) curve is computed. This process is repeated in a tiled pattern throughout the volume. This provides a resolution for each sampled location in the volume. This procedure involves a trade-off: with smaller Gaussians, FSC curves are less precise, but with larger Gaussians the resolution estimate is less localized. The Gaussians overlap to provide better sampling of the volume. The resulting resolution map is similar to those produced by ResMap80, but it measures FSC, which is filter-independent, unlike ResMap (80) that requires unfiltered volumes and would not work well on sub-tomogram averages. The global resolution of

our cryo-ET map at the standard FSC0.143 cutoff is ~28Å and the local resolution distribution ranges from 20 to 38Å, with the inner ring being in the 20–25Å range (Extended Data Fig. 5a–d). This local resolution estimate was used to locally filter the density map, to produce a map with the appropriate level of detail in each area (Extended Data Fig. 5). The size of the Gaussian window was 140Å, indicating the smallest region over which the resolution is considered to vary. Although this may seem large, it is small compared to the size of the overall NPC (Extended Data Fig. 5). CTF phase-flipping was applied during tomographic reconstruction and a final approximate amplitude correction was applied to the averaged NPC ring. Therefore, theoretical CTF curves for the mean defocus values present in the tomograms were averaged assuming 10% amplitude contrast. The reciprocal of this curve was then applied as a filter to the final uncorrected map. The cryo-ET density map was refined at 5.3Å per pixel on the basis of a recalibration of the map with known structures.

In parallel, and as an additional validation of our final map, we also carried out a tomographic analysis of the yeast NPC dataset (the same 6,416 sub-tomograms) using RELION 1.4, and incorporated a CTF model (81, 82). In brief, we calculated back-projection tomograms without phase-flipping corrections for the individual tilted images, and binned the output sub-tomograms twofold to 10.6Å per pixel. The datasets underwent sequential rounds of 2D classification using Z-projections of the sub-tomograms to eliminate poor particles. A subsequent 2D classification identified near-top, tilted and side views; the latter provided an independent estimate of NPC thickness perpendicular to the nuclear envelope (620–640Å). A 3D reconstruction using the best sub-tomograms with RELION produced a map at ~35Å resolution (data not shown) with similar features to

those obtained with the e2spt in the EMAN2 single particle tomography package (72, 73), including distinct connections between each spoke and the transporter, further validating these features in our cryo-ET map (**Fig. 8.3**). Finally, Z-projections of original sub-tomograms that were roughly aligned along the C8-symmetry axis were used for an additional unsupervised 2D classification, which produced classes with central transporters without using the C8-symmetry restraint (Extended Data Fig. 6a, b). Differences in the apparent resolution of the class averages in Extended Data Fig. 6a, b reflect different particle numbers in the classes. As mentioned above, the dataset of particles has a strong orientational bias, in which the NPCs tend to bind to the carbon support film with a range of 0–30 degrees of tilt. The class averages are based on 2D projections along the z axis of the original sub-tomograms, to avoid issues with the missing wedge, and there is therefore a disparity in particle numbers in the classes. Tilting in the tomographic data collection helped to fill in the missing data, but we took great care to ensure an equal coverage of Fourier space in the calculation of our final map, to avoid distortions, and also took a number of other steps to ensure that radiation damage and loss of data quality in later tilts was minimized by using only information in Fourier space at ±45° from each particle sub-tomogram when they were combined to form the final map. The RELION map serves as a strong validation of our final map, because if our map was flawed, a reconstruction with RELION would have resulted in a different map (Extended Data Fig. 5e). Additionally, the fact that a RELION reconstruction resulted in a ~35Å resolution map—virtually the same resolution as obtained in our 'intermediate' map described above (~38Å)—validates our methodology and the quality of our final map (Extended Data Fig. 5e). An additional point that provides prima facie evidence that our

cryo-ET map was calculated correctly is that local two-fold symmetry (C2-symmetry), which was expected in the inner ring of the NPC, emerges without any enforcement, whereas the overall map shows a clear asymmetry with large and distinctly different features on the nuclear and cytoplasmic face of the yeast NPC (which were also observed in the RELION map) and a slightly tapered appearance, as is shown in **Fig. 8.3a** and Extended Data Figs 4d, 5b.

Small angle X-ray scattering.

Small angle X-ray scattering (SAXS) measurements for 147 constructs of 18 Nups (9, 12, 23, 83–86) (Supplementary Table 6; S.J.K. et al., manuscript in preparation; and Source Data) were carried out both at the Stanford Synchrotron Radiation Lightsource Beamline 4-2 in the SLAC National Accelerator Laboratory and at the SIBYLS Beamline 12.3.1 of the Advanced Light Source in the Lawrence Berkeley National Laboratory. SAXS data were collected at concentrations ranging from 0.5 to 5.0 (or higher, depending on the sample) mg/ml, using the previously defined standard protocol (12, 23, 83); approximately 20 one-second exposures were used for each sample and buffers at 15°C. Further details of the SAXS experiments have previously been published (9, 12, 23, 83–86).

Phenotypic analysis by one-cell doubling evaluation by living arrays of yeast

Yeast growth phenotypes were quantified using the one-cell doubling evaluation by living arrays of yeast (ODELAY) assay, as previously described14. In brief, yeast was cultured in YPD medium in 96-well plates overnight. Cultures were diluted to an OD600 of 0.09 and allowed to grow for 6 h at 30°C. The cultures were then diluted again to an OD600 of 0.02 and spotted onto YPD agarose medium. The resulting cultures were then observed using time-lapse microscopy for 48h with 30min intervals between images. All images

were collected on Leica DMI6000 microscopes with a 10× 0.3NA lens using bright field microscopy. MATLAB scripts using the Micro-Manager interface controlled the image collection process (87). Six independent experiments were performed. The population growth rates were scored against each other using the following equation:

$$z_{mean} = \frac{1}{n}\sum_{i}^{n}\frac{d_i - \mu_i}{\sigma_i}$$

in which d_i is the ith decile of doubling time of the query population, μ_i is the mean of the ith decile of the doubling time of the parent strain and σ_i is the standard deviation of the ith decile of the doubling time of the parent strain. The mean and standard deviation deciles were calculated from at least 4 separate populations containing at least 200–300 individuals. All calculations were performed using MATLAB scripts. Following Z-scoring of the populations, an additional weight was added to the scoring for truncation strains that occurred in haploid versus diploid strains of yeast.

Negative-stain electron microscopy of the native Nic96 complex

An affinity-captured and natively eluted sample of the endogenous Nic96 complex (composed of Nic96, Nsp1, Nup49 and Nup57) was applied to a glow-discharged grid and stained with 1% uranyl formate. Images were collected on a Philips CM200 transmission electron microscope (FEI) operating at 200 kV at 50,000× magnification and a defocus of ~1.5µm (2.03 pixels per Å). Images were recorded on a Gatan UltraScan 1000 2k×2k CCD camera (Gatan). Particles were selected using Boxer from EMAN (88), normalized and then phase-flipped using ctfit from EMAN. In total, 34 class averages (selected classes shown in Extended Data Fig. 7g) were generated through ISAC (89) that classified ~86% of the original set of 5,458 particles.

Integrative structure determination of the S. cerevisiae NPC

The structure of the *S. cerevisiae* NPC, including the scaffold, membrane rings, cytoplasmic export platform and nuclear baskets in the context of the pore membrane—but excluding the flexible FG regions—was solved by integrative structure determination (see 'Integrative structure determination of the *S. cerevisiae* NPC scaffold, membrane rings, cytoplasmic export platform and nuclear basket'). Moreover, the distributions of the FG regions and the cargo-bound NTFs, comprising the central transporter, were computed by Brownian dynamics simulation (see 'Brownian dynamics simulation of FG repeats and NTFs').

Integrative structure determination of the *S. cerevisiae* NPC scaffold, membrane rings, cytoplasmic export platform and nuclear basket

Integrative structure determination of the *S. cerevisiae* NPC proceeded through four stages (8, 90–92) (Extended Data Fig. 1, Supplementary Table 3 and Supplementary Videos 1–3): (1) gathering data, (2) representing subunits and translating data into spatial restraints, (3) configurational sampling to produce an ensemble of structures that satisfies the restraints and (4) analysing and validating the ensemble structures and data (Extended Data Figs 1, 7, 8 and Supplementary Tables 2–4). The integrative structure modelling protocol (stages 2, 3 and 4) was scripted using the Python modelling interface (PMI) package version 4d97507, which is a library for modelling macromolecular complexes based on our opensource integrative modelling platform (IMP) package (90) version 2.6 (https:// integrativemodeling.org). The current procedure is an updated version of previously described protocols (9, 10, 12, 93–96).

Stage 1: gathering data

The stoichiometry of Nups in the NPC was determined using native mass spectrometry and biochemical quantification of the purified NPC complex (**Fig. 8.1** and Extended Data Figs 2, 3). In total, 3,077 intra and intermolecular DSS and EDC unique cross-links were identified using mass spectrometry (**Fig. 8.2** and Supplementary Table 1), which informed the spatial proximities among the 32 Nups and their conformations. The density map of the entire NPC was determined by cryo-ET at an average resolution of 28Å, with the local resolution as high as ~20Å for the inner ring, which informed the shape of the NPC (**Fig. 8.3** and Extended Data Figs 4–6). Re-interpreted immunoelectron microscopy data (3, 8) informed the positions of 29 Nups. Predictions of the transmembrane domains obtained from the Saccharomyces Genome Database (97) (http://yeastgenome.org) and predictions of MBMs from the HeliQuest webserver (12, 98) informed about their respective proximities to the pore membrane. Previous immuno-electron microscopy measurements (99) informed the end-to-end distance for Mlp1 and Mlp2. Low-resolution electron microscopy images of the NPC44 informed the diameter of the distal basket ring formed by Mlp1 and Mlp2.

Representations of individual Nups and some of their sub-complexes (Supplementary Table 2 and references therein) relied on (1) atomic structures of 21 yeast Nup domains and 3 sub-complexes determined by X-ray crystallography or nuclear magnetic resonance spectroscopy; (2) our previously determined structures of Nup116, Nup133, Nup145N, Nup192 and Pom152, as well as the Nup82 and Nup84 sub-complexes solved by integrative structure determination (9, 10, 12, 23, 83–86); (3) 29 comparative models built with MODELLER 9.13 (100) on the basis of known structure(s) detected by HHpred (101, 102); (4) SAXS profiles for 147 constructs of 18 Nups (9, 10,

12, 23, 83–86); (Supplementary Table 6; S.J.K. et al., manuscript in preparation); (5) secondary structure, disordered regions, and domain boundaries predicted by PSIPRED (103, 104), DISOPRED (105), and DomPred (106), respectively; (6) coiled-coil regions of Nup82, Nup159, Nsp1, Nup49, Nup57, Mlp1, and Mlp2 predicted by COILS/PCOILS (107) and Multicoil2 (108); (7) an atomic structure of the $Nup53^{229–365}$ RRM domain from *S. cerevisiae* determined by X-ray crystallography (P. Sampathkumar et al., manuscript in preparation); and (8) negative-stain electron micropscopy density maps of full-length Nup192 (EMD-5556 (86)) and Pom152 (EMD-8543 (23)). See Supplementary Table 2 and references therein for all above (1) to (8).

Our previously published topological map of the NPC (3) and the 82 composites determined by affinity purification and overlay assay (8) were not used for computing the current NPC structure, but were used for validating the current NPC structure.

Stage 2: representing subunits and translating data into spatial restraints

Information about the modelled system (see ‘Stage 1: gathering data’) can in general be used for defining the system’s representation, defining the scoring function that guides sampling of alternative structural models, limiting sampling, filtering of good-scoring structures obtained by sampling and final validation of the structures. Here the NPC representation relies primarily on stoichiometry as well as atomic structures, integrative structures, comparative models and SAXS profiles of Nups and their sub-complexes (Supplementary Tables 2 and 6, and references therein); the scoring function relies on chemical cross-links, the cryo-ET density map, immuno-electron microscopy localizations, excluded volume, sequence connectivity, the shape of the pore membrane and four types of sequence-based localization relative to the membrane (below); the

sampling benefits from symmetry constraints (below); and the validation of the final structure relies in part on the SAXS profiles (Supplementary Table 6) and composites determined by affinity purification and overlay assays8 (below).

To improve computational efficiency and avoid a representation that was too coarse, we represented the NPC in a multi-scale fashion. A rigid body consisting of multiple beads was defined for each X-ray structure, NMR structure, comparative model and integrative structure of the NPC components (Supplementary Table 2). The remainders of the Nup sequences not in rigid bodies (36.8% of residues, excluding FG repeats) were represented as flexible strings of beads. In a rigidbody, the beads have their relative distances constrained during configurational sampling, whereas in a flexible string the beads are restrained by the sequence connectivity, excluded volume and potentially additional restraints, such as chemical cross-links, as exemplified in previous publications (9, 10, 23, 93, 109).

Rigid bodies (63.2% of residues, excluding FG repeats) were coarse-grained using two resolutions, in which beads represented either individual residues or segments of up to ten residues. The coordinates of a 1-residue bead were those of the corresponding Cα atom. The coordinates of a 10-residue bead were the center of mass of the ten constituent 1-residue beads. Finally, the remaining regions without an atomic representation (that is, the predicted transmembrane and disordered regions) were represented by a flexible string of beads encompassing 25 to 100 residues each; the low-resolution representation of these regions is justified because their conformations are likely to be ‘decoupled’ from the structure of the rest of the NPC (3, 110). We used the SAXS data to confirm the rigid body representations of eight Nups with X-ray structures, comparative models, and

previously published atomic integrative structures (9, 10, 12, 23, 83–86) (Extended Data Fig. 7f and Supplementary Tables 2, 6). The rigid-body representation of a Nup construct was validated by a χ-value that quantifies the difference between the computed (from an atomic rigid-body representation using FoXS (111)) and experimental SAXS profiles, except for several constructs of Nup133 and Nup192 that were flexible during integrative modelling and were thus evaluated as previously described (12, 86, 112). The χ-value validation assumes that each Nup construct, corresponding in most cases to a single domain (not the whole protein), has the same conformation in solution and in complex; this assumption is consistent with other data (for example, the chemical cross-links and cryo-ET map). The SAXS validation is necessarily limited to *S. cerevisiae* constructs of Nups that exist as a rigid monomer in solution and do not contain FG repeats; rigid body representations of the constructs from other species, constructs that oligomerize in solution and constructs that include FG repeats cannot be easily used for validation, because of the sensitivity of a computed SAXS profile to the differences in the sequence and stoichiometry, as well as to potential errors in comparative modelling (especially of insertion and deletion).

After producing this validated representation, we next encoded the spatial restraints on the basis of information gathered in Stage 1, according to the following steps (Supplementary Table 4; for the definition of the scoring function consisting of these restraints, see ‘Scoring function’):

(1) Cross-link restraints: 1,643 of the 3,077 unique cross-links (**Fig. 8.2** and Supplementary Table 1a) were used to restrain the distances spanned by the crosslinked residues, relying on a Bayesian scoring function (10). The evaluation takes into account

the ambiguity due to multiple copies of identical subunits and, for crosslinks involving the same protein type, due to the lack of knowledge of whether they are intra- or intermolecular (9, 93, 109); the ambiguous cross-link restraint considers all intra- and intermolecular assignments in multiple copies of identical subunits, with only the least violated distance contributing to the score. The remaining 1,434 DSS and EDC cross-links (**Fig. 8.2** and Supplementary Table 1b–f) were already used as restraints to build the integrative structures of the Nup84 (10) and Nup82 subcomplexes (9), represented here as rigid bodies. The two homo-dimer DSS crosslinks between two copies of residue 62 of Pom152 (23) and two copies of residue 151 in Nup60 (22) were transformed into harmonic upper-distance bounds, enforcing the homo-dimer configuration.

(2) Cryo-ET density restraint: the cryo-ET density restraint was applied, which corresponded to the cross-correlation between the Gaussian mixture model (GMM) representation of most Nups and the GMM representation of the cryo-ET density map (95, 113–115) (**Fig. 8.3** and Extended Data Figs 4–6); we used a GMM representation for the sake of computational efficiency, necessitated by the large size of the NPC. An assessment of a given structure against a density map is much faster when both are represented with a mixture model (because the number of components in a mixture model is much smaller than the number of grid points covering the maps). However, these two scores are very strongly correlated. Thus, the structures obtained with a grid representation, if we had sufficient computational power, would certainly be indistinguishable from the current NPC structures (115).

A 90° arc of the cryo-ET density map was approximated by the GMM, which contained 1,750 components computed using the expectation-maximization algorithm implemented

in scikit-learn (http://scikit-learn.org); the cryo-ET GMM appeared to be sufficient to reproduce the complete features of the density map (excluding the central transporter region). To use a comparable number of GMM components for Nups, a Nup was approximated by a GMM component for each of its 100 to 500 residues. The cross-correlation quantified the degree of overlap between the Nup GMM components and the cryo-ET GMM components.

(3) Immuno-electron microscopy localization restraints: the immuno-electron microscopy localization restraint was used to localize the C-terminal beads of 29 of the 32 Nups, on the basis of previous immuno-electron microscopy data (3, 8, 116). This goal was achieved by imposing upper and lower harmonic bounds on the axial and radial coordinates of the restrained bead, reflecting the uncertainty in the immuno-electron microscopy data (8). The three remaining Nups (Nsp1, Sec13, and Seh1) were not restrained by the immuno-electron microscopy data because of their high uncertainty, presumably due to the positional heterogeneity of the tagged Nup in the multiple superposed electron microscopy images of the NPC. This heterogeneity is more likely to occur for Nups with multiple copies per C2- symmetry unit, which are unlikely to share the same radial and axial coordinates.

(4) Excluded volume restraints: the protein excluded volume restraints were applied to each 10-residue bead, using the statistical relationship between the volume and the number of residues that it covered (8–10, 117).

(5) Sequence connectivity restraints: we applied sequence connectivity restraints, using a harmonic upper bound on the distance between consecutive beads in a subunit, with a threshold distance equal to three times the sum of the radii of the two connected beads.

The bead radius was calculated from the excluded volume of the corresponding bead, assuming standard protein density (8–10, 117).

(6) Membrane exclusion restraints: the membrane exclusion restraints were applied to beads in the non-membrane-spanning Nups or to their segments to prevent these beads from penetrating the pore membrane. A lower harmonic bound at 0Å was applied to the distance between a bead and the closest point on the pore-side membrane surface (3, 8) (modelled as a half-torus with the large and small radii of 390 and 150Å, respectively; Supplementary Table 3), for all coarse beads (10 residues or more per bead) in all Nups but Pom152, Ndc1, and Pom34; the restraint was also applied to all non-membrane coarse beads of Pom152$^{1–110}$, Ndc1$^{1–28}$, Ndc1$^{248–655}$, Pom34$^{1–63}$, and Pom34$^{151–299}$.

(7) Transmembrane domain restraints: the transmembrane domain restraint was used to localize the coarse beads in the predicted transmembrane domains (Pom152$^{111–200}$, Ndc1$^{29–247}$, and Pom34$^{64–150}$; Supplementary Table 2 and references therein) within the pore membrane, which is 45Å thick (3, 8). This aim was achieved by imposing an upper harmonic bound at 45Å and a lower harmonic bound at 0Å on the distance between the bead and the closest point on the poreside membrane surface.

(8) Membrane surface binding restraints: the membrane surface binding restraint was used to localize the coarse beads in the predicted MBMs (Nup1$^{1–32}$, Nup60$^{27–47}$, Nup120$^{135–152}$, Nup120$^{197–216}$, Nup133$^{252–270}$, Nup157$^{310–334}$, Nup170$^{320–344}$, Nup53^{475} and Nup59^{528}; Supplementary Table 2 and references therein), within the pore membrane up to 12 Å from the pore-side membrane surface (118). This aim was achieved by imposing an upper harmonic bound at 12 Å within the pore membrane and a lower harmonic bound at 0 Å on the distance between the bead and the closest point on the pore-side membrane

surface. For Nup120, only the best satisfied of the Nup120$^{135–152}$ and Nup120$^{197–216}$ restraints were used (10, 12) (conditional restraint).

(9) Pom152 perinuclear volume restraint: only the C-terminal region of Pom152 (residues 201–1337) was restrained to the perinuclear lumen of the pore membrane (23). This aim was achieved by imposing a lower harmonic bound at 0Å on the distance between the Pom152 beads and the closest point on the perinuclear side of the membrane surface.

(10) Distal basket ring restraints: the conformations of Mlp1 and Mlp2 were restrained by an upper harmonic bound at 350Å and a lower harmonic bound at 230Å on the distance between the N-terminal and C-terminal beads, on the basis of immuno-electron microscopy measurements (99). In addition, the radius of the distal basket ring was restrained by an upper harmonic bound at 170Å and a lower harmonic bound at 130Å on the radial coordinates of the C-terminal beads of Mlp1 and Mlp2, on the basis of low-resolution electron microscopy images of the NPC (44). The nuclear basket was also informed by cross-linking restraints and the C8-symmetry constraint (see 'Sampling space with symmetry constraints').

Stage 3: Configurational sampling

We used the configurational sampling to produce an ensemble of structures that satisfies the restraints, as described below.

Sampling space with symmetry constraints

We aimed to maximize the efficiency of the configurational sampling: more specifically, we aimed to maximize the precision at which the sampling of good-scoring solutions was exhaustive (see 'Stage 4: analysing and validating the ensemble structures and data'). Therefore, we reduced the number of independently moving parts in the NPC structure

by explicitly considering the C8- and C2-symmetries of the NPC, as follows. The entire NPC consists of 8 clones of the C8-symmetry unit, related by multiples of a 45° rotation around the z axis (**Fig. 8.3** and Extended Data Figs 4–6). The C8-symmetry unit was further broken into two C2-symmetry units and non-C2-symmetric Nups (Supplementary Table 2a); the C2-symmetry unit contains Nups that occur equally on both the cytoplasmic and nucleoplasmic sides (3, 8, 116). For computational efficiency, we defined the coordinate system such that the C2-symmetry is imposed simply by cloning a bead in the C2-symmetry unit at (x, y, z) to (x, −y, −z) (equivalent to a rotation of 180° around the x axis). This aim was achieved by fitting both copies of Pom152 (23) into the cryo-ET density map, followed by moving the centre of the map to the origin of the coordinate system and orienting the map such that the x, −y, −z transformation applies to Pom152. With these symmetries in hand, we sampled only the positions of rigid-bodies and beads corresponding to the Nups in the C2-symmetry unit and non-C2-symmetric Nups. There are no Nups that occur on both the cytoplasmic and nuclear sides and are not related by the C2-symmetry; there are no Nups that occur with a different stoichiometry on both sides. In addition, the luminal domain of Pom152 was considered already well-positioned given its fit into the cryo-ET density map (**Fig. 8.3e**) and peripheral location in the NPC (23), and was not sampled further.

Scoring function

The scoring function included restraints on the sampled Nups and the Pom152 luminal domain as well as restraints across the interfaces with neighbouring symmetry units: (1) the cryo-ET density restraint and distal basket ring restraint applied to the Nups in the sampled C8-symmetry unit; (2) sequence connectivity, immuno-electron microscopy

localization, and the four types of sequence-based localizations relative to the membrane applied to the Nups in the sampled C2-symmetry unit and non-C2-symmetric Nups; and (3) cross-link and excluded volume restraints applied to the pairs of beads for Nups within the sampled C8-symmetry unit and across the interfaces with neighbouring symmetry units.

Sampling algorithm

The search for good-scoring structures relied on replica exchange Gibbs sampling, based on the Metropolis Monte Carlo algorithm (9, 10) (Supplementary Table 3). The Monte Carlo moves included random translation and rotation of rigid-bodies (up to 4Å and 0.04 radians, respectively) and random translation of individual beads in the flexible segments (up to 4Å). As indicated above, these operations were applied only to the sampled rigid bodies and beads. The remaining, symmetry-constrained rigid-bodies and units were moved in lockstep to maintain the exact C8- and C2-symmetries at each sampling step, as described above. Up to 64 replicas were used, with a 1.0–5.0 temperature range. Forty-two independent sampling calculations were performed, each one starting with a random initial configuration. The coordinates were saved every 10 Gibbs sampling steps, each consisting of a cycle of Monte Carlo steps that moved every rigid-body and flexible bead once.

To further increase the efficiency of sampling, we first applied the above Monte Carlo algorithm separately to the following four subsets of Nups, which are co-localized on the basis of previous characterizations (3, 8, 9, 23) and the current cryo-ET density map: (1) Nup82 and Nup84 sub-complexes, (2) Pom152, (3) inner-ring Nups (Nup157, Nup170, Nup188, Nup192, Pom34, Ndc1, Nup53, Nup59 and Nic96$^{205–839}$), and (4) Mlp1

and Mlp2. Next, the best-scoring solutions from sampling each of the first three subsets were combined; they were already in the same reference frame, because they were all obtained by fitting the same cryo-ET density map and immuno-electron microscopy data. The rest of the Nups and the Mlp1– Mlp2 heterodimer were then added in random positions and orientations, followed by another application of the above Monte Carlo algorithm to all sampled Nups. This sampling produced a total of 100,453 modelled structures in 42 independent runs (the score ranges from 88,545.0 to 103,589.5, with the mean and standard deviation of 88,831.5 and 187.4, respectively), requiring ~10 weeks on a cluster of ~2,500 CPU cores. For the most detailed specification of the sampling procedure, see the IMP modelling script (https://salilab.org/npc2018). We considered for further analysis only the 5,529 modelled structures with the scores better than 88,644.1 (1 standard deviation below the mean value); this threshold implies satisfaction of the input datasets within their uncertainties (Supplementary Table 4; see 'Fit to input information'). These structures are already superposed because they were fit into the same Cryo-ET map and sampling did not move the luminal domain of Pom152 (see 'Sampling space with symmetry constraints').

Stage 4, analysing and validating the ensemble structures and data

Input information and output structures need to be analyzed to estimate structure precision and accuracy, detect inconsistent and missing information, and to suggest more informative future experiments. We used the previously published analysis and validation protocol (8, 9). Assessment began with a test of the thoroughness of structural sampling, followed by structural clustering of the modelled structures and estimating their precision based on the variability in the ensemble of good-scoring structures, quantification of the

structure fit to the input information and structure assessment by data not used to compute it; structure assessment by cross-validation was not performed in this case, because it takes ~10 weeks on approximately 2,500 CPU cores to compute an ensemble of structures for a single set of input datasets. These validations are based on the nascent wwPDB effort (92) toward archiving, validating and disseminating integrative structures. We now discuss each one of these validations in turn.

Thoroughness of the configurational sampling

We must first estimate the precision at which sampling found the most good-scoring solutions (sampling precision); the sampling precision must be at least as high as the precision of the structure ensemble that is consistent with the input data (structure precision). As a proxy for testing the thoroughness of sampling, we performed four tests of sampling convergence (119), as follows.

The first convergence test confirmed that the scores of refined structures do not continue to improve as more structures are computed, essentially independently of each other (Extended Data Fig. 1c).

The second convergence test confirmed that the good-scoring structures in independent sampling runs 1–21 (structure sample 1; nsample1=2,359 structures) and 22–42 (structure sample 2; nsample2=3,170 structures) satisfied the data equally well. The non-parametric Kolmogorov–Smirnov two-sample test (120, 121) (two-sided) indicates that the difference between the two score distributions is insignificant (*P* value (1.0) > 0.05). In addition, the magnitude of the difference is small, as demonstrated by the Kolmogorov–Smirnov two-sample test statistic, *D*, of 0.045 (Extended Data Fig. 1d). Thus, the two score distributions are effectively equal.

Next, we considered the 5,529 good-scoring structures themselves (not their scores as in the two tests described above). For stochastic sampling methods, thoroughness of sampling can be assessed by showing that multiple independent runs (for example, using random starting configurations and different random number generator seeds, as is the case for structure samples 1 and 2) do not result in noticeably different structures (8–10, 13). We tested the similarity between structure samples 1 and 2 in the following two ways.

The third convergence test (119) relied on the χ^2-test (one-sided) for homogeneity of proportions (122) between structure samples 1 and 2 (Extended Data Fig. 1e, f). The test involves clustering structures from both samples, followed by comparing the proportions of structures from each sample in each cluster. No adjustment was made for multiple comparisons. A comparison of two NPC structures considered only the beads representing Nups with a single copy per C2- symmetry unit and the Nic96 complex (including all Nups in the inner, outer and membrane rings, but excluding Nup100, Nup116, Nup145N, Nup1, Nup60, Gle1, Nup42, Mlp1 and Mlp2), to avoid the combinatorial explosion in identification of topologically equivalent Nup copies. The sampling precision is defined as the largest root-mean-square deviation (r.m.s.d.) between a pair of NPC structures within any cluster, in the finest clustering for which each sample contributes structures proportionally to its size (considering both the significance and magnitude of the difference) and for which a sufficient proportion of all structures occur in sufficiently large clusters. The sampling precision for our NPC structure is 9 Å (Extended Data Fig. 1e).

Threshold-based clustering (123) results in a single dominant cluster containing 80.3% of the good-scoring structures (Extended Data Fig. 1e, f) with a root-meansquare fluctuation (r.m.s.f.) of 9 Å (cluster precision). The remaining 19.7% of the structures are similar to those in the dominant cluster; the largest r.m.s.d. value from a structure in the dominant cluster is 17 Å (the mean and standard deviation of the r.m.s.d. values are 13.3 and 1.3Å, respectively). Therefore, there is effectively a single good-scoring solution, at the structure precision of 9Å (equal to the cluster precision). The sampling precision of 9Å (r.m.s.d.) is sufficiently high for computing a structure at 9 Å precision (r.m.s.f.; r.m.s.d. is approximately $\sqrt{2}$ × r.m.s.f. (124)). For the remainder of our analysis, we use only the structures in the dominant cluster.

The fourth convergence test relied on a comparison of two localization probability density maps for each Nup, obtained for dominant cluster structures in samples 1 and 2. A localization probability density map defines the probability of any voxel (here, 6×6×6 $Å^3$) being occupied by a specific protein in a set of structure densities, which in turn are obtained by convolving superposed structures with a Gaussian kernel (here, with a standard deviation of 5.4Å). The average cross-correlation coefficient between the two maps for each Nup is 0.90, indicating that the positions of most Nups in the two samples are nearly identical at the structure precision of 9Å.

In conclusion, all four sampling tests indicate that the sampling was exhaustive at 9Å precision (Supplementary Table 3). The caveat is that passing these tests is necessary but not sufficient evidence of thorough sampling; a positive outcome of the tests may be misleading if, for example, the landscape contains only a narrow— and thus difficult to find—pathway to the pronounced minimum corresponding to the correct

structure. Moreover, our sampling was not completely stochastic because it proceeded in two steps, the first of which prepared the starting configuration for the second step. As a result, the actual structure precision might be worse (125–128) than the estimated 9Å.

Clustering and structure precision

An ensemble of good-scoring structures needs to be analyzed in terms of the precision of its structural features (3, 8, 9). The precision of a component position can be quantified by its variation in an ensemble of superposed good-scoring structures. It can also be visualized by the localization probability density for each of the components of the NPC structure. As described above, integrative structure determination of the NPC resulted in effectively a single good-scoring solution, at the precision of ~9Å. This precision is sufficiently high to pinpoint the locations and orientations of the constituent Nups, demonstrating the quality of the input data, including the chemical crosslinks (**Fig. 8.2** and Extended Data Fig. 7a–c) and the cryo-ET density map (**Fig. 8.3** and Extended Data Fig. 8).

Fit to input information

An accurate structure needs to satisfy the input information used to compute it. Because the sampling was exhaustive at ~9Å precision, overfitting is not a problem at this precision; all structures at this precision that are consistent with the data are provided in the ensemble.

The dominant cluster satisfies 90% of the DSS cross-links (Extended Data Fig. 7a–c and Supplementary Tables 1, 4); a cross-link restraint is satisfied by a cluster of structures if the corresponding Cα–Cα distance in any of the structures in the cluster (considering restraint ambiguity) is < 35Å (Extended Data Fig. 7a–c; shown in blue).

Therefore, the dominant cluster essentially satisfies the cross-linking data within its uncertainty (the false detection rate is approximately 5% to 10% (129, 130)). Most of the cross-link violations are small, and can be rationalized by local structural fluctuations, coarse-grained representations of some Nup domains, and/or finite structural sampling, as shown in Extended Data Fig. 7a (a histogram presenting the distribution of the cross-linked Cα–Cα distances).

The localization probability densities for the dominant cluster overlap well with the cryo-ET density map, with the cross-correlation coefficient of 0.92 (**Fig. 8.3**, Extended Data Fig. 8 and Supplementary Table 4). Additional density is present in the cryo-ET map for the Nup82 complex (cytoplasm) and basket attachment sites (nucleoplasm). This density may arise from local flexibility of these modules or may be due to the presence of cargo or transport factors associated with the NPC (**Fig. 8.1b, c** and Extended Data Fig. 3c). For visualization, the localization probability densities are typically smoothed and contoured at the threshold that results in approximately twice the protein volume estimated from its sequence (**Fig. 8.4**).

The remainder of the restraints are harmonic, with a specified standard deviation. The dominant cluster generally satisfied at least 95% of restraints of each type (Supplementary Table 4); a restraint is satisfied by a cluster of structures if the restrained distance in any structure in the cluster (considering restraint ambiguity) is violated by less than 3 standard deviations, specified for the restraint. Most of the violations are small, and can be rationalized by local structural fluctuations, coarse-grained representations of some Nup domains and/or finite structural sampling.

Satisfaction of data and considerations that were not used to compute structures

The most direct test of a modelled structure is by comparing it to the data that were not used to compute it (a generalization of cross-validation).

First, our current NPC structure is consistent with our previously published data and topological map (3, 8) (Extended Data Fig. 7d, e). Our current structure satisfies all 82 composites determined by affinity purification and overlay assays (3, 8), even though these were not used in this calculation. For example, Pom152, Pom34, Ndc1, Nup157 and Nup170 are connected with each other (left panel in Extended Data Fig. 7e), consistent with the composites determined in a previous publication using the affinity purification data (3, 8) (right panel in Extended Data Fig. 7e). Moreover, the position of each Nup in the current structure is generally similar to that in the previous topological map (3, 8), although the current structure is determined at a precision that is an order of magnitude higher than in the previous map (Extended Data Fig. 7d).

Second, the atomic structures of eight Nups are consistent with the corresponding SAXS profiles for their constructs (Extended Data Fig. 7f and Supplementary Tables 2, 6), as discussed in 'Stage 2: representing subunits and translating data into spatial restraints'. For example, the SAXS profile calculated from the atomic structure of $Pom152^{718–1148}$ (red curve in Extended Data Fig. 7f) using FoXS (111) is well matched (χ=1.48) to the corresponding experimental SAXS profile (23) (black dots in Extended Data Fig. 7f; n=20 exposures). For visualization purposes, the $Pom152^{718–1148}$ structure (represented as a ribbon) is shown along with the best fit of the ab initio shape (represented as a transparent envelope) computed from the experimental SAXS profile, in Extended Data Fig. 7f.

Third, the structures of the Nic96 complex (composed of Nic96, Nsp1, Nup49 and Nup57) in the dominant cluster can be projected well on most of the 2D class averages obtained for the natively isolated complex (Extended Data Fig. 7g; see 'Negative-stain electron microscopy of the native Nic96 complex'). More specifically, the electron microscopy 2D validation fits the structure of the Nic96 complex in the whole NPC context to the electron microscopy class averages of the Nic96 complex, and computes a score that quantifies the match. The computation proceeds in three stages: (1) generation of alternative model projections, (2) alignment of the class average and each model projection, and (3) calculation of the fitting score for each projection, as follows. First, 1,000 uniformly distributed projections of the low-pass-filtered structure of the NPC on the sphere (stage 1) were generated. Second, each projection was optimally aligned to each of the class averages in Fourier space (stage 2). Finally, a score corresponding to the cross-correlation coefficient was computed (stage 3). For example, the experimental class averages were satisfied by the structure with cross-correlation coefficients of 0.85 and 0.80, respectively (Extended Data Fig. 7g).

Fourth, the structure was also validated by its comparison to the core scaffold maps of the Homo sapiens NPC, which are based primarily on electron microscopy density maps (6, 7, 17, 29) (Extended Data Fig. 12). Overall, the inner-ring architecture is similar in both yeast and vertebrates, consistent with it being the most conserved part of the NPC (39).

Finally, the structure allows us to rationalize the functional fitness (**Fig. 8.5**), the transport through the NPC (**Fig. 8.6** and Extended Data Fig. 11) and the evolution

(Extended Data Fig. 10), therefore increasing our confidence in the structure compared to not being able to rationalize these aspects of the NPC (8, 131, 132).

Brownian dynamics simulation of FG repeats and NTFs

Distributions of the FG repeats and NTFs were computed by Brownian dynamics simulation (133), using our previously published protocol (34) implemented in IMP (90) version 2.6. The simulated system included the static NPC ring determined in this study, the pore membrane, disordered and flexible FG-repeat domains as well as freely diffusing NTFs and inert macromolecules, all enclosed within a bounding box of 2,000×2,000×2,000Å^3.

The pore membrane was represented as a 250Å slab with a cylindrical pore of radius 375 Å that contains the static NPC ring (this pore membrane representation is simplified compared to the toroidal pore used for solving the structure of the static NPC ring, for reasons of computational efficiency). Each of the FG-repeat domains was represented as a flexible string of beads; a bead had a radius of 6Å and encompassed 20 residues to achieve a compromise between computational efficiency and accuracy (34, 110, 134–136). Consecutive beads were restrained by a bond with an equilibrium length of 18Å and a constant force of 1.0 kcal per mol per Å, approximating the spring-like nature of flexible polymers (137) in general and FG-repeat domains (110, 134–136, 138–140) in particular. The freely diffusing molecules included 1,600 NTFs and 1,600 inert macromolecules (0.33 mM each), each consisting of 8 subgroups of 200 macromolecules, ranging in radius from 4 to 28Å in 2Å increments (10 to 75kDa, assuming constant protein density of 1.38 g/cm^2). Excluded volume interactions were applied to pairs of overlapping beads and to beads penetrating the pore membrane or the

bounding box, using a constant repulsive force of 10.0 kcal per mol perÅ. The potential binding energy between a binding site on an FG motif and a binding site on an NTR was modelled by an anisotropic harmonic potential dependent on the distance and orientation between the two sites that reproduces measured apparent dissociation constants in our simulations (B.R. et al., manuscript in preparation).

The Brownian dynamics of the entire system were simulated at 297.15K for 40 microseconds with a time step of 1,047 femtoseconds, independently 400 times; the first 10 microseconds of each trajectory were considered equilibration time and ignored in subsequent analysis. The distributions of the FG Nup and NTR positions were then computed from the total of 12 milliseconds of simulations over a cubic grid with a voxel size of 10×10×10$Å^3$, at time intervals of 9.5 picoseconds, from all 400 trajectories; these distributions were averaged by relying on the C8-symmetry of the NPC.

Code availability.

Files containing integrative structure modelling scripts, as well as the input data and output results are available at http://salilab.org/npc2018. The source code for calibrated imaging is available at https://github.com/jayunruh/ Jay_Plugins.

Data availability.

Source Data for Fig. 1a are provided as an excel file. Original data underlying the image calibration data (source data for **Fig. 8.1** and Extended Data Fig. 3) can be accessed from the Stowers Original Data Repository at http://www.stowers.org/publications/LIBPB-1267.

Raw data for the chemical cross-links (source data for **Fig. 8.2** and Supplementary Table 1) are available via the Zenodo data repository, at http://doi.org/10.5281/zenodo.1149746.

The cryo-ET density map (source data for **Fig. 8.3**) is deposited in the Electron Microscopy Data Bank (EMDB) with the accession code EMD-7321.

The cryo-ET raw data (120 tilt series; source data for Extended Data Figs 4–6) are deposited in the Electron Microscopy Public Image Archive (EMPIAR) (https:// www.ebi.ac.uk/pdbe/emdb/empiar/) with the accession code EMPIAR-10155 (see Supplementary Power Point Presentation slides 1, 2 as examples).

The integrative NPC structure (source data for Fig. 4) is deposited in the nascent public Protein Data Bank (PDB) repository, PDB-dev (https://pdb-dev.wwpdb. org/), under the accession codes PDBDEV_00000010, PDBDEV_00000011 and PDBDEV_00000012.

Source data for Extended Data Fig. 2b are provided in the Supplementary Information (Supplementary Fig. 1). SAXS data (source data for Extended Data Fig. 7f) are deposited in the Small Angle Scattering Biological Data Bank (SASBDB; https://www.sasbdb.org/), under the accession codes SASDBV9, SASDBW9, SASDBX9, SASDBY9 and SASDBZ9. In addition, all SAXS data (Supplementary Table 6) are provided as source data with the article.

Raw data for the negative-stain electron microscopy of the native Nic96 complex (source data for Extended Data Fig. 7g) are deposited in the Electron Microscopy Public Image Archive (EMPIAR) (https://www.ebi.ac.uk/pdbe/emdb/empiar/) with the accession code EMPIAR-10162. All other data are available from the corresponding author upon reasonable request.

Supplementary Information

Supplementary Information is available in the online version of the paper.

Acknowledgements

We thank B. Webb (UCSF) for help with the Integrative Modelling Platform, the Rockefeller University Outreach Program for support for A.S.C., the NYULMC OCS Microscopy Core, K. Uryu and the EMRC Resource Center (Rockefeller University) for assistance with negative-stain electron microscopy, F. Alber, M. C. Field, N. Ketaren, S. Obado, R. Hayama and D. Simon for feedback and critical reading of the manuscript, and L. Herlands for support and encouragement. The work was supported by a NSF GRF 1650113 (I.E.C.), a NSF grant CHE-1531823 (M.F.J.), the SIMR (J.L.G.), NIH grants R01 GM080477 (J.L.G.), U54 GM103511 (B.T.C., A.S., J.D.A. and M.P.R.), R01 GM112108 (M.P.R. and J.D.A.), P41 GM109824 (M.P.R., A.S., J.D.A. and B.T.C.), P50 GM076547 (J.D.A.), R01 GM063834 (C.W.A.), R01 GM080139 (S.J.L.), P41 GM103314 (B.T.C.), R01 GM083960 (A.S.) and U54 DK107981 (M.P.R. and J.D.A.). We are grateful for the support provided by G. Blobel, who inspired the work presented here.

Author Contributions

The order of first co-authors was determined through a random selection process. I.N., J.F.-M., A.S.C., R.W. and M.P.R. performed the affinity purifications; W.Z., J.F.-M., R.W., R.M., E.Y.J., M.P.R. and B.T.C. performed the quantitative mass spectrometry; M.S., B.D.S., J.R.U. and J.L.G. performed the calibrated imaging; J.A.H.. B.T.C. and M.F.J. performed the charge detection mass spectrometry; Y.S., J.F.-M., R.W., I.N., J.W. and B.T.C. performed the chemical crosslinking with mass spectrometry; C.W.A., S.J.L., I.N., Z.Y. and M.J.d.l.C. performed the cryo-ET; S.J.K. performed the small-angle X-ray

scattering; T.H., J.F.-M. and J.D.A. performed the phenotypic profiling; P.U. and D.L.S. performed the negative-stain electron microscopy; S.J.K., B.R., I.E.C., R.P., I.E., C.H.G. and A.S. performed the integrative structure computations; S.J.L., C.W.A., B.T.C., A.S. and M.P.R. supervised the project; S.J.K., J.F.-M., I.N., Y.S., W.Z., B.R., S.J.L., C.W.A., B.T.C., A.S. and M.P.R. wrote the manuscript.

References

1. C. Ptak, J. D. Aitchison, R. W. Wozniak, The multifunctional nuclear pore complex: a platform for controlling gene expression. *Curr Opin Cell Biol* **28**, 46–53 (2014).
2. V. Nofrini, D. Di Giacomo, C. Mecucci, Nucleoporin genes in human diseases. *Eur J Hum Genet* **24**, 1388–95 (2016).
3. F. Alber, *et al.*, The molecular architecture of the nuclear pore complex. *Nature* **450**, 695–701 (2007).
4. G. J. Stanley, A. Fassati, B. W. Hoogenboom, Biomechanics of the transport barrier in the nuclear pore complex. *Semin Cell Dev Biol* (2017).
5. C. W. Akey, D. S. Goldfarb, Protein import through the nuclear pore complex is a multistep process. *J Cell Biol* **109**, 971–82 (1989).
6. J. Kosinski, *et al.*, Molecular architecture of the inner ring scaffold of the human nuclear pore complex. *Science* **352**, 363–5 (2016).
7. D. H. Lin, *et al.*, Architecture of the symmetric core of the nuclear pore. *Science* **352**, aaf1015 (2016).
8. F. Alber, *et al.*, Determining the architectures of macromolecular assemblies. *Nature* **450**, 683–694 (2007).
9. J. Fernandez-Martinez, *et al.*, Structure and Function of the Nuclear Pore Complex Cytoplasmic mRNA Export Platform. *Cell* **167**, 1215–1228 (2016).
10. Y. Shi, *et al.*, Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. *Mol Cell Proteomics* **13**, 2927–43 (2014).

11. J. A. Briggs, Structural biology in situ–the potential of subtomogram averaging. *Curr Opin Struct Biol* **23**, 261–7 (2013).
12. S. J. Kim, *et al.*, Integrative structure-function mapping of the nucleoporin Nup133 suggests a conserved mechanism for membrane anchoring of the nuclear pore complex. *Mol Cell Proteomics* (2014).
13. J. Fernandez-Martinez, *et al.*, Structure-function mapping of a heptameric module in the nuclear pore complex. *J Cell Biol* **196**, 419–34 (2012).
14. T. Herricks, *et al.*, One-Cell Doubling Evaluation by Living Arrays of Yeast, ODELAY! *G3 (Bethesda)* **7**, 279–288 (2017).
15. J. D. Aitchison, M. P. Rout, M. Marelli, G. Blobel, R. W. Wozniak, Two novel related yeast nucleoporins Nup170p and Nup157p: complementation with the vertebrate homologue Nup155p and functional interactions with the yeast nuclear pore-membrane protein Pom152p. *J Cell Biol* **131**, 1133–48 (1995).
16. J. Fischer, R. Teimer, S. Amlacher, R. Kunze, E. Hurt, Linker Nups connect the nuclear pore complex inner ring with the outer ring and transport channel. *Nat Struct Mol Biol* **22**, 774–81 (2015).
17. A. von Appen, *et al.*, In situ structural analysis of the human nuclear pore complex. *Nature* **526**, 140–3 (2015).
18. M. Marelli, C. P. Lusk, H. Chan, J. D. Aitchison, R. W. Wozniak, A link between the synthesis of nucleoporins and the biogenesis of the nuclear envelope. *J Cell Biol* **153**, 709–24 (2001).
19. B. Vollmer, *et al.*, Dimerization and direct membrane interaction of Nup53 contribute to nuclear pore complex assembly. *EMBO J* **31**, 4072–84 (2012).

20. H. S. Seo, *et al.*, Structural and functional analysis of Nup120 suggests ring formation of the Nup84 complex. *Proc Natl Acad Sci U S A* **106**, 14281–6 (2009).
21. G. Drin, *et al.*, A general amphipathic alpha-helical motif for sensing membrane curvature. *Nat Struct Mol Biol* **14**, 138–46 (2007).
22. N. Meszaros, *et al.*, Nuclear pore basket proteins are tethered to the nuclear envelope and can regulate membrane curvature. *Dev Cell* **33**, 285–98 (2015).
23. P. Upla, *et al.*, Molecular Architecture of the Major Membrane Ring Component of the Nuclear Pore Complex. *Structure* **25**, 434–445 (2017).
24. A. C. Meinema, *et al.*, Long unfolded linkers facilitate membrane protein import through the nuclear pore complex. *Science* **333**, 90–3 (2011).
25. K. E. Knockenhauer, T. U. Schwartz, The Nuclear Pore Complex as a Flexible and Dynamic Gate. *Cell* **164**, 1162–71 (2016).
26. A. W. Folkmann, K. N. Noble, C. N. Cole, S. R. Wente, Dbp5, Gle1-IP6 and Nup159: a working model for mRNP export. *Nucleus* **2**, 540–8 (2011).
27. M. A. Saroufim, *et al.*, The nuclear basket mediates perinuclear mRNA scanning in budding yeast. *J Cell Biol* **211**, 1131–40 (2015).
28. R. A. Meseroll, O. Cohen-Fix, The Malleable Nature of the Budding Yeast Nuclear Envelope: Flares, Fusion, and Fenestrations. *J Cell Physiol* **231**, 2353–60 (2016).
29. M. Eibauer, *et al.*, Structure and gating of the nuclear pore complex. *Nat Commun* **6**, 7532 (2015).
30. A. Paradise, M. K. Levin, G. Korza, J. H. Carson, Significant proportions of nuclear transport proteins with reduced intracellular mobilities resolved by fluorescence correlation spectroscopy. *J Mol Biol* **365**, 50–65 (2007).

31. R. L. Adams, S. R. Wente, Uncovering nuclear pore complexity with innovation. *Cell* **152**, 1218–21 (2013).
32. S. S. Patel, B. J. Belmont, J. M. Sante, M. F. Rexach, Natively unfolded nucleoporins gate protein diffusion across the nuclear pore complex. *Cell* **129**, 83–96 (2007).
33. R. L. Adams, L. J. Terry, S. R. Wente, Nucleoporin FG domains facilitate mRNP remodeling at the cytoplasmic face of the nuclear pore complex. *Genetics* **197**, 1213–24 (2014).
34. B. L. Timney, *et al.*, Simple rules for passive diffusion through the nuclear pore complex. *J Cell Biol* **215**, 57–76 (2016).
35. J. Yamada, *et al.*, A bimodal distribution of two distinct categories of intrinsically disordered structures with separate functions in FG nucleoporins. *Mol Cell Proteomics* **9**, 2205–24 (2010).
36. D. Devos, *et al.*, Components of coated vesicles and nuclear pore complexes share a common molecular architecture. *PLoS Biol* **2**, e380 (2004).
37. M. Faini, R. Beck, F. T. Wieland, J. A. Briggs, Vesicle coats: structure, function, and general principles of assembly. *Trends Cell Biol* **23**, 279–88 (2013).
38. M. P. Rout, M. C. Field, The Evolution of Organellar Coat Complexes and Organization of the Eukaryotic Cell. *Annu Rev Biochem* **86**, 637–657 (2017).
39. S. O. Obado, *et al.*, Interactome Mapping Reveals the Evolutionary History of the Nuclear Pore Complex. *Plos Biology* **14** (2016).
40. M. Iwamoto, *et al.*, Compositionally distinct nuclear pore complexes of functionally distinct dimorphic nuclei in the ciliate Tetrahymena. *J Cell Sci* **130**, 1822–1834 (2017).

41. S. O. Obado, M. C. Field, M. P. Rout, Comparative interactomics provides evidence for functional specialization of the nuclear pore complex. *Nucleus* **8**, 340–352 (2017).
42. K. H. Bui, *et al.*, Integrated structural analysis of the human nuclear pore complex scaffold. *Cell* **155**, 1233–43 (2013).
43. E. W. Debler, *et al.*, A fence-like coat for the nuclear pore membrane. *Mol Cell* **32**, 815–26 (2008).
44. M. P. Rout, G. Blobel, Isolation of the yeast nuclear pore complex. *J Cell Biol* **123**, 771–83 (1993).
45. M. P. LaCava J. ;. Fernandez-Martinez, J. ;. Hakhverdyan, Z. ;. Rout, "Protein Complex Purification by Affinity Capture" in *Budding Yeast: A Laboratory Manual*, S. Andrews B. ;. Boone, C. ;. Davis, T. N. ;. Fields, Ed. (Cold Spring Harbor Laboratory Press, 2016), pp. 383–400.
46. J. LaCava, J. Fernandez-Martinez, Z. Hakhverdyan, M. P. Rout, Optimized Affinity Capture of Yeast Protein Complexes. *Cold Spring Harb Protoc* **2016**, pdb prot087932 (2016).
47. J. LaCava, J. Fernandez-Martinez, M. P. Rout, Native Elution of Yeast Protein Complexes Obtained by Affinity Capture. *Cold Spring Harb Protoc* **2016**, pdb prot087940 (2016).
48. M. Oeffinger, *et al.*, Comprehensive analysis of diverse ribonucleoprotein complexes. *Nat Methods* **4**, 951–6 (2007).
49. Z. Hakhverdyan, *et al.*, Rapid, optimized interactomic screening. *Nat Methods* **12**, 553–60 (2015).

50. R. J. Beynon, M. K. Doherty, J. M. Pratt, S. J. Gaskell, Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nat Methods* **2**, 587–9 (2005).

51. M. Shivaraju, *et al.*, Cell-cycle-coupled structural oscillation of centromeric nucleosomes in yeast. *Cell* **150**, 304–16 (2012).

52. D. Z. Keifer, T. Motwani, C. M. Teschke, M. F. Jarrold, Measurement of the accurate mass of a 50 MDa infectious virus. *Rapid Commun Mass Spectrom* **30**, 1957–62 (2016).

53. J. M. Pratt, *et al.*, Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes. *Nat Protoc* **1**, 1029–43 (2006).

54. K. Kito, K. Ota, T. Fujita, T. Ito, A synthetic protein approach toward accurate mass spectrometric quantification of component stoichiometry of multiprotein complexes. *J Proteome Res* **6**, 792–800 (2007).

55. C. Ding, *et al.*, Quantitative analysis of cohesin complex stoichiometry and SMC3 modification-dependent protein interactions. *J Proteome Res* **10**, 3652–9 (2011).

56. M. P. Rout, J. V. Kilmartin, Components of the yeast spindle and spindle pole body. *J Cell Biol* **111**, 1913–27 (1990).

57. C. Strambio-de-Castillia, G. Blobel, M. P. Rout, Isolation and characterization of nuclear envelopes from the yeast Saccharomyces. *J Cell Biol* **131**, 19–31 (1995).

58. M. P. Rout, C. Strambio-De-Castillia, “Isolation of yeast nuclear pore complexes and nuclear envelopes” in *Cell Biology: A Laboratory Handbook 2*, J. E. Celis, Ed. (Academic Press, 1998), pp. 143–151.

59. M. Cadene, B. T. Chait, A robust, detergent-friendly method for mass spectrometric analysis of integral membrane proteins. *Anal Chem* **72**, 5655–8 (2000).

60. D. Fenyo, *et al.*, MALDI sample preparation: the ultra thin layer method. *J Vis Exp*, 192 (2007).

61. H. I. Field, D. Fenyo, R. C. Beavis, RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics* **2**, 36–47 (2002).

62. J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367–72 (2008).

63. B. Schwanhausser, *et al.*, Global quantification of mammalian gene expression control. *Nature* **473**, 337–42 (2011).

64. N. C. Contino, E. E. Pierson, D. Z. Keifer, M. F. Jarrold, Charge detection mass spectrometry with resolved charge states. *J Am Soc Mass Spectrom* **24**, 101–8 (2013).

65. Y. Shi, *et al.*, A strategy for dissecting the architectures of native macromolecular assemblies. *Nat Methods* **12**, 1135–8 (2015).

66. A. Leitner, *et al.*, Expanding the chemical cross-linking toolbox by the use of multiple proteases and enrichment by size exclusion chromatography. *Mol Cell Proteomics* **11**, M111 014126 (2012).

67. B. Yang, *et al.*, Identification of cross-linked peptides from complex samples. *Nat Methods* **9**, 904–6 (2012).

68. M. A. Cevher, *et al.*, Reconstitution of active human core Mediator complex reveals a critical role of the MED14 subunit. *Nat Struct Mol Biol* **21**, 1028–34 (2014).

69. J. Sun, *et al.*, The architecture of a eukaryotic replisome. *Nat Struct Mol Biol* **22**, 976–82 (2015).

70. S. Q. Zheng, *et al.*, UCSF tomography: an integrated software suite for real-time electron microscopic tomographic data collection, alignment, and reconstruction. *J Struct Biol* **157**, 138–47 (2007).

71. J. R. Kremer, D. N. Mastronarde, J. R. McIntosh, Computer visualization of three-dimensional image data using IMOD. *J Struct Biol* **116**, 71–6 (1996).

72. J. G. Galaz-Montoya, J. Flanagan, M. F. Schmid, S. J. Ludtke, Single particle tomography in EMAN2. *J Struct Biol* **190**, 279–90 (2015).

73. J. G. Galaz-Montoya, *et al.*, Alignment algorithms and per-particle CTF correction for single particle cryo-electron tomography. *J Struct Biol* **194**, 383–94 (2016).

74. Q. Yang, M. P. Rout, C. W. Akey, Three-dimensional architecture of the isolated yeast nuclear pore complex: functional and evolutionary implications. *Mol Cell* **1**, 223–34 (1998).

75. M. Beck, *et al.*, Nuclear pore complex structure and dynamics revealed by cryoelectron tomography. *Science* **306**, 1387–90 (2004).

76. A. von Appen, M. Beck, Structure Determination of the Nuclear Pore Complex with Three-Dimensional Cryo electron Microscopy. *J Mol Biol* **428**, 2001–10 (2016).

77. S. J. Ludtke, Single-Particle Refinement and Variability Analysis in EMAN2.1. *Methods Enzymol* **579**, 159–89 (2016).

78. J. Iwanczyk, *et al.*, Structure of the Blm10-20 S proteasome complex by cryo-electron microscopy. Insights into the mechanism of activation of mature yeast proteasomes. *J Mol Biol* **363**, 648–59 (2006).

79. N. Elad, *et al.*, The dynamic conformational landscape of gamma-secretase. *J Cell Sci* **128**, 589–98 (2015).

80. A. Kucukelbir, F. J. Sigworth, H. D. Tagare, Quantifying the local resolution of cryo-EM density maps. *Nat Methods* **11**, 63–5 (2014).

81. T. A. Bharat, S. H. Scheres, Resolving macromolecular structures from electron cryo-tomography data using subtomogram averaging in RELION. *Nat Protoc* **11**, 2054–65 (2016).

82. T. A. Bharat, C. J. Russo, J. Lowe, L. A. Passmore, S. H. Scheres, Advances in Single-Particle Electron Cryomicroscopy Structure Determination applied to Sub-tomogram Averaging. *Structure* **23**, 1743–53 (2015).

83. P. Sampathkumar, *et al.*, Atomic structure of the nuclear pore complex targeting domain of a Nup116 homologue from the yeast, Candida glabrata. *Proteins* **80**, 2110–6 (2012).

84. P. Sampathkumar, *et al.*, Structure of the C-terminal domain of Saccharomyces cerevisiae Nup133, a component of the nuclear pore complex. *Proteins* **79**, 1672–7 (2011).

85. P. Sampathkumar, *et al.*, Structures of the autoproteolytic domain from the Saccharomyces cerevisiae nuclear pore complex component, Nup145. *Proteins* **78**, 1992–8 (2010).

86. P. Sampathkumar, *et al.*, Structure, dynamics, evolution, and function of a major scaffold component in the nuclear pore complex. *Structure* **21**, 560–71 (2013).

87. A. D. Edelstein, *et al.*, Advanced methods of microscope control using μManager software. *J Biol Methods* **1** (2014).

88. S. J. Ludtke, P. R. Baldwin, W. Chiu, EMAN: semiautomated software for high-resolution single-particle reconstructions. *J Struct Biol* **128**, 82–97 (1999).

89. Z. Yang, J. Fang, J. Chittuluru, F. J. Asturias, P. A. Penczek, Iterative stable alignment and clustering of 2D transmission electron microscope images. *Structure* **20**, 237–47 (2012).

90. D. Russel, *et al.*, Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. *PLOS Biology* **10**, e1001244 (2012).

91. D. Schneidman-Duhovny, R. Pellarin, A. Sali, Uncertainty in integrative structural modeling. *Current Opinion in Structural Biology* **28**, 96–104 (2014).

92. A. Sali, *et al.*, Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure* **23**, 1156–67 (2015).

93. J. LoPiccolo, *et al.*, Assembly and Molecular Architecture of the Phosphoinositide 3-Kinase p85alpha Homodimer. *J Biol Chem* **290**, 30390–405 (2015).

94. J. Luo, *et al.*, Architecture of the human and yeast general transcription and DNA repair factor TFIIH. *Mol Cell* **59**, 794–806 (2015).

95. P. Robinson, *et al.*, Molecular architecture of the yeast Mediator complex. *eLife* **4**, e08719 (2015).

96. B. Webb, *et al.*, Modeling of proteins and their assemblies with the Integrative Modeling Platform. *Methods Mol Biol* **1091**, 277–95 (2014).

97. J. M. Cherry, *et al.*, SGD: Saccharomyces Genome Database. *Nucleic Acids Res* **26**, 73–9 (1998).

98. R. Gautier, D. Douguet, B. Antonny, G. Drin, HELIQUEST: a web server to screen sequences with specific alpha-helical properties. *Bioinformatics* **24**, 2101–2 (2008).

99. M. Niepel, *et al.*, The nuclear basket proteins Mlp1p and Mlp2p are part of a dynamic interactome including Esc1p and the proteasome. *Mol Biol Cell* **24**, 3920–38 (2013).

100. A. Sali, T. L. Blundell, Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779–815 (1993).

101. J. Soding, A. Biegert, A. N. Lupas, The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* **33**, W244-8 (2005).

102. J. Soding, Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951–60 (2005).

103. D. W. Buchan, F. Minneci, T. C. Nugent, K. Bryson, D. T. Jones, Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res* **41**, W349-57 (2013).

104. D. T. Jones, Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**, 195–202 (1999).

105. J. J. Ward, L. J. McGuffin, K. Bryson, B. F. Buxton, D. T. Jones, The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**, 2138–9 (2004).

106. R. L. Marsden, L. J. McGuffin, D. T. Jones, Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci* **11**, 2814–24 (2002).

107. A. Lupas, M. Van Dyke, J. Stock, Predicting coiled coils from protein sequences. *Science* **252**, 1162–4 (1991).

108. J. Trigg, K. Gutwin, A. E. Keating, B. Berger, Multicoil2: predicting coiled coils and their oligomerization states from sequence in the twilight zone. *PLoS One* **6**, e23519 (2011).

109. R. Algret, *et al.*, Molecular architecture and function of the SEA complex, a modulator of the TORC1 pathway. *Mol Cell Proteomics* **13**, 2855–70 (2014).

110. B. Raveh, *et al.*, Slide-and-exchange mechanism for rapid and selective transport through the nuclear pore complex. *Proc Natl Acad Sci U S A* **113**, E2489-97 (2016).

111. D. Schneidman-Duhovny, M. Hammel, A. Sali, FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res* **38**, W540-4 (2010).

112. D. Schneidman-Duhovny, S. J. Kim, A. Sali, Integrative structural modeling with small angle X-ray scattering profiles. *BMC Struct Biol* **12**, 17 (2012).

113. T. Kawabata, Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a gaussian mixture model. *Biophys J* **95**, 4643–58 (2008).

114. S. Jonic, *et al.*, Denoising of high-resolution single-particle electron-microscopy density maps by their approximation using three-dimensional Gaussian functions. *J Struct Biol* **194**, 423–33 (2016).

115. S. Hanot, *et al.*, Multi-scale Bayesian modeling of cryo-electron microscopy density maps. *https://doi.org/10.1101/113951* **Preprint** (2017).

116. M. P. Rout, *et al.*, The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J Cell Biol.* **148**, 635–51 (2000).

117. M. Y. Shen, A. Sali, Statistical potential for assessment and prediction of protein structures. *Protein Sci* **15**, 2507–24 (2006).

118. F. Campelo, M. M. Kozlov, Sensing membrane stresses by protein insertions. *PLoS Comput Biol* **10**, e1003556 (2014).

119. S. Viswanath, I. E. Chemmama, P. Cimermancic, A. Sali, Assessing Exhaustiveness of Stochastic Sampling for Integrative Modeling of Macromolecular Structures. *Biophysical Journal* **113**, 2344–2353 (2017).

120. S. Siegel, Nonparametric statistics for the behavioral sciences (McGraw-Hill, 1956).

121. D. McCarroll, *Simple Statistical Tests for Geography* (CRC Press, 2016).

122. J. H. McDonald, *Handbook of biological statistics* (Sparky House Publishing Baltimore, MD, 2009).

123. X. Daura, *et al.*, Peptide folding: When simulation meets experiment. *Angewandte Chemie-International Edition* **38**, 236–240 (1999).

124. A. Kuzmanic, B. Zagrovic, Determination of ensemble-average pairwise root mean-square deviation from experimental B-factors. *Biophysical journal* **98**, 861–871 (2010).

125. R. J. Read, *et al.*, A new generation of crystallographic validation tools for the protein data bank. *Structure* **19**, 1395–412 (2011).

126. G. T. Montelione, *et al.*, Recommendations of the wwPDB NMR Validation Task Force. *Structure* **21**, 1563–70 (2013).

127. R. Henderson, *et al.*, Outcome of the first electron microscopy validation task force meeting. *Structure* **20**, 205–14 (2012).

128. J. Trewhella, *et al.*, Report of the wwPDB Small-Angle Scattering Task Force: data requirements for biomolecular modeling and the PDB. *Structure* **21**, 875–81 (2013).

129. A. Leitner, *et al.*, Chemical cross-linking/mass spectrometry targeting acidic residues in proteins and protein complexes. *Proc Natl Acad Sci U S A* **111**, 9455–60 (2014).

130. J. P. Erzberger, *et al.*, Molecular architecture of the 40SeIF1eIF3 translation initiation complex. *Cell* **158**, 1123–35 (2014).

131. F. Alber, F. Forster, D. Korkin, M. Topf, A. Sali, Integrating diverse data for structure determination of macromolecular assemblies. *Annu Rev Biochem* **77**, 443–77 (2008).

132. F. Alber, B. T. Chait, M. P. Rout, A. Sali, “Integrative Structure Determination of Protein Assemblies by Satisfaction of Spatial Restraints” in *Protein-Protein Interactions and Networks: Identification, Characterization and Prediction.*, A. Panchenko, T. Przytycka, Eds. (Springer-Verlag, 2008), pp. 99–114.

133. D. L. Ermak, J. A. Mccammon, Brownian Dynamics with Hydrodynamic Interactions. *Journal of Chemical Physics* **69**, 1352–1360 (1978).

134. L. E. Hough, *et al.*, The molecular mechanism of nuclear transport revealed by atomic-scale measurements. *Elife* **4** (2015).

135. S. Milles, *et al.*, Plasticity of an ultrafast interaction between nucleoporins and nuclear transport receptors. *Cell* **163**, 734–45 (2015).

136. Y. Sakiyama, A. Mazur, L. E. Kapinos, R. Y. Lim, Spatiotemporal dynamics of the nuclear pore complex transport barrier resolved by high-speed atomic force microscopy. *Nat Nanotechnol* **11**, 719–23 (2016).

137. van der Maarel, *Introduction to Biopolymer Physics* (World Scientic, 2008).

138. D. P. Denning, S. S. Patel, V. Uversky, A. L. Fink, M. Rexach, Disorder in the nuclear pore complex: the FG repeat regions of nucleoporins are natively unfolded. *Proc Natl Acad Sci U S A* **100**, 2450–5 (2003).

139. E. A. Lemke, The Multiple Faces of Disordered Nucleoporins. *J Mol Biol* **428**, 2011–24 (2016).

140. R. Y. Lim, *et al.*, Flexible phenylalanine-glycine nucleoporins as entropic barriers to nucleocytoplasmic transport. *Proc Natl Acad Sci U S A* **103**, 9512–7 (2006).

Chapter IX - Reconstruction of 3D structures of MET antibodies from electron microscopy 2D class averages

Contributing authors

Qi Chen[1,*], Michal Vieth[1,*], David E. Timm[1], Christine Humblet[1], Dina Schneidman-Duhovny[2], Ilan E. Chemmama[2], Andrej Sali[2], Wei Zheng[3], Jirong Lu[1], Ling Liu[1]

[1]Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, Indiana, United States of America

[2]Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, University of California, San Francisco, San Francisco, California 94158, USA.

[*]Contacts: chen_qi_qc@lilly.com (QC); vieth_michal@lilly.com (MV)

Abstract

Dynamics of three MET antibody constructs (IgG1, IgG2, and IgG4) and the IgG4-MET antigen complex was investigated by creating their atomic models with an integrative experimental and computational approach. In particular, we used two-dimensional (2D) Electron Microscopy (EM) images, image class averaging, homology modeling, Rapidly exploring Random Tree (RRT) structure sampling, and fitting of models to images, to find the relative orientations of antibody domains that are consistent with the EM images. We revealed that the conformational preferences of the constructs depend on the extent of the hinge flexibility. We also quantified how the MET antigen impacts on the

conformational dynamics of IgG4. These observations allow to create testable hypothesis to investigate MET biology. Our protocol may also help describe structural diversity of other antigen systems at approximately 5 Å precision, as quantified by Root-Mean-Square Deviation (RMSD) among good-scoring models.

Introduction

Antibodies are among the most specific biomedicines. They are important therapeutic agents, both as biomolecular drugs and as delivery vehicles of drugs in antibody drug conjugates (1). Antibodies usually contain three domains, i.e., two Fab domains and one Fc domain, connected by two short peptidic hinges (**Fig 9.1**). The 3D atomic structure of each full length antibody exists as an ensemble of multiple conformational states (2), although the three domains are almost always arranged into a Y or T-shaped 3D object as shown in their X-ray crystal structures (3–5). Due to the flexibility of the two hinges, the Cα RMSD between antibody structures could be higher than 30 Å, despite a similar overall arrangement of the three domains and the structural similarity among the individual Fab and Fc domains (**Fig 9.1**). This diverse structural space of antibodies makes the structure determination by X-ray crystallography and the application of structure-based design approaches extremely challenging.

Multiple techniques have been used to study full length antibody structures, including X-ray crystallography that gave the structures of three full length constructs (3–5), 3D Individual Particle Electron Tomography (IPET) (2, 6) and EM imaging (7). The IPET maps at 10–15 Å resolution combined with molecular dynamics simulations demonstrated a vast structural space represented by 120 diverse structure models (2) available to the mouse IgG1 construct. The model construction in the IPET study used a single starting structure from X-ray crystallography (3), allowing flexibility in the hinge region while keeping the individual domains rigid. The antibody structural space resulting from different arrangements of the rigid domains referred to as the “domain conformations” revealed by the IPET study serves as a starting point for our study.

To model the MET domain conformations, we used the EM2D module (8) of the open source Integrative Modeling Package (IMP) (9, 10) to construct the models of three MET isotypes (IgG1, IgG2, IgG4) from the low resolution (~20 Å, see “Stage 3: Scoring domain conformation” of Materials and Methods section for details) 2D class averages of individual particle EM images. We found that for all examined antibody constructs, every good quality 2D class average could be uniquely represented by a single model of domain conformation selected from a diverse conformational ensemble at model precision of 5 Å RMSD. The variability among the generated models that sufficiently satisfy the experimental 2D class averages is quantified by model precision, defined as the largest RMSD value of a model that still satisfies the 2D class average to the best scoring model for that 2D class average (see “Stage 4: Analysis and Assessment of the Ensemble” in “Materials and Methods” section for details). The determination of domain conformations at 5 Å RMSD precision increases our understanding of antibody structural dynamics. Furthermore, it allows us to relate the biological profile of constructs to the inter-domain interactions, location and orientation of complementarity determining regions (CDRs) as well as the overall shape of the antibodies. The methodology presented here can be applied for future exploration of dynamics of antibodies in general.

Our modeling effort focused on MET antibody constructs. MET, the receptor for hepatocyte growth factor (HGF), has been implicated in driving tumor proliferation and metastasis. Given the critical roles of the MET/HGF pathway in tumor growth and development, various groups developed MET blocking antibodies (11–14). However, bivalent anti-MET antibodies that inhibit both HGF-dependent and HGF-independent activation were largely unsuccessful as these constructs tended to have agonistic rather

than antagonistic activity (12–14). The first reported construct with no agonistic activity was LY2875358 (11). LY2875358 is a humanized IgG4 antibody against the MET receptor, currently being evaluated in Phase II clinical trials for non-small cell lung cancer (NSCLC). It has high neutralization and internalization activities, resulting in inhibition of ligand dependent and ligand independent MET pathway inhibition. While LY2875358 does not have agonist activity in the IgG4 format, the IgG1 version of the same MET antibody shows increased agonist activity. This observation suggests that the agonistic activity of the antibody might depend on the IgG isotypes on the top of the differences in the variable regions. Knowledge of the 3D structural differences between different IgG isotypes could help us better understand the above-mentioned functional outcomes.

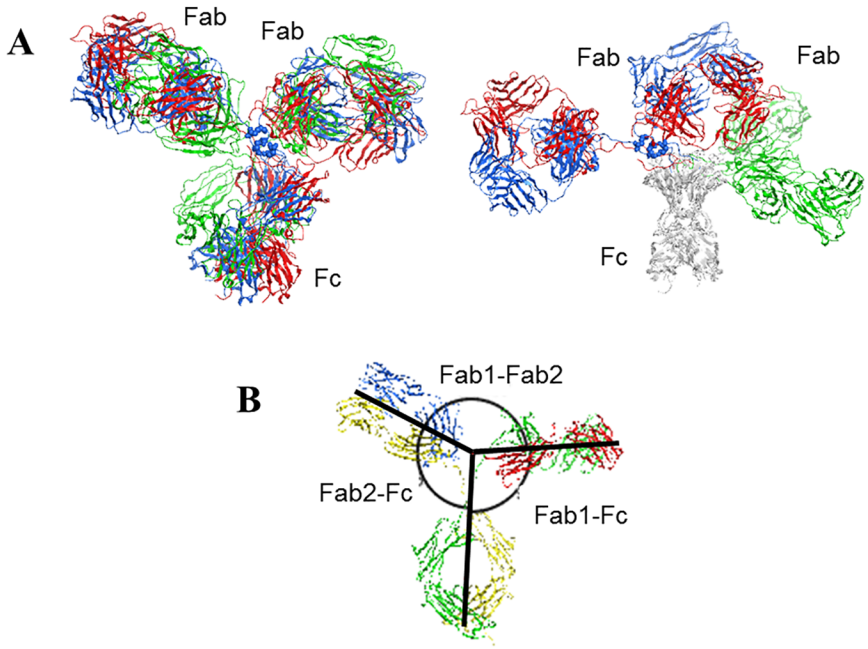

Figure 9.1 | The antibody structure and variability.
(**A**) X-ray crystal structures of three full length antibodies: mouse IgG2 in blue (PDB code 1IGT) (3), human IgG1 in red (1HZH) (4), and mouse IgG1 in green (1IGY) (5), superposed on all aligned Cα atoms from all three domains (left) for the overall shape comparison or only Fc domains (right) to highlight the differences in Fab domains. The pairwise Cα RMSD values range from 20 to 34 Å with an average of 27.6 Å. The pairwise Cα RMSD values of the Fab domains range from 1.1 to 3.9 Å when superposed only on the Fab domain residues, with an average of 2.0 Å. The pairwise Cα RMSD values of the

Fc domains range from 2.2 to 2.4 Å when superposed only on the Fc domain residues. Superposition was done using MOE 2014.09 (15). (**B**) An example to show the definition of the domain angles between every two domains measured from 3D structures as described by Zhang et. al. (2). The lines follow the longest axis of each domain.

Results

Isotype dependence of agonist activity

We have previously reported (11) multiple *in vitro* bioassays to characterize the agonist properties of LY2875358, using HGF and agonist bivalent MET antibody 5D5 as positive controls with Phospho-AKT as the most sensitive agonist assay. Here we established that LY2875358 (IgG4 isotype) induced only a weak and transient phosphorylation of pan-AKT upon binding to MET (**Fig. 9.2A**), and this weak phosphorylation of pan-AKT did not stimulate biologic activity in seven functional MET agonist assays (11). However, the higher levels of AKT phosphorylation induced by MET antibody 5D5 and HGF (**Fig. 9.2A**) correlated well with cell proliferation, mobility and anti-apoptosis in the same functional assays (11). We further compared IgG1, IgG2 and IgG4 MET antibodies in the phospho-AKT assay, showing that IgG1 MET antibody significantly increased phospho-AKT activity levels by more than eight-fold, close to the levels from agonist 5D5 and HGF. In the same assay, the IgG2 isotype displayed a slight increase in phosphorylation of AKT as compared to the IgG4 isotype (**Fig. 9.2B**). Because these antibodies have shown comparable binding affinity to the MET extracellular domain (ECD) by Biacore and have nearly identical sequences except in the hinges (**Fig. 9.2C**), we hypothesized that the hinge flexibility of antibodies of different isotype might contribute to how antibodies engage MET on the cell surface, hence impacting the agonistic activity.

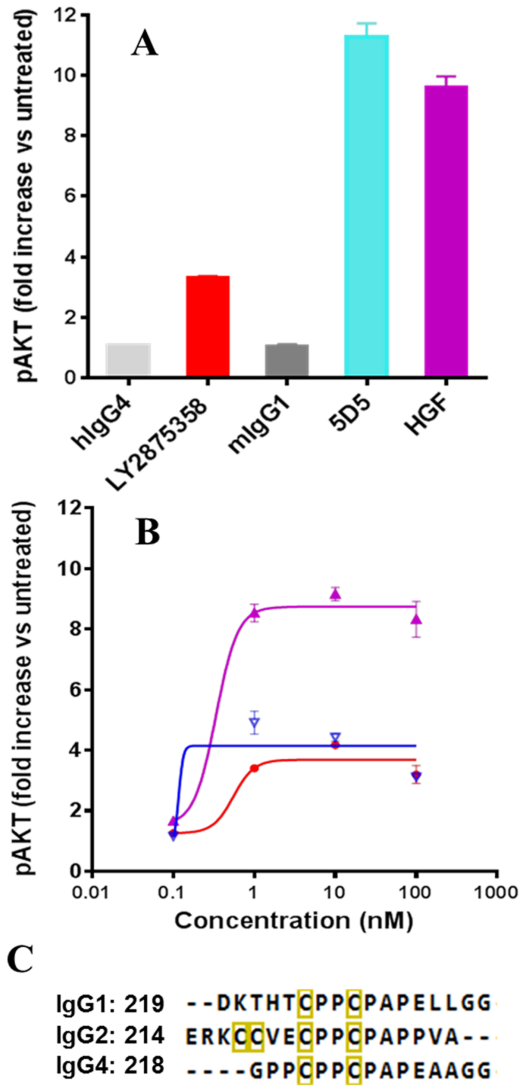

Figure 9.2 | The effect of MET antibody isotypes on pAKT in Caki-1 cells.
(**A**) The humanized IgG4 MET antibody (LY2875358), hIgG4 and mIgG1 induce weak phosphorylation of pan-AKT as compared to the strong phosphorylation of pan-AKT by agonist antibody 5D5 and HGF (11). (**B**) Comparison of IgG1 (purple), IgG2 (blue) and IgG4 MET antibodies (red). (**C**) The sequences in the hinges. The numbering of the first residue is shown in each sequence. The inter-heavy-chain disulfide bonded cysteine residues are indicated in yellow boxes.

Reconstruction of 3D domain conformation models and their relative diversity

Fig. 9.3 shows an example of EM particle images and their 2D class averages for the IgG1 MET antibody construct. We found that 2D class averages of antibody samples often had the Y/T shape, although in some cases two of the "arms" could be very close to each other. Therefore, we excluded the class averages with images that didn't display the Y/T shape. **Fig. 9.4** shows the observed good quality 2D class average, the corresponding best scoring simulated 2D images, together with the ribbon diagrams of corresponding 3D domain conformation models for all four antibody samples. Eight (IgG1), five (IgG2), nine (IgG4) and seven (IgG4-MET antigen complex) domain conformations were obtained with our protocol. The modeled conformations are quite different within each construct, with pairwise RMSD values ranging from 5.7 to 37.0 Å (Table B in S1 File and **Fig. 9.5**).

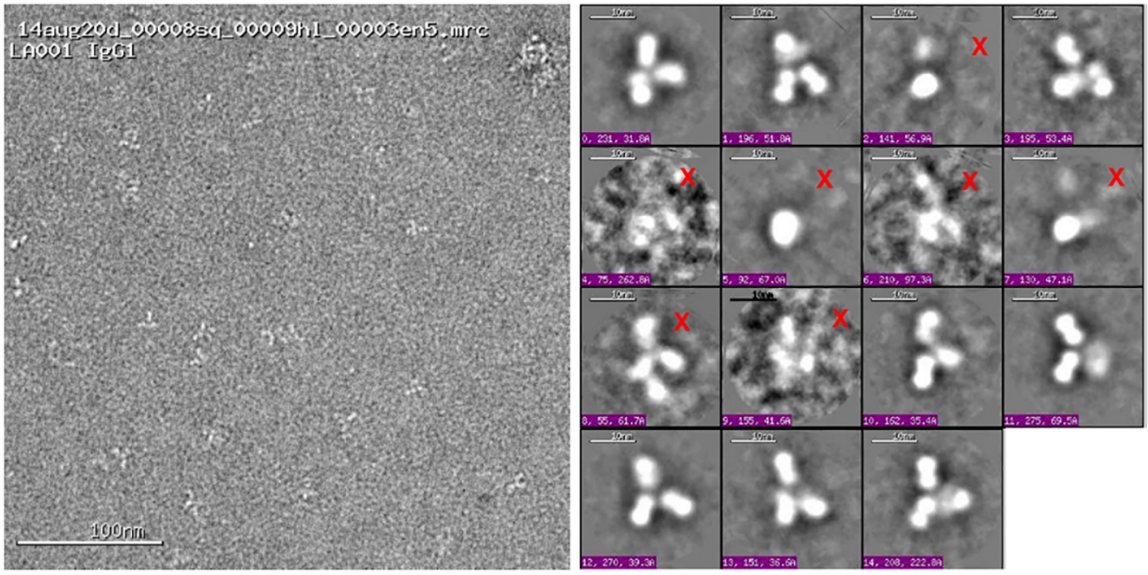

Figure 9.3 | Examples of EM image thumbnails.
An EM micrograph with a number of IgG1 particles (left) and EM 2D class average images (right). Not all 2D class averages contained recognizable Y/T-shaped antibody particles; those marked by a red "X" were not used in modeling.

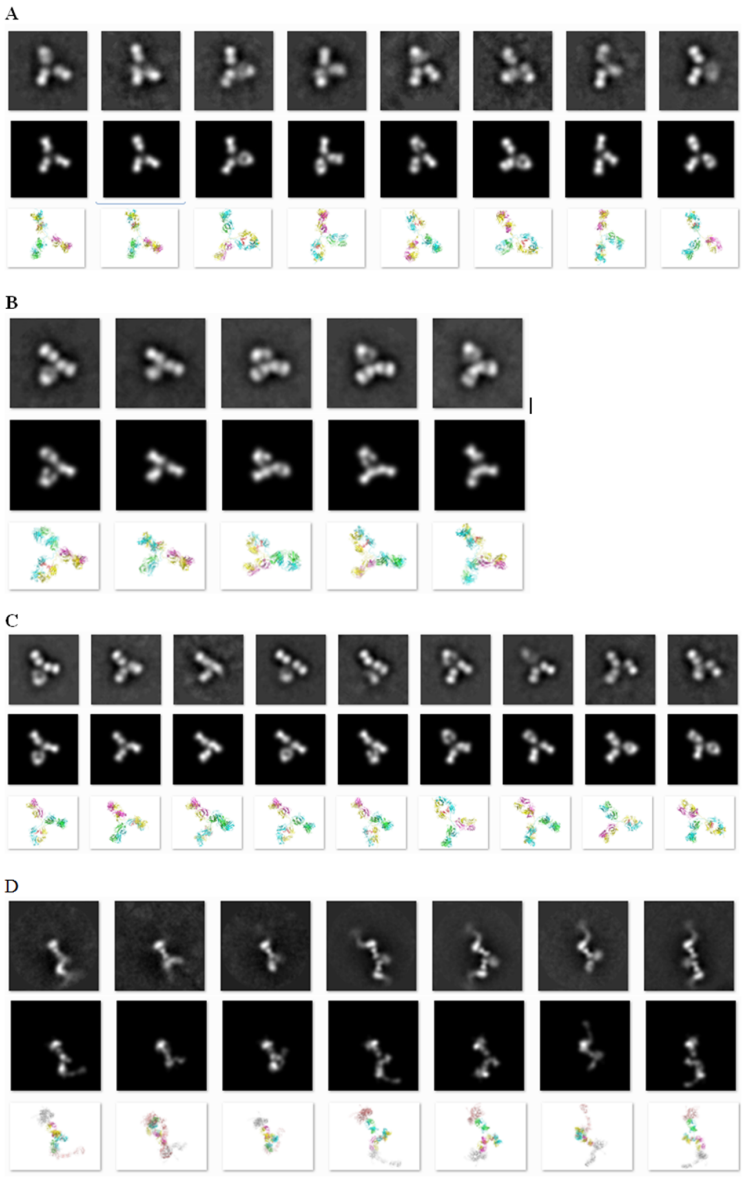

Figure 9.4 | Observed class averages, resulting 3D models, and class averages computed from the models.
Images of observed 2D class average (first row), class averages computed from the resulting 3D models (second row) and the ribbon views from the 3D model (light chains in green and magenta, heavy chains in yellow and cyan, glycoside heavy atoms in red), and the models (second row). (**A**) IgG1: image dimension 160x160 pixels, 2.0 Å/pixel. (**B**) IgG2: image dimension 160x160 pixels, 2.0 Å/pixel. (**C**) IgG4: image dimension 106x106 pixels, 3.0 Å/pixel. (**D**) IgG4-MET antigen complex: image dimension 160x160 pixels, 3.24 Å/pixel.

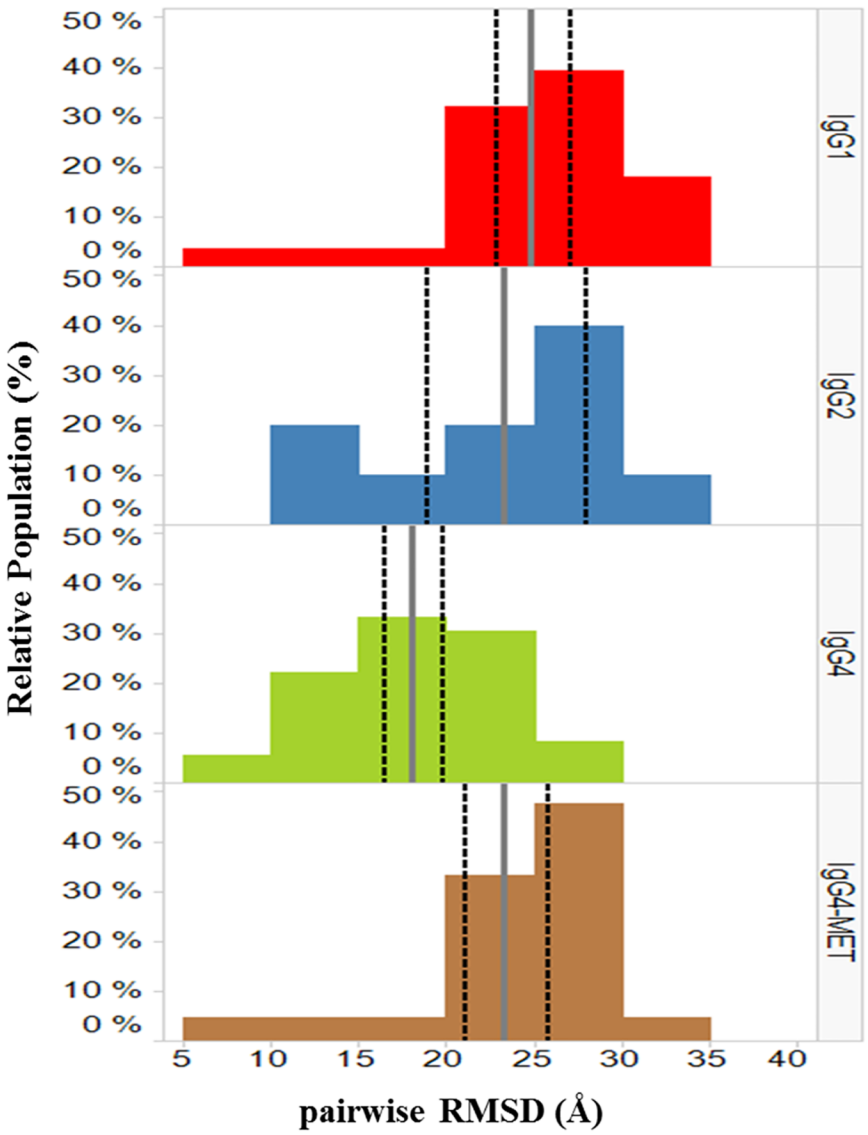

Figure 9.5 | Distributions of pairwise RMSD values for IgG1, IgG2, IgG4 and IgG4-MET complex models.
The average values of pairwise RMSD are indicated by the gray vertical solid bars. The 95% confidence intervals of the average are indicated by the black vertical dashed bars. The plots were created using TIBCO Spotfire 6.5.3 (16).

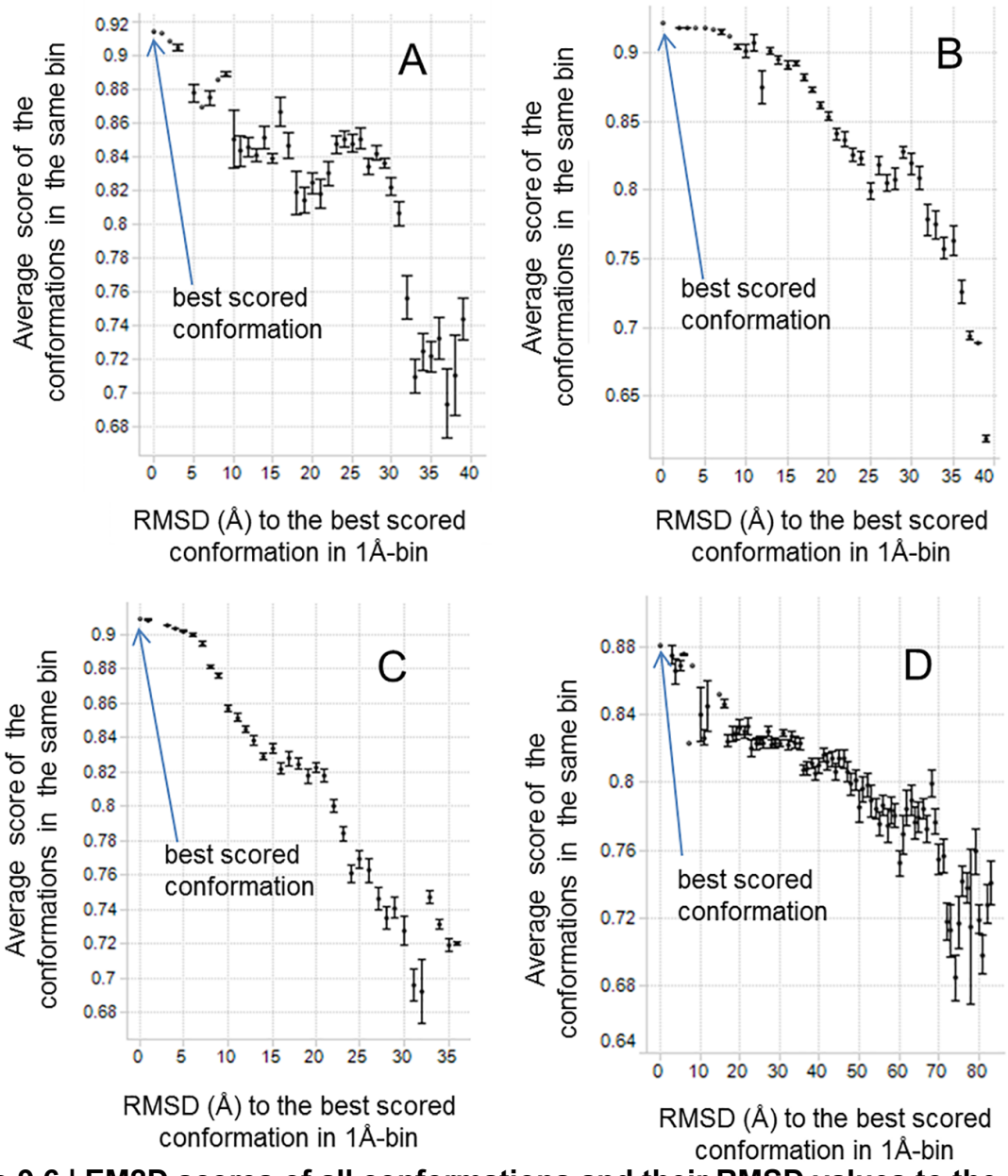

Figure 9.6 | EM2D scores of all conformations and their RMSD values to the highest scoring conformation.
The conformations are binned into groups of 1 Å size according to the RMSD values. The standard error is shown as the error bar for each bin. The samples are (**A**) IgG1, (**B**) IgG2, (**C**) IgG4, and (**D**) IgG4 antigen complex.

EM2D scores (**Fig. 9.6**) and the visual examination (as described in Materials and Methods) indicated that only the best scoring conformation and its close conformational analogs (within 5 Å RMSD) provided satisfactory representation of the 2D class average. Other domain conformations (more distant than 5 Å RMSD from the best scoring one) had lower EM2D scores and less satisfactory matches to the 2D class average. This observation suggested that the 3D domain conformational models within 5 Å RMSD from the best scoring model were a good explanation of a given 2D class average. It further suggested that the model precision is approximately 5 Å RMSD. This relatively high "model precision" demonstrates that our modeling protocol adds significant structural information to the initial 2D images, which were produced at ~20 Å image resolution.

Approximately 75% of 2D class averages were excluded from our analysis, mostly due to the lack of recognizable Y/T shape (see **Fig. 9.3** for the examples of the excluded images) or because we could not find a model matching the class average (**Fig. 9.4**; **Materials and Methods**.) These exclusions likely stemmed from the sample preparation and imaging process. Our protocol, while offering model precision of 5 Å RMSD, could only be used to create 3D models for domain conformations that have clearly recognizable antibody shapes in 2D images. As a result of this process limitation, our comparison of domain conformations only focused on the 2D class averages that have recognizable stems and arms of an antibody, potentially resulting in an underestimation of conformational flexibility.

The IgG1 isotype had the highest internal diversity of 3D domain conformations, with average pairwise RMSD of 24.8 Å (**Fig. 9.5** and **Table 1**). The other two MET constructs, IgG2 and IgG4, had lower conformational diversity with average RMSD of

23.3 and 18.0 Å, respectively. The difference of the internal diversity between IgG1 and IgG4 is statistically significant at the 95% confidence level (**Fig. 9.5**). Both IgG1 and IgG4 have two hinge disulfide bonds, whereas IgG2 has four disulfide bonds. The hinge length is 17 residues for both IgG1 and IgG2, and 15 for IgG4. Both the longer hinge length and lower number of disulfide bonds are expected to lead to higher conformational diversity. The internal conformational diversity observed in our 3D models is more consistent with the increased hinge length than with the number of hinge disulfide bonds.

In contrast, the number of the modeled conformational states (as opposed to internal conformational diversity) increases as the number of disulfide bonds decreases. A lower number of disulfide bonds resulted in more 2D class averages and a higher number of the corresponding modeled 3D domain conformations. For example, only five unique conformational states were observed for the four-hinge-disulfide-bonded IgG2, while eight and nine conformational states were found for the two-hinge-disulfide-bonded IgG1 and IgG4, respectively.

The angular range of domains relative to one another (2), i.e., the angles between two Fab domains and the angles between the Fab and Fc domains (as defined in **Fig. 9.1B**), are another commonly used characteristics of antibody flexibility. These parameters, together with other metrics characterizing the modeled constructs, are shown in **Table 1** and Figure A in S1 File. Overall, IgG2 and IgG4 constructs had similar average values of all three domain angles (Fab-Fab and two Fab1/2-Fc) while IgG1 and IgG4 had a similar span of the values. The narrower span of the Fab-Fab angles from IgG2 (46 degrees) than that for IgG1 and IgG4 (57 degrees for both) is consistent with the lower expected hinge flexibility of IgG2.

3D models for antigen bound IgG4 MET complex

The EM images of the complex between the MET antigen and IgG4 allowed us to compute seven models of domain conformations. While the antigen presence did not change the average values of the domain angles, it led to a larger span of these angles and higher pairwise RMSD values (**Table 1** and **Fig. 9.5**) than found of IgG4 on its own. The domain conformations of the complex were also significantly different from those of the isolated IgG4 with an average pairwise RMSD of 25.7 ± 4.2 Å (Table C in S1 File). This observation suggests that angular measurement alone is not sufficient to determine 3D domain conformational differences. The model differences between the antigen-bound and isolated IgG4 indicate that the conformational space of the free form might differ from that for the antigen bound form.

Table 9.1 | Flexibility of domain arrangements.
The differences in average values of the domain angles are not statistically significant at 95% confidence level (Figure B in S1 File).

	IgG1	IgG2	IgG4	IgG4-MET
Hinge				
number of residues	17	17	15	15
number of disulfide bonds	2	4	2	2
EM Experiments				
magnification ratio	110,000	110,000	110,000	67,000
number of recognizable particles	1,688	2,557	4,685	4,000
number of 2D class averages	8	5	9	7
Domain Angles from 3D Structures (°)				
average Fab1-Fab2 (population weighted)	114 (114)	139 (139)	138 (140)	139 (144)
average Fab1-Fc (population weighted)	113 (115)	101 (99)	95 (95)	92 (90)
average Fab2-Fc (population weighted)	130 (129)	119 (121)	122 (121)	122 (118)
range of Fab-Fab	82–139	118–164	108–165	108–177
span of Fab-Fab	57	46	57	69

	IgG1	IgG2	IgG4	IgG4-MET
Domain Angles from 3D Structures (°)				
range of Fab1-Fc	71–135	72–142	71–127	63–127
span of Fab1-Fc	64	70	56	64
range of Fab2-Fc	108–163	85–138	104–163	83–154
span of Fab2-Fc	55	53	59	71
Pairwise RMSD of 3D Structures (Å)				
Average ± sd	24.8 ± 5.7	23.3 ± 7.3	18.0 ± 5.0	23.3 ± 5.6
Median	26.1	24.1	18.1	25.0
Minimum	8.3	11.4	5.7	8.0
Maximum	33.8	34.5	27.2	32.5

Discussion

It has been shown that antibodies can adopt multiple conformational states. For example, as many as 120 different conformations were reported for the mouse IgG1 from an IPET study (2). Our results support the existence of multiple conformational states for all our MET antibody constructs. In particular, we found that the IgG1 construct had much wider diversity of domain conformations (**Fig. 9.5** and **Table 1**) and showed the smallest angle between the two Fab domains (**Table 1**). This behavior is in contrast to the other two isotypes (IgG2 and IgG4) that showed little or low agonist activity. It is tempting to speculate that the agonist activity of MET IgG1 (**Fig. 9.2**) is related to the extra domain conformational space and/or smaller Fab-Fab angles. Further structural characterization on additional samples with a range of biological agonist profiles will be required to substantiate our hypothesis about the MET antibody agonist activity. Although these results may not be generalizable to all antibodies, our protocol can be readily applied to study dynamics and/or heterogeneity of other antibody constructs.

Our results also indicate that antigen free IgG4 MET conformations are different from those observed in the complex with the antigen. This observation suggests the conformational states required for MET antigen recognition might be different from those accessible to the free antibody. Therefore, the complexes rather than free antibodies are more suited to study structures relevant to the biological activity. Our results demonstrate that the conformational domain variability in our models was dependent primarily on the hinge length, while the number of conformational states revealed by 2D class averages depends on the hinge disulfide patterns.

It is commonly understood that negative staining may introduce artifacts, such as particle flattening (17). Our analysis is clearly limited by any such artifacts. However, it seems conceivable that staining does not significantly affect conformational variability in the sample, because antibody adherence to the carbon coated grid likely traps the observed conformations prior to application of the negative stain. Also, negative staining with uranyl formate was reported to fix protein samples so rapidly that the overall protein conformations or protein complex formation were not affected (17). Indeed, the degree of flexibility we observed is consistent with other studies of antibody conformation (2).

The resolution of 2D images and the resulting class averages is lower when compared to the near atomic resolution from the state-of-the art single particle cryo-EM reconstruction (18–24). But for some applications, in particular to systems containing multiple functional states, the practicality of obtaining single-particle reconstructions at atomic resolution might be time and cost prohibitive. In such cases, our approach may offer a reasonable and economical start before pursuing the more involved experiments. In recent review articles (25, 26), it was estimated that higher resolution data comes at a cost of UK£1,000/day for cryo-EM-related computing works, in addition to US$5 million instrument cost and half million dollars only for annual operational expenses. A number of studies reported the use of 2D class averages to investigate protein conformation flexibility(27, 28). As cryo-EM methods become more routinely applicable to particles smaller than 150 kDa, it will be informative to compare cryo-EM class averages of antibody conformations in vitreous ice with those observed by single particle and tomographic analyses of negative-stained samples at similar magnification on the same EM camera.

Our integrative experimental and computational approach was able to provide models of four MET antibody systems with 5 Å RMSD precision. It is expected to be applicable to the determination of low-resolution 3D models of many other antibodies and for testing certain structural hypothesis at a fraction of cost needed for single-particle reconstruction. While 2D images likely reveal only a fraction of all conformational states, they can serve as a good starting point to understanding a relative structural diversity of a particular system. Therefore, lower resolution 2D imaging and resulting class averages could be complementary to 3D Cryo-EM for biological systems that have multiple conformations differing by more than 5 Å RMSD, such as in the case demonstrated here for the MET antibodies.

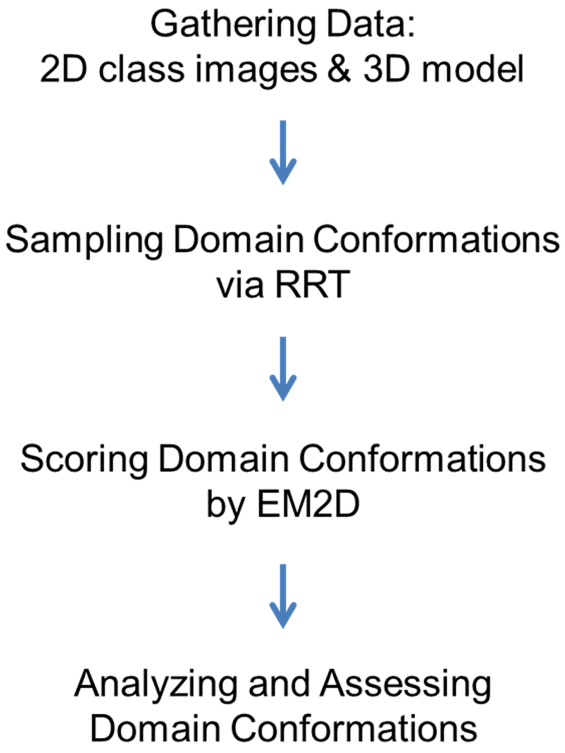

Figure 9.7 | Flowchart of integrative multi-state modeling.

Materials and methods

Antibody samples and preparation

The variable regions from LY2875358 (11) were cloned into human IgG1 and IgG2 constant regions. The antibodies with human IgG1, IgG2 and IgG4 were expressed in Chinese Hamster Ovary cells and purified to greater than 95% homogeneity using Protein A followed by preparative Size Exclusion Chromatography. Human MET ECD containing a FLIS tag was expressed and purified from Chinese Hamster Ovary cells. The IgG4 antibody in complex with human MET ECD antigen was prepared by Size Exclusion Chromatography.

Samples of IgG1, IgG2, IgG4, and IgG4 MET antigen complex were diluted 1:500 with HBS-N pH 7.4 prior to imaging. The samples were then prepared using continuous carbon grid method. Grids were nitrocellulose supported 400-mesh copper. The samples were prepared by applying 3 µL of sample suspension to a cleaned grid, blotting away with filter paper, and immediately staining with Uranyl Formate.

Phosphorylation of pan-AKT assay

Caki-1 Cells were starved overnight in serum-free medium with 0.5% BSA and then treated with various doses of MET antibodies for 15 minutes. Cell lysates were analyzed for phosphorylation of pan-AKT by MSD ELISA. HGF and agonist bivalent MET antibody 5D5 were used as positive controls.

EM imaging

EM experiments were performed (29) using an FEI Tecnai T12 electron microscope, operating at 120 keV equipped with an FEI Eagle 4k x 4k CCD camera. Negative stain grids were transferred into the electron microscope using a room temperature stage.

Images of each grid were acquired at multiple scales to assess the overall distribution of the specimen. After identifying potentially suitable target areas for imaging at lower magnifications, high magnification images were acquired at nominal magnifications of 110,000X (0.10 nm/pixel) or 67,000X (0.16 nm/pixel). The images were acquired at a nominal underfocus of -2 μm (110,000X) or -3 μm to -2 μm (67,000X), and electron doses of ~25–40 e/Å^2.

Integrative modeling

Our integrative structure modeling proceeds through four stages (9, 10): (1) gathering the data, (2) sampling the domain conformations, (3) scoring the domain conformations, and (4) analyzing and assessing the domain conformations (**Fig. 9.7**):

Stage 1: Gathering the data for the initial models

EM 2D class averages

The individual particles were identified in the high magnification images prior to the alignment and classification. The individual particles were then selected, boxed out, and individual sub-images were combined into a stack to be processed using a reference-free classification method (30). Individual particles in the 67,000X or 110,000X high magnification images were selected using automated picking protocols (31). An initial round of alignments was done on each sample, followed by selecting recognizable particles for additional rounds of alignment. Only classes with three recognizable domains were kept for further analysis and 3D structure modeling. Particle alignment and classification were carried out using a reference-free alignment strategy based on the XMIPP (30) processing package. Algorithms in this package aligned the selected particles and sorted them into self-similar groups of classes.

Antibody comparative modeling

Initial comparative structure models of the antibody constructs were built by MOE modeling package (15) using an X-ray crystal structure (PDB code 1IGT) (3) as a template. In all cases, disulfide bridges were added (if not created automatically). The structures were then minimized using the Amber10:EHT force field (15) to avoid atomic clashes and resolve any strain created by disulfide addition. The glycosides were also grafted from the 1IGT template structure. The models were then minimized with restraints for non-hydrogen atoms to their original positions, and served as starting points for hinge flexibility exploration. The antibody-antigen complex structure for IgG4 was constructed by grafting the aligned MET-Fab complex (PDB code 4K3J) (32) onto the comparative model of MET IgG4, using the Fab domains' backbones for superposition. In addition, the complete structure of the MET antigen (PDB code 2UZY) (33) was grafted to the existing model based on the alignment of the overlapping MET region. The final model of the complex was then energy minimized with backbone restraints to avoid the clashes produced by superposition.

Stage 2: Sampling domain conformation

We used the Rapidly exploring Random Tree (RRT) algorithm (34, 35) to explore the domain conformational space of the full length antibody, using the optimized comparative model as a starting conformation. A modified version of the RRT algorithm implemented in IMP (9, 10) was used, which sampled the dihedral angles of the protein under the closure constraint to keep the disulfide connections among different chains in the hinge region intact. During the search, two Fab domains and the Fc domain were treated as rigid bodies. Up to 100,000 iterations of RRT were performed, and typically about 2,000

diverse domain conformations were generated. The diversity of the domain conformations was characterized by the RMSD values of C_α atoms between every pair of generated models as well as the domain angles as defined in **Fig. 9.1B**. In depth discussion of the sampling exhaustiveness is present in S1 File.

Stage 3: Scoring domain conformation

EM2D scoring function

The EM2D module of the IMP program (8–10) was used to compare all antibody domain conformation models to every experimental 2D class average. The 2D class average resolution was estimated to be about 20 Å based on the simulation results. Specifically, the EM2D score of a 3D model reaches a maximum when the model is projected to generate 2D images using the resolution that matches the actual resolution of the 2D class average. We tested a range of resolution values from 2–30 Å and found that the maximal EM2D scores occurred at about 20 Å, thus defining the resolution of the 2D class averages. For every experimental 2D class average, 1,000 different orientations of every domain conformation were projected onto the 2D class image plane to produce the simulated images. The simulated images were then optimized and scored based on the Gaussian-weighted cross-correlation coefficient (i.e., EM2D score) (8) between the observed and the simulated 2D images. For every domain conformational model, the score, simulated 2D image, and parameters of the best-scored orientation were recorded as the final solutions. The PDB file of the conformational model in the final orientation was generated along with the protein ribbon-view image created using a PyMOL script (36).

A scoring function that can rank alternative models by their accuracy is an essential part of any structure modeling. Good scoring functions can correctly rank models over a broad

range of RMSD values from the native state. We examined whether or not such “scoring funnels” existed for EM2D scoring (37) and determined their shape for the specific antibody constructs. The EM2D scores of all RRT generated candidate domain conformations are plotted against the RMSD values from the highest scoring model in **Fig. 9.6**. When fitting to the same experimental image, differences were often observed from two simulated images even when the EM2D scores differed by only 0.01. We thus expanded the score threshold window to 0.03 (from the best scoring one) to broaden the pool of candidates for the visual inspection and final model refinement.

Stage 4: Analysis and assessment of the ensemble

For every experimental 2D class average, all domain conformational models were sorted in descending order by their EM2D scores (see Stage 3) and clustered by a “leader” algorithm (38, 39) to facilitate visual examination. In depth discussion on the EM2D scoring as it relates to image selection is present in S1 File. The highest scoring domain conformation was selected as the first leader object and removed from the list. The all-residue Cα RMSD values of every subsequent domain conformation in the sorted list to the already-found leader conformation determined whether the conformation was designated as a new leader or a member of an existing cluster represented by a previously-selected leader. A new leader conformation was selected with the highest EM2D score and the RMSD to existing leaders exceeding the threshold value. Four thresholds of the RMSD values were used for the leader clustering, 5, 10, 15, and 20 Å. For each threshold, up to 20 top scoring leaders and several other top scoring cluster members were visually examined for their fit to the experimental 2D class average.

The visual inspection, aided by PyMOL (36) protein ribbon representation, was used only to select the cutoff for the EM2D score that ensures consistency between the 2D class average, the simulated image, and the 3D model ribbon view, including the assignment of the Fc domain. Specifically, visual inspection of matches and mismatches confirmed that the EM2D score can indeed be used for ranking models based on class averages. However, this visual examination also suggested that some 3D models with slightly lower EM2D scores also match the 2D class averages well, if these models were within 5 Å RMSD to the best scoring model, resulting in the estimate of model precision of 5 Å RMSD. Visualization further revealed that the EM2D score of at least 0.83 was needed to ensure such consistency; otherwise, mismatches between the models and the class averages could occur due to similar shapes of the three domains. Therefore, we elected to discard the class averages without models with the EM2D score of at least 0.83. In summary, a good fit between an experimental 2D class average and a 3D antibody domain conformation has two attributes: (1) EM2D score of at least 0.83; and (2) the observed 2D image, the simulated 2D image, and the ribbon view image are consistent with each other in the overall shape and Fc domain assignment.

The final 3D structures representing the best matches to the 2D class averages were subsequently transferred to the MOE modelling package (15), and minimized with constraints for Fab and Fc domain backbone heavy atoms while allowing the hinge atoms to be flexible. The glycosides were attached to the glycosylation sites on the Fc domain heavy chains. The resulting minimized structures were transferred into Maestro15.3 (40), optimized with the protein preparation wizard and solvated in orthorhombic TIP3P water box for the NPT Molecular Dynamics (MD) local refinement with Langevin thermostat and

PME using Desmond package (40). The refinement protocol consisted of three steps of constrained minimization: 5,000 steps with heavy atom restraints, followed by unrestrained 50,000 steps, restrained heating at 30 ps at temperature 100 K, 200 K and 300 K equilibration. The 5 ns MD production run for each 3D structure with backbone heavy atom constraints for Fab and Fc domains allowed for equilibration of side chains, glycosides and hinge residues. The resulting structures were then minimized with 1,000 steps of steepest descent minimization to produce the final 3D models.

Data Availability

All relevant data are within the paper and its Supporting Information files.

S1 File. Diversity of candidate conformation states for quality reconstructions.

The file also contains two figures and three tables. https://doi.org/10.1371/journal.pone.0175758.s001

Funding

Eli Lilly and Company provided support in the form of research materials for authors (QC, MV, DET, CH, WZ, JL, LL). The IMP software development was funded in part by NIH grants to AS, including P41 GM109824 and R01 GM083960. IEC is supported by the NSF Graduate Student Research Fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

The IMP software development was funded in part by NIH grants to AS, including P41 GM109824 and R01 GM083960. IEC is supported by the NSF Graduate Student Research Fellowship. The affiliation of authors (QC, MV, DET, CH, WZ, JL, LL) with Eli

Lilly and Company does not alter our adherence to PLOS ONE policies on sharing data and materials.

Acknowledgments

The authors would like to thank Dr. Jeremy Desaphy for technical supports of protein structure analysis and Dr. Phil Hipskind for his help to initiate the EM studies. The authors would like to thank Dr. Bridget Carrager (NanoImaging Services Inc.) for helpful discussions on interpreting EM images. The EM data collection was performed by NanoImaging Services, Inc., 10835 Road to the Cure, Suite 150, San Diego, CA 92121. The IMP software development was funded in part by NIH grants to AS, including P41 GM109824 and R01 GM083960. I.E.C is supported by the NSF Graduate Student Research Fellowship.

Author Contributions

Conceptualization: QC MV DET CH. Data curation: QC MV DET WZ. Formal analysis: QC MV WZ. Funding acquisition: DET CH AS. Investigation: DET WZ QC MV. Methodology: QC MV DET CH DSD IEC AS. Project administration: QC DET CH. Resources: DET DSD IEC WZ LL JL. Software: QC DSD IEC. Supervision: QC MV DET CH AS. Validation: DET IEC WZ. Visualization: QC MV WZ.Writing – original draft: QC MV. Writing – review & editing: CH DSD IEC AS LL JL DET.

References

1. J. M. Lambert, Antibody-Drug Conjugates (ADCs): Magic Bullets at Last! *Mol. Pharm.* **12**, 1701–1702 (2015).
2. X. Zhang, *et al.*, 3D Structural Fluctuation of IgG1 Antibody Revealed by Individual Particle Electron Tomography. *Sci Rep* **5**, 9803 (2015).
3. L. J. Harris, S. B. Larson, K. W. Hasel, A. McPherson, Refined structure of an intact IgG2a monoclonal antibody. *Biochemistry* **36**, 1581–1597 (1997).
4. E. O. Saphire, *et al.*, Crystal structure of a neutralizing human IGG against HIV-1: a template for vaccine design. *Science* **293**, 1155–1159 (2001).
5. L. J. Harris, E. Skaletsky, A. McPherson, Crystallographic structure of an intact IgG1 monoclonal antibody. *J. Mol. Biol.* **275**, 861–872 (1998).
6. H. Tong, *et al.*, Peptide-conjugation induced conformational changes in human IgG1 observed by optimized negative-staining and individual-particle electron tomography. *Sci Rep* **3**, 1089 (2013).
7. I. Correia, *et al.*, The structure of dual-variable-domain immunoglobulin molecules alone and bound to antigen. *MAbs* **5**, 364–372 (2013).
8. J. Velazquez-Muriel, *et al.*, Assembly of macromolecular complexes by satisfaction of spatial restraints from electron microscopy images. *Proc Natl Acad Sci U S A* **109**, 18821–6 (2012).
9. F. Alber, B. T. Chait, M. P. Rout, A. Sali, “Integrative Structure Determination of Protein Assemblies by Satisfaction of Spatial Restraints” in *Protein-Protein Interactions and Networks: Identification, Characterization and Prediction.*, A. Panchenko, T. Przytycka, Eds. (Springer-Verlag, 2008), pp. 99–114.

10. D. Russel, *et al.*, Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. *PLOS Biology* **10**, e1001244 (2012).
11. L. Liu, *et al.*, LY2875358, a neutralizing and internalizing anti-MET bivalent antibody, inhibits HGF-dependent and HGF-independent MET activation and tumor growth. *Clin. Cancer Res.* **20**, 6059–6070 (2014).
12. P. Morton, *et al.*, In vitro and in vivo activity of fully-human monoclonal antibodies to c-Met protein tyrosine kinase. in (2003).
13. M. Prat, T. Crepaldi, S. Pennacchietti, F. Bussolino, P. M. Comoglio, Agonistic monoclonal antibodies against the Met receptor dissect the biological responses to HGF. *Journal of cell science* **111**, 237–247 (1998).
14. Z. Zheng, C. Adams, B. Moffat, R. Schwall, editors, A chimeric Fab antibody serves as an antagonist to the HGF/SF receptor cMet. in (2003).
15. CCG, *Molecular Operating Environment (MOE)* (Chemical Computing Group Inc., 2016).
16. TIBCO Software Inc., *TIBCO*.
17. D. S. Booth, A. Avila-Sakar, Y. Cheng, Visualizing proteins and macromolecular complexes by negative stain EM: from grid preparation to image acquisition. *JoVE (Journal of Visualized Experiments)*, e3227 (2011).
18. A. Bartesaghi, *et al.*, 2.2 Å resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor. *Science* **348**, 1147–1151 (2015).

19. M. Fischer, *et al.*, Cryo-EM structure of fatty acid synthase (FAS) from Rhodosporidium toruloides provides insights into the evolutionary development of fungal FAS. *Protein Sci.* **24**, 987–995 (2015).
20. N. Fischer, *et al.*, Structure of the E. coli ribosome-EF-Tu complex at <3 Å resolution by Cs-corrected cryo-EM. *Nature* **520**, 567–570 (2015).
21. L. Sun, *et al.*, Cryo-EM structure of the bacteriophage T4 portal protein assembly at near-atomic resolution. *Nat Commun* **6**, 7548 (2015).
22. A. Bartesaghi, D. Matthies, S. Banerjee, A. Merk, S. Subramaniam, Structure of β-galactosidase at 3.2-Å resolution obtained by cryo-electron microscopy. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 11709–11714 (2014).
23. X. C. Bai, *et al.*, An atomic structure of human gamma-secretase. *Nature* **525**, 212–7 (2015).
24. S. Banerjee, *et al.*, 2.3 Å resolution cryo-EM structure of human p97 and mechanism of allosteric inhibition. *Science* **351**, 871–875 (2016).
25. M. Eisenstein, The field that came in from the cold. *Nat. Methods* **13**, 19–22 (2016).
26. R. M. Glaeser, How good can cryo-EM become? *Nat Methods* **13**, 28–32 (2016).
27. E. J. Brignole, S. Smith, F. J. Asturias, Conformational flexibility of metazoan fatty acid synthase enables catalysis. *Nat. Struct. Mol. Biol.* **16**, 190–197 (2009).
28. J. Zhang, P. Minary, M. Levitt, Multiscale natural moves refine macromolecules using single-particle electron microscopy projection images. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 9845–9850 (2012).
29. , *NanoImaging* (NanoImaging Services Inc).

30. C. O. S. Sorzano, *et al.*, XMIPP: a new generation of an open-source image processing package for electron microscopy. *J. Struct. Biol.* **148**, 194–204 (2004).
31. G. C. Lander, *et al.*, Appion: an integrated, database-driven pipeline to facilitate EM image processing. *J. Struct. Biol.* **166**, 95–102 (2009).
32. M. Merchant, *et al.*, Monovalent antibody design and mechanism of action of onartuzumab, a MET antagonist with anti-tumor activity as a therapeutic agent. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E2987-2996 (2013).
33. H. H. Niemann, *et al.*, Structure of the human receptor tyrosine kinase met in complex with the Listeria invasion protein InlB. *Cell* **130**, 235–246 (2007).
34. S. LaValle, “Rapidly-exploring random trees: A new tool for path planning.” (Computer Science Department, Iowa State University, 1998).
35. S. M. LaValle, J. J. Kuffner Jr, Randomized kinodynamic planning. *The international journal of robotics research* **20**, 378–400 (2001).
36. The PyMOL Molecular Graphics System (Schrödinger, LLC.).
37. N. London, O. Schueler-Furman, Funnel hunting in a rough terrain: learning and discriminating native energy funnels. *Structure* **16**, 269–279 (2008).
38. L. Hodes, Clustering a large number of compounds. 1. Establishing the method on an initial sample. *J Chem Inf Comput Sci* **29**, 66–71 (1989).
39. L. Hodes, A. Feldman, Clustering a large number of compounds. 3. The limits of classification. *J Chem Inf Comput Sci* **31**, 347–350 (1991).
40. Schrodinger, *Schrodinger* (Schrodinger LLC, 2015).

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:

7A176E47D7AD4F9...

Author Signature

3/16/2020

Date