Modeling of the Nup84 complex

Seung Joong Kim,
Benjamin Webb

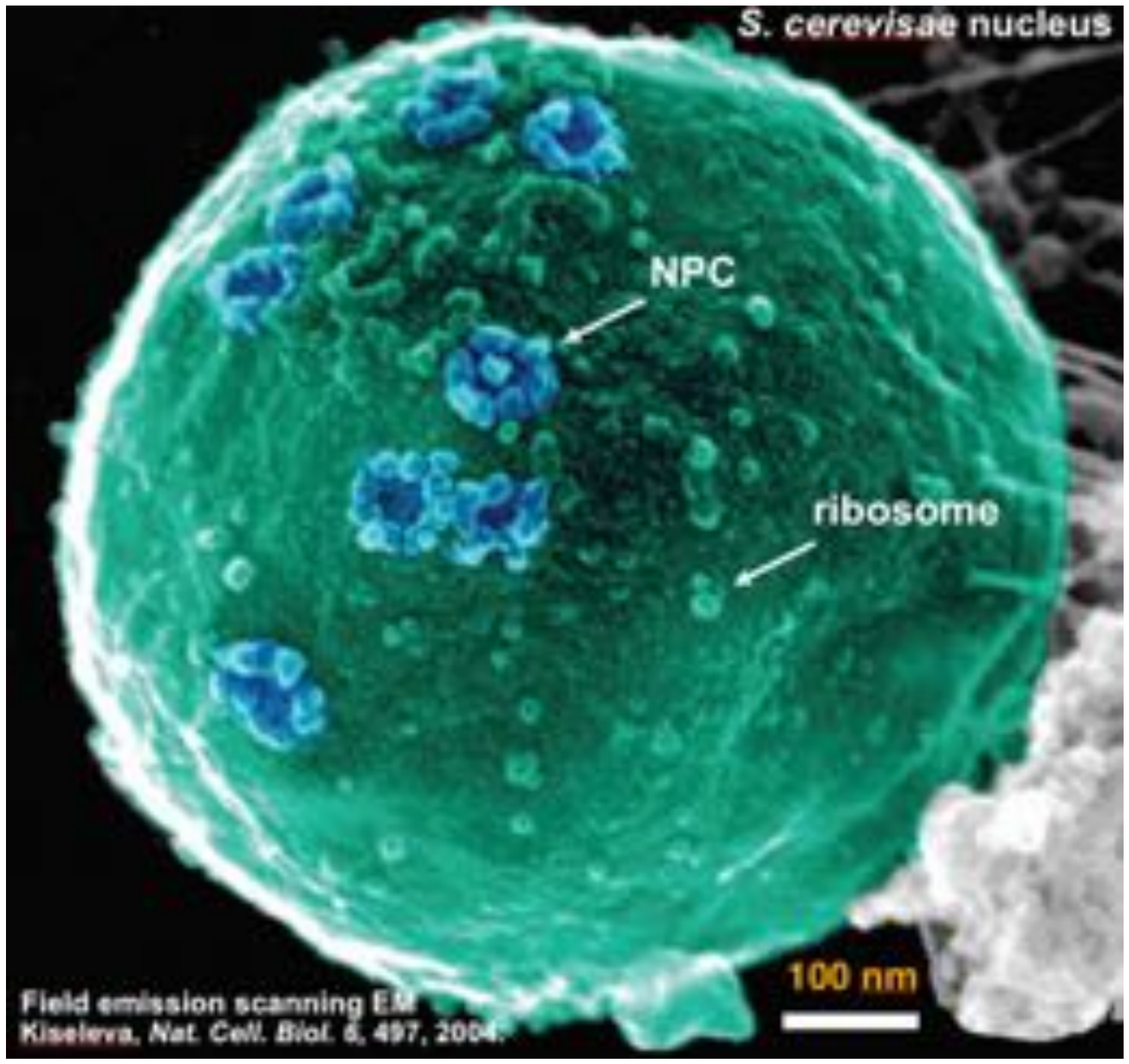
Sali Lab

Ask questions…

Thu/Fri workshop overview

- Thursday morning
  - Lecture on integrative modeling (Andrej Sali)
  - Lecture on a recent integrative modeling study (this talk)
- Thursday afternoon
  - Guided tutorial introducing the IMP software and its use
- Friday
  - More advanced topics (mix of lectures and guided tutorials)
- Free time
  - Can continue with VMD/NAMD tutorials and/or look at IMP tutorials on the workshop website

Nuclear Pore Complex (NPC)



1. Structure
2. Evolution
3. Mechanism of transport
4. Mechanism of assembly
5. Interactions with other systems
6. Modulation and therapy

A large collaborative effort with Mike Rout and Brian Chait at Rockefeller University, also involving many other collaborators (Acknowledgments).

NCDIR *National Center for Dynamic Interactome Research* NIH TCNP

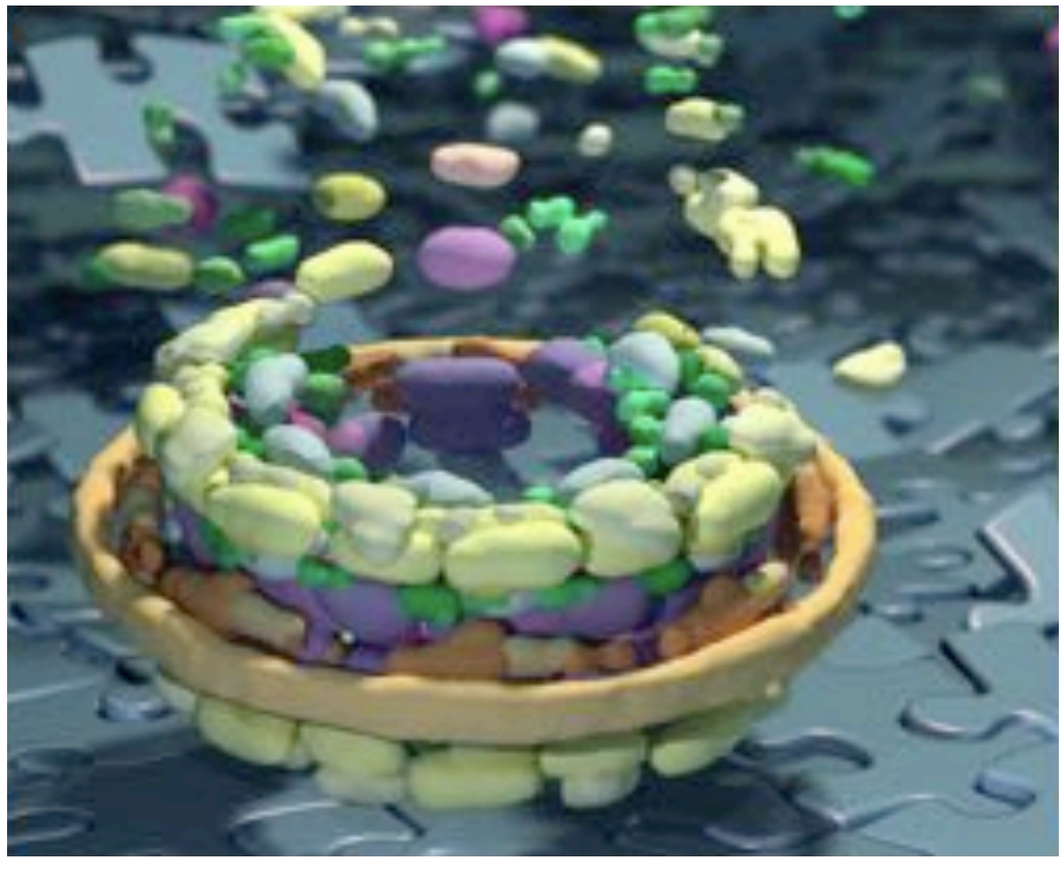
Consists of broadly conserved nucleoporins (nups).

50 MDa complex: ~480 proteins of 30 different types.

Mediates all known nuclear transport, *via* cognate transport factors (karyoferins or kaps)

Final 2007 NPC model

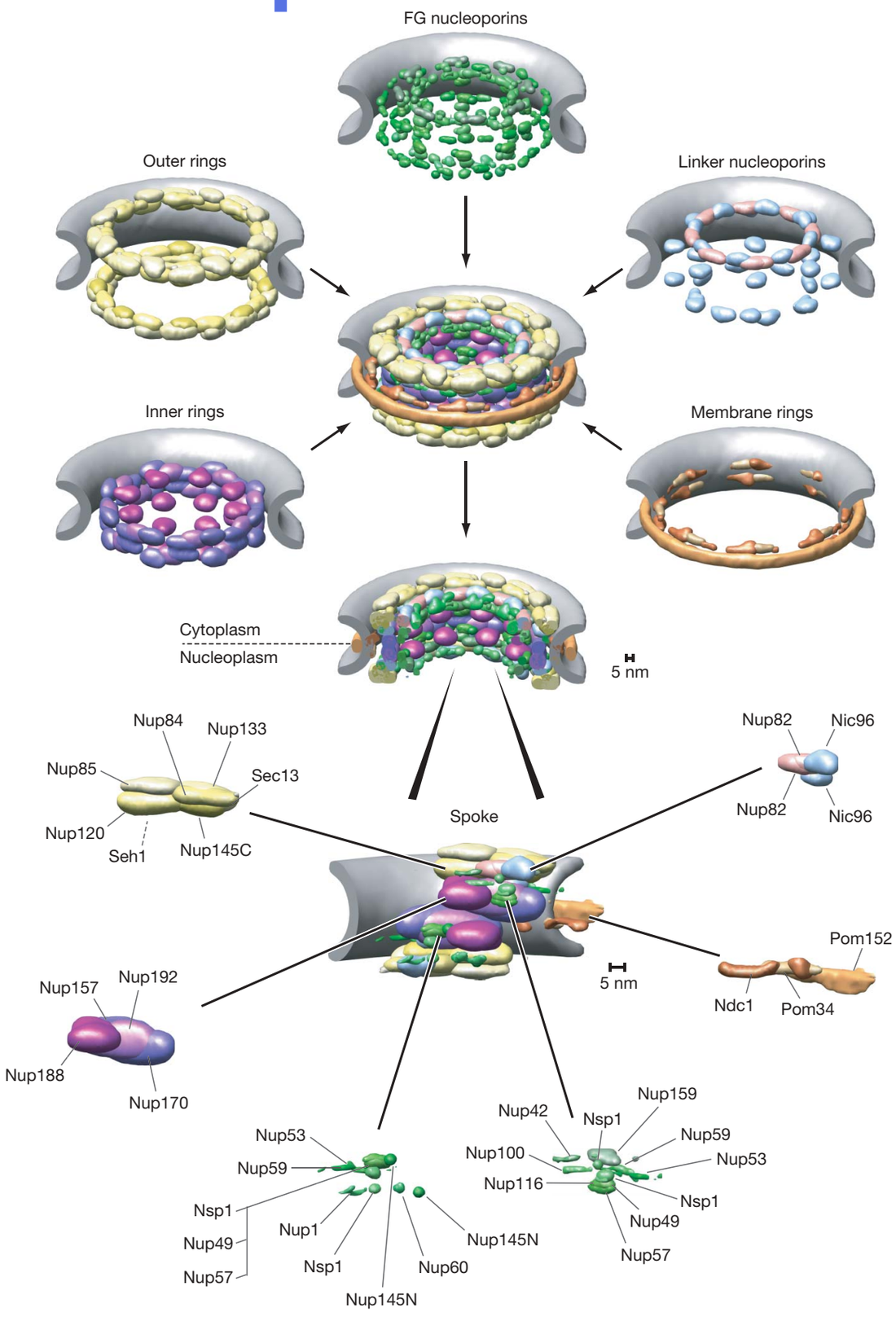
- Sufficient to place each *protein* within the entire complex
- Very coarse-grained model; no atomic information



Alber *et al.* Nature 450, 684-694, 2007
Alber *et al.* Nature 450, 695-702, 2007
with M. Rout & B. Chait

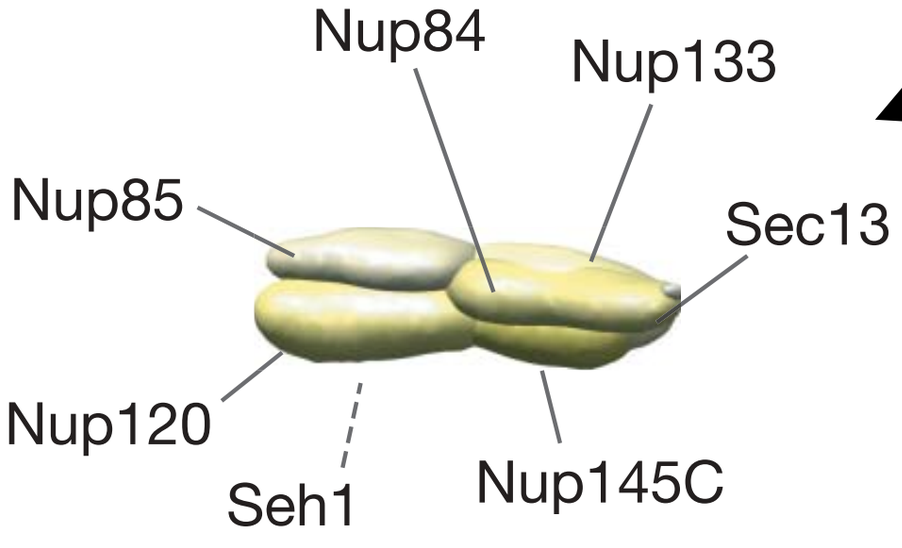
Nup84 subcomplex

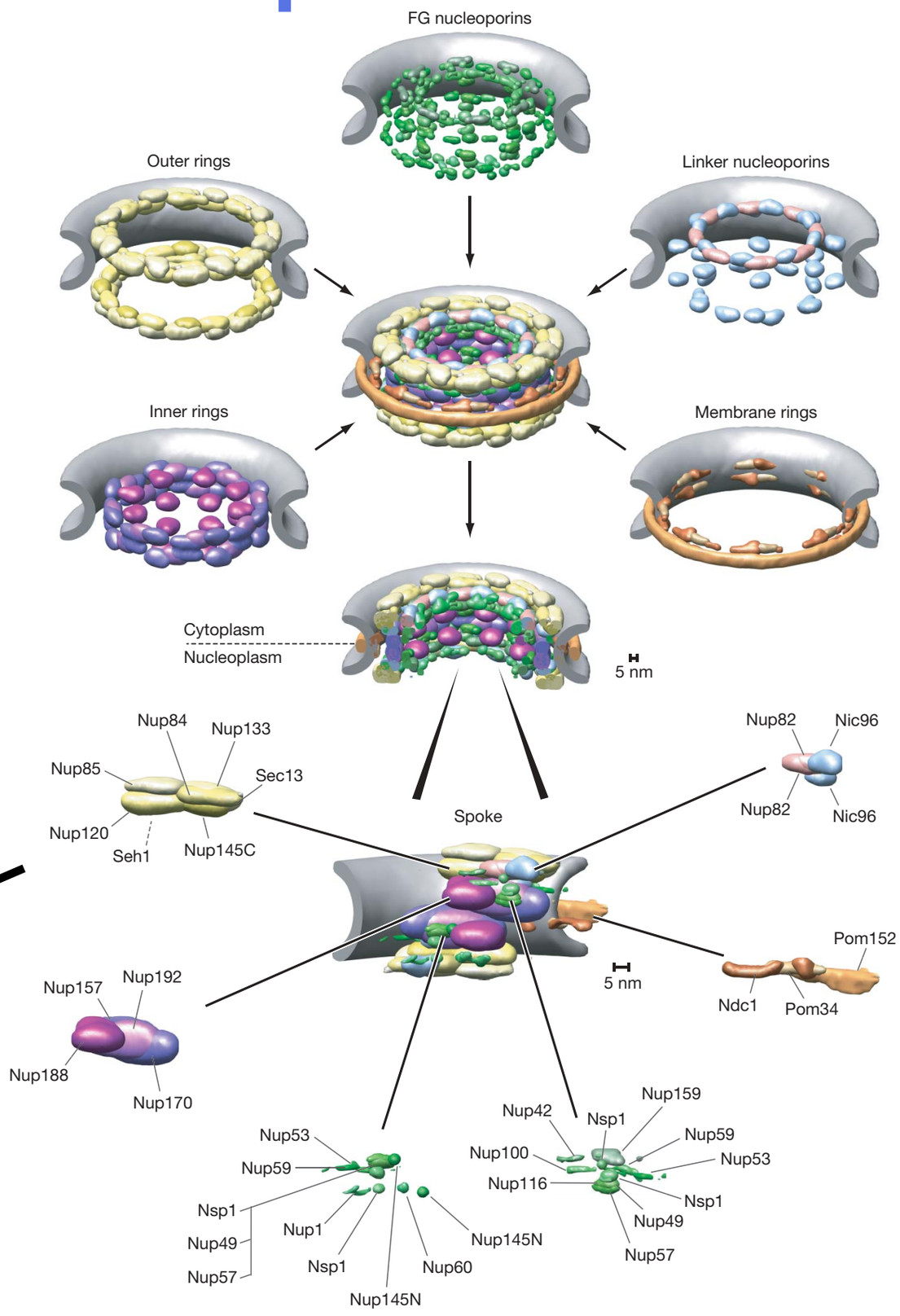
- Look at subcomplexes, towards a higher resolution structure of the entire NPC
- Nup84 is one such subcomplex of 7 proteins, 16 copies of which form the outer ring



Nup84 subcomplex

- Look at subcomplexes, towards a higher resolution structure of the entire NPC
- Nup84 is one such subcomplex of 7 proteins, 16 copies of which form the outer ring

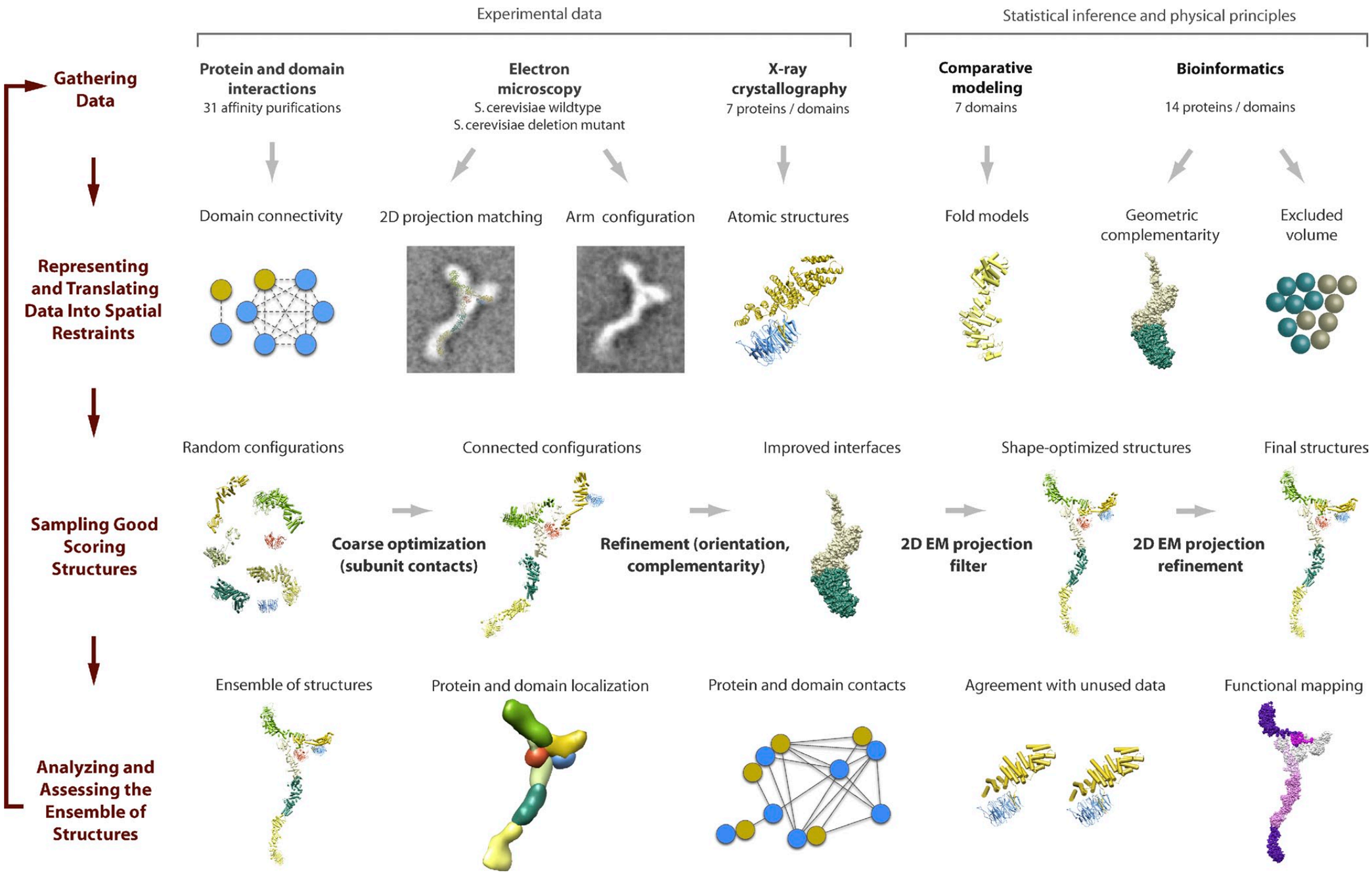




Modeling Nup84 with IMP

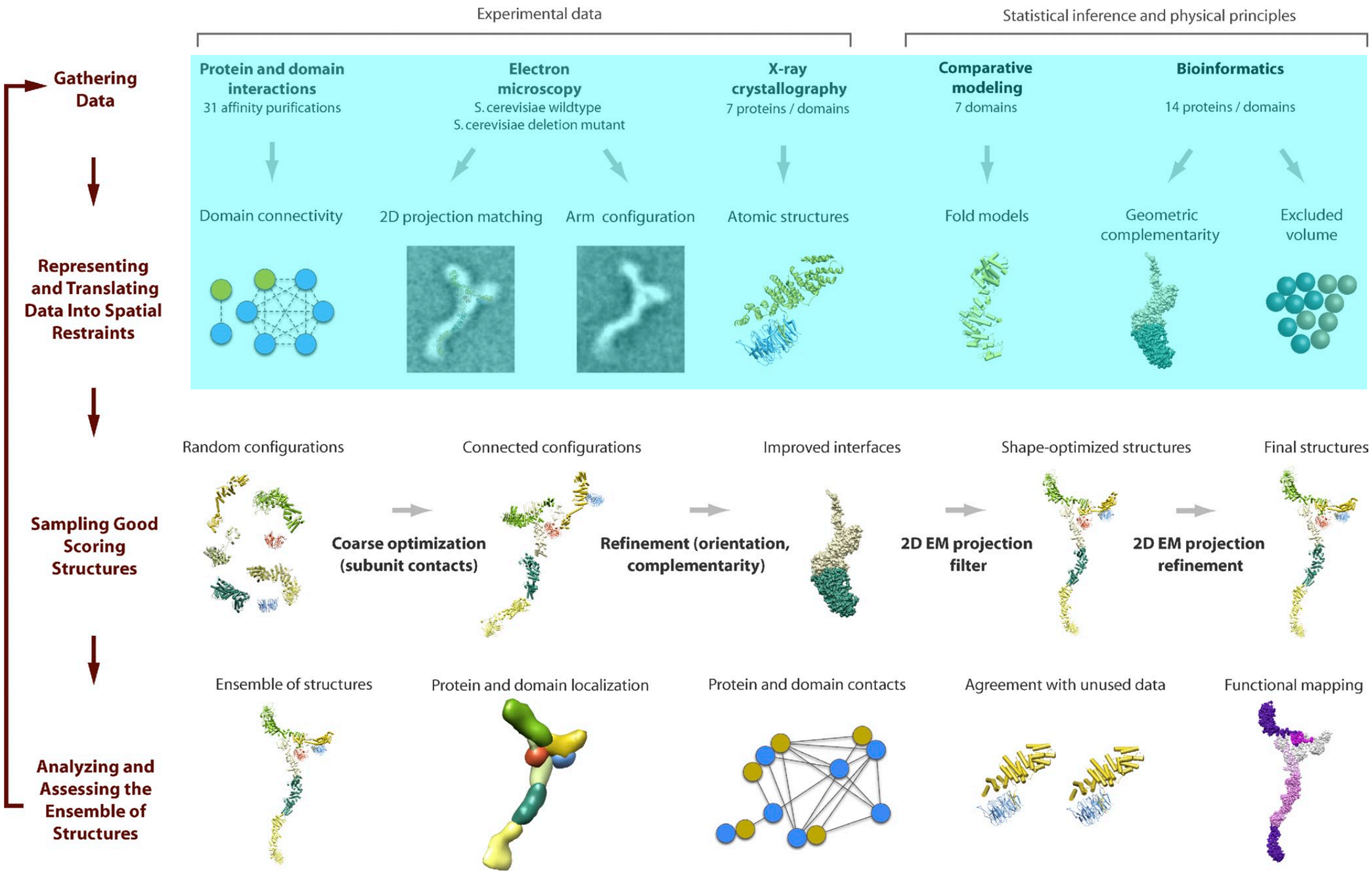
- The Nup84 complex has been well studied, experimentally and computationally
- However, an X-ray structure of the whole complex was not available as of 2014 (only fragments, totaling about 50% of the sequence; their arrangement was only partly known)
- We'll look at two studies here using the IMP software:
  - A medium resolution model using domain deletion mapping and negative stain electron microscopy data (2012)
  - A higher resolution model using cross-linking data and the same electron microscopy data (2014)
- Similar approaches used, so we'll summarize the 2012 study briefly then look at the 2014 study in detail

Modeling Nup84 with IMP (2012)



J Cell Biol 196, 419-434, 2012

Modeling Nup84 with IMP (2012)



J Cell Biol 196, 419-434, 2012

How can we use input data in modeling?

How can we use input data in modeling?

- As *representation* (e.g. rigid body X-ray structure)

How can we use input data in modeling?

- As *representation* (e.g. rigid body X-ray structure)
- In *sampling* (e.g. only propose system perturbations that are consistent with the data)

How can we use input data in modeling?

- As *representation* (e.g. rigid body X-ray structure)
- In *sampling* (e.g. only propose system perturbations that are consistent with the data)
  - as a *constraint*: ensure the data are consistent (e.g. SHAKE)

How can we use input data in modeling?

- As *representation* (e.g. rigid body X-ray structure)
- In *sampling* (e.g. only propose system perturbations that are consistent with the data)
  - as a *constraint*: ensure the data are consistent (e.g. SHAKE)
- As a mathematical function that scores how closely our model (representation) matches the input data:

How can we use input data in modeling?

- As *representation* (e.g. rigid body X-ray structure)
- In *sampling* (e.g. only propose system perturbations that are consistent with the data)
  - as a *constraint*: ensure the data are consistent (e.g. SHAKE)
- As a mathematical function that scores how closely our model (representation) matches the input data:
  - *restraint*: sample configurations guided by the value of this function

How can we use input data in modeling?

- As *representation* (e.g. rigid body X-ray structure)
- In *sampling* (e.g. only propose system perturbations that are consistent with the data)
  - as a *constraint*: ensure the data are consistent (e.g. SHAKE)
- As a mathematical function that scores how closely our model (representation) matches the input data:
  - *restraint*: sample configurations guided by the value of this function
  - *filter*: build models (guided by some other function) then only keep those that score well against the filter

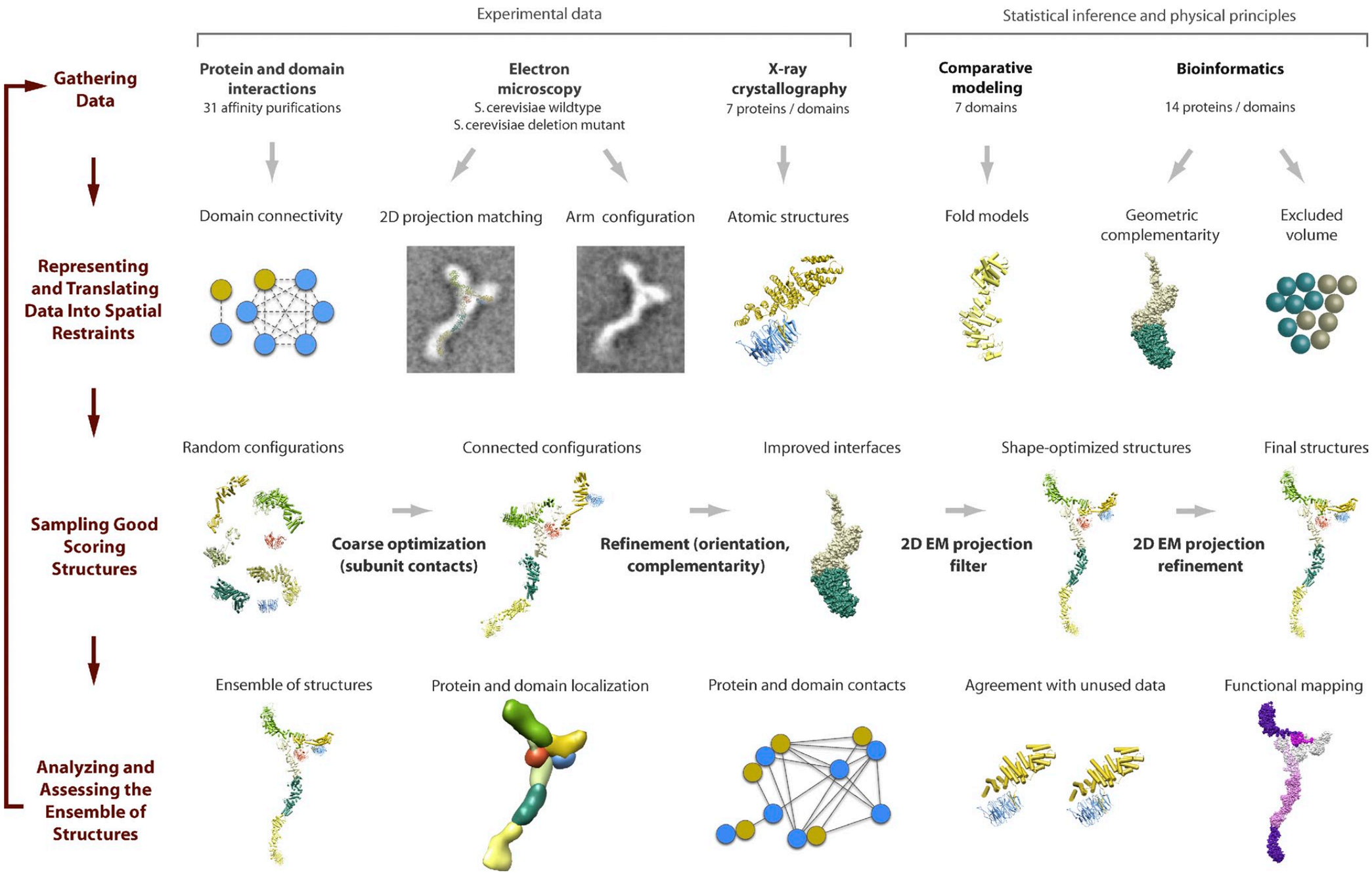
How can we use input data in modeling?

- As *representation* (e.g. rigid body X-ray structure)
- In *sampling* (e.g. only propose system perturbations that are consistent with the data)
  - as a *constraint*: ensure the data are consistent (e.g. SHAKE)
- As a mathematical function that scores how closely our model (representation) matches the input data:
  - *restraint*: sample configurations guided by the value of this function
  - *filter*: build models (guided by some other function) then only keep those that score well against the filter
  - restraint is preferred over filter (but slower)

How can we use input data in modeling?

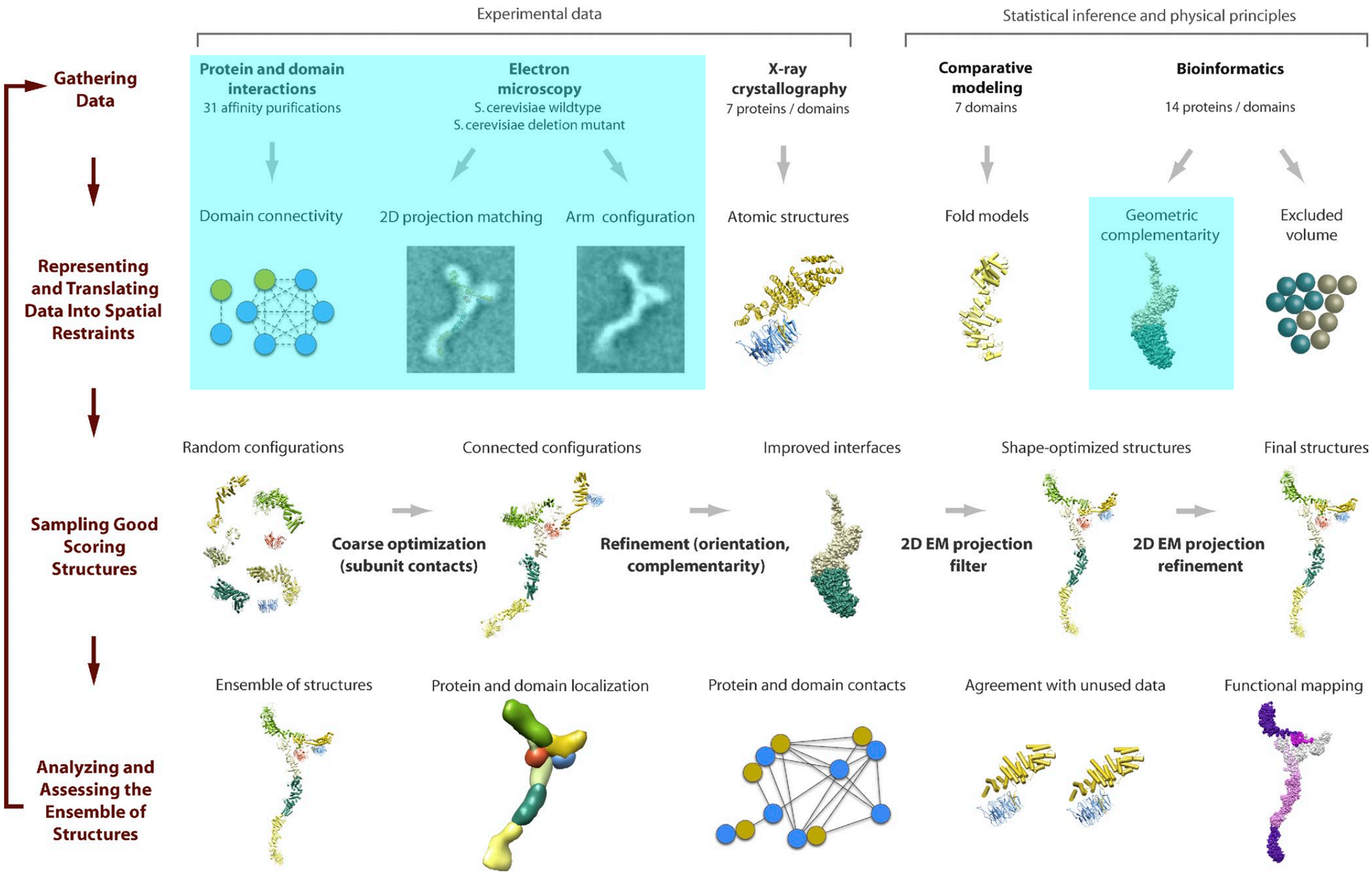
- As *representation* (e.g. rigid body X-ray structure)
- In *sampling* (e.g. only propose system perturbations that are consistent with the data)
  - as a *constraint*: ensure the data are consistent (e.g. SHAKE)
- As a mathematical function that scores how closely our model (representation) matches the input data:
  - *restraint*: sample configurations guided by the value of this function
  - *filter*: build models (guided by some other function) then only keep those that score well against the filter
  - restraint is preferred over filter (but slower)
- As *validation*: don't use the data at all, until after we've selected the 'best' models

Modeling Nup84 with IMP (2012)



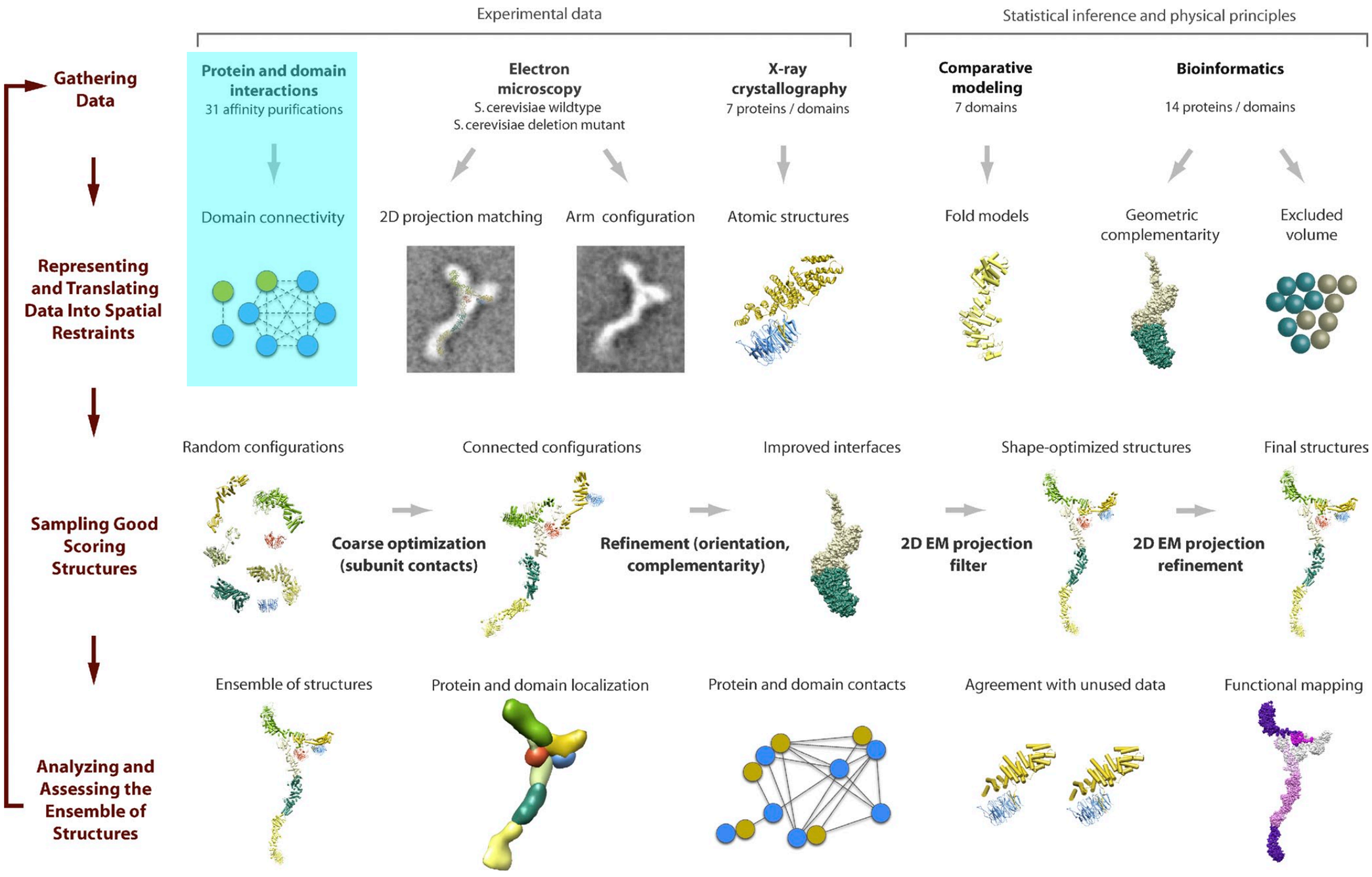
J Cell Biol 196, 419-434, 2012

Modeling Nup84 with IMP (2012)



J Cell Biol 196, 419-434, 2012

Modeling Nup84 with IMP (2012)

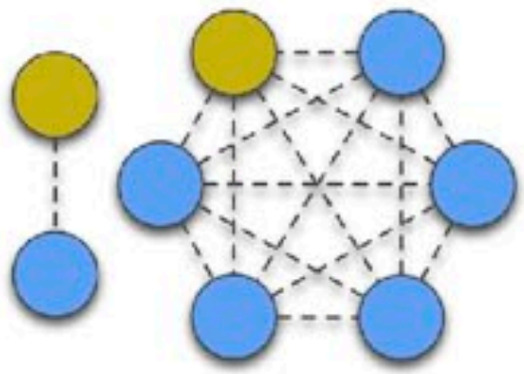


J Cell Biol 196, 419-434, 2012

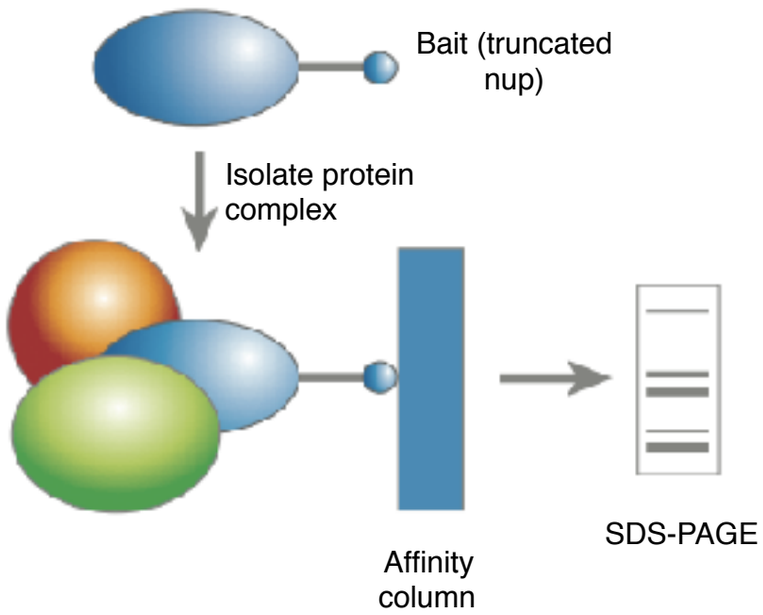
Protein and domain interactions

31 affinity purifications

Domain connectivity

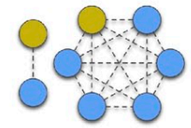


Affinity purification data

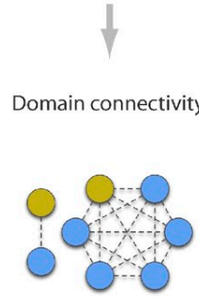


Protein and domain interactions
31 affinity purifications

Domain connectivity



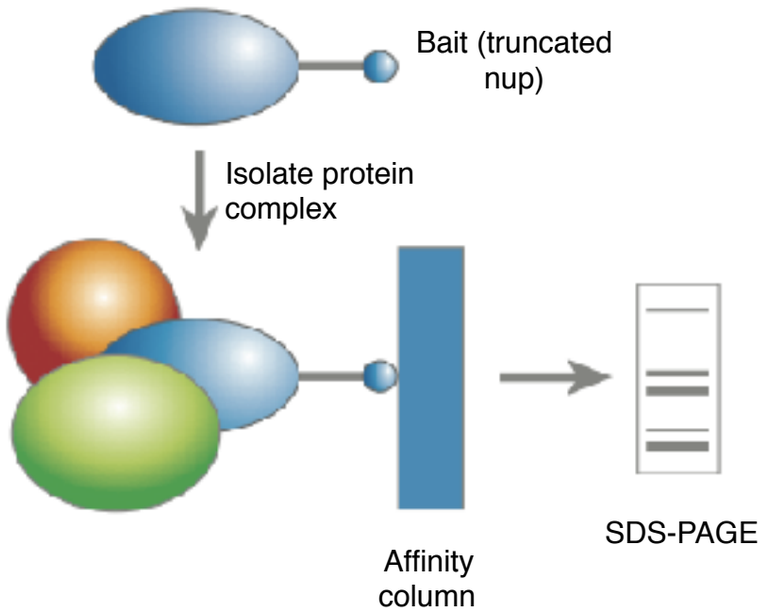
Protein and domain interactions
31 affinity purifications



Domain connectivity

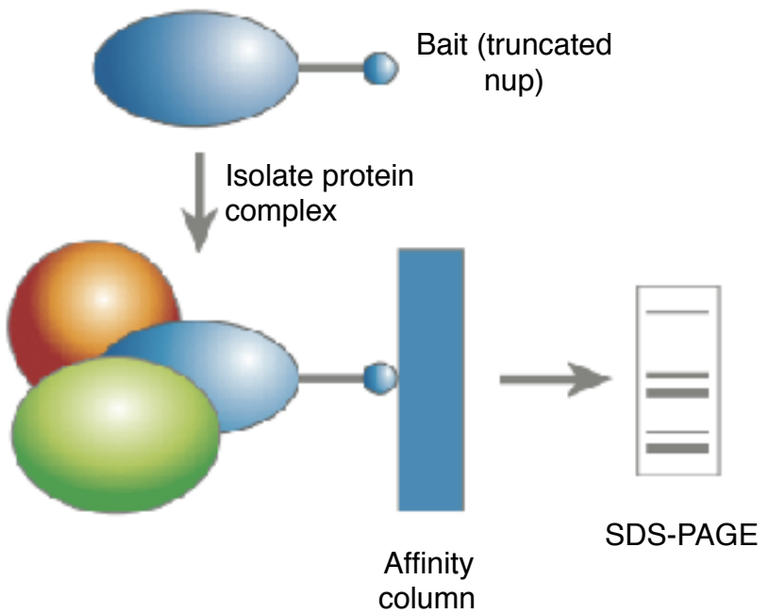
Affinity purification data

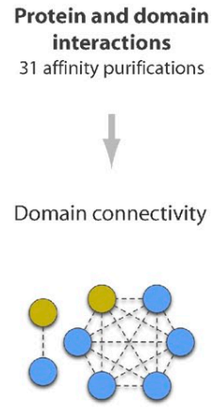
- Domain-truncated nups (NPC subunits) were tagged



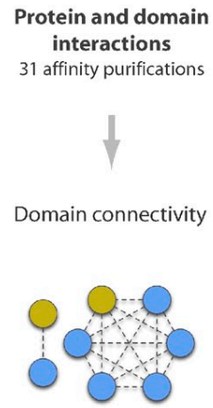
Affinity purification data

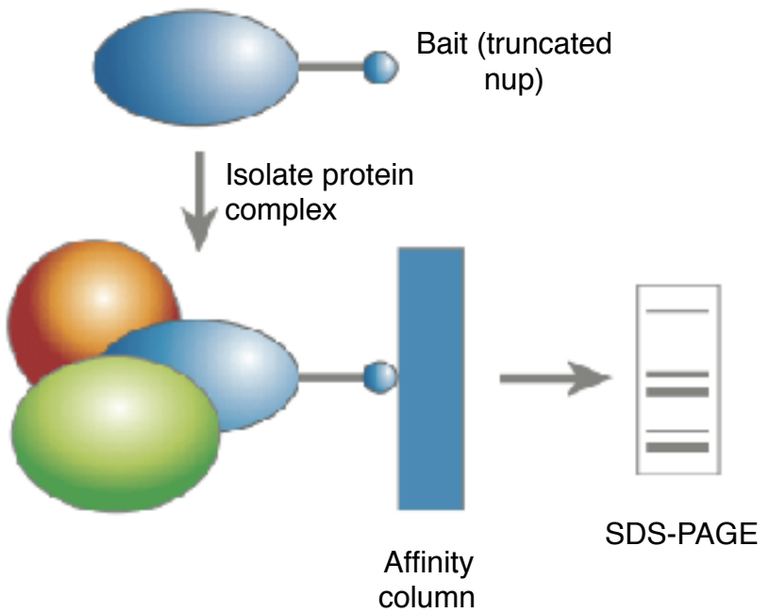
- Domain-truncated nups (NPC subunits) were tagged
- Resulting subcomplexes were affinity purified followed by mass spectrometry





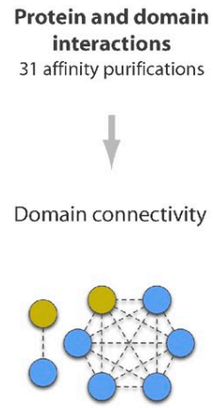
Affinity purification data

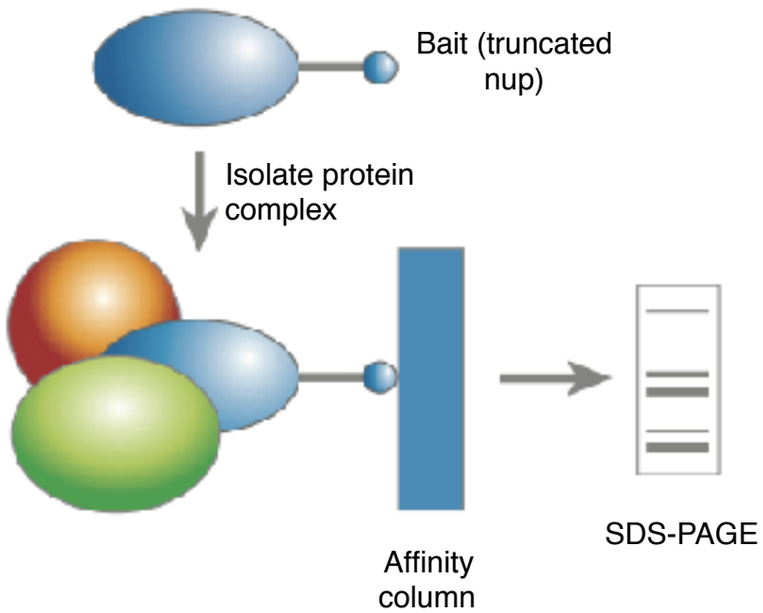




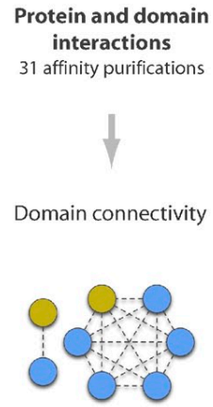
- Domain-truncated nups (NPC subunits) were tagged
- Resulting subcomplexes were affinity purified followed by mass spectrometry
- Any nup that is affinity purified with a truncated nup bait must necessarily interact (directly or indirectly) *in vivo* with the remaining domains

Affinity purification data

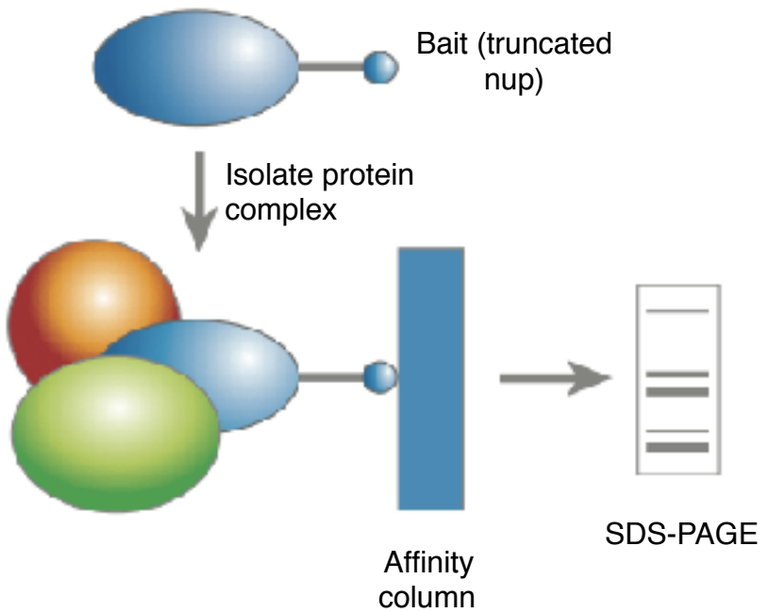




- Domain-truncated nups (NPC subunits) were tagged
- Resulting subcomplexes were affinity purified followed by mass spectrometry
- Any nup that is affinity purified with a truncated nup bait must necessarily interact (directly or indirectly) *in vivo* with the remaining domains

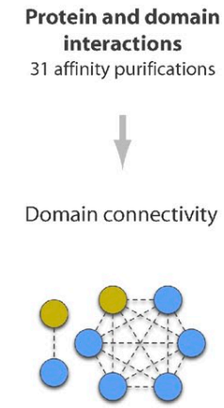


Affinity purification data

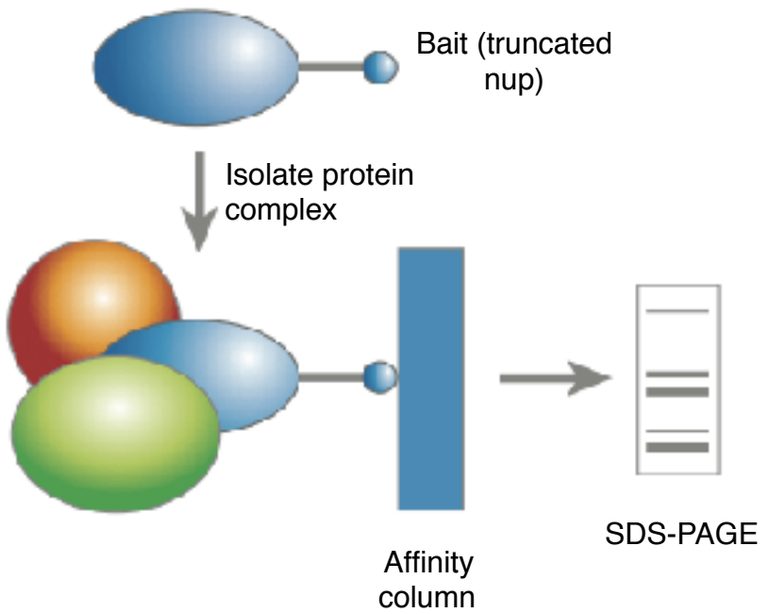


- Domain-truncated nups (NPC subunits) were tagged
- Resulting subcomplexes were affinity purified followed by mass spectrometry
- Any nup that is affinity purified with a truncated nup bait must necessarily interact (directly or indirectly) *in vivo* with the remaining domains

- Computationally, this tells us “Domain A is found in proximity to domains B, C, and D”



Affinity purification data

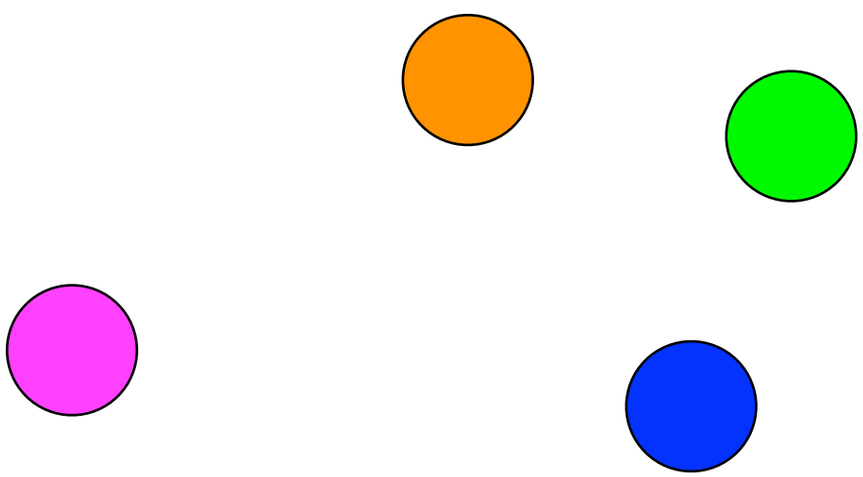


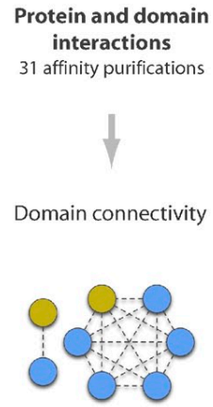
- Domain-truncated nups (NPC subunits) were tagged
- Resulting subcomplexes were affinity purified followed by mass spectrometry
- Any nup that is affinity purified with a truncated nup bait must necessarily interact (directly or indirectly) *in vivo* with the remaining domains

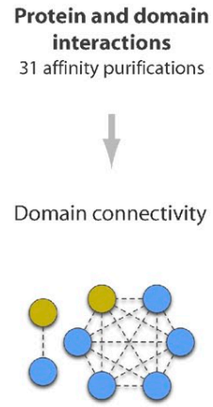
- Computationally, this tells us “Domain A is found in proximity to domains B, C, and D”
- We encode this with a composite (connectivity) restraint

Composite restraint

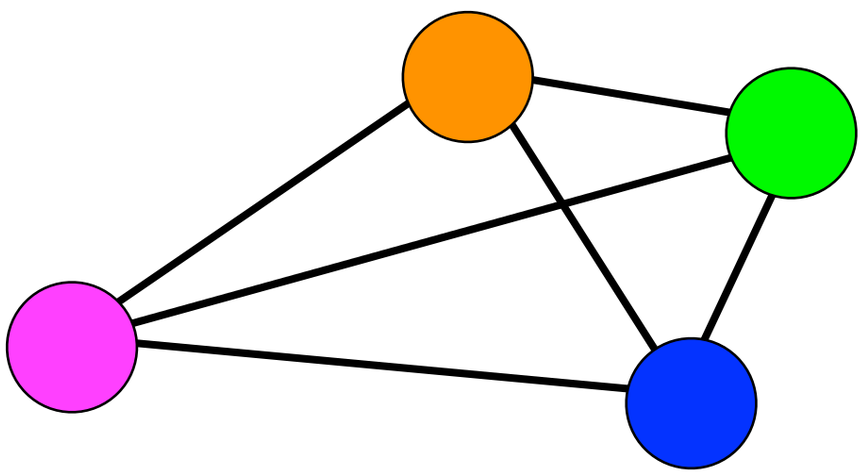
At each scoring step:





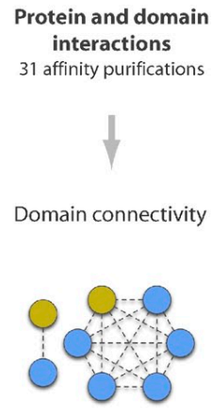


Composite restraint

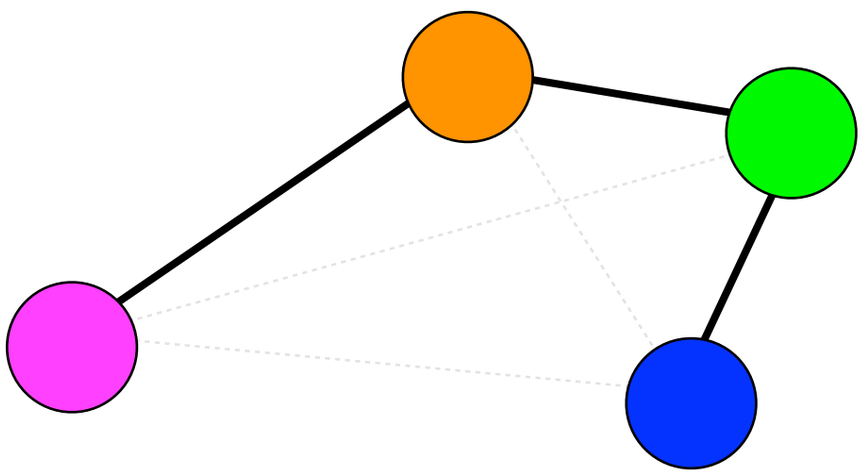


At each scoring step:

1. Determine the fully-connected graph (all pairwise distances between domains in the composite)

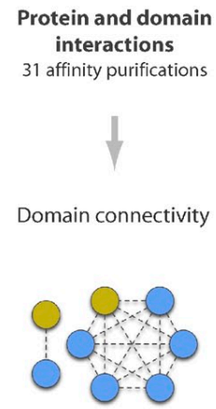


Composite restraint

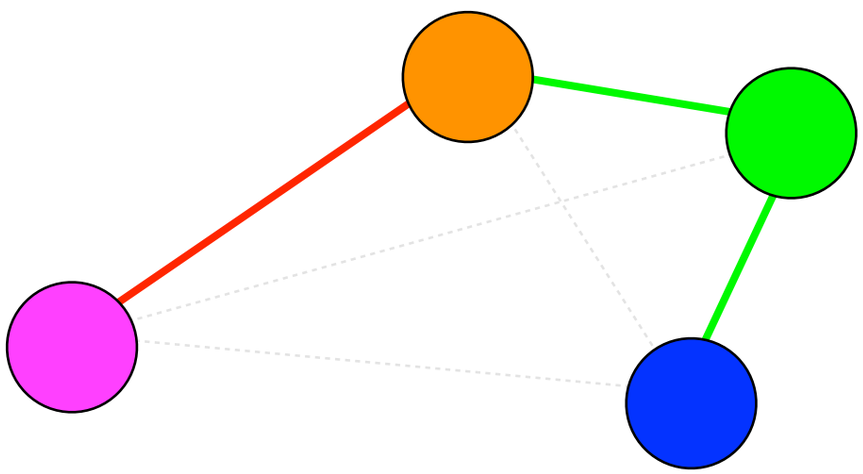


At each scoring step:

1. Determine the fully-connected graph (all pairwise distances between domains in the composite)
2. Reduce to the minimum spanning tree (set of distances that ensures everything is connected and minimizes total distance)

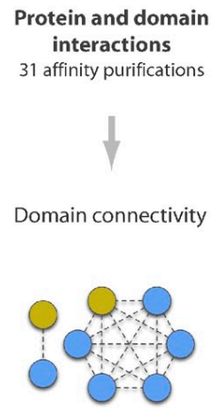


Composite restraint

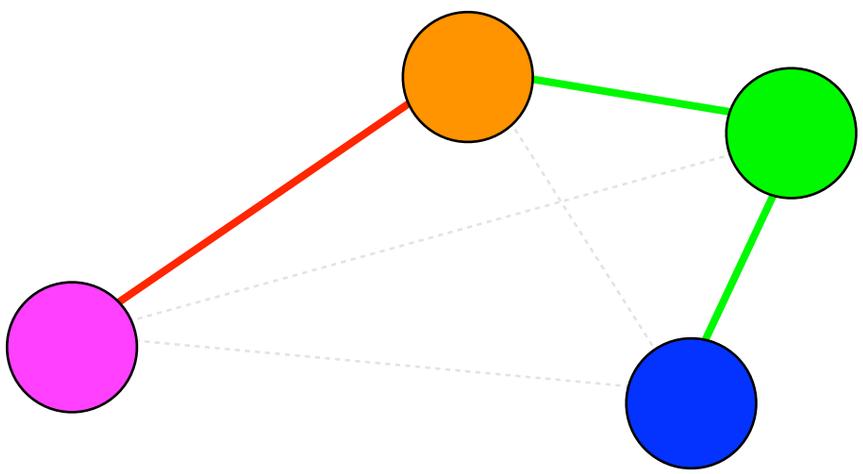


At each scoring step:

1. Determine the fully-connected graph (all pairwise distances between domains in the composite)
2. Reduce to the minimum spanning tree (set of distances that ensures everything is connected and minimizes total distance)
3. Score the resulting set with simple harmonic (“spring”) restraints



Composite restraint

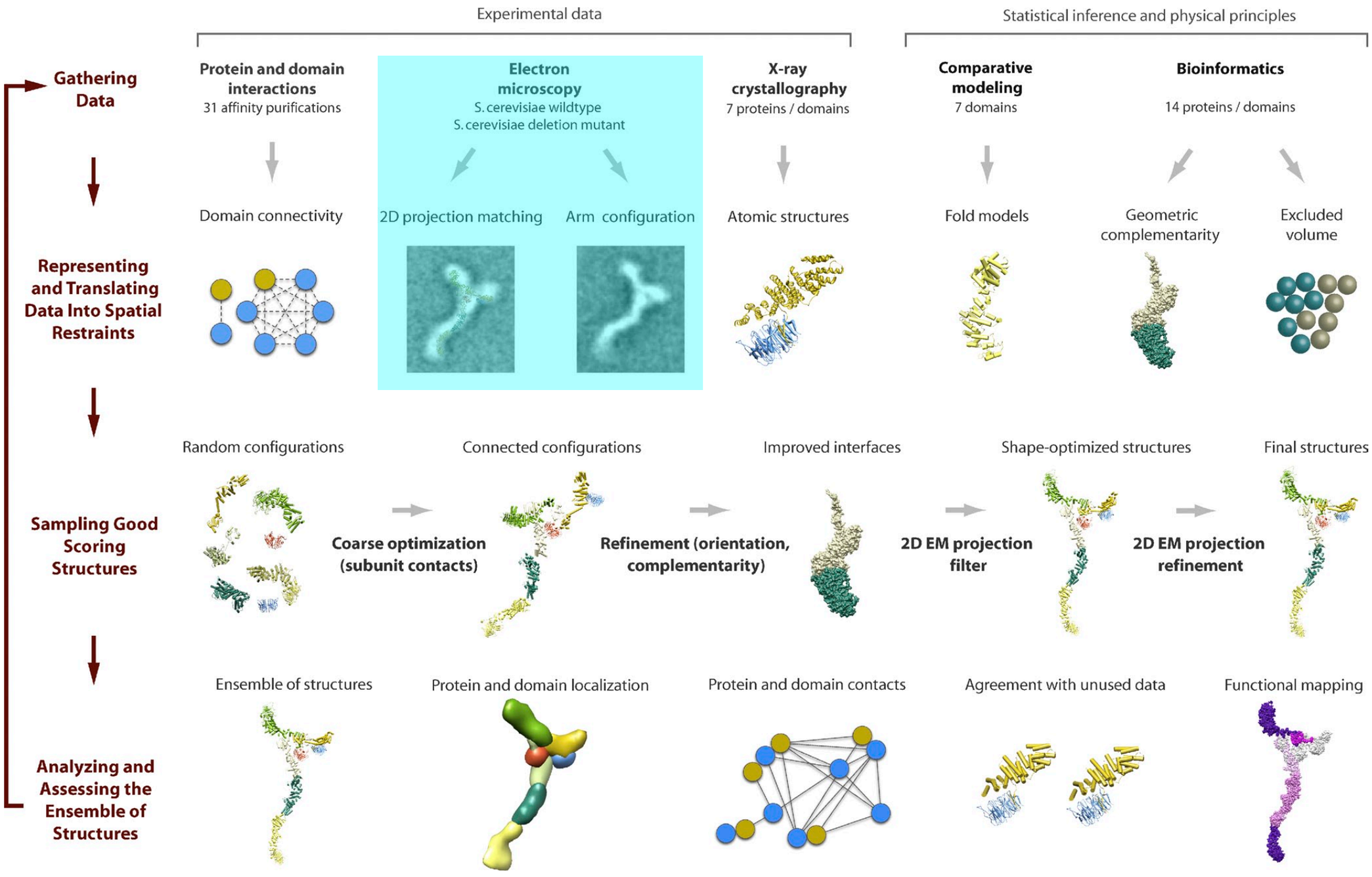


At each scoring step:

1. Determine the fully-connected graph (all pairwise distances between domains in the composite)
2. Reduce to the minimum spanning tree (set of distances that ensures everything is connected and minimizes total distance)
3. Score the resulting set with simple harmonic (“spring”) restraints

Thus, the composite restraint is satisfied when the set of domains is connected, without enforcing the *order* in which they are connected (and this order can change from step to step during a simulation).

Modeling Nup84 with IMP (2012)



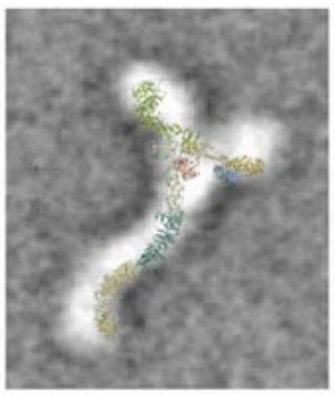
J Cell Biol 196, 419-434, 2012

Electron microscopy

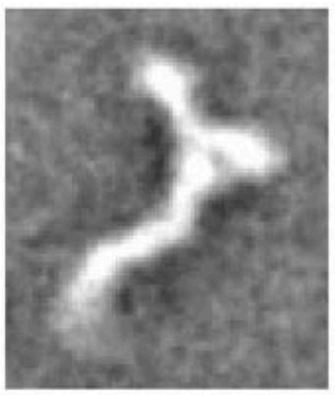
S. cerevisiae wildtype

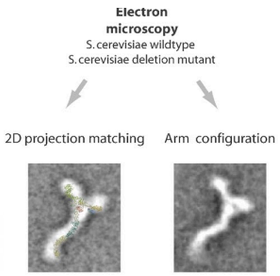
S. cerevisiae deletion mutant

2D projection matching

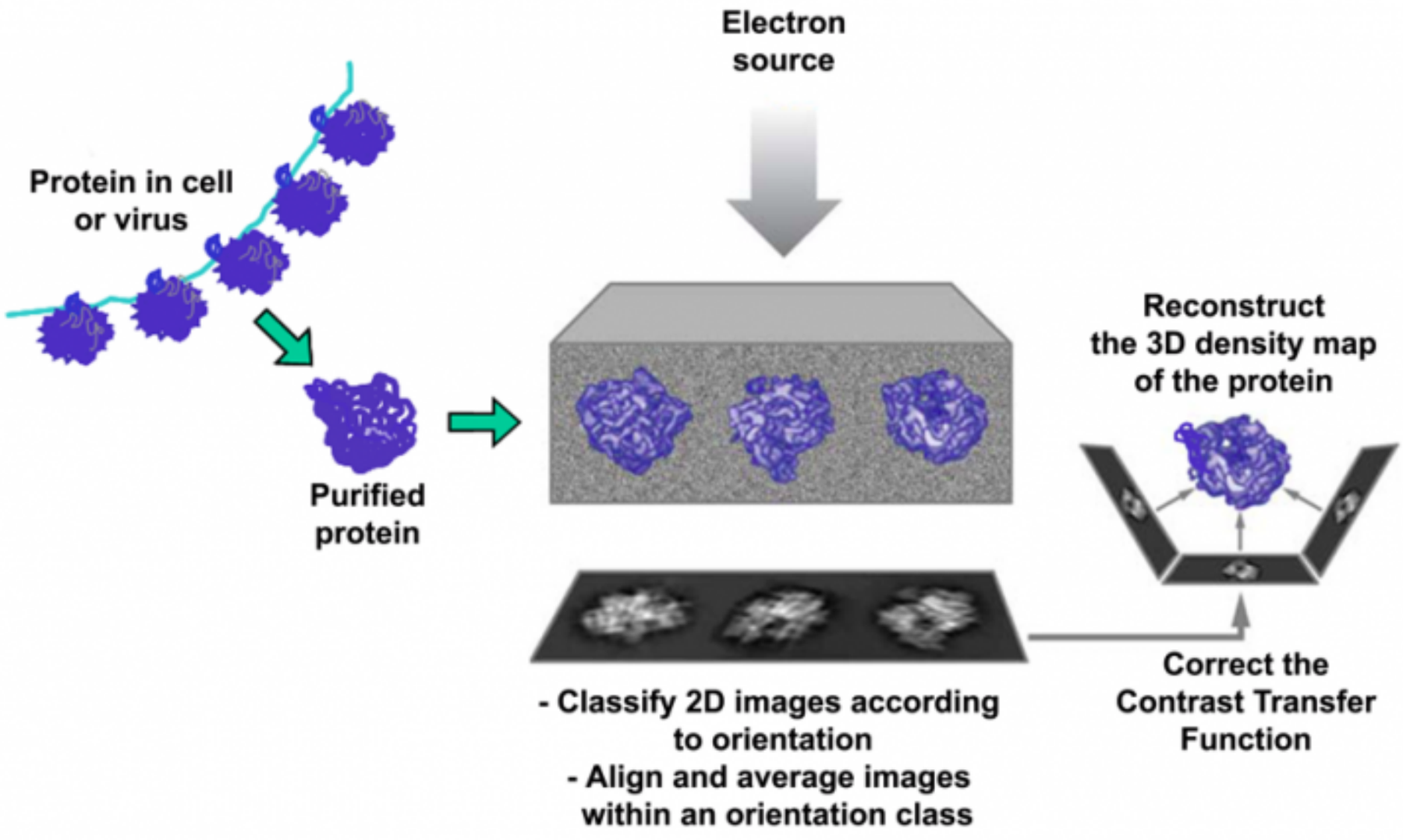


Arm configuration



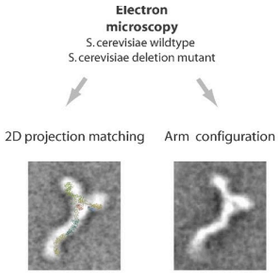


EM

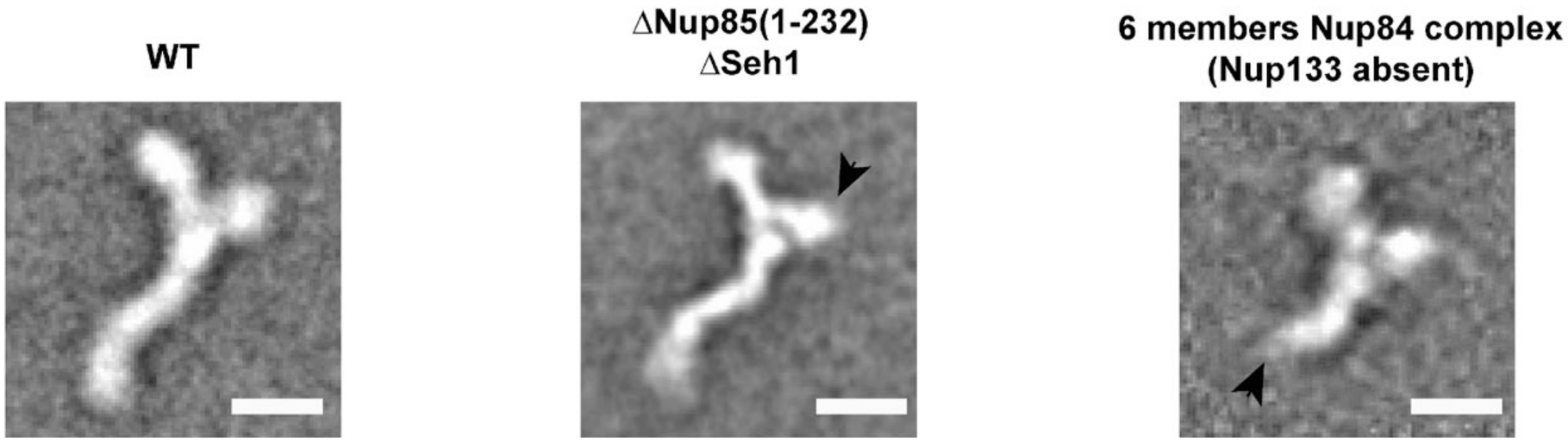


- Significant processing required to generate a 3D map
- Here we skip the last step and so end up with 2D class averages

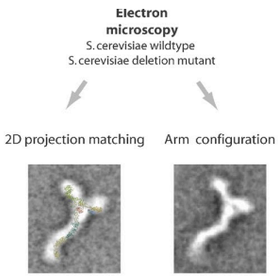
EM data for Nup84



- 2D class average of the entire complex shows overall “Y” shape
- Class averages of truncation complexes suggest locations of the removed subunits by identification of the missing density

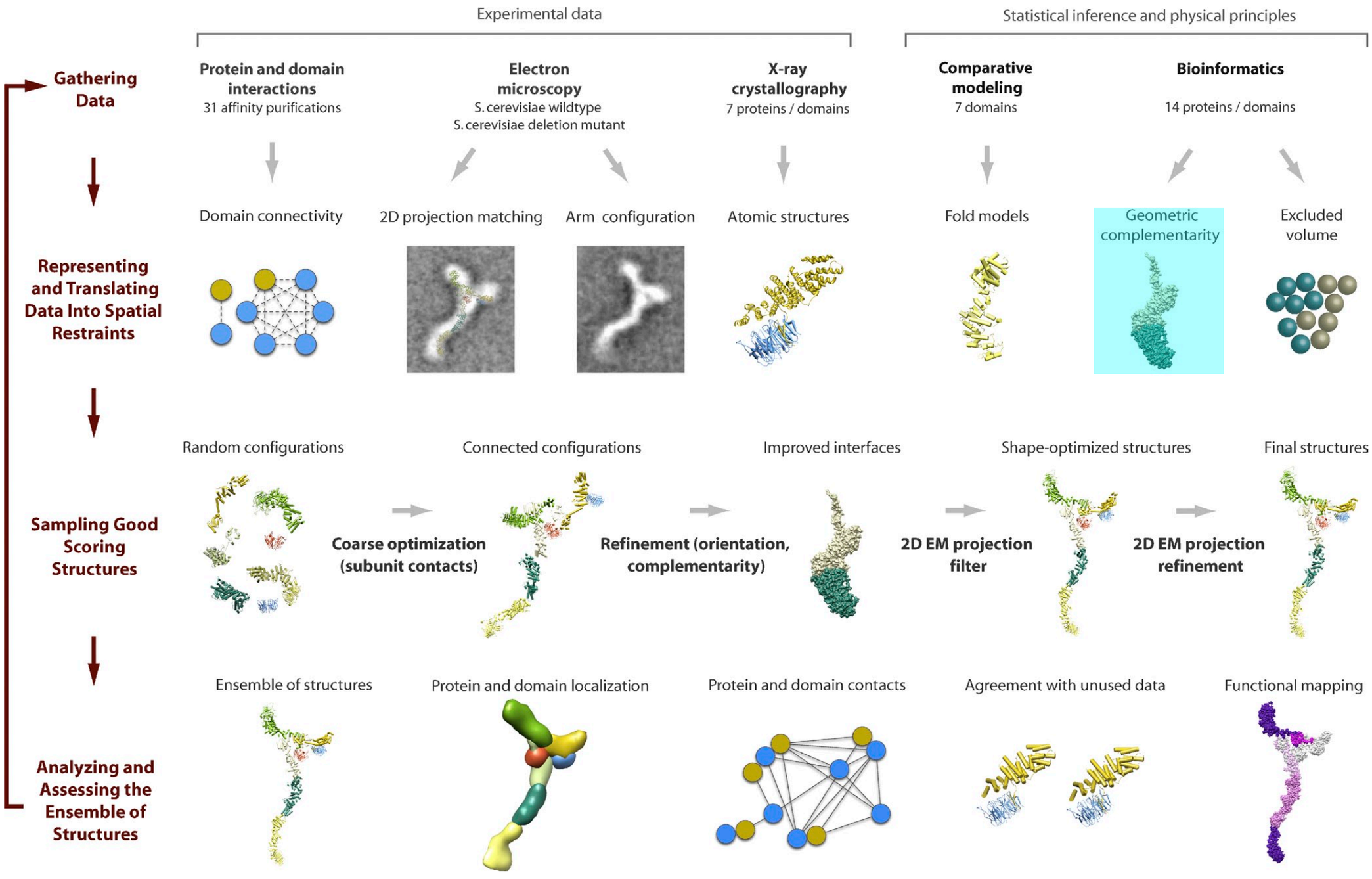


EM2D score



- Complete 2D EM map used as a *restraint*:
  - Calculate a number of evenly-spaced 2D projections of the model
  - Align each class average with each projection
  - Calculate cross correlation coefficient for each alignment
    - Final score is the negative log of the highest cross-correlation coefficient
- Truncated maps used as a *filter*:
  - Discard models that place subunits in impossible regions, as determined by the difference maps

Modeling Nup84 with IMP (2012)



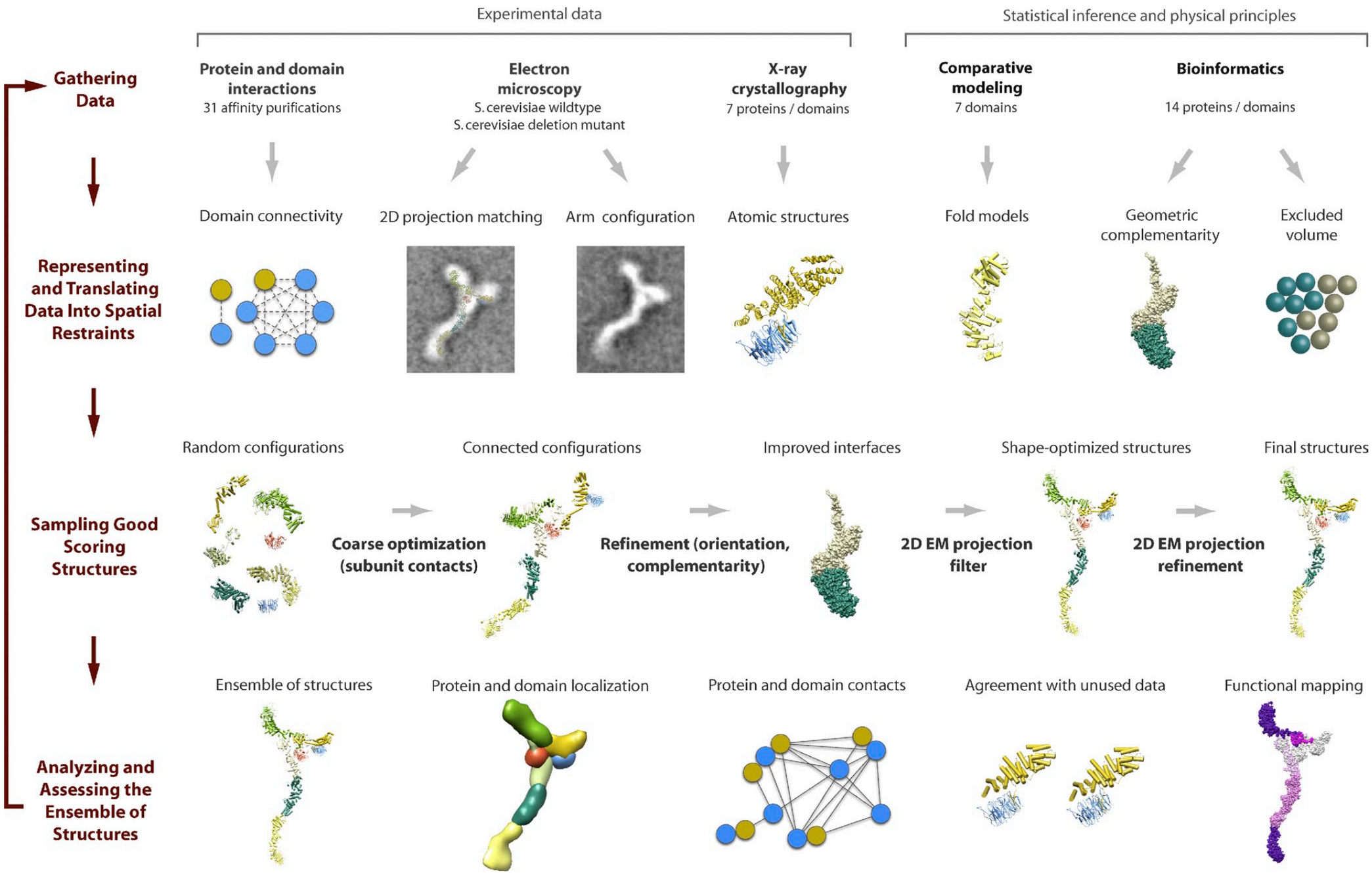
J Cell Biol 196, 419-434, 2012

Geometric
complementarity

Geometric complementarity

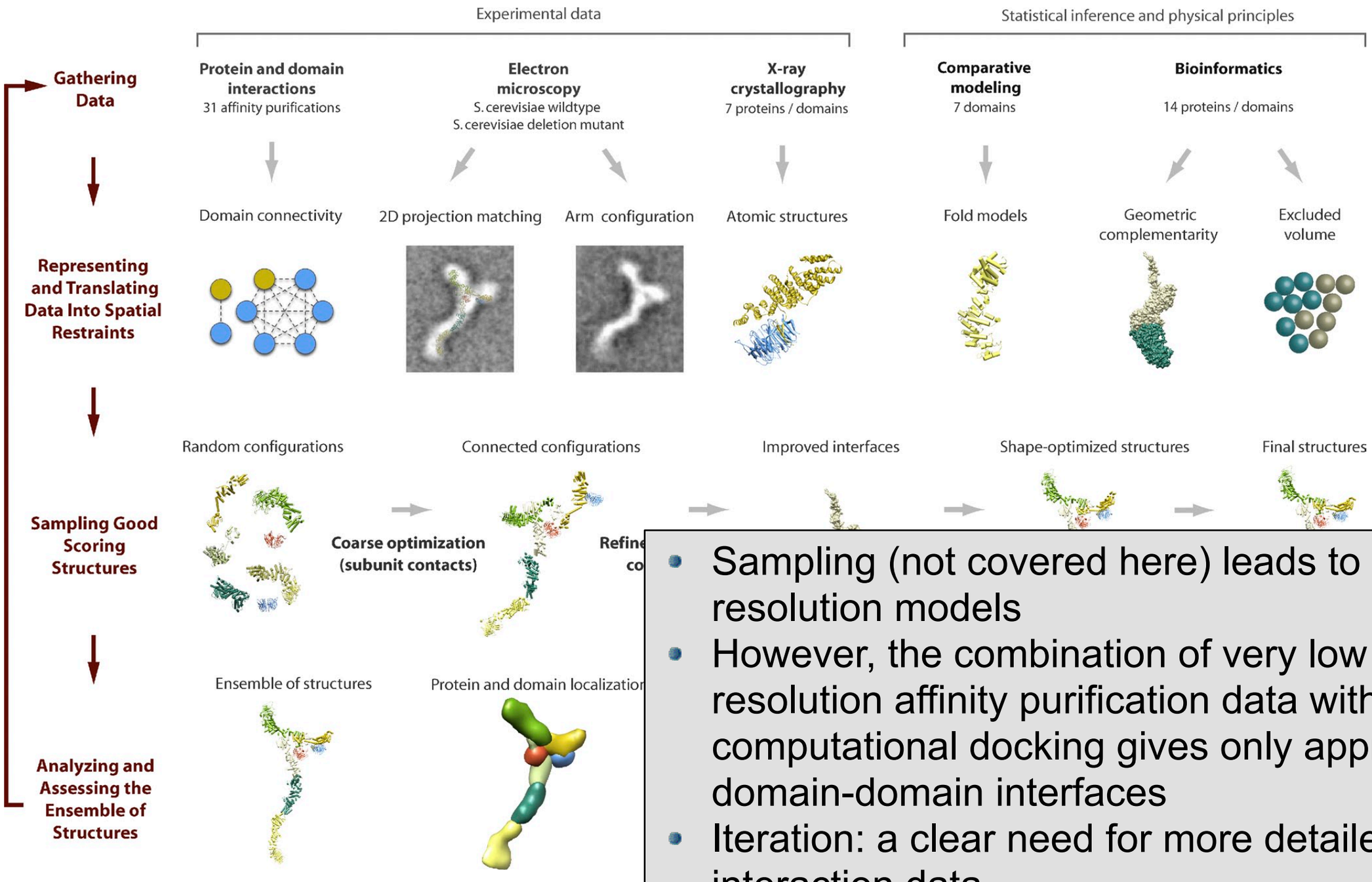
- Basically, computational docking
- Score pairs of proteins based on their shape complementarity and penalize any overlaps
- Helps to get correct orientation of subunits, which is not strongly constrained by the other data

Modeling Nup84 with IMP (2012)



J Cell Biol 196, 419-434, 2012

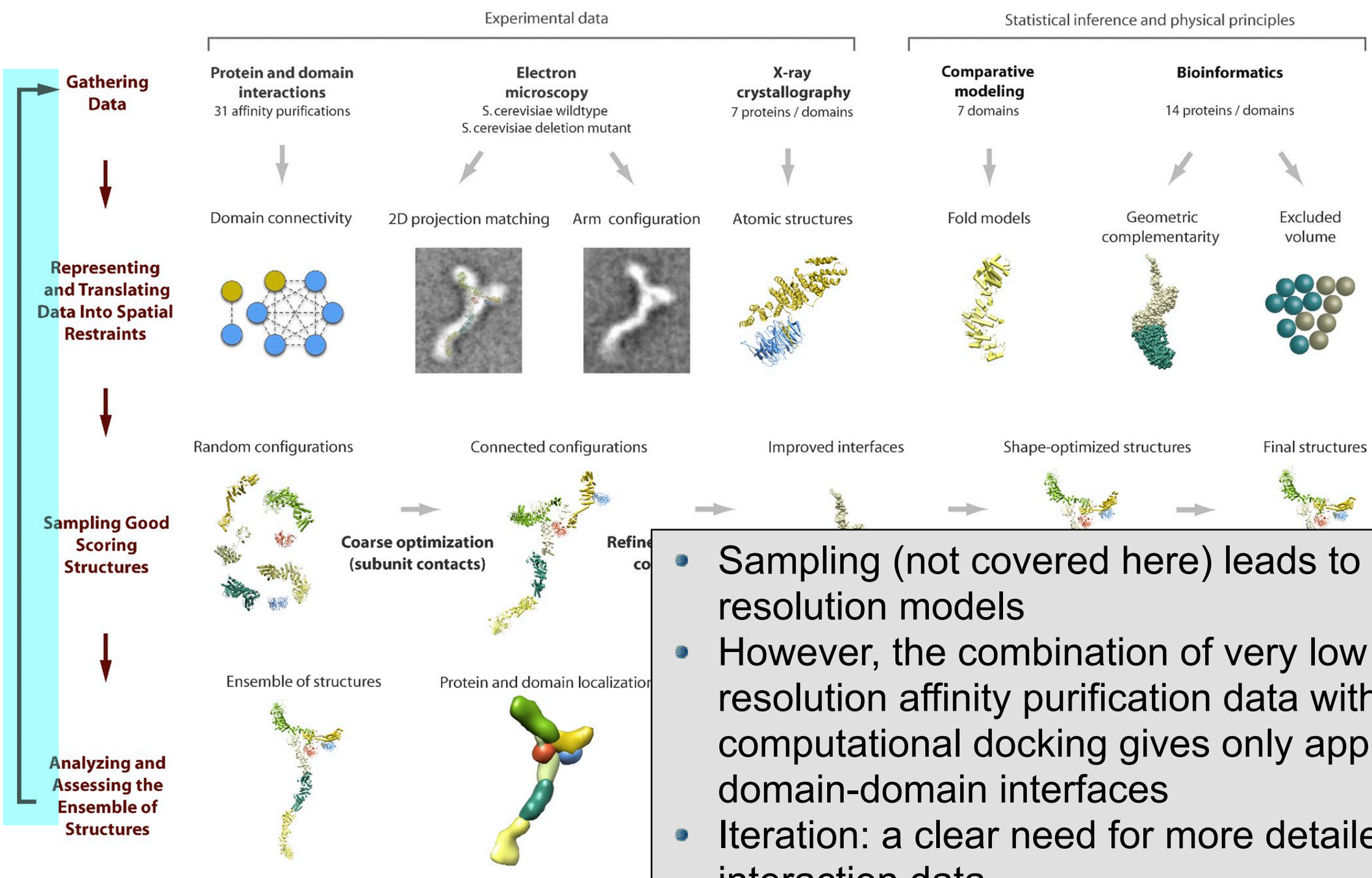
Modeling Nup84 with IMP (2012)



- Sampling (not covered here) leads to medium resolution models
- However, the combination of very low resolution affinity purification data with computational docking gives only approximate domain-domain interfaces
- Iteration: a clear need for more detailed interaction data

J Cell Biol 196, 419-434, 2012

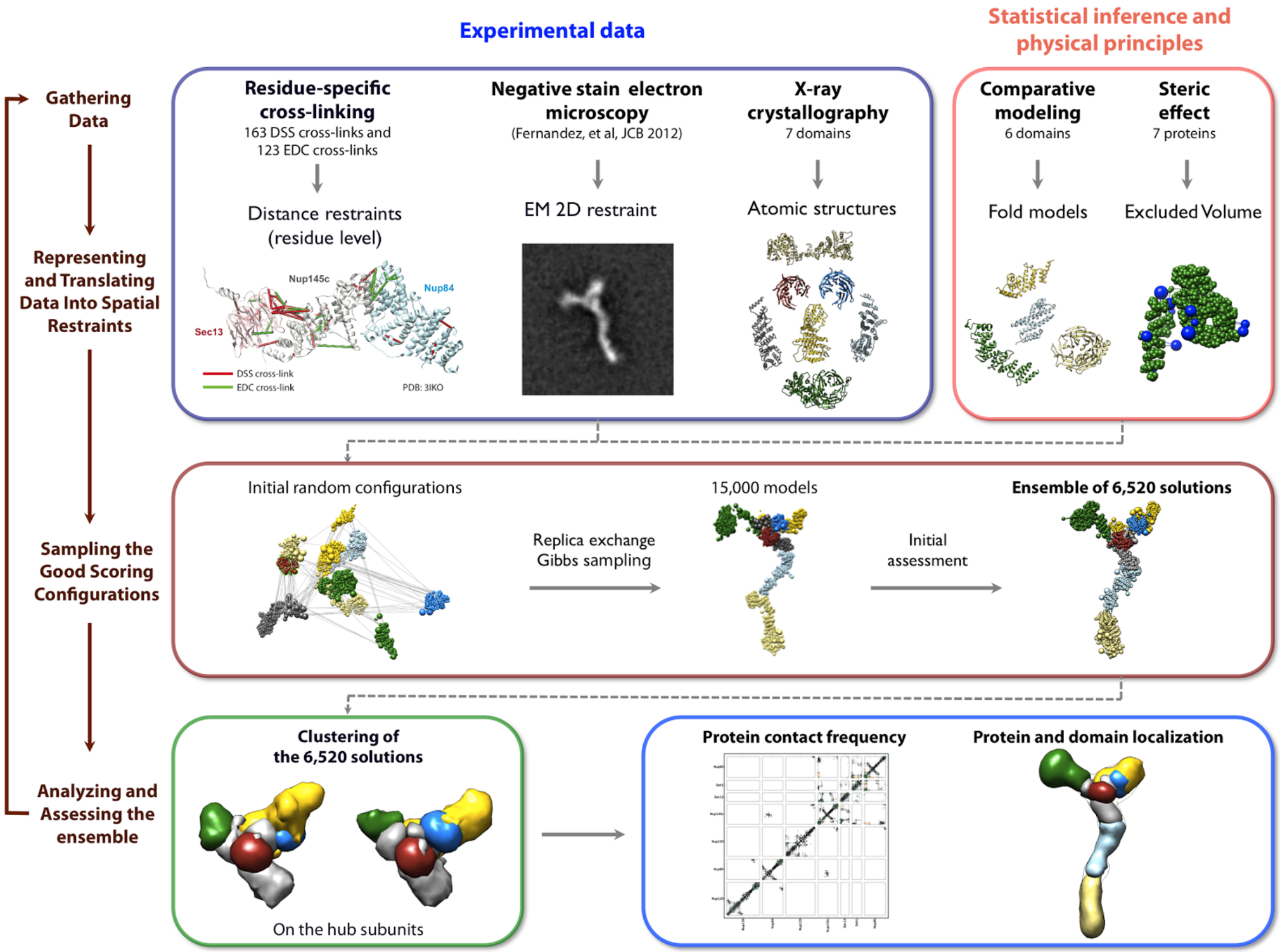
Modeling Nup84 with IMP (2012)



- Sampling (not covered here) leads to medium resolution models
- However, the combination of very low resolution affinity purification data with computational docking gives only approximate domain-domain interfaces
- Iteration: a clear need for more detailed interaction data

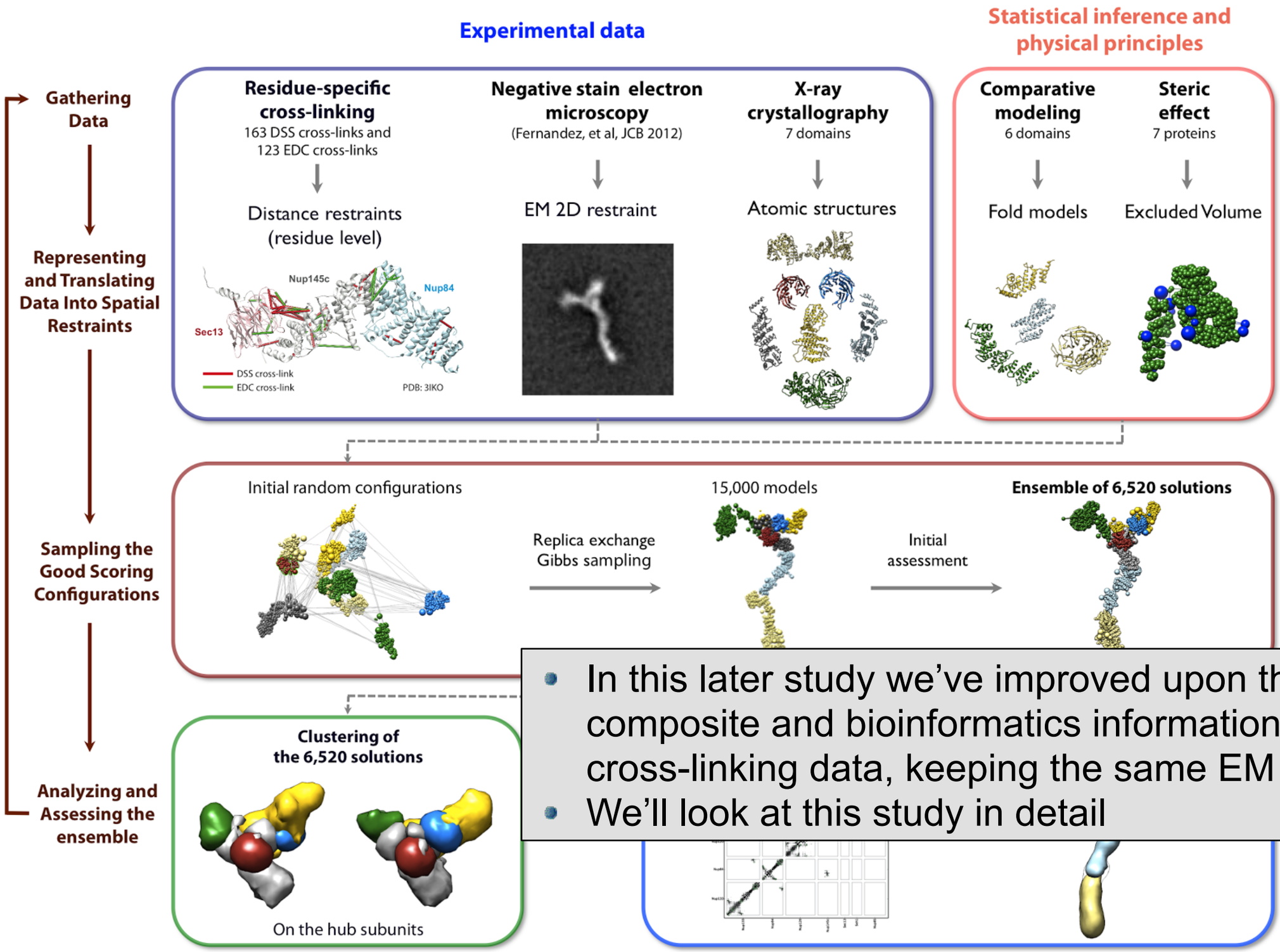
J Cell Biol 196, 419-434, 2012

Modeling Nup84 with IMP (2014)



Mol Cell Proteomics 13, 2927-2943, 2014

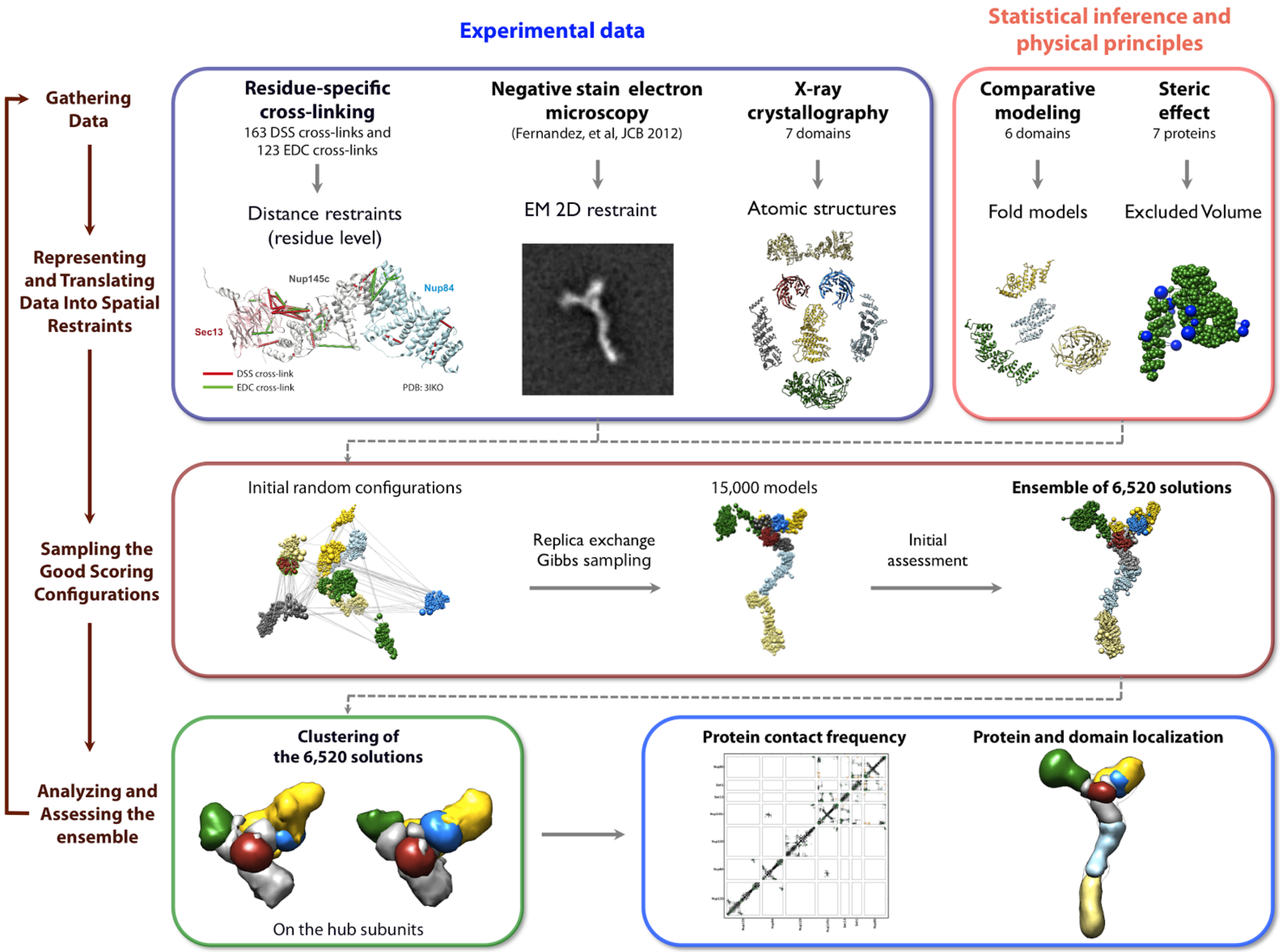
Modeling Nup84 with IMP (2014)



- In this later study we've improved upon the composite and bioinformatics information with cross-linking data, keeping the same EM image
- We'll look at this study in detail

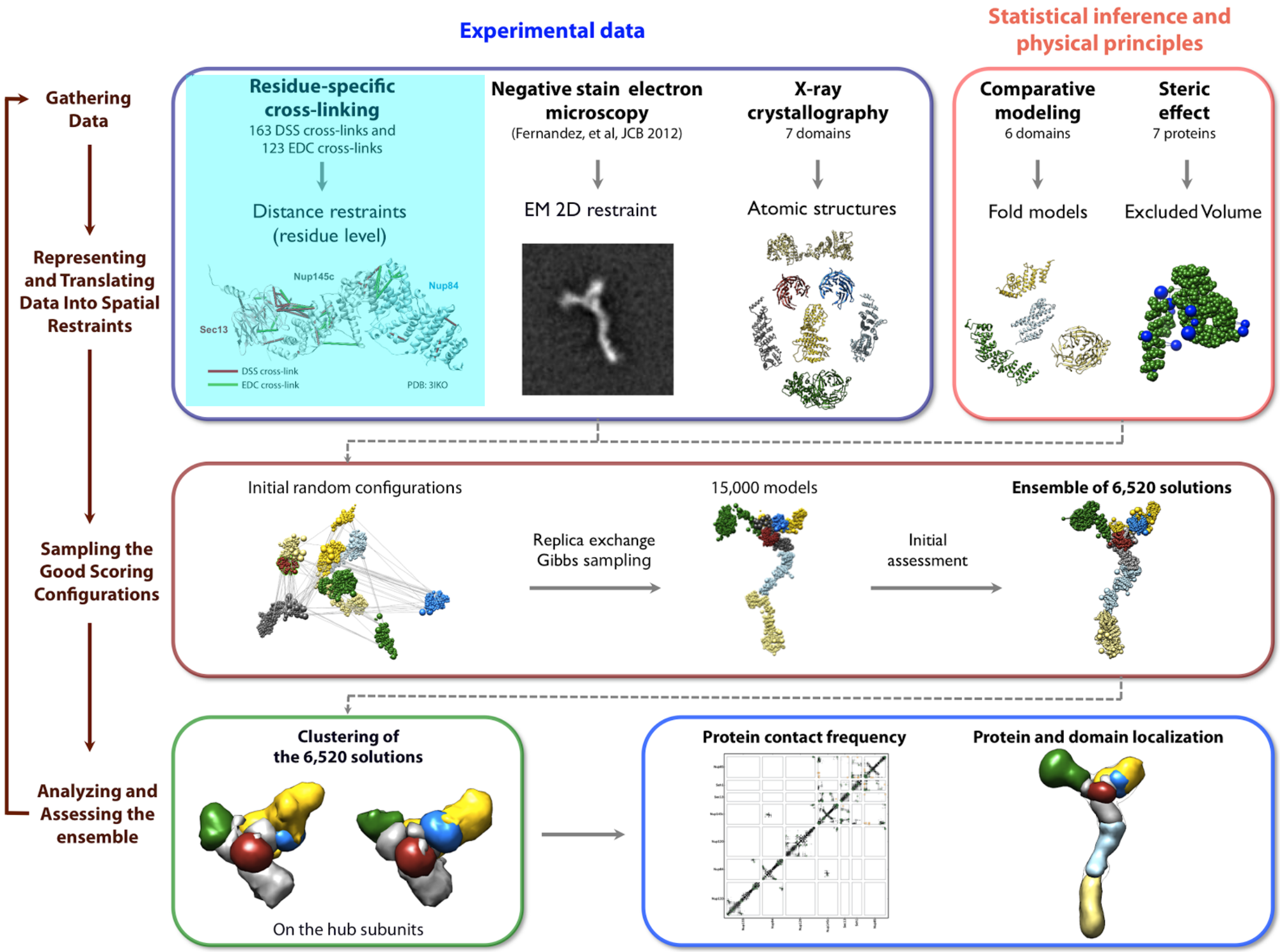
Mol Cell Proteomics 13, 2927-2943, 2014

Modeling Nup84 with IMP (2014)



Mol Cell Proteomics 13, 2927-2943, 2014

Modeling Nup84 with IMP (2014)



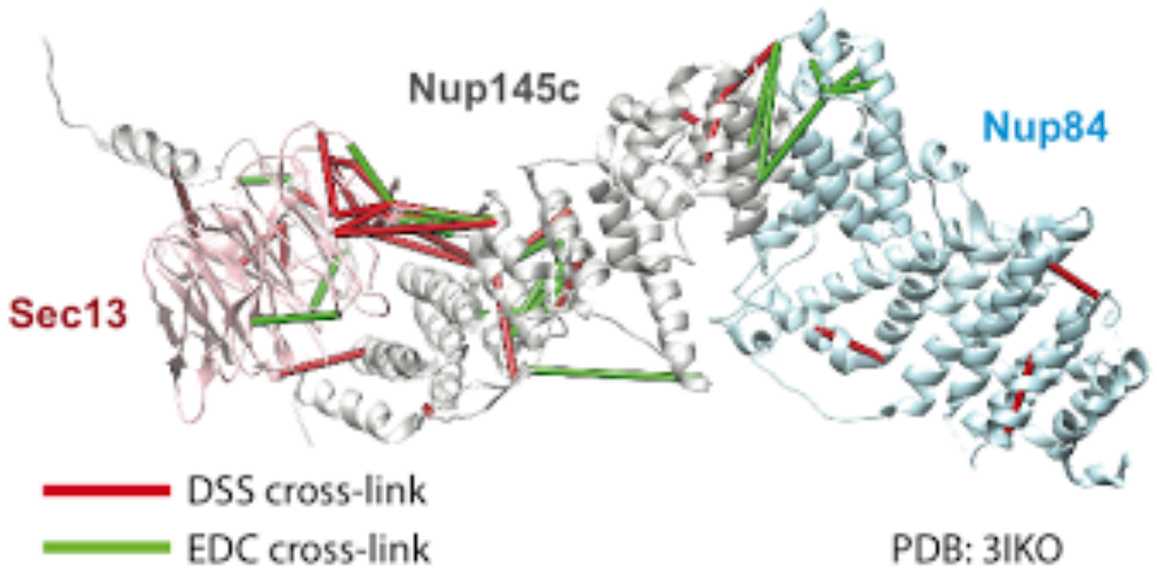
Mol Cell Proteomics 13, 2927-2943, 2014

Residue-specific cross-linking

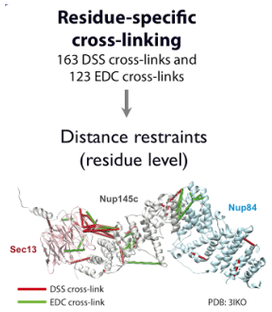
163 DSS cross-links and
123 EDC cross-links

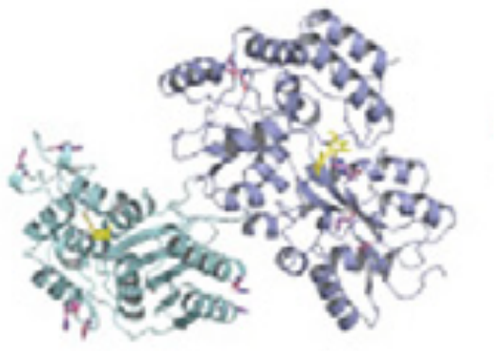
↓

Distance restraints
(residue level)



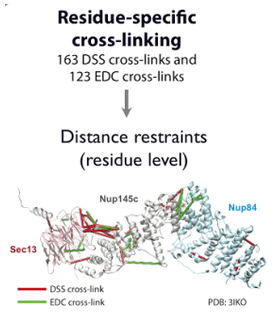
Cross-linking coupled with mass spectrometry (CX-MS)

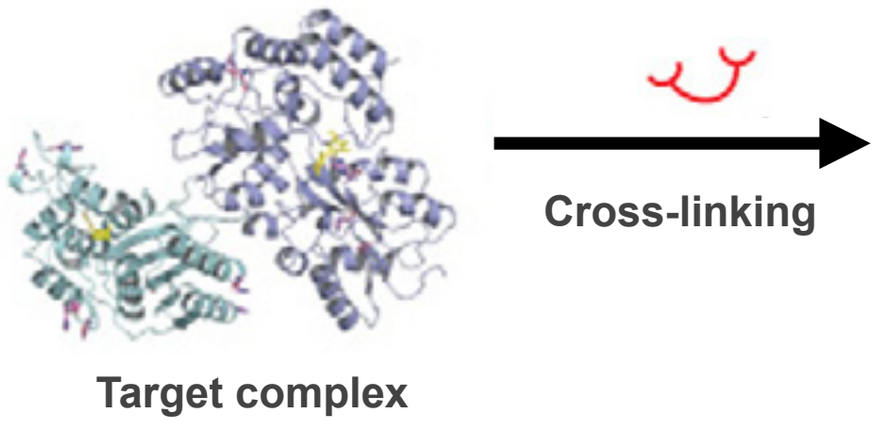




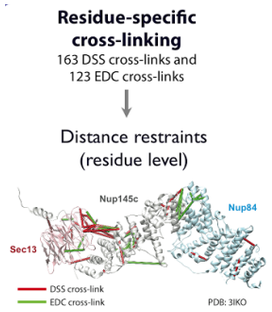
Target complex

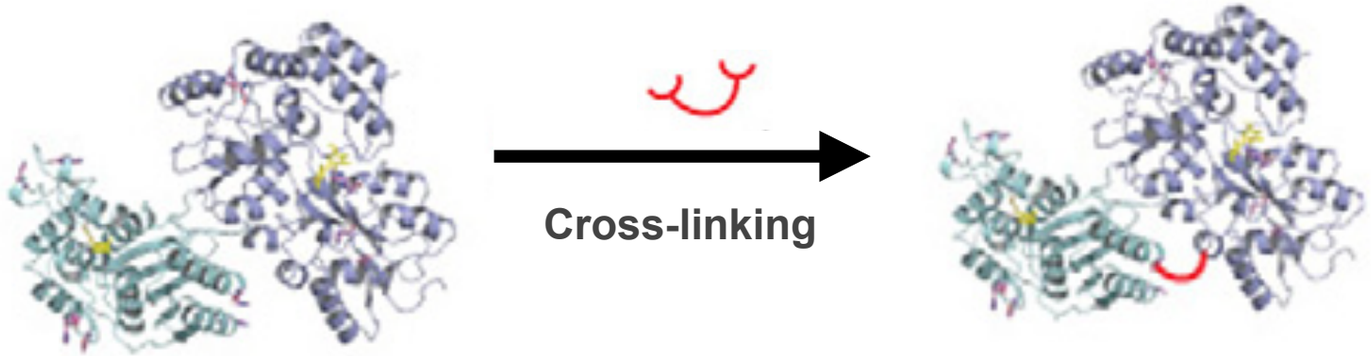
Cross-linking coupled with mass spectrometry (CX-MS)



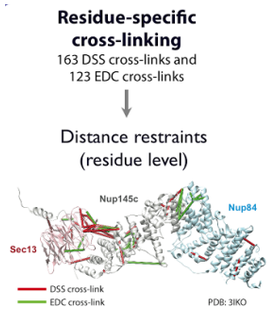


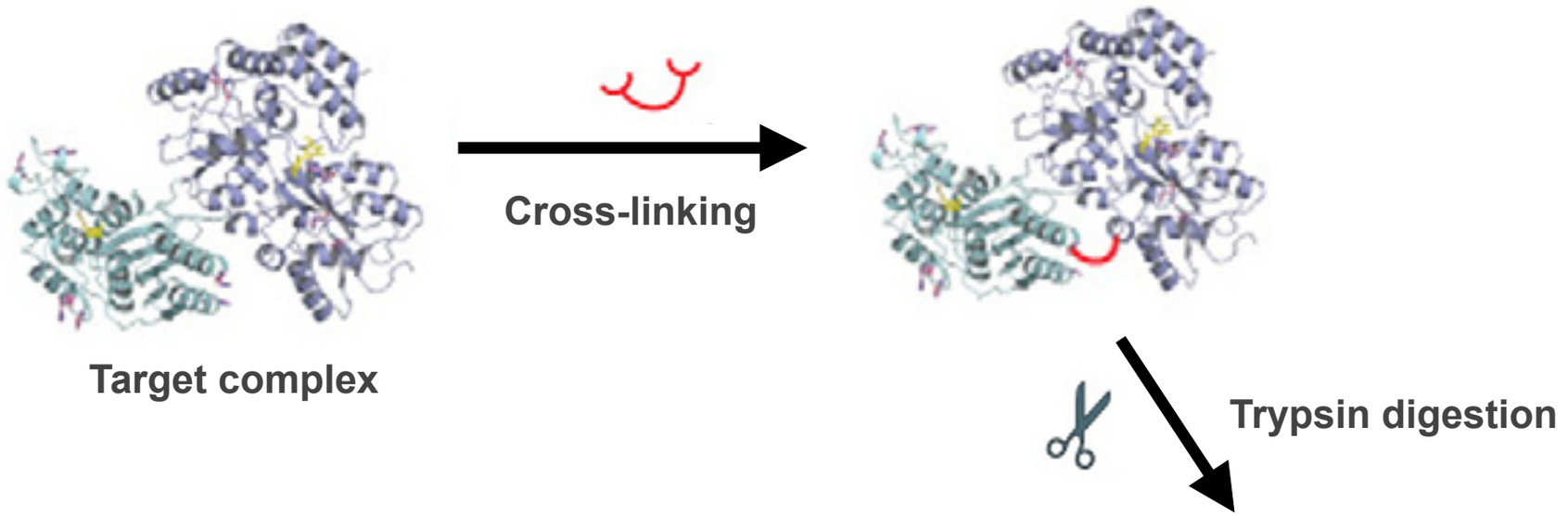
Cross-linking coupled with mass spectrometry (CX-MS)



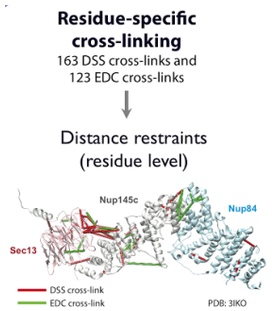


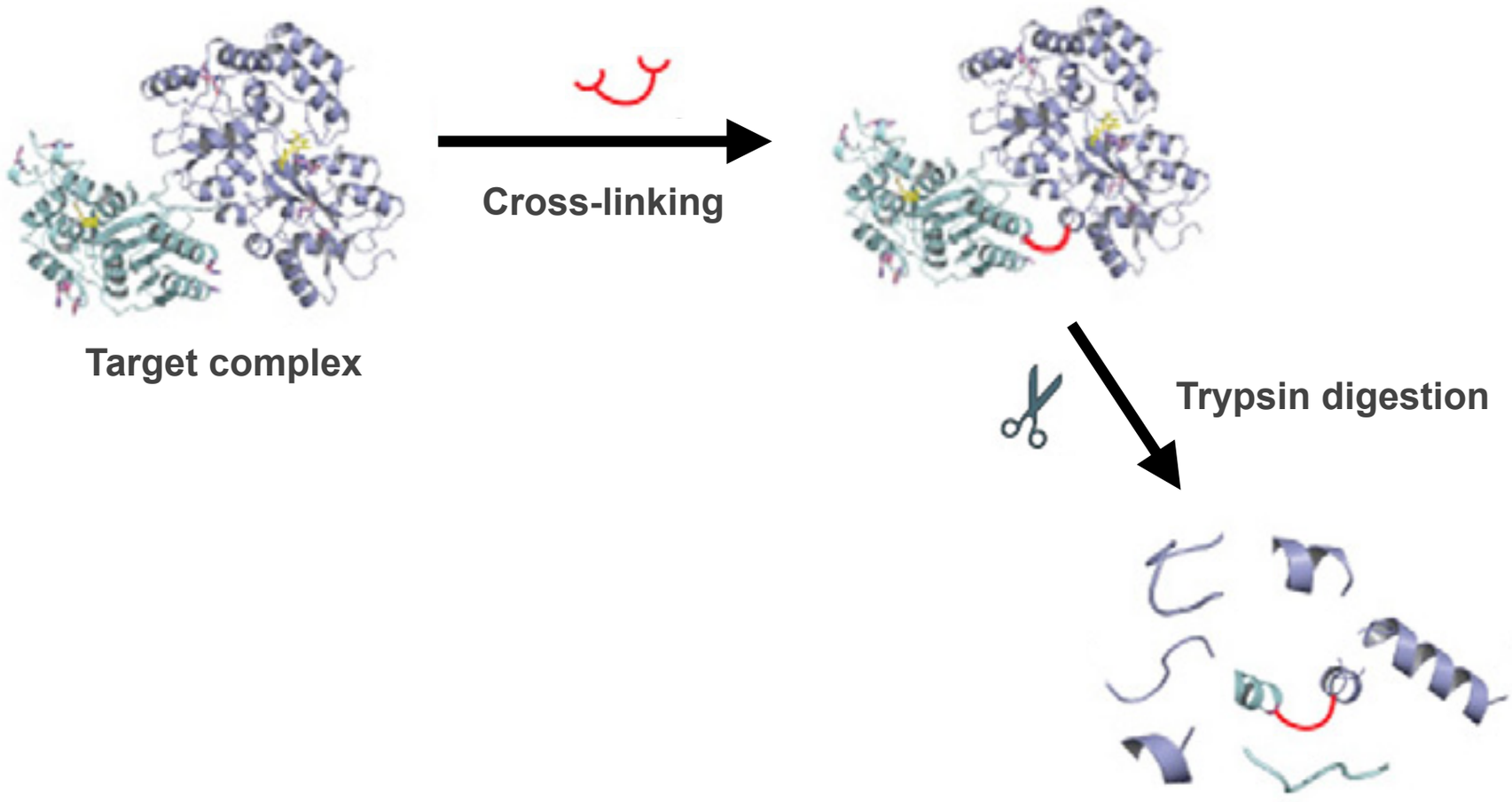
Cross-linking coupled with mass spectrometry (CX-MS)



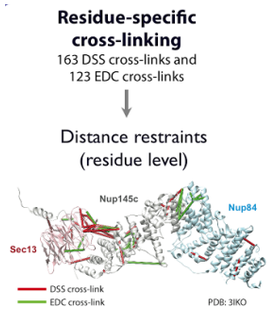


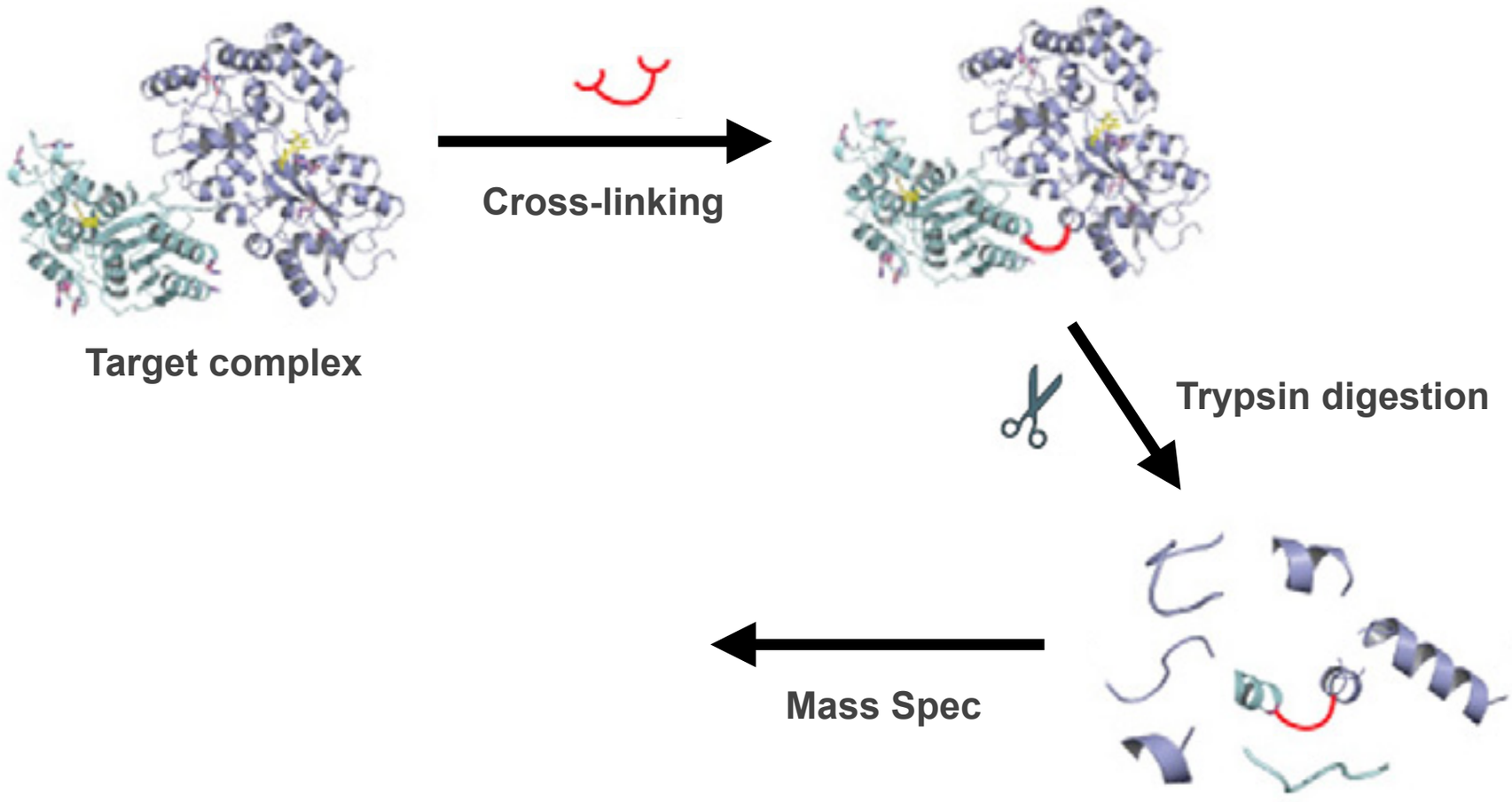
Cross-linking coupled with mass spectrometry (CX-MS)



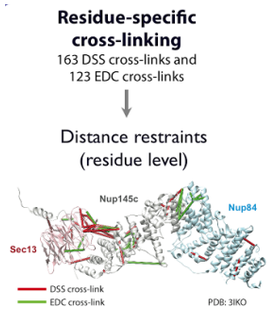


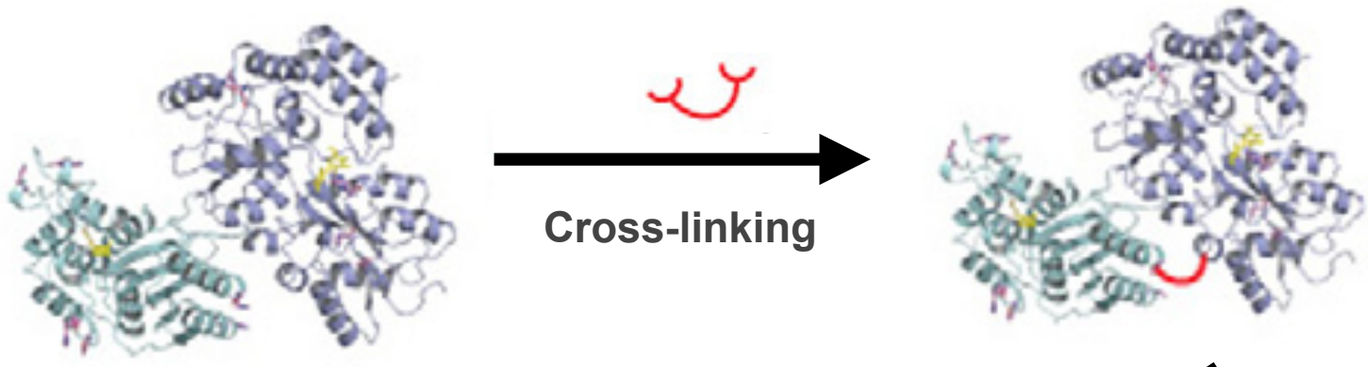
Cross-linking coupled with mass spectrometry (CX-MS)

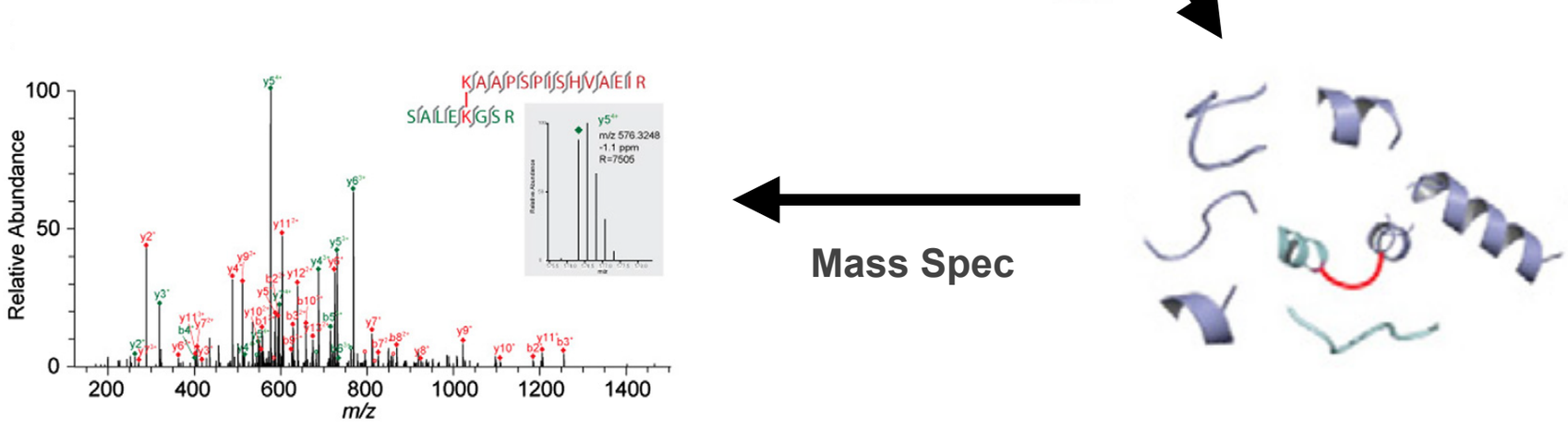




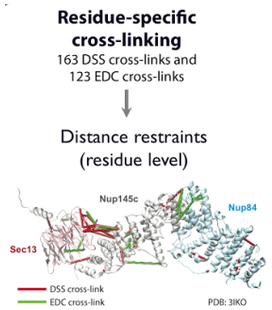
Cross-linking coupled with mass spectrometry (CX-MS)

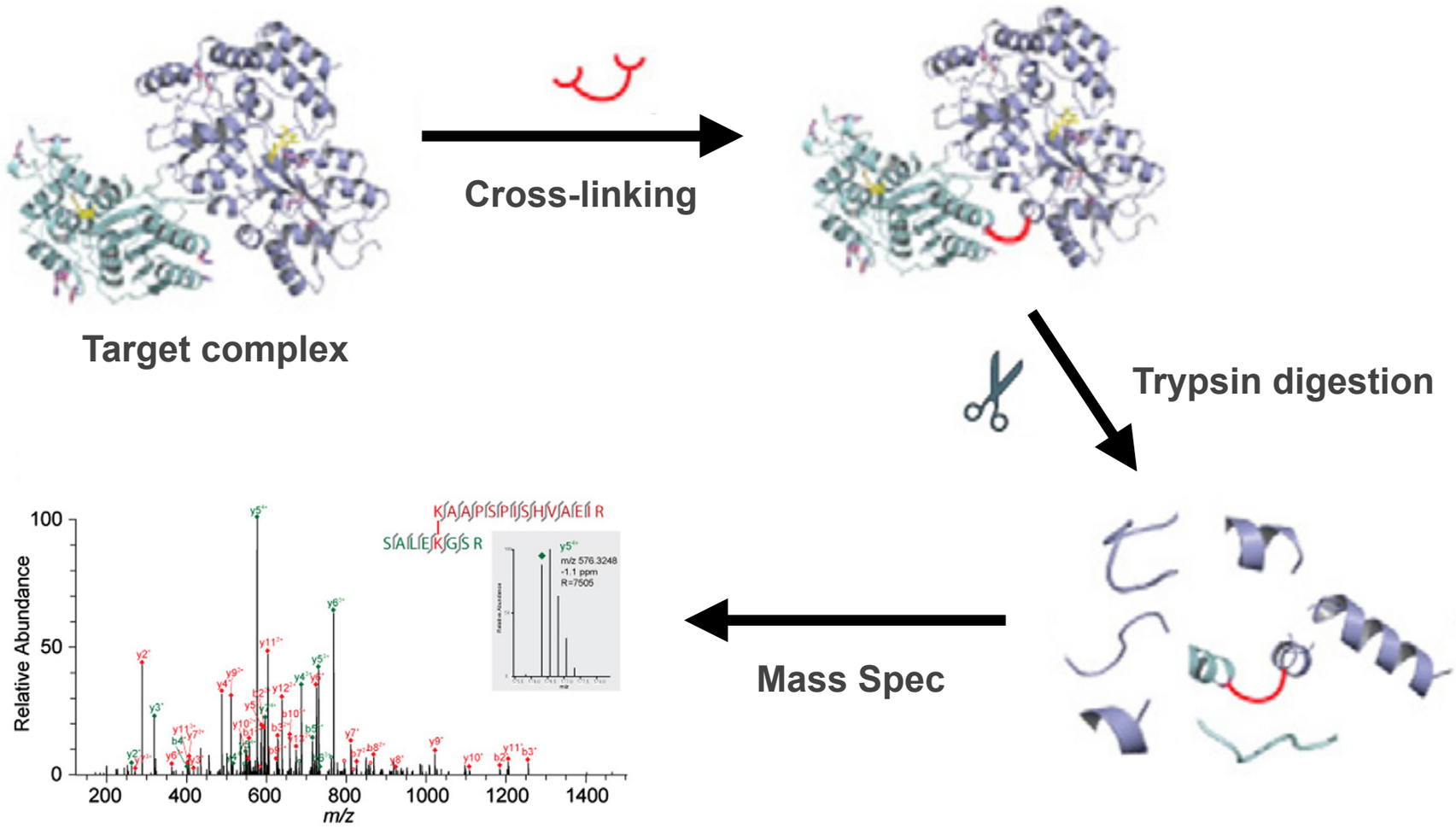






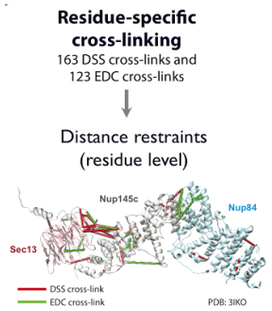
Cross-linking coupled with mass spectrometry (CX-MS)





Output, essentially, is a list of proximal residue pairs (again, after processing)

Cross-linking Nup84 protocol



Cell cryolysis of tagged strain
(Oeffinger *et al.* 2007)

↓

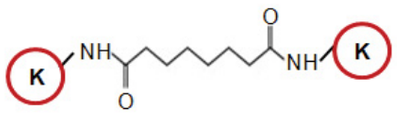
Affinity capture and **native elution** of endogenous complex (10-20 µg)
(Fernandez-Martinez *et. al.* 2012)

Endogenous complex purification

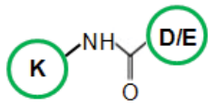
↓

Cross-linking of the purified complex with **complementary crosslinking reagents**

DSS
(primary amine to primary amine)



EDC
(primary amine to carboxilic acid)



Cross-linking and MS

In-solution digestion | **In-gel digestion**

Peptide size exclusion chromatography
(Leitner *et. al.* 2012)

↓

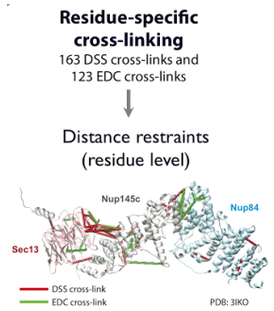
LC/MS
(fragmentation by HCD)

↓

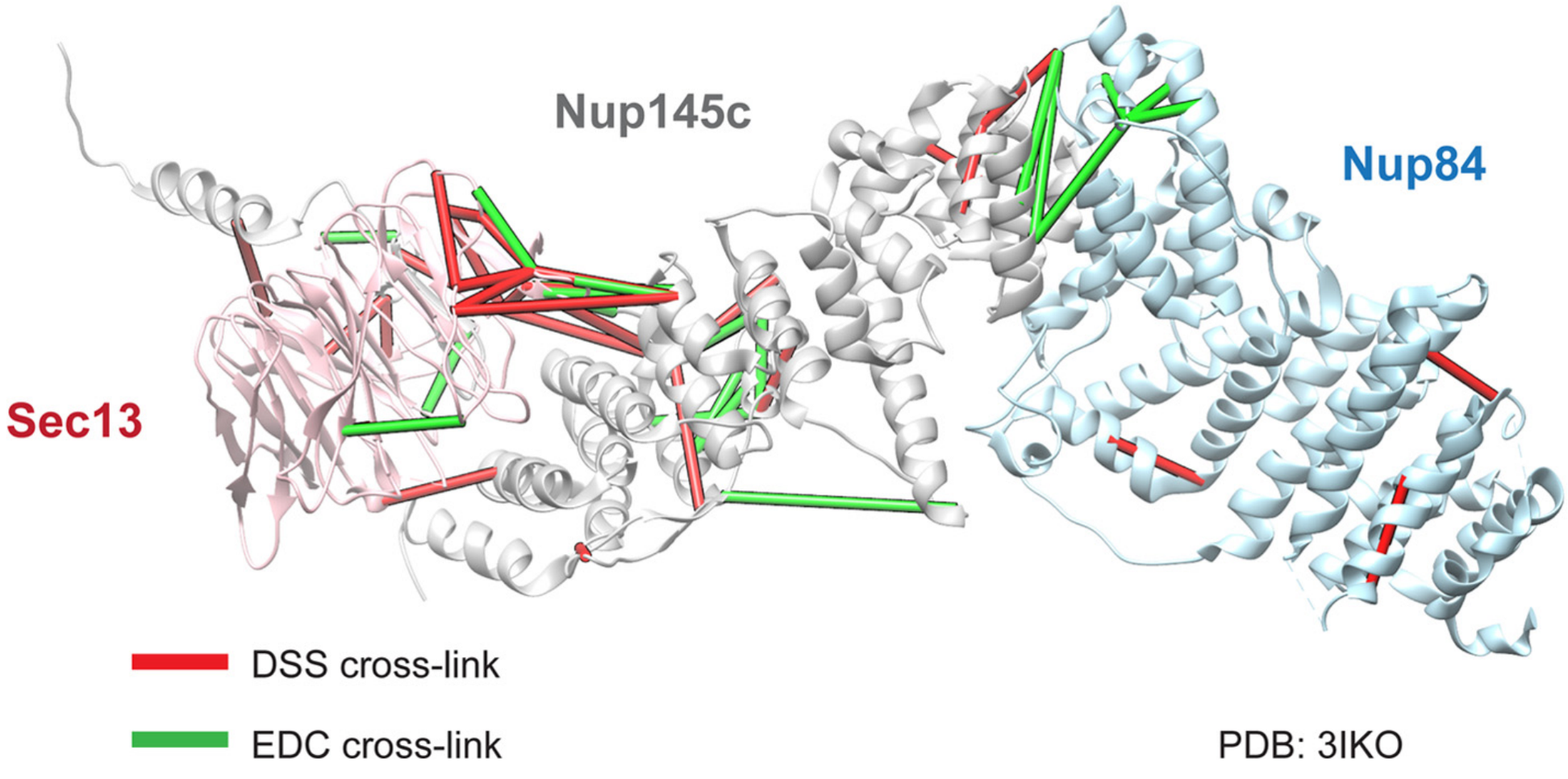
Cross-link identification by pLink
(Yang *et. al.* 2012)
and **spectra verification**

- Two cross-linkers used to probe the Nup84 complex: DSS and EDC
  - Both ends of DSS react with lysine (or N-terminus)
  - EDC cross-links amines to carboxylic acids (aspartic acid, glutamic acid, C-terminus)

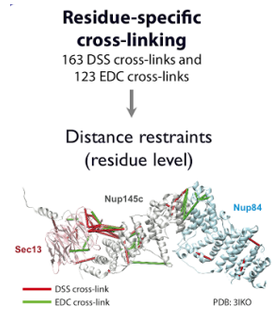
Cross-linkers are complementary



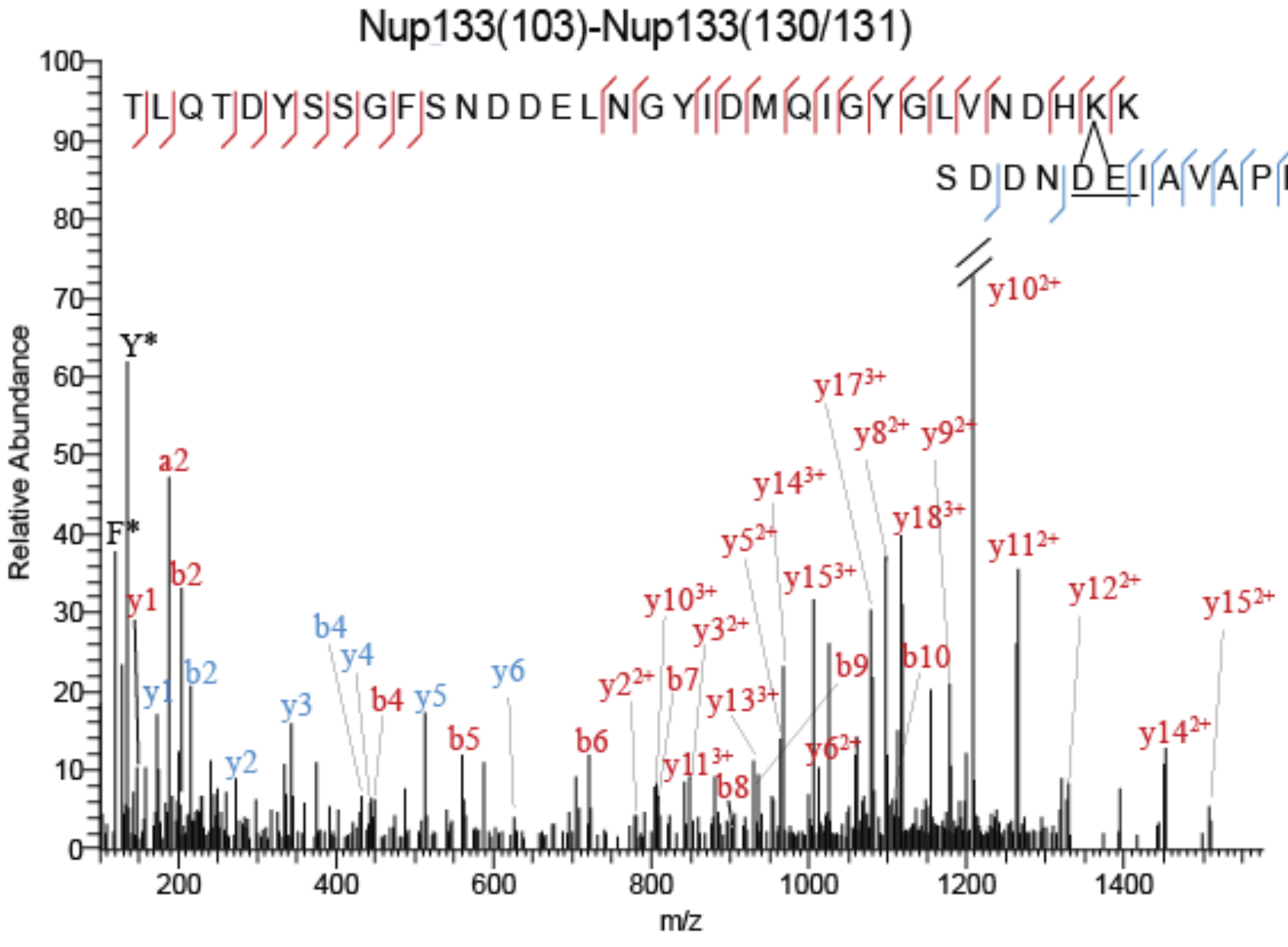
- Since EDC and DSS act on different sidechains, they yield complementary information



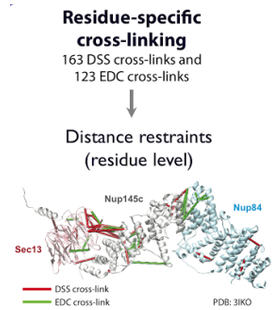
Ambiguity (1/3)



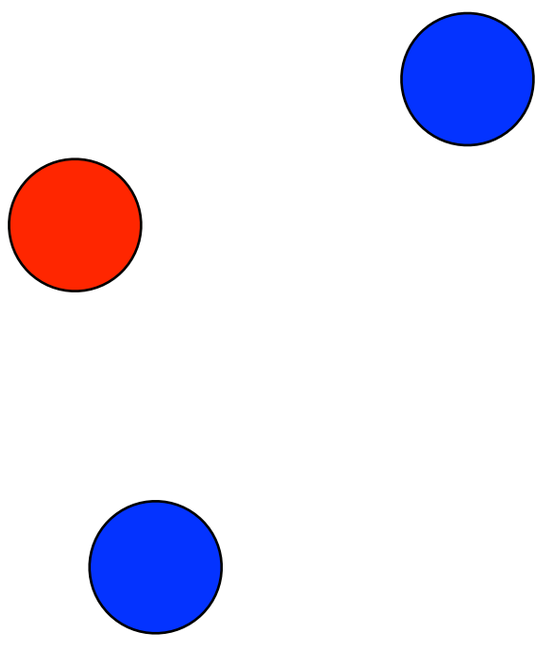
- Not always possible to uniquely *identify* a cross-link from the spectra:



Ambiguity (2/3)

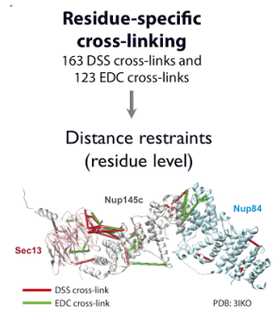


- *Compositional* ambiguity can also occur if there are multiple copies of a protein available (not the case for Nup84):

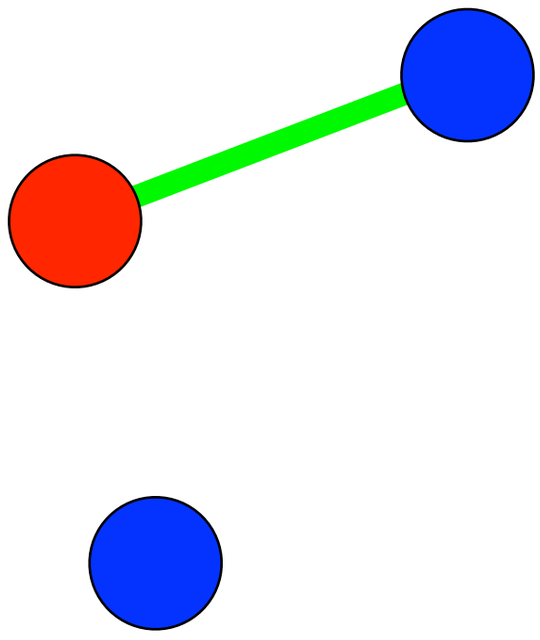


A cross-link observed between the red and blue proteins does not identify *which* blue protein is interacting with red

Ambiguity (2/3)

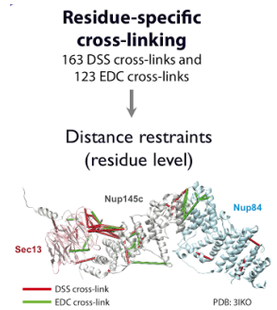


- *Compositional* ambiguity can also occur if there are multiple copies of a protein available (not the case for Nup84):

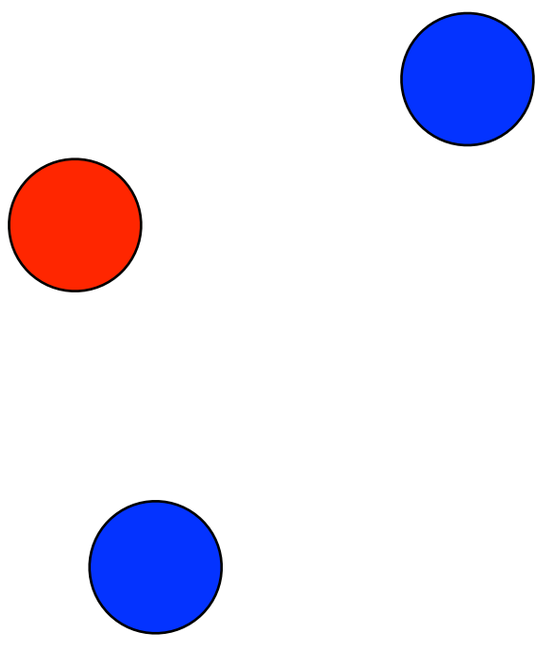


A cross-link observed between the red and blue proteins does not identify *which* blue protein is interacting with red

Ambiguity (2/3)

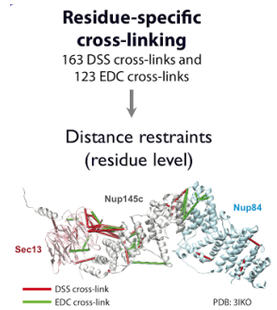


- *Compositional* ambiguity can also occur if there are multiple copies of a protein available (not the case for Nup84):

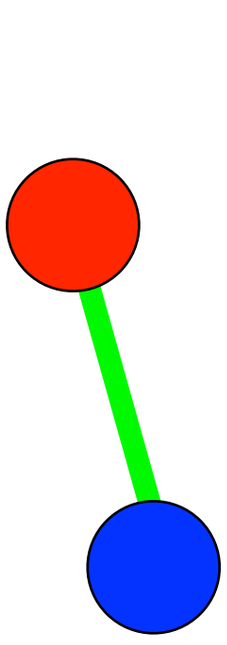


A cross-link observed between the red and blue proteins does not identify *which* blue protein is interacting with red

Ambiguity (2/3)

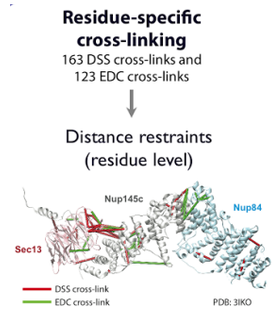


- *Compositional* ambiguity can also occur if there are multiple copies of a protein available (not the case for Nup84):

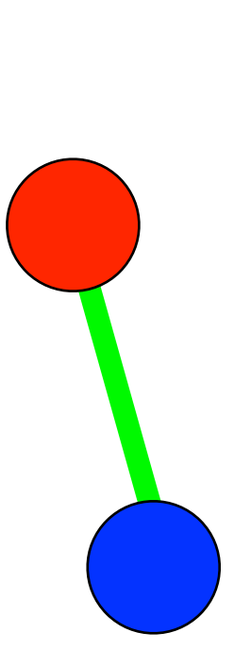


A cross-link observed between the red and blue proteins does not identify *which* blue protein is interacting with red

Ambiguity (2/3)

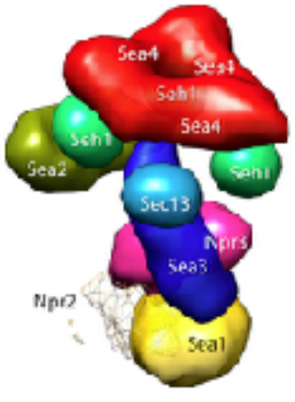


- *Compositional* ambiguity can also occur if there are multiple copies of a protein available (not the case for Nup84):

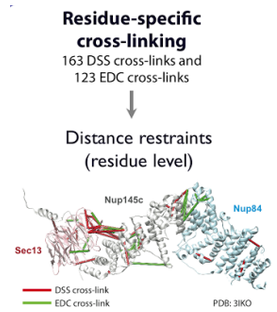


A cross-link observed between the red and blue proteins does not identify *which* blue protein is interacting with red

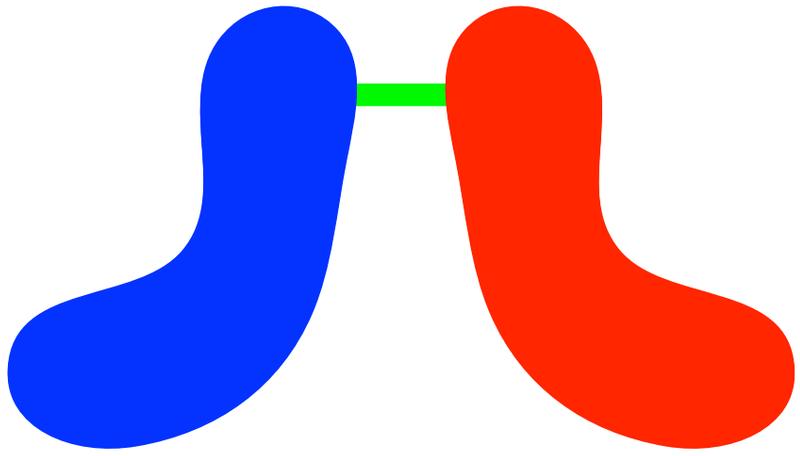
https://salilab.org/sea



Ambiguity (3/3)

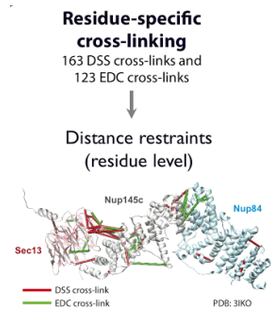


- *State* ambiguity can also occur if there are multiple states of the complex present (heterogeneity):

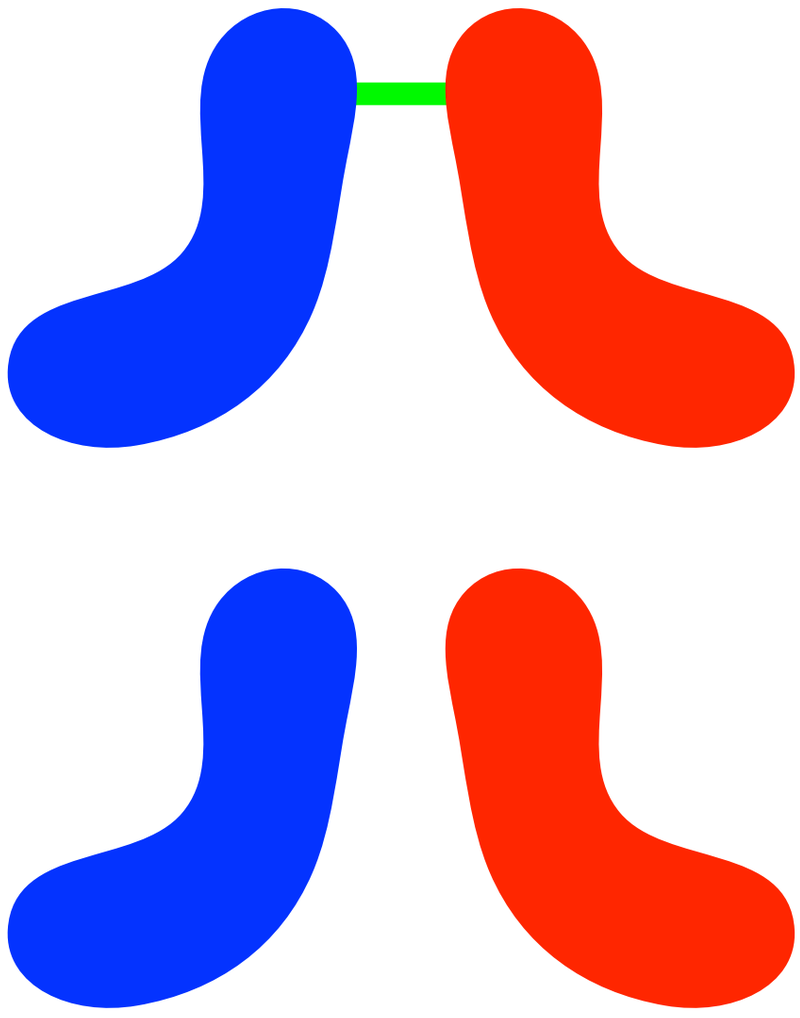


State 1

Ambiguity (3/3)

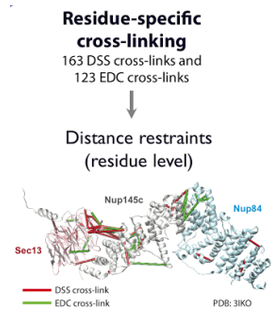


- *State* ambiguity can also occur if there are multiple states of the complex present (heterogeneity):

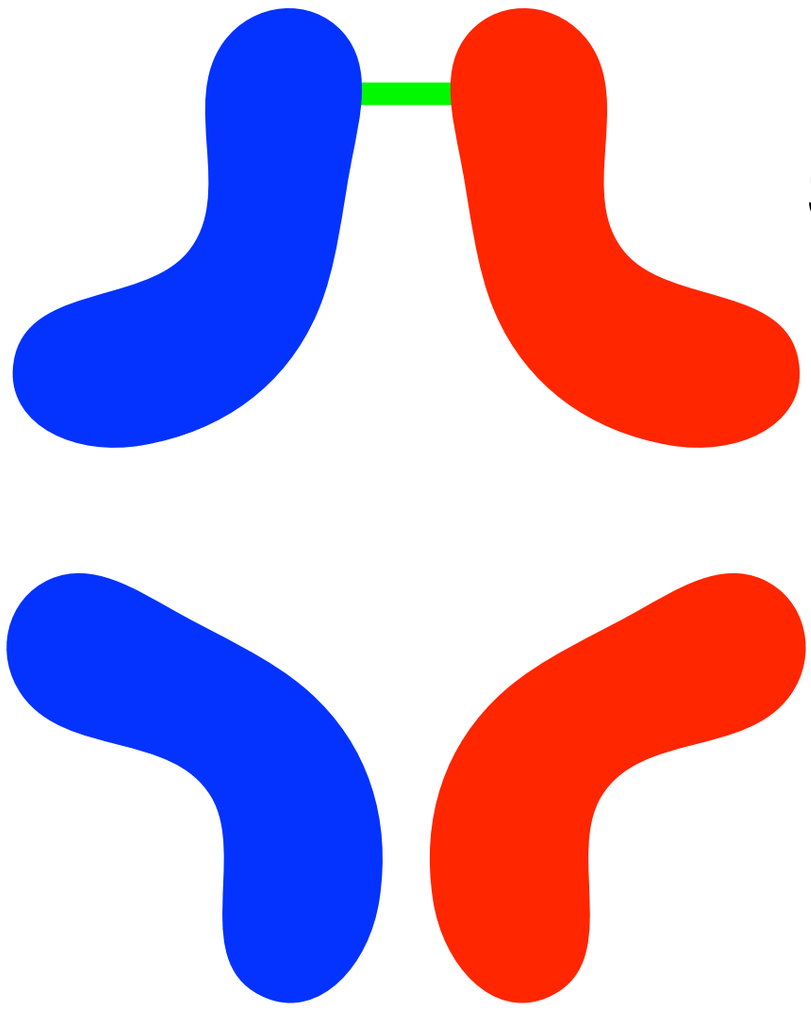


State 1

Ambiguity (3/3)

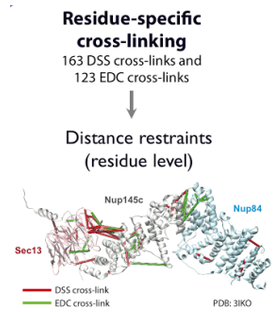


- *State* ambiguity can also occur if there are multiple states of the complex present (heterogeneity):

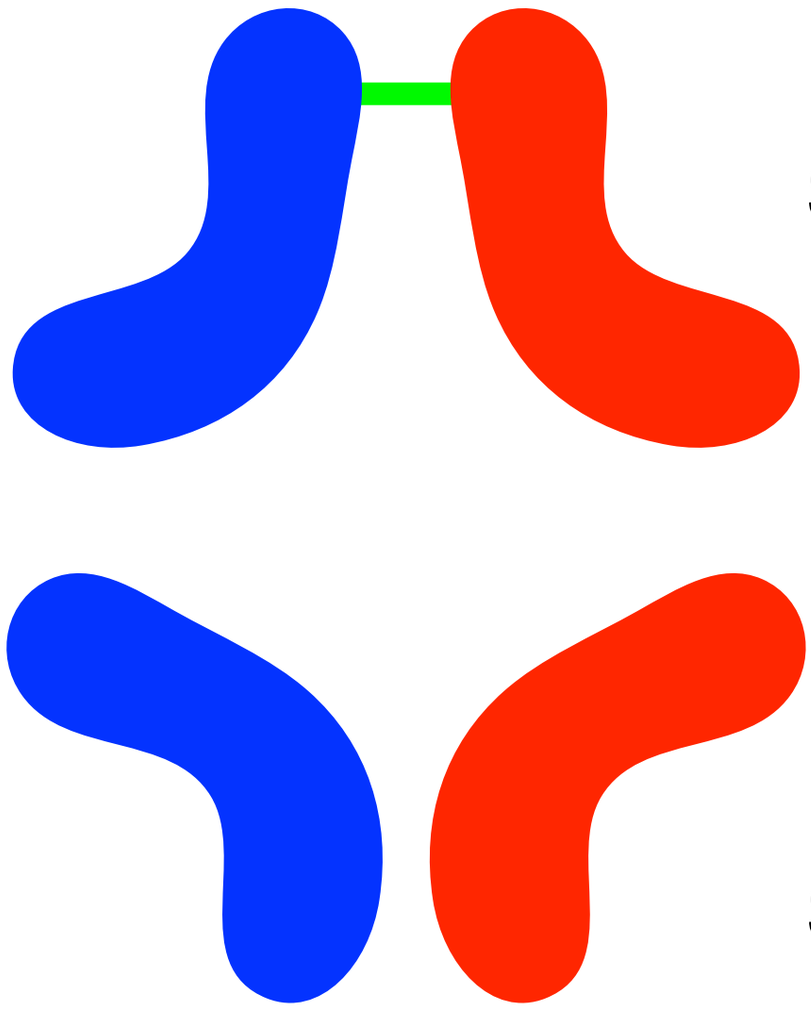


State 1

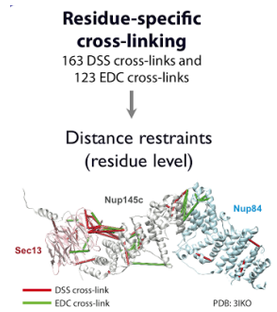
Ambiguity (3/3)



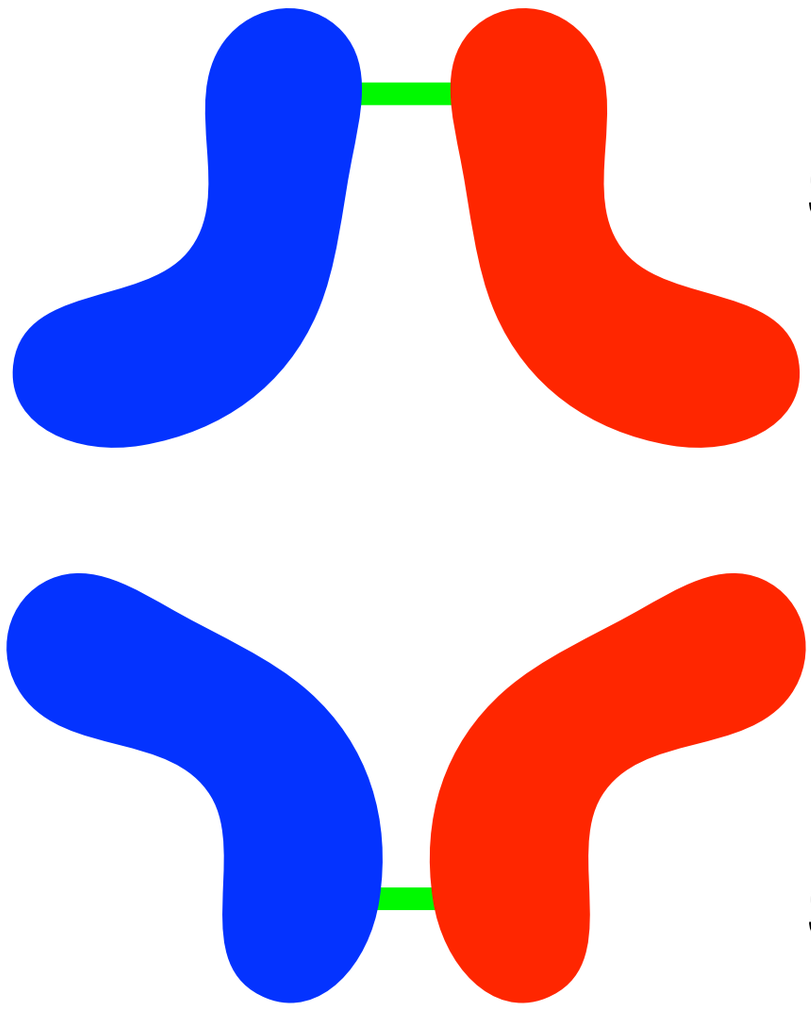
- *State* ambiguity can also occur if there are multiple states of the complex present (heterogeneity):



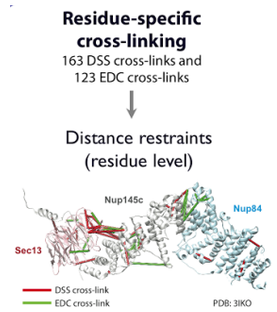
Ambiguity (3/3)



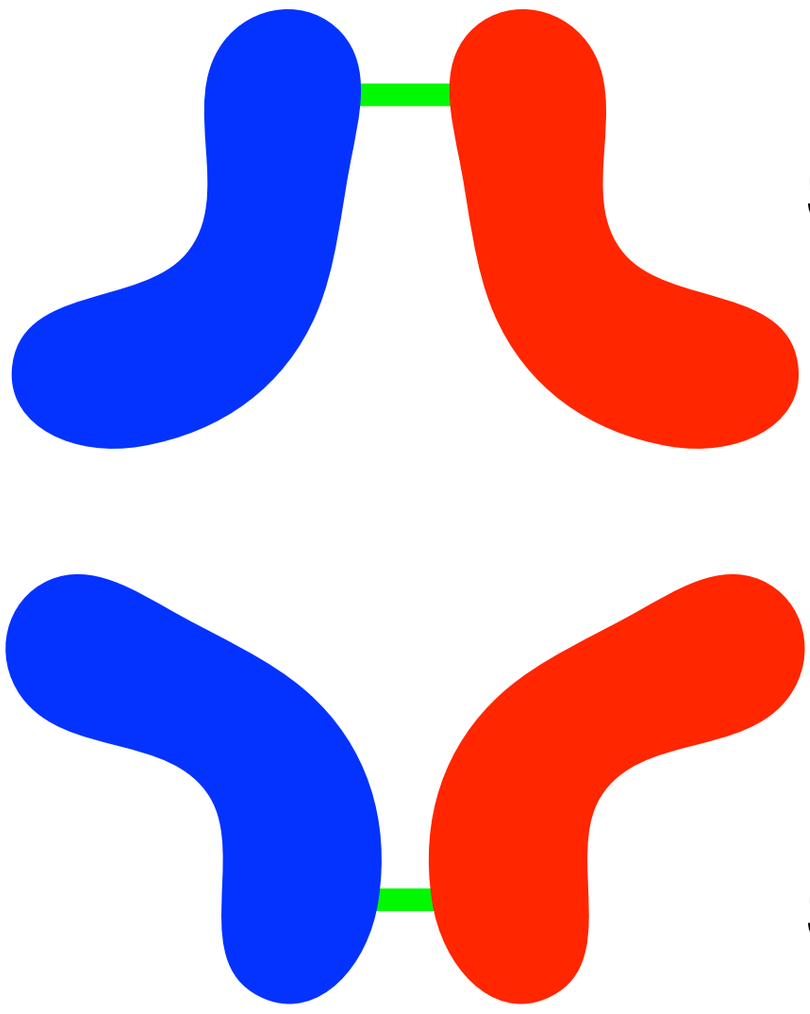
- *State* ambiguity can also occur if there are multiple states of the complex present (heterogeneity):



Ambiguity (3/3)

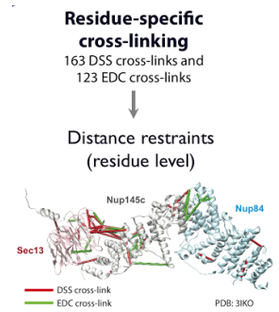


- *State* ambiguity can also occur if there are multiple states of the complex present (heterogeneity):

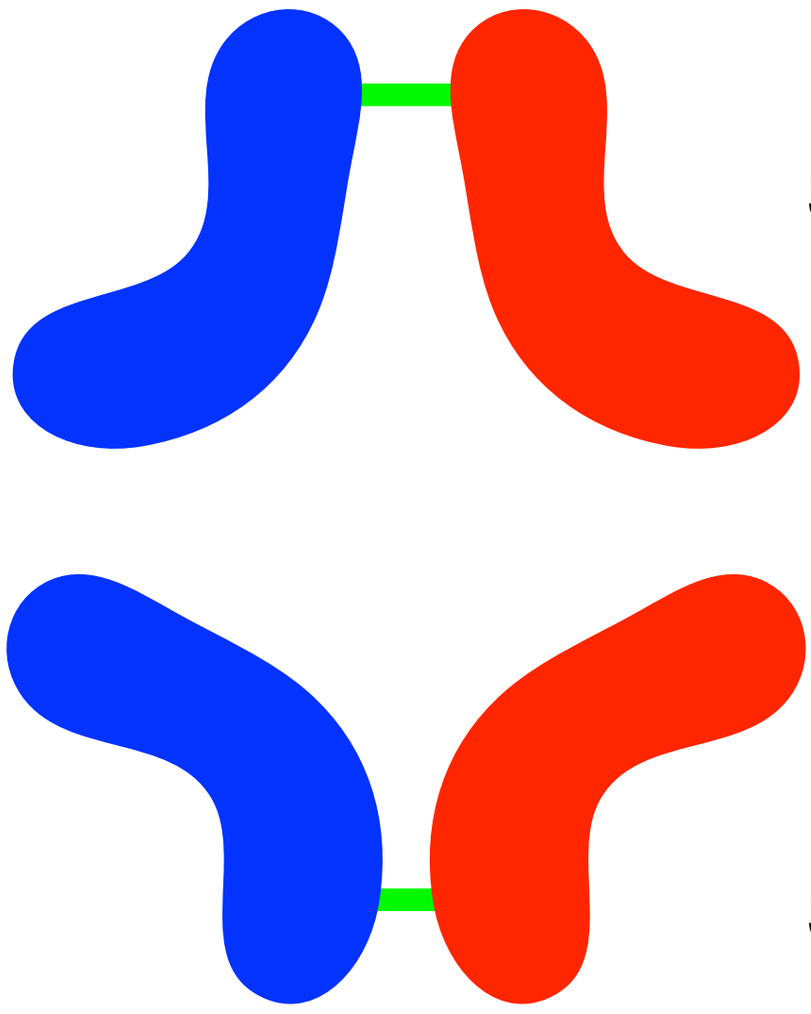


- The cross-linking experiment will yield cross-links representative of both states

Ambiguity (3/3)

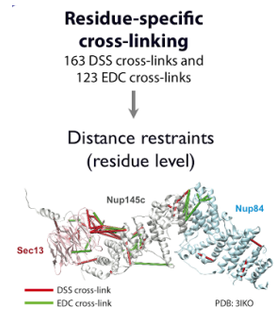


- *State* ambiguity can also occur if there are multiple states of the complex present (heterogeneity):

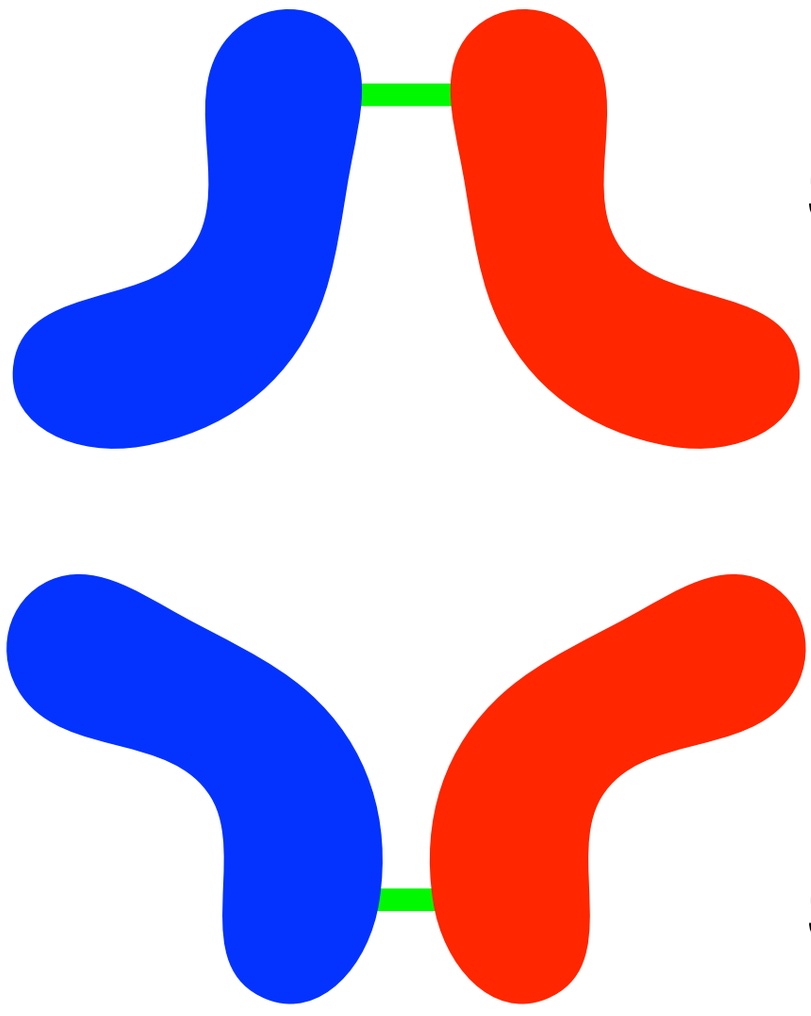


- The cross-linking experiment will yield cross-links representative of both states
- A single model cannot satisfy both cross-links simultaneously

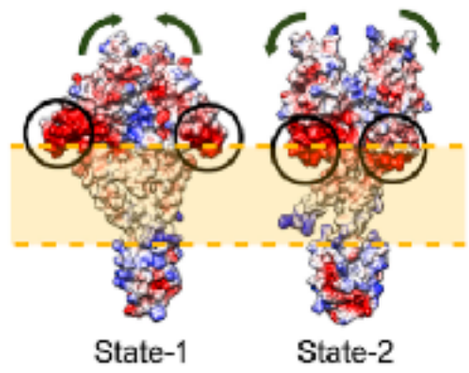
Ambiguity (3/3)



- *State* ambiguity can also occur if there are multiple states of the complex present (heterogeneity):

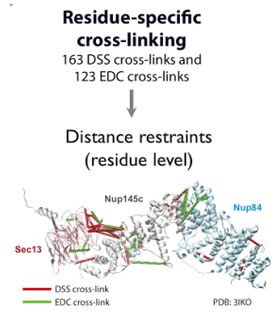


- The cross-linking experiment will yield cross-links representative of both states
- A single model cannot satisfy both cross-links simultaneously

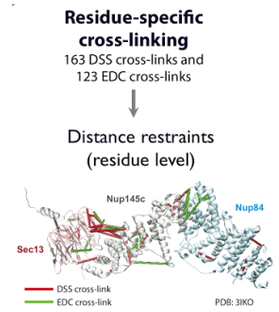


https://salilab.org/phoq

Scoring function for XL-MS

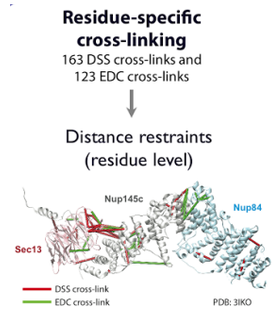


Scoring function for XL-MS

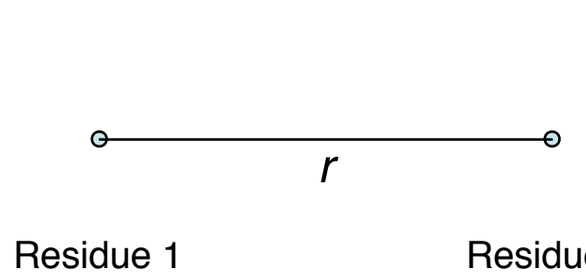


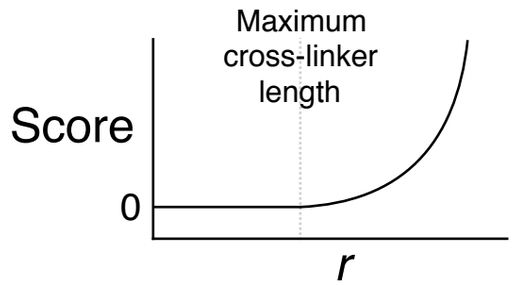
- Simplest way to score a cross-link would be as an upper bound harmonic on the interresidue distance:

Scoring function for XL-MS

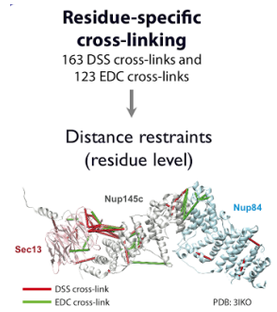


- Simplest way to score a cross-link would be as an upper bound harmonic on the interresidue distance:

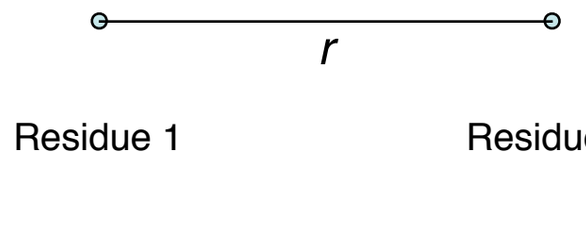


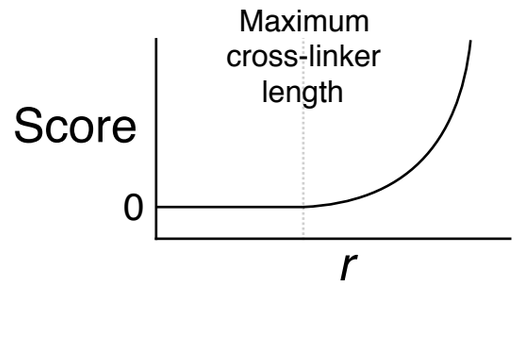


Scoring function for XL-MS



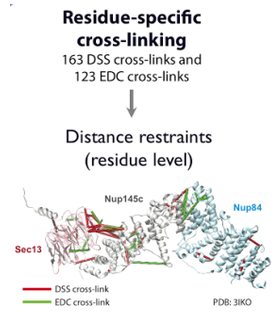
- Simplest way to score a cross-link would be as an upper bound harmonic on the interresidue distance:



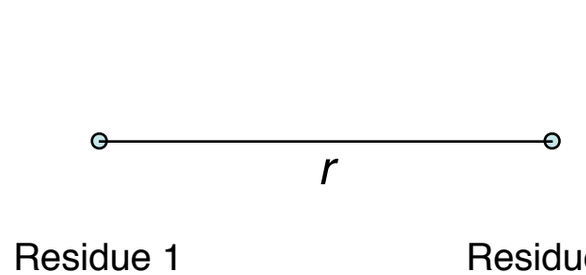


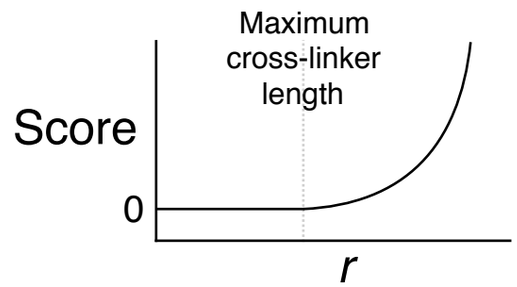
- We account for uncertainty in position by instead restraining intersphere distance (sphere radius, σ ≈ uncertainty)

Scoring function for XL-MS

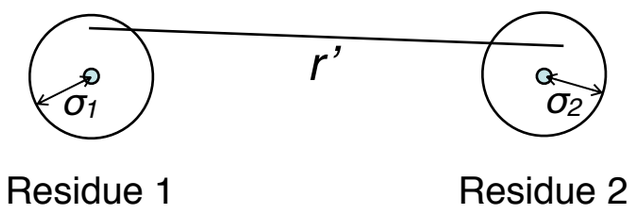


- Simplest way to score a cross-link would be as an upper bound harmonic on the interresidue distance:

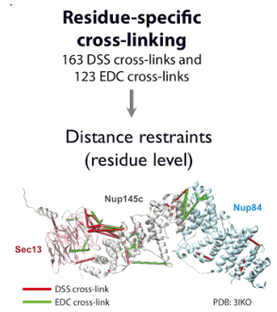




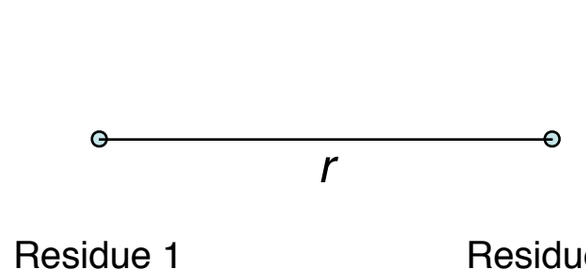
- We account for uncertainty in position by instead restraining intersphere distance (sphere radius, σ ≈ uncertainty)

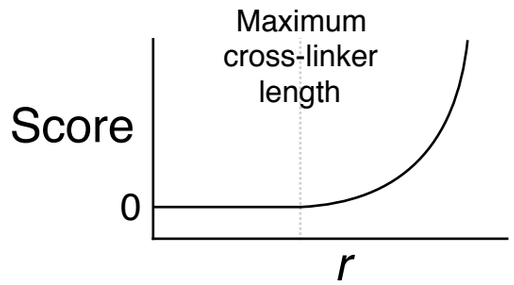


Scoring function for XL-MS

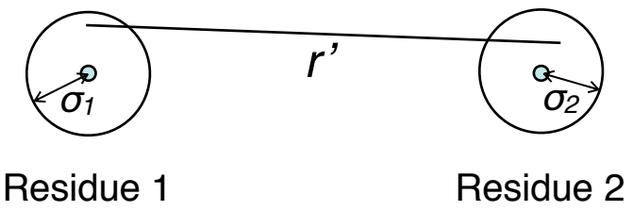


- Simplest way to score a cross-link would be as an upper bound harmonic on the interresidue distance:



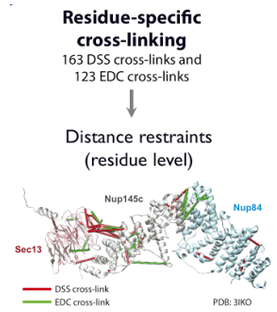


- We account for uncertainty in position by instead restraining intersphere distance (sphere radius, σ ≈ uncertainty)

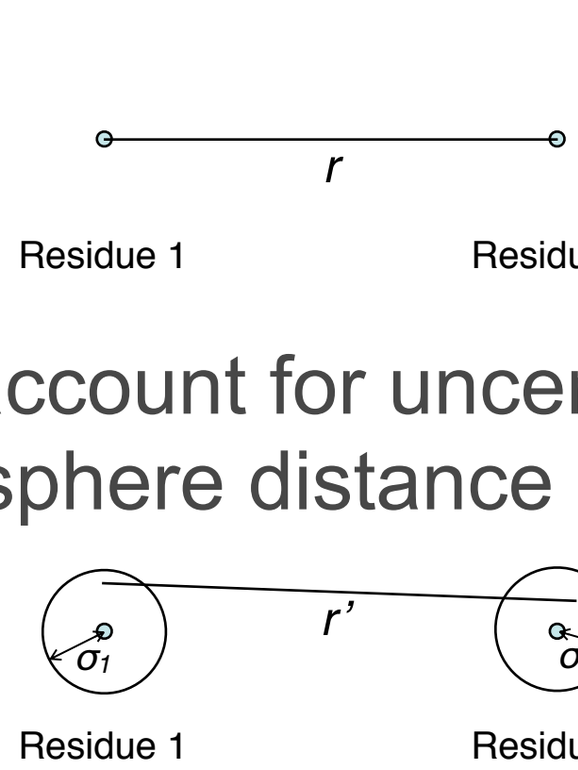


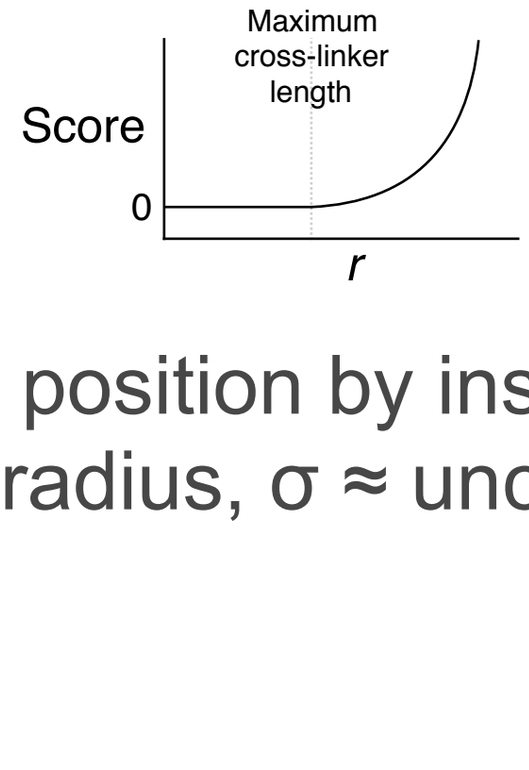
- We account for confidence in the cross-links themselves with another parameter, ψ

Scoring function for XL-MS



- Simplest way to score a cross-link would be as an upper bound harmonic on the interresidue distance:

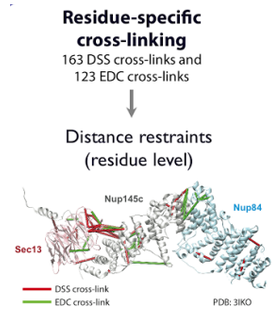




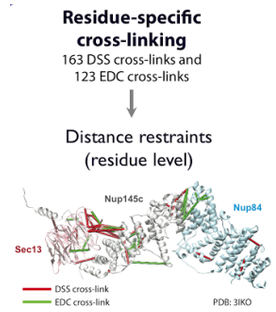
- We account for uncertainty in position by instead restraining intersphere distance (sphere radius, σ ≈ uncertainty)

- We account for confidence in the cross-links themselves with another parameter, ψ
- The score is Bayesian and the σ and ψ parameters are optimized to best fit the data

Bayesian scoring function for XL-MS



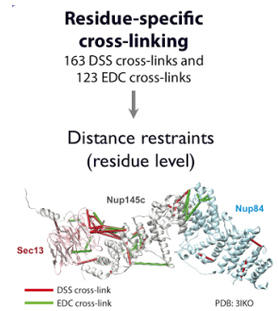
Bayesian scoring function for XL-MS



Bayes' rule

$$p(M \mid \{XL\}) \propto p(\{XL\} \mid M) \bullet p(M)$$

Bayesian scoring function for XL-MS

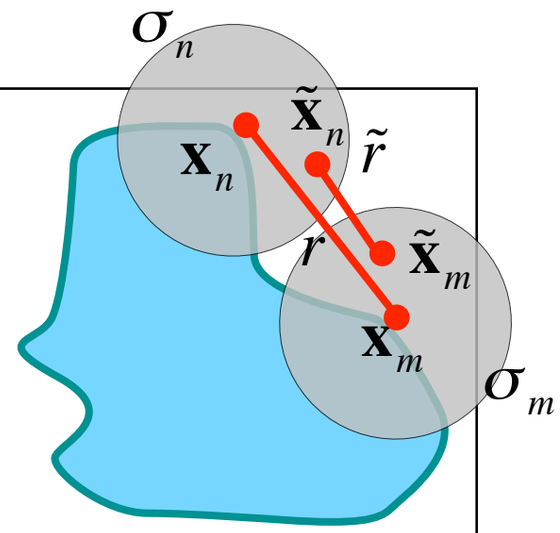


Bayes' rule

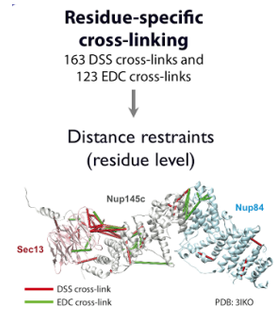
$$p(M \mid \{XL\}) \propto p(\{XL\} \mid M) \bullet p(M)$$

Forward model

$d(X,\sigma)$ = probability of having two residues at distance less than the cross-linker length, when residue positions are uncertain.



Bayesian scoring function for XL-MS



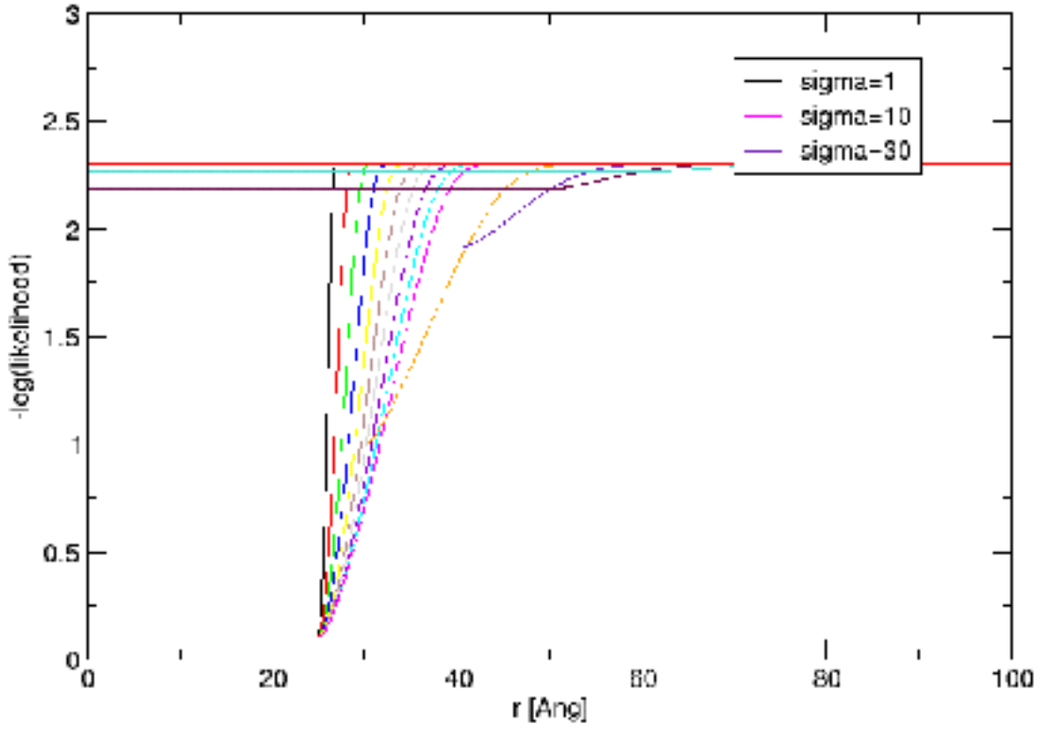
Bayes' rule

$$p(M \mid \{XL\}) \propto p(\{XL\} \mid M) \bullet p(M)$$

Likelihood

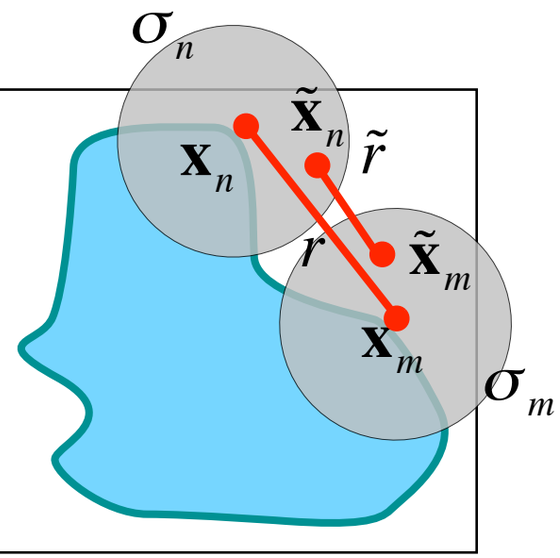
$$p(XL \mid X) = (1 - d(X,\sigma)) \cdot \psi + d(X,\sigma) \cdot (1 - \psi)$$

uncertainty ψ=[0,0.5]

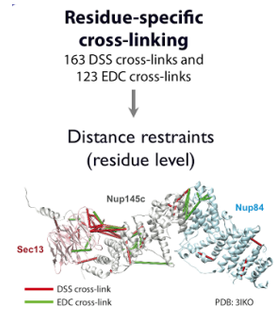


Forward model

$d(X,\sigma)$ = probability of having two residues at distance less than the cross-linker length, when residue positions are uncertain.



Bayesian scoring function for XL-MS



Bayes' rule

$$p(M \mid \{XL\}) \propto p(\{XL\} \mid M) \bullet p(M)$$

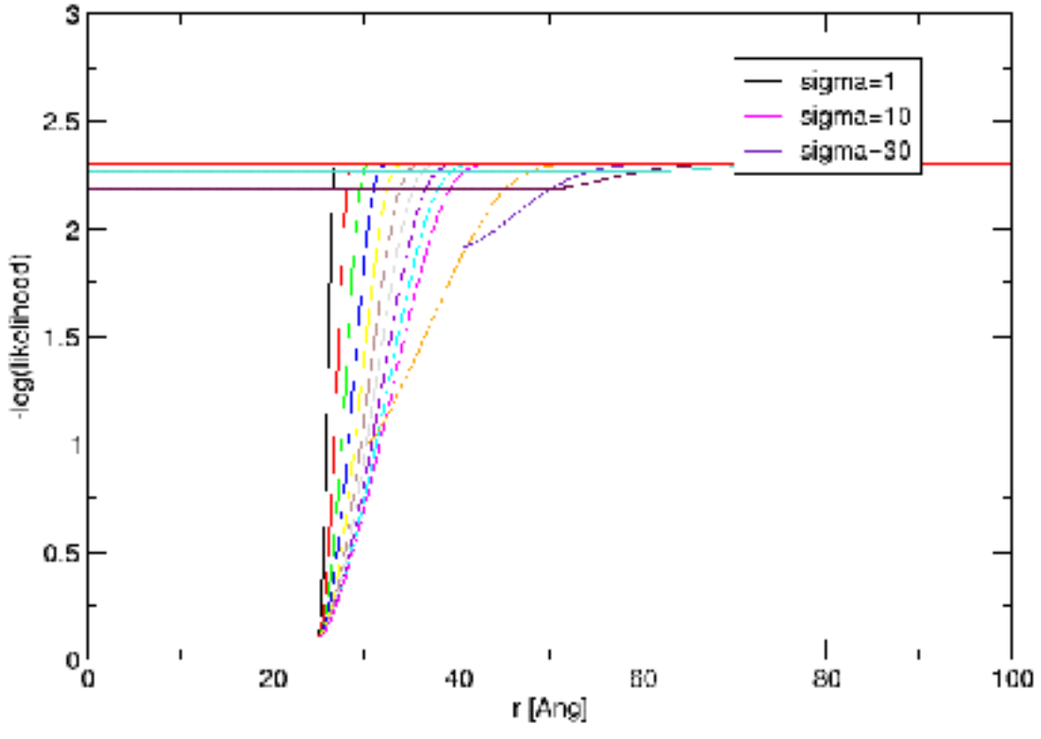
Prior

- Excluded volume restraint (pairwise hard-sphere repulsive potential)
- Sequence connectivity terms

Likelihood

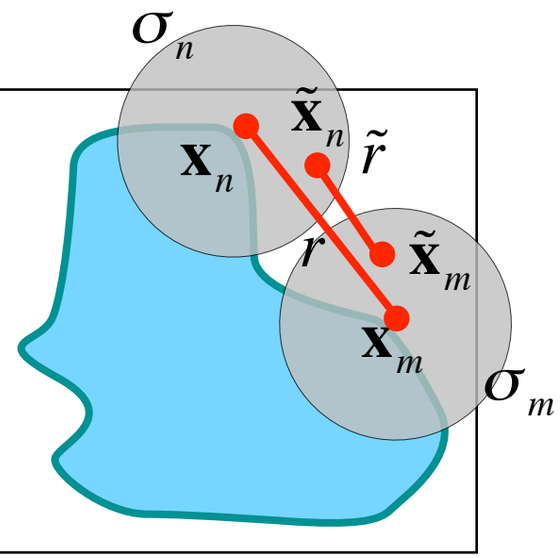
$$p(XL \mid X) = (1 - d(X,\sigma)) \cdot \psi + d(X,\sigma) \cdot (1 - \psi)$$

uncertainty ψ=[0,0.5]

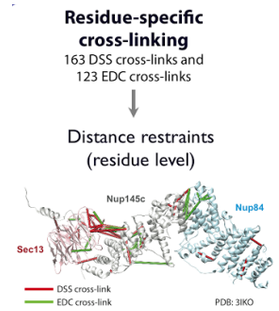


Forward model

$d(X,\sigma)$ = probability of having two residues at distance less than the cross-linker length, when residue positions are uncertain.



Bayesian scoring function for XL-MS



Bayes' rule

$$p(M \mid \{XL\}) \propto p(\{XL\} \mid M) \bullet p(M)$$

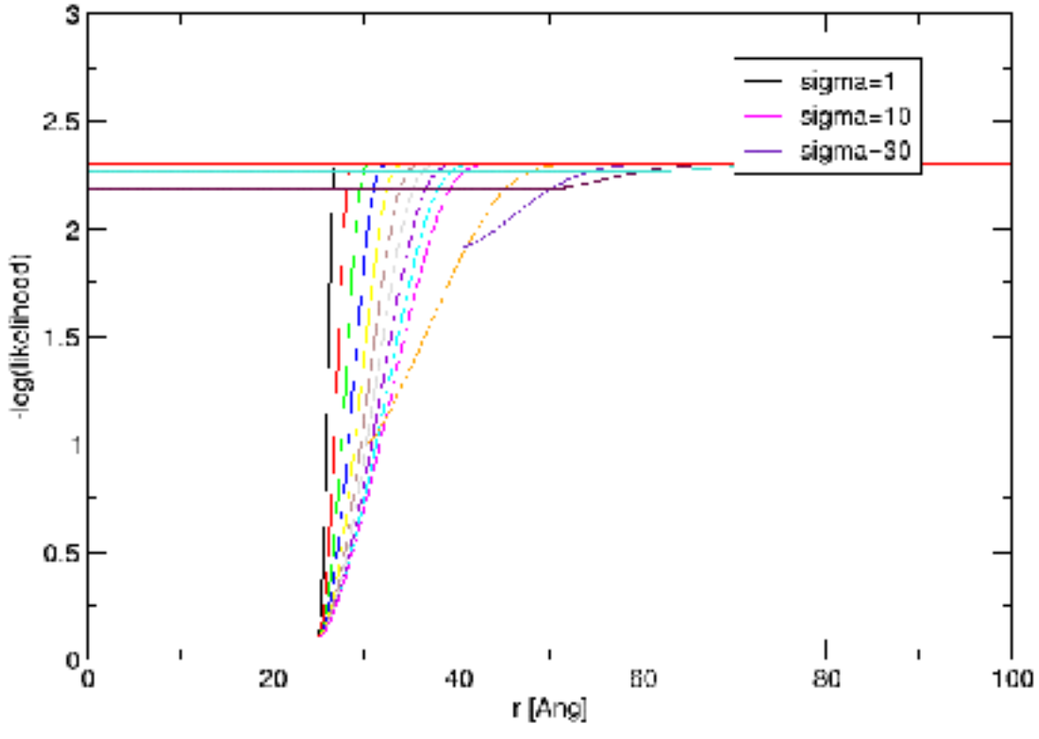
Prior

- Excluded volume restraint (pairwise hard-sphere repulsive potential)
- Sequence connectivity terms

Likelihood

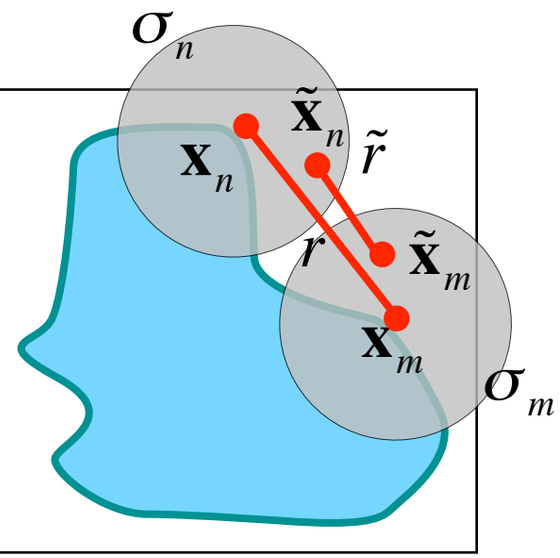
$$p(XL \mid X) = (1 - d(X,\sigma)) \cdot \psi + d(X,\sigma) \cdot (1 - \psi)$$

uncertainty ψ=[0,0.5]



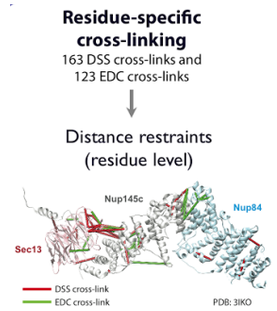
Forward model

$d(X,\sigma)$ = probability of having two residues at distance less than the cross-linker length, when residue positions are uncertain.



- Each cross-link subdataset (class) can be given a different ψ parameter
- Cross-links identified multiple times are weighed proportionally
- Score allows us to estimate the position uncertainties, as well as the noise in the cross-link class (weight)

Bayesian scoring function for XL-MS

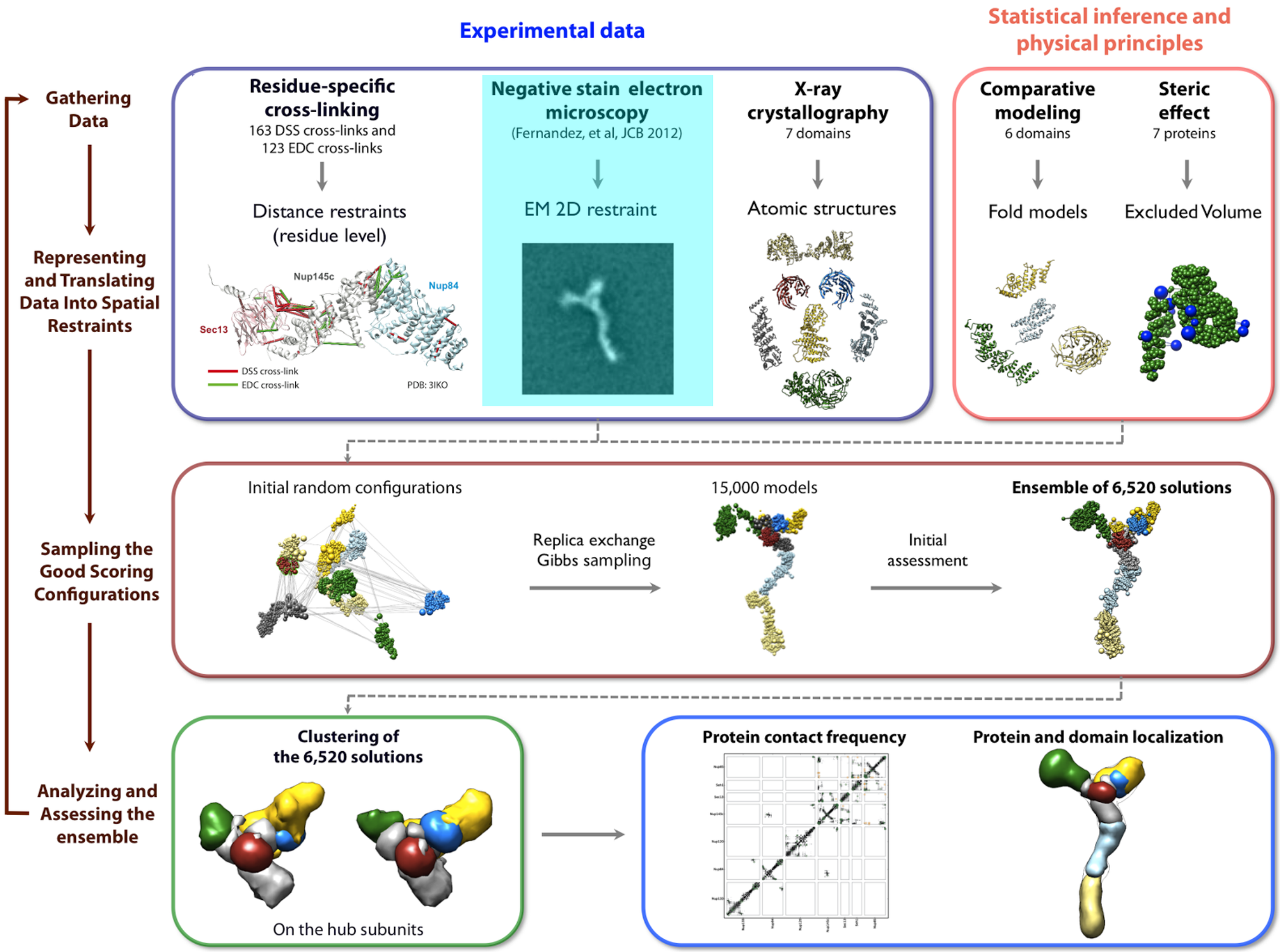


- *Identification* and *compositional* ambiguity of residue pairs (*n*, *n*+1, …) are handled by means of a compound likelihood function:

$$p(d_{n,n+1,\ldots} \mid X, I) = 1 - \prod_{n} (1 - p(d_n \mid X, I))$$

- For Nup84, we optimize a single σ for all residues and fix ψ (approximately equal to the fraction of false-positive cross-links) at 5% for all cross-links

Modeling Nup84 with IMP (2014)



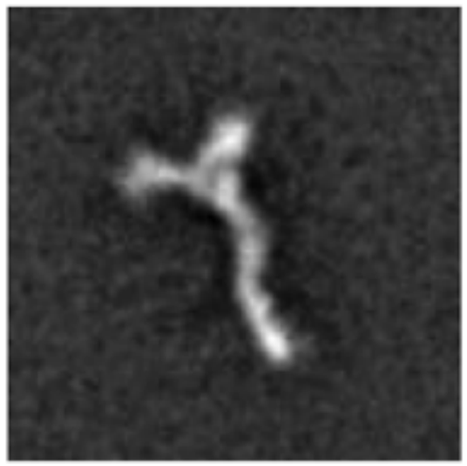
Mol Cell Proteomics 13, 2927-2943, 2014

Negative stain electron microscopy

(Fernandez, et al, JCB 2012)

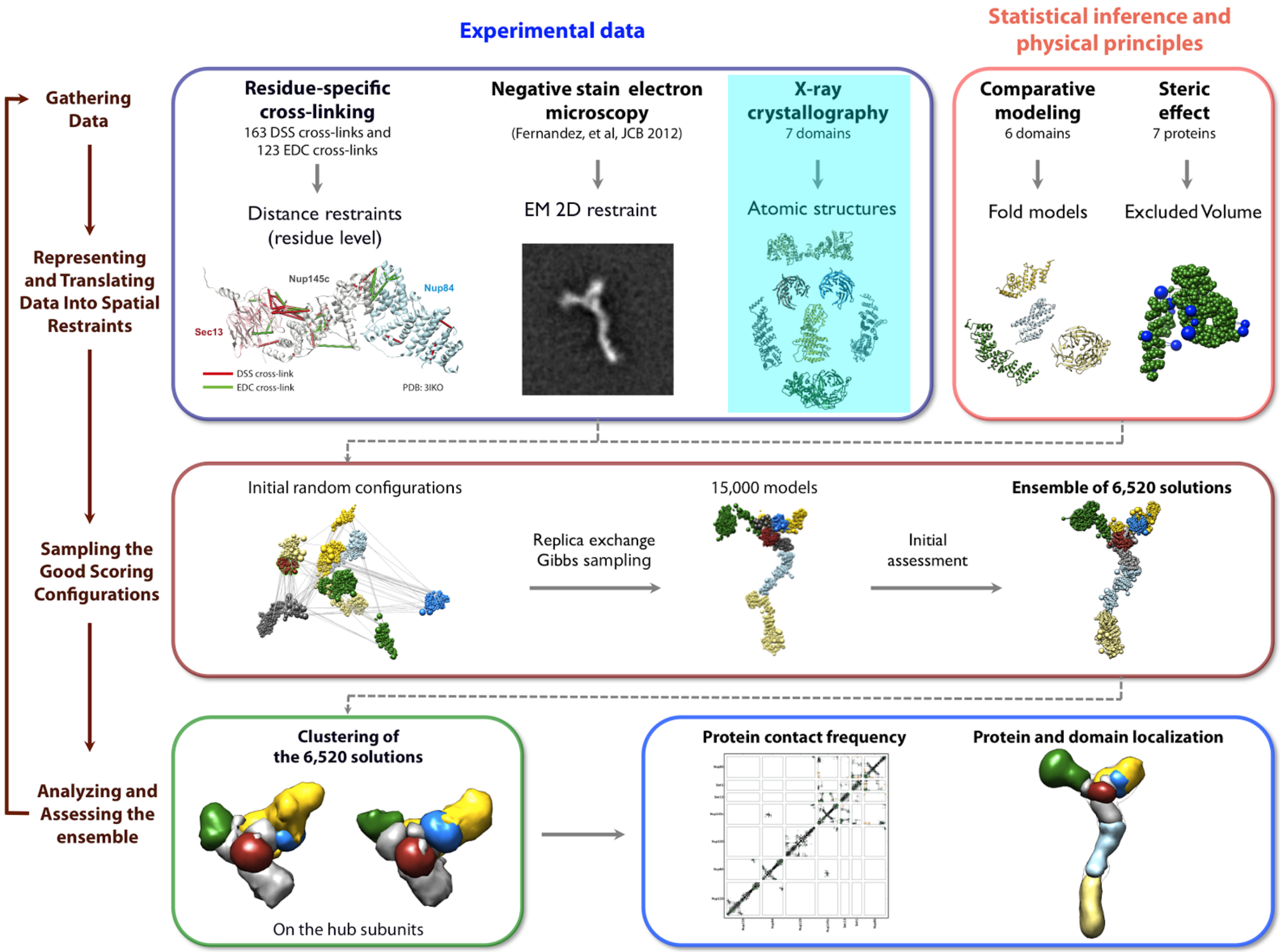
↓

EM 2D restraint



- The same class average of the entire complex used in the 2012 study was used here
- The class averages of the truncated complexes were not used as restraints, but as validation (later)
- Data used via a similar restraint as in the 2012 study

Modeling Nup84 with IMP (2014)



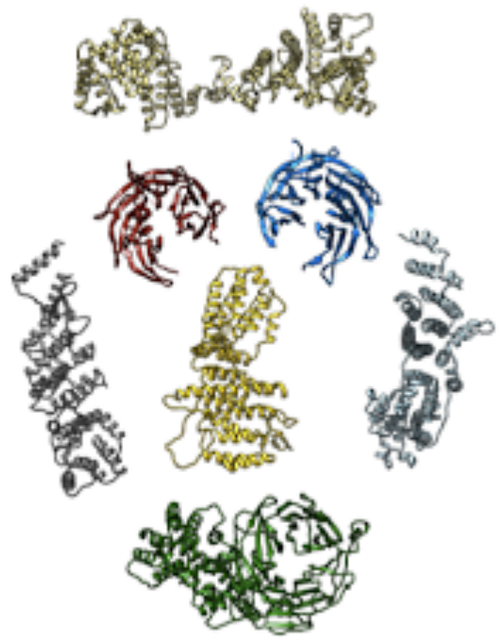
Mol Cell Proteomics 13, 2927-2943, 2014

X-ray crystallography

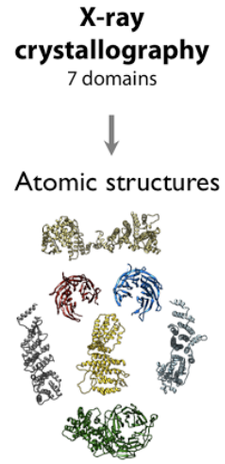
7 domains

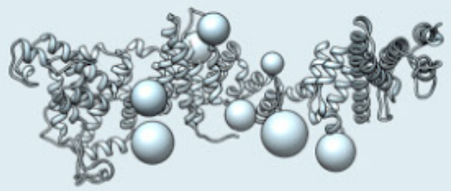
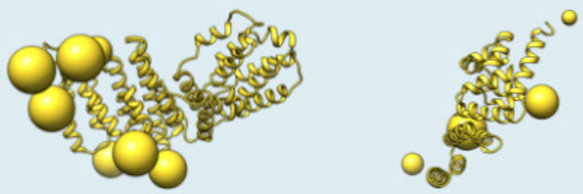
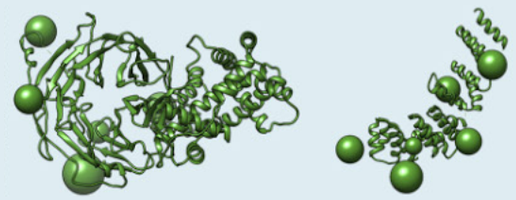
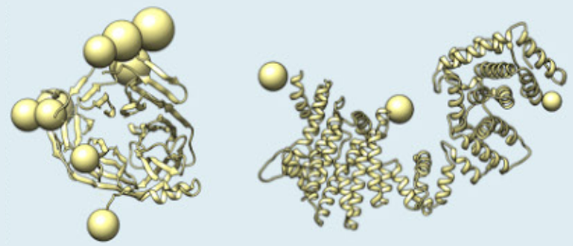
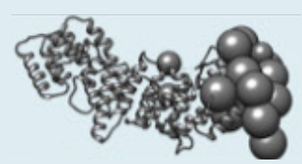
↓

Atomic structures

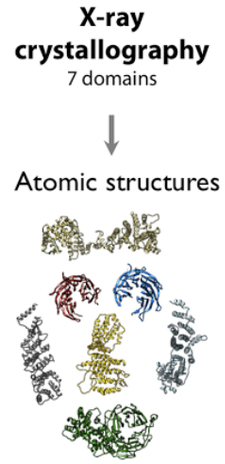


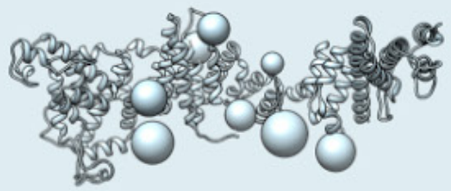
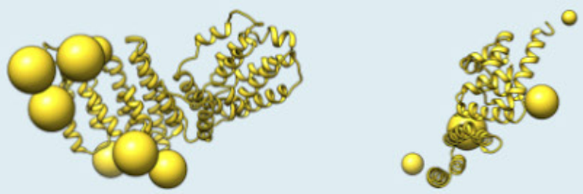
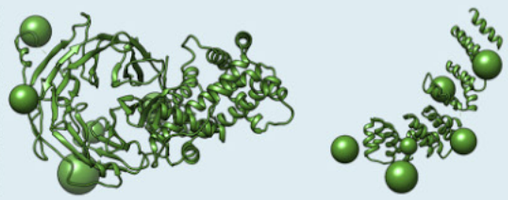
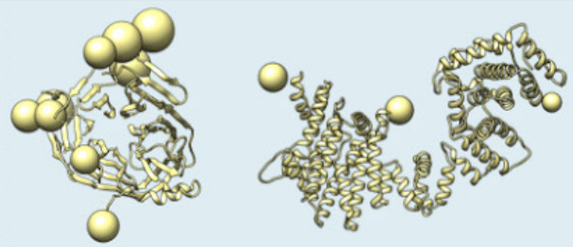
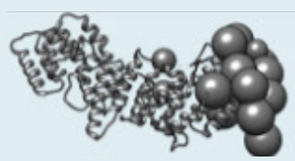
X-ray structures



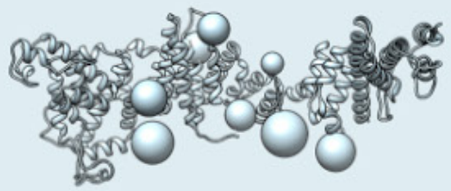
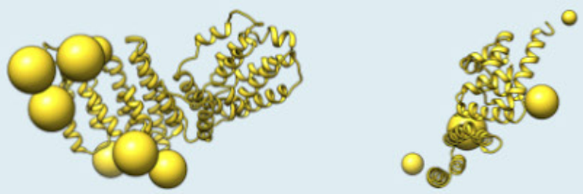
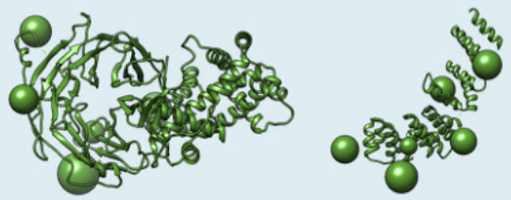
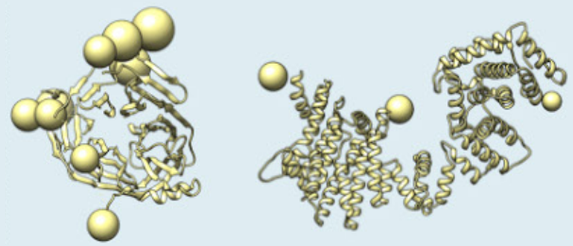
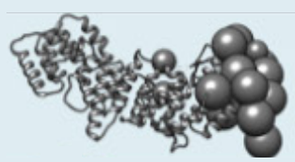
Nup84		1-6: flexible string of a bead (DISOPRED) 7-436: X-ray 3JRO_C, 3IKO_C (100% seq id) 429-488: Model 3F3F_G (10% seq id, HHpred) 489-505: flexible string of beads (DISOPRED) 506-726: Model 3CQC_A (18% seq id, HHpred)
Nup85		1-43: flexible string of beads (DISOPRED) 44-555: X-ray 3F3F_D, 3EWE_D (100% seq id) Linker (flexible string of a bead) 532-655: Model 2QX5_B (18% seq id, HHpred) 532-743: Model 4LCT_A (14% seq id, HHpred) 744: flexible string of a bead (PSIPRED)
Nup120		1-712: X-ray 3F7F_A, 3HXR_A (100% seq id) Linker (flexible string of a bead) 727-1037: Model 4FHN_B, 4FHN_D (14% seq id, HHpred)
Nup133		1-55: flexible string of beads (DISOPRED) 56-480: Model VpNup133 (46% seq id, Muscle) Linker (flexible string of a bead) 490-945: Model 3I4R (15% seq id, HHpred) 946-1157: X-ray 3KFO_A (100% seq id)
Nup145c		1-125: flexible string of beads (DISOPRED) 126-553: X-ray 3IKO_B, 3JRO_A, 3BG1_B, 3BGO_B (100% seq id) 554-712: flexible string of beads (PSIPRED)
Seh1		1-346: X-ray 3F3F (100% seq id) 347-349: flexible string of a bead (DISOPRED)
Sec13		1: flexible string of a bead (DISOPRED) 2-296: X-ray 2PM7_D (100% seq id) 297: flexible string of a bead (DISOPRED)

X-ray structures

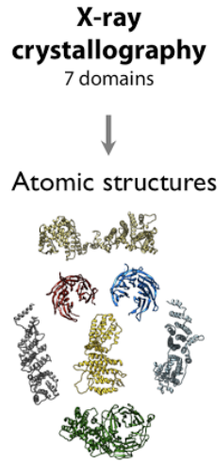


Nup84		1-6: flexible string of a bead (DISOPRED) 7-436: X-ray 3JRO_C, 3IKO_C (100% seq id) 429-488: Model 3F3F_G (10% seq id, HHpred) 489-505: flexible string of beads (DISOPRED) 506-726: Model 3CQC_A (18% seq id, HHpred)
Nup85		1-43: flexible string of beads (DISOPRED) 44-555: X-ray 3F3F_D, 3EWE_D (100% seq id) Linker (flexible string of a bead) 532-655: Model 2QX5_B (18% seq id, HHpred) 532-743: Model 4LCT_A (14% seq id, HHpred) 744: flexible string of a bead (PSIPRED)
Nup120		1-712: X-ray 3F7F_A, 3HXR_A (100% seq id) Linker (flexible string of a bead) 727-1037: Model 4FHN_B, 4FHN_D (14% seq id, HHpred)
Nup133		1-55: flexible string of beads (DISOPRED) 56-480: Model VpNup133 (46% seq id, Muscle) Linker (flexible string of a bead) 490-945: Model 3I4R (15% seq id, HHpred) 946-1157: X-ray 3KFO_A (100% seq id)
Nup145c		1-125: flexible string of beads (DISOPRED) 126-553: X-ray 3IKO_B, 3JRO_A, 3BG1_B, 3BGO_B (100% seq id) 554-712: flexible string of beads (PSIPRED)
Seh1		1-346: X-ray 3F3F (100% seq id) 347-349: flexible string of a bead (DISOPRED)
Sec13		1: flexible string of a bead (DISOPRED) 2-296: X-ray 2PM7_D (100% seq id) 297: flexible string of a bead (DISOPRED)

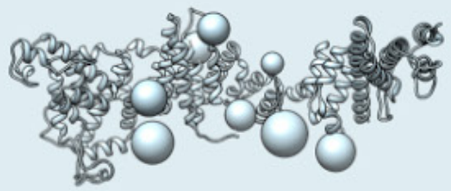
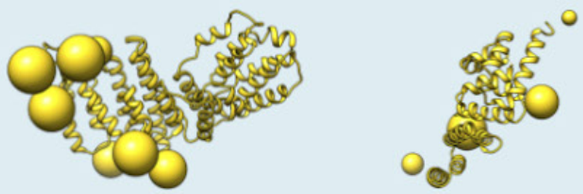
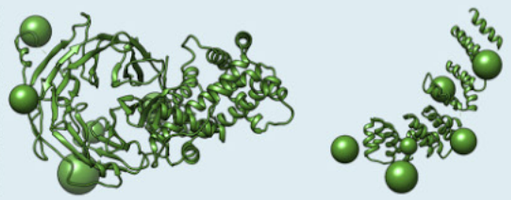
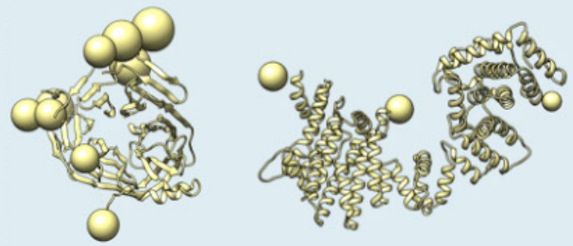
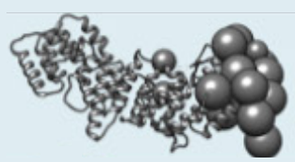
- Atomic structures were only available for 7 domains, 54% of the total sequence

Nup84		1-6: flexible string of a bead (DISOPRED) 7-436: X-ray 3JRO_C, 3IKO_C (100% seq id) 429-488: Model 3F3F_G (10% seq id, HHpred) 489-505: flexible string of beads (DISOPRED) 506-726: Model 3CQC_A (18% seq id, HHpred)
Nup85		1-43: flexible string of beads (DISOPRED) 44-555: X-ray 3F3F_D, 3EWE_D (100% seq id) Linker (flexible string of a bead) 532-655: Model 2QX5_B (18% seq id, HHpred) 532-743: Model 4LCT_A (14% seq id, HHpred) 744: flexible string of a bead (PSIPRED)
Nup120		1-712: X-ray 3F7F_A, 3HXR_A (100% seq id) Linker (flexible string of a bead) 727-1037: Model 4FHN_B, 4FHN_D (14% seq id, HHpred)
Nup133		1-55: flexible string of beads (DISOPRED) 56-480: Model VpNup133 (46% seq id, Muscle) Linker (flexible string of a bead) 490-945: Model 3I4R (15% seq id, HHpred) 946-1157: X-ray 3KFO_A (100% seq id)
Nup145c		1-125: flexible string of beads (DISOPRED) 126-553: X-ray 3IKO_B, 3JRO_A, 3BG1_B, 3BGO_B (100% seq id) 554-712: flexible string of beads (PSIPRED)
Seh1		1-346: X-ray 3F3F (100% seq id) 347-349: flexible string of a bead (DISOPRED)
Sec13		1: flexible string of a bead (DISOPRED) 2-296: X-ray 2PM7_D (100% seq id) 297: flexible string of a bead (DISOPRED)

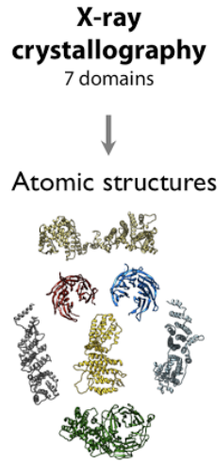
X-ray structures



- Atomic structures were only available for 7 domains, 54% of the total sequence

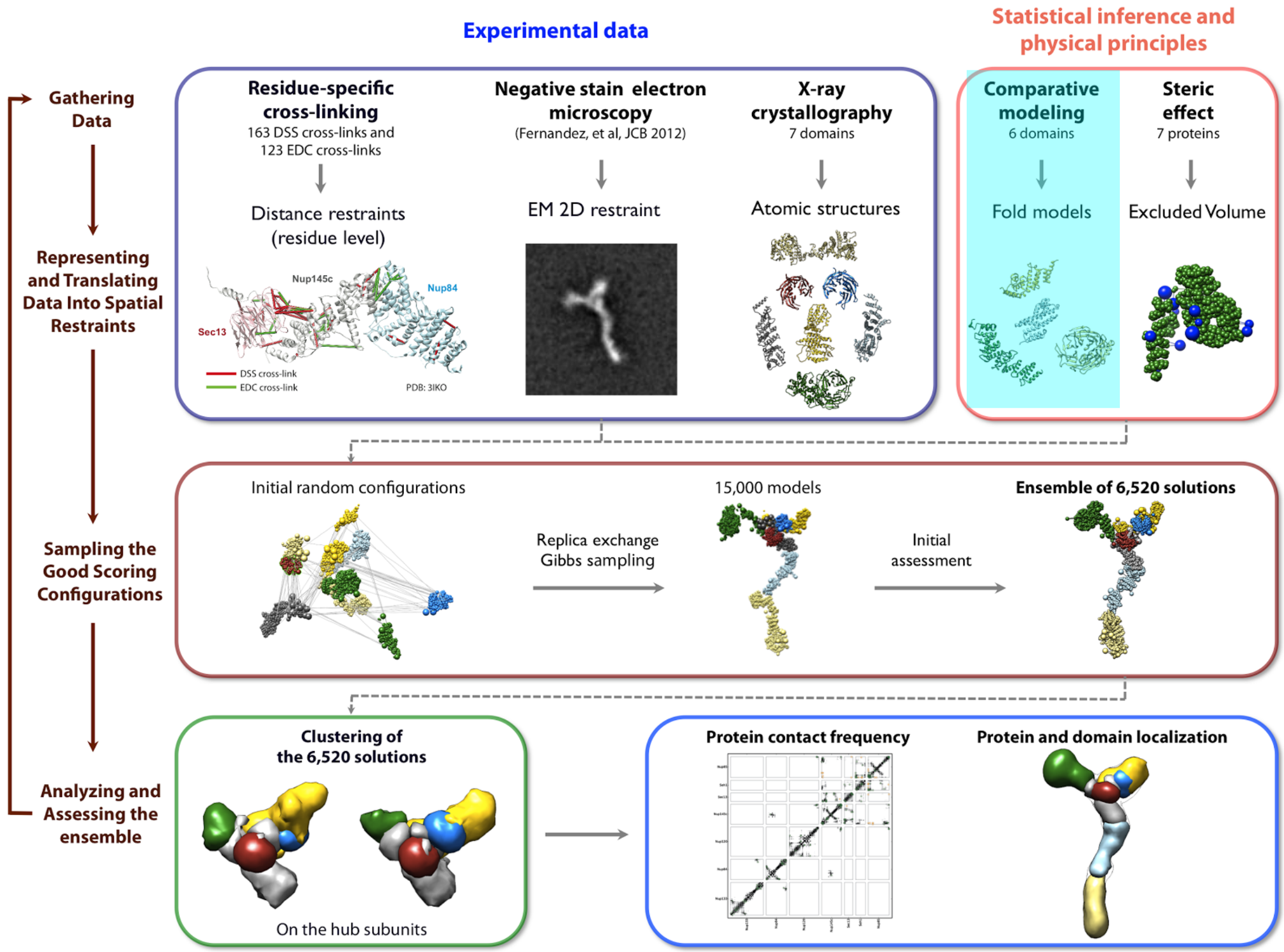
Nup84		1-6: flexible string of a bead (DISOPRED) 7-436: X-ray 3JRO_C, 3IKO_C (100% seq id) 429-488: Model 3F3F_G (10% seq id, HHpred) 489-505: flexible string of beads (DISOPRED) 506-726: Model 3CQC_A (18% seq id, HHpred)
Nup85		1-43: flexible string of beads (DISOPRED) 44-555: X-ray 3F3F_D, 3EWE_D (100% seq id) Linker (flexible string of a bead) 532-655: Model 2QX5_B (18% seq id, HHpred) 532-743: Model 4LCT_A (14% seq id, HHpred) 744: flexible string of a bead (PSIPRED)
Nup120		1-712: X-ray 3F7F_A, 3HXR_A (100% seq id) Linker (flexible string of a bead) 727-1037: Model 4FHN_B, 4FHN_D (14% seq id, HHpred)
Nup133		1-55: flexible string of beads (DISOPRED) 56-480: Model VpNup133 (46% seq id, Muscle) Linker (flexible string of a bead) 490-945: Model 3I4R (15% seq id, HHpred) 946-1157: X-ray 3KFO_A (100% seq id)
Nup145c		1-125: flexible string of beads (DISOPRED) 126-553: X-ray 3IKO_B, 3JRO_A, 3BG1_B, 3BGO_B (100% seq id) 554-712: flexible string of beads (PSIPRED)
Seh1		1-346: X-ray 3F3F (100% seq id) 347-349: flexible string of a bead (DISOPRED)
Sec13		1: flexible string of a bead (DISOPRED) 2-296: X-ray 2PM7_D (100% seq id) 297: flexible string of a bead (DISOPRED)

X-ray structures



- Atomic structures were only available for 7 domains, 54% of the total sequence
- Data used as *representation*: structures kept rigid during the simulation

Modeling Nup84 with IMP (2014)



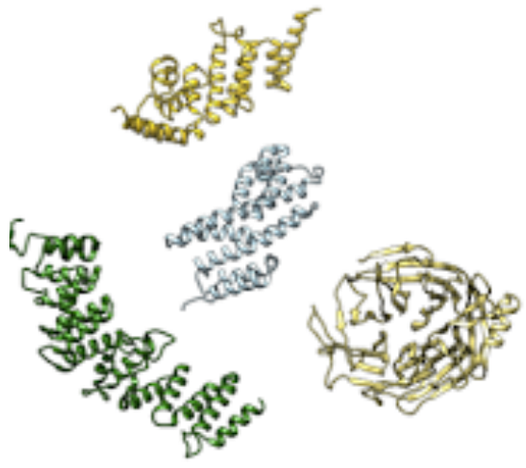
Mol Cell Proteomics 13, 2927-2943, 2014

Comparative modeling

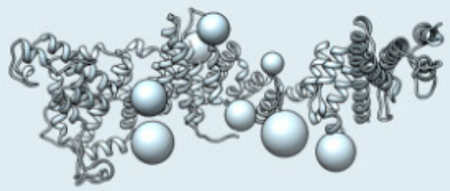
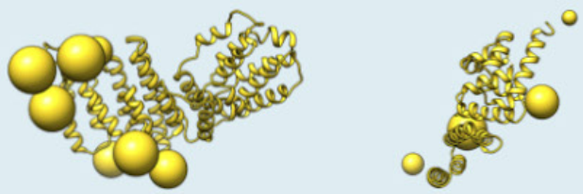
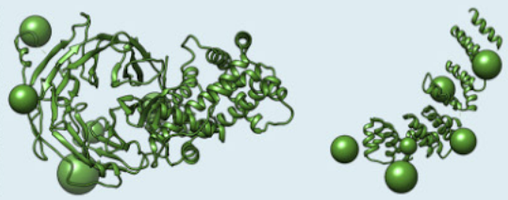
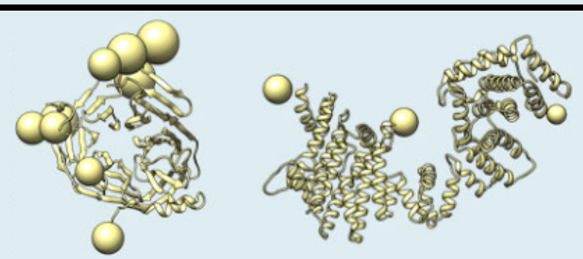
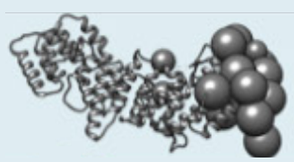
6 domains

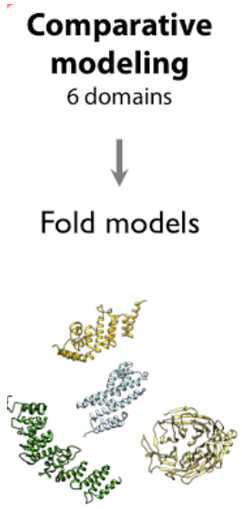
↓

Fold models

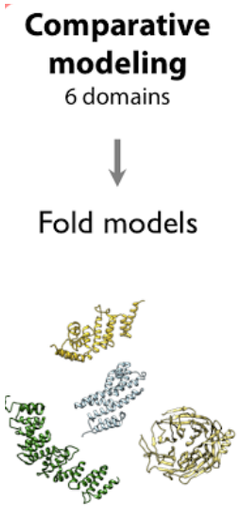


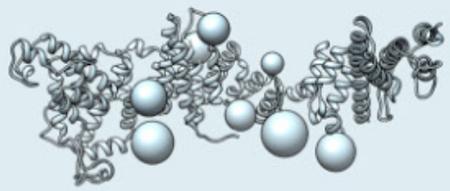
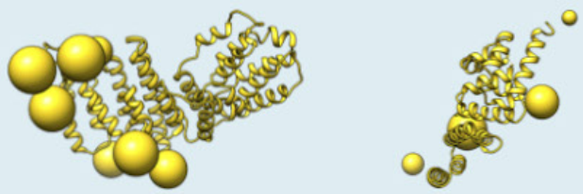
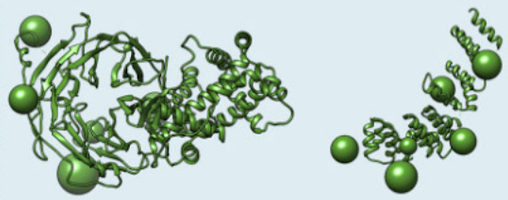
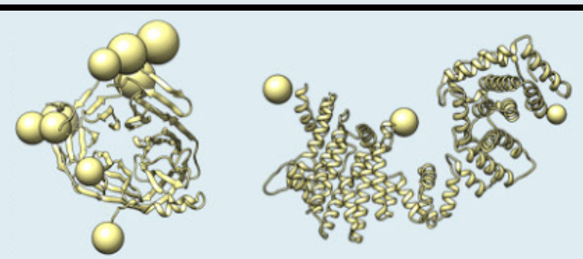
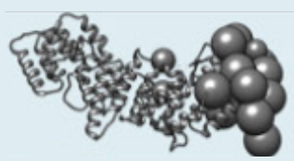
Comparative models

Nup84		1-6: flexible string of a bead (DISOPRED) 7-436: X-ray 3JRO_C, 3IKO_C (100% seq id) 429-488: Model 3F3F_G (10% seq id, HHpred) 489-505: flexible string of beads (DISOPRED) 506-726: Model 3CQC_A (18% seq id, HHpred)
Nup85		1-43: flexible string of beads (DISOPRED) 44-555: X-ray 3F3F_D, 3EWE_D (100% seq id) Linker (flexible string of a bead) 532-655: Model 2QX5_B (18% seq id, HHpred) 532-743: Model 4LCT_A (14% seq id, HHpred) 744: flexible string of a bead (PSIPRED)
Nup120		1-712: X-ray 3F7F_A, 3HXR_A (100% seq id) Linker (flexible string of a bead) 727-1037: Model 4FHN_B, 4FHN_D (14% seq id, HHpred)
Nup133		1-55: flexible string of beads (DISOPRED) 56-480: Model VpNup133 (46% seq id, Muscle) Linker (flexible string of a bead) 490-945: Model 3I4R (15% seq id, HHpred) 946-1157: X-ray 3KFO_A (100% seq id)
Nup145c		1-125: flexible string of beads (DISOPRED) 126-553: X-ray 3IKO_B, 3JRO_A, 3BG1_B, 3BGO_B (100% seq id) 554-712: flexible string of beads (PSIPRED)
Seh1		1-346: X-ray 3F3F (100% seq id) 347-349: flexible string of a bead (DISOPRED)
Sec13		1: flexible string of a bead (DISOPRED) 2-296: X-ray 2PM7_D (100% seq id) 297: flexible string of a bead (DISOPRED)



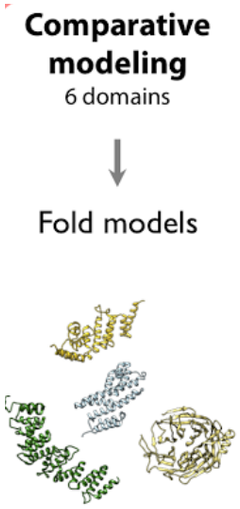
Comparative models

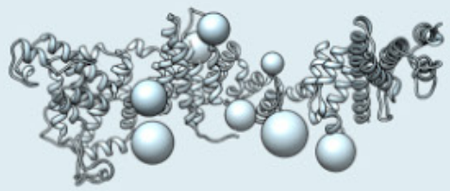
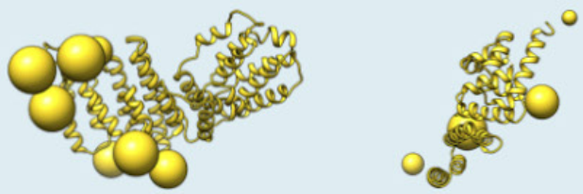
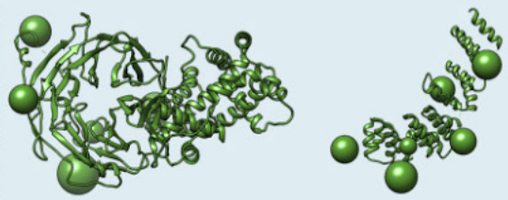
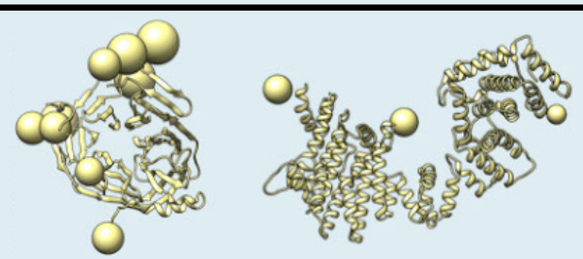
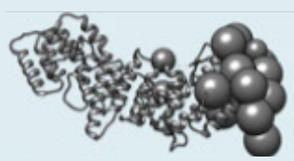


Nup84		1-6: flexible string of a bead (DISOPRED) 7-436: X-ray 3JRO_C, 3IKO_C (100% seq id) 429-488: Model 3F3F_G (10% seq id, HHpred) 489-505: flexible string of beads (DISOPRED) 506-726: Model 3CQC_A (18% seq id, HHpred)
Nup85		1-43: flexible string of beads (DISOPRED) 44-555: X-ray 3F3F_D, 3EWE_D (100% seq id) Linker (flexible string of a bead) 532-655: Model 2QX5_B (18% seq id, HHpred) 532-743: Model 4LCT_A (14% seq id, HHpred) 744: flexible string of a bead (PSIPRED)
Nup120		1-712: X-ray 3F7F_A, 3HXR_A (100% seq id) Linker (flexible string of a bead) 727-1037: Model 4FHN_B, 4FHN_D (14% seq id, HHpred)
Nup133		1-55: flexible string of beads (DISOPRED) 56-480: Model VpNup133 (46% seq id, Muscle) Linker (flexible string of a bead) 490-945: Model 3I4R (15% seq id, HHpred) 946-1157: X-ray 3KFO_A (100% seq id)
Nup145c		1-125: flexible string of beads (DISOPRED) 126-553: X-ray 3IKO_B, 3JRO_A, 3BG1_B, 3BGO_B (100% seq id) 554-712: flexible string of beads (PSIPRED)
Seh1		1-346: X-ray 3F3F (100% seq id) 347-349: flexible string of a bead (DISOPRED)
Sec13		1: flexible string of a bead (DISOPRED) 2-296: X-ray 2PM7_D (100% seq id) 297: flexible string of a bead (DISOPRED)

- For 6 domains (30% of overall sequence) no structure was available, but the structure of a related protein existed

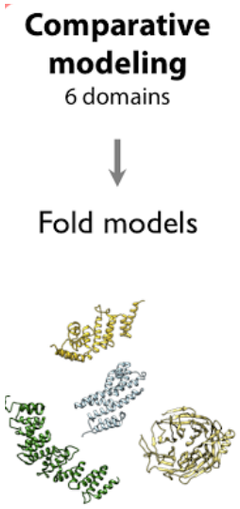
Comparative models

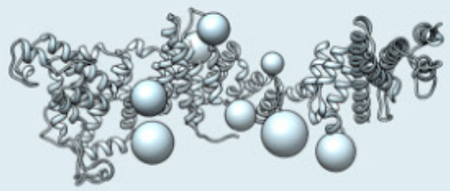
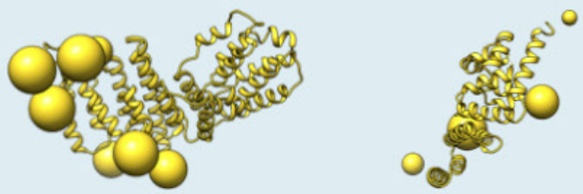
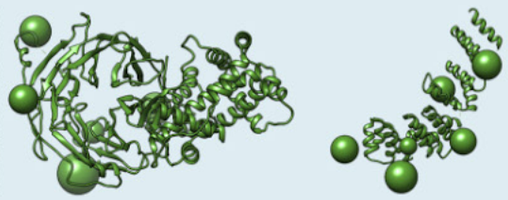
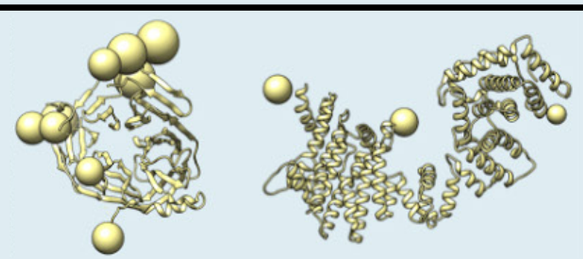
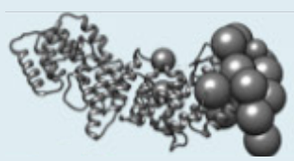


Nup84		1-6: flexible string of a bead (DISOPRED) 7-436: X-ray 3JRO_C, 3IKO_C (100% seq id) 429-488: Model 3F3F_G (10% seq id, HHpred) 489-505: flexible string of beads (DISOPRED) 506-726: Model 3CQC_A (18% seq id, HHpred)
Nup85		1-43: flexible string of beads (DISOPRED) 44-555: X-ray 3F3F_D, 3EWE_D (100% seq id) Linker (flexible string of a bead) 532-655: Model 2QX5_B (18% seq id, HHpred) 532-743: Model 4LCT_A (14% seq id, HHpred) 744: flexible string of a bead (PSIPRED)
Nup120		1-712: X-ray 3F7F_A, 3HXR_A (100% seq id) Linker (flexible string of a bead) 727-1037: Model 4FHN_B, 4FHN_D (14% seq id, HHpred)
Nup133		1-55: flexible string of beads (DISOPRED) 56-480: Model VpNup133 (46% seq id, Muscle) Linker (flexible string of a bead) 490-945: Model 3I4R (15% seq id, HHpred) 946-1157: X-ray 3KFO_A (100% seq id)
Nup145c		1-125: flexible string of beads (DISOPRED) 126-553: X-ray 3IKO_B, 3JRO_A, 3BG1_B, 3BGO_B (100% seq id) 554-712: flexible string of beads (PSIPRED)
Seh1		1-346: X-ray 3F3F (100% seq id) 347-349: flexible string of a bead (DISOPRED)
Sec13		1: flexible string of a bead (DISOPRED) 2-296: X-ray 2PM7_D (100% seq id) 297: flexible string of a bead (DISOPRED)

- For 6 domains (30% of overall sequence) no structure was available, but the structure of a related protein existed
- Comparative models were built for these regions, mostly using MODELLER

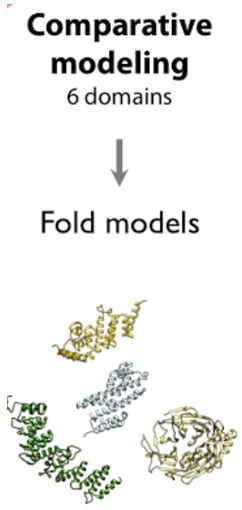
Comparative models



Nup84		1-6: flexible string of a bead (DISOPRED) 7-436: X-ray 3JRO_C, 3IKO_C (100% seq id) 429-488: Model 3F3F_G (10% seq id, HHpred) 489-505: flexible string of beads (DISOPRED) 506-726: Model 3CQC_A (18% seq id, HHpred)
Nup85		1-43: flexible string of beads (DISOPRED) 44-555: X-ray 3F3F_D, 3EWE_D (100% seq id) Linker (flexible string of a bead) 532-655: Model 2QX5_B (18% seq id, HHpred) 532-743: Model 4LCT_A (14% seq id, HHpred) 744: flexible string of a bead (PSIPRED)
Nup120		1-712: X-ray 3F7F_A, 3HXR_A (100% seq id) Linker (flexible string of a bead) 727-1037: Model 4FHN_B, 4FHN_D (14% seq id, HHpred)
Nup133		1-55: flexible string of beads (DISOPRED) 56-480: Model VpNup133 (46% seq id, Muscle) Linker (flexible string of a bead) 490-945: Model 3I4R (15% seq id, HHpred) 946-1157: X-ray 3KFO_A (100% seq id)
Nup145c		1-125: flexible string of beads (DISOPRED) 126-553: X-ray 3IKO_B, 3JRO_A, 3BG1_B, 3BGO_B (100% seq id) 554-712: flexible string of beads (PSIPRED)
Seh1		1-346: X-ray 3F3F (100% seq id) 347-349: flexible string of a bead (DISOPRED)
Sec13		1: flexible string of a bead (DISOPRED) 2-296: X-ray 2PM7_D (100% seq id) 297: flexible string of a bead (DISOPRED)

- For 6 domains (30% of overall sequence) no structure was available, but the structure of a related protein existed
- Comparative models were built for these regions, mostly using MODELLER

Comparative modeling by satisfaction of spatial restraints: MODELLER



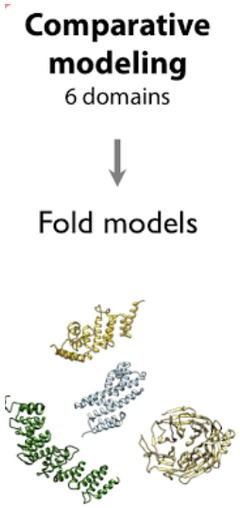
3D GKITFYERGFQGHCYESDC-NLQP…

SEQ GKITFYERG---RCYESDCPNLQP…

A. Šali & T. Blundell. *J. Mol. Biol.* **234**, 779, 1993.
J.P. Overington & A. Šali. *Prot. Sci*. **3**, 1582, 1994.
A. Fiser, R. Do & A. Šali, *Prot. Sci*., **9**, 1753, 2000.

https://salilab.org/modeller/

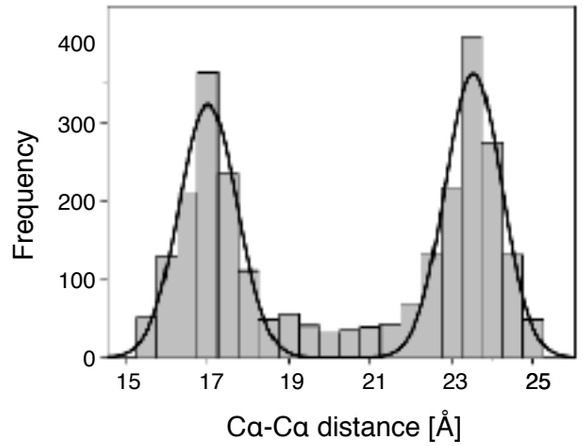
Comparative modeling by satisfaction of spatial restraints: MODELLER



3D GKITFYERGFQGHCYESDC-NLQP…

SEQ GKITFYERG---RCYESDCPNLQP…

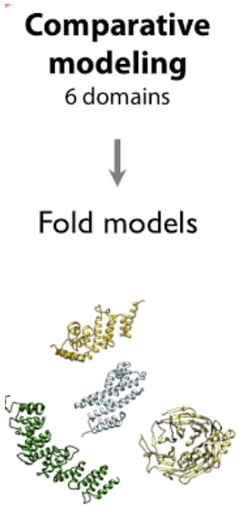
1. Extract spatial restraints



A. Šali & T. Blundell. *J. Mol. Biol.* **234**, 779, 1993.
J.P. Overington & A. Šali. *Prot. Sci*. **3**, 1582, 1994.
A. Fiser, R. Do & A. Šali, *Prot. Sci*., **9**, 1753, 2000.

https://salilab.org/modeller/

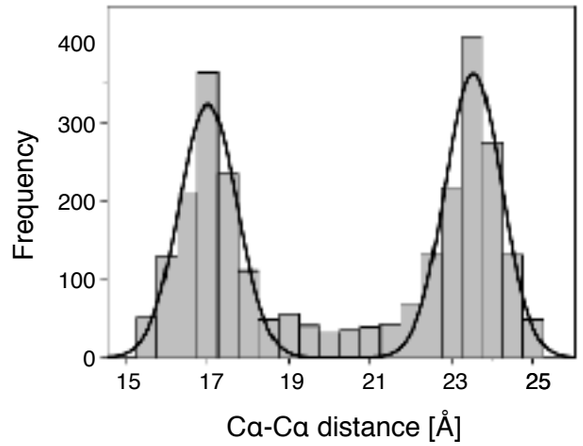
Comparative modeling by satisfaction of spatial restraints: MODELLER



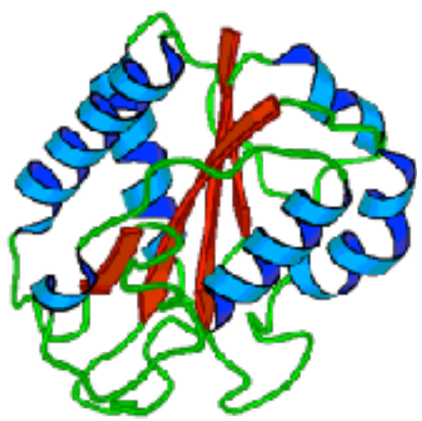
3D GKITFYERGFQGHCYESDC-NLQP…

SEQ GKITFYERG---RCYESDCPNLQP…

1. Extract spatial restraints



2. Satisfy spatial restraints

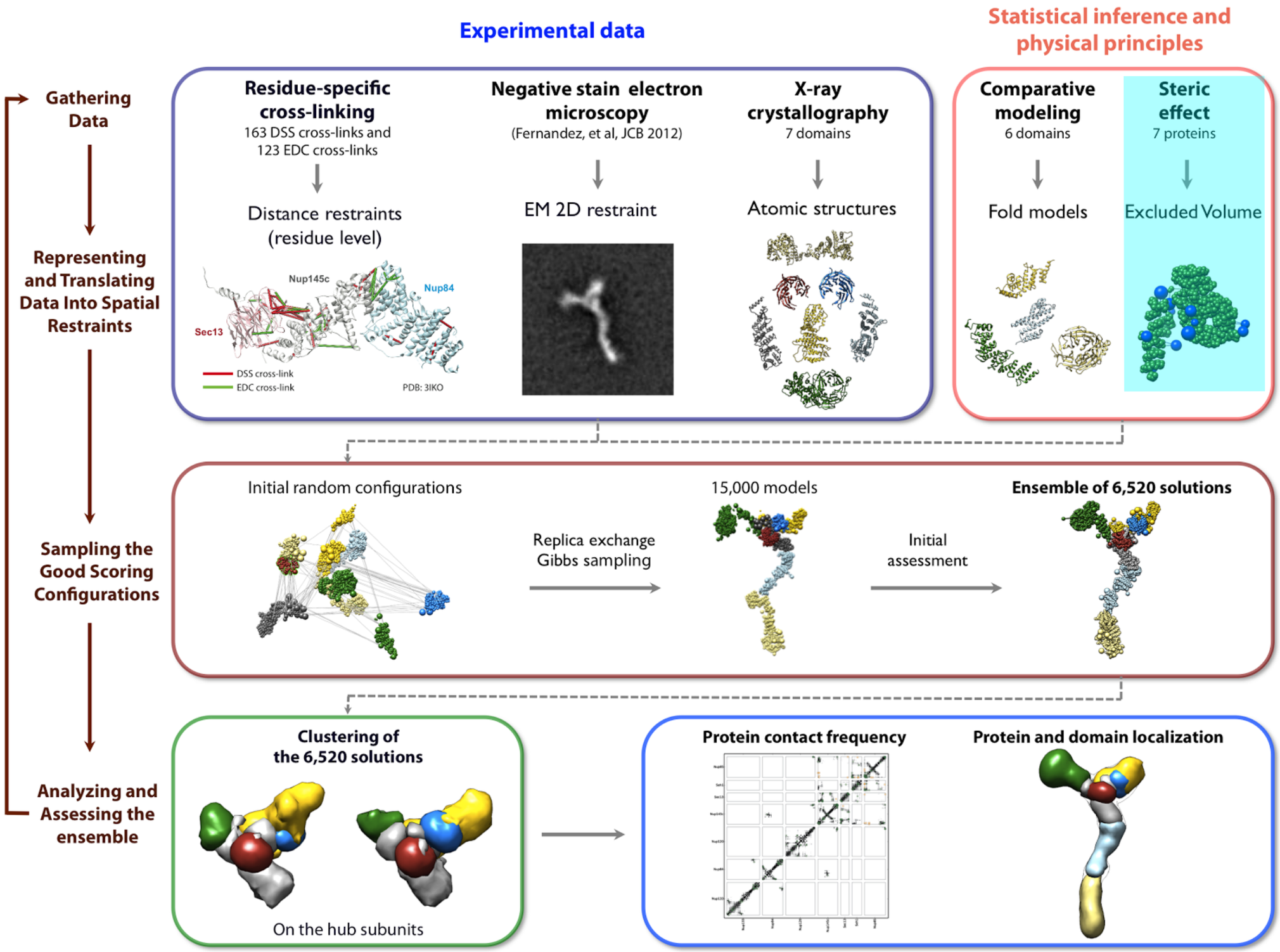


$$F(\mathbf{R}) = \prod_{i} p_i(f_i/I)$$

A. Šali & T. Blundell. *J. Mol. Biol.* **234**, 779, 1993.
J.P. Overington & A. Šali. *Prot. Sci*. **3**, 1582, 1994.
A. Fiser, R. Do & A. Šali, *Prot. Sci*., **9**, 1753, 2000.

https://salilab.org/modeller/

Modeling Nup84 with IMP (2014)



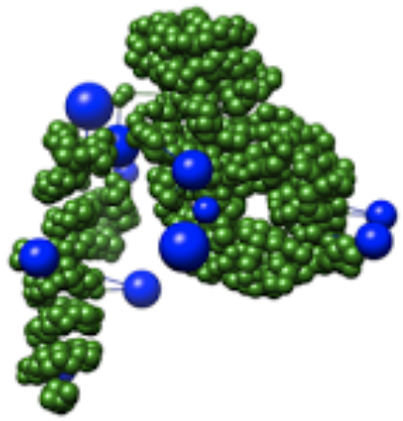
Mol Cell Proteomics 13, 2927-2943, 2014

Steric effect

7 proteins

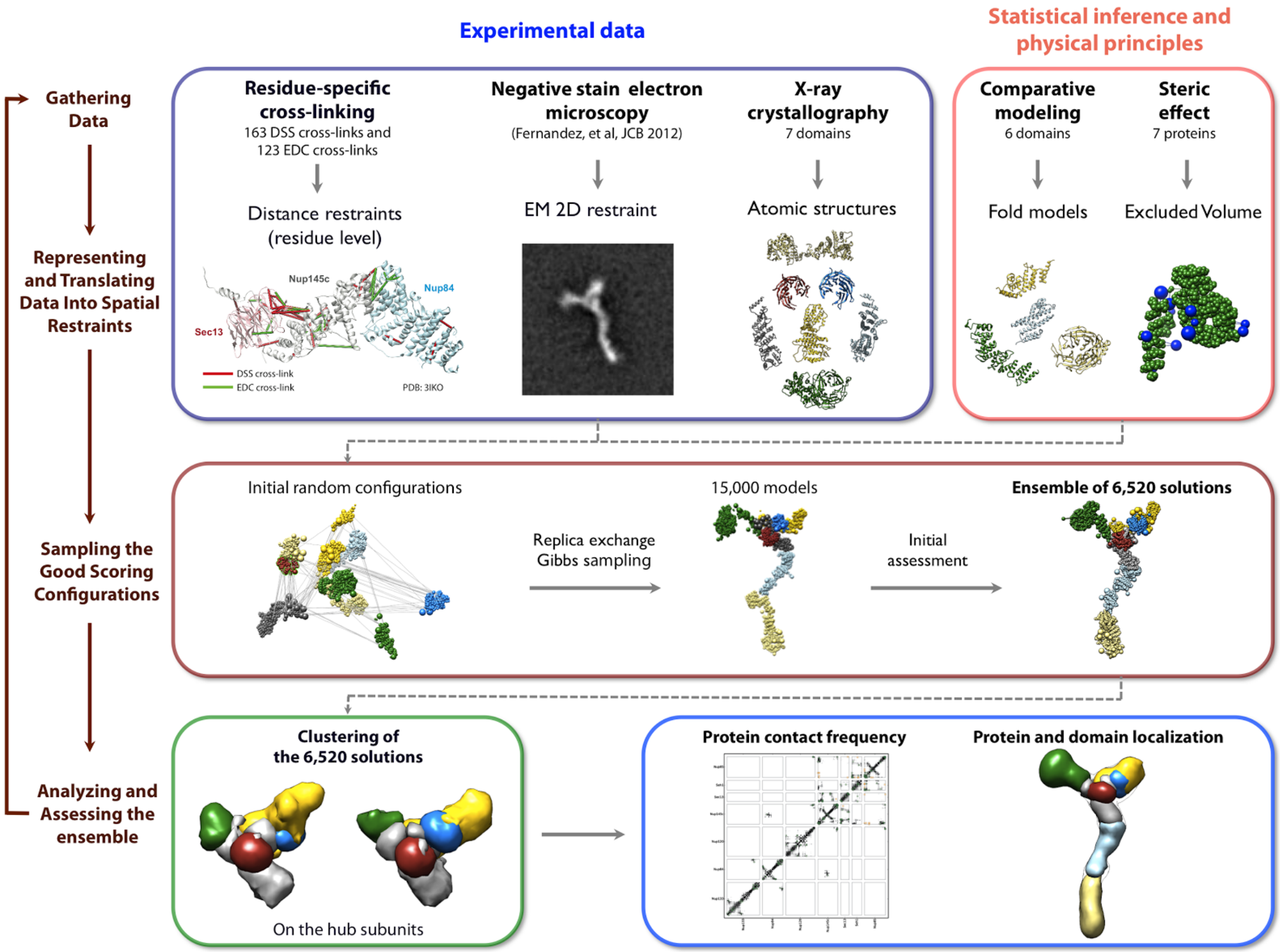
↓

Excluded Volume



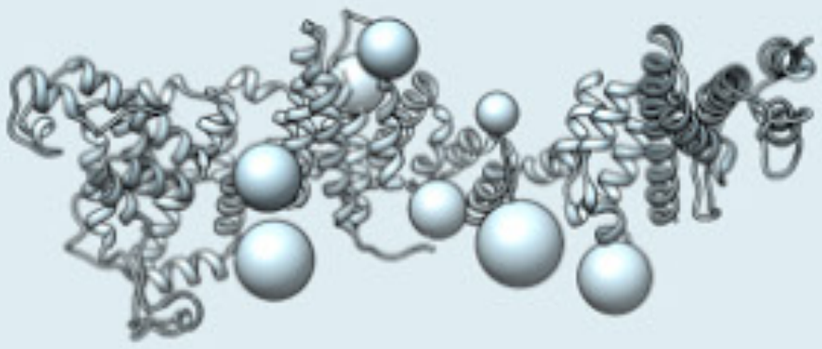
- Since we know proteins cannot occupy the same volume, we added a simple excluded volume restraint
  - Repulsive spring (soft sphere) between pairs of particles

Modeling Nup84 with IMP (2014)



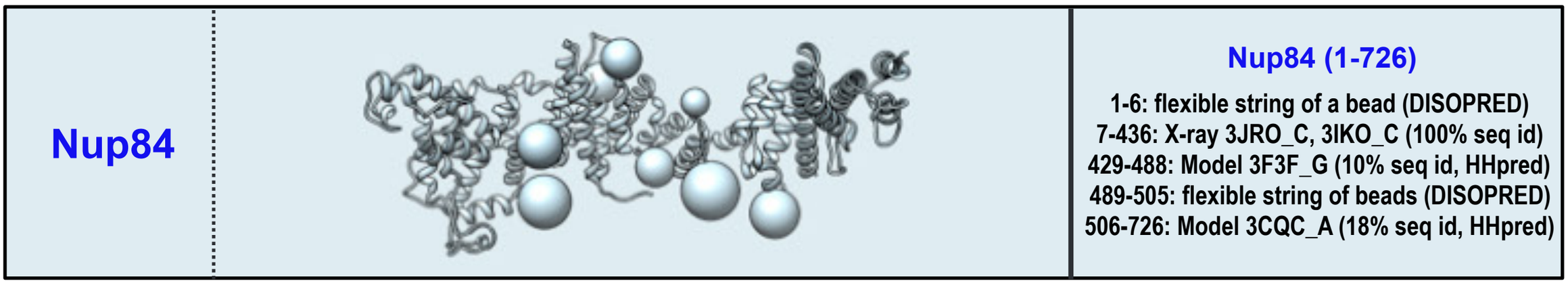
Mol Cell Proteomics 13, 2927-2943, 2014

Representation

Nup84		**Nup84 (1-726)** 1-6: flexible string of a bead (DISOPRED) 7-436: X-ray 3JRO_C, 3IKO_C (100% seq id) 429-488: Model 3F3F_G (10% seq id, HHpred) 489-505: flexible string of beads (DISOPRED) 506-726: Model 3CQC_A (18% seq id, HHpred)

http://bioinf.cs.ucl.ac.uk/psipred/

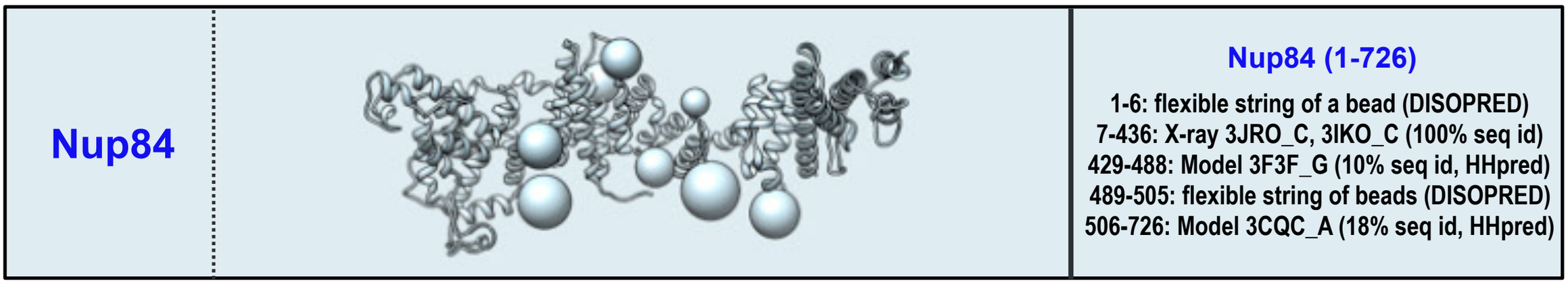
Representation



- Regions with known structures from X-ray or comparative modeling: represent each residue as a sphere, treat the entire region as a rigid body

http://bioinf.cs.ucl.ac.uk/psipred/

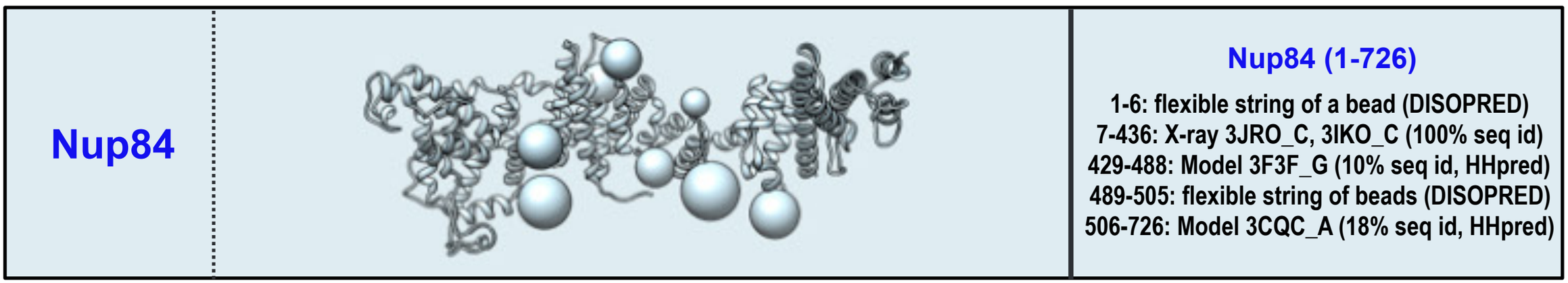
Representation



- Regions with known structures from X-ray or comparative modeling: represent each residue as a sphere, treat the entire region as a rigid body
- Unknown structure (or predicted disordered by DISOPRED): represent up to 20 residues as a single sphere “bead”, allow them to move, add simple spring restraints between consecutive beads to maintain sequence connectivity (flexible string of beads)

http://bioinf.cs.ucl.ac.uk/psipred/

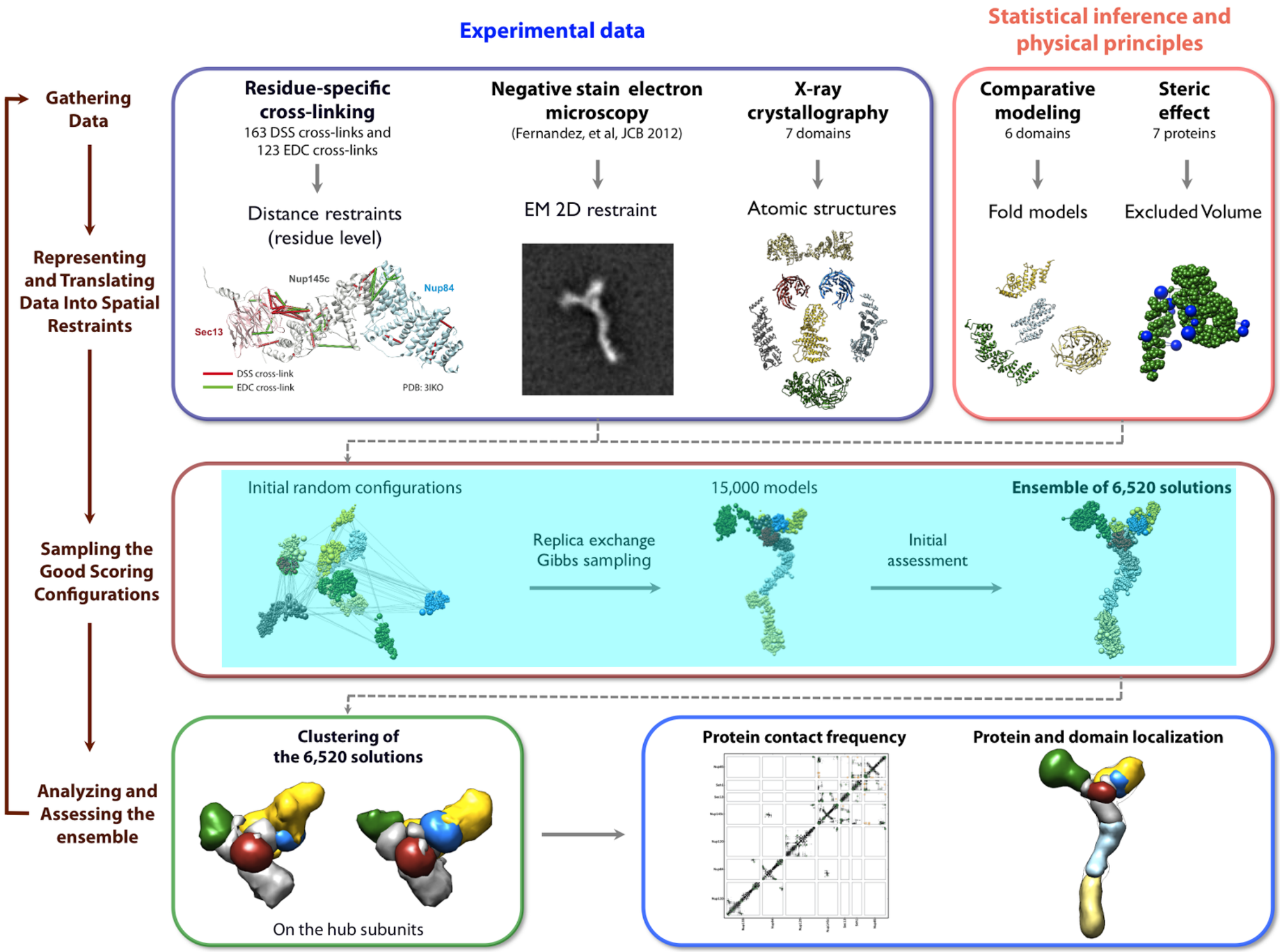
Representation



- Regions with known structures from X-ray or comparative modeling: represent each residue as a sphere, treat the entire region as a rigid body
- Unknown structure (or predicted disordered by DISOPRED): represent up to 20 residues as a single sphere “bead”, allow them to move, add simple spring restraints between consecutive beads to maintain sequence connectivity (flexible string of beads)
- Not coarse-graining for *speed* but to avoid *overinterpretation* of the data

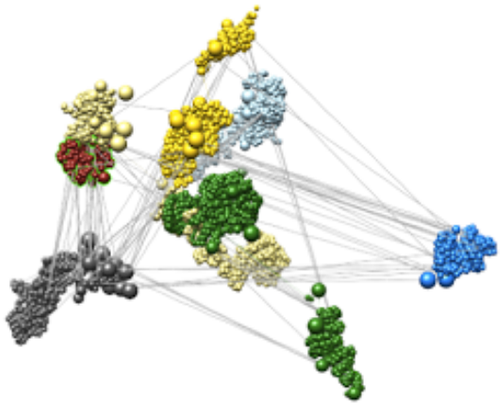
http://bioinf.cs.ucl.ac.uk/psipred/

Modeling Nup84 with IMP (2014)



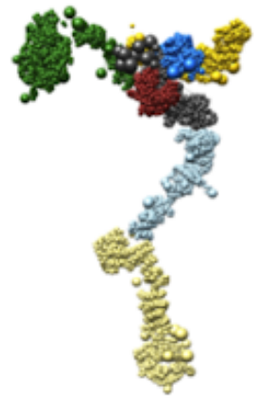
Mol Cell Proteomics 13, 2927-2943, 2014

Initial random configurations



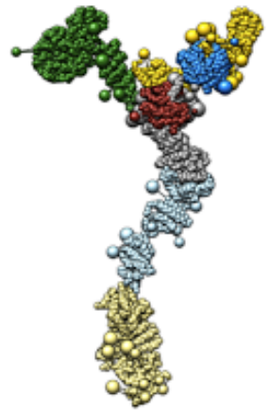
Replica exchange
Gibbs sampling

15,000 models

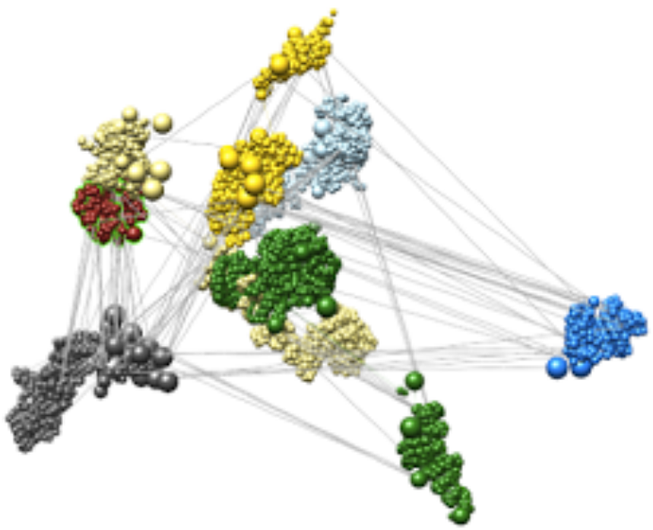


Initial
assessment

Ensemble of 6,520 solutions

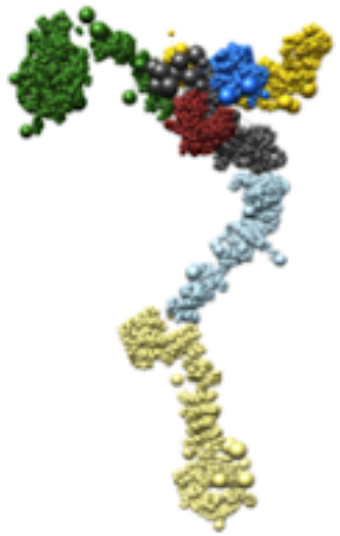


Sampling



Initial random configuration of subunits

Replica exchange Gibbs sampling using Metropolis Monte Carlo with 64 replicas



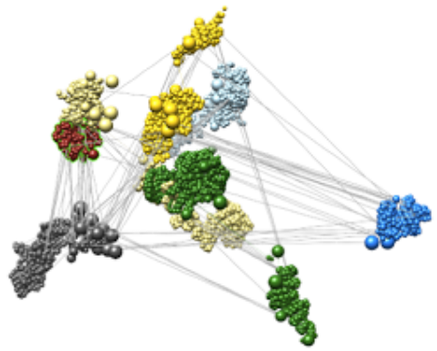
~750 candidate models

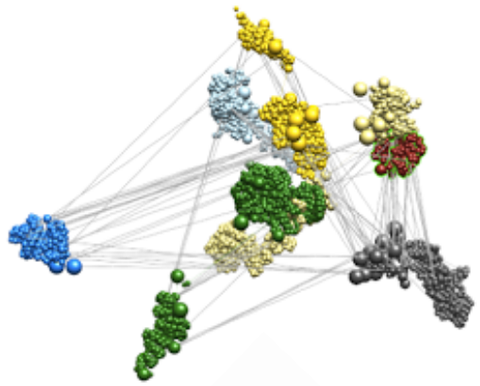
At each step, perturb the system by

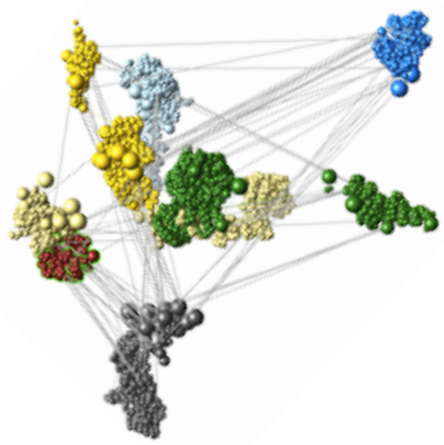
- translating and rotating rigid bodies
- translating beads
- adjusting Bayes σ parameter

Evaluate score as sum of all restraints

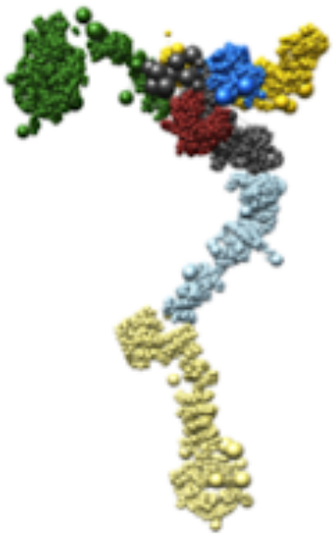
Sampling







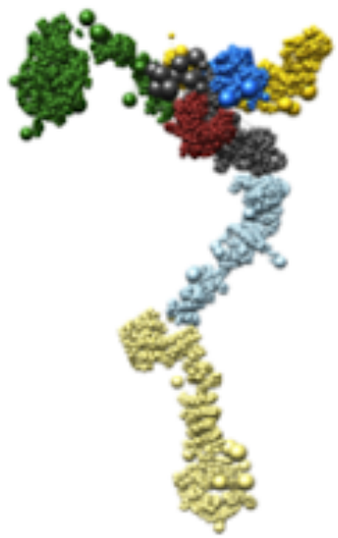
20 independent runs with different starting configurations



Total of 15000 candidate models

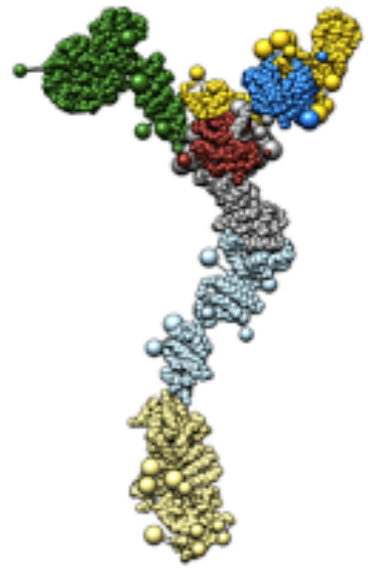
Total time: about 2 weeks on 1280 cores (in 2014)

Filtering



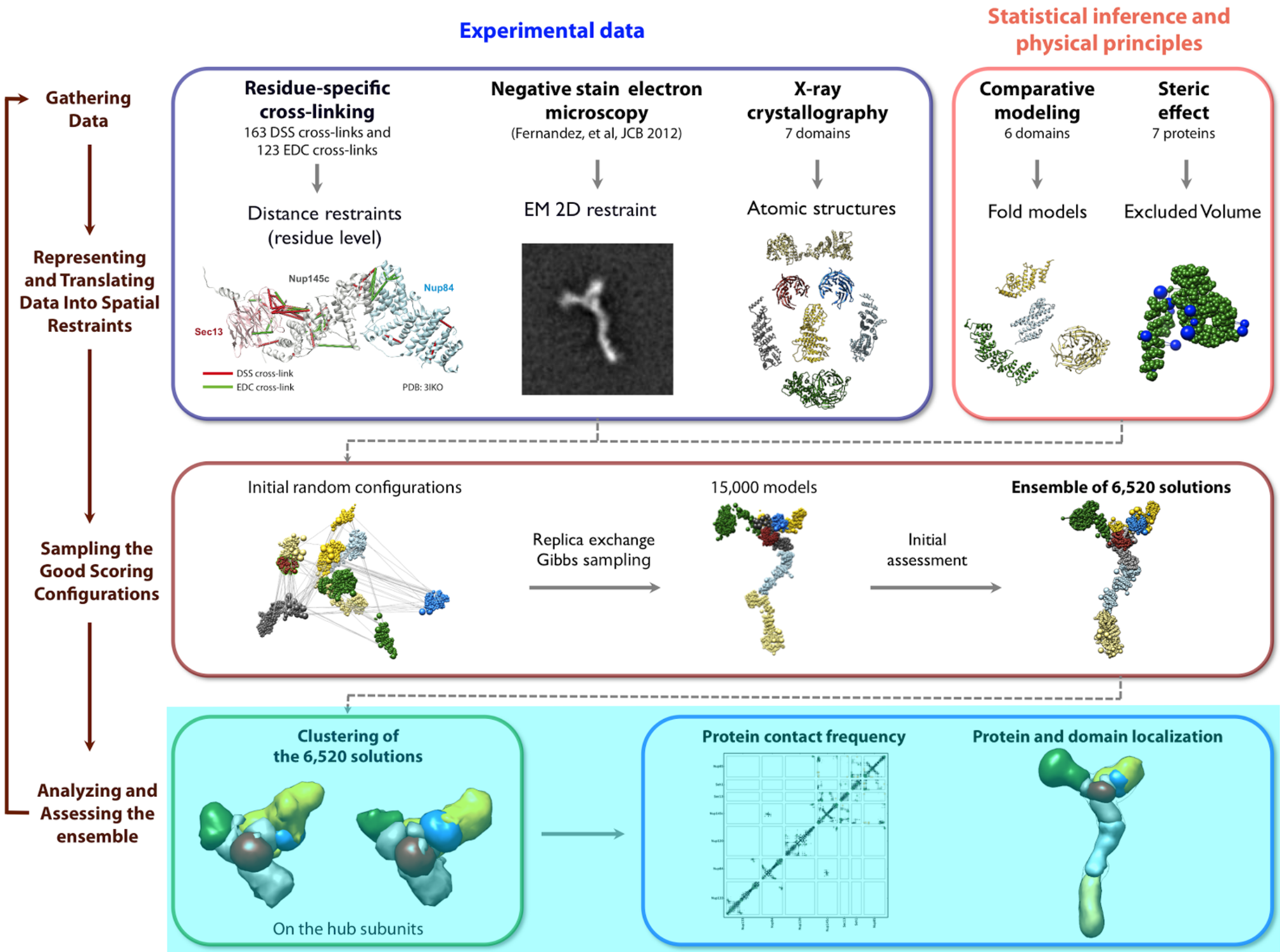
Total of 15000 candidate models

Keep only good-scoring models



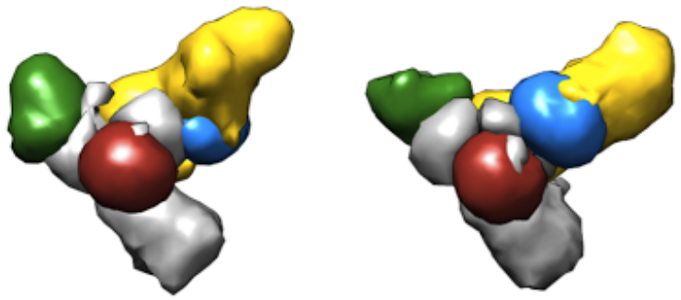
Ensemble of 6520 models

Modeling Nup84 with IMP (2014)



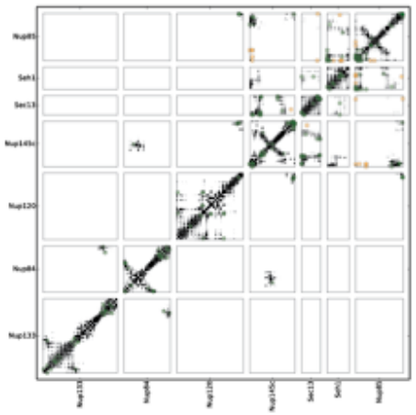
Mol Cell Proteomics 13, 2927-2943, 2014

Clustering of the 6,520 solutions

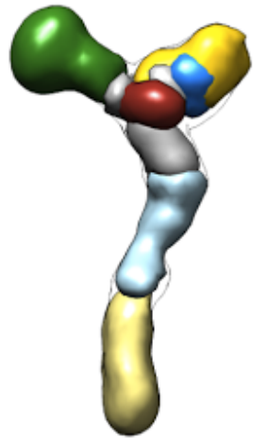


On the hub subunits

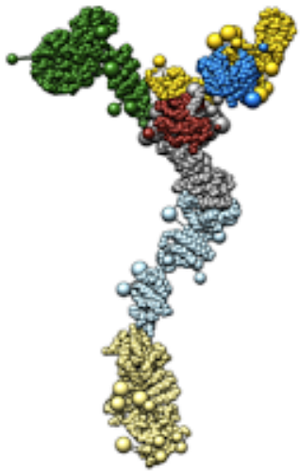
Protein contact frequency



Protein and domain localization

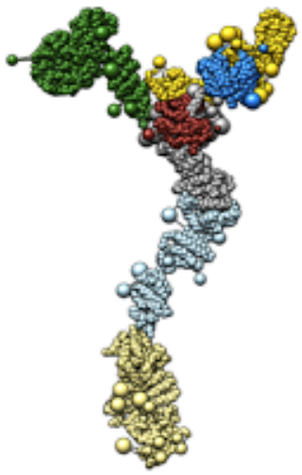


Clustering



Ensemble of 6520 models

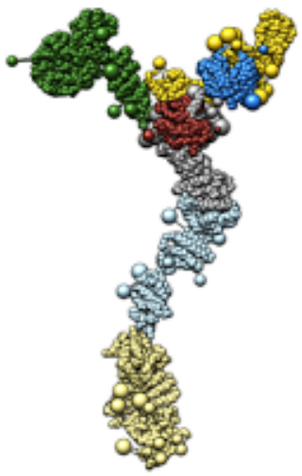
Clustering



Ensemble of 6520 models

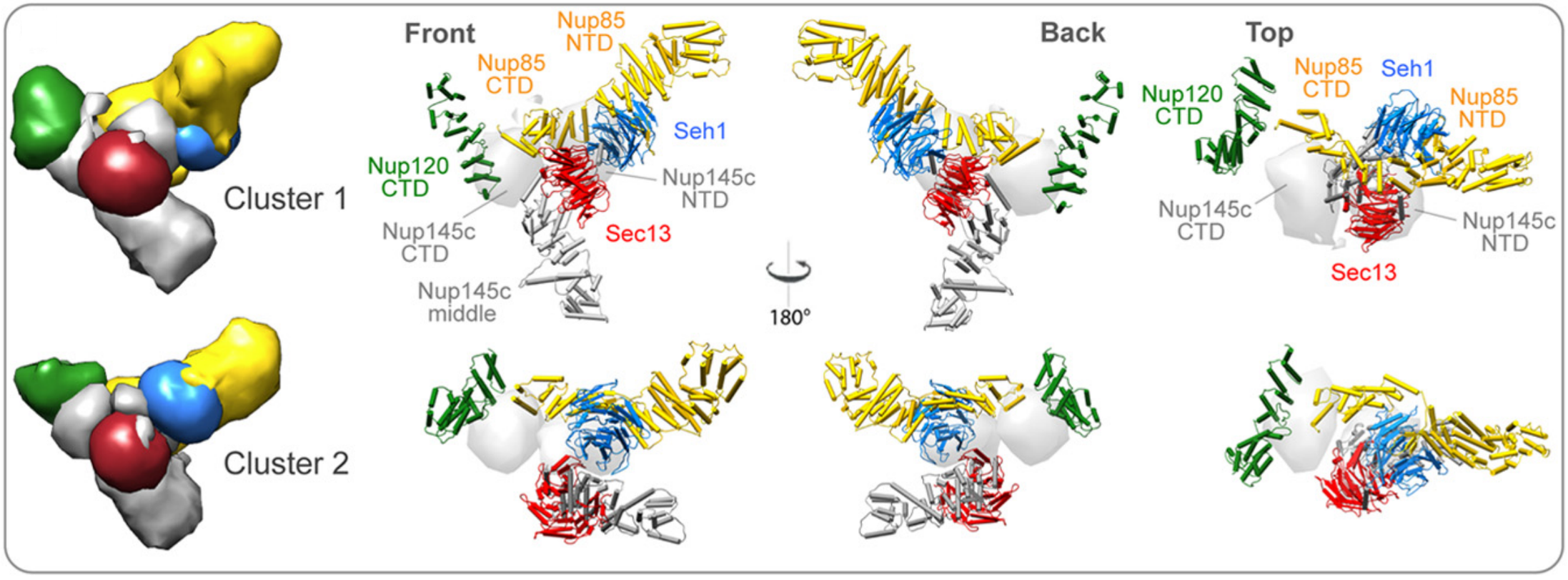
Cluster by RMSD

Clustering

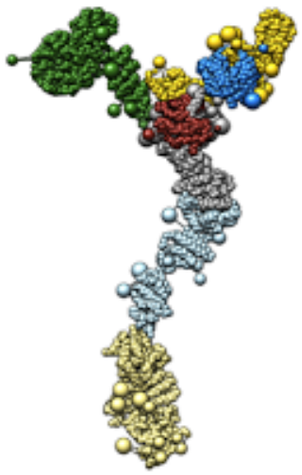


Ensemble of 6520 models

Cluster by RMSD



Clustering

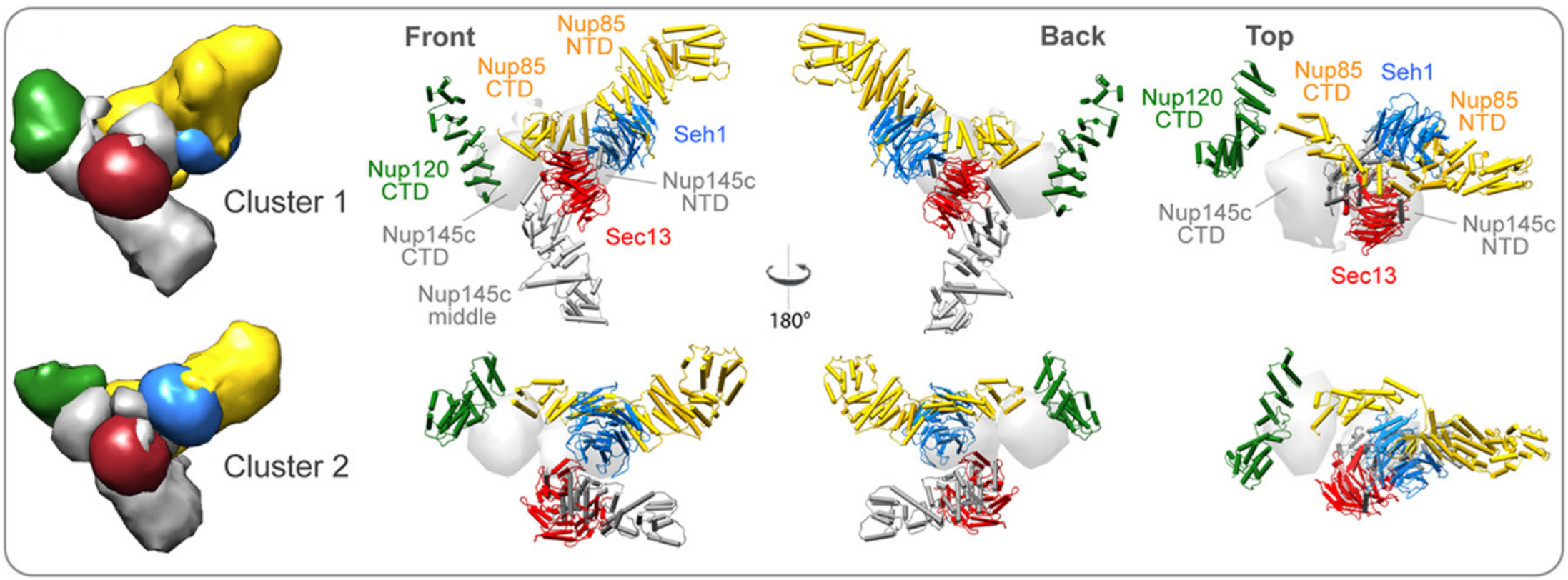


Ensemble of 6520 models

Cluster by RMSD

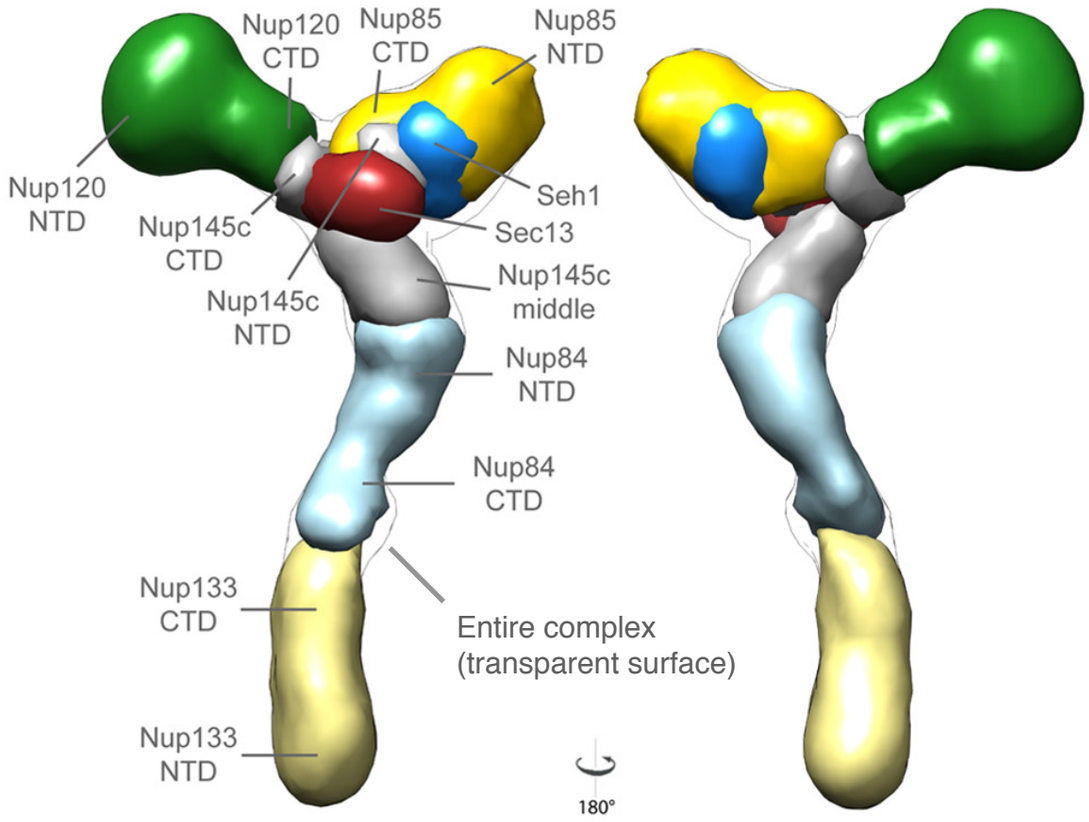
Two dominant clusters are found (of 1257 and 1010 models)

- Much improved ‘hub’ detail from 2012
- Two possible positions of Seh1 are supported by the data
- May suggest flexibility in the complex (Nup85-Seh1 arm)



Localization densities

- The probability of any volume element being occupied by a given protein, over the entire ensemble



Representative single structure overlaid on the localization density maps

Validation

How confident can we be in the final models?

- Completeness of sampling
- Agreement with input data
- Agreement with other data not used in the modeling

Completeness of sampling

Completeness of sampling

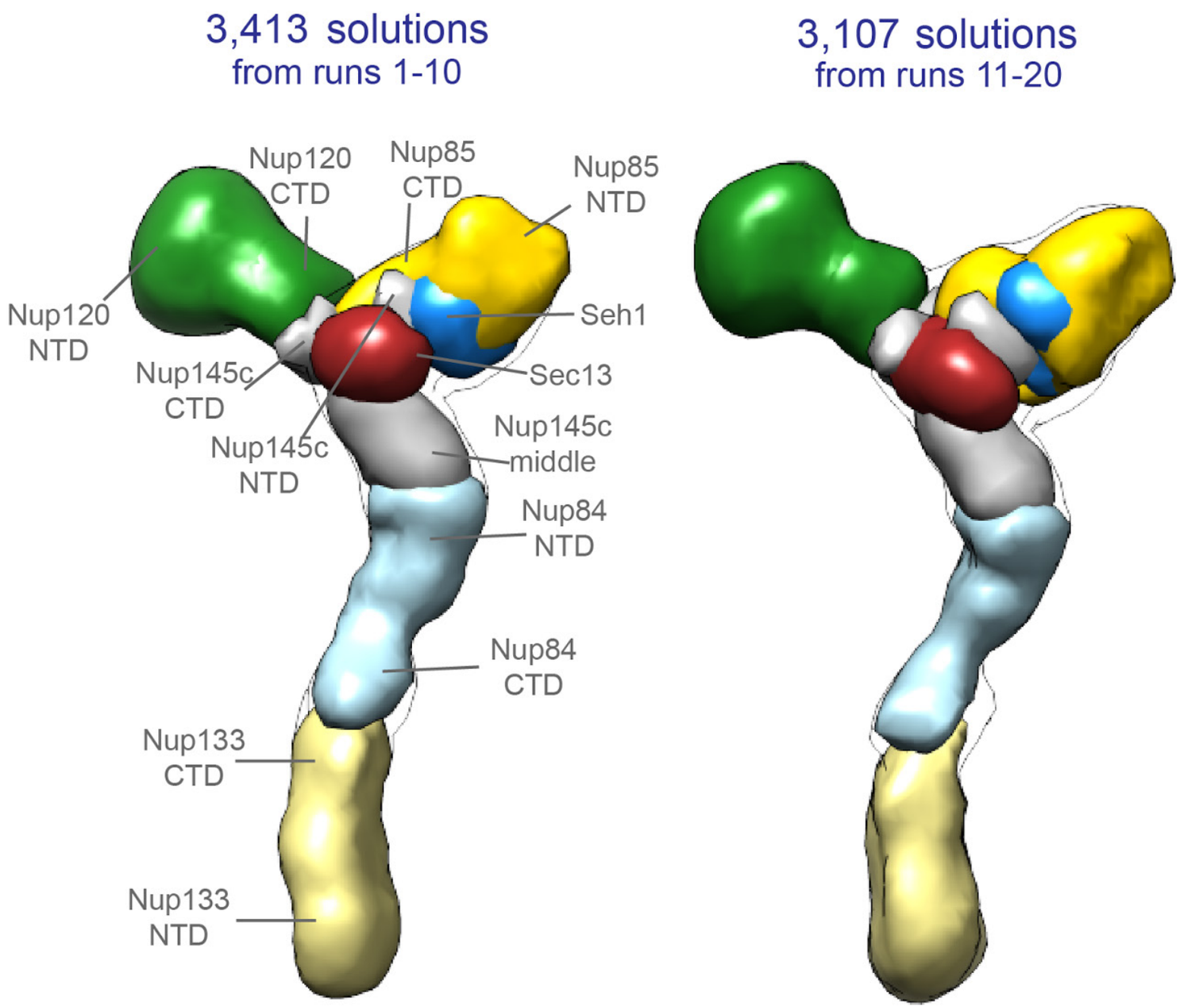
- Recall that 20 independent runs were carried out

Completeness of sampling

- Recall that 20 independent runs were carried out
- Split in half and compare results

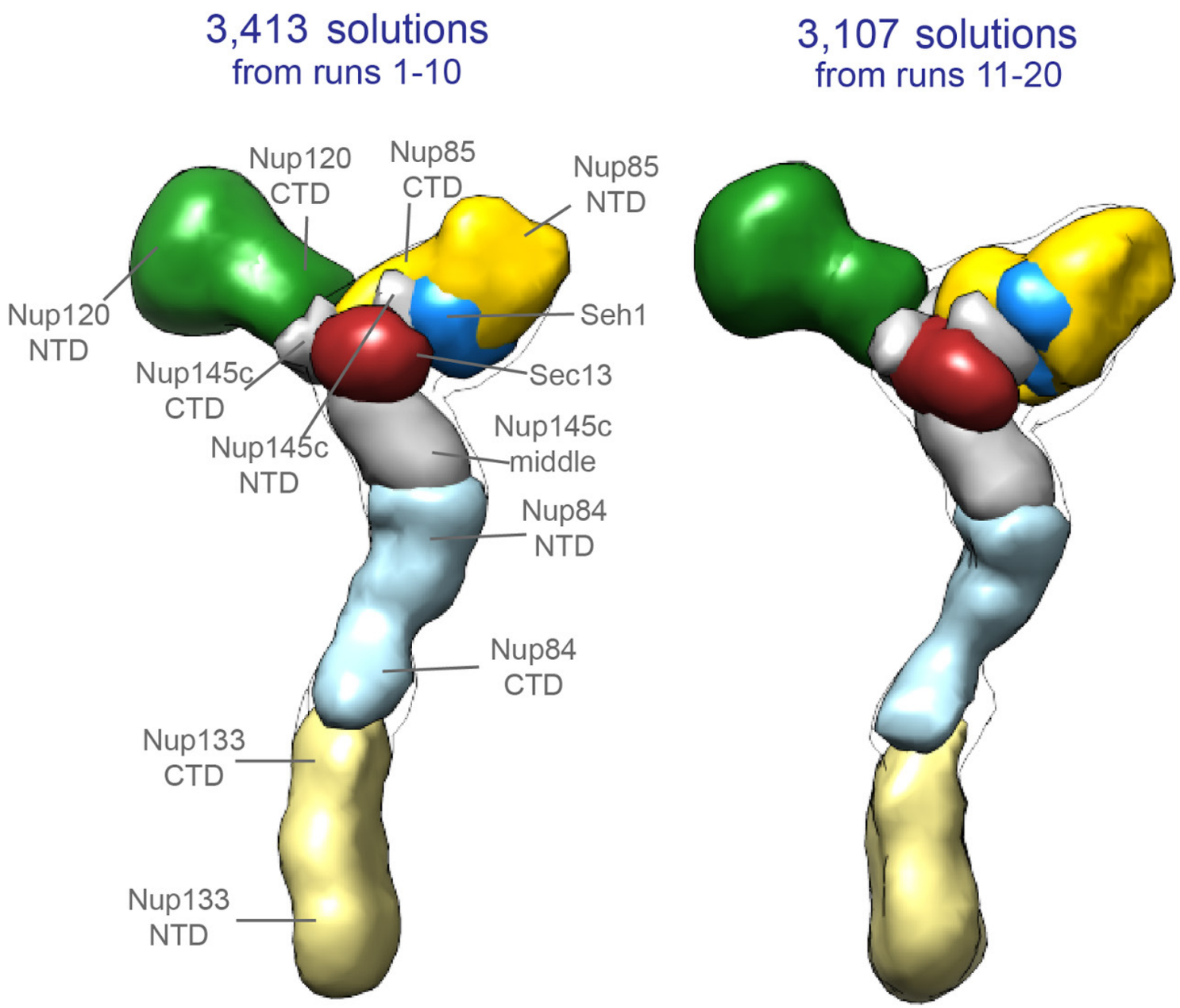
Completeness of sampling

- Recall that 20 independent runs were carried out
- Split in half and compare results



Completeness of sampling

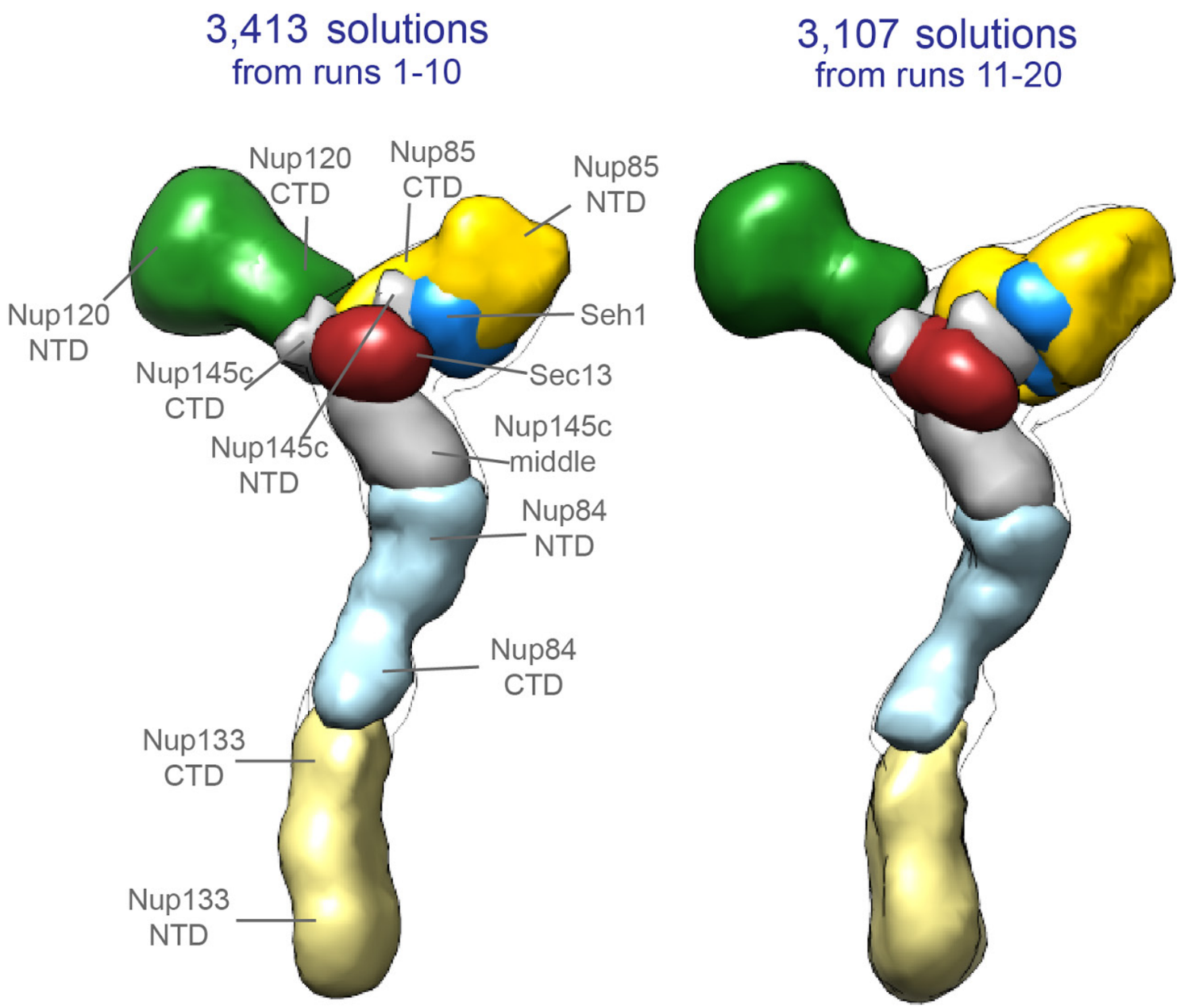
- Recall that 20 independent runs were carried out
- Split in half and compare results
- Ensemble derived from runs 1-10 is very similar to that from 11-20



Completeness of sampling

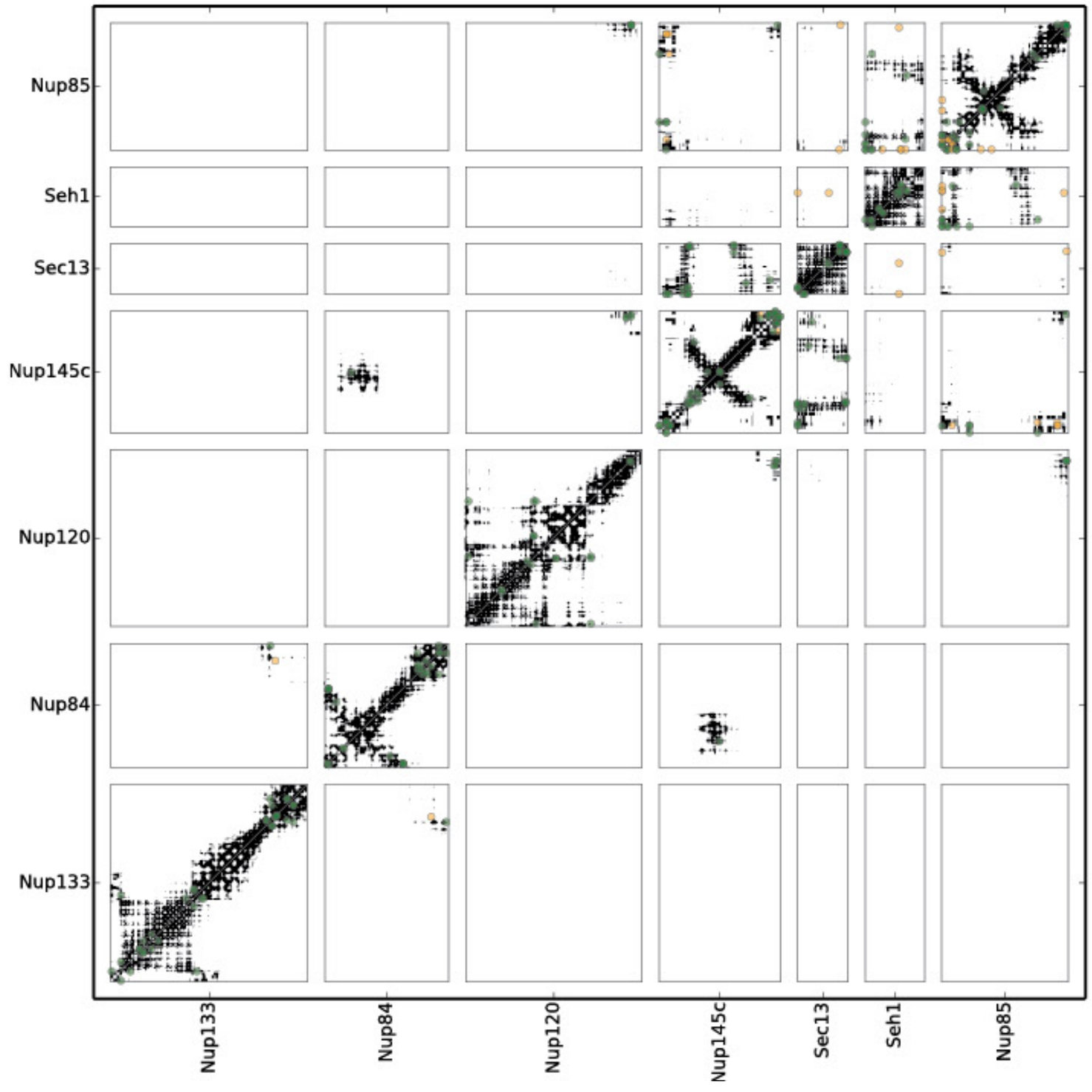
- Recall that 20 independent runs were carried out
- Split in half and compare results

- Ensemble derived from runs 1-10 is very similar to that from 11-20
- Strongly suggests that sampling is complete



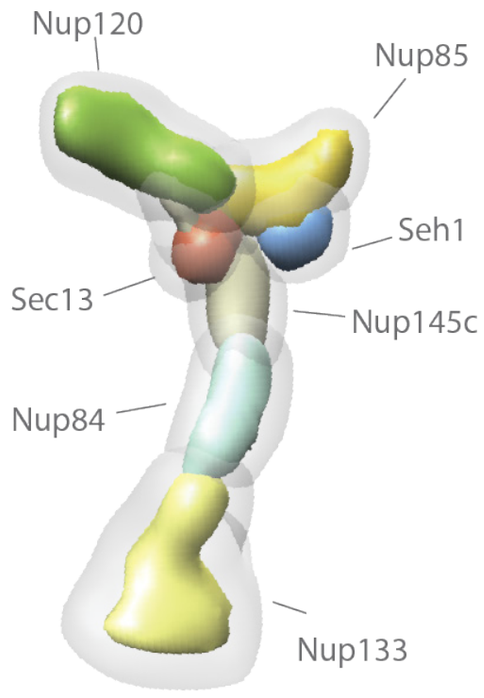
Agreement with input data

- 86.5% of DSS and 83.6% of EDC cross-links satisfied
- 99% of solutions satisfied excluded volume and sequence connectivity
- Solutions fit the EM class average with an average cross-correlation coefficient of 0.9

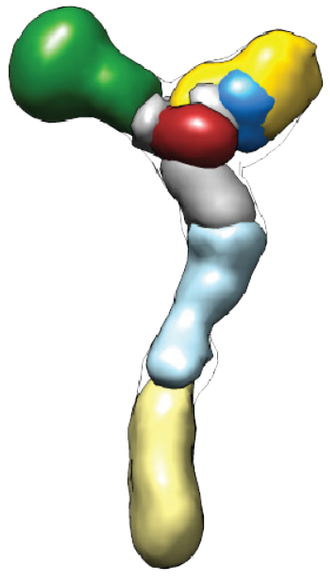


Agreement with data not used in the modeling

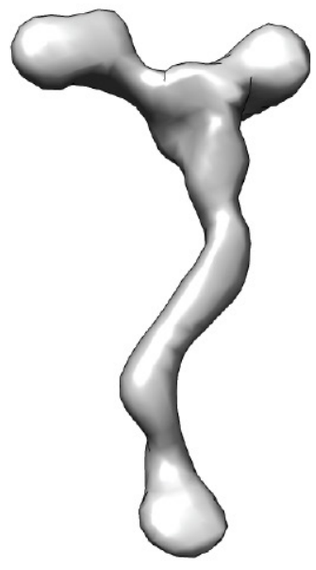
- Models are in good agreement with those from the 2012 study, and also closely match a 3D EM map that was not used in the modeling



2012 IMP study



2014 IMP study



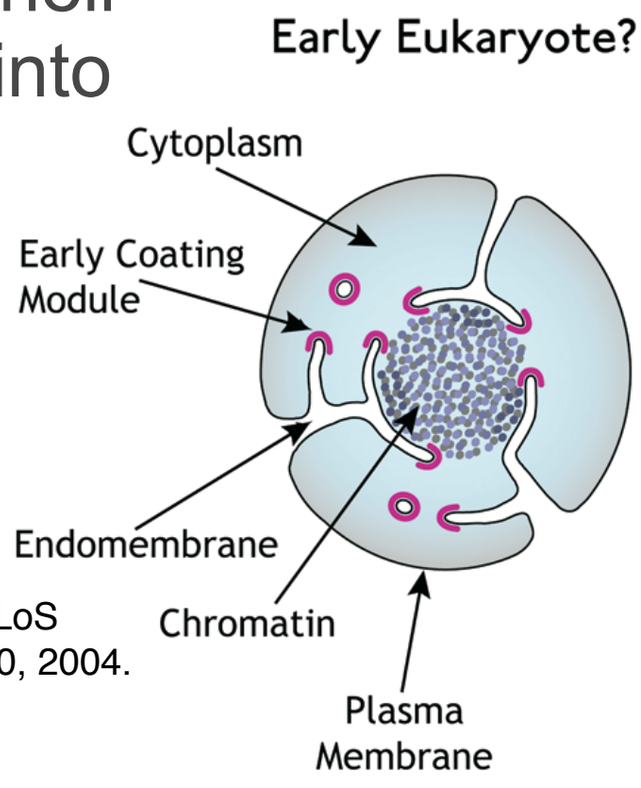
Negative stain 3D EM
(Kampmann *et al.* 2009)

Dimer interfaces

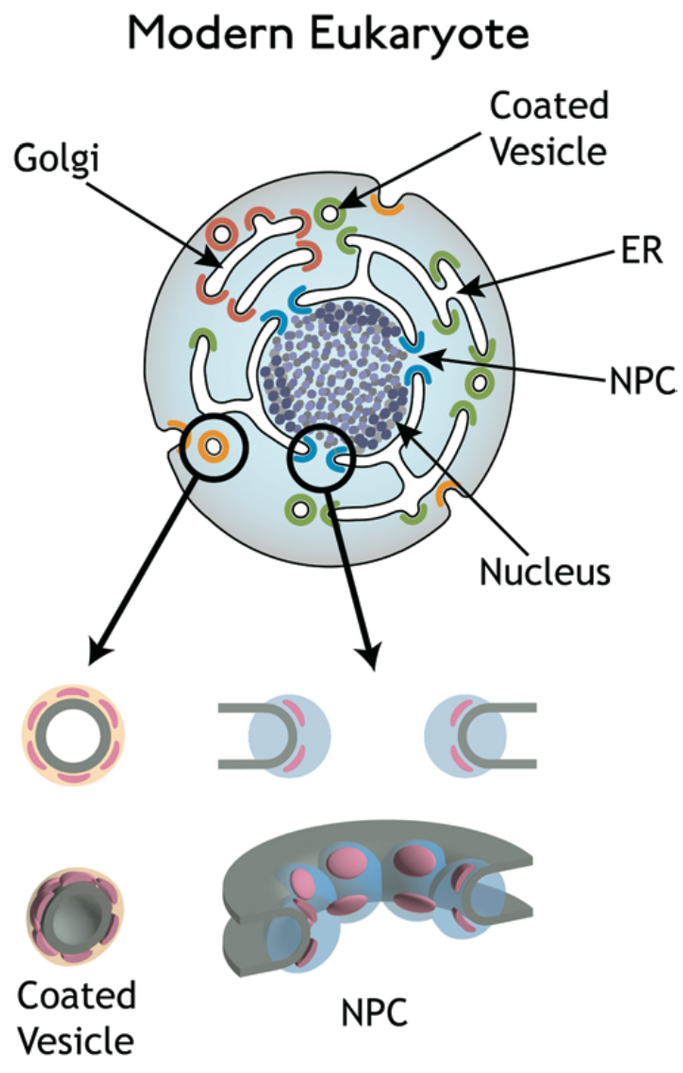
- X-ray structures include three crystallographic interfaces:
  - Nup145c–Sec13
  - Nup85–Seh1
  - Nup84–Nup145c
- These were included as input data (by virtue of keeping the X-ray structure regions as rigid bodies)
- However, repeating the modeling with these dimers broken up into monomers still yields the correct dimer interfaces (with an accuracy of 4.0, 12.0 and 7.5 Å respectively)
  - i.e. if the data were not used, the structures still end up being consistent with the information
  - most likely enough cross-links span the dimer interfaces

Protocoatamer hypothesis

- Similar protein folds are found in the NPC and coating related complexes
- Protocoatamer hypothesis: early eukaryotes developed simple coating modules to fold their membranes, developed into complex structures in modern eukaryotes
- Some Nup84 subunits (Seh1, Sec13) are also found in other coating related complexes
- IMP-generated structure supports this hypothesis (similar dimer arrangements to coating complexes)

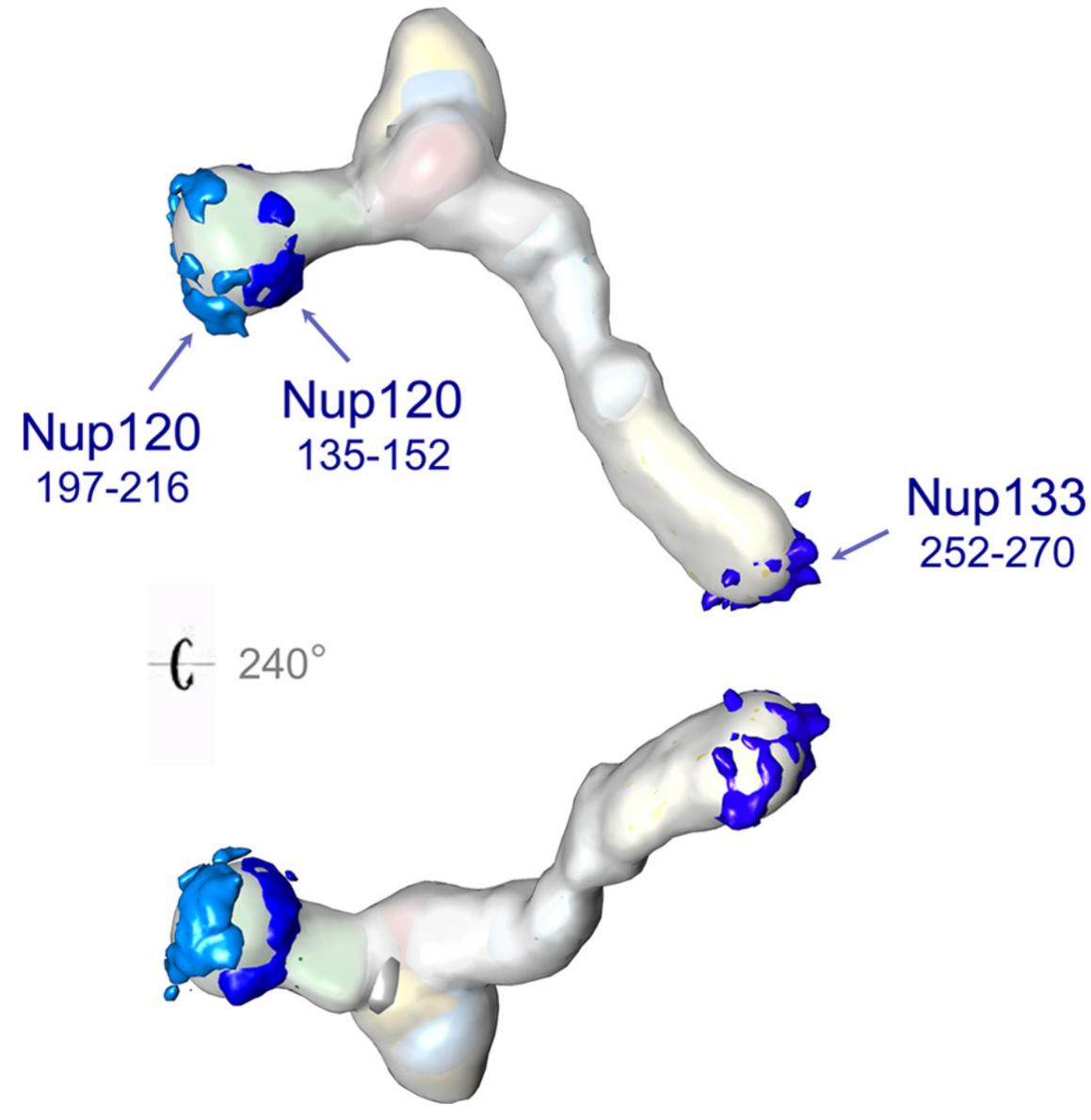


Devos *et* al.,PLoS Biology 2, e380, 2004.

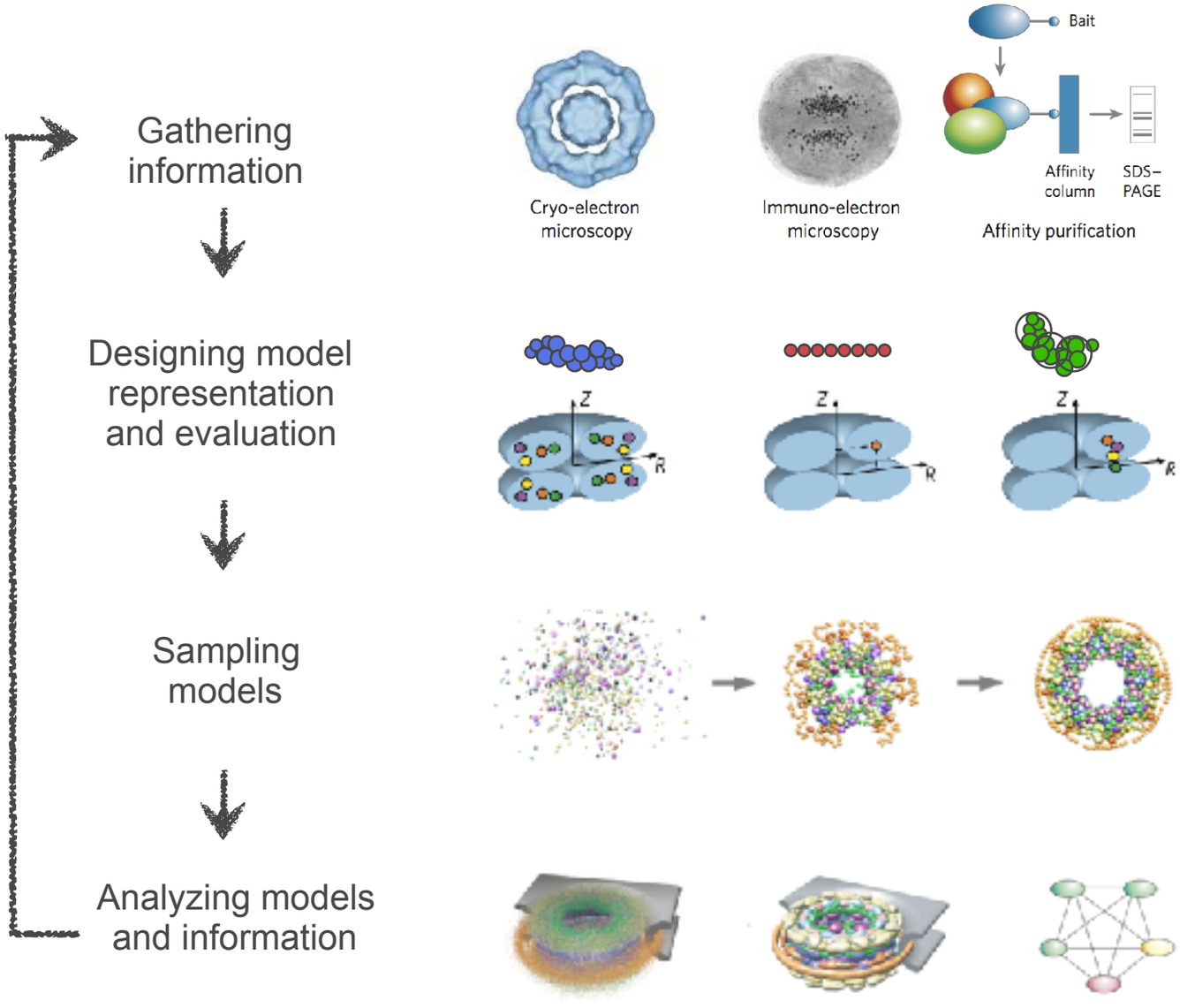


ALPS motifs

- Believed that ArfGAP1 lipid packing sensor (ALPS) motifs help anchor protein to membrane
- Mapping ALPS motifs onto Nup84 structure shows them at the periphery
- Consistent with prediction that Nup120 and Nup133 contact the nuclear membrane and stabilize the curvature of the entire NPC

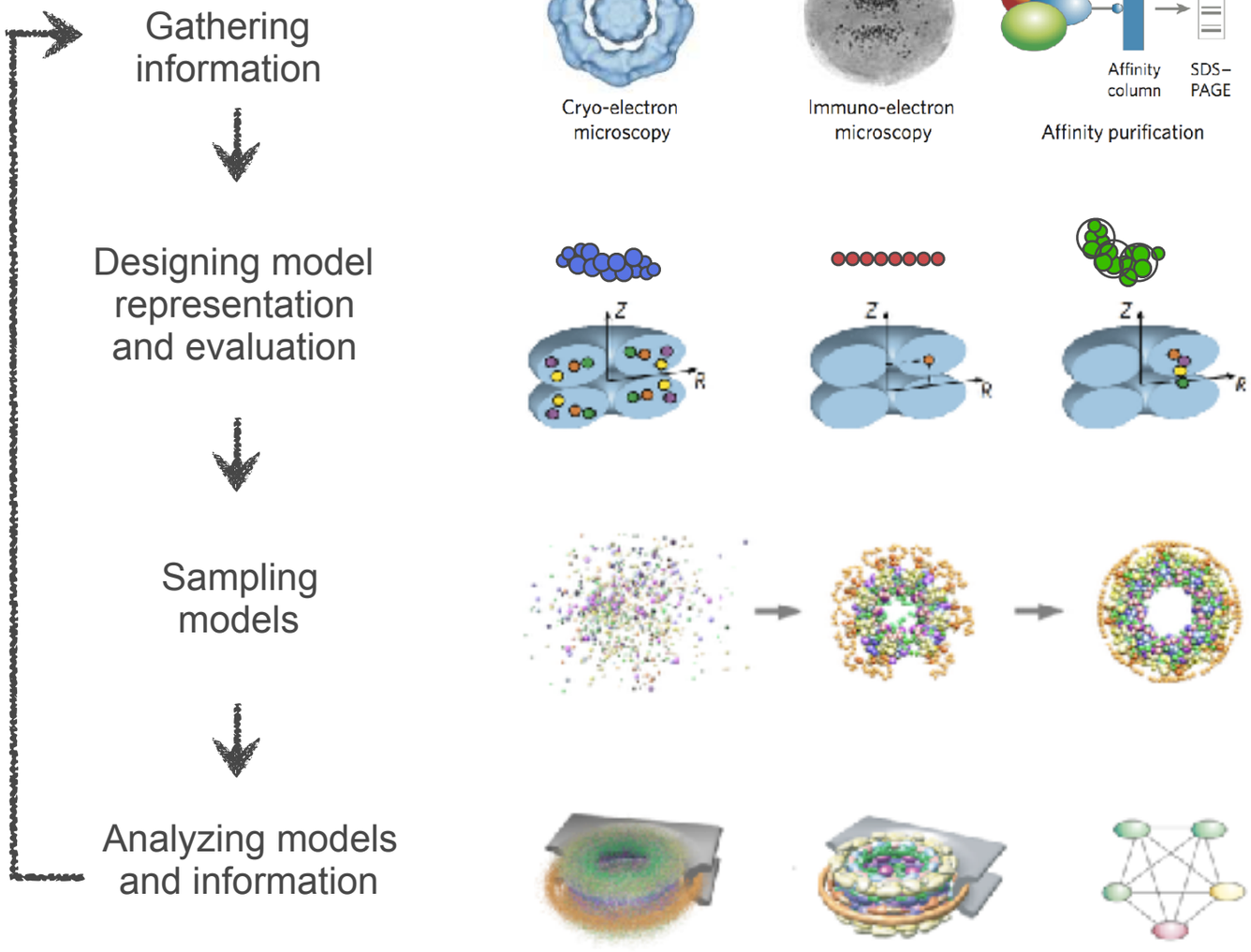


Reproducibility/Deposition



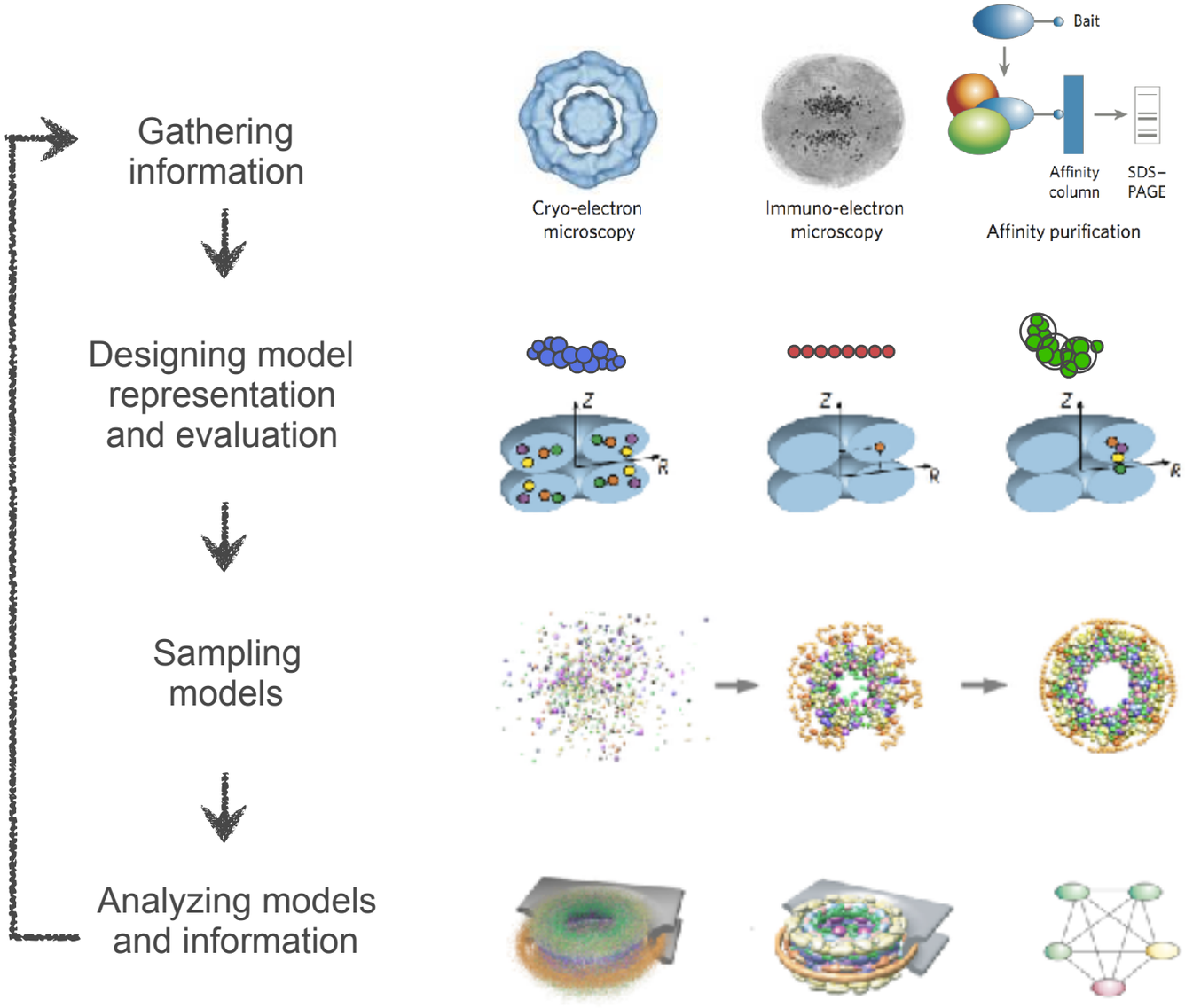
Reproducibility/Deposition

- To allow others to improve upon any integrative model, it must be *reproducible*



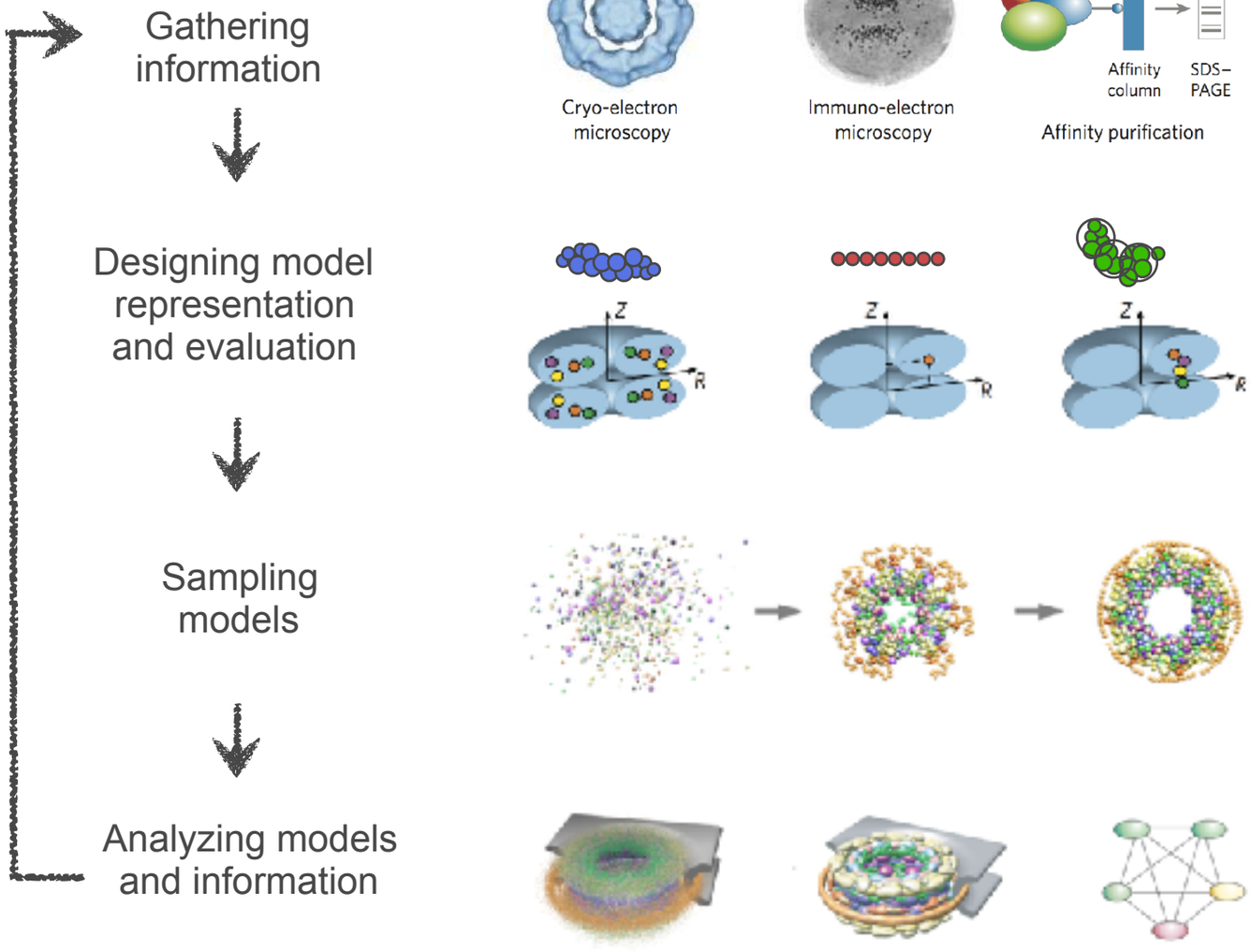
Reproducibility/Deposition

- To allow others to improve upon any integrative model, it must be *reproducible*
- This can't happen unless the data are *deposited* somewhere in a public location



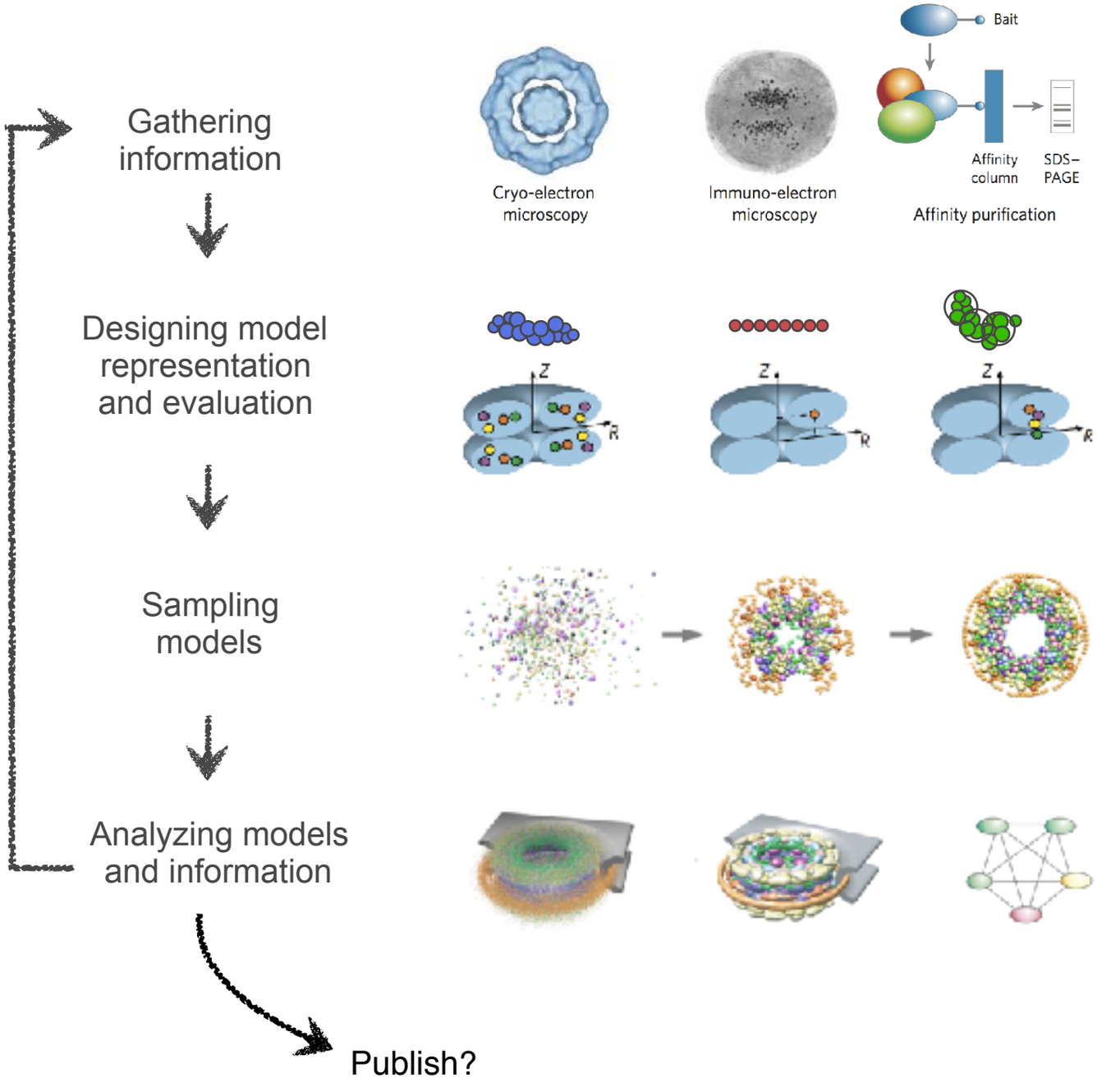
Reproducibility/Deposition

- To allow others to improve upon any integrative model, it must be *reproducible*
- This can't happen unless the data are *deposited* somewhere in a public location
- "Data" include inputs, the protocol, and the outputs, not just a final set of coordinates



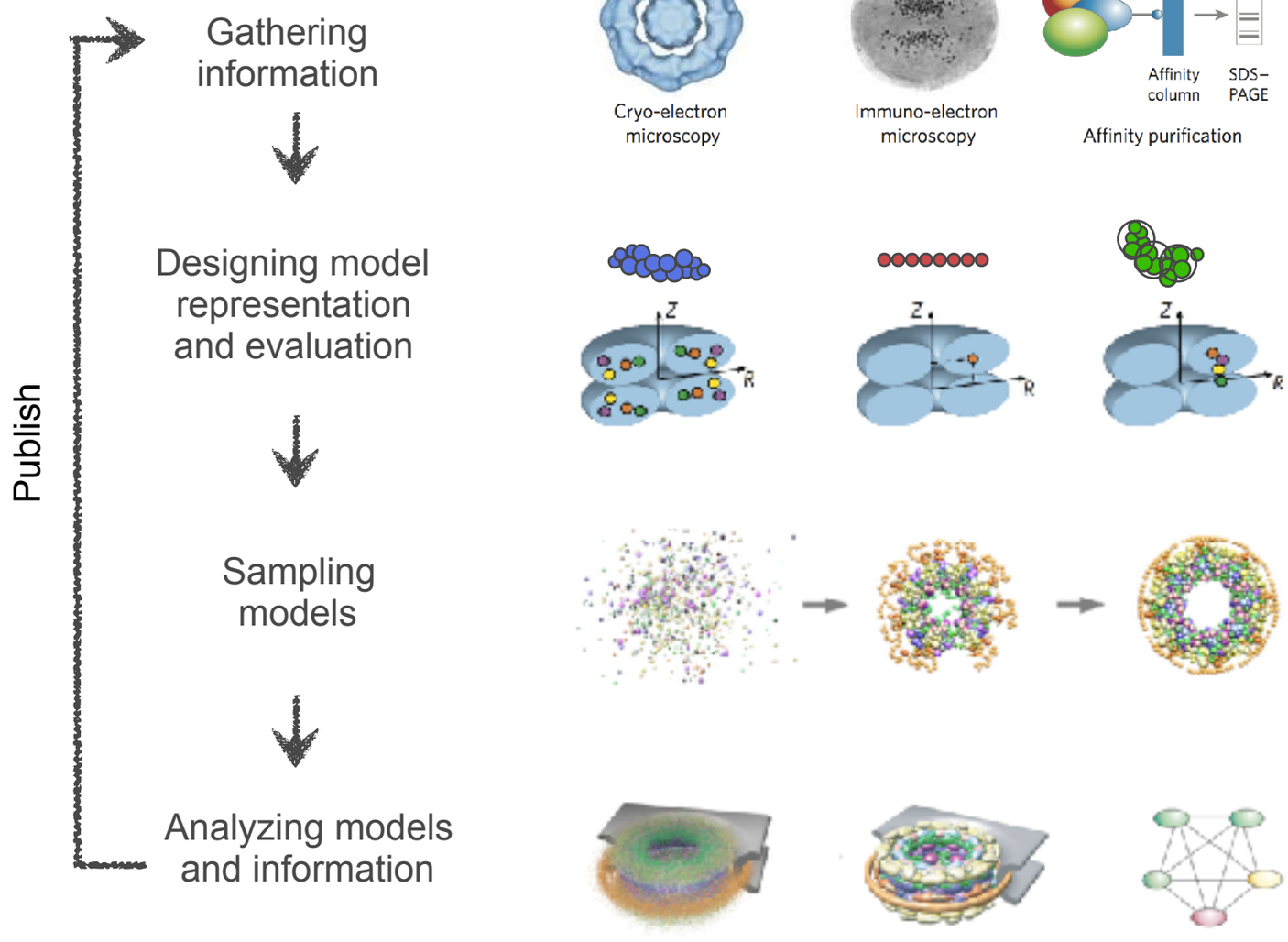
Reproducibility/Deposition

- To allow others to improve upon any integrative model, it must be *reproducible*
- This can't happen unless the data are *deposited* somewhere in a public location
- "Data" include inputs, the protocol, and the outputs, not just a final set of coordinates



Reproducibility/Deposition

- To allow others to improve upon any integrative model, it must be *reproducible*
- This can't happen unless the data are *deposited* somewhere in a public location
- "Data" include inputs, the protocol, and the outputs, not just a final set of coordinates



Quantitative reproduction

Quantitative reproduction

- GitHub repository containing all files used in the Nup84 modeling (and their history): https://github.com/integrativemodeling/nup84/

Quantitative reproduction

- GitHub repository containing all files used in the Nup84 modeling (and their history): https://github.com/integrativemodeling/nup84/
- README file pointing to data, and describing how to reproduce the modeling

Quantitative reproduction

- GitHub repository containing all files used in the Nup84 modeling (and their history): https://github.com/integrativemodeling/nup84/
- README file pointing to data, and describing how to reproduce the modeling
- Permissive license (CC BY-SA or LGPL)

Quantitative reproduction

- GitHub repository containing all files used in the Nup84 modeling (and their history): https://github.com/integrativemodeling/nup84/
- README file pointing to data, and describing how to reproduce the modeling
- Permissive license (CC BY-SA or LGPL)
- Permanently archived via a DOI (a snapshot of the GitHub repository; should outlive any single lab or company, including GitHub itself): 10.5281/zenodo.46266

Quantitative reproduction

- GitHub repository containing all files used in the Nup84 modeling (and their history): https://github.com/integrativemodeling/nup84/
- README file pointing to data, and describing how to reproduce the modeling
- Permissive license (CC BY-SA or LGPL)
- Permanently archived via a DOI (a snapshot of the GitHub repository; should outlive any single lab or company, including GitHub itself): 10.5281/zenodo.46266
- Periodically retested against newer software

Quantitative reproduction

- GitHub repository containing all files used in the Nup84 modeling (and their history): https://github.com/integrativemodeling/nup84/
- README file pointing to data, and describing how to reproduce the modeling
- Permissive license (CC BY-SA or LGPL)
- Permanently archived via a DOI (a snapshot of the GitHub repository; should outlive any single lab or company, including GitHub itself): 10.5281/zenodo.46266
- Periodically retested against newer software
- A number of such systems collated at https://integrativemodeling.org/systems/

Deposition

- Coarse-grained models using non-X-ray data are not supported by PDB
- However, PDB's next-generation file format, mmCIF, *does* support these structures (via an Integrative/Hybrid Modeling extension dictionary)
- Nup84 and some other systems are already deposited as mmCIF files at PDB-dev: https://pdb-dev.rcsb.rutgers.edu/
- Designed to allow "qualitative reproduction" (e.g. use similar input data but a modified protocol)

Data stored in mmCIF

- Input data (e.g. cross-links)
- Our interpretation of the data (e.g. ambiguity, segmentation)
- Output models (cluster representatives)
  - Atomic information
  - Coarse-grained coordinates (beads)
- Non-Cartesian data (e.g. Bayesian noise parameters)
- Ensemble info (number & size of clusters)
- Other metadata (e.g. publications, software used; as for regular atomic PDB structures)
- Validation (how well do the models fit the data)

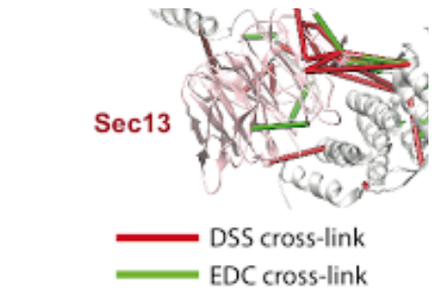
Link out to other data

- mmCIF links out to data where available via standard database IDs or DOIs

mmCIF
https://pdb-dev.rcsb.rutgers.edu/

Link out to other data

- mmCIF links out to data where available via standard database IDs or DOIs

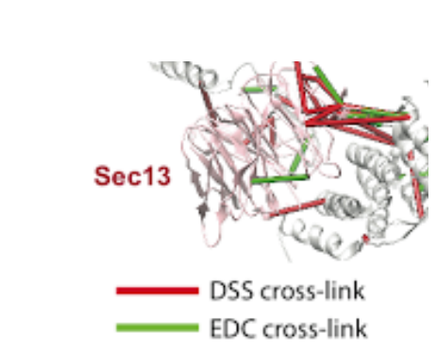


Cross-links,
DOI 10.5281/zenodo.46266

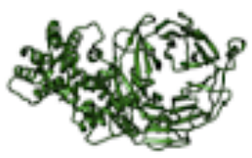
mmCIF
https://pdb-dev.rcsb.rutgers.edu/

Link out to other data

- mmCIF links out to data where available via standard database IDs or DOIs



Cross-links,
DOI 10.5281/zenodo.46266

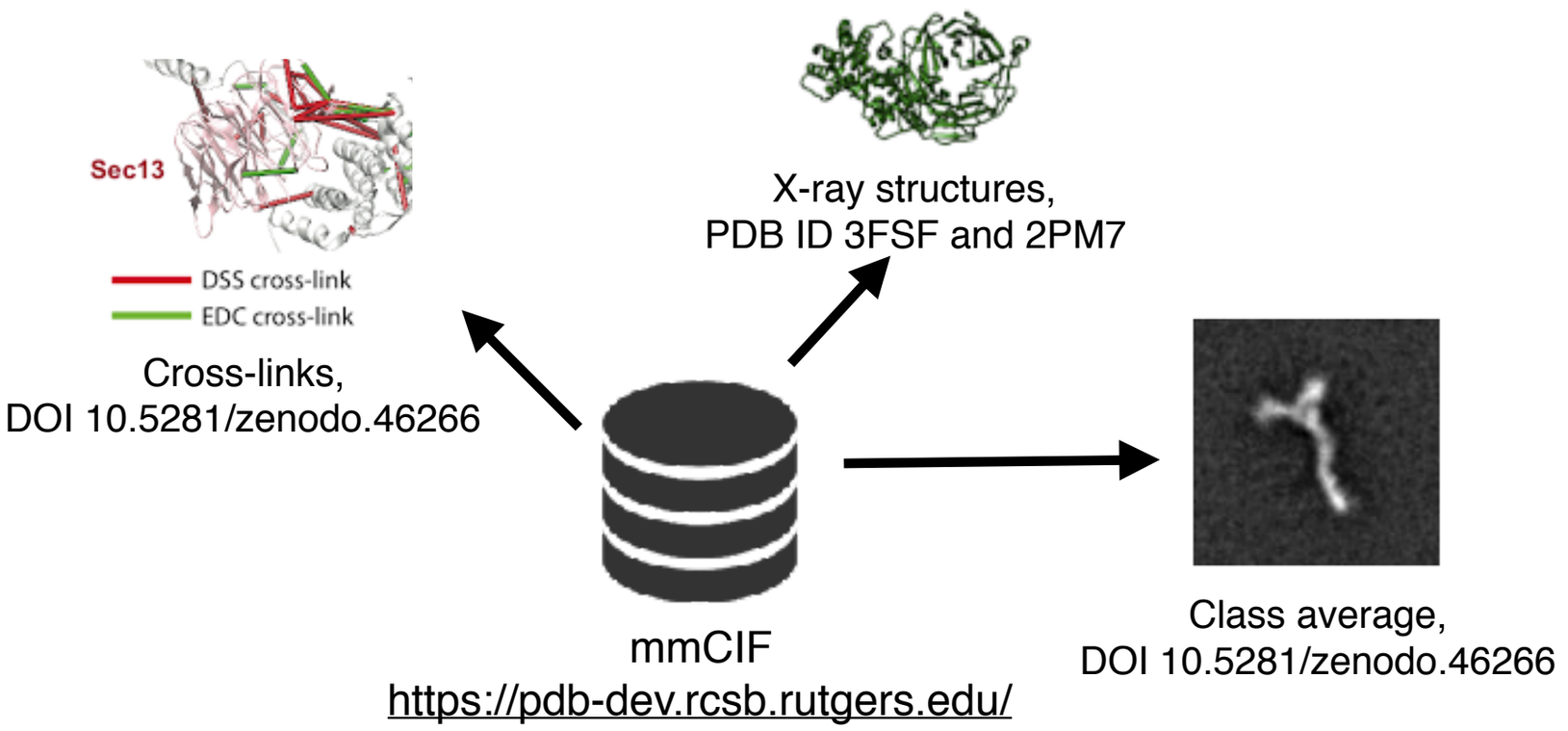


X-ray structures,
PDB ID 3FSF and 2PM7

mmCIF
https://pdb-dev.rcsb.rutgers.edu/

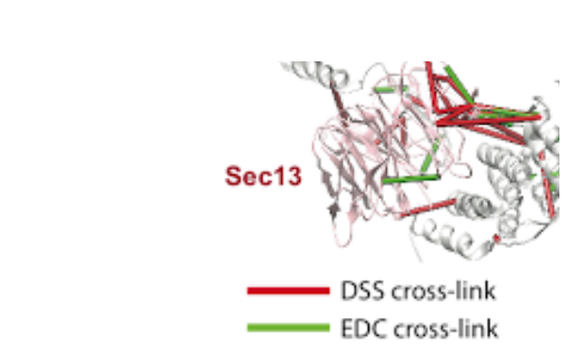
Link out to other data

- mmCIF links out to data where available via standard database IDs or DOIs

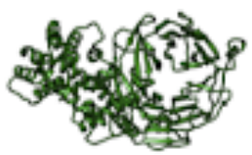


Link out to other data

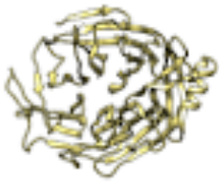
- mmCIF links out to data where available via standard database IDs or DOIs



Cross-links,
DOI 10.5281/zenodo.46266



X-ray structures,
PDB ID 3FSF and 2PM7



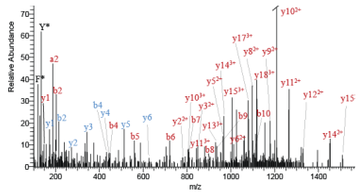
Comparative models,
DOI 10.5281/zenodo.46266
(or PMP ID)

mmCIF
https://pdb-dev.rcsb.rutgers.edu/

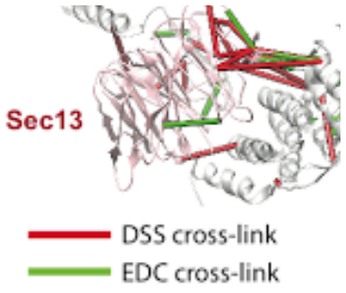
Class average,
DOI 10.5281/zenodo.46266

Link out to other data

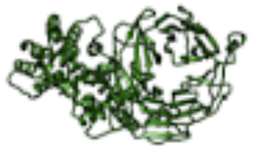
- mmCIF links out to data where available via standard database IDs or DOIs



Raw spectra, peaklists,
MASSIVE ID



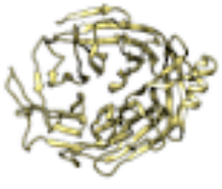
Cross-links,
DOI 10.5281/zenodo.46266



X-ray structures,
PDB ID 3FSF and 2PM7

mmCIF
https://pdb-dev.rcsb.rutgers.edu/

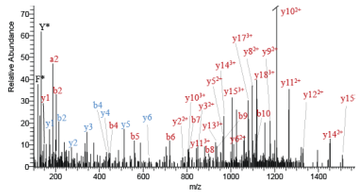
Class average,
DOI 10.5281/zenodo.46266



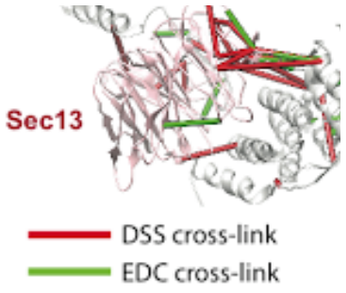
Comparative models,
DOI 10.5281/zenodo.46266
(or PMP ID)

Link out to other data

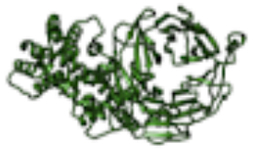
- mmCIF links out to data where available via standard database IDs or DOIs



Raw spectra, peaklists,
MASSIVE ID



Cross-links,
DOI 10.5281/zenodo.46266

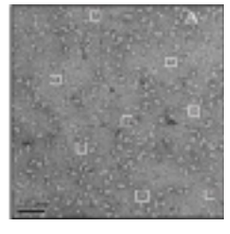


X-ray structures,
PDB ID 3FSF and 2PM7

mmCIF
https://pdb-dev.rcsb.rutgers.edu/

Comparative models,
DOI 10.5281/zenodo.46266
(or PMP ID)

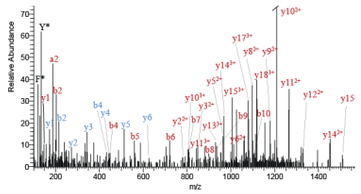
Class average,
DOI 10.5281/zenodo.46266



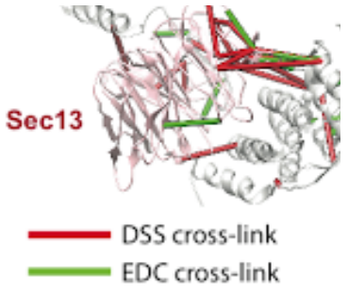
Raw EM micrographs,
DOI 10.5281/zenodo.58025
(or EMPIAR ID)

Link out to other data

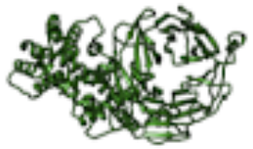
- mmCIF links out to data where available via standard database IDs or DOIs



Raw spectra, peaklists,
MASSIVE ID

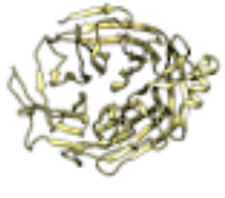


Cross-links,
DOI 10.5281/zenodo.46266



X-ray structures,
PDB ID 3FSF and 2PM7

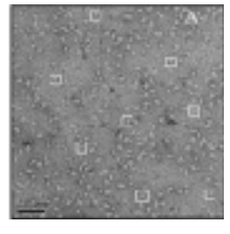
Comparative models,
DOI 10.5281/zenodo.46266
(or PMP ID)



Template structures,
PDB IDs

mmCIF
https://pdb-dev.rcsb.rutgers.edu/

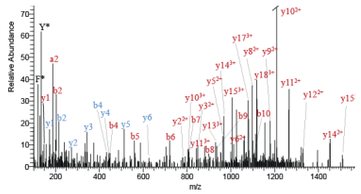
Class average,
DOI 10.5281/zenodo.46266



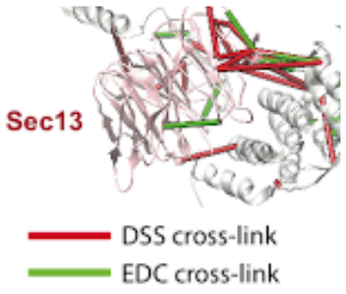
Raw EM micrographs,
DOI 10.5281/zenodo.58025
(or EMPIAR ID)

Link out to other data

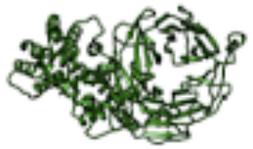
- mmCIF links out to data where available via standard database IDs or DOIs



Raw spectra, peaklists,
MASSIVE ID



Cross-links,
DOI 10.5281/zenodo.46266



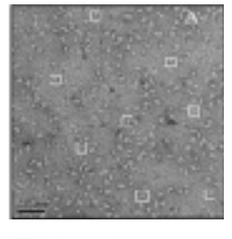
X-ray structures,
PDB ID 3FSF and 2PM7

mmCIF
https://pdb-dev.rcsb.rutgers.edu/

Comparative models,
DOI 10.5281/zenodo.46266
(or PMP ID)

Template structures,
PDB IDs

Class average,
DOI 10.5281/zenodo.46266

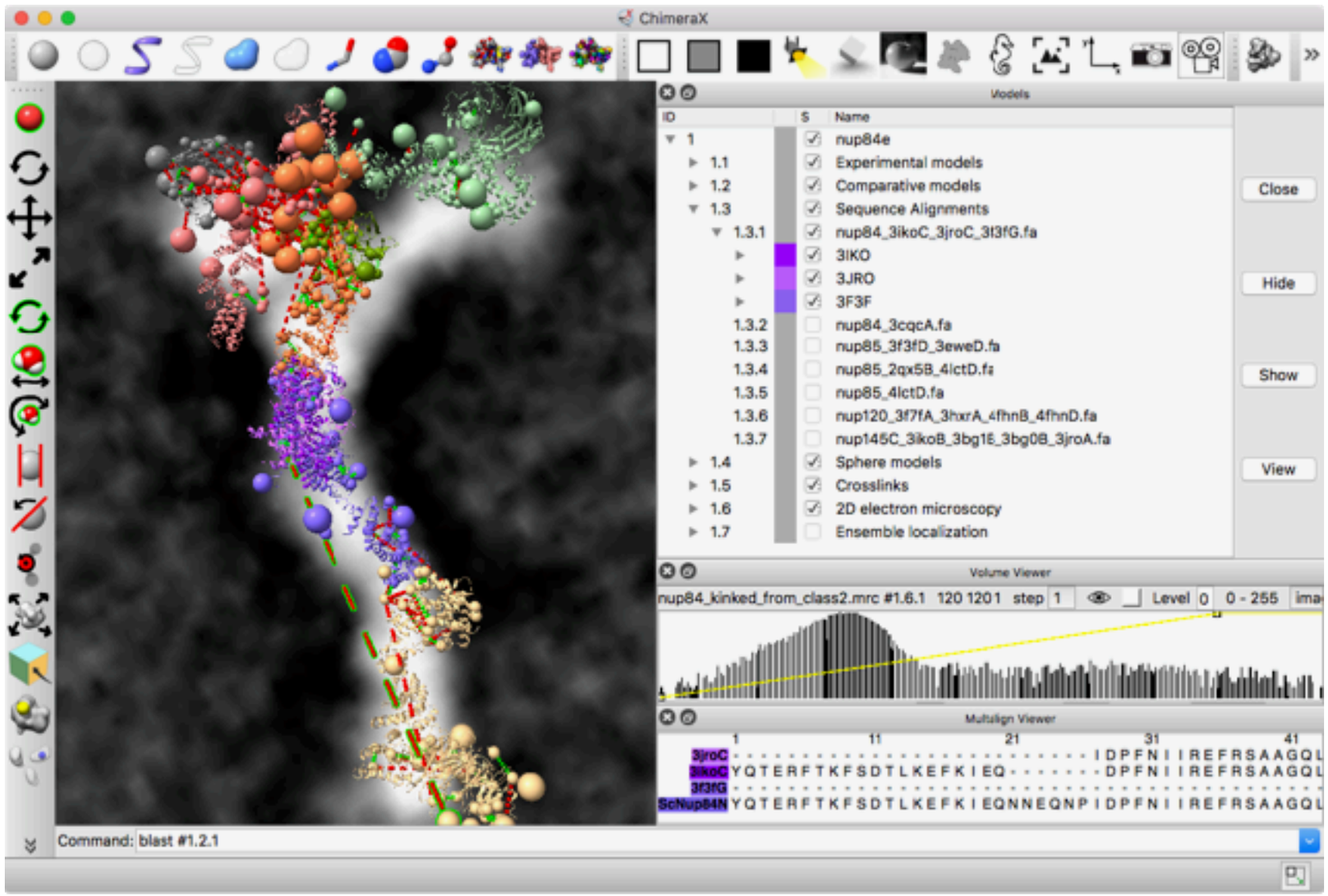


Raw EM micrographs,
DOI 10.5281/zenodo.58025
(or EMPIAR ID)

Protocol for quantitative reproduction,
DOI pointing to GitHub repository

Visualization

- Any viewer that supports mmCIF can add support for the Integrative/Hybrid Modeling dictionary
- e.g. ChimeraX has basic support



Useful resources

- IMP, https://integrativemodeling.org/
- Nup84, https://salilab.org/nup84/
- PDB-dev, https://pdb-dev.rcsb.rutgers.edu/
- Applications of IMP to varied biological systems, https://integrativemodeling.org/systems/