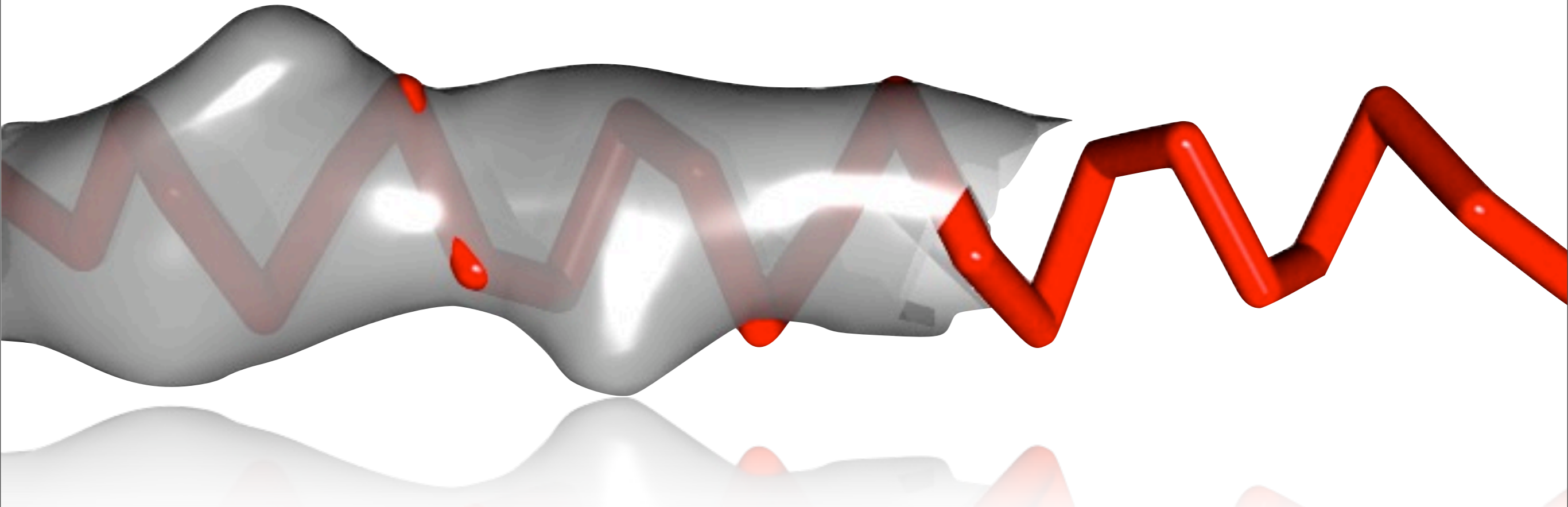# De Novo Modeling with Cryo-EM Density Maps

Matthew Baker

National Center for Macromolecular Imaging

Baylor College of Medicine

ELECTRON CRYOMICROSCOPY

- Image single molecules >200kDa

- Small amount of sample

- Sample can be frozen and imaged at chemically defined states

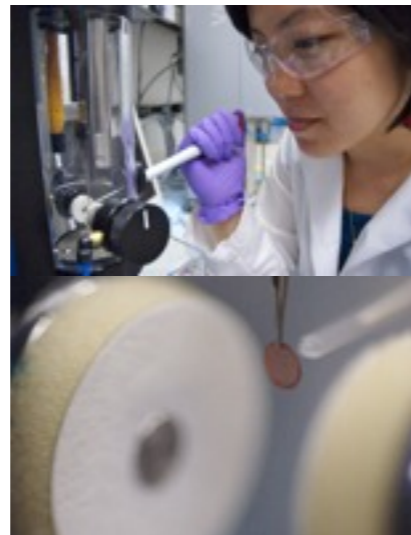- Capable of resolving biological samples to near-atomic resolutions resolutions
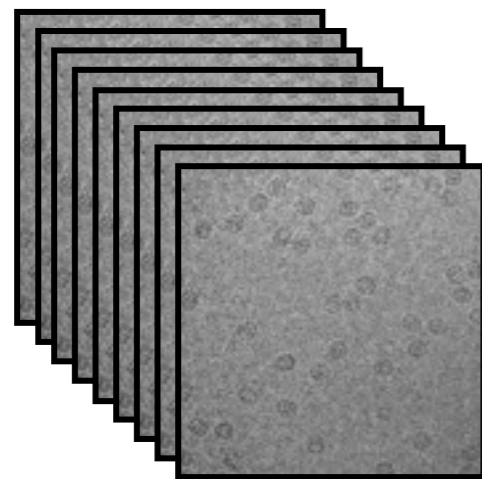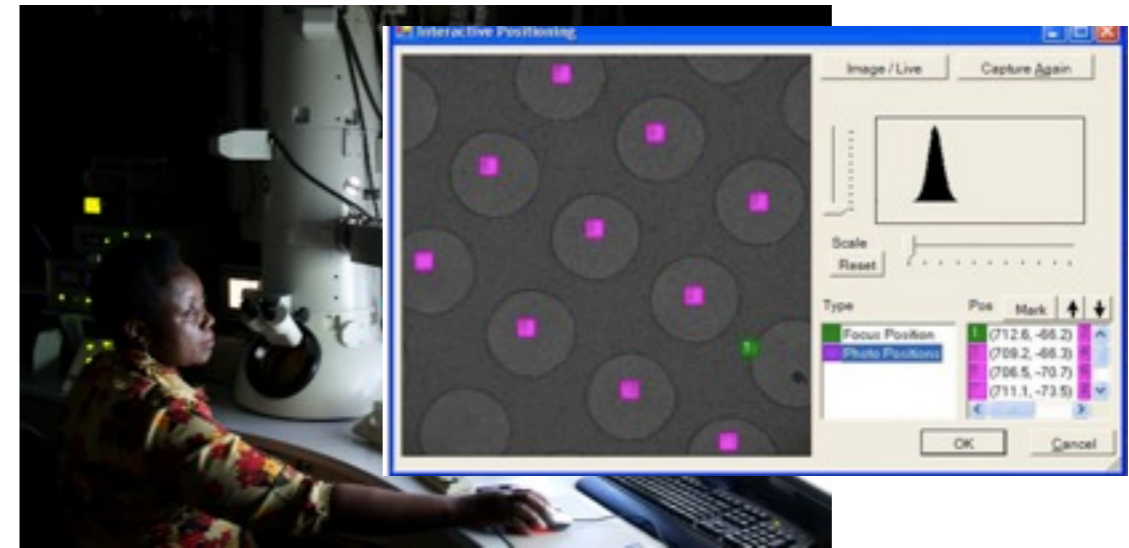
THE CRYO-EM "EXPERIMENT"
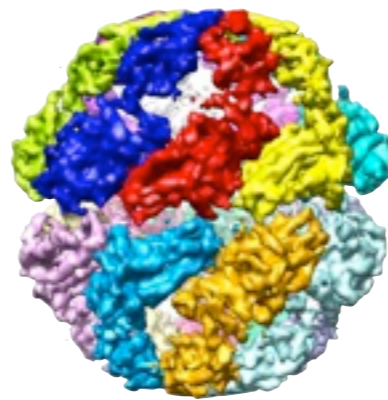
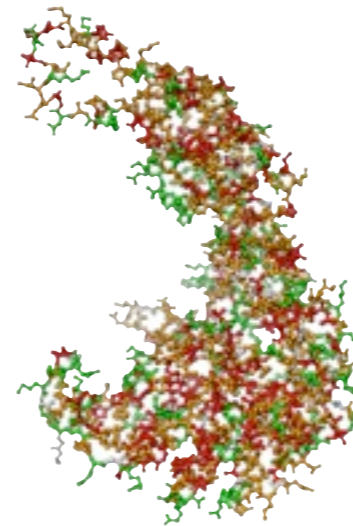Biochemical
Preparation

Cryo-EM Sample
Preparation

Data Collection

Data Processing

3D Reconstruction

Model Building & Validation

Structure Deposition

CRYO-EM DENSITY MAPS IN THE EMDB

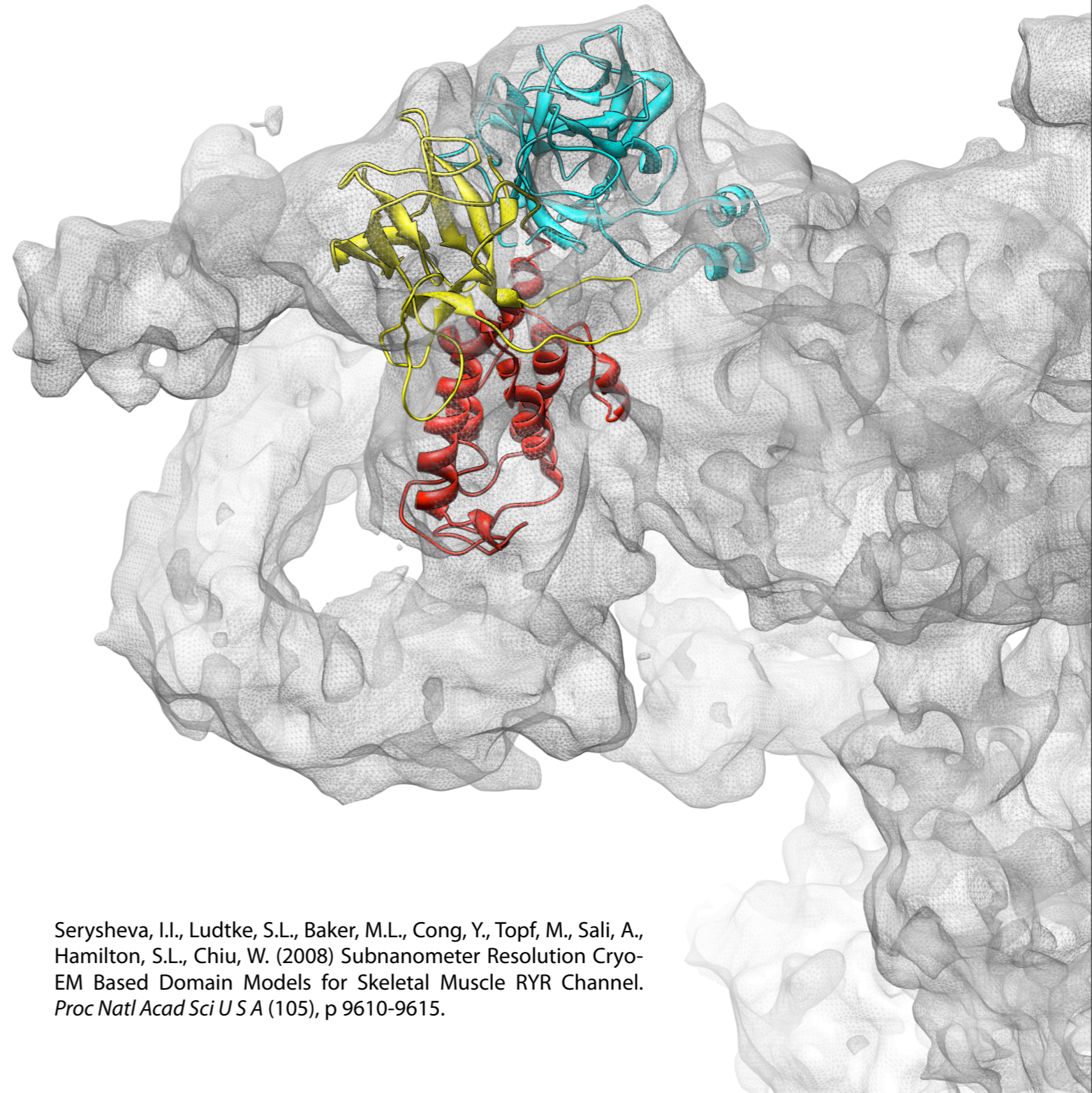Legend: >5Å | 5Å to 10Å | 10Å to 20Å | 20Å to 40Å | <40Å | undisclosed
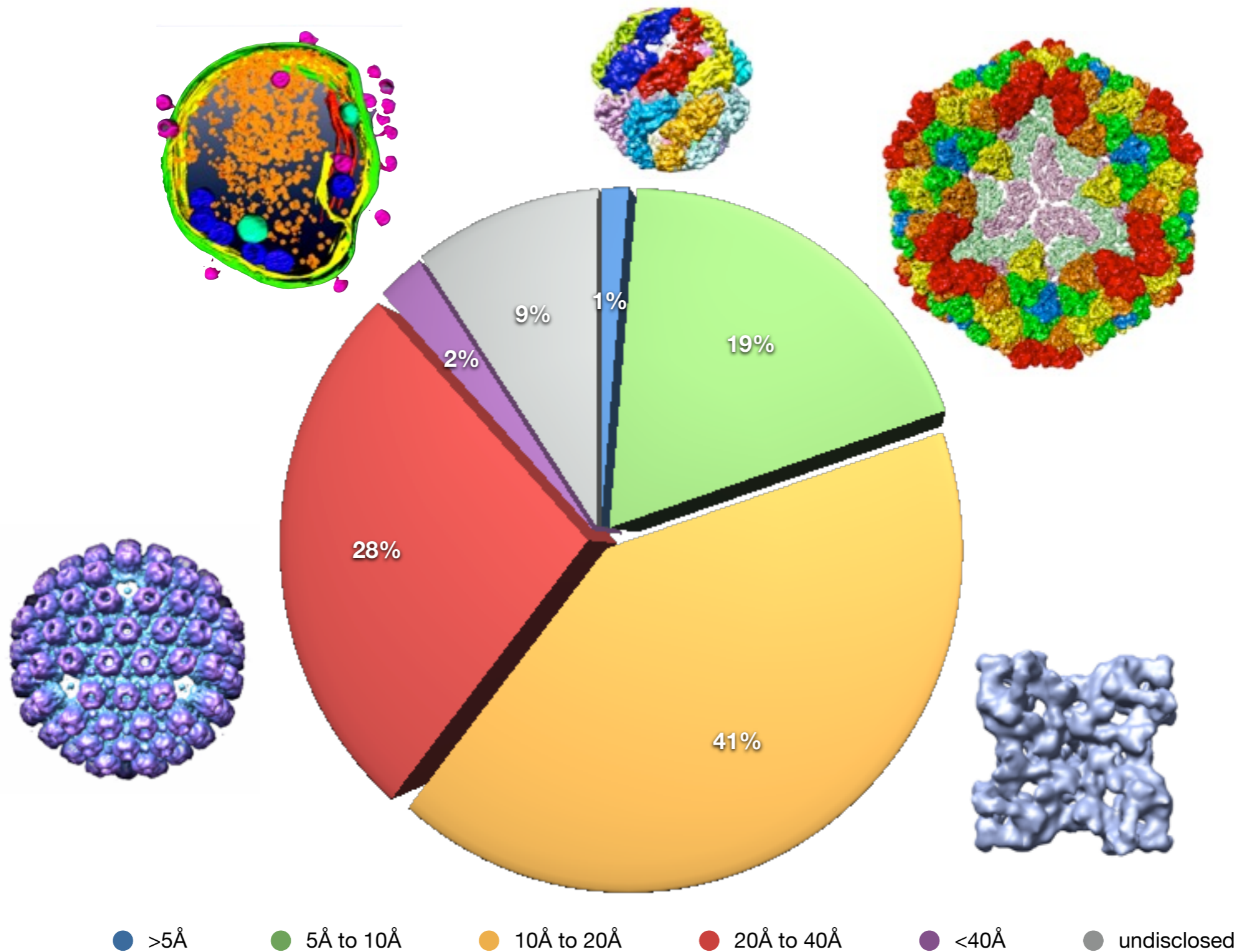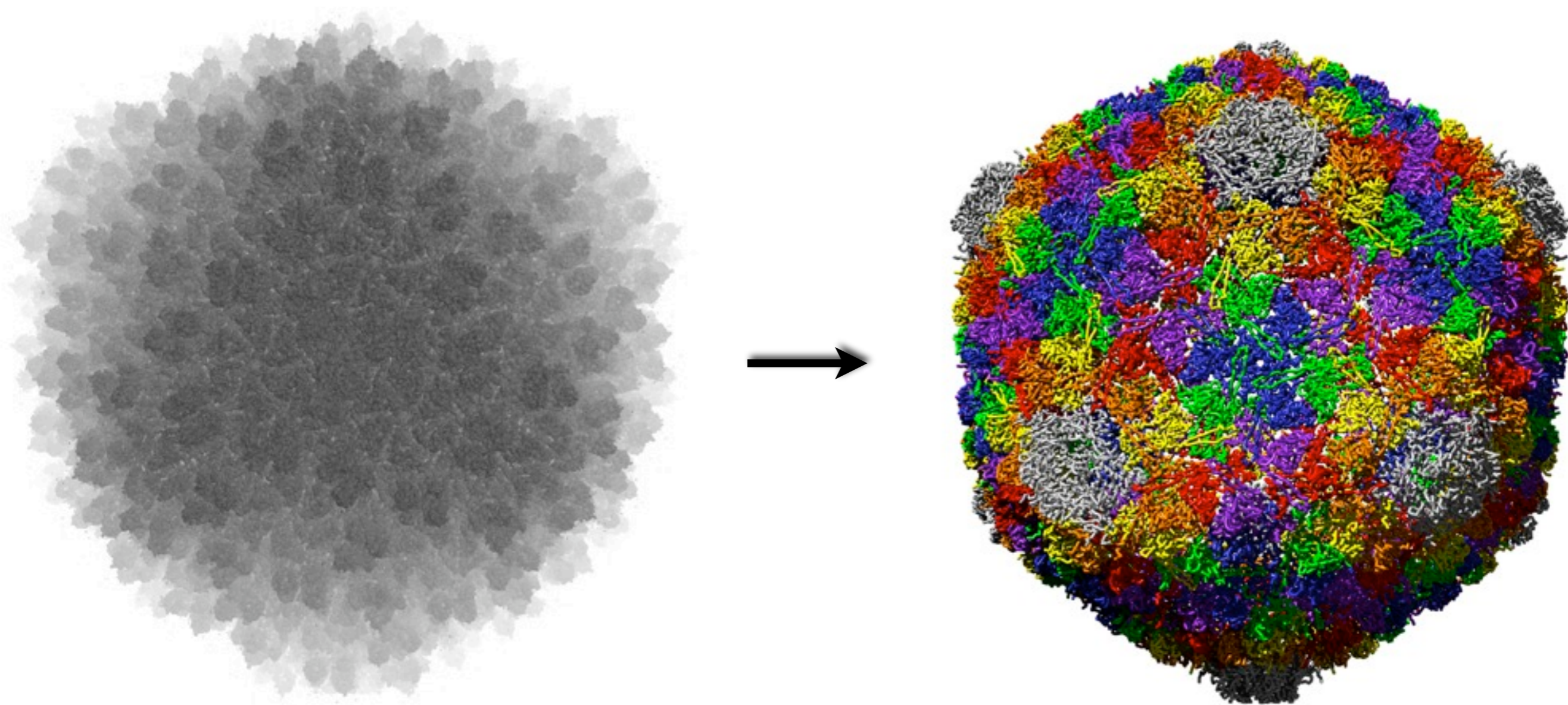
MODELING IN CRYO-EM

MODELING WITH KNOWN STRUCTURES

- Fitting atomic models

- Homology modeling

- Flexible fitting

Serysheva, I.I., Ludtke, S.L., Baker, M.L., Cong, Y., Topf, M., Sali, A., Hamilton, S.L., Chiu, W. (2008) Subnanometer Resolution Cryo-EM Based Domain Models for Skeletal Muscle RYR Channel. *Proc Natl Acad Sci U S A* (105), p 9610-9615.

CRYO-EM DENSITY AS A CONSTRAINT

- CryoEM density can discriminate amongst a gallery of models

- Evaluation of models is resolution dependent

- Low resolution provides basic shape

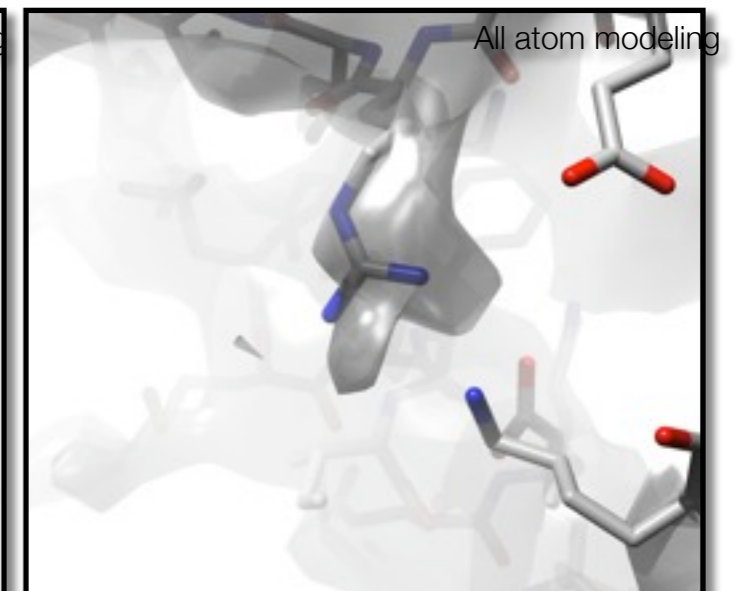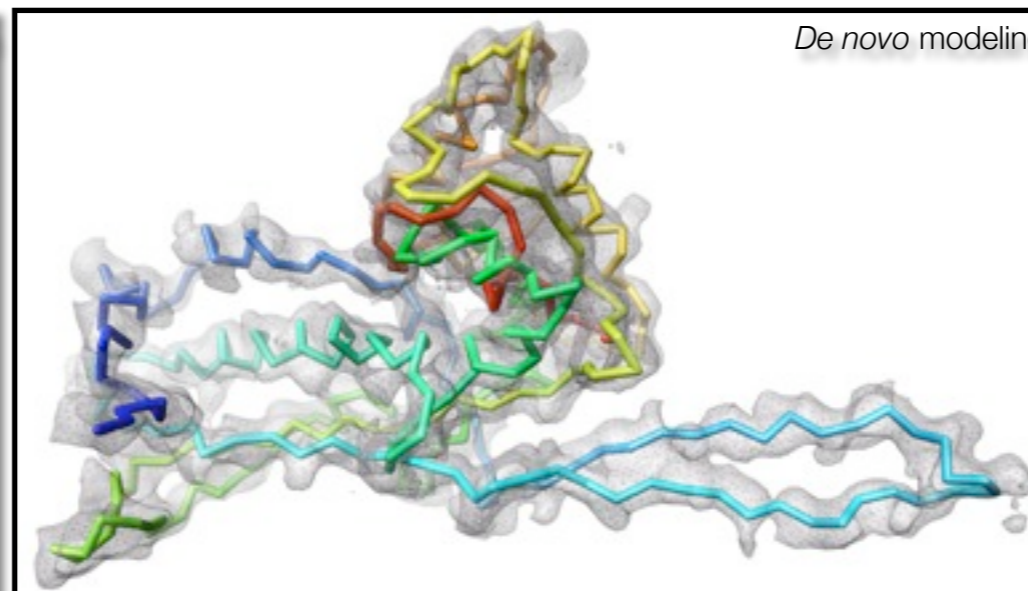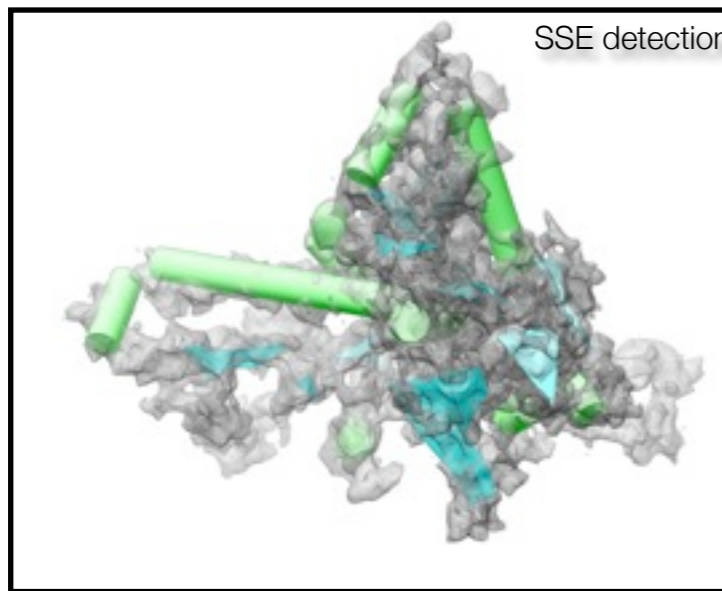- High resolution provides finer structural details (loops, SSE)
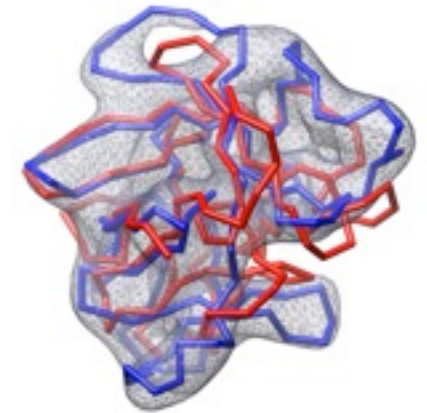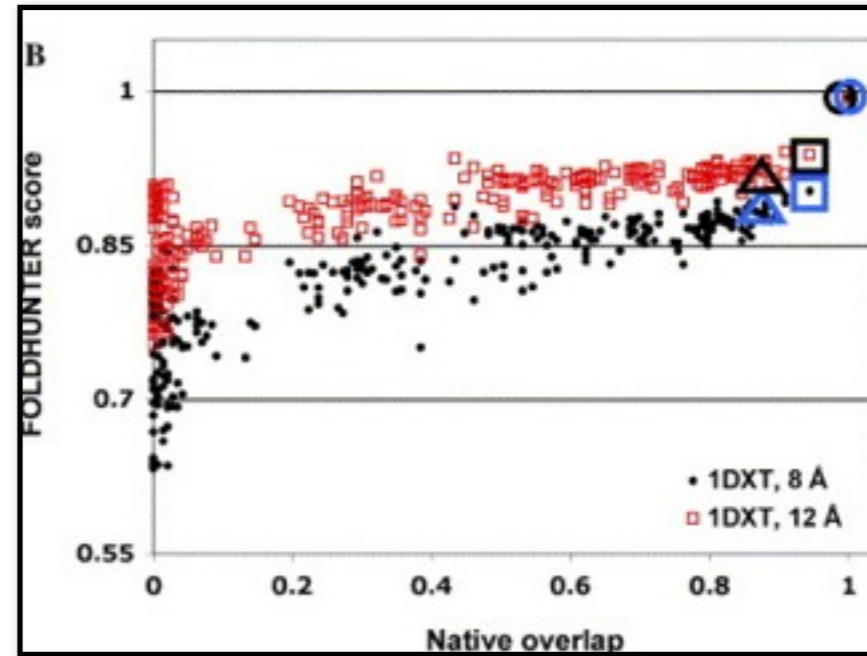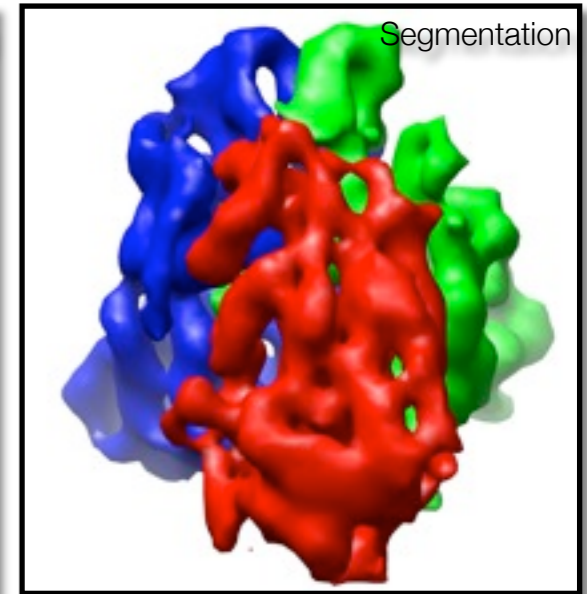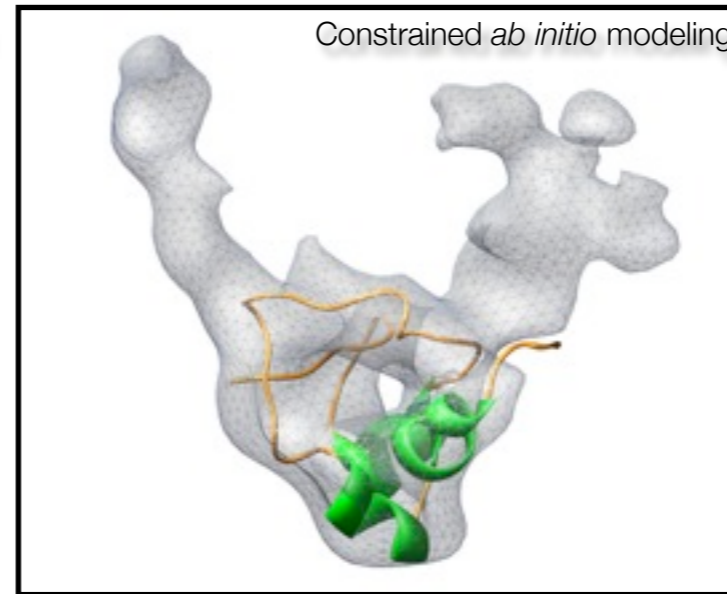

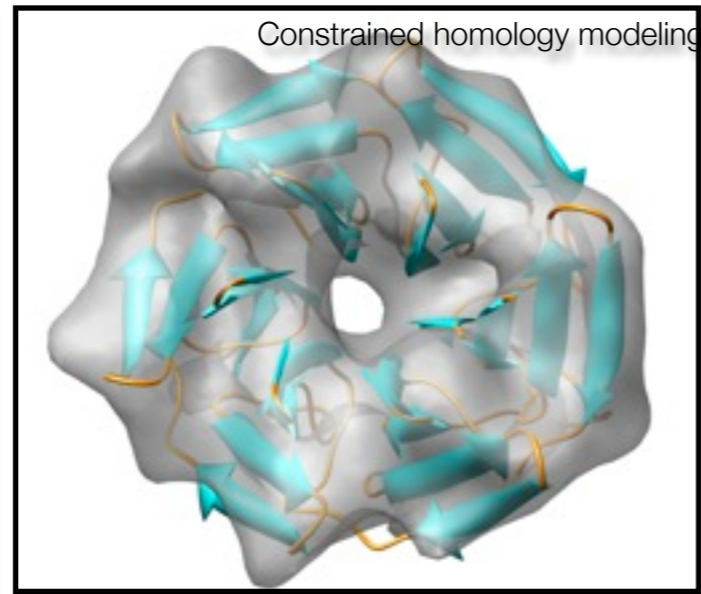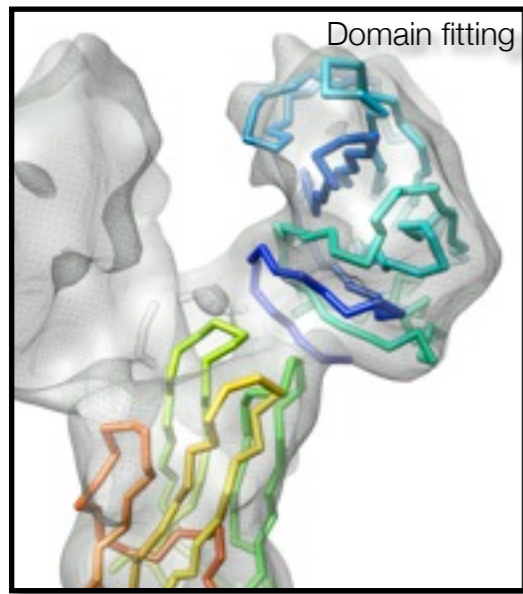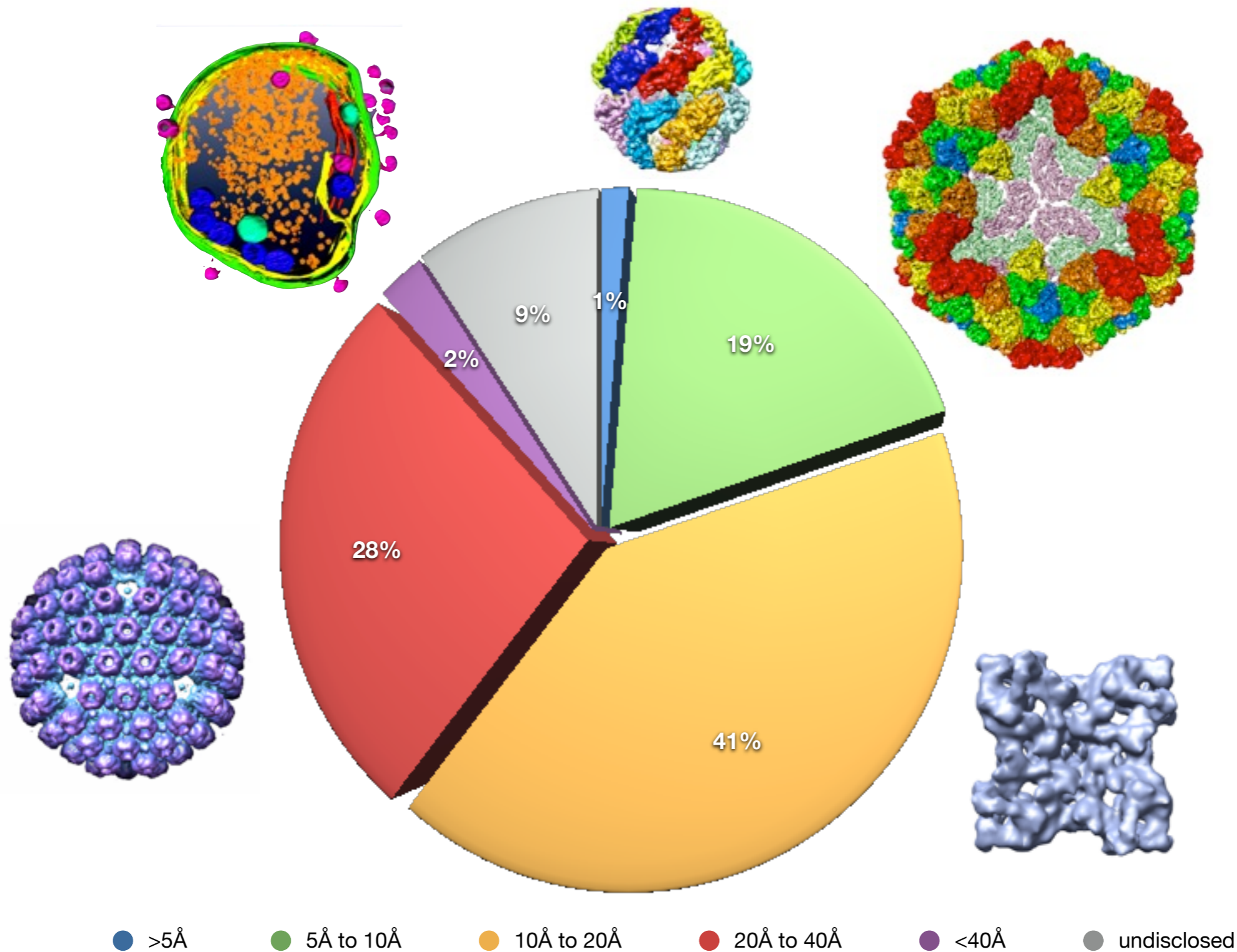Topf, M. , Baker, M.L., John, B., Chiu, W., Sali, A. (2005) JSB
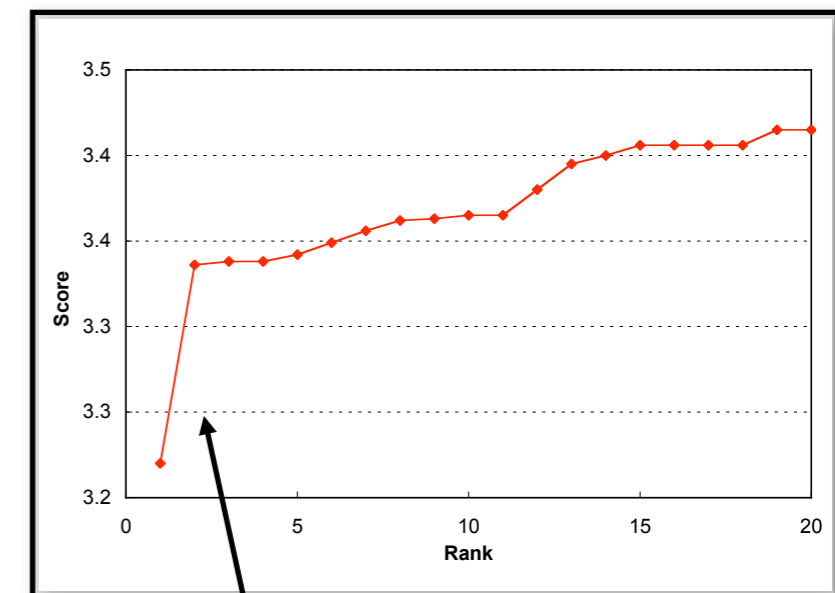
Decoy 6072

Baker, M.L., Jiang, W., Wedemeyer, W., Rixon, F., Baker, D., Chiu, W. (2006) PLoS Comp Biol
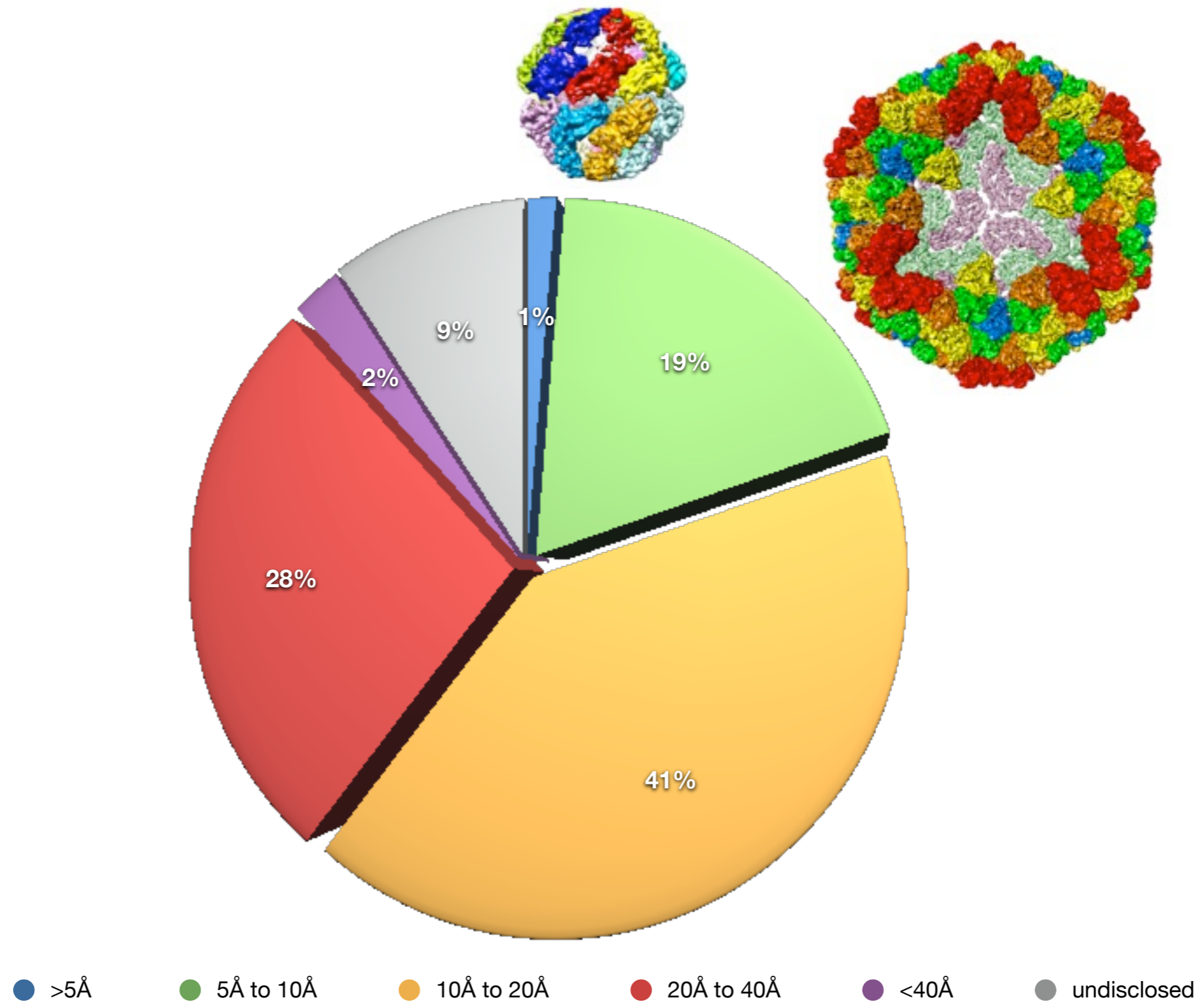
MODELING IN CRYO-EM

CRYO-EM DENSITY MAPS IN THE EMDB

Legend: >5Å | 5Å to 10Å | 10Å to 20Å | 20Å to 40Å | <40Å | undisclosed

CRYO-EM DENSITY MAPS IN THE EMDB

Legend: >5Å | 5Å to 10Å | 10Å to 20Å | 20Å to 40Å | <40Å | undisclosed

INTERMEDIATE RESOLUTION (5-10Å)

Features

- Accurate segmentation

- Secondary structure elements (SSE)

  - Cylindrical helices

  - Plane-like sheets

- Integration of sequence information

- Topology

- Flexible fitting

- Computational modeling

Limitations

- "Lumpy/noisy" density maps

- Ambiguities in helix direction

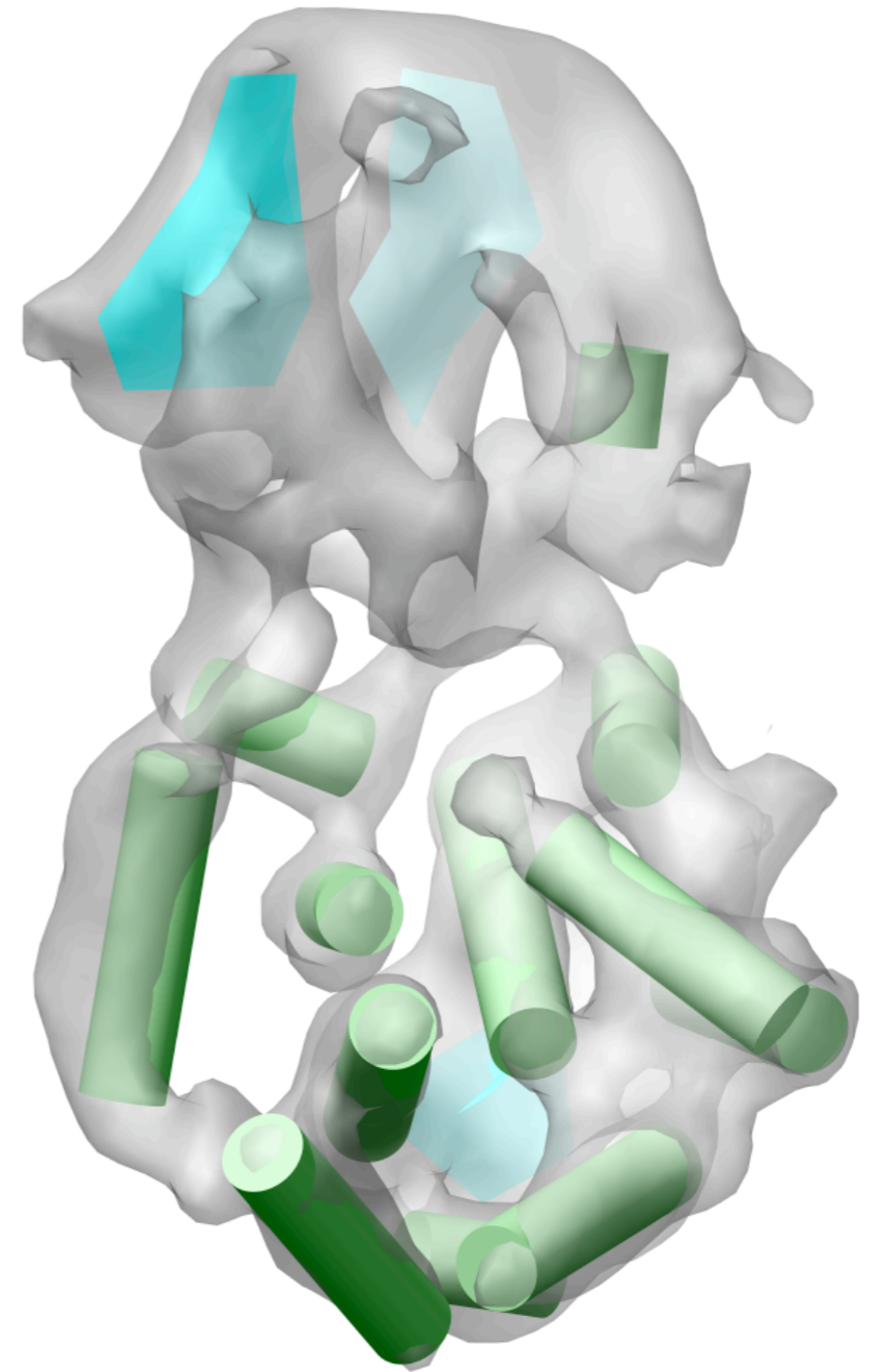- No strand separation

- Ambiguous connectivity

DETECTING SECONDARY STRUCTURE ELEMENTS

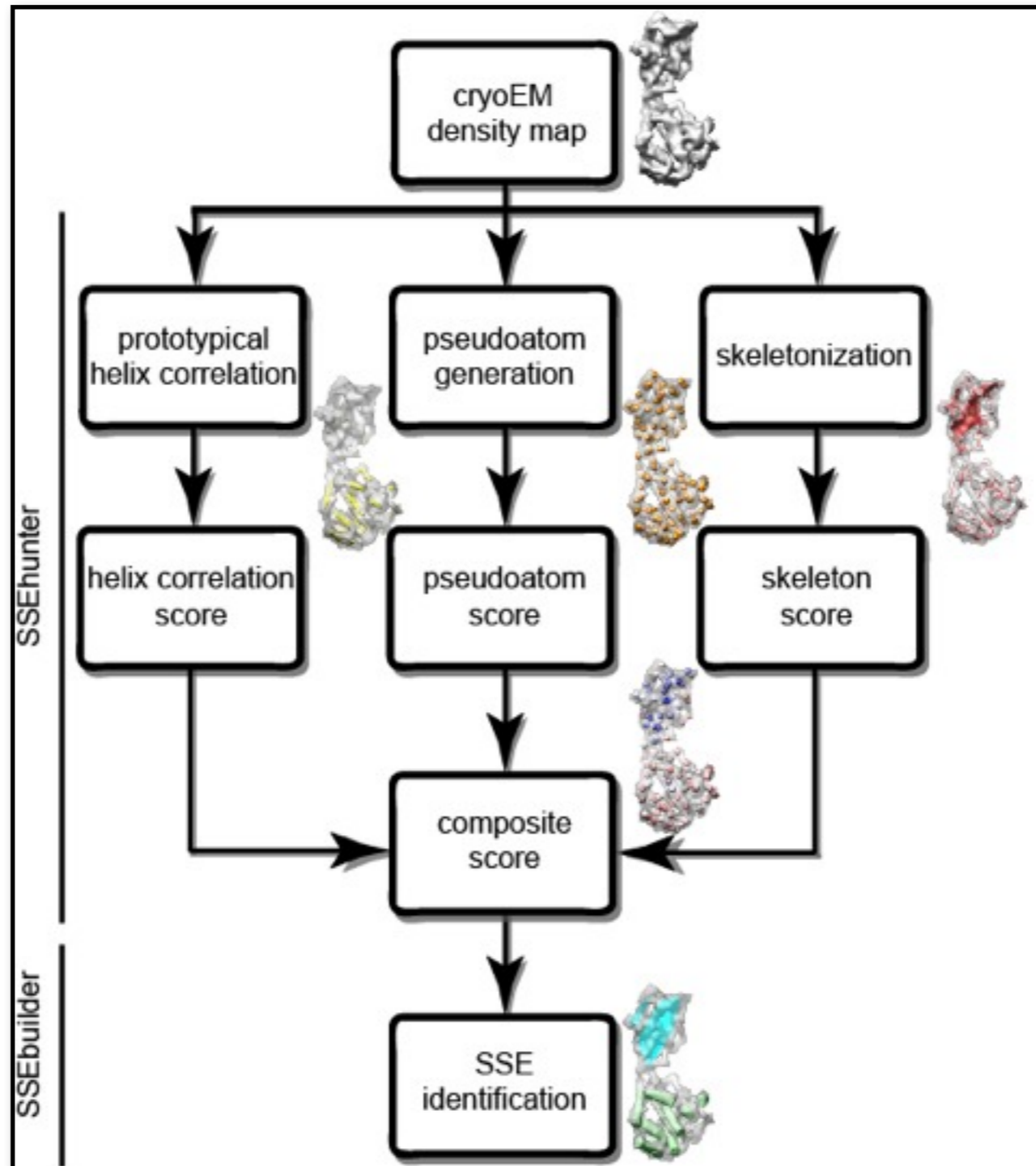SSEHunter: simultaneous alpha helix and beta sheet detection

Uses cross correlation, density skeletonization and local geometry calculations at discrete points
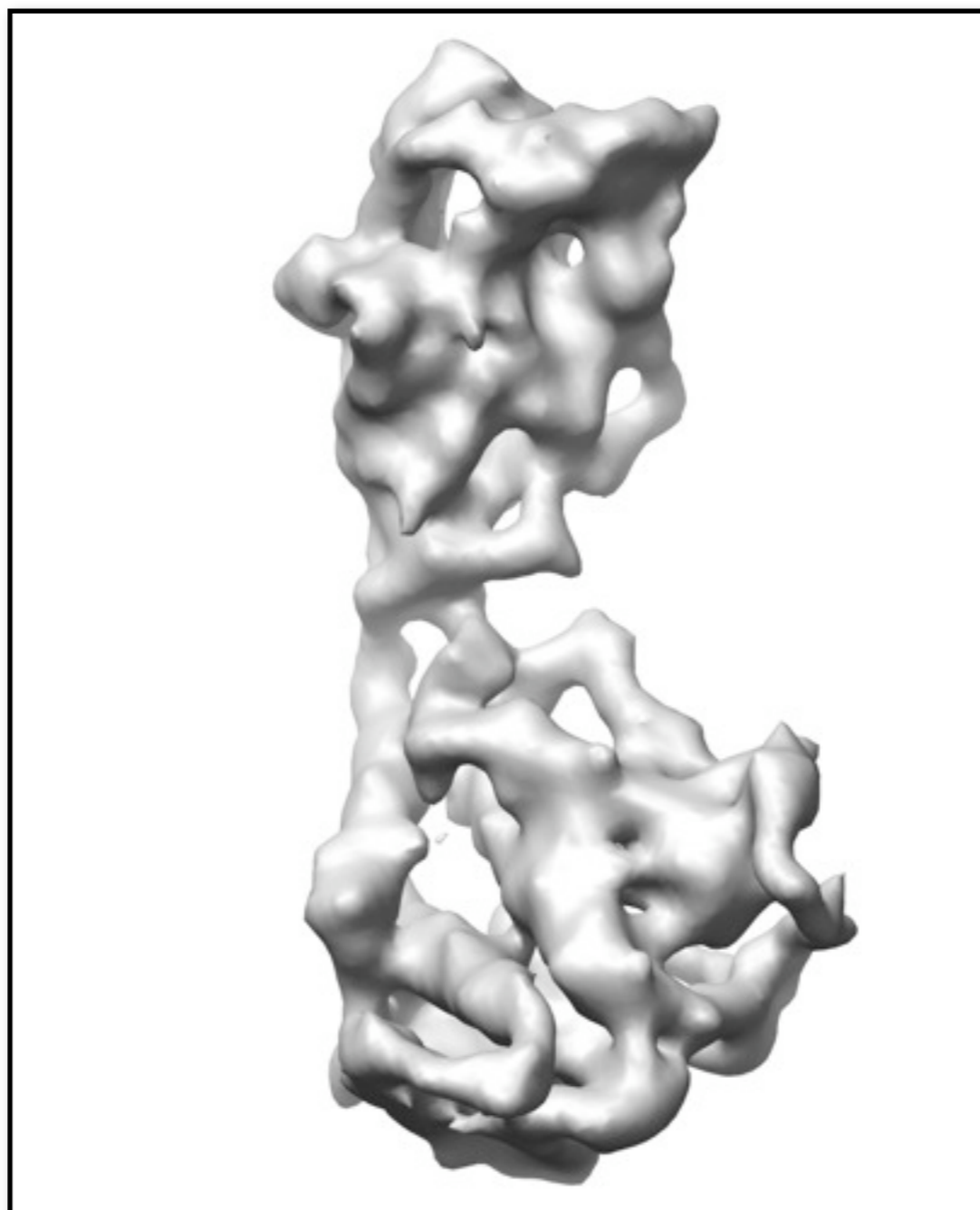
Distributed with EMAN and Gorgon

Baker, M.L., Ju, T., Chiu, W. (2007) Identification of Secondary Structure Elements in Intermediate Resolution Density Maps. Structure (15), p 7-19.
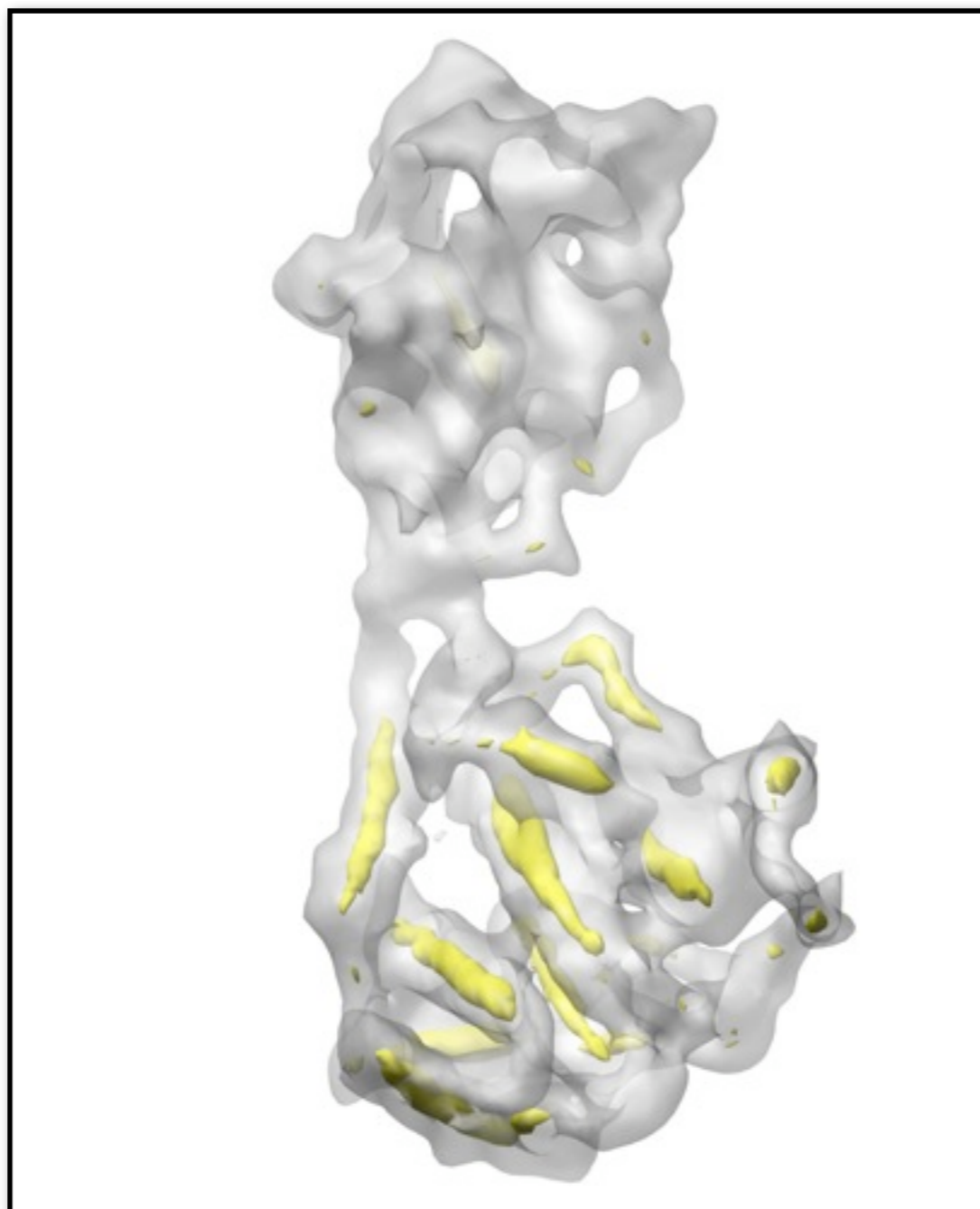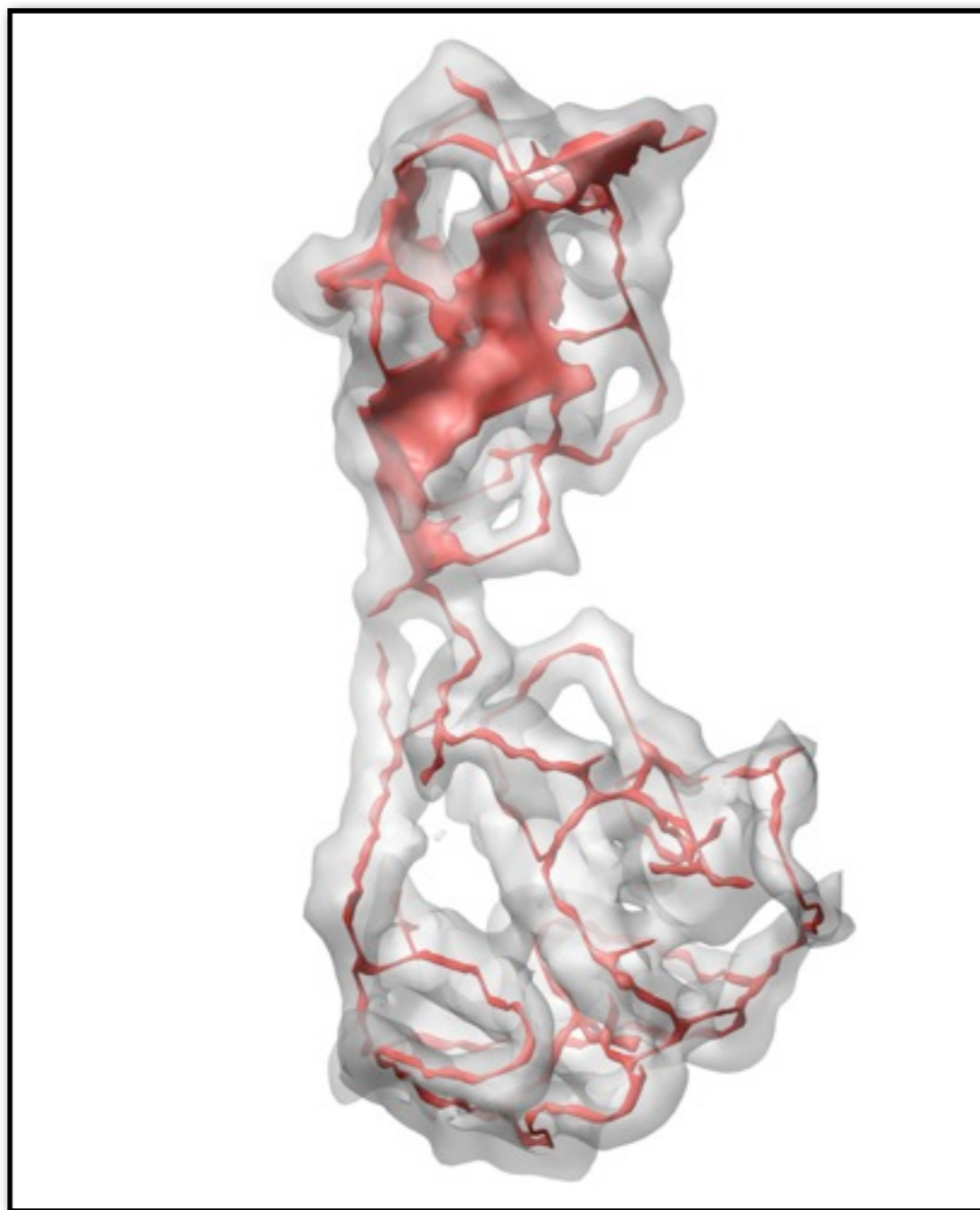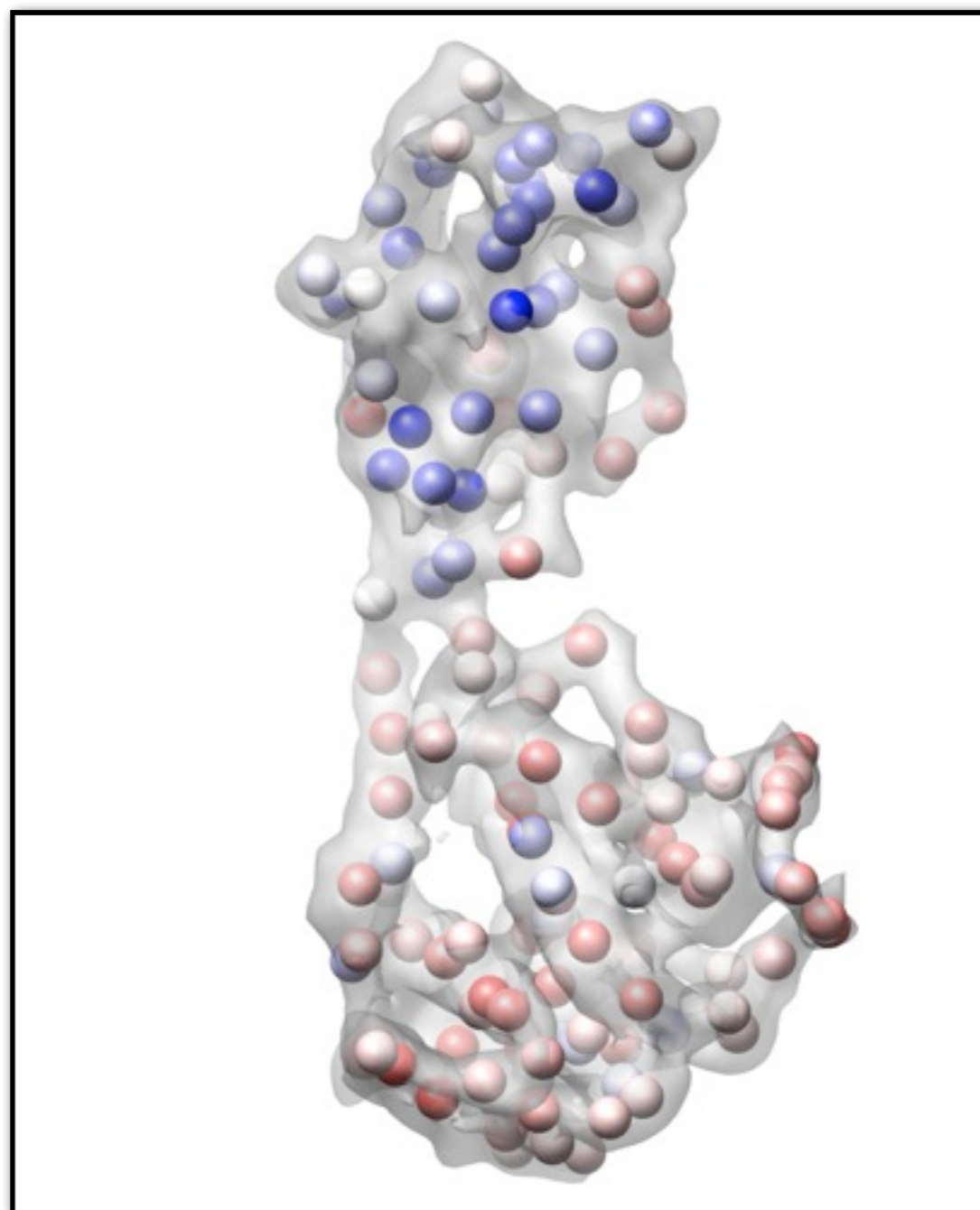
SSEHUNTER: METHODOLOGY

SSEHUNTER: METHODOLOGY
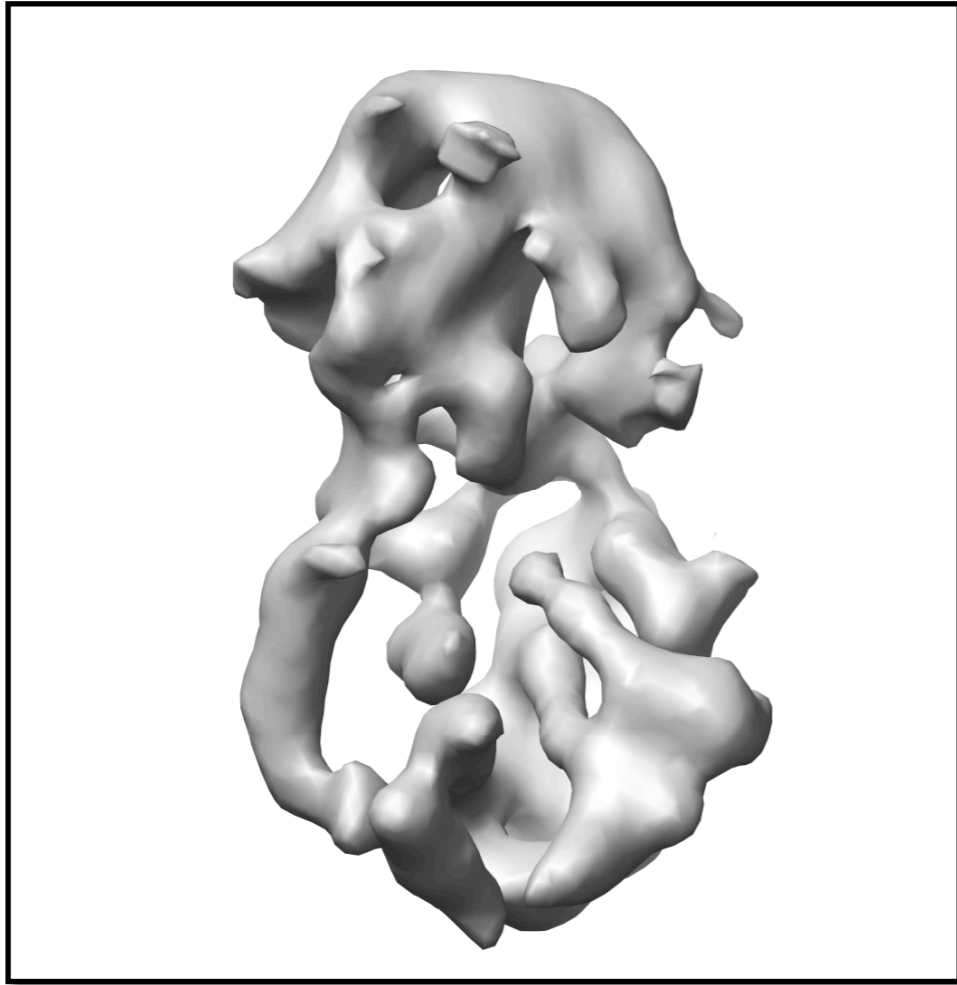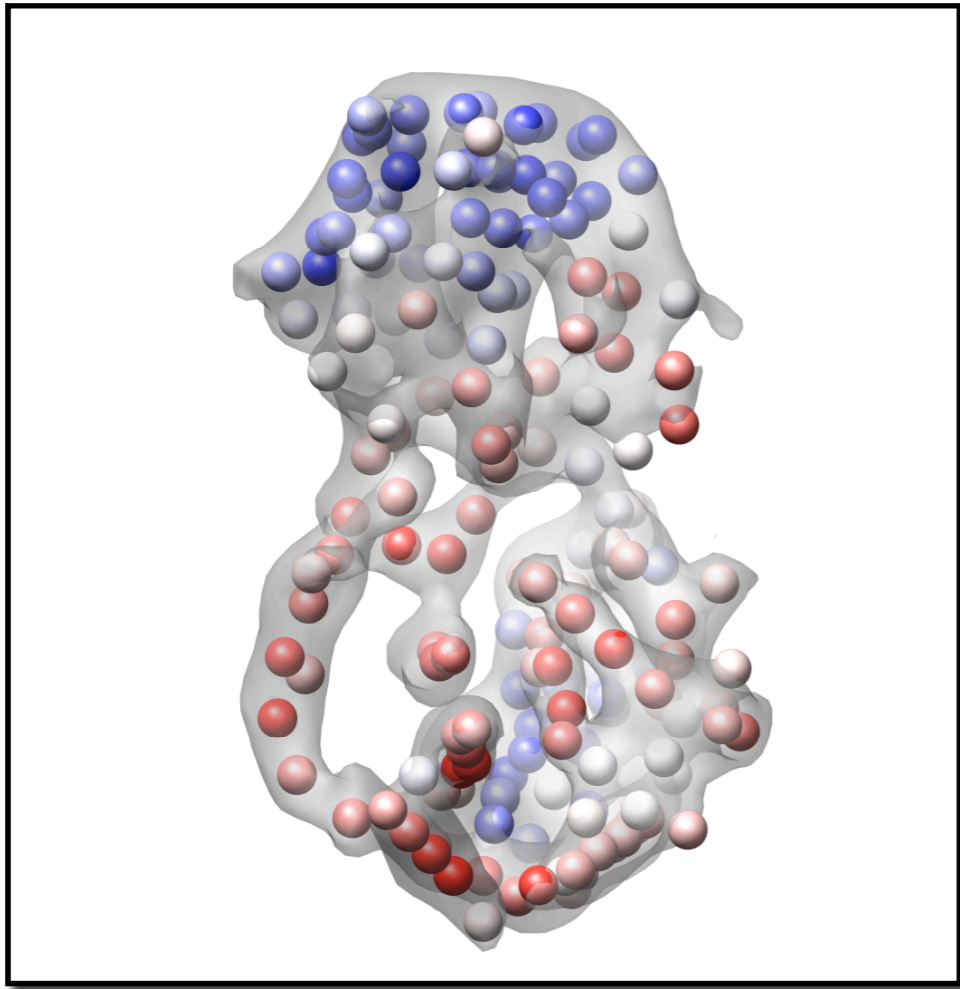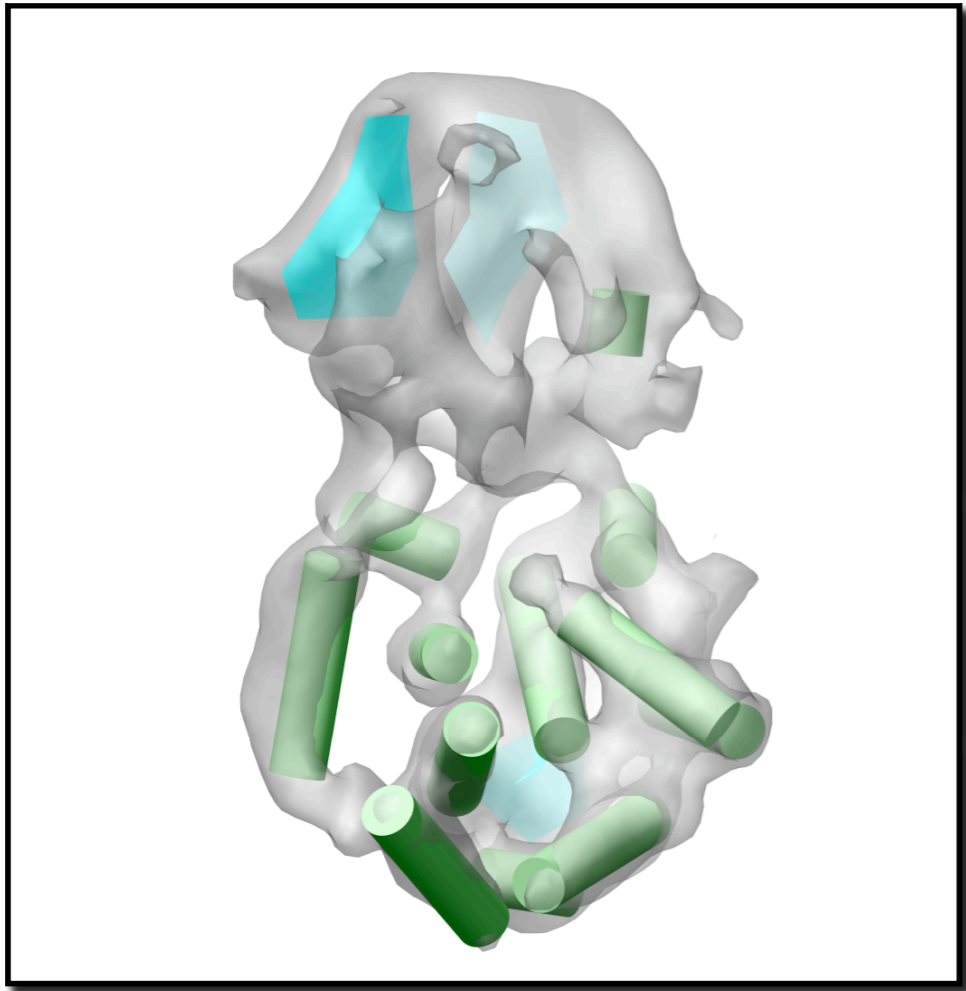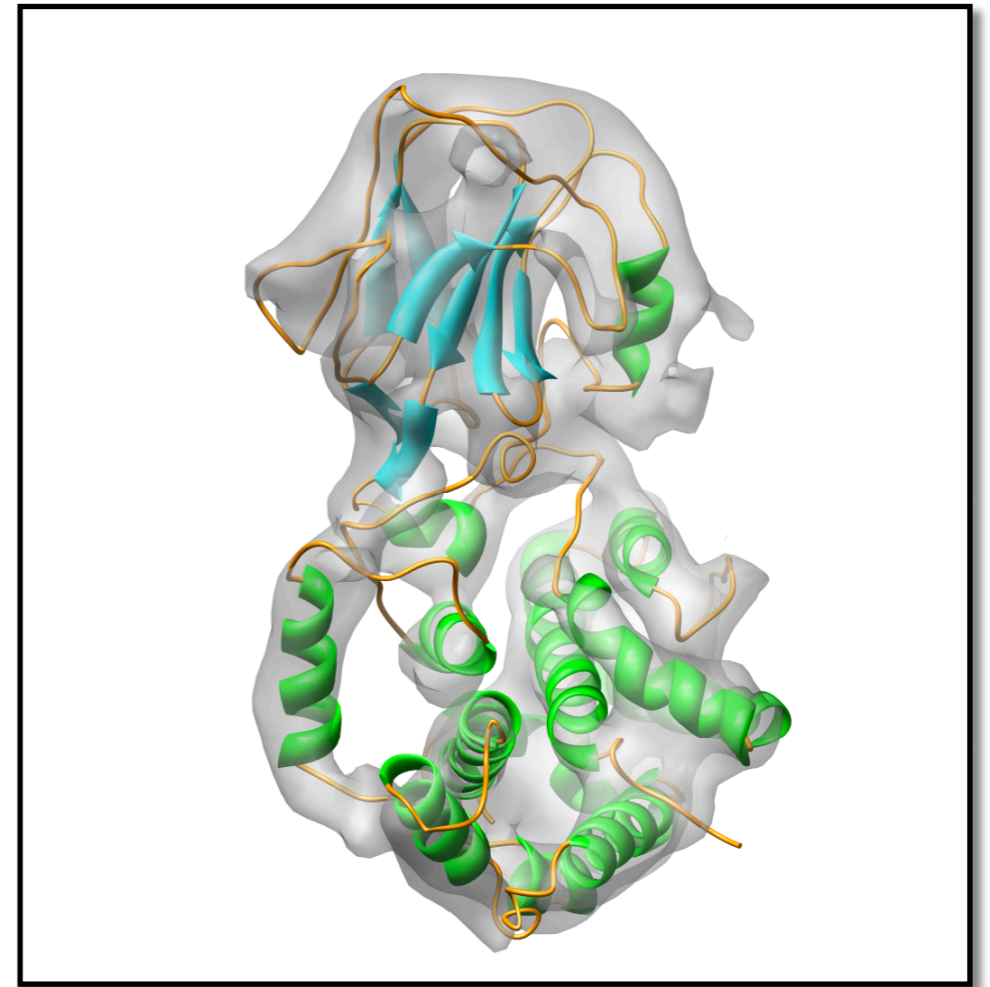
SSEHUNTER: METHODOLOGY
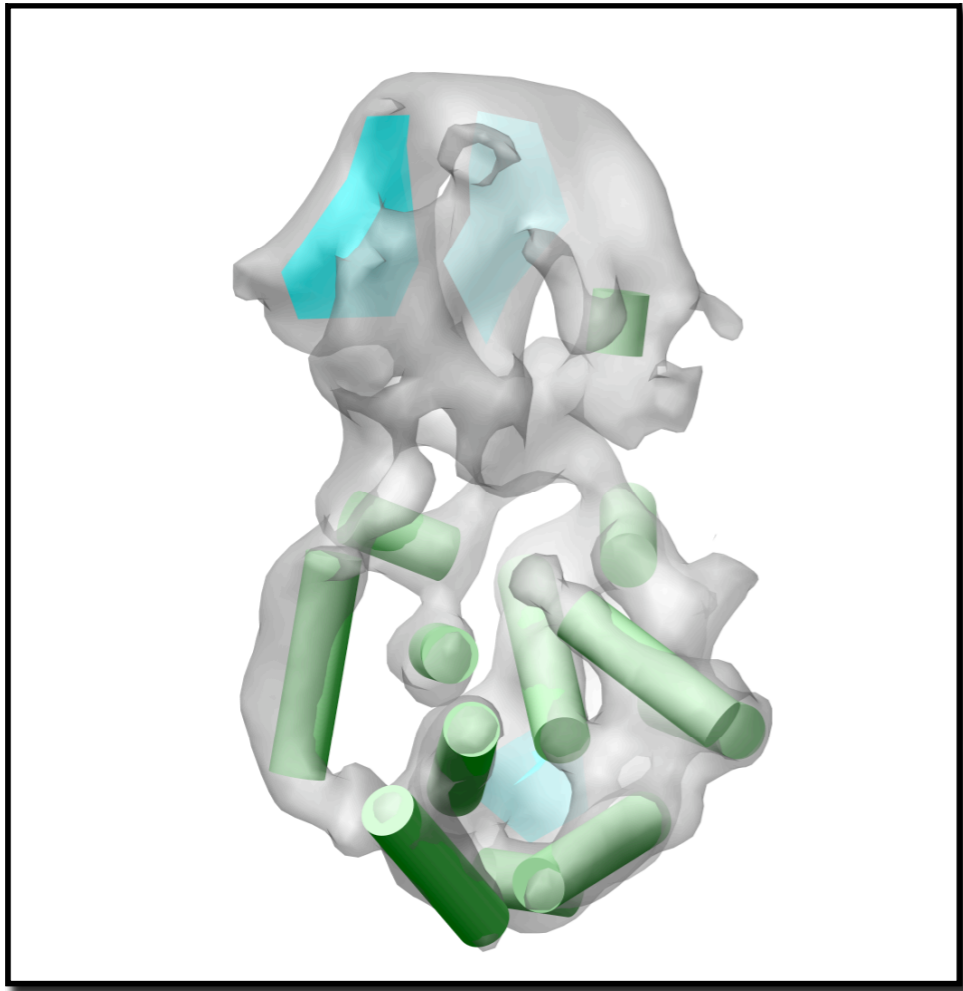
SSEHUNTER: METHODOLOGY

SSEHUNTER: 6.8Å RESOLUTION RDV P8

SSEHUNTER: 6.8Å RESOLUTION RDV P8

SSEHUNTER: 6.8Å RESOLUTION RDV P8
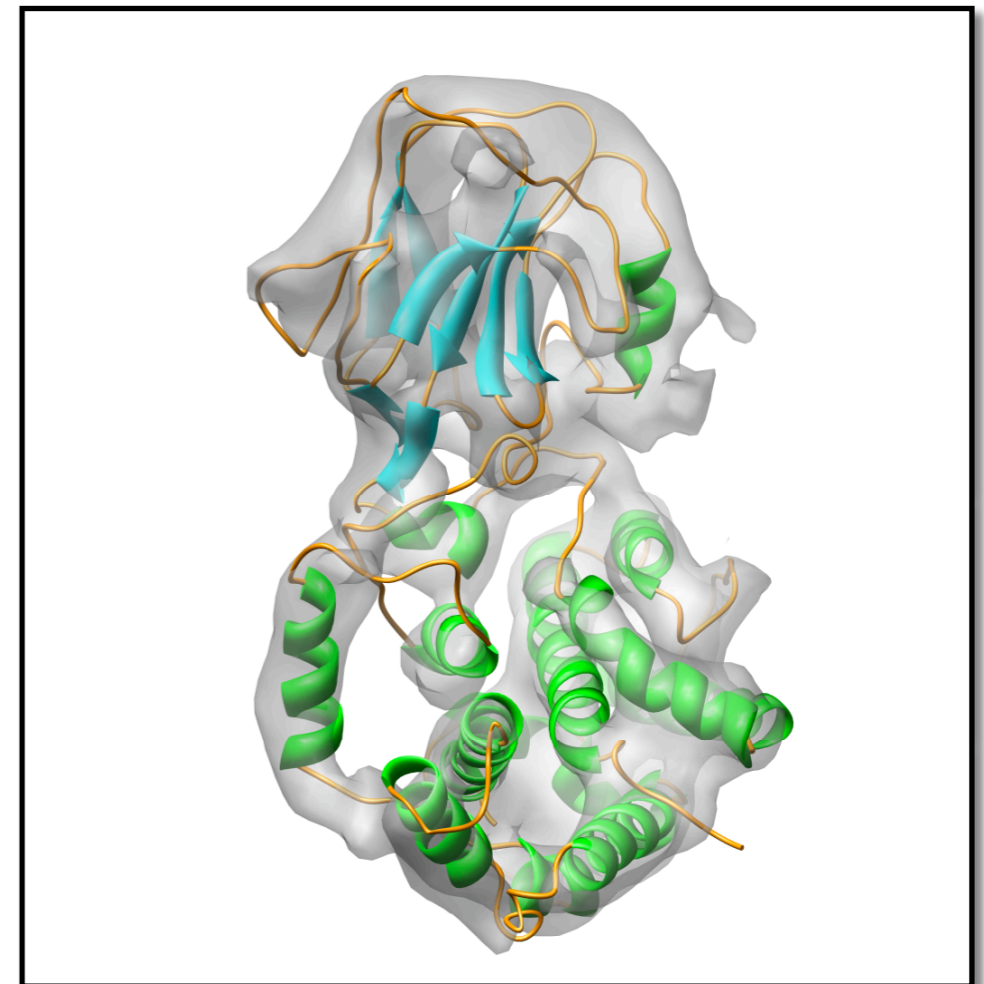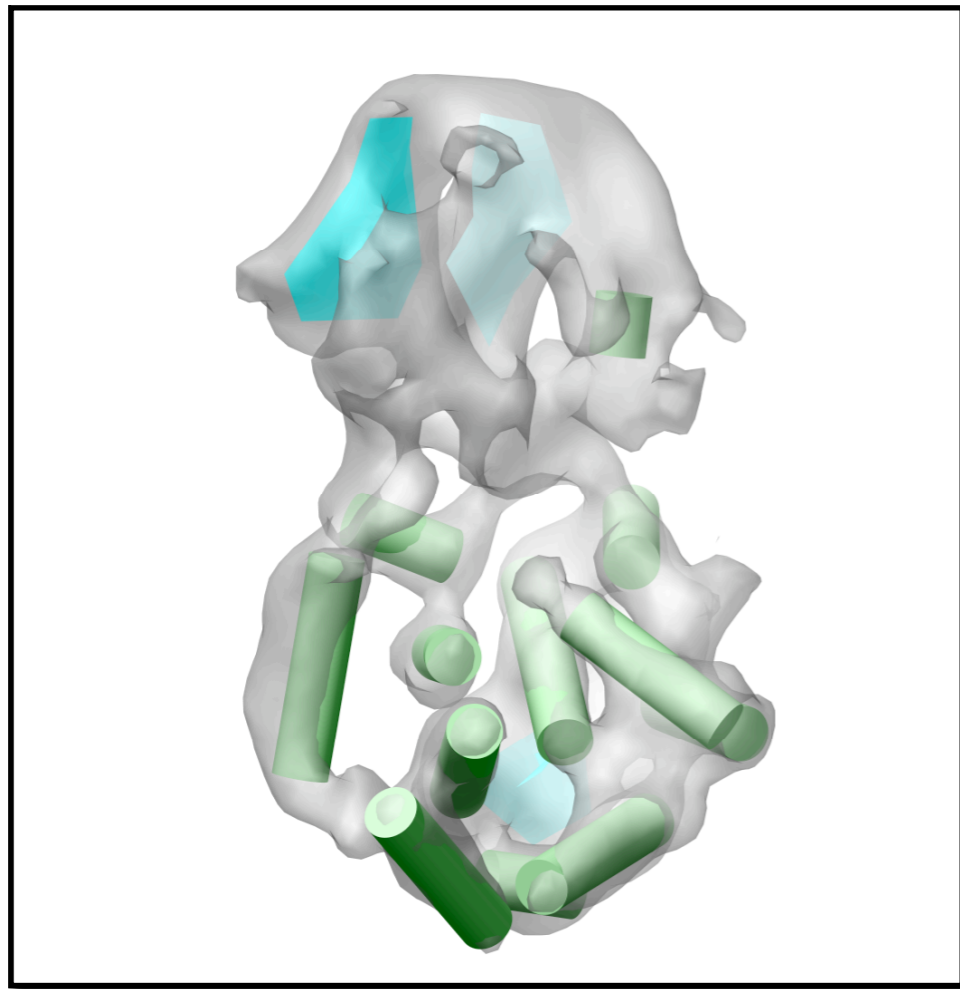
structure type	helix <=4aa	helix 5-8aa	helix >8aa	2 stranded sheet	3+ stranded sheet
SSEs detected	7/24 (29.2%)	39/58 (67.2%)	133/134 (99.3%)	7/24 (29.2%)	25/25 (100%)

TOPOLOGY ASSIGNMENT
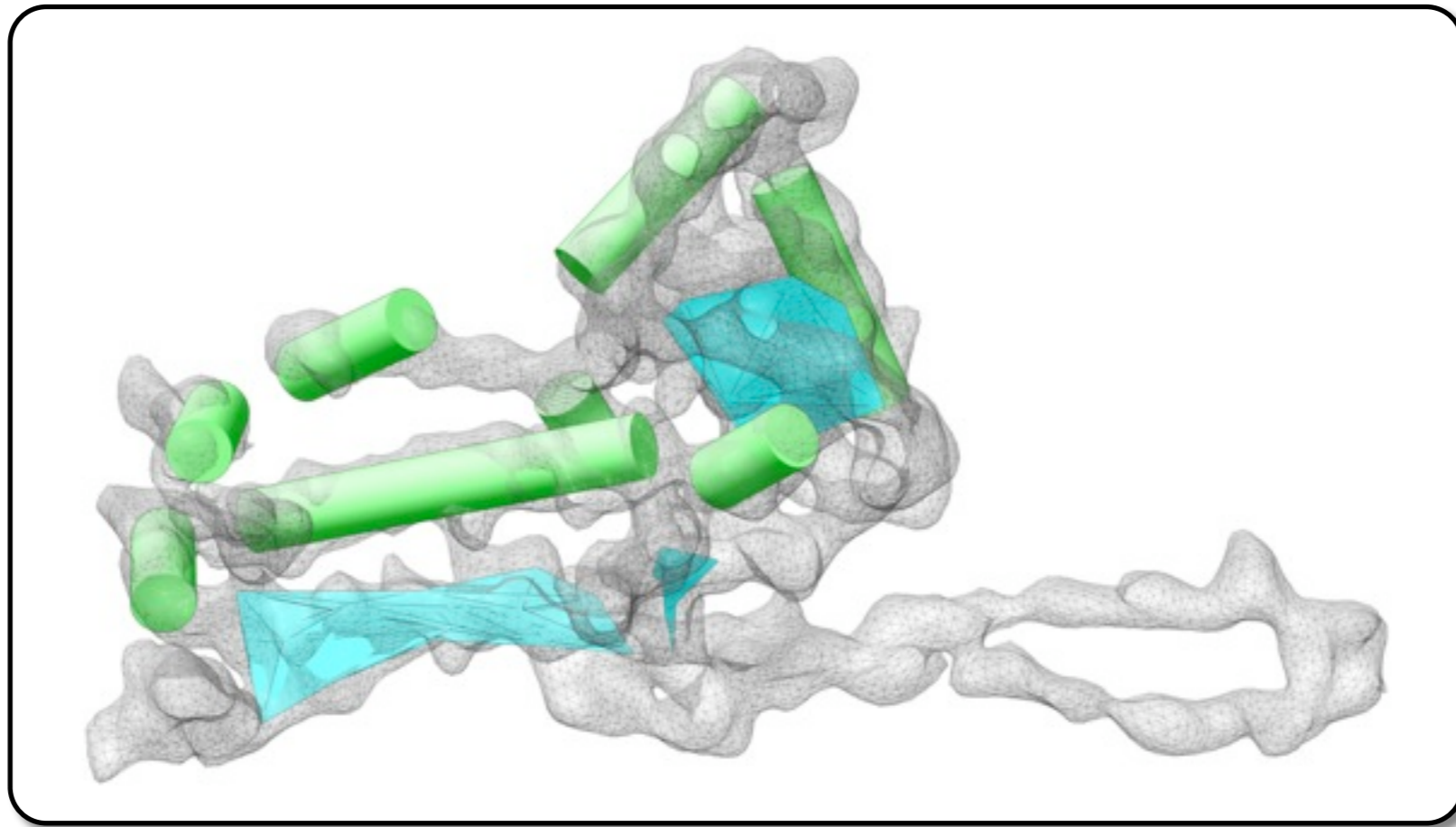

```
  1  MALIGQTLPSLLDIYNRTDKNGRIARIVEQLAKTNDILTDAIYVPCNDGSKHKTTIRAGIPEPVWRRYNQ
 71  GVQPTKTQTVPVTDTTGMLYDLGFVDKALADRSNNAAAFRVSENMGKLQGFNNKVARYSIYGNTDAEPEA
141  FMGLAPRFNTLSTSKAASAENVFSAGGSGSTNTSIWFMSWGENTAHMIYPEGMVAGFQHEDLGDDLVSDG
211  NGGQFRAYRDEFKWDIGLSVRDWRSISRICNIDVTTLTKDASTGADLISMMVDAYYARDVAMLGDGKEVI
281  YANKTIHAWLHKQAMNAKNVNLTIEEYGGKKIVSFLGIPIRRVDAILNTESAVTA
```

HSV-1 VP5

- VP5/VP26 Hexon subunit from 8.5Å resolution HSV-1 reconstruction

  - Secondary structure elements in upper domain correspond to VP5ud crystal structure

  - Helix bundle in middle domain have annexin fold

  - Floor domain has same architecture as bacteriophages
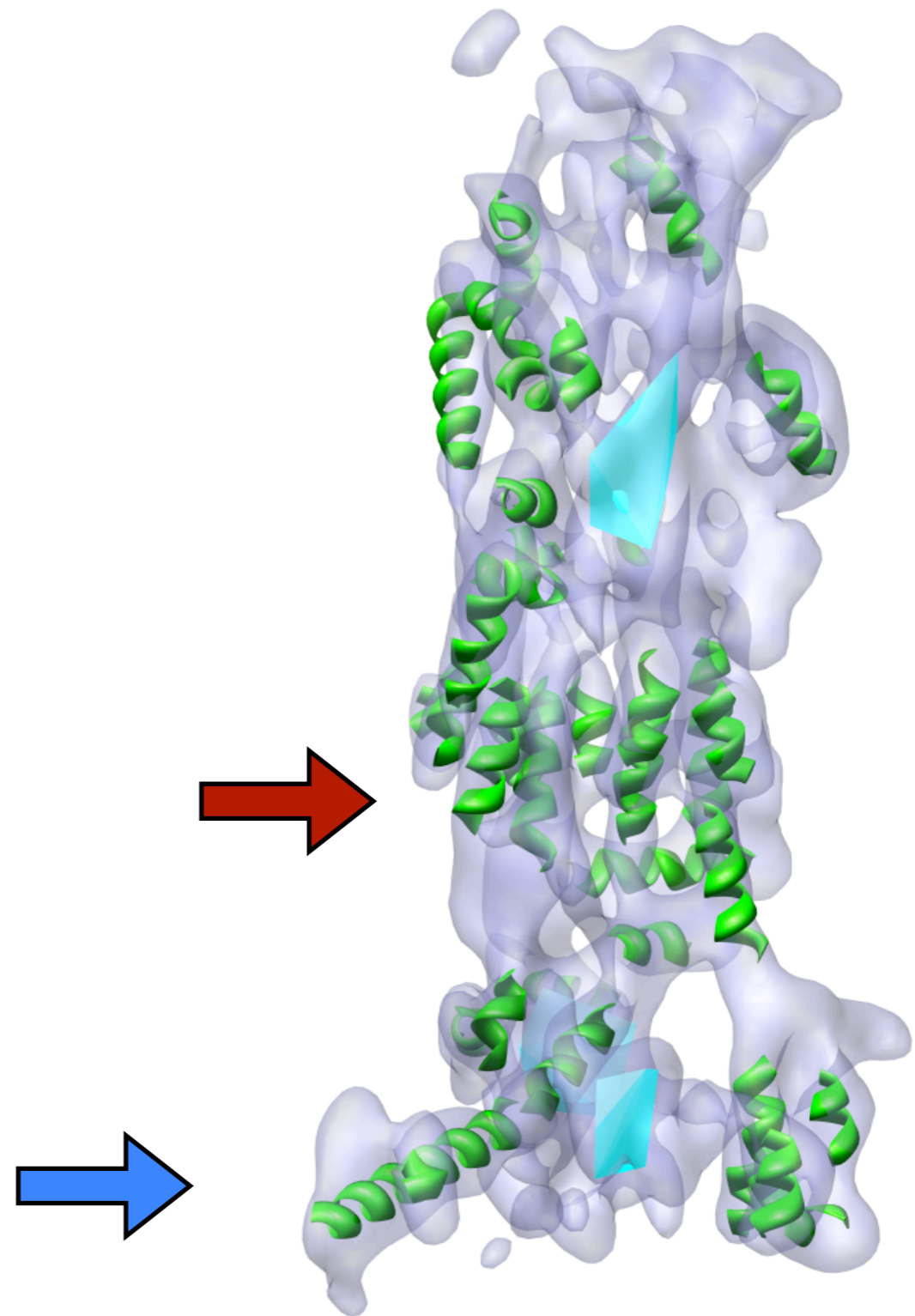
  - Helix re-arrangement in floor domain of penton subunit

Baker, M.L., Jiang, W., Zhou, Z.H., Rixon, F., Chiu, W. (2003) Architecture of the Herpes Simplex Virus Major Capsid Protein Derived from Electron Cryomicroscopy and Bioinformatics. *J Mol Biol* (331), p 447-456

Baker, M.L., Jiang, W., Rixon, F., Chiu, W. (2005) Common Ancestry of Herpesviruses and Tailed DNA Bacteriophages. *J Virol* (79), p 14967-14970.

15

NEAR ATOMIC RESOLUTIONS (3.5-5Å)

Features

- Subunit interfaces

- Helix pitch visible

- Partial strand separation

- Clear connections between secondary structure elements

- Protrusions at bulky amino acid positions

- Traceable with *de novo* methods

- Secondary structure element anchors

Limitations

- "Lost in the forest": complicated segmentations

- Difficult to distinguishing between features and noise

- Limited sidechain density

- Modeling depends on accuracy of secondary structure prediction
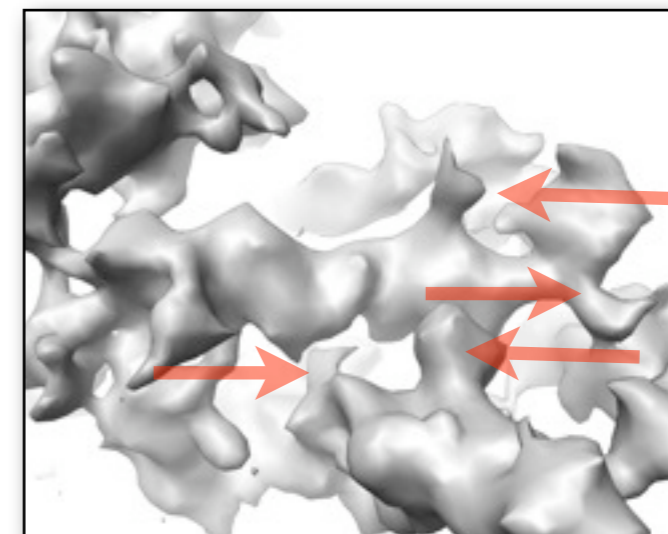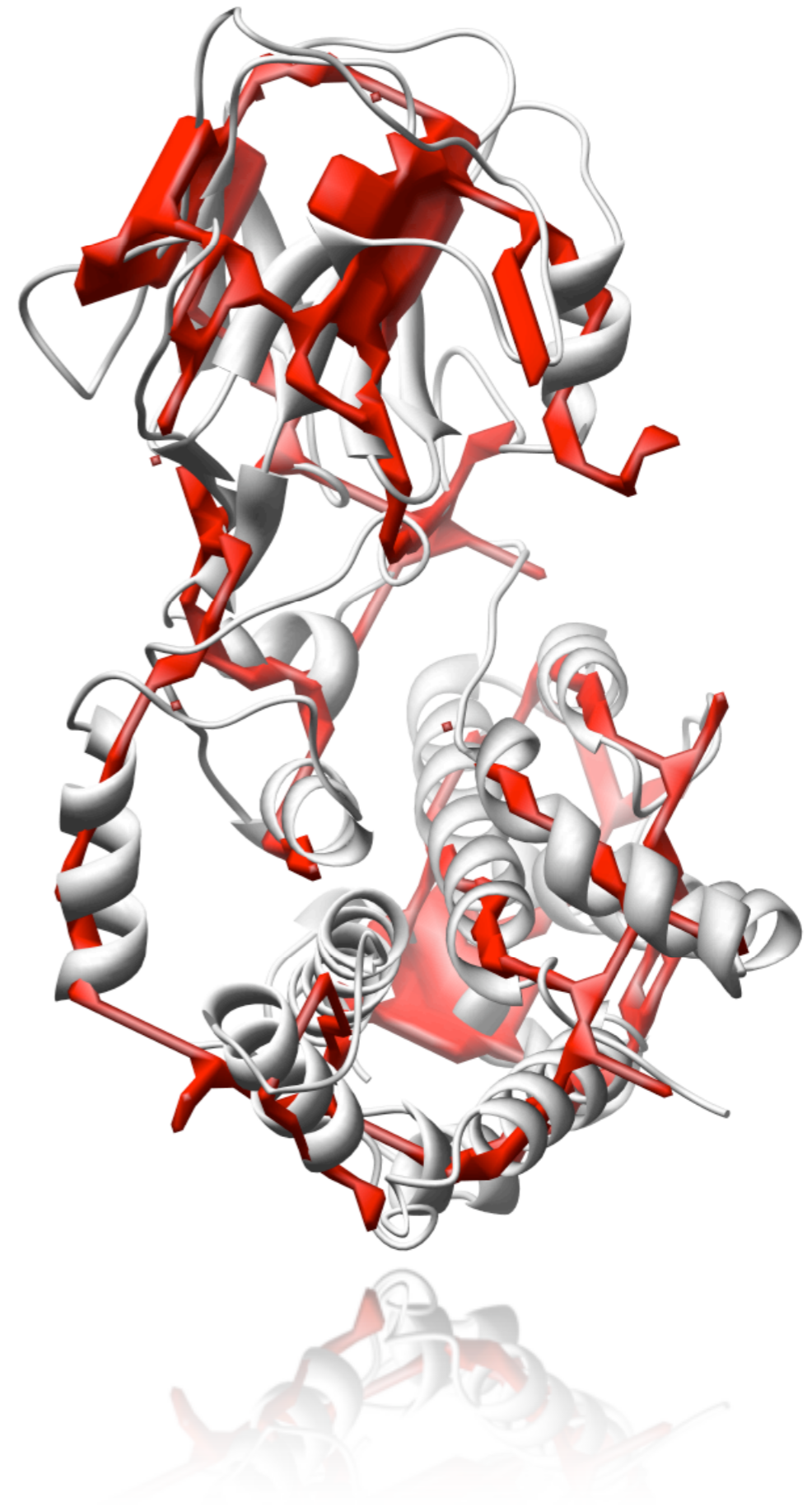
CONNECTING SSE

- Density skeleton: a compact geometric representation of a volume

- Feature preserving

  - Sheets are flat surfaces

  - Helices and loops are curves

- Topology preserving

  - Maintains density connectivity while minimizing number of branches and breaks
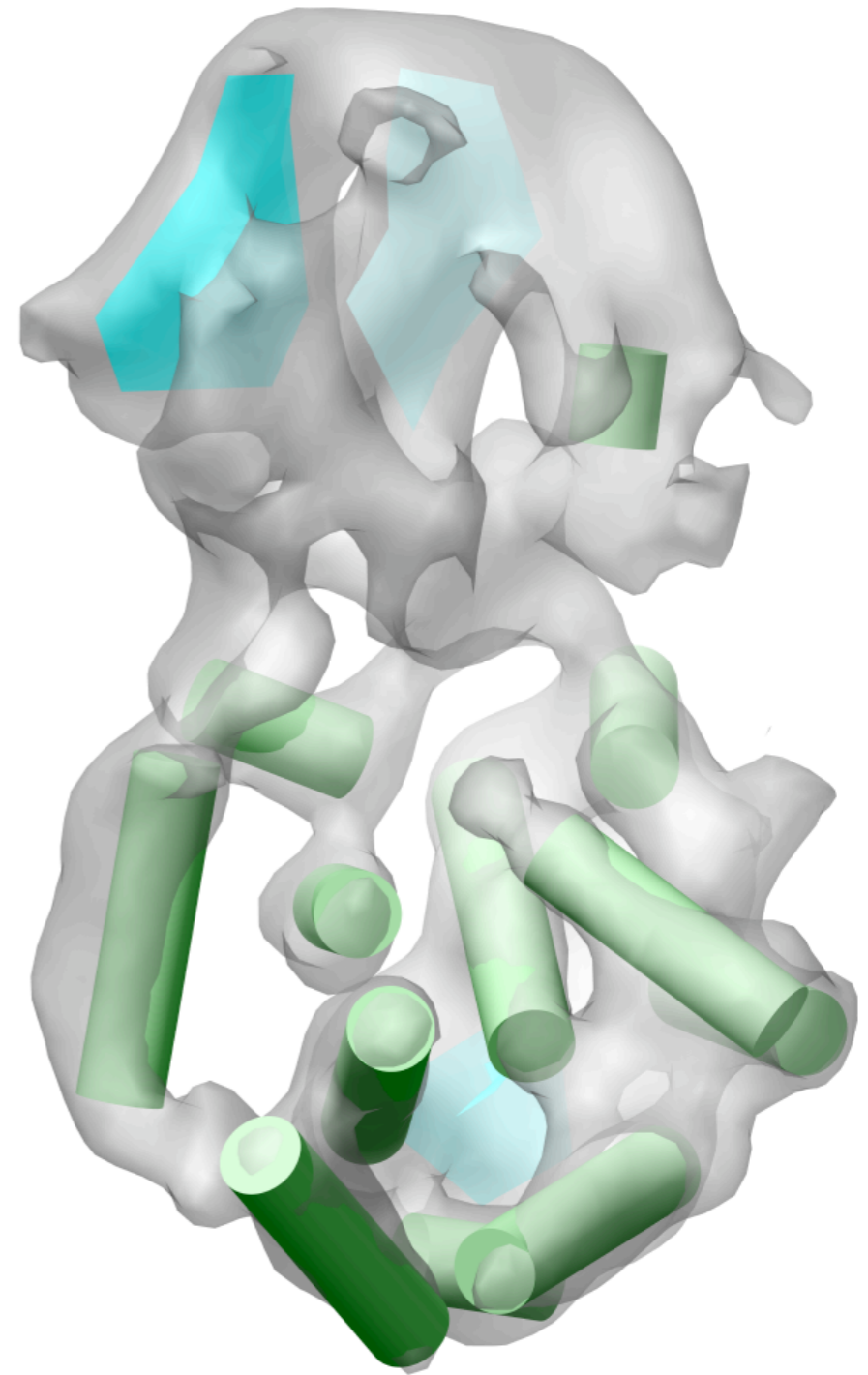
Ju, T., Baker, M.L., Chiu, W. (2007). Computing a Family of Skeletons of Volumetric Models for Shape Description. Computer-AIDED Design (39), p 352-360.
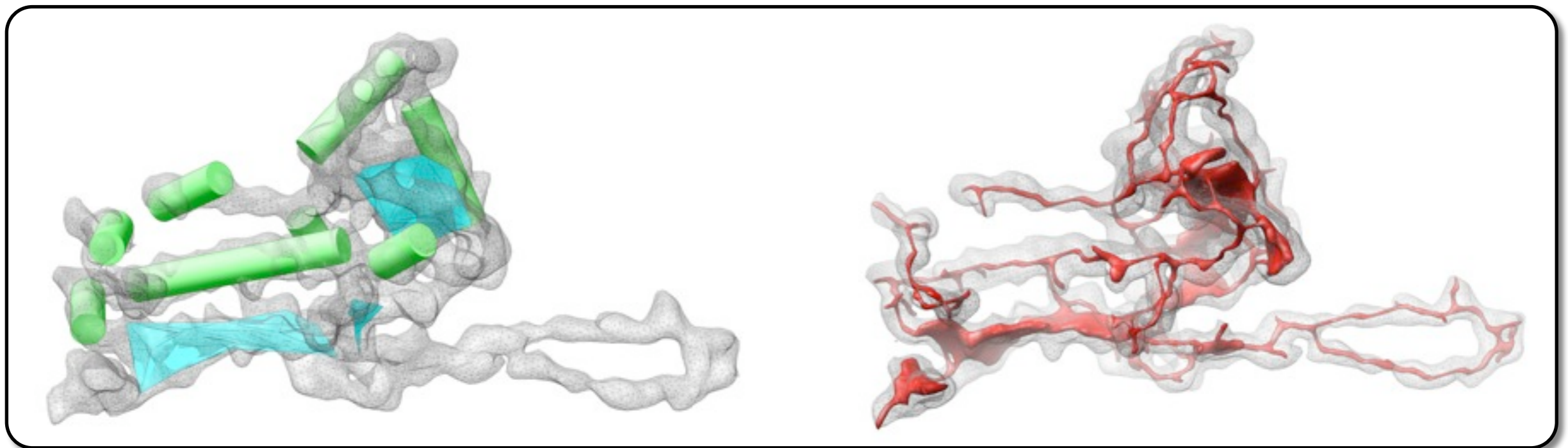
CONNECTING SSE

- Density skeleton: a compact geometric representation of a volume

- Feature preserving

  - Sheets are flat surfaces

  - Helices and loops are curves

- Topology preserving

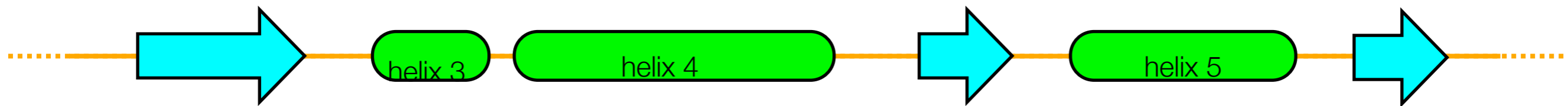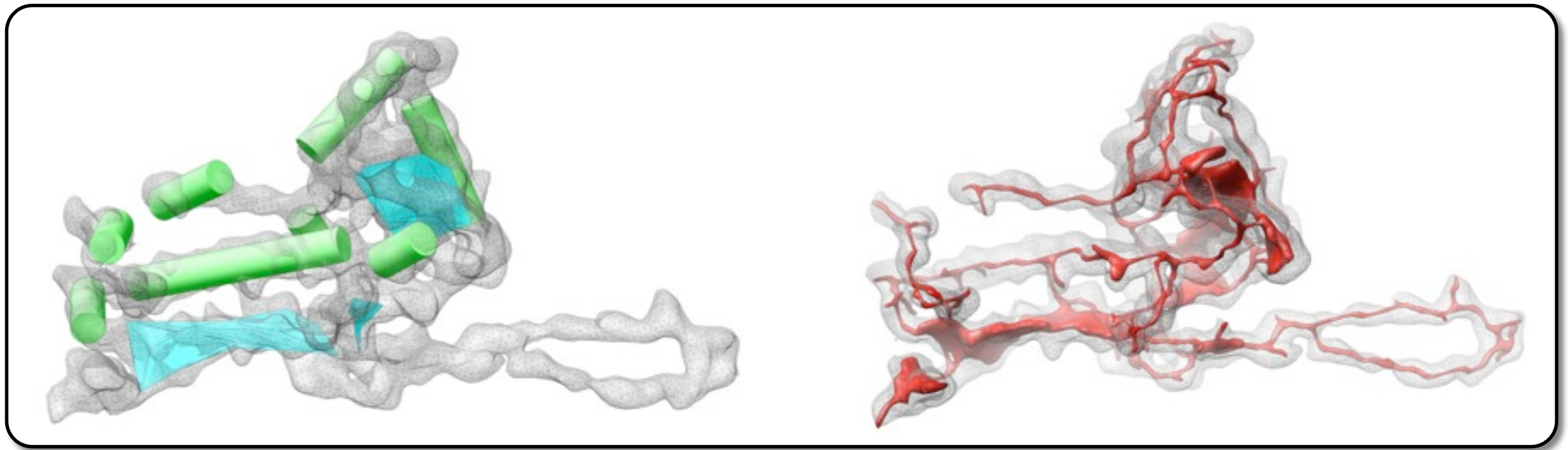 - Maintains density connectivity while minimizing number of branches and breaks

Ju, T., Baker, M.L., Chiu, W. (2007). Computing a Family of Skeletons of Volumetric Models for Shape Description. Computer-AIDED Design (39), p 352-360.
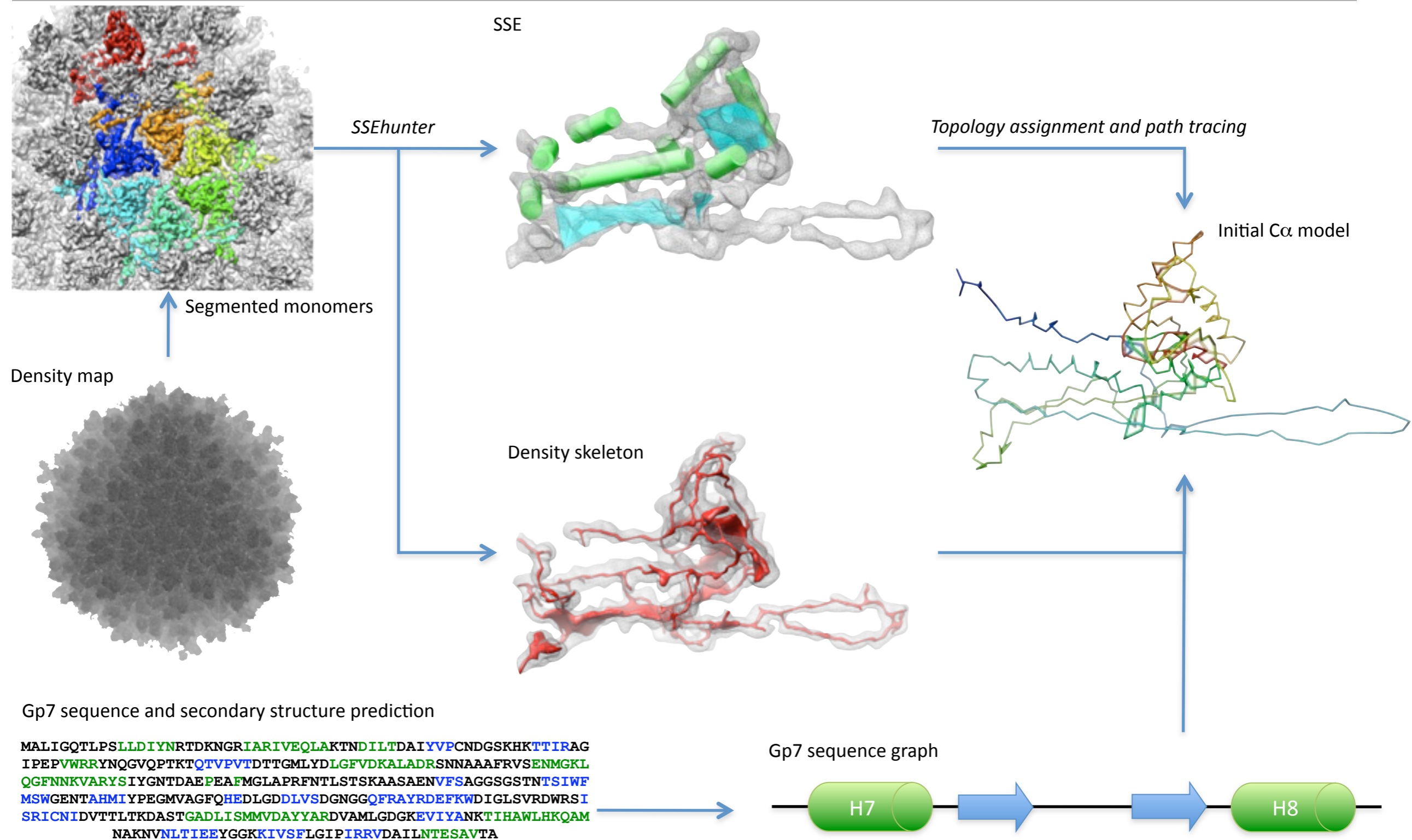
```
  1  MALIGQTLPSLLDIYNRTDKNGRIARIVEQLAKTNDILTDAIYVPCNDGSKHKTTIRAGIPEPVWRRYNQ
 71  GVQPTKTQTVPVTDTTGMLYDLGFVDKALADRSNNAAAFRVSENMGKLQGFNNKVARYSIYGNTDAEPEA
141  FMGLAPRFNTLSTSKAASAENVFSAGGSGSTNTSIWFMSWGENTAHMIYPEGMVAGFQHEDLGDDLVSDG
211  NGGQFRAYRDEFKWDIGLSVRDWRSISRICNIDVTTLTKDASTGADLISMMVDAYYARDVAMLGDGKEVI
281  YANKTIHAWLHKQAMNAKNVNLTIEEYGGKKIVSFLGIPIRRVDAILNTESAVTA
```
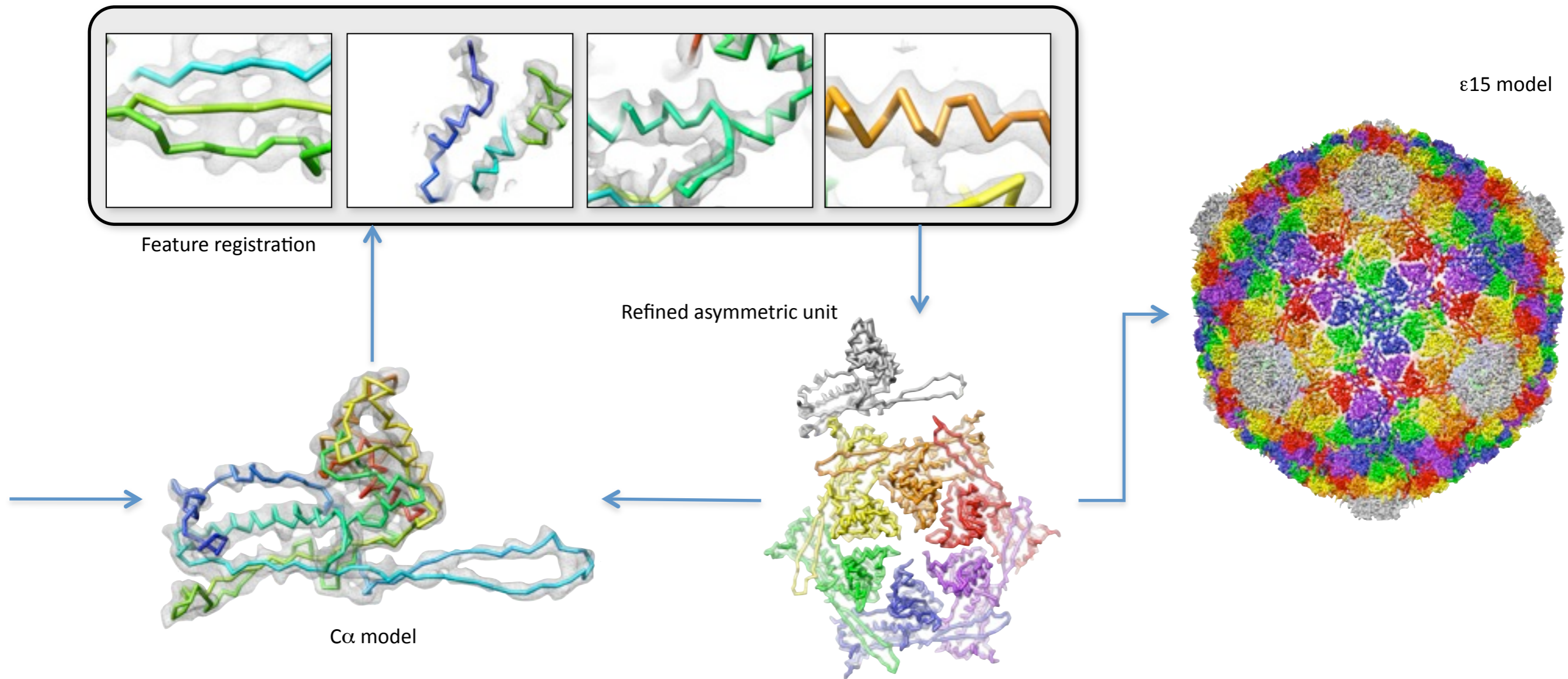
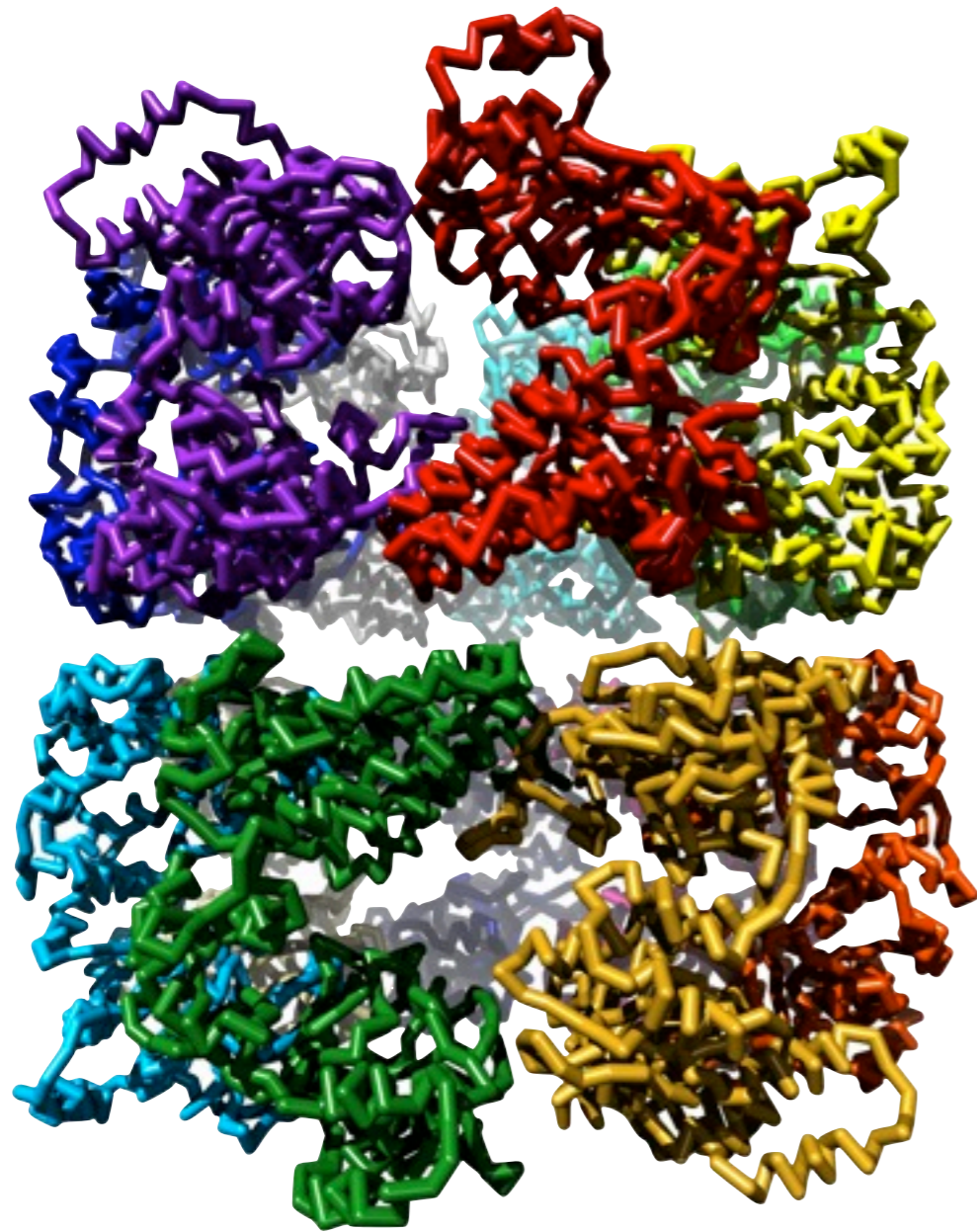# MAPPING SEQUENCE TO STRUCTURE

DE NOVO MODELING: INITIAL MODEL

SSE

SSEhunter

*Topology assignment and path tracing*

Initial Cα model

Segmented monomers

Density map

Density skeleton

Gp7 sequence and secondary structure prediction

MALIGQTLPS**LLDIYN**RTDKNGR**IARIVEQLA**KTND**ILT**DAI**YVP**CNDGSKHK**TTIR**AG
IPEP**VWRR**YNQGVQPTKT**QTVPVT**DTTGMLYD**LGFVDKALADR**SNNAAAFRVS**ENMGKL**
QGFNNKVARYSIYGNTDAE**PEAF**MGLAPRFNTLSTSKAASAEN**VFS**AGGSGSTN**TSIWF**
MSWGENT**AHMI**YPEGMVAGFQ**HE**DLGD**DLVS**DGNGG**QFRAYRDEFKW**DIGLSVRDWRS**I**
SRICNIDVTTLTKDAST**GADLIS**MMVDAYYARDVAMLGDGK**EVIYA**NK**TIHAWLHKQAM**
NAKNV**NLTIEE**YGGK**KIVSF**LGIP**IRRV**DAIL**NTESAV**TA

Gp7 sequence graph

H7 → → H8

DE NOVO MODELING

Model optimization

Feature registration

Refined asymmetric unit

ε15 model

Cα model

4.2Å resolution GroEL

4.5Å resolution ε15

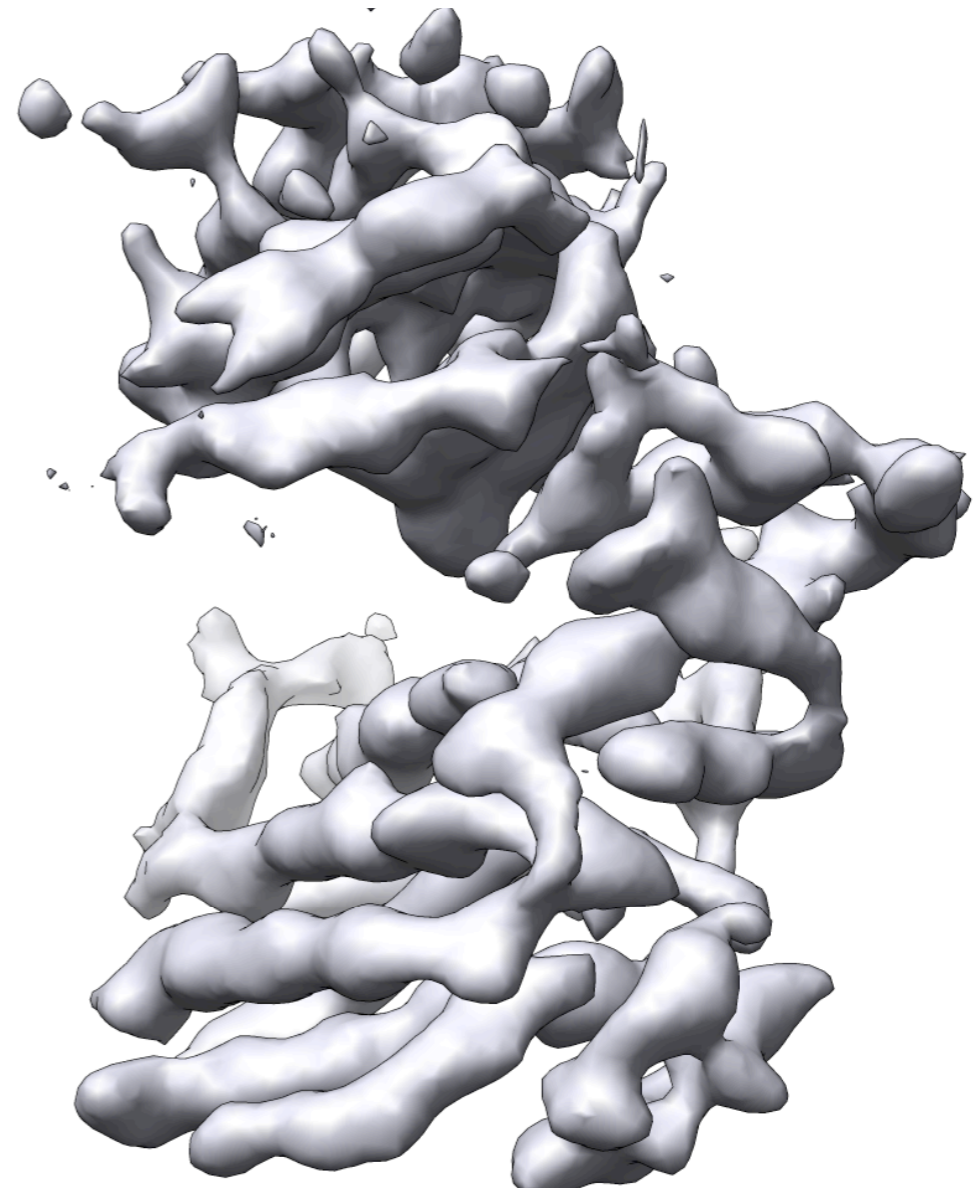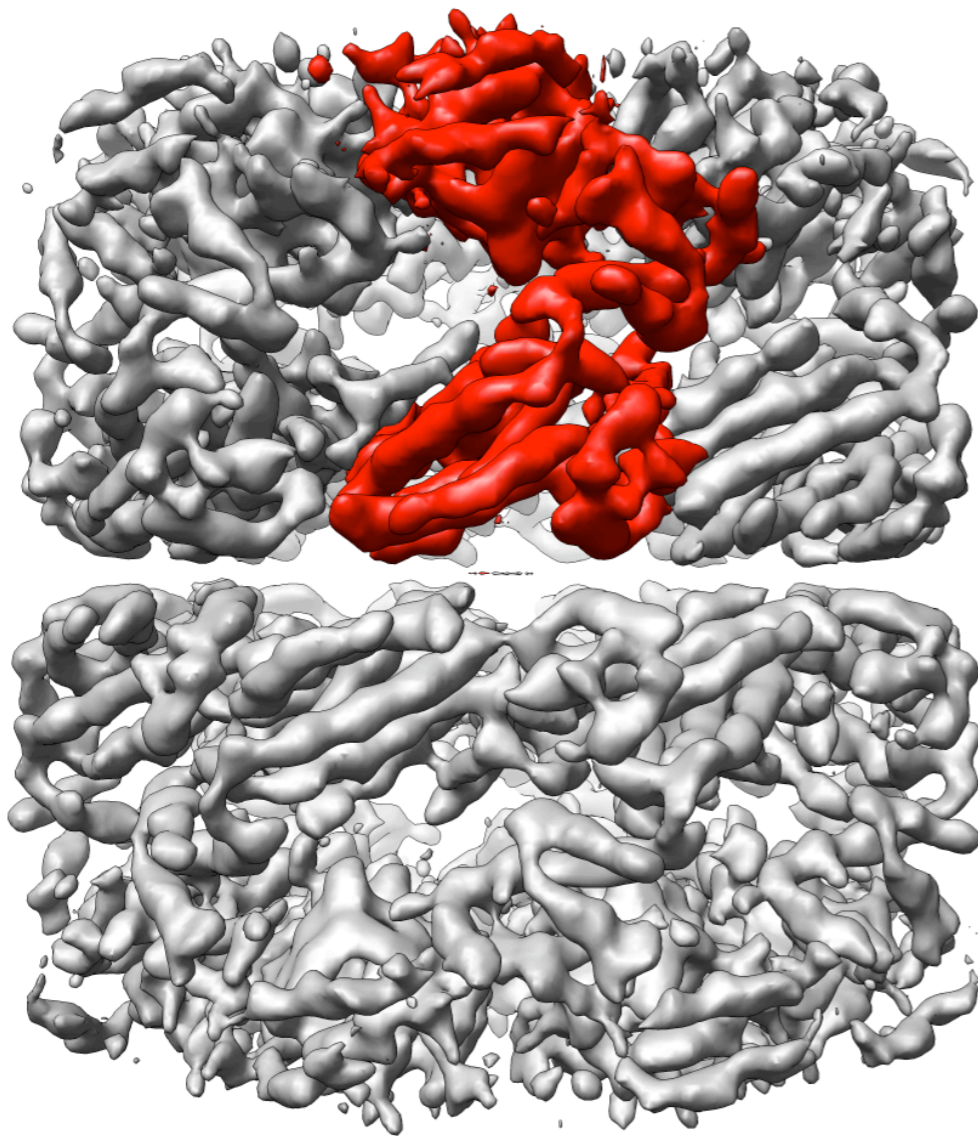Ludtke, S.J.*, Baker, M.L.*, Chen, D.H. Song, J.L., Chuang, D.T., Chiu, W. (2008) *Structure* (16), p 441-448.
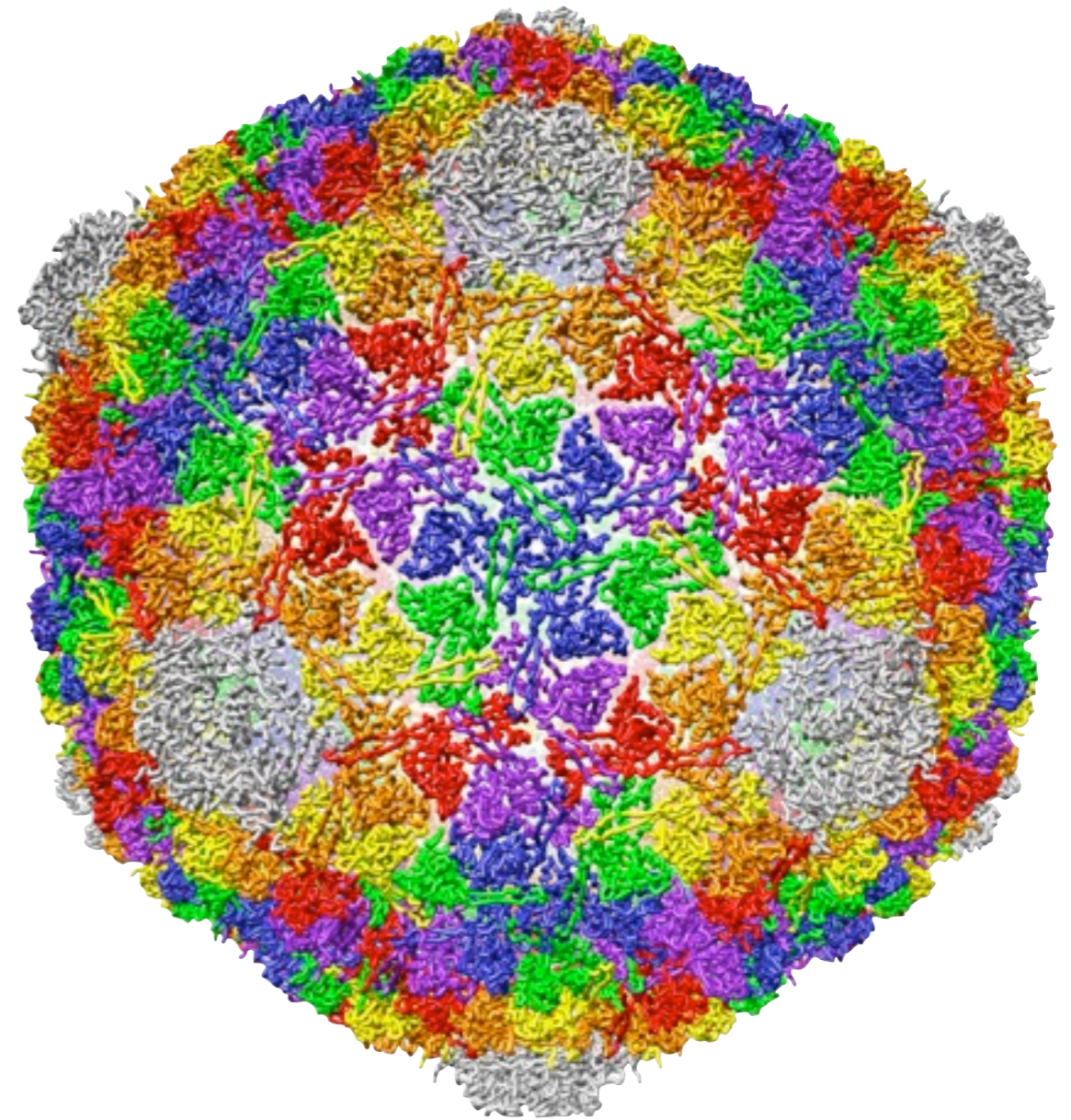
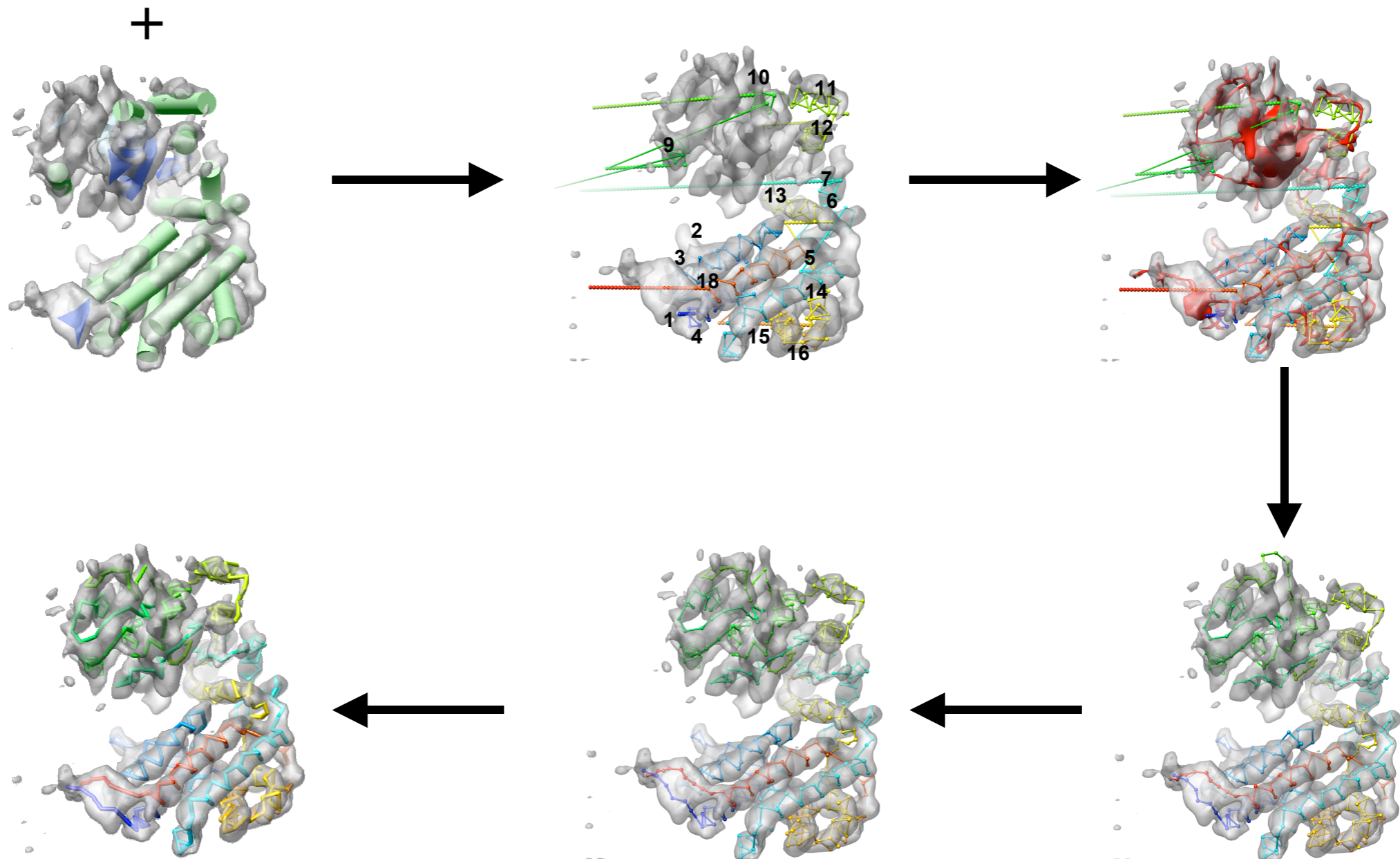Jiang, W.*, Baker, M.L.*, Jakana, J Weigele, P.R., King, J., Chiu W. (2008) *Nature* (451), p 1130-1135.
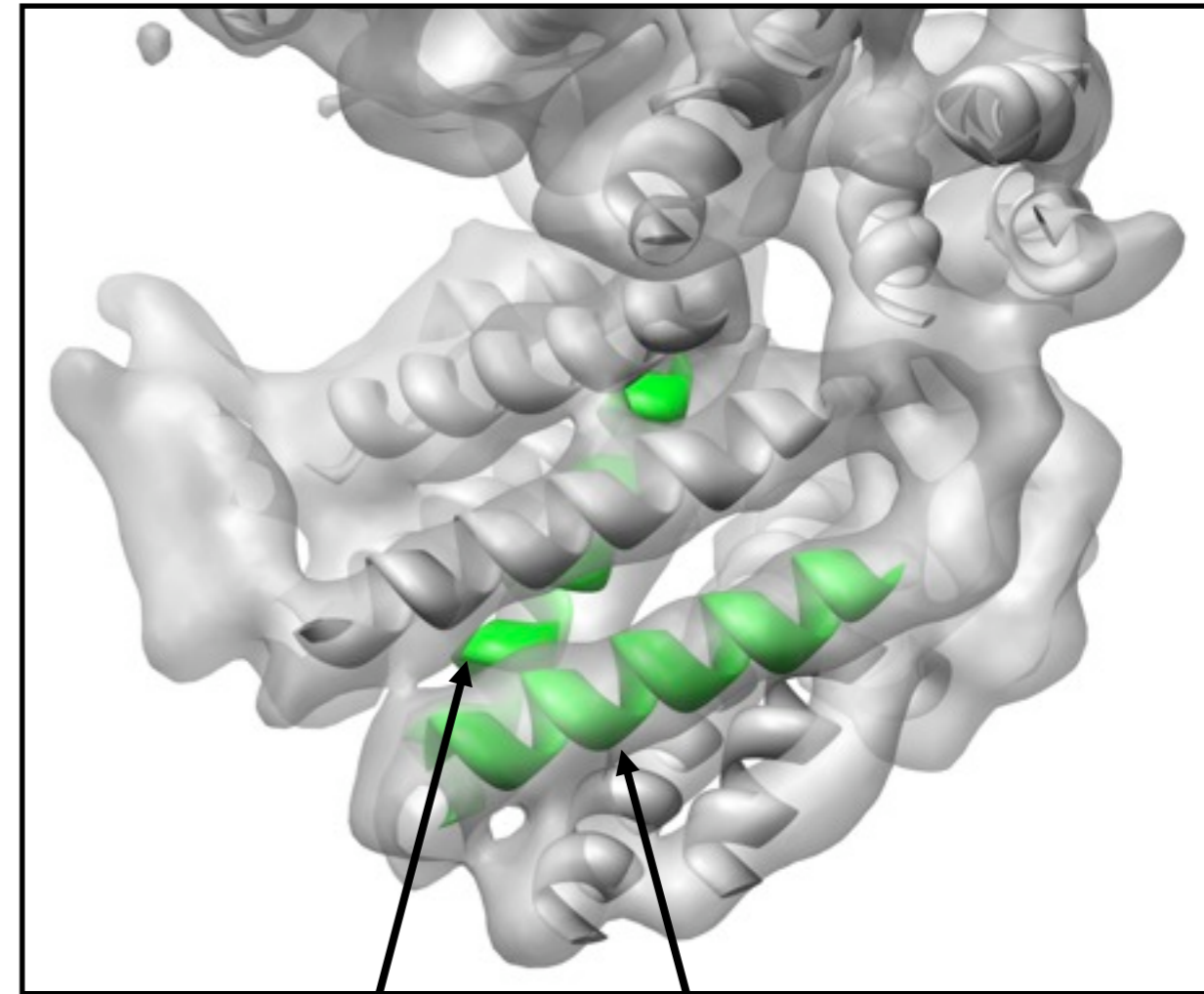
Cα MODELS

GroEL AT 4.2Å RESOLUTION

GroEL: *DE NOVO* MODELING

..S**LLDIYN**RTDKNG**DILT**DAI**YVP**CNDGSK..

STARTING THE TRACE

20 aa   19 aa

STARTING THE TRACE


```
1    AAKDVKFGNDAGVKMLRGVNVLADAVKVTLGPKGRNVVLDKSFGAPTITKDGVSVAREIE
61   LEDKFENMGAQMVKEVASKANDAAGDGTTTATVLAQAIITEGLKAVAAGMNPMDLKRGID
121  KAVTVAVEELKALSVPCSDSKAIAQVGTISANSDETVGKLIAEAMDKVGKEGVITVEDGT
181  GLQDELDVVEGMQFDRGYLSPYFINKPETGAVELESPFILLADKKISNIREMLPVLEAVA
241  KAGKPLLIIAEDVEGEALATAVVNTIRGIVKVAAVKAPGFGDRRKAMLQDIATLTGGTVI
301  SEEIGMELEKATLEDLGQAKRVVINKDTTTIIDGVGEEAAIQGRVAQIRQQIEEATSDYD
361  REKLQERVAKLAGGVAVIKVGAATEVEMKEKKARVEDALHATRAAVEEGVVAGGGVALIR
421  VASKLADLRGQNEDQNVGIKVALRAMEAPLRQIVLNCGEEPSVVANTVKGGDGNYGYNAA
481  TEEYGNMIDMGILDPTKVTRSALQYAASVAGLMITTECMVTDLPKNDAADLGAAGGMGGM
541  GGMGGMM
```
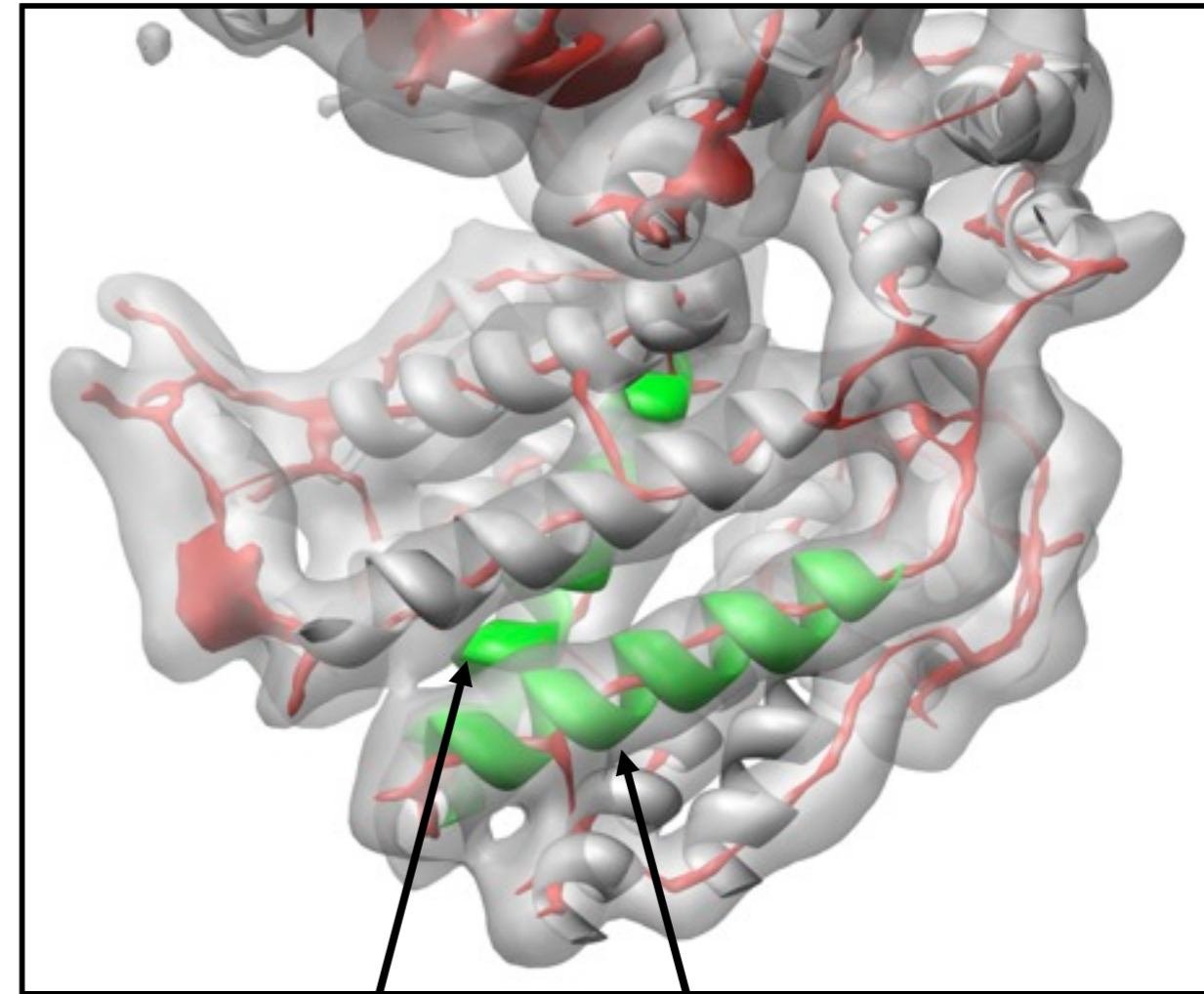
20 aa   19 aa

STARTING THE TRACE


```
1    AAKDVKFGNDAGVKMLRGVNVLADAVKVTLGPKGRNVVLDKSFGAPTITKDGVSVAREIE

61   LEDKFENMGAQMVKEVASKANDAAGDGTTTATVLAQAIITEGLKAVAAGMNPMDLKRGID

121  KAVTVAVEELKALSVPCSDSKAIAQVGTISANSDETVGKLIAEAMDKVGKEGVITVEDGT

181  GLQDELDVVEGMQFDRGYLSPYFINKPETGAVELESPFILLADKKISNIREMLPVLEAVA

241  KAGKPLLIIAEDVEGEALATAVVNTIRGIVKVAAVKAPGFGDRRKAMLQDIATLTGGTVI

301  SEEIGMELEKATLEDLGQAKRVVINKDTTTIIDGVGEEAAIQGRVAQIRQQIEEATSDYD

361  REKLQERVAKLAGGVAVIKVGAATEVEMKEKKARVEDALHATRAAVEEGVVAGGGVALIR

421  VASKLADLRGQNEDQNVGIKVALRAMEAPLRQIVLNCGEEPSVVANTVKGGDGNYGYNAA

481  TEEYGNMIDMGILDPTKVTRSALQYAASVAGLMITTECMVTDLPKNDAADLGAAGGMGGM

541  GGMGGMM
```
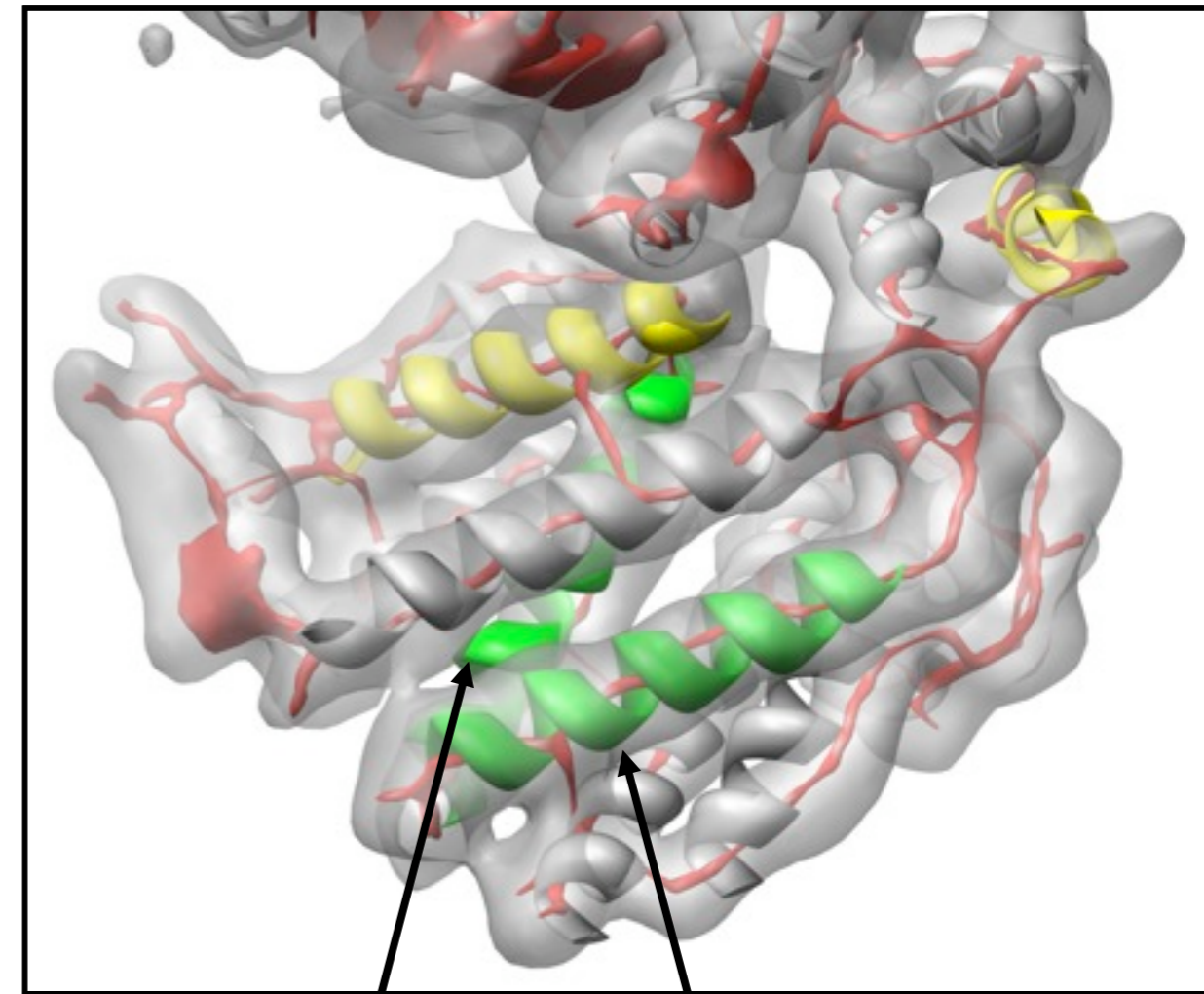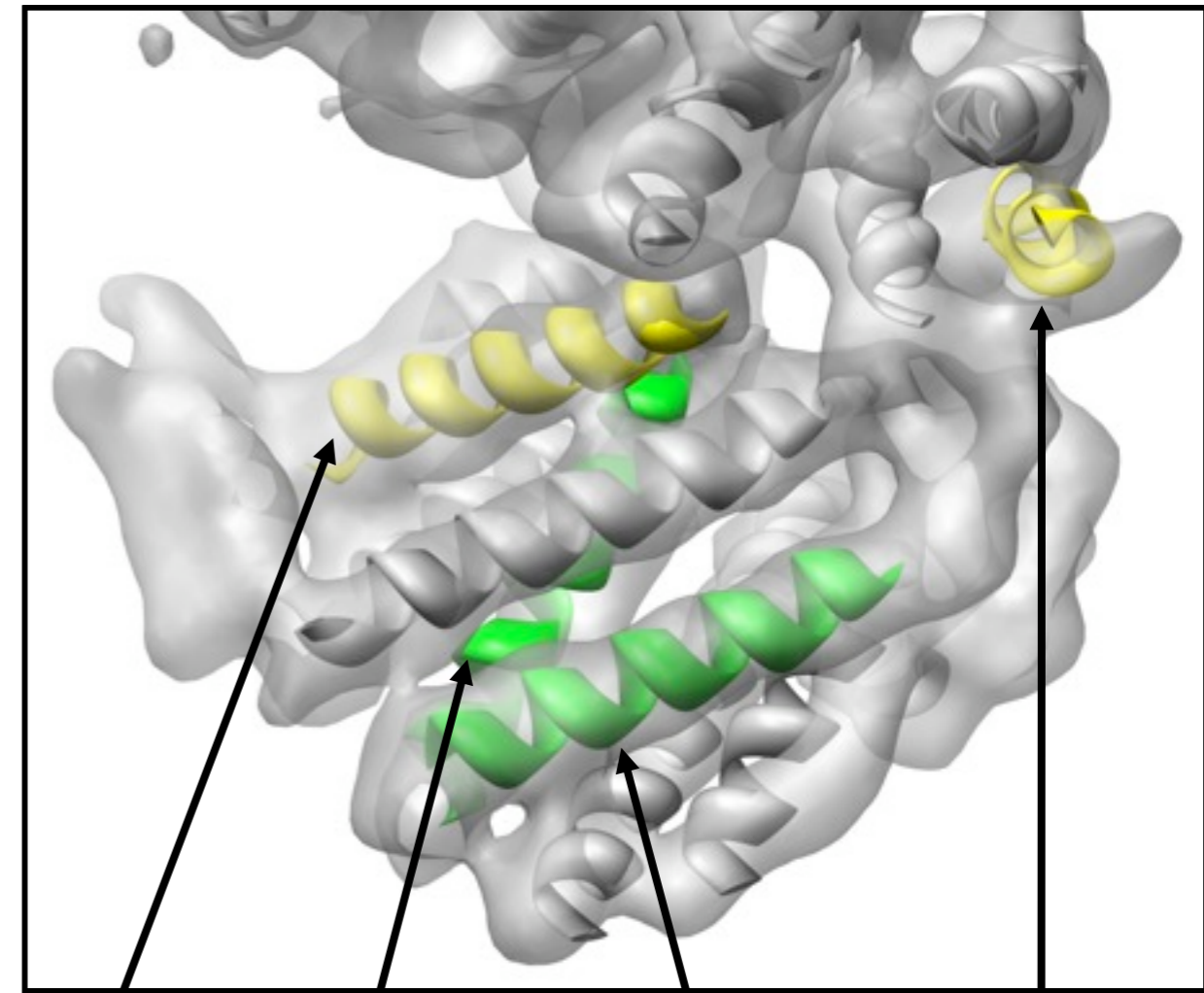
20 aa   19 aa

```
1     AAKDVKFGNDAGVKMLRGVNVLADAVKVTLGPKGRNVVLDKSFGAPTITKDGVSVAREIE

61    LEDKFENMGAQMVKEVASKANDAAGDGTTTATVLAQAIITEGLKAVAAGMNPMDLKRGID

121   KAVTVAVEELKALSVPCSDSKAIAQVGTISANSDETVGKLIAEAMDKVGKEGVITVEDGT

181   GLQDELDVVEGMQFDRGYLSPYFINKPETGAVELESPFILLADKKISNIREMLPVLEAVA

241   KAGKPLLIIAEDVEGEALATAVVNTIRGIVKVAAVKAPGFGDRRKAMLQDIATLTGGTVI

301   SEEIGMELEKATLEDLGQAKRVVINKDTTTIIDGVGEEAAIQGRVAQIRQQIEEATSDYD

361   REKLQERVAKLAGGVAVIKVGAATEVEMKEKKARVEDALHATRAAVEEGVVAGGGVALIR

421   VASKLADLRGQNEDQNVGIKVALRAMEAPLRQIVLNCGEEPSVVANTVKGGDGNYGYNAA

481   TEEYGNMIDMGILDPTKVTRSALQYAASVAGLMITTECMVTDLPKNDAADLGAAGGMGGM

541   GGMGGMM
```
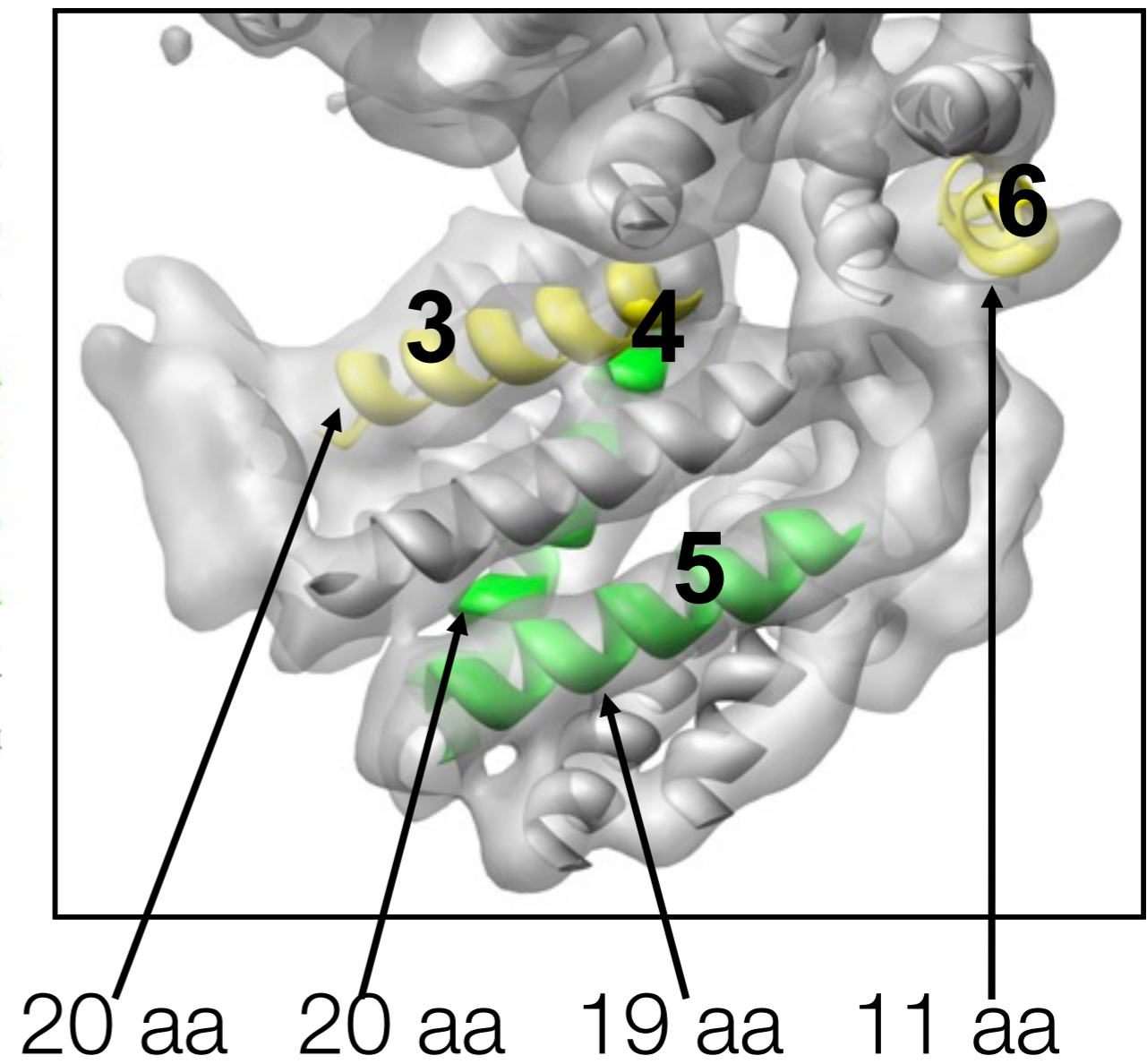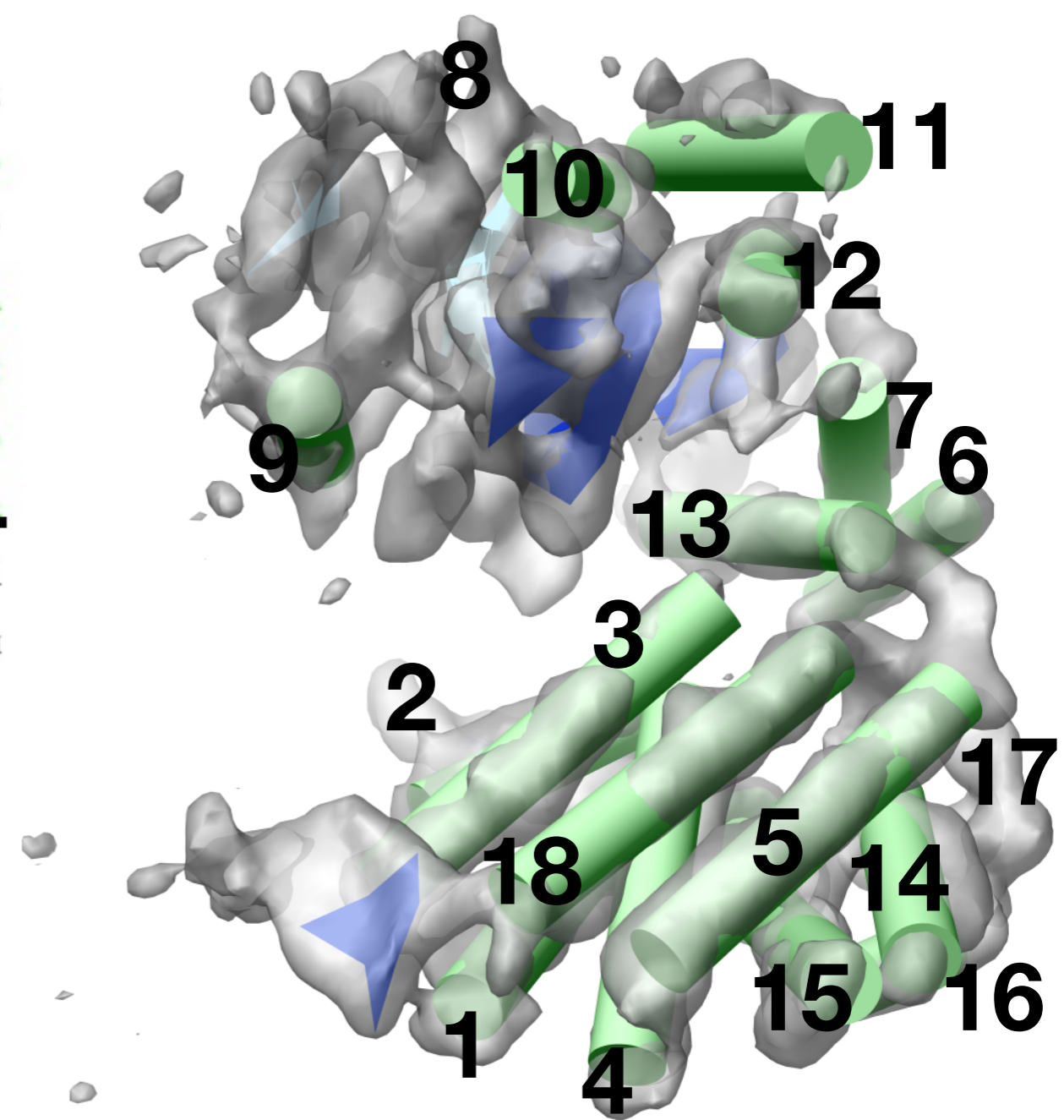
20 aa   19 aa

STARTING THE TRACE


```
1    AAKDVKFGNDAGVKMLRGVNVLADAVKVTLGPKGRNVVLDKSFGAPTITKDGVSVAREIE
61   LEDKFENMGAQMVKEVASKANDAAGDGTTTATVLAQAIITEGLKAVAAGMNPMDLKRGID
121  KAVTVAVEELKALSVPCSDSKAIAQVGTISANSDETVGKLIAEAMDKVGKEGVITVEDGT
181  GLQDELDVVEGMQFDRGYLSPYFINKPETGAVELESPFILLADKKISNIREMLPVLEAVA
241  KAGKPLLIIAEDVEGEALATAVVNTIRGIVKVAAVKAPGFGDRRKAMLQDIATLTGGTVI
301  SEEIGMELEKATLEDLGQAKRVVINKDTTTIIDGVGEEAAIQGRVAQIRQQIEEATSDYD
361  REKLQERVAKLAGGVAVIKVGAATEVEMKEKKARVEDALHATRAAVEEGVVAGGGVALIR
421  VASKLADLRGQNEDQNVGIKVALRAMEAPLRQIVLNCGEEPSVVANTVKGGDGNYGYNAA
481  TEEYGNMIDMGILDPTKVTRSALQYAASVAGLMITTECMVTDLPKNDAADLGAAGGMGGM
541  GGMGGMM
```

20 aa   19 aa

24

20 aa 20 aa 19 aa 11 aa

HELIX CORRESPONDENCE

COMPUTING A CORRESPONDENCE

- Exhaustive search scales poorly
  - ➡ GroEL has 6.4×10^{15} possible helix assignments
- Skeleton provides partial connectivity to reduce complexity
  - ➡ GroEL has 1000+ possible helix assignments that satisfy skeleton connectivity
- Graph matching can rapidly compute a gallery of correspondences
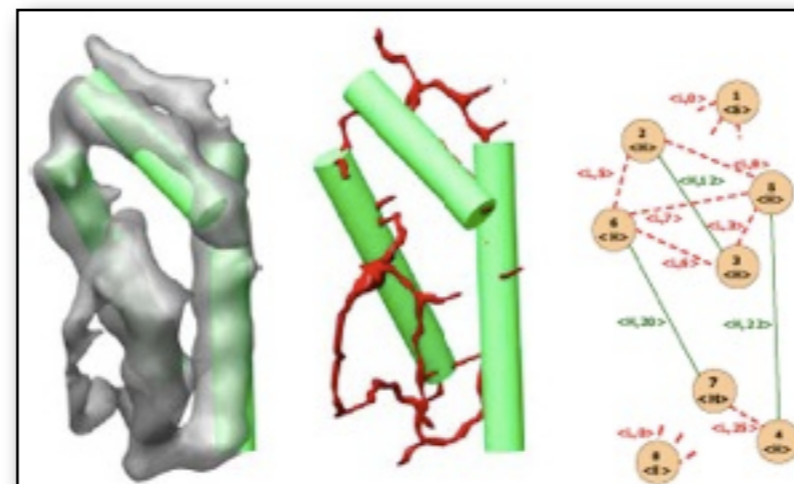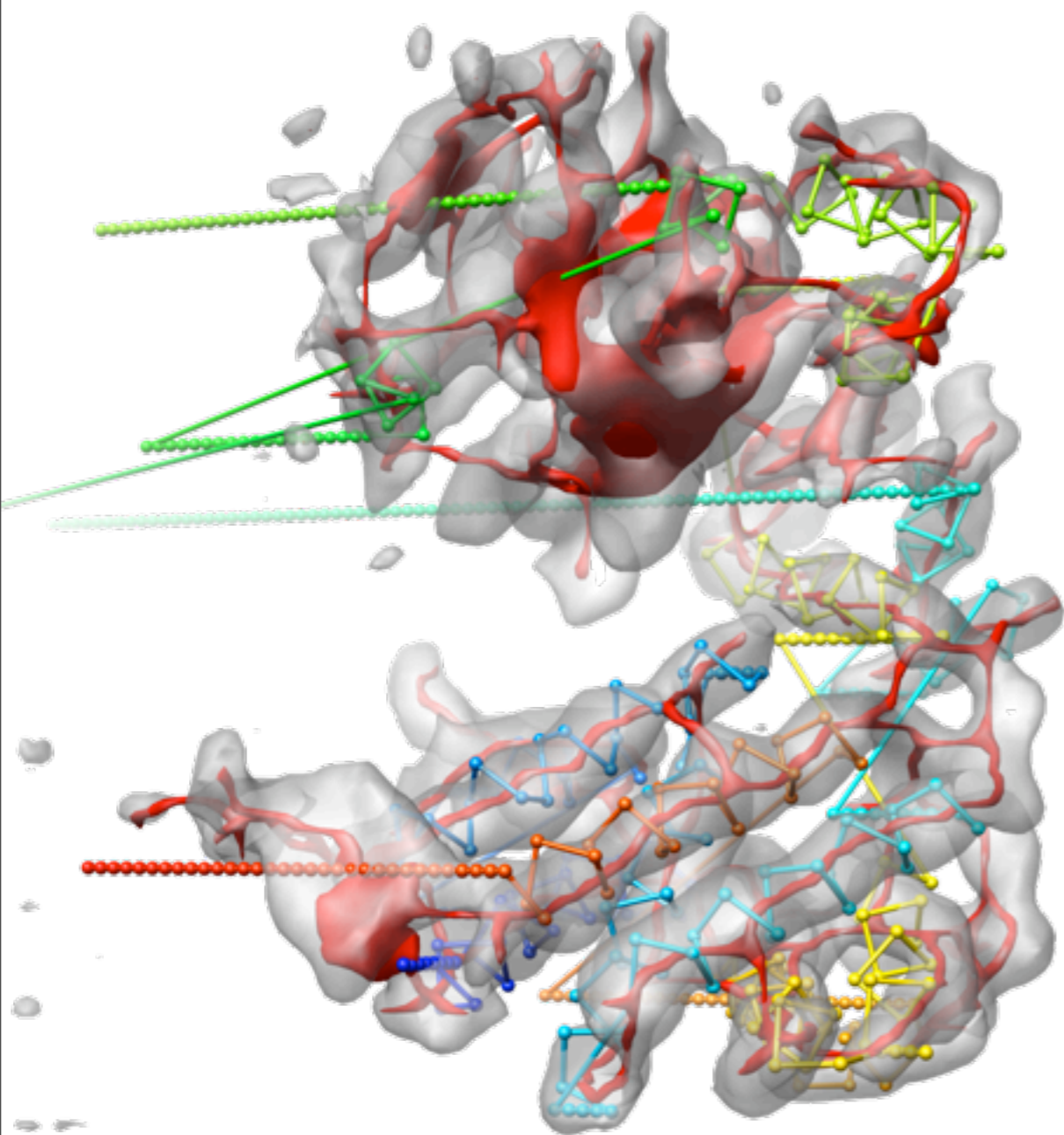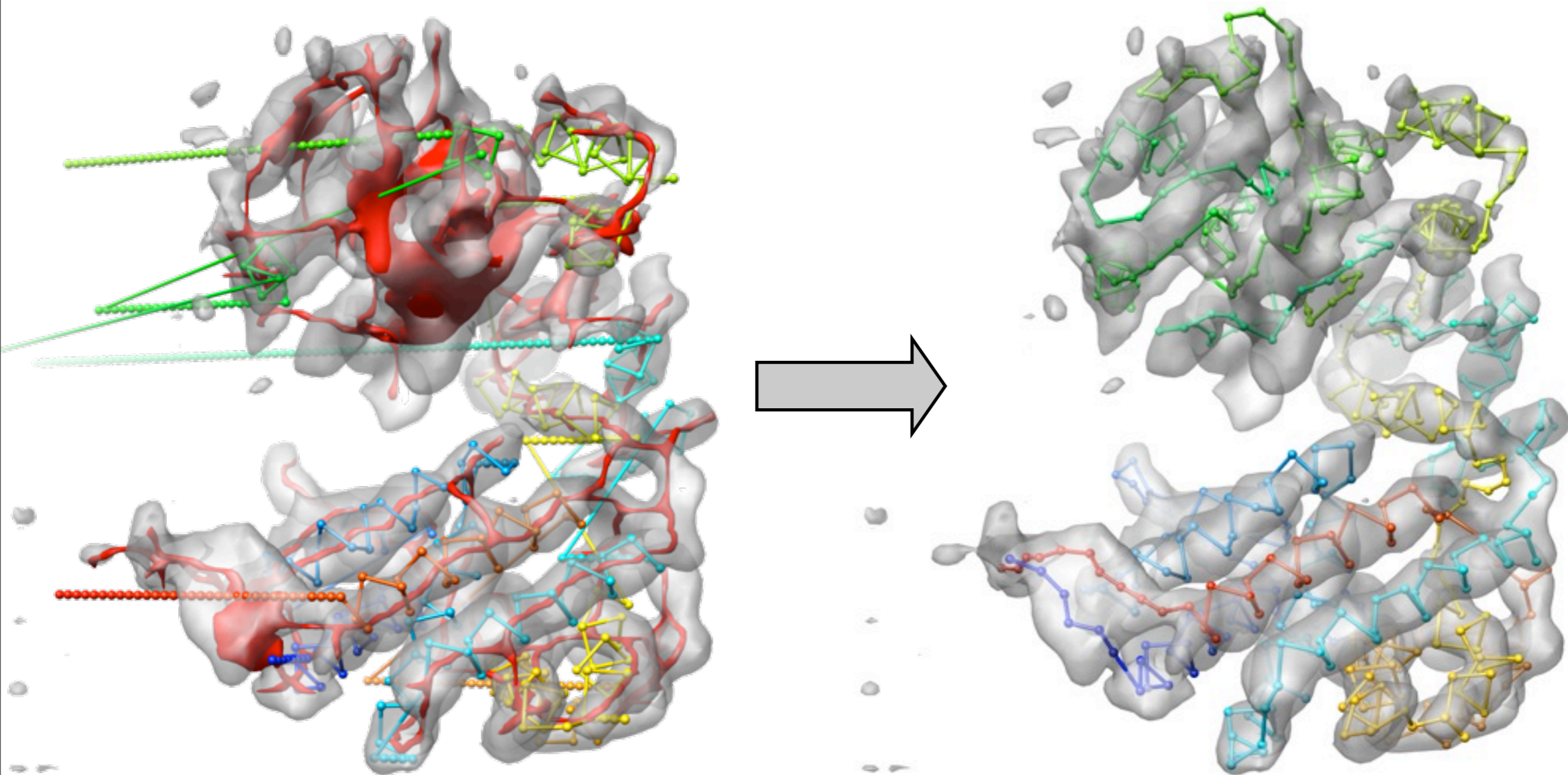- Can utilize user constraints

sequence graph

structure graph

correspondence
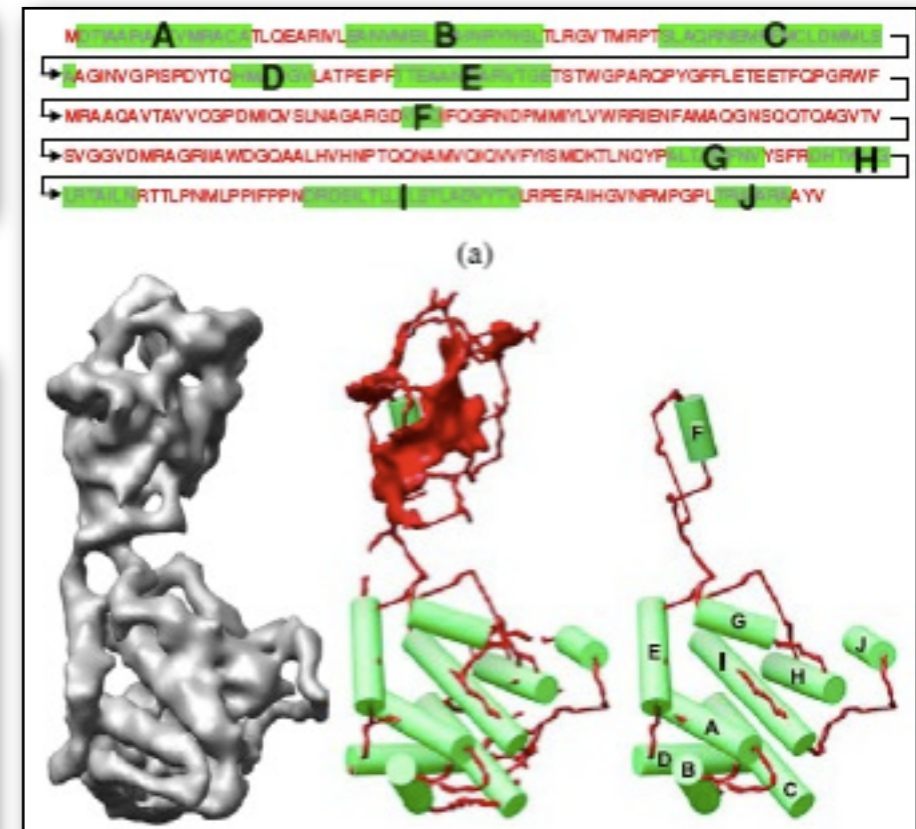
Abeysinghe, S., Ju, T., Baker, M.L., Chiu, W. (2008) Shape Modeling and Matching in Identifying Protein Structure from Low-Resolution Images. Computer-AIDED Design.

TRACING THE BACKBONE

MODEL BUILDING

- Optimize atom placement in density at/or near skeleton
- Optimize distances
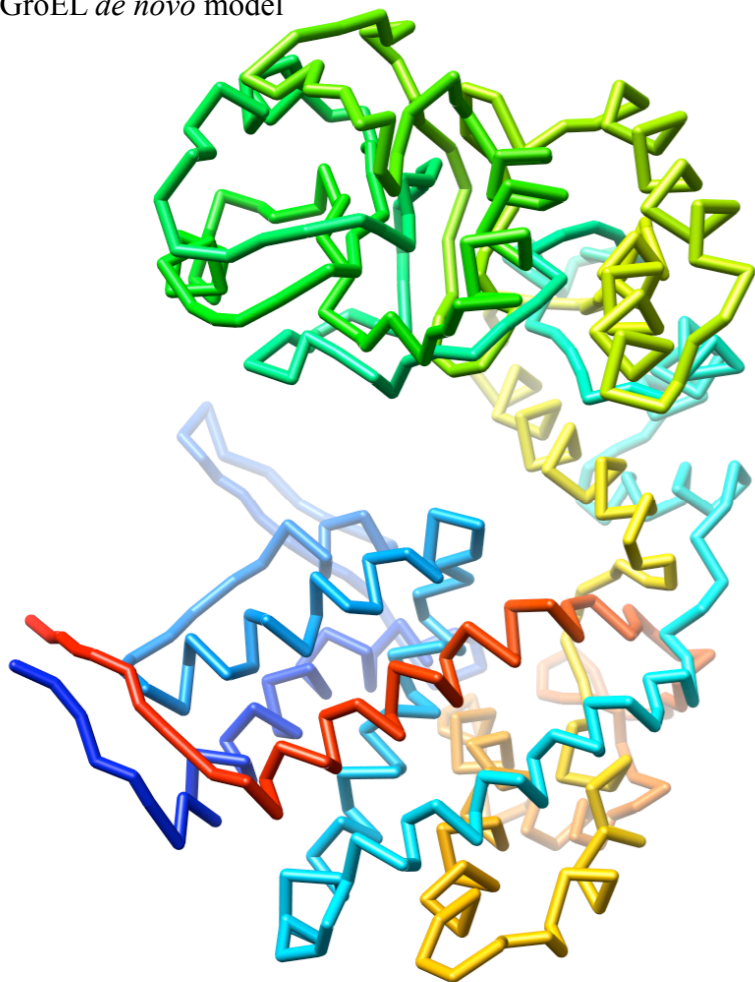- Maintain secondary structure
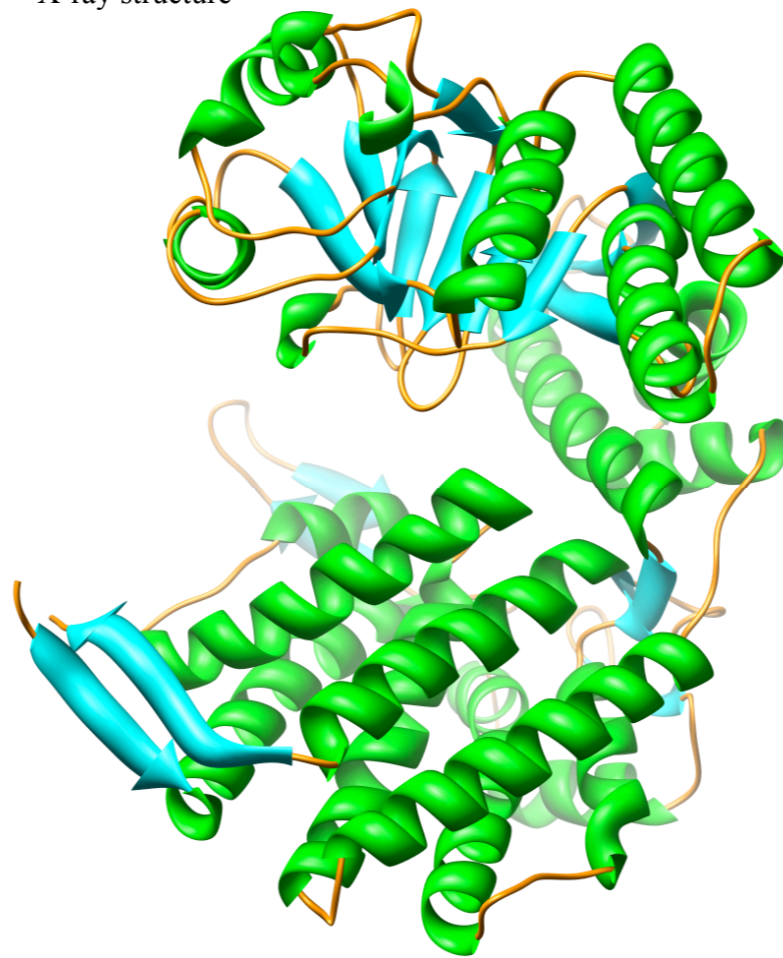- Minimize clashes
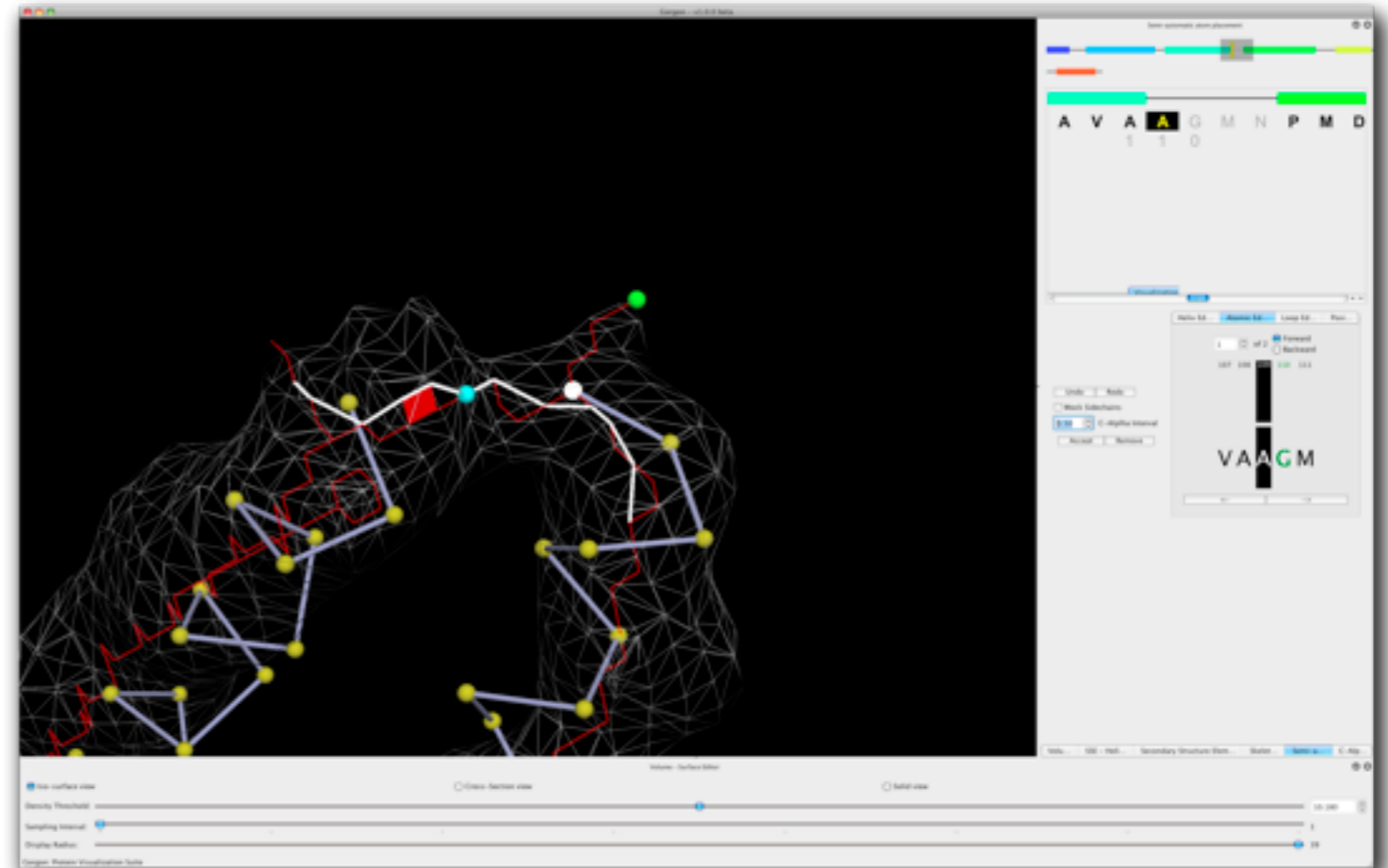
GroEL Cα BACKBONE MODEL
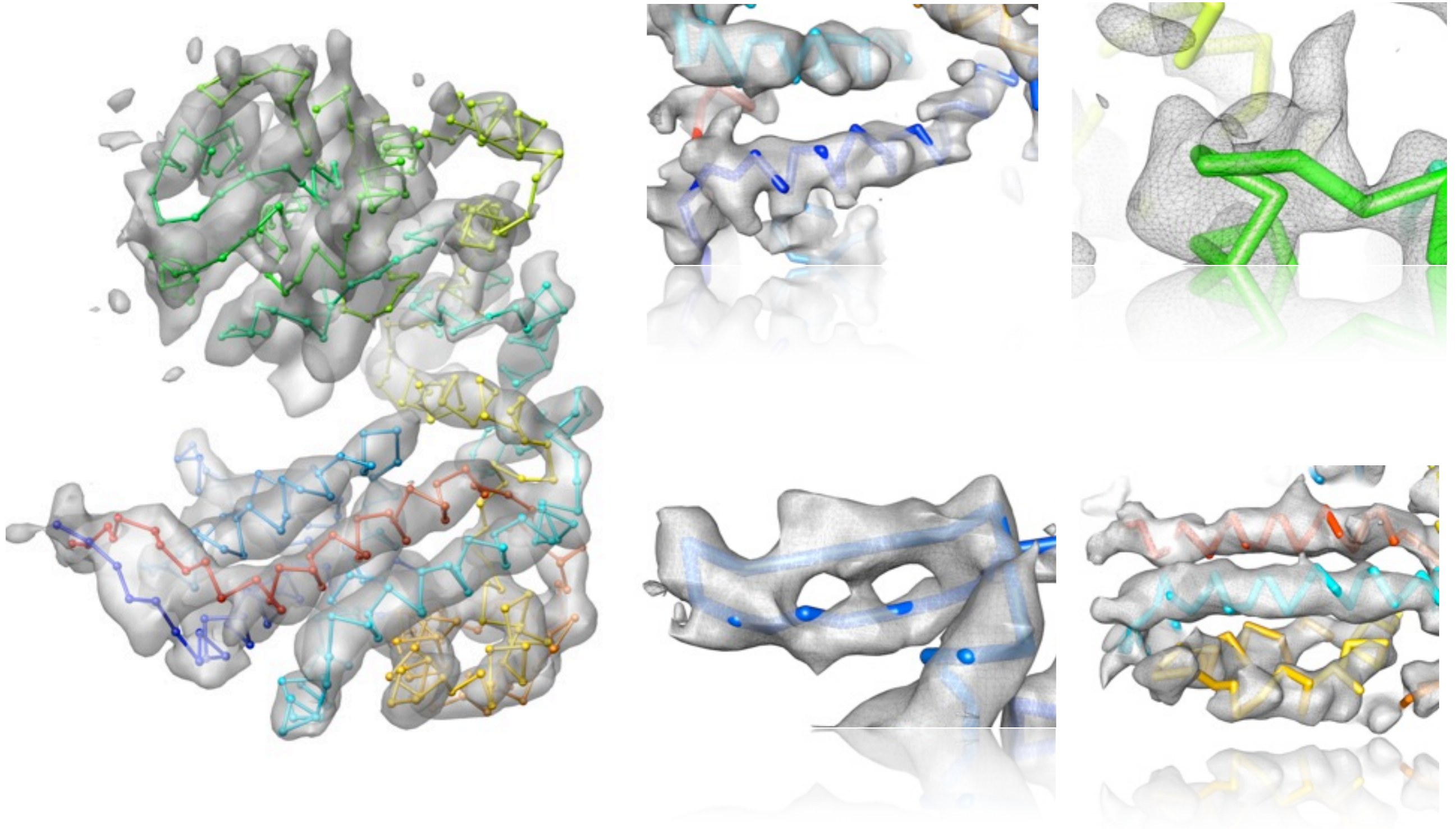
MODEL ACCURACY

GroEL *de novo* model

X-ray structure

- ~2.1Å nearest neighbor RMSD

  ➡ Cα positioning errors due to ambiguous density

- ~4.2Å position specific RMSD

  ➡ SSEHunter lengths

  ➡ Secondary structure prediction errors
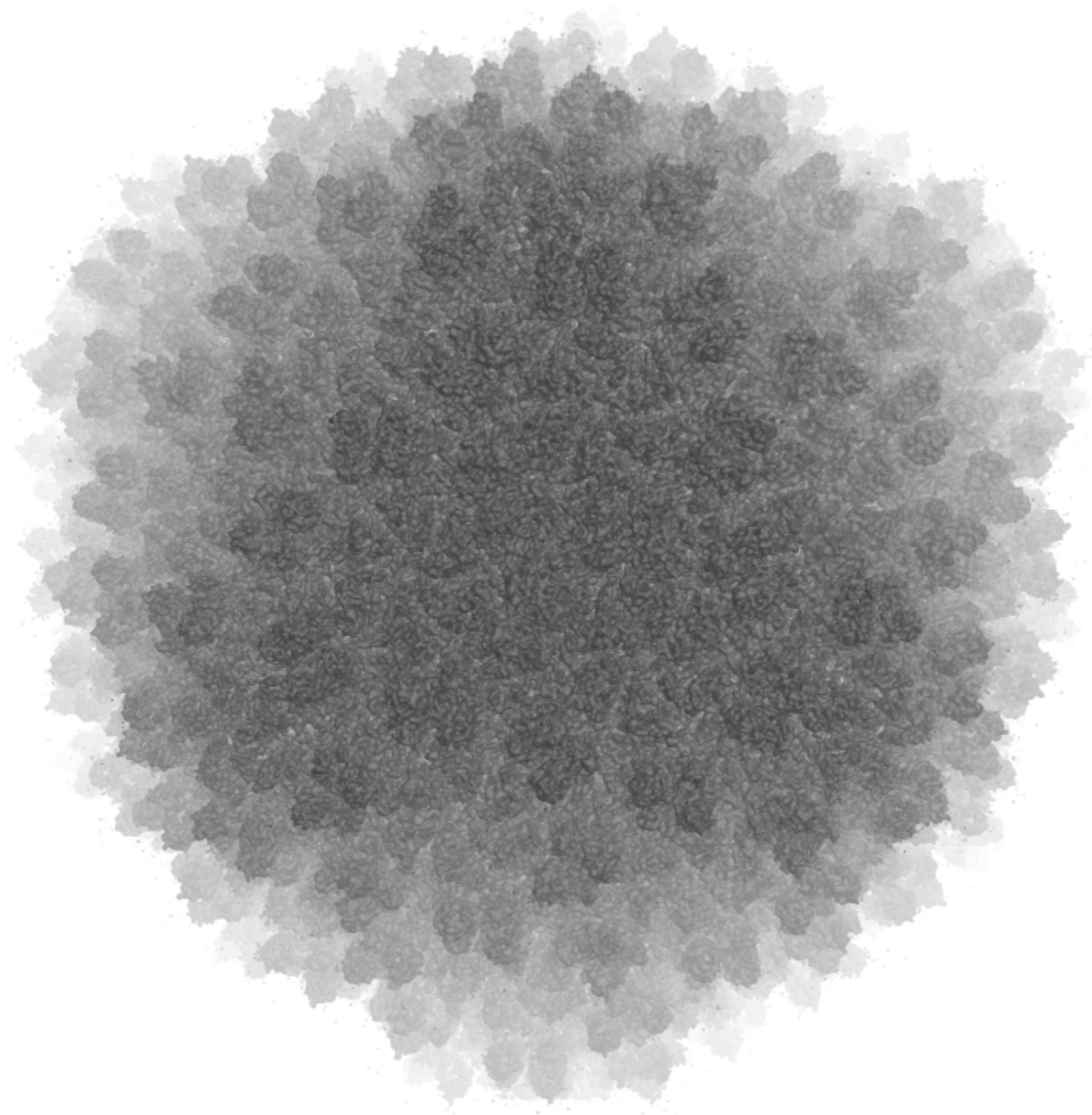
  ➡ Poorly resolved density

MODEL REFINEMENT WITH ROSETTA

Low-resolution, cryo-EM based energy function in Rosetta

4.2 Å map in the equatorial domain of GroEL
- All atom model with native side chain packing
- Improved RMSD of Cα trace (3.6 to 3.4 Å)
- Improve assignments
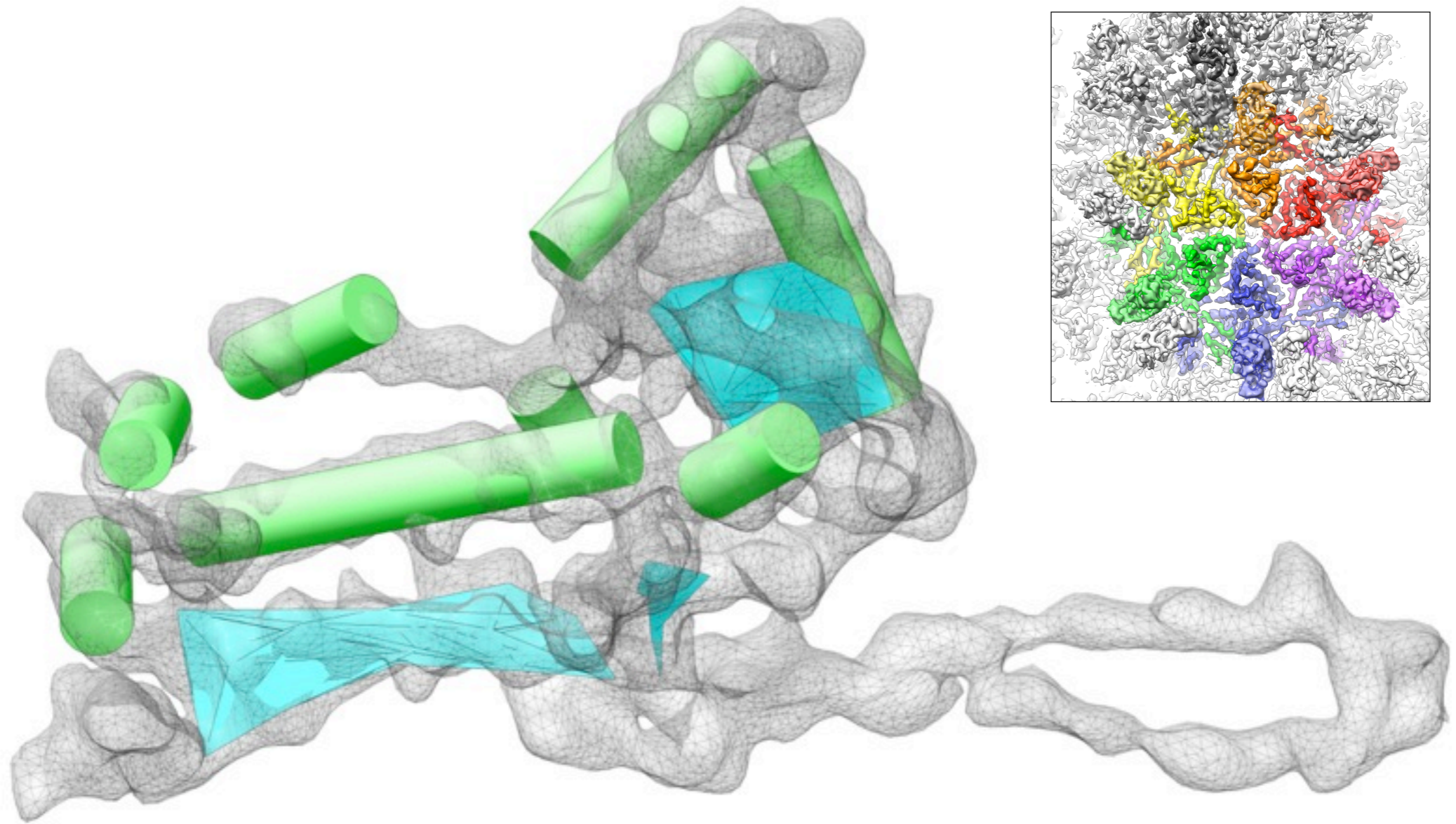 - Cα RMSD in helices reduced to 2.23Å from 3.41Å
  - Correct β-strand pairings
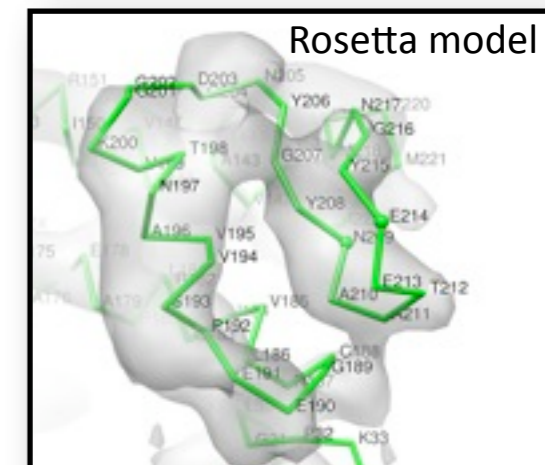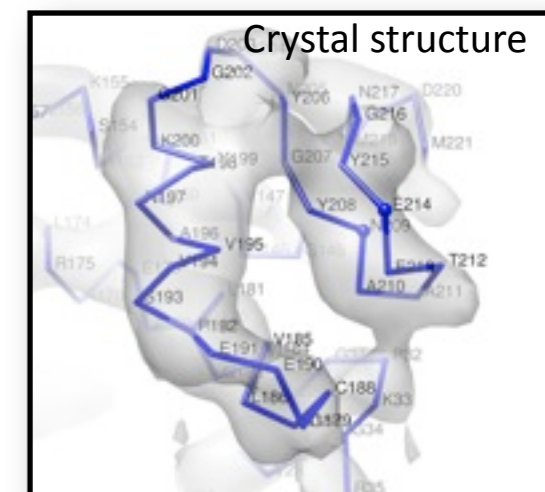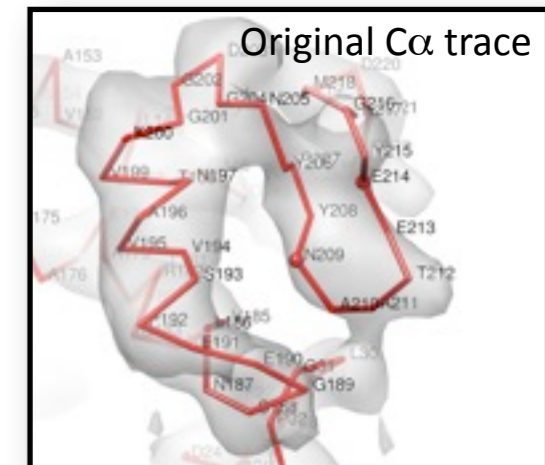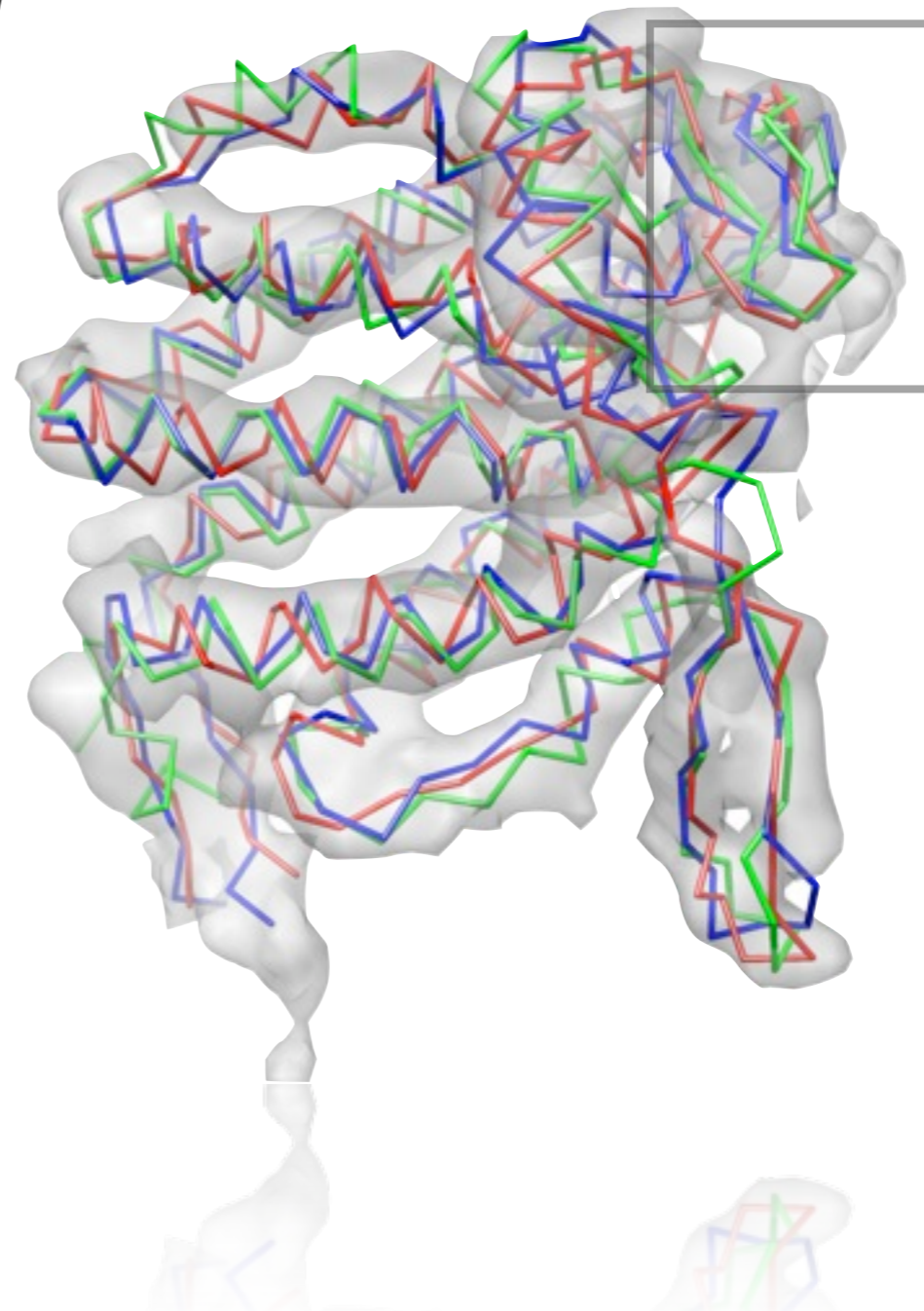
DiMaio, F, Tyka, M.D., Baker, M.L., Chiu, W., Baker, D. (2009) Refinement of Protein Structures into Low-Resolution Density Maps using Rosetta. JMB
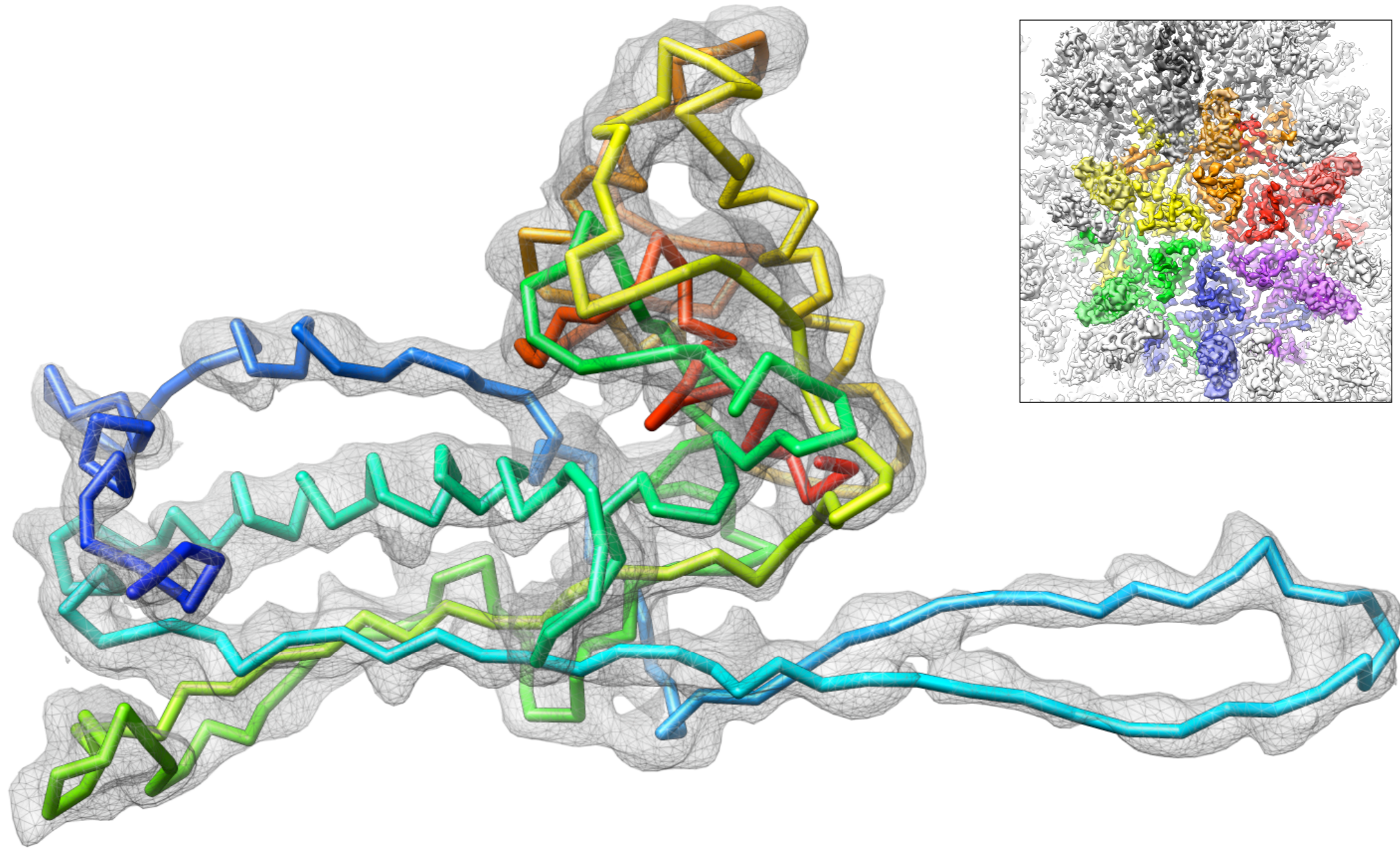
Original Cα trace

Crystal structure

Rosetta model

BACTERIOPHAGE ε15

4.5Å resolution

GP7

Jiang, W.*, Baker, M.L.*, Jakana, J Weigele, P.R., King, J., Chiu W. (2008) Backbone Structure of the Infectious ε15 Virus Capsid Revealed by Cryo-EM. Nature (451), p 1130-1135. *contributed equally

GP7

Jiang, W.*, Baker, M.L.*, Jakana, J Weigele, P.R., King, J., Chiu W. (2008) Backbone Structure of the Infectious ε15 Virus Capsid Revealed by Cryo-EM. Nature (451), p 1130-1135. *contributed equally
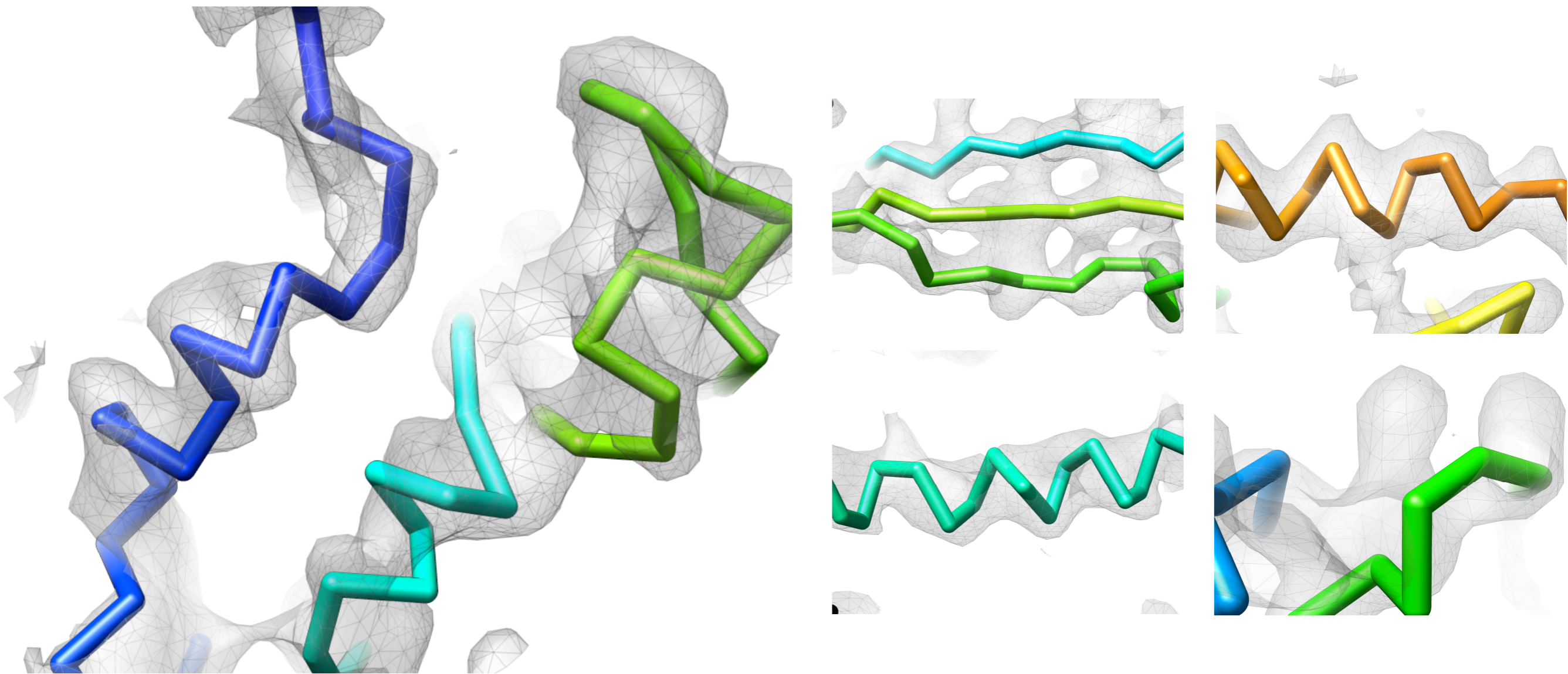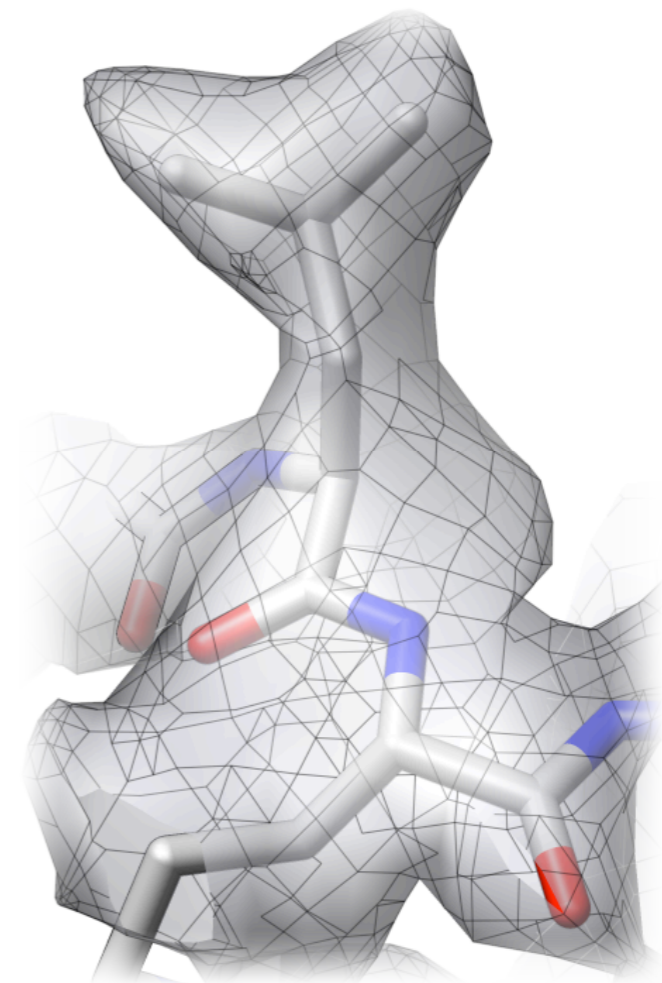
GP7 MODEL
FEATURES

SIDECHAIN RESOLUTION (~3.5+Å)

Features

- Strand separation

- Sidechain like protrusions

  - "lollipop" aromatics

  - extended shapes

- Atomic models

- Good stereochemistry

- Optimized subunit interfaces

Limitations

- Noisy density maps

- Difficult to segment properly

- Individual atoms not resolved

BUILDING AN ATOMIC MODEL

0. Cα model (Gorgon)

1. Full-backbone (SABBAC)

2. Model repair (Coot)

3. Rotamer assignment

6. Loop optimization

7. Rotamer and Ramachandran angle refinement

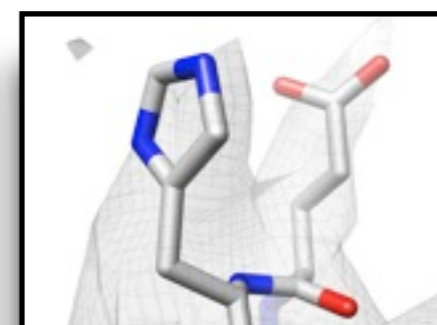5. Atom assignment in SSE

4. Clash resolution
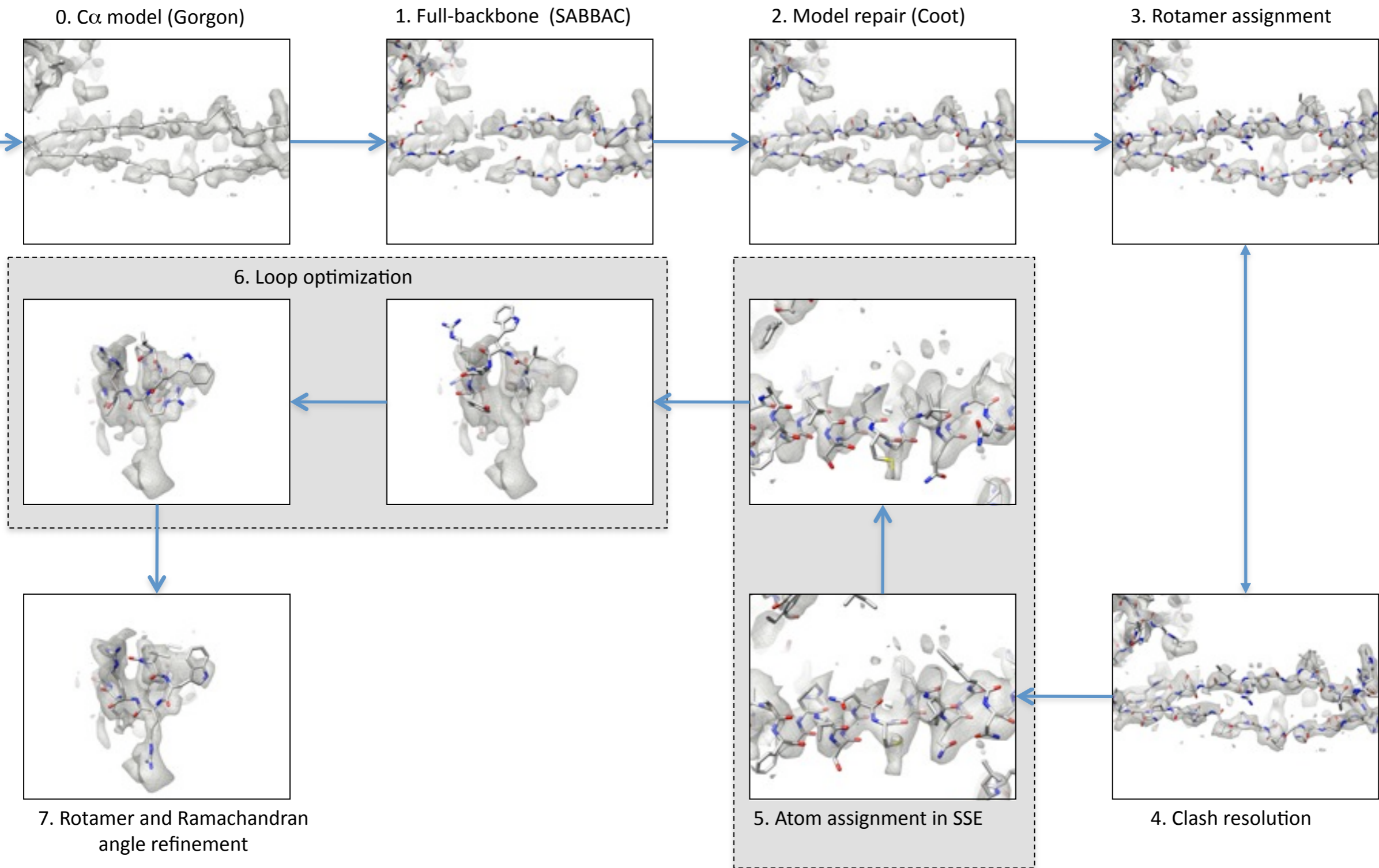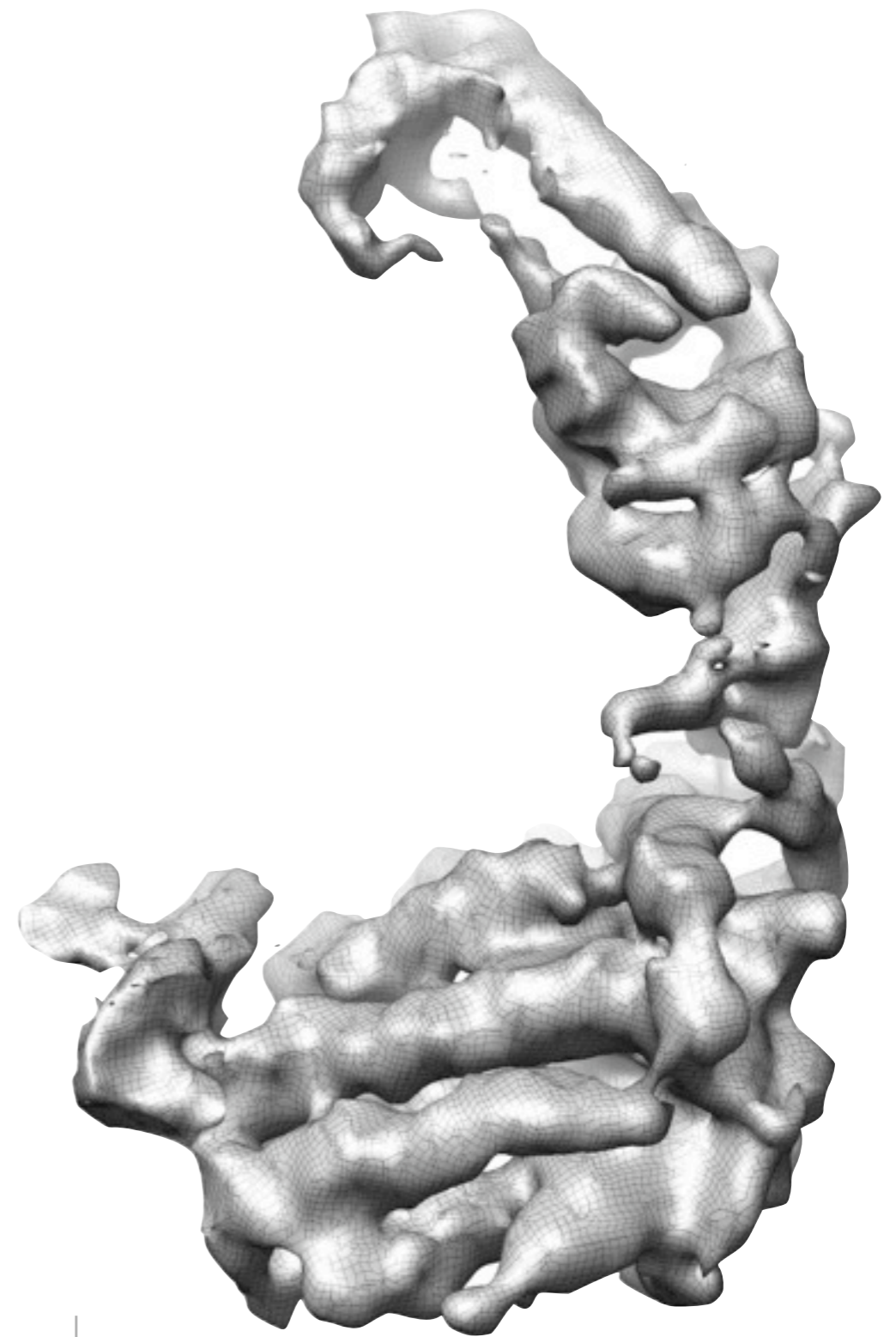
MM-CPN | 4.2Å resolution

MM-CPN ATOMIC MODEL

➡ residues 1-532
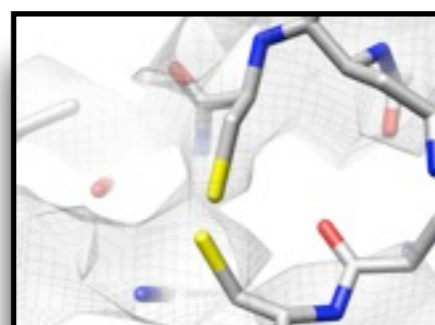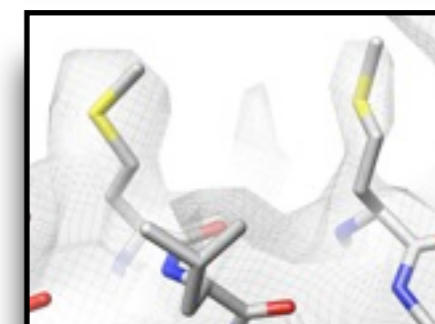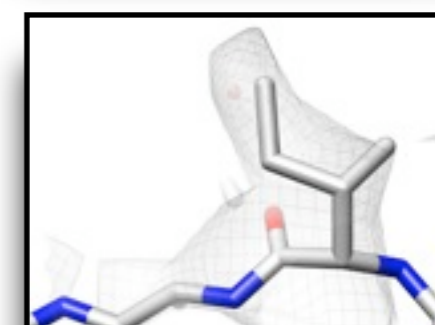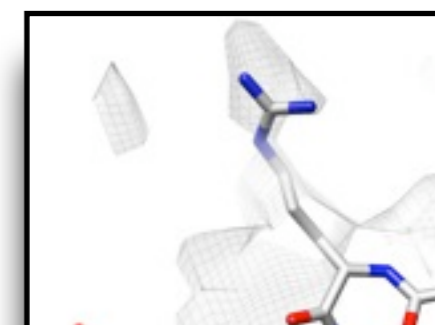➡ 85% of residues are have favorable Phi-Psi angles, >99% acceptable
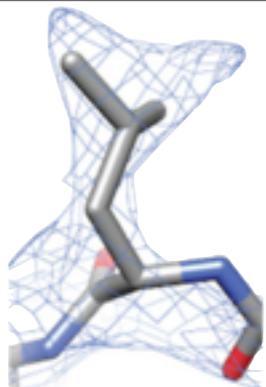➡ >70% visible sidechain densities
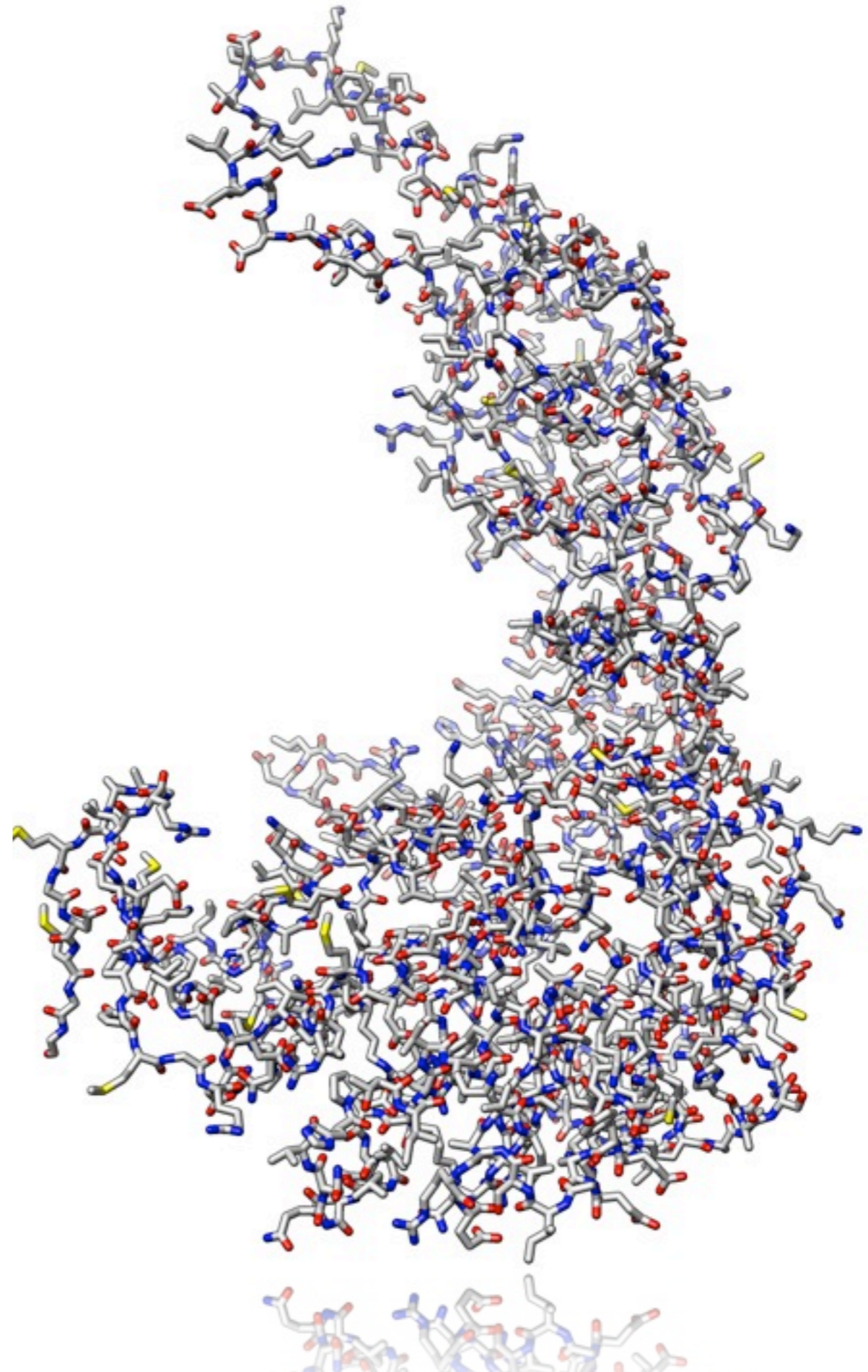
MM-CPN: SIDECHAINS
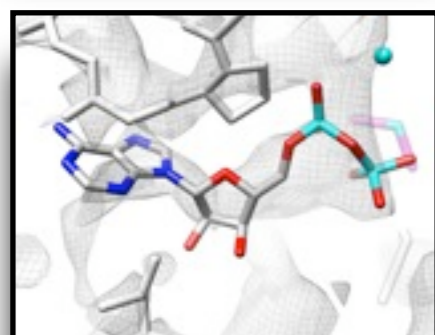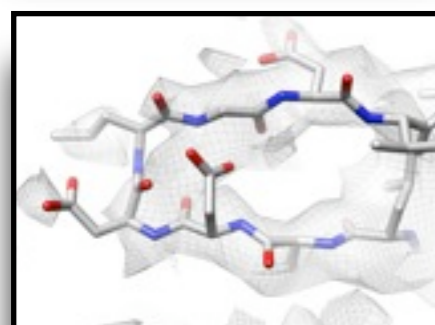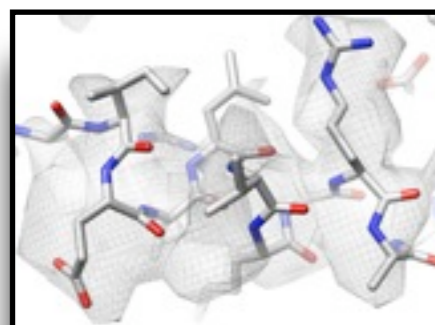
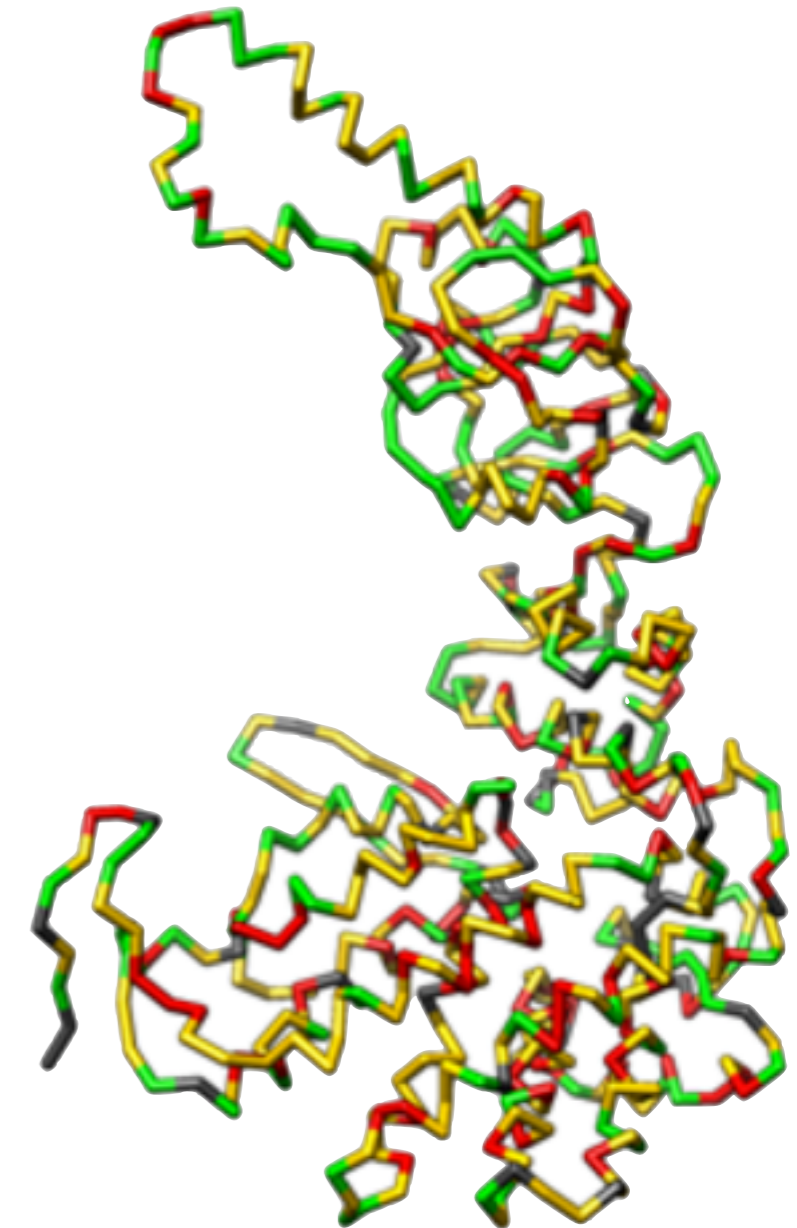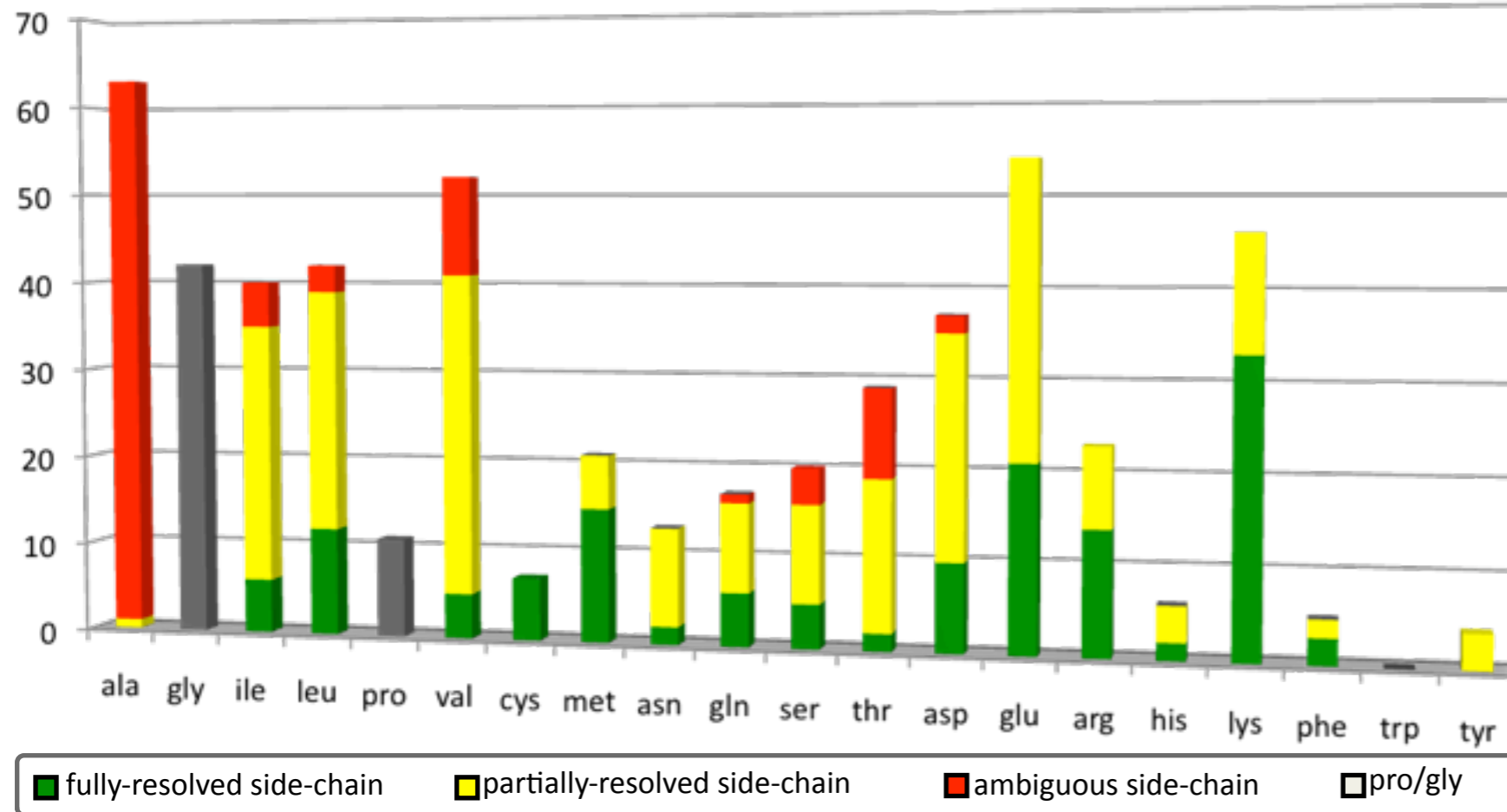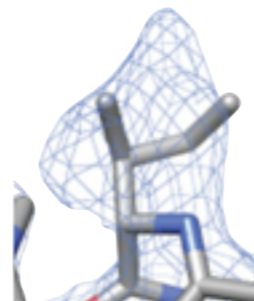fully-resolved side-chain
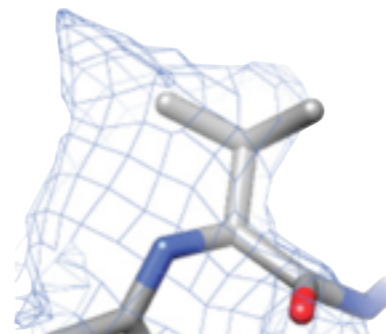partially-resolved side-chain
ambiguous side-chain
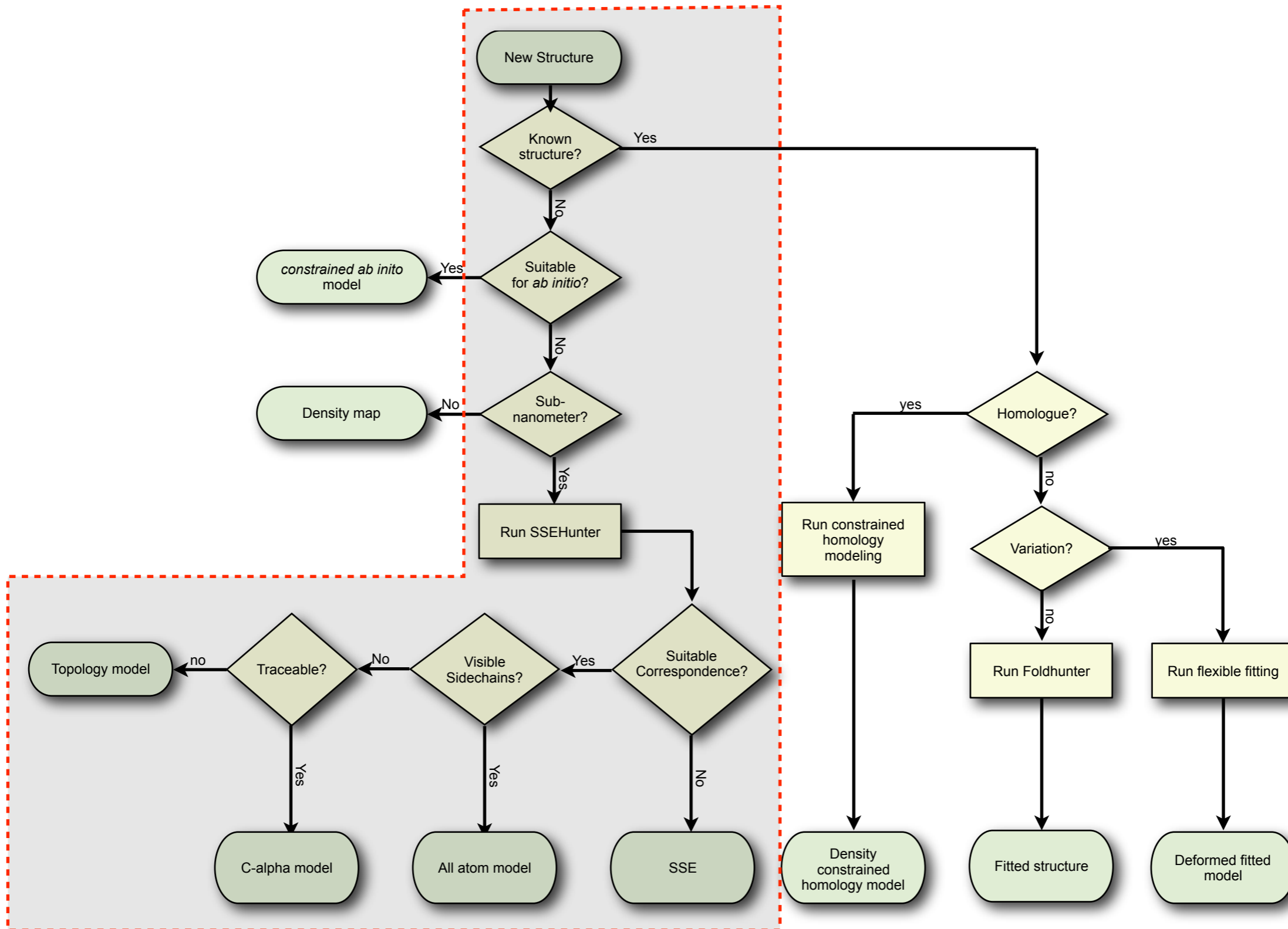pro/gly
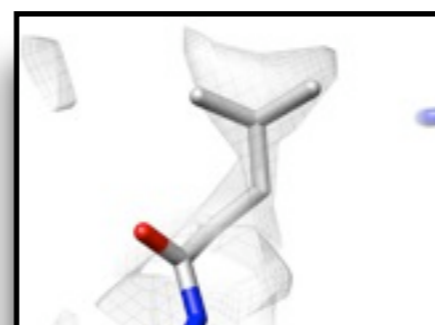
fully-resolved
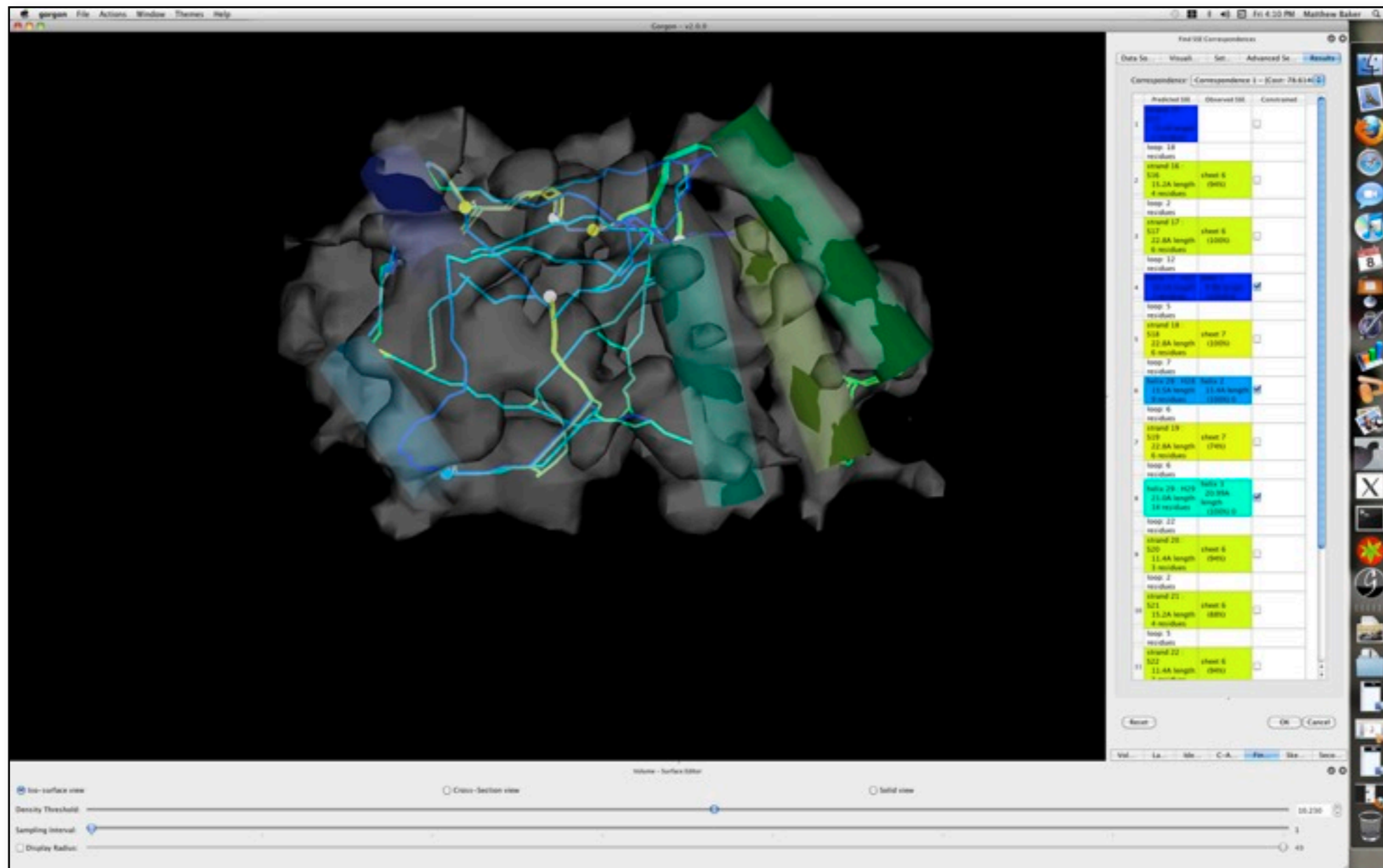143aa, 27%

partially-resolved
238aa, 45%

ambiguous
109aa, 20%

DE NOVO MODELING IN CRYO-EM

GORGON

Interactive molecular modeling system for subnanometer to near-atomic resolution structures

http://www.cs.wustl.edu/~ssa1/gorgon/

GORGON FEATURES

Version 2.0.0

- Greyscale and interactive density skeletonization
- SSE identification and building with SSEHunter/SSEBuilder
- SSE correspondence with helices and sheets
- Semi-automated atom placement
- Plugins
- Sessions

- Cross platform (Windows 32/64 bit, Linus 32/64 bit, OS X 10.5+)
- On-line videos and tutorials with sample data
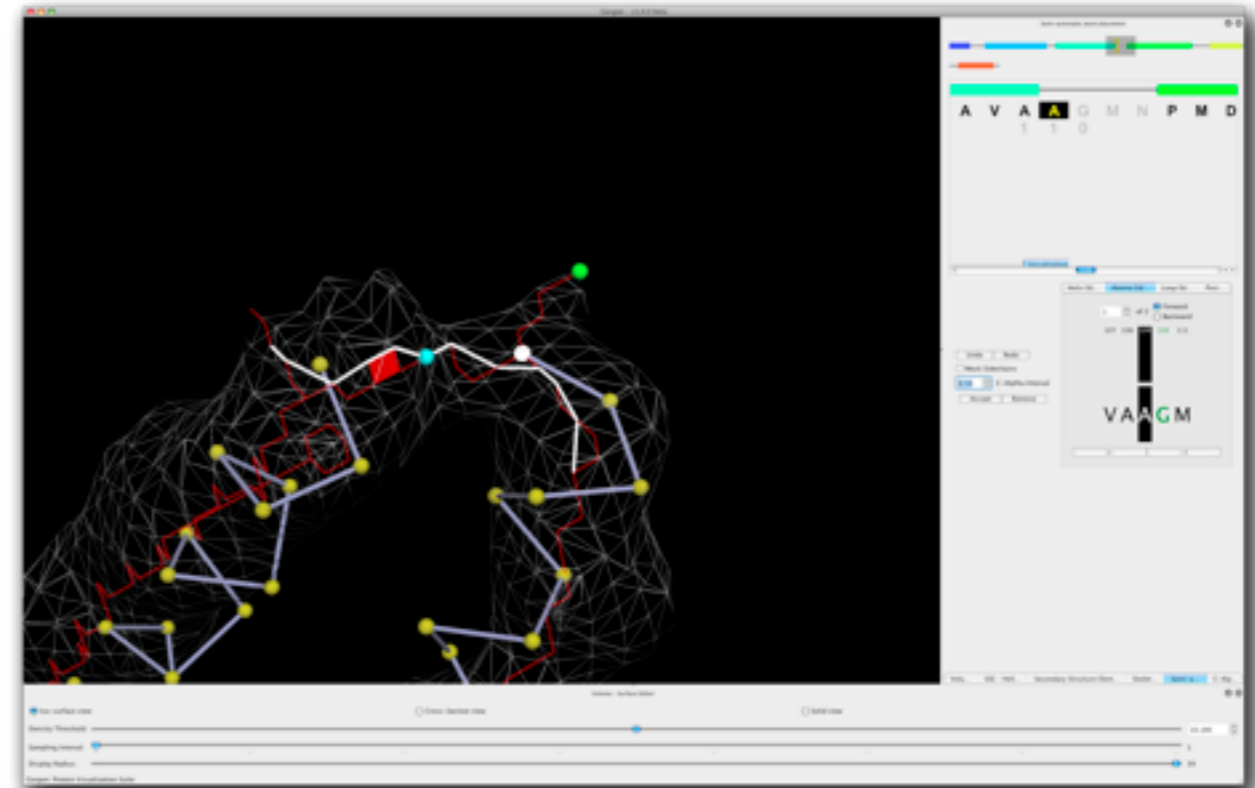
Interactive, semi-automated model building
Auto-build of SSE
SSE fitting
Interactive, sketching of loops
Manual editing with local fitting

GORGON: FEATURES IN DEVELOPMENT

- Flexible fitting

- Extensions for EMAN and Rosetta

- Pathwalking

- Enhanced documentation and tutorials

- Improve interface

- Foldhunter integration

- Improved sheet and strand modeling

- Cα backbone conversion to full model

- Rotamer refinement

ACKNOWLEDGEMENTS

NCMI
National Center for Macromolecular Imaging

BCM
Baylor College of Medicine