

# PISA the Crystallographic Time Machine

*... or a story about perceptions, expectations, naivety,  
macromolecular complexes and their complexity in  
bioinformatics and crystallography*

Eugene Krissinel

CCP4, STFC Research Complex at Harwell  
Didcot, United Kingdom

[eugene.krissinel@stfc.ac.uk](mailto:eugene.krissinel@stfc.ac.uk)

*E. Krissinel and K. Henrick (2007) J. Mol. Biol. 372, 774-797*

*E. Krissinel (2010) J. Comp. Chem. 31, 133-143*

Biozentrum der Universität Basel, December 2, 2013, Basel, Switzerland

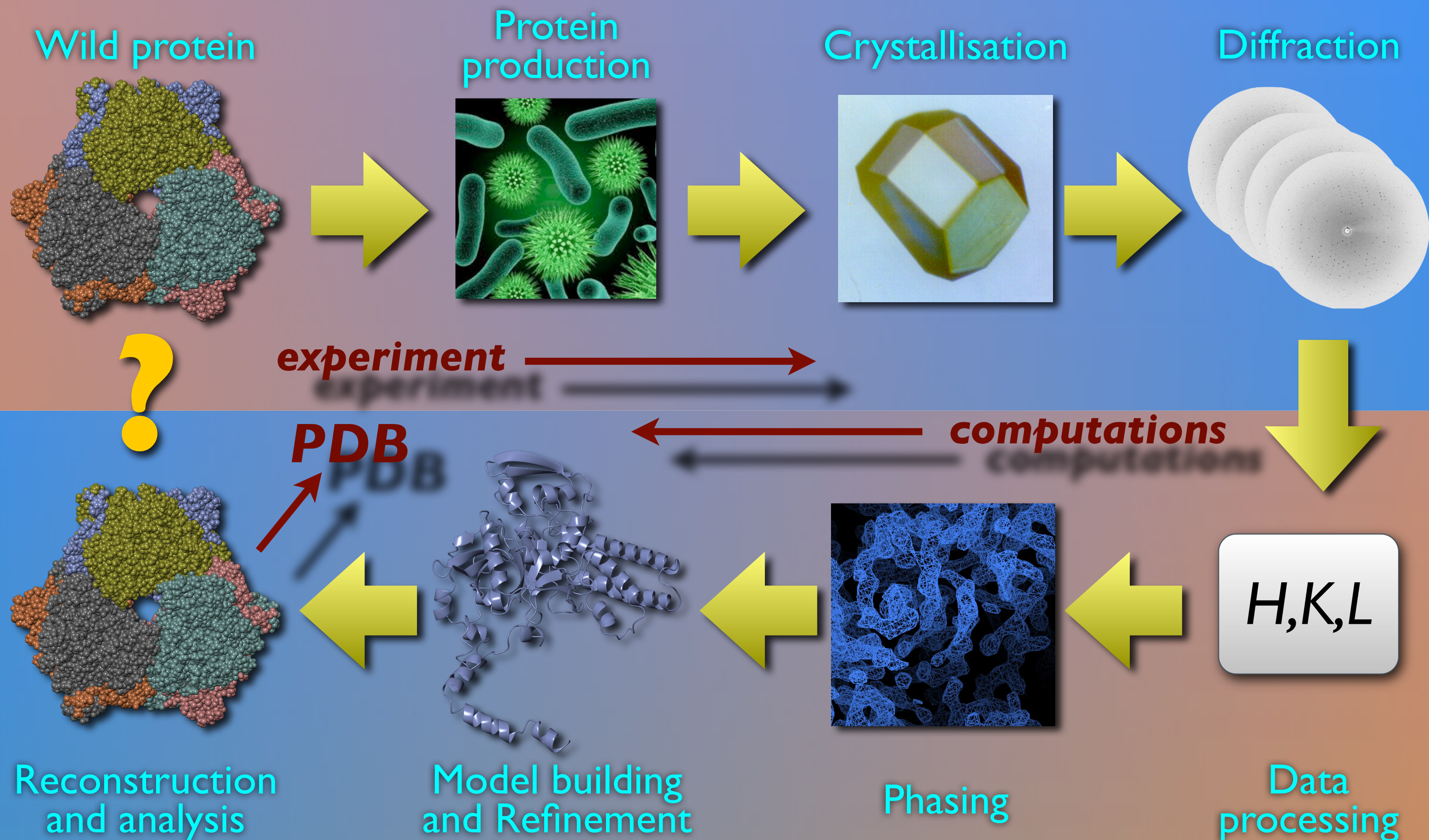


# Today's Roadmap

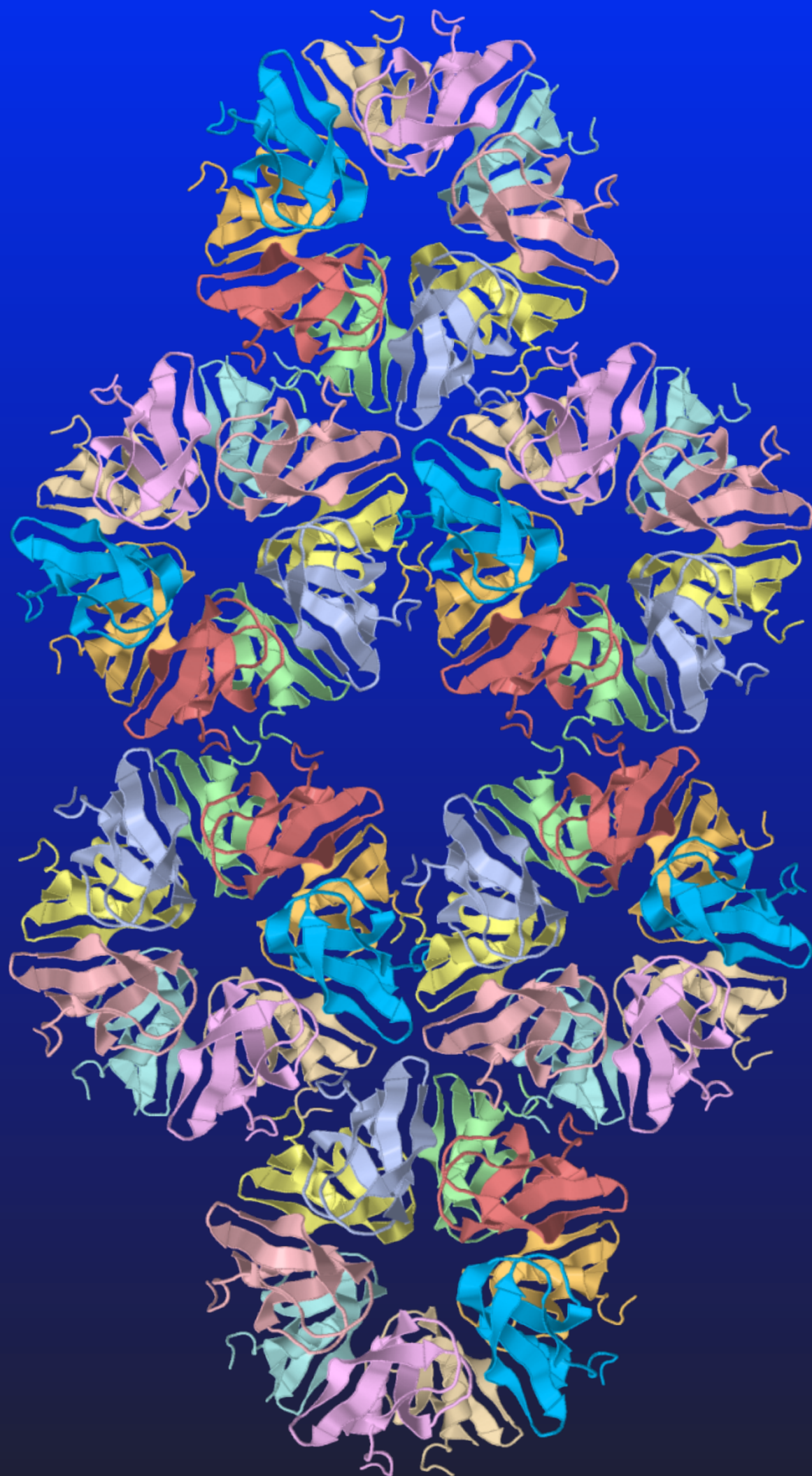
1. *Reconstruction problem in crystallography*
2. *What's wrong with bioinformatics?*
3. *What is PISA and why I call it a Time Machine?*
4. *Where PISA **does not** work?*
5. *Why PISA works at all? (**strange informatics**)*
6. *Where is the limit of possible? (**lying crystals**)*
7. *Beyond the limit of possible with bioinformatics (**as a suggestion**)*



# MX Schematic Loop

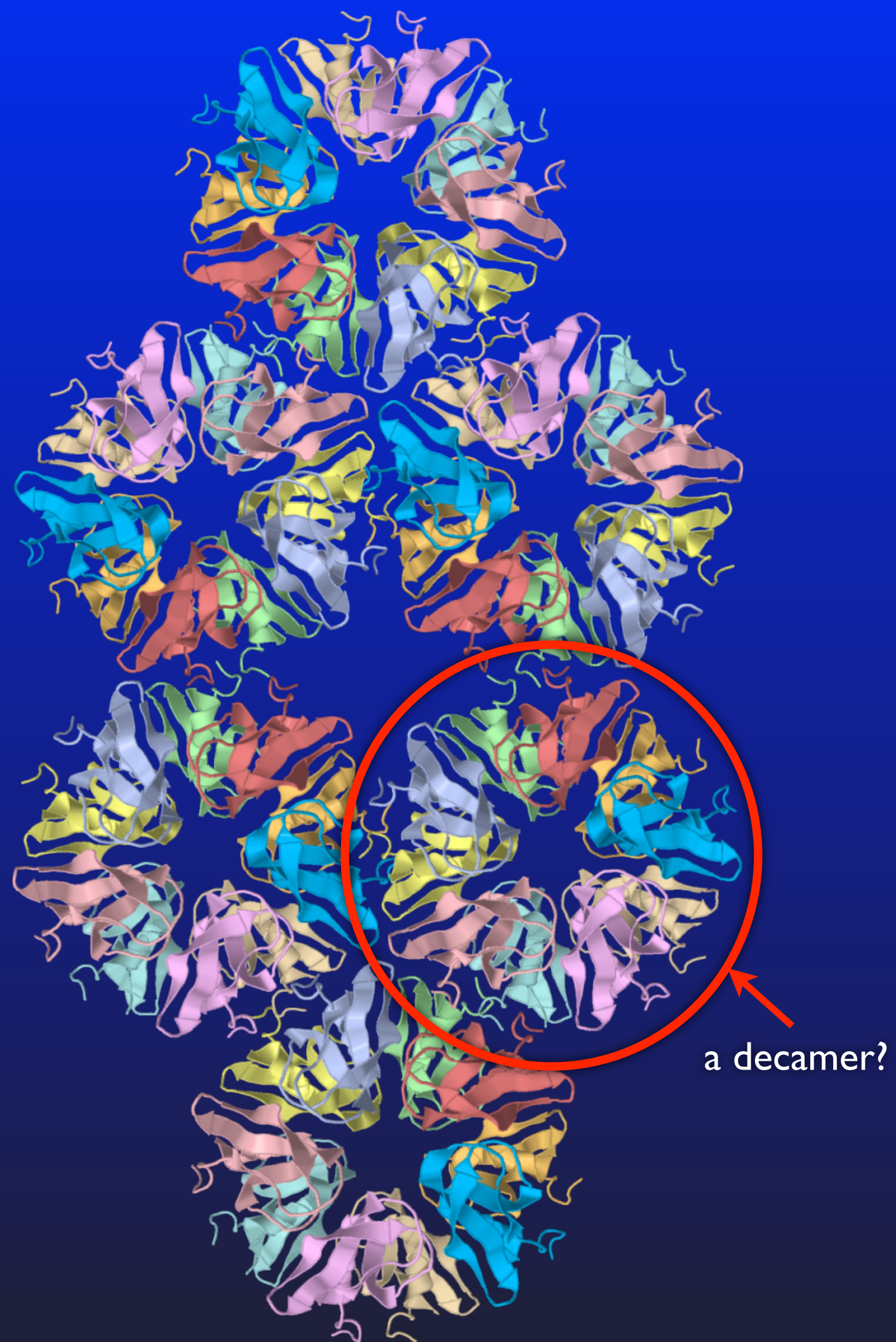




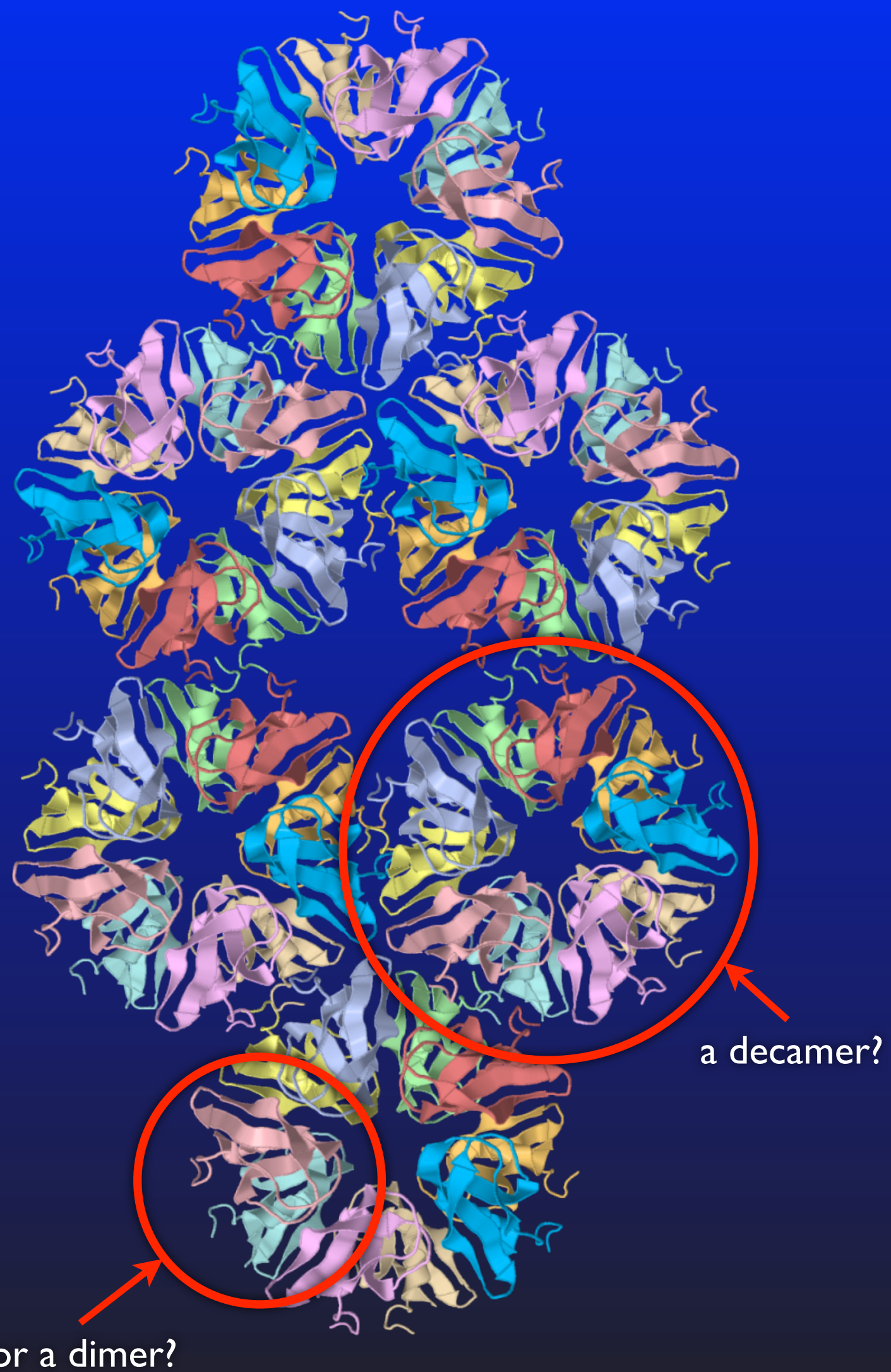


*Research Complex at Harwell*

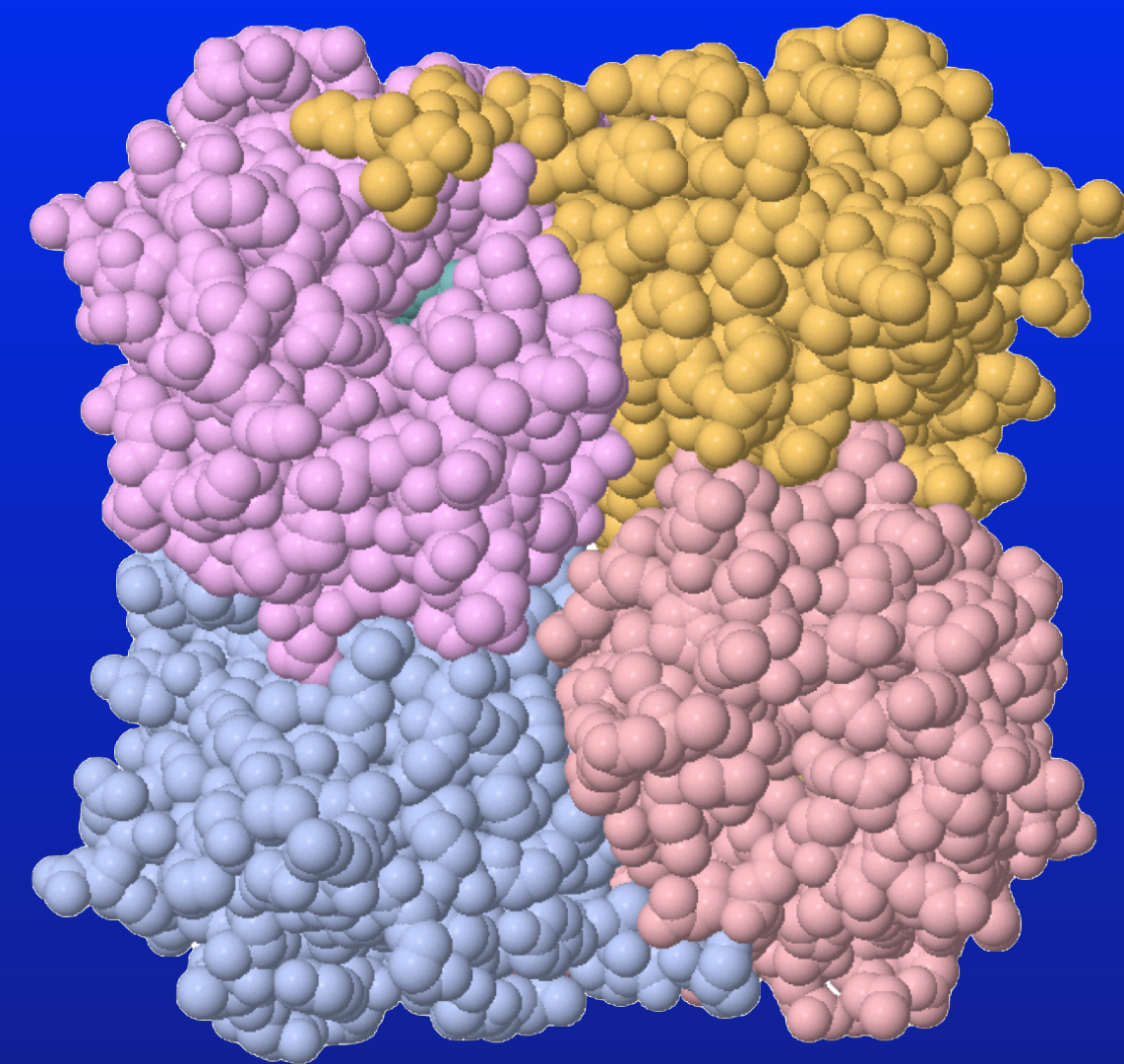
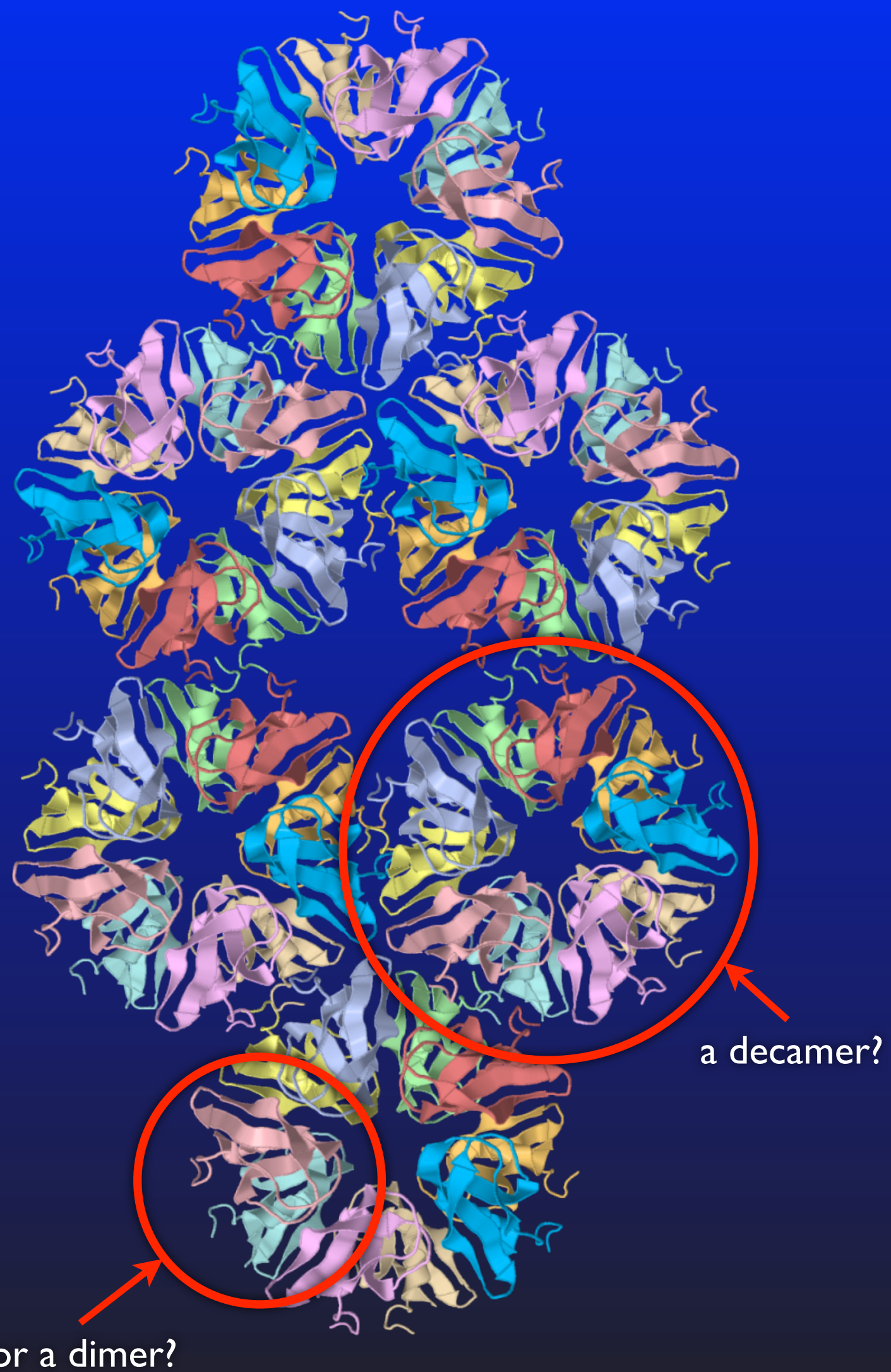








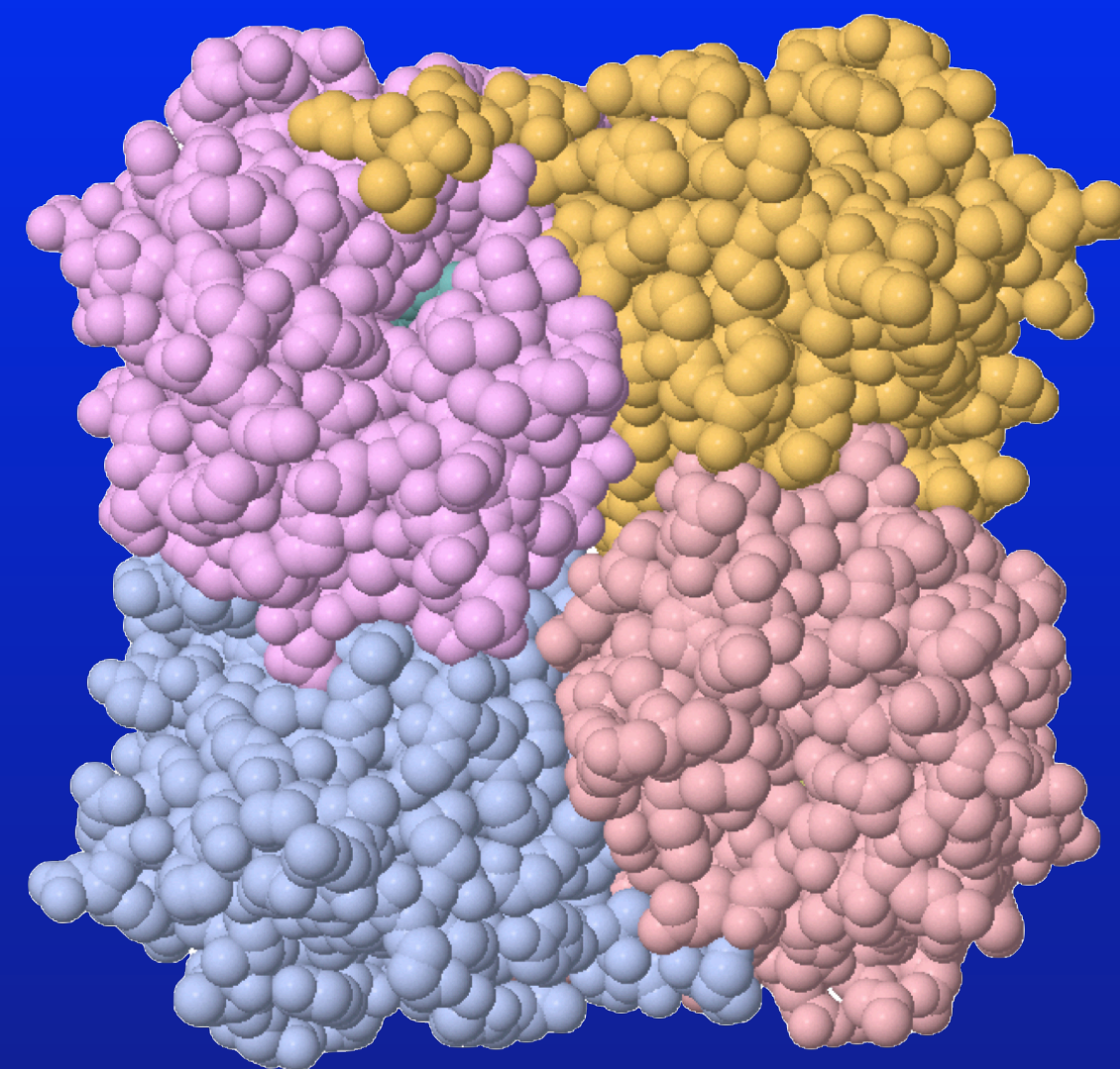
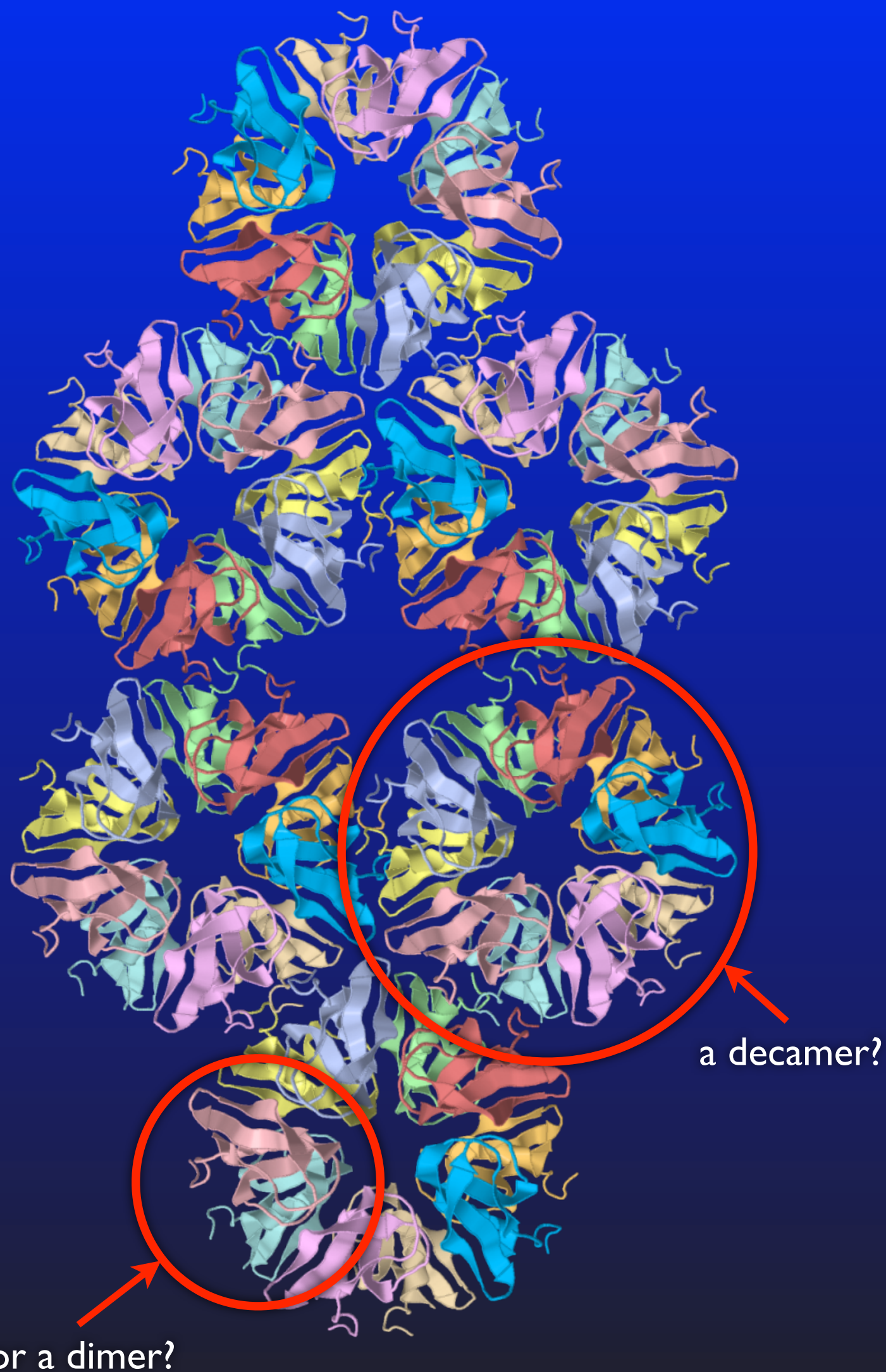




3bxc







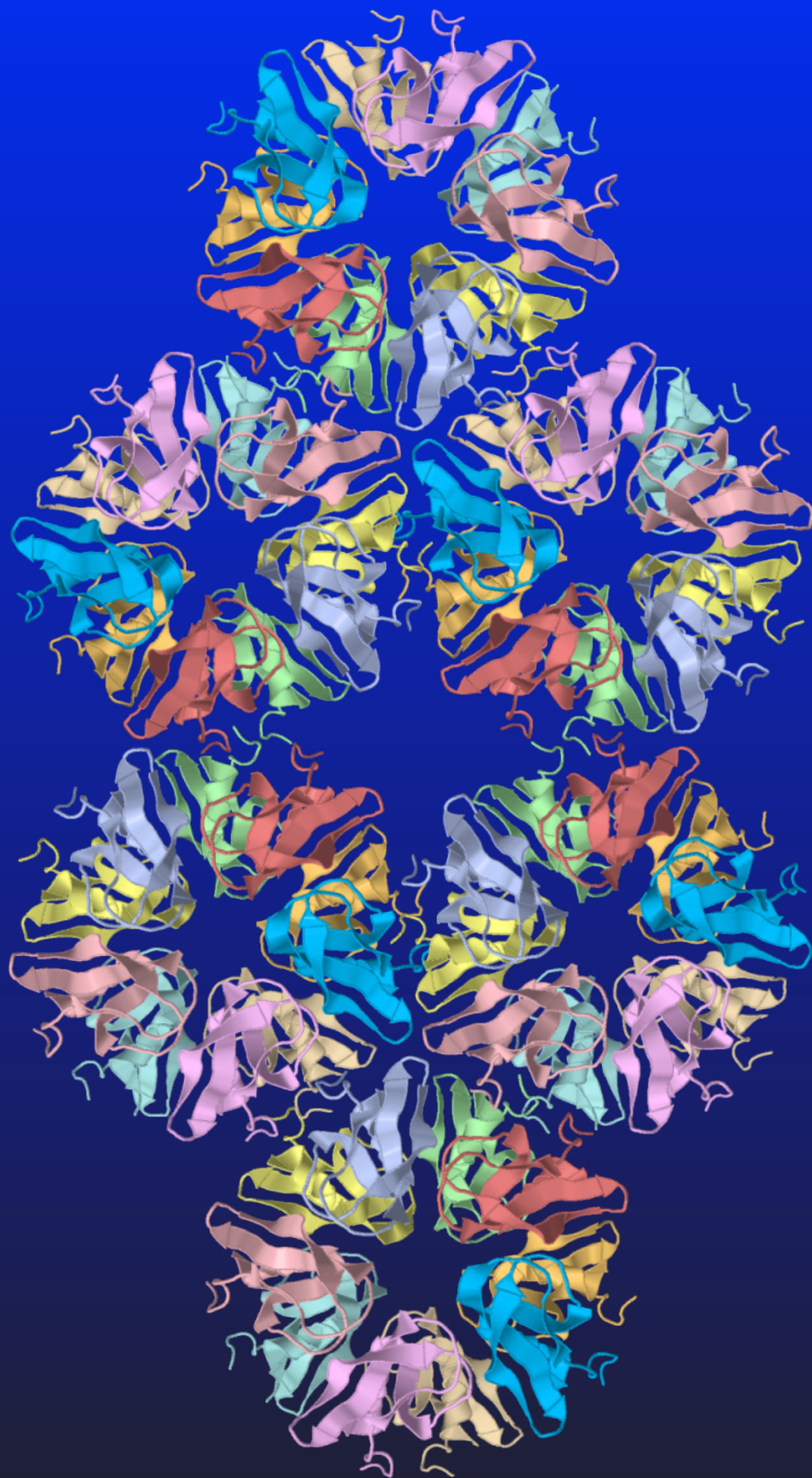
3bxc

*“Monomer Association—According to gel filtration data, mKate exists in solution in the monomeric state at concentration as high as 10 mg/ml (13). However, in the crystalline state, which corresponds to a much higher protein concentration, mKate adopts at all pH values tetrameric arrangement with 222 symmetry, typically seen in GFP-like proteins.”*

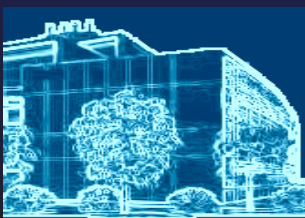
Pletnev et.al. (2008), J. Biol. Chem. 283, 28980-7





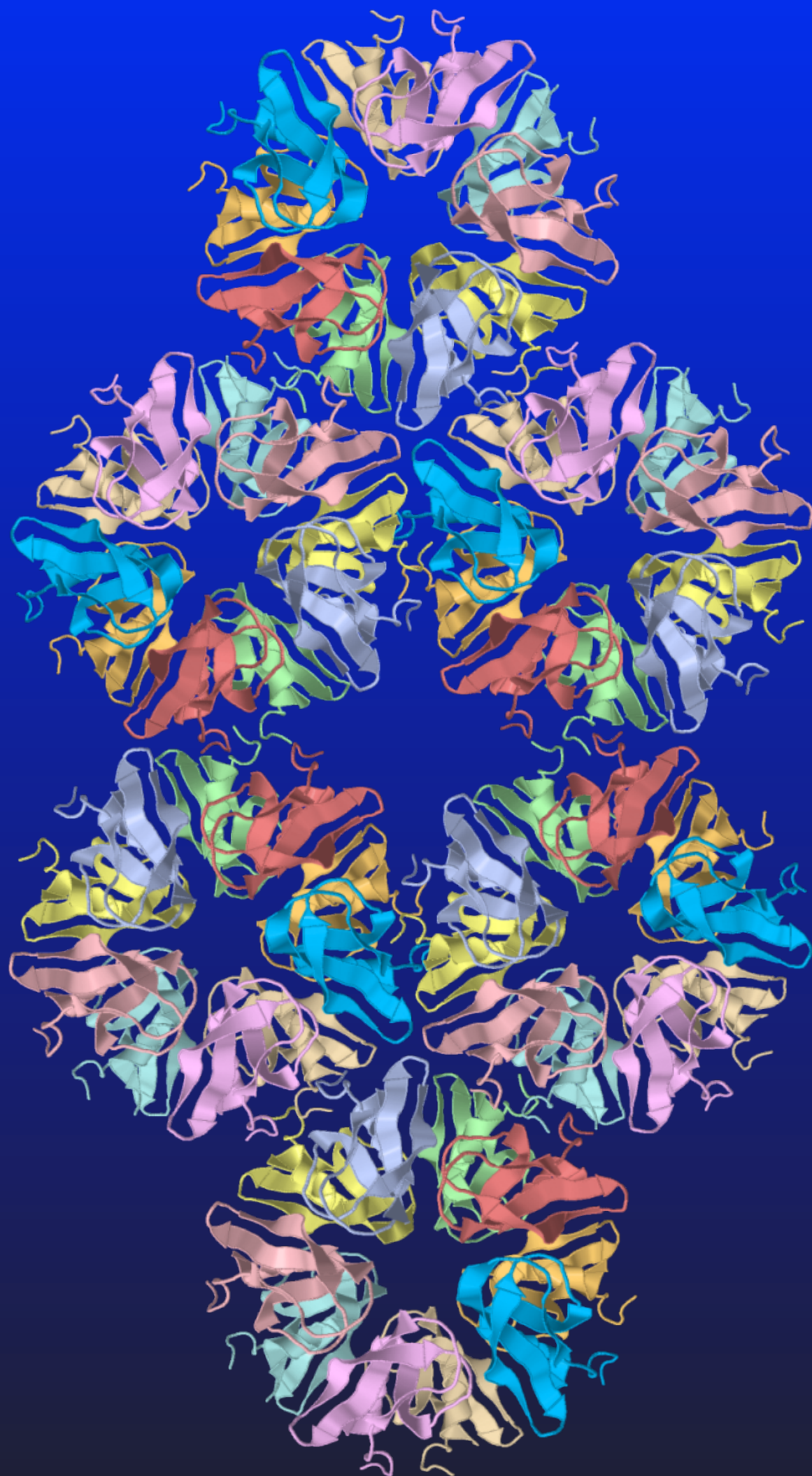


★ Why would we want to know the structure of a macromolecule?



Research Complex at Harwell



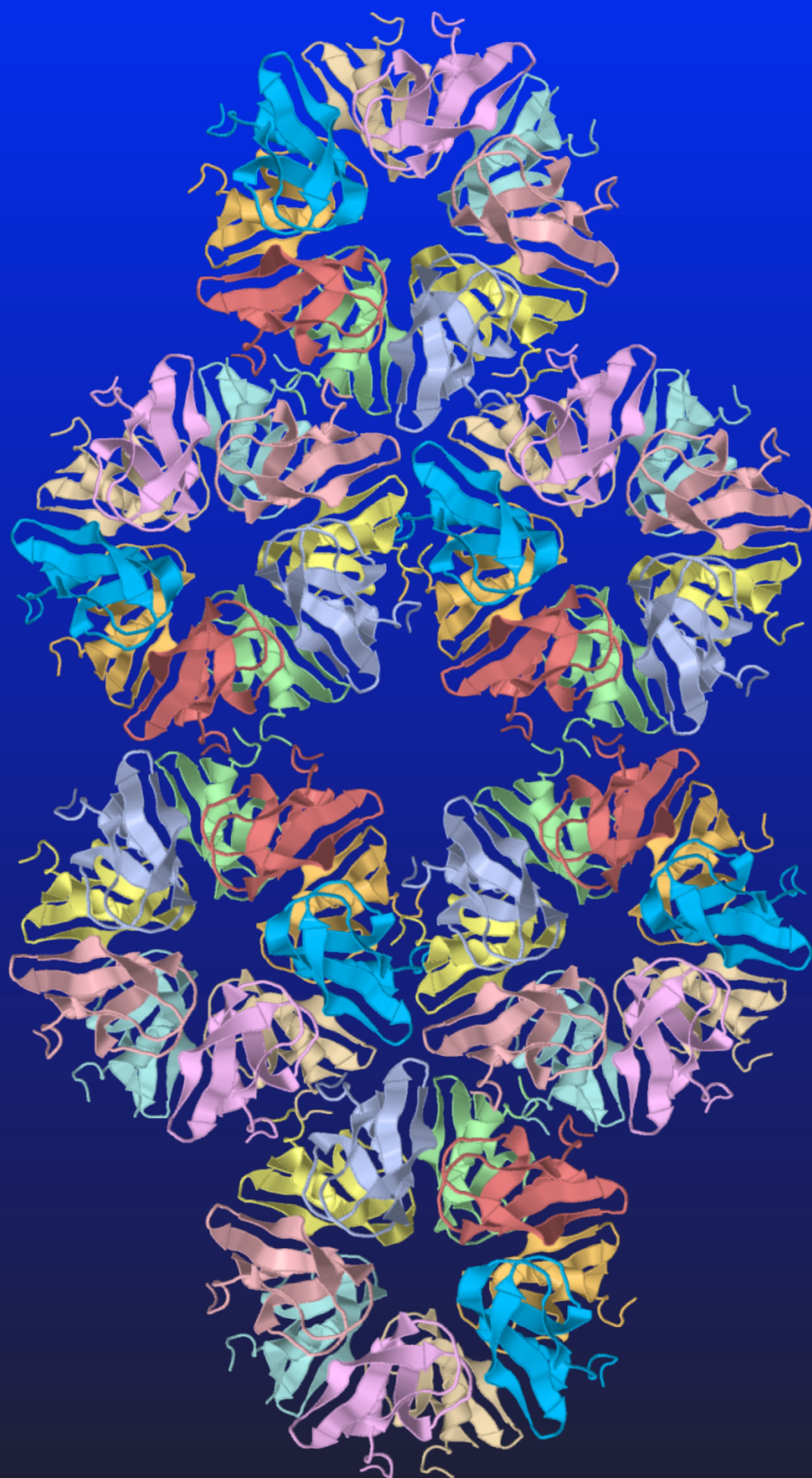


★ Why would we want to know the structure of a macromolecule?

- ➔ for many reasons, but probably firstly for finding out how it interacts with other molecules





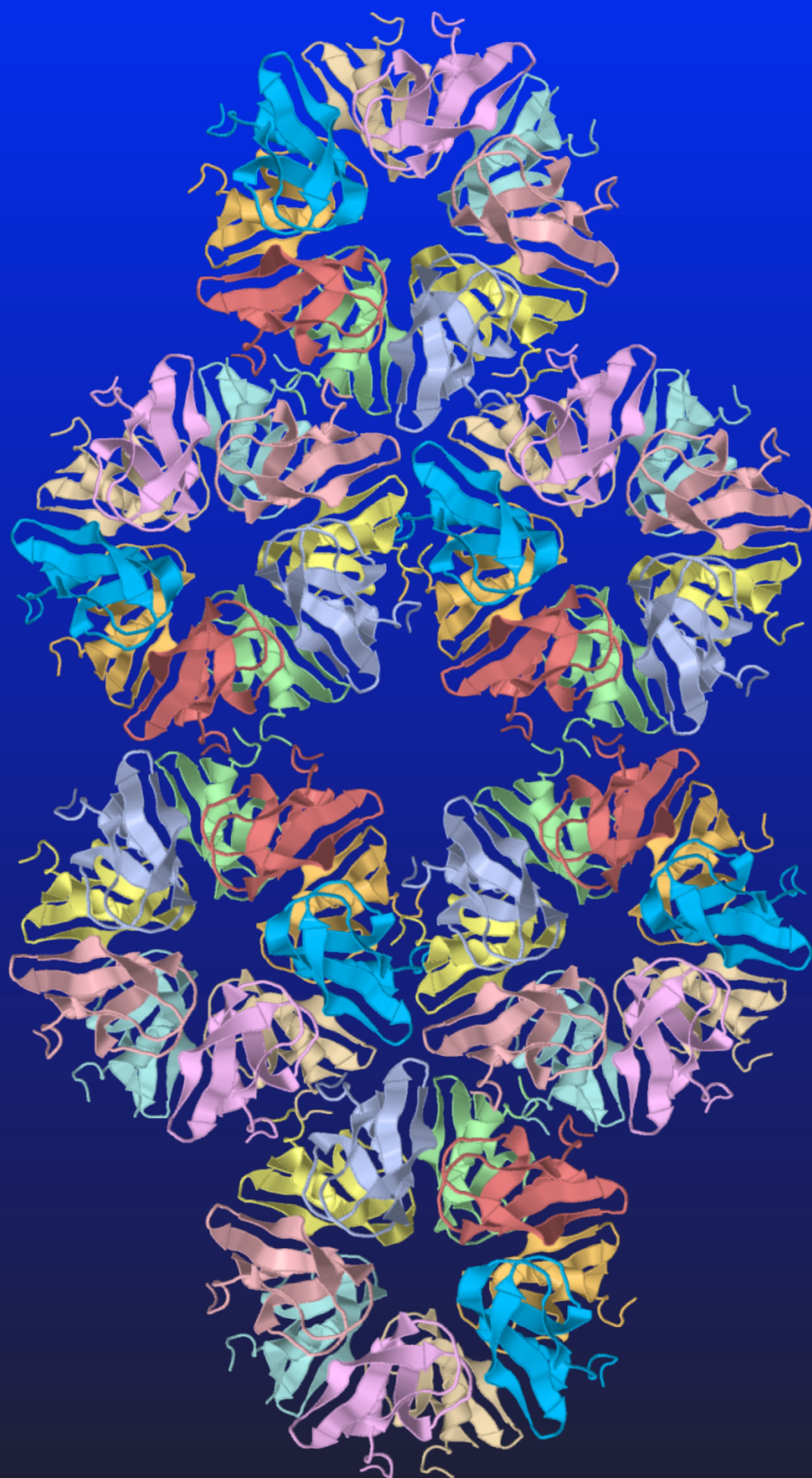


- ★ Why would we want to know the structure of a macromolecule?
  - ➔ for many reasons, but probably firstly for finding out how it interacts with other molecules
  
- ★ Macromolecular crystals present us with **models** of biological structures and interactions between them
  - ➔ “if you want to know how A interacts with B - crystallise them together!” (crystallographer’s sweet dream) But does this always work?



Research Complex at Harwell

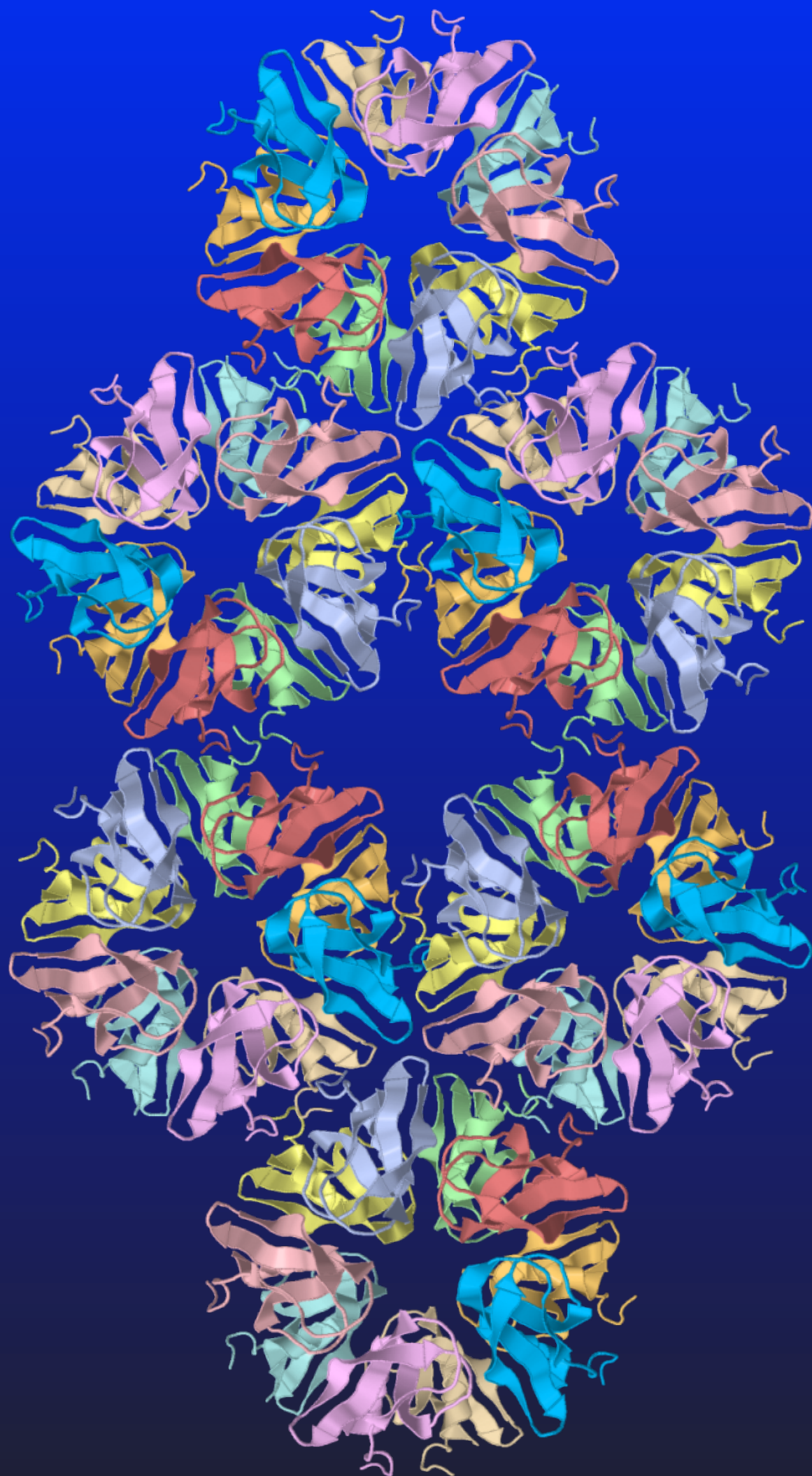




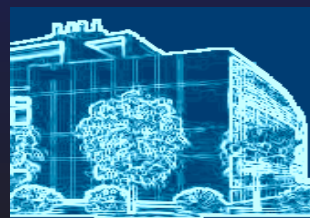
- ★ Why would we want to know the structure of a macromolecule?
  - ➔ for many reasons, but probably firstly for finding out how it interacts with other molecules
  
- ★ Macromolecular crystals present us with **models** of biological structures and interactions between them
  - ➔ “if you want to know how A interacts with B - crystallise them together!” (crystallographer’s sweet dream) But does this always work?
    - ➔ interactions make complexes







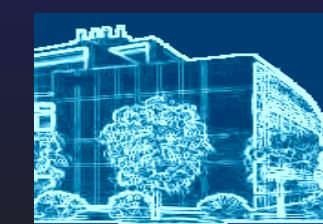
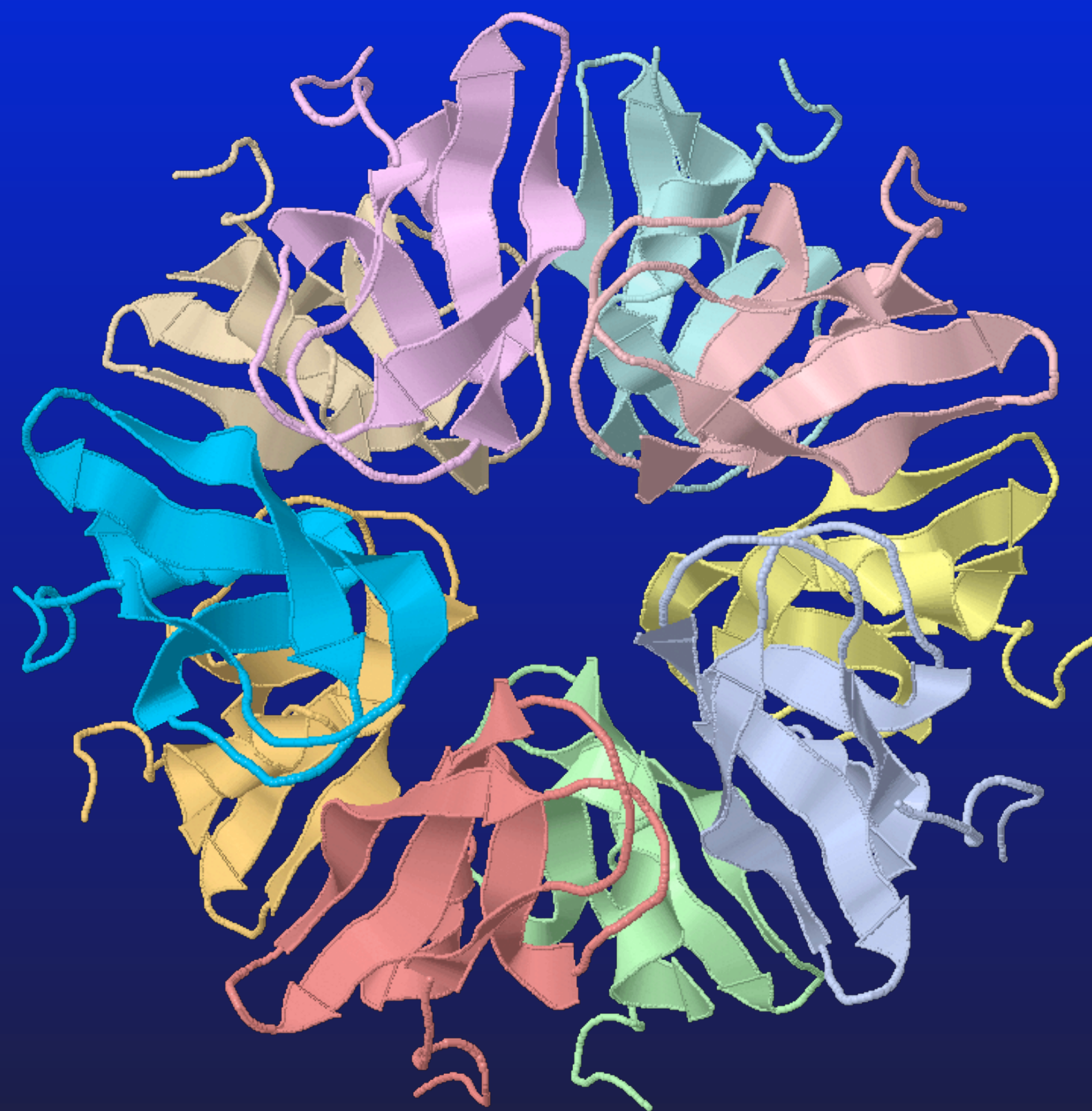
- ★ Why would we want to know the structure of a macromolecule?
  - ➔ for many reasons, but probably firstly for finding out how it interacts with other molecules
  
- ★ Macromolecular crystals present us with models of biological structures and interactions between them
  - ➔ “if you want to know how A interacts with B - crystallise them together!” (crystallographer’s sweet dream) But does this always work?
    - ➔ interactions make complexes
    - ➔ complexes make biology





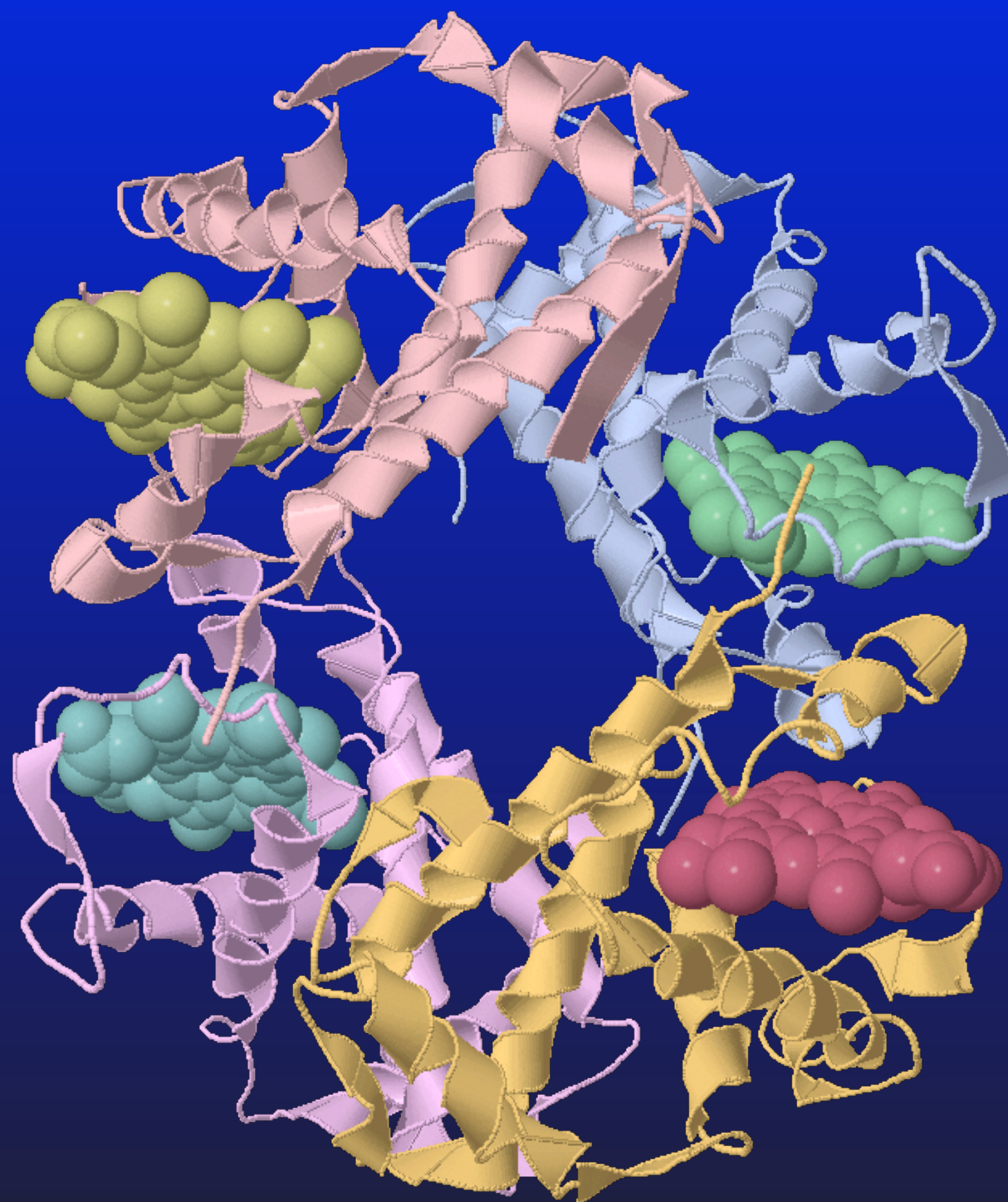
# Macromolecular Assemblies

- ◆ Complexes of protein, DNA/RNA chains and ligands, stable in native environment
- ◆ The way the chains assemble represents the [Protein] Quaternary Structure (PQS)
- ◆ Macromolecular assemblies are often the Biological Units, performing certain biochemical functions
- ◆ Biological significance of macromolecular assemblies is truly immense





# Complex example: Haemoglobin



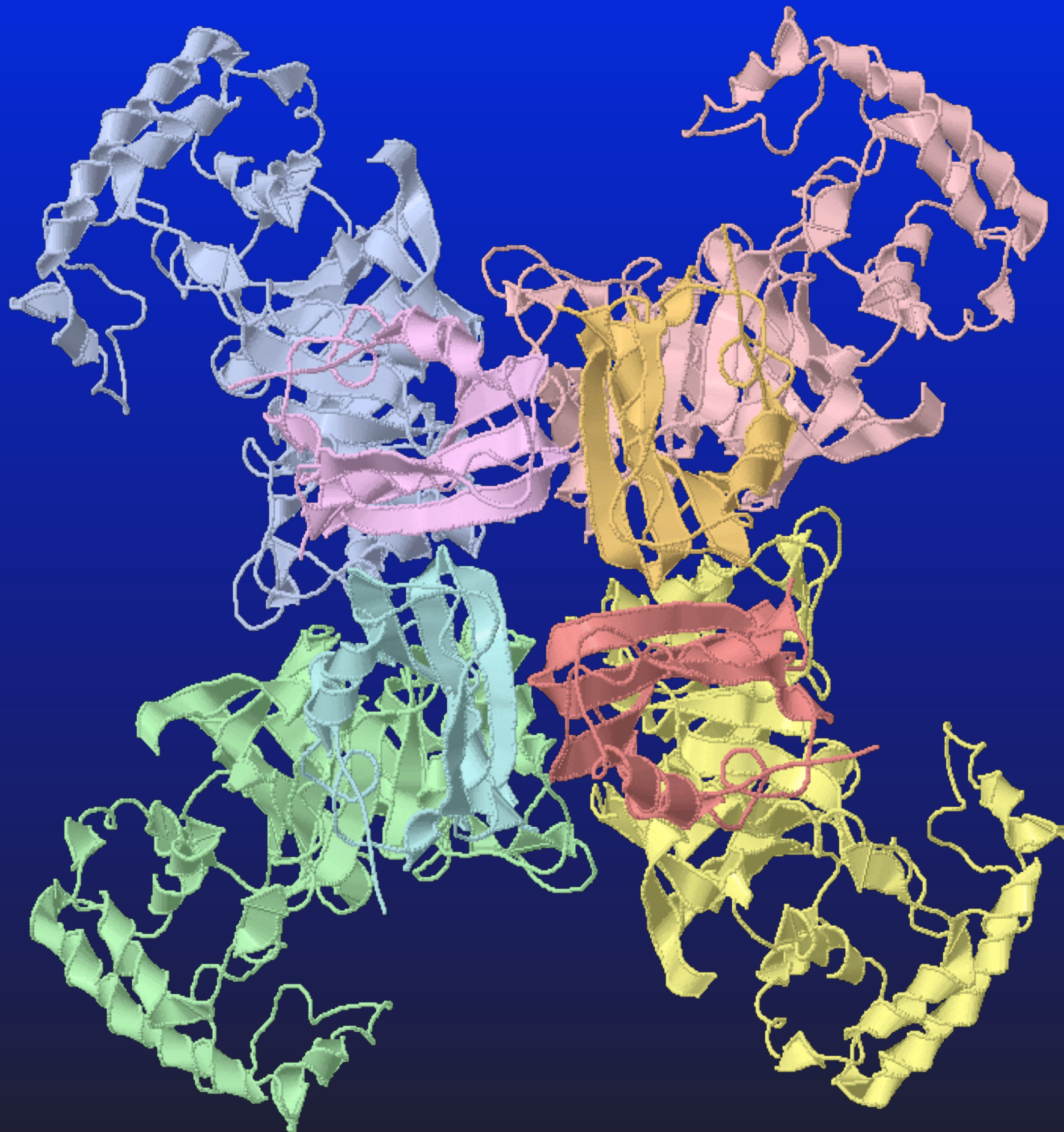
PDB code: 1a00

- Iron-containing oxygen-transport tetrameric protein complex
- Oxygen gets bound to  $\text{Fe}^{+2}$  in heme ligand
- Binding and release of ligands induces structural changes





# Complex example: Ion Channel



- Pore-forming proteins that help to establish and control a voltage gradient across the plasma membrane of cells.
- Change conformation as they operate

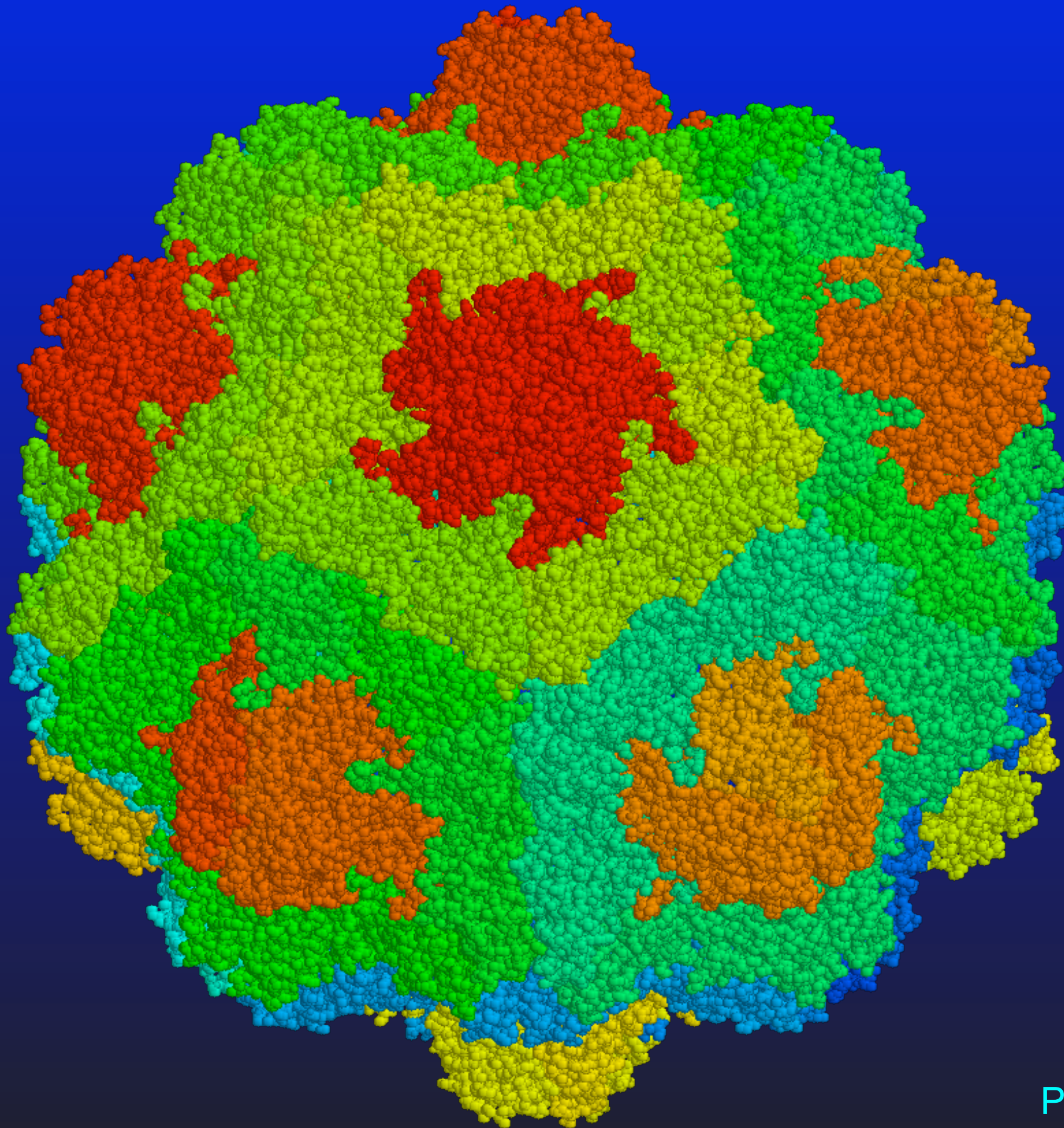
PDB code: 1exb



Research Complex at Harwell



# Complex example: Viral Capsid



- Outer shell of a virus
- Protects genetic material of the virus
- Determines if a cell is suitable for infection
- Starts the actual infection by attaching and opening the target cell and injecting the genetic material into the cell

PDB code: 1w39

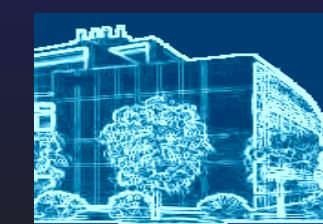
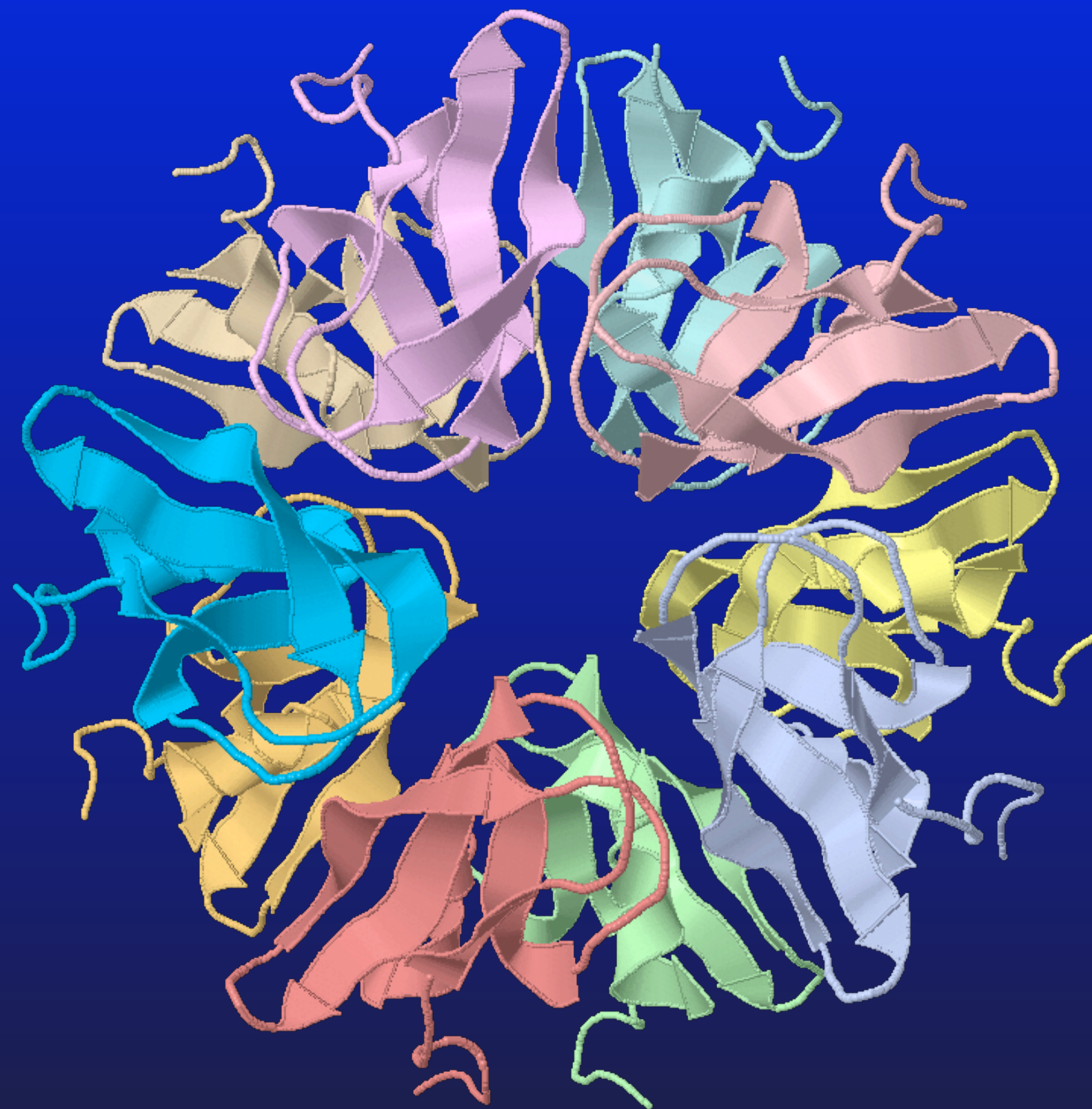


Research Complex at Harwell



# PQS is a difficult object for experimental studies

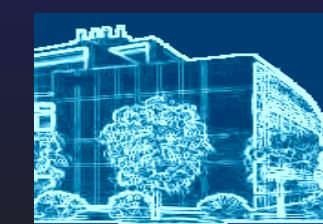
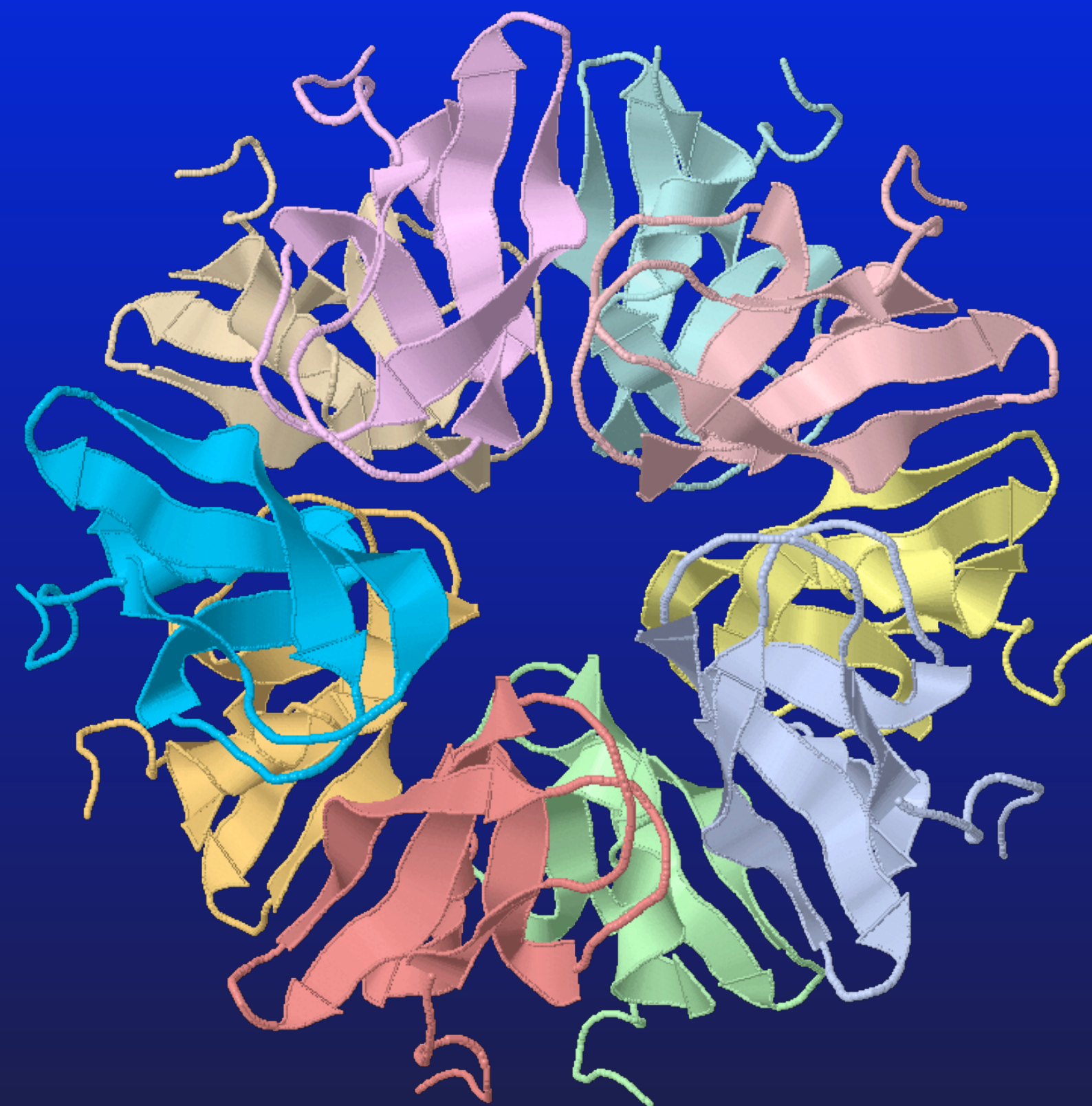
- ◆ *Light / Neutron / X-ray / Small angle scatterings*: mainly composition and multimeric state may be found. 3D shape may be guessed from mobility measurements.





# PQS is a difficult object for experimental studies

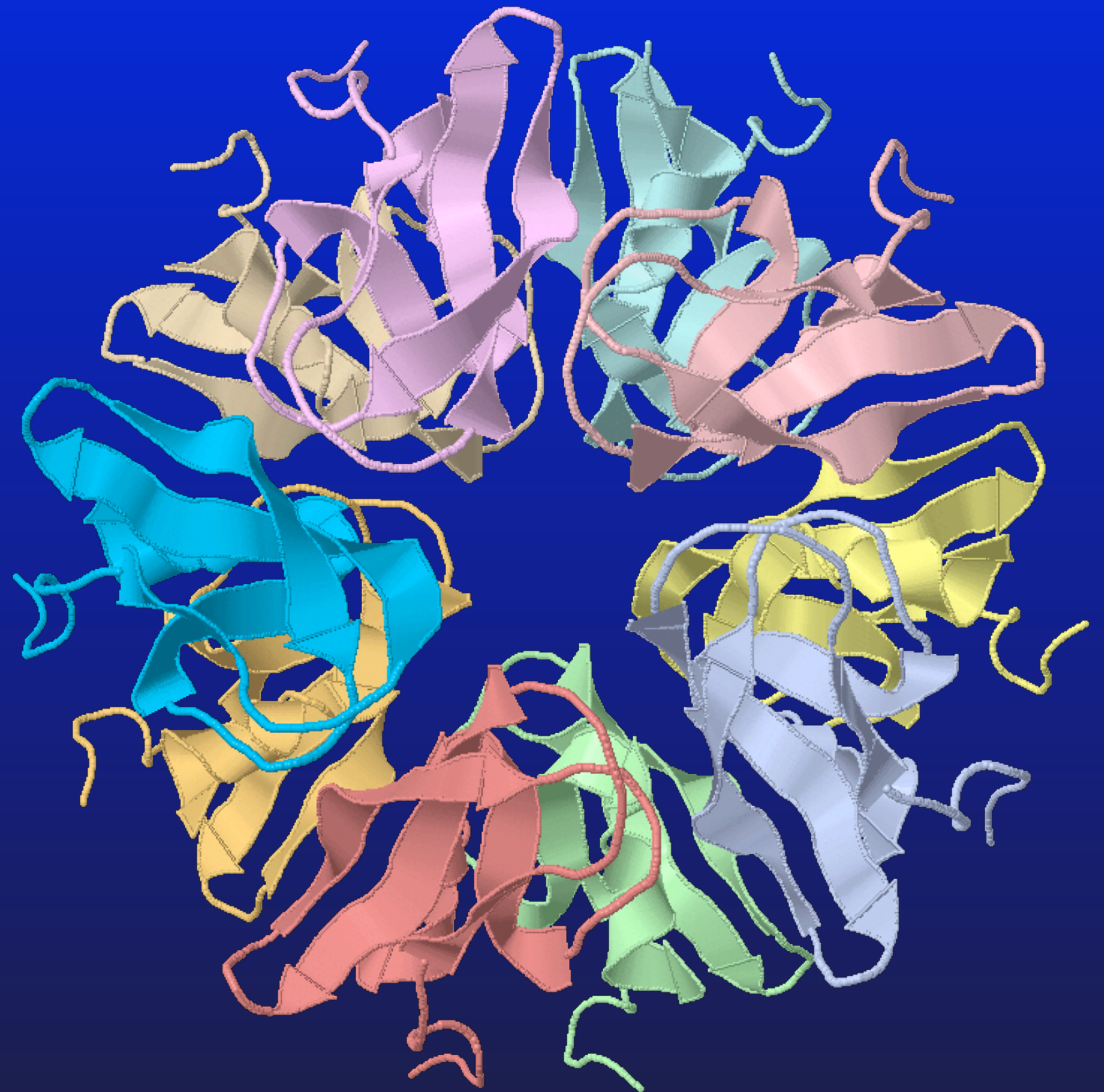
- ◆ *Light / Neutron / X-ray / Small angle scatterings*: mainly composition and multimeric state may be found. 3D shape may be guessed from mobility measurements.
- ◆ *Electron microscopy*: not a fantastic resolution and not applicable to all objects





# PQS is a difficult object for experimental studies

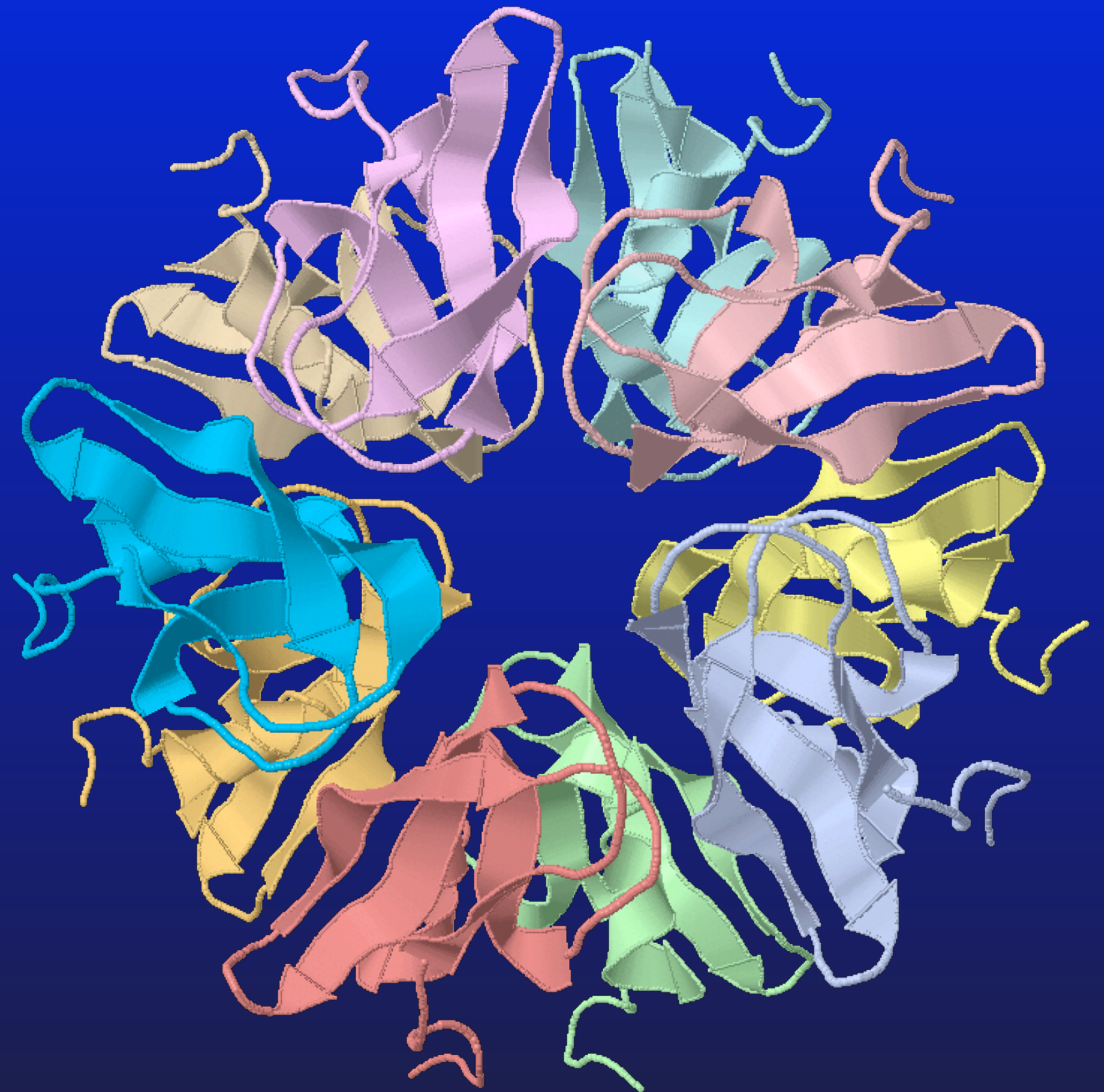
- ◆ *Light / Neutron / X-ray / Small angle scatterings*: mainly composition and multimeric state may be found. 3D shape may be guessed from mobility measurements.
- ◆ *Electron microscopy*: not a fantastic resolution and not applicable to all objects
- ◆ *NMR* is not good for big chains, even less so for protein assemblies.





# PQS is a difficult object for experimental studies

- ◆ *Light / Neutron / X-ray / Small angle scatterings*: mainly composition and multimeric state may be found. 3D shape may be guessed from mobility measurements.
- ◆ *Electron microscopy*: not a fantastic resolution and not applicable to all objects
- ◆ *NMR* is not good for big chains, even less so for protein assemblies.



*Not so many quaternary structures have been identified outside the crystal context, while crystal models may or may not be correct*





# PQS are difficult to calculate



*Research Complex at Harwell*



# PQS are difficult to calculate

*If we know the sequence ...*

**1** **VNKERTFLAVKPDGVARGLVGEIIARYEKKGFVLVGLKQLVPTKDLAESHYA EHKERPFF**  
*then we can calculate ...*



*Research Complex at Harwell*



# PQS are difficult to calculate

*If we know the sequence ...*

**1** VNKERTFLAVKPDGVARGLVGEIIARYEKKGFVLVGLKQLVPTKDLAESHYA~~EHK~~ERPFF

*then we can calculate ...*



**50 - 90%** Secondary Structure (CASP 5), depending on *method*




Research Complex at Harwell

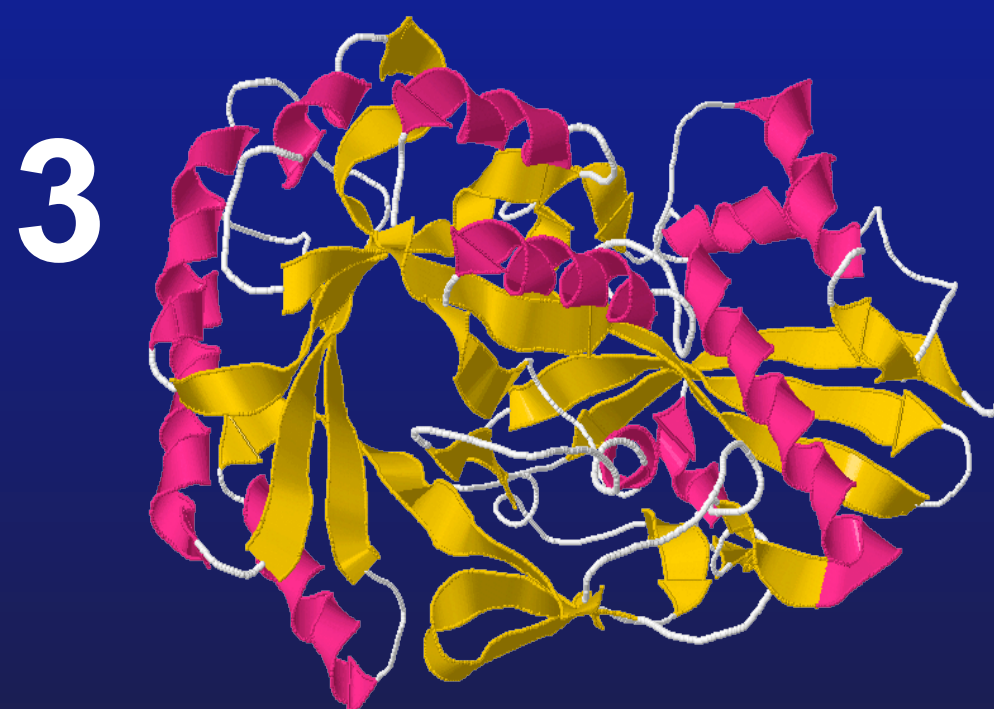


# PQS are difficult to calculate

*If we know the sequence ...*

**1** **VNKERTFLAVKPDGVARGLVGEIIARYEKKGFVLVGLKQLVPTKDLAESHYA EHKERPFF**  
*then we can calculate ...*

**2**   
**50 - 90%** Secondary Structure (CASP 5), depending on *method*



**10 - 90%** Tertiary Structure  
(CASP 5), depending on  
method and *target*




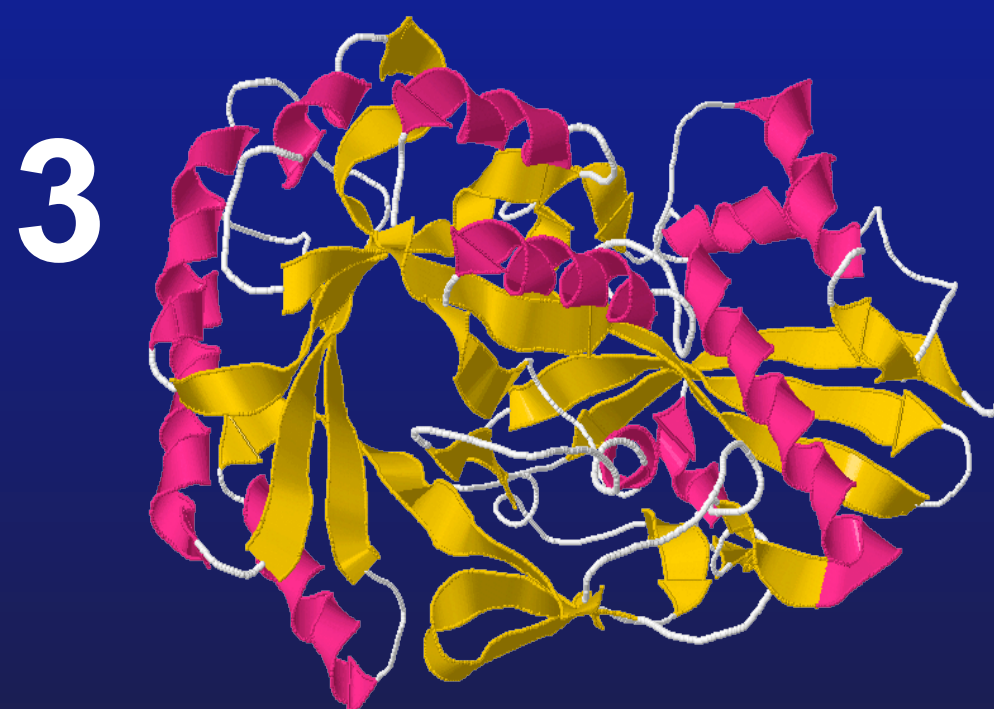


# PQS are difficult to calculate

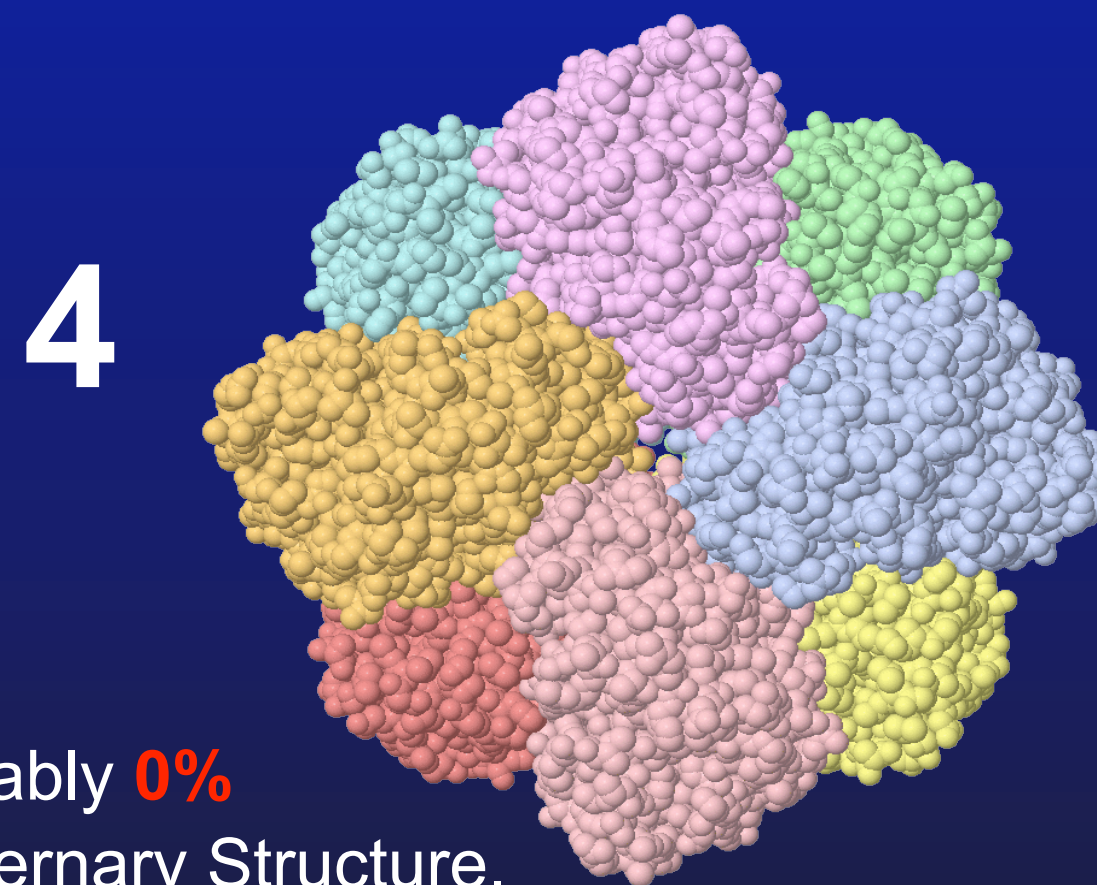
*If we know the sequence ...*

**1** **VNKERTFLAVKPDGVARGLVGEIIARYEKKGFVLVGLKQLVPTKDLAESHYA EHKERPFF**  
*then we can calculate ...*

**2**   
**50 - 90%** Secondary Structure (CASP 5), depending on *method*



**10 - 90%** Tertiary Structure  
(CASP 5), depending on  
method and *target*



Probably **0%**  
Quaternary Structure.  
Docking of given number of  
given structures: **5 - 20%**  
success (CAPRI 5)





But PQS are assigned to almost all entries in the PDB!



[www.pdb.org](http://www.pdb.org)



Research Complex at Harwell



But PQS are assigned to almost all entries in the PDB!



[www.pdb.org](http://www.pdb.org)

Most of those are **PROBABLE** Quaternary **S**tructures.



Research Complex at Harwell



# But PQS are assigned to almost all entries in the PDB!



[www.pdb.org](http://www.pdb.org)

Most of those are **PROBABLE** Quaternary **S**tructures.

The PDB “rules” are:

1. Depositor's say prevails.
2. Accept everything which passes formal validation checks.
3. No experimental evidence for PQS is required.
4. If a depositor does not know or does not care (60-80% of instances for PQS), the curator is to decide.
5. The curator may or may not use computing/modeling tools to assist the PQS annotation.



Research Complex at Harwell



# PDB does indeed contain a wealth of experimental data on PQS

More than 80% of macromolecular structures are solved by means of X-ray diffraction on crystals.

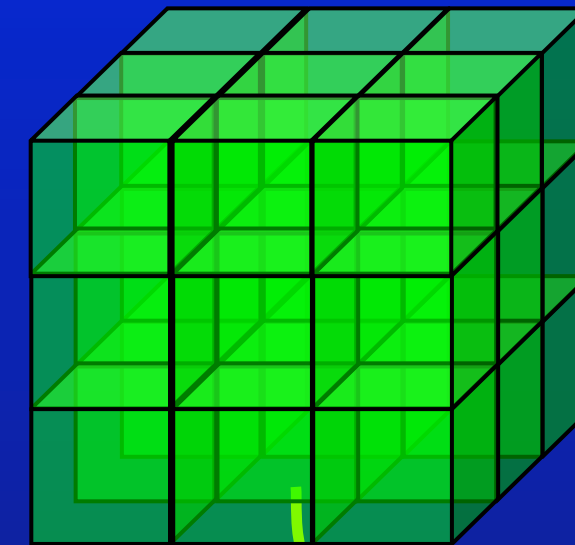
Any crystal represents macromolecular interactions and associations through inter-molecular interfaces

An X-ray diffraction experiment produces atomic coordinates of the Asymmetric Unit (ASU), which is stored as a PDB file.

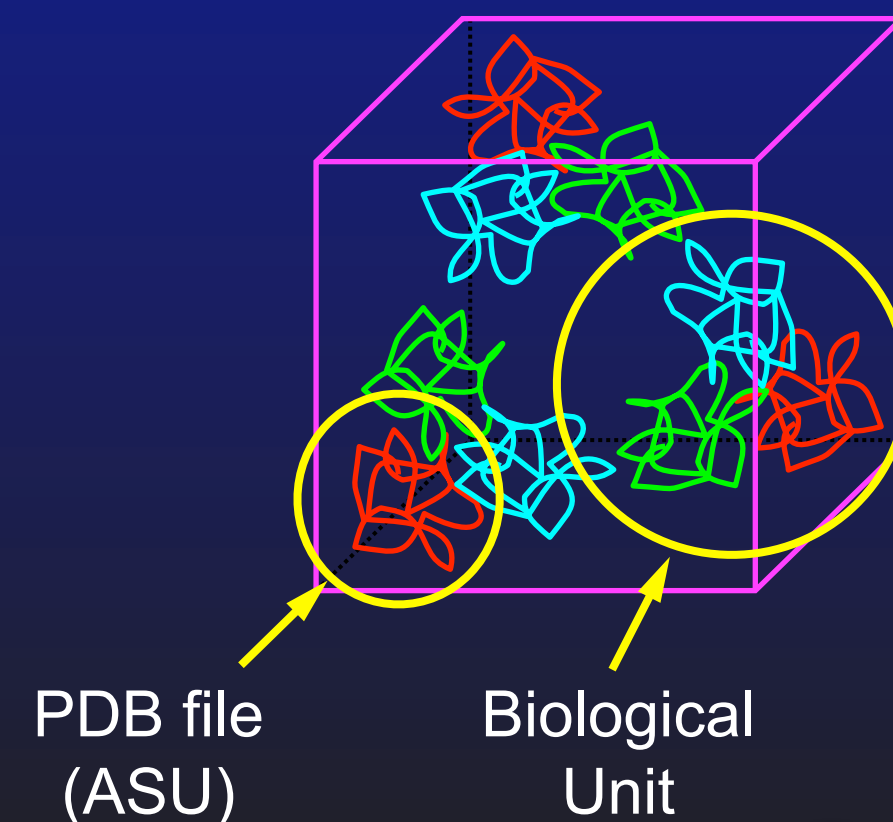
In general, neither ASU nor Unit Cell has any direct relation to PQS. The PQS may be made of

- a single ASU
- a part of ASU
- several ASU
- several ASU parts

Crystal = translated Unit Cells



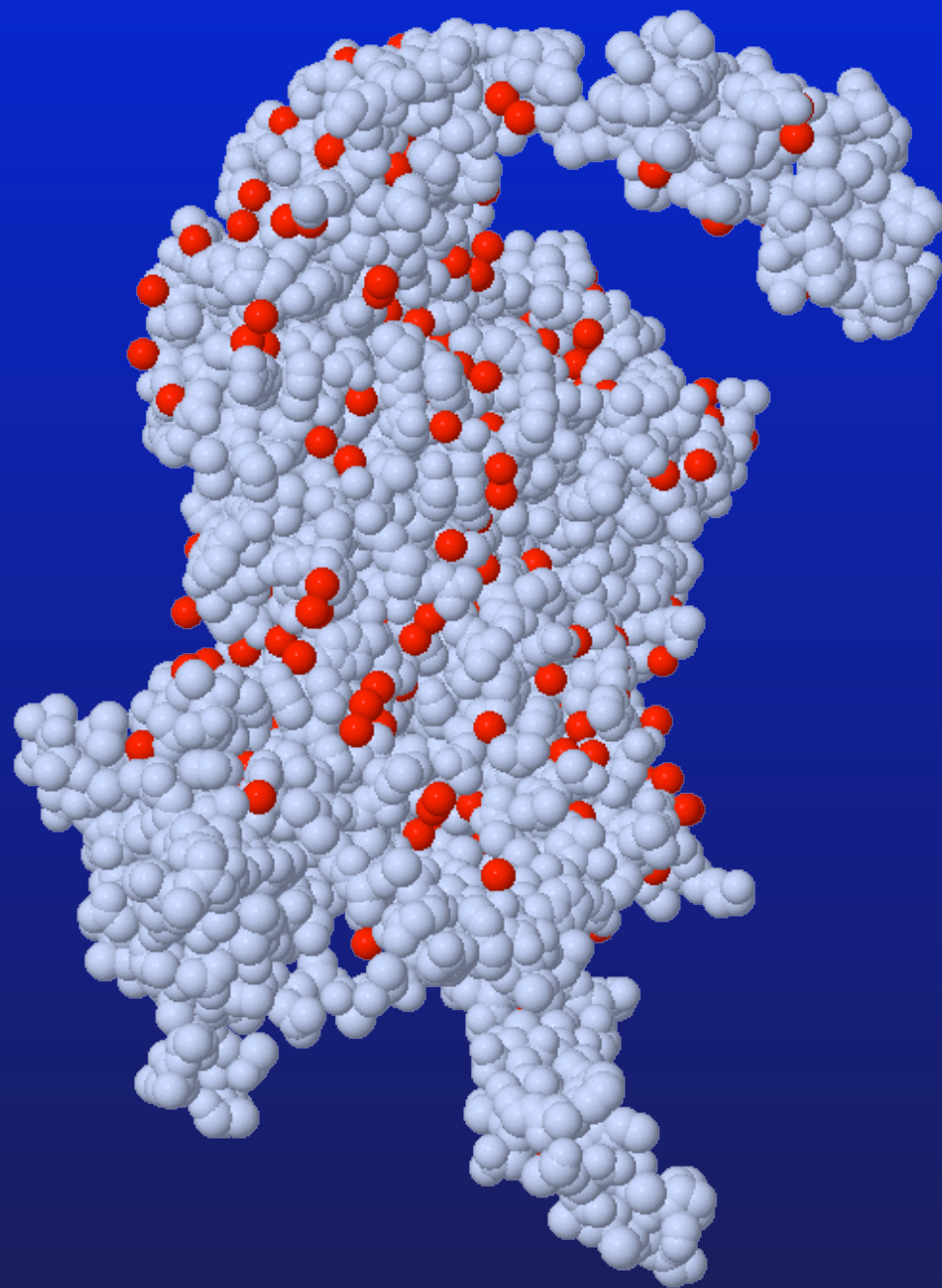
Unit Cell = all space symmetry group mates of ASU



Research Complex at Harwell



# PDB entries and Biological Units



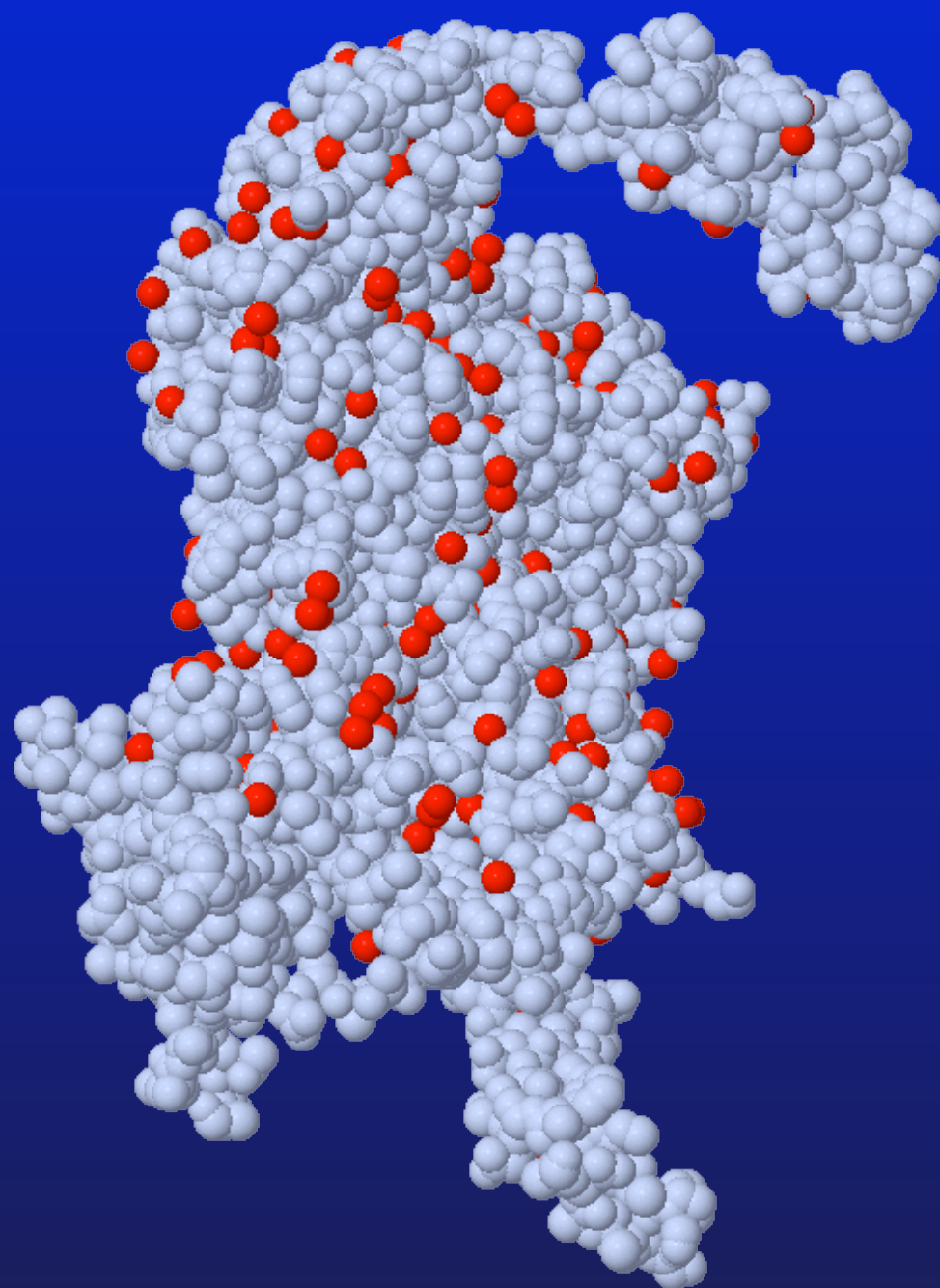
PDB entry 1P30  
A monomer?



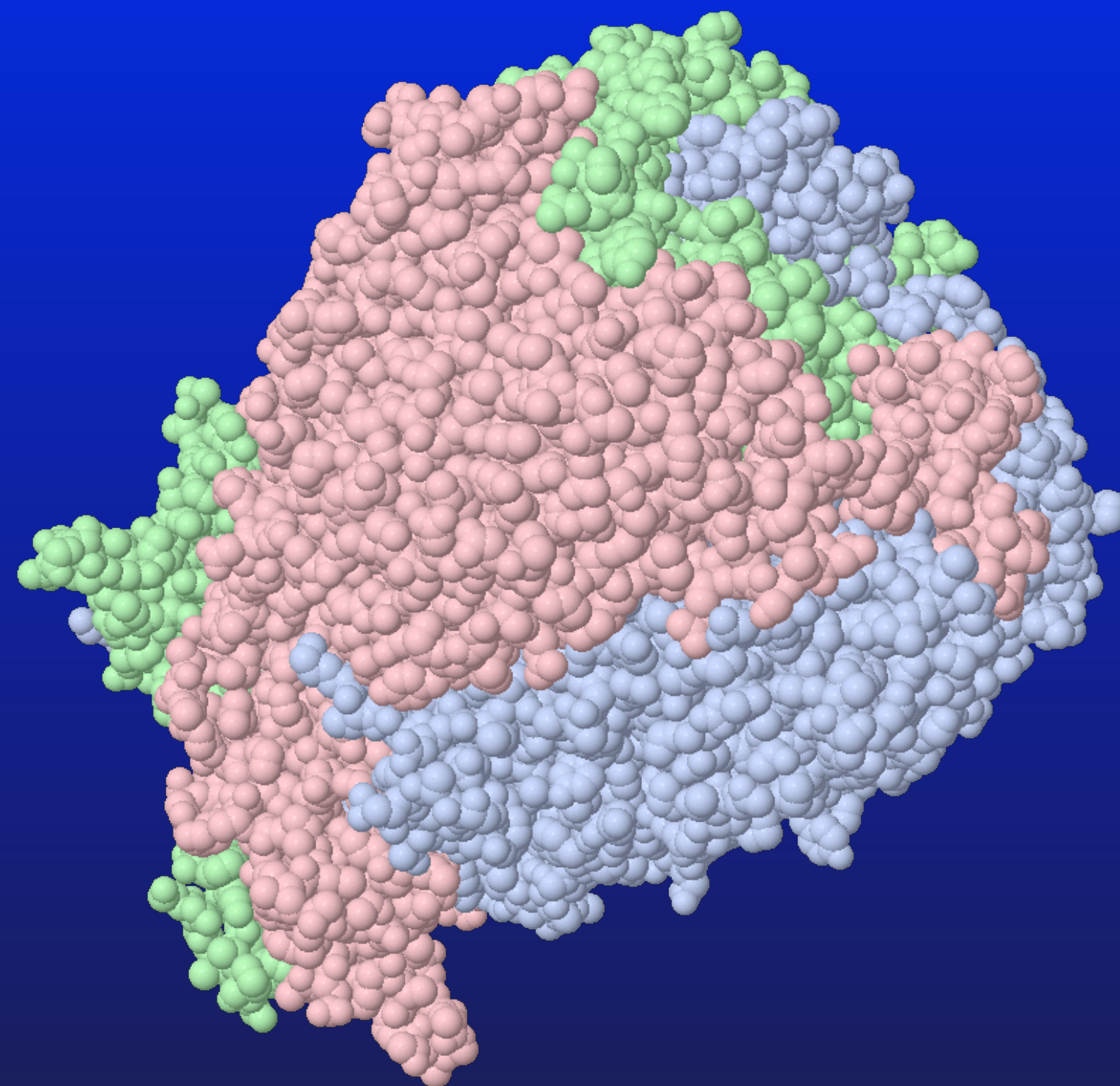
Research Complex at Harwell



# PDB entries and Biological Units



PDB entry 1P30  
A monomer?

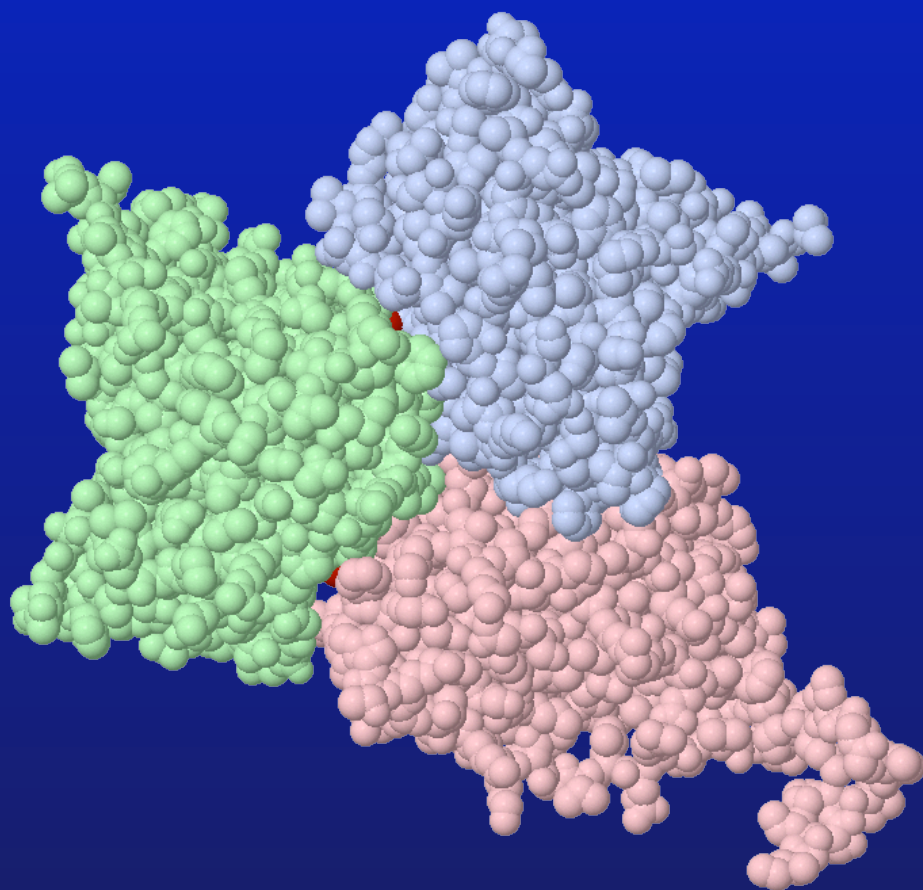


Biological unit 1P30  
Homotrimer!





# PDB entries and Biological Units



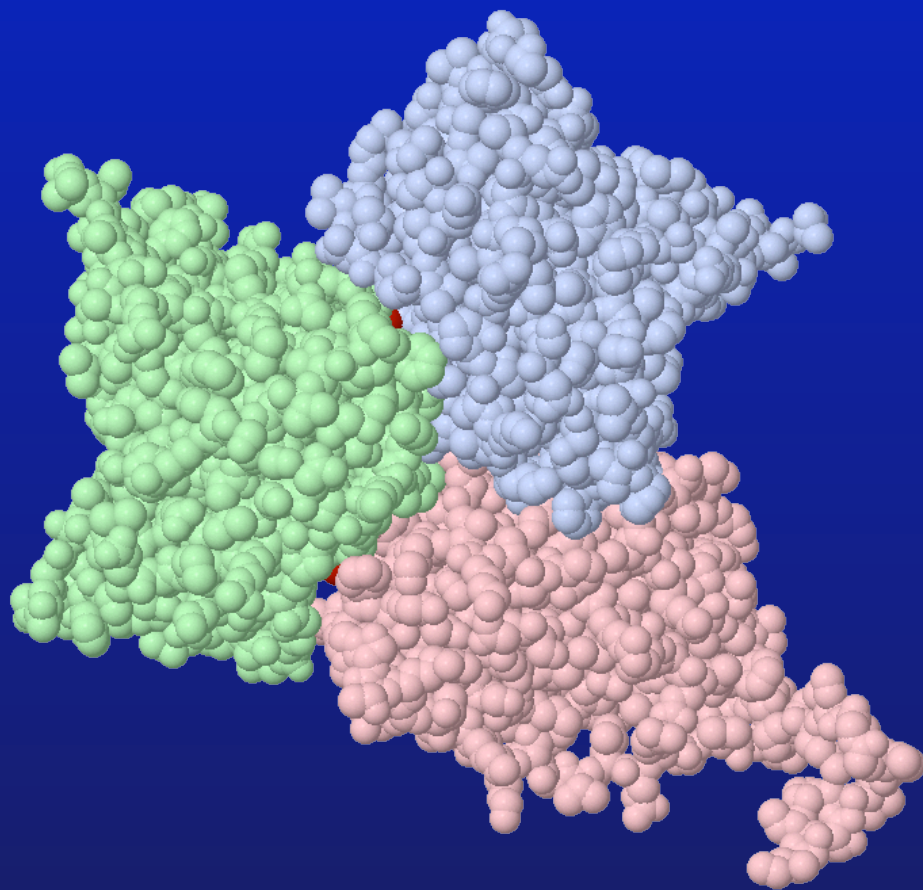
PDB entry 2TBV  
A trimer?



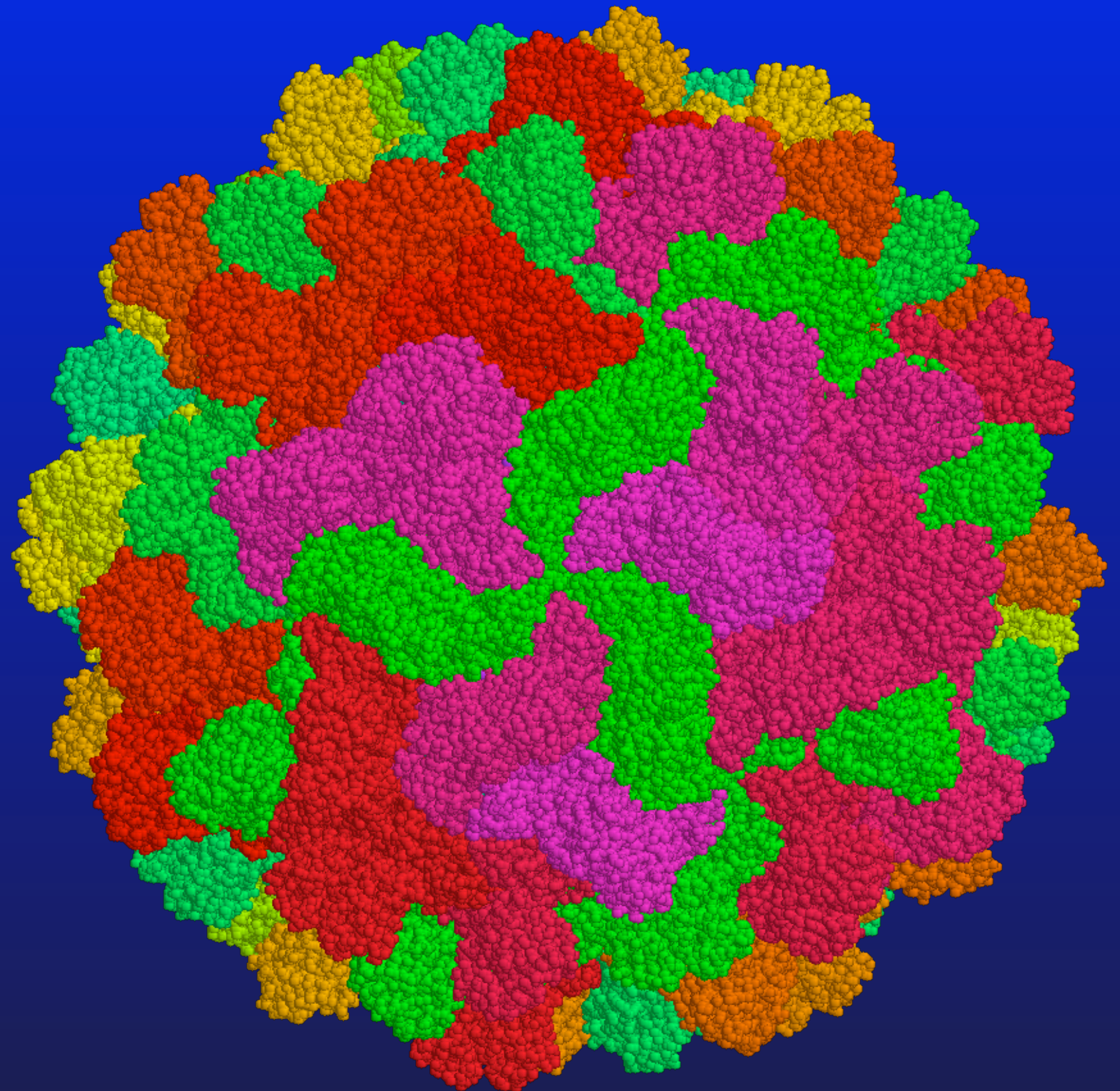
Research Complex at Harwell



# PDB entries and Biological Units



PDB entry 2TBV  
A trimer?

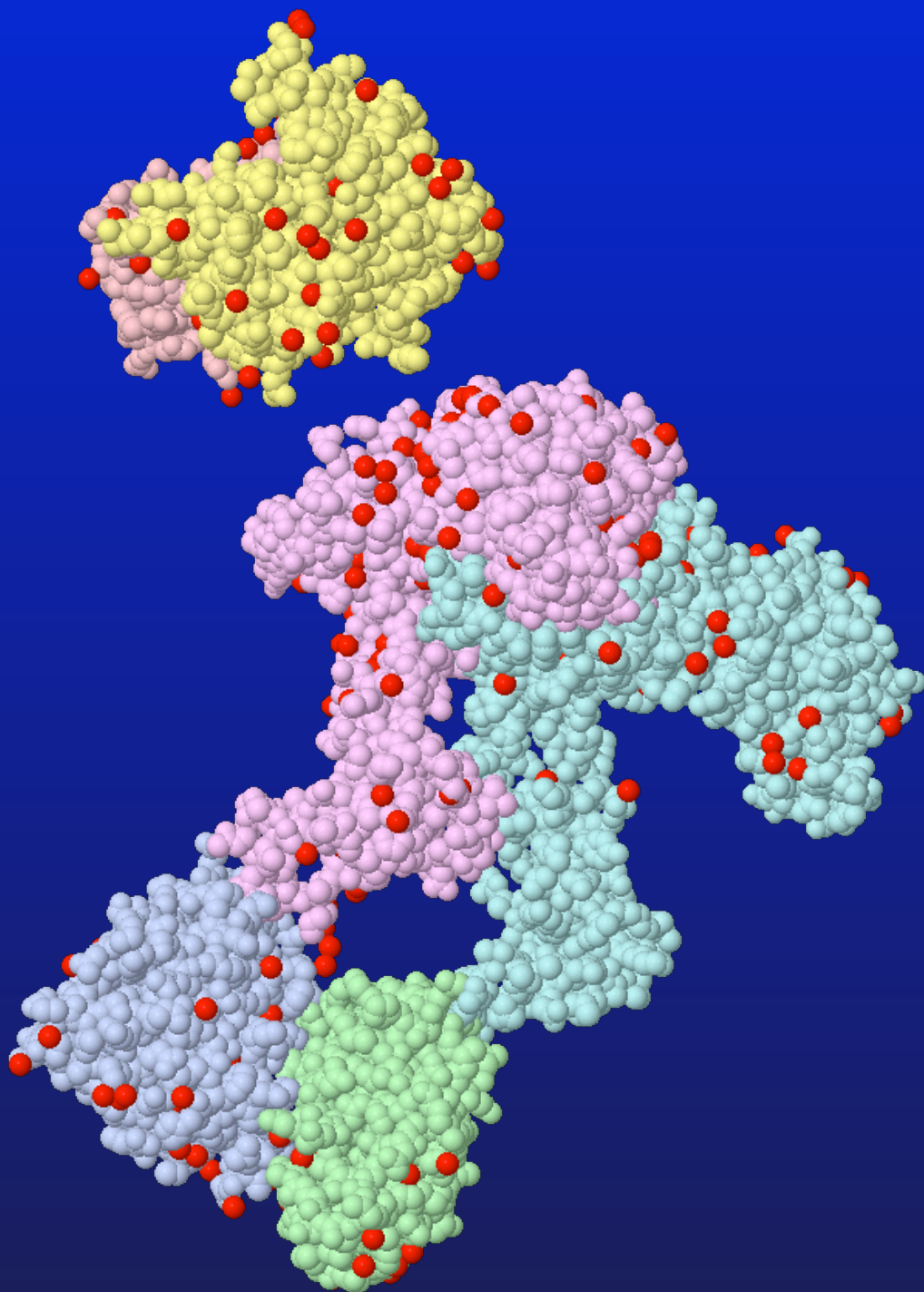


Biological Unit 2TBV  
180-mer!





# PDB entries and Biological Units



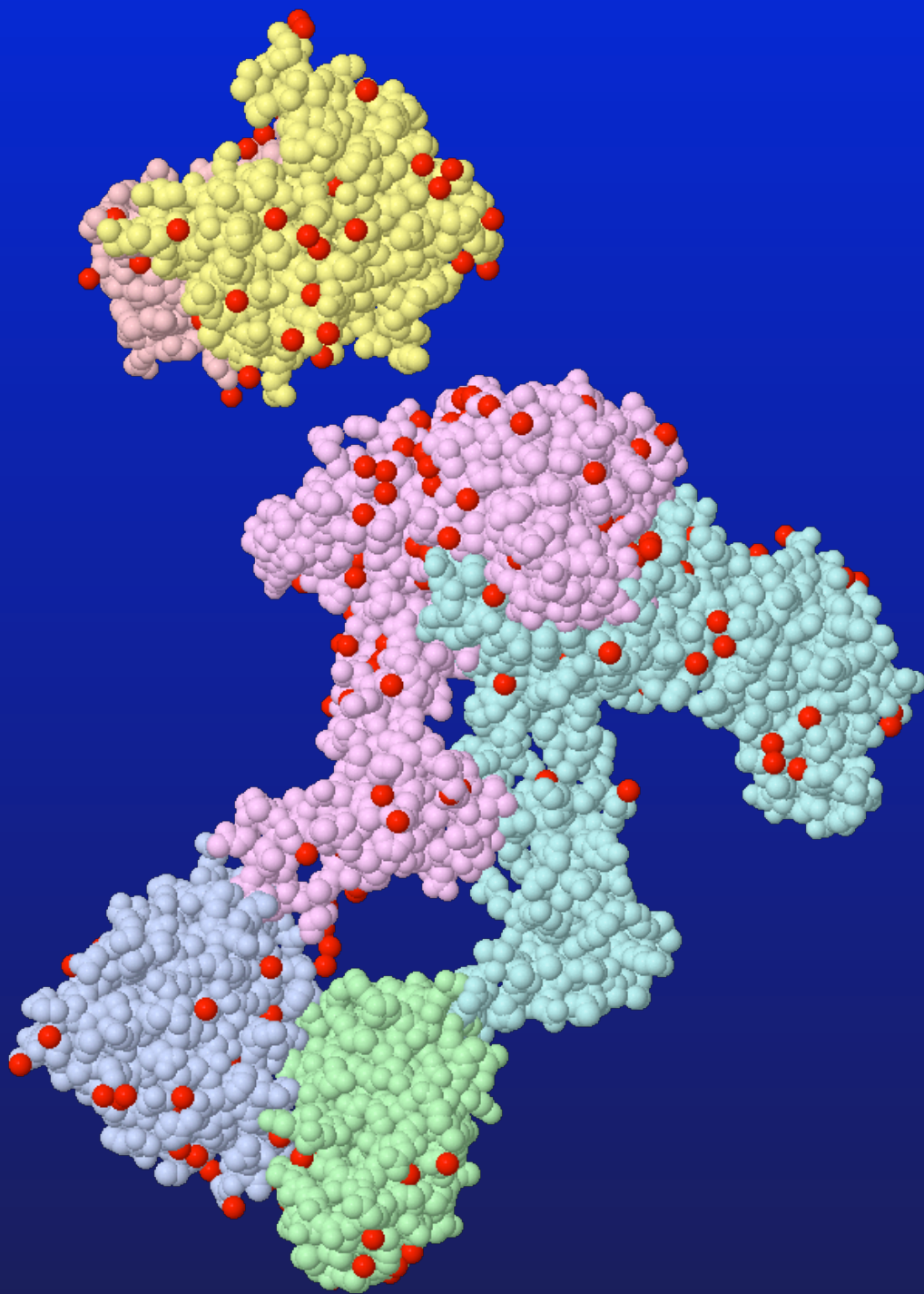
PDB entry 1E94  
?????????



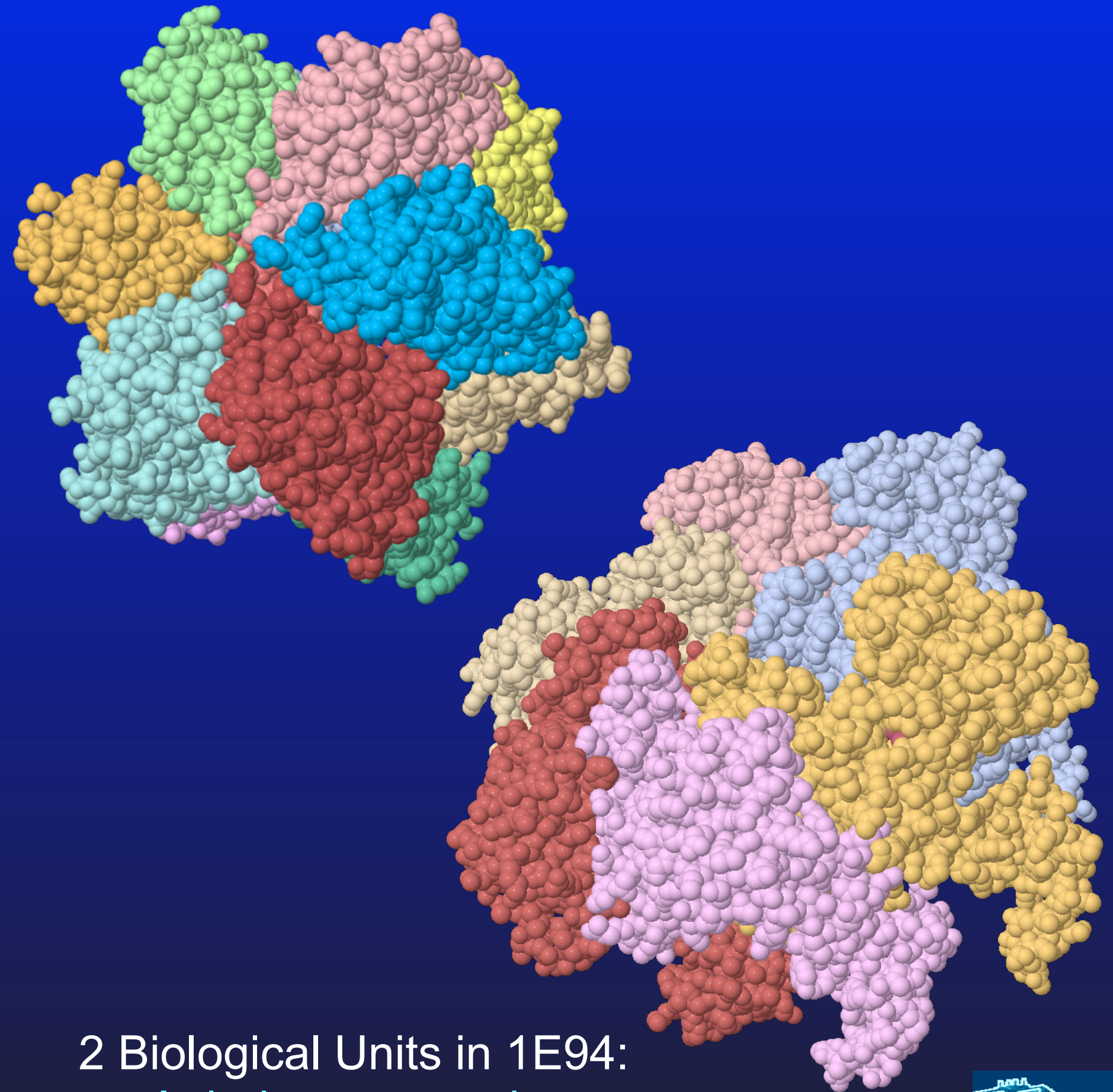
Research Complex at Harwell



# PDB entries and Biological Units



PDB entry 1E94  
???????



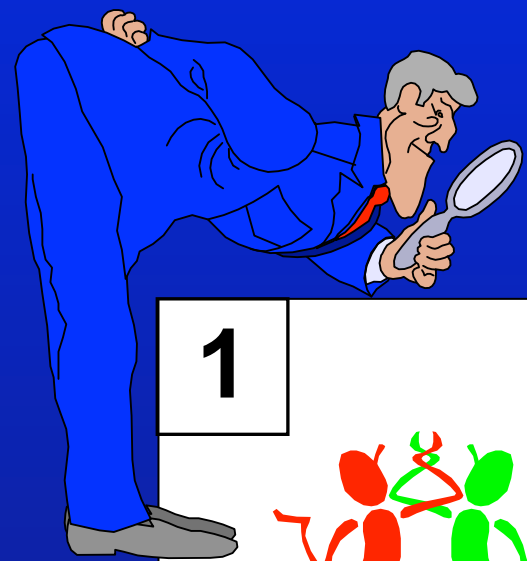
## 2 Biological Units in 1E94: A dodecamer and a hexamer!



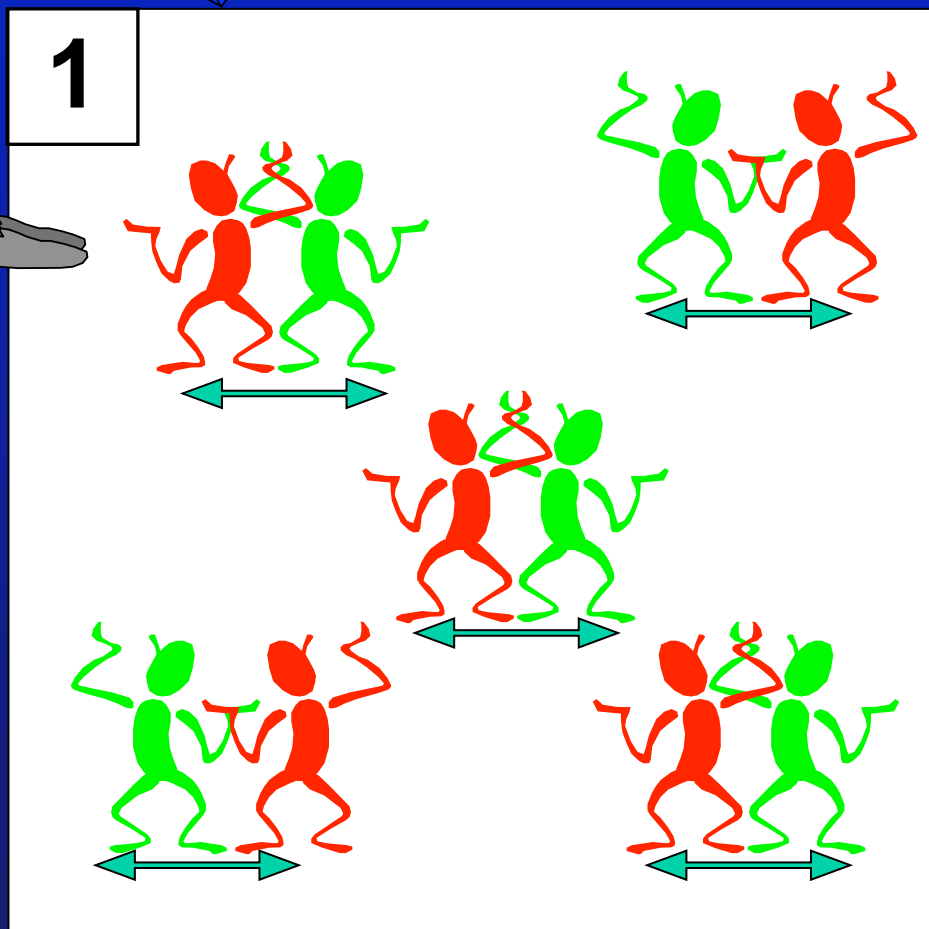
**Research Complex at Harwell**



# In (very) simple words ...



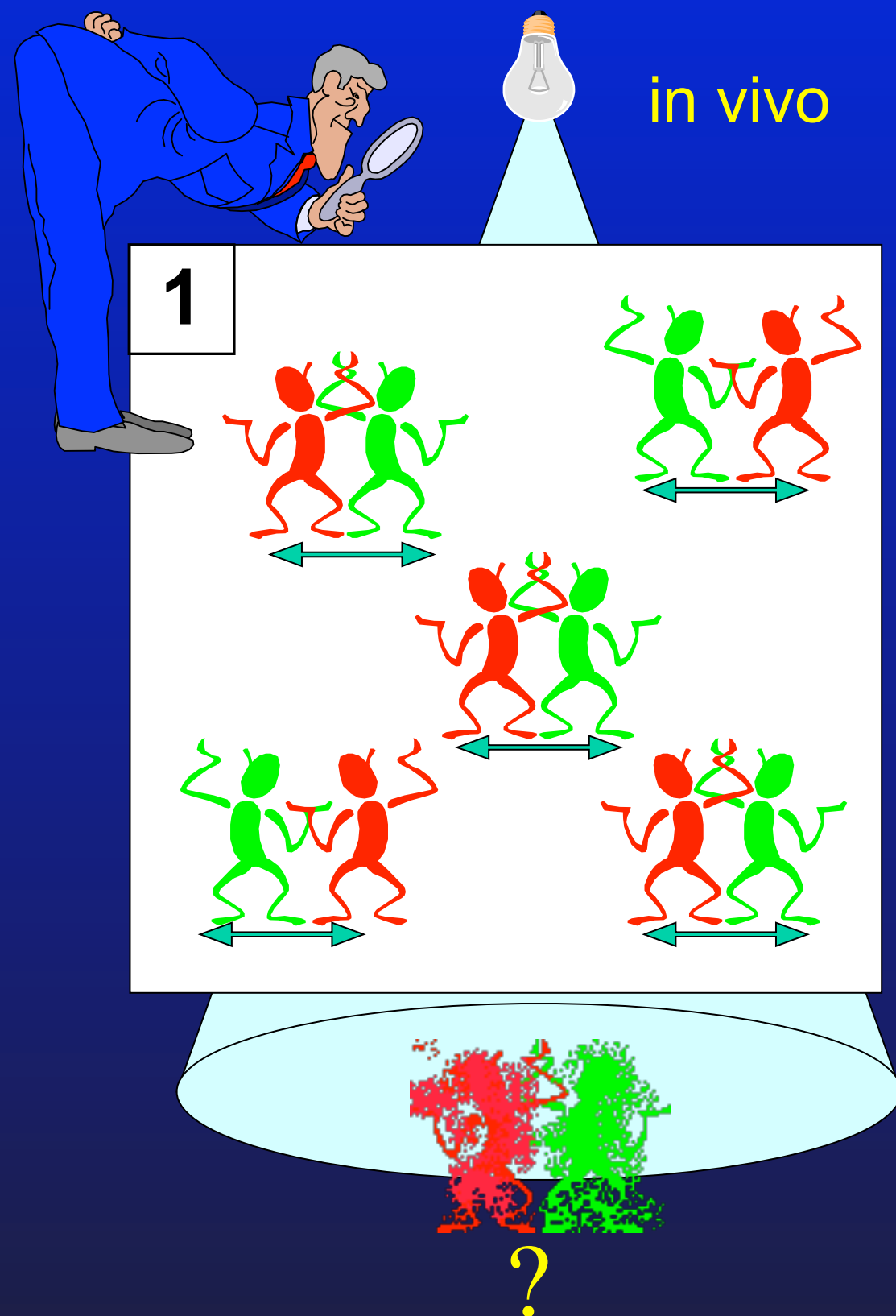
in vivo



Research Complex at Harwell

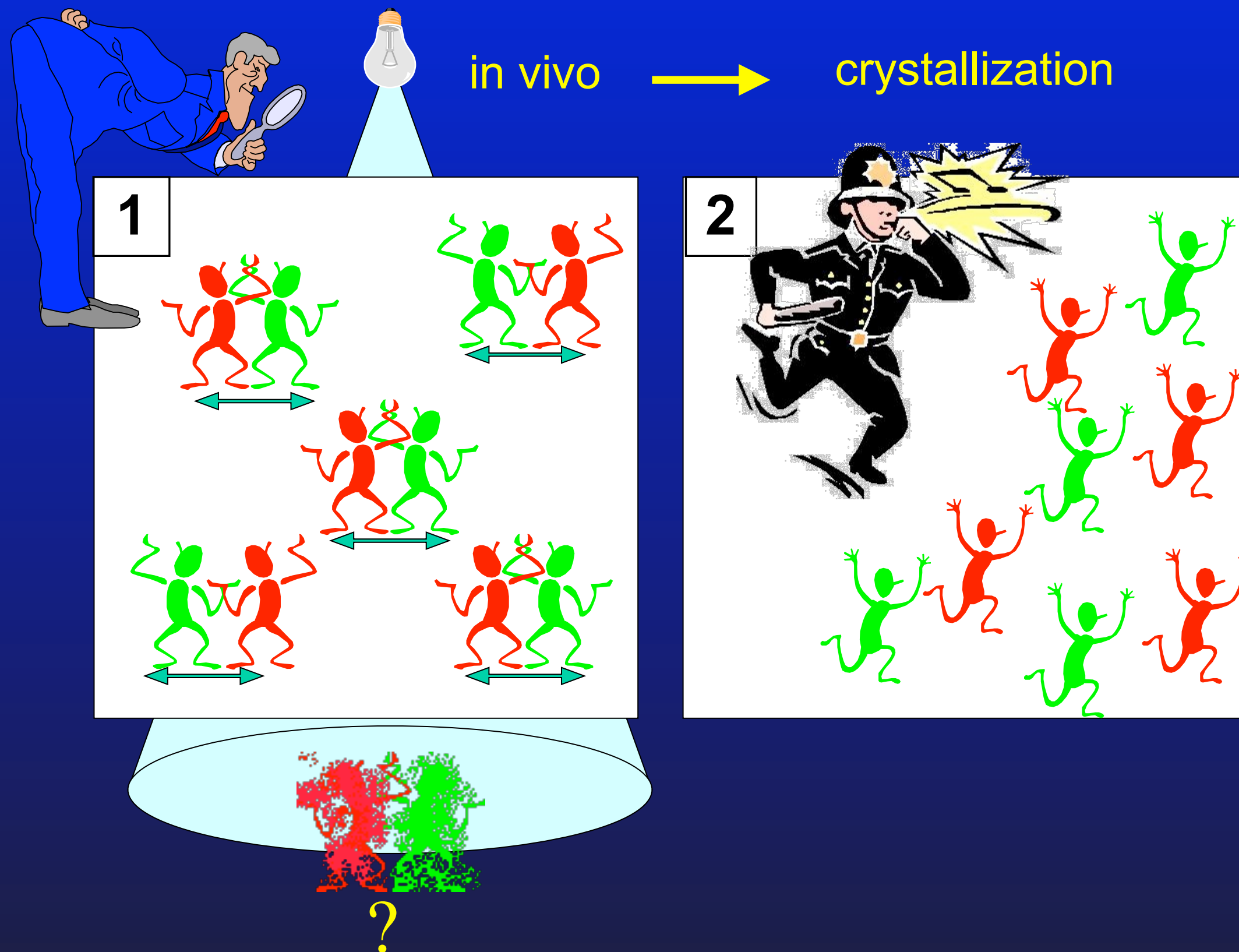


# In (very) simple words ...





# In (very) simple words ...

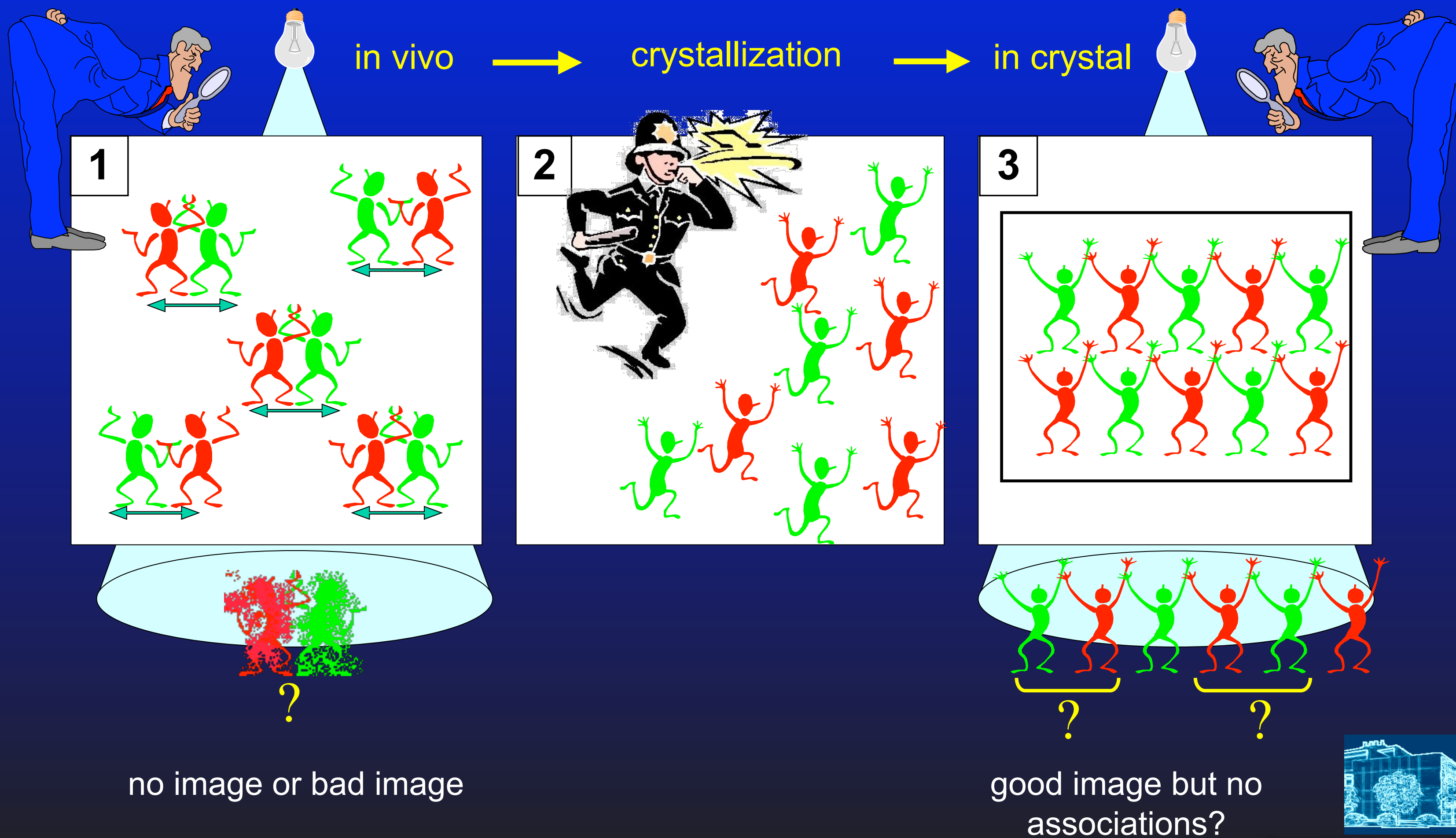


no image or bad image





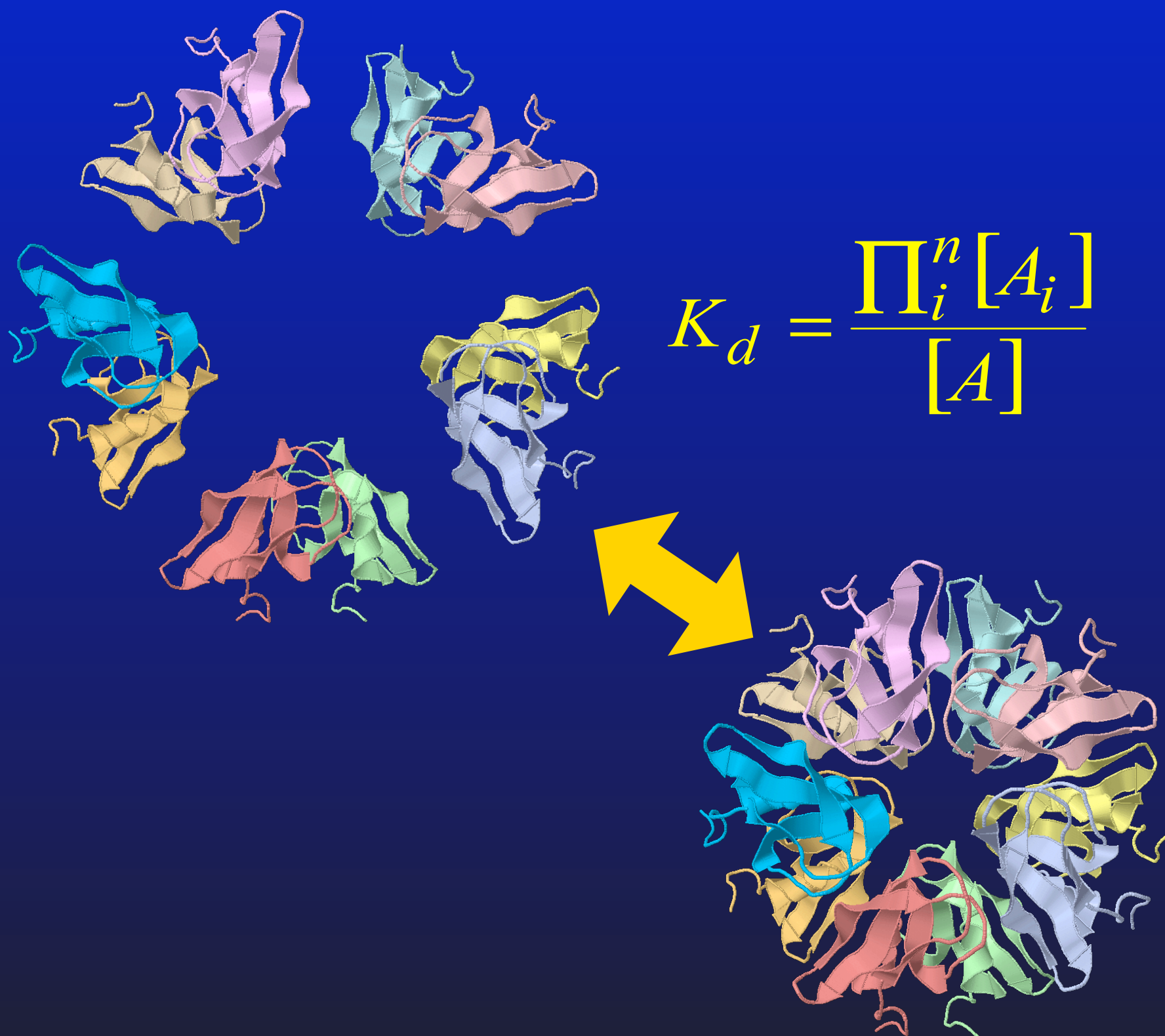
# In (very) simple words ...



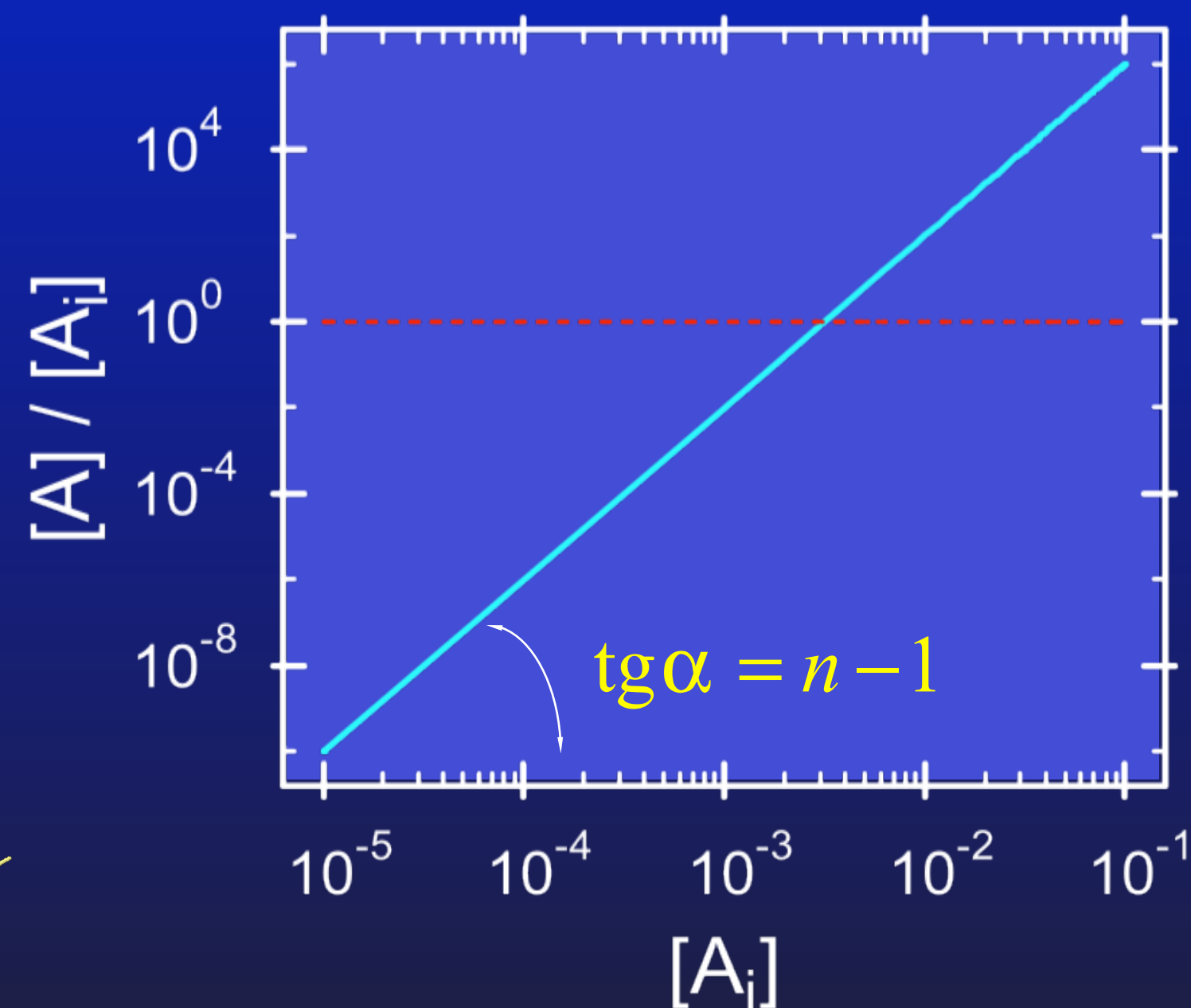


# Things are not that bad, though

Because: A) crystal is made of complexes



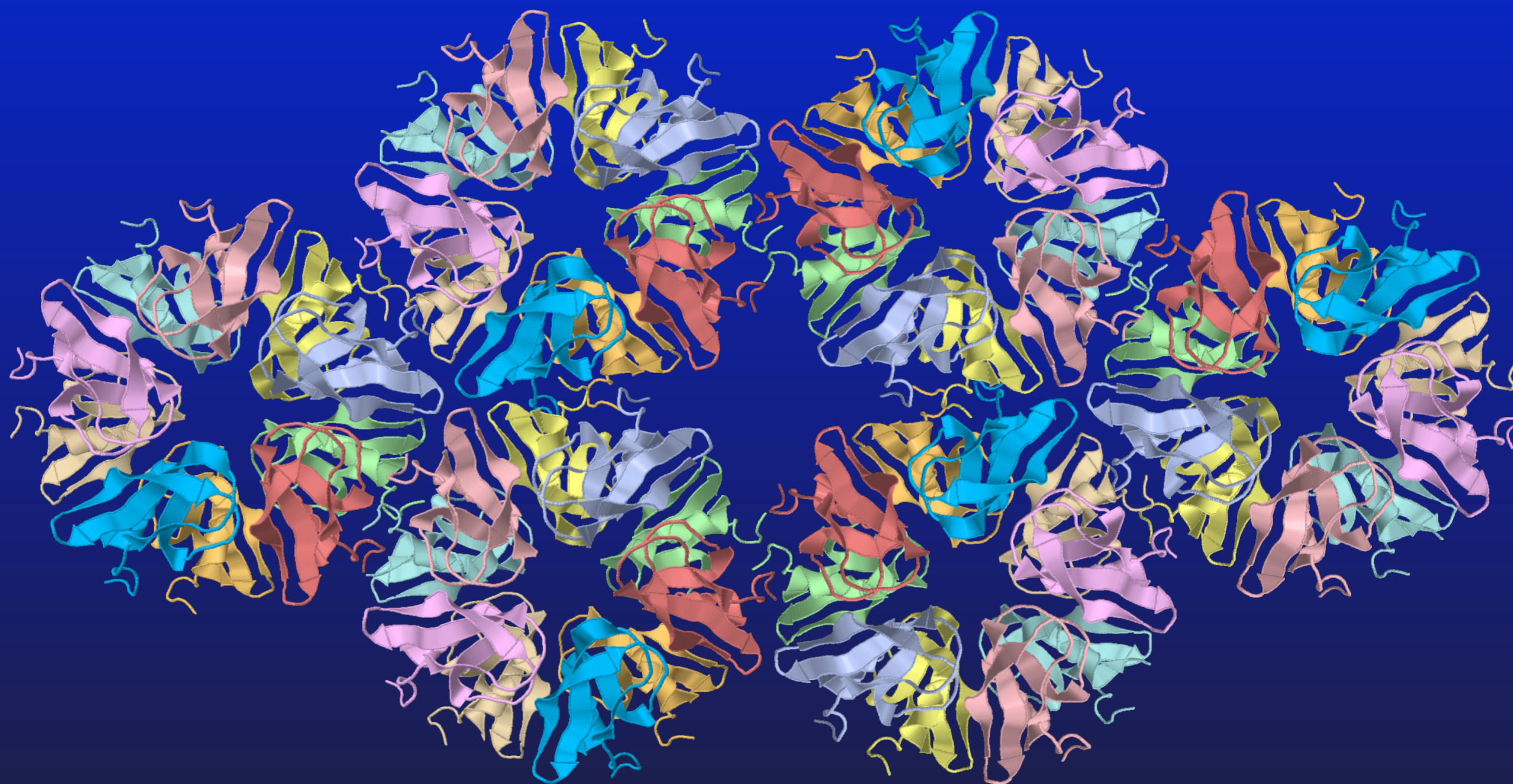
$$[A]/[A_i] = [A_i]^{n-1} / K_d$$





# Things are not that bad, though

Because: B) there is no need to dock subunits for predicting complexes  
– the docking is given by crystal structure



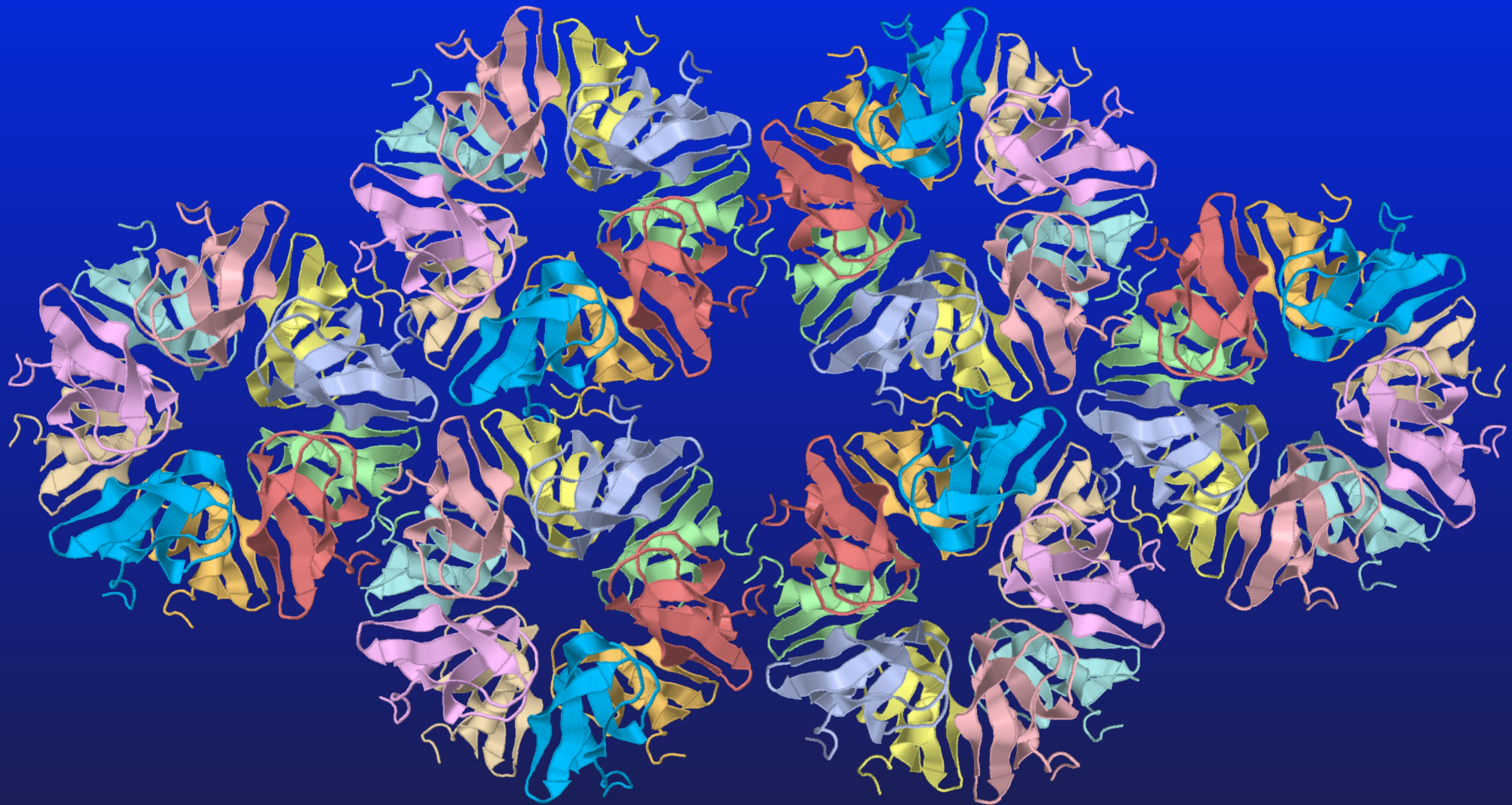
Macromolecular interfaces should be viewed as an additional important product of protein crystallography



Research Complex at Harwell



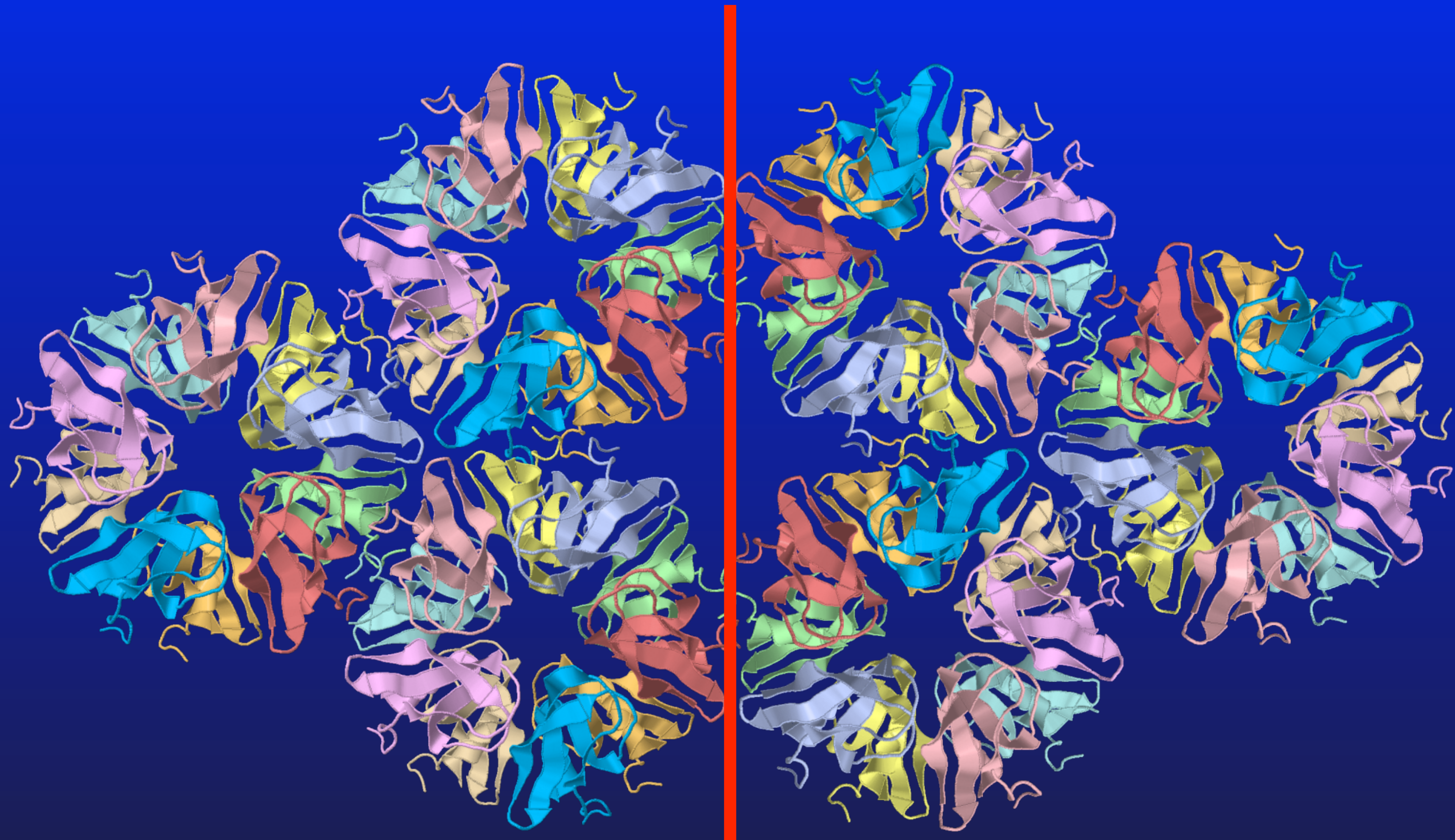
# A simple thing to do



Research Complex at Harwell

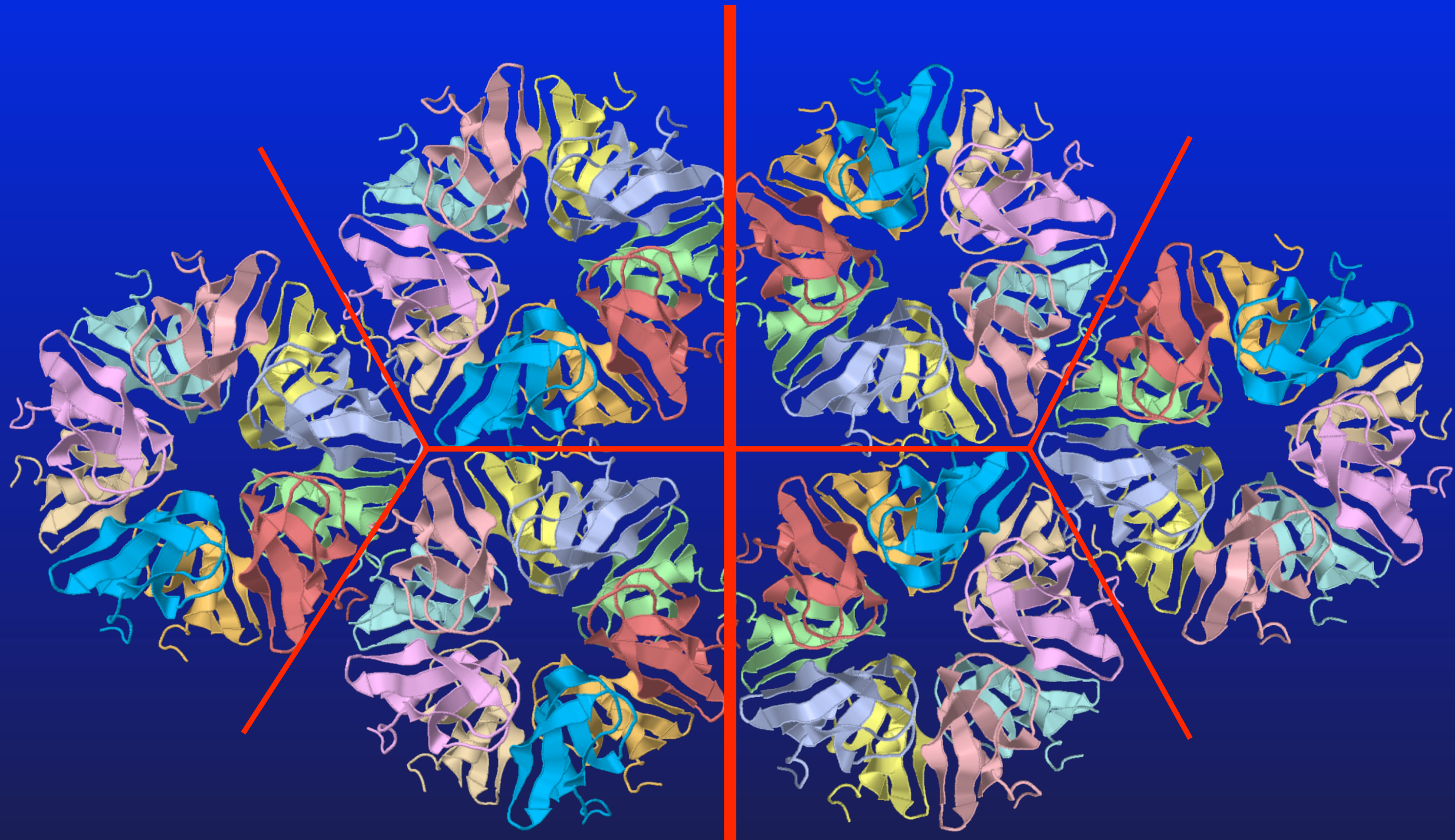


# A simple thing to do



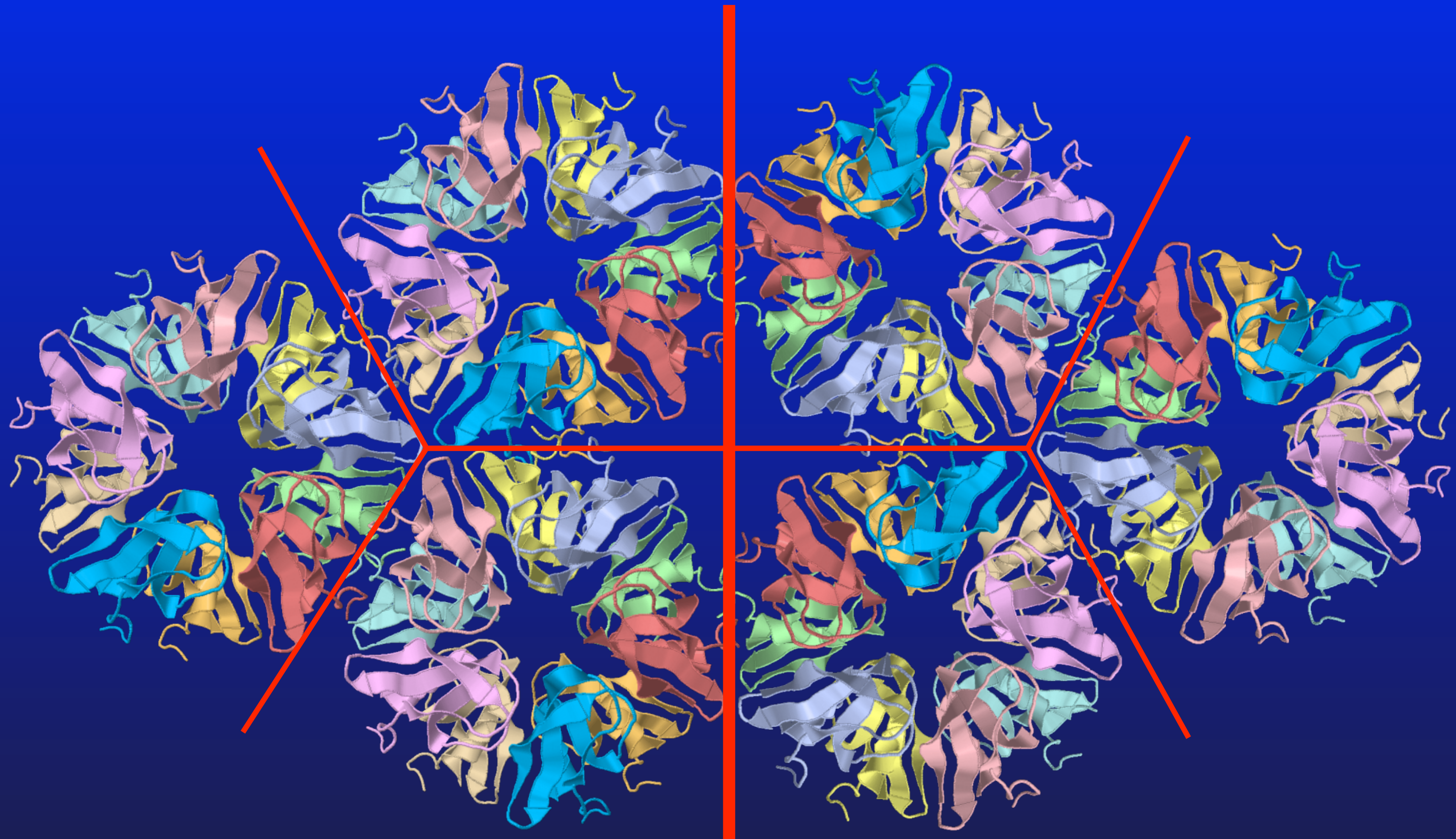


# A simple thing to do

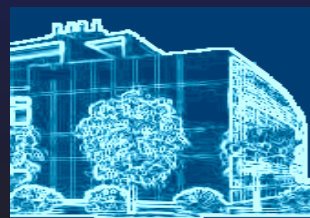




# A simple thing to do



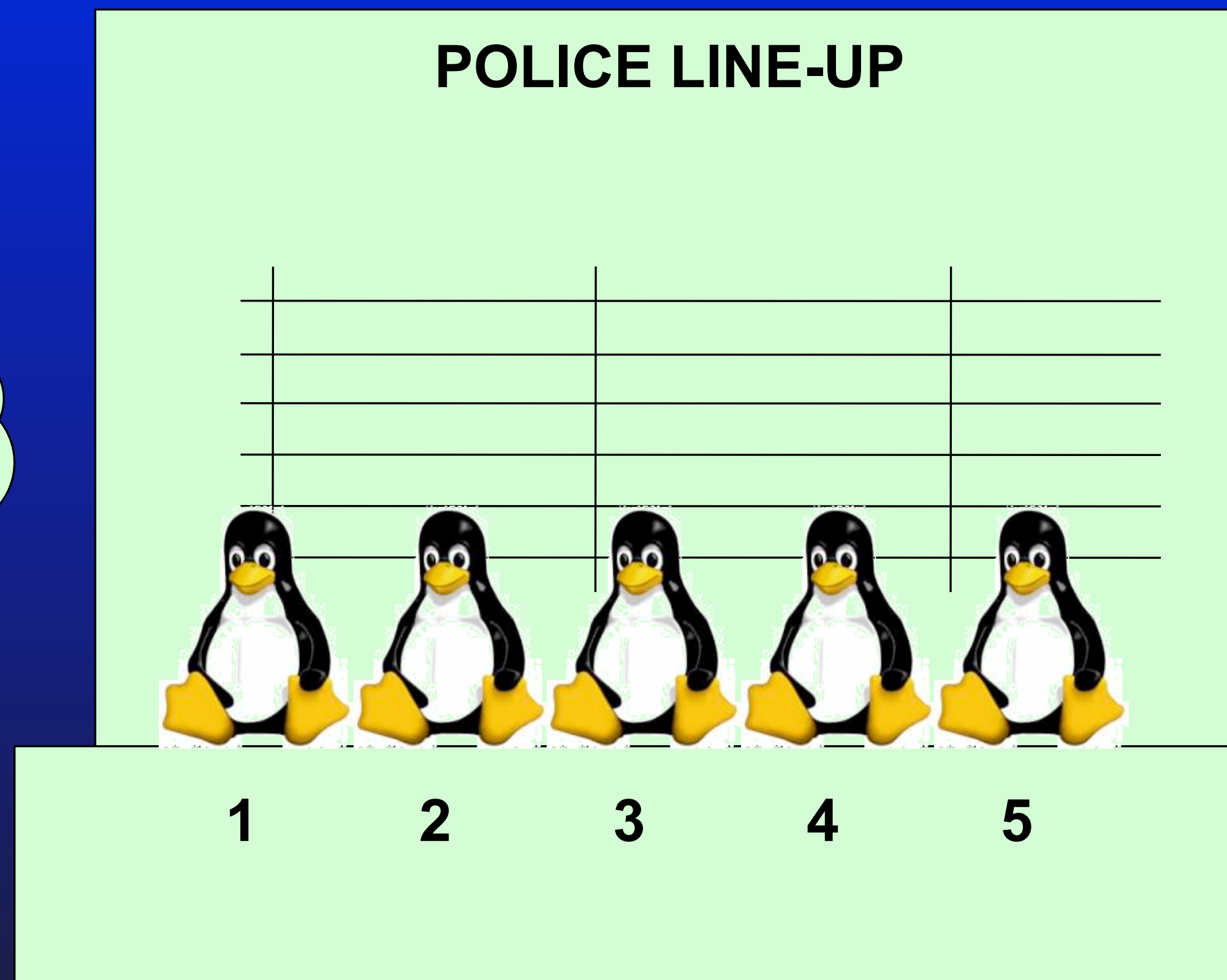
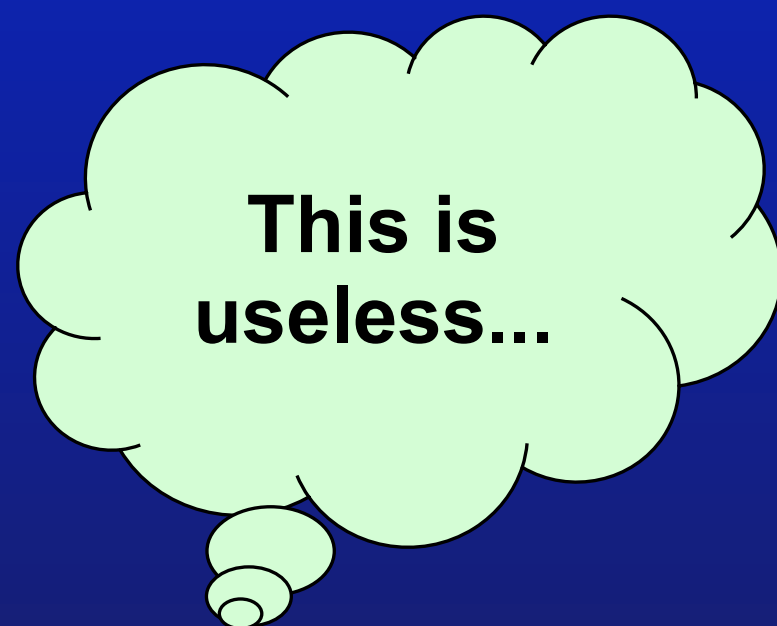
PQS server @ EBI (Kim Henrick) *Trends in Biochem. Sci.* (1998) 23, 358  
PITA server @ EBI (Hannes Ponstingl) *J. Appl. Cryst.* (2003) 36, 1116



Research Complex at Harwell



# A Common Identification Problem





# What is a significant interface?

Depends on the problem.

Protein functionality: the interface should be engaged in *any* sort of **interaction**, including transient short-living protein-ligand and protein-protein etc. associations. Obviously important properties:

- Affinity (comes from area, hydrophobicity, electrostatics, H-bond density etc.)

and properties that may be important for *reaction pathway and dynamics*:

- Aminoacid composition
- Geometrical complementarity
- Overall shape, compactness
- Charge distribution
- etc.



Research Complex at Harwell



# What is a significant interface?

Depends on the problem.

Protein functionality: the interface should be engaged in *any* sort of **interaction**, including transient short-living protein-ligand and protein-protein etc. associations. Obviously important properties:

- Affinity (comes from area, hydrophobicity, electrostatics, H-bond density etc.)

and properties that may be important for *reaction pathway and dynamics*:

- Aminoacid composition
- Geometrical complementarity
- Overall shape, compactness
- Charge distribution
- etc.

Stable macromolecular complexes, PQS: the interface should make a sound **binding**. Important properties:

- Sufficient free energy of binding
- something else?



Research Complex at Harwell




# What is a significant interface?

Depends on the problem.

Protein functionality: the interface should be engaged in *any* sort of **interaction**, including transient short-living protein-ligand and protein-protein etc. associations. Obviously important properties:

- Affinity (comes from area, hydrophobicity, electrostatics, H-bond density etc.)

and properties that may be important for *reaction pathway and dynamics*:

- 
- Aminoacid composition
  - Geometrical complementarity
  - Overall shape, compactness
  - Charge distribution
  - etc.

Stable macromolecular complexes, PQS: the interface should make a sound **binding**. Important properties:

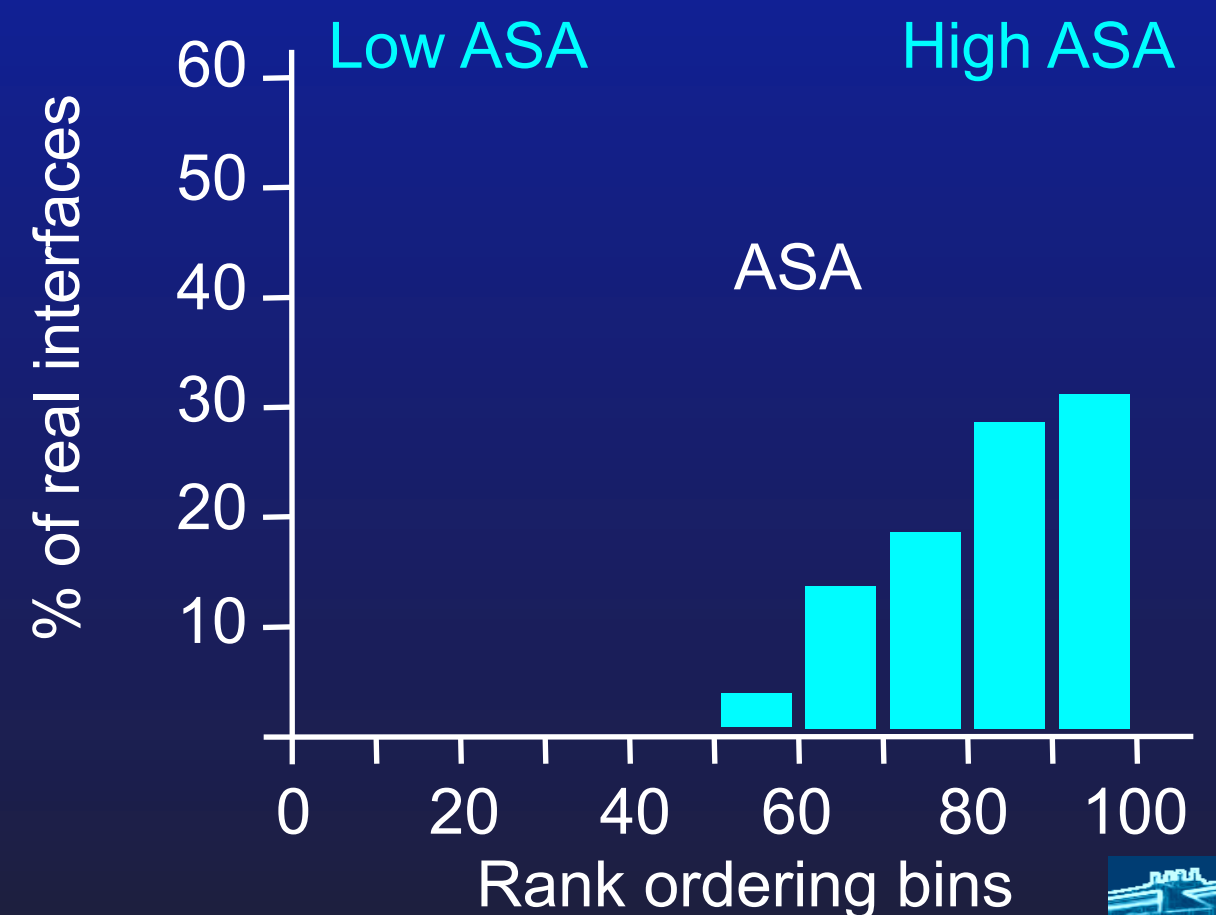
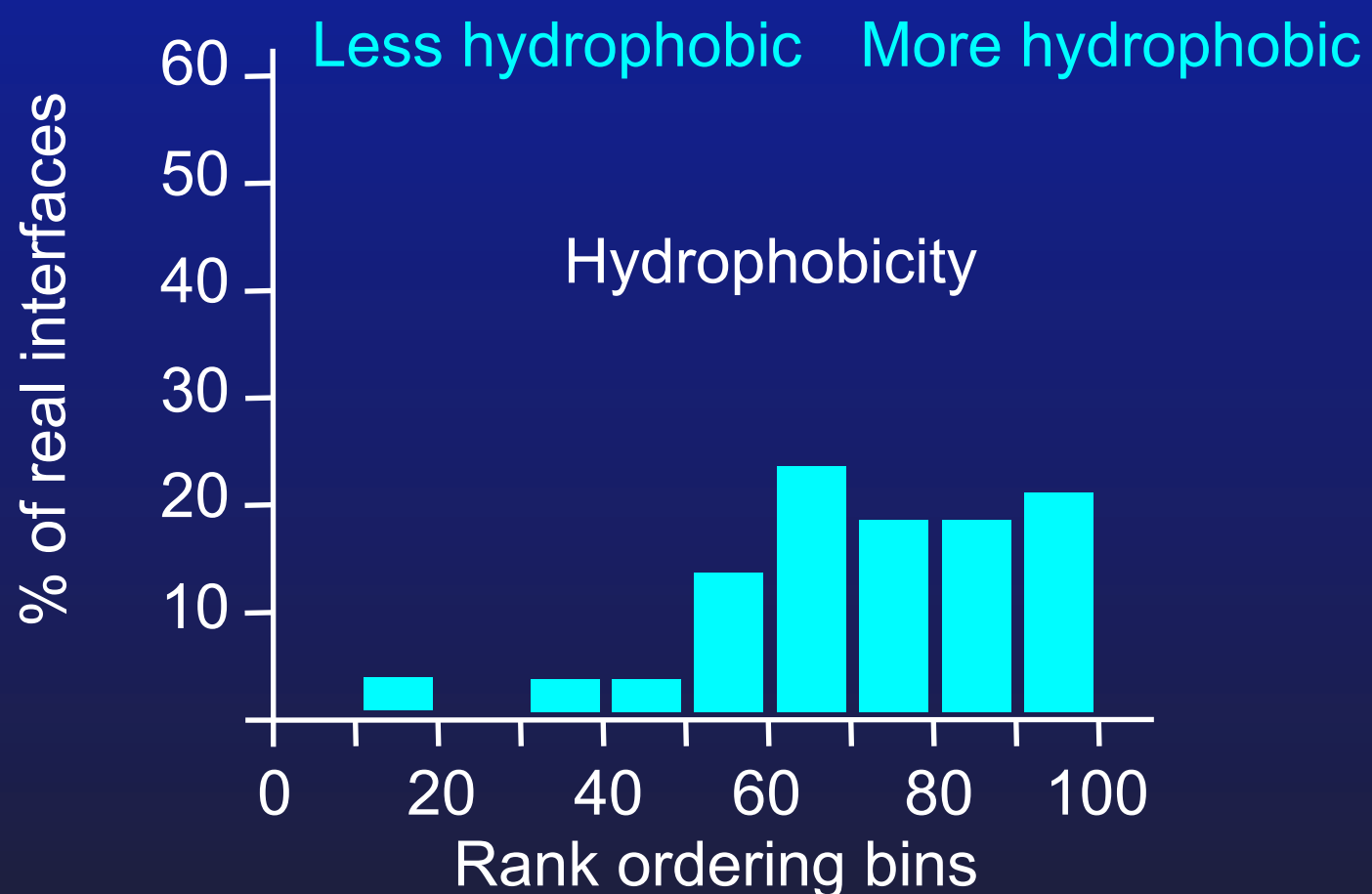
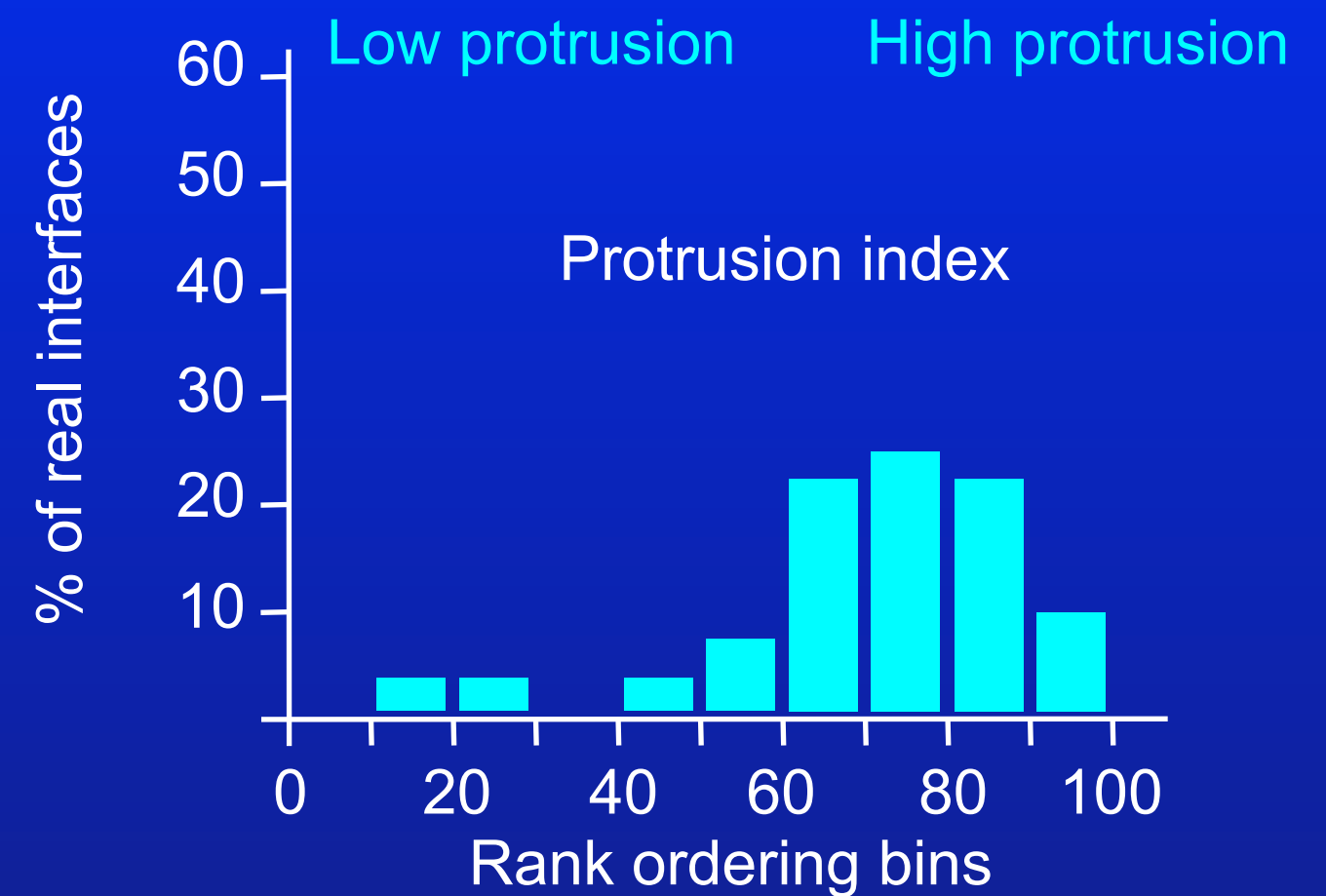
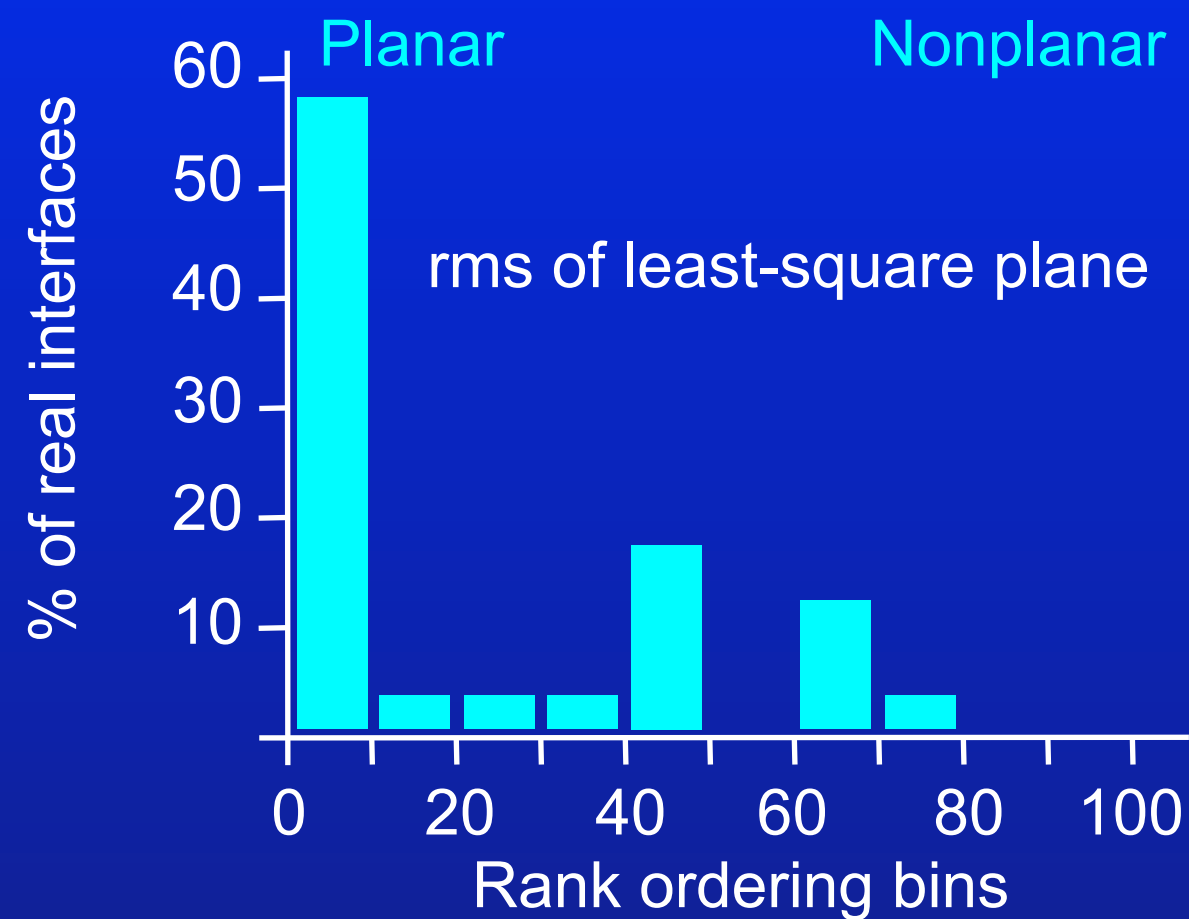
- Sufficient free energy of binding
- something else?



Research Complex at Harwell



# Real and superficial protein interfaces



Jones, S. & Thornton, J.M. (1996) Principles of protein-protein interactions, *Proc. Natl. Acad. Sci. USA*, **93**, 13-20.



Research Complex at Harwell



# Real and superficial protein interfaces

◆ “No single parameter absolutely differentiates the interfaces from all other surface patches”

Jones, S. & Thornton, J.M. (1996) Principles of protein-protein interactions, *Proc. Natl. Acad. Sci. USA*, 93, 13-20.



Research Complex at Harwell

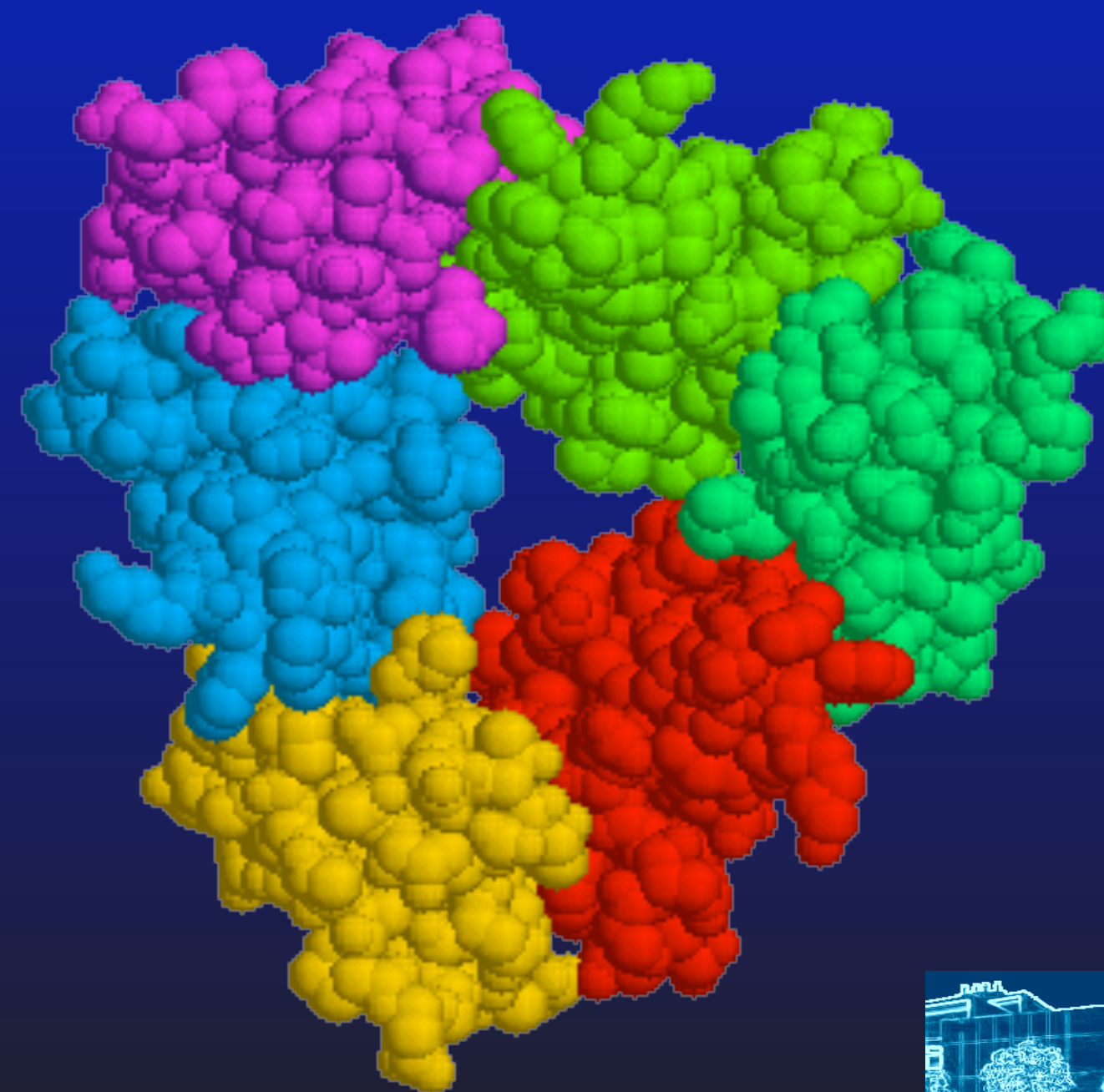


# Real and superficial protein interfaces

- ◆ “No single parameter absolutely differentiates the interfaces from all other surface patches”

Jones, S. & Thornton, J.M. (1996) Principles of protein-protein interactions, *Proc. Natl. Acad. Sci. USA*, 93, 13-20.

- ◆ Formation of  $N > 2$  -meric complexes is most probably a corporate process involving a set of interfaces. Therefore significance of an interface should not be detached from the context of macromolecular complex





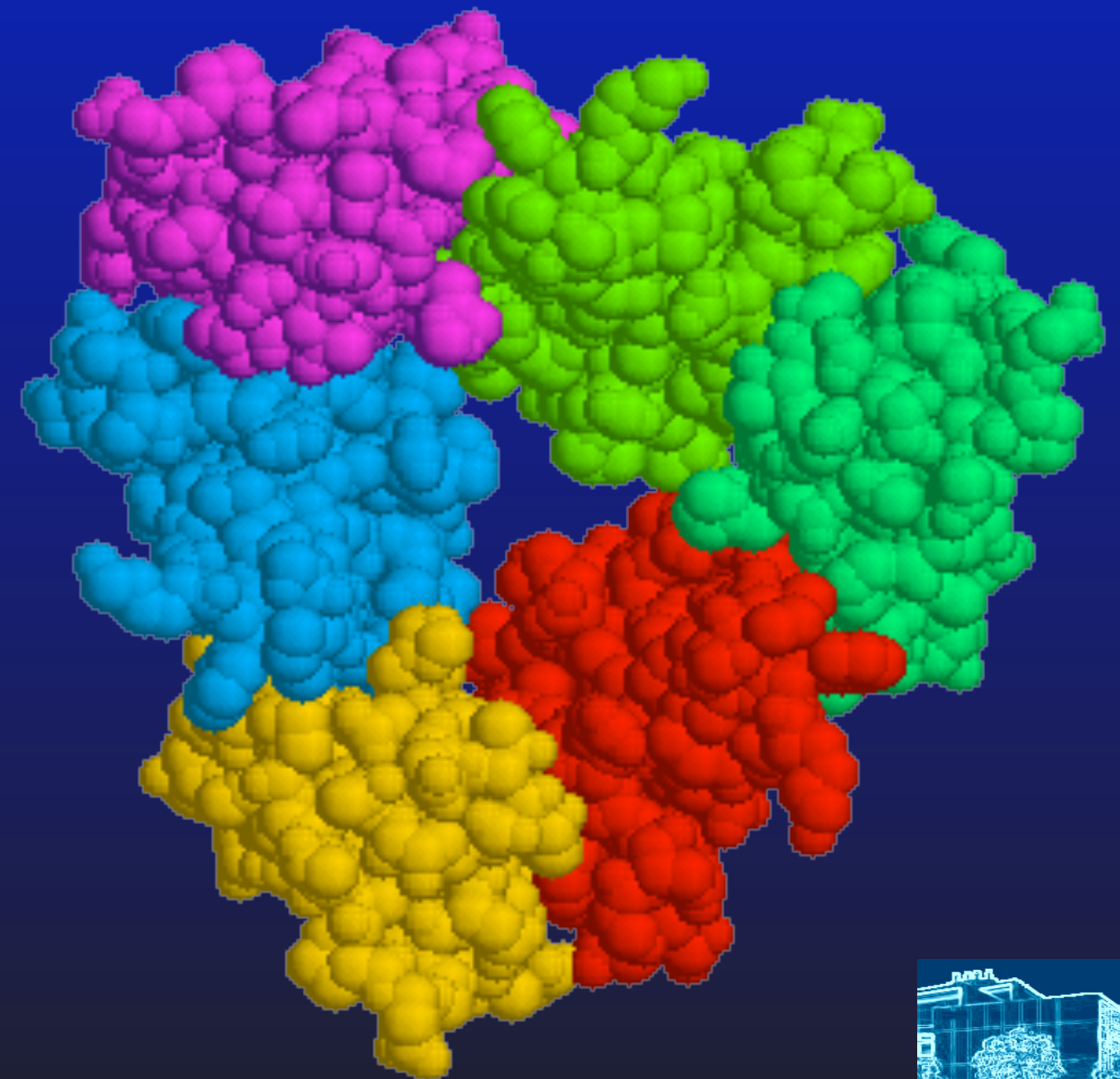
# Real and superficial protein interfaces

- ◆ “No single parameter absolutely differentiates the interfaces from all other surface patches”

Jones, S. & Thornton, J.M. (1996) Principles of protein-protein interactions, *Proc. Natl. Acad. Sci. USA*, 93, 13-20.

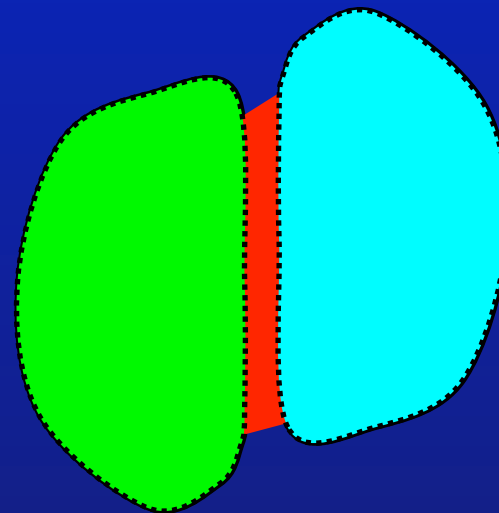
- ◆ Formation of  $N > 2$  -meric complexes is most probably a corporate process involving a set of interfaces. Therefore significance of an interface should not be detached from the context of macromolecular complex
- ◆ “...the type of complexes need to be taken into account when characterising interfaces between them.”

Jones, S. & Thornton, J.M., *ibid.*



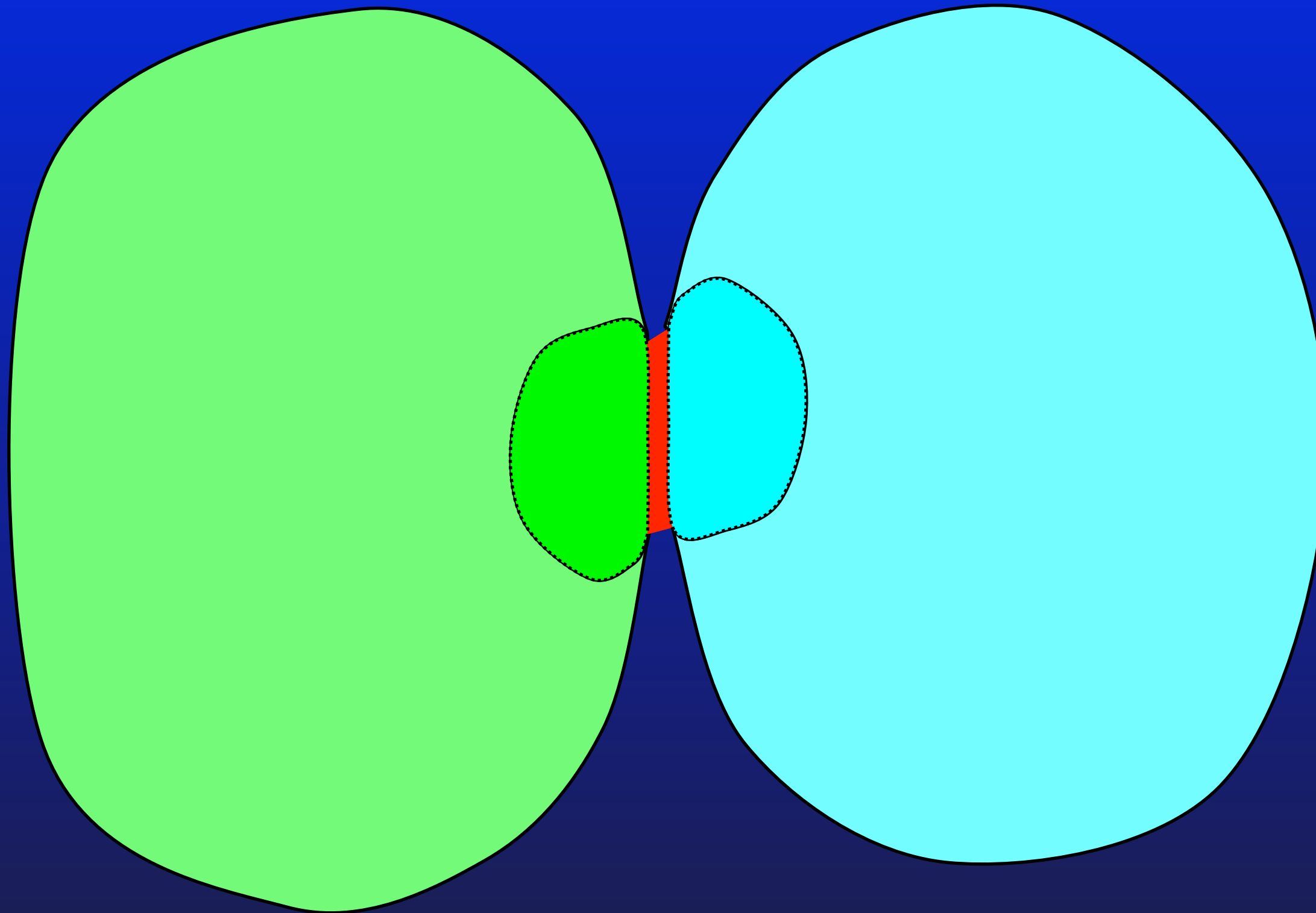


# Is there a (good) measure of interface significance at all?



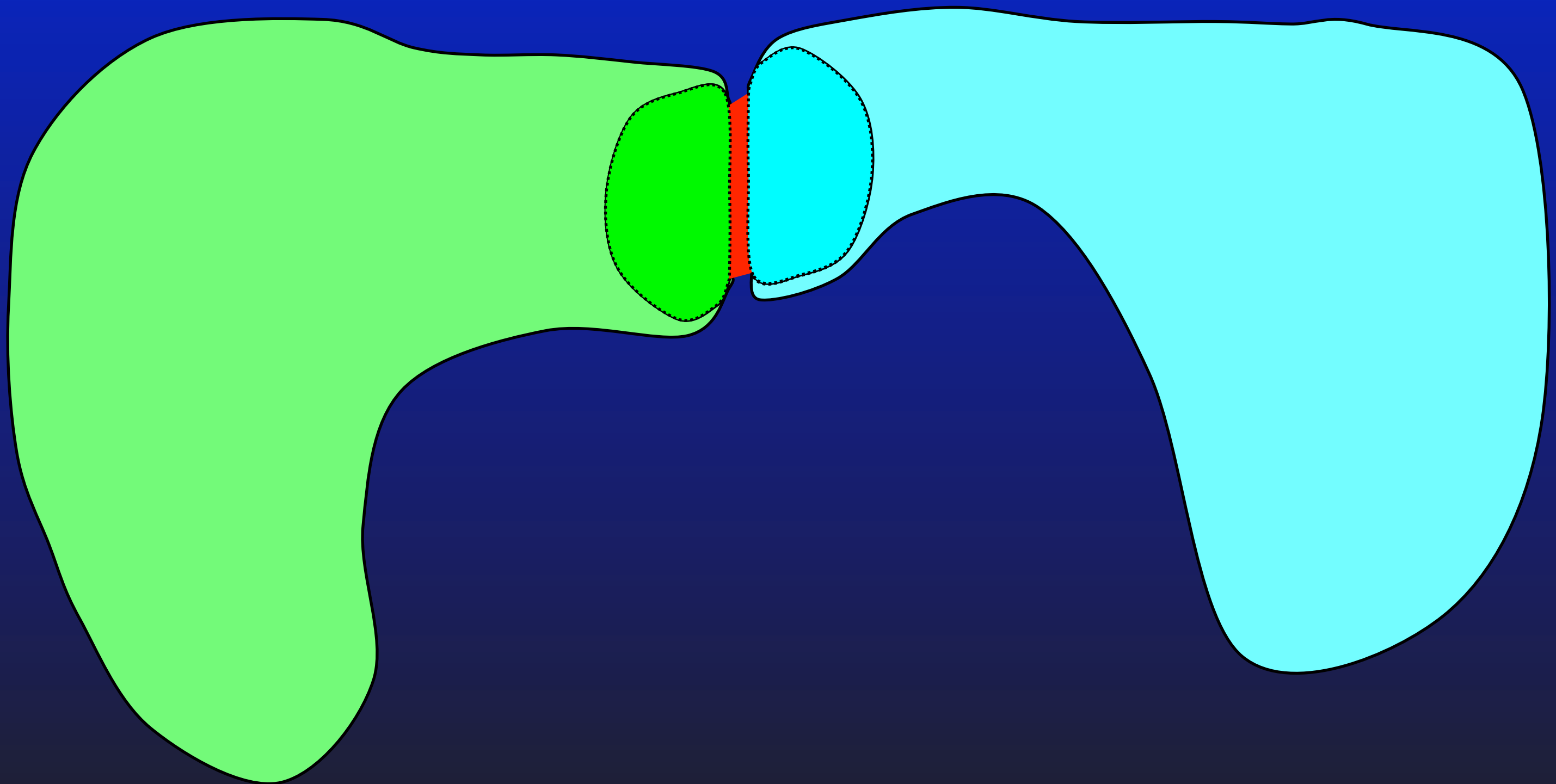
Research Complex at Harwell

# Is there a (good) measure of interface significance at all?



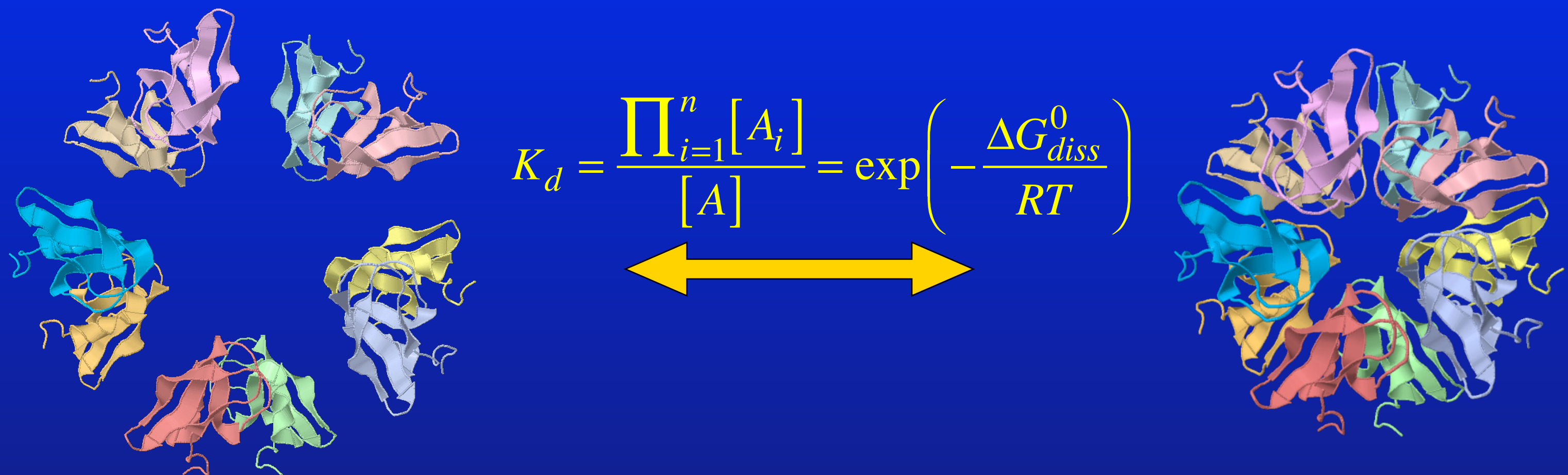


# Is there a (good) measure of interface significance at all?



Research Complex at Harwell

# Chemistry Say



- ◆ It is not properties of individual interfaces but rather chemical stability of complexes in general that really matters
- ◆ *In solution, macromolecular units will most likely associate into largest complexes that are still stable*
- ◆ A complex is stable if its Gibbs free energy of dissociation is positive:

$$\Delta G^0_{diss} = -\Delta G_{int} - T \Delta S > 0$$



Research Complex at Harwell



$$\Delta G_0 = -\Delta G_{\text{int}} - T\Delta S > 0$$



# Protein affinity

*Solvation energy of  
protein complex*



*Solvation energies of  
dissociated subunits*



*Free energy  
of H-bond  
formation*



*Free energy  
of salt bridge  
formation*



$$\Delta G_{\text{int}} = \Delta G_{\text{sol}}(A_1, A_2 \dots A_n) - \sum_{i=1}^n \Delta G_{\text{sol}}(A_i) - E_{hb} N_{hb} - E_{sb} N_{sb}$$

*Number of  
H-bonds  
between  
dissociated  
subunits*



*Number  
of salt  
bridges  
between  
dissociate  
d subunits*



Research Complex at Harwell



# Protein affinity

Solvation energy of  
protein complex

Solvation energies of  
dissociated subunits

Free energy  
of H-bond  
formation

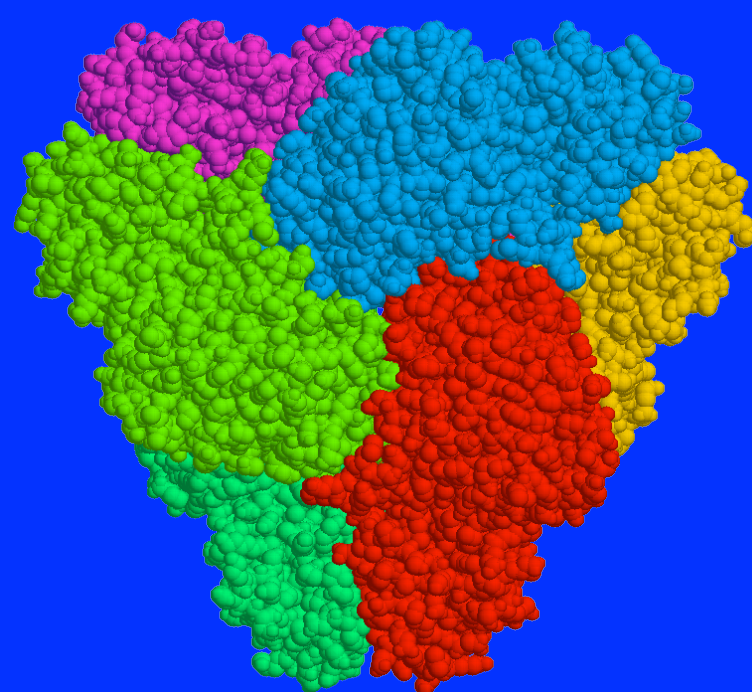
Free energy  
of salt bridge  
formation

$$\Delta G_{\text{int}} = \Delta G_{\text{sol}}(A_1, A_2 \dots A_n) - \sum_{i=1}^n \Delta G_{\text{sol}}(A_i) - E_{hb} N_{hb} - E_{sb} N_{sb}$$

Number of  
H-bonds  
between  
dissociated  
subunits

Number  
of salt  
bridges  
between  
dissociate  
d subunits

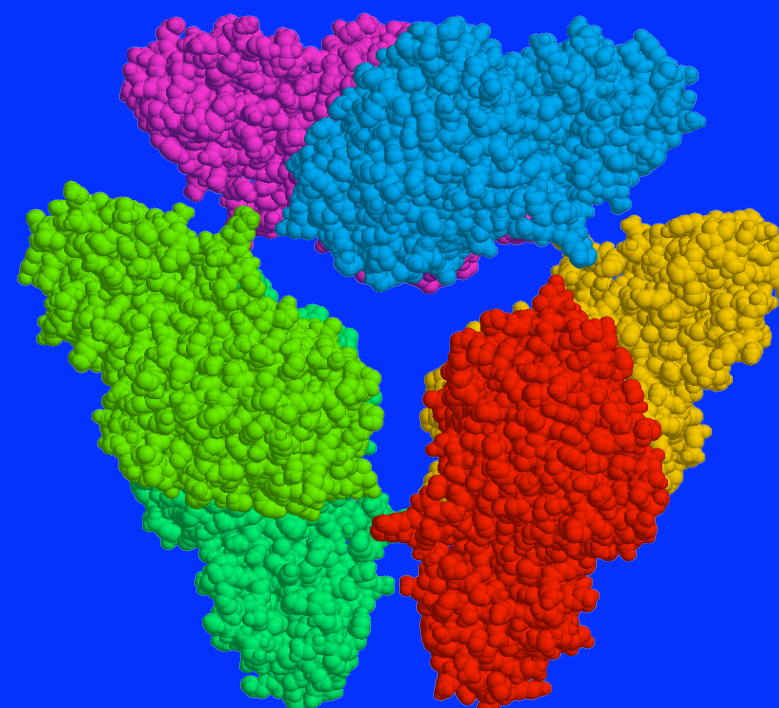
Choice of dissociation subunits:



$(A_1 A_2 A_3)$

Dissociation into  
stable subunits with  
minimum

$\Delta G_0$



$A_1 + A_2 + A_3$



# Solvation free energy

Eisenberg, D. & McLachlan, A.D. (1986)  
*Nature* 319, 199-203.

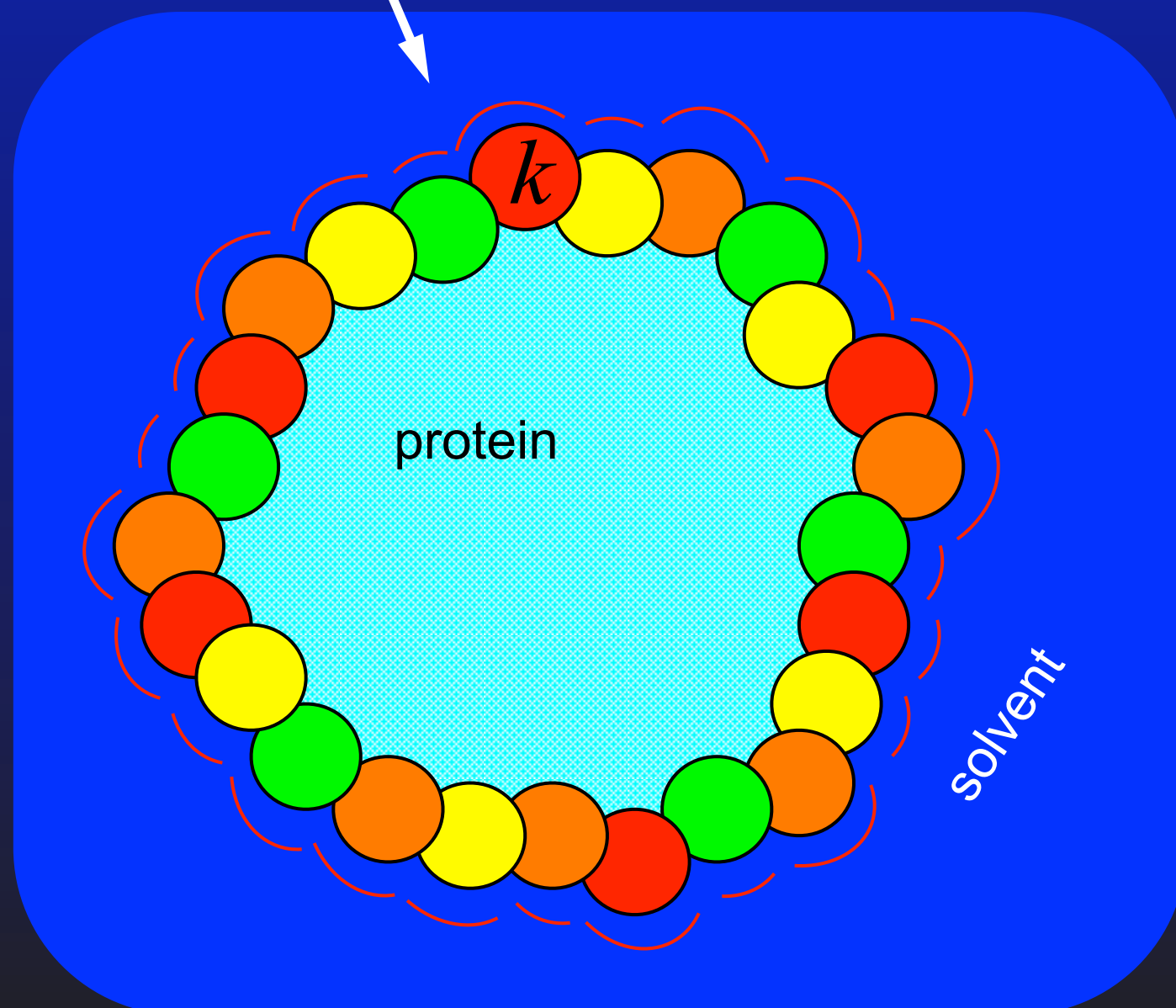
Atomic solvation  
parameters

Atom's accessible  
surface area

$$\Delta G_{sol}(A) = \sum_k \Delta \sigma_k (a_k - a_k^r)$$

Atom's accessible surface area  
in the reference (unfolded) state

$a_k$

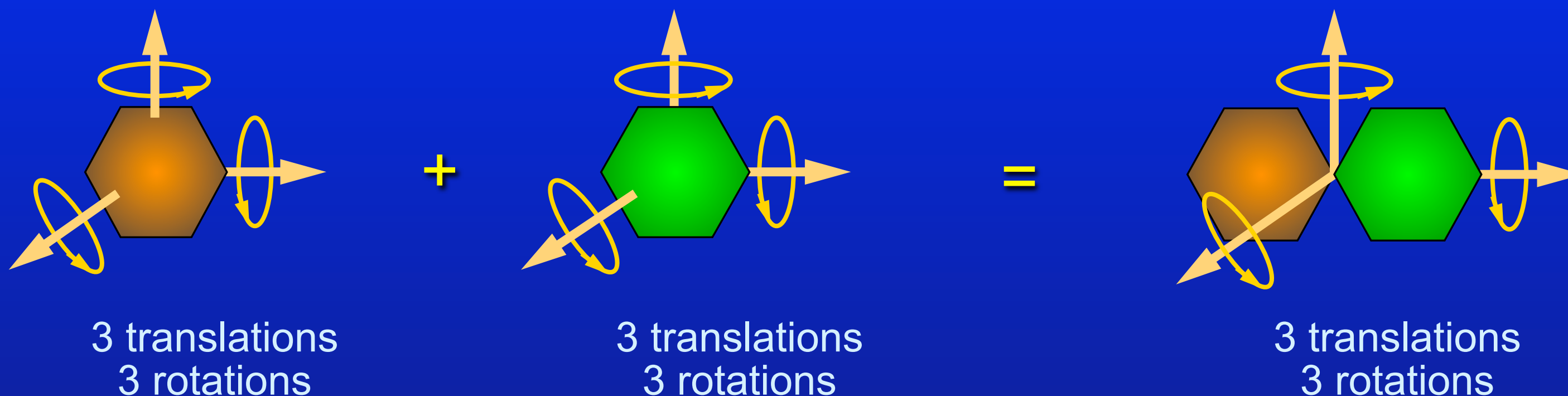


In this model, binding energy is a function  
of individual interfaces.





# Entropy of macromolecules in solution



Cost of translations:  $S_{trans}(m) \approx c_t + \frac{3R}{2} \log(m)$

*Mass*

Murray C.W. and Verdonk M.L.  
(2002) *J. Comput.-Aided Mol. Design* 16, 741-753.

Cost of rotations:  $S_{rot}(\hat{I}, \sigma_S) \approx c_r + \frac{R}{2} \log(I_1 I_2 I_3 / \sigma_S^2)$

*Symmetry number*  
*Moments of inertia*

Cost of side chain motion:  $S_{surf}(a) \approx Fa$

*Solvent-accessible surface area*

$c_t$ ,  $c_r$  and  $F$  are semiempirical parameters



# Entropy of macromolecular dissociation

$$\Delta S = \sum_{i=1}^n S(A_i) - S(A_1, A_2, \dots, A_n)$$



Research Complex at Harwell



# Entropy of macromolecular dissociation

$$\begin{aligned} \Delta S &= \sum_{i=1}^n S(A_i) - S(A_1, A_2, \dots, A_n) \\ &= (n-1)C + \frac{3R}{2} \log \left( \frac{\prod_i m_i}{\sum_i m_i} \right) + \\ &\quad \frac{R}{2} \log \left( \frac{\prod_i \prod_k I_k(A_i) / \sigma_S^2(A_i)}{\prod_k I_k(A_1, \dots, A_n) / \sigma_S^2(A_1, \dots, A_n)} \right) + Fa_{buried} \end{aligned}$$

*Mass of i-th subunit*

*k-th principal moment of inertia of i-th subunit*

*Empirical parameter*

*Empirical parameter*



Research Complex at Harwell

# Entropy of macromolecular dissociation

$$\begin{aligned}
 \Delta S &= \sum_{i=1}^n S(A_i) - S(A_1, A_2, \dots, A_n) \\
 &= (n-1)C + \frac{3R}{2} \log \left( \frac{\prod_i m_i}{\sum_i m_i} \right) + \\
 &\quad \frac{R}{2} \log \left( \frac{\prod_i \prod_k I_k(A_i) / \sigma_S^2(A_i)}{\prod_k I_k(A_1, \dots, A_n) / \sigma_S^2(A_1, \dots, A_n)} \right) + Fa_{buried}
 \end{aligned}$$

*Mass of i-th subunit*  
*k-th principal moment of inertia of i-th subunit*  
*Empirical parameter*  
*Empirical parameter*

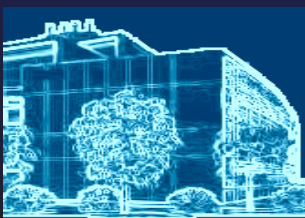
By its very nature, entropy of dissociation is function of protein complex rather than that of individual interfaces.





# Entropy of dissociation

- ◆ Drives thermodynamic systems towards most disordered (dissolved) state.
- ◆ Makes bias towards less symmetric states. However, in practice, this is outweighed by binding energy, which usually maximises in most symmetric states.
- ◆ Responsible for complex “instability”.



Research Complex at Harwell











# Detection of Biological Units in Crystals: PISA summary

1. Enumerate all possible assemblies in crystal packing, subject to crystal properties: space symmetry group, geometry and composition of the Asymmetric Unit

- Achieved with graph-theoretical techniques, by representing crystal as an infinite periodic graph of connected macromolecules

2. Evaluate assemblies for chemical stability:

$$\Delta G_0 = -\Delta G_{\text{int}} - T\Delta S > 0$$

3. Leave only sets of stable assemblies in the list, and range by chances to be a biological unit:

- Larger assemblies take preference
- Single-assembly sets take preference
- Otherwise, assemblies with higher  $\Delta G_0$  take preference





Contents

Data

▶ Monomers

▶ Interfaces

Assembly stock

▼ Crystal splits

▶ Stable splits

▶ Metastable splits

▶ Unstable splits

▶ Unsplitted

QtPISA [1E94.pisa]

File:

/Users/Eugene/Projects/PISA/setup/1E94.pdb

Title:

HSLV-HSLU FROM E.COLI

Space group:

P 63 2 2

Resolution:

2.8

Cell:

172.022 172.022 276.569 90 90 120

Cell volume [Å<sup>3</sup>]:

7.088e+06

▼ ASU (File) contents:

Protein chains:

6 (6)

DNA/RNA chains:

0 (0)

Ligands:

2 (2)

NCS-mates:

0

Excluded ligands: None

Ligand processing: Auto ☐ change



Contents

- Data
- Monomers
- Interfaces
- Assembly stock
- Crystal splits
  - Stable splits
  - Metastable splits
  - Unstable splits
  - Unsplitted

File: /Users/Eugen

Title: HSLV-HSLU FR

Space group: P 63 2 2

Resolution: 2.8

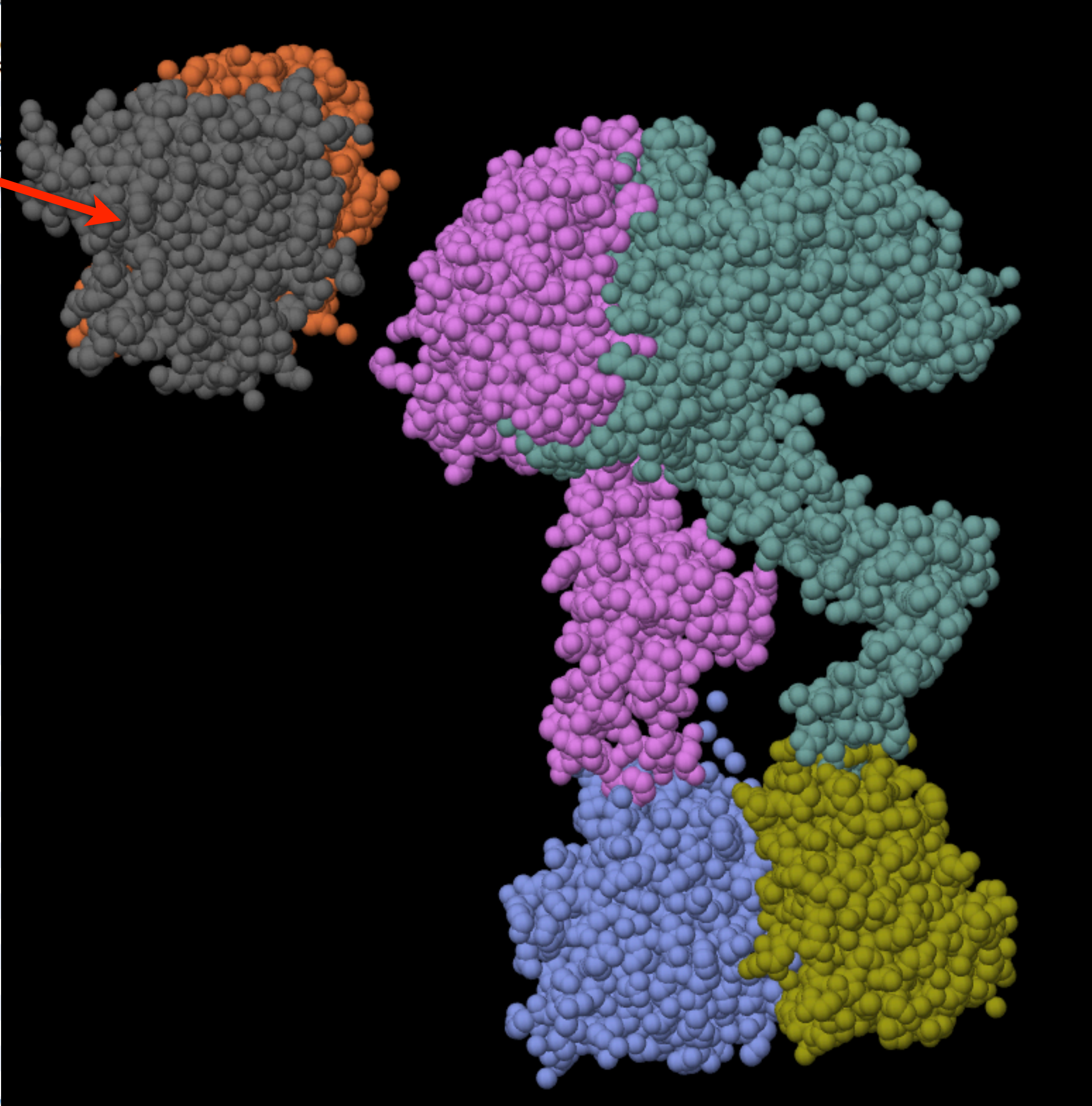
Cell: 172.022 172.022 172.022

Cell volume [Å<sup>3</sup>]: 7.088e+06

ASU (File) contents:

- Protein chains: 6 (6)
- DNA/RNA chains: 0 (0)
- Ligands: 2 (2)
- NCS-mates: 0

CCP4MG version 2.7.3



Excluded ligands: None



QtPISA [1E94.pisa]


Contents

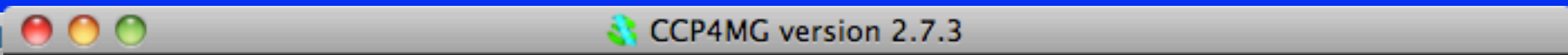
- Data
- Monomers
- Interfaces
- Assembly stock
- ▼ Crystal splits
  - Stable splits
  - Metastable splits
  - Unstable splits
  - Unsplitted

The following quaternary structures appear to be stable in solution

	Split No.	Size	Type	ASA	BSA	dG_diss	dG0	Formula	Composition
1	1	12	1	77218.4	25619.7	31.6	2.6	A(12)	A(6)B(6)
2		12	1	77204.3	25139.2	22.4	3.7	A(12)	C(6)D(6)
3		6	2	104504.1	32216.1	69.2	11.5	A(6)a(6)	E(3)F(3)[ANP](6)
4	2	12	1	77218.4	25619.7	31.6	2.6	A(12)	A(6)B(6)
5		12	1	77204.3	25139.2	22.4	3.7	A(12)	C(6)D(6)
6		1	3	21852.2	852.8	0.0	0.0	Aa	E[ANP]
7		1	3	22032.3	836.1	0.0	0.0	Aa	F[ANP]
8	3	6	2	104504.1	32216.1	69.2	11.5	A(6)a(6)	E(3)F(3)[ANP](6)
9		1	4	8561.7	0.0	0.0	0.0	A	B
10		1	4	8459.0	0.0	0.0	0.0	A	C
11		1	4	8598.2	0.0	0.0	0.0	A	D
12		1	4	8578.0	0.0	0.0	0.0	A	A





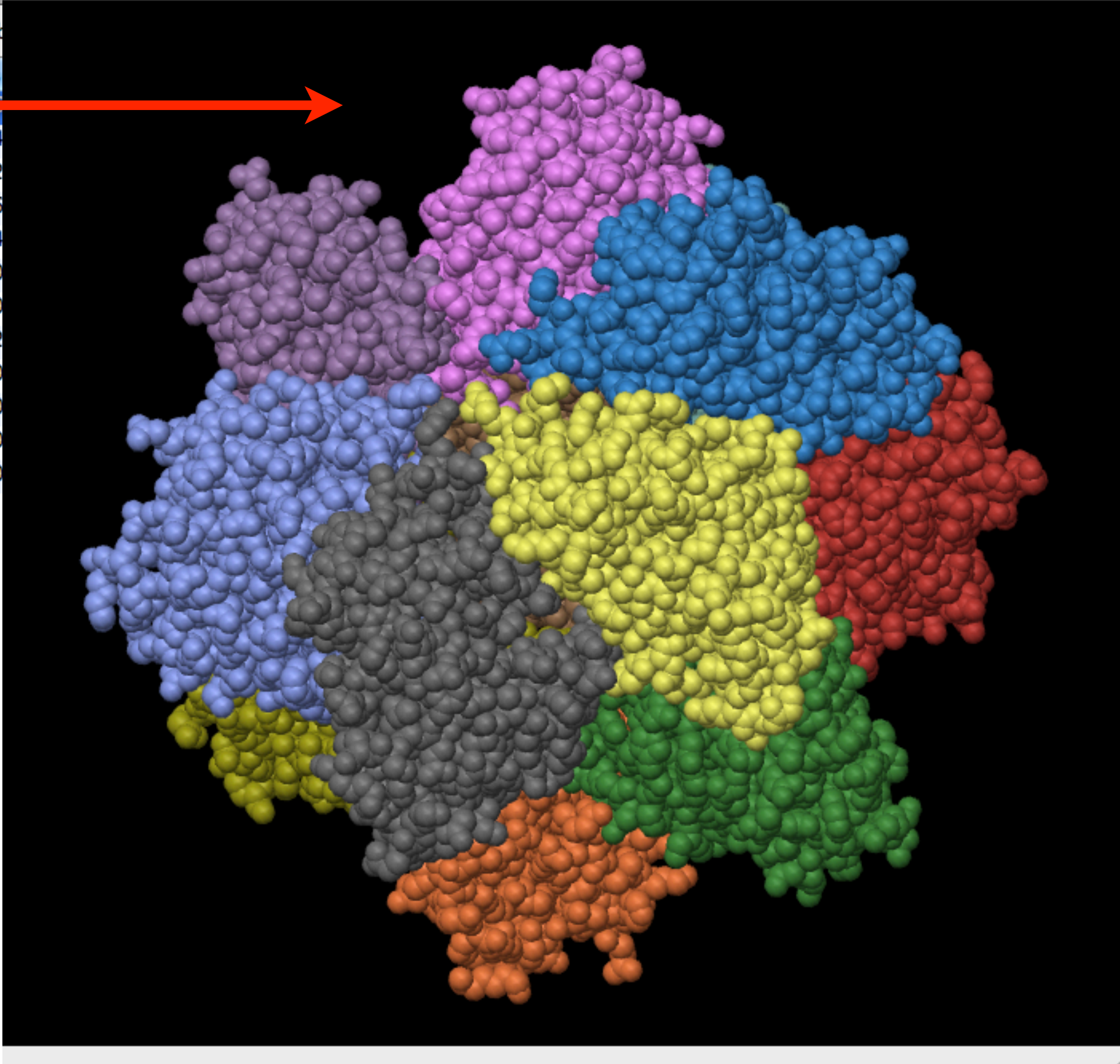


Contents

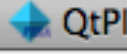

- Data
- Monomers
- Interfaces
- Assembly stock
- ▼ Crystal splits
  - **Stable splits**
  - Metastable splits
  - Unstable splits
  - Unsplitted







The following quaternary structures appear to be stable

	Split No.	Size	Type	ASA	BSA	dG_diss
1	1	12	1	77218.4	25619.7	31.6
2		12	1	77204.3	25139.2	22.4
3		6	2	104504.1	32216.1	69.2
4	2	12	1	77218.4	25619.7	31.6
5		12	1	77204.3	25139.2	22.4
6		1	3	21852.2	852.8	0.0
7		1	3	22032.3	836.1	0.0
8	3	6	2	104504.1	32216.1	69.2
9		1	4	8561.7	0.0	0.0
10		1	4	8459.0	0.0	0.0
11		1	4	8598.2	0.0	0.0
12		1	4	8578.0	0.0	0.0





 QtP
  CCP4MG version 2.7.3

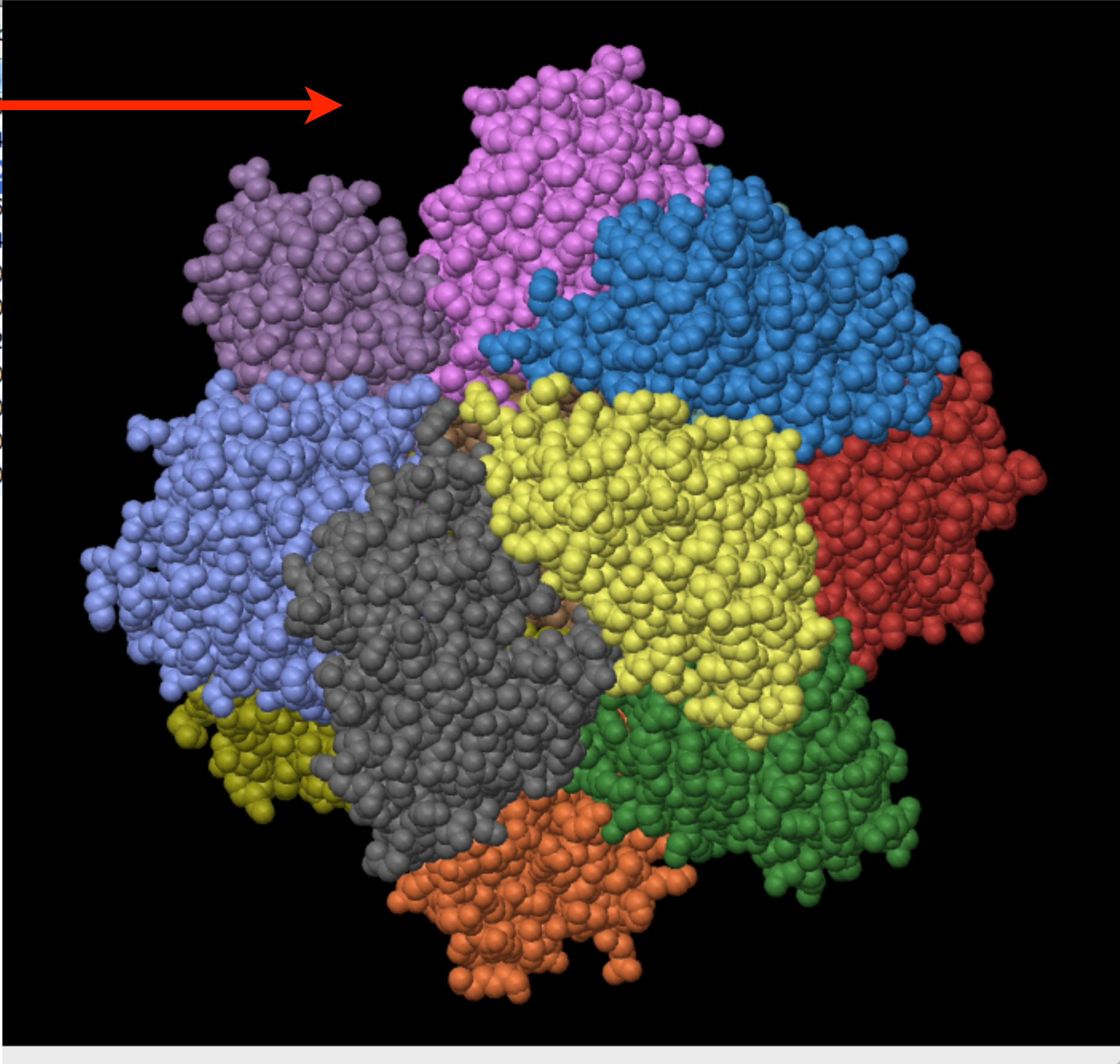







Contents

- Data
- Monomers
- Interfaces
- Assembly stock
- ▼ Crystal splits
  - Stable splits
  - Metastable splits
  - Unstable splits
  - Unsplitted

The following quaternary structures appear to be stable

	Split No.	Size	Type	ASA	BSA	dG_diss
1	1	12	1	77218.4	25619.7	31.6
2	12	1	1	77204.3	25139.2	22.4
3	6	2	104504.1	32216.1	69.2	
4	2	12	1	77218.4	25619.7	31.6
5	12	1	1	77204.3	25139.2	22.4
6	1	3	21852.2	852.8	0.0	
7	1	3	22032.3	836.1	0.0	
8	3	6	104504.1	32216.1	69.2	
9	1	4	8561.7	0.0	0.0	
10	1	4	8459.0	0.0	0.0	
11	1	4	8598.2	0.0	0.0	
12	1	4	8578.0	0.0	0.0	







QtP...

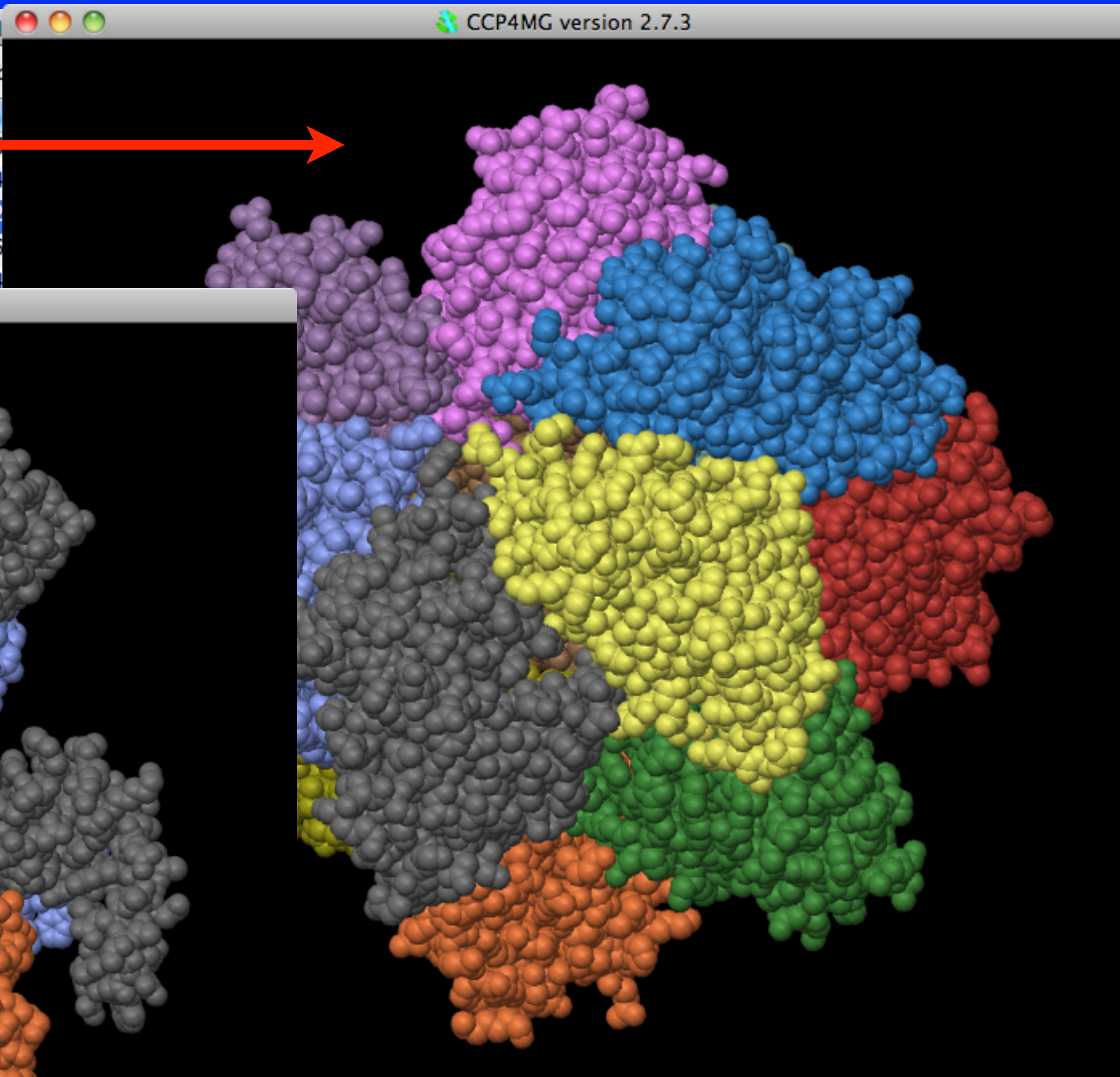
The following quaternary structures appear to be stable

	Split No.	Size	Type	ASA	BSA	dG_diss
1	1	12	1	77218.4	25619.7	31.6
2	12	1	77204.3	25139.2	22.4	
3	6	2	104504.1	32216.1	69.2	
4	2	12	1	77218.4	25619.7	31.6
5	12	1	77204.3	25139.2	22.4	

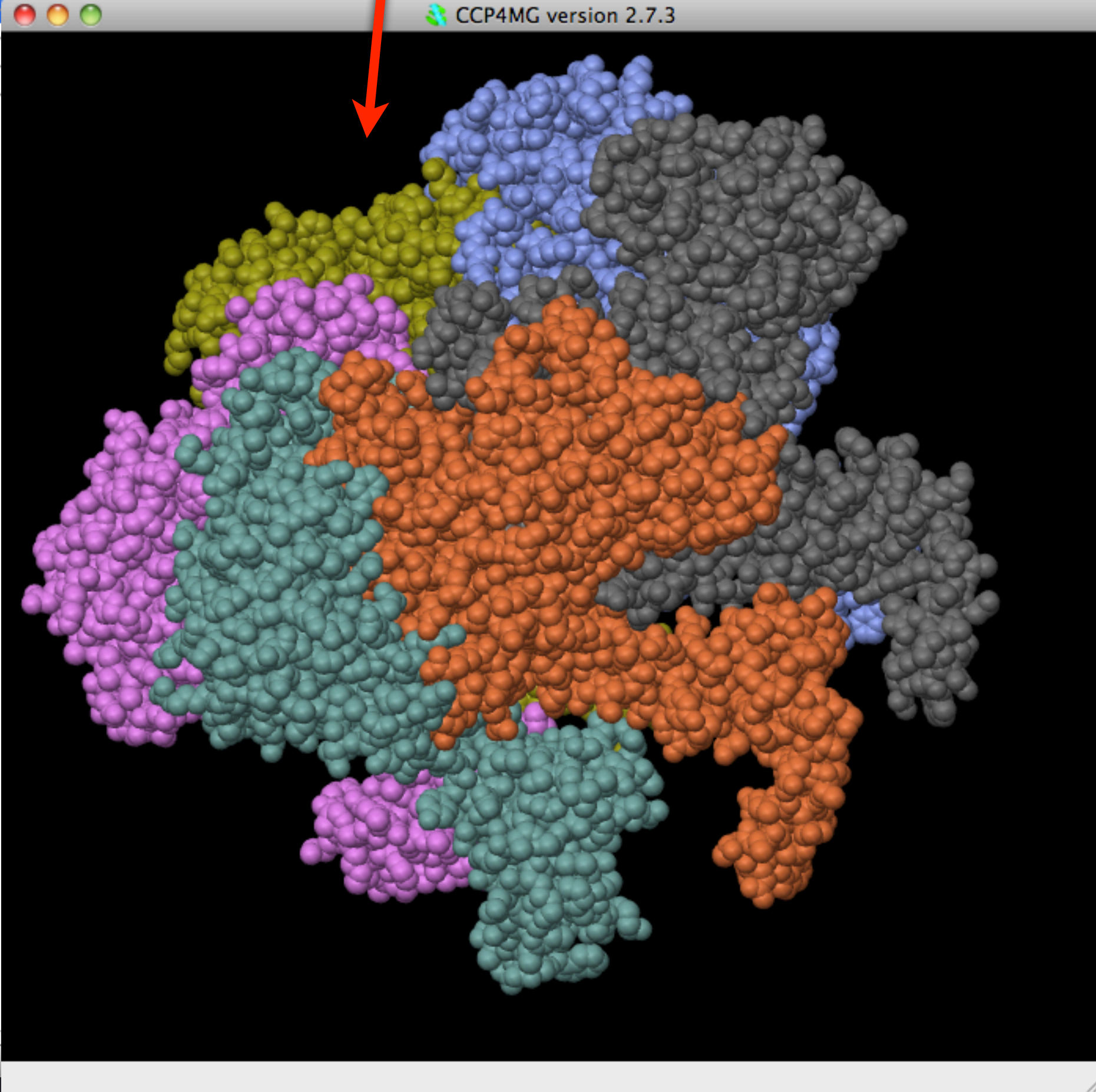
Contents


- Data
- ▶ Monomers
- ▶ Interfaces
- Assembly stock
- ▼ Crystal splits

CCP4MG version 2.7.3



CCP4MG version 2.7.3

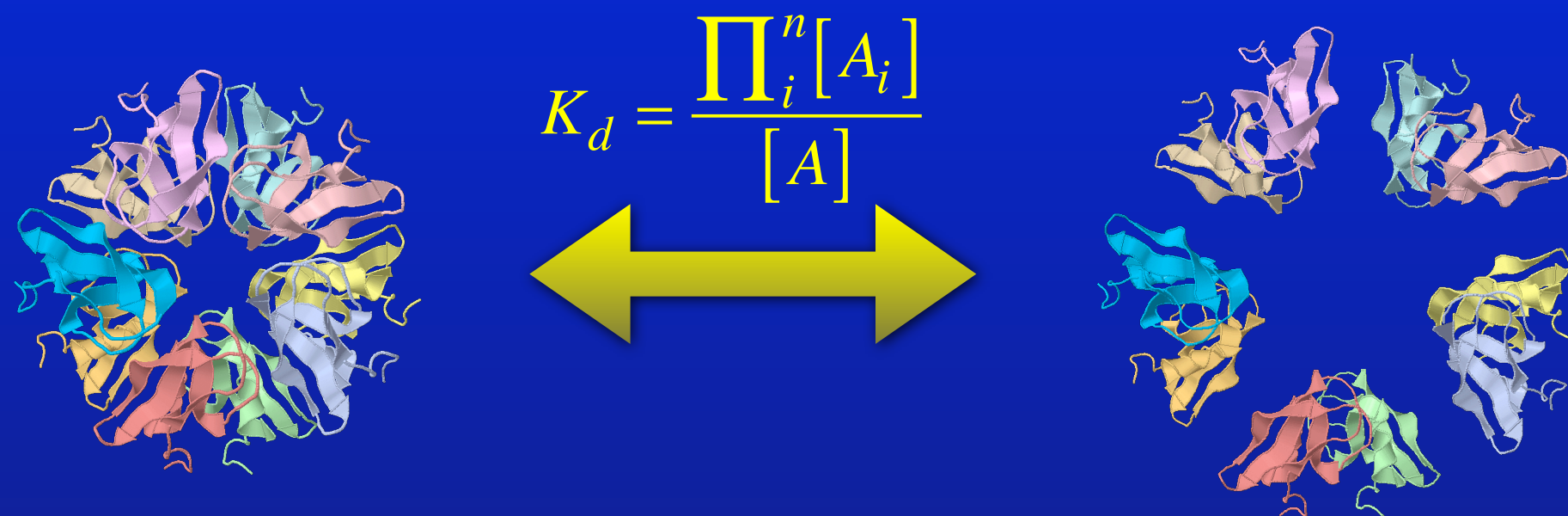






# What is “A Stable Complex?”

- ★ Chemical systems always move towards equilibrium:



- ★ PISA reports  $\Delta G_0 = -RT \log K_d$ , how to interpret?

♦ *In general, if equilibrium is shifted to the left ( $K_d < 1$ ), the complex is stable.*

- But does this always mean that stable complex has higher concentration than the dissociates? - no it does not
- And this depends on the concentration anyway? - yes it does
- And it also depends on the dissociation pattern (dissociation into monomers, dimers, trimers etc.)? How to know the pattern?

- *the pattern is, in essence, the minimum free energy route*

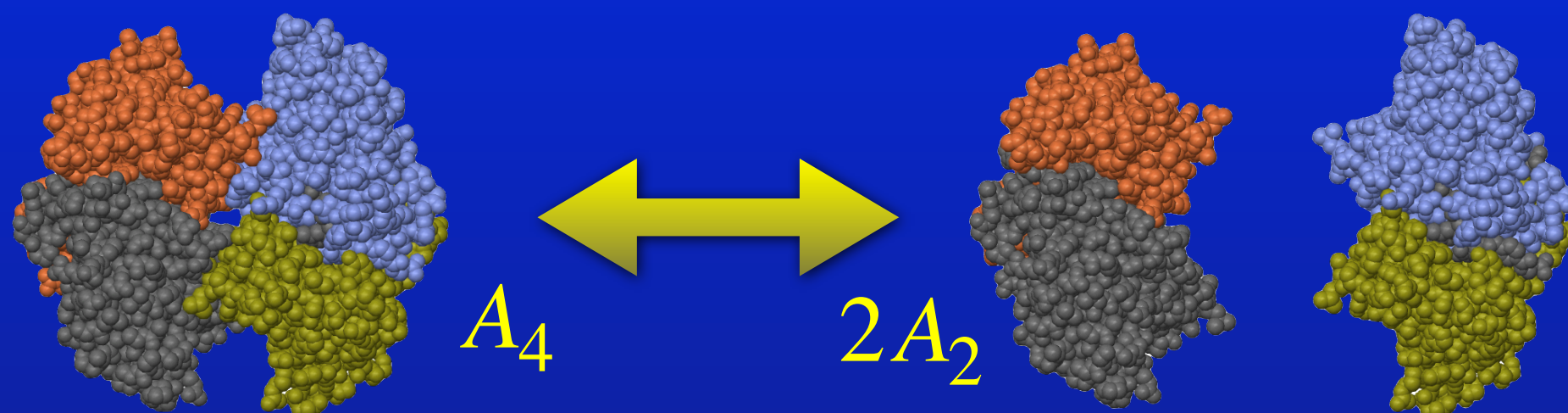
- *does it not depend on concentration (temperature, pH, etc.), too?*



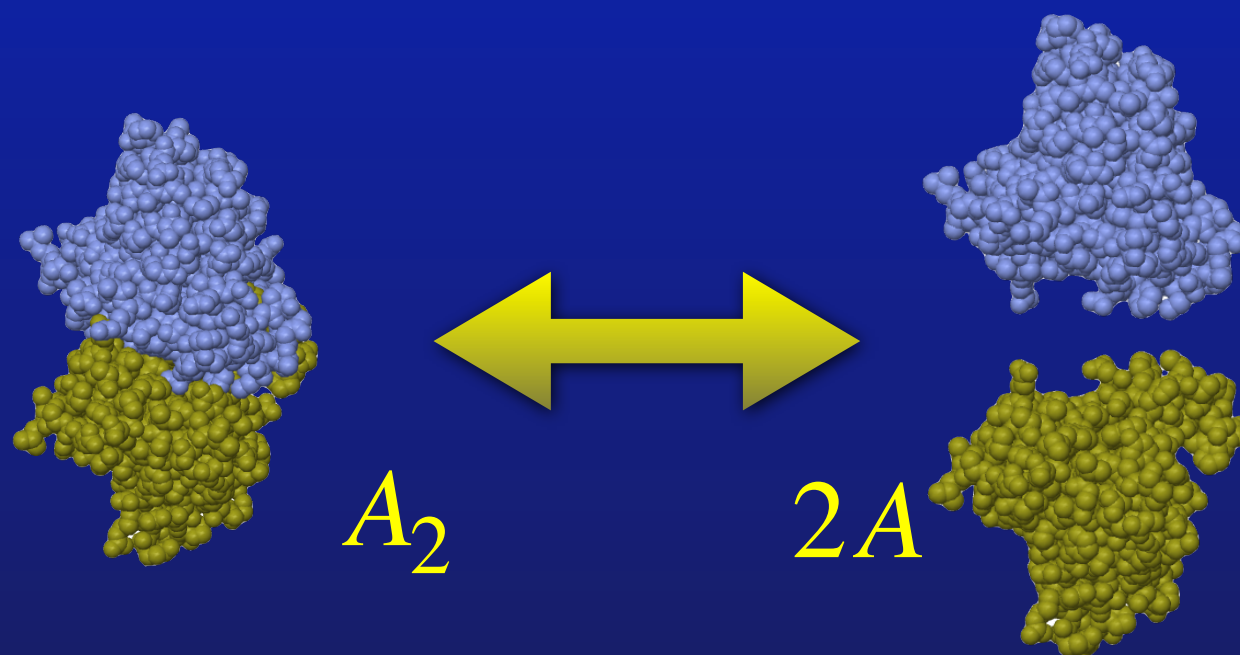
Research Complex at Harwell

# Is $\Delta G_0$ Sufficient An Indicator?

★ Consider PDB entry 3LT5:



$$\Delta G_0 = 3 \text{ kcal / M}$$



$$\Delta G_0 = 10 \text{ kcal / M}$$

The tetramer is weaker than the dimer, so one may think that the structure is dimeric  
But the tetramer is equilibrated with the dimer, so that their concentrations can be comparable

What is the correct answer?





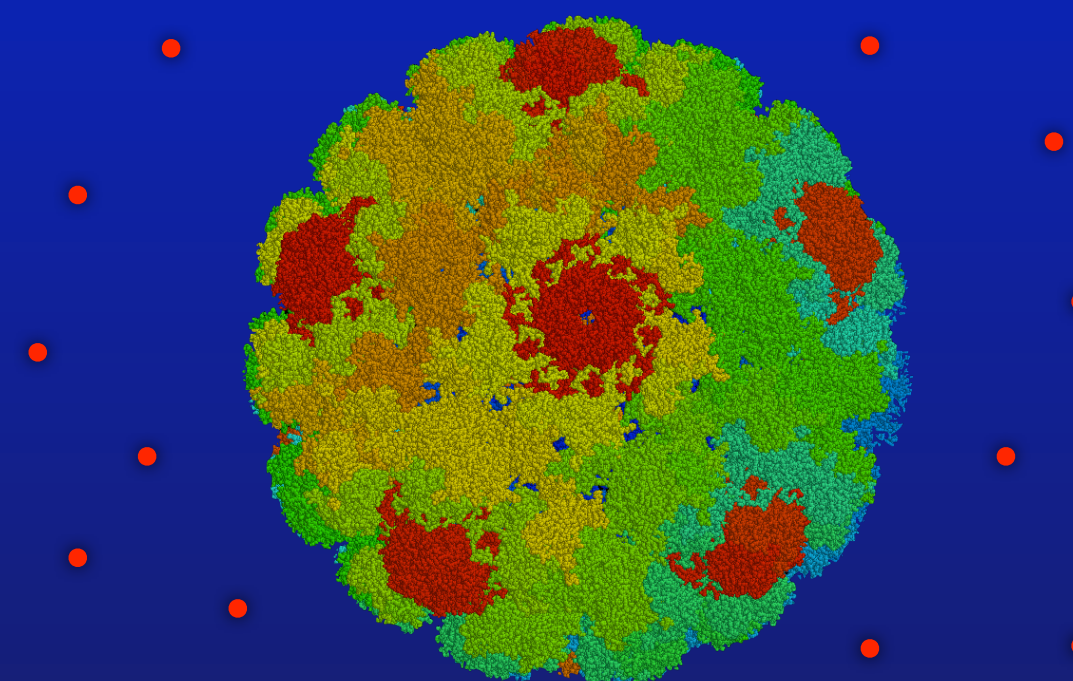
# The Stock

- ★ All possible complexes co-exist in dynamic equilibrium and form a “stock”  
- *PISA's Stock is limited to complexes formed by crystal interfaces*
- ★ Their stock concentration do vary
- ★ Concentrations depend on free energy of dissociation *and stock composition*
- ★ Concentration-based analysis is not very indicative:

*for the equilibrium between large complex and its monomeric units on the right,*

$$[A_{360}] \ll [A]$$

*from which one could conclude that the complex is unstable;  
but obviously, the protein is highly aggregated*



- ★ Aggregated states are better indicated by the aggregation index:

$$A_i = \frac{m_i}{\sum_j m_j}$$

$m_i$  mass of  $i$ th species in the Stock

$$0 < A_i < 1$$

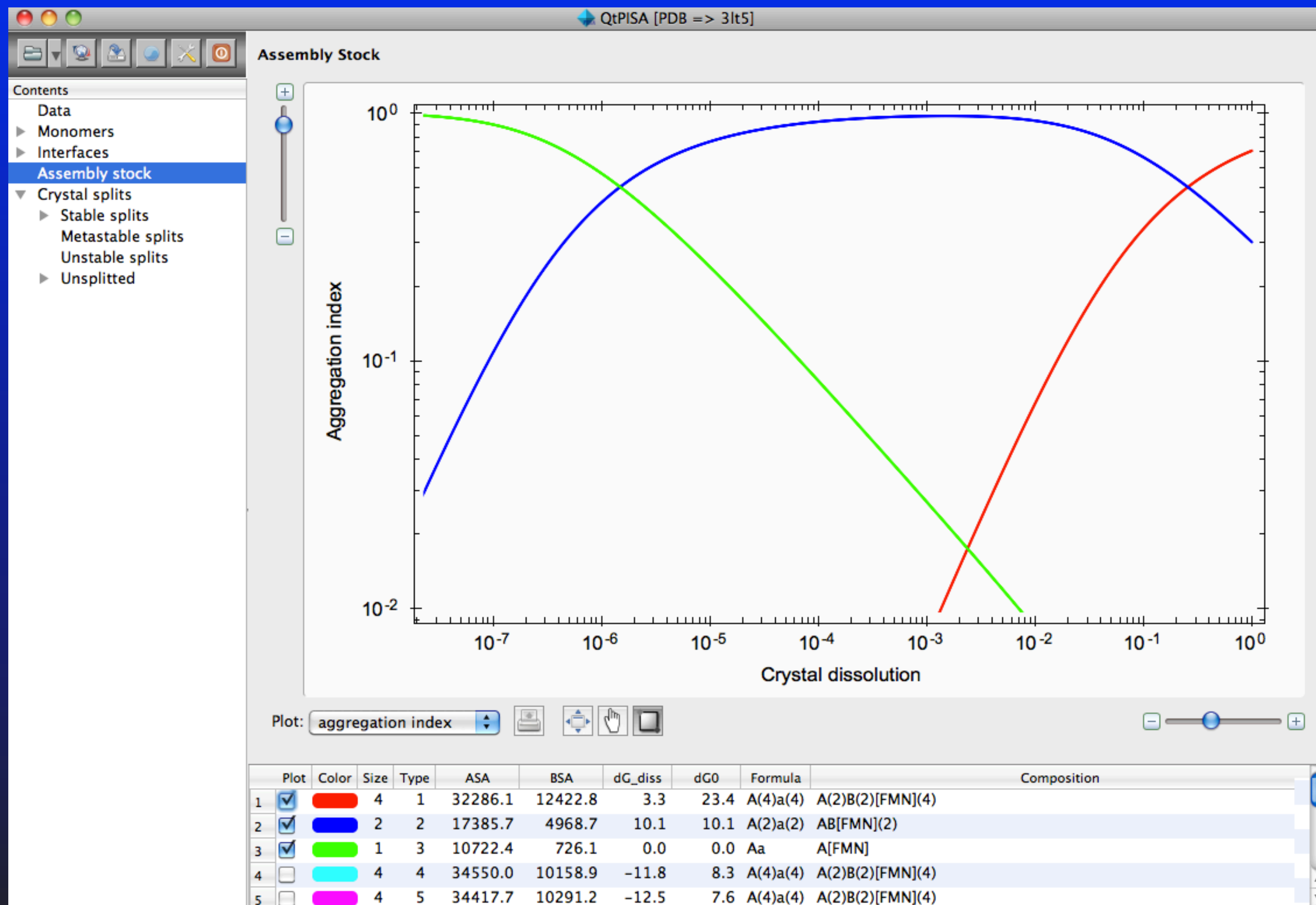
fully dissolved

fully aggregated



Research Complex at Harwell

# Assembly Stock for 3LT5 ( $A_4 \Leftrightarrow A_2 \Leftrightarrow A$ )





# Classification of Protein Assemblies

Assembly classification on the benchmark set of 218 protein structures published in

*Ponstingl, H., Kabir, T. and Thornton, J. (2003) Automatic inference of protein quaternary structures from crystals. J. Appl. Cryst. 36, 1116-1122.*

	1mer	2mer	3mer	4mer	6mer	Other	Sum	Correct
1mer	49	3	0	1	1	1	55	89%
2mer	3	71+11	0	2+1	0	0	76+12	93%
3mer	1	0	22	0	1	0	24	92%
4mer	2	2+1	0	26+6	0	1	31+7	84%
6mer	0	0	0	1	10+2	0	10+3	92%
196+22 $\Leftrightarrow$ 196 homomers and 22 heteromers							Total: 196+22	90%

Classification error in  $\Delta G_0$  :  $\pm 5$  kcal/mol



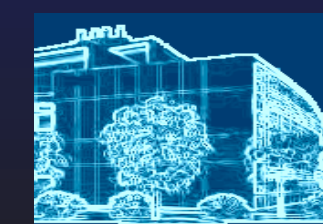
# Classification of Protein-DNA Complexes

Assembly classification on the benchmark set of 212 protein-DNA complexes published in

*Luscombe, N.M., Austin, S.E., Berman, H.M. and Thornton, J. (2000) An overview of the structures of protein-DNA complexes. Genome Biol. 1, 1-37.*

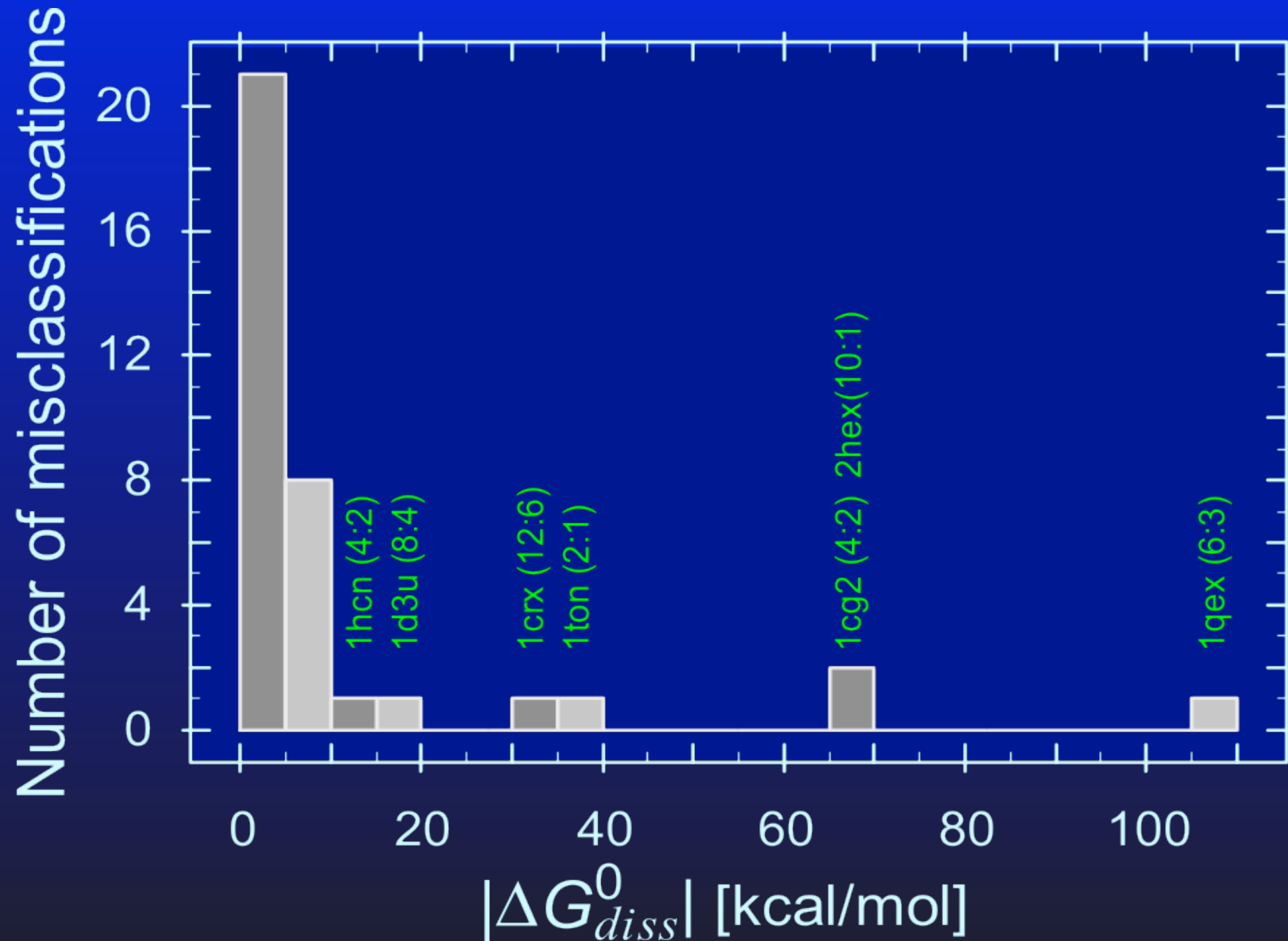
	2mer	3mer	4mer	5mer	6mer	10mer	Other	Sum	Correct
2mer	1	0	0	0	0	0	0	1	100%
3mer	6	96	0	0	1	0	2	105	91%
4mer	0	2	83	0	0	0	0	85	98%
5mer	0	0	2	3	0	0	0	5	60%
6mer	1	0	0	0	13	0	1	15	87%
10mer	0	0	0	0	0	1	0	1	100%
Total:								212	93%

Classification error in  $\Delta G_0$  :  $\pm 5$  kcal/mol





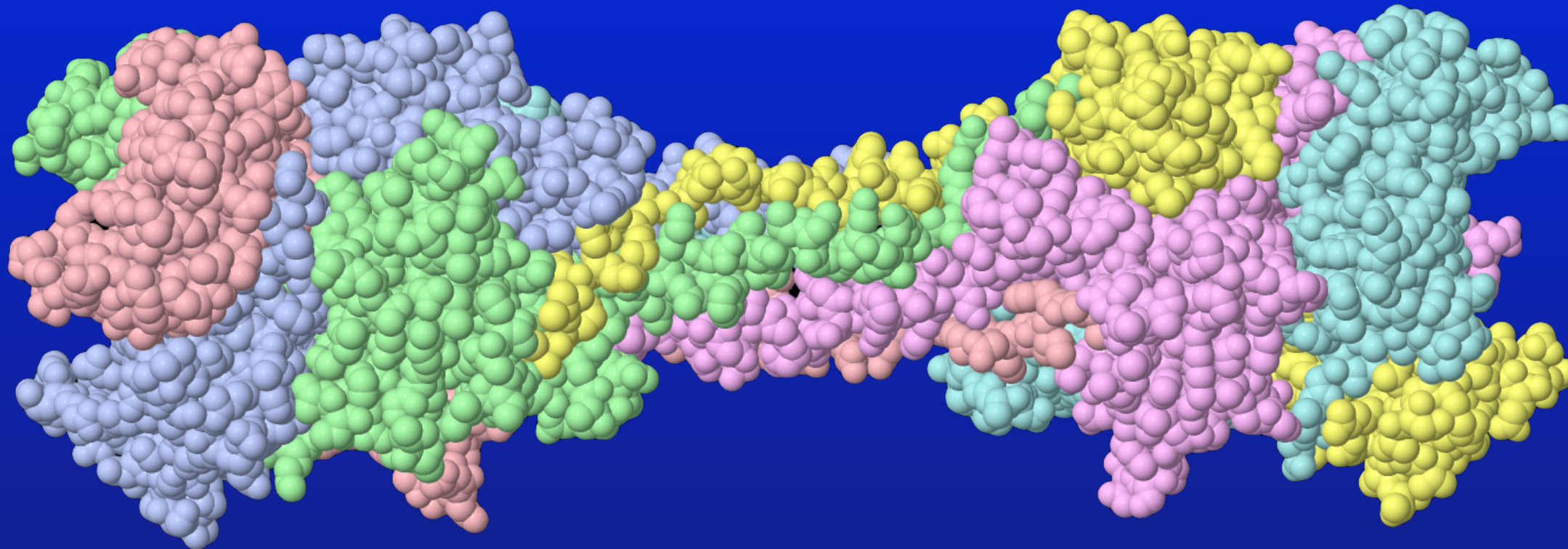
# Free Energy Distribution of Misclassifications



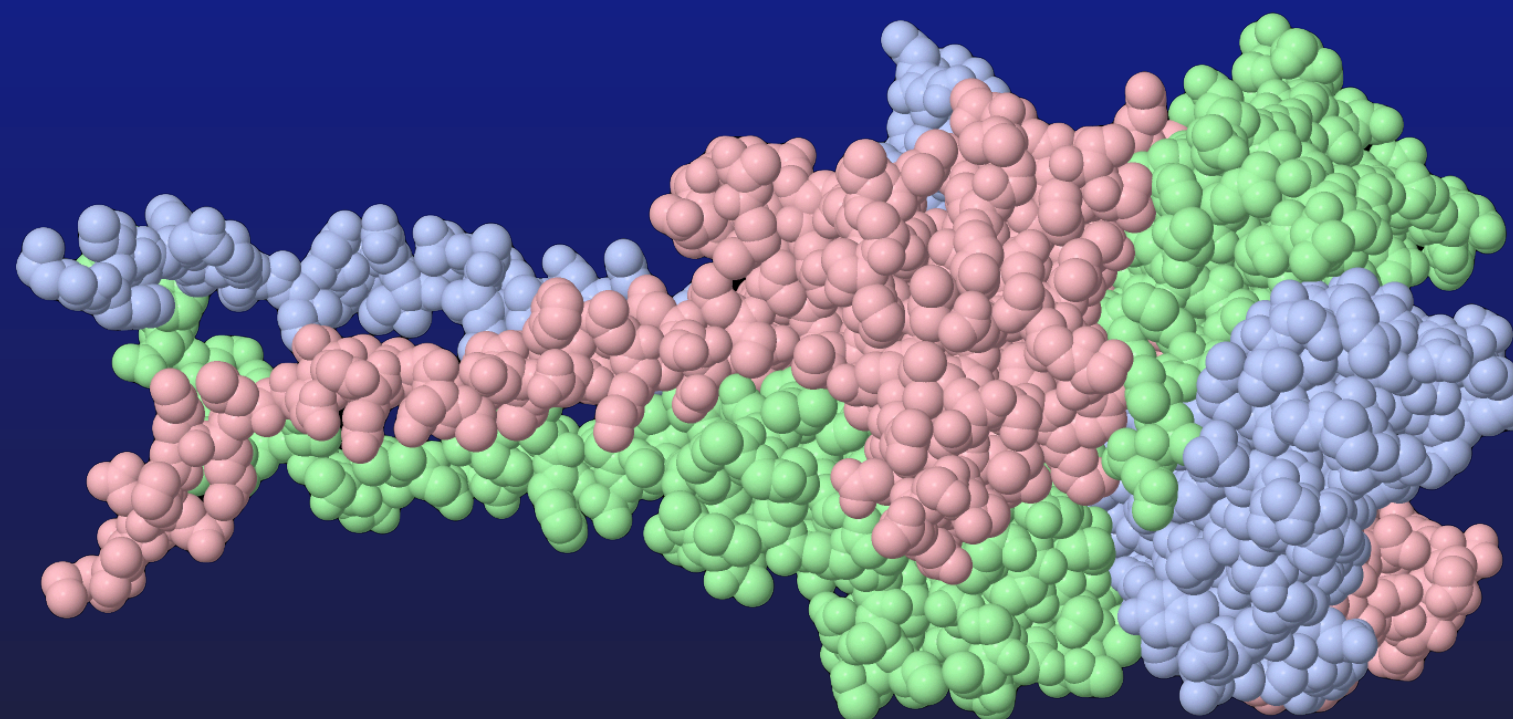


# Example of misclassification: 1QEX

BACTERIOPHAGE T4 GENE PRODUCT 9 (GP9), THE TRIGGER OF TAIL CONTRACTION AND THE LONG TAIL FIBERS CONNECTOR



**Predicted:** homohexamer  
Dissociates into 2 trimers  
 $\Delta G_0 \approx 106$  kcal/mol



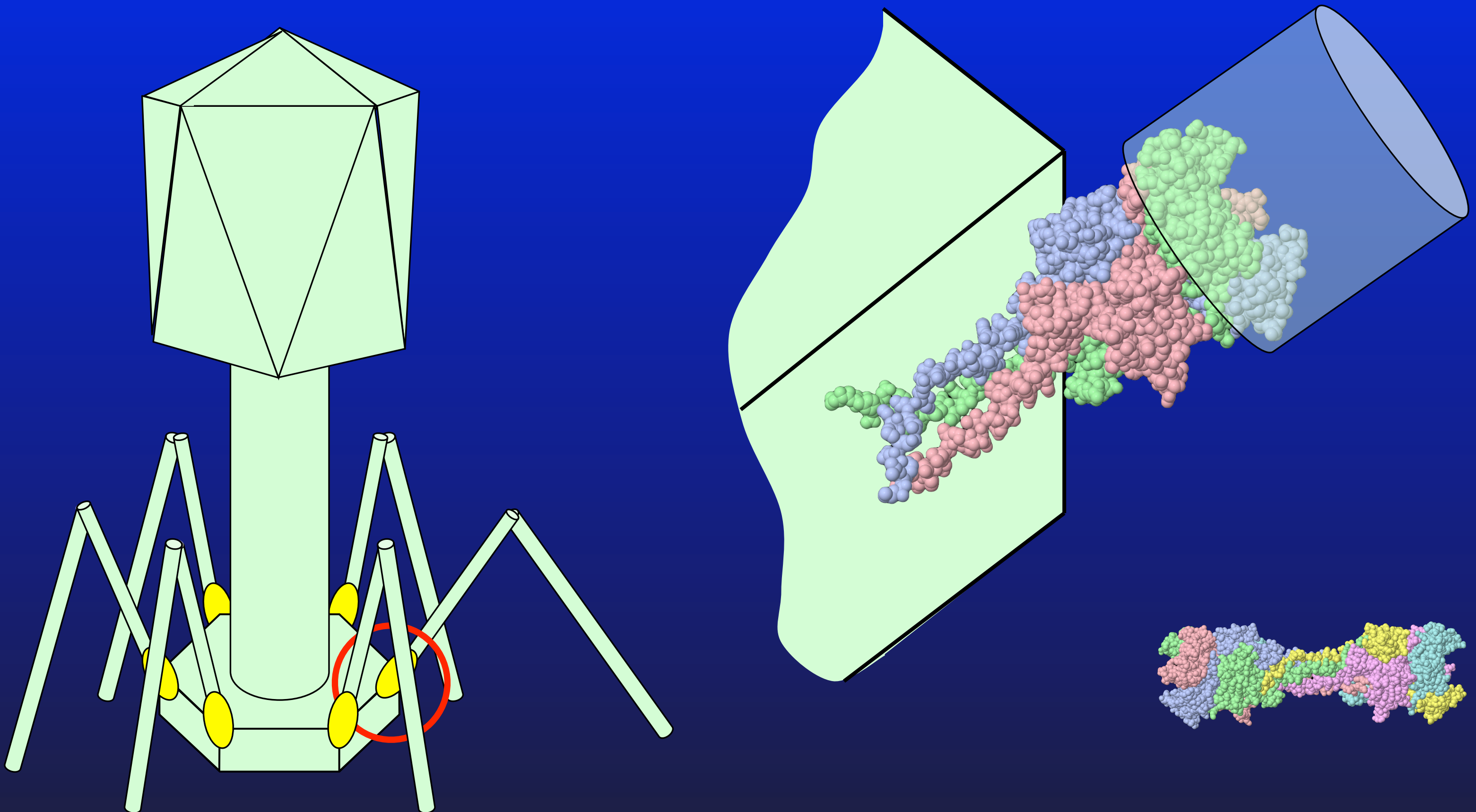
**Biological unit:** homotrimer  
Dissociates into 3 monomers  
 $\Delta G_0 \approx 90$  kcal/mol





# Example of misclassification: 1QEX

BACTERIOPHAGE T4 GENE PRODUCT 9 (GP9), THE TRIGGER OF TAIL CONTRACTION AND THE LONG TAIL FIBERS CONNECTOR



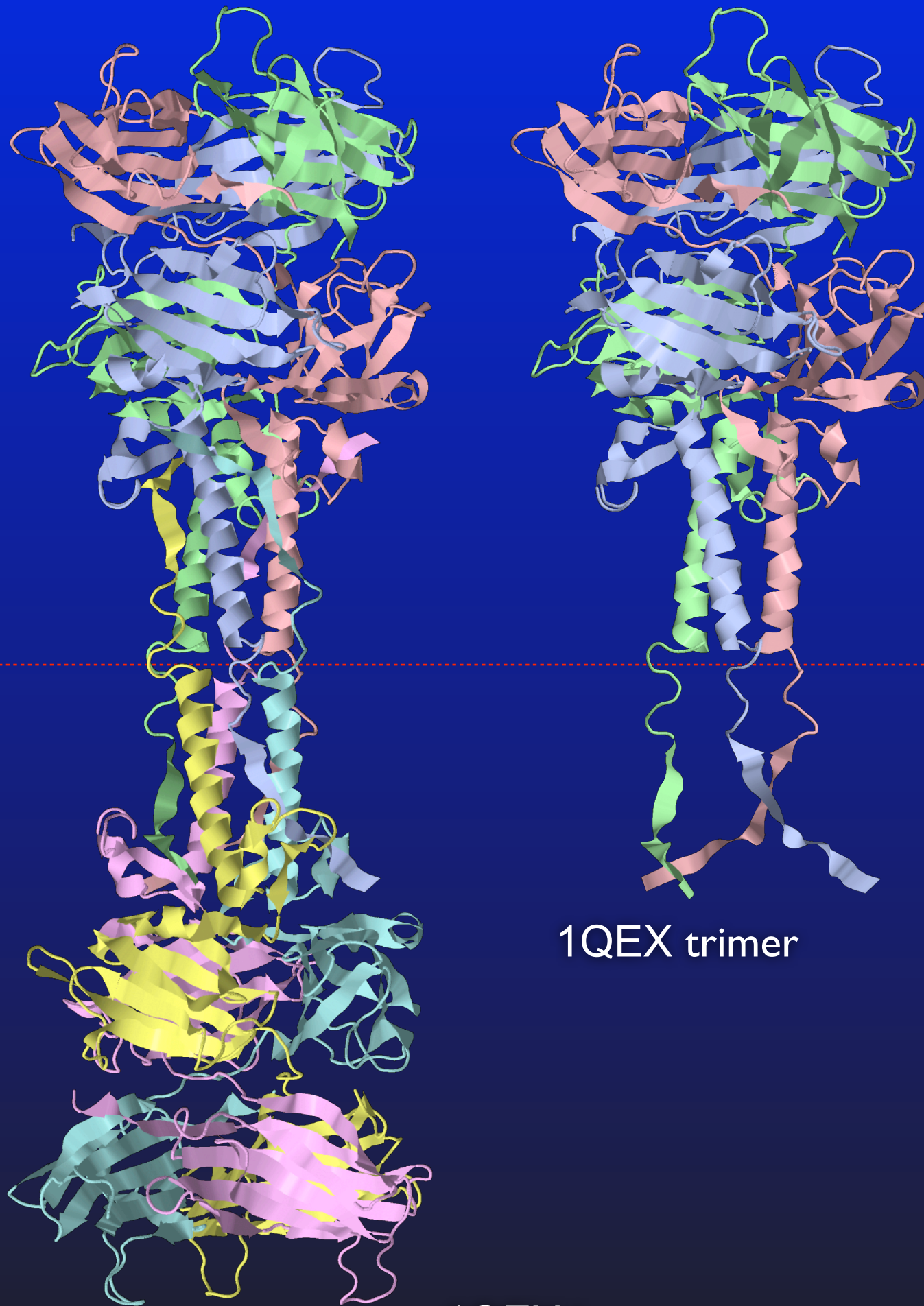
Rossmann M.G., Mesyanzhinov V.V., Arisaka F and Leiman P.G. (2004) *The bacteriophage T4 DNA injection machine*.  
Curr. Opin Struct. Biol. 14:171-180.



Research Complex at Harwell

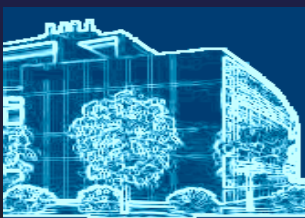
# Example of misclassification: 1QEX

BACTERIOPHAGE T4 GENE PRODUCT 9 (GP9), THE TRIGGER OF TAIL CONTRACTION AND THE LONG TAIL FIBERS CONNECTOR



1QEX trimer

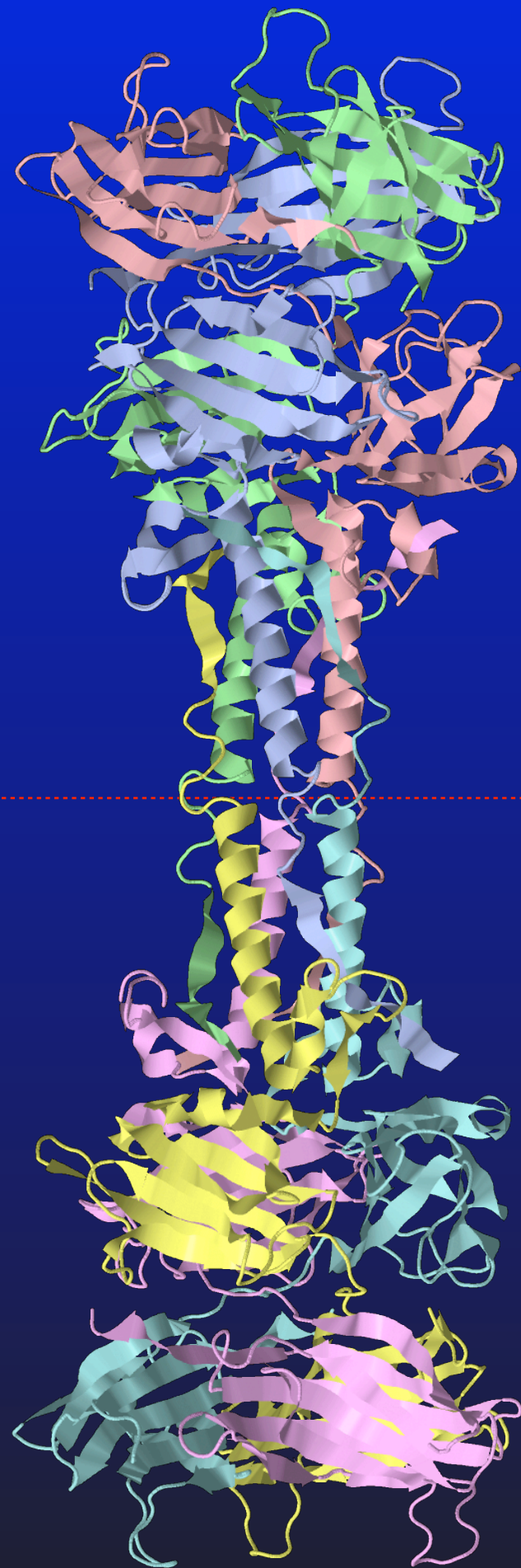
1QEX hexamer



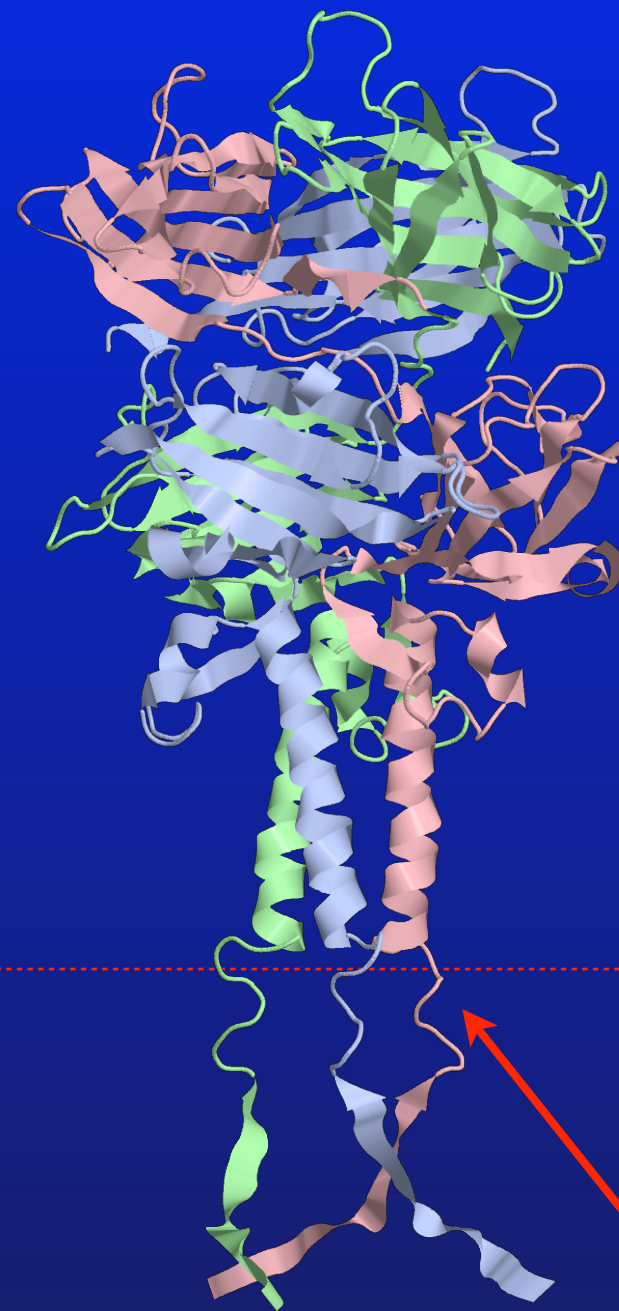


# Example of misclassification: 1QEX

BACTERIOPHAGE T4 GENE PRODUCT 9 (GP9), THE TRIGGER OF TAIL CONTRACTION AND THE LONG TAIL FIBERS CONNECTOR



1QEX hexamer



1QEX trimer

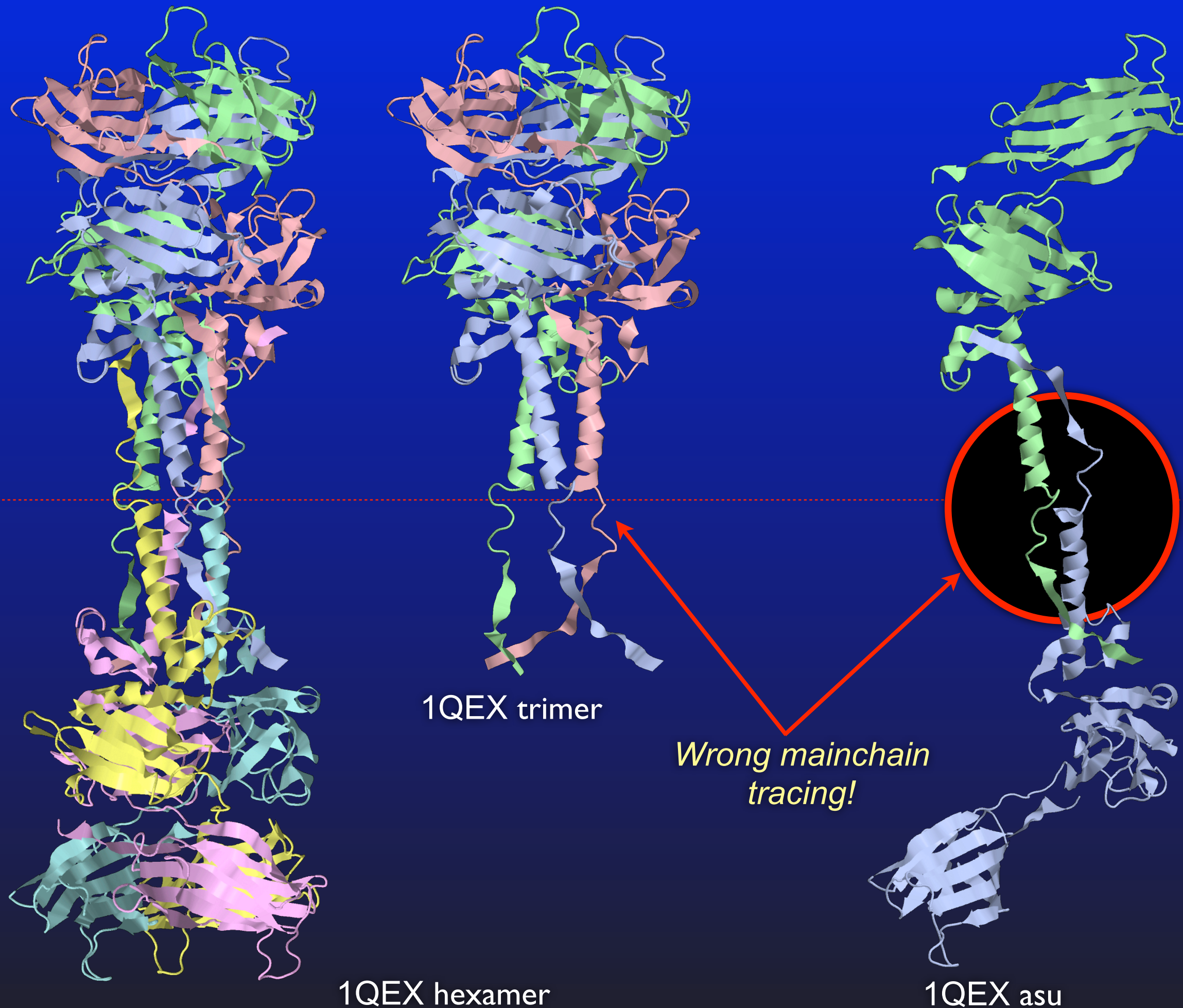
*Wrong mainchain  
tracing!*





# Example of misclassification: 1QEX

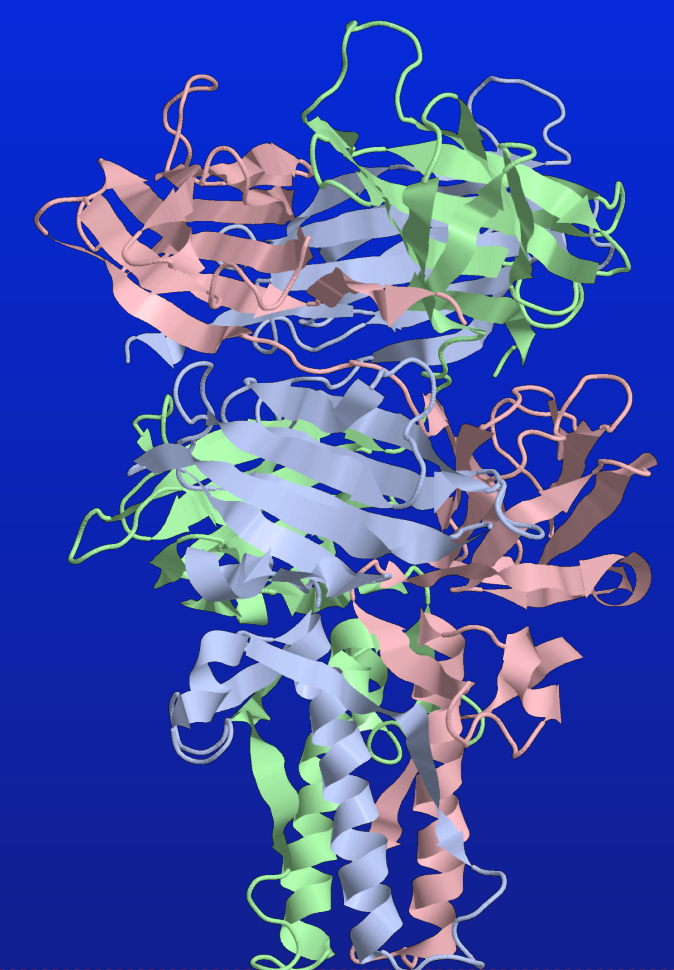
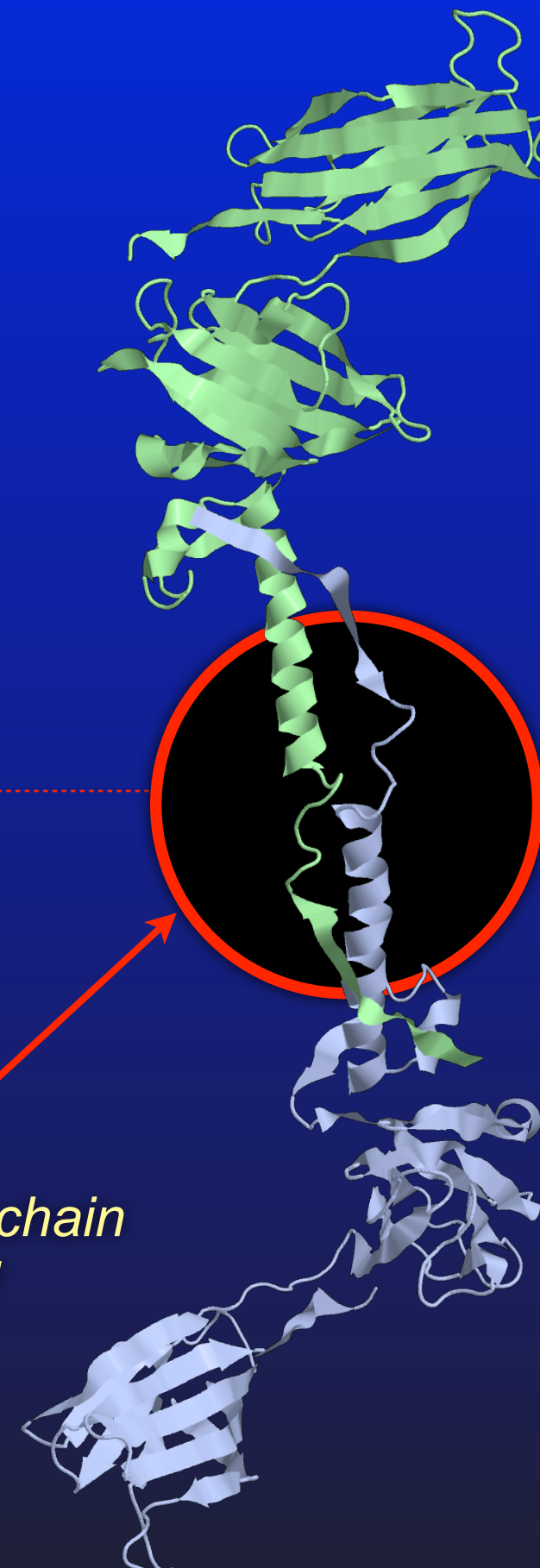
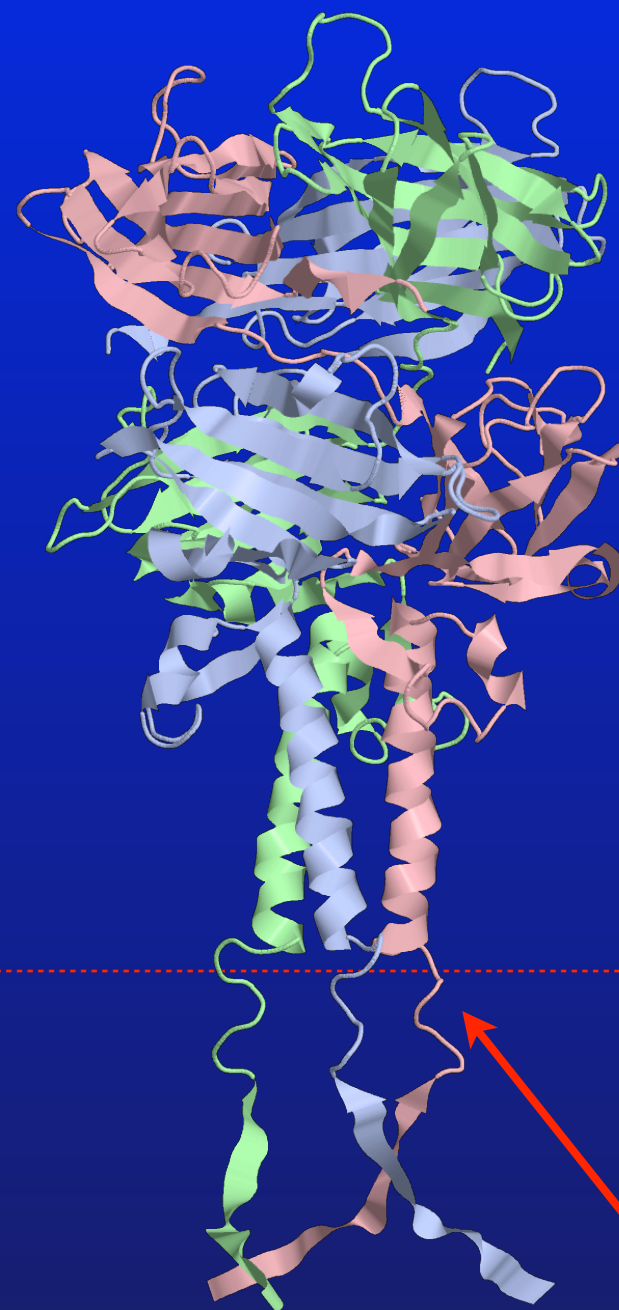
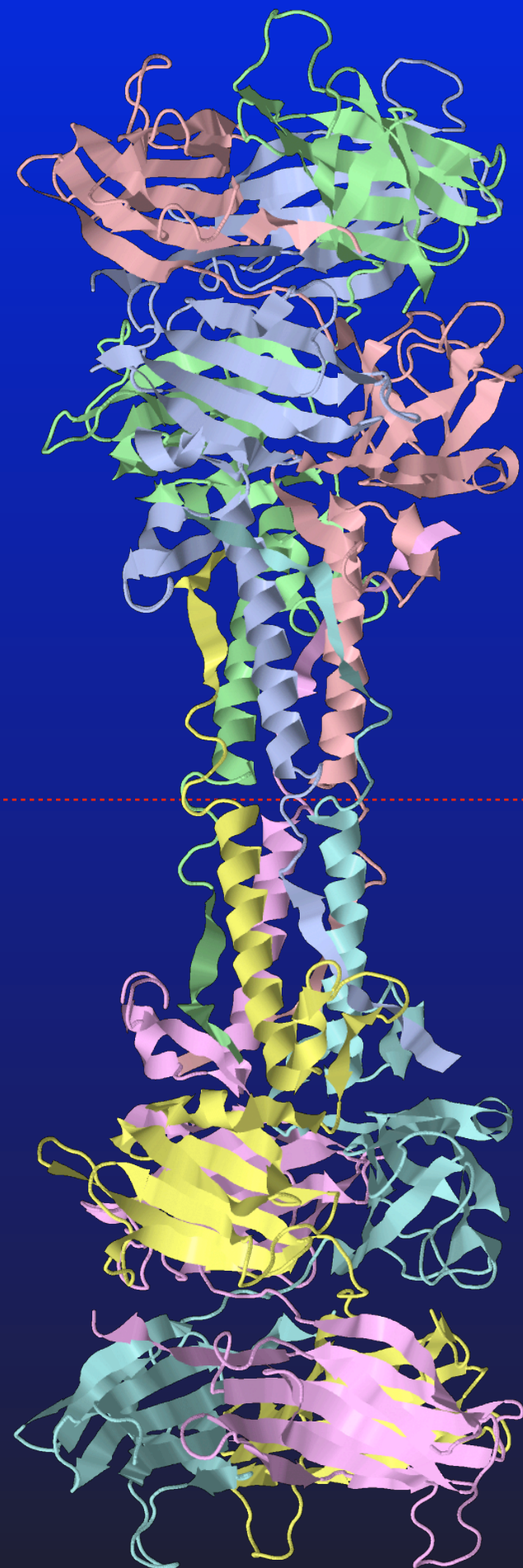
BACTERIOPHAGE T4 GENE PRODUCT 9 (GP9), THE TRIGGER OF TAIL CONTRACTION AND THE LONG TAIL FIBERS CONNECTOR





# Example of misclassification: 1QEX

BACTERIOPHAGE T4 GENE PRODUCT 9 (GP9), THE TRIGGER OF TAIL CONTRACTION AND THE LONG TAIL FIBERS CONNECTOR



Correct mainchain tracing  
*Identified correctly*

*Wrong mainchain tracing!*



Research Complex at Harwell



# Example of misclassification: 1D3U

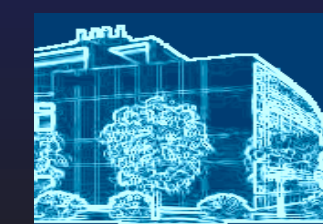
TATA-BINDING PROTEIN / TRANSCRIPTION FACTOR

Predicted: octamer

Dissociates into 2 tetramers

$$\Delta G_0 \approx 20 \text{ kcal/mol}$$

Functional unit:  
tetramer

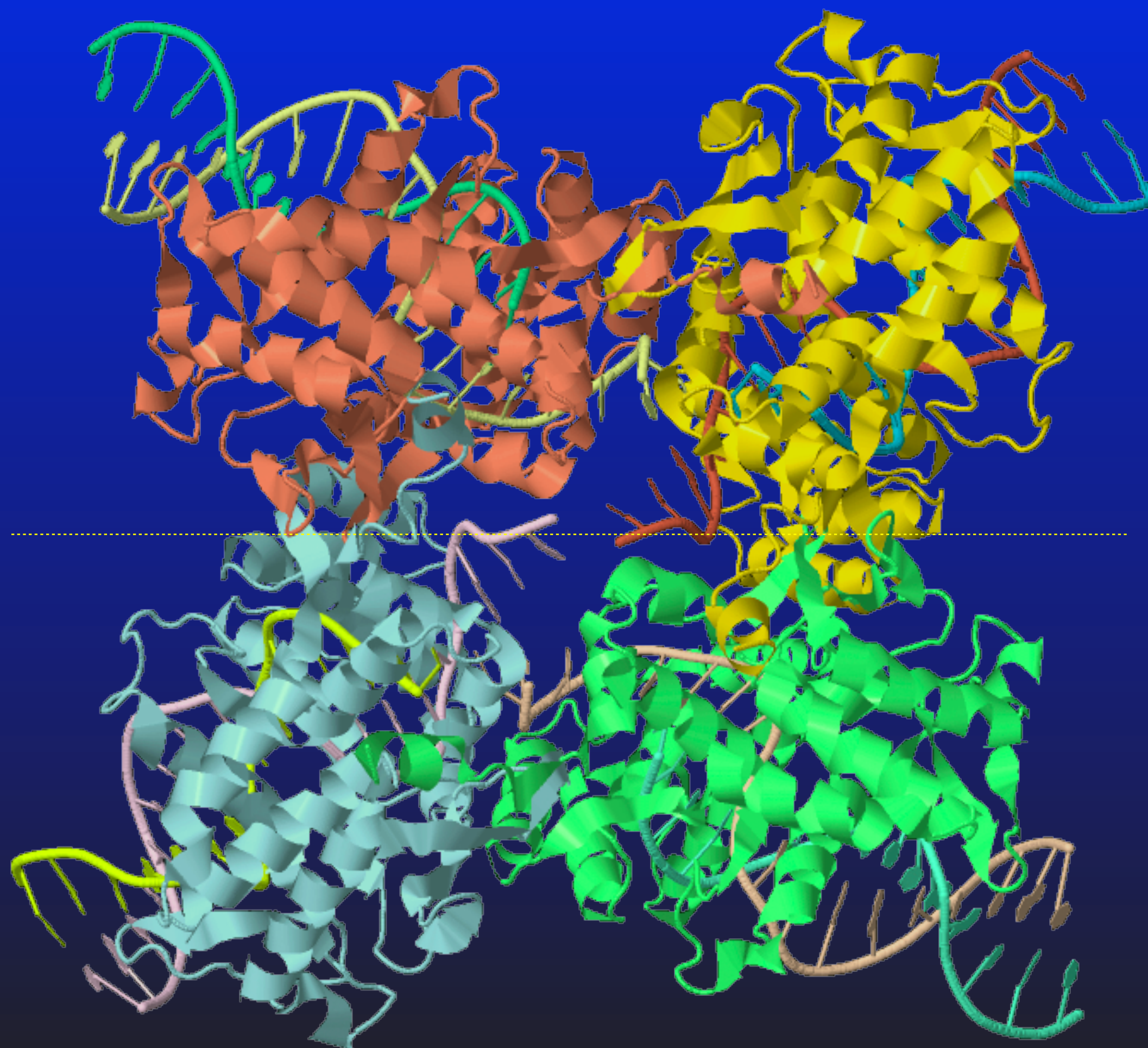


Research Complex at Harwell



# Example of misclassification: 1CRX

CRE RECOMBINASE / DNA COMPLEX REACTION INTERMEDIATE



Predicted: dodecamer

Dissociates into 2 hexamers

$$\Delta G_0 \approx 28 \text{ kcal/mol}$$

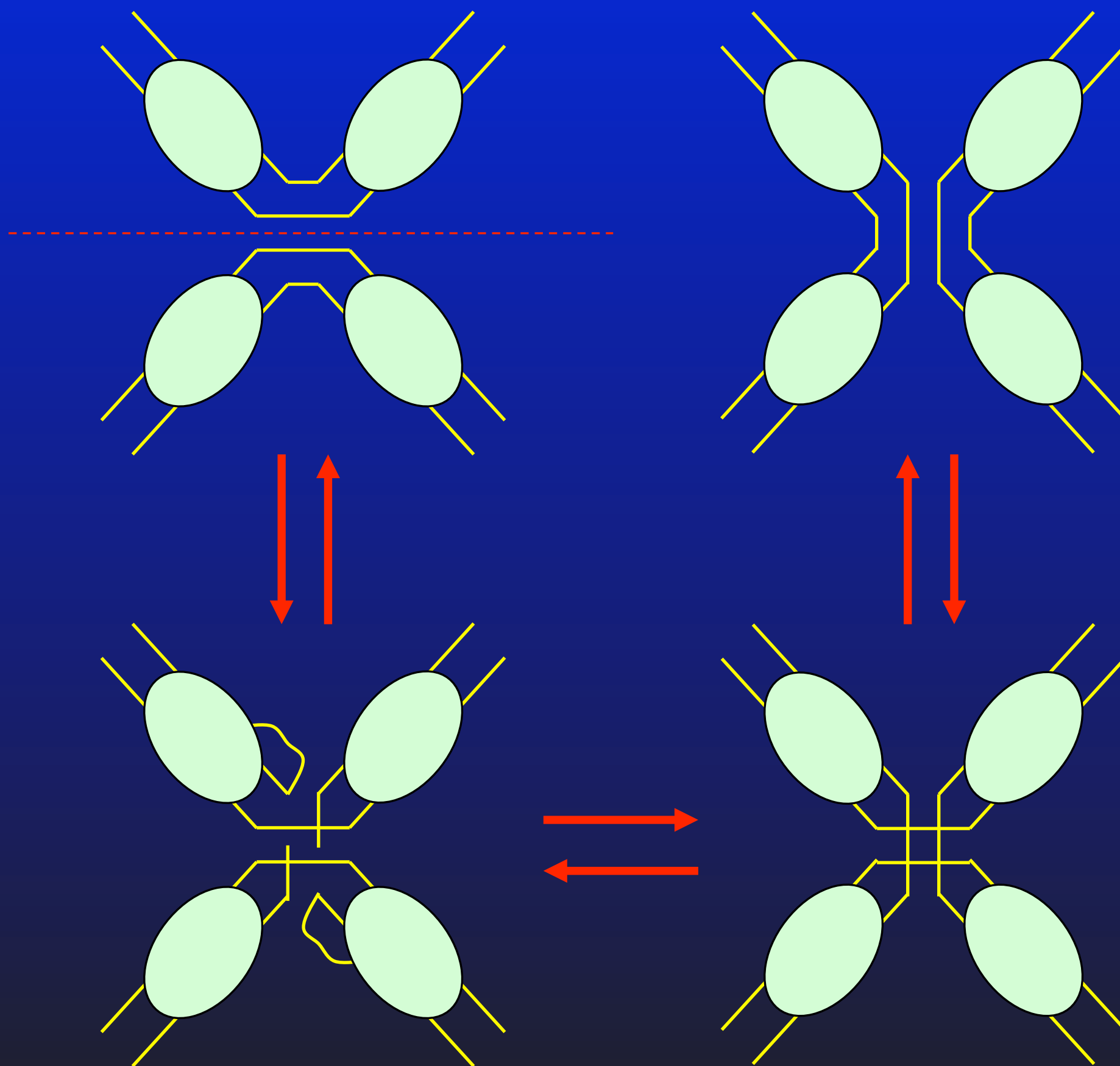
Functional unit: trimer



Research Complex at Harwell

# Example of misclassification: 1CRX

CRE RECOMBINASE / DNA COMPLEX REACTION INTERMEDIATE



Guo F., Gopaul D.N. and van  
Duyne G.D. (1997)

*Structure of Cre recombinase  
complexed with DNA in a site-  
specific recombination  
synapse.*

Nature 389:40-46.

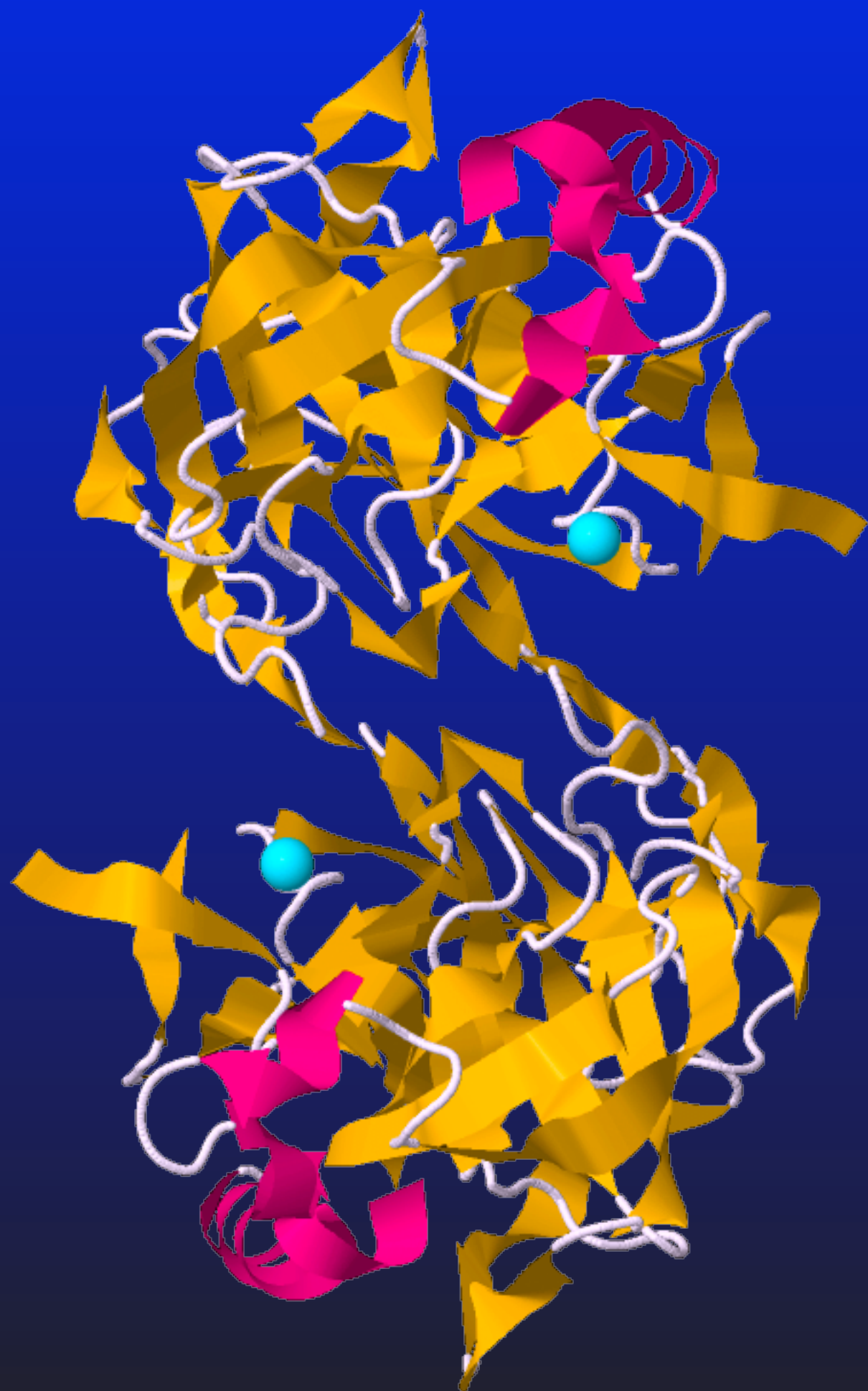


Research Complex at Harwell



# Example of misclassification: 1TON

TONIN



Predicted: dimer

Dissociates at

$$\Delta G_0 \simeq 37 \text{ kcal/mol}$$

Biological unit: monomer

Apparent dimerization is an artefact due to the presence of  $\text{Zn}^{+2}$  ions added to the buffer to aid crystallization. Removal of Zn from the file results in  $\Delta G_0 \simeq 3 \text{ kcal/mol}$

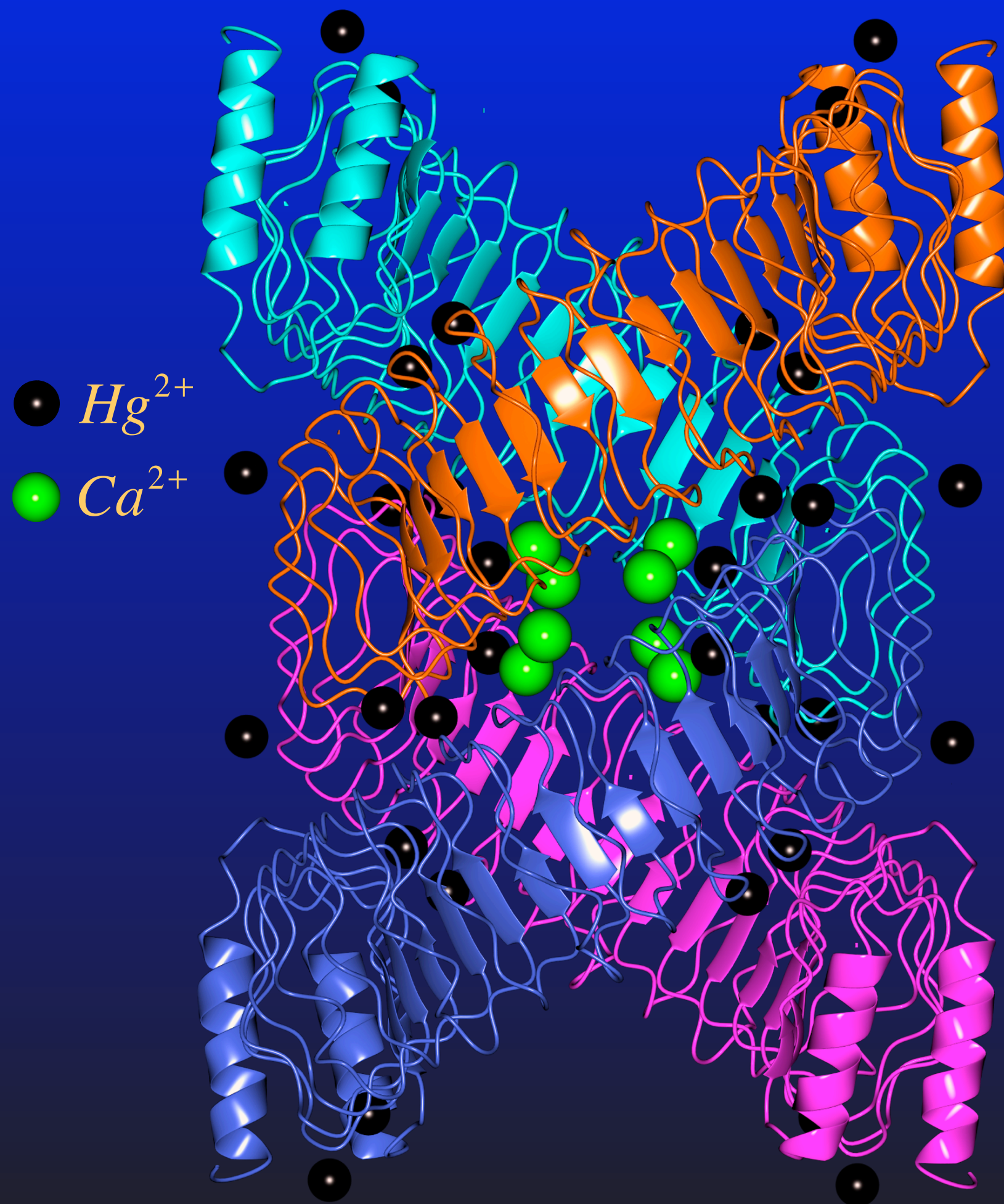
Fujinaga M., James M.N.G. (1997) *Rat submaxillary gland serine protease, tonin structure solution and refinement at 1.8 Å resolution.* J.Mol.Biol. 195:373-396.





# Example of ion effect: 1G9U vs 1JL5

Y. PESTIS CYTOXIN YopM



**Predicted:** homotetramer in form of a superhelix featuring a hollow cylinder with an inner diameter of  $\sim 35$  Å.

	<b>1G9U</b>	<b>1JL5</b>
Space Group	$P4_222$	$I4_122$
$\Delta G_0$ , kcal/mol	37	3
Number of ions	40	16

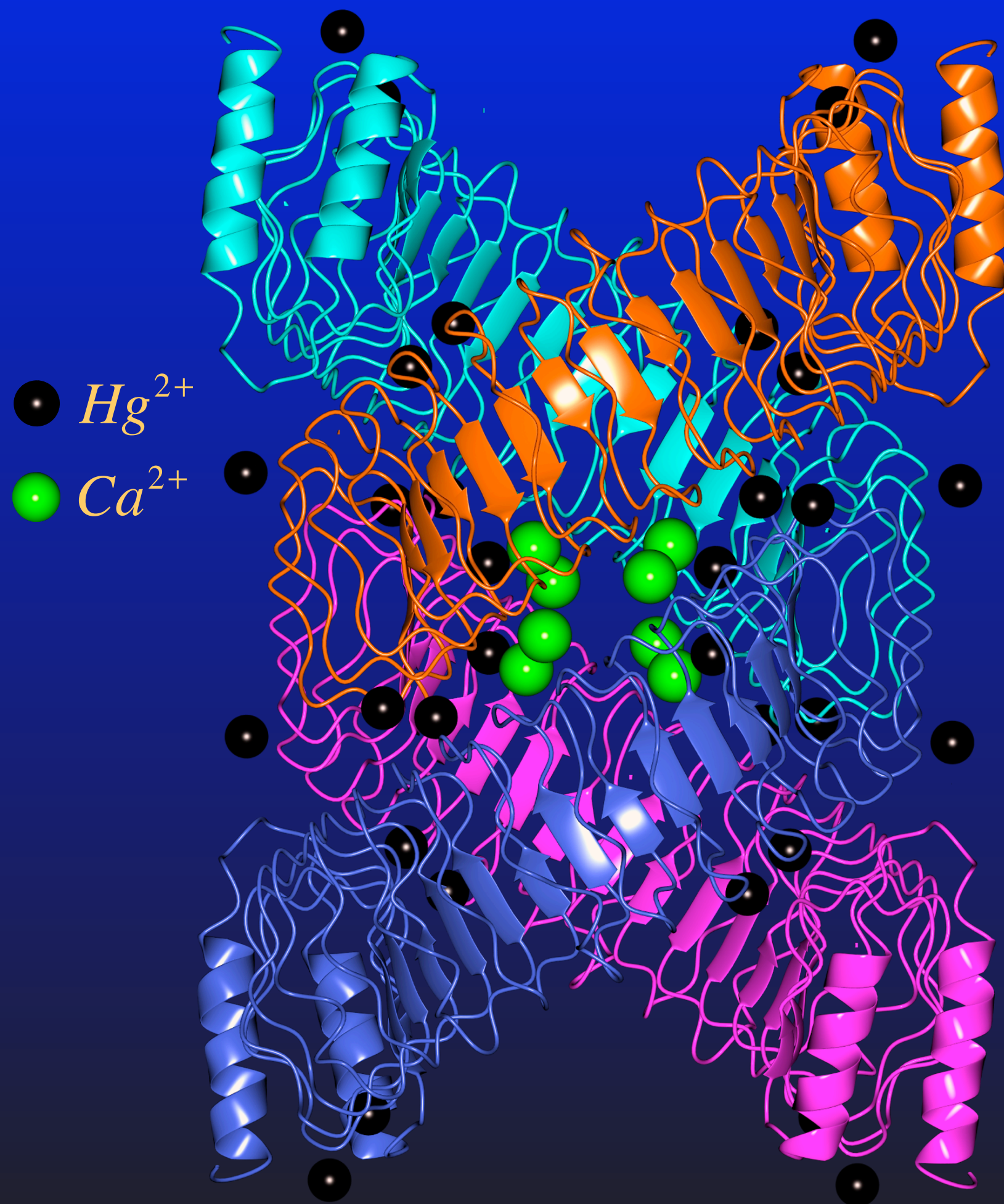


Research Complex at Harwell



# Example of ion effect: 1G9U vs 1JL5

Y. PESTIS CYTOXIN YopM



**Predicted:** homotetramer in form of a superhelix featuring a hollow cylinder with an inner diameter of  $\sim 35$  Å.

	<b>1G9U</b>	<b>1JL5</b>
Space Group	$P4_222$	$I4_122$
$\Delta G_0$ , kcal/mol	37	3
Number of ions	40	16

**Biological unit:** monomer

Evdokimov, A. G., Anderson, D. E., Routzahn, K. M. & Waugh, D. S. (2001). J. Mol. Biol. 312, 807–821

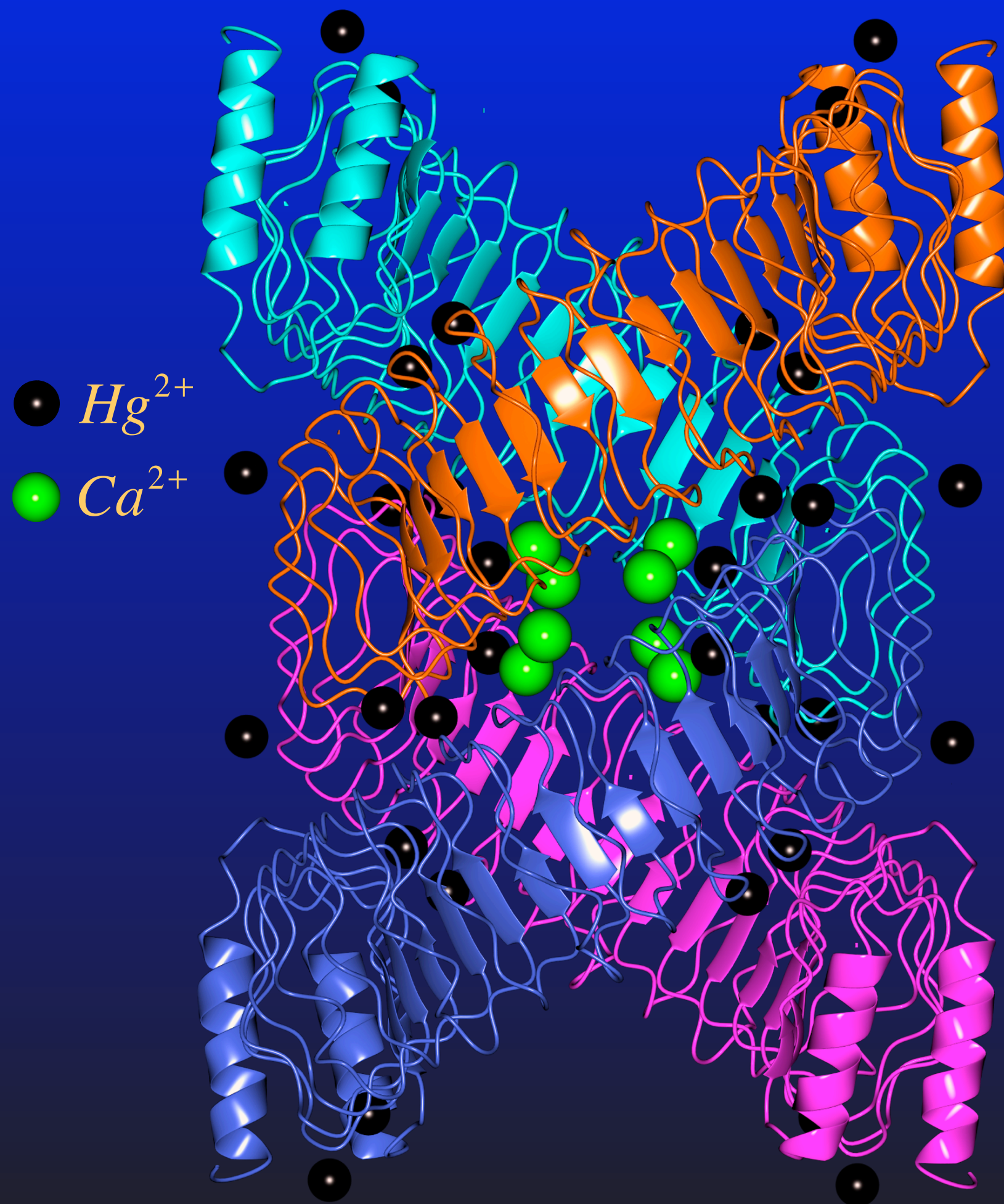


Research Complex at Harwell



# Example of ion effect: 1G9U vs 1JL5

Y. PESTIS CYTOXIN YopM



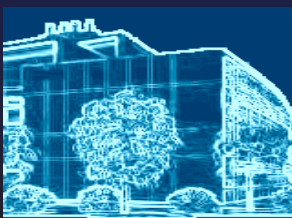
**Predicted:** homotetramer in form of a superhelix featuring a hollow cylinder with an inner diameter of  $\sim 35$  Å.

	<b>1G9U</b>	<b>1JL5</b>
Space Group	$P4_222$	$I4_122$
$\Delta G_0$ , kcal/mol	37	3
Number of ions	40	16

**Biological unit:** monomer

Evdokimov, A. G., Anderson, D. E., Routzahn, K. M. & Vaughn, D. S. (2001). J. Mol. Biol. 312, 807–821

*Removal of ions makes the structure monomeric in PISA estimates*





# Does it really work?

- ★ PISA appears to work quite well, which seems to be a “problem”
  - ➡ 90% success rate achieved on the benchmark set
  - ➡ in 2007, wwPDB adopted PISA as a mandatory processing tool for all depositions
  - ➡ since that, feedback from wwPDB curators suggests that up to 95% of classifications made by PISA agree with experimental data on oligomeric state, where available, and with intuitive and common-sense considerations where experimental evidence is not given



Research Complex at Harwell

# Does it really work?

- ★ PISA appears to work quite well, which seems to be a “problem”
  - ➔ 90% success rate achieved on the benchmark set
  - ➔ in 2007, wwPDB adopted PISA as a mandatory processing tool for all depositions
  - ➔ since that, feedback from wwPDB curators suggests that up to 95% of classifications made by PISA agree with experimental data on oligomeric state, where available, and with intuitive and common-sense considerations where experimental evidence is not given

## ★ Why it should work well? Two reasons:

Energy models and calculations are quite accurate



Research Complex at Harwell



# Does it really work?

- ★ PISA appears to work quite well, which seems to be a “problem”
  - ➔ 90% success rate achieved on the benchmark set
  - ➔ in 2007, wwPDB adopted PISA as a mandatory processing tool for all depositions
  - ➔ since that, feedback from wwPDB curators suggests that up to 95% of classifications made by PISA agree with experimental data on oligomeric state, where available, and with intuitive and common-sense considerations where experimental evidence is not given

★ Why it should work well? Two reasons:

Energy models and calculations are quite accurate

Obviously wrong



Research Complex at Harwell

# Does it really work?

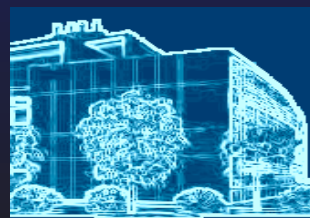
- ★ PISA appears to work quite well, which seems to be a “problem”
  - ➔ 90% success rate achieved on the benchmark set
  - ➔ in 2007, wwPDB adopted PISA as a mandatory processing tool for all depositions
  - ➔ since that, feedback from wwPDB curators suggests that up to 95% of classifications made by PISA agree with experimental data on oligomeric state, where available, and with intuitive and common-sense considerations where experimental evidence is not given

- ★ Why it should work well? Two reasons:

Energy models and calculations are quite accurate

Obviously wrong

PISA relies on geometry of interactions given by crystal packing. PISA does not dock monomeric units; rather, it uses crystal contacts as “nature’s dockings” assuming that they are correct.



Research Complex at Harwell



# Does it really work?

- ★ PISA appears to work quite well, which seems to be a “problem”
  - ➔ 90% success rate achieved on the benchmark set
  - ➔ in 2007, wwPDB adopted PISA as a mandatory processing tool for all depositions
  - ➔ since that, feedback from wwPDB curators suggests that up to 95% of classifications made by PISA agree with experimental data on oligomeric state, where available, and with intuitive and common-sense considerations where experimental evidence is not given

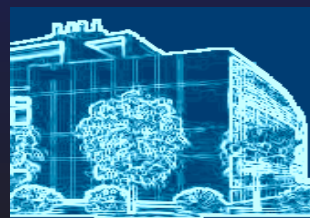
- ★ Why it should work well? Two reasons:

Energy models and calculations are quite accurate

PISA relies on geometry of interactions given by crystal packing. PISA does not dock monomeric units; rather, it uses crystal contacts as “nature’s dockings” assuming that they are correct.

Obviously wrong

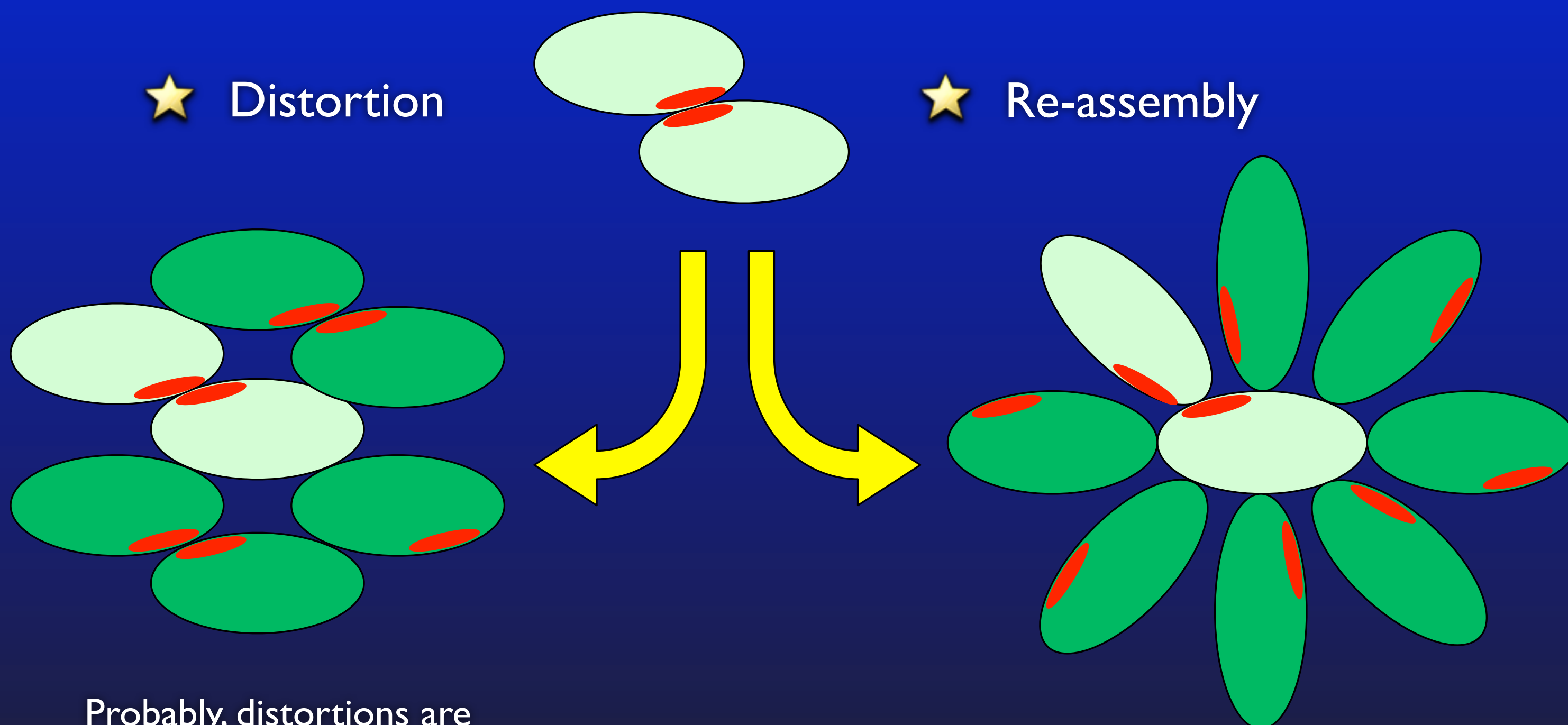
Probably correct



Research Complex at Harwell

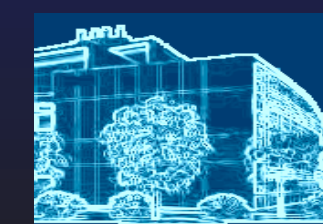
# Distortions and Re-assembly

- ★ Crystal optimizes energy globally, therefore it may sacrifice biologically relevant interaction in favour of unspecific crystal contacts



Probably, distortions are always there

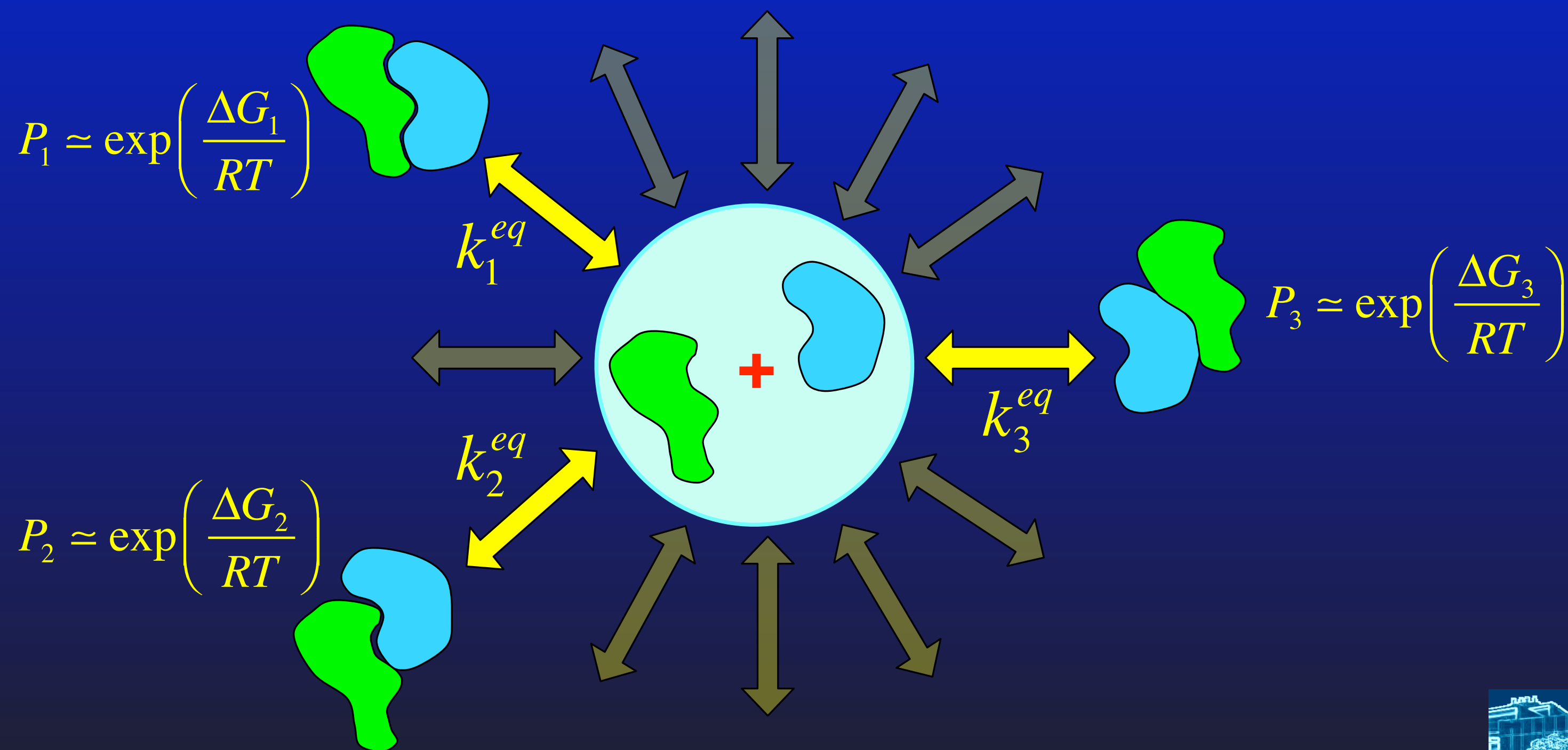
There is a chance for re-assembly if interaction is weak





# Alternative assemblies

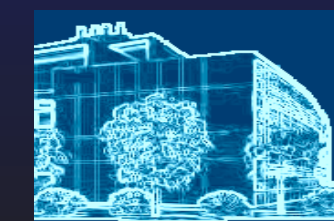
- ★ All complexes (assemblies) have right to exist in solvent, however with different occurrence probabilities. These probabilities may differ of those in crystal environment, e.g., in case of substantially assisted crystallisation.



# Real and superficial crystal contacts

- ★ If a crystal contact remains thermodynamically preferential in solution, the chances are that it represents a biochemically relevant interaction
- ★ Experimental data on structure of complexes in solution is sparse
- ★ One can hope to get some clues using computational docking, assuming that docking approximates in-solvent situation
- ★ Being applied to 4065 non-redundant dimers from the PDB, docking **fails** to arrive at crystal interface in **38%** of instances

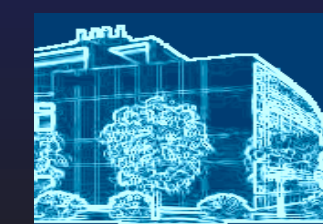
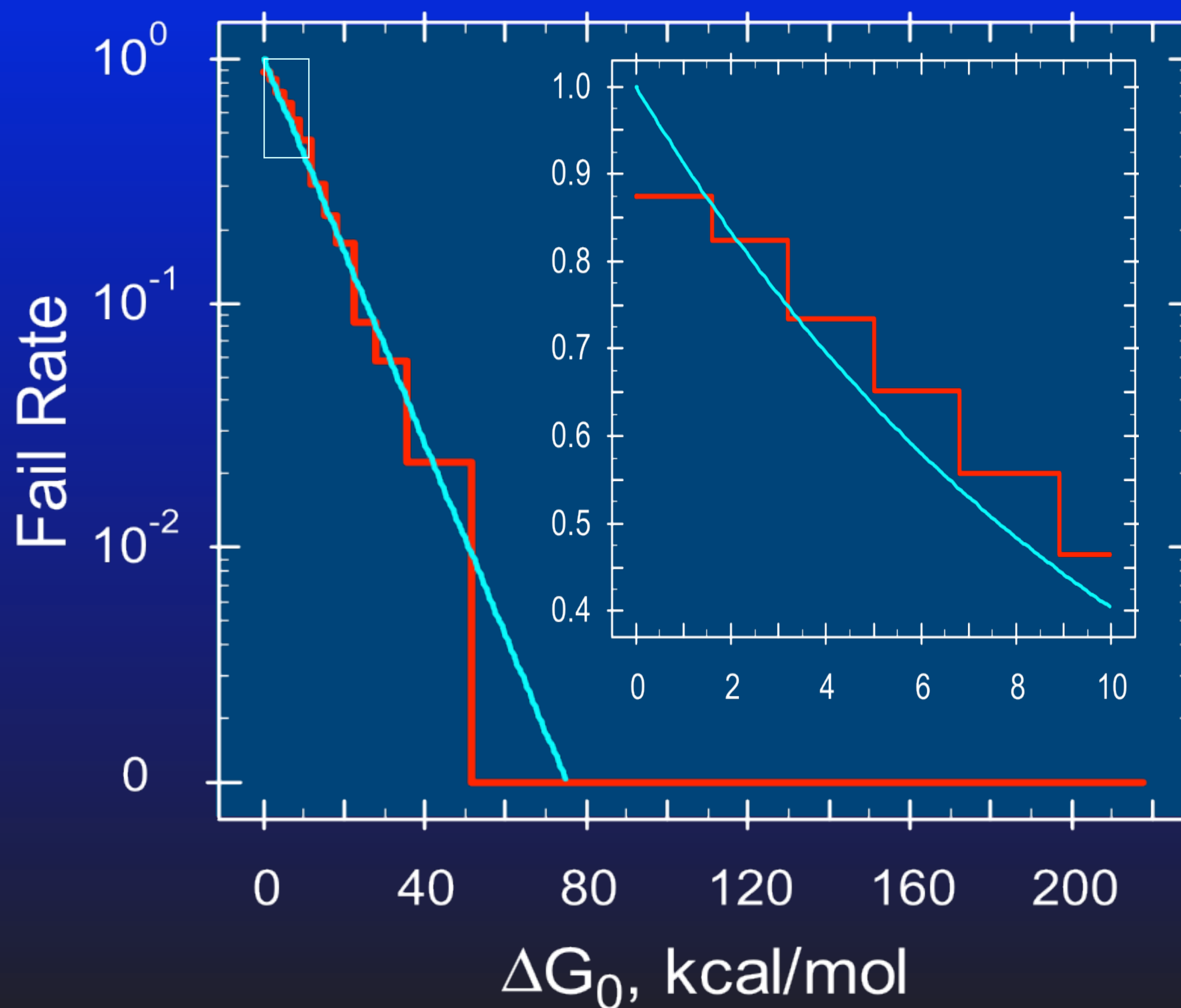
E. Krissinel (2010) J. Comp. Chem. 31, 133-143



Research Complex at Harwell

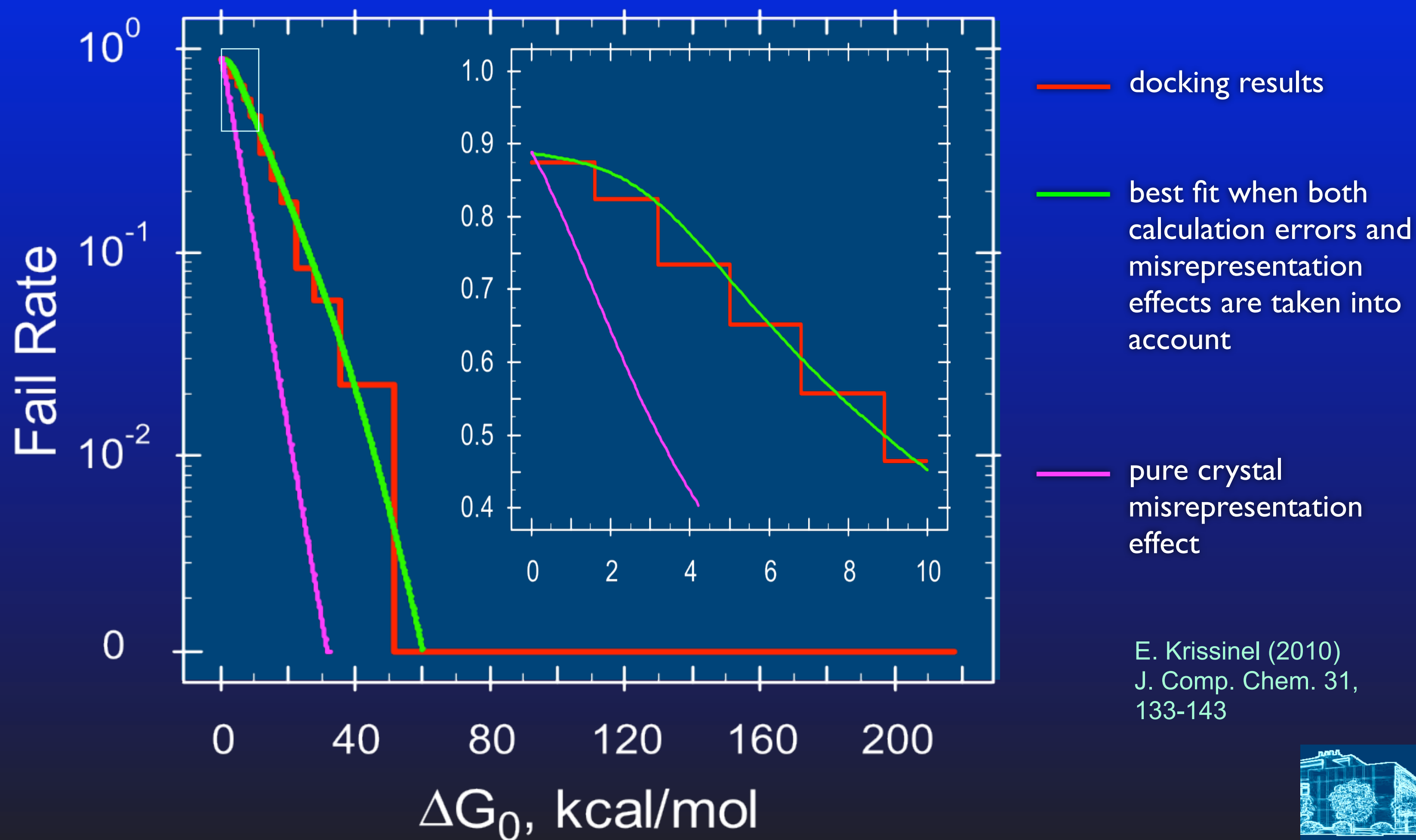


# Fail rate of docking



Research Complex at Harwell

# Calculation errors and crystal misrepresentation effects





# CAPRI Competition

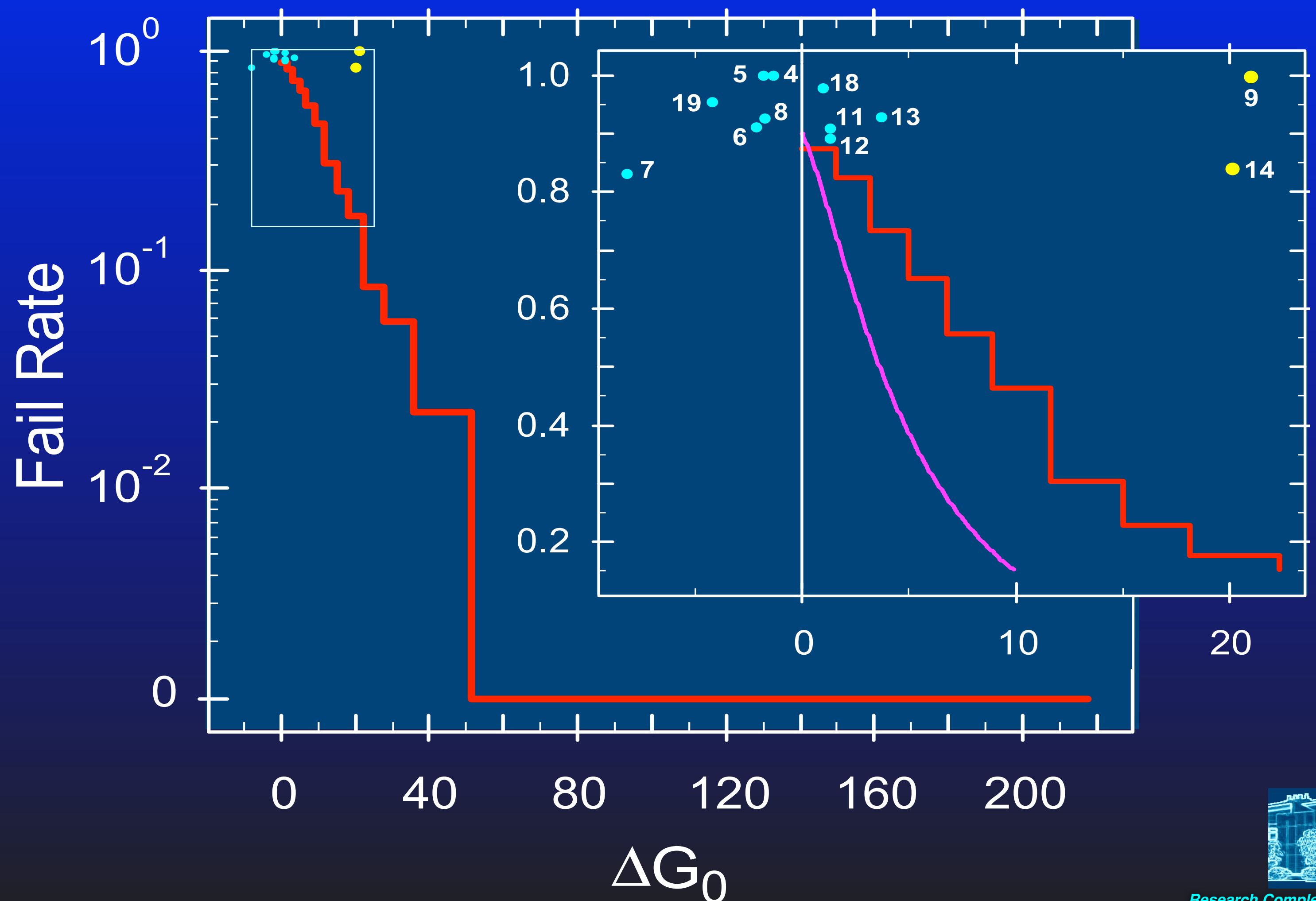
CAPRI success data from: S. Vajda (2005) *Classification of protein complexes based on docking difficulty*, Proteins, **60**:176-180



*Research Complex at Harwell*

# CAPRI Competition

CAPRI success data from: S. Vajda (2005) *Classification of protein complexes based on docking difficulty*, *Proteins*, **60**:176-180



Research Complex at Harwell



# Conclusions

- ★ Chemical-thermodynamical models allow one to reconstruct protein complexes from crystal data at ca. 90% success rate
- ★ However, this high success rate owes to structural information given by crystal packing, rather than to the accuracy of physical models and computations
- ★ Considerable part of errors come from the difference between experimental environment and in-wild conditions, as well as artificial interactions, induced by crystal packing
- ★ It is a question whether functional significance of an interface can be reliably inferred solely from interface properties. However, it should be more identifiable from the context of functional biological units, or complexes
- ★ Crystals are likely to misrepresent weak interactions and weakly-bound complexes
- ★ Structural bioinformatics may be most useful where interactions are weak. ???



# Acknowledgements

<b>Kim Henrick</b> <i>European Bioinformatics Institute</i>	General introduction and PQS expertise
<b>Mark Shenderovich</b> <i>Structural Bioinformatics Inc.</i>	Helpful discussion
<b>Hannes Ponstingl</b> <i>Sanger Centre</i>	Sharing expertise and benchmark data
<b>Sergei Strelkov</b> <i>University of Leuven</i>	“Mystery” of bacteriophage T4
<b>MSD &amp; PDB teams</b> <i>EBI &amp; Rutgers</i>	Everyday use of PISA, examples, verification and feedback
<b>CCP4</b> <i>Daresbury-York-Oxford</i>	Encouragement and publicity
<b>~10,000 PISA users</b> <i>Worldwide</i>	Using PISA and feedback
<b>Biotechnology and Biological Sciences Research Council</b> <i>(BBSRC) UK</i>	Research grant No. 721/B19544



Research Complex at Harwell