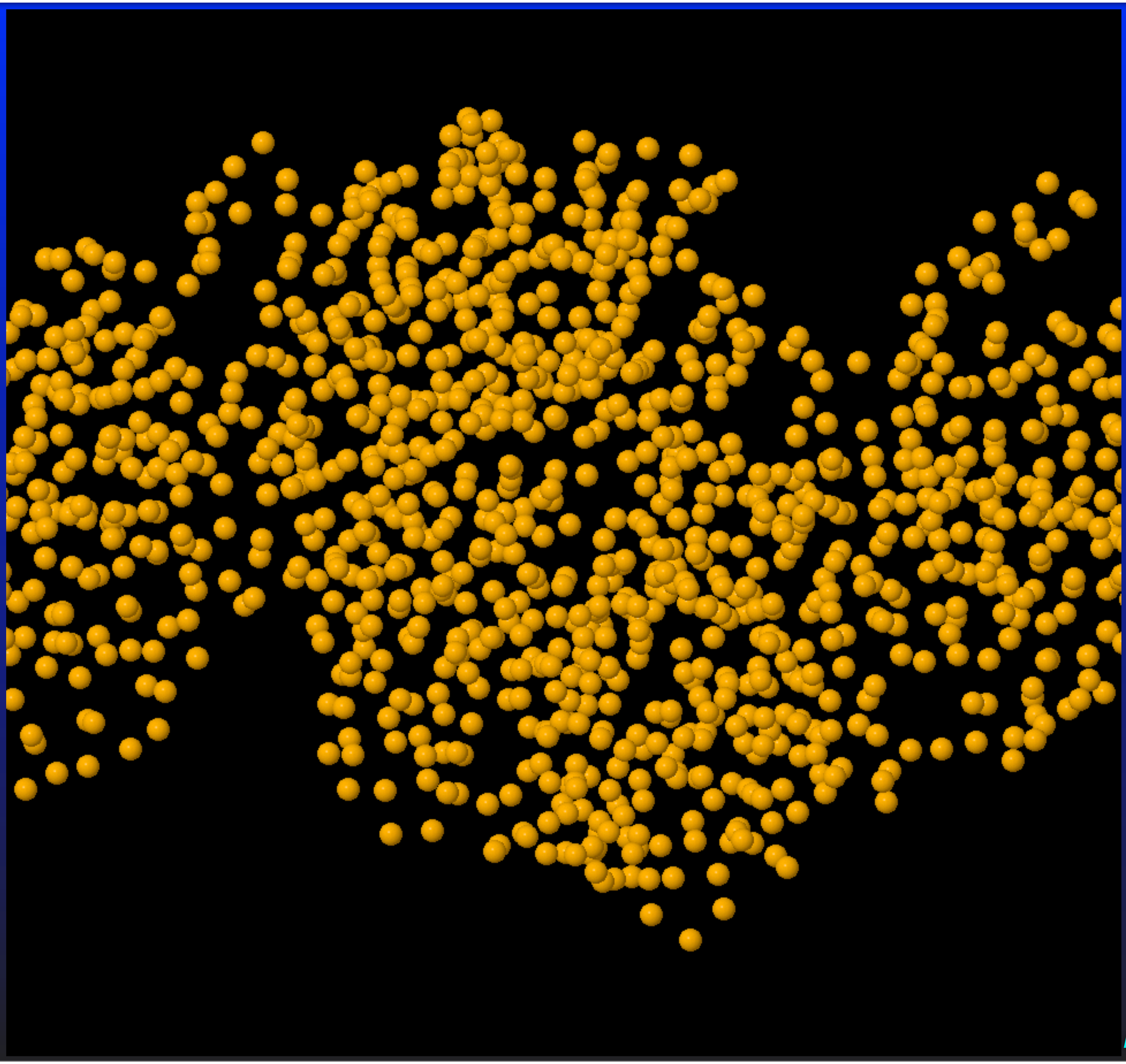
Macromolecular Complexes in Crystals and Solutions

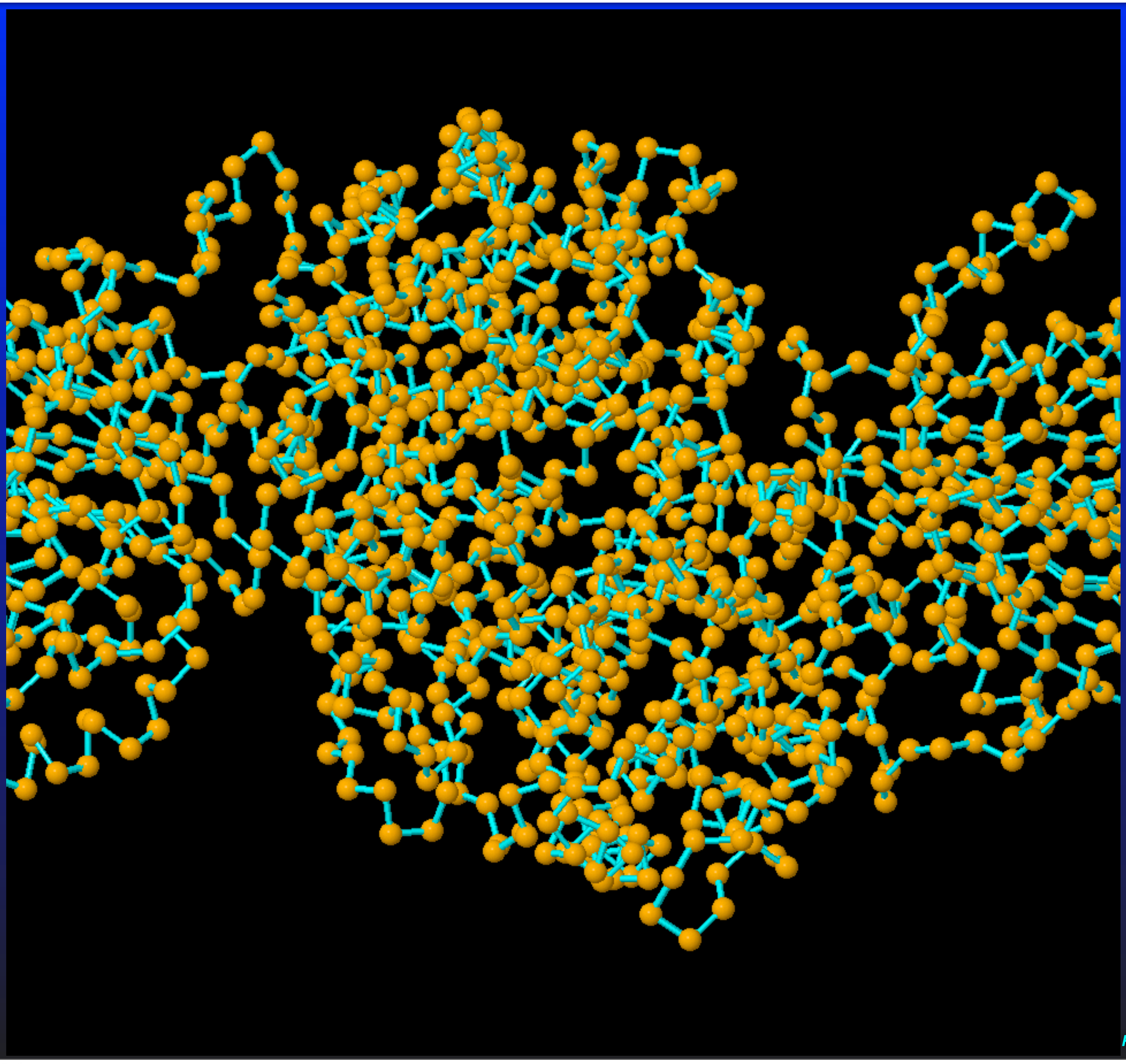
Eugene Krissinel
CCP4, STFC Research Complex at Harwell
Didcot, United Kingdom
eugene.krissinel@stfc.ac.uk

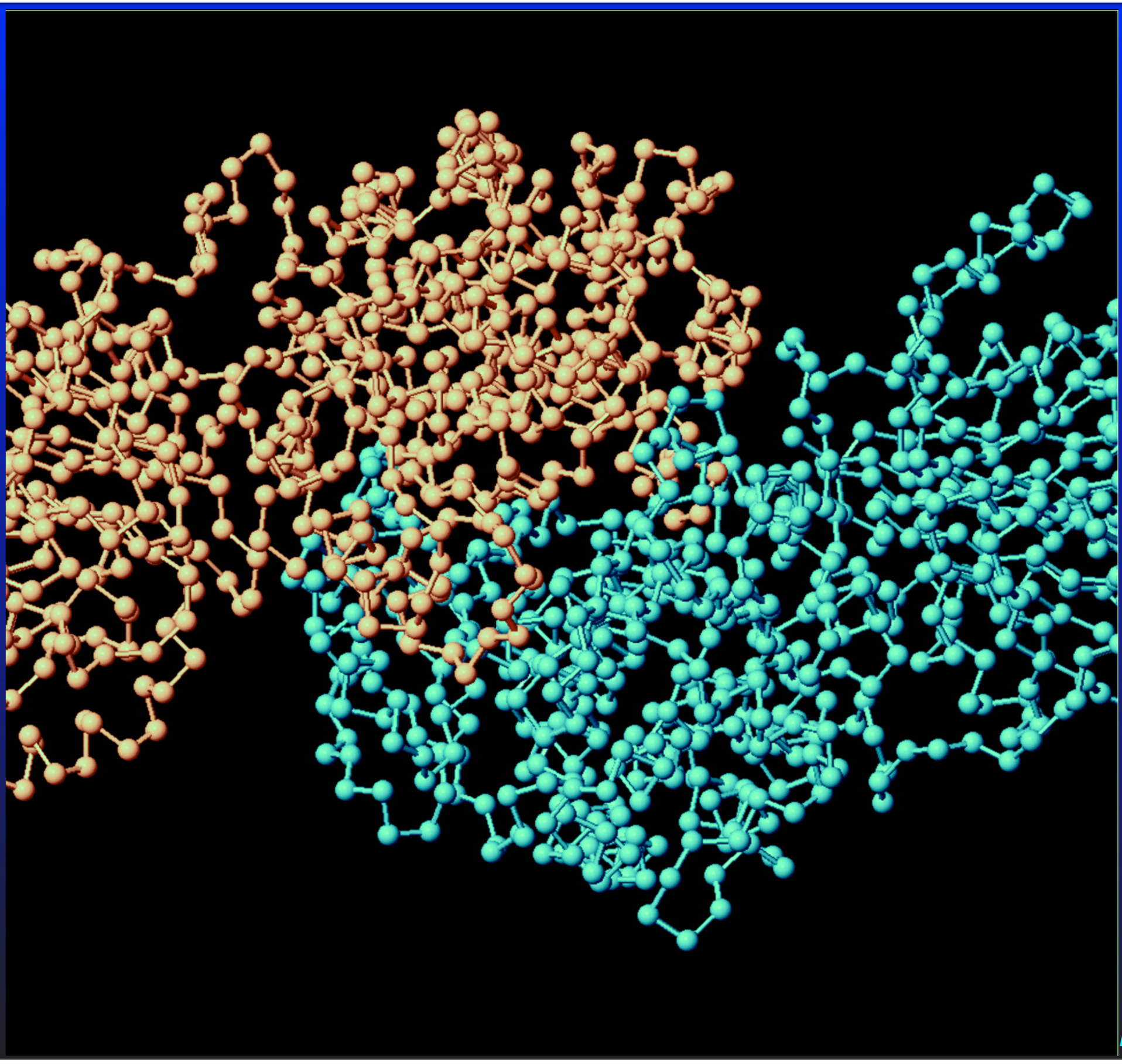
E. Krissinel and K. Henrick (2007) J. Mol. Biol. 372, 774-797
E. Krissinel (2010) J. Comp. Chem. 31, 133-143

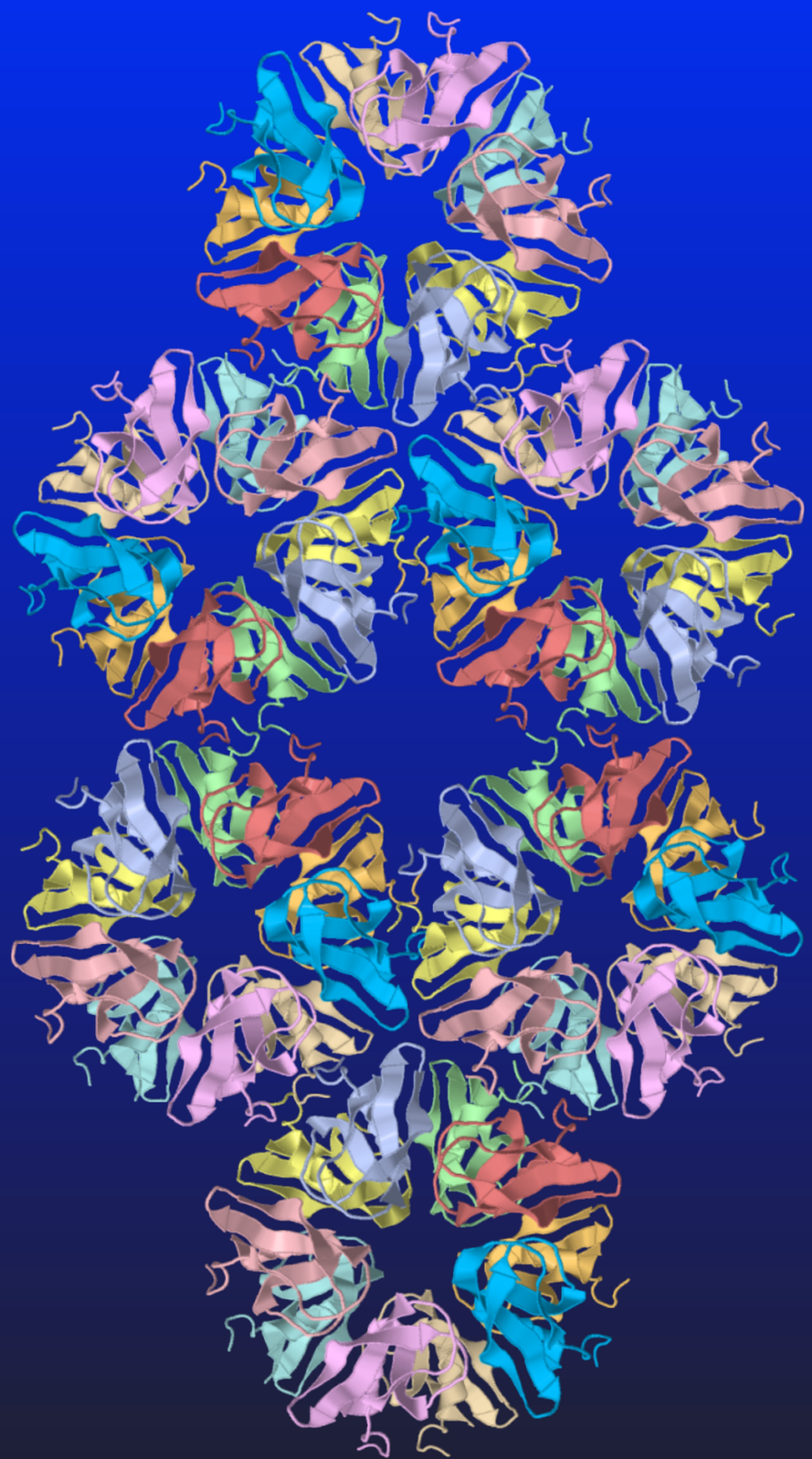
CCP4 Crystallography School in Fukuoka, October 29 - November 2, 2012, Fukuoka, Japan

Research Complex at Harwell









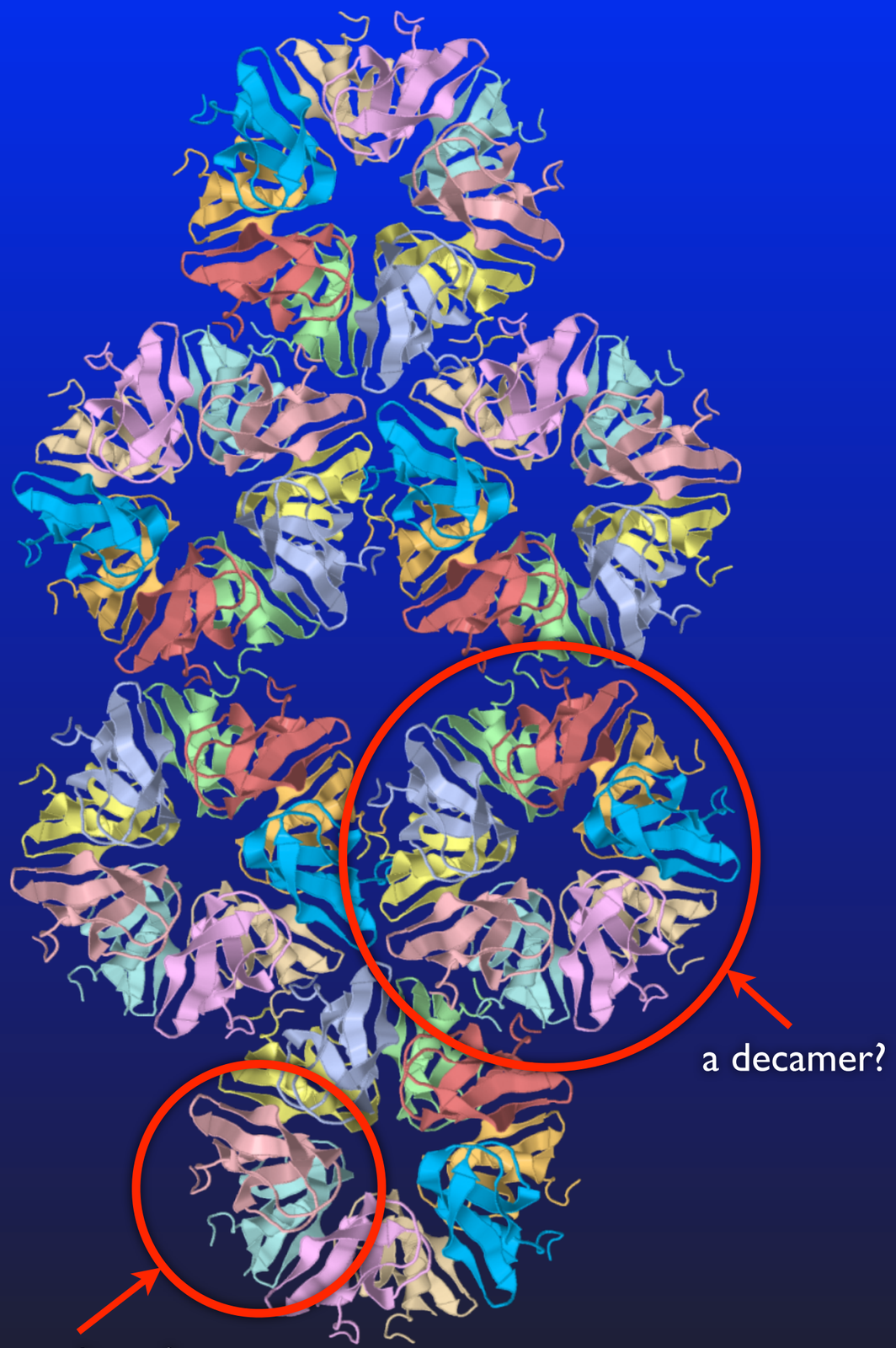
★ Why would we want to know structure of a macromolecule?

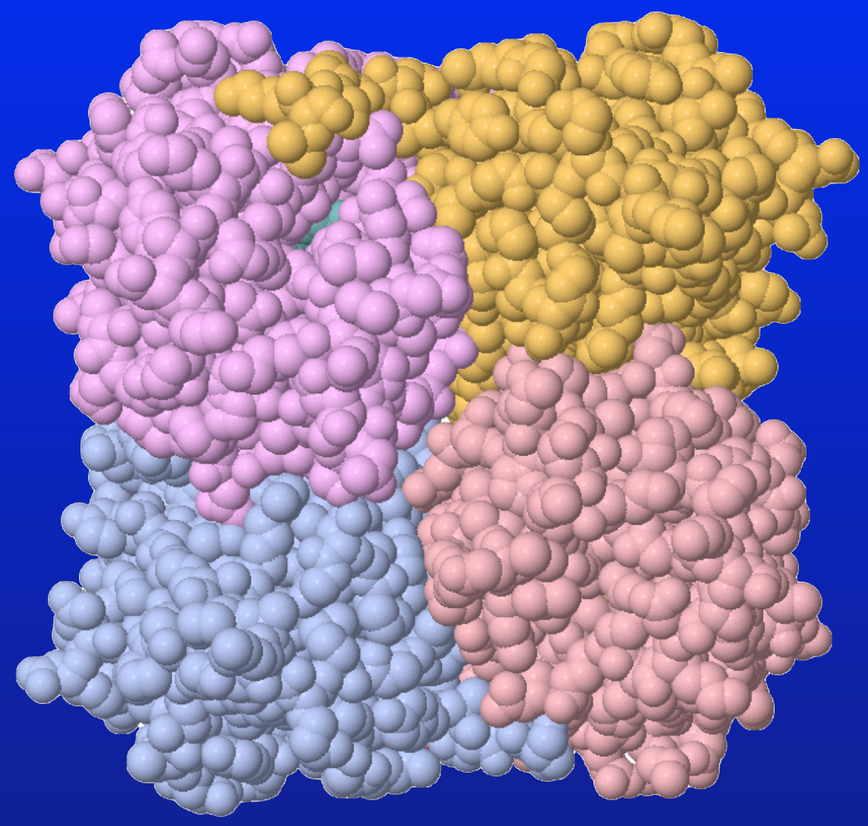
➡ for many reasons, but probably firstly for finding out how it interacts with other molecules

★ Macromolecular crystals present us with models of biological structures and interactions between them

➡ "if you want to know how A interacts with B - crystallize them together!" (crystallographer's sweet dream)

Research Complex at Harwell

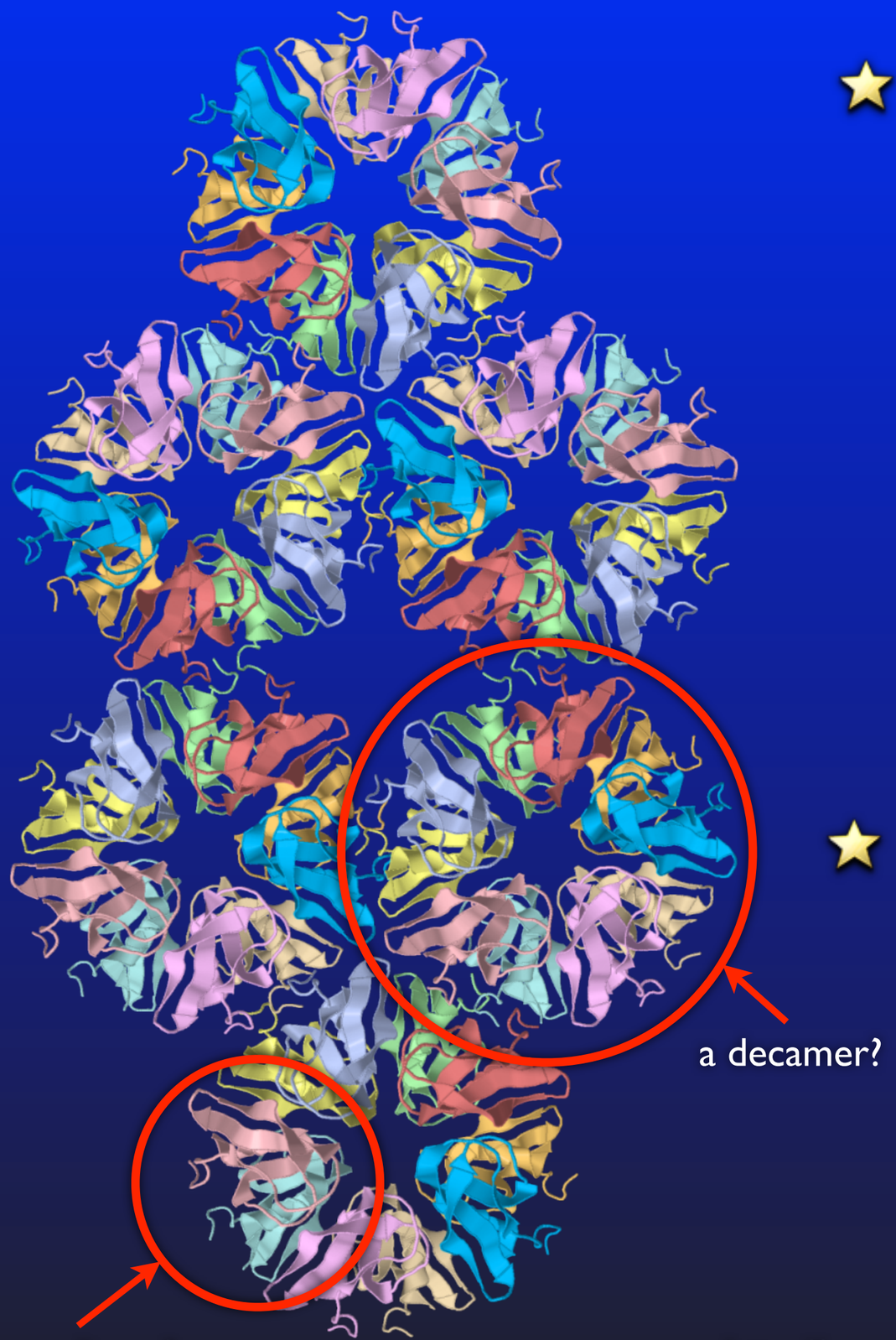




3bxc

"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



★ Crystals present us with both real and artifactual interactions, which may be difficult to differentiate. Frequently used techniques:

Theoretical:	sharp eye and scientific authority
Rules of thumb:	e.g. manifestation in different crystal forms
Experimental:	complementing studies (MS, EM, NMR, scattering)
Bioinformatical:	homology and interface similarity analysis
Computational:	energy estimates and modelling

★ PISA software infers significant interactions and macromolecular assemblies from crystals by evaluating their free Gibbs energy:

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

http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

Research Complex at Harwell

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

Protein interfaces

http://www.ebi.ac.uk/msd-srv/prot_int/cgi-bin/piserver

Home RAL Mail Google Mail Yahoo! Mail Staff Central Google Maps Wikipedia Apple YouTube

EMBL-EBI EB-eye Search All Databases

Databases Tools EBI Groups Training Indus

Home › Databases › PDBe › Services › PISA

Submission Form fo

Protein structure to be ed:

Coordinate file
PDB entry

1e94.pdb uploade
Specify another fi

Wait for page to update after you change

6 aminoacid chains and **2** ligands

Cell parameters:

A:	172.022	Alpha:	90
B:	172.022	Beta:	90
C:	276.569	Gamma:	120

You may change the values of the cell
Your input will supe

Process ligands: ANP

Processing mode: Auto

interfaces

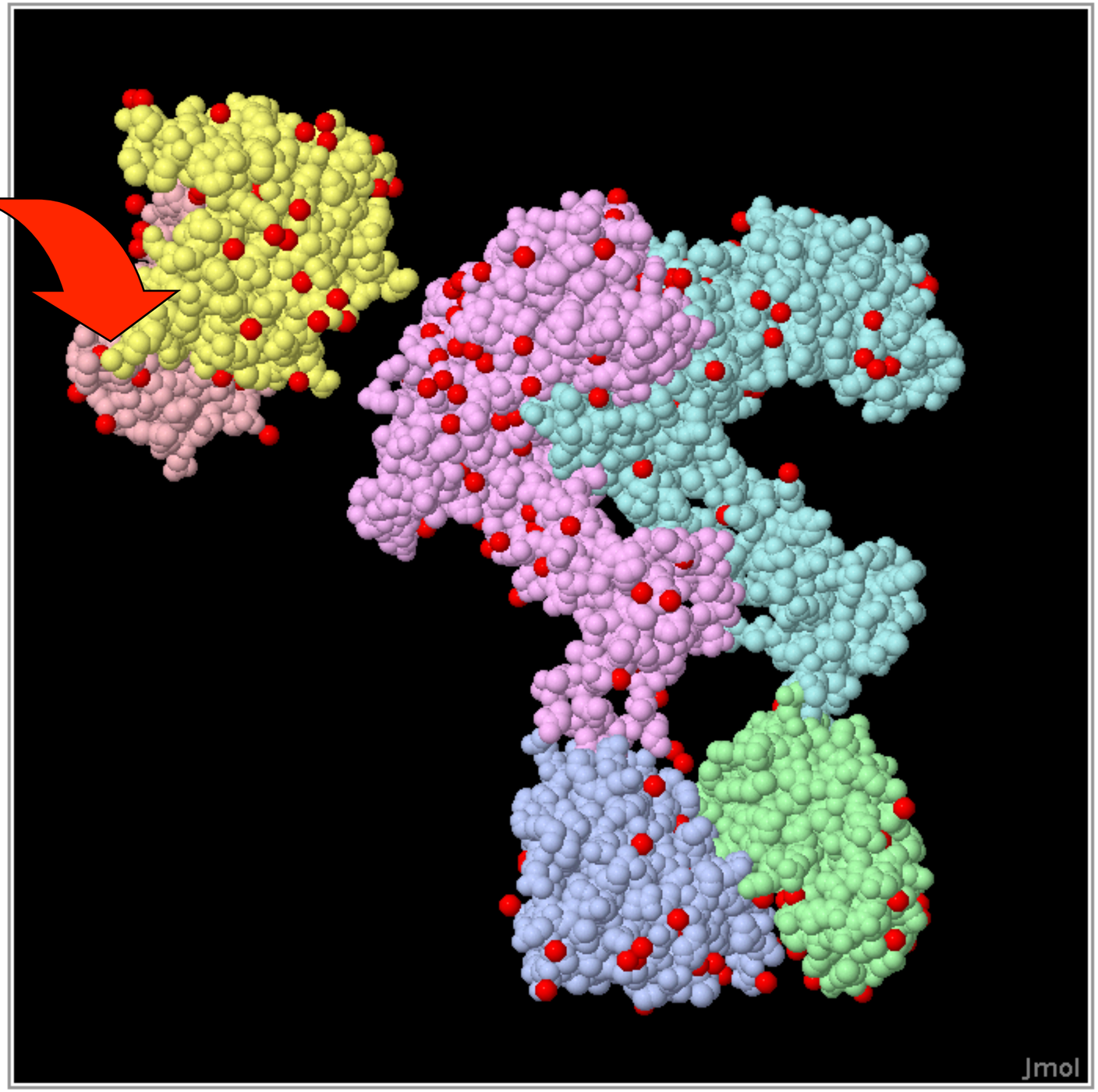
PISA v1.18 << 30/03/2009 >>

developer: Dr. Eugene Krissinel
last modified: 30/03/2009

the C
of the Biotechnology and Biological Sciences Research Council

Terms of Use | EBI Funding | Contact EBI | © European Bioinformatics Institute 2010. EBI is an Outstation of the European Molecular Biology Laboratory.

1e94.pdb



Rendering: Spacefill
Background: Black
Zoom: 100%
Screen: 60%
spin

Close

Protein interfaces

http://www.ebi.ac.uk/msd-srv/prot_int/cgi-bin/piserver

Home RAL Mail Google Mail Yahoo! Mail Staff Central Google M

EMBL-EBI EB-eye Search All Databases Enter Text Here

Databases Tools EBI Groups Training Industry About U

Home › Databases › PDBe › Services › PISA

Session **766-54-O99** map

query **1e94.pdb** ⇨ interfaces • monomers • **assemblies** • ⇨ *interface search results* *interfaces* *monomers* *assemblies*

PQS sets 1 to 10 of total 10

Analysis of complex represented *As Is* by uploaded file is found

NOTE: In one or more solutions (PQS sets), not all chains appear in their origi to the set number.

Analysis of protein interfaces suggests that the following qu

PQS set NN	«»	mm Size	Formula	Composition	Id	Biomol R350	Stable	a	
1	◉	12	A_{12}	A_6B_6	1	1	yes	77220	25620
	○	12	A_{12}	C_6D_6	1	2	yes	77200	25140
	○	6	A_6a_6	$E_3F_3[ANP]_6$	2	3	yes	104500	32220
2	○	12	A_{12}	A_6B_6	1	1	yes	77220	25620
	○	12	A_{12}	C_6D_6	1	2	yes	77200	25140
	○	1	Aa	E[ANP]	3	–	yes	21850	850
	○	1	Aa	F[ANP]	3	–	yes	22030	840
3	○	6	A_6a_6	$E_3F_3[ANP]_6$	2	3	yes	104500	32220
	○	1	A	B	4	–	yes	8560	0
	○	1	A	C	4	–	yes	8460	0
	○	1	A	D	4	–	yes	8600	0
	○	1	A	A	4	–	yes	8580	0

Assembly 1e94.pdb//1:1

Rendering: Spacefill
Background: Black
Zoom: 100%
Screen: 60%
spin
Close

Assembly 1e94.pdb//1:3

Rendering: Spacefill
Background: Black
Zoom: 100%
Screen: 60%
spin
Close

Jmol

Terminal — bash — 88×11

```
rcccp4mw001:setup Eugene$ ./pisa 1e94 -analyse 1e94.pdb pi

 ... session '1e94' created
 ... structure read and understood
 ... structural equivalence analysed
 ... molecular surfaces analysed
 ... 30 potential contacts identified, now processing:
     ||||||||||||||||||||||||||||||
 ... total 24 interfaces calculated
 ... assembly analysis: done
 ... results written
```

Terminal — bash — 90×25

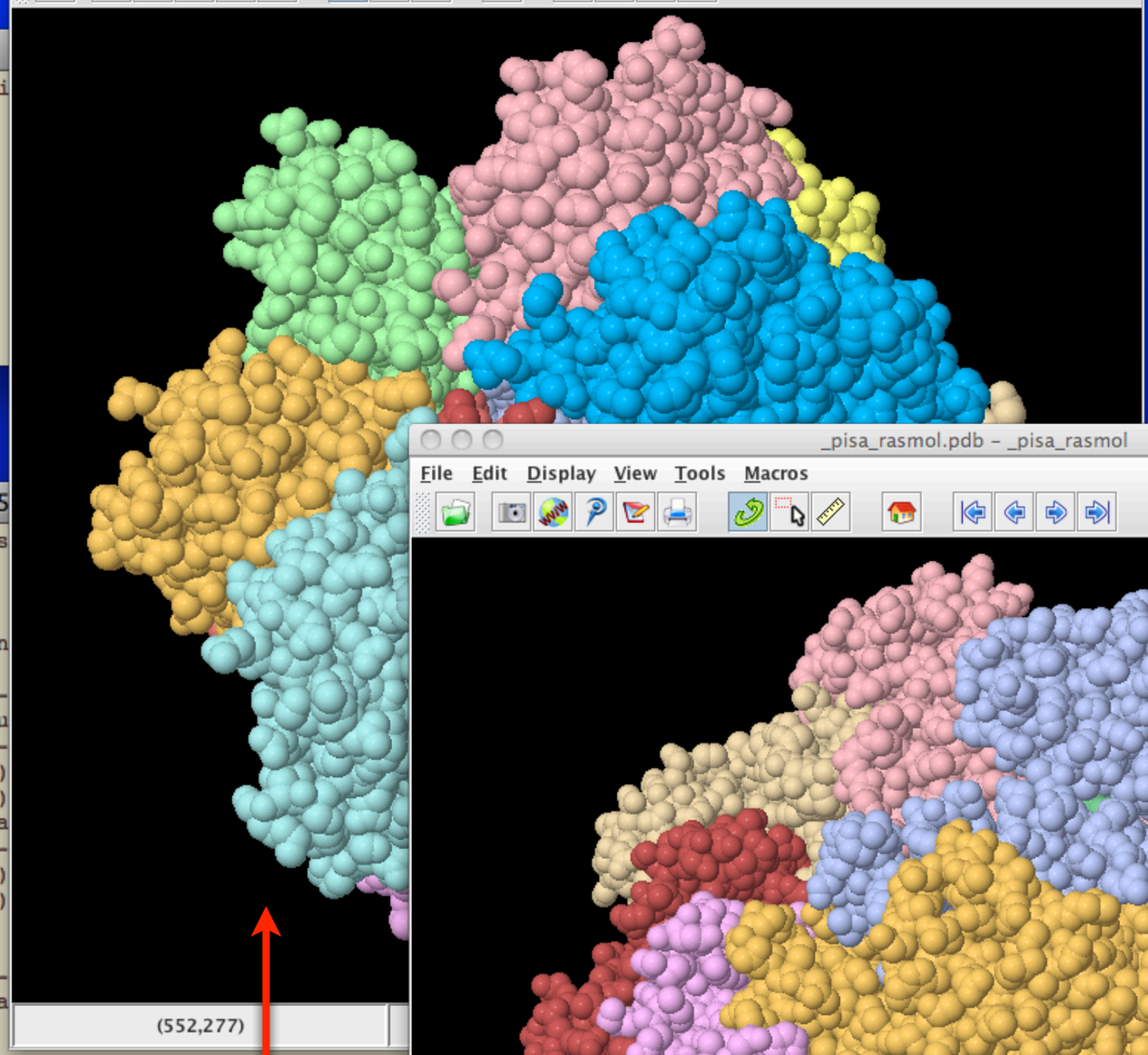
```
rcccp4mw001:setup Eugene$ ./pisa 1e94 -list assemblies pis

 PISA v1.06 << 16/01/2010 >>   Session 1e94

 Analysis of protein interfaces suggests that the followin
 quaternary structures are stable in solution
 ----.-----.---------------------------------------.------
 Set |  No | Size  Id      ASA       BSA     DGdiss | Formu
 ----+-----+---------------------------------------+------
   1 |   1 |   12    1   77218.0   25619.7     31.6 | A(12)
     |   2 |   12    1   77204.0   25139.2     22.4 | A(12)
     |   3 |    6    2  104504.0   32216.1     69.2 | A(6)a
 ----+-----+---------------------------------------+------
   2 |   4 |   12    1   77218.0   25619.7     31.6 | A(12)
     |   5 |   12    1   77204.0   25139.2     22.4 | A(12)
     |   6 |    1    3   21852.2     852.8     -0.0 | Aa
     |   7 |    1    3   22032.3     836.1     -0.0 | Aa
 ----+-----+---------------------------------------+------
   3 |   8 |    6    2  104504.0   32216.1     69.2 | A(6)a
     |   9 |    1    4    8561.7       0.0     -0.0 | A
     |  10 |    1    4    8459.0       0.0     -0.0 | A
     |  11 |    1    4    8598.2       0.0     -0.0 | A
     |  12 |    1    4    8578.0       0.0     -0.0 | A
 ----'-----'---------------------------------------'--------------
```

_pisa_rasmol.pdb – _pisa_rasmol

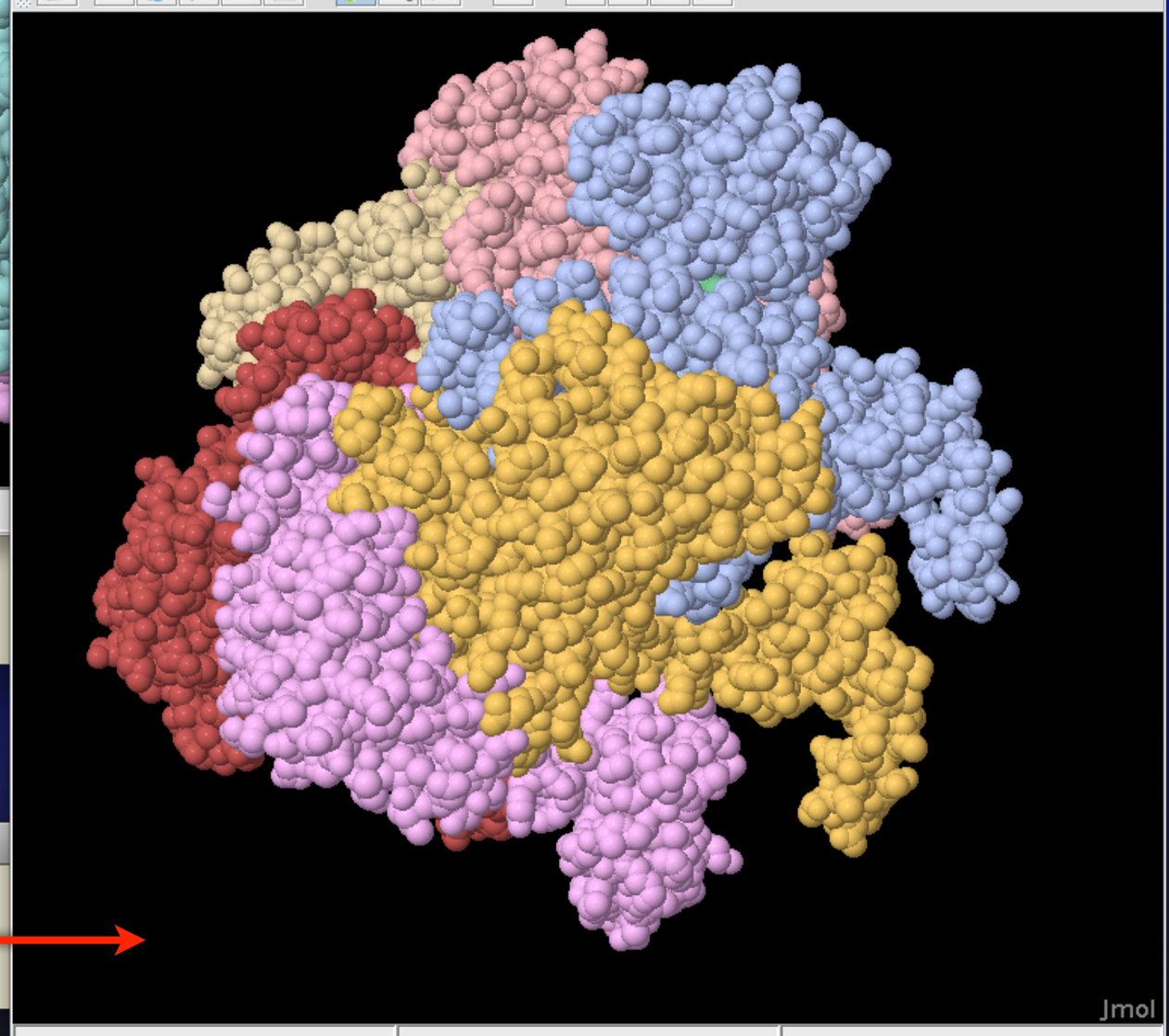
File Edit Display View Tools Macros Help



(552,277)

_pisa_rasmol.pdb – _pisa_rasmol

File Edit Display View Tools Macros Help



(379,297) | 762 x 670 | 63.1/105.6 Mb; 38/36 ms

Terminal — bash — 91×5

```
rcccp4mw001:setup Eugene$ ./pisa 1e94 -view assembly 1 pisa.cfg > /dev/null &
[1] 1189
rcccp4mw001:setup Eugene$ ./pisa 1e94 -view assembly 3 pisa.cfg > /dev/null &
[2] 1202
rcccp4mw001:setup Eugene$
```

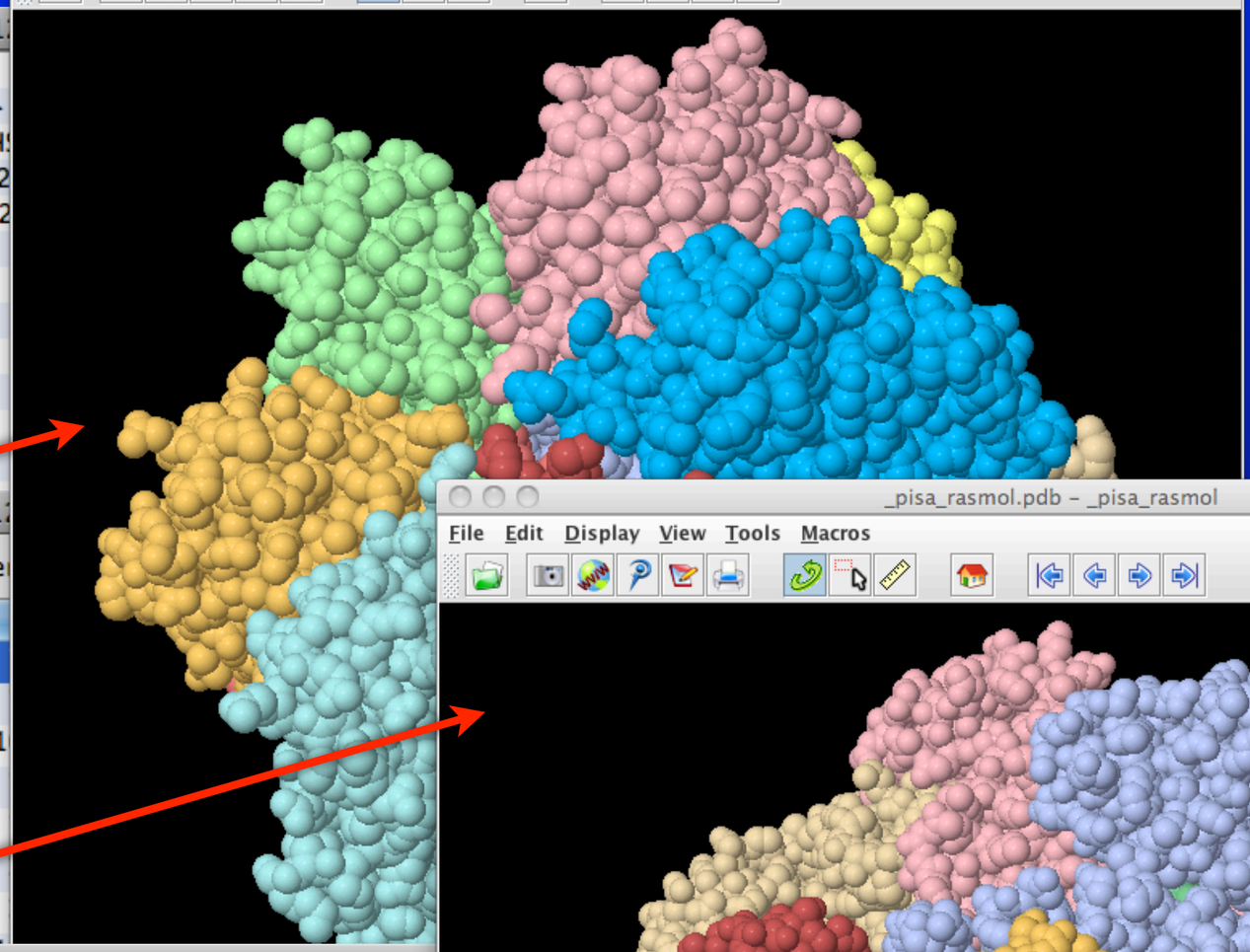
QtPISA 1.12 25/06/201

Contents
- Data
- Monomers
- Interfaces
- Assemblies
  - Stable
  - Grey area
  - Unstable
  - As given

File:	PDB =>
Title:	HSLV-H
Space group:	P 63 2 2
Cell:	172.022
ASU (File) contents:	
Protein chains:	6 (6)
DNA/RNA chains:	0 (0)
Ligands:	2 (2)
NCS-mates:	0

_pisa_rasmol.pdb – _pisa_rasmol

File Edit Display View Tools Macros Help



QtPISA 1.1

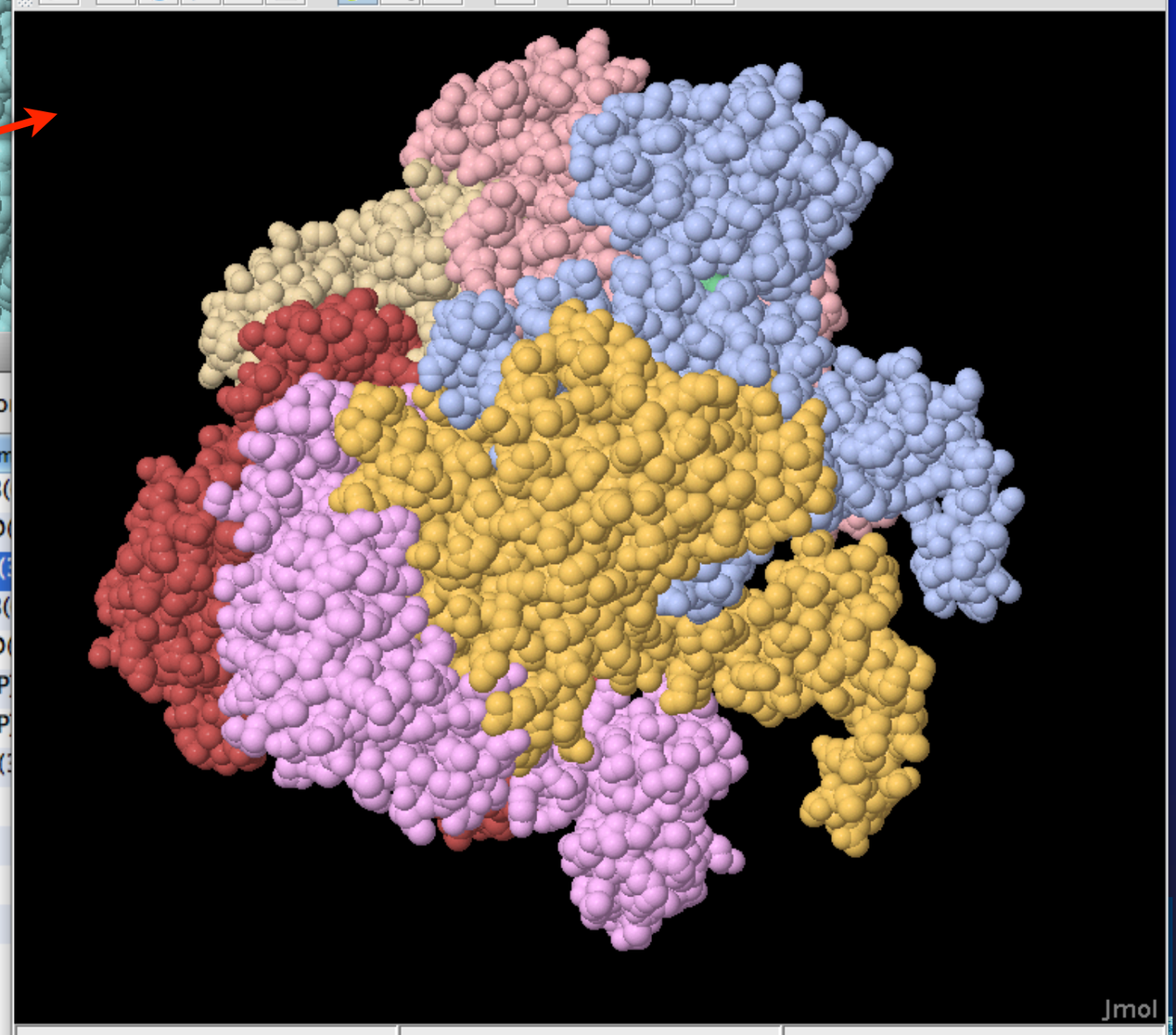
The following quate

Contents
- Data
- Monomers
- Interfaces
- Assemblies
  - Stable
    - 1. A(12)
    - 2. A(12)
    - 3. A(6)a(6)

	Set	Size	Type
1	1	12	1
2		12	1
3		6	2
4	2	12	1
5		12	1
6		1	3
7		1	3

_pisa_rasmol.pdb – _pisa_rasmol

File Edit Display View Tools Macros Help



Jmol

(379,297) | 762 x 670 | 63.1/105.6 Mb; 38/36 ms

QtPISA 1.12 25/06/2012 [PDB => 1e94]

The following quaternary structures appear to be stable in solutio

Contents
- Data
- Monomers
  1. A
  2. B
  3. C
  4. D
  5. E
  6. F
  7. [ANP]E:500
  8. [ANP]F:501
- Interfaces
- Assemblies
  - Stable
  - Grey area
  - Unstable
  - As given

	Set	Size	Type	ASA	BSA	Delta G	Formula	Com
1	1	12	1	77218.4	25619.7	31.6	A(12)	A(6)B(
2		12	1	77204.3	25139.2	22.4	A(12)	C(6)D(
3		6	2	104504.1	32216.1	69.2	A(6)a(6)	E(3)F(
4	2	12	1	77218.4	25619.7	31.6	A(12)	A(6)B(
5		12	1	77204.3	25139.2	22.4	A(12)	C(6)D(
6		1	3	21852.2	852.8	-0.0	Aa	E[ANP
7		1	3	22032.3	836.1	-0.0	Aa	F[ANP
8	3	6	2	104504.1	32216.1	69.2	A(6)a(6)	E(3)F(
9		1	4	8561.7	0.0	-0.0	A	B
10		1	4	8459.0	0.0	-0.0	A	C
11		1	4	8598.2	0.0	-0.0	A	D
12		1	4	8578.0	0.0	-0.0	A	A

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:

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

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 ΔG_0 take preference

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

Research Complex at Harwell

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 ⇔ 196 homomers and 22 heteromers					Total:	196+22	**90%**

Classification error in ΔG_0 : ± 5 kcal/mol

Research Complex at Harwell

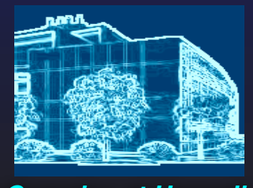
Classification of Protein-DNA Complexes

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.

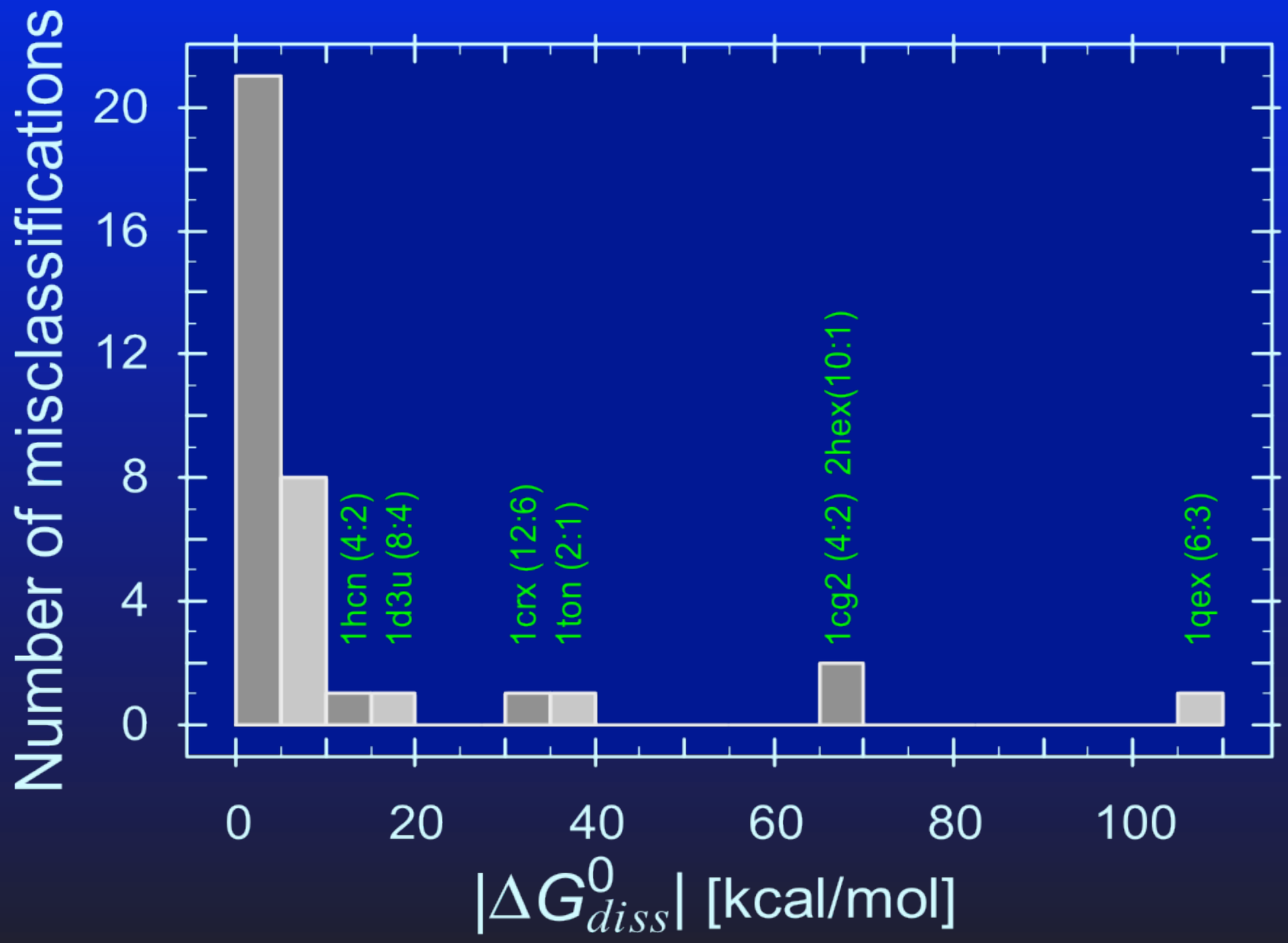
	2mer	3mer	4mer	5mer	6mer	10mer	Other	Sum	Correct
2mer	1	0	0	0	0	0	0	1	100%
3mer	6	96	0	0	1	0	2	105	91%
4mer	0	2	83	0	0	0	0	85	98%
5mer	0	0	2	3	0	0	0	5	60%
6mer	1	0	0	0	13	0	1	15	87%
10mer	0	0	0	0	0	1	0	1	100%
							Total:	212	**93%**

Classification error in ΔG_0 : ± 5 kcal/mol



Research Complex at Harwell

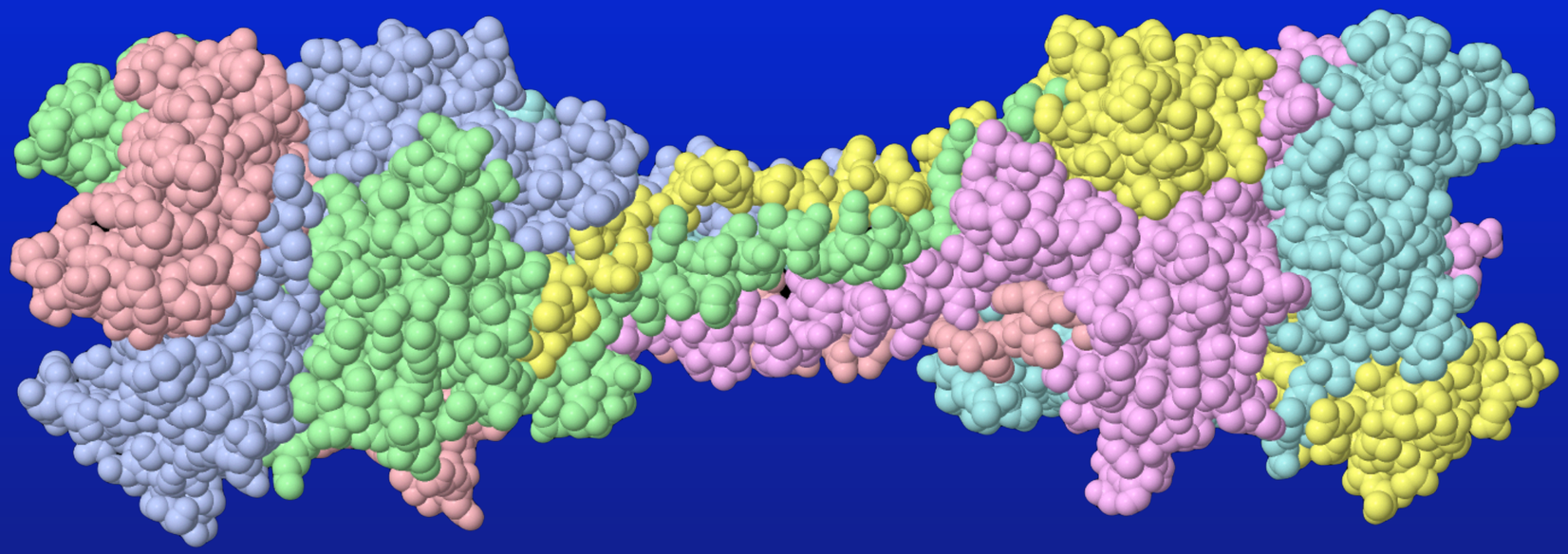
Free Energy Distribution of Misclassifications



Research Complex at Harwell

Example of misclassification: 1QEX

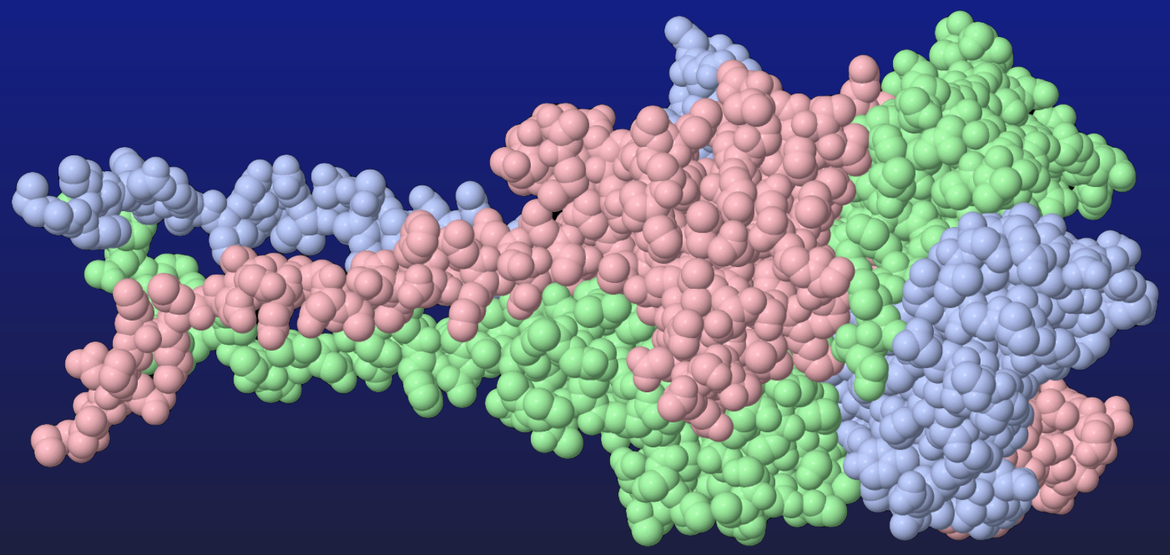
BACTERIOPHAGE T4 GENE PRODUCT 9 (GP9), THE TRIGGER OF TAIL CONTRACTION AND THE LONG TAIL FIBERS CONNECTOR



Predicted: homohexamer

Dissociates into 2 trimers

$\Delta G_0 \simeq 106$ kcal/mol



Biological unit: homotrimer

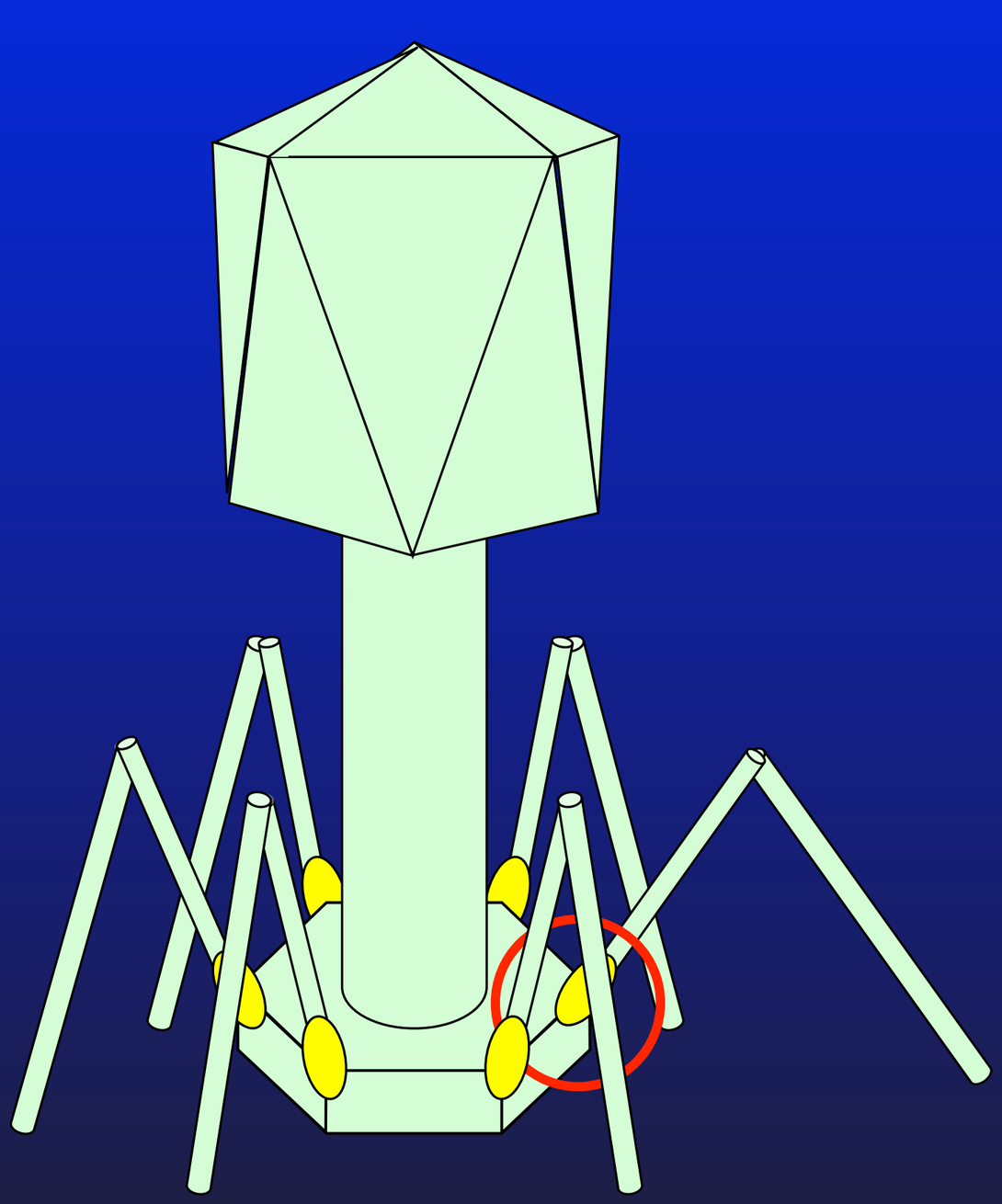
Dissociates into 3 monomers

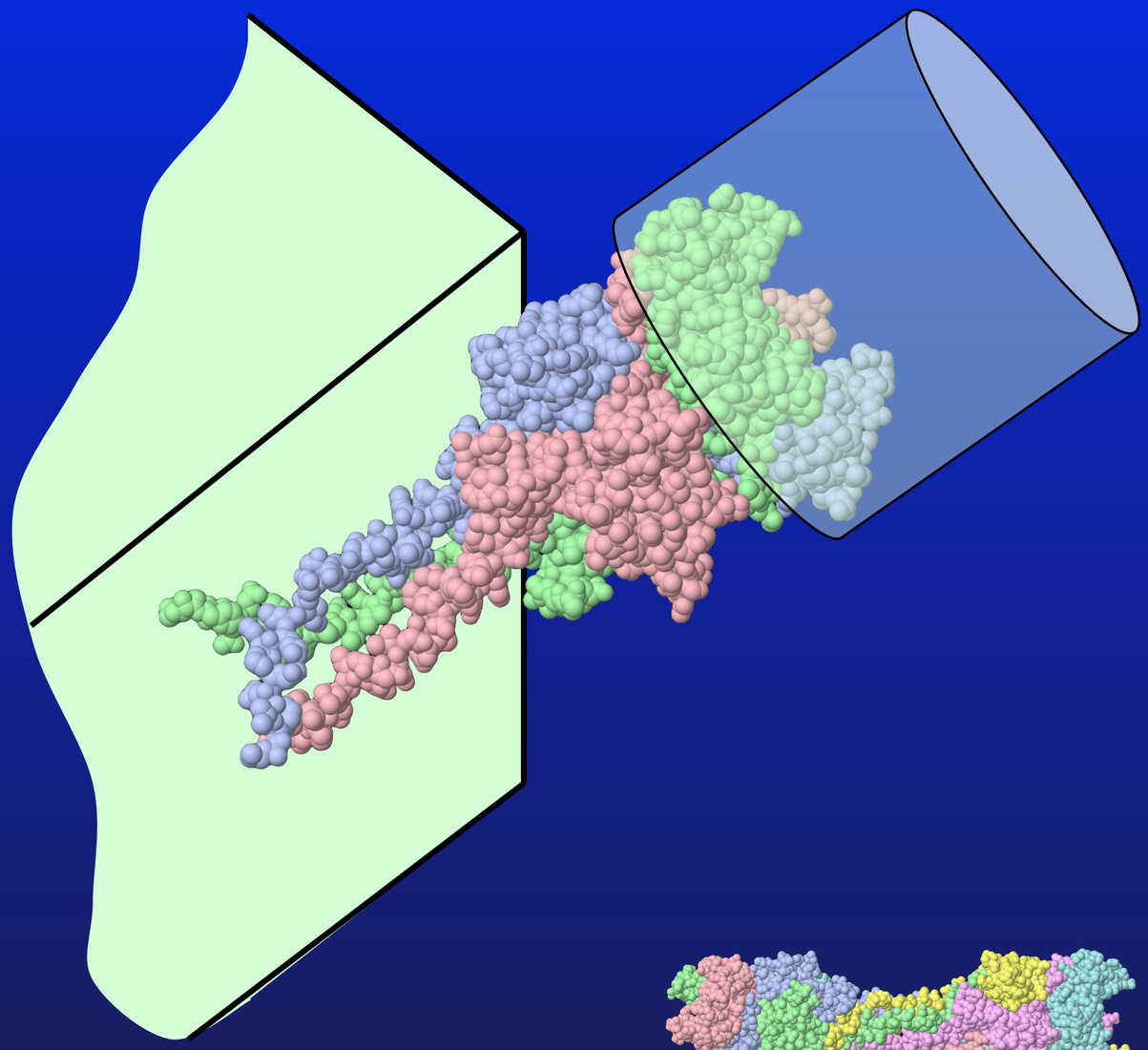
$\Delta G_0 \simeq 90$ kcal/mol

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

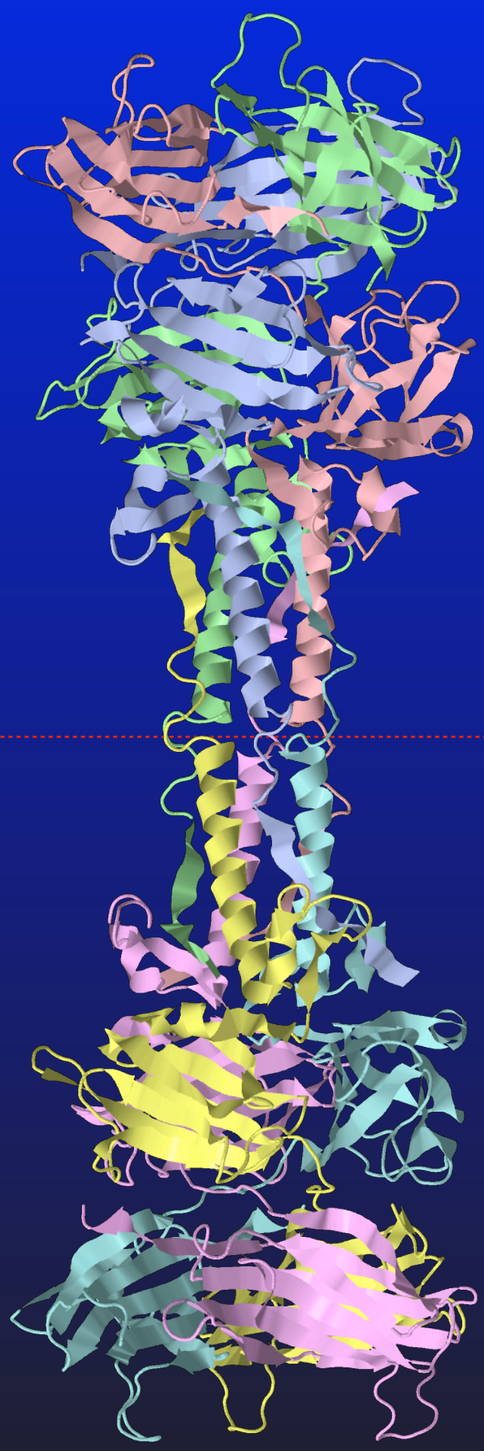




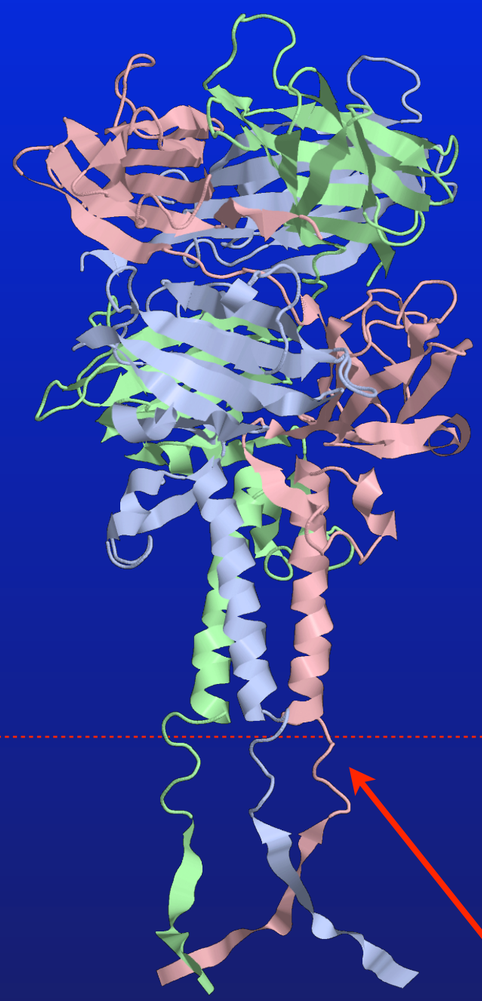
Rossmann M.G., Mesyanzhinov V.V., Arisaka F and Leiman P.G. (2004) *The bacteriophage T4 DNA injection machine*. Curr. Opinion Struct. Biol. 14:171-180.

Example of misclassification: 1QEX

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

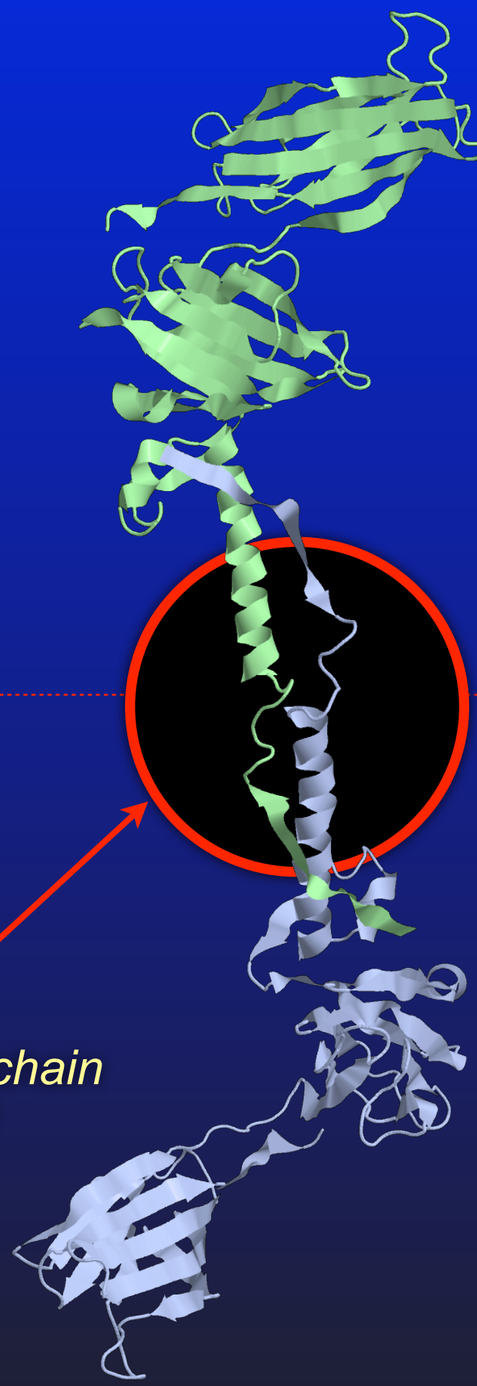


1QEX hexamer

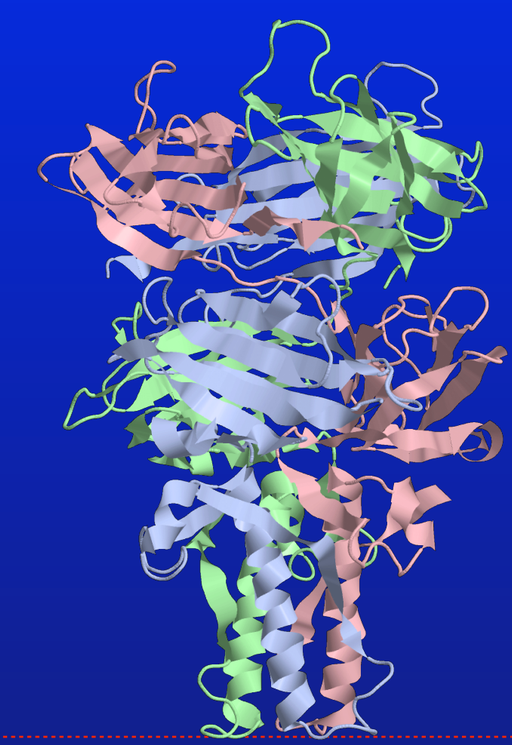


1QEX trimer

Wrong mainchain tracing!



1QEX asu

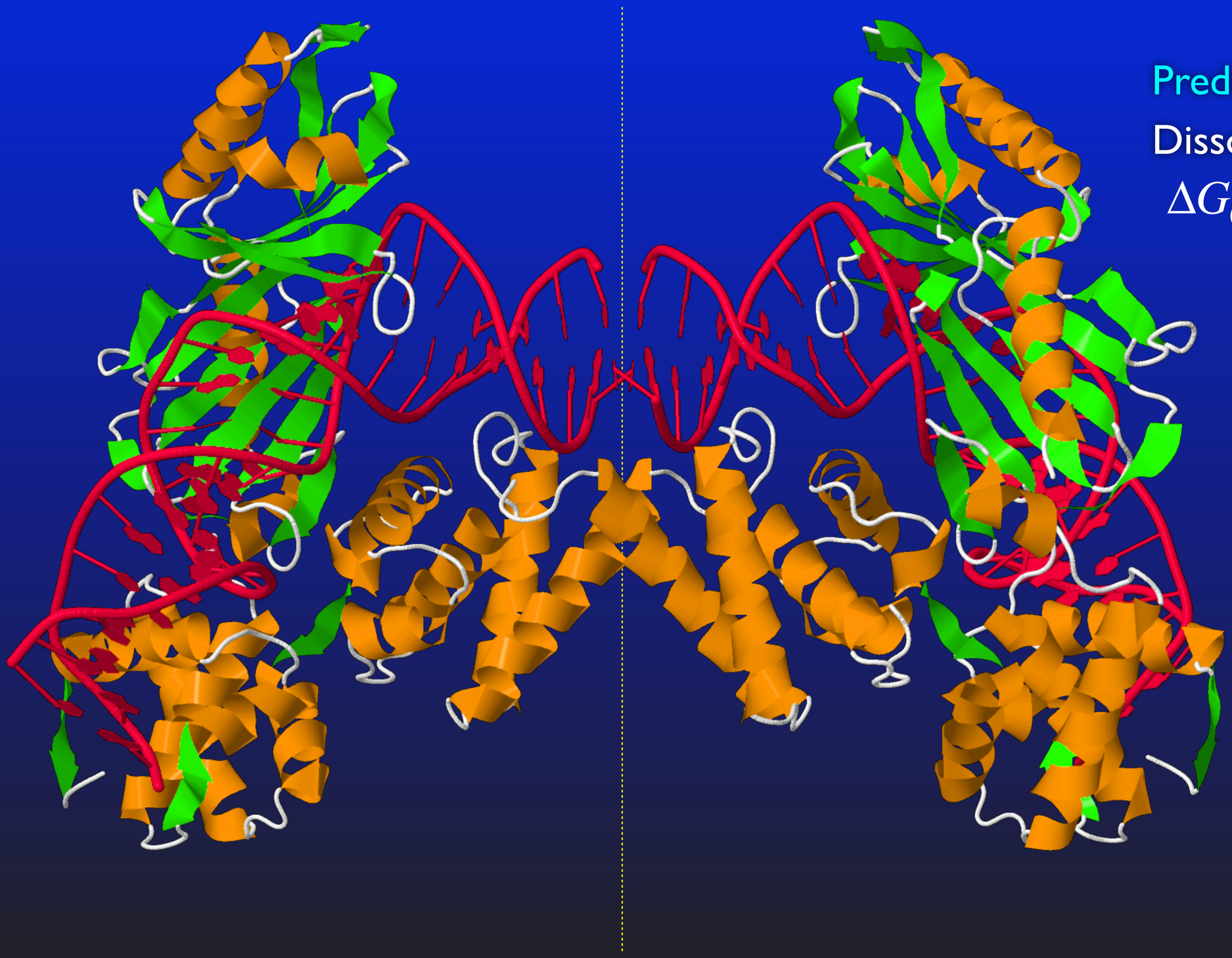


1S2E trimer
Correct mainchain tracing
Identified correctly

Research Complex at Harwell

Example of misclassification: 1D3U

TATA-BINDING PROTEIN / TRANSCRIPTION FACTOR



Predicted: octamer

Dissociates into 2 tetramers

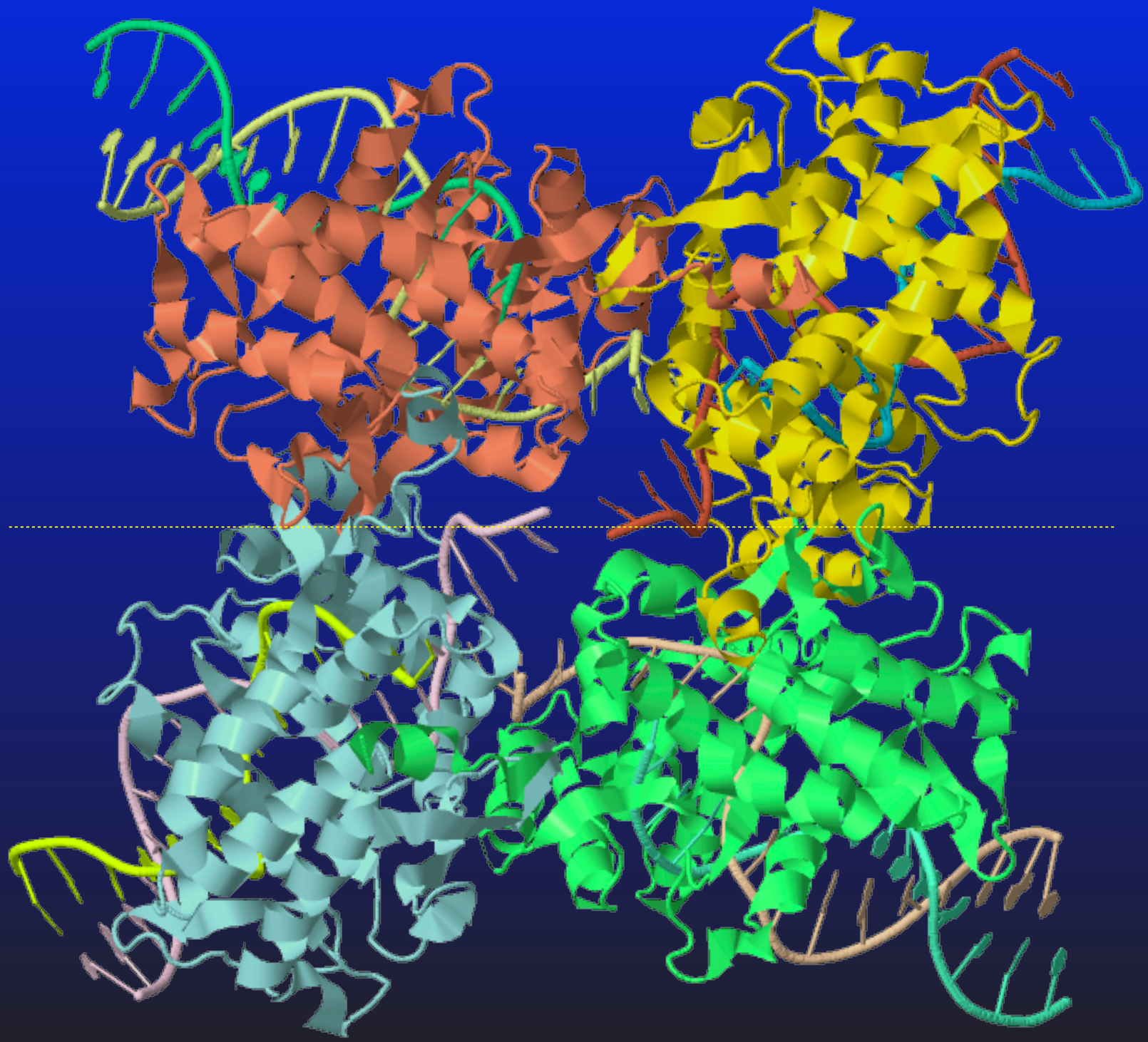
$\Delta G_0 \simeq 20$ kcal/mol

Functional unit:
tetramer

Research Complex at Harwell

Example of misclassification: 1CRX

CRE RECOMBINASE / DNA COMPLEX REACTION INTERMEDIATE



Predicted: dodecamer

Dissociates into 2 hexamers

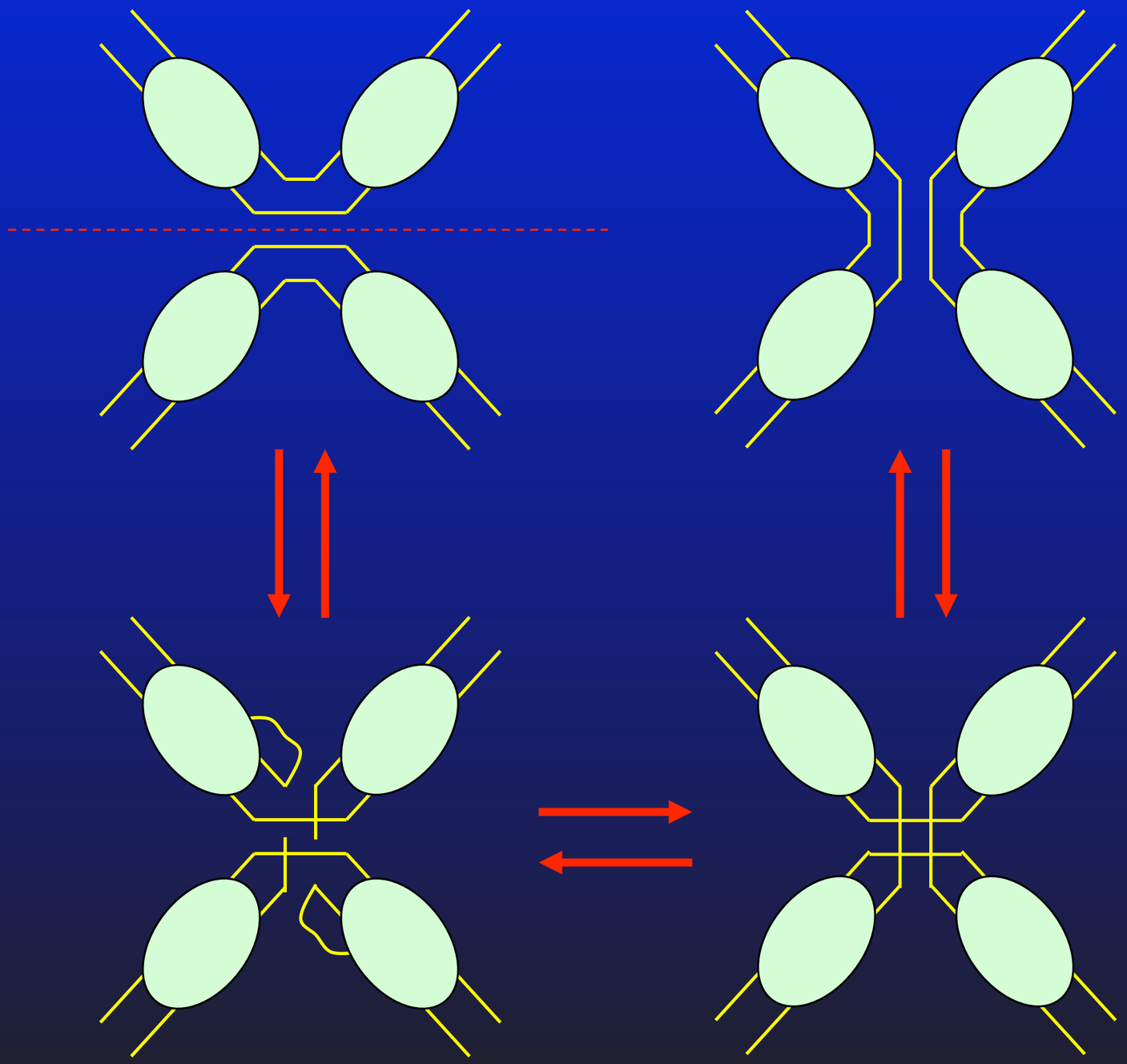
$\Delta G_0 \simeq 28$ kcal/mol

Functional unit: trimer

Research Complex at Harwell

Example of misclassification: 1CRX

CRE RECOMBINASE / DNA COMPLEX REACTION INTERMEDIATE



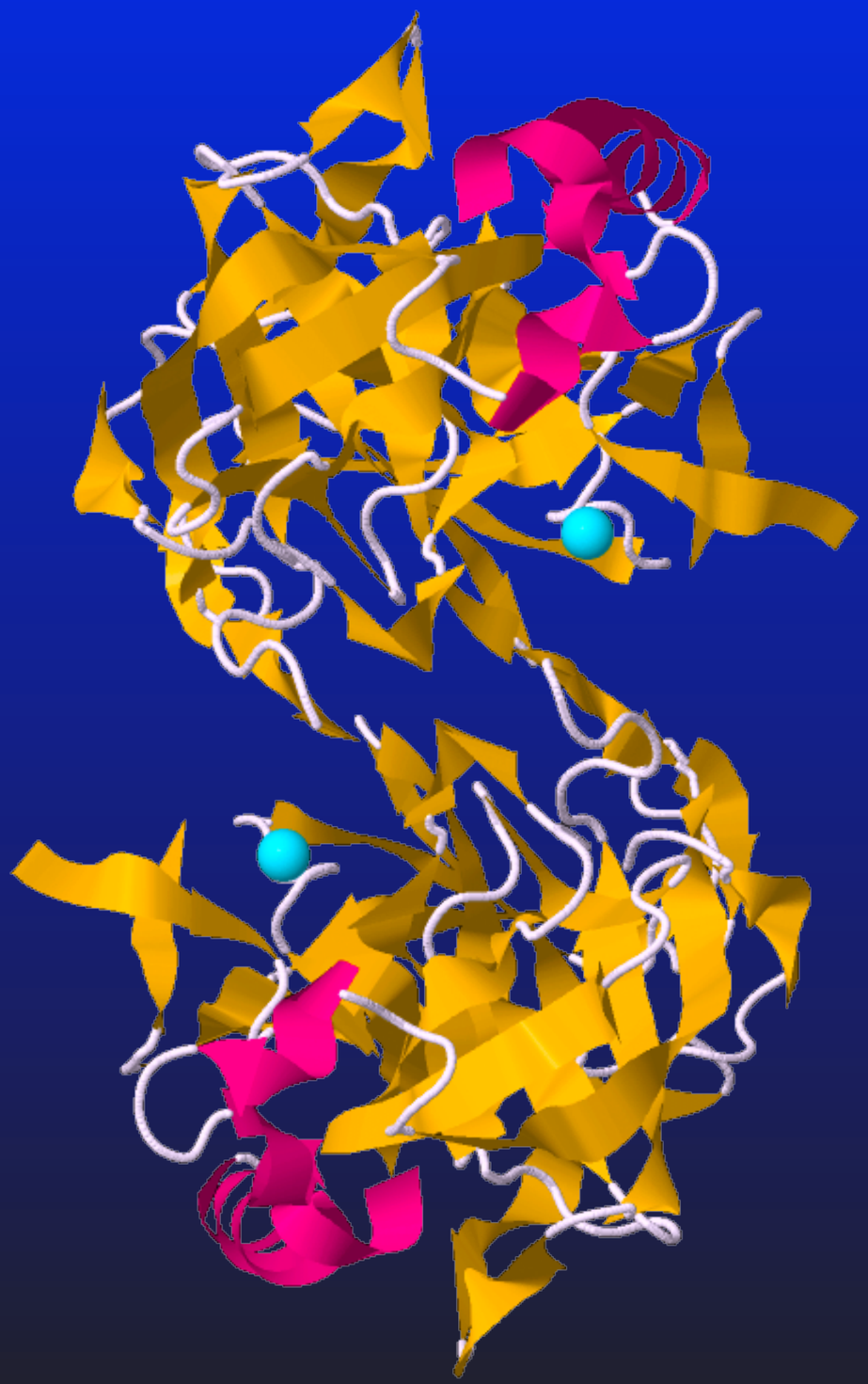
Guo F., Gopaul D.N. and van Duyne G.D. (1997)

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

Nature 389:40-46.

Example of misclassification: 1TON

TONIN



Predicted: dimer

Dissociates at

$\Delta G_0 \simeq 37$ kcal/mol

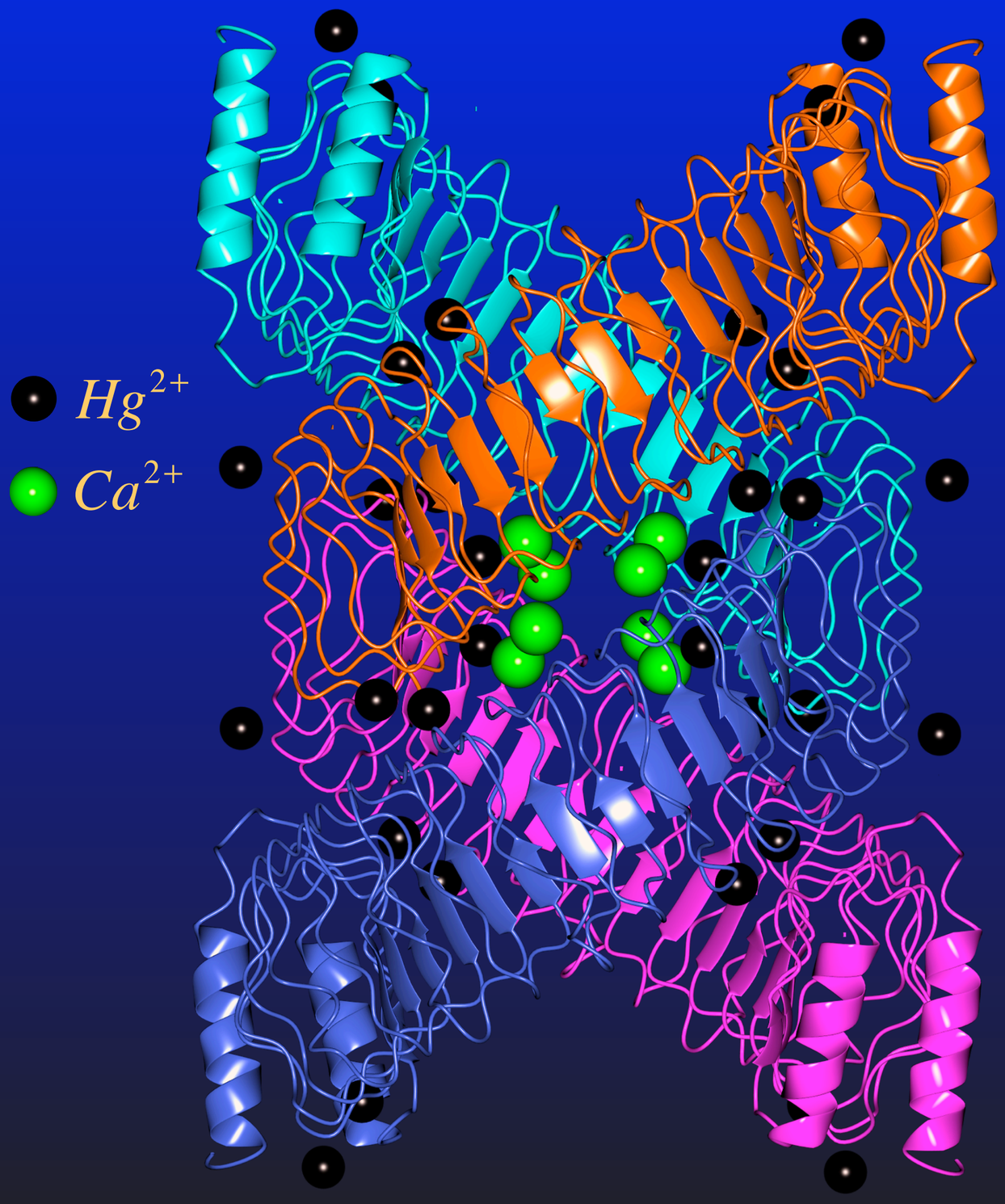
Biological unit: monomer

Apparent dimerization is an artefact due to the presence of Zn^{+2} ions added to the buffer to aid crystallization. Removal of Zn from the file results in $\Delta G_0 \simeq 3$ 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 ~35 Å.

	1G9U	1JL5
Space Group	$P4_22$	$I4_122$
ΔG_0, kcal/mol	37	3
Number of ions	40	16

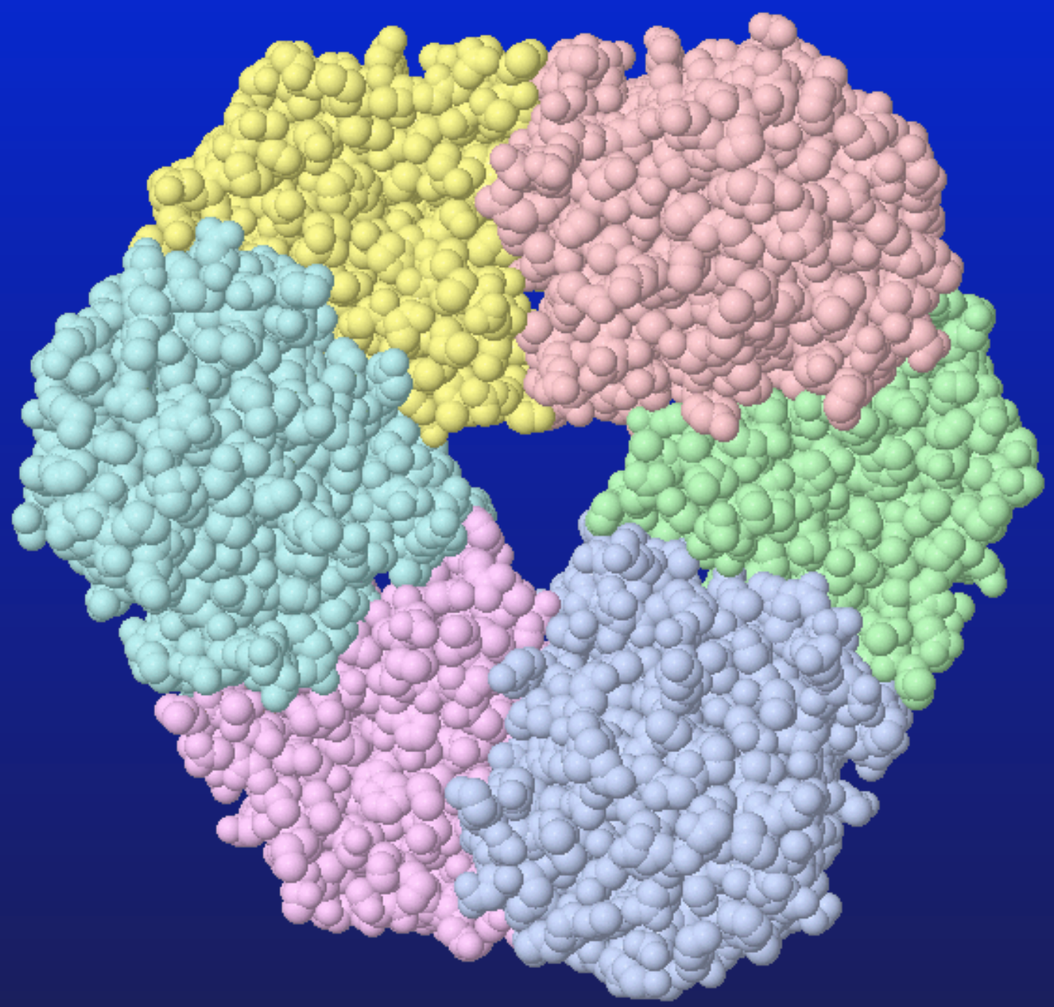
Biological unit: monomer

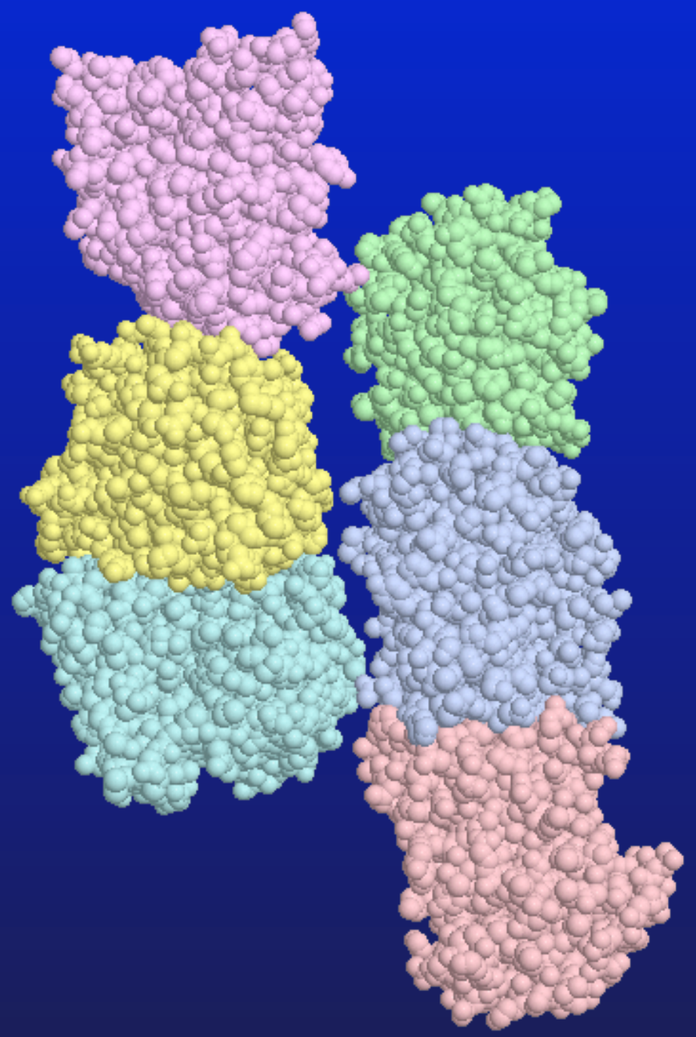
Evdokimov, A. G., Anderson, D. E., Routzahn, K. M. & Waugh, D. S. (2001). J. Mol. Biol. 312, 807–821

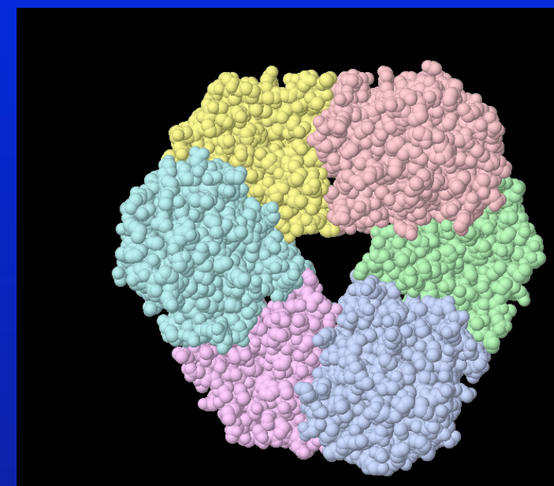
Removal of ions makes the structure monomeric in PISA estimates

Research Complex at Harwell

Example of misclassification: 1YWK







Structural homologue 1XRU:

RMSD = 0.9 A

Seq.Id = 50%

Homohexameric with

$\Delta G_0 \simeq$ 20 kcal/mol

Predicted: homohexameric

Dissociates into 3 dimers at

$\Delta G_0 \simeq 4.4$ kcal/mol

Believed to be:

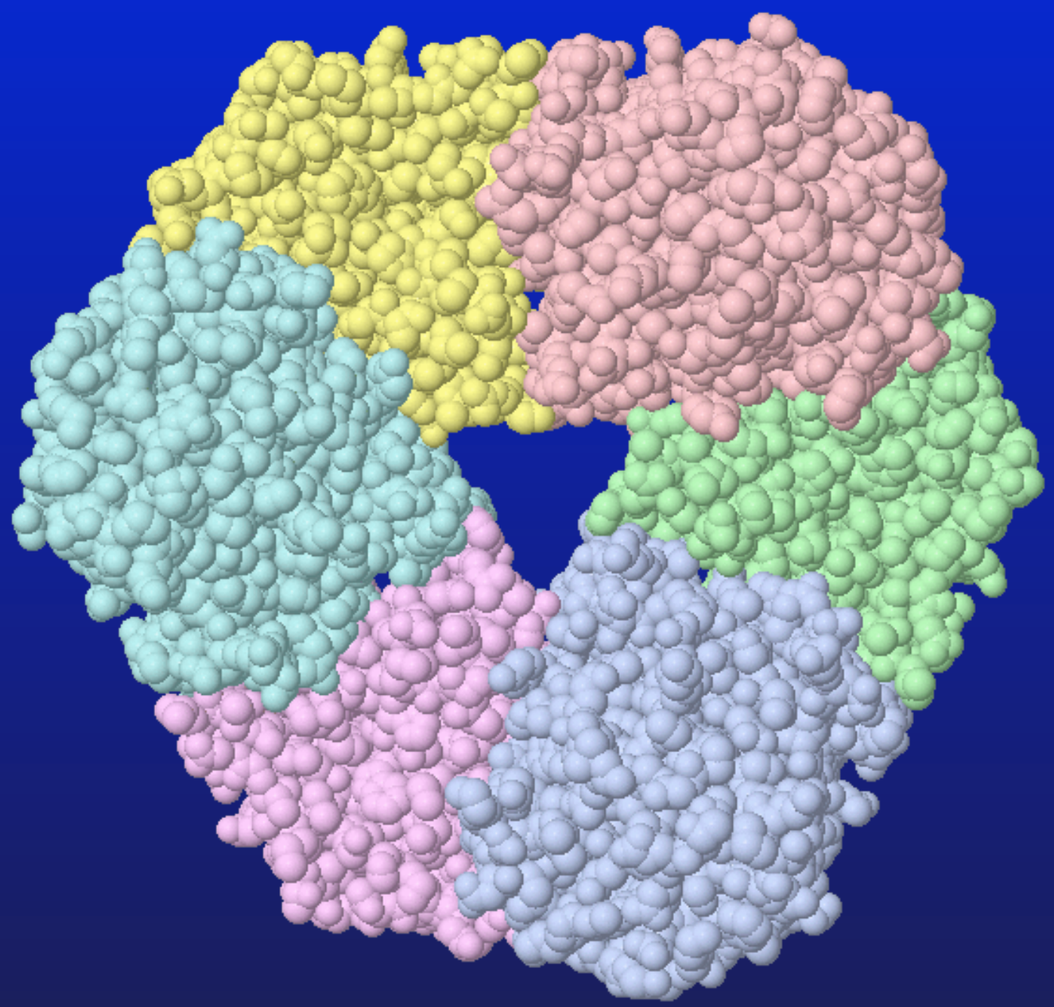
monomeric,

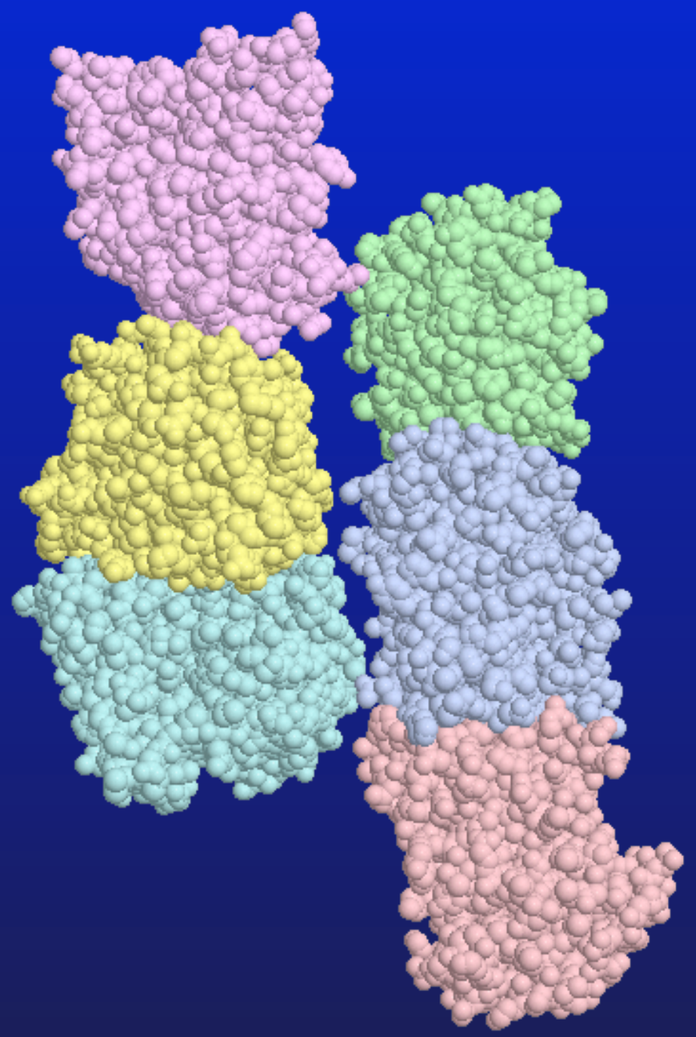
6 units in ASU

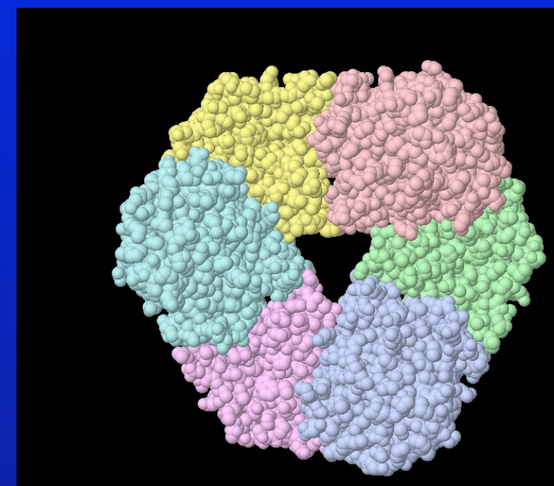
Research Complex at Harwell

Choice of ASU

Example of misclassification: 1YWK







Structural homologue 1XRU:

RMSD = 0.9 A

Seq.Id = 50%

Homohexameric with

$\Delta G_0 \simeq$ 20 kcal/mol

Predicted: homohexameric

Dissociates into 3 dimers at

$\Delta G_0 \simeq 4.4$ kcal/mol

Believed to be:

monomeric,

6 units in ASU

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

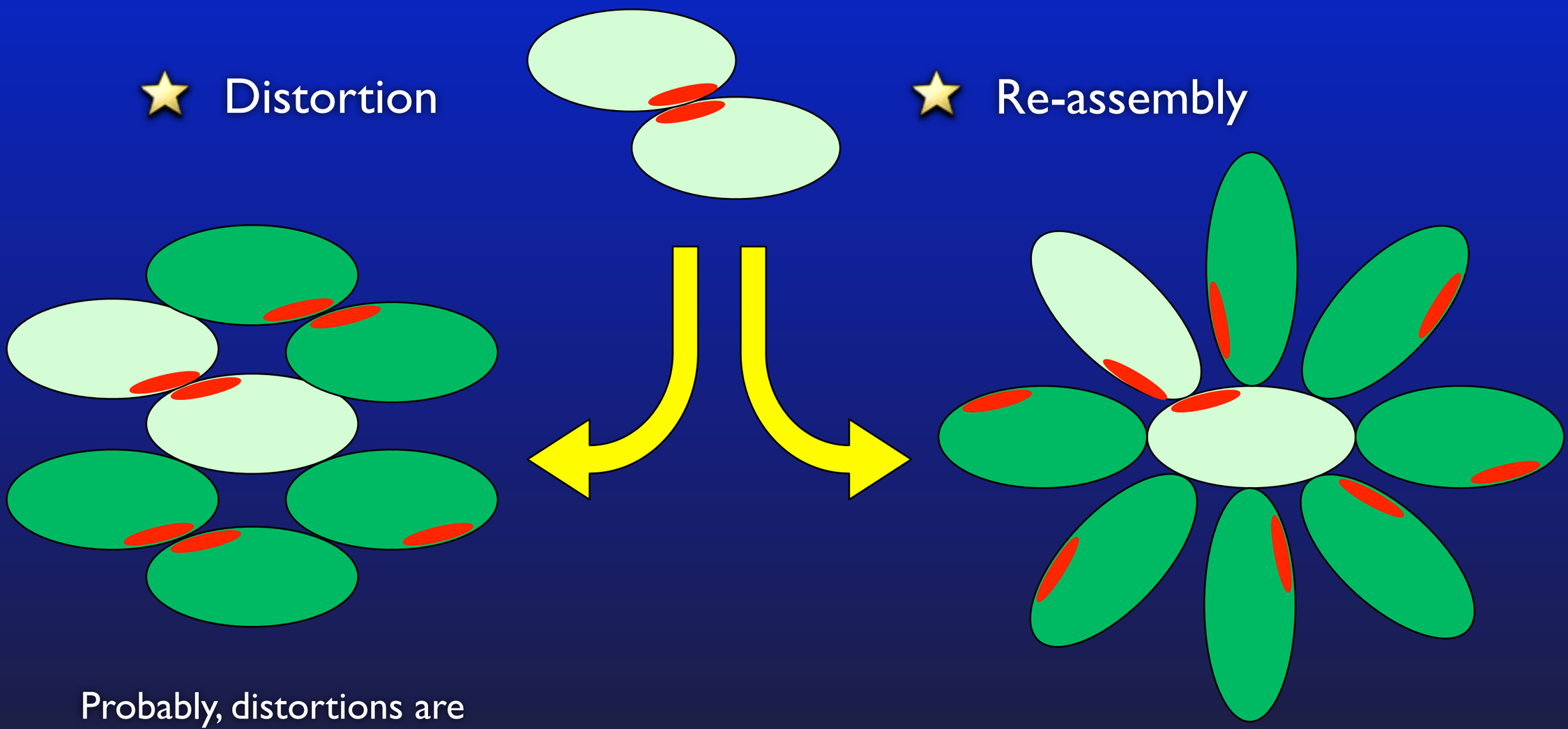
PISA relies on geometry of interactions given by crystal packing. PISA does not dock monomeric units; rather, it uses crystal contacts as "nature's dockings" assuming that they are correct. — Probably correct

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

- Distortion



- Re-assembly

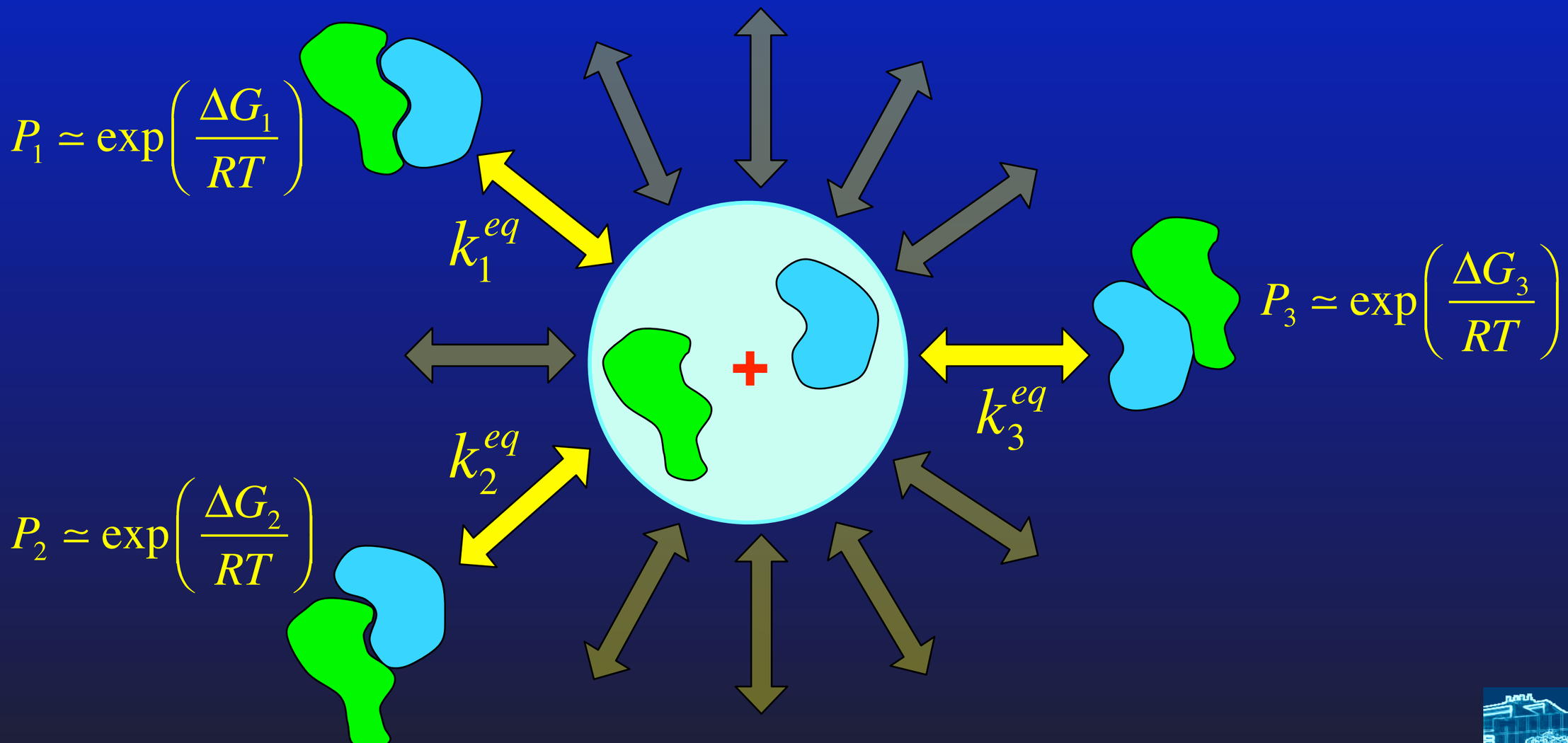
Probably, distortions are always there

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

Research Complex at Harwell

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.




Research Complex at Harwell

Real and superficial crystal contacts

- If a crystal contact remains thermodynamically preferential in solution, the chances are that it represents a biochemically relevant interaction

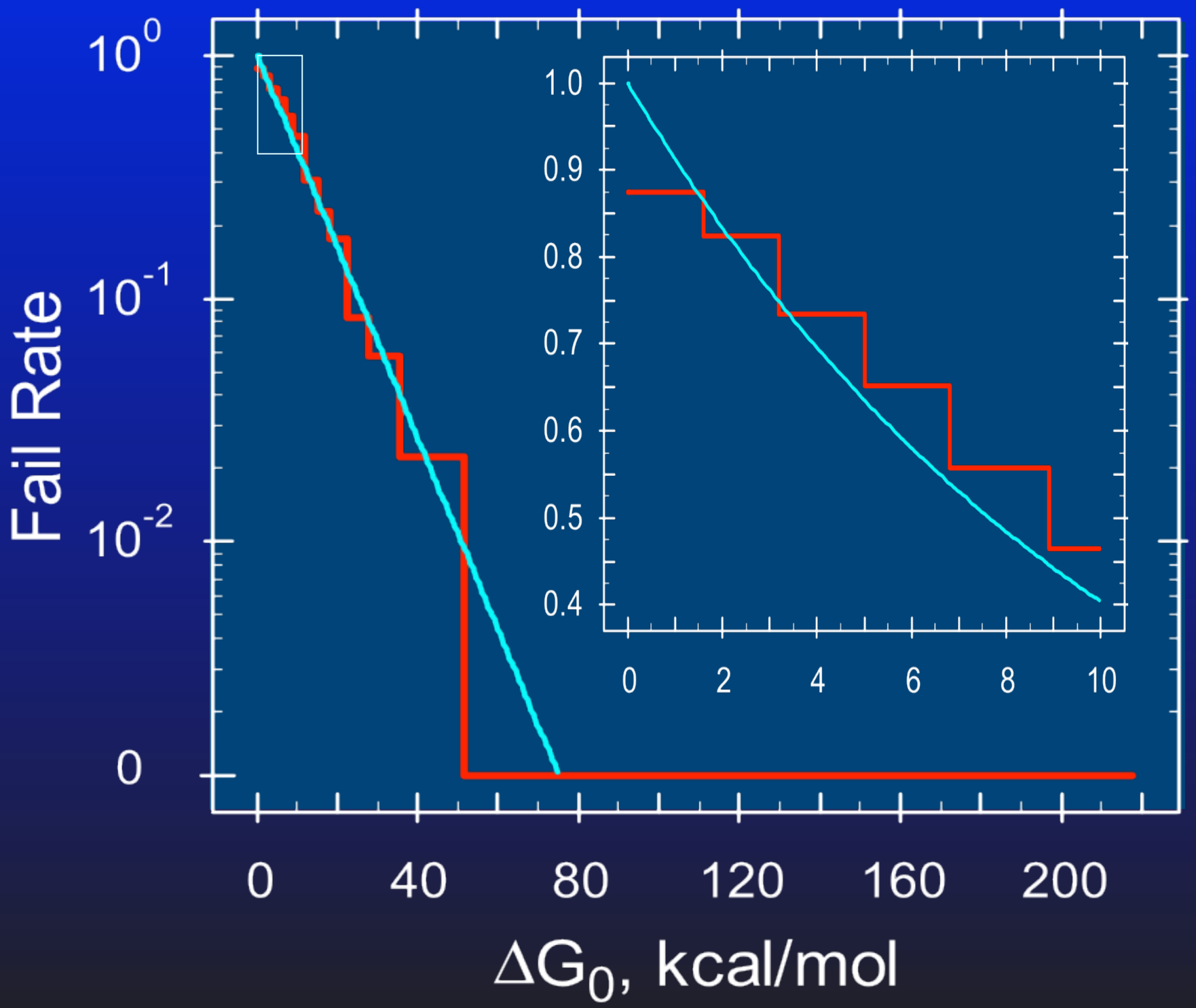
- Experimental data on structure of complexes in solution is rather sparse

- 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

Fail rate of docking



The plot shows the probability of docking not to arrive at crystal interface, as a function of interface free energy.

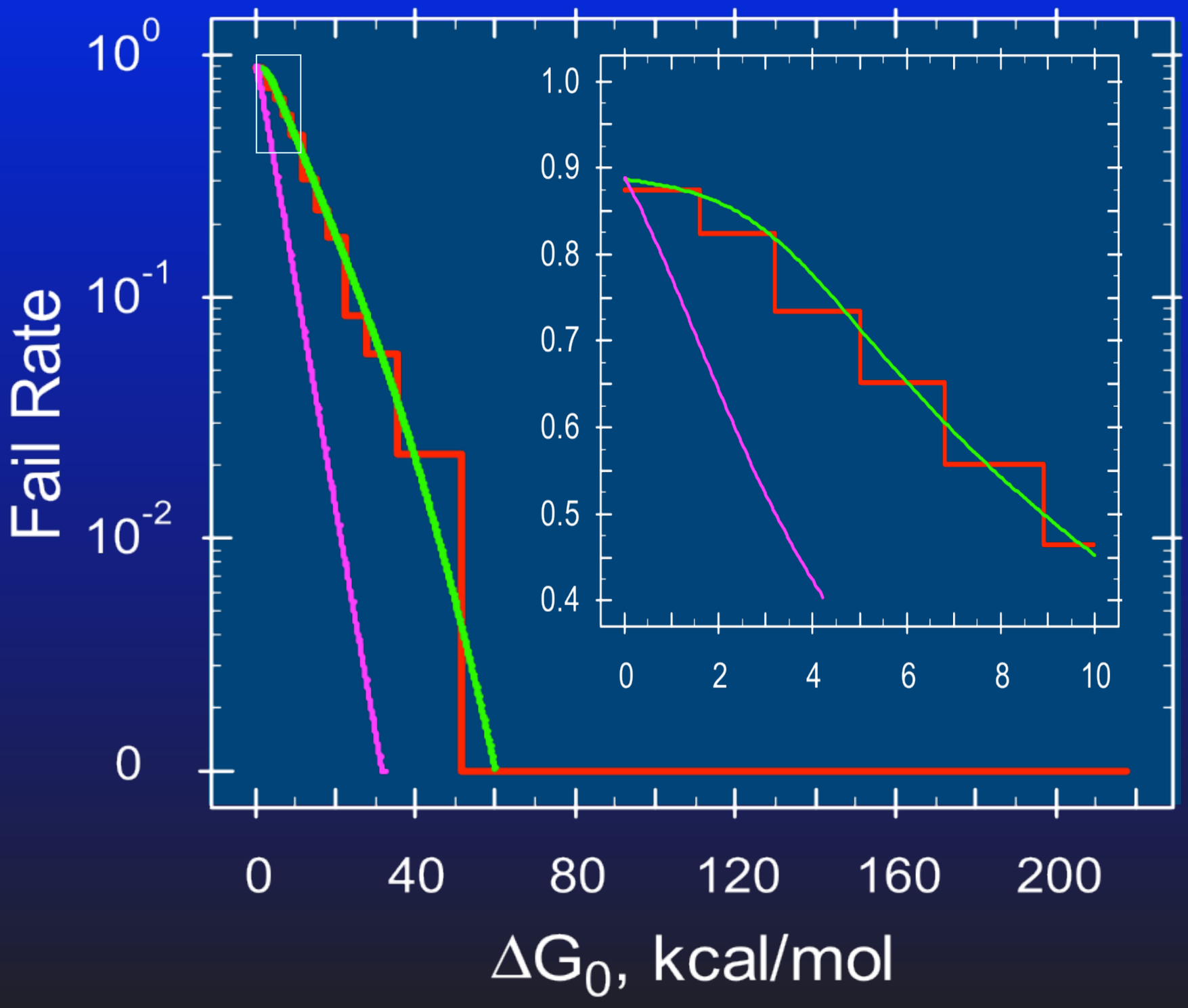
The probabilities are calculated using equipopulated bins.

Overall, 38% of failures.

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

Research Complex at Harwell

Calculation errors and crystal misrepresentation effects



docking results

best fit when both calculation errors and misrepresentation effects are taken into account

pure crystal misrepresentation effect (same as the green curve but with errors set to none)

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

Research Complex at Harwell

So what is the practicality of all this?

- PISA *is not* a substitution for experiments on the identification of protein's oligomeric state
  - both the software and (much less likely) experiment may give wrong results
  - but, in difference of experimental results, calculations do not make a scientific evidence!

- PISA may be used for choosing complex models for molecular replacement
  - already done in CCP4's BALBES and MrBUMP automatic molecular replacement pipelines

- PISA may be used for interpretation of experimental results when evidence is not sufficient for a definite answer
  - which dimer?
  - inconclusive evidence (e.g. oligomeric state highly dependent on concentration/ temperature/ion presence etc.)

- PISA may be used for sanity checks, comparative analysis and flag raising
  - is proposed complex structure compatible with crystal packing?
  - is proposed complex different from close homologs?
  - is there a strong disagreement with biological/biochemical expectations?

Research Complex at Harwell

Acknowledgements

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

Research Complex at Harwell