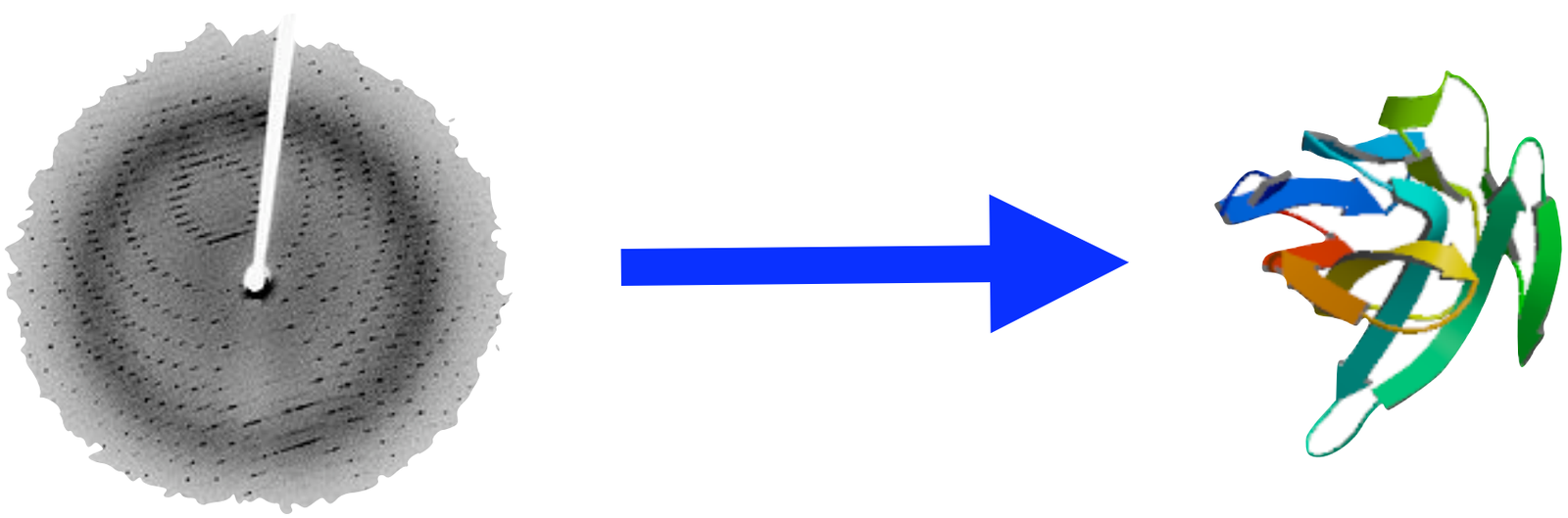
Global Molecular Replacement for Protein Structure Determination

Ian Stokes-Rees
SBGrid - Harvard Medical School



SBGrid and NEBioGrid

Washington U. School of Med.
T. Ellenberger
D. Fremont

U. Washington
T. Gonen

UC Davis
H. Stahlberg

UCSF
JJ Miranda
Y. Cheng

Stanford
A. Brunger
K. Garcia
T. Jardetzky

CalTech
P. Bjorkman
W. Clemons
G. Jensen
D. Rees

WesternU
M. Swairjo

UCSD
T. Nakagawa
H. Viadiu

Rosalind Franklin
D. Harrison

NIH
M. Mayer

U. Maryland
E. Toth

Cornell U.
R. Cerione
B. Crane
S. Ealick
M. Jin
A. Ke

NE-CAT
R. Oswald
C. Parrish
H. Sondermann

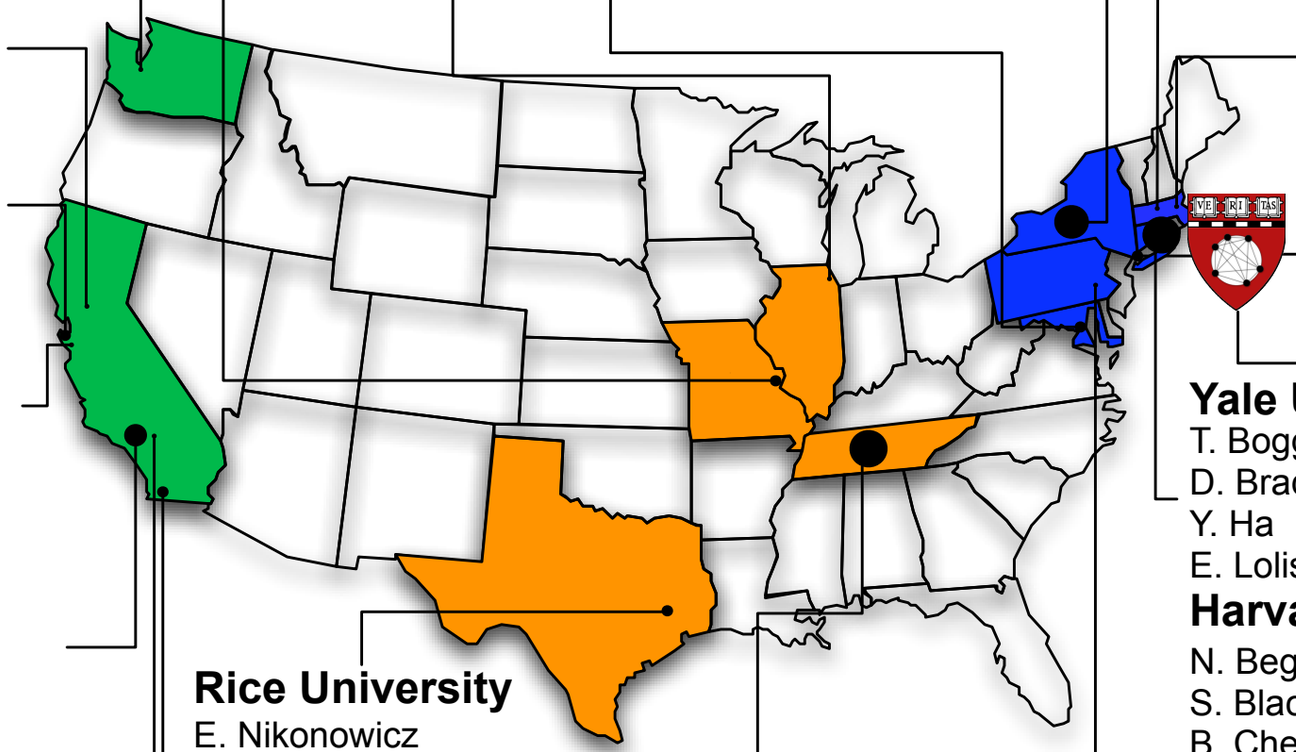
UMass Medical
W. Royer

Brandeis U.
N. Grigorieff

Tufts U.
K. Heldwein

Columbia U.
Q. Fan

Rockefeller U.
R. MacKinnon



Yale U.
T. Boggon
D. Braddock
Y. Ha
E. Lolis
K. Reinisch
J. Schlessinger
F. Sigworth
F. Zhou

Harvard and Affiliates
N. Beglova
S. Blacklow
B. Chen
J. Chou
J. Clardy
M. Eck
B. Furie
R. Gaudet
M. Grant
S.C. Harrison
J. Hogle
D. Jeruzalmi
D. Kahne
T. Kirchhausen
A. Leschziner
K. Miller
A. Rao
T. Rapoport
M. Samso
P. Sliz
T. Springer
G. Verdine
G. Wagner
L. Walensky
S.Walker
T.Walz
J. Wang
S. Wong

Rice University
E. Nikonowicz
Y. Shamoo
Y.J. Tao

Vanderbilt
Center for Structural Biology
W. Chazin
B. Eichman
M. Egli
B. Lacy
C. Sanders
B. Spiller
M. Stone
M. Waterman

Thomas Jefferson
J. Williams

Not Pictured:
University of Toronto: L. Howell, E. Pai, F. Sicheri; NHRI (Taiwan): G. Liou; Trinity College, Dublin: Amir Khan

Primary thesis:

Molecular replacement, used to solve over 60% of known structures, can benefit from novel computationally intensive techniques to identify search models, including those with low sequence identity or a lack of previous association with the unknown structure.

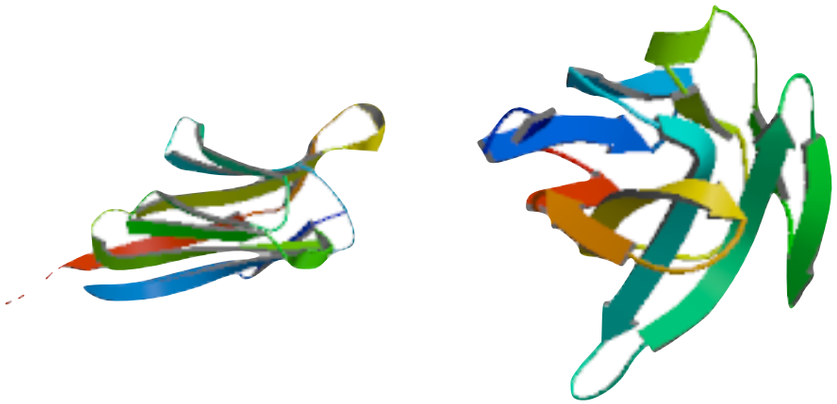
Expected benefits:

identify search models which would otherwise be missed;
faster bootstrapping of MR search model selection;
broaden range of structures amenable to MR, avoiding more costly phasing techniques;
allow greater parameter tuning of MR stage;

Transferable infrastructure:

framework developed to support 20,000 CPU-hour computation with 10 GB of data, 100,000 invocations of a scientific application, and the consequent results filtering, aggregation, and analysis can be re-used for other applications.

Traditional Molecular Replacement



one, or maybe more carefully selected search model

0.1 CPUh

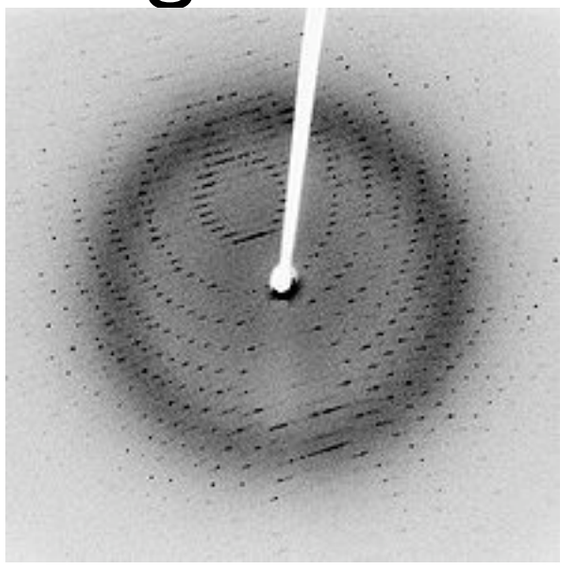
10-20 Solutions

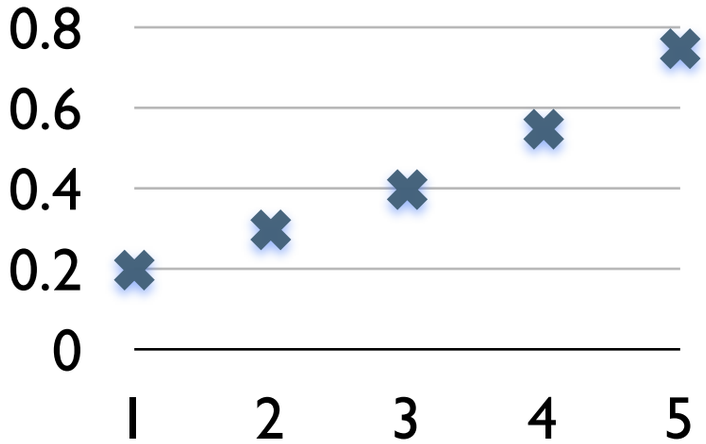
Internal Validation

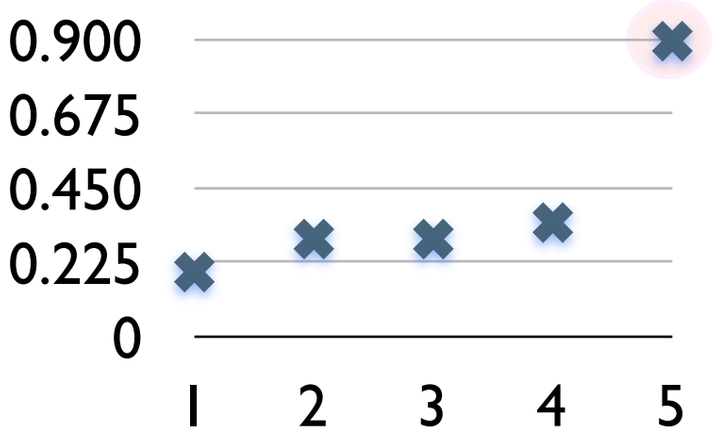
+ Refinement Validation

Hit

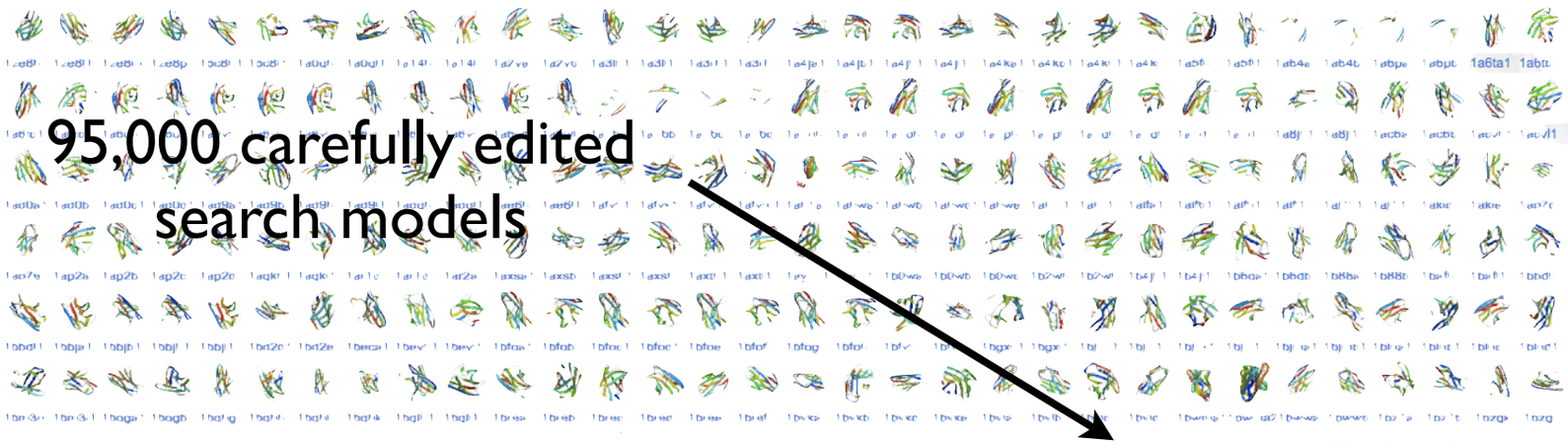
Target Data

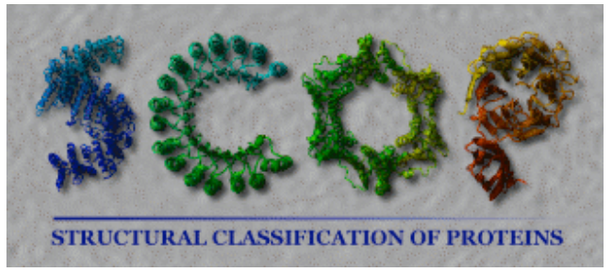






Global Molecular Replacement





9500 CPUh

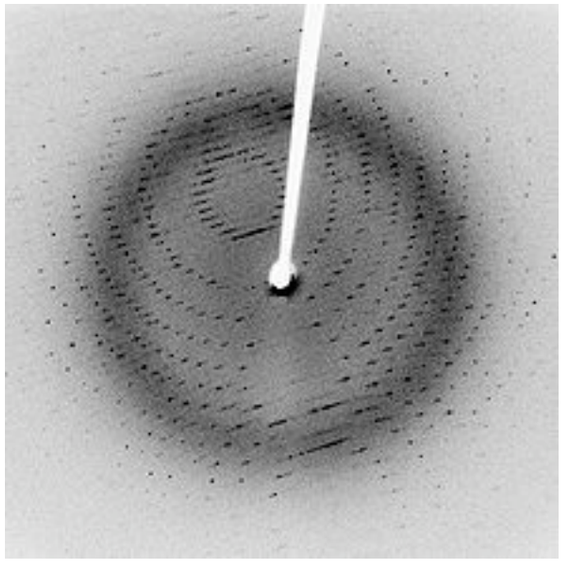
~50K
Solutions

External Validation

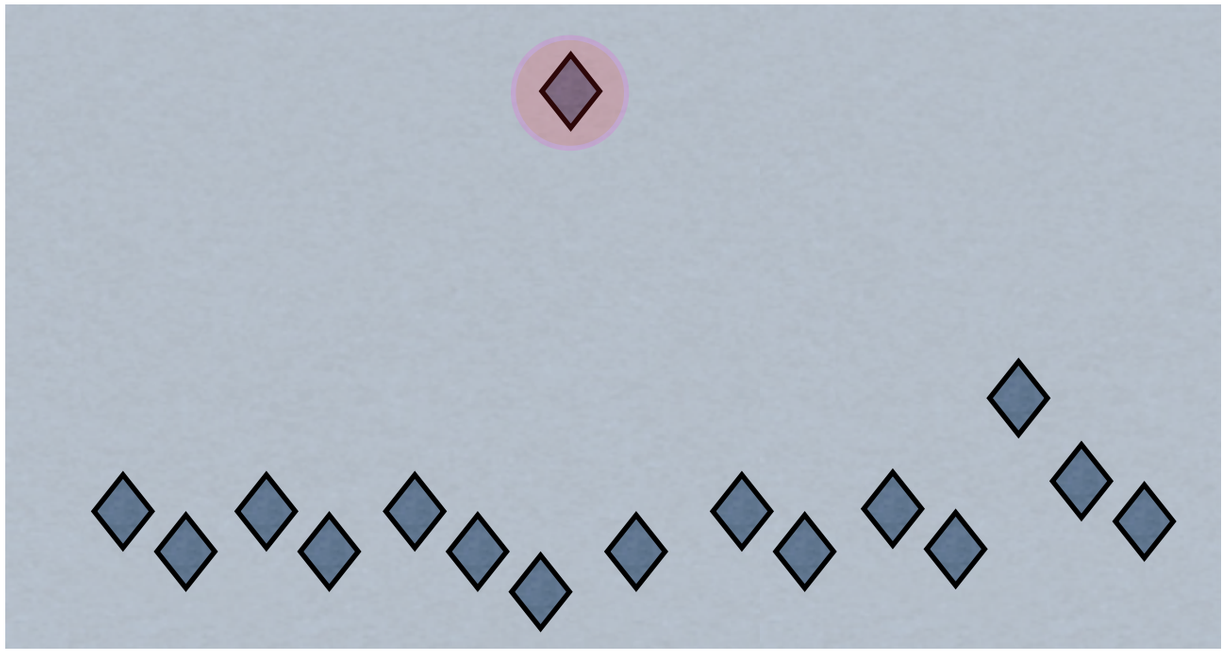
+ Refinement Validation

Hit

Target Data



Score



Individual Models

Small Physical Differences, Big Impact On Results

TARGET

MODEL A

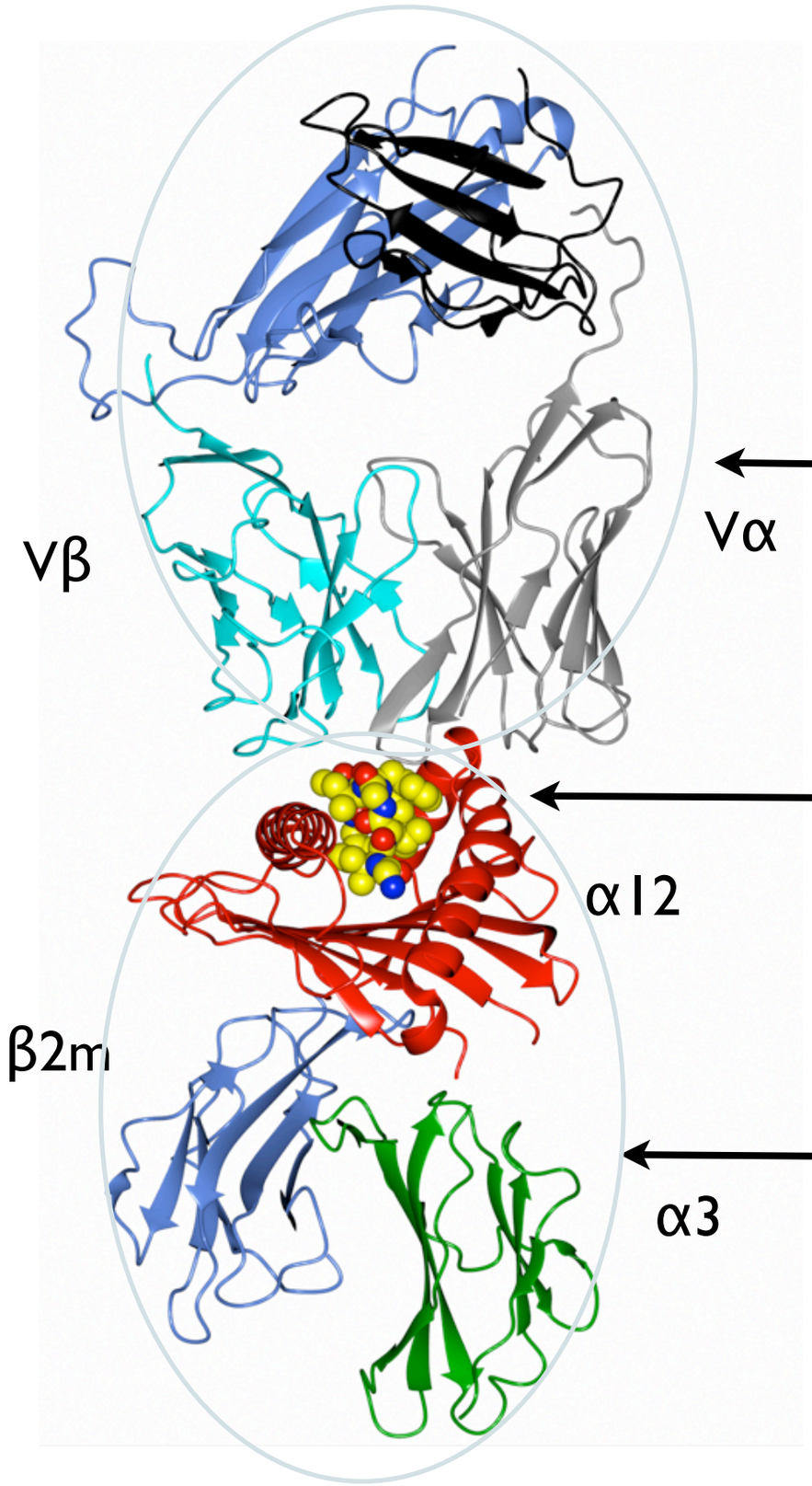
MODEL B

MODEL C

differences in loops, and shifts of the secondary structure elements degrade results

- Would global search work? What are the boundaries of global search method?
- What is the best scoring function?
- Is MR Score related to RMSD/Sequence Identity of target molecule
- Real Life example

Target I: 2VLJ



The Structural Dynamics and Energetics of an Immunodominant T Cell Receptor Are Programmed by Its Vβ Domain

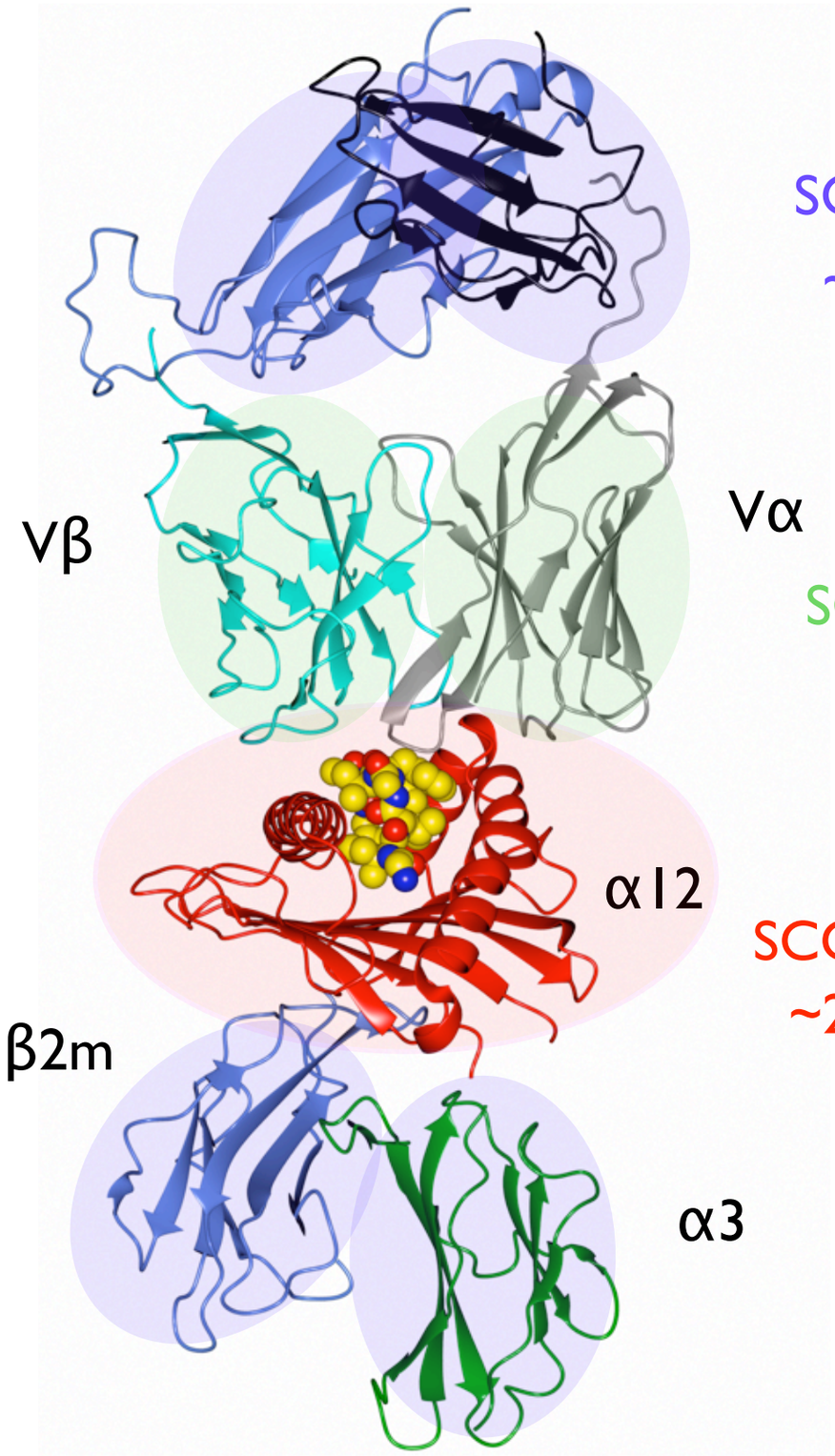
Jeffrey Ishizuka,[1,4] Guillaume B.E. Stewart-Jones,[1,2,4] Anton van der Merwe,[3] John I. Bell,[1] Andrew J. McMichael,[1,*] and E. Yvonne Jones[2,*]
[1]MRC Human Immunology Unit, Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK
[2]Division of Structural Biology, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, University of Oxford, Oxford OX3 7BN, UK
[3]Sir William Dunn School of Pathology, Oxford OX1 3RE, UK
[4]These authors contributed equally to this work.
*Correspondence: andrew.mcmichael@ndm.ox.ac.uk (A.J.M.), yvonne.jones@strubi.ox.ac.uk (E.Y.J.)
DOI 10.1016/j.immuni.2007.12.018

T cell receptor
(4 Immunoglobulin Domains)

influenza-virus matrix peptide

presentation of the peptide
by the major Histocompatibility
Complex (MHC) molecule
(2 Immunoglobulin Domains + peptide
binding domain)

1 x MHC domain + 6 x Ig domain



SCOP **b.1.1.2** - antibody constant domain-like, 2535 domains
~12.5% by MW

SCOP **b.1.1.1** - antibody variable domain-like, 2001 domains

SCOP **d.19.1.1** - MHC antigen recognition domain, 568 domains
~22% by MW

Molecular Weight of the complex: 94.495 kDa

Selection Criteria:

- a multidomain protein
- wide range of models

Bjorkman et al. Structure of the human class I histocompatibility antigen, HLA-A2. Nature (1987) vol. 329 (6139) pp. 506-12

Garboczi et al. Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. Nature (1996) vol. 384 (6605) pp. 134-41

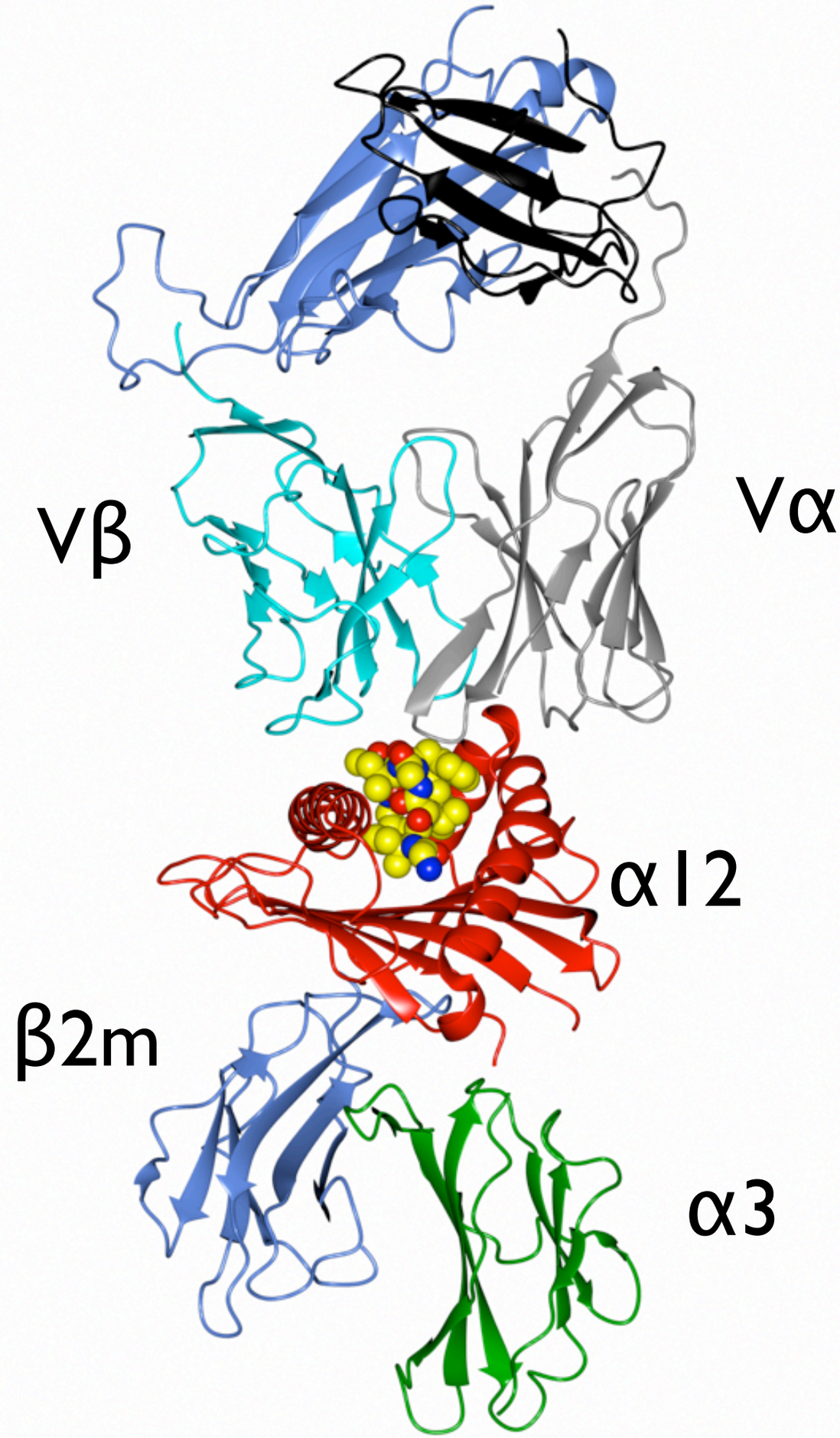
Phaser - round I

Search with 95K SCOP models

5 min timeout

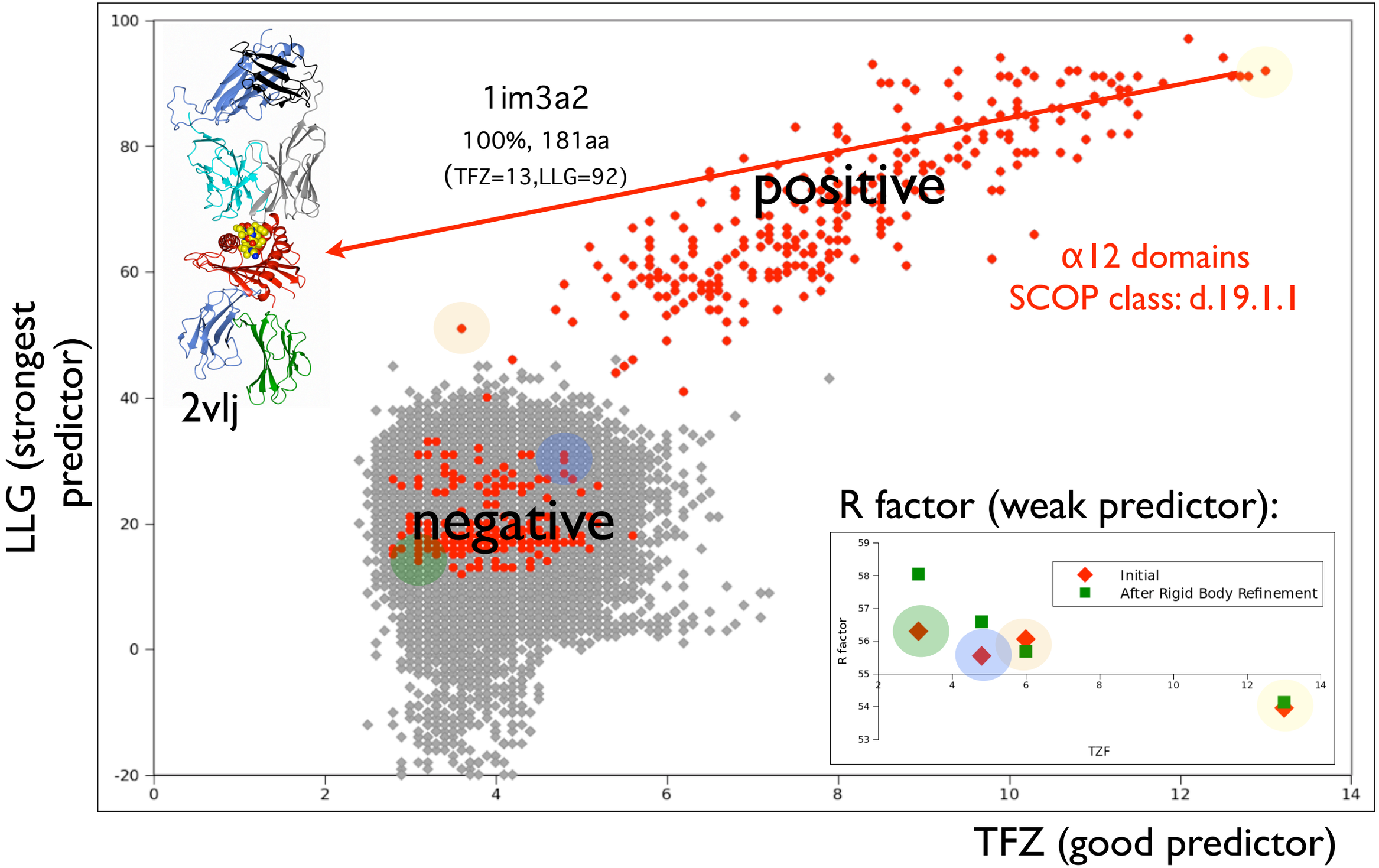
2000 CPU cores on OSG

24h



2D representation of MR results

Top Scoring Solution: 1im3a2



Phaser - round II

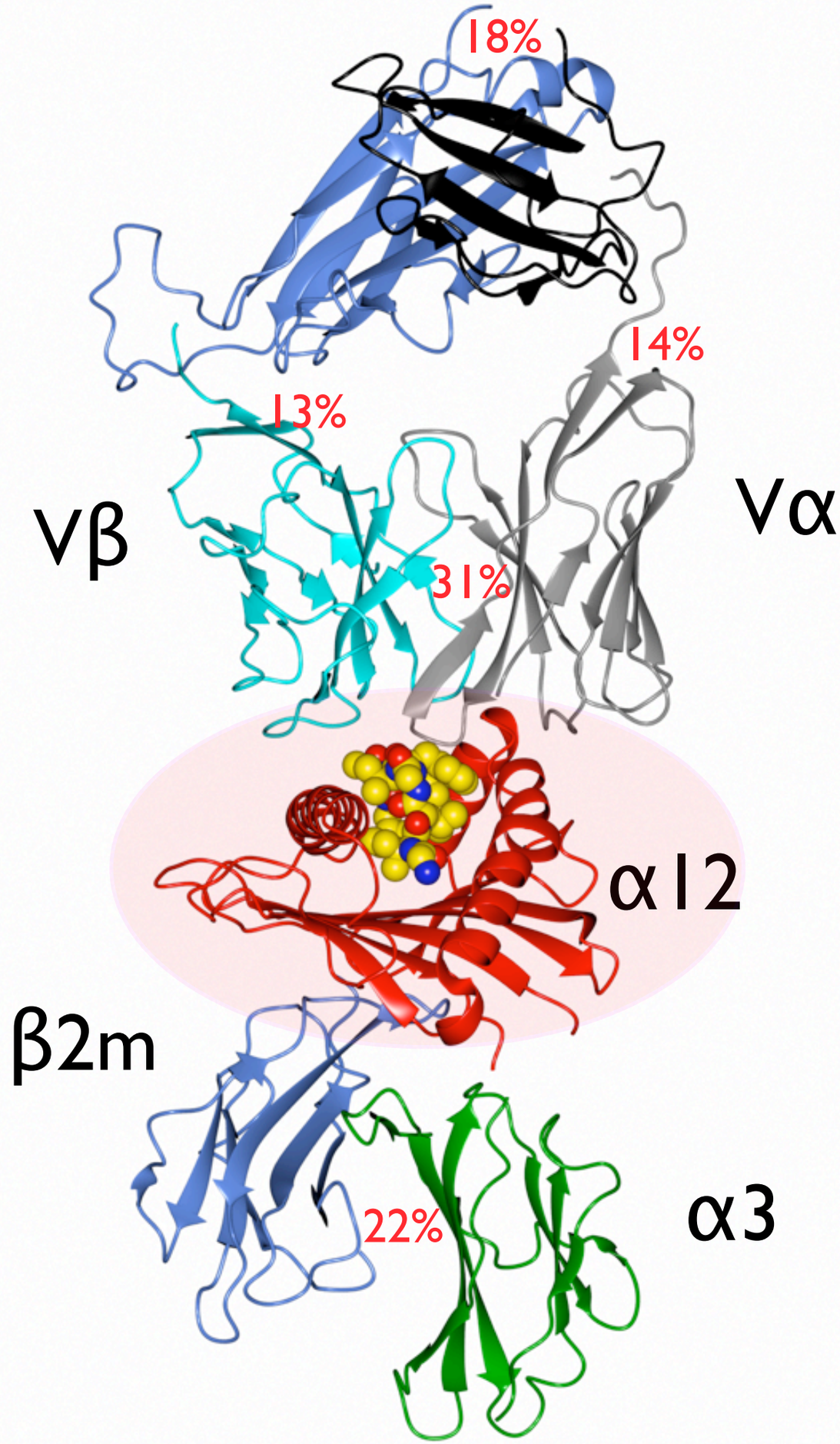
Fix the α12 domain

Repeat MR search with the 95K SCOP dataset

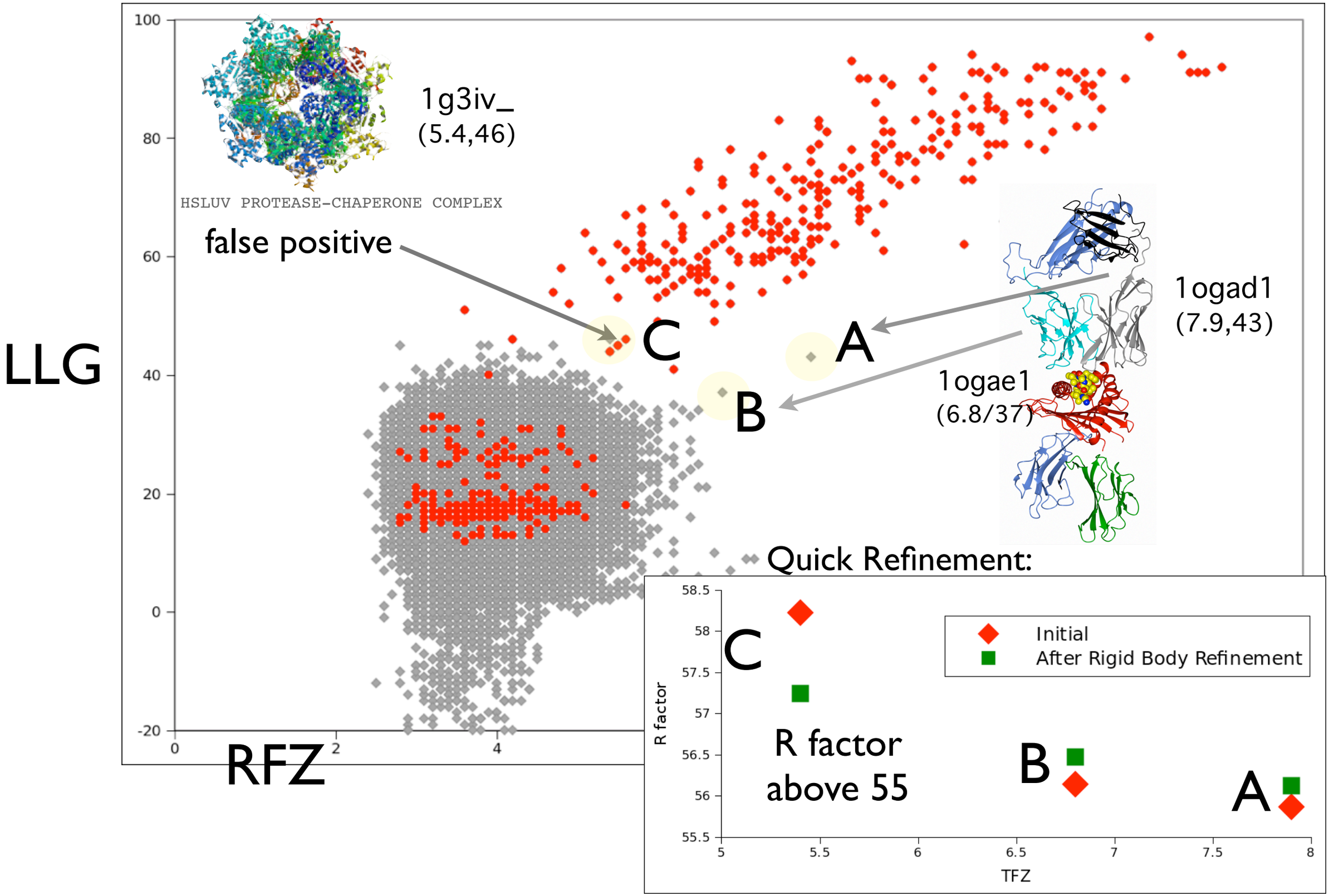
5 min timeout

2000 CPU cores on OSG

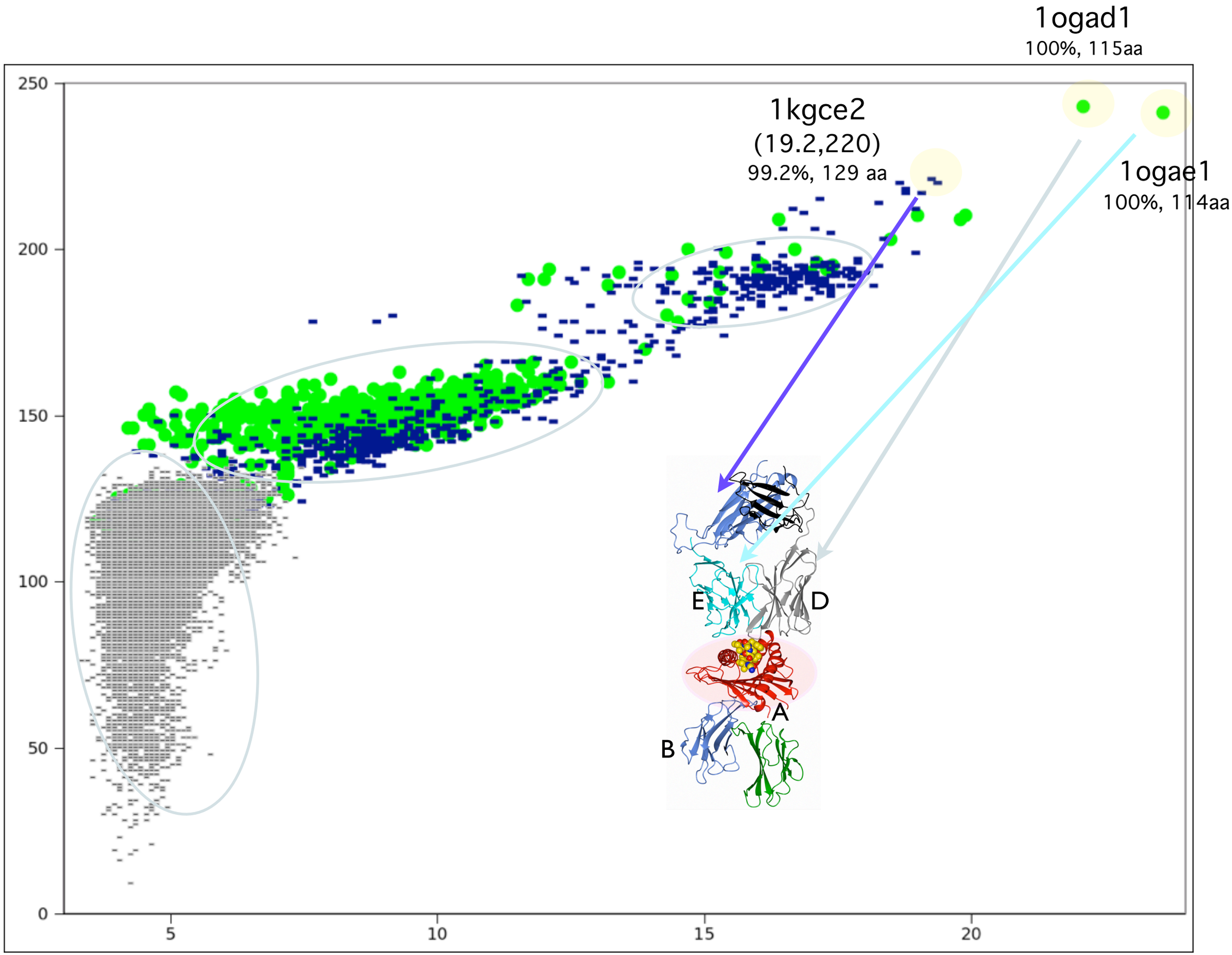
24h



Two solutions for Ig domains from TCR

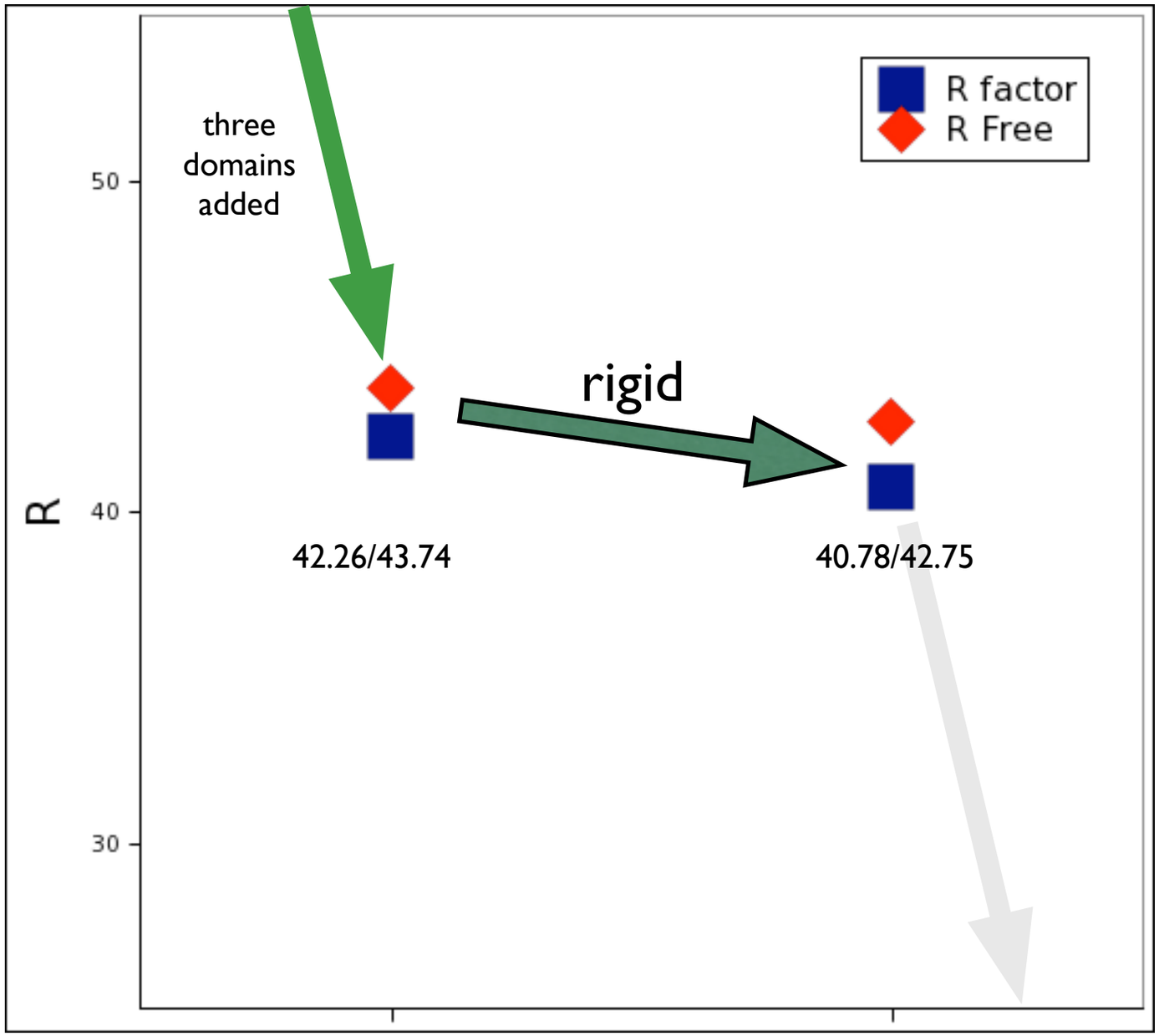


Domain A12 placed, searching for next domain

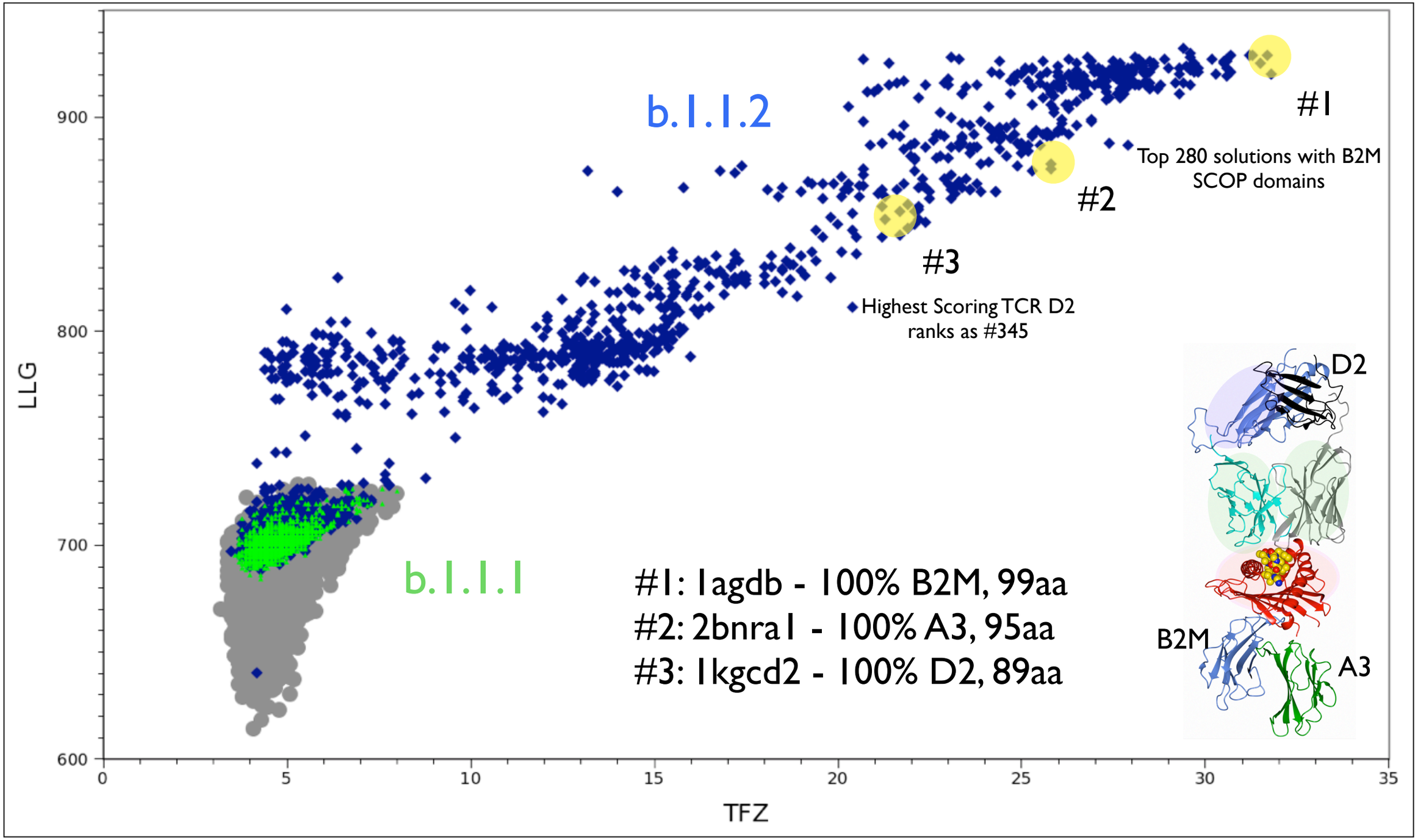


Refinement

3 cycles of
Rigid Body

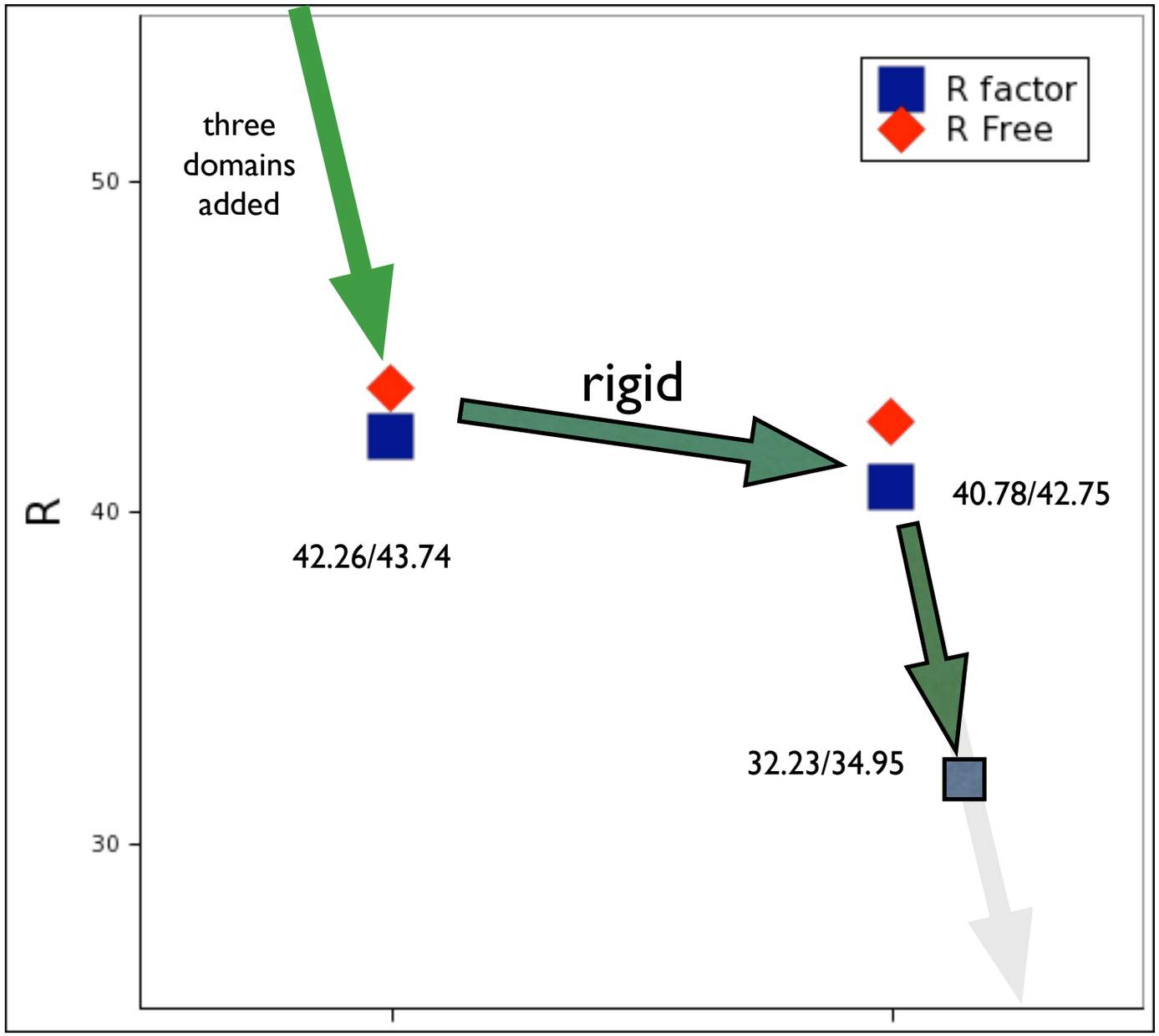


4 domains placed, searching for 3 remaining domains



Refinement

3 cycles of
Rigid Body



Solved!

- Would global search work? What are the boundaries of global search method?
- What is the best MR scoring function?
- Is MR Score related to RMSD/Sequence Identity of target molecule
- Real Life example

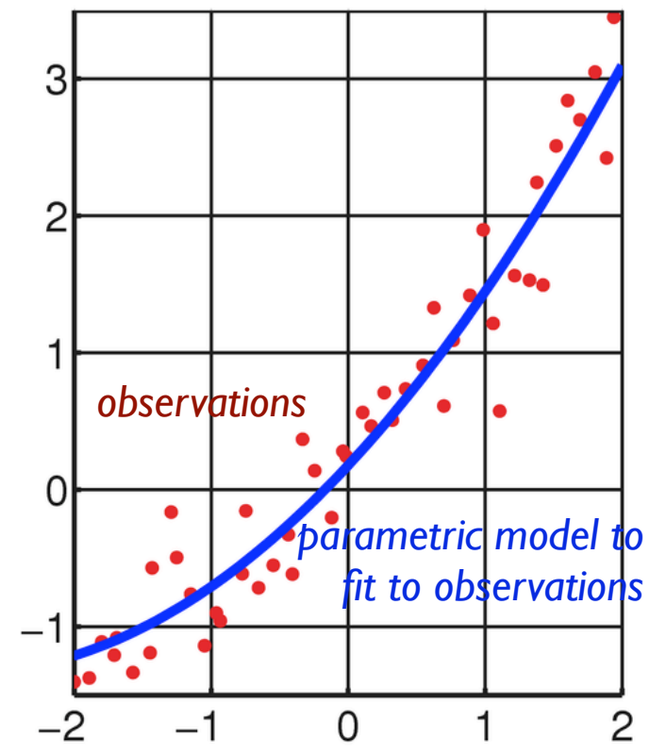
Common approach to molecular replacement: Least Squares match

difference between scalar amplitudes

$$SS = \sum_{\mathbf{h}} \frac{1}{\sigma^2} \left(\left| \mathbf{F}_o(\mathbf{h}) \right| - \left| \mathbf{F}_c(\mathbf{h}) \right| \right)^2$$

Least Squares: commonly used for molecular replacement model quality measure

select model with minimum error between *observed* amplitudes $|\mathbf{F}_O|$ and *calculated* amplitudes $|\mathbf{F}_C|$



real-space equivalent

magnitude of vector difference

$$SS = \sum_{\mathbf{h}} \frac{1}{\sigma^2} \left\| \mathbf{F}_o(\mathbf{h}) \right| \exp(i\alpha_c) - \mathbf{F}_c(\mathbf{h}) \right|^2$$

Problem: Implicitly biased towards model to select **h** (structure parameters) based on model phasing

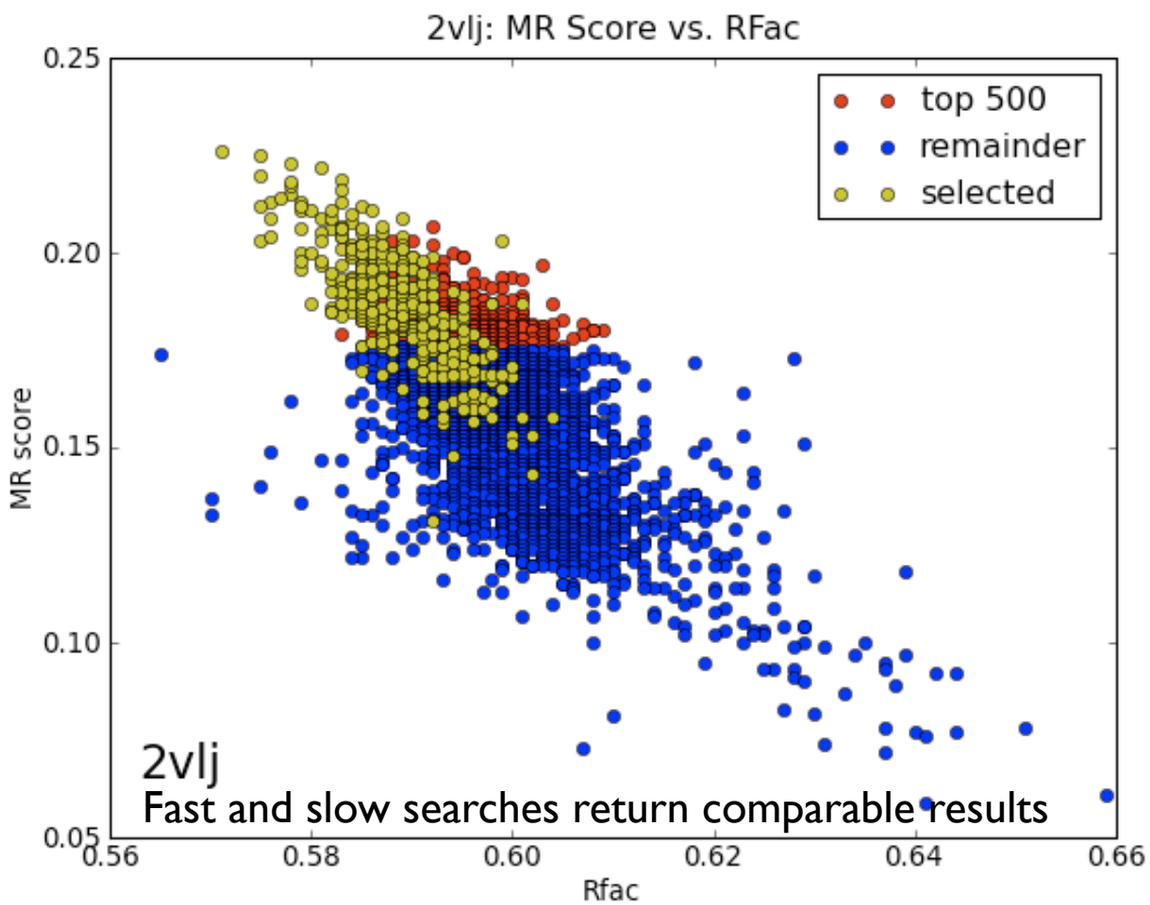
Iterative Convergence: Rotate search model (3D RF) then translate (3D TF) to find best (lowest) least squares fit

Solution Quality: Typically measured by heuristic score, or residual factor (measure of agreement between solution and experimental observations)

Phaser performs better (although more CPU demanding)

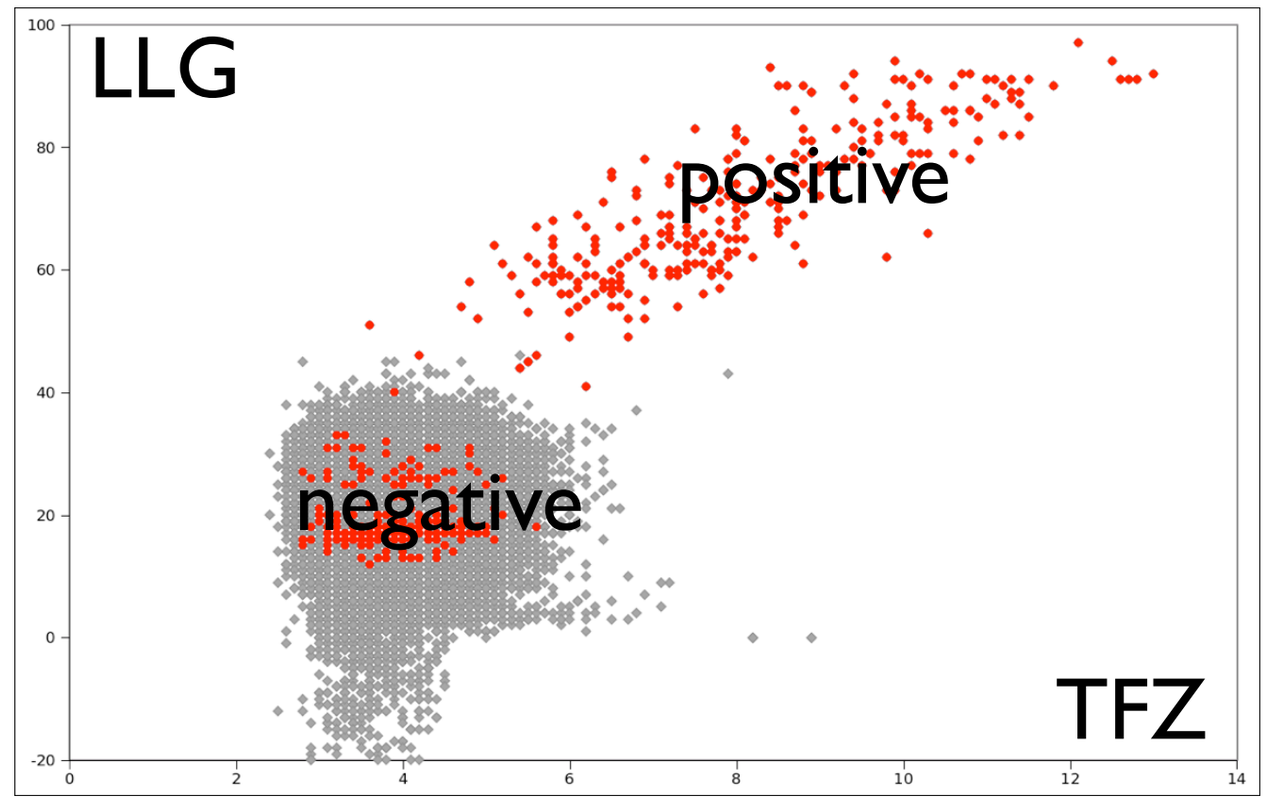
Molrep

(Crowther rotation + FFT in reciprocal space)



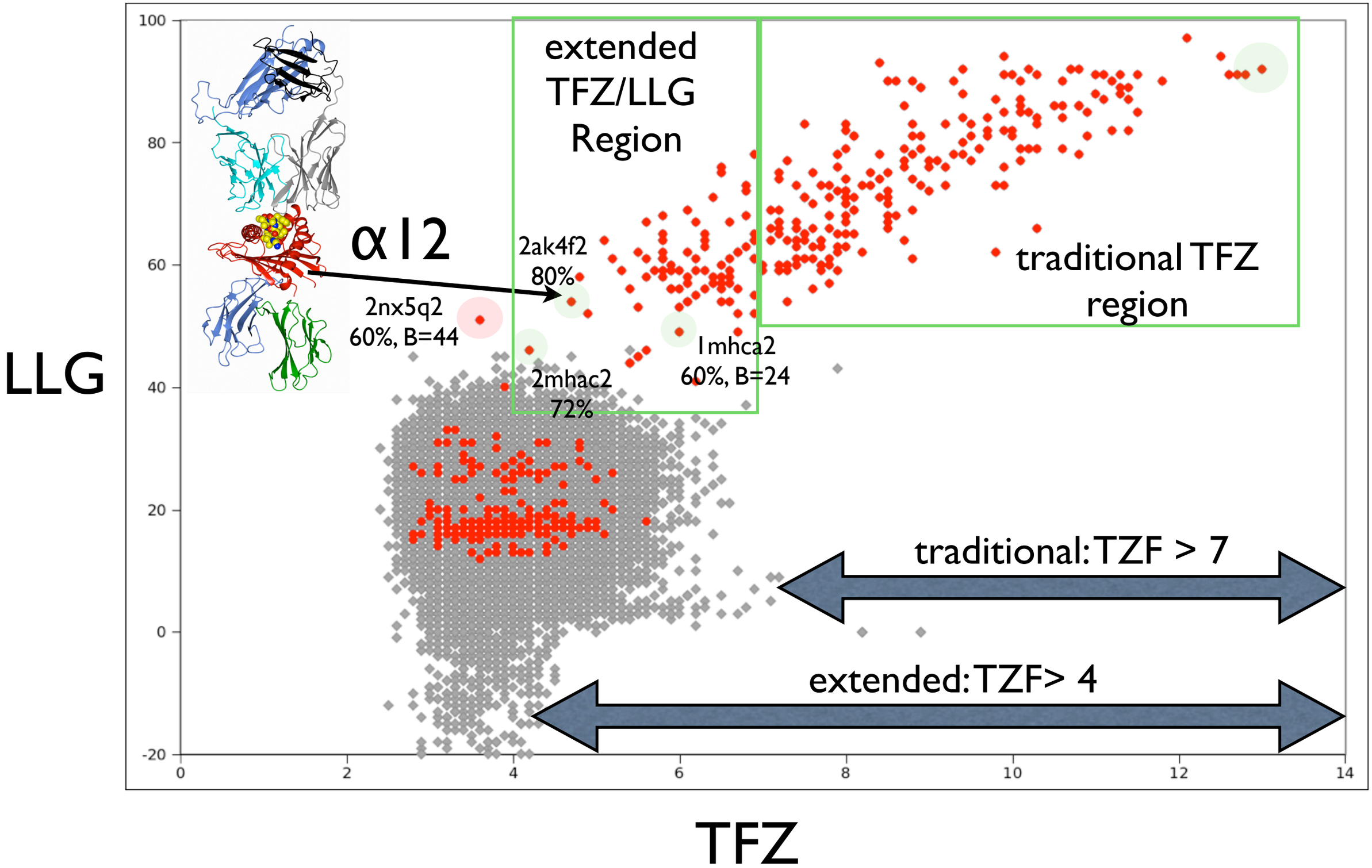
Phaser

(maximum likelihood)

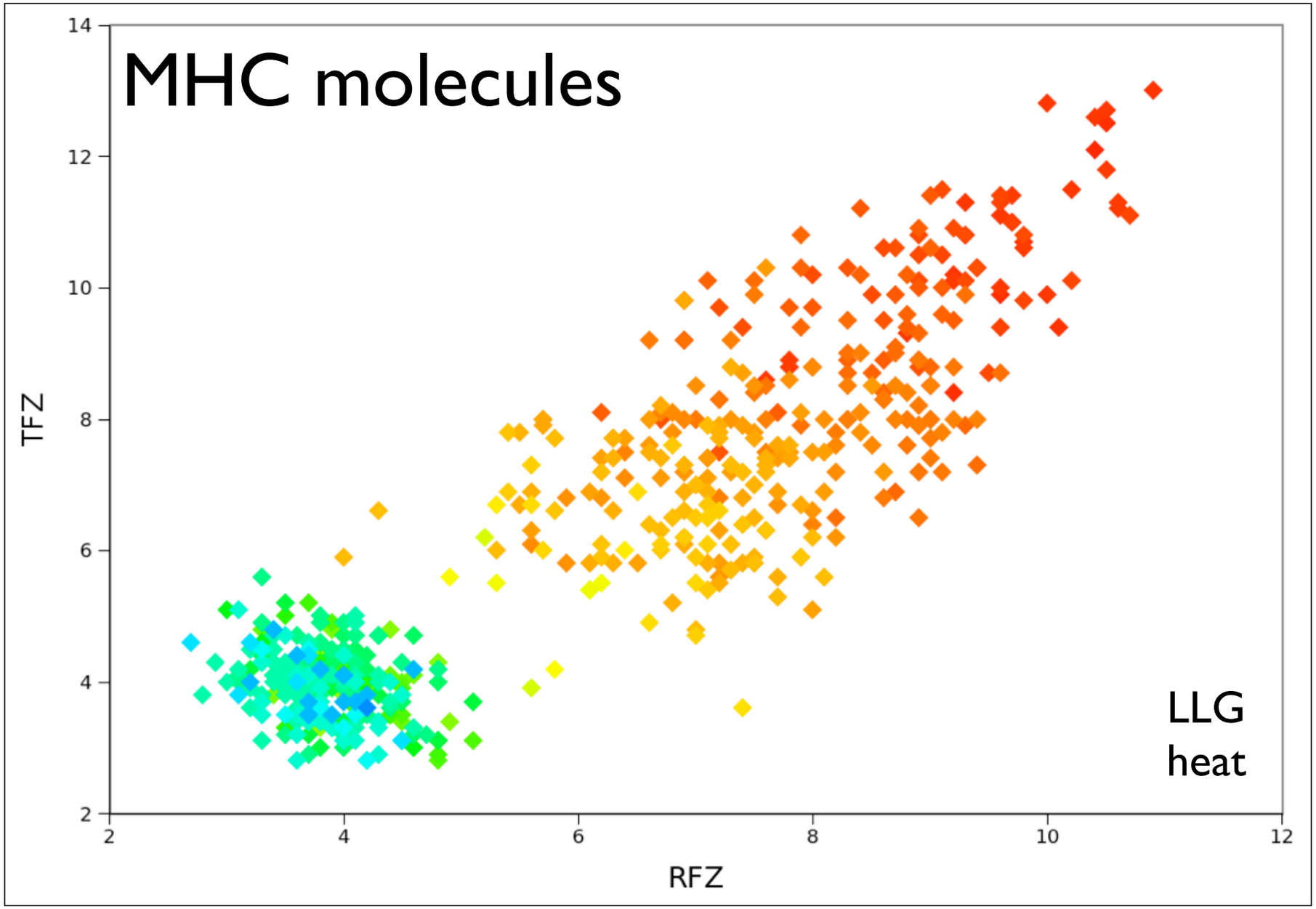


Clear separation between two populations!

Extended range of correct solutions!

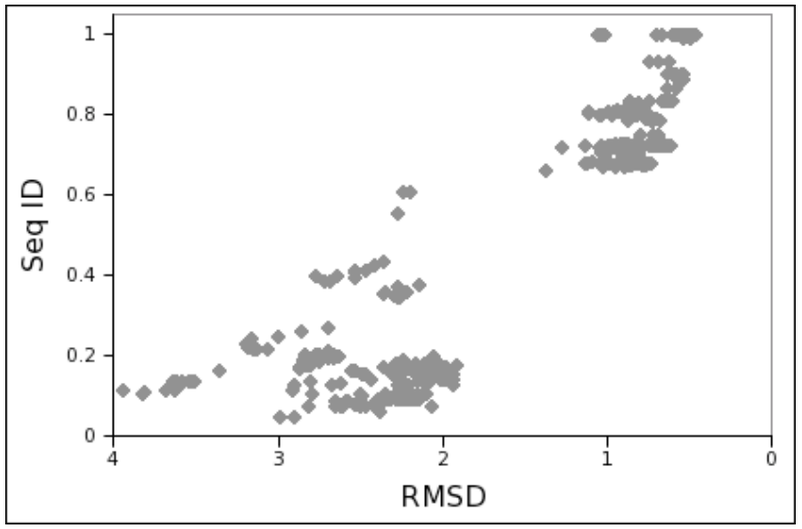


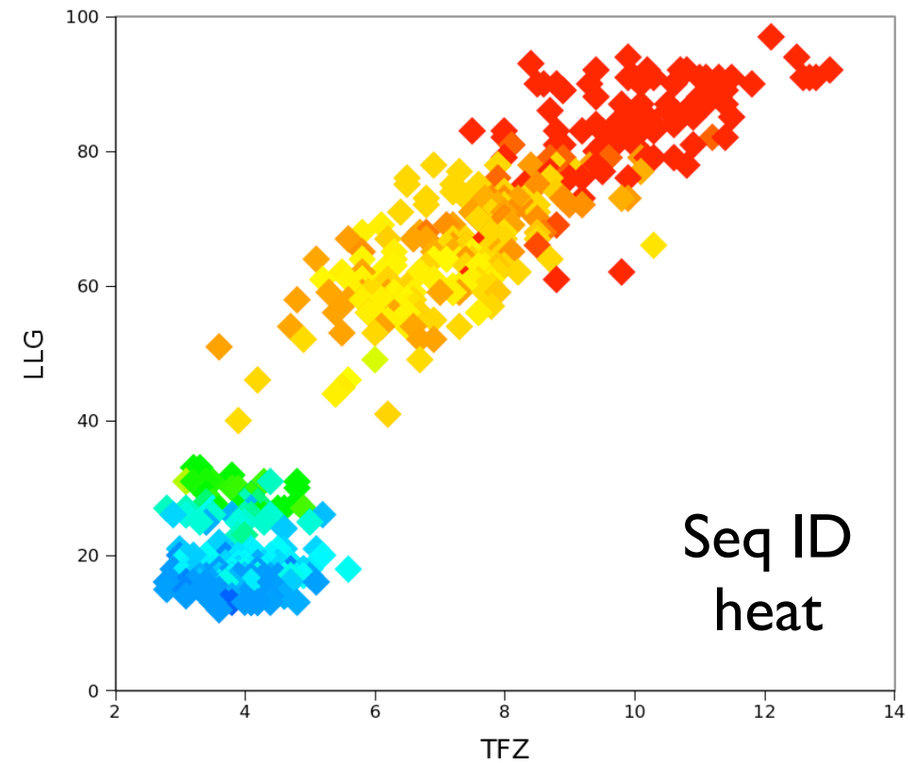
Rotation Function Score

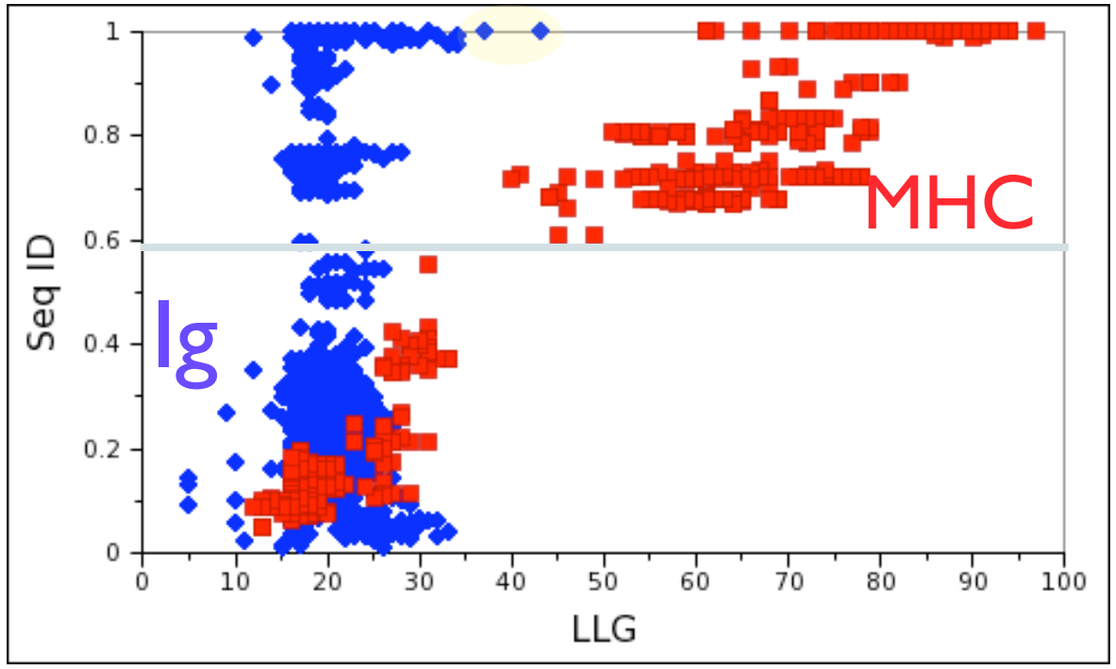


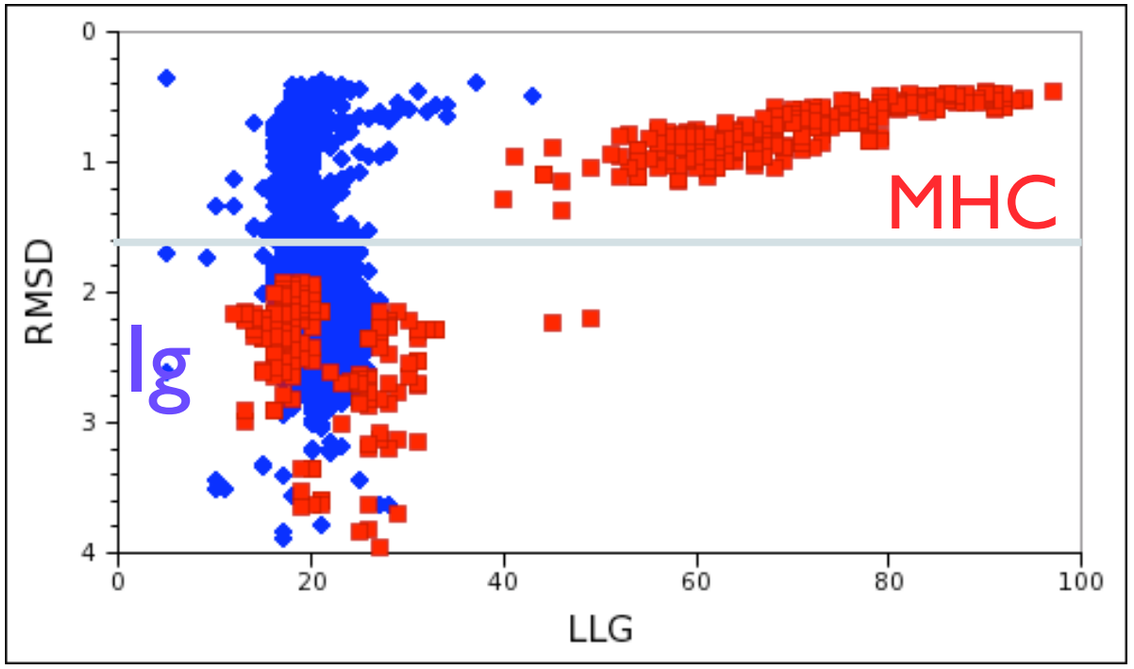
- Would global search work? What are the boundaries of global search method?
- What is the best MR scoring function?
- Is MR Score related to RMSD/Sequence Identity of target molecule
- Real Life example

Search for the first molecule:





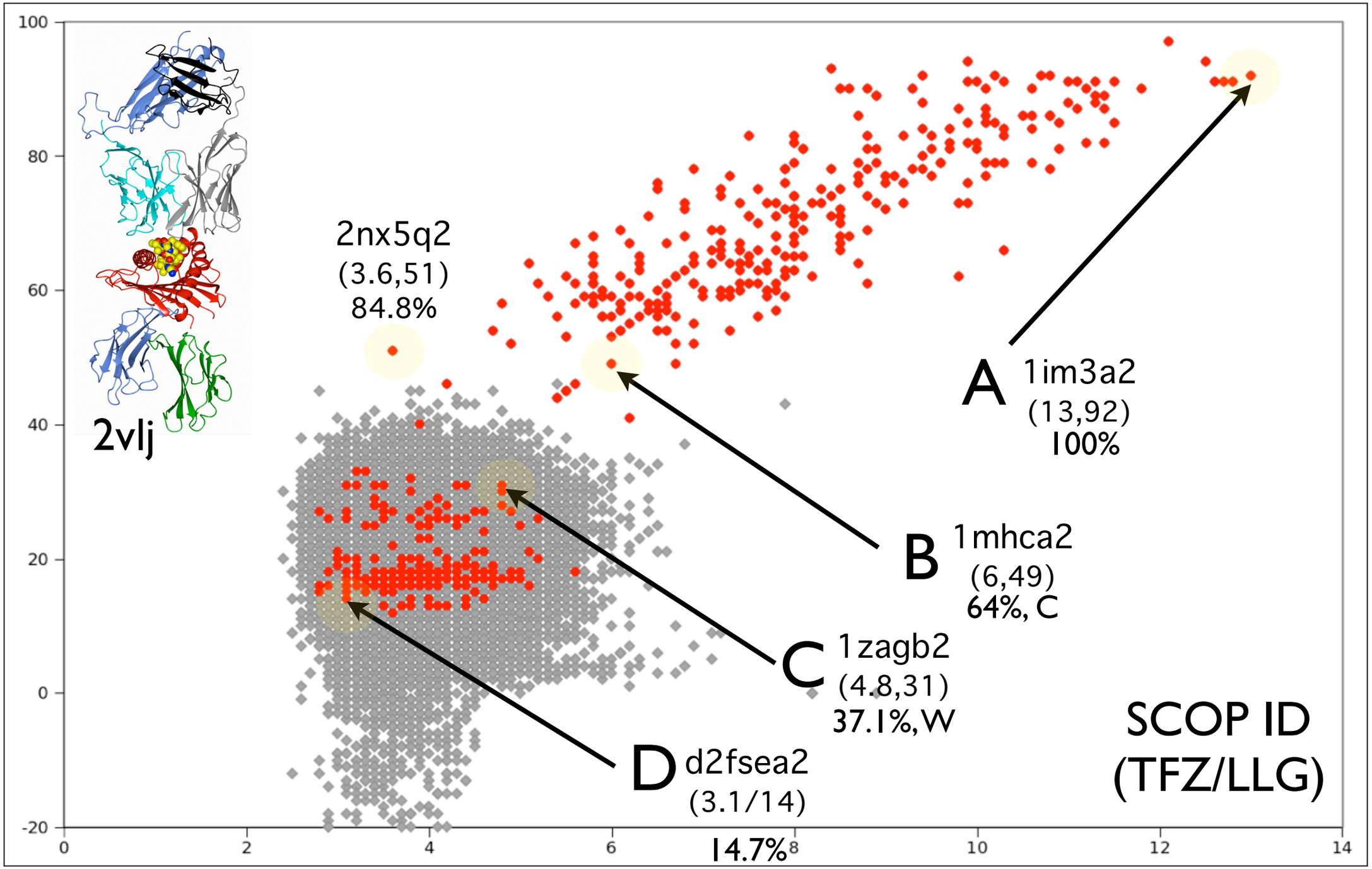




With small fraction of target (~22%) sequence identity > 60% (rmsd < 1.5) required

For Ig domains (~12%) even 100% is barely sufficient

Differences between A12 solutions



Structure Superimposition

TARGET

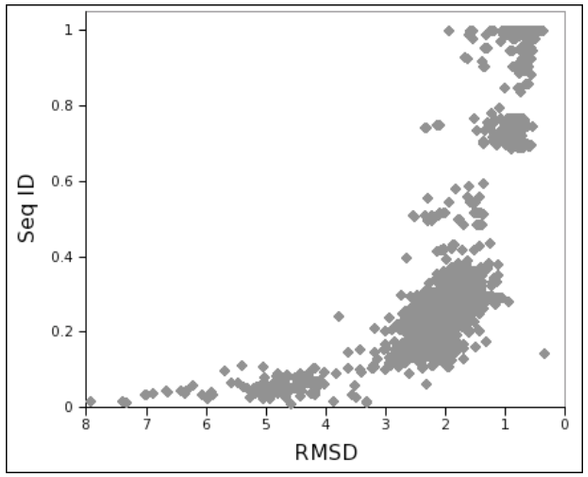
MODEL A

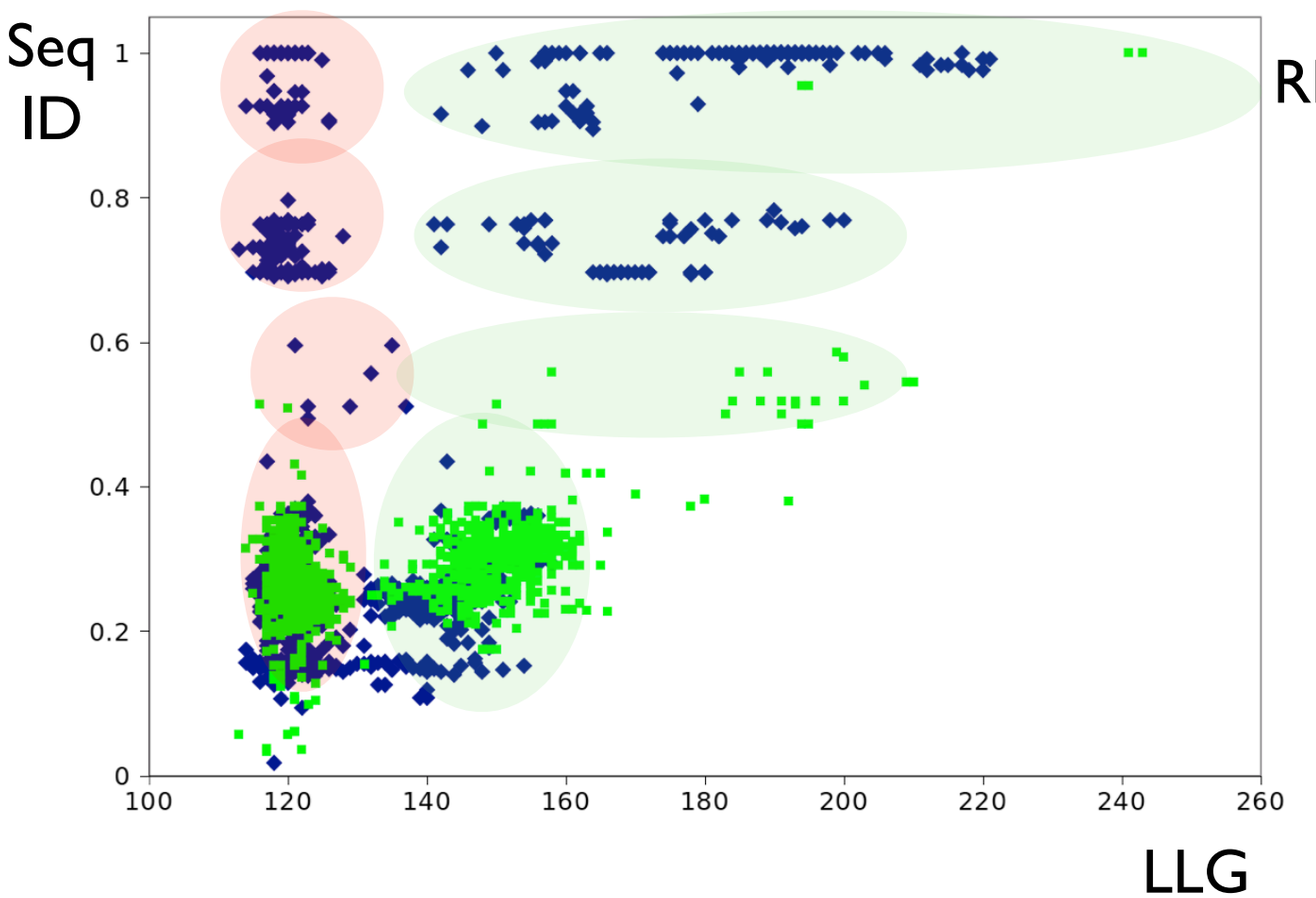
MODEL B

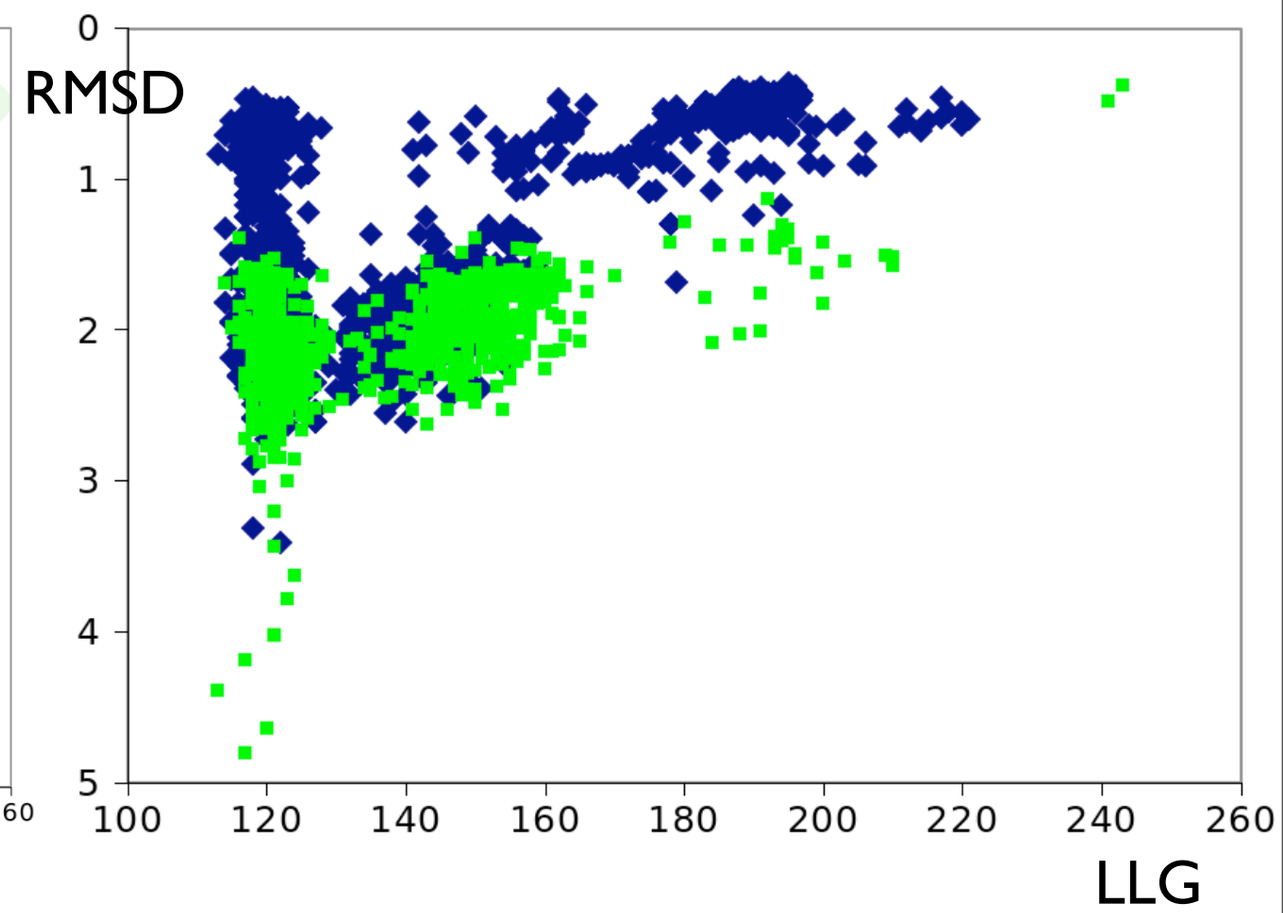
MODEL C

differences in loops, and shifts of the secondary structure elements degrade results

Ig Domains
variable and constant







- Would global search work? What are the boundaries of global search method?
- What is the best MR scoring function?
- Is MR Score related to RMSD/Sequence Identity of target molecule
- Real Life example

2VZF was solved by MAD - MR failed

Phasing and Refinement—Initial phases of the EmoB crystal structure were determined by the MAD phasing method (30) using the software SOLVE (31) **after prior approaches with the molecular replacement method**.

72% Solvent

ABLE 1

rystallographic data for the apo-form and FMN·FMN and FMN·NADH complexes of EmoB

peak; I, inflection; R, remote; r.m.s.d., root mean square deviation.

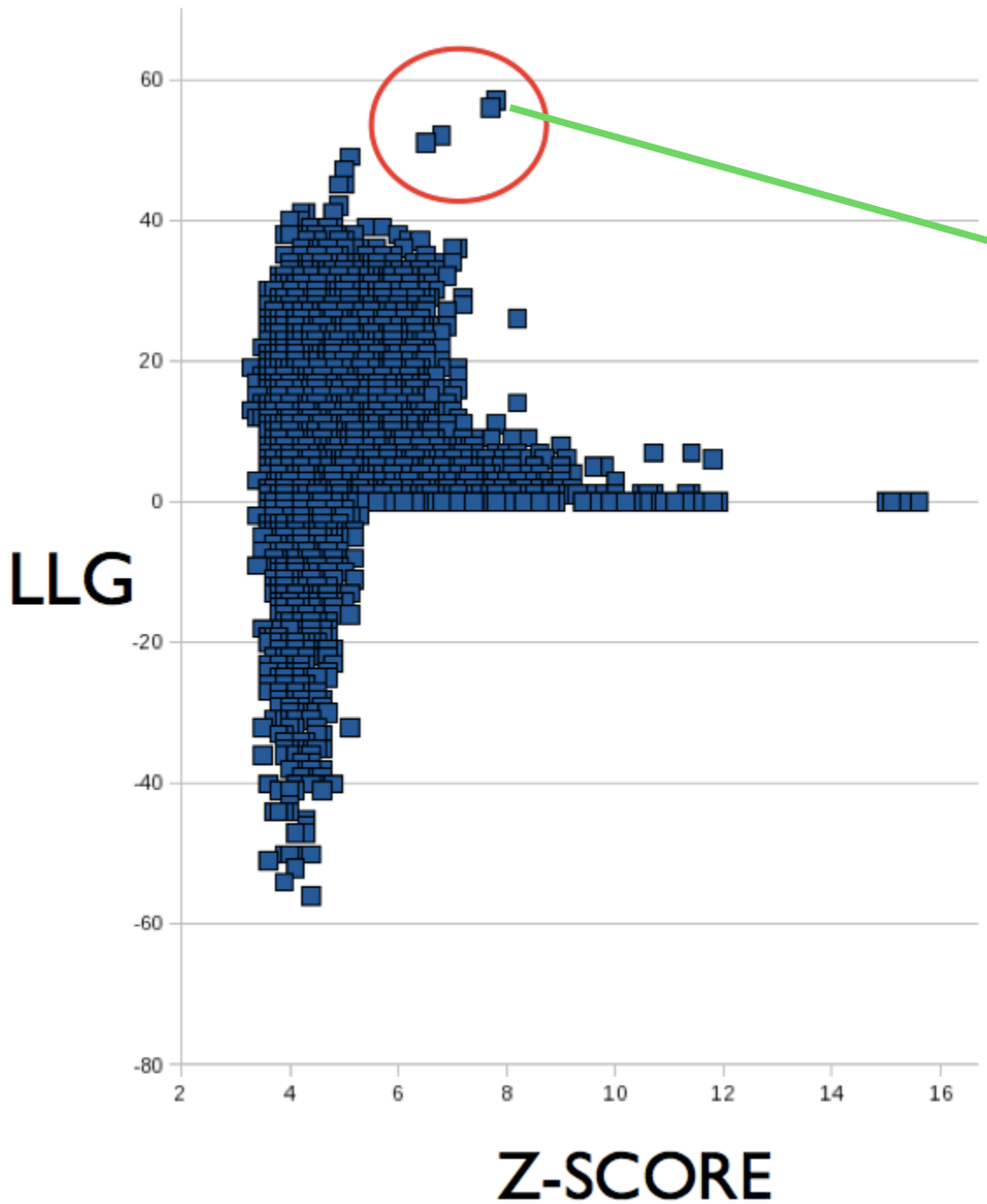
	Apo		FMN·FMN complex	FMN·NADH complex
	Native	Se-MAD		
Data				
Wavelength (Å)	1.0332	0.97925 (P), 0.97942 (I), 0.91162 (R)	1.54	1.54
Resolution (Å)	20-2.5	47.4-2.66 (P), 47.4-2.66 (I), 47.4-2.62 (R)	20-2.5	20-2.5
Space group	$P6_422$	$P6_422$	$P6_422$	$P6_422$
Cell dimensions (Å)	$a = 101.59, c = 130.16$	$a = 101.78, c = 130.07$	$a = 101.27, c = 130.22$	$a = 101.18, c = 129.71$
Asymmetric unit	1 molecule	1 molecule	1 molecule	1 molecule
Total observations	233,238	133,802 (P), 136,160 (I), 141,191 (R)	233,187	233,200
Completeness (%)	99.9 (99.7)	100.0 (100.0)	99.9 (99.3)	99.9 (99.5)
R_{sym} [a,b]	5.5 (13.6)	8.3 (13.9) (P), 9.5 (14.4) (I), 9.8 (16.0) (R)	5.7 (11.4)	4.5 (9.5)
Refinement				
Resolution (Å)	12-2.5		12-2.5	12-2.5
No. of reflections (>2σ)	12,415 (89%)		13,826 (99%)	13,264 (96%)
R_{cryst} [c]	20.2		20.3	20.2
R_{free} [d]	23.6		23.2	23.8
r.m.s.d. bonds (Å)	0.014		0.016	0.016
r.m.s.d. angles	3.185°		3.85°	3.82°
No. of atoms				
Protein and ligand	1418		1480	1493
Water	127		121	122

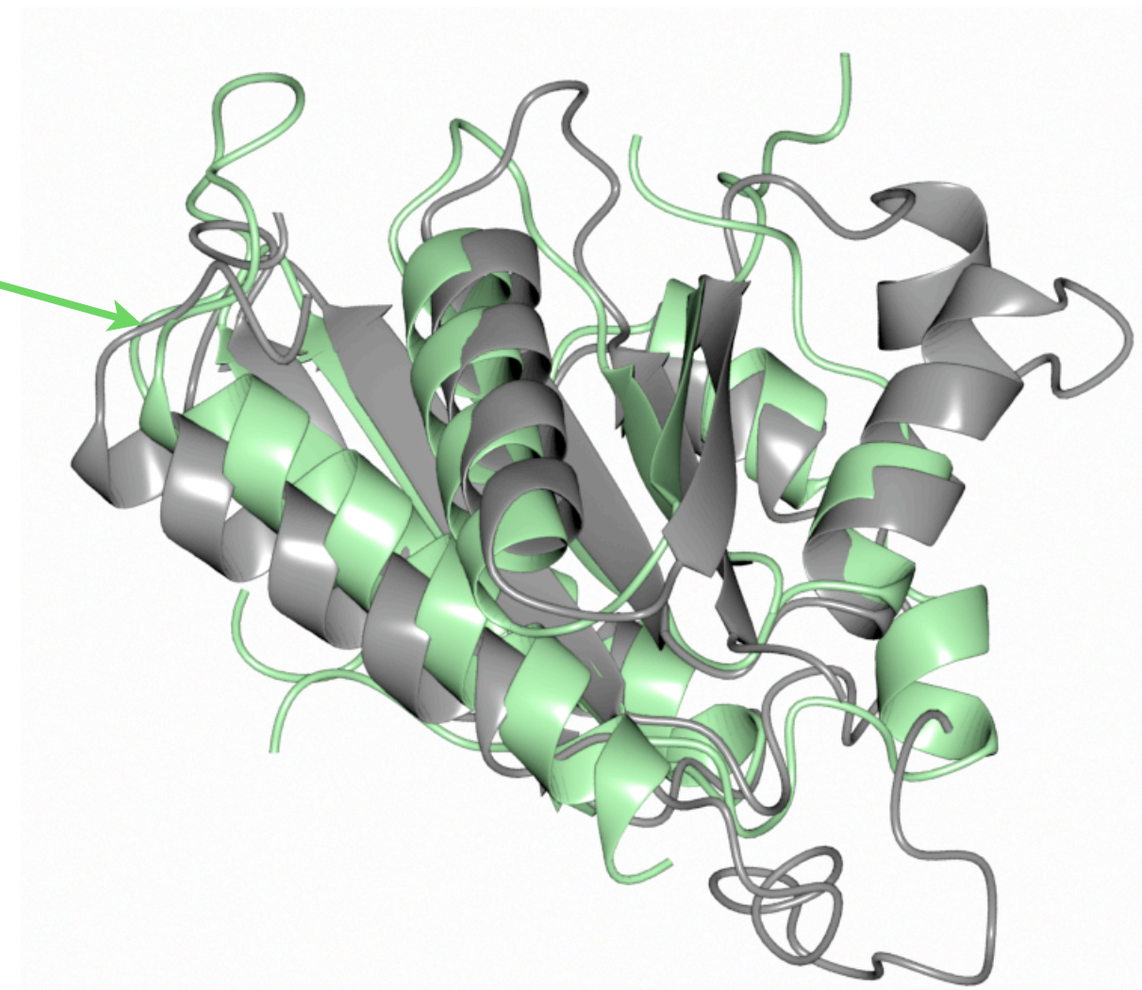
[a] Numbers in parentheses refer to the highest resolution shell.

[b] $R_{sym} = \Sigma|I_h - \langle I_h \rangle|/\Sigma I_h$, where $\langle I_h \rangle$ is the average intensity over symmetry equivalent reflections.

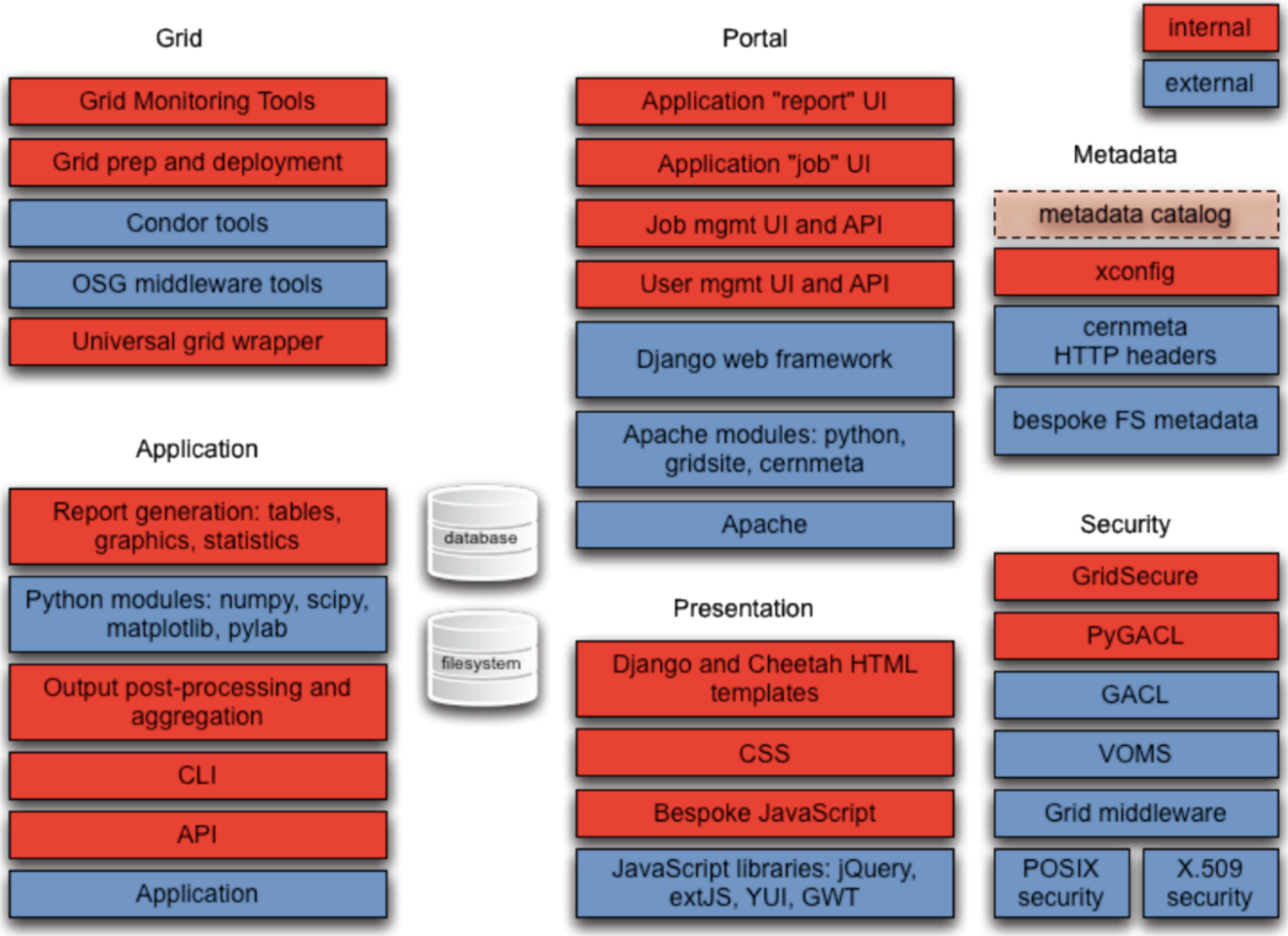
[c] $R_{cryst} = \Sigma|F_o - F^c|/\Sigma F^o$, where summation is over the data used for refinement.

[d] R_{free} was calculated as for R_{cryst} using 5% of the data that were excluded from refinement.





Sequence Identity < 20%
3 cycles of refinement in Phenix shift
secondary structure elements
and lower Rfac to 43%


Grid
Grid Monitoring Tools
Grid prep and deployment
Condor tools
OSG middleware tools
Universal grid wrapper
Application
Report generation: tables, graphics, statistics
Python modules: numpy, scipy, matplotlib, pylab
Output post-processing and aggregation
CLI
API
Application
database
filesystem
Portal
Application "report" UI
Application "job" UI
Job mgmt UI and API
User mgmt UI and API
Django web framework
Apache modules: python, gridsite, cernmeta
Apache
Presentation
Django and Cheetah HTML templates
CSS
Bespoke JavaScript
JavaScript libraries: jQuery, extJS, YUI, GWT
internal
external
Metadata
metadata catalog
xconfig
cernmeta HTTP headers
bespoke FS metadata
Security
GridSecure
PyGACL
GACL
VOMS
Grid middleware
POSIX security
X.509 security

- **NEBioGrid Django Portal**

  Interactive dynamic web portal for workflow definition, submission, monitoring, and access control

- **NEBioGrid Web Portal**

  GridSite based web portal for file-system level access (raw job output), meta-data tagging, X.509 access control/sharing, CGI

- **PyCCP4**

  Python wrappers around CCP4 structural biology applications

- **PyCondor**

  Python wrappers around common Condor operations

  enhanced Condor log analysis

- **PyOSG**

  Python wrappers around common OSG operations

- **PyGACL**

  Python representation of GACL model and API to work with GACL files

- **osg_wrap**

  Swiss army knife OSG wrapper script to handle file staging, parameter sweep, DAG, results aggregation, monitoring

- **sbanalysis**

  data analysis and graphing tools for structural biology data sets

- **osg.monitoring**

  tools to enhance monitoring of job set and remote OSG site status

- **shex**

  Write bash scripts in Python: replicate commands, syntax, behavior

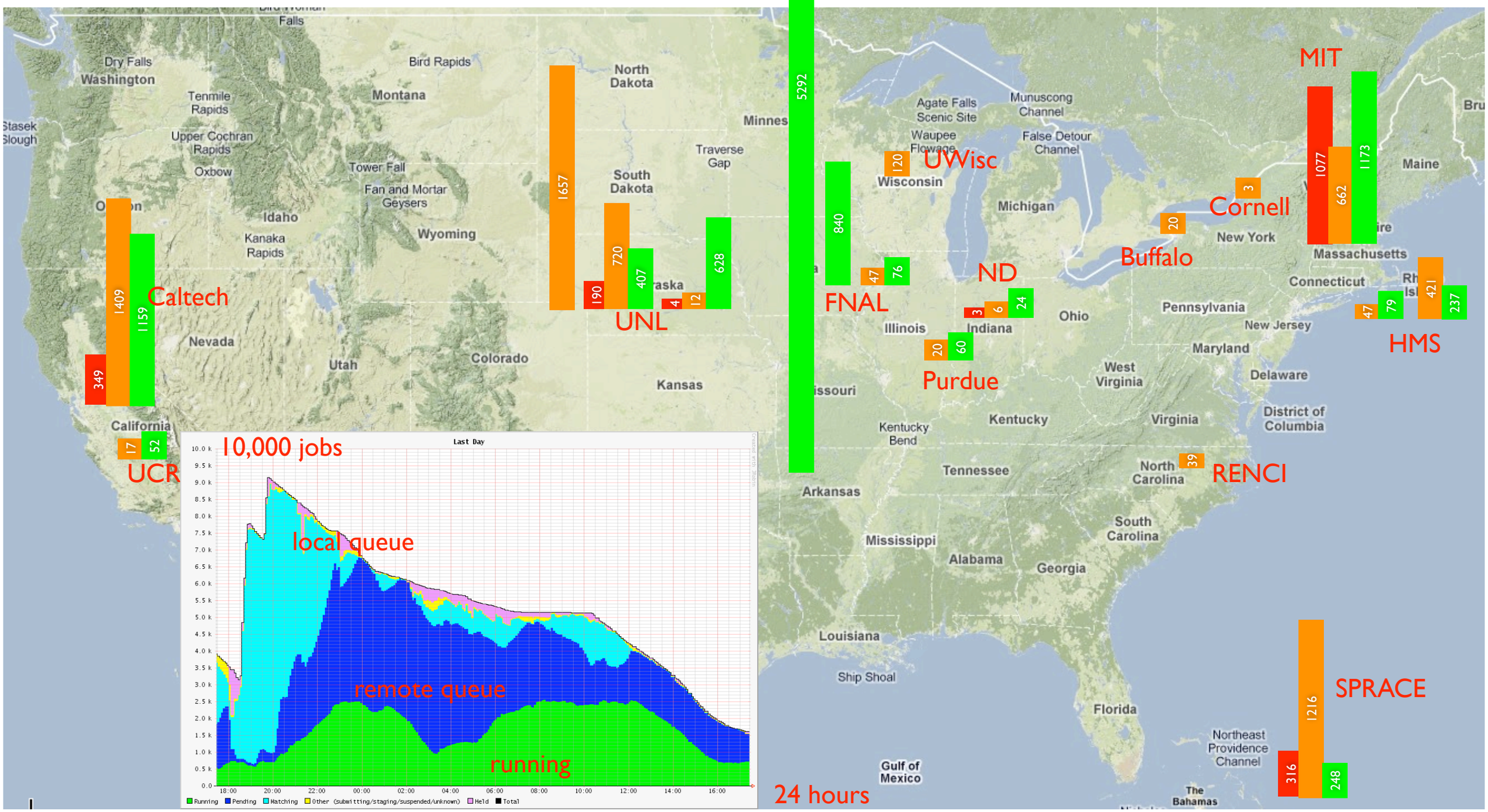
- **xconfig**

  Universal configuration

Example Job Set

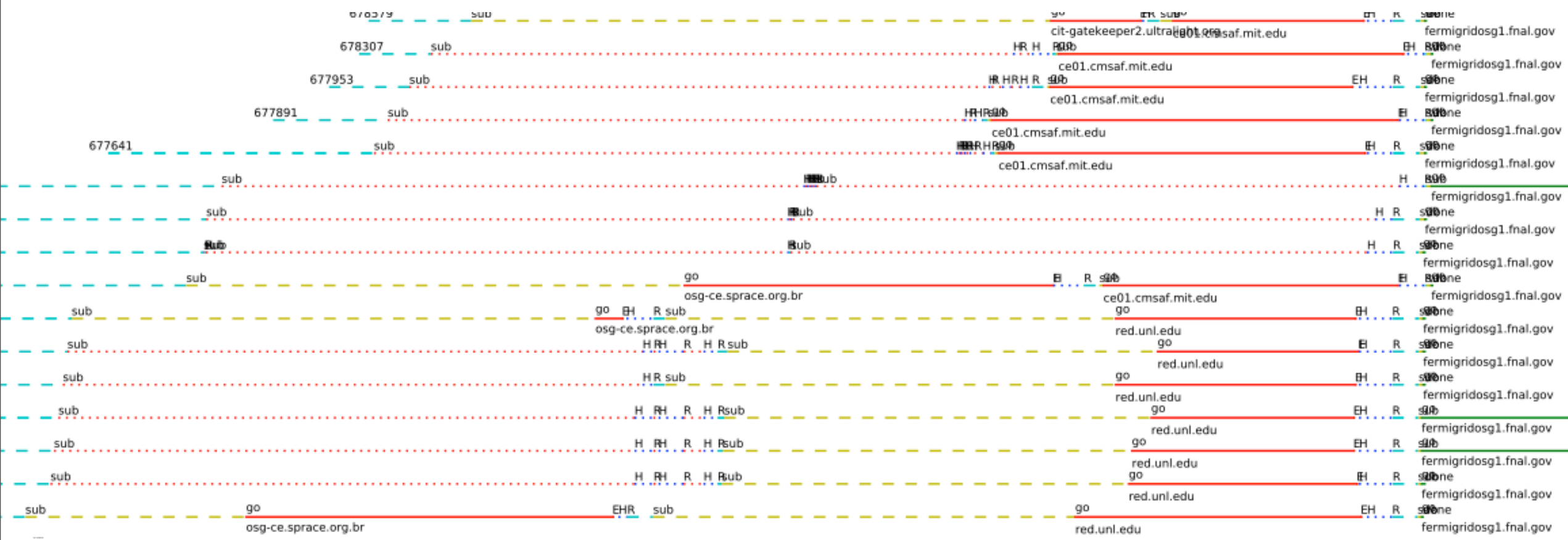
10k grid jobs
approx 30k CPU hours
99.7% success rate
24 wall clock hours

evicted - red
held - orange
completed - green



Job Lifelines

- local submit
- s - grid submit
- g - go
- d - done
- e - evict (running)
- e - evict (queued)
- h - hold (running)
- h - hold (queued)
- r - release
- ? - unknown



Typical Layered Environment

Map-Reduce

- Command line application (e.g. Fortran)
- Friendly application API wrapper
- Batch execution wrapper for N-iterations
- Results extraction and aggregation
- Grid job management wrapper
- Web interface
- forms, views, static HTML results
- GOAL eliminate shell scripts
- often found as “glue” language between layers

Acknowledgements

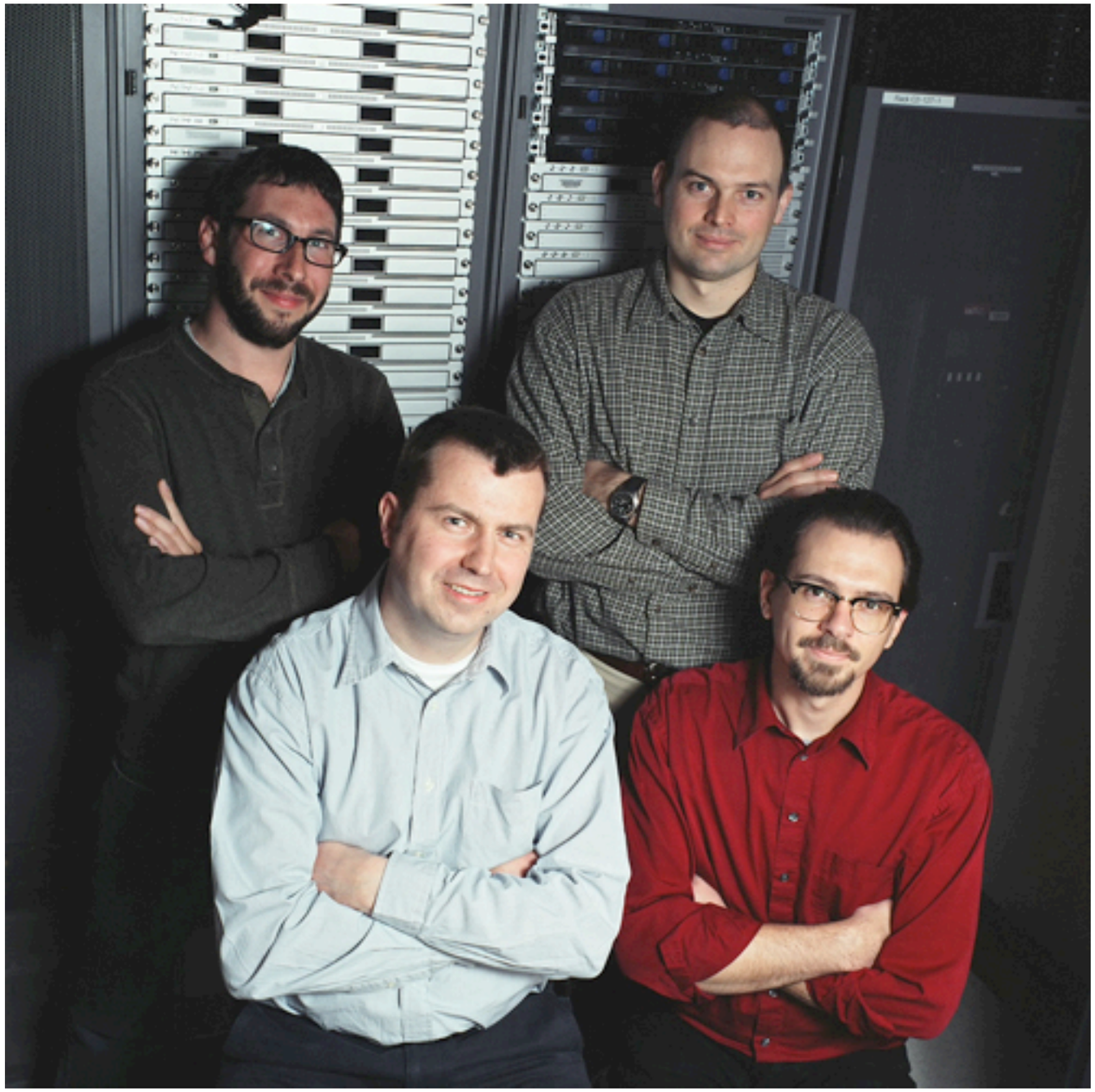
Piotr Sliz
PI and SBGrid team leader

Peter Doherty
Grid Administrator

Ian Levesque
Systems Architect

Ben Eisenbraun
Software Curator

Steve Jahl
System Administrator

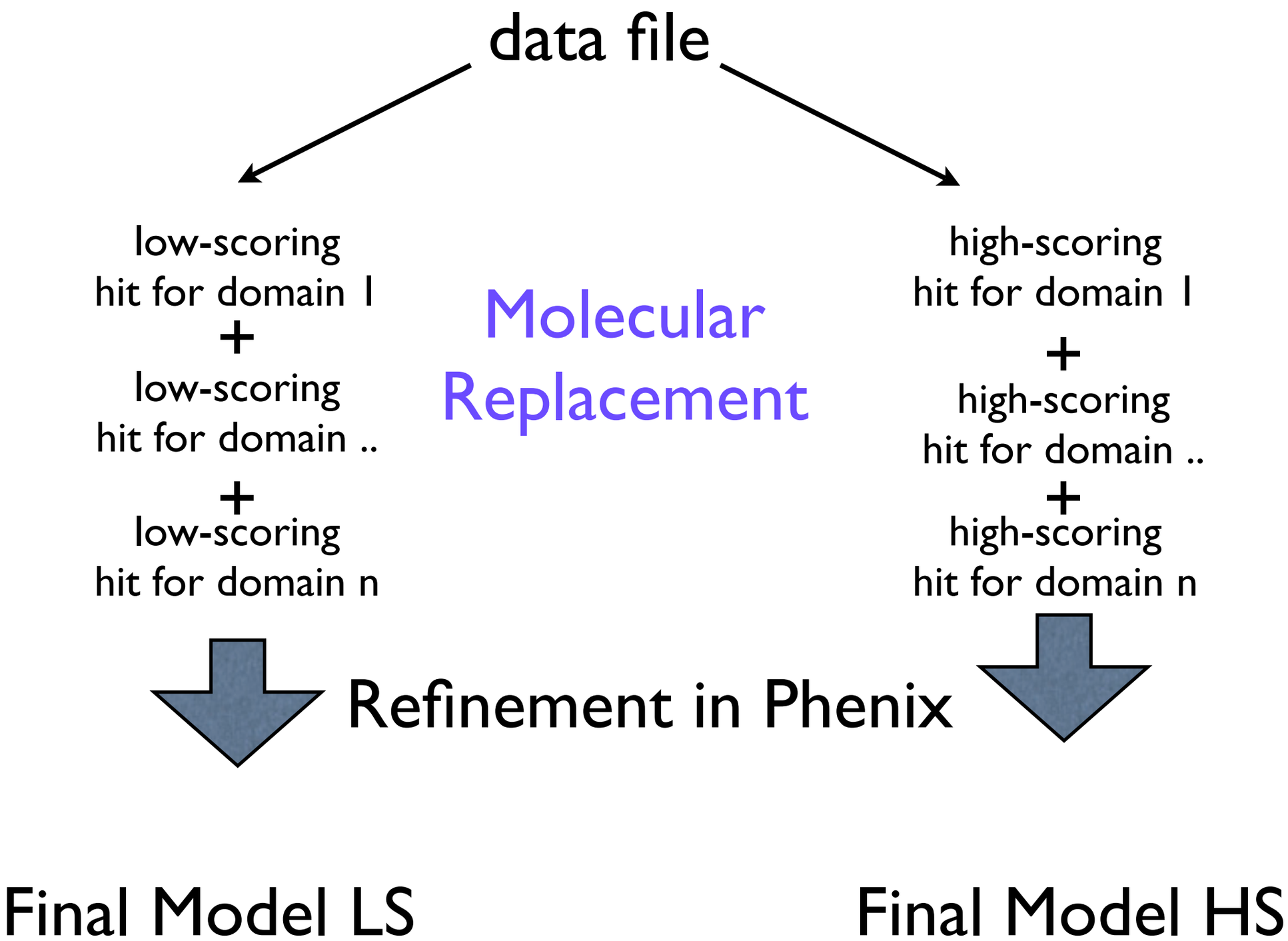


http://abitibi.sbgrid.org
http://www.nebiogrid.org

The End

for later

Structure Determination Strategy:



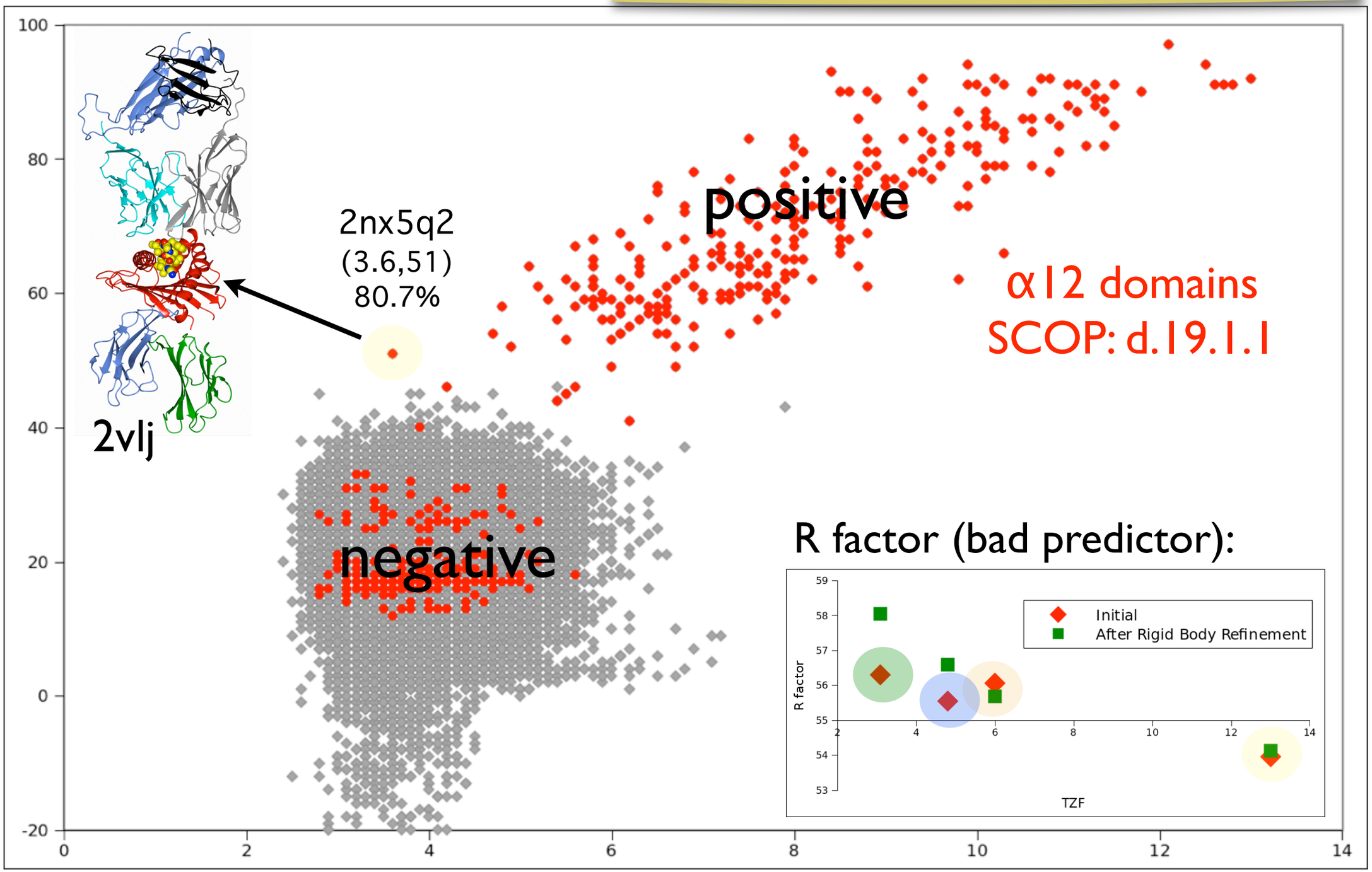
color all Ig domains

2D representat

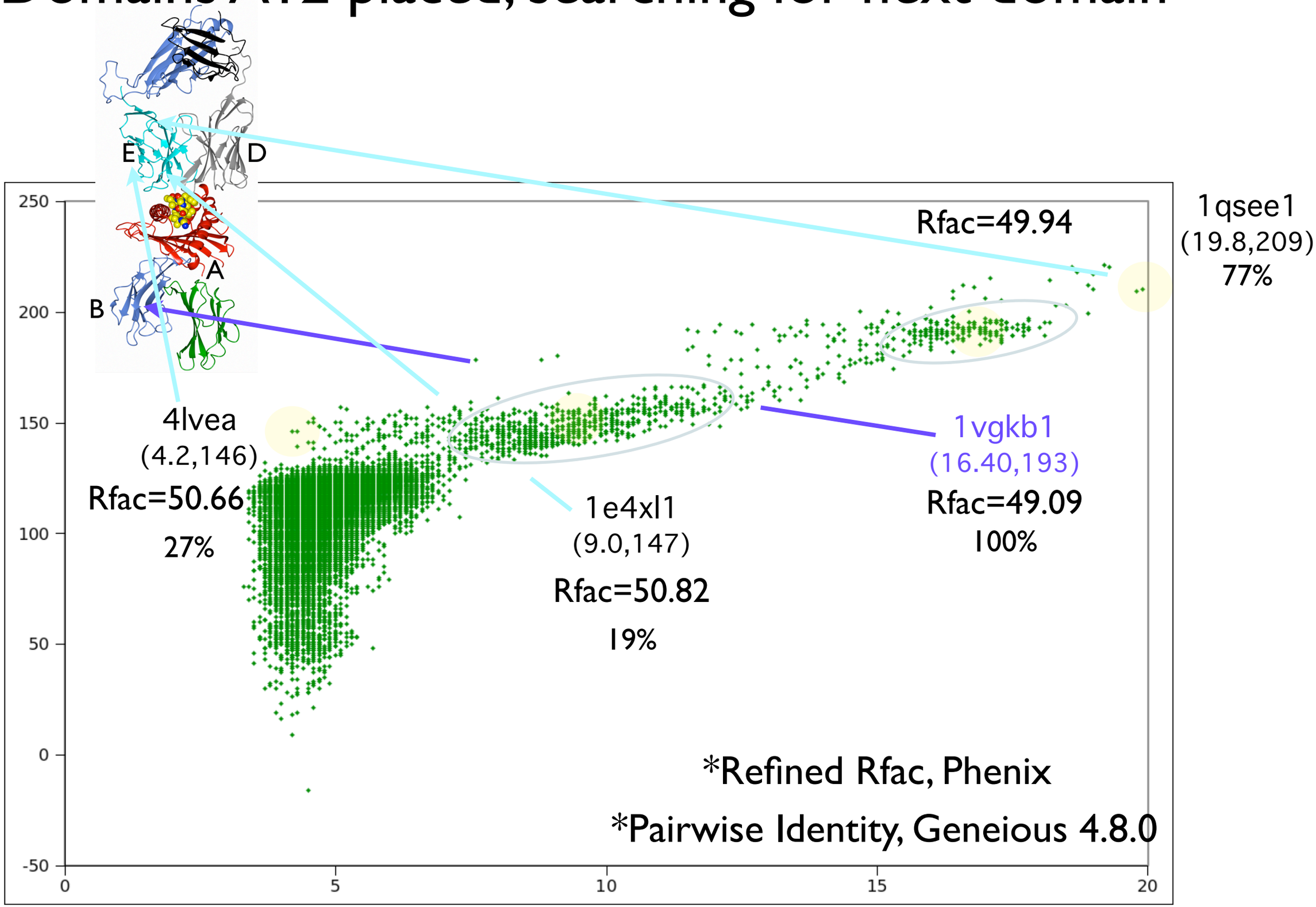
Top Scoring S

- wide TFZ range of solutions (from 3.5 to 14) which overlaps with missed searches
LLG score does not overlap with failed searches

- both TFZ and LLG scores predict the most likely MR candidate



Domains A12 placed, searching for next domain

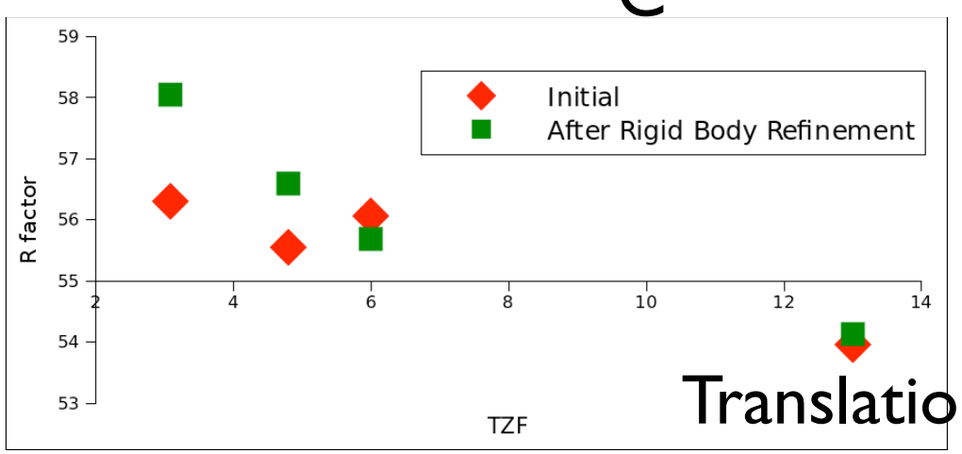


3 cycles of refinement in Phenix
Rigid Body + ADP

D

B

C



A

Translation Function Z Score

Discriminating Solutions