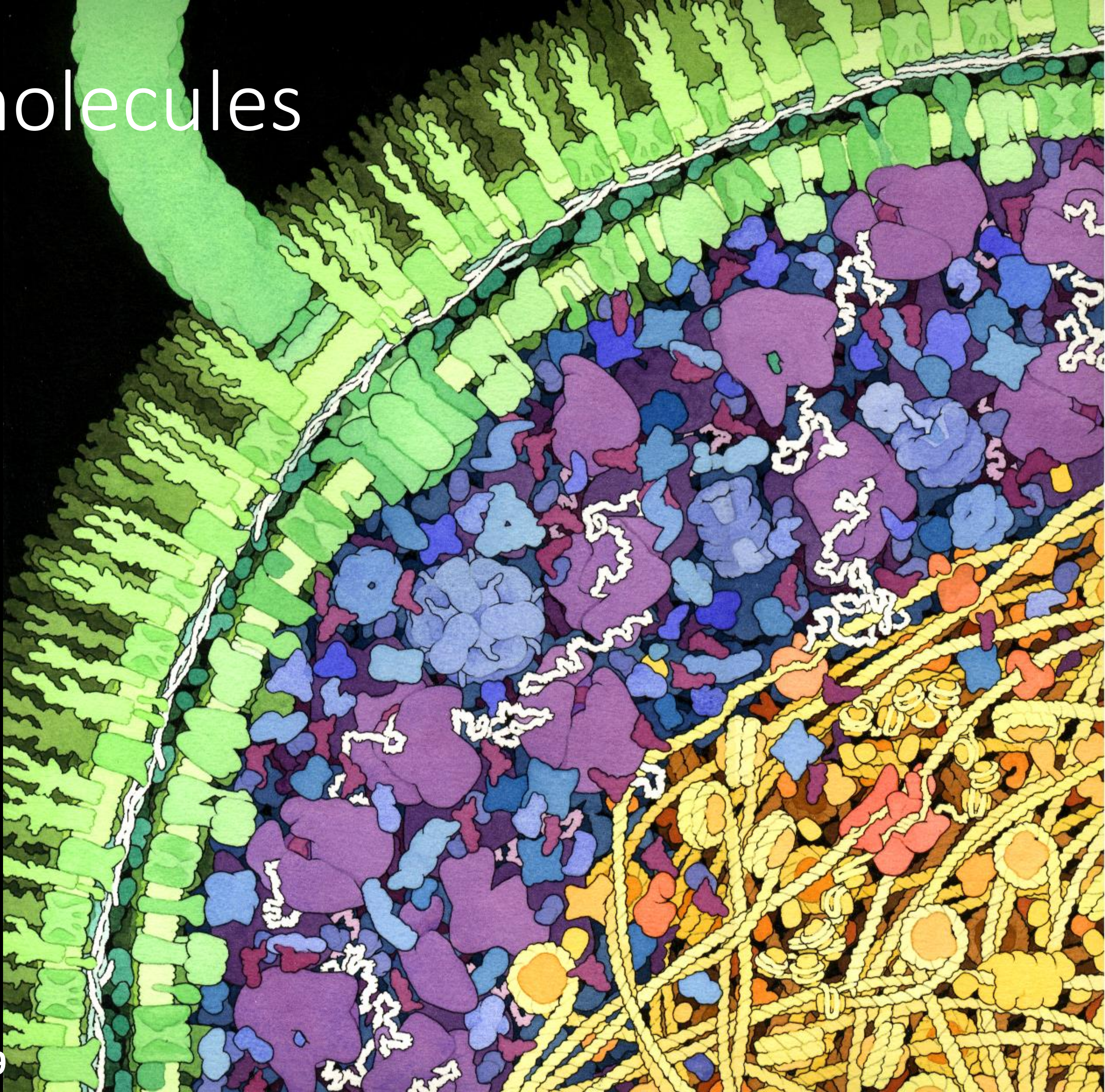
HTCondor and macromolecular structure validation

Vincent Chen

John Markley/Eldon Ulrich, NMRFAM/BMRB, UW@Madison

David & Jane Richardson, Duke University

Macromolecules



David S. Goodsell 1999

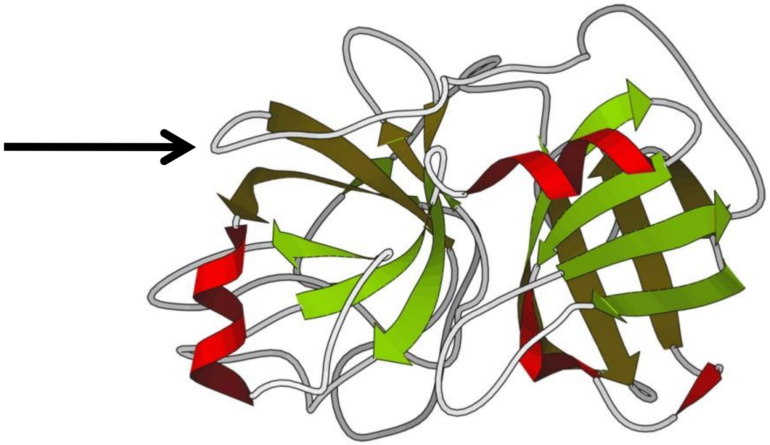
Two questions of structural biology

Sequence

IVGGTSASAGDFPFI
VSISRNGGPWCGGSL
LNANTVLTAAHCVSG
YAQSGFQIRAGSLSR
TSGGITSSLSSVRVH
PSYSGNNNDLAILKL
STSIPSGGNIGYARL
AASGSDPVAGSSATV
AGWGATSEGGSSTPV
NLLKVTVPIVSRATC
RAQYGTSAITNQMFC
AGVSSGGKDSCQGDS
GGPIVDSSNTLIGAV
SWGNGCARPNYSGVY
ASVGALRSFIDTYA

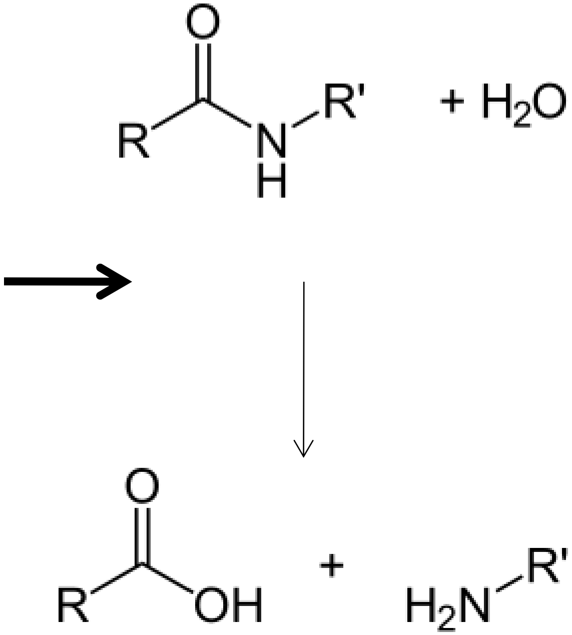
Trypsin sequence

3D structure



Trypsin structure
PDB: 1pq7

Function



Trypsin reaction
Hydrolysis of peptide bond

How do we solve structures?

X-ray crystallography

- X-ray diffraction of crystals
- Provides a picture of the electron density of a macromolecular structure
- Overall shape, but no atom identities
- Lower numbers on resolution means more data

NMR Spectroscopy

- NMR spectra of solutions
- Provides relationships (distances, angles, dihedral angles) between atoms
- Information about specific atoms, but no overall shape

Protein Data Bank (PDB)

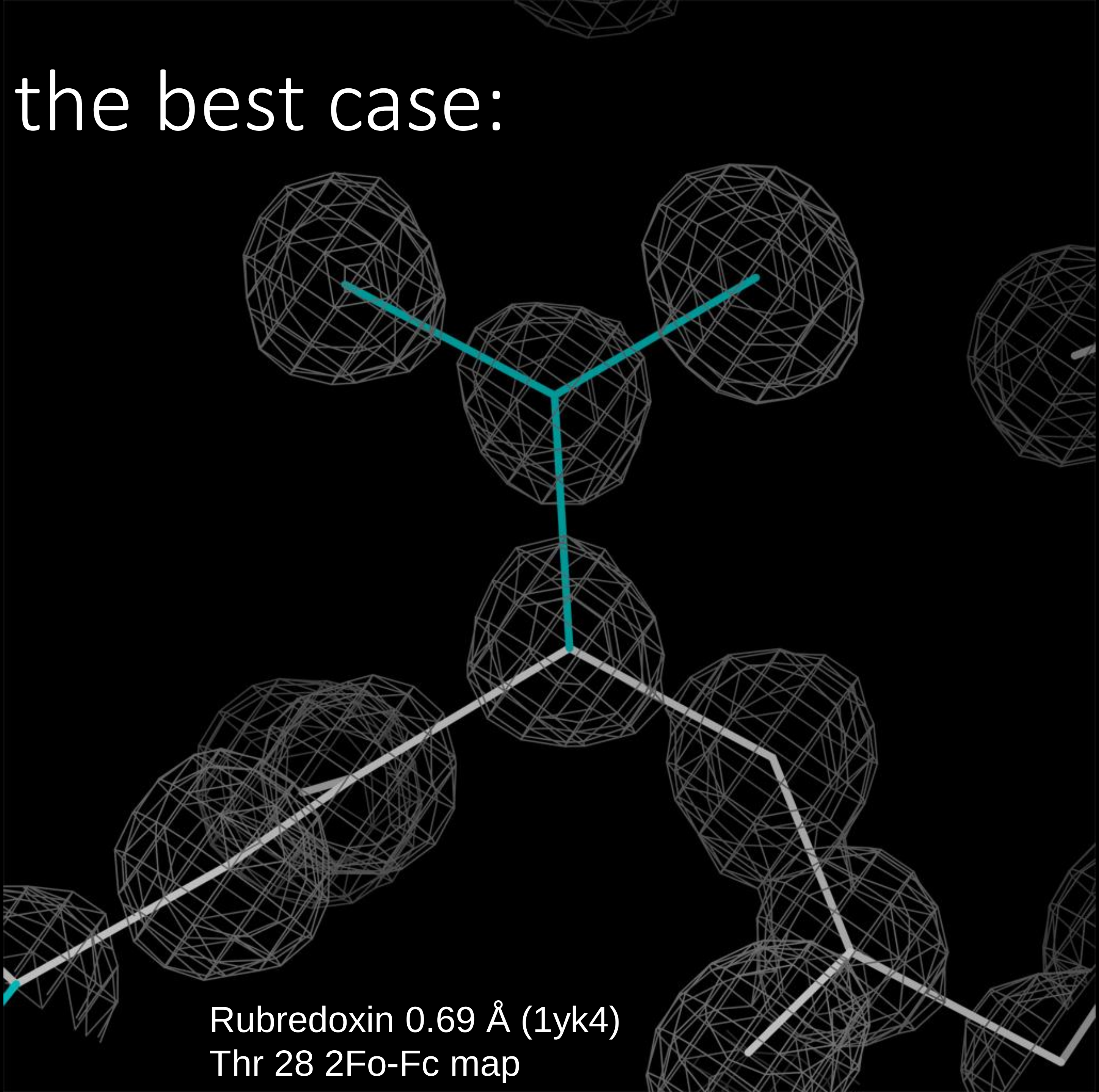
- Repository for 3D structures and data
- Also refers to the file format
- 88,247 X-ray structures vs 10,451 NMR structures deposited
- 92,283 protein structures vs 2,557 nucleic acid structures (~4600 protein-nucleic acid complexes)
- We make extensive use of the structures deposited in the PDB

Berman et al. (2000) NAR, 28, 235-242

Building high-quality models is difficult

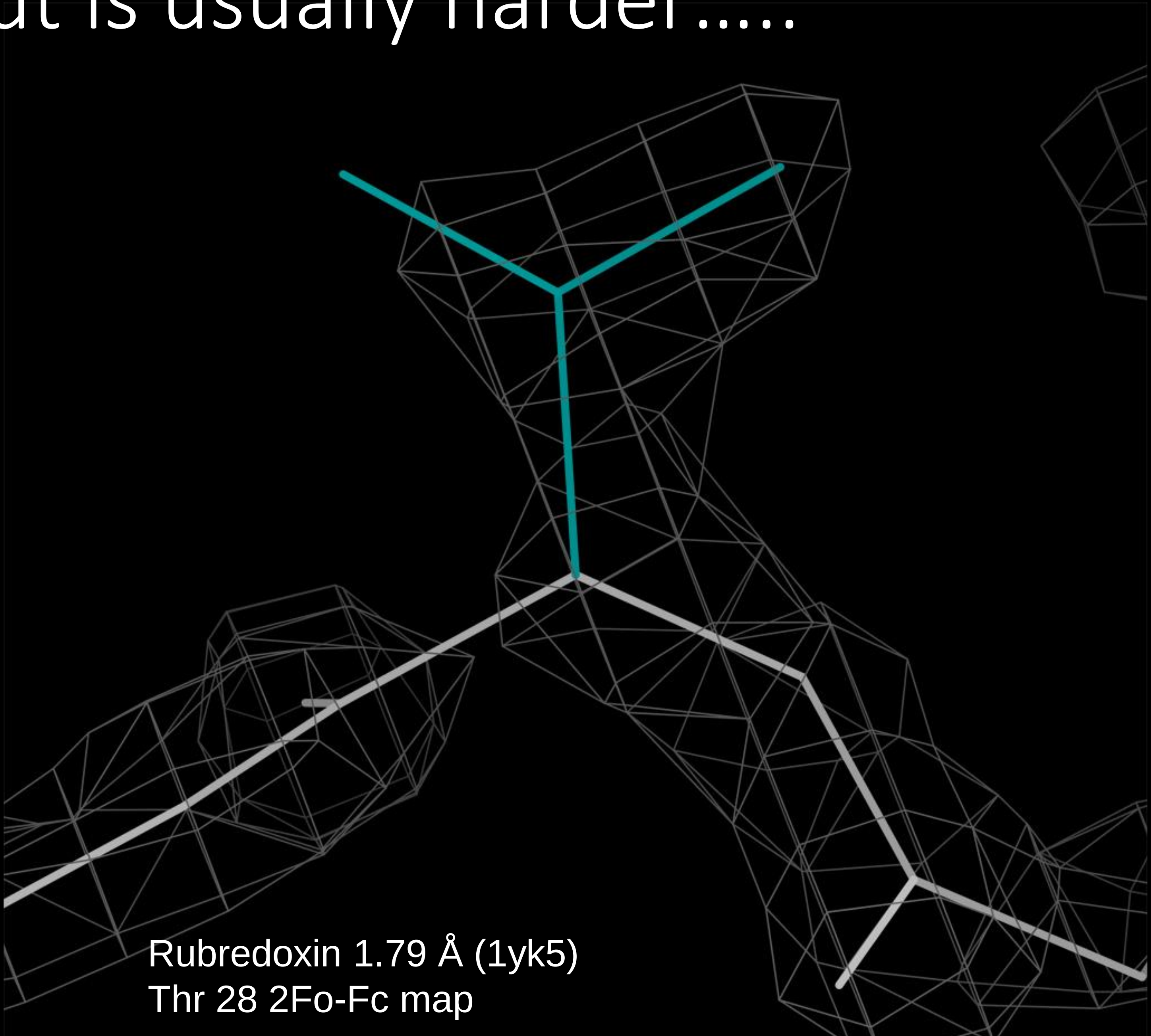
- No way to directly see atom positions
- X-ray crystallography and NMR spectroscopy provide *models* of structures
    - Structural biologists should build the highest quality models possible
    - Data is limited
    - Have to use other knowledge (chemistry, algorithms, etc) to fill in for lack of data
    - Subjectivity in interpreting data

In the best case:



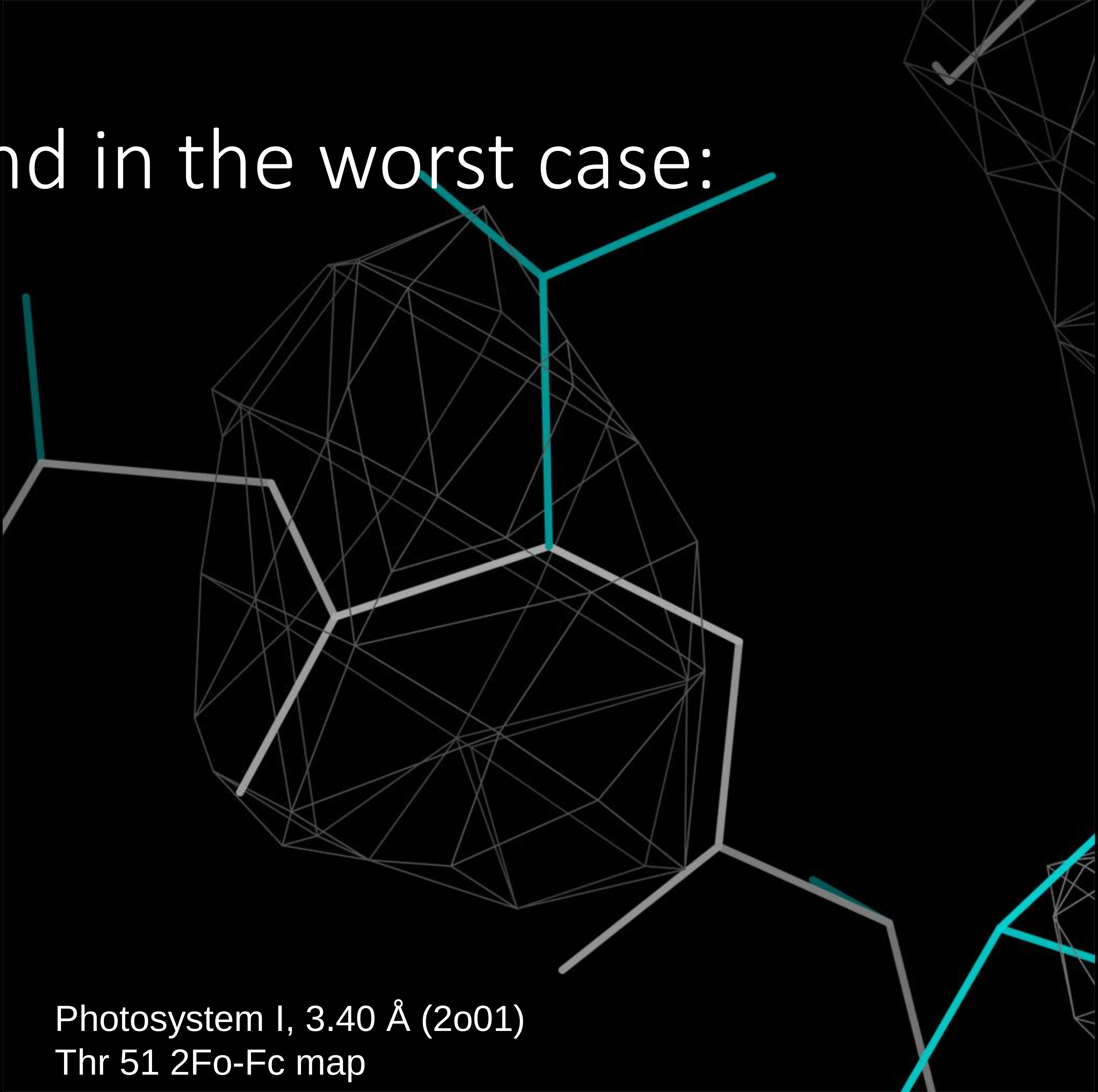
Rubredoxin 0.69 Å (1yk4)
Thr 28 2Fo-Fc map

But is usually harder…..



Rubredoxin 1.79 Å (1yk5)
Thr 28 2Fo-Fc map

And in the worst case:

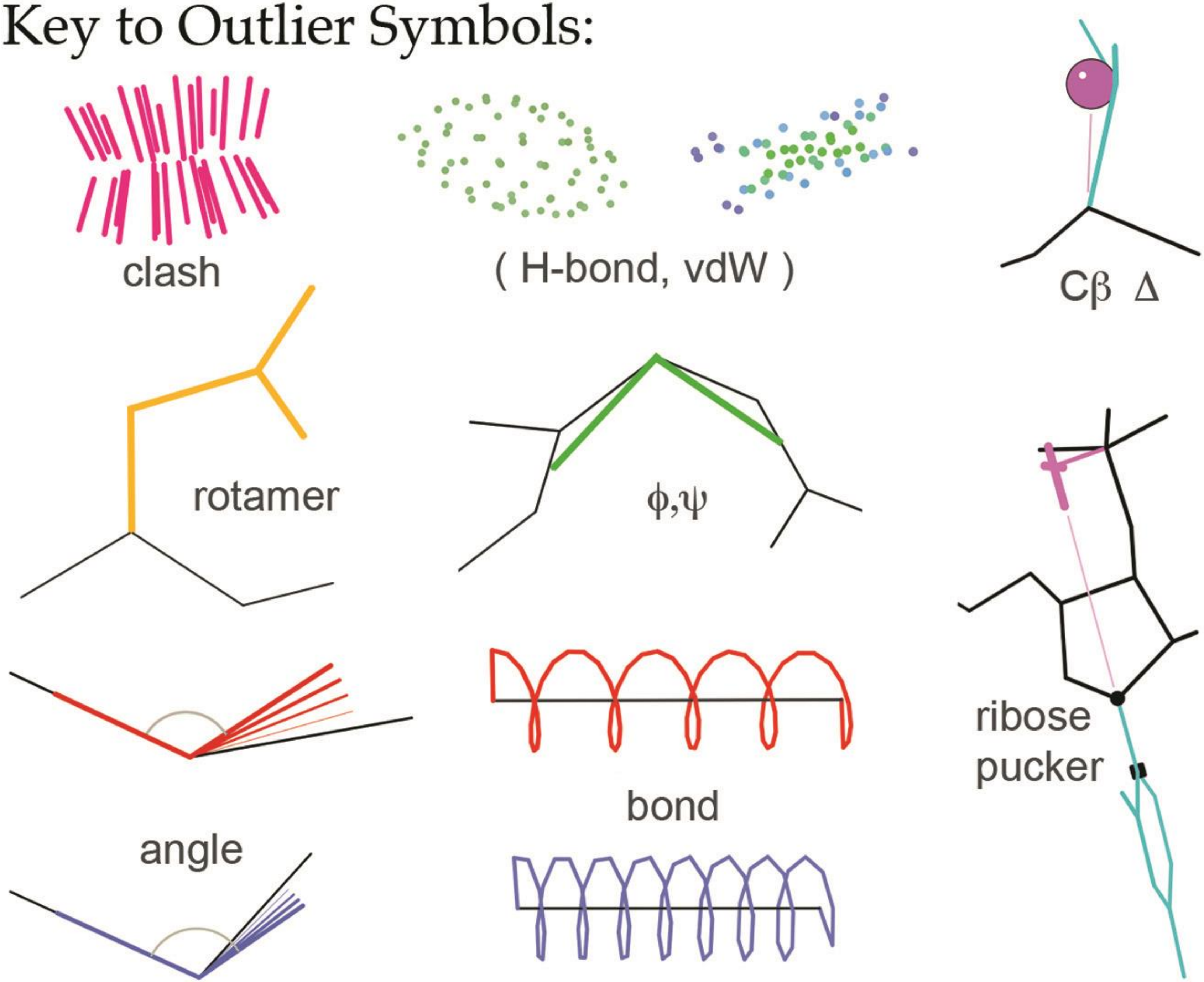


Photosystem I, 3.40 Å (2o01)
Thr 51 2Fo-Fc map

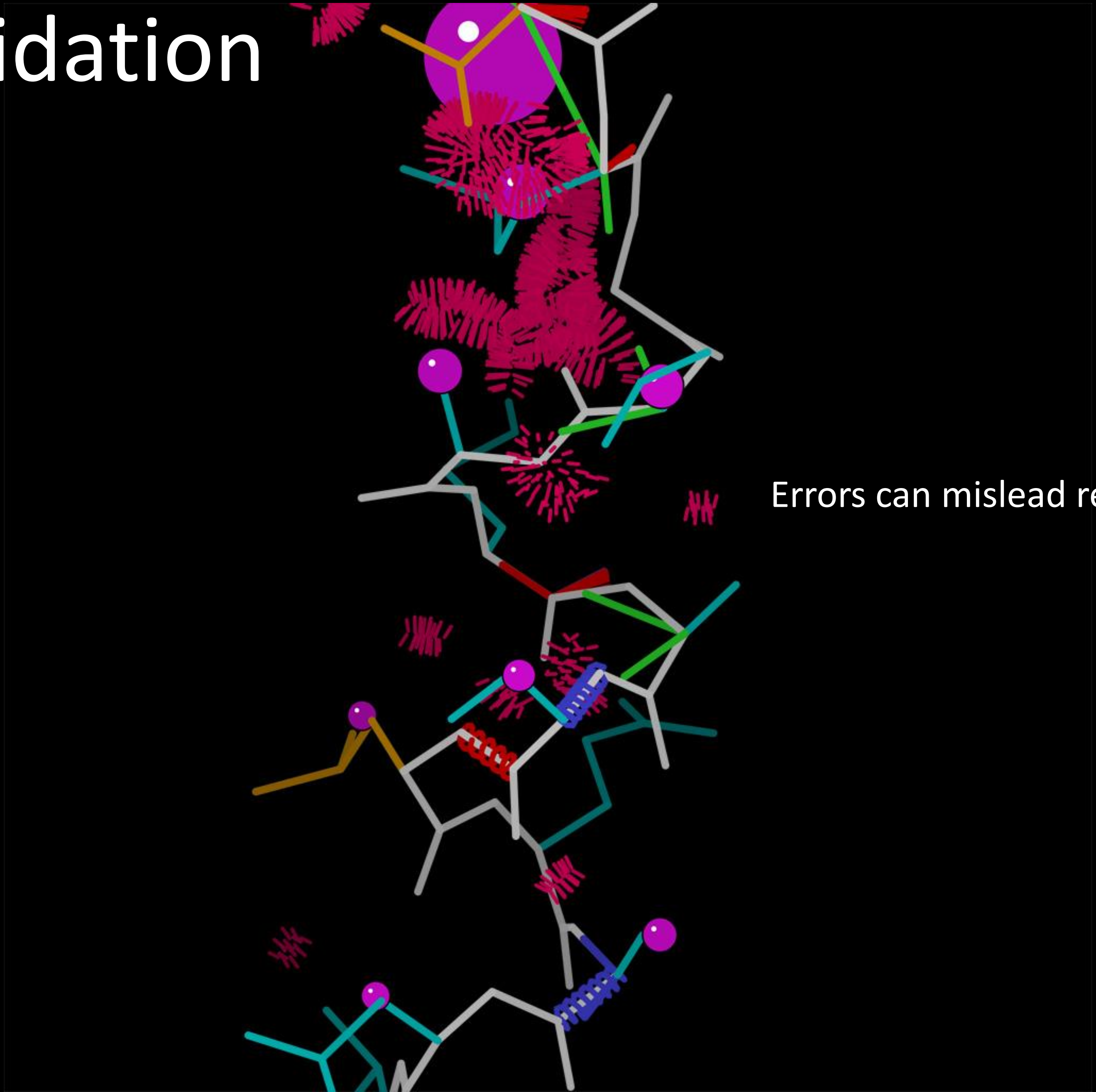
Errors in models

- Steric clashes, Ramachandran outliers, poor sidechain rotamers, bad bond geometry
- Sequence register shifts, underpacking
- Structural validation is needed!
- Users and scientists should filter (i.e. remove errors) from models before use
- MolProbity website for structure validation (i.e. finding errors)
- Errors presented in visual and tabular formats

Key to Outlier Symbols:

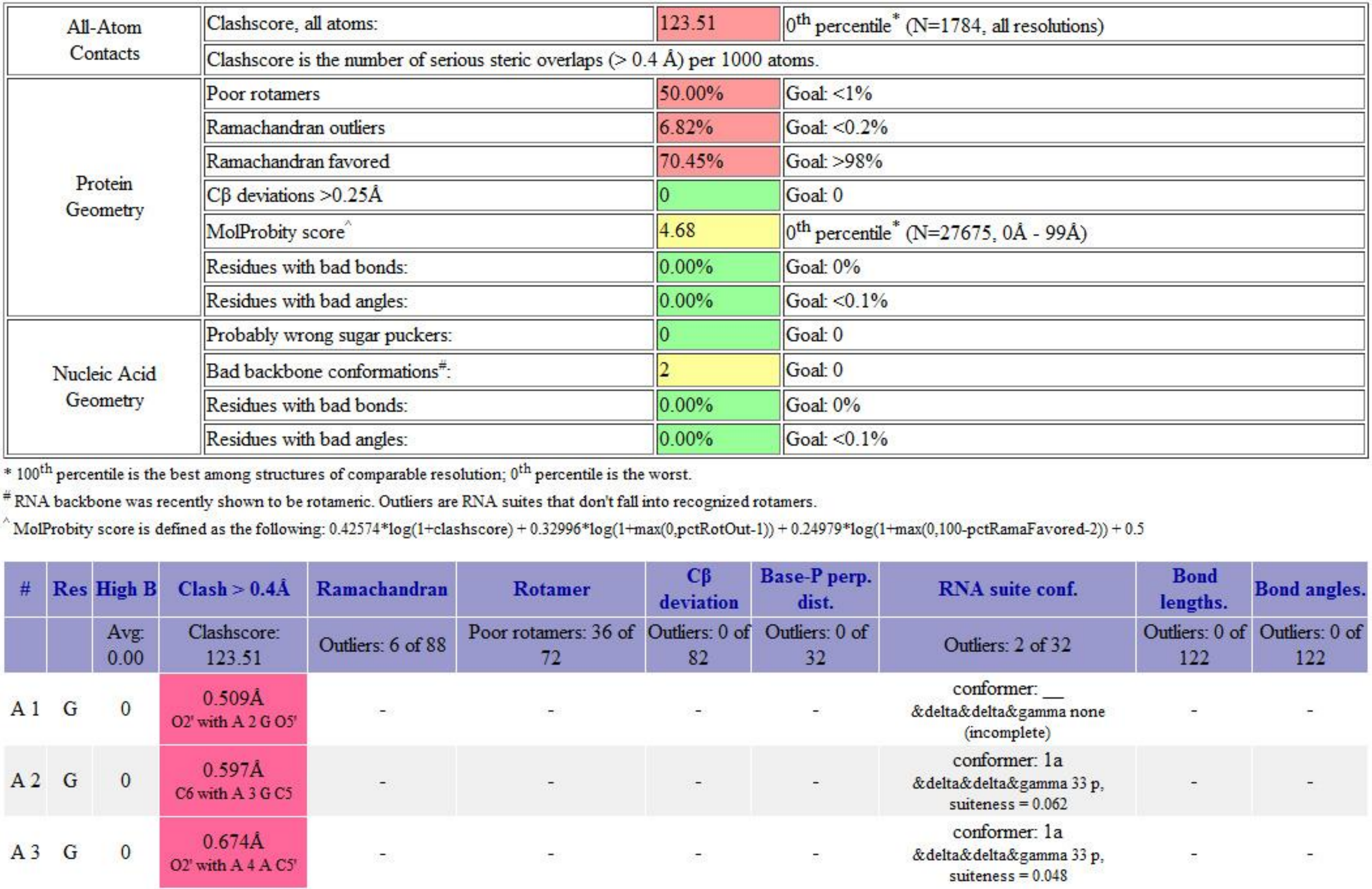


Visualizing a structure with validation



Errors can mislead research

Validation report table



All-Atom Contacts	Clashscore, all atoms:	123.51	0th percentile* (N=1784, all resolutions)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.		
Protein Geometry	Poor rotamers	50.00%	Goal: <1%
	Ramachandran outliers	6.82%	Goal: <0.2%
	Ramachandran favored	70.45%	Goal: >98%
	Cβ deviations >0.25Å	0	Goal: 0
	MolProbity score^	4.68	0th percentile* (N=27675, 0Å - 99Å)
	Residues with bad bonds:	0.00%	Goal: 0%
	Residues with bad angles:	0.00%	Goal: <0.1%
Nucleic Acid Geometry	Probably wrong sugar puckers:	0	Goal: 0
	Bad backbone conformations#:	2	Goal: 0
	Residues with bad bonds:	0.00%	Goal: 0%
	Residues with bad angles:	0.00%	Goal: <0.1%

* 100th percentile is the best among structures of comparable resolution; 0th percentile is the worst.

RNA backbone was recently shown to be rotameric. Outliers are RNA suites that don't fall into recognized rotamers.

^ MolProbity score is defined as the following: 0.42574*log(1+clashscore) + 0.32996*log(1+max(0,pctRotOut-1)) + 0.24979*log(1+max(0,100-pctRamaFavored-2)) + 0.5

#	Res	High B	Clash > 0.4Å	Ramachandran	Rotamer	Cβ deviation	Base-P perp. dist.	RNA suite conf.	Bond lengths.	Bond angles.
		Avg: 0.00	Clashscore: 123.51	Outliers: 6 of 88	Poor rotamers: 36 of 72	Outliers: 0 of 82	Outliers: 0 of 32	Outliers: 2 of 32	Outliers: 0 of 122	Outliers: 0 of 122
A 1	G	0	0.509Å O2' with A 2 G O5'	-	-	-	-	conformer: __ &delta&delta&gamma none (incomplete)	-	-
A 2	G	0	0.597Å C6 with A 3 G C5	-	-	-	-	conformer: 1a &delta&delta&gamma 33 p, suiteness = 0.062	-	-
A 3	G	0	0.674Å O2' with A 4 A C5'	-	-	-	-	conformer: 1a &delta&delta&gamma 33 p, suiteness = 0.048	-	-

MolProbity at BMRB/NMRFAM

- Biological Magnetic Resonance Data Bank – archives and disseminates NMR data on biological molecules
- National Magnetic Resonance Facility at Madison – developing software to facilitate biomolecular NMR spectroscopy
- Incorporate MolProbity validation software into the BMRB/NMRFAM software
    - Improve compatibility of MolProbity with NMR PDB files

MolProbity on large datasets

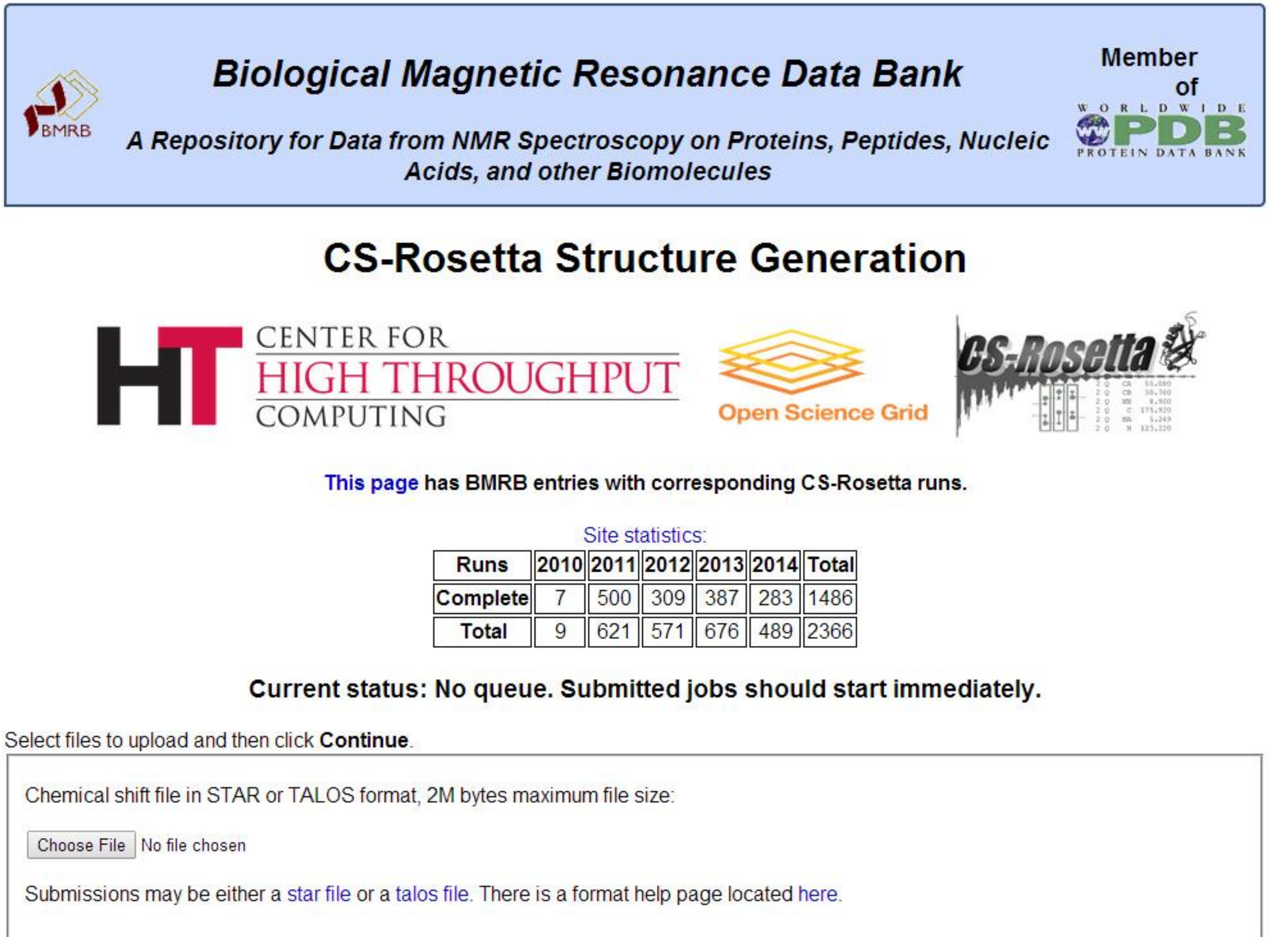
- Command-line tools available:
    - Add hydrogens to files
    - Scripts for generating summary scores for models
- Analyzing 10,000 NMR PDB files
    - 10 batches
    - 2 weeks to analyze
    - Numerous bugs
- High-throughput computing?

HTCondor

HT
CENTER FOR
HIGH THROUGHPUT
COMPUTING

HTCondor @ BMRB

- Pool of 66 slots
- Experience running CS-Rosetta on HTCondor
- Thanks Jon!


BMRB
Member of
WORLDWIDE PDB PROTEIN DATA BANK
CENTER FOR HIGH THROUGHPUT COMPUTING
Open Science Grid
CS-Rosetta

Biological Magnetic Resonance Data Bank

A Repository for Data from NMR Spectroscopy on Proteins, Peptides, Nucleic Acids, and other Biomolecules

CS-Rosetta Structure Generation

This page has BMRB entries with corresponding CS-Rosetta runs.

Site statistics:

Runs	2010	2011	2012	2013	2014	Total
Complete	7	500	309	387	283	1486
Total	9	621	571	676	489	2366

Current status: No queue. Submitted jobs should start immediately.

Select files to upload and then click **Continue**.

Chemical shift file in STAR or TALOS format, 2M bytes maximum file size:

Choose File No file chosen

Submissions may be either a star file or a talos file. There is a format help page located here.

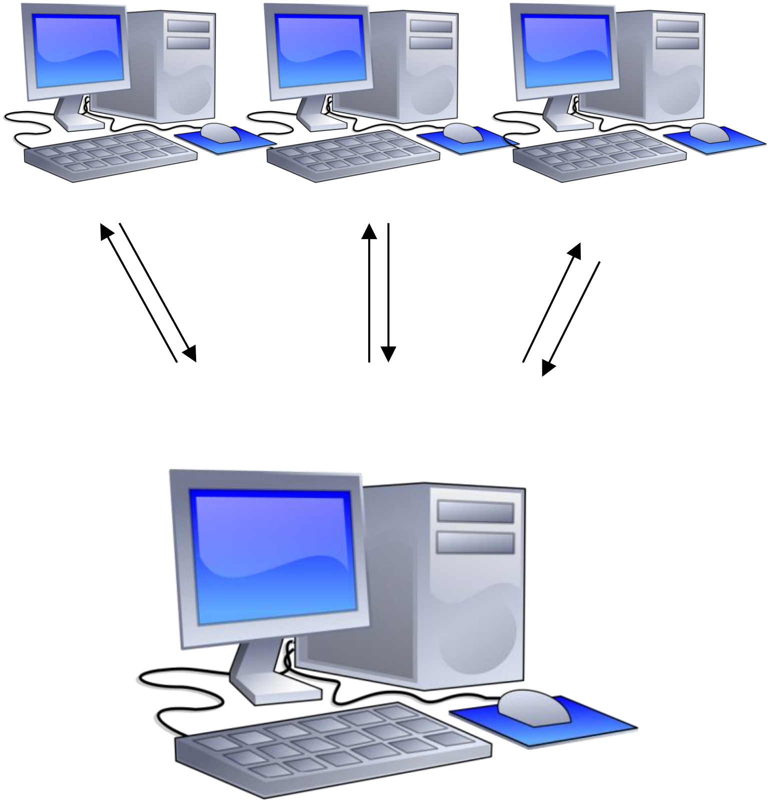
MolProbity = many programs/languages

- C, C++, Java, PHP, shell, Perl, AWK…
  - Reduce – addition of hydrogens
  - Probe – calculates and draws clashes
  - Chiropraxis – calculates rotamer and Ramachandran outliers
  - Dangle – calculates bond geometry outliers
  - Suitename – calculates RNA backbone conformers

  …..

MolProbity runs each program on each PDB file one at a time

HTCondor + MolProbity?



- HTCondor distributes software/input files to available machines
- Runs the jobs, then returns the results
- Impractical to send whole MolProbity package (30 MB)
- Rewrote analysis as a Python script
    - HTCondor sends individual programs/pdb files to compute nodes

HTCondor novice pitfalls

- Things which are easy to do with HTCondor, and are **bad**:
  - Spawning 100s of local compute jobs within a few seconds on one machine
  - Trying to write output to directories that don’t exist
  - Having multiple jobs trying to write to the same log file at the same time
  - Storing 100,000+ PDB/result/log files in one directory

MolProbity + PDB files pitfalls

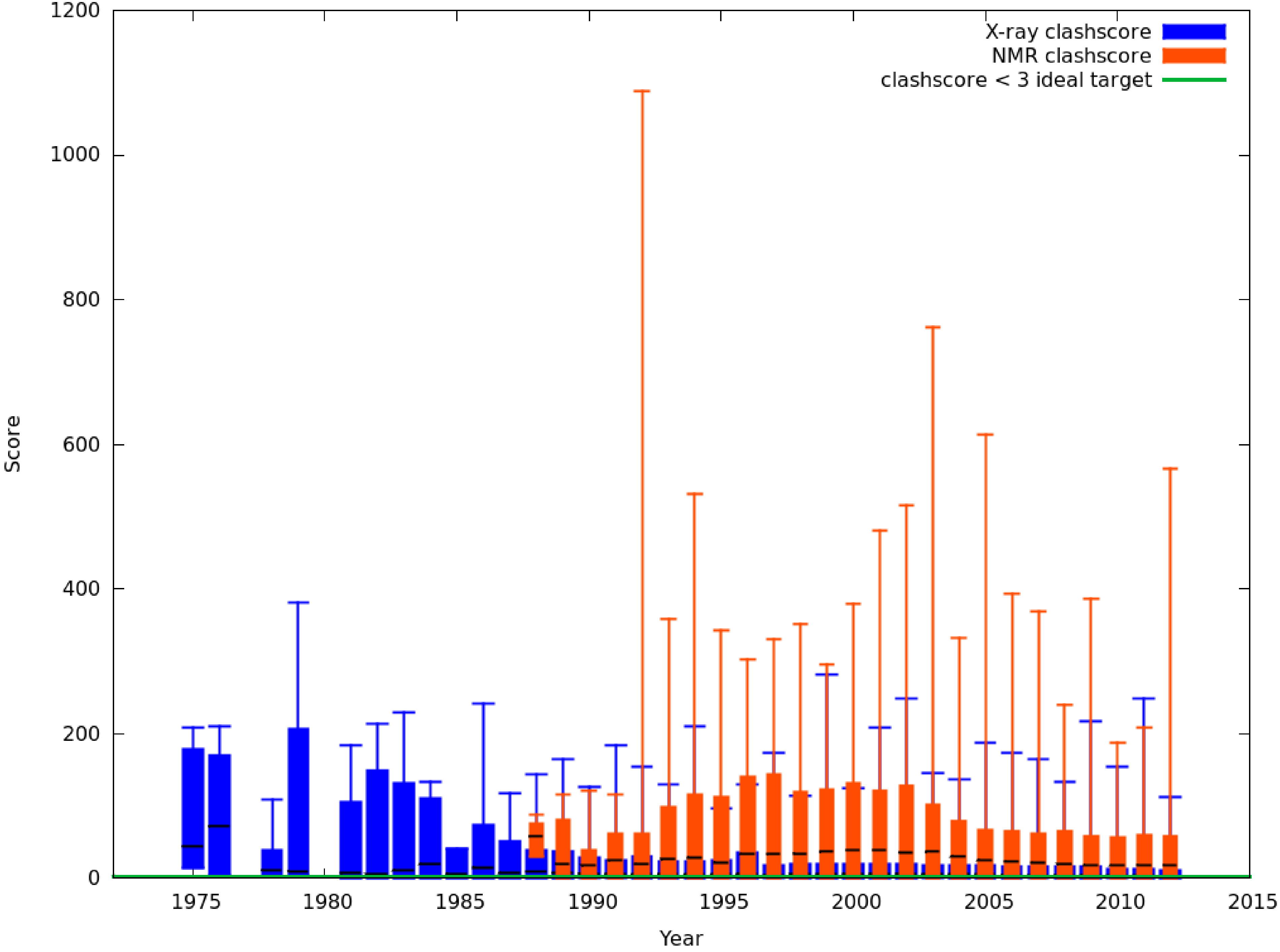
- Multiple model PDB files
    - NMR structures are typically ensembles of models that are most consistent with data
- PDB format doesn’t have many constraints
    - Calpha only models and models missing sidechains
    - Structures with no standard protein or nucleic acid residues

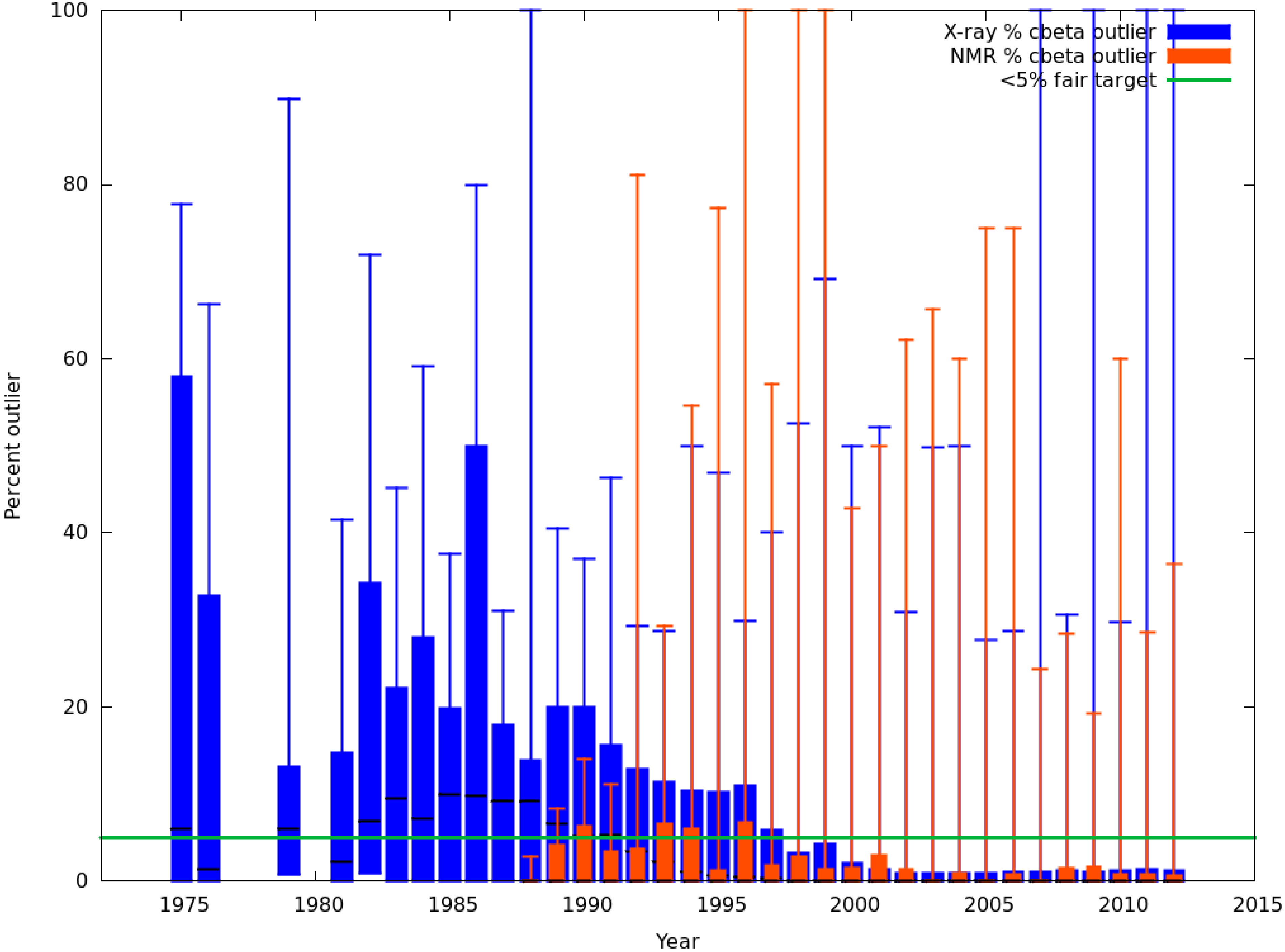
HTCondor + MolProbity

- Python script input: directory of PDB files
  - Divides up PDB files into separate directories
  - Prepares output directories
  - Writes dag and submit files
- Uses DAGMan to manage jobs
- Output:
  - MolProbity overall summary scores for models
  - *Scores for residue-level in models*

Results of HTCondor + MolProbity

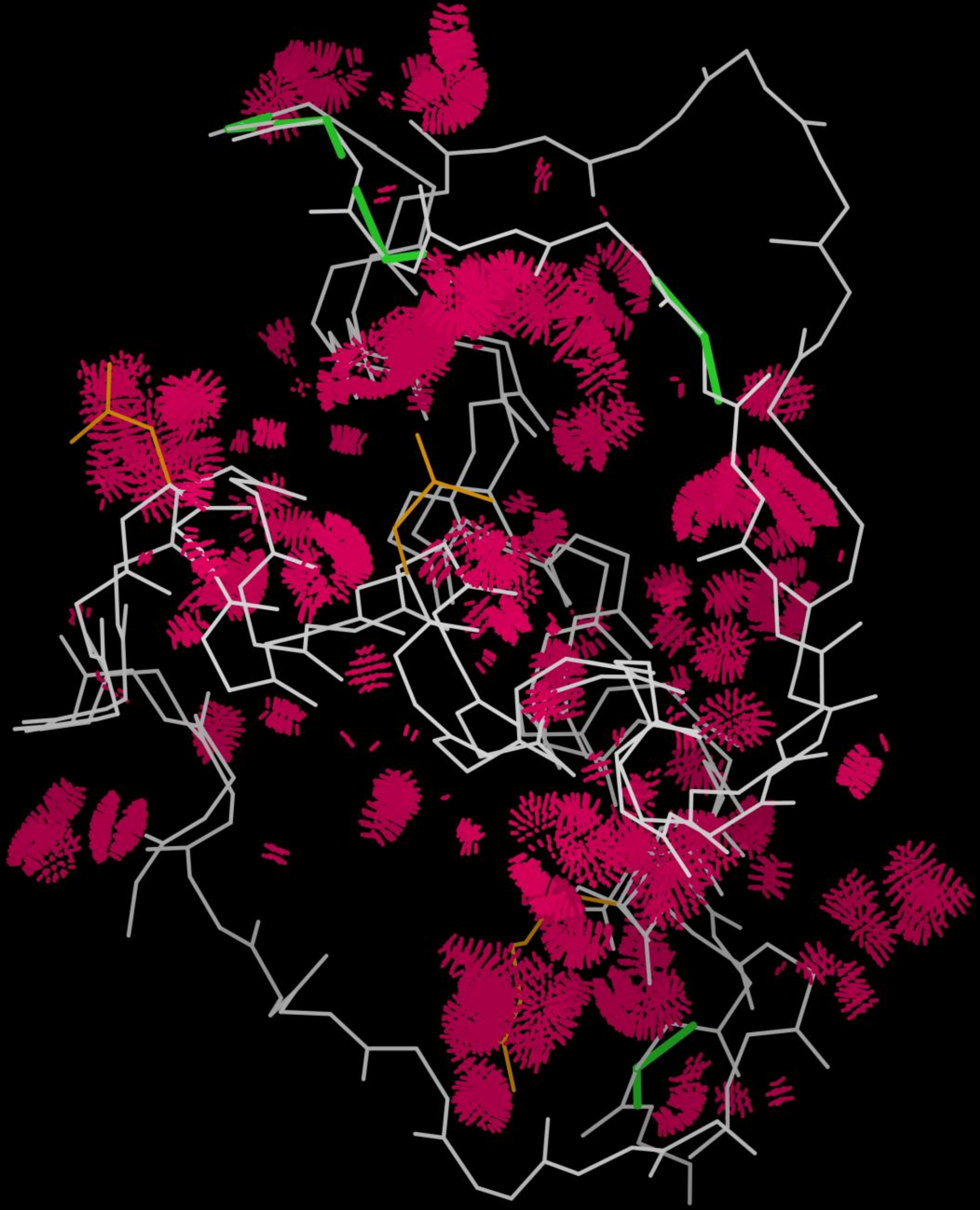
- Running MolProbity analysis on 10,000 NMR PDBs (170,000 models)
- Before condor:
  - ~240 hours over 2 weeks
- After condor:
  - 8 hours
- How do NMR and X-ray structures compare overall?


X-ray clashscore
NMR clashscore
clashscore < 3 ideal target
1200
1000
800
600
400
200
0
Score
1975
1980
1985
1990
1995
2000
2005
2010
2015
Year


X-ray % cbeta outlier
NMR % cbeta outlier
<5% fair target
Percent outlier
Year
0
20
40
60
80
100
1975
1980
1985
1990
1995
2000
2005
2010
2015

Odd PDBs

- 2 homologous domains in 1 model, superimposed
- ~280 clashscore



Conclusions for high-throughput MolProbity

- High-throughput version of MolProbity is powerful!
  - Deals with NMR ensembles
  - Allows analysis of large structural datasets
  - Allows us to test different methods of validation
- Check your structures before you use them!

Acknowledgements

Duke University

David and Jane Richardson

Bryan Arendall

Jeremy Block

Ian Davis

Lindsay Deis

Jeff Headd

Bradley Hintze

Bob Immormino

Swati Jain

Gary Kapral

Daniel Keedy

Laura Murray

Michael Prisant

Lizbeth Videau

Christopher Williams

J. Michael Word

UW Madison

John Markley

Eldon Ulrich

Miron Livny

Dimitri Maziuk

Jon Wedell

R. Kent Wenger

Hongyang Yao