

De novo Protein folding

Rosetta Workshop

August 2020

Diego del Alamo

del.alamo@vanderbilt.edu

Meiler Laboratory, Vanderbilt University

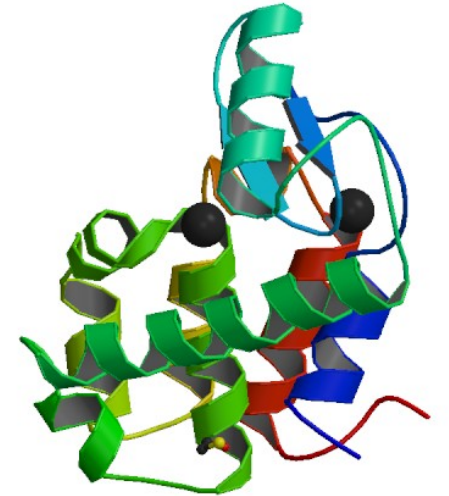
<http://www.meilerlab.org>

De novo Protein folding

- Determination of protein structure from sequence data alone
- No template structure available, unlike homology modeling
- Practical for small, simple proteins (<200 residues)

MNIFEMLRIDEGLRLKIYK
DTEGYTIGIGHLLTKSPS
LNASKSELDKAIGRNTNGV
ITKDEAEKLFNQDVDAVR
GILRNAKLKPVYDSLDAVR
RAALINMVFQMGETGVAGF
TNSLRMLQQKRWDEAAVNL
AKSRWYNQTPNRAKRVITT
FRTGTWDAYKNL

Primary Sequence



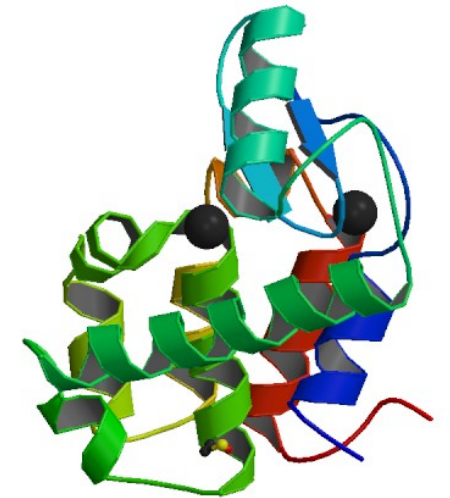
Tertiary Structure

De novo Protein folding

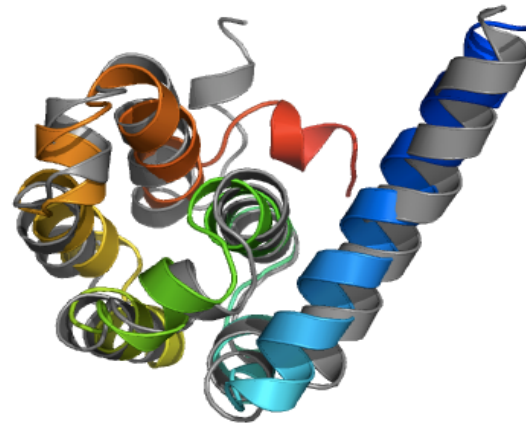
- Determination of protein structure from sequence data alone
- No template structure available, unlike homology modeling
- Practical for small, simple proteins (<200 residues)

MNIFEMLRIDEGLRLKIYK
 DTEGYTIGIGHLLTKSPS
 LNASKSELDKAIGRNTNGV
 ITKDEAEKLFNQDVDAVR
 GILRNAKLKPVYDSLDAVR
 RAALINMVFQMGETGVAGF
 TNSLRMLQQKRWDEAAVNL
 AKSRWYNQTPNRAKRVITT
 FRTGTWDAYKNL

Primary Sequence

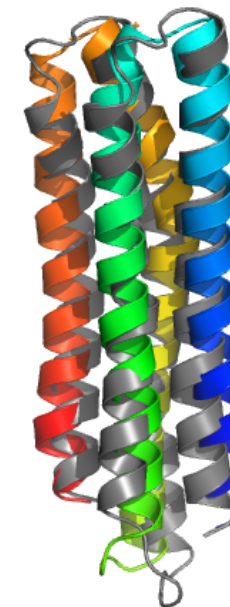


Tertiary Structure



Small, globular,
soluble proteins

**T4-lysozyme
C-terminal domain**



Small, simple
membrane proteins

**V-type Na⁺ ATP
synthase subunit**



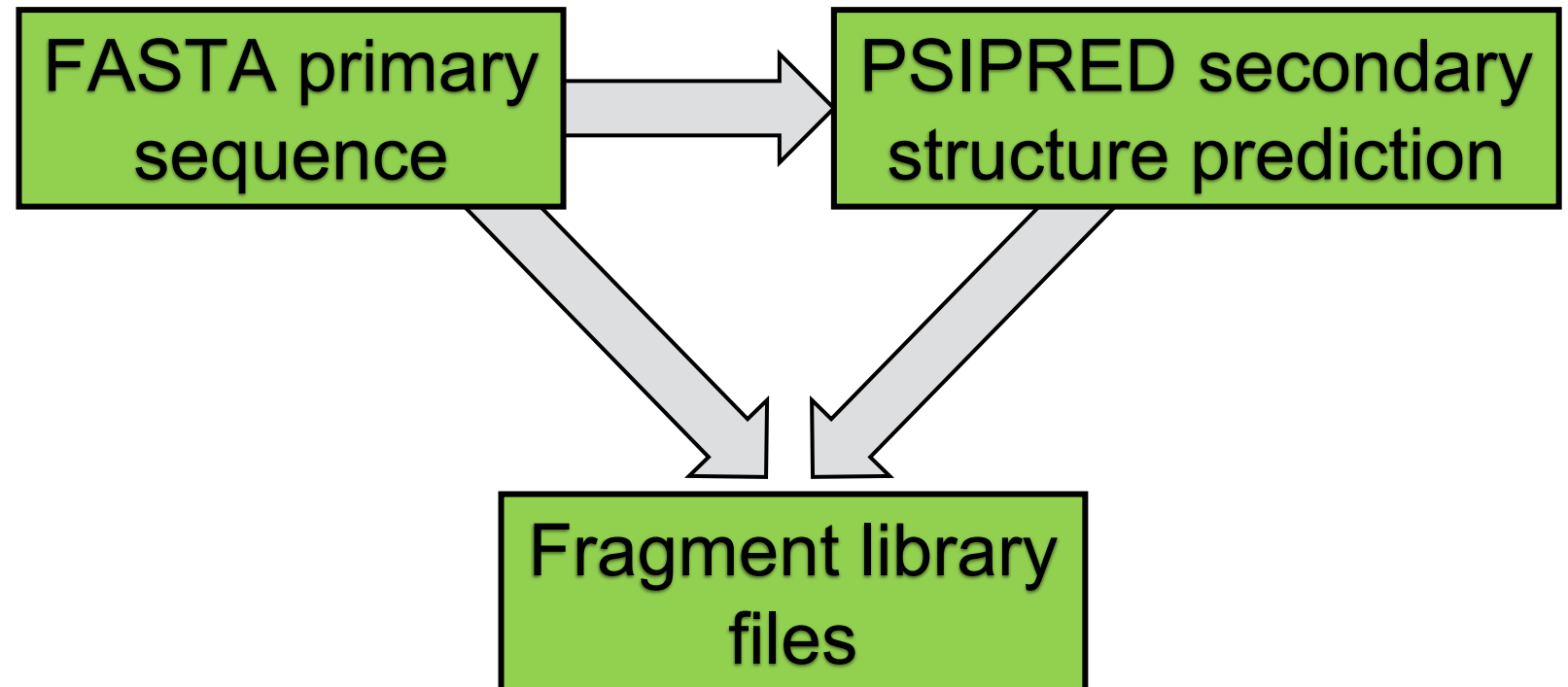
...but not large,
complex proteins

Rhodopsin

Inputs

- **Necessary:**

- Protein primary sequence (FASTA)
- Secondary structure prediction
- Fragment library files
- Rosetta options file

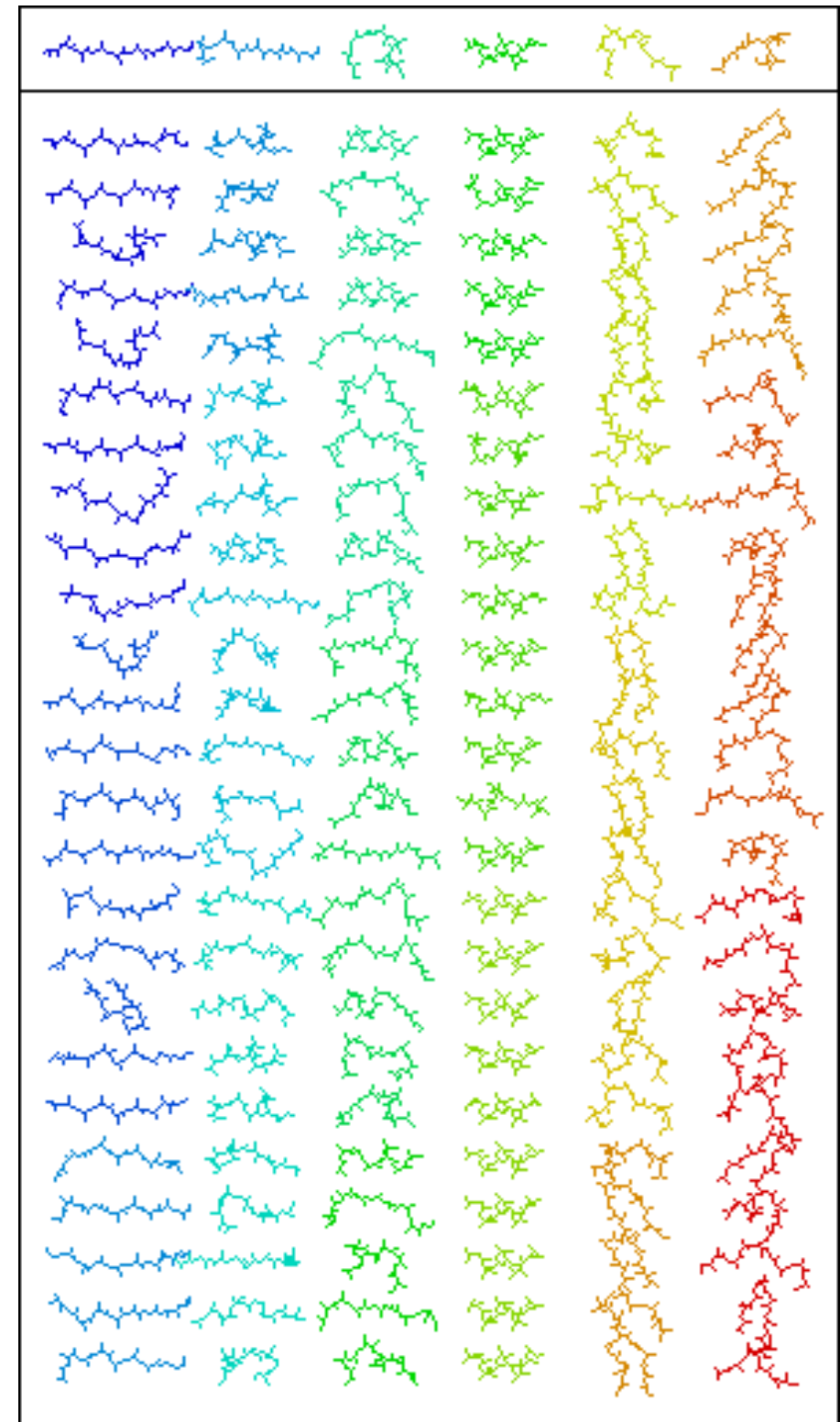


- **Optional:**

- Native structure (PDB)
- Constraints file

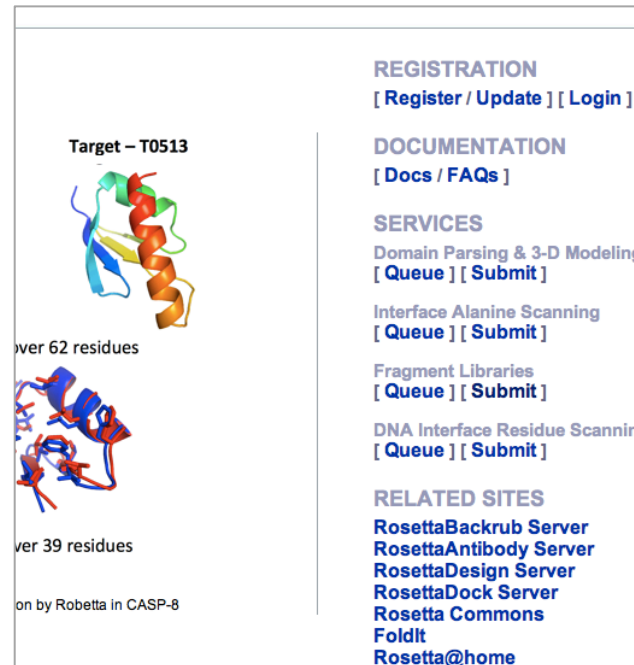
Inputs: fragment library files

- Short peptide backbone conformations generated from known protein structures in PDB (typically 3 or 9 residues long)
- Introduced to backbone during folding to change protein structure
- Fragments are picked prior to running the folding simulation
- Fragment selection can be guided by secondary structure prediction (PSIPRED, Jufo) and experimental data e.g. from NMR (chemical shifts, residual dipolar couplings)



Getting fragments: Robetta

- Online server for fragment selection, two simultaneous jobs per user
- Academic users only
- Alternatives:
 - *make_fragments.pl*
 - *fragment_picker* application
 - require setting up specific environment (discussed in the tutorial)



<http://old.robetta.org/>

ROBETTA BETA
Full-chain Protein Structure Prediction Server

Structure Prediction Fragment Libraries Alanine Scanning DNA Interface Scanning
[Queue] [Submit] [Queue] [Submit] [Queue] [Submit] [Queue] [Submit]
[Register / Update] [Docs / FAQs] [Login]

Submit a job to the Fragment Server
*Please submit one job at a time

• Identifier must be at least 5 alphanumeric characters

Required
Registered Username: or Registered Email Address:
Target Name:
Paste Fasta
> 2LZM Sequence
ITKDEAEKLFNQDVDAAVRGILRNALKPYYDSLDAVRRCALINMVFQMGETGV
AGFTNSLRMLQQKRWDEAAVNLAQSRWYNQTPNRAKRVITTFRTCTWDAYKNL
or Upload Fasta: no file selected

Optional
Identifier:
Exclude Homologues: ☐
Rosetta NMR (click links below for input format)
Chemical Shifts: no file selected
NOE Constraints: no file selected
Dipolar Constraints: no file selected

ROBETTA BETA
Full-chain Protein Structure Prediction Server

Structure Prediction Fragment Libraries Alanine Scanning DNA Interface Scanning
[Queue] [Submit] [Queue] [Submit] [Queue] [Submit] [Queue] [Submit]
[Register / Update] [Docs / FAQs] [Login]

Fragment Server Queue
0 Job(s) Queued
Username: Target: Host:

Page 1 2 3 4 5 6

ID	Status	Date (PST)	Username	Length	Target	Host
18162	Complete	02/10/11 10:48:48 AM	vji4	226	anceu	dhcp128036156
18161	Complete	02/10/11 10:44:02 AM	jamsmad	26	1GZLIEnd	titan
18160	Complete	02/10/11 10:14:27 AM	jamsmad	22	1GZLShort	titan
18159	Complete	02/10/11 09:36:01 AM	zwenhor	38	2I2V4	Farid-HP.
18158	Complete	02/10/11 09:15:17 AM	zwenhor	41	1K1V	Farid-HP.
18157	Complete	02/10/11 09:11:09 AM	zwenhor	41	1K1V	Farid-HP.
18156	Complete	02/10/11 08:31:07 AM	jamsmad	26	1GZLFront	titan
18155	Complete	02/10/11 08:30:07 AM	jamsmad	26	1GZLAdd	titan
18154	Complete	02/10/11 07:51:36 AM	jamsmad	24	1GZLAdd	titan
18153	Complete	02/10/11 07:12:33 AM	jamsmad	24	1GZLDel	titan
18152	Complete	02/10/11 04:32:04 AM	Only Dym	441	PAN	wisweb2-out.v
18151	Complete	02/09/11 08:03:47 PM	maruti	56	GB1	142.1
18150	Complete	02/09/11 09:27:59 AM	drx	176	f9	128.2
18149	Complete	02/09/11 08:35:47 AM	gise	126	Nav beta-2 extra	139.1
18148	Complete	02/09/11 08:33:55 AM	zwenhor	208	1EOG	129.1

Options for *AbInitioRelax*

options.txt (1/2)

-in

-file

-native <native PDB file>

native PDB file (optional)

-fasta <primary sequence in FASTA format>

or command line: -in:file:fasta

-frag3 <3mer fragment file>

protein 3-residue fragments file

-frag9 <9mer fragment file>

```
# protein 9-residue fragments file
```

```
-psipred_ss2 <PSIPRED secondary structure prediction file>
```

```
# psipred_ss2 secondary structure
definition file (required for -use_filters)
```

-broker

```
-setup <topology broker file>
```

```
# topology broker file
```

-abinitio

```
-increase_cycles 10
```

Increase the number of cycles at each stage in AbinitioRelax by this factor

```
-rg reweight 0.5
```

Reweight contribution of radius of gyration to total score by this factor

-rsd wt helix 0.5

Reweight env, pair, and cb scores for helix residues by this factor

```
-rsd wt loop 0.5
```

Reweight env, pair, and cb scores for loop residues by this factor

-relax

At the end of de novo folding, do a relax step

```
-use filters true
```

Discontinues refinement if a filter fails

-relax

-fast

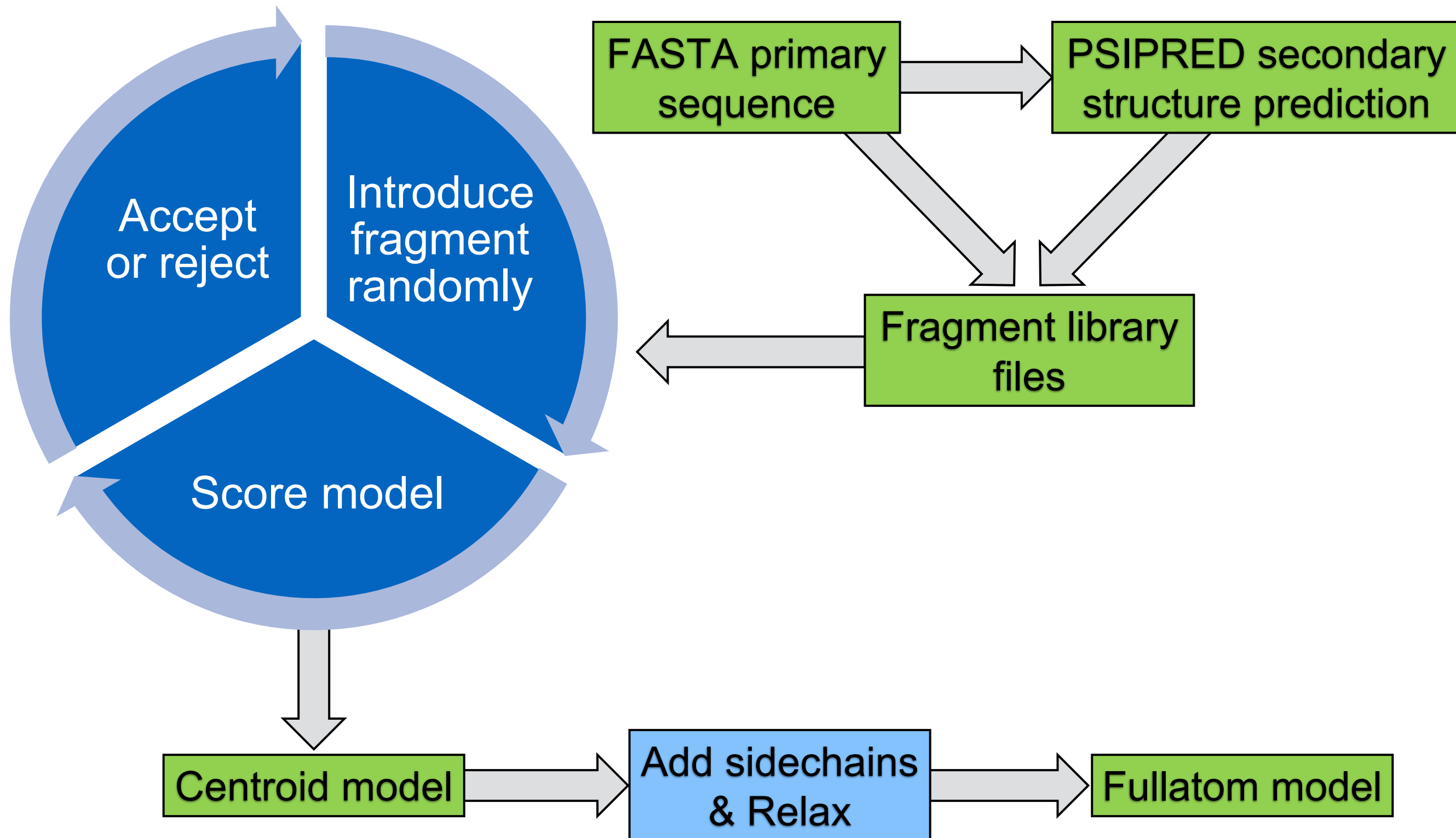
Type of relax protocol. This has been shown to be the best deal for speed and robustness.

Options for *AbInitioRelax*

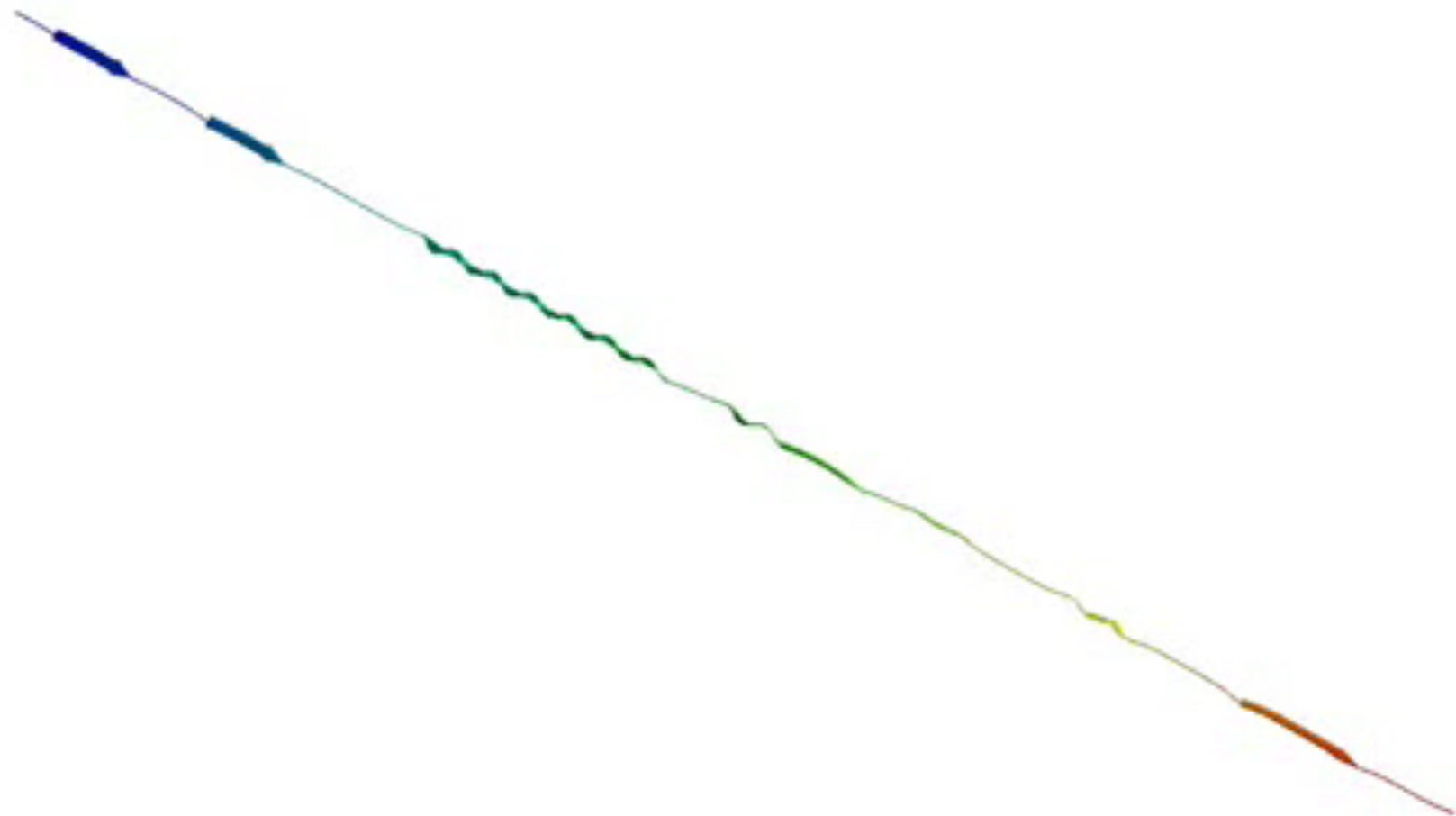
options.txt (2/2)

```
-run
    -reinitialize_mover_for_each_job # Job distributor generates fresh copy of its mover
                                     before each apply (once per job)
-protocol broker # run the topology_broker protocol for de novo folding
-score
    -find_neighbors_3dgrid # Use a 3D lookup table for doing neighbor calculations. For
                           spherical, well-distributed conformations
-evaluation
    -rmsd <file to compute RMSD against> <column name> <file defining residues over which to compute RMSD>
        # compute CA-RMSD for model comparing to native structure
-out
    -output # use this to tell Rosetta you actually want output
    -nstruct 1 # how many structures do you want to generate?
    -file
        -silent <silent output file> # full path to silent file output
        -silent_struct_type binary # we want binary silent files
        -scorefile <scorefile> # full path to scorefile
-overwrite # overwrite any existing output with the same name you may
           have generated
```

Running *AbInitioRelax*



Running *AbInitioRelax*



Adding experimental data

Constraint info

Constraint Function info

	<cst_type>	<atom1>	<res1>	<atom2>	<res2>	<cst_func>	<spline_type>	<Dcb>	<weight>	<bin>
EPR	AtomPair	CB	32	CB	36	SPLINE	EPR_DISTANCE	16.0	1.0	0.5
	AtomPair	CB	59	CB	74	SPLINE	EPR_DISTANCE	19.0	1.0	0.5
	AtomPair	CB	62	CB	71	SPLINE	EPR_DISTANCE	19.0	1.0	0.5
	<cst_type>	<atom1>	<res1>	<atom2>	<res2>	<cst_func>	<lb>	<ub>	<sd>	<tag>
NMR	AtomPair	HB	32	HG	36	BOUNDED	11.5	21.5	1.0	NOE
	AtomPair	H	59	HA	74	BOUNDED	11.5	21.5	1.0	NOE
	AtomPair	H	62	HB	71	BOUNDED	11.5	21.5	1.0	NOE

- Constraint type (AtomPair, Angle, Dihedral, etc.)
- Constraint function (Harmonic, Bounded, Gaussian, Spline, etc.)

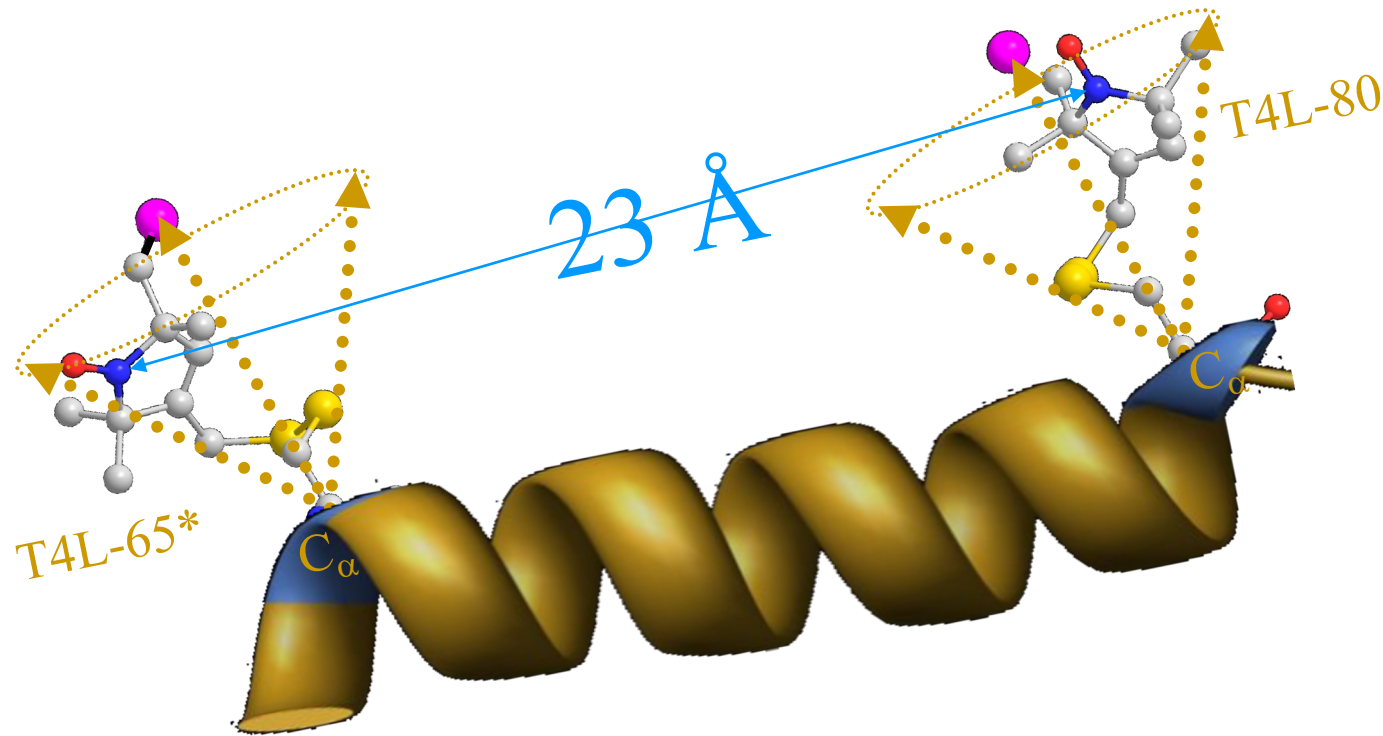
options_cst.txt

```

-fold_cst                # use FoldConstraints protocol
    -force_minimize      # minimize in FoldConstraints protocol
-constraints
    -cst_file <file>     # path to your cst file
    -cst_weight 4         # factor by which to multiply cst score
    -cst_fa_file <file>  # path to your cst file
    -cst_fa_weight 4     # factor by which cst score multiplied by
    -epr_distance        # Use RosettaEPR knowledge-based potential

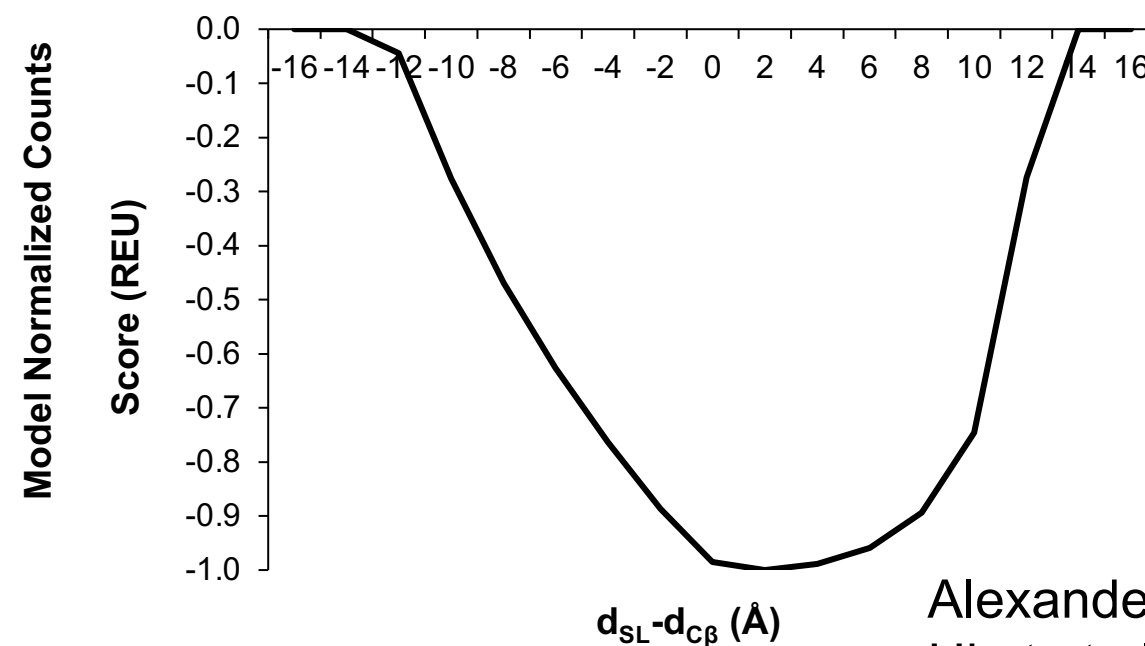
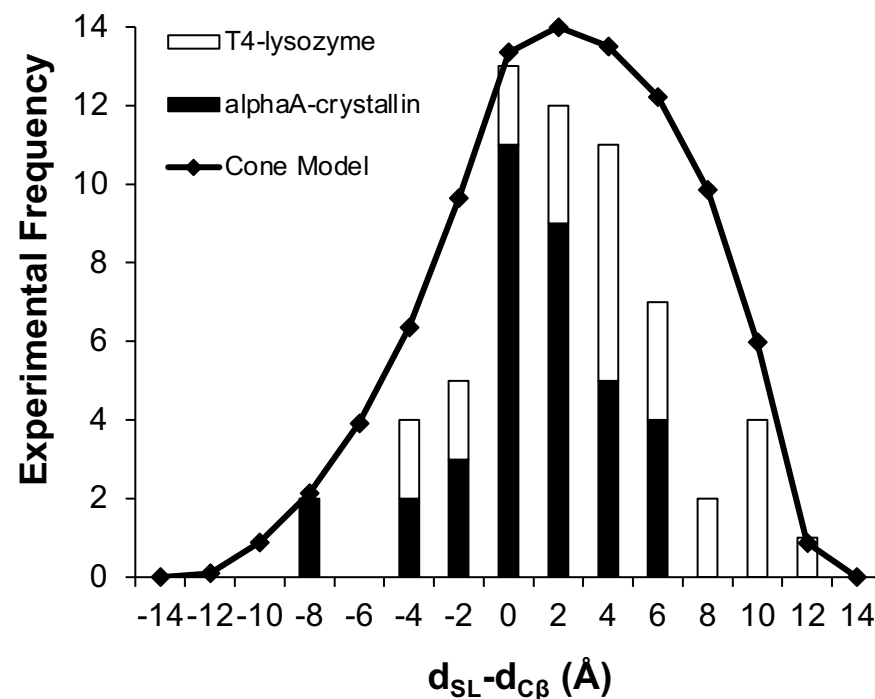
```

RosettaEPR



- Rosetta uses a knowledge-based potential to evaluate how a protein model agrees with the EPR distance data

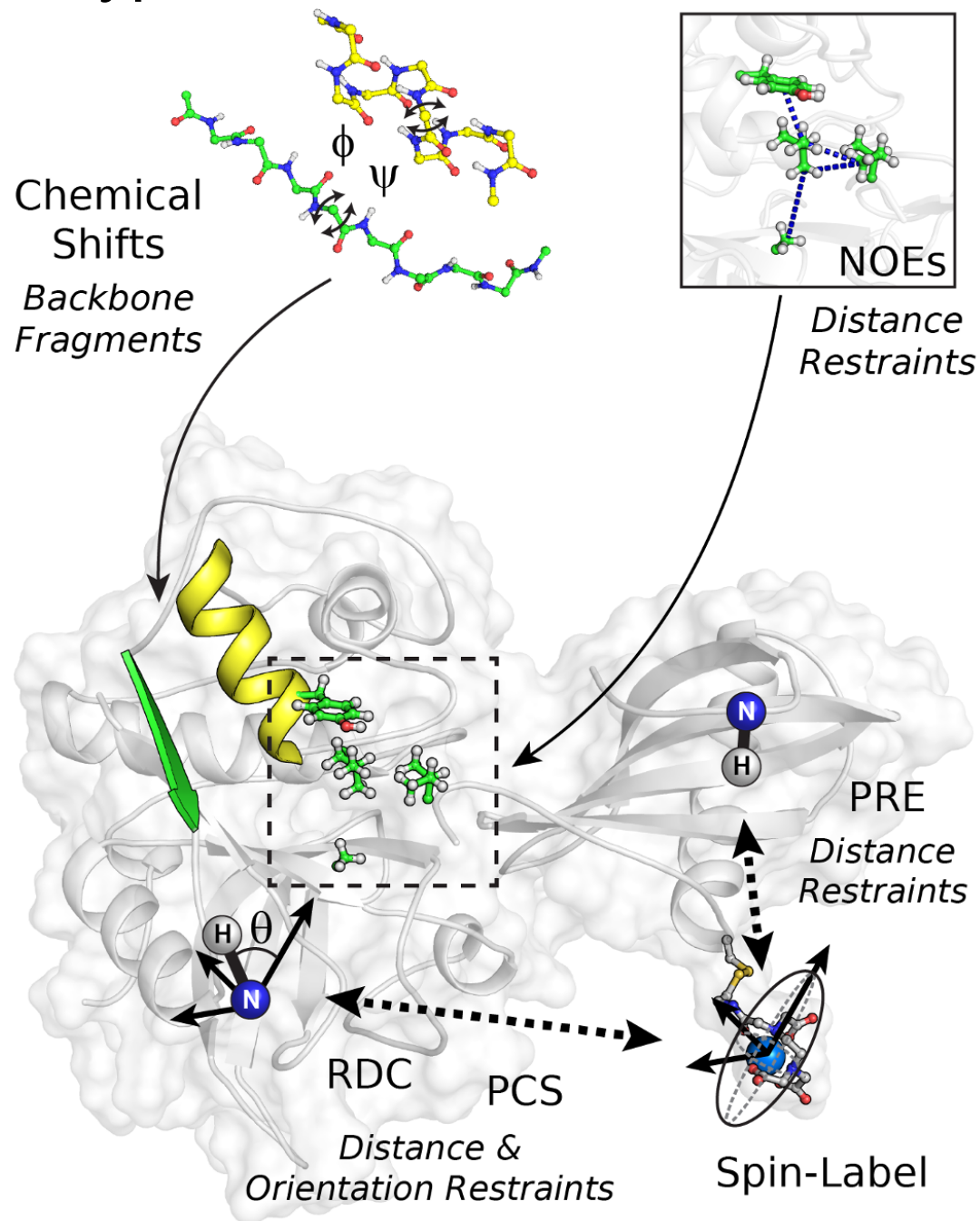
Borbat, Mchaourab, Freed, *JACS*, 2002



Alexander et al., *Structure*, 2008
Hirst et al., *J Struct Biol*, 2011

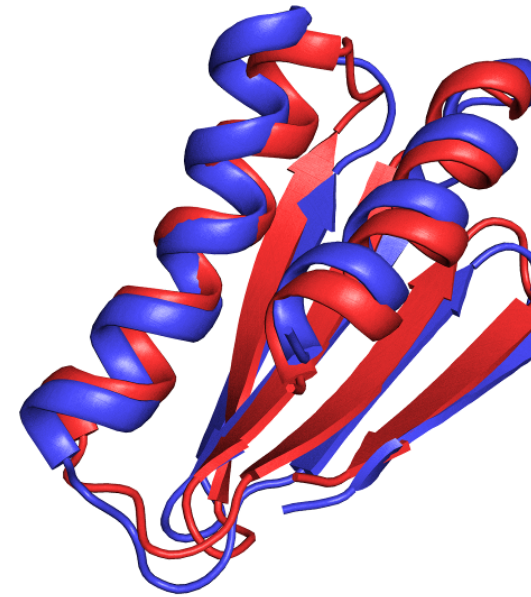
RosettaNMR

- Rosetta allows using different types of NMR data

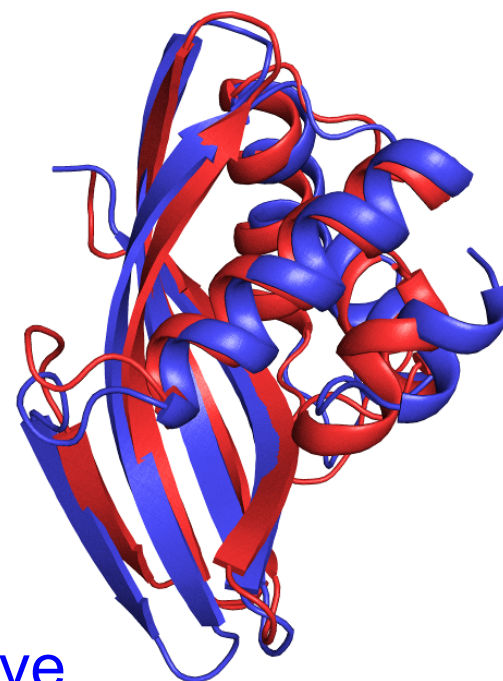
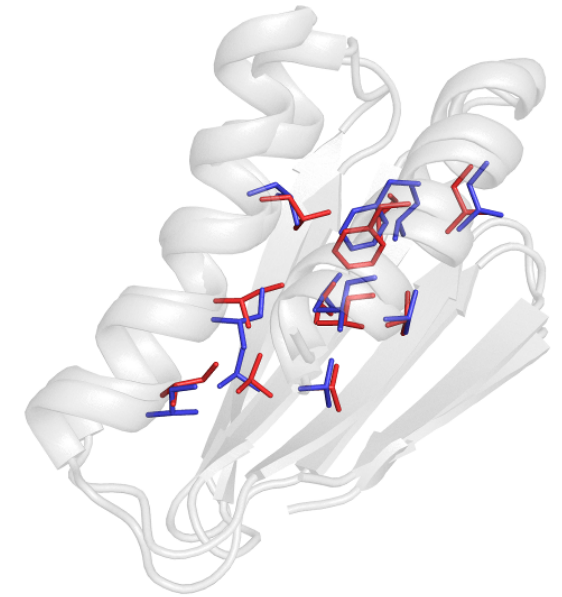


Kuenze et al., *Structure*, 2019

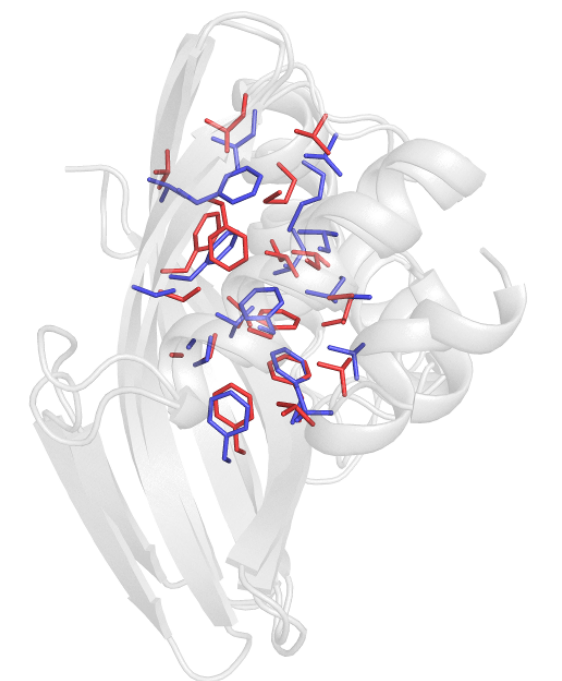
- Meiler lab submissions in CASP13



N1008 (C α -Rmsd = 2.1 Å)



N0968s1 (C α -Rmsd = 2.9 Å)



Native
Rosetta

Evolutionary coupling constraints

- Residue contact constraints from coevolution data

GREMLIN.BAKERLAB.org Powered by OPENSEQ.org

Home **GREMLIN** Complexes PFAM New Projects Contact Us

Submit Sequence for Coevolution Analysis

Note, due to limited resources your submission may take forever to complete (Jobs Running: 47).



Email (Will only be used to send you notice of Completion/Error).

Job name

Starting alignment
Alignment must be in FASTA format. Alternatively, you may provide a single sequence in plain text or in FASTA format.

No file chosen

NOTE: We treat your input as a "starting alignment" which is then enriched using HHblits. If Jackhmmer is selected only the first (or query) sequence is used. [20 < query_length < 1000] [contact us](#) if you wish to submit a sequence larger than the limit.

NEW: We have a separate [submission page](#) for protein-protein coevolution analysis.

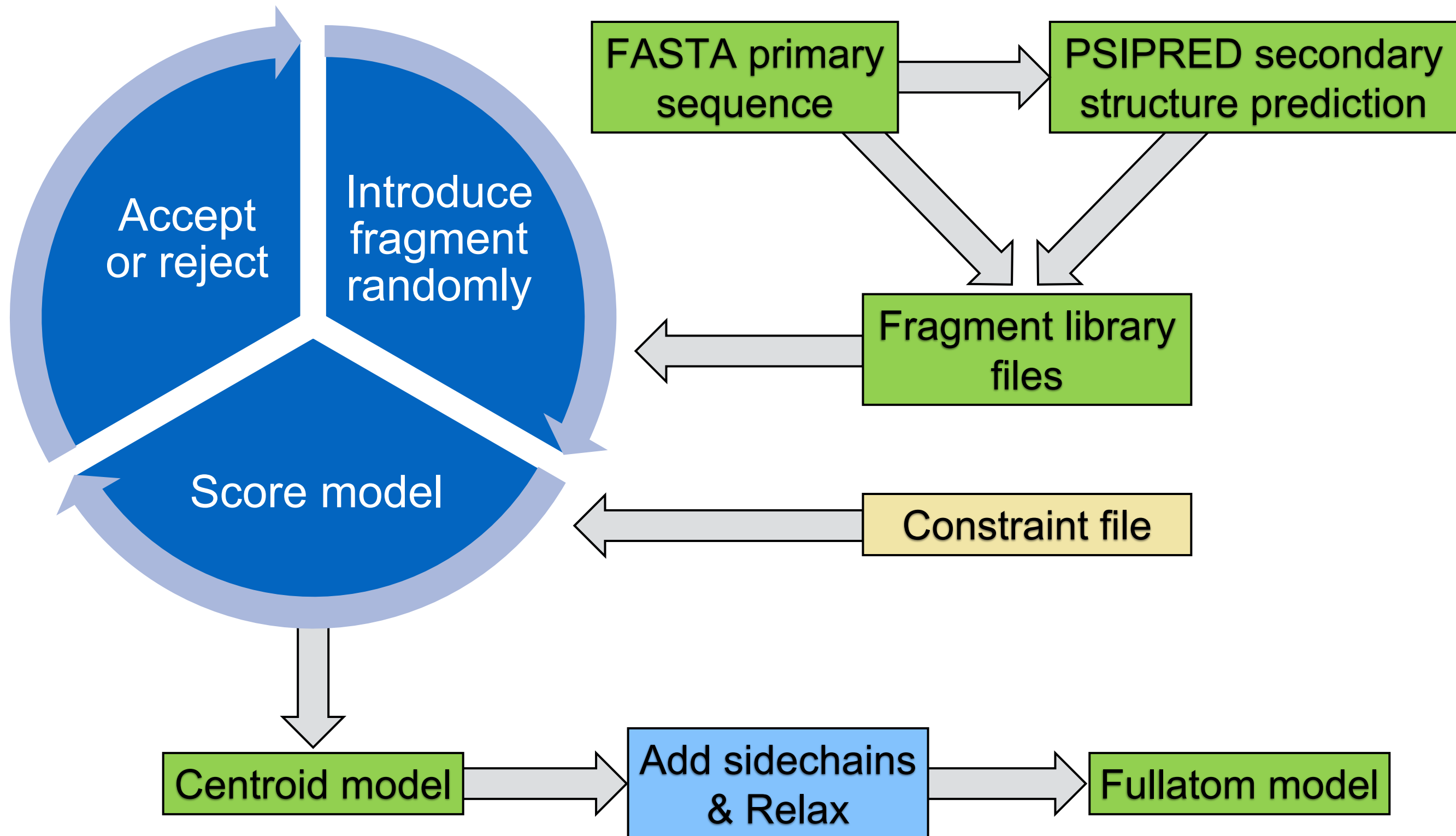
Options

Generate MSA using: HHblits	Filter MSA	GREMLIN
E-value: 1E-10	Iterations: 4	Coverage: 75
	Remove gaps: 75	Prior: Vanilla

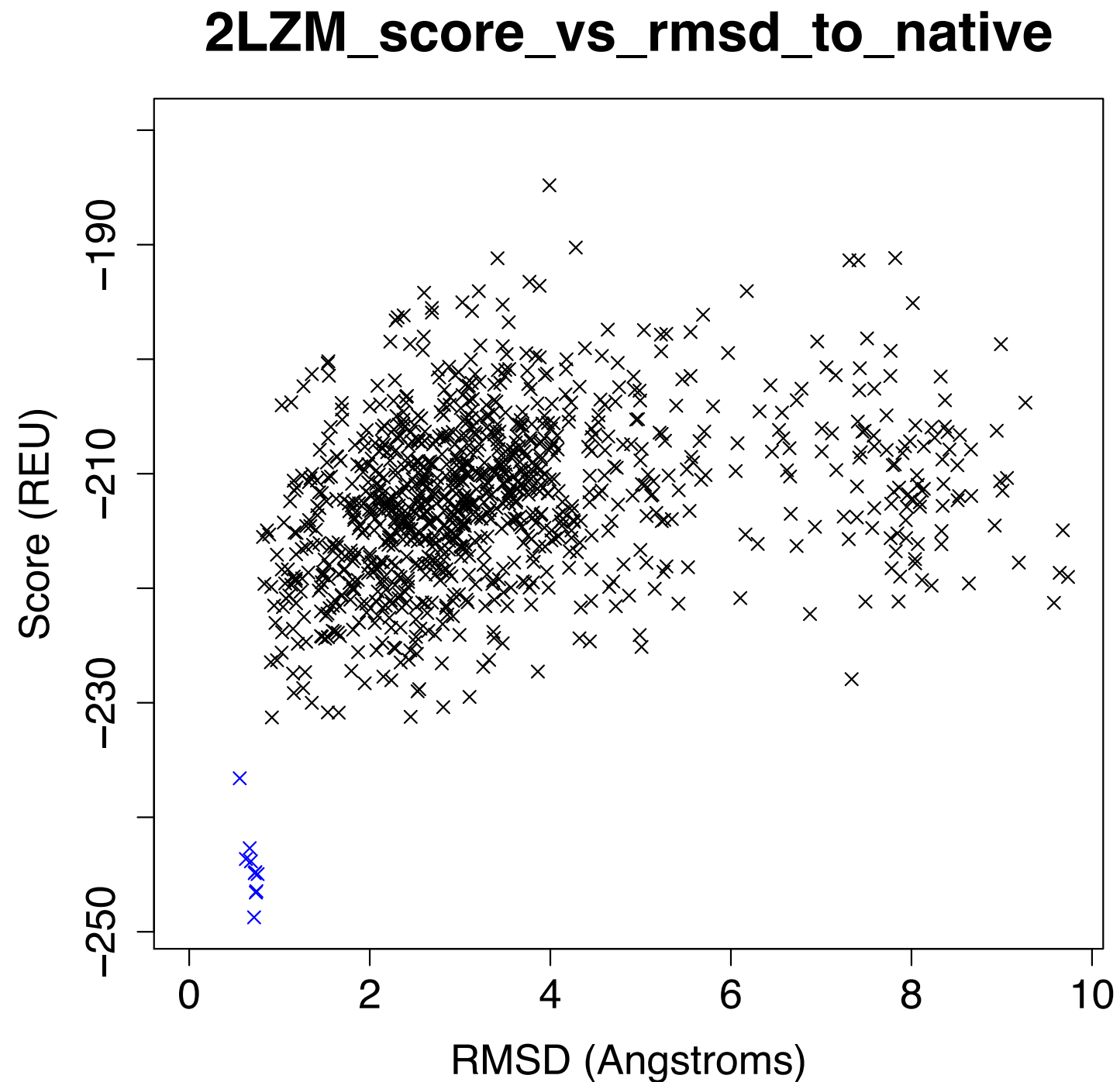
- HHblits/Jackhmmer**
 - HHblits and Jackhmmer are methods used to generate the Multiple Sequence Alignment (MSA).
 - <E-value/Iterations> options are used to control the diversity of MSA generation.
 - To use your own alignment, without MSA enrichment, set Iterations to 0.
- Filter MSA**
 - <Coverage> Filter MSA to remove sequences that don't cover at least (%) of query.
 - <Remove gaps> After coverage filter, remove positions in the alignment that have (%) gaps.
- GREMLIN**
 - only **Vanilia** works right now, other options are under construction.
 - Vanilla - Pure sequence based analysis (no prior).
 - SS - Add Secondary Structure prior, based on PSIPRED predictions.
 - SVM - Add SVM-CON prior.

1 + 2 = ☒ List results for others to see.

Running *AbInitioRelax* + Constraints

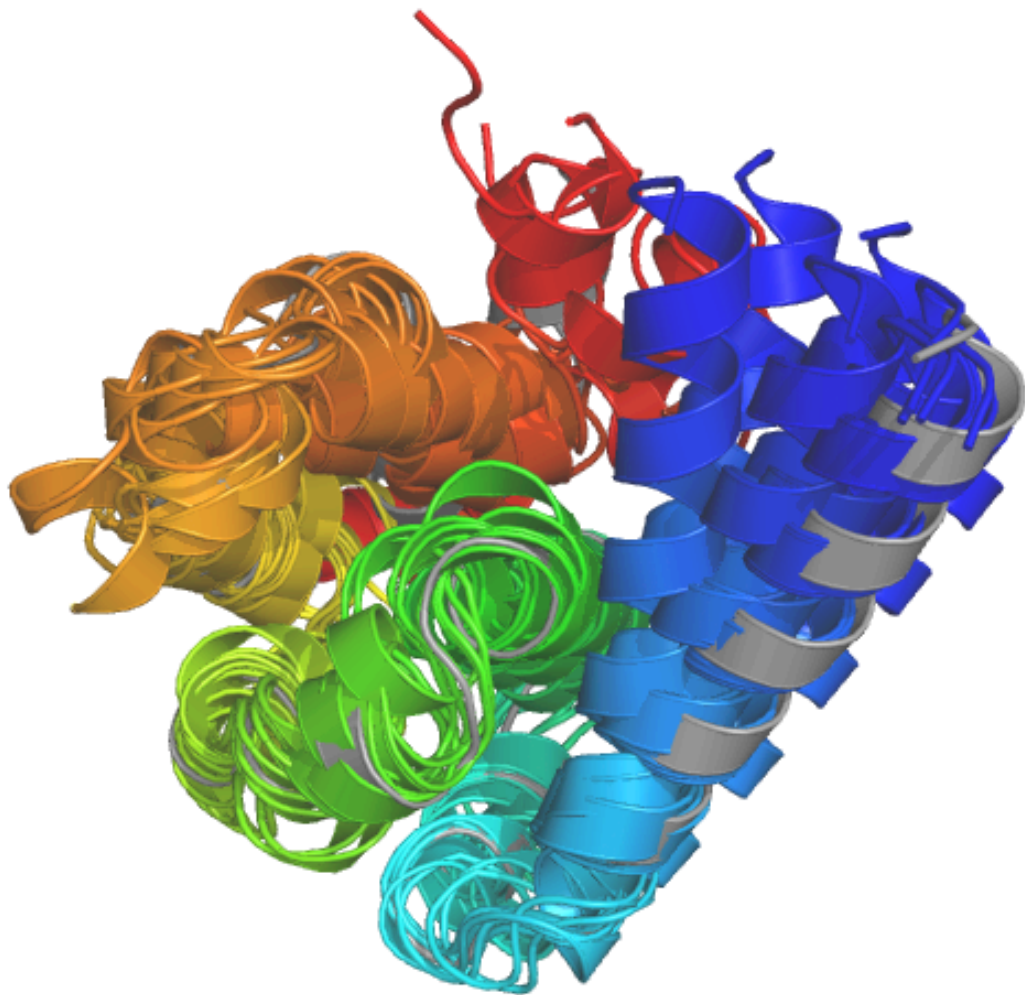


Evaluating model quality

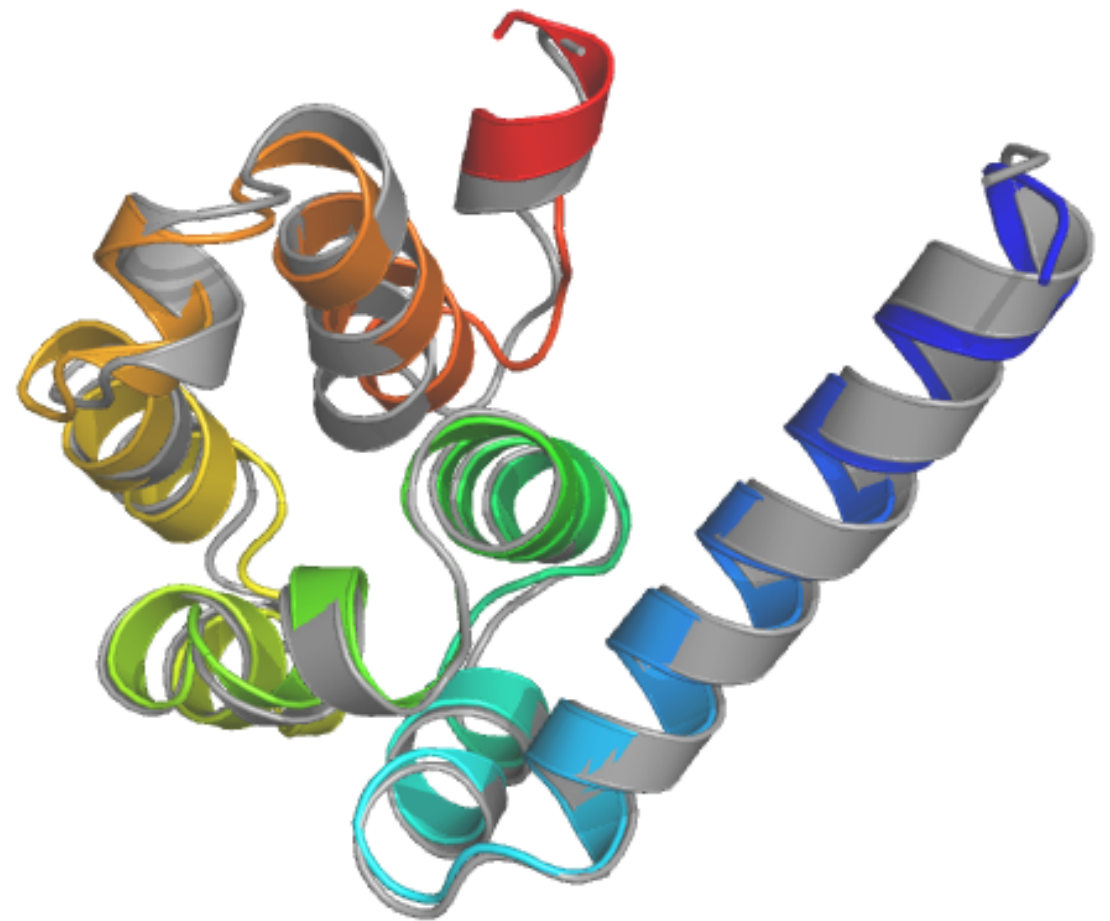


Evaluating model quality

Run: `pymol *.pdb`



top 10 scoring models



top scoring model

Conclusions & Resources

- *AbInitioRelax* is useful for predicting small single-domain proteins
- Experimental data can improve prediction outcomes
- Further Reading
 - *De novo* folding
Simons et al 1997, *JMB*, 268:209-225
Rohl et al., 2004, *Methods Enzymol*, 383:66-93
 - RosettaEPR
Alexander et al., 2008, *Structure*, 16:181-195
Hirst et al., 2011, *J Struct Biol*, 173:506-514
 - RosettaNMR
Meiler & Baker, 2003, *Proc Natl Acad Sci U S A.*, 100:12105-12110
Nerli & Sgourakis, 2019, *Methods Enzymol*, 614:321-362
Kuenze et al., 2019, *Structure*, 27:1721-1734
 - Fragment Picker
Gront et al., 2011, *PLoS ONE*, 6(8): e23294
https://www.rosettacommons.org/docs/latest/application_documentation/utilities/app-fragment-picker