Biologics Modelling
Proteins & Peptides

Annette Höglund, Jianxin Duan

Helsingfors

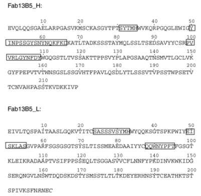
November 12, 2015

BioLuminate Features

- Protein-protein docking
- Antibody structure prediction from sequence
- Antibody humanization
- Fast homology model generation
- Accurate long loop predictions
- Residue scanning
- Affinity Maturation
- Cysteine scanning
- Crosslink design
- Peptide QSAR
- Aggregation hot spot ID
- Free energy perturbation

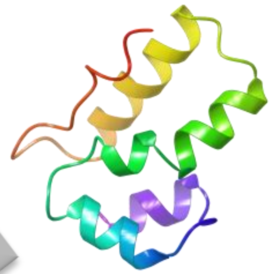
Using BioLuminate to Go From Sequence to Model of Antibody/Antigen Complex

- Starting point:
  - Sequence of antibody
    - FAB13B5

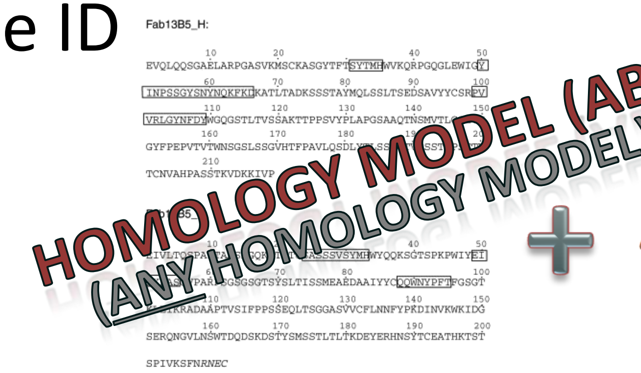


```
Fab13B5_H:

          10        20                  40        50
EVQLQQSGAELARPGASVKMSCKASGYTFTSYTMHWVKQRPGQGLEWIGY
         60        70        80        90       100
INPSSGYSNYNQKFKDKATLTADKSSSTAYMQLSSLTSEDSAVYYCSRPV
        110       120       130       140       150
VRLGYNFDYWGQGSTLTVSSAKTTPPSVYPLAPGSAAQTNSMVTLGCLVK
        160       170       180       190       200
GYFPEPVTVTWNSGSLSSGVHTFPAVLQSDLYTLSSSVTVPSSTWPSETV
        210
TCNVAHPASSTKVDKKIVP

Fab13B5_L:

          10        20                  40        50
EIVLTQSPAITAASLGQKVTITCSASSSVSYMHWYQQKSGTSPKPWIYEI
         60        70        80                 100
SKLASGVPARFSGSGSGTSYSLTISSMEAEDAAIYYCQQWNYPFTFGSGT
        110       120       130       140       150
KLEIKRADAAPTVSIFPPSSEQLTSGGASVVCFLNNFYPKDINVKWKIDG
        160       170       180       190       200
SERQNGVLNSWTDQDSKDSTYSMSSTLTLTKDEYERHNSYTCEATHKTST

SPIVKSFNRNEC
```

  - Crystal structure of known antigen (unbound)
    - HIV-1 Capsid Protein (P24) (Dimerization Domain)
    - 1A43

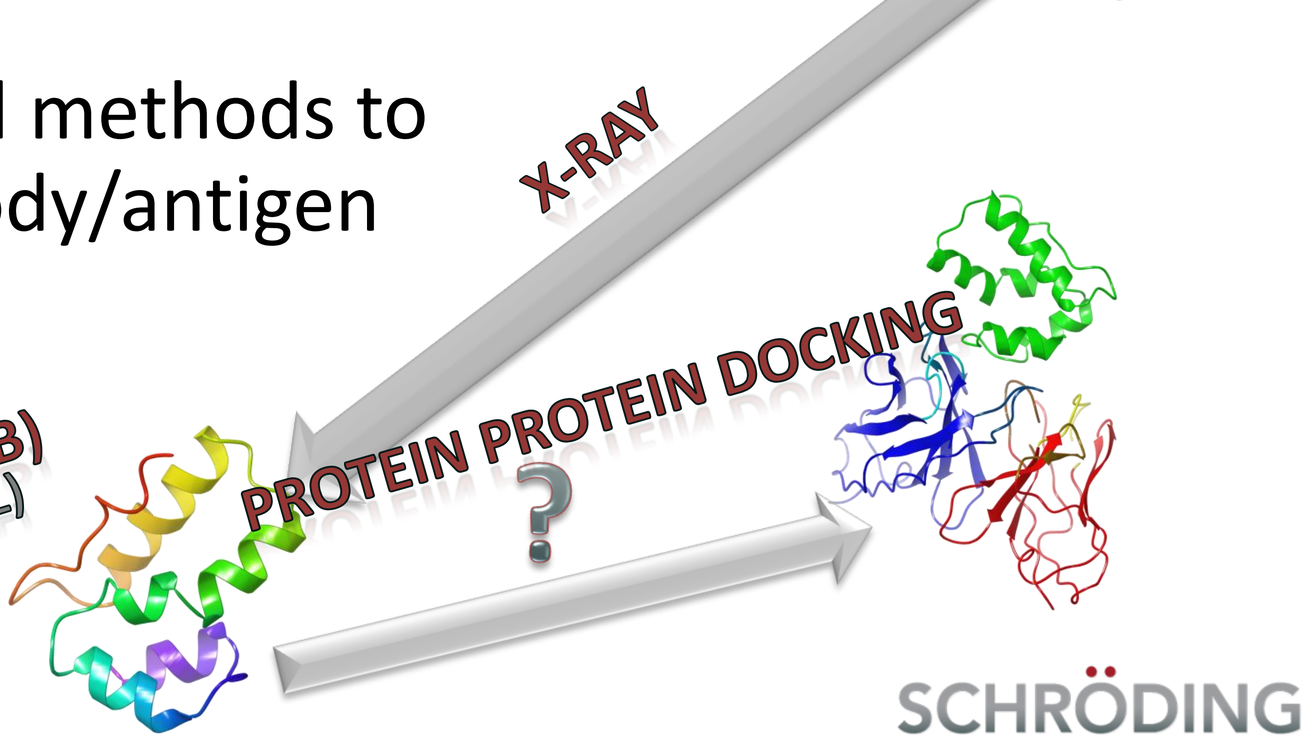


- Can we use computational methods to predict structure of antibody/antigen complex?
  - Epitope ID



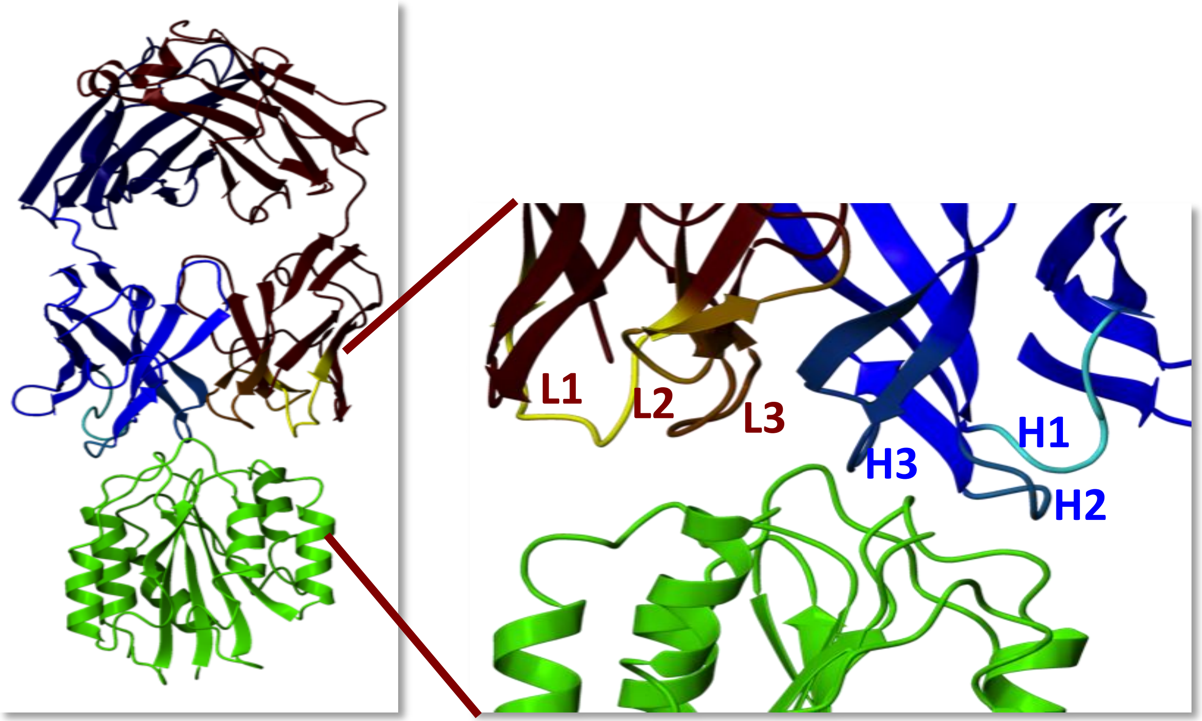
HOMOLOGY MODEL (AB)
(ANY HOMOLOGY MODEL)
+

```
Fab13B5_H:

          10        20                  40        50
EVQLQQSGAELARPGASVKMSCKASGYTFTSYTMHWVKQRPGQGLEWIGY
         60        70        80        90       100
INPSSGYSNYNQKFKDKATLTADKSSSTAYMQLSSLTSEDSAVYYCSRPV
        110       120       130       140       150
VRLGYNFDYWGQGSTLTVSSAKTTPPSVYPLAPGSAAQTNSMVTLGCLVK
        160       170       180       190       200
GYFPEPVTVTWNSGSLSSGVHTFPAVLQSDLYTLSSSVTVPSSTWPSETV
        210
TCNVAHPASSTKVDKKIVP

Fab13B5_L:

          10        20                  40        50
EIVLTQSPAITAASLGQKVTITCSASSSVSYMHWYQQKSGTSPKPWIYEI
         60        70        80                 100
SKLASGVPARFSGSGSGTSYSLTISSMEAEDAAIYYCQQWNYPFTFGSGT
        110       120       130       140       150
KLEIKRADAAPTVSIFPPSSEQLTSGGASVVCFLNNFYPKDINVKWKIDG
        160       170       180       190       200
SERQNGVLNSWTDQDSKDSTYSMSSTLTLTKDEYERHNSYTCEATHKTST

SPIVKSFNRNEC
```



Antibody Modeling Using BioLuminate…

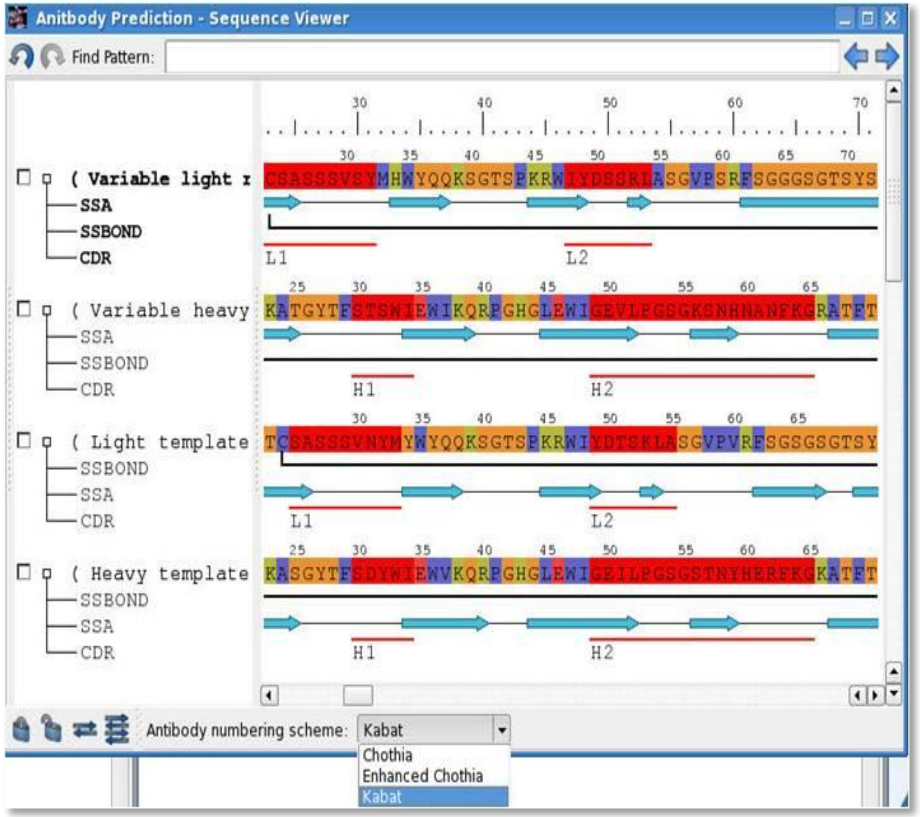
Workspace:



Recognizes & colors chains and CDRs

Antibody Aware Environment

Sequence Viewer:



Anitbody Prediction - Sequence Viewer

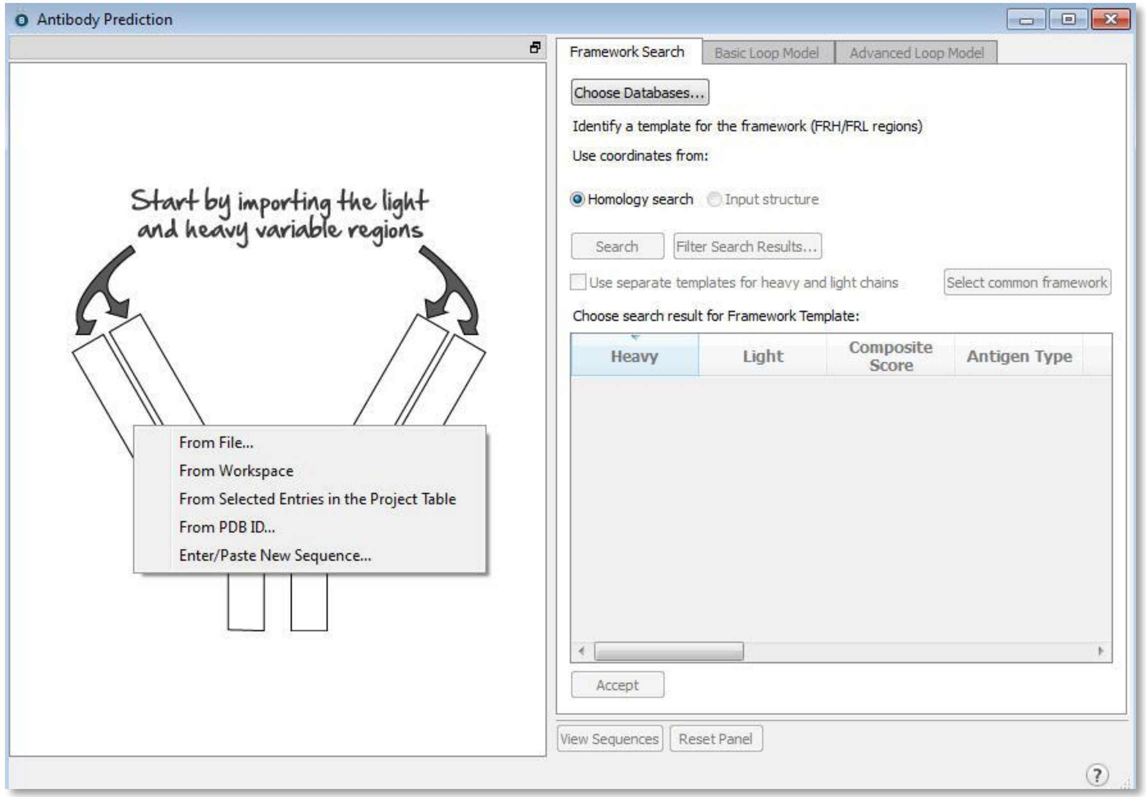
Find Pattern:

```
                        30        40        50        60        70
                     30   35   40   45   50   55   60   65   70
□ ( Variable light r  CSASSSVSYMHWYQQKSGTSPKRWIYDSSRLASGVPSRFSGGGSGTSYS
    SSA
    SSBOND
    CDR               L1                    L2
                     25   30   35   40   45   50   55   60   65
□ ( Variable heavy    KATGYTFSTSWIEWIKQRPGHGLEWIGEVLPGSGKSNHNANFKGRATFT
    SSA
    SSBOND
    CDR                     H1                   H2
                        30   35   40   45   50   55   60   65
□ ( Light template    TCSASSSVNYMYWYQQKSGTSPKRWIYDTSKLASGVPVRFSGSGSGTSY
    SSBOND
    SSA
    CDR                 L1                      L2
                     25   30   35   40   45   50   55   60   65
□ ( Heavy template    KASGYTFSDYWIEWVKQRPGHGLEWIGEILPGSGSTNYHERFKGKATFT
    SSBOND
    SSA
    CDR                     H1                   H2
```

Antibody numbering scheme: Kabat

- Detects and labels CDR loops
- Multiple numbering schemes

Antibody Modeling – CDR Prediction: Input & Framework



Identify a template for the framework (FRH/FRL regions)

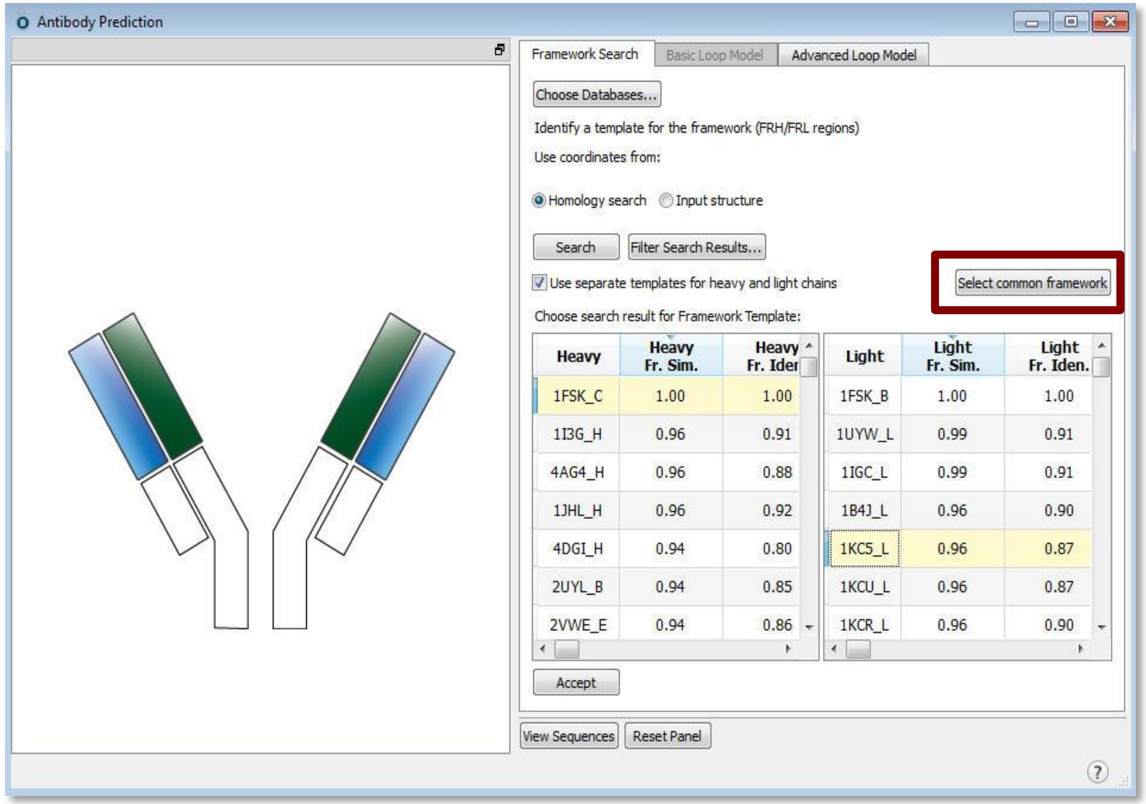
Use coordinates from:

● Homology search ○ Input structure

☐ Use separate templates for heavy and light chains

Choose search result for Framework Template:

Heavy	Light	Composite Score	Antigen Type



Identify a template for the framework (FRH/FRL regions)

Use coordinates from:

● Homology search ○ Input structure

☑ Use separate templates for heavy and light chains

Choose search result for Framework Template:

Heavy	Heavy Fr. Sim.	Heavy Fr. Ider
1FSK_C	1.00	1.00
1I3G_H	0.96	0.91
4AG4_H	0.96	0.88
1JHL_H	0.96	0.92
4DGI_H	0.94	0.80
2UYL_B	0.94	0.85
2VWE_E	0.94	0.86

Light	Light Fr. Sim.	Light Fr. Iden.
1FSK_B	1.00	1.00
1UYW_L	0.99	0.91
1IGC_L	0.99	0.91
1B4J_L	0.96	0.90
1KC5_L	0.96	0.87
1KCU_L	0.96	0.87
1KCR_L	0.96	0.90

- Antibody-specific workflow
- Search public or in-house structures for templates

- Framework selection:
  - Separate control over L/H chain framework templates
  - Control over framework used to align chains

H3 Antibody Loop Prediction Remains Difficult Using Homology Methods

Program	<L1>	<L2>	<L3>	<H1>	<H2>	<H3>	# Best	# Worst
Accelrys	1.2	0.7	1.4	1.1	1.6	3.0	1	3
CCG/MOE	0.7	0.5	1.5	1.3	1.1	3.6	1	1
PIGS server	1.0	0.4	1.4	1.1	0.8	3.2	3	0
Rosetta	1.0	0.5	1.6	1.7	1.1	3.3	0	2
BioLuminate	**1.0**	**0.5**	**1.2**	**1.1**	**1.1**	**2.2**	**3**	**0**

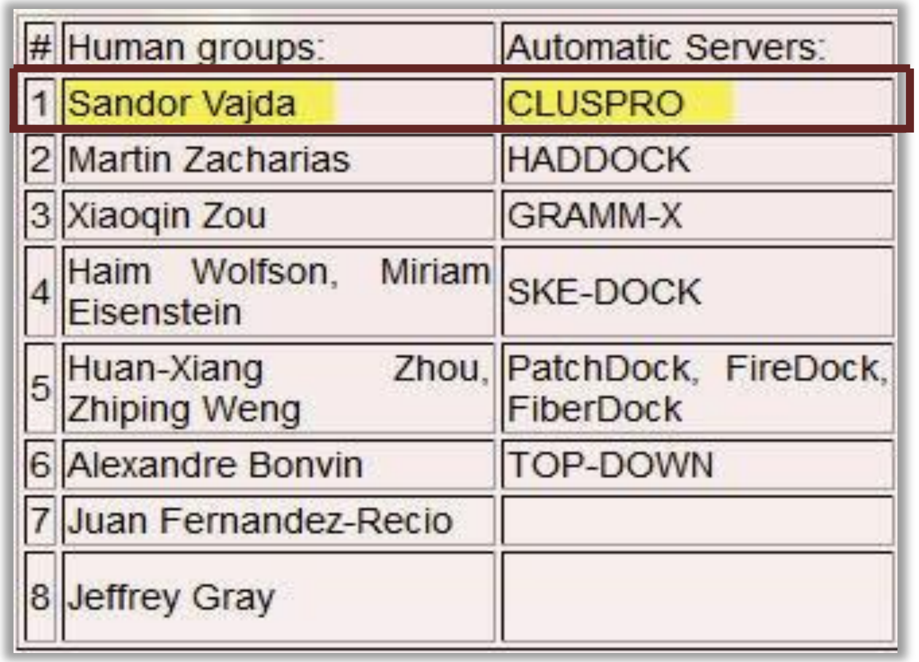
Red - Best average RMSD for loop

Gray - Worst average RMSD for loop

→ Use *de novo* approach to predict H3

Protein-Protein Docking: How do Two Proteins Best Fit Together?

- Licensed from Vajda group at Boston University
  - – Kozakov et al. (2006) *Proteins: Struct, Funct, Bioinf* **65** 392-406
- #1 server in most recent CAPRI competition
  - – Competitive with human groups

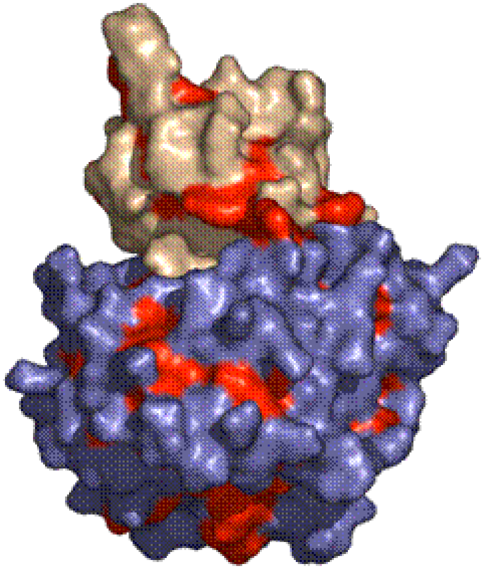


#	Human groups:	Automatic Servers:
1	Sandor Vajda	CLUSPRO
2	Martin Zacharias	HADDOCK
3	Xiaoqin Zou	GRAMM-X
4	Haim Wolfson, Miriam Eisenstein	SKE-DOCK
5	Huan-Xiang Zhou, Zhiping Weng	PatchDock, FireDock, FiberDock
6	Alexandre Bonvin	TOP-DOWN
7	Juan Fernandez-Recio	
8	Jeffrey Gray	

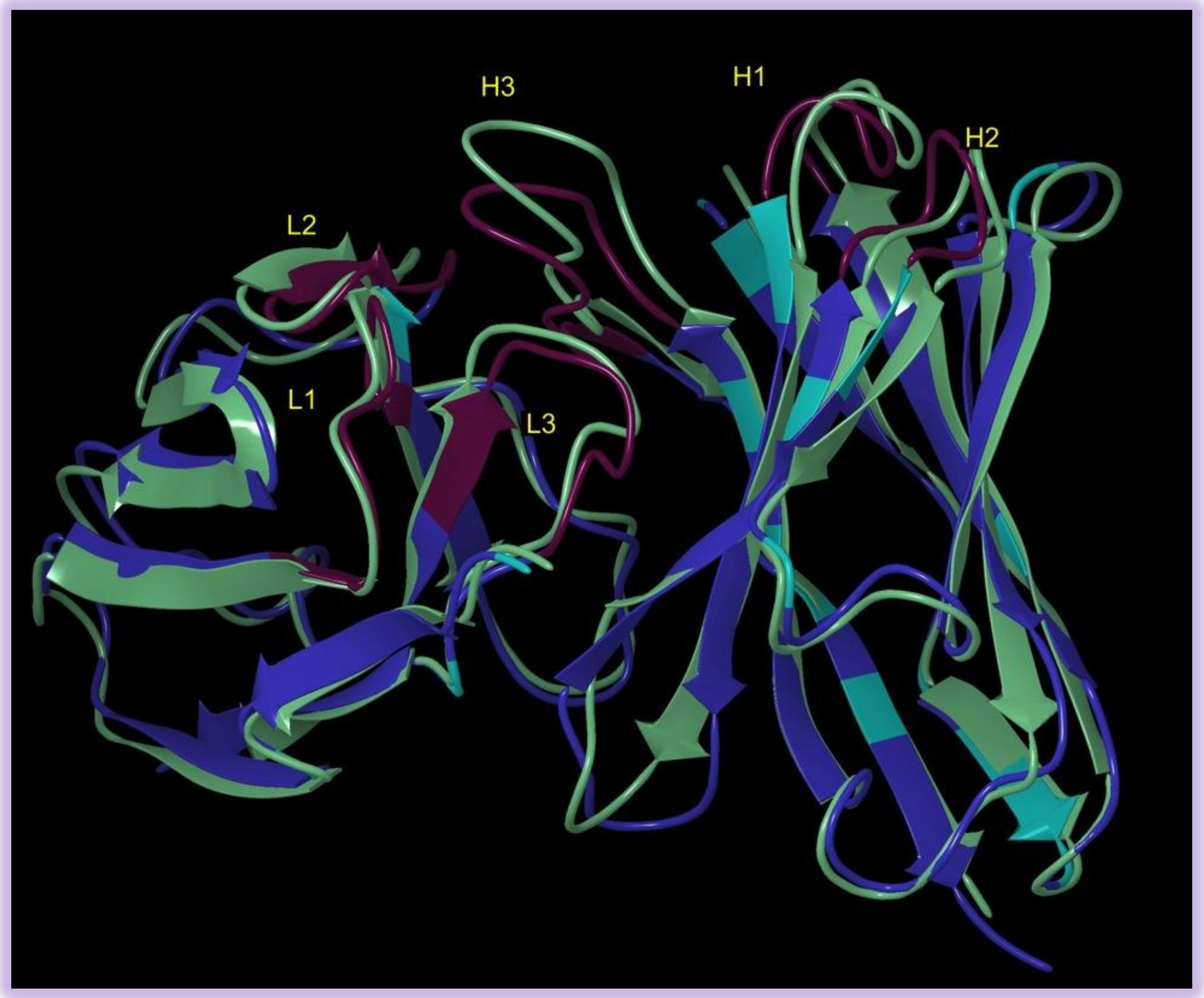
CAPRI rankings
(Nir London, Rosetta Design
Group, 2010)

Piper/Cluspro:

- **#1 group**
- **#1 server**

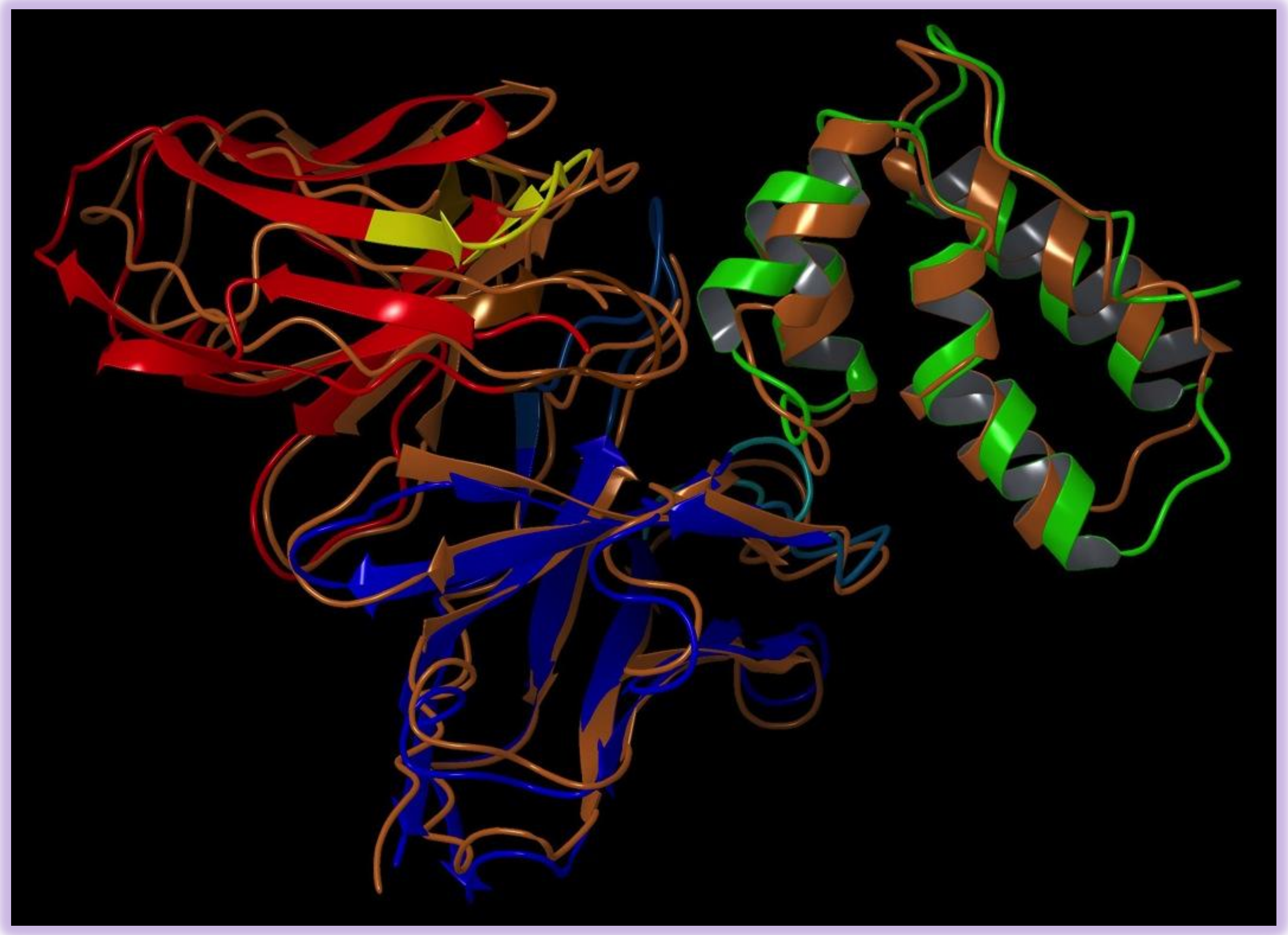


Antibody Prediction Using BioLuminate



Predicted CDR region FAB13B5 versus experiment (1E6J, light green)

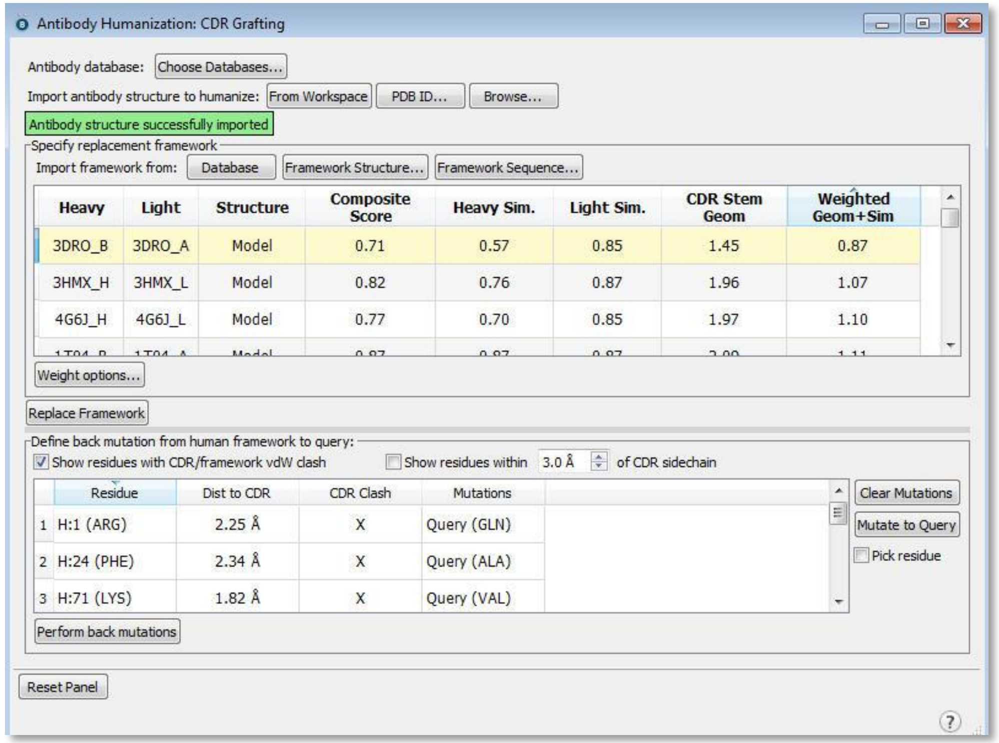
Antibody/Antigen Complex: Predicted Versus Experiment



Modeled FAB13B5 CDR docked with crystal structure of unbound antigen P24 (orange) versus x-ray complex 1E6J. 3rd ranked complex shown.

Antibody Modeling – Humanization

CDR-Grafting (Framework Replacement)



Antibody Humanization: CDR Grafting

Antibody database: Choose Databases...

Import antibody structure to humanize: From Workspace | PDB ID... | Browse...

Antibody structure successfully imported

Specify replacement framework

Import framework from: Database | Framework Structure... | Framework Sequence...

Heavy	Light	Structure	Composite Score	Heavy Sim.	Light Sim.	CDR Stem Geom	Weighted Geom+Sim
3DRO_B	3DRO_A	Model	0.71	0.57	0.85	1.45	0.87
3HMX_H	3HMX_L	Model	0.82	0.76	0.87	1.96	1.07
4G6J_H	4G6J_L	Model	0.77	0.70	0.85	1.97	1.10
1T04_B	1T04_A	Model	0.87	0.87	0.87	2.00	1.11

Weight options...

Replace Framework

Define back mutation from human framework to query:

☑ Show residues with CDR/framework vdW clash   ☐ Show residues within 3.0 Å of CDR sidechain

	Residue	Dist to CDR	CDR Clash	Mutations
1	H:1 (ARG)	2.25 Å	X	Query (GLN)
2	H:24 (PHE)	2.34 Å	X	Query (ALA)
3	H:71 (LYS)	1.82 Å	X	Query (VAL)

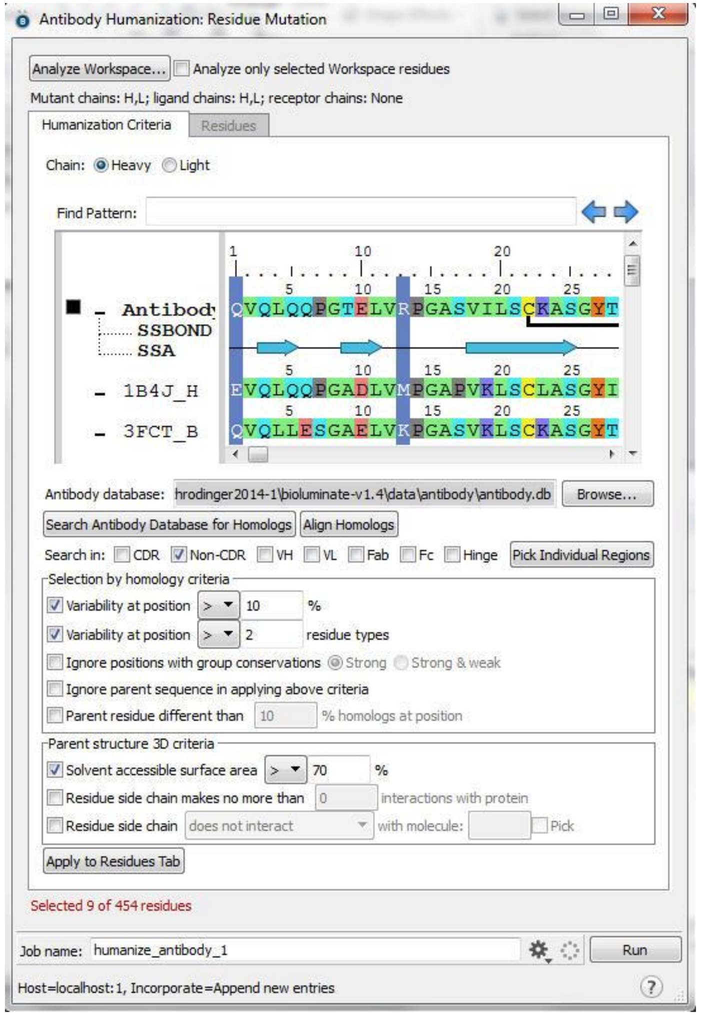
Clear Mutations | Mutate to Query | ☐ Pick residue

Perform back mutations

Reset Panel

- Easy to use
- Automatically IDs clashing residues for back mutation

Homology-based suggestions



Antibody Humanization: Residue Mutation

Analyze Workspace... ☐ Analyze only selected Workspace residues

Mutant chains: H,L; ligand chains: H,L; receptor chains: None

Humanization Criteria | Residues

Chain: ● Heavy ○ Light

Find Pattern:

```
Antibody   QVQLQQPGTELVRPGASVILSCKASGYT
 SSBOND
 SSA
1B4J_H     EVQLQQPGADLVMPGAPVKLSCLASGYI
3FCT_B     QVQLLESGAELVKPGASVKLSCKASGYT
```

Antibody database: hrodinger2014-1\bioluminate-v1.4\data\antibody\antibody.db | Browse...

Search Antibody Database for Homologs | Align Homologs

Search in: ☐ CDR ☑ Non-CDR ☐ VH ☐ VL ☐ Fab ☐ Fc ☐ Hinge | Pick Individual Regions

Selection by homology criteria

☑ Variability at position > 10 %

☑ Variability at position > 2 residue types

☐ Ignore positions with group conservations ● Strong ○ Strong & weak

☐ Ignore parent sequence in applying above criteria

☐ Parent residue different than 10 % homologs at position

Parent structure 3D criteria

☑ Solvent accessible surface area > 70 %

☐ Residue side chain makes no more than 0 interactions with protein

☐ Residue side chain does not interact with molecule: ☐ Pick

Apply to Residues Tab

Selected 9 of 454 residues

Job name: humanize_antibody_1 | Run

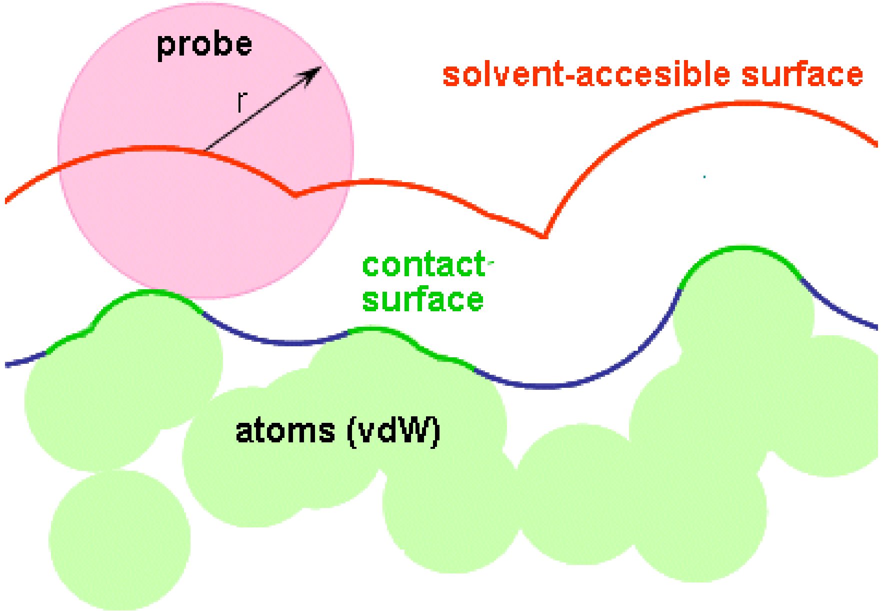
Host=localhost:1, Incorporate=Append new entries

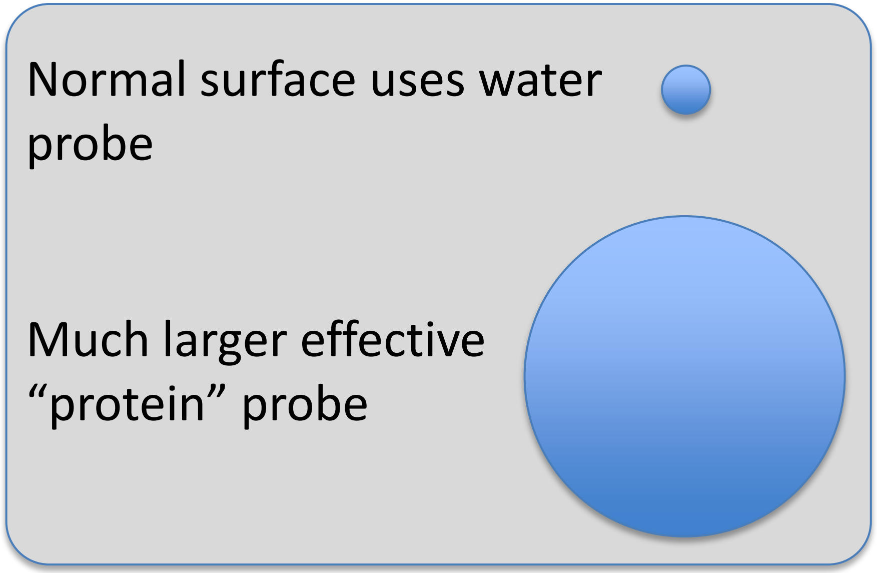
Compare to human sequences

- Degree of variability
- 3D information

Aggregation Prediction

- Aggregation can be viewed as recognition by a large sphere
  - Roll large probe sphere
  - Detect patches of exposed hydrophobic residues
- Reference: SAP (Spatial Aggregation Propensity)
  - Validated through collaboration with Novartis
  - Chennamsetty et al. (2009) *PNAS* 106 11937

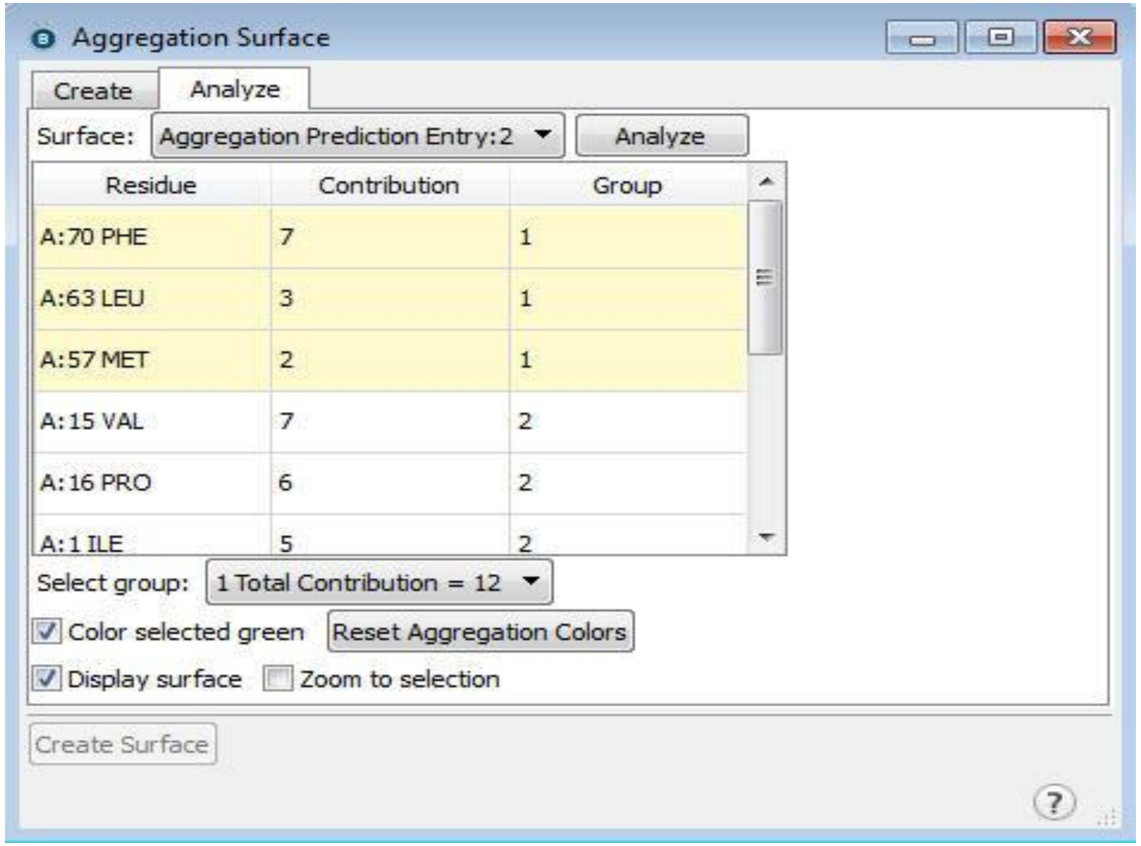



Normal surface uses water probe

Much larger effective “protein” probe

Color surface red to reflect hydrophobicity of contributing residues
Red hydrophobic “hot spots” are likely aggregation regions

Aggregation Surface Analysis



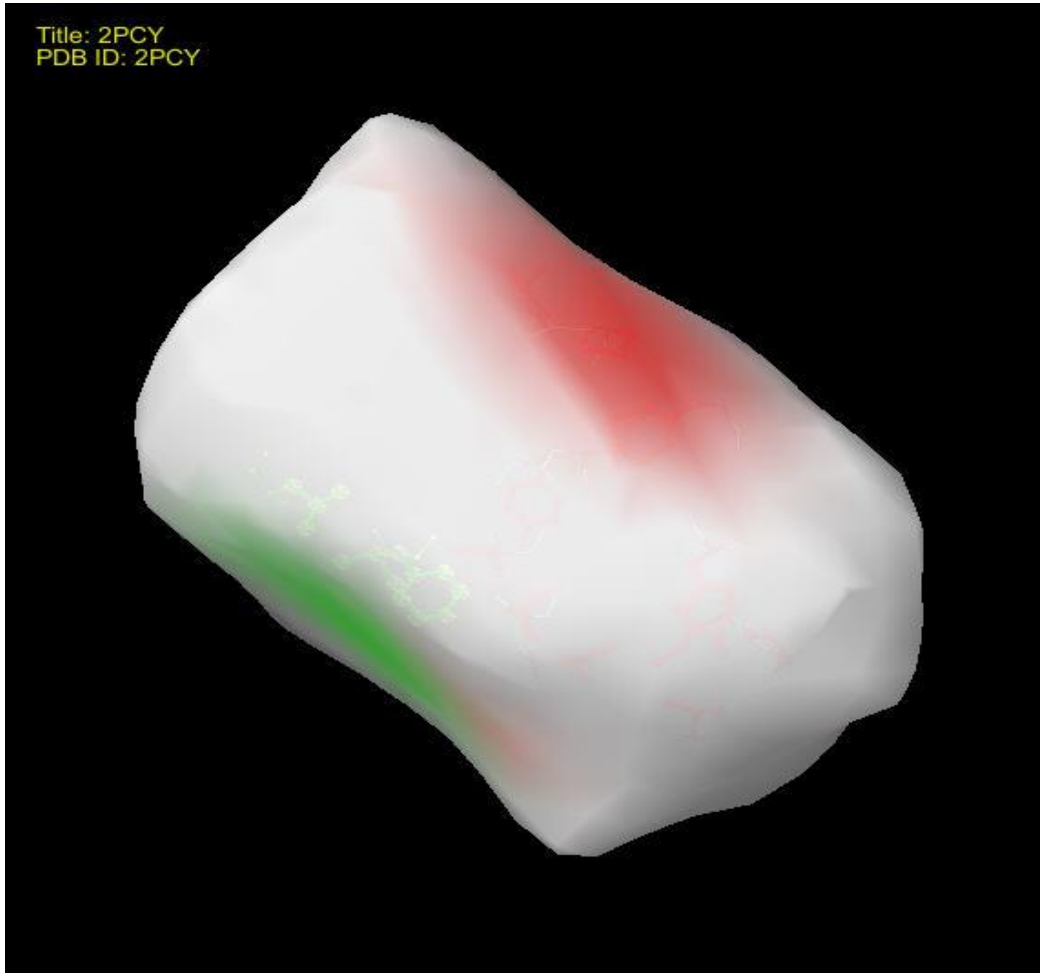
Surface: Aggregation Prediction Entry:2 ▾

Residue	Contribution	Group
A:70 PHE	7	1
A:63 LEU	3	1
A:57 MET	2	1
A:15 VAL	7	2
A:16 PRO	6	2
A:1 ILE	5	2

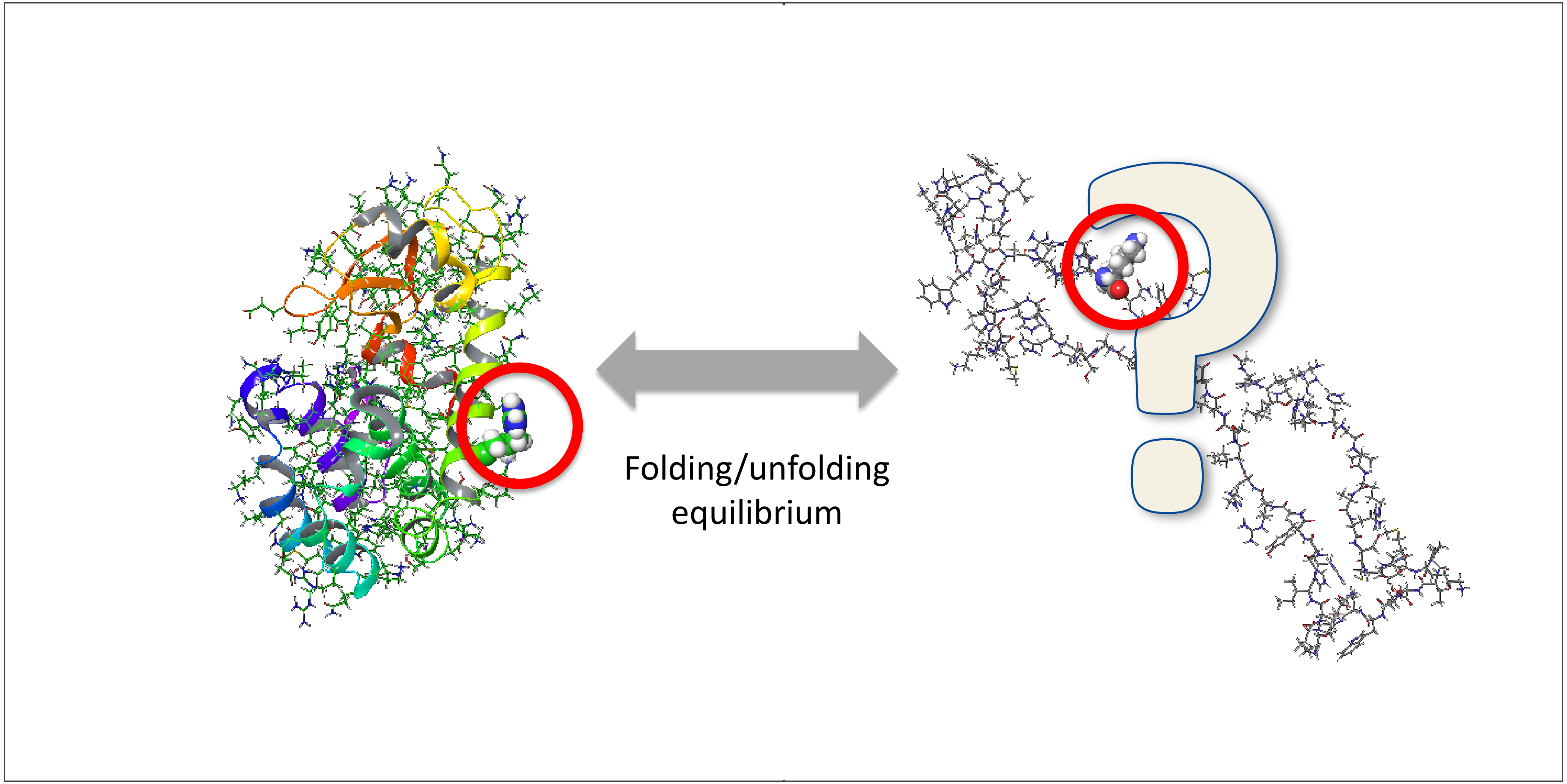
Select group: 1 Total Contribution = 12 ▾

☑ Color selected green

☑ Display surface ☐ Zoom to selection



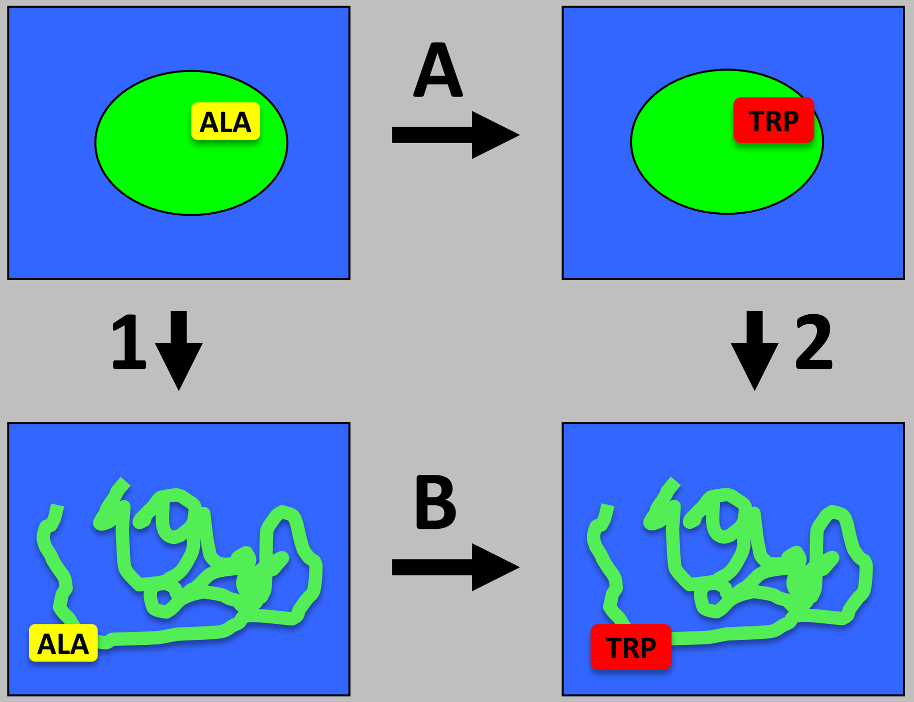
Protein Stability Thermodynamics



Folding/unfolding equilibrium

But what does the unfolded state really look like?

Schematic Thermodynamic Cycle

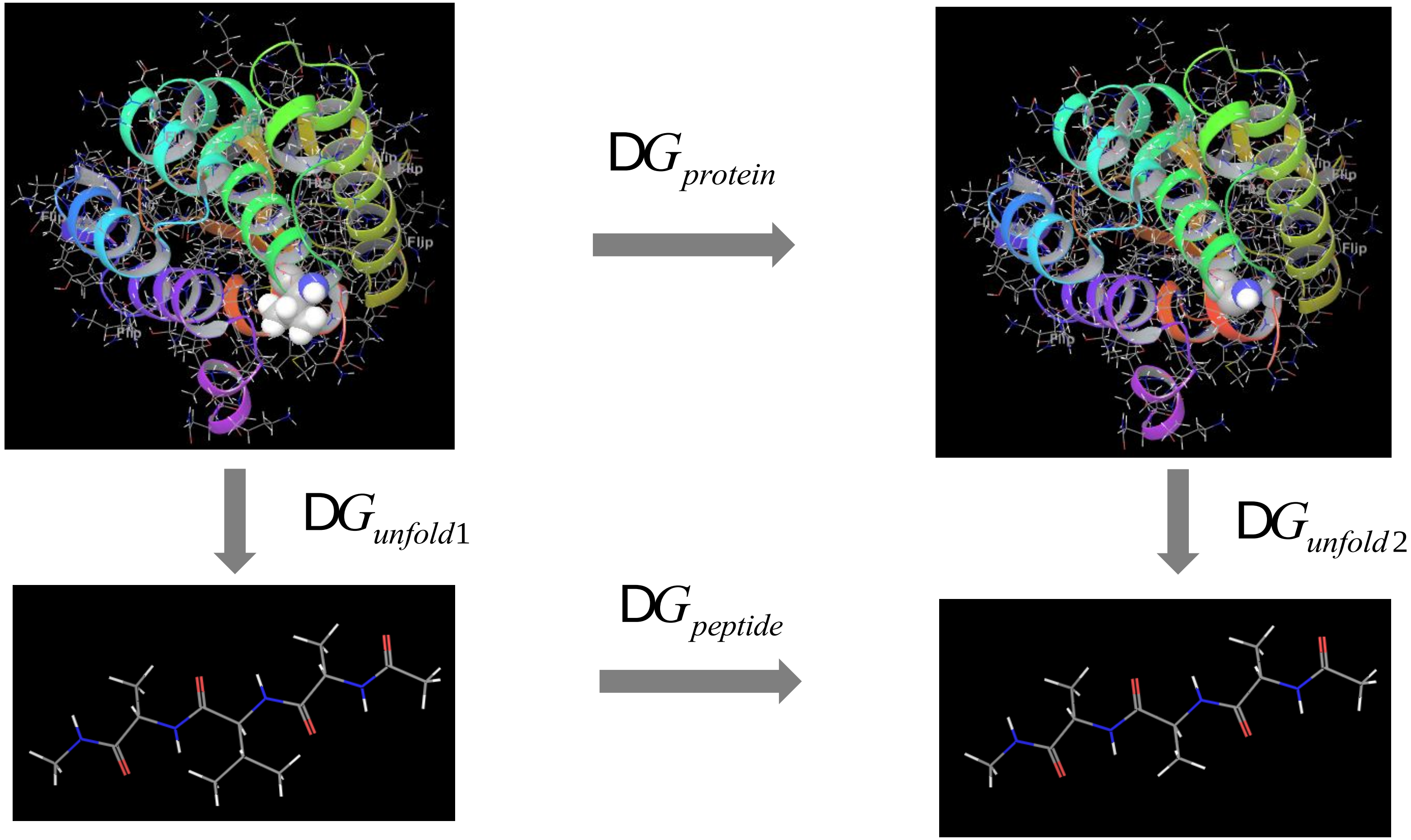


$$\begin{aligned}\Delta\Delta G_{stability} &= \Delta G_1 - \Delta G_2 \\ &= \Delta G_A - \Delta G_B\end{aligned}$$

- Simulate the (non-physical) protein side chain transformation
- Standard approach in FEP
- All non-physical terms cancel in the final result

Setting up a Cycle

Unfolded state is modeled by capped peptide



$$DDG = DG_{unfold1} - DG_{unfold2} = DG_{protein} - DG_{peptide}$$

Current Results, Part of the Fold-X Test Set

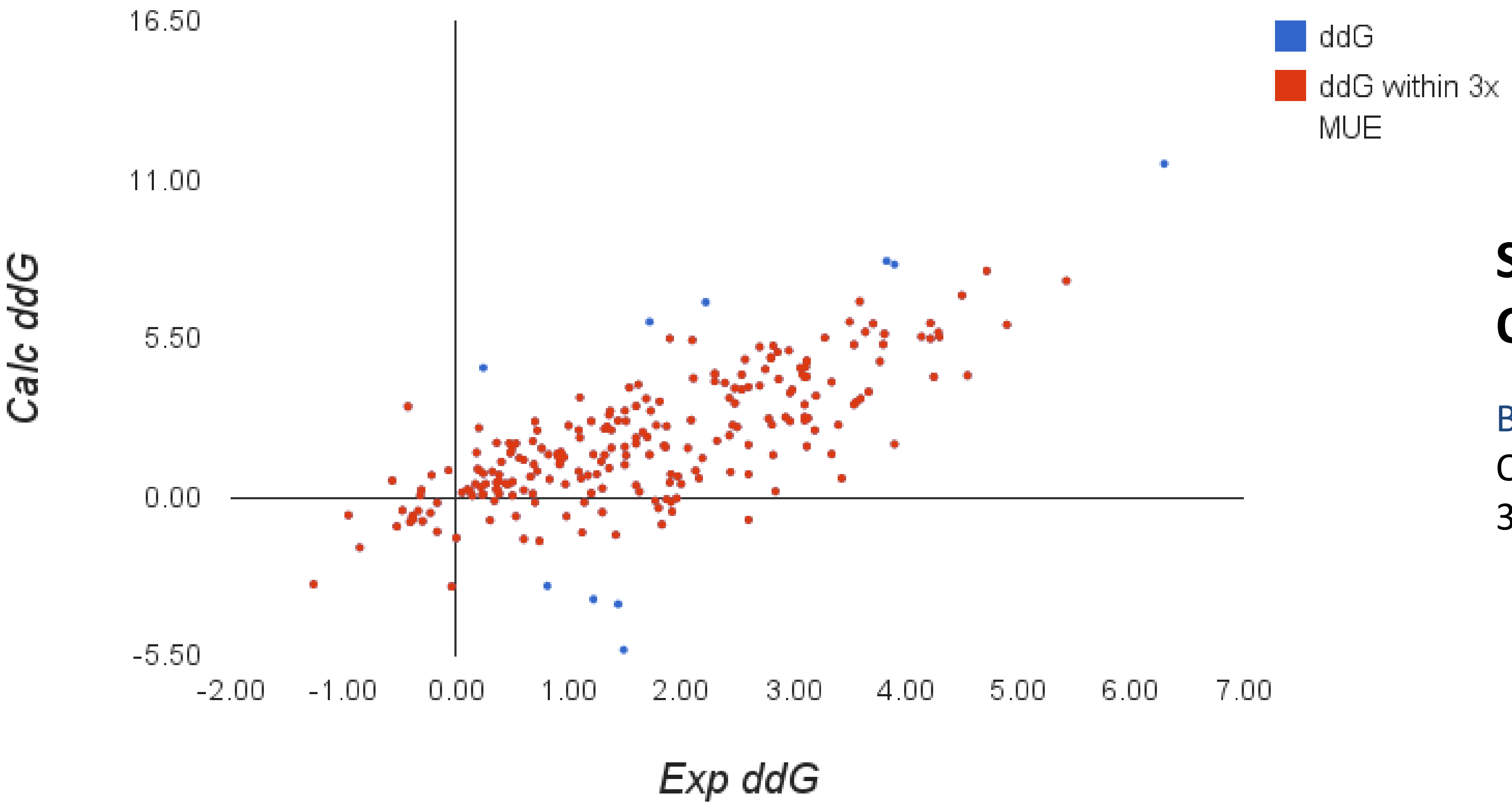
System	PDB ID	# Mutations	R^2-value	MUE	RMSE	ΔΔG Sign correct
T4-Lysozyme	2LZM	66	0.67	1.2	1.6	92%
Human Lysozyme	1REX	45	0.66	1.3	1.8	80%
Peptostrept. Magn. Prot. L	1HZ6	44	0.59	1.1	1.3	89%
B1 IG binding protein G	1PGA	24	0.37	1.1	1.4	79%
Fibronectin II domain	1TEN	32	0.60	1.4	1.7	91%
FK506 BP	1FKB	27	0.4	1.6	4.9*	85%
All		238	0.55	1.2	1.7	87%

Errors in kcal/mol

*Result strongly influenced by some outliers

Note: No charge changes in this set!

Correlation Plot



Slope = 1.3
Offset = -0.2

Blue:
Outliers of more than 3x MUE

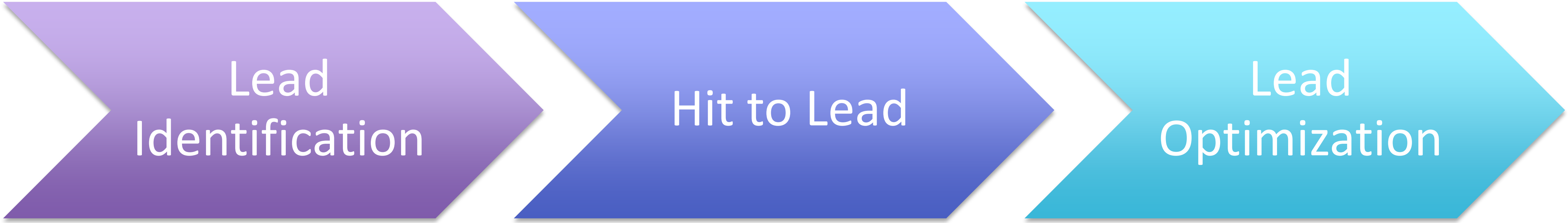
Comparison to Other Tools

- FEP+ performs well, but comparable to other tools
- For FEP+, no parameterisation was necessary, so results are more transferable

Software	R^2-value achieved*	Stabilizing/destabilizing % correct	MUE [kcal/mol]
CC/PBSA	0.31	79%	1.0
EGAD	0.35	71%	1.0
FoldX	0.25	70%	1.3
Hunter	0.20	69%	1.1
I-Mutant2.0	0.29	78%	1.1
Rosetta	0.07	73%	1.7
FEP+ (smaller data set!)	0.55	87%	1.2

* Calculated from R-values given in Tab I of Potapov, 2009, Prot. Eng. Des. Sel., 22, 553

The development of peptide therapeutics



Lead Identification

Hit to Lead

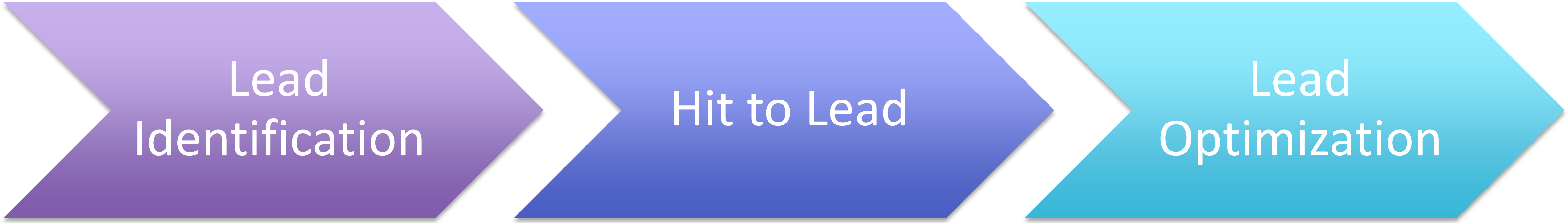
Lead Optimization

- Literature search
- Library screening

- Affinity and selectivity optimization
- Property Optimization

- Half-life extension

Computational tools can accelerate each step


Lead Identification

Hit to Lead

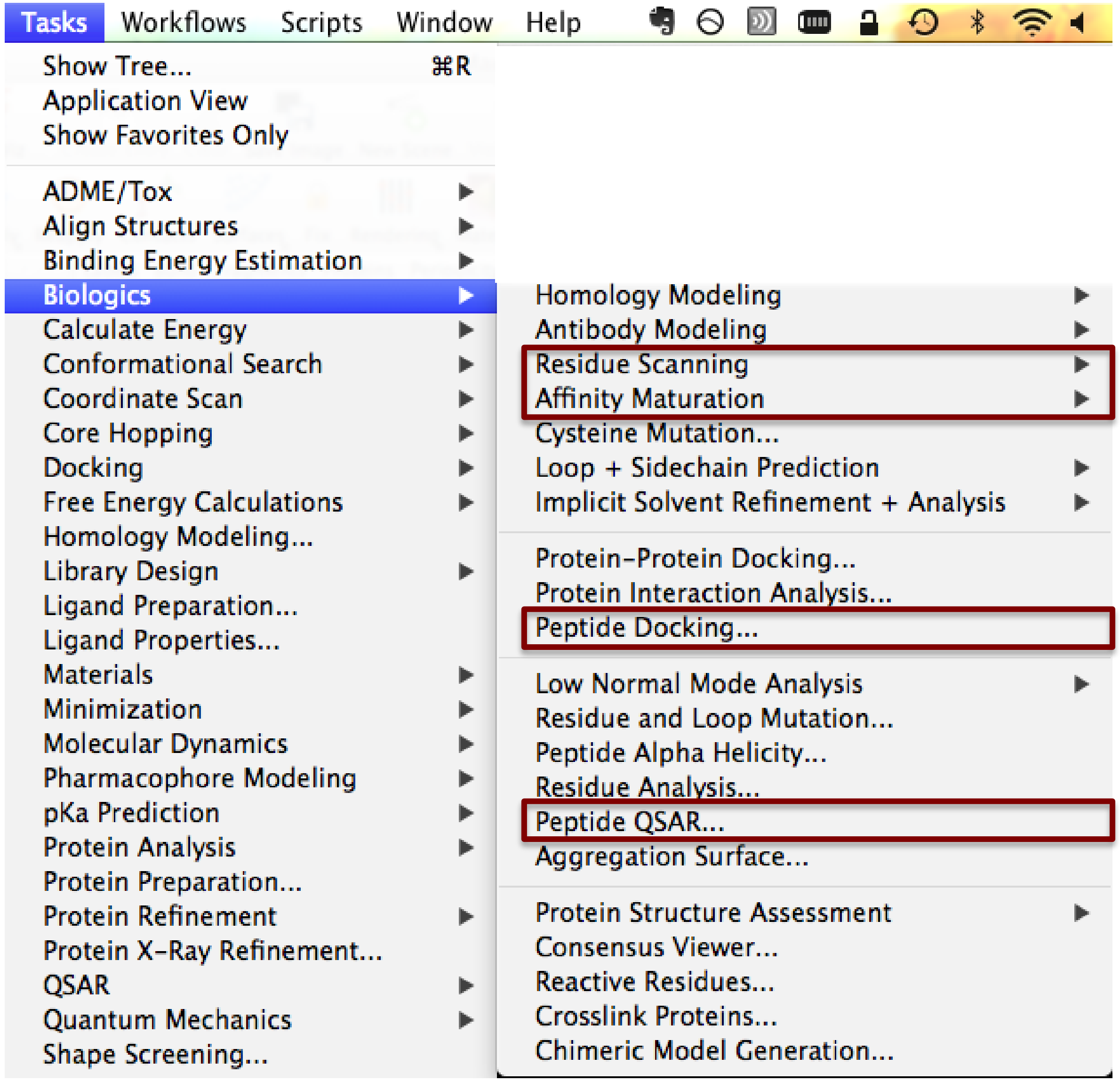
Lead Optimization

- Peptide docking

- Residue Scanning
- Affinity Maturation
- Peptide QSAR/QSAM

- Peptide QSAR/QSAM

Peptide Modeling with the Biologics Suite



Lead Identification

Peptide Docking

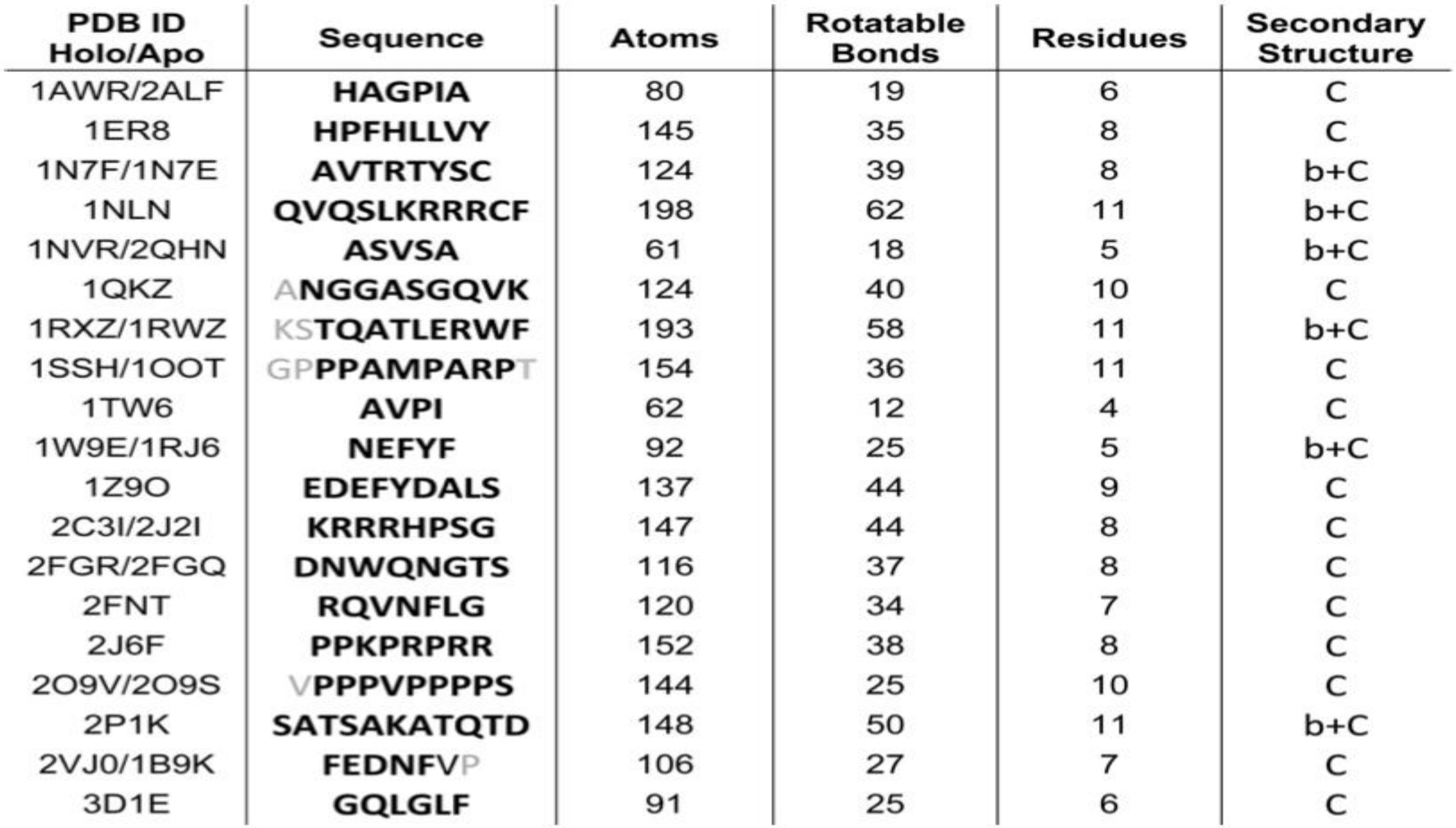
Polypeptide docking and Glide

- Several brute force sampling methods performs well in polypeptide docking but require hundreds or thousands of CPU hours per polypeptide docking
  - Rosetta FlexPepDock *ab-initio* (Raveh et al., *PloS one* **2011**, *6*, e18934)
  - DynaDock (Antes, *Proteins*, **2010**, *78*, 1084)
  - HADDOCK (Trellet et al., *PloS one* **2013**, *8*, e58769)
- Small molecule docking programs such as Glide are comparatively fast and accurate for docking of small molecules

Question: Can Glide SP dock 4-11 residue polypeptides well?

Raveh B et al. *PloS one.* **2011**; *6*: e18934.

Performance tested on a dataset from Raveh et al.



PDB ID Holo/Apo	Sequence	Atoms	Rotatable Bonds	Residues	Secondary Structure
1AWR/2ALF	HAGPIA	80	19	6	C
1ER8	HPFHLLVY	145	35	8	C
1N7F/1N7E	AVTRTYSC	124	39	8	b+C
1NLN	QVQSLKRRRCF	198	62	11	b+C
1NVR/2QHN	ASVSA	61	18	5	b+C
1QKZ	ANGGASGQVK	124	40	10	C
1RXZ/1RWZ	KSTQATLERWF	193	58	11	b+C
1SSH/1OOT	GPPPAMPARPT	154	36	11	C
1TW6	AVPI	62	12	4	C
1W9E/1RJ6	NEFYF	92	25	5	b+C
1Z9O	EDEFYDALS	137	44	9	C
2C3I/2J2I	KRRRHPSG	147	44	8	C
2FGR/2FGQ	DNWQNGTS	116	37	8	C
2FNT	RQVNFLG	120	34	7	C
2J6F	PPKPRPRR	152	38	8	C
2O9V/2O9S	VPPPVPPPPS	144	25	10	C
2P1K	SATSAKATQTD	148	50	11	b+C
2VJ0/1B9K	FEDNFVP	106	27	7	C
3D1E	GQLGLF	91	25	6	C

Raveh B et al. *PloS One.* **2011**; *6*: e18934.

Regular Glide Performance is Poor

- Metric of success: iRMSD of any of top 10 poses < 2.0 Å

  iRMSD: RMSD of peptide backbone atoms within 8 Å of protein

- Only **21%** of systems have an accurate pose (iRMSD ≤ 2.0 Å) within top 10 ranked poses by Glide SP (as compared to **63%** with Rosetta FlexPepDock)

- α-helical polypeptides not considered (ConfGen does not generate such conformations)

Optimized SP-PEP parameters improve results

Parameters	ConfGen	Rough Scoring	Refinement	Minimization	Final Pose
Glide Default	17	10	9	5	4
SP-PEP	17	11	10	8	7

24 experiments

Number of systems with at least one iRMSD < 2.0 Å pose

SP-PEP Parameters: 10 conformers generated using ConfGen, dock each conformer using Glide SP

Tubert-Brohman I et al. *Chem. Inf. Model.* 2013; 53(7): 1689-99

Classification of Complexes Based on Accuracy

PDB	Highest ranking of accurate pose	PDB	Highest ranking of accurate pose	PDB	Highest ranking of accurate pose	PDB	Highest ranking of accurate pose
1N7F	2	1AWR	1	2J6F	321	1QKZ	92
1NLN	1	1ER8	5	2O9V	13	1RXZ	-
1NVR	1	1SSH	1			1Z9O	-
1TW6	1	1W9E	18			2C3I	-
2FNT	1	1P1K	9			2FGR	-
3D1E	1	1VJ0	14				
Easy		**Medium**		**Hard**		**Very Hard**	

Conclusions

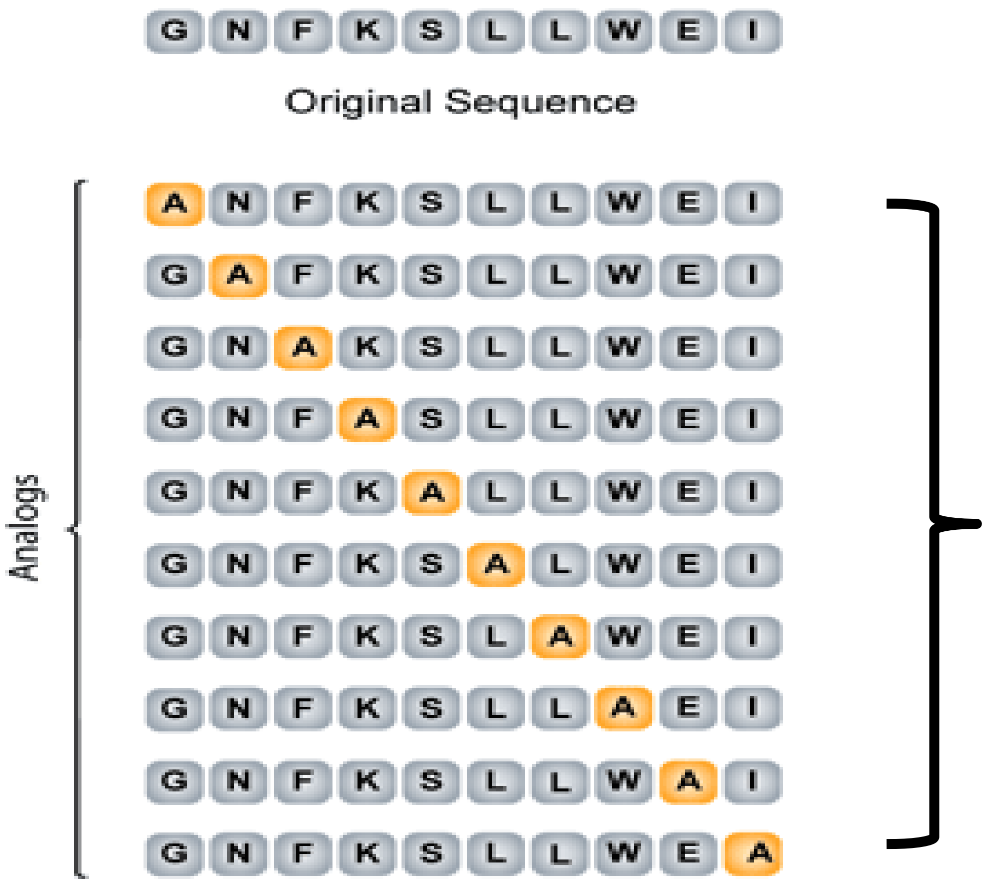
	Standard SP	SP-PEP	SP-PEP + MMGBSA	Rosetta FlexPepDock
% cases where top 10 iRMSD < 2Å	21%	41%	58%	63% (but 100x slower)

- ConfGen performed well finding <2Å RMSD pose in 100% of cases
- α-helical peptides cannot be docked with Glide
- Performance best on short, extended, non-ionizable peptides
- More work is needed to achieve small-molecule like accuracy

Hit to Lead

Residue Scanning, Affinity Maturation

Residue Scanning: Overview



G N F K S L L W E I

Original Sequence

A N F K S L L W E I

G A F K S L L W E I

G N A K S L L W E I

G N F A S L L W E I

G N F K A L L W E I

G N F K S A L W E I

G N F K S L A W E I

G N F K S L L A E I

G N F K S L L W A I

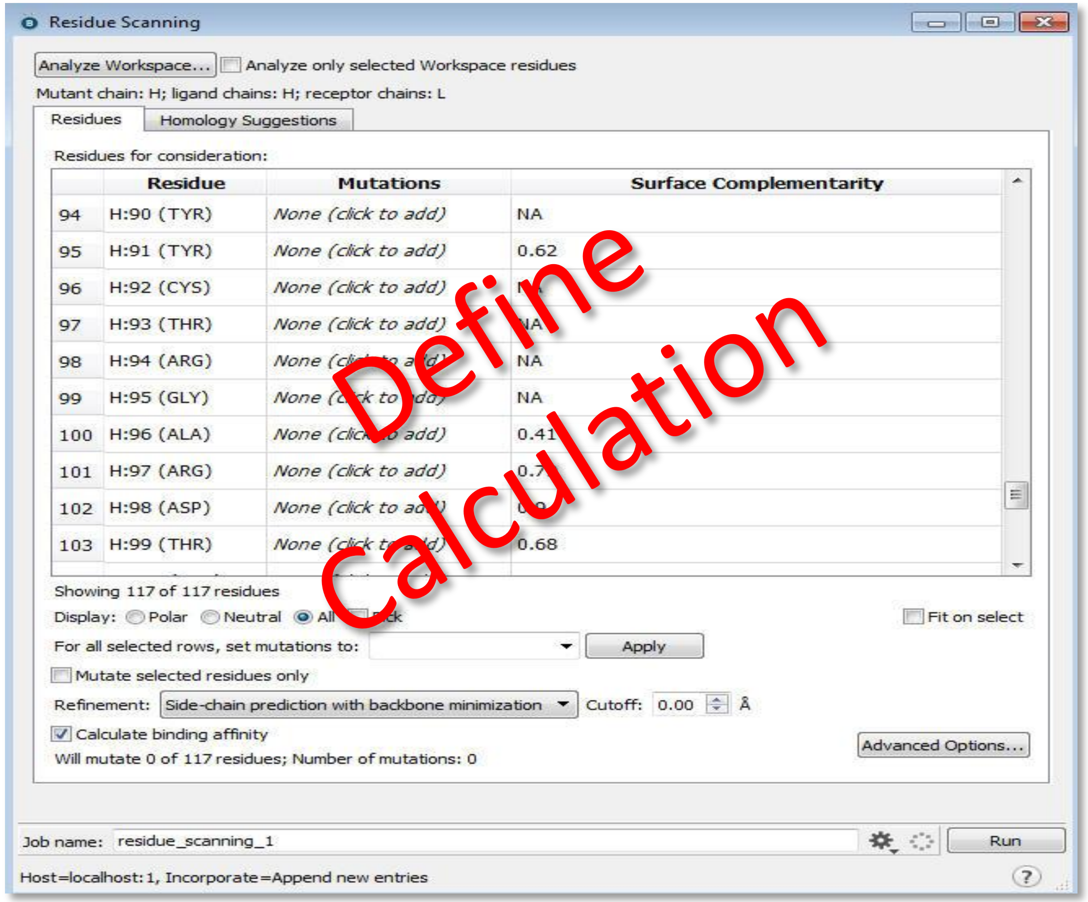
G N F K S L L W E A

Measure Δ

Affinity
Stability
pKa
Etc.

- Used to determine what effect specific amino acid positions have on properties such as binding, stability, etc.
- Tells us what mutations may be beneficial, and what may be harmful
- Can be a very laborious and difficult task to do in the lab.

Residue Scanning in BioLuminate



Residue Scanning

Analyze Workspace... □ Analyze only selected Workspace residues

Mutant chain: H; ligand chains: H; receptor chains: L

Residues | Homology Suggestions

Residues for consideration:

	Residue	Mutations	Surface Complementarity
94	H:90 (TYR)	*None (click to add)*	NA
95	H:91 (TYR)	*None (click to add)*	0.62
96	H:92 (CYS)	*None (click to add)*	NA
97	H:93 (THR)	*None (click to add)*	NA
98	H:94 (ARG)	*None (click to add)*	NA
99	H:95 (GLY)	*None (click to add)*	NA
100	H:96 (ALA)	*None (click to add)*	0.41
101	H:97 (ARG)	*None (click to add)*	0.7
102	H:98 (ASP)	*None (click to add)*	0.9
103	H:99 (THR)	*None (click to add)*	0.68

Showing 117 of 117 residues

Display: ○ Polar ○ Neutral ● All □ Pick □ Fit on select

For all selected rows, set mutations to: [Apply]

□ Mutate selected residues only

Refinement: Side-chain prediction with backbone minimization Cutoff: 0.00 Å

☑ Calculate binding affinity

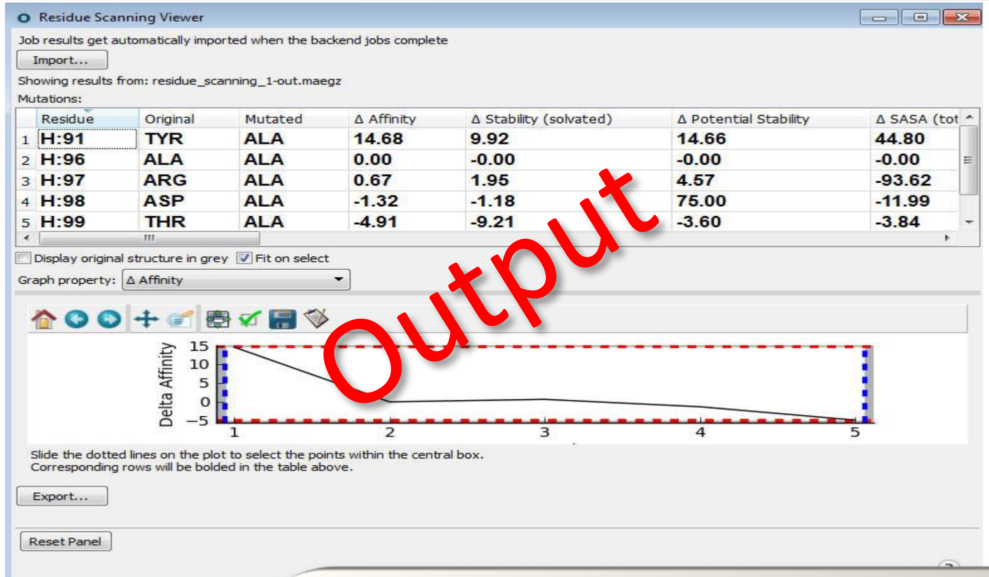
Will mutate 0 of 117 residues; Number of mutations: 0

Advanced Options...

Job name: residue_scanning_1 [Run]

Host=localhost:1, Incorporate=Append new entries

Define Calculation



Residue Scanning Viewer

Job results get automatically imported when the backend jobs complete

Import...

Showing results from: residue_scanning_1-out.maegz

Mutations:

	Residue	Original	Mutated	Δ Affinity	Δ Stability (solvated)	Δ Potential Stability	Δ SASA (tot
1	H:91	TYR	ALA	14.68	9.92	14.66	44.80
2	H:96	ALA	ALA	0.00	-0.00	-0.00	-0.00
3	H:97	ARG	ALA	0.67	1.95	4.57	-93.62
4	H:98	ASP	ALA	-1.32	-1.18	75.00	-11.99
5	H:99	THR	ALA	-4.91	-9.21	-3.60	-3.84

□ Display original structure in grey ☑ Fit on select

Graph property: Δ Affinity

Slide the dotted lines on the plot to select the points within the central box.
Corresponding rows will be bolded in the table above.

Export...

Reset Panel

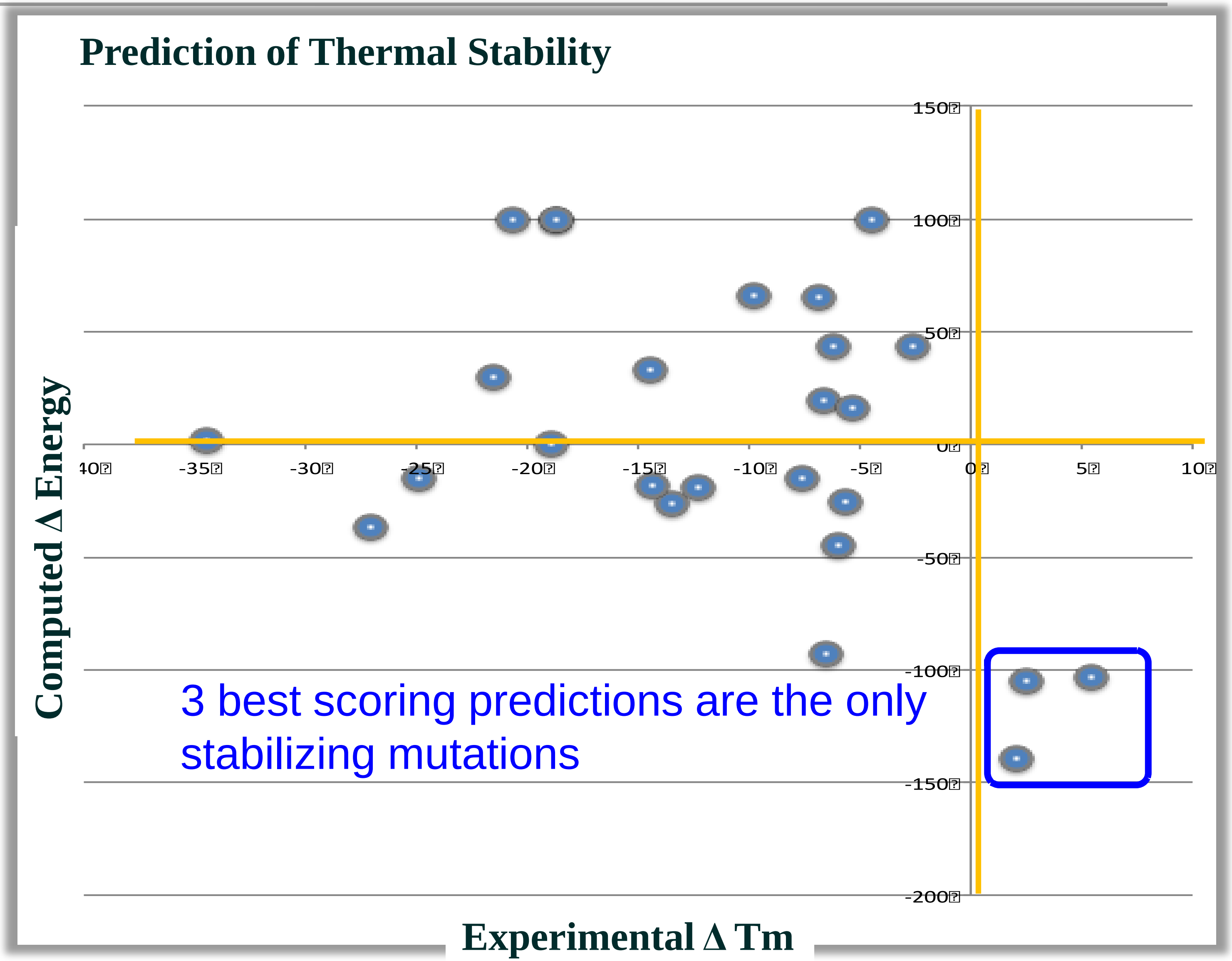
Output

Workspace Interactive

- Select any protein residues to be mutated
- Run time ~30sec/mutation
- See how properties change
  - Affinity, stability, hydrophobicity, SASA, etc.

Prospective Example: Thermal Stability of SH3 Domain Mutants

- 2 mutation locations
  - Glu107
  - Ser124
- 25 mutations made and tested experimentally
- Only 3 mutations lead to increased thermal stability
  - E107D
  - S124K
  - S124R
- Residue scanning IDs these 3 mutations

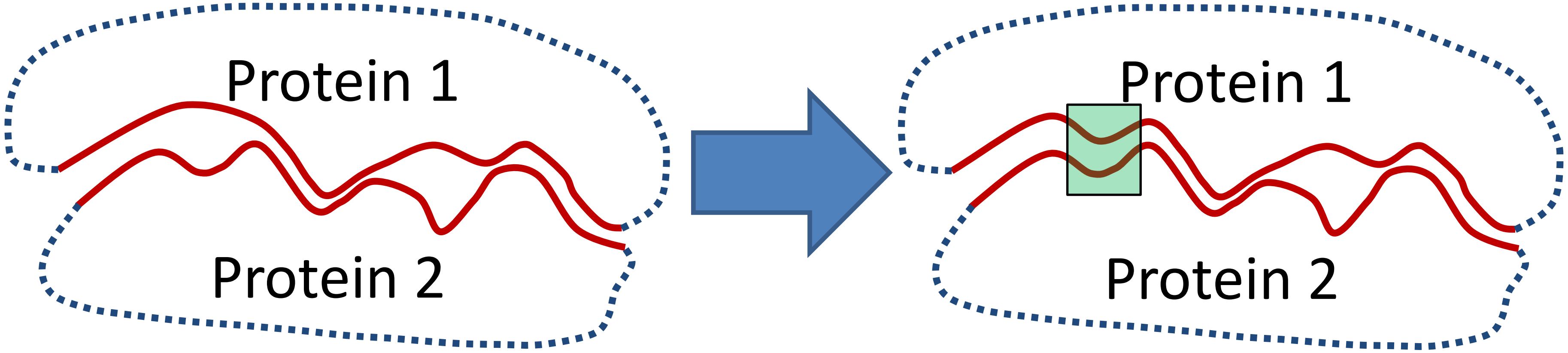

Prediction of Thermal Stability

3 best scoring predictions are the only stabilizing mutations

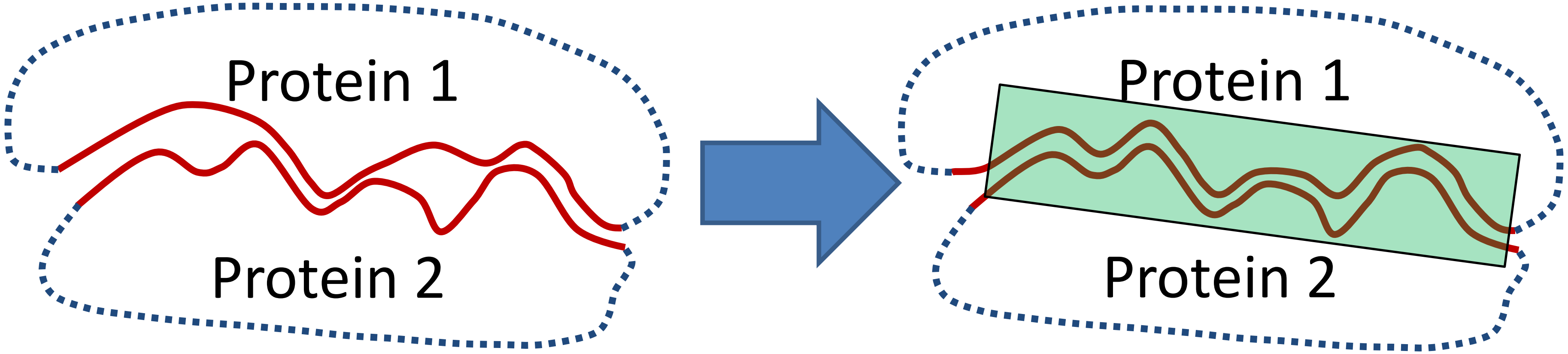
SCHRÖDINGER.

From Residue Scanning to Affinity Maturation

Residue Scanning (single mutations):

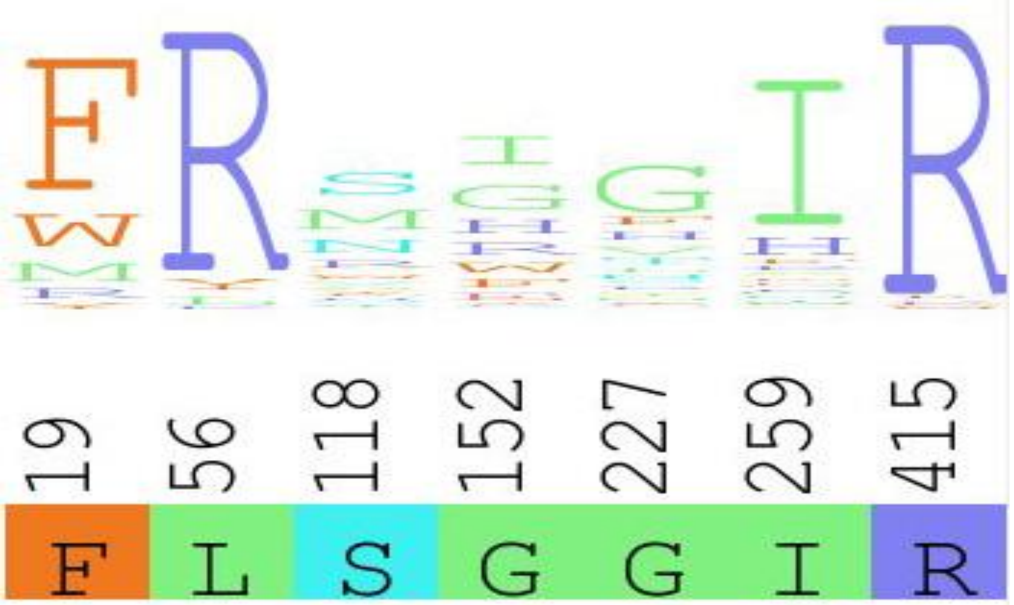


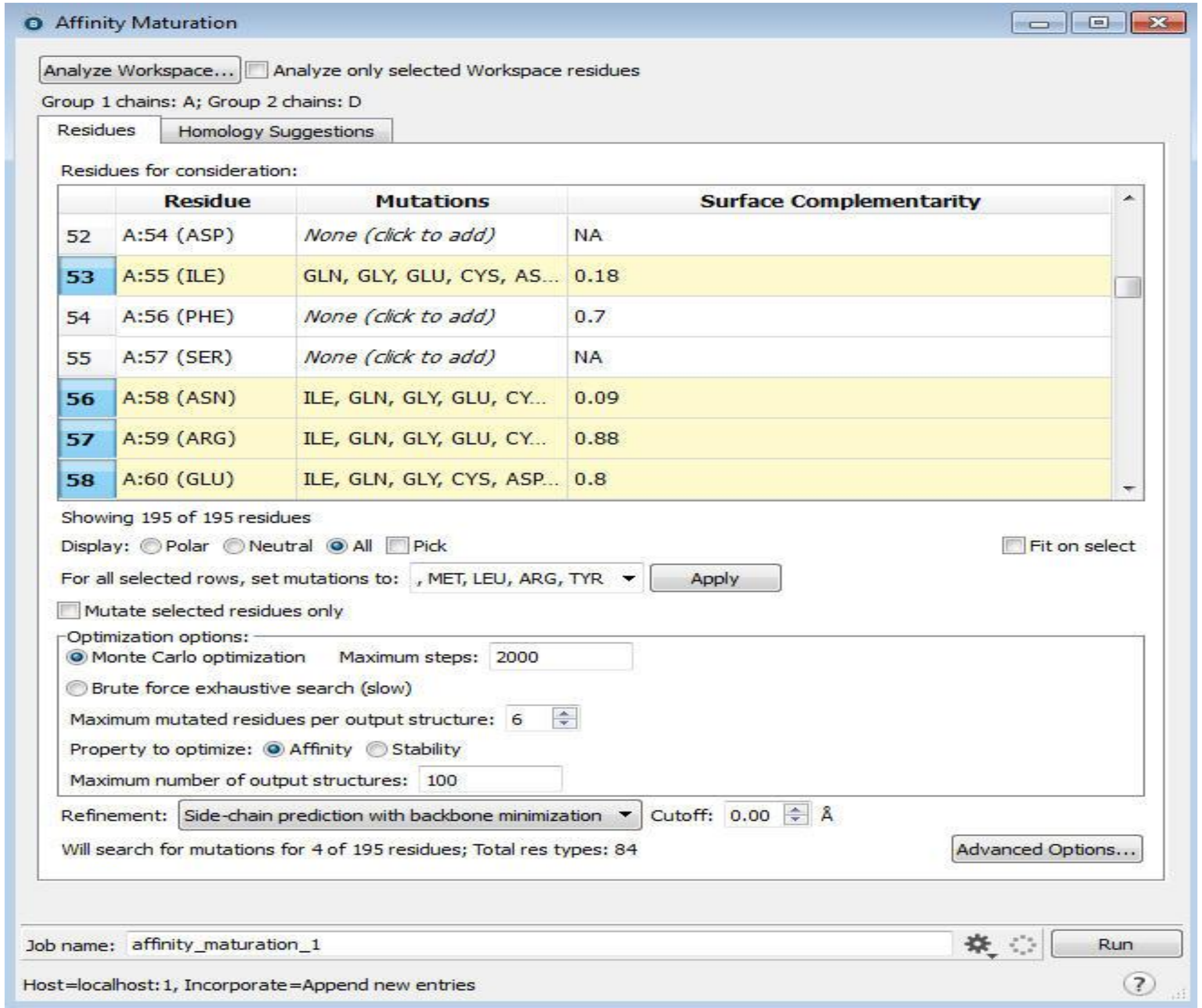
Affinity Maturation/Protein Design (multiple simultaneous mutations):



Affinity Maturation in BioLuminate

- Search multiple residue pos simultaneously for changes
- Use to suggest new sequences, or to influence random library design




Affinity Maturation

Analyze Workspace... □ Analyze only selected Workspace residues

Group 1 chains: A; Group 2 chains: D

Residues | Homology Suggestions

Residues for consideration:

	Residue	Mutations	Surface Complementarity
52	A:54 (ASP)	*None (click to add)*	NA
53	A:55 (ILE)	GLN, GLY, GLU, CYS, AS...	0.18
54	A:56 (PHE)	*None (click to add)*	0.7
55	A:57 (SER)	*None (click to add)*	NA
56	A:58 (ASN)	ILE, GLN, GLY, GLU, CY...	0.09
57	A:59 (ARG)	ILE, GLN, GLY, GLU, CY...	0.88
58	A:60 (GLU)	ILE, GLN, GLY, CYS, ASP...	0.8

Showing 195 of 195 residues

Display: ○ Polar ○ Neutral ● All □ Pick □ Fit on select

For all selected rows, set mutations to: , MET, LEU, ARG, TYR Apply

□ Mutate selected residues only

Optimization options:

● Monte Carlo optimization Maximum steps: 2000

○ Brute force exhaustive search (slow)

Maximum mutated residues per output structure: 6

Property to optimize: ● Affinity ○ Stability

Maximum number of output structures: 100

Refinement: Side-chain prediction with backbone minimization Cutoff: 0.00 Å

Will search for mutations for 4 of 195 residues; Total res types: 84

Advanced Options...

Job name: affinity_maturation_1 Run

Host=localhost:1, Incorporate=Append new entries

SCHRÖDINGER.

Lead Optimization

Peptide QSAR

What is QSAM modeling?

- In traditional QSAR modeling, structural features of biomolecules are used to develop models for activity
  - i.e. Activity = $\boldsymbol{f}$ (molecular structure)

- QSAM stands for **Q**uantitative **S**equence **A**ctivity **M**odeling:
  - As compared to small molecule QSAR approaches, QSAM models **sequence** information directly using **sequence descriptors**
  - i.e. Activity = $\boldsymbol{f}$ (peptide sequence)

Sequence descriptors are similar to molecular descriptors

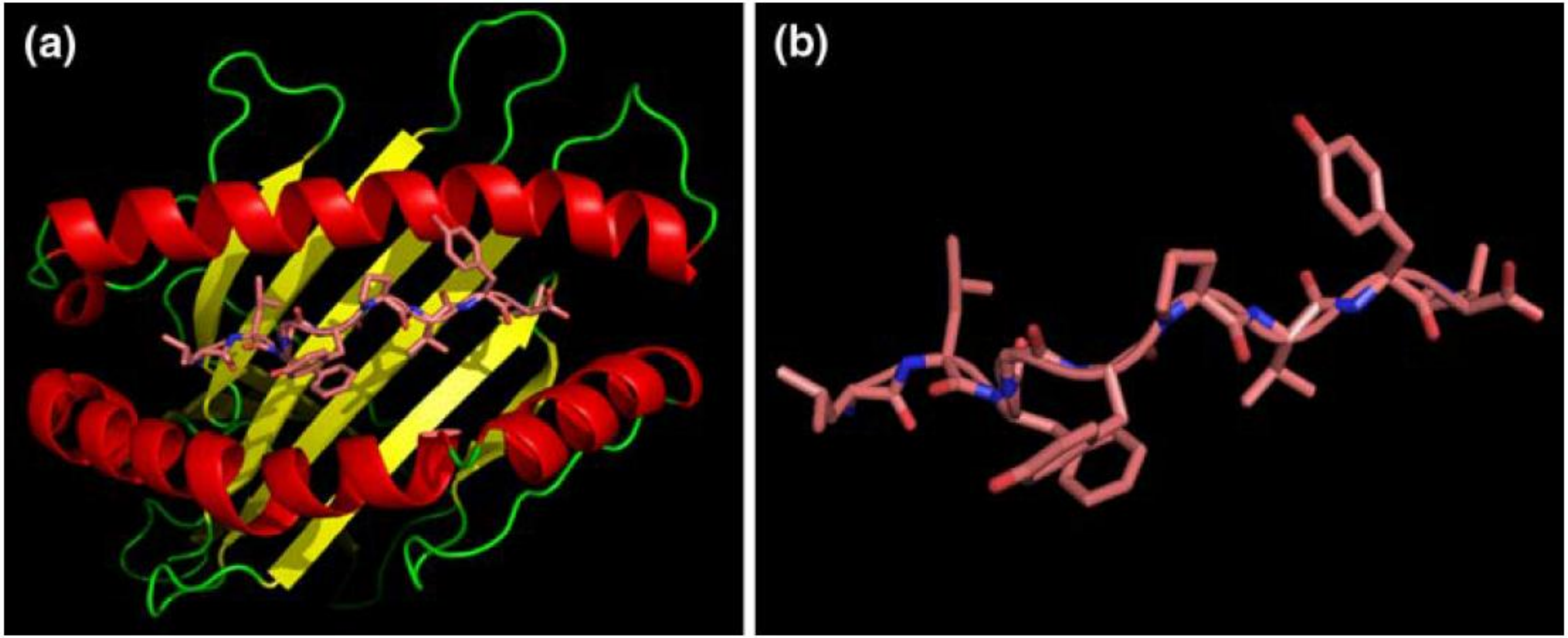
- They are based on physicochemical properties of the individual amino acids that comprise the sequence
  - i.e. size, shape, charge, etc
- Three Examples:
  - Zvalue: derived from principle components analysis (PCA) of 29 physicochemical properties of the 20 natural AAs
    - Hellberg et al. *J Med Chem*. 1987; 30: 1126-1135.
  - EZvalue: derived from principle components analysis (PCA) of 29 physicochemical properties for 87 AAs (natural and modified)
    - Sandberg et al. J *Med Chem*. 1998; 41: 2481-2491
  - DPPS: 10 score vectors derived from PCA of 109 properties of the 20 natural Aas
    - Properties include 23 electronic properties, 37 steric properties, 54 hydrophobic properties and 5 hydrogen bond properties
    - Tian et al. *Amino Acids*. 2009; 36: 535-554

QSAM Modelling: Pros and Cons

- Pros:
  - **Very quick calculation**
    - There is no need for any sort of 3D-structure
      - And certainly no requirement for alignment/docking
    - Can be used to filter through large lists of sequences very rapidly
- Cons:
  - **Immediate interpretation is difficult**
    - The underlying amino-acid descriptors do have physical interpretability
      - Theoretically it is possible to understand what residues are required at each position*
  - **Success depends on having descriptors for each amino acid present**
    - Handling un-natural amino-acids can be difficult

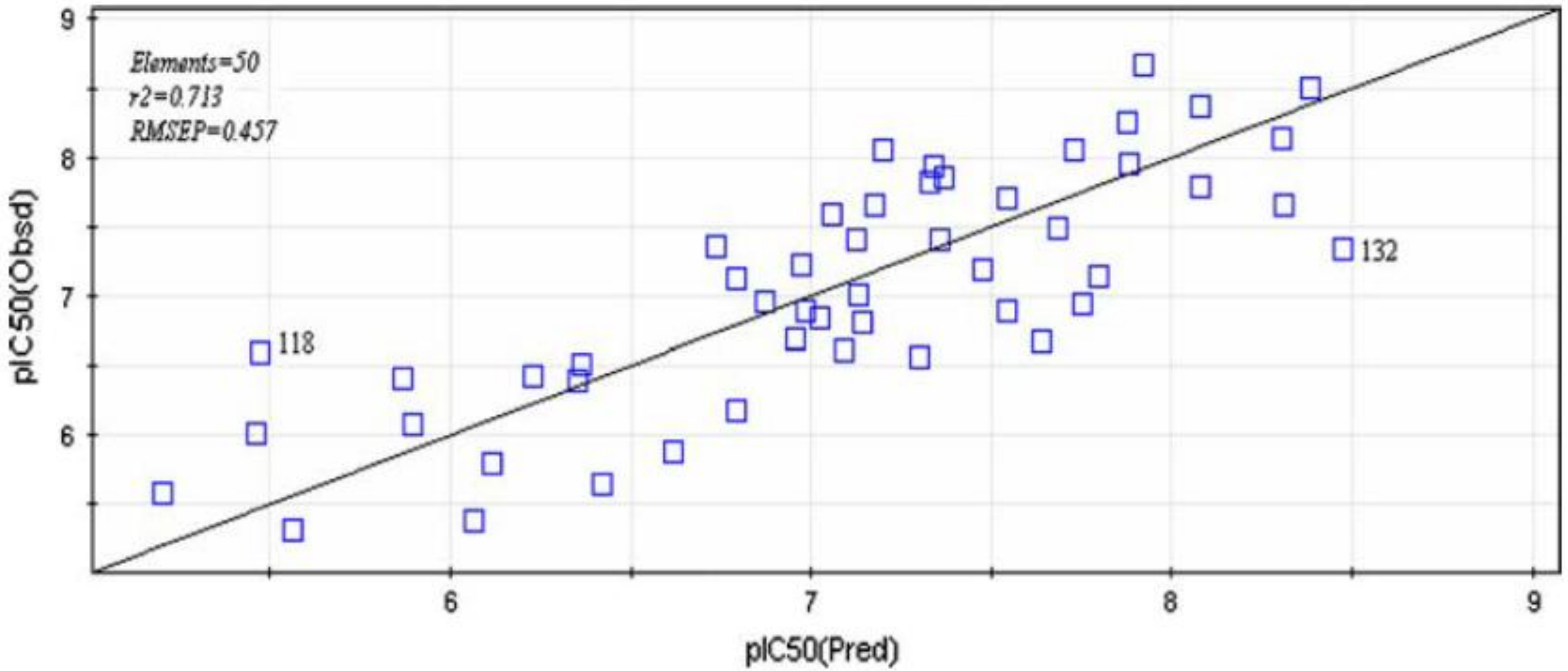
* This is as much a limitation of the underlying Canvas PLS implementation as it is of QSAM and the Bioluminate panel. More advanced PLS tools (e.g. Umetric’s SIMCA package) would enable a more detailed analysis to be performed.

Example: Modeling antigenic peptide binding to MHC



Tian et al. *Amino Acids*. 2009; 36: 535-554

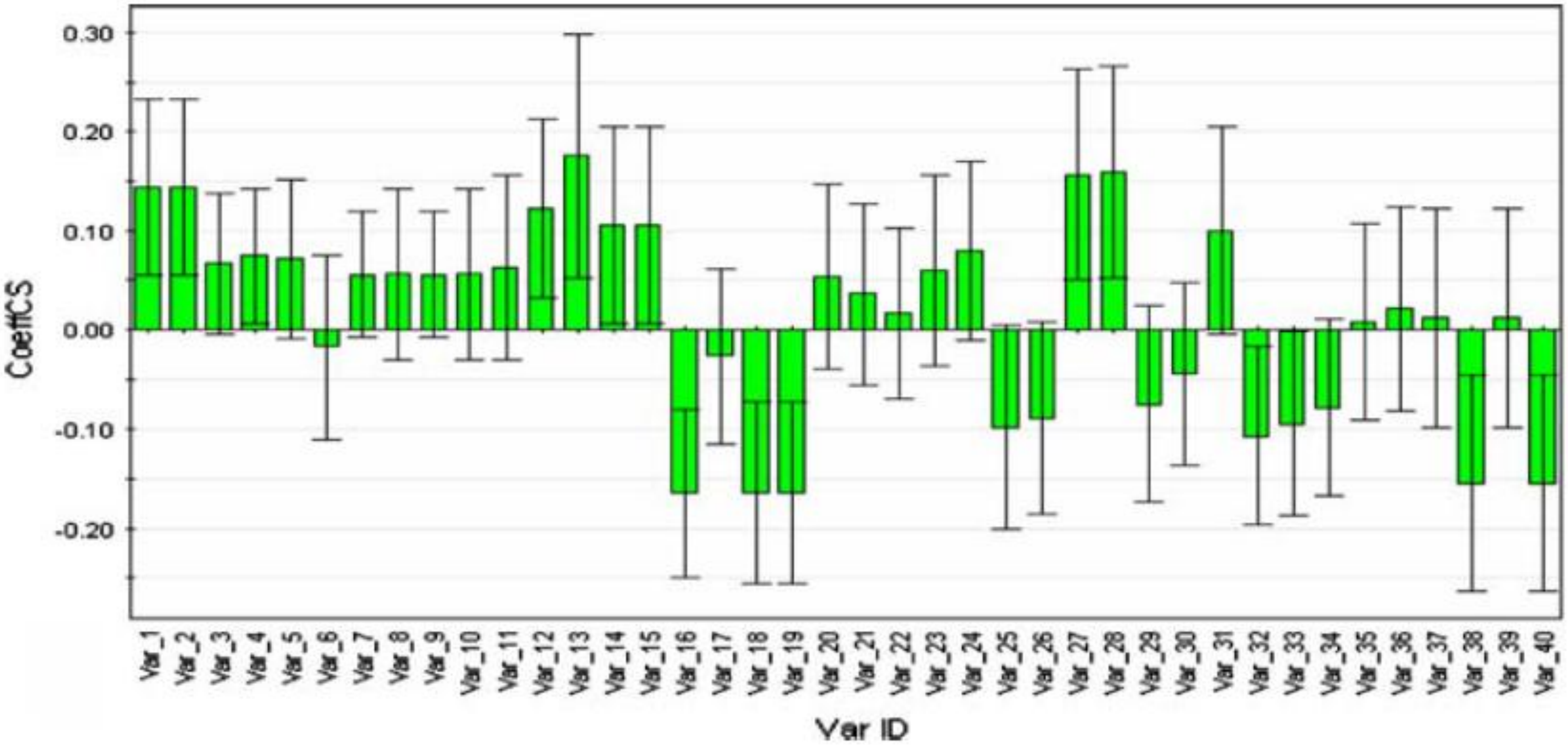
The model performs very well



- The model was derived using the DPPS descriptor on a dataset of 152 sequences
- Partial least squares (PLS) regression was used to generate the model

Tian et al. *Amino Acids*. 2009; 36: 535-554

But physical interpretation of the model is tricky …



- Standardized coefficients of 40 selected variables from the model.
- Each variable corresponds to a peptide sequence position.

Tian et al. *Amino Acids*. 2009; 36: 535-554

Acknowledgement

- David Pearlman
- Kai Zhu
- Johannes Maier
- Eric Feyfant
- Thijs Beuming
- Thomas Steinbrecher