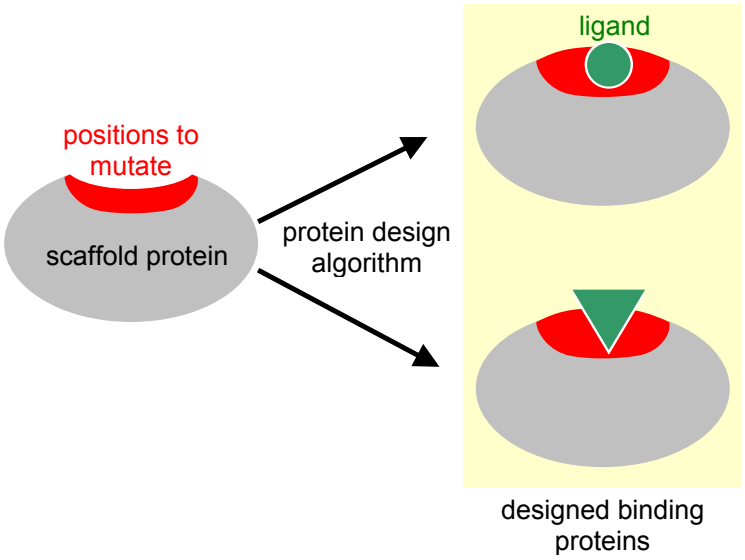
Physics-based engineering of protein-ligand binding

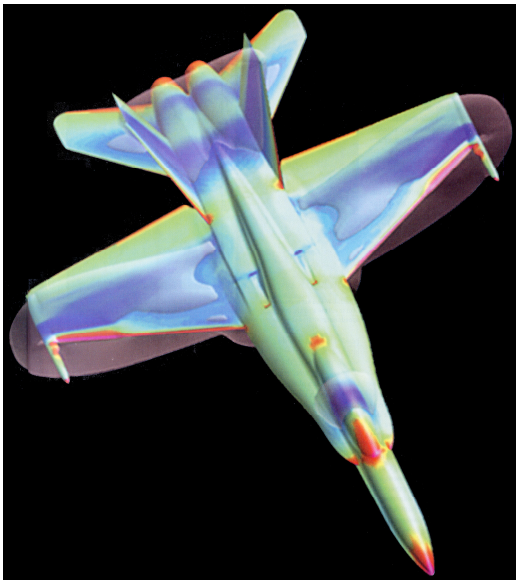


F. Edward Boas
Harbury Lab
Thesis defense
2006-06-19

Engineering approach to protein design

Most of biochemistry research has been focused on *reverse* engineering proteins, pathways, and cells. We would like to turn this around and start *forward* engineering new functions into proteins.

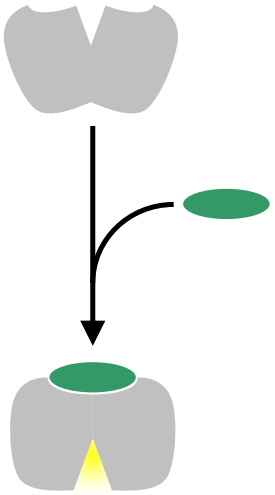
Airplanes, buildings, and circuits are all designed in a computer using accurate models. Can we do the same thing for protein design?



Computer model of an airplane flying at Mach 0.9. Warmer colors denote higher air pressures; the transparent region indicates where local airflow velocity is Mach 1.

Steven Ashley. (Nov. 2003) "Flying on flexible wings" *Scientific American*. 84-91.
Image credit: Richard Snyder, Wright-Patterson Air Force Base AFRL.

Application: Biosensor



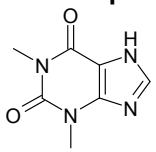
Binding → conformational change → change in fluorescence or enzymatic activity

Subcellular fluorescence imaging of signaling molecules

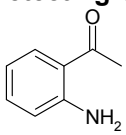
IP_3, cAMP, leukotrienes

Drugs with a low therapeutic index

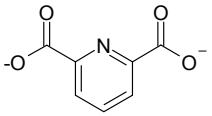
theophylline



Detecting bacteria



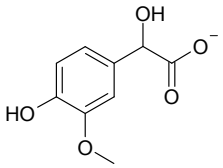
o-aminoacetophenone (*P. aeruginosa*)



dipicolinic acid (*Bacillis*)

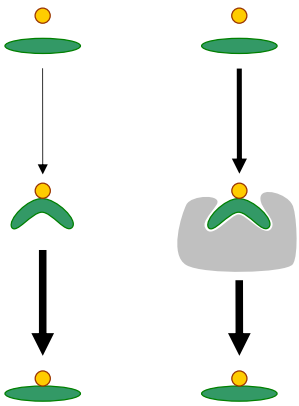
Cancer diagnosis

vanillylmandelic acid (presence in urine indicates pheochromocytoma or neuroblastoma)



Application: Custom enzymes

Binding to the transition state of a reaction catalyzes that reaction:



- Custom proteases and restriction enzymes
- Chemical synthesis
- Degrade toxins, bacterial biofilm matrix

Outline

Protein design technology

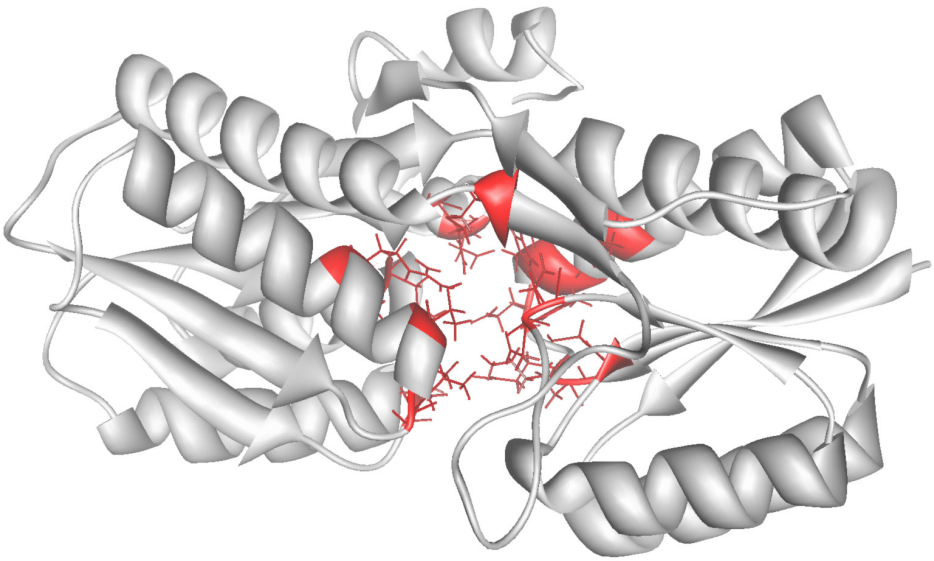
Initial tests

- Structural prediction
- Energetic prediction
- Redesigning a natural binding site

Application: Designing new binding sites

Protein design technology

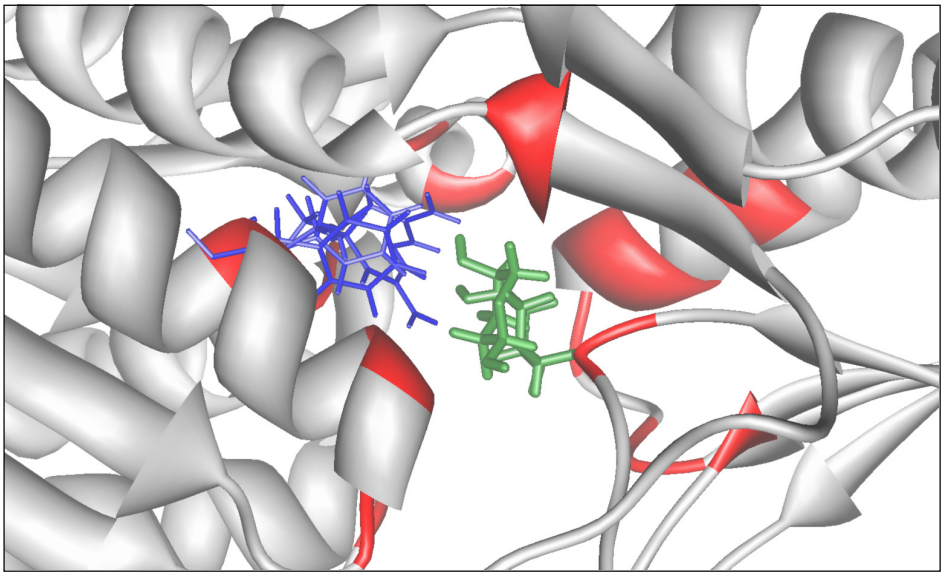
Setting up a design calculation



❶ Pick **scaffold** with known crystal structure; pick **design positions**.

Protein design technology

Setting up a design calculation

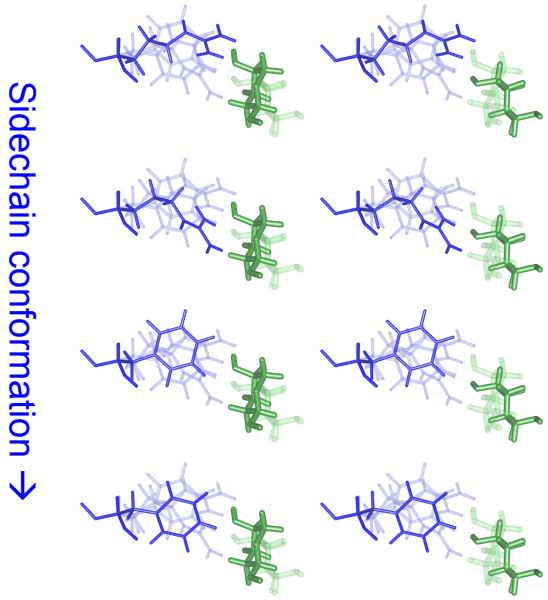


❷ Construct possible **ligand poses** and **sidechain conformations** for each amino acid at each **position**.

Protein design technology

Setting up a design calculation

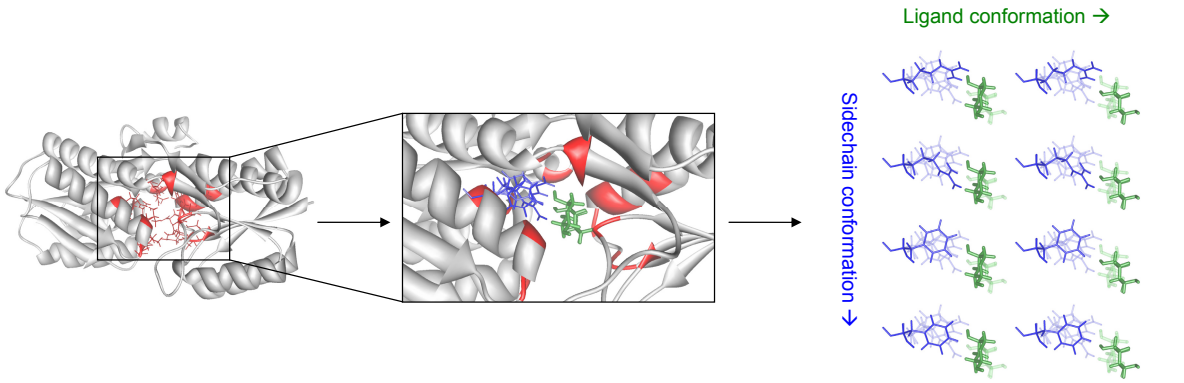
Ligand conformation →



❸ Calculate matrix of interaction energies between possible conformations.

Protein design technology

Setting up a design calculation



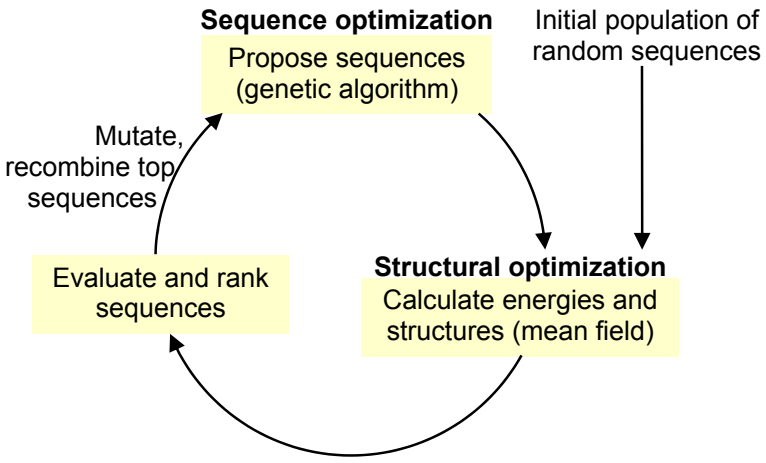
Pick **scaffold** with known crystal structure; pick **design positions**.

Construct possible **ligand poses** and **sidechain conformations** for each amino acid at each **position**.

Calculate matrix of interaction energies between possible conformations.

Protein design technology

Running a design calculation



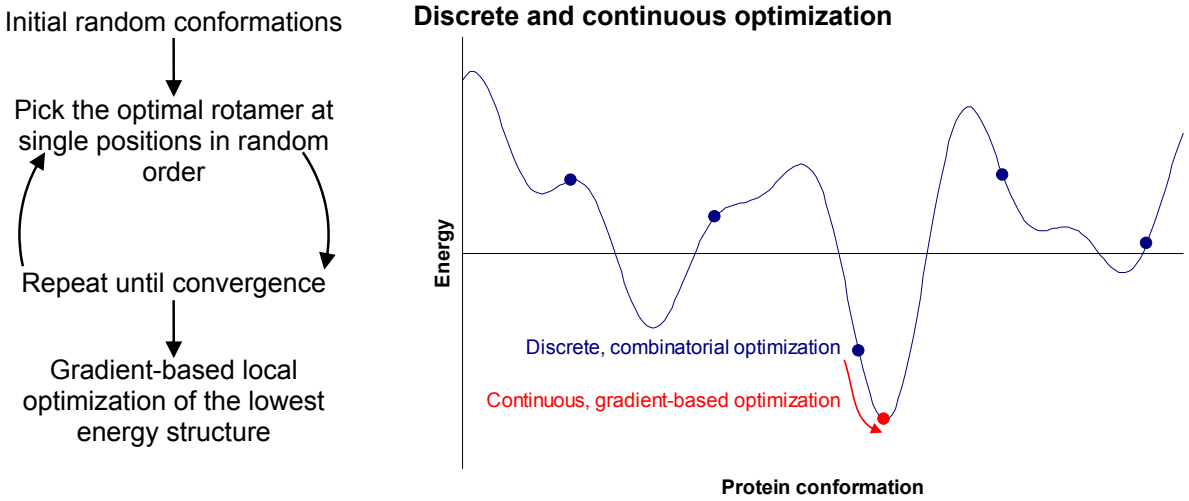
The design procedure involves separate sequence optimization (to find sequences with the desired properties) and structural optimization (to determine the properties of each proposed sequence). In a typical protein design calculation, we search a space of 5000 alternative conformations at 10 positions ≈ 10^{37} possibilities.

Protein design technology

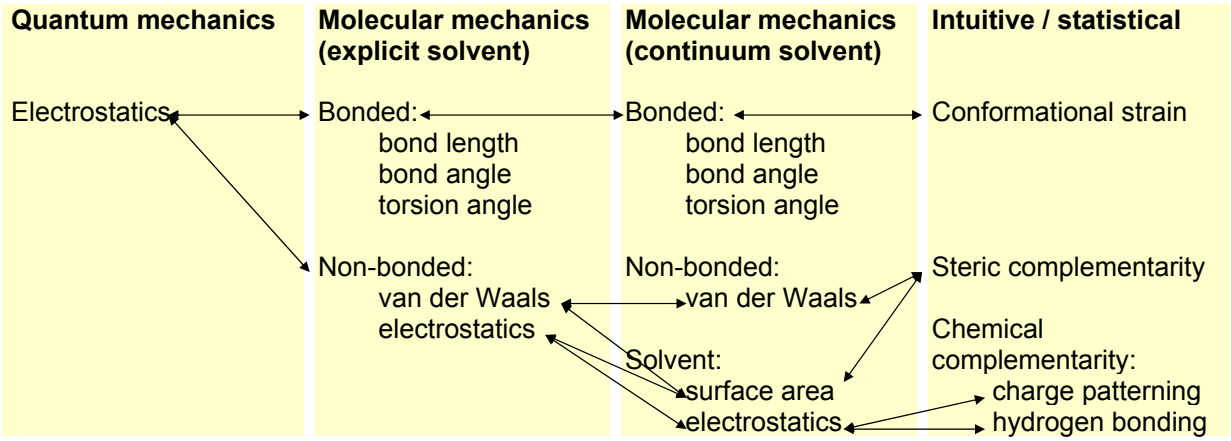
Sequence optimization: Genetic algorithm

Protein design technology

Structural optimization



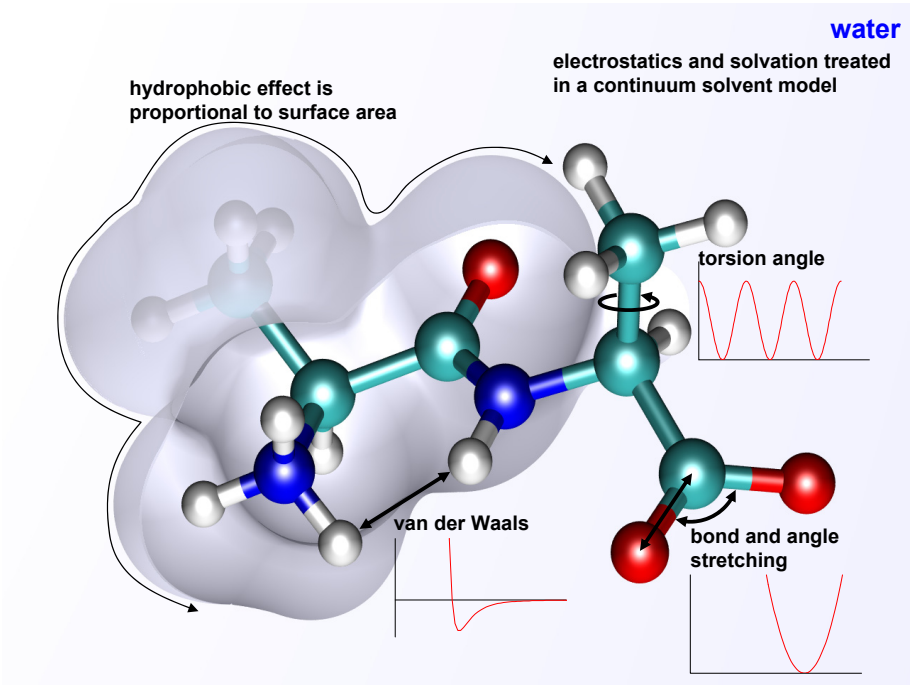
Potential energy functions



Physics-based molecular model

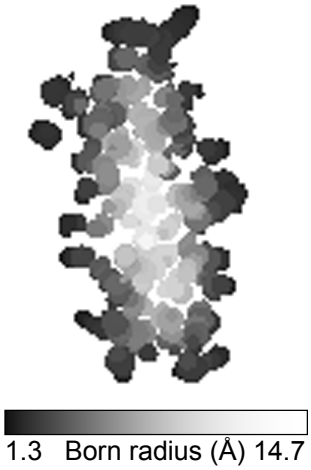
Energy = Molecular mechanics + generalized Born + surface area + protonation energy + entropy

(bond length + angle + torsion + VDW + Coulomb) (interaction with solvent) (hydrophobic effect)



Continuum solvent model

Generalized Born radius



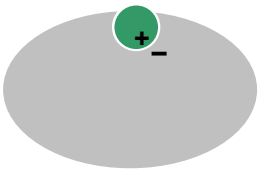
Charged atoms closer to the protein's surface have:

- more favorable solvation energy
- smaller charge-charge interactions

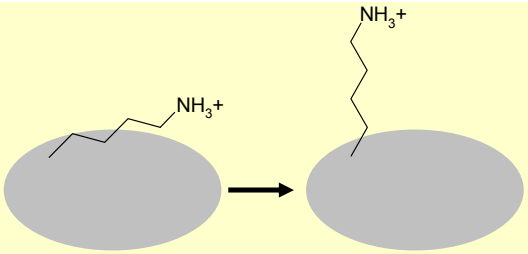
Hydrophobic effect is proportional to exposed surface area.

Examples of protein behaviors treated by our model

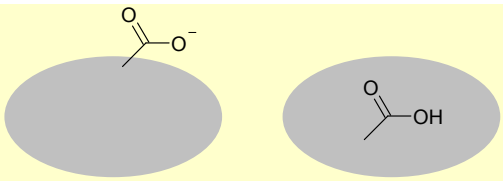
(factors typically ignored in design calculations are highlighted)



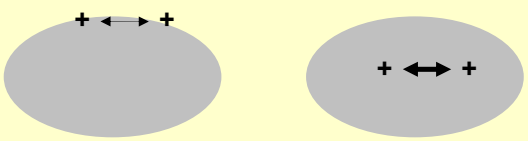
Steric / electrostatic / hydrogen bond complementarity



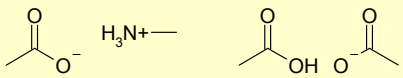
Charges are pulled towards solvent



Asp/Glu buried in a hydrophobic environment tends to be protonated



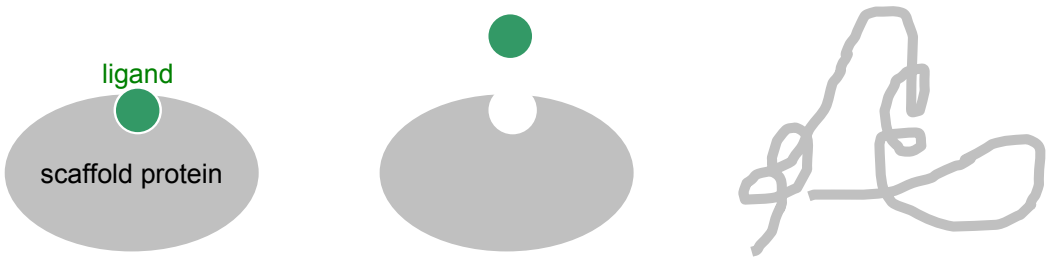
Buried charges interact more strongly than surface charges



Electrostatics affect protonation state

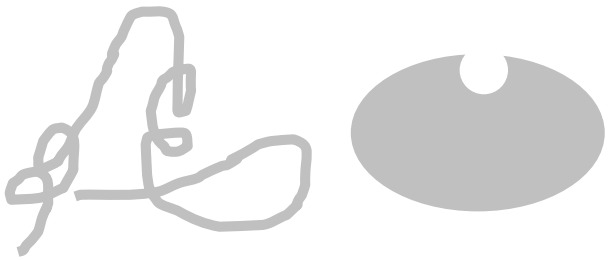
Examples of protein behaviors treated by our model

We explicitly model the bound, unbound, and unfolded states. This allows us to model conformational changes, and also to optimize for stability, binding, and specificity separately.



docked **undocked** **unfolded**

Stability and dissociation energy



Stability = unfolded protein energy – protein energy

Dissociation energy = protein energy + ligand energy – protein·ligand complex energy

Outline

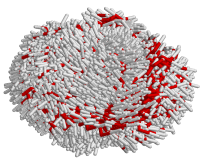
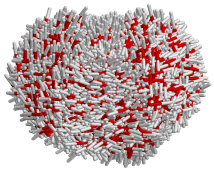
Protein design technology

Initial tests

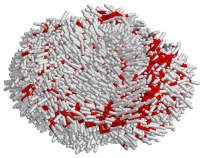
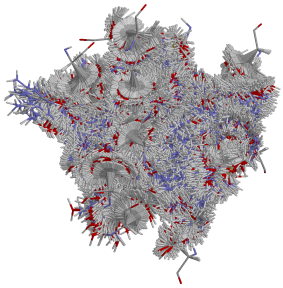
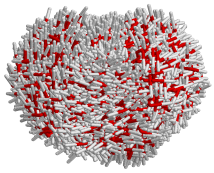
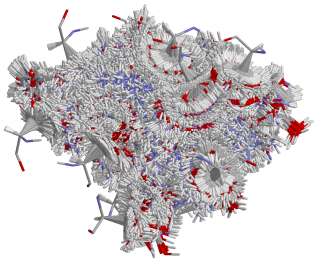
- Structural prediction
- Energetic prediction
- Redesigning a natural binding site

Application: Designing new binding sites

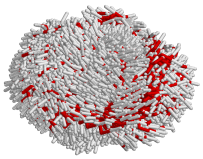
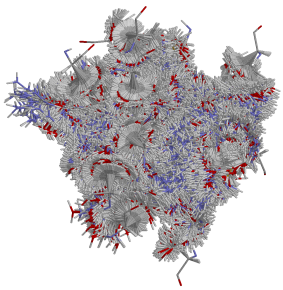
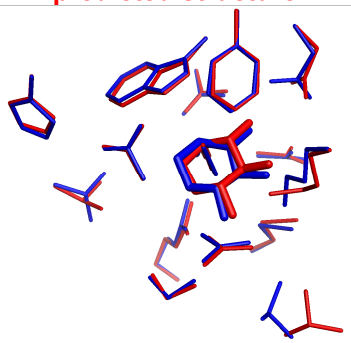
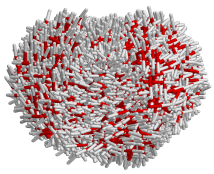
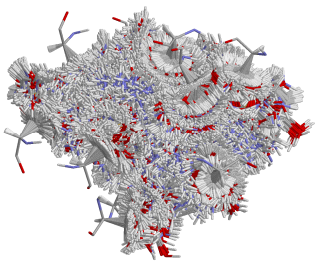
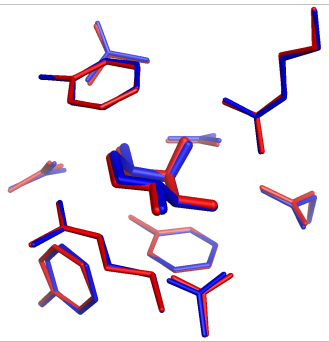
Predicting binding site coordinates

	Ligand poses	Sidechain rotamers	crystal structure / predicted structure	RMS error
ABP-arabinose	 4111 ligand poses			
RBP-ribose	 4639 ligand poses			

Predicting binding site coordinates

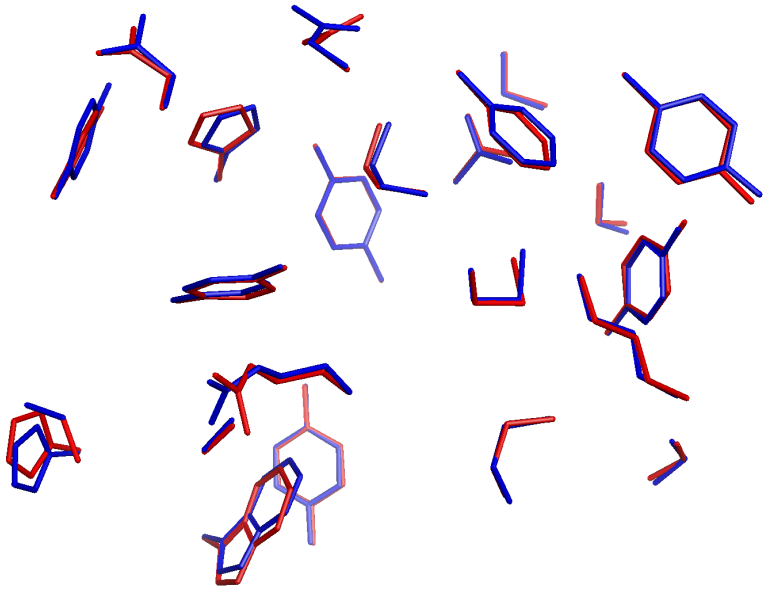
	Ligand poses	Sidechain rotamers	crystal structure / predicted structure	RMS error
ABP-arabinose	 4111 ligand poses	 6028 rotamers / position		
RBP-ribose	 4639 ligand poses	 5449 rotamers / position		

Predicting binding site coordinates

	Ligand poses	Sidechain rotamers	crystal structure / predicted structure	RMS error
ABP-arabinose	 4111 ligand poses	 6028 rotamers / position		0.677 Å
RBP-ribose	 4639 ligand poses	 5449 rotamers / position		0.148 Å

Predicting binding site coordinates

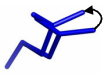
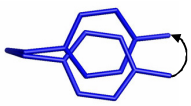
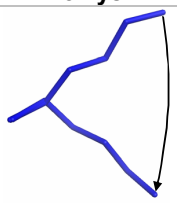
Structural prediction for Avastin-VEGF



Crystal structure (1BJ1) / **Predicted structure**
RMS error: 0.621

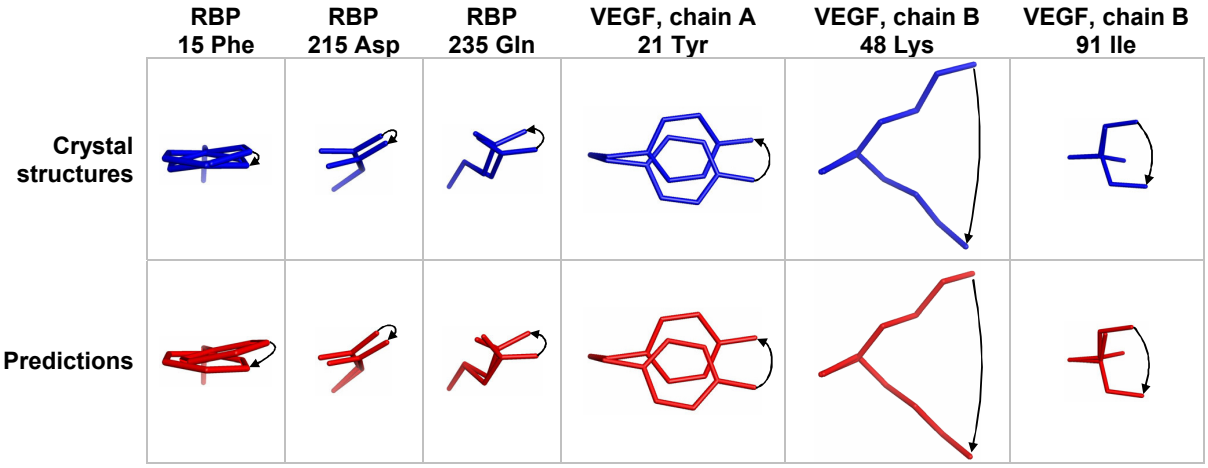
Predicting conformational shifts upon binding

RBP binding ribose
VEGF binding Avastin

	RBP 15 Phe	RBP 215 Asp	RBP 235 Gln	VEGF, chain A 21 Tyr	VEGF, chain B 48 Lys	VEGF, chain B 91 Ile
Crystal structures						
Predictions						

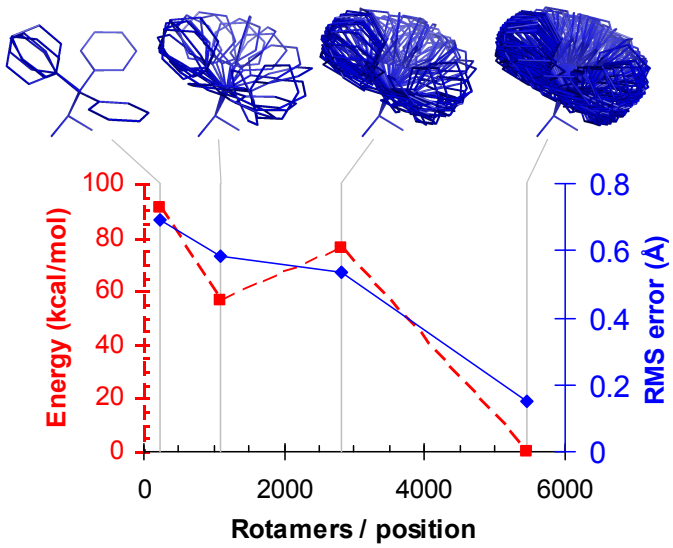
Predicting conformational shifts upon binding

RBP binding ribose
VEGF binding Avastin



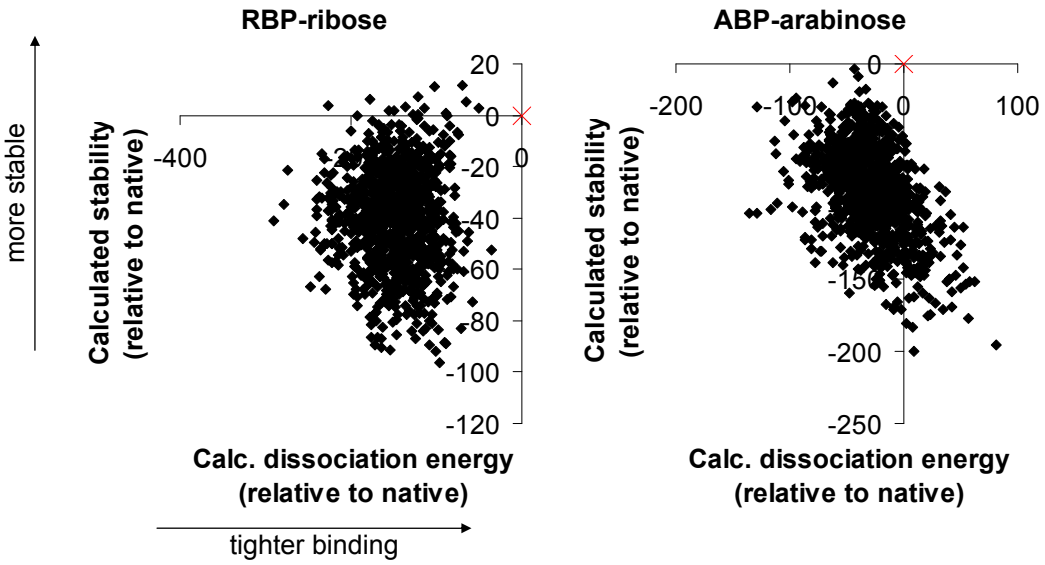
Higher rotamer resolution improves the structural prediction

RBP binding ribose

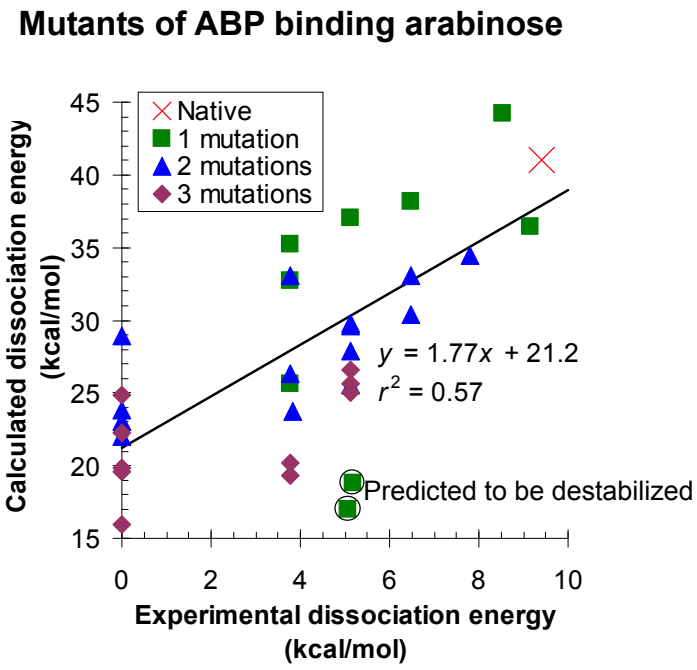


Highest resolution rotamer library has a rotamer within 0.3 Å of 99.9% of rotamers seen in high resolution crystal structures.

Energetic predictions: Scrambled sequences

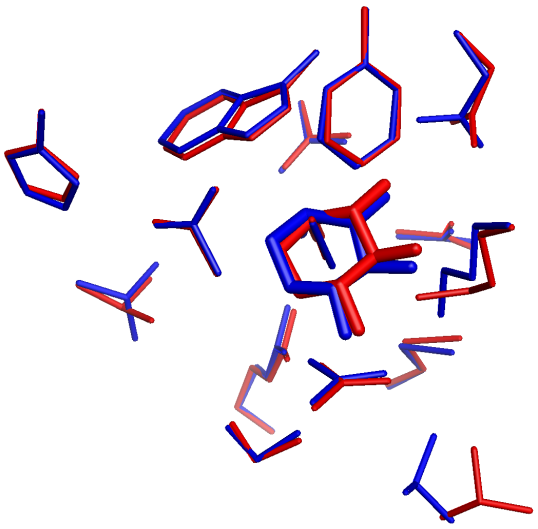


Energetic predictions: Relative binding energies of mutants



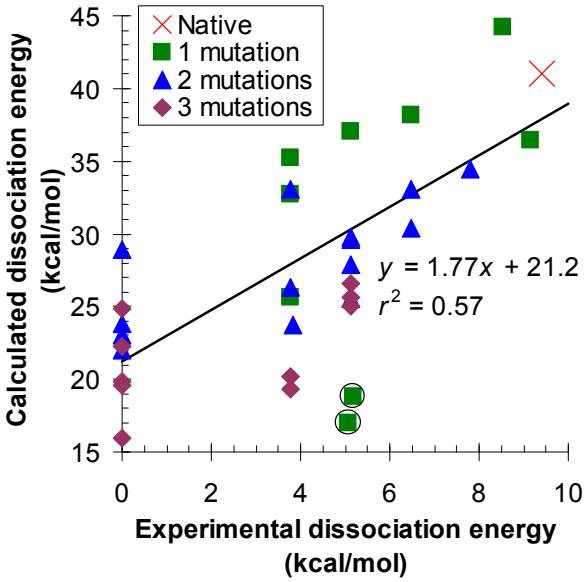
Why are structures easier to predict than energies?

Structural prediction

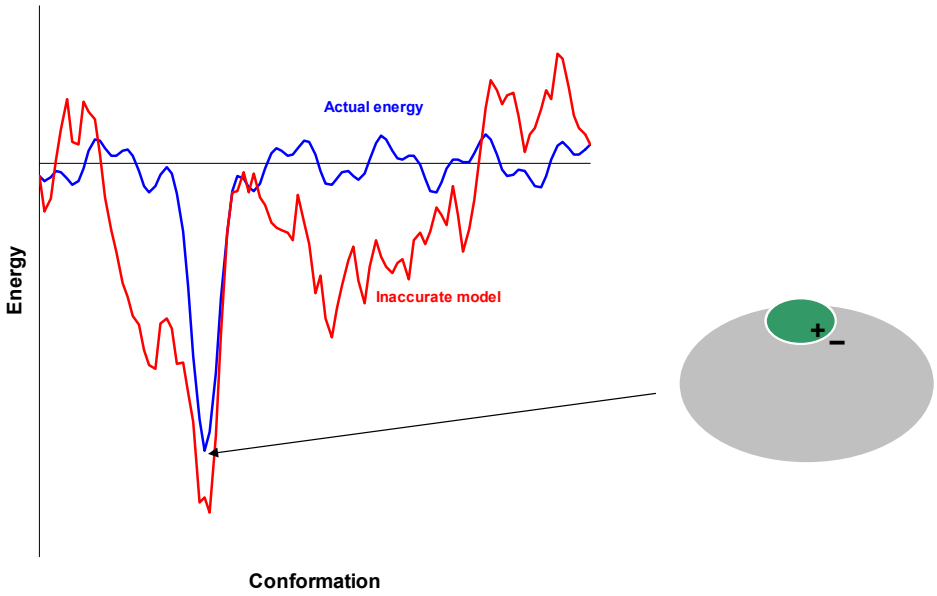


crystal structure / repacked structure

Dissociation energy predictions

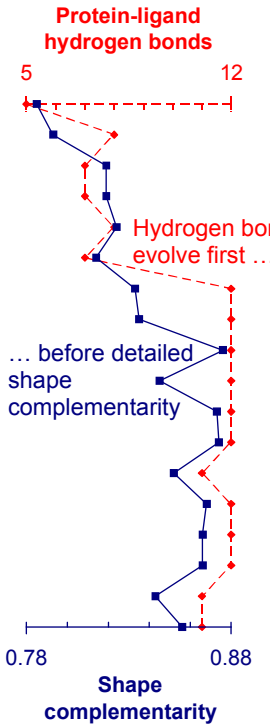


Why are structures easier to predict than energies?



Redesigned ribose binding protein

Seq. #	Best sequence
2709	QFLMRQFNAQ
2764	RALDLRFNSQ
4384	RYEDRRFNAV
4619	RYEDRRFNAV
4640	RYLDARFNAQ
4663	MILDRRFNSQ
5030	MYLDRRFNSQ
5474	LFLDRRFNSQ
5727	QYFDRRFNSQ
5916	MYMDRRYNSQ
5928	QYFDRRYNSQ
6461	QFFDRRFNSQ
6496	SYFDRRYNSQ
6559	NYFDRRYNSQ
8782	NFFDRRFNSQ
8864	NFFDRRFNSQ
8871	NFFDRRFSSQ
native (8964)	NFFDRRFNDQ

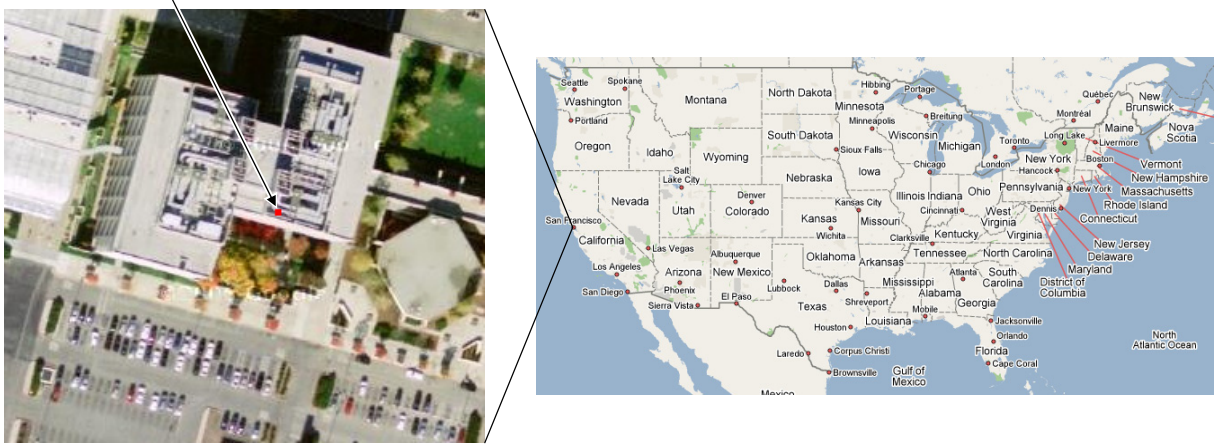


Positions identical to the native are highlighted in yellow.

Redesigned ribose binding protein

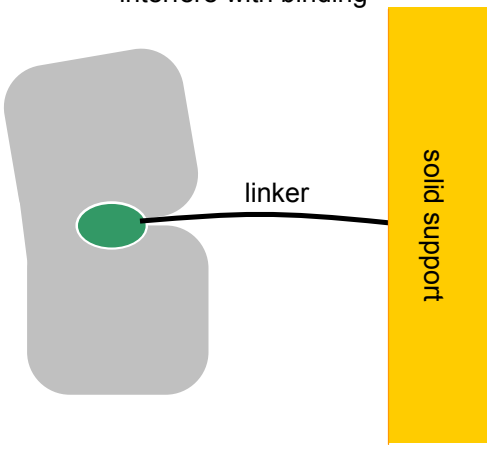
Out of $17^{10} = 2.0 \times 10^{12}$ possible sequences, the design algorithm picked a point mutant of the native as the top sequence, and the native as the second best sequence.

To give an idea of the size of the sequence space considered, this is equivalent to locating a 4.6 m^2 region (red square) in the entire United States.



Protein design vs. in vitro evolution

In vitro evolution requires a linker that can interfere with binding



Protein design models the protein-ligand interaction without a linker

Redesigned ribose binding protein

Essential elements of the design algorithm:

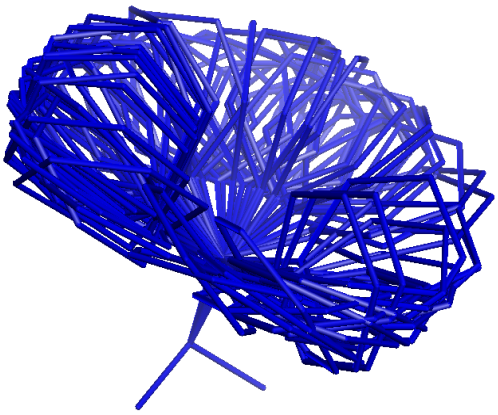
- High resolution rotamer library
- Final gradient-based local minimization step
- Accurate solvation model

Design calc.	Rotamers / position	Local minimization	Solvation model	Rank	Identity to native	K_d (exptl.)	Sequence (10 primary contacts)
A	2800	no	Brooks	1	3	> 330 mM	NMMMIMFNAN
				2	2		NMMMLMFTAN
				3	4		NFMLVMFNAN
B	5449	no	Brooks	1	8	690 mM	NFFDRRFSSQ
				2	9		NFFDRRFNSQ
				3	8		NMFDRRFNSQ
C	5449	yes	Still	1	6	> 900 mM	NYYDRRYNAQ
				2	6		NYMDRRYNSQ
				3	7		NYFDRRYNAQ
★ D	5449	yes	Brooks	1	9	17.2 μM	LFFDRRFNDQ
				2	10	210 nM	NFFDRRFNDQ
				3	9		NTFDRRFNDQ
Native					10	210 nM	NFFDRRFNDQ

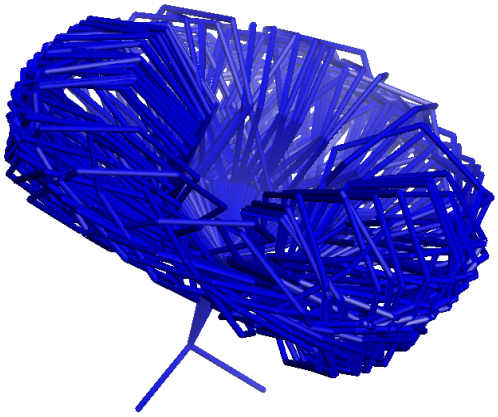
Note: The Brooks solvation model is only 2% off from the Poisson-Boltzmann equation. The Still solvation model is less accurate.

High resolution is critical for protein design

2800 rotamers / position
(only Phe shown)



5449 rotamers / position
(only Phe shown)



Outline

Protein design technology

Initial tests

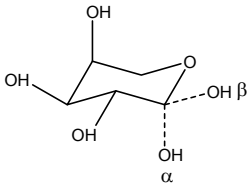
- Structural prediction
- Energetic prediction
- Redesigning a natural binding site

Application: Designing new binding sites

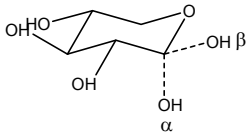
Design targets

Redesign ribose binding protein to bind …

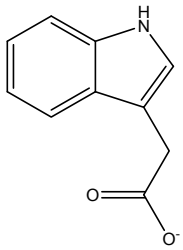
L-arabinose



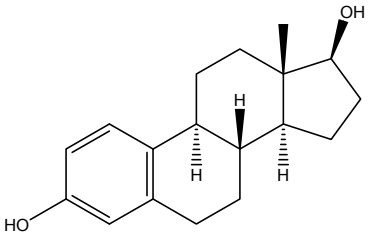
D-xylose



indole-3-acetic acid (IAA)



estradiol

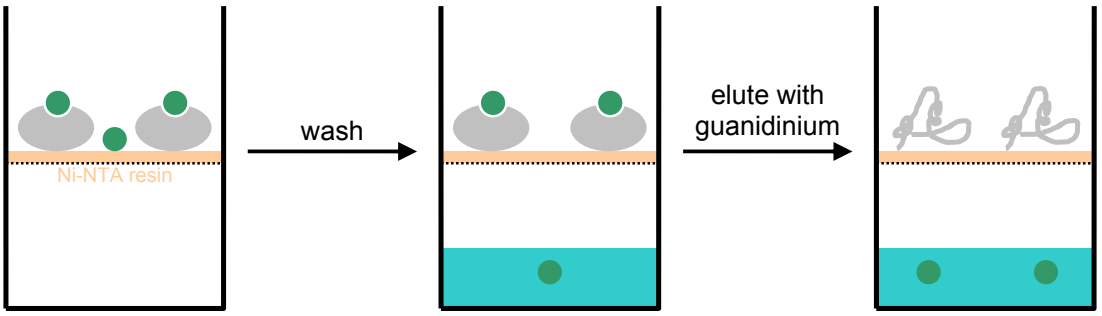


Designed binding proteins

	9	13	15	16	89	90	103	105	141	164	190	215	235
RBP (native)	SER	ASN	PHE	PHE	ASP	ARG	SER	ASN	ARG	PHE	ASN	ASP	GLN
RBP→arabinose 2	GLN	ASN	MET	TYR	VAL	MET	GLN		MET	PHE	ASN	SER	VAL
RBP→xylose 1		ASN	PHE	PHE	GLN	GLN			MET	PHE	ASN	SER	MET
RBP→xylose 2		MET	TYR	PHE	GLN	HIS			MET	PHE	ASN	SER	GLN
RBP→estradiol 4	SER	ASN	VAL	MET	ALA	ASN	ASN		MET	PHE	ASN	SER	ILE
RBP→IAA 1		ARG	THR	MET	VAL	MET	HIS	TYR	MET	PHE	ASN	ALA	SER
RBP→IAA 2		ARG	THR	MET	ALA	MET	HIS	TYR	MET	PHE	ASN	SER	SER
RBP→IAA 3		ARG	THR	MET	VAL	ASN	HIS	TYR	MET	PHE	ASN	ALA	SER

Sequence is only shown at positions being designed.

Solid phase radioligand binding assay

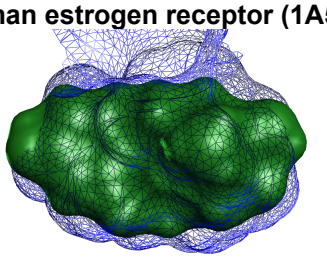


Experimental characterization of designed binding proteins

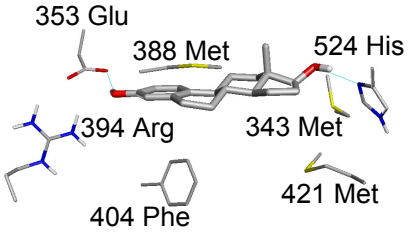
Protein	Ligand	K_d	Protein stability (kcal/mol)
RBP→arabinose 2	arabinose	250 mM	
RBP→xylose 1	xylose	160 mM	4.2
RBP→xylose 2	xylose	270 mM	
RBP→estradiol 4	estradiol	46 mM	2.0
RBP→IAA 1	IAA	11 mM	2.5
RBP→IAA 2	IAA	14 mM	1.0
RBP→IAA 3	IAA	16 mM	1.9

Designed estradiol binding protein

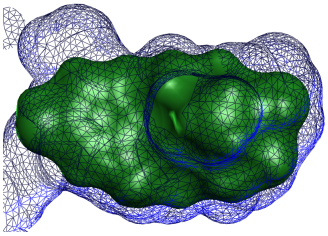
human estrogen receptor (1A52)



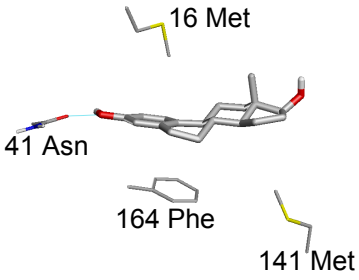
Shape complementarity = 0.72



RBP→estradiol 4

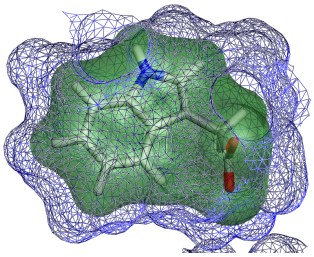


Shape complementarity = 0.75

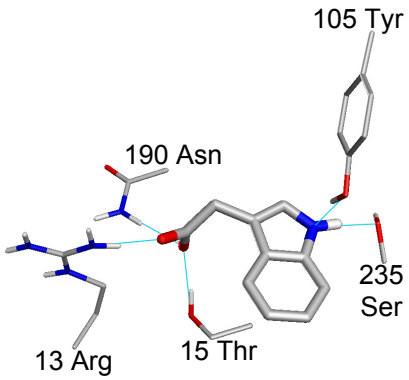


Designed indole-3-acetic acid binding protein

RBP→IAA 1

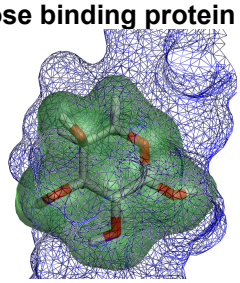


Shape complementarity = 0.81



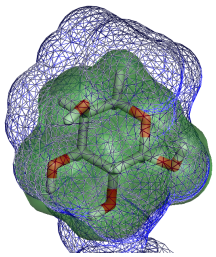
Designed arabinose binding protein

arabinose binding protein (1ABE)

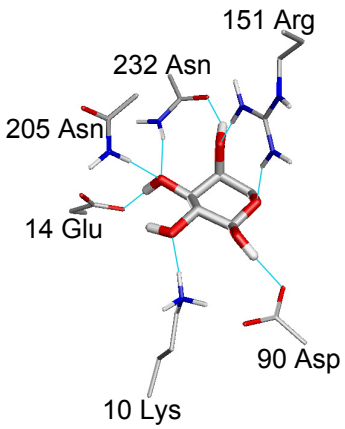


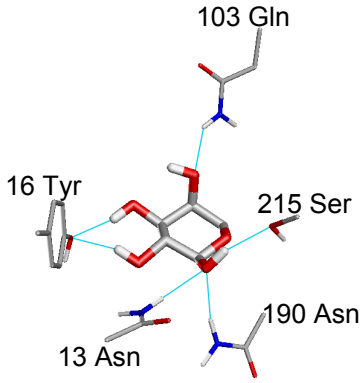
Shape complementarity = 0.81

RBP→arabinose 2



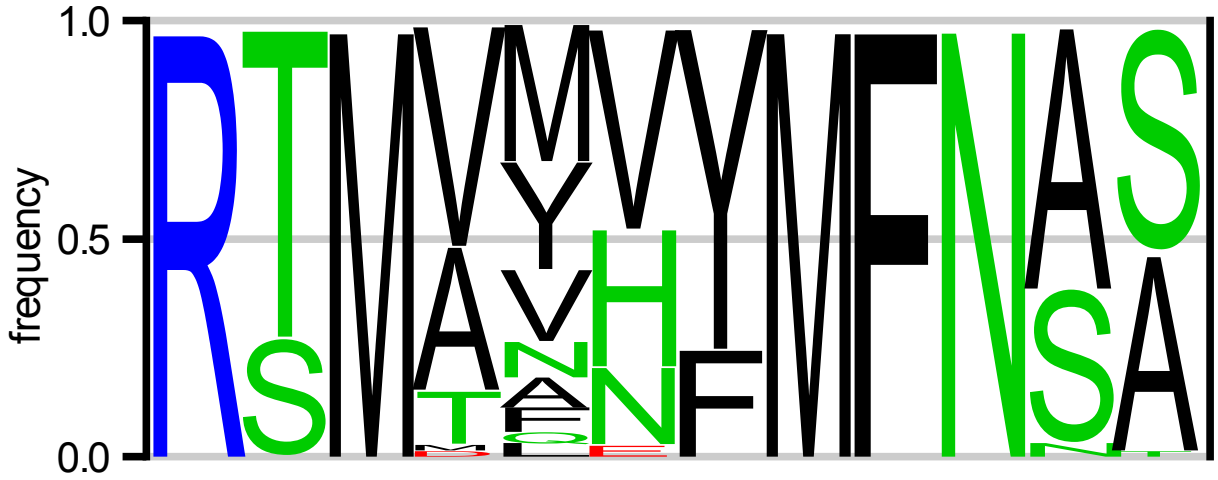
Shape complementarity = 0.79





RBP→IAA library

We generated a library of RBP variants based on amino acid frequencies in the top 72 sequences from the RBP→IAA design.

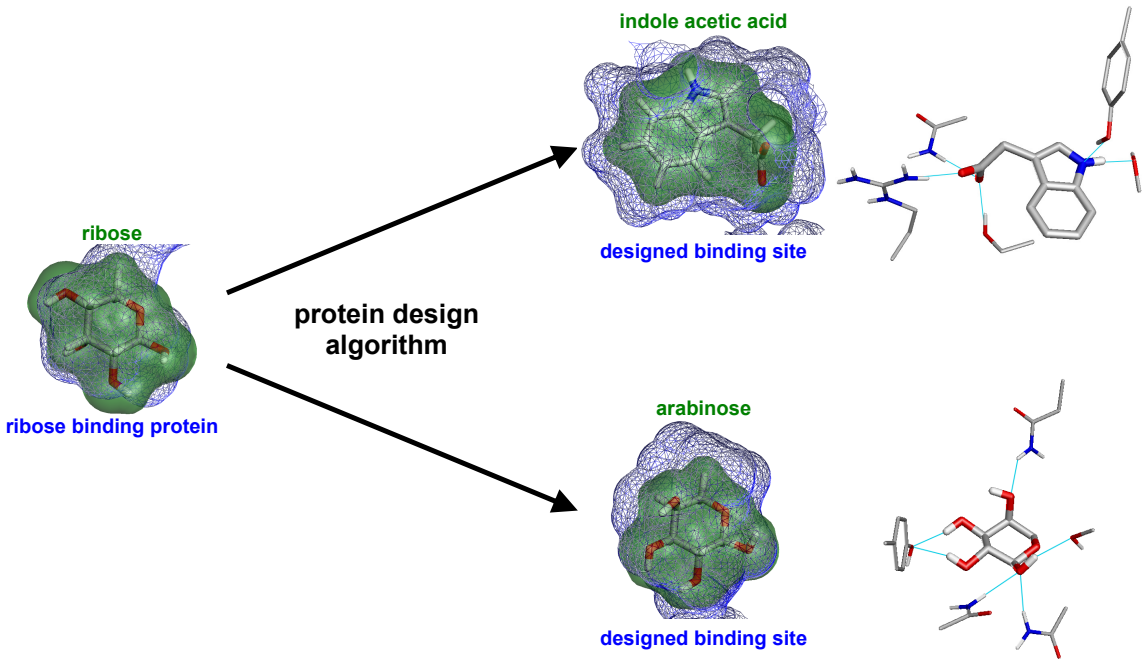


RBP→IAA library

We screened 279 sequences from the library for binding to IAA, and found sequences with 10-fold improved affinity 10-fold (highlighted):

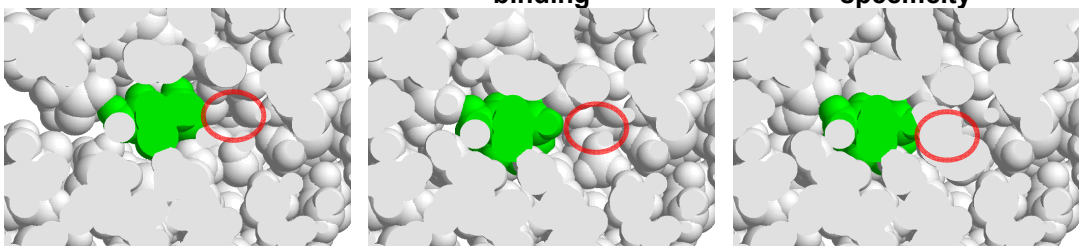
Protein	K_d (IAA)	13	15	16	89	90	103	105	141	164	190	215	235
RBP→IAA 1	11 mM	ARG	THR	MET	VAL	MET	HIS	TYR	MET	PHE	ASN	ALA	SER
RBP→IAA 2	14 mM	ARG	THR	MET	ALA	MET	HIS	TYR	MET	PHE	ASN	SER	SER
RBP→IAA 3	16 mM	ARG	THR	MET	VAL	ASN	HIS	TYR	MET	PHE	ASN	ALA	SER
RBP→IAA 101A-F11	1.4 mM	ARG	SER	MET	GLY	CYS	HIS	TYR	MET	PHE	ASN	ALA	SER
RBP→IAA 95A-C1	1.1 mM	ARG	SER	MET	ILE	CYS	HIS	TYR	MET	PHE	ASN	ALA	SER

Redesigning binding sites



Designing for specificity

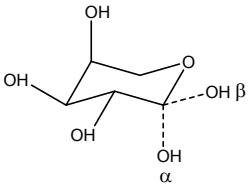
native | **design for arabinose binding** | **design for arabinose specificity**



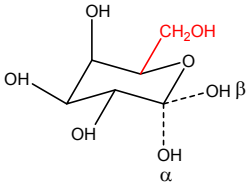
protein / arabinose

Space for galactose CH_2OH is seen in the native and arabinose binding design (both of which bind galactose), but not in the specificity design.

arabinose

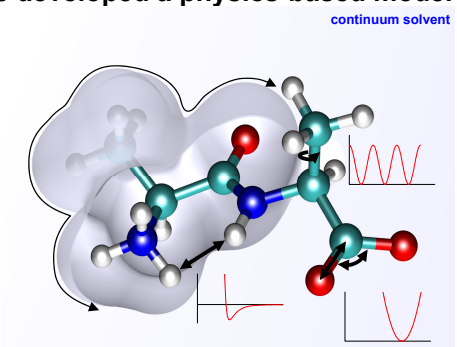


galactose

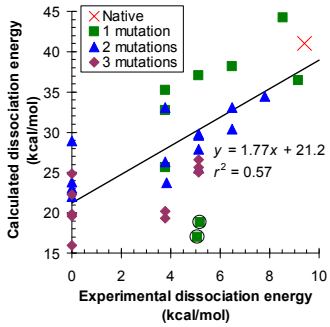


Summary

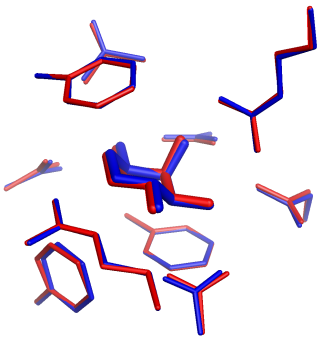
We developed a physics-based model …



that predicts binding constants …

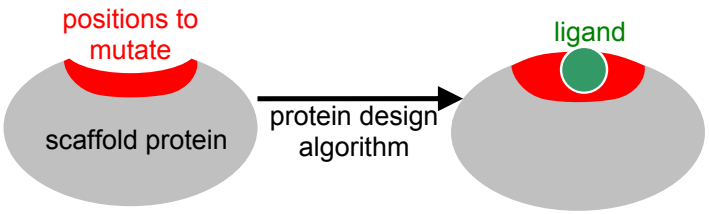


structures …



crystal structure / **predicted structure**

and designs new binding sites …



Acknowledgements

Harbury Lab
Erica Raffauf
Jim Havranek
Jarrett Wrenn
Lance Martin
Rebecca Weisinger
Becky Fenn
Kierstin Schmidt

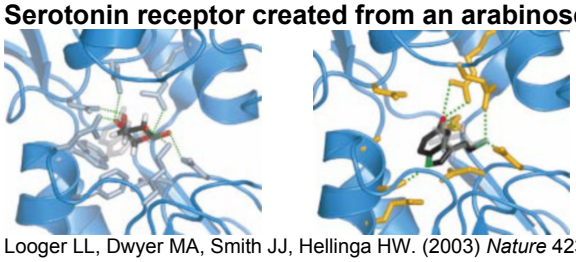
Dan Herschlag
Axel Brunger
Tom Wandless

Michael Levitt
Buzz Baldwin
Loren Looger
Pavel Strop

End

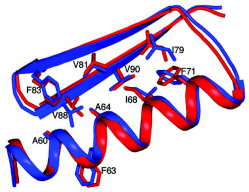
Examples of computational protein design

Serotonin receptor created from an arabinose receptor



Looger LL, Dwyer MA, Smith JJ, Hellinga HW. (2003) *Nature* 423: 185-90.

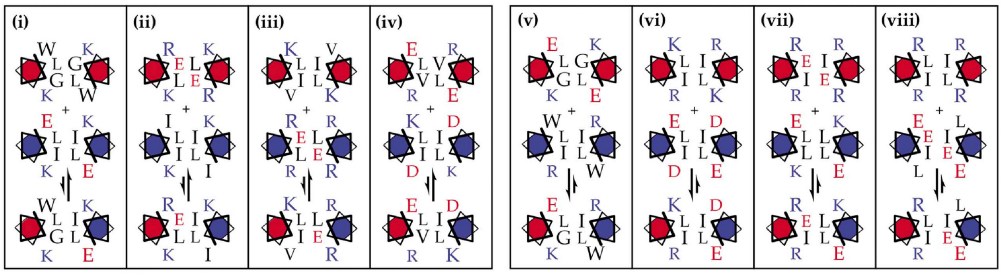
Protein with a new α/β fold



Calculated / Observed

Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. (2003) *Science*. 302:1364-8.

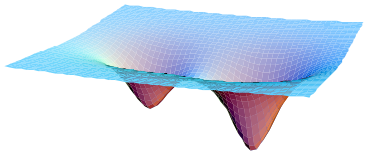
Specific coiled-coil interactions



Havranek JJ, Harbury PB. (2003) *Nature structural biology*. 10(1): 45-52.

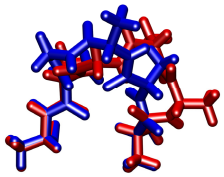
Finding a protein's low energy conformations

Molecular dynamics is slow because crossing large energetic barriers is a rare event. Can we skip these barriers and just find the low energy conformations?



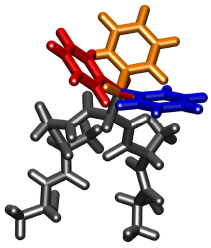
Low energy loop conformations

- Database search
- Systematic search
- Randomly perturb and splice together existing conformations



Low energy sidechain conformations

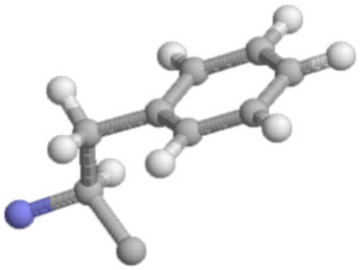
- Rotamers seen in crystal structures



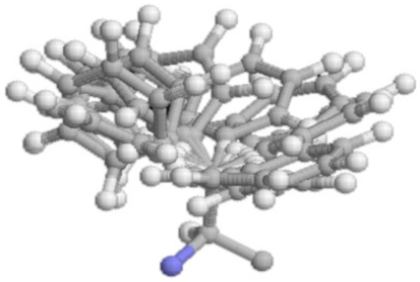
Rotamer library

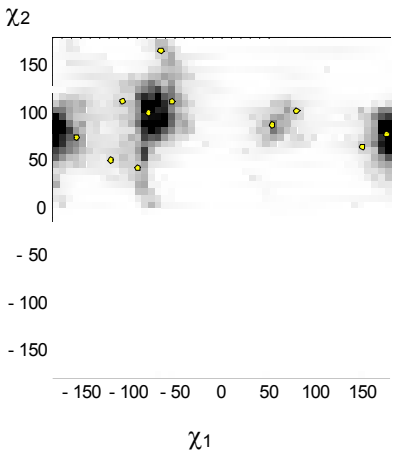
Example: Phenylalanine

Most common rotamer



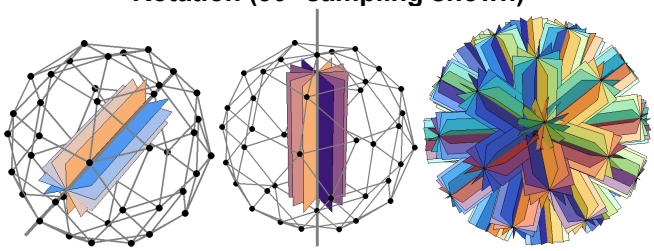
All 11 rotamers



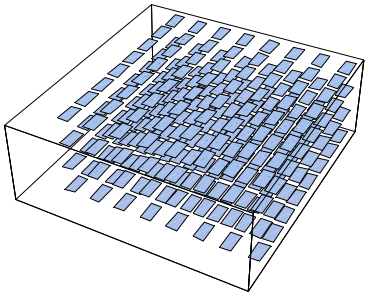


Ligand sampling and filters

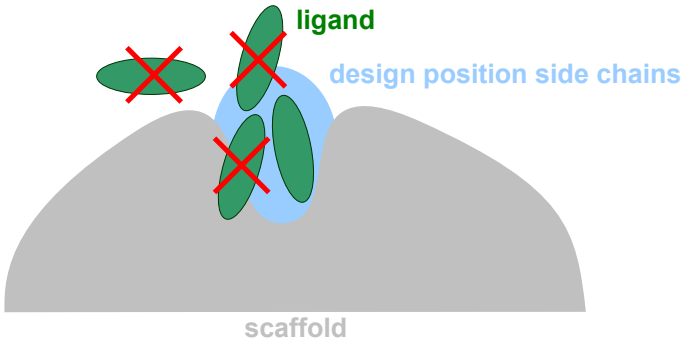
Rotation (30° sampling shown)



Translation



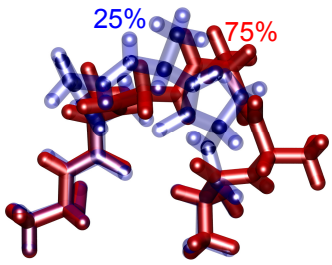
Eliminate ligand orientations that clash with the scaffold, or do not make sufficient contact with design position side chains



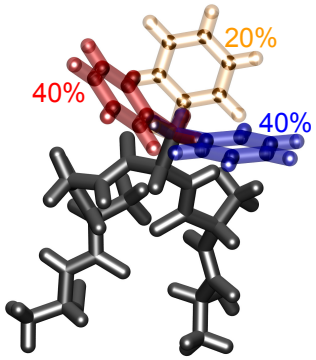
Probabilistic description of protein conformation

We represent the protein/ligand system as a probabilistic ensemble of different backbone, side chain, and ligand conformations. This allows us to model conformational changes and thermal fluctuations.

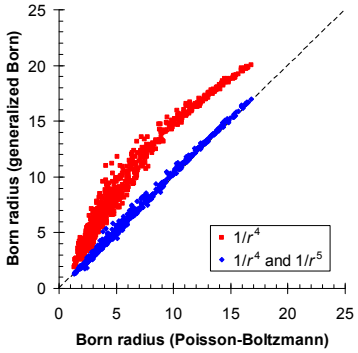
Loop conformations



Sidechain conformations

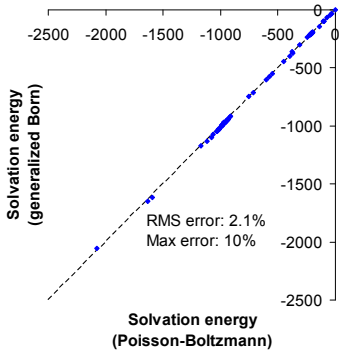


Generalized Born solvation energy

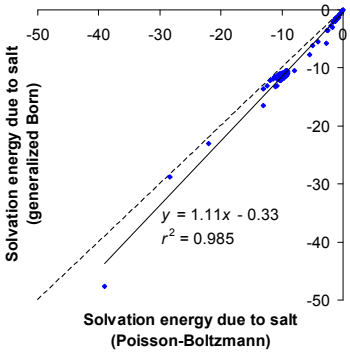


Brooks' empirical expression with a $1/r^5$ term gives much more accurate Born radii for atoms in a test protein (protein tyrosine phosphatase 1B).

No salt

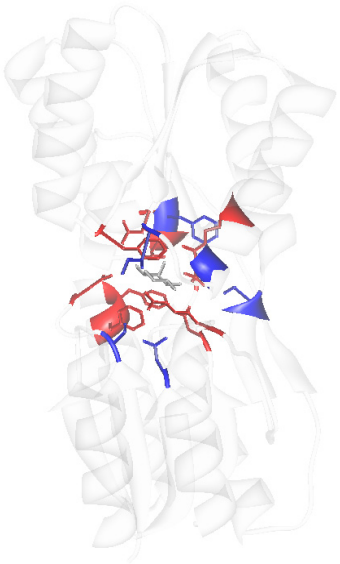


100 mM salt



For a set of small molecules, peptides, and proteins, solvation energy calculated using the $1/r^5$ term closely matches the gold-standard Poisson-Boltzmann equation (each point corresponds to a single structure).

Design positions for RBP



Primary contacts:
13,15,16,89,90,141,164,190,215,235
Secondary contacts:
9,64,103,132,137,214

For ribose binding protein, we only picked primary contacts.

Primary contacts with ligand:

h-bonding	13	Asn
	89	Asp
	90	Arg
	141	Arg
	190	Asn
	215	Asp
	235	Gln
hydrophobic	15	Phe
	16	Phe
	164	Phe

Computational point mutagenesis

Predicted dissociation energy (kcal/mol, relative to native) of RBP-ribose

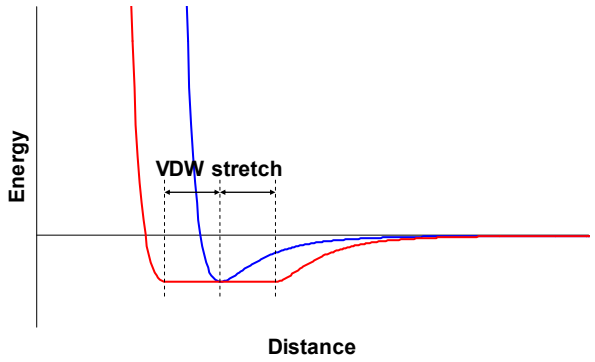
		Residue											
		9	13	15	16	89	90	103	141	164	190	215	235
Native sequence		SER	ASN	PHE	PHE	ASP	ARG	SER	ARG	PHE	ASN	ASP	GLN
Mutation	ALA		-3.5		-9.8	-50	-24	-0.1				-9.9	-20
	ARG	-71	-48	-37			0	-52	0				-37
	ASN	-20	0	-11	-13	-30	-25	-17			0	-16	-30
	ASP			-12	-14	0	-19					0	-29
	GLN	-21	-4.8		-5.4	-44	-32	-36			-22		0
	GLU		-15	-13		-44		-30			-44	-30	
	HIS	-17			-17	-62	-64	3.46			-13	-75	-36
	ILE		-8.2	-20		-52	-24	-0.2				-41	-27
	LEU		1.57		-5.2	-67	-28					-40	-22
	LYS		-39	-47	-47	-115	-44	-82					-36
	MET		-4.1		-1.2	-53	-29						
	PHE		-18	0	0	-101	-60			0		-109	
	SER	0	-2.4		-11	-44	-26	0				-5.3	-22
	THR			-0.6		-55		1			-11	-19	-27
	TRP		-26			-160	-72						
	TYR		-24	-22	-14	-129	-65				-48	-81	
	VAL		-3			-56	-17					-23	-23

Stability reduced by more than 5 kcal/mol
1.57 High affinity mutant
-1.2 Affinity reduced by more than 5 kcal/mol

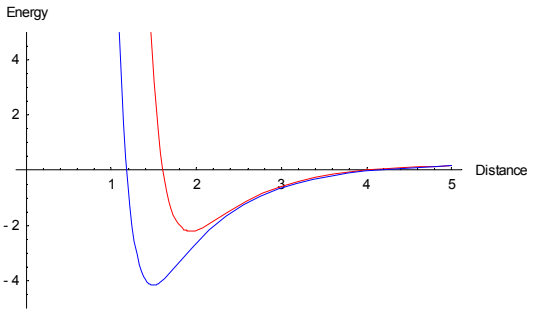
Effect of softening the van der Waals energy

The van der Waals energy is frequently softened so as not to penalize the small steric clashes resulting from limited sampling resolution. A side effect of this is to make hydrogen bonds appear stronger than they actually are.

VDW stretch



C=O···H-N backbone hydrogen bond energy using CHARMM22 parameters



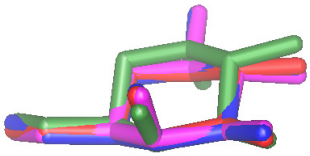
— VDW stretch = 0.0 Å, energy = –2.2 kcal/mol, distance = 1.9 Å
— VDW stretch = 0.3 Å, energy = –4.1 kcal/mol, distance = 1.5 Å

Effect of softening the van der Waals energy

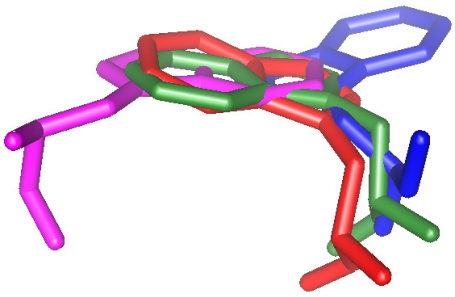
A softened van der Waals energy encourages the design algorithm to bury charges and polar residues at a designed hydrophobic interface:

	13	15	16	89	90	141	164	190	215	235
Designed estradiol binding site in RBP (VDW stretch = 0.3)	ARG	GLU	MET	SER	ALA	MET	PHE	ASN	SER	SER
	ARG	GLU	MET	TYR	SER	MET	TYR	ASN	GLN	ASN
	ARG	GLU	LEU	ALA	LEU	ILE	PHE	ASN	ASN	SER
	ARG	GLU	THR	ALA	ASN	MET	ASN	GLU	ALA	ALA
	ARG	GLU	THR	ALA	ASN	MET	ASN	ASN	ASN	VAL
	HIS	GLU	LEU	MET	ASN	GLU	PHE	ASN	GLU	THR
	ARG	GLU	MET	TYR	SER	MET	TYR	ASP	GLN	ASN
	ARG	ASP	ALA	MET	ASN	MET	ASN	ASN	ASN	THR
	ARG	GLU	THR	ALA	ASN	MET	ASN	GLU	ASN	ALA
	ARG	PHE	LEU	ALA	THR	MET	PHE	ASN	SER	VAL
Designed estradiol binding site in RBP (VDW stretch = 0.0)	ASN	ALA	MET	ALA	ASN	MET	PHE	ASN	ALA	ALA
	ASN	VAL	MET	ALA	ASN	MET	PHE	ASN	ALA	ALA
	ASN	VAL	MET	ALA	ASN	MET	PHE	ASN	ALA	SER
	ASN	SER	MET	ALA	ASN	MET	PHE	ASN	ALA	ALA
	ASN	VAL	MET	SER	ASN	MET	PHE	ASN	ALA	ALA
	MET	VAL	MET	ALA	ALA	MET	PHE	ASN	ALA	ALA
	ASN	VAL	MET	ALA	ASN	MET	PHE	ASN	SER	ALA
	ASN	ILE	MET	ALA	ASN	MET	PHE	ASN	ALA	ALA
	SER	VAL	MET	ALA	ASN	MET	PHE	ASN	ALA	ALA
	ASN	VAL	LEU	ALA	ASN	MET	PHE	ASN	ALA	ALA

Tryptophan stacking in galactose binding proteins



galactose

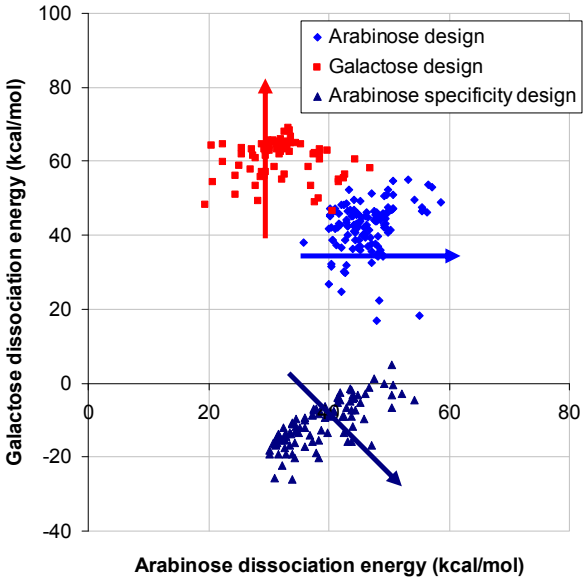


tryptophan

natural proteins:
- Arabinose binding protein (5ABP), K_d = 231 nM
- Galactose/glucose binding protein (1GLG), K_d = 310 nM
- C-type lectin (1TLG), K_d = 443 μM

designed protein:
- ABP→galactose 4 (model)

Designing for specificity



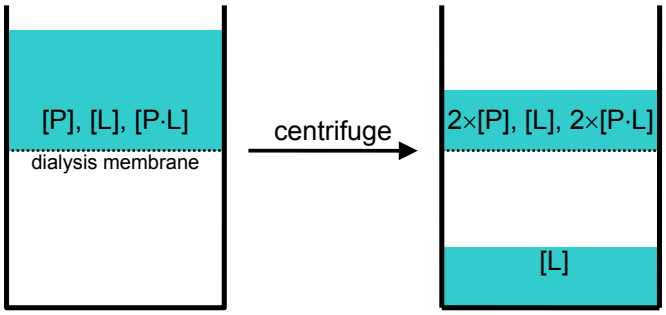
(Each point represents a single sequence)

Assays for measuring K_d

Technique	Protein concentration	Limitations
tryptophan fluorescence	$< 10 \times K_d$	Small signal, can't use absorbant ligand
calorimetry	$5 – 500 \times K_d$	ΔH must be non-zero
spin concentrator	$> K_d$	Need radioligand
solid phase binding assay	40 μM to detect mM binders	Need radioligand

Assays for measuring K_d

Spin concentrator
(quick alternative to radioligand equilibrium dialysis)



Future directions

Scaffold selection

Potential energy function

- explicit water
- lone pairs
- quantum effects
- better hydrogen bond model
- protein polarization

Structural sampling

- backbone flexiblity
- more design positions

Future application: Therapeutic proteins

Limitations of monoclonal antibodies

- Unable to bind to some targets, such as deep grooves in proteins
- Some antibodies are unstable or aggregation-prone
- Non-human antibodies are immunogenic in humans
- Therapeutic antibodies are glycosylated and thus more expensive to manufacture
- Large size of whole antibodies may limit their tissue distribution