

*Structure-based
Protein Engineering*

Hongmin Zhang

Outlines

I. General introduction

- ❖ Protein structure and function
- ❖ Application of protein engineering

II. Strategies and approaches of protein engineering research

- ❖ Directed evolution
- ❖ Rational design
- ❖ De novo design

III. Rational design and De novo design in research and industrial applications

- ❖ Stability improvement
- ❖ Activity switch and enhancement
- ❖ Enzyme specificity

Proteins

Proteins

- are polymers constructed from amino acid monomers,
- account for more than 50% of the dry weight of most cells,
- perform most of the tasks required for life, and form enzymes, chemicals that change the rate of a chemical reaction without being changed in the process.

MAJOR TYPES OF PROTEINS

**Structural Proteins
(provide support)**



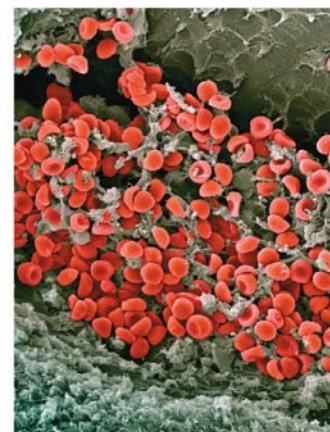
**Storage Proteins
(provide amino acids for growth)**



**Contractile Proteins
(help movement)**



**Transport Proteins
(help transport substances)**

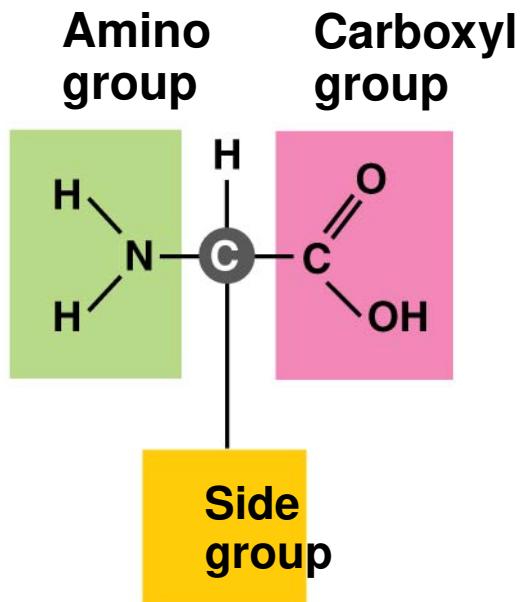


**Enzymes
(help chemical reactions)**

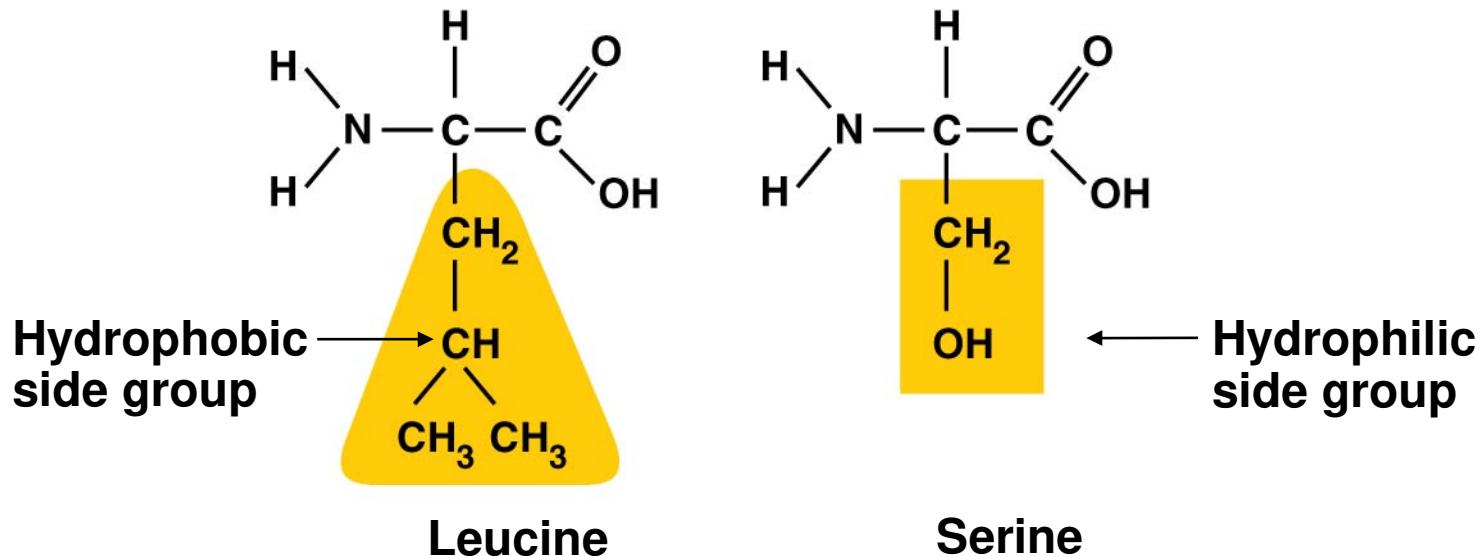


The Monomers of Proteins: Amino Acids

- All proteins are macromolecules constructed from a common set of 20 kinds of amino acids.
- Each amino acid consists of a central carbon atom bonded to four covalent partners.
- Three of those attachment groups are common to all amino acids:
 - a carboxyl group (-COOH),
 - an amino group (-NH_2), and
 - a hydrogen atom.



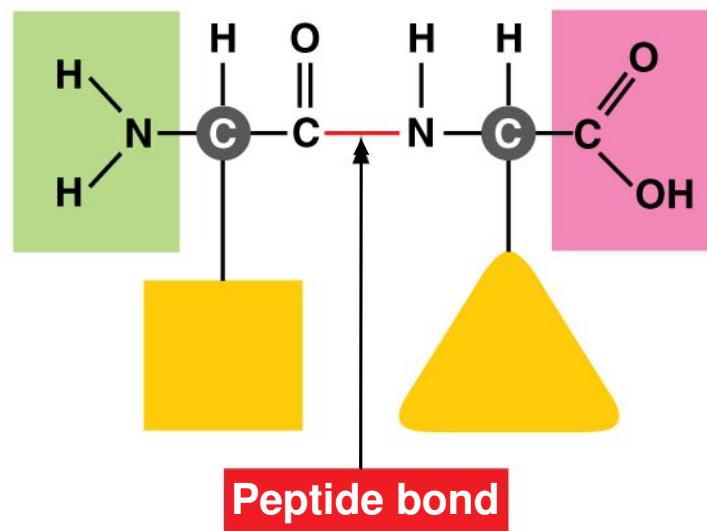
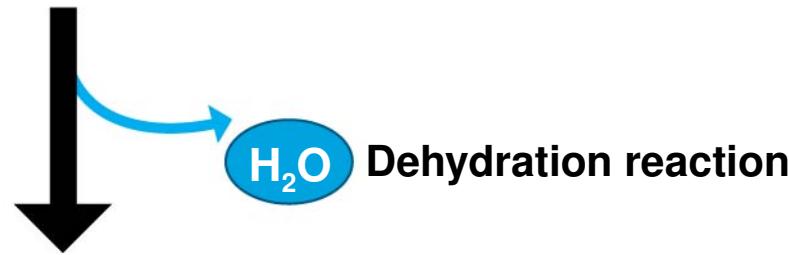
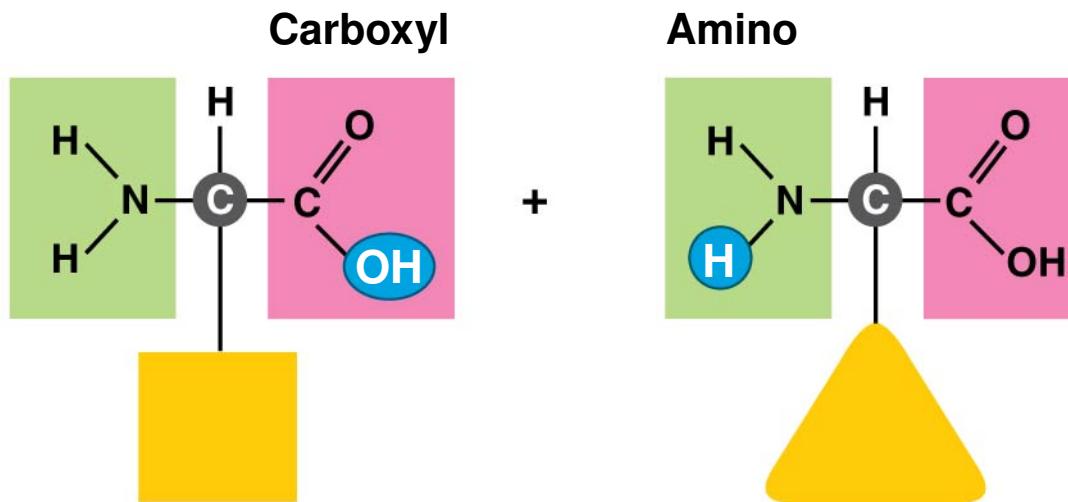
The general structure of an amino acid

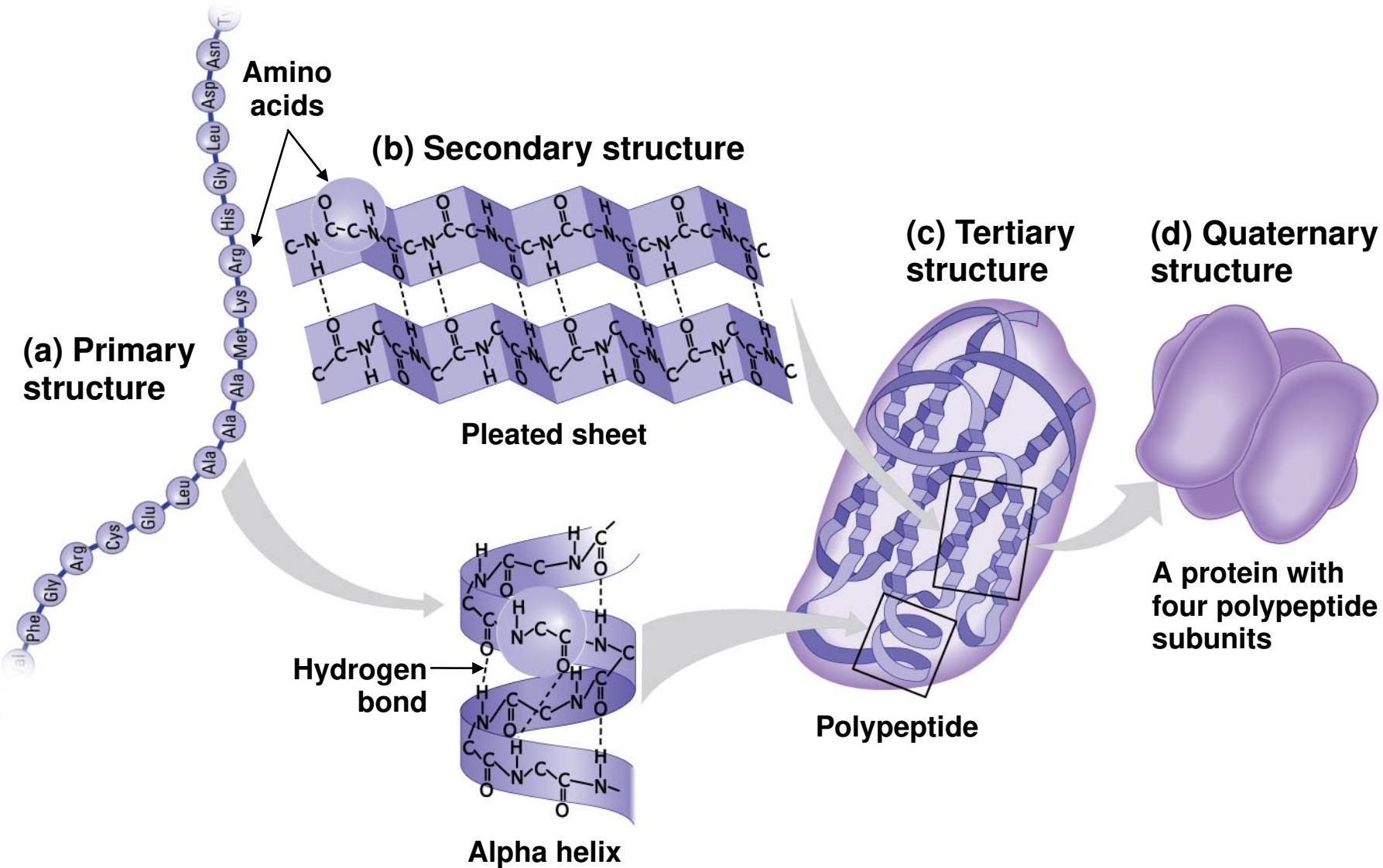


Proteins as Polymers

Cells link amino acids together

- *by dehydration reactions,*
- *forming peptide bonds, and*
- *creating long chains of amino acids called polypeptides.*

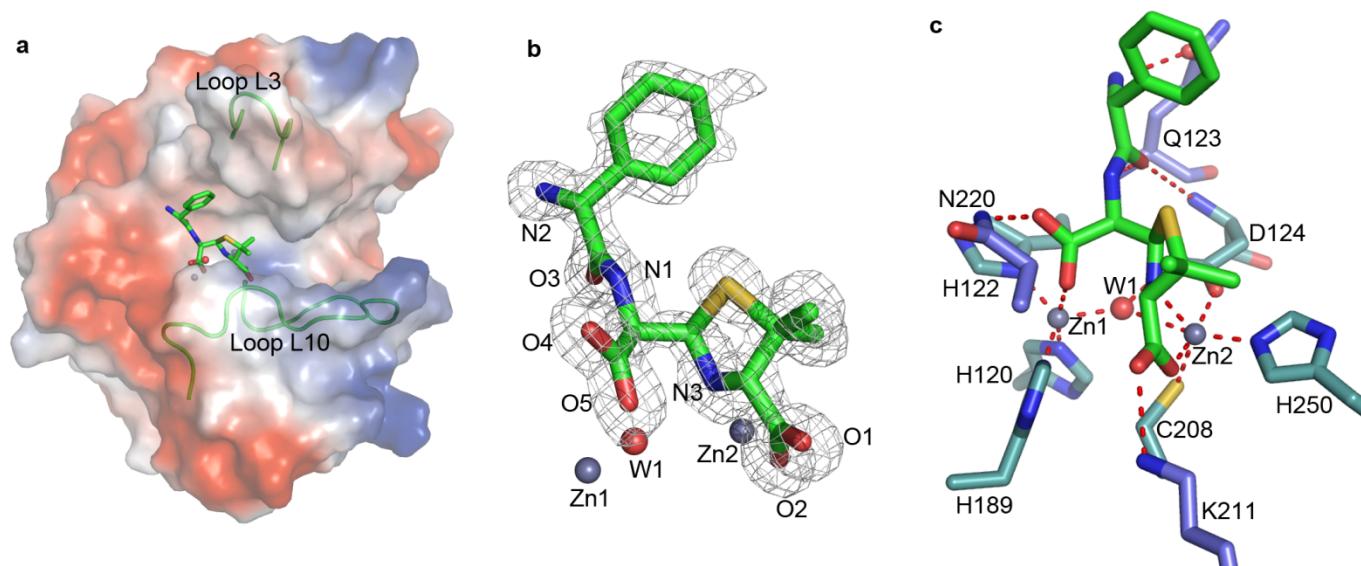




Protein Shape

A protein's three-dimensional shape

- typically recognizes and binds to another molecule and
- enables the protein to carry out its specific function in a cell.



Applications of protein engineering

- *Medicine*

Insulin, growth factors, antibodies...

- *Industry*

Enzymes

Cellulase: alcohol and glucose making

Glucose isomerase: manufacture of high-fructose syrups

Lactase: hydrolysis of lactose

Lipase: cheese making, preparation of flavorings

Applications of protein engineering

1. Detergents (> 40 years of experience)

 lower laundry wash temperatures incre^{ase}  need for adding enzymes.

remove 95% of a stain, how about the remaining 5% ?

“total cleaning” concept

“white becomes whiter!!”

- *Detergent enzymes : (hydrolases)*

- *proteases : a. alkaline optimum, pH 8-11*

- b. bleach stable

- c. low wash temp. (10-20 °C)

- *lipases*

- *amylases*

- *cellulases (clean, overall fabric care, and color maintenance)*

Applications of protein engineering

2. Baking

- *Flour consists of gluten, starch, non-starch polysaccharides, lipids and trace amounts of minerals.*
- *Optimize a combination of lower dosages of enzymes to achieve optimum dough consistency, stability, and bread quality.*

Fungal alpha-amylase:

Maximizes the fermentation process to obtain An even crumb structure and a high loaf volume

Glucose oxidase:

Oxidizes free sulphydryl groups in gluten to smaller crumb cells and a silkier texture elastic

Lipase:

Dough conditioning by producing more uniform, make weak dough stronger and whiter crumb color.

Lipoxygenase:

Bleaching and strengthening dough

Xylanase:

Dough conditioning, Easier dough handling and improved crumb structure

Protease:

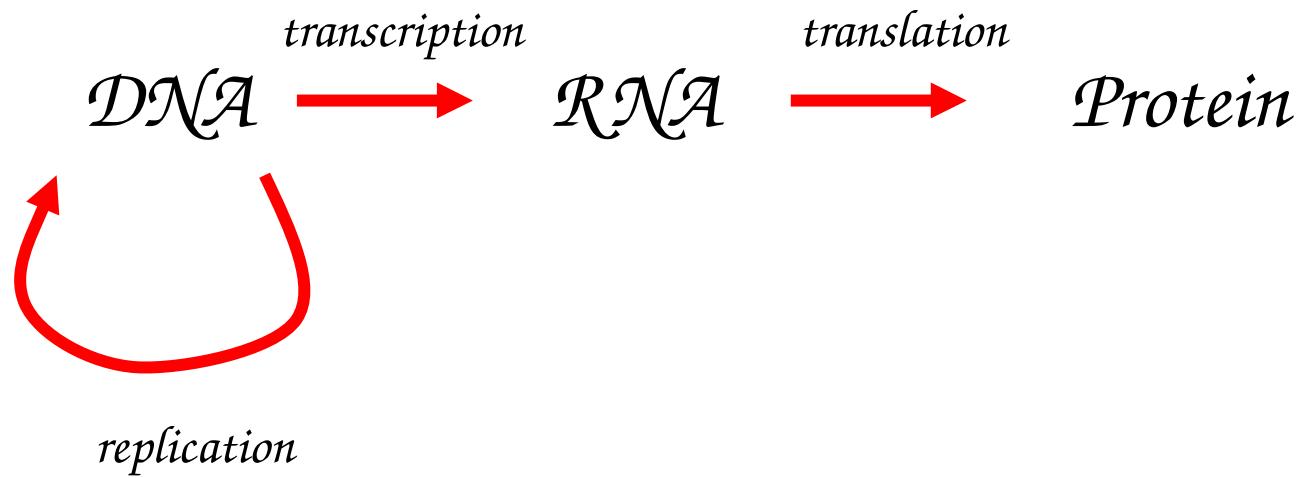
Weakens the gluten to provide the plastic properties required in dough for biscuits

Strategies of protein engineering research

- *How to improve or change the properties of proteins?*

Strategies of protein engineering research

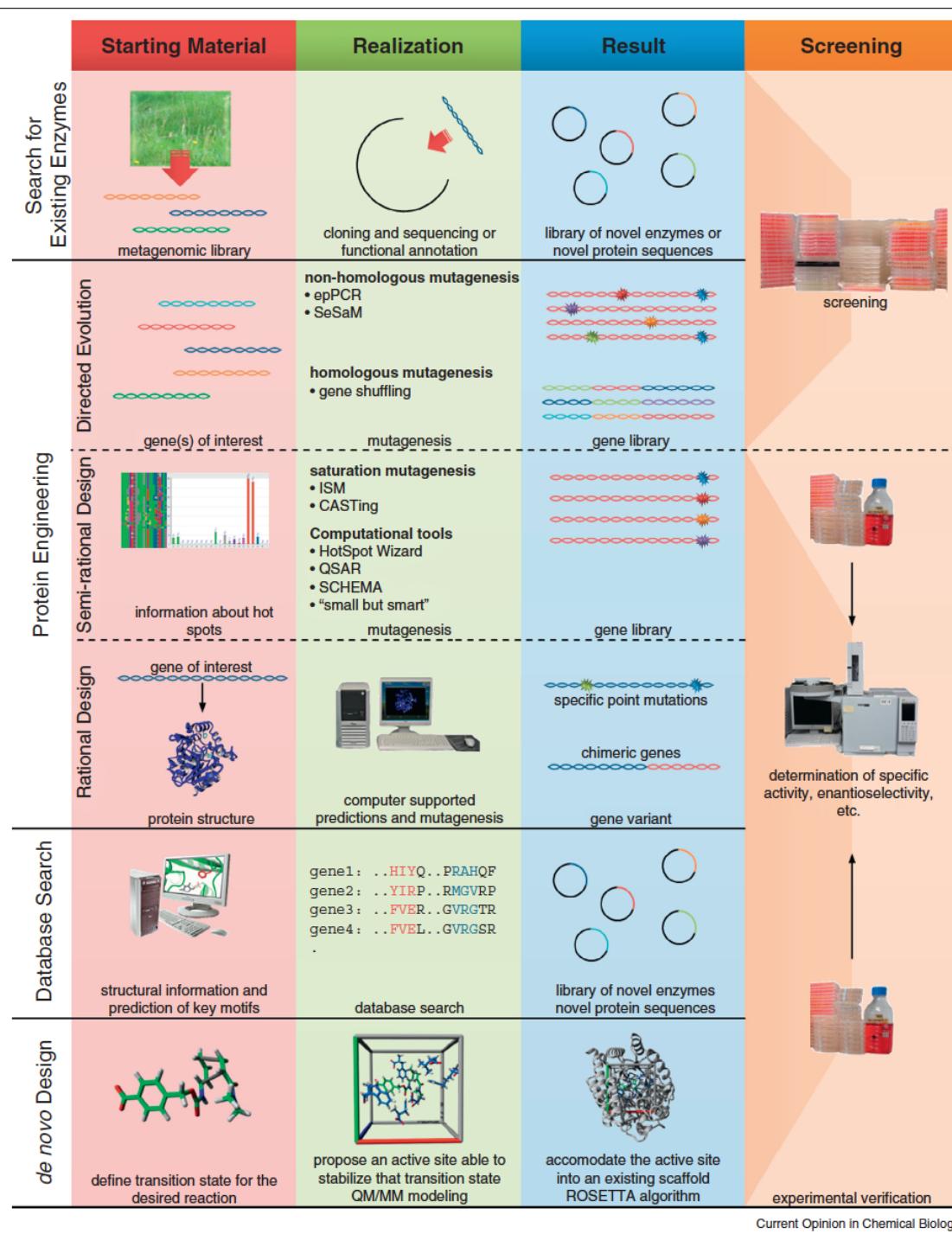
Central Dogma: Information flow in cells



Second base of RNA codon

	U	C	A	G		
U	UUU UUC UUA UUG	UCU UCC UCA UCG	UAU UAC UAA UAG	Tyrosine (Tyr) Stop Stop	Cysteine (Cys) Stop Tryptophan (Trp)	U C A G
C	CUU CUC CUA CUG	CCU CCC CCA CCG	CAU CAC CAA CAG	Histidine (His) Glutamine (Gln)	CGU CGC CGA CGG	U C A G
A	AUU AUC AUA AUG Met or start	ACU ACC ACA ACG	AAU AAC AAA AAG	Asparagine (Asn) Lysine (Lys)	AGU AGC AGA AGG	U C A G
G	GUU GUC GUA GUG	GCU GCC GCA GCG	GAU GAC GAA GAG	Aspartic acid (Asp) Glutamic acid (Glu)	GGU GGC GGA GGG	U C A G

Overview of concepts used to identify or create enzymes with desired properties



Directed Evolution – Random mutagenesis

- > based on the process of natural evolution
- *NO structural information required*
- *NO understanding of the mechanism required*

General Procedure:

Generation of genetic diversity
⇒ *Random mutagenesis*
(Error-prone PCR and DNA shuffling)

Identification of successful variants
⇒ *Screening and selection*

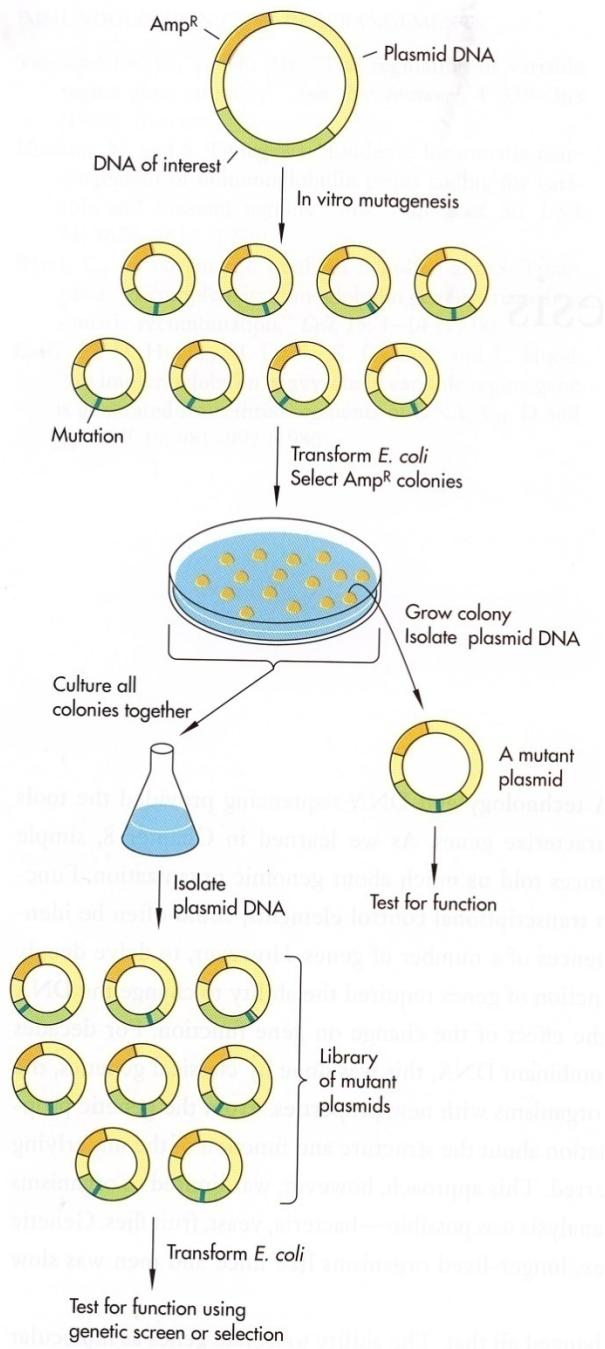
Rational Protein Design

⇒ *Site-directed mutagenesis !!!*

Requirements:

- > *Knowledge of sequence and preferable Structure
(active site,....)*
- > *Understanding of mechanism
(knowledge about structure – function relationship)*
- > *Identification of cofactors.....*

General strategy for directed mutagenesis

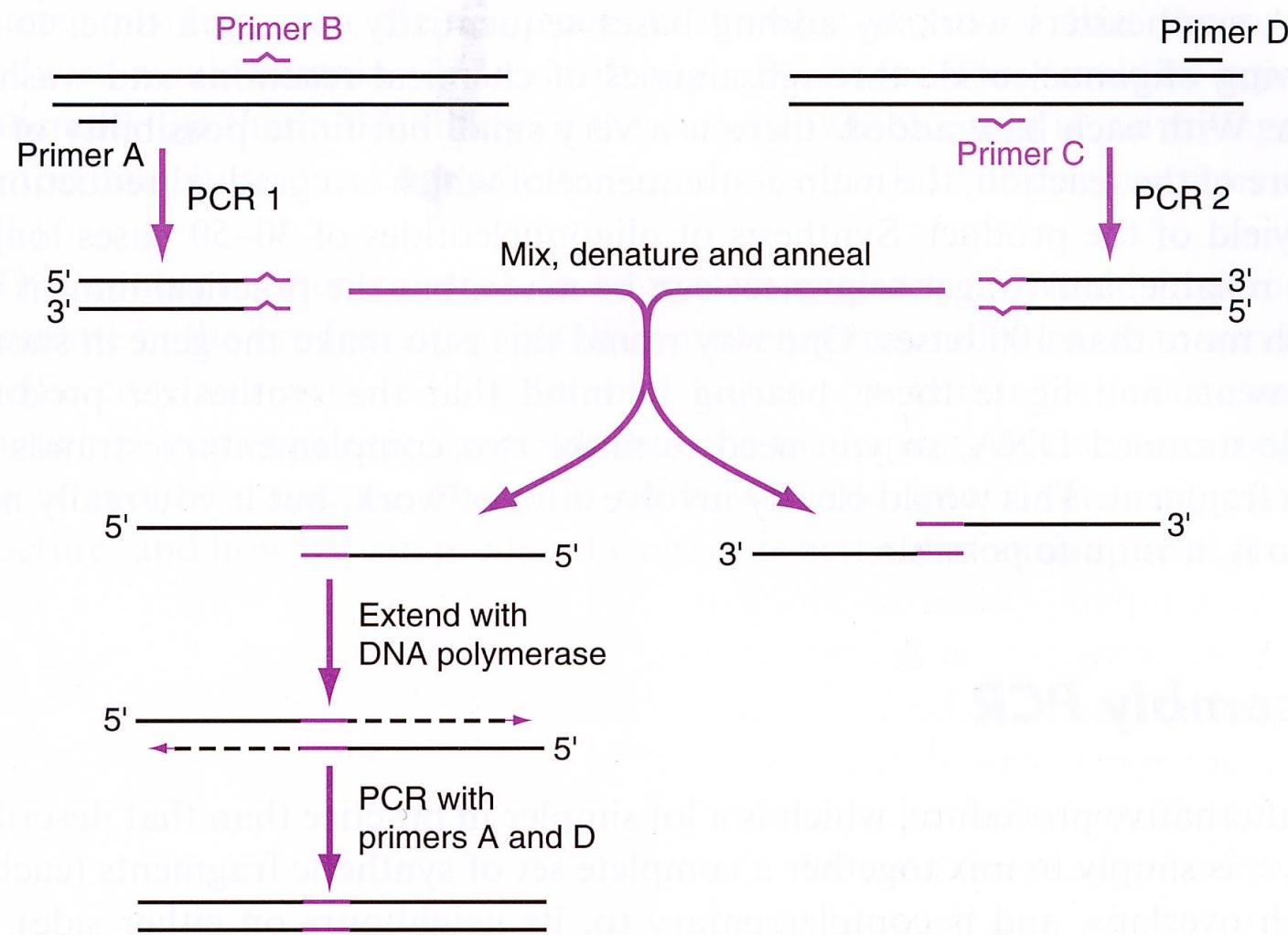


Requirements:

- *DNA of interest (gene or promoter) must be cloned*
- *Expression system must be available -> for testing phenotypic change*

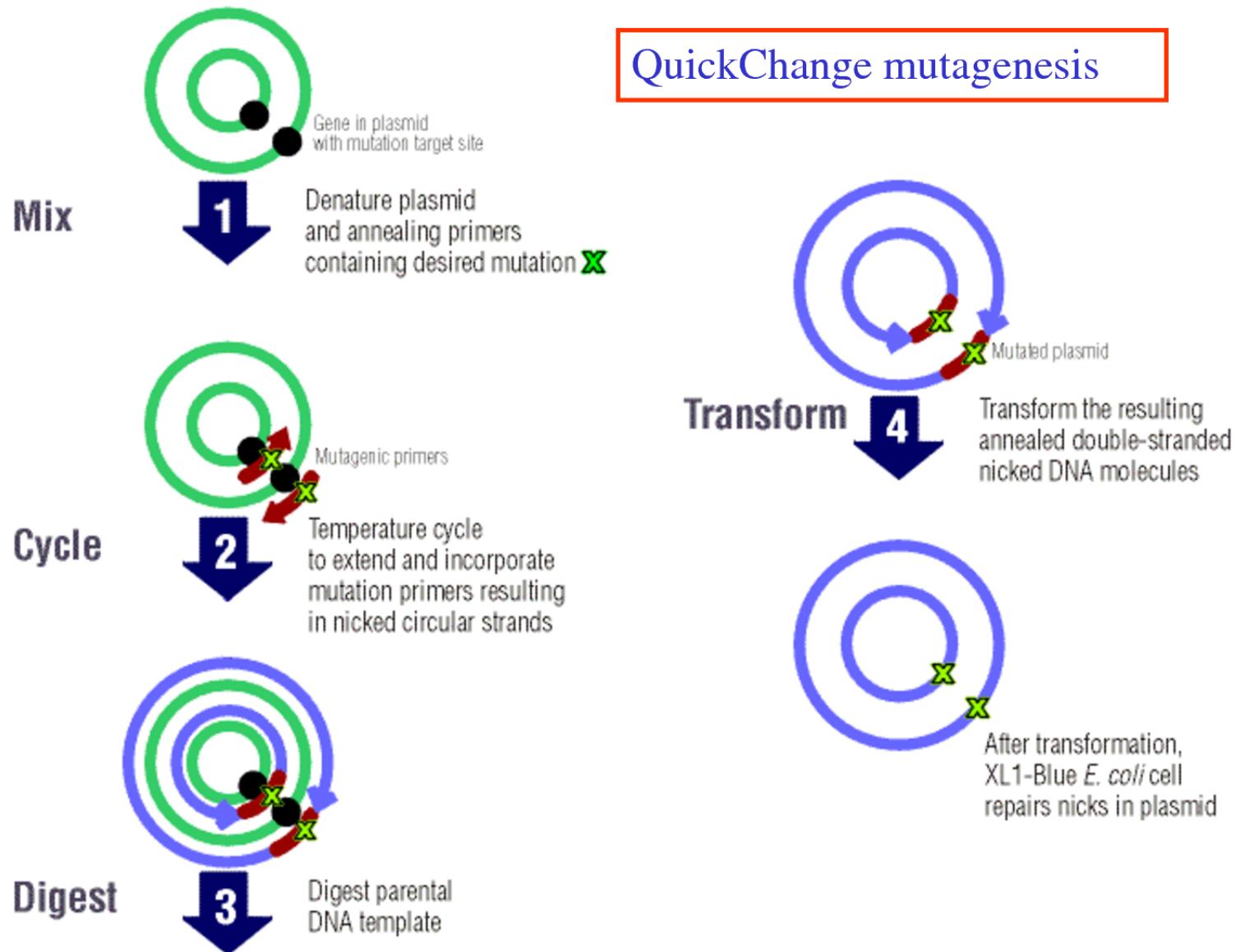
Site-directed mutagenesis methods – PCR based

1. Overlapping PCR



Site-directed mutagenesis methods – PCR based

2. Quick-change and related methods



Site-directed mutagenesis methods – PCR based

2. Quick-change and related methods

Genetailor

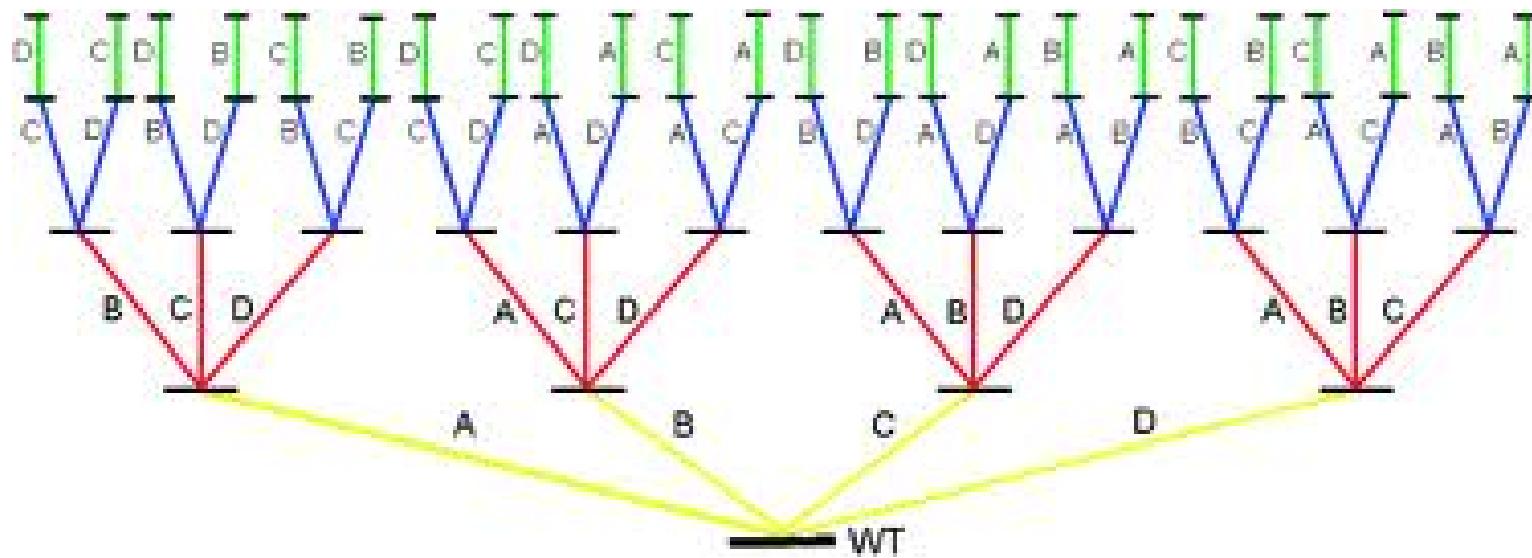
Mutagenic forward primer:

Mutagenic Reverse primer:



Site-directed mutagenesis methods – PCR based

3. Iterative saturated mutation



Site-directed mutagenesis methods – PCR based

3. Iterative saturated mutation

SELECT DEGENERATE CODON			
CODON	N	N	K
DNA DEGENERACIES - IUB code	W		
Base	Name	Bases	K
T	Thymine	T	M
C	Cytosine	C	B
A	Adenine	A	H
G	Guanine	G	V
Y	pYrimidine	C T	R
R	puRine	A G	Y
S	Strong (3H)	G C	S
W	Weak (2H)	A T	W
K	Keto	T G	M
M	aMino	A C	K
B	not A	C G T	V
D	not C	A G T	H
H	not G	A C T	D
V	not T	A C G	B
N	Unknown	A C G T	N

Screening effort [95% Coverage]*		
Number of colonies		
AA Sites	Codons	Colonies
1	32	94
2	1024	3066
3	32768	98163
4	1048576	3141251
5	33554432	100520093
10	1.13E+15	3.4E+15

Calculated Amino acid and Codon Distribution										
Genetic Code - Codon Distribution										
SECOND POSITION OF THE CODON										
	T	C	A	G						
F	TTT	Phe [F]	TCT	Ser [S]	TAT	Tyr [Y]	TGT	Cys [C]	T	T
I	TTC	Phe [F]	TCC	Ser [S]	TAC	Tyr [Y]	TGC	Cys [C]	C	H
R	TTA	Leu [L]	TCA	Ser [S]	TAA	[Ter end]	TGA	[Ter end]	A	I
S	TTG	Leu [L]	TCG	Ser [S]	TAG	[Ter end]	TGG	Trp [W]	G	R
T	CTT	Leu [L]	CCT	Pro [P]	CAT	His [H]	CGT	Arg [R]	T	D
C	CTC	Leu [L]	CCC	Pro [P]	CAC	His [H]	CGC	Arg [R]	C	P
P	CTA	Leu [L]	CCA	Pro [P]	CAA	Gln [Q]	CGA	Arg [R]	A	O
O	CTG	Leu [L]	CCG	Pro [P]	CAG	Gln [Q]	CGG	Arg [R]	G	O
S	ATT	Ile [I]	ACT	Thr [T]	AAT	Asn [N]	AGT	Ser [S]	T	S
I	ATC	Ile [I]	ACC	Thr [T]	AAC	Asn [N]	AGC	Ser [S]	C	I
A	ATA	Ile [I]	ACA	Thr [T]	AAA	Lys [K]	AGA	Arg [R]	A	T
T	ATG	Met [M]	ACG	Thr [T]	AAG	Lys [K]	AGG	Arg [R]	G	O
O	GTT	Val [V]	GCT	Ala [A]	GAT	Asp [D]	GGT	Gly [G]	T	O
N	GTC	Val [V]	GCC	Ala [A]	GAC	Asp [D]	GGC	Gly [G]	C	N
G	GTA	Val [V]	GCA	Ala [A]	GAA	Glu [E]	GGA	Gly [G]	A	
	GTG	Val [V]	GCG	Ala [A]	GAG	Glu [E]	GGG	Gly [G]	G	

Amino acid Distribution - Chemical Properties		
Side Chain	Amino acids	%
Acidic [-]	Asp, Glu	6,3
Basic [+]	Arg, His, Lys	15,6
Non-polar aliphatic	Ala, Ile, Leu, Met, Val	28,1
Aromatic	Phe, Trp, Tyr	9,4
Polar	Asn, Cys, Gln, Ser, Thr	25,0
Special Features	Gly, Pro	12,5

De novo design

“Knowledge-based protein design”

A highly challenging approach, offering the broadest possibility for new structure

- *To find the best amino acid sequence to ensure folding in a selected structure (α -helical bundle or α/β -barrel) **VS structure prediction***
- *To design and construct a synthetic protein using the information on the 3-D structures of natural protein accumulates and folding rules of proteins*

Methionine: prefers rigid segments, the central part of helical segments, and buried regions of natural proteins

Threonine: prefers flexible segments

Lysine: a good helix former and prefers at the C-terminal, etc.

Leucine: stabilizes helices, prefers buried regions, and is usually found in the middle positions in a helix

De novo design

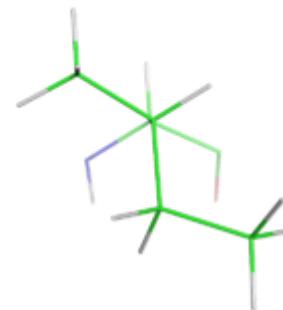
Questions to be addressed:

- *Target structure*
 - ✓ Function-structure relationship
 - ✓ Compared to direction evolution
 - ✓ Often based on known structure, novel folds are possible
- *Sequence space*
- *Structural flexibility*
- *Energy function*

De novo design

Questions to be addressed:

- *Target structure*
- *Sequence space*
- *Structural flexibility*
 - Side-chain*
 - Main chain (backbone)*
- *Energy function*

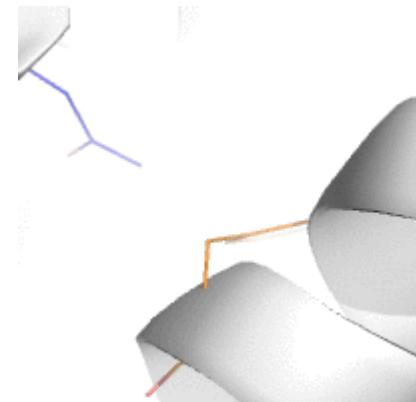


De novo design

Questions to be addressed:

- *Energy function*

- ✓ Rank and score sequences how well they fold to the target structure
- ✓ Accurate but simple
- ✓ Knowledge-based energy functions
- ✓ Physics-based energy functions (*AMBER, CHARMM, decomposable*)
- ✓ Statistical potentials (fast)
- ✓ Challenges
- ✓ single sequence vs multi-sequences, Combined, Rosetta
- ✓ Water



De novo design

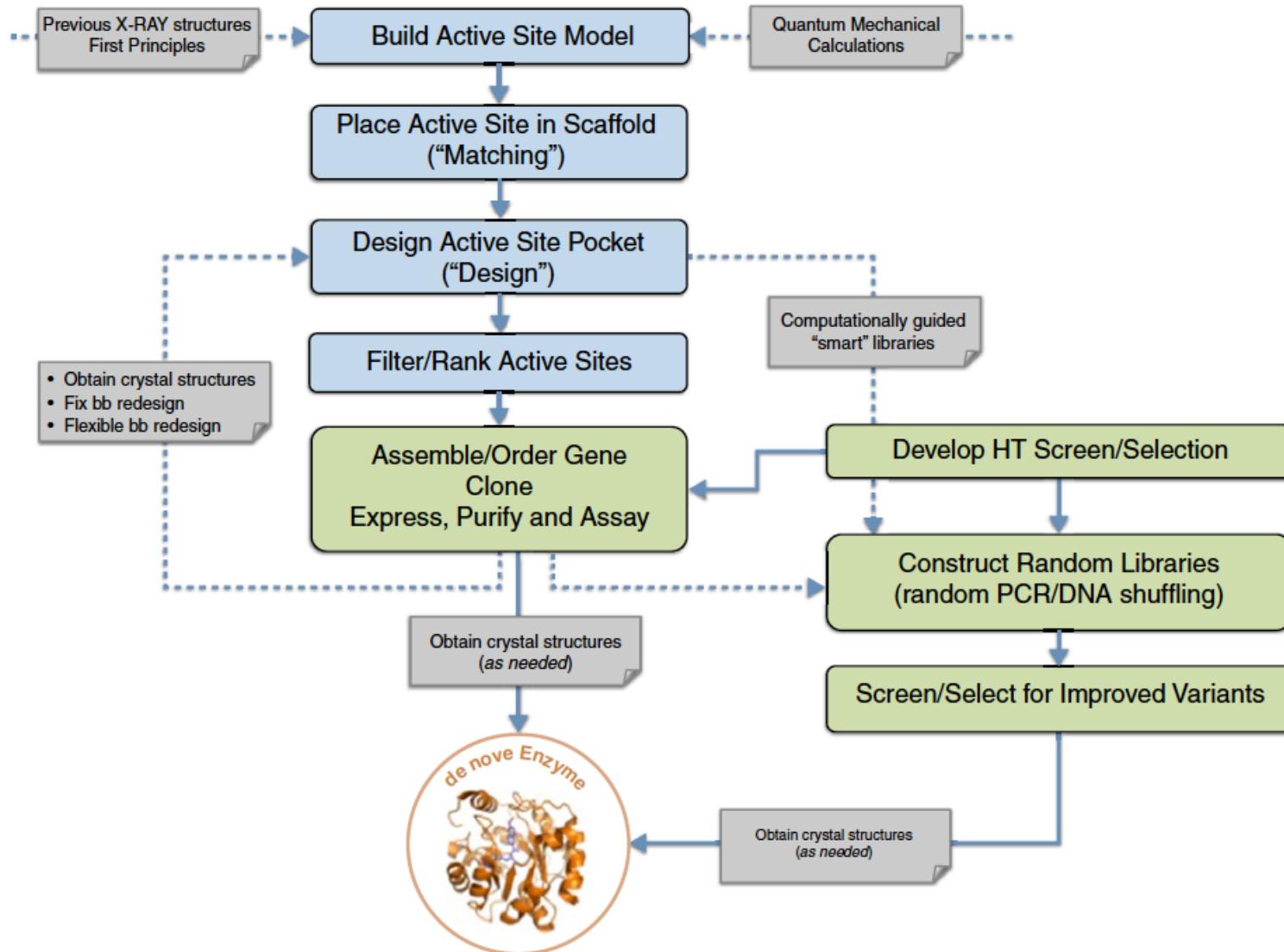
- *Protein design as an optimization problem*

$$\min E_T = \sum_i \left[E_i(r_i) + \sum_{i \neq j} E_{ij}(r_i, r_j) \right]$$

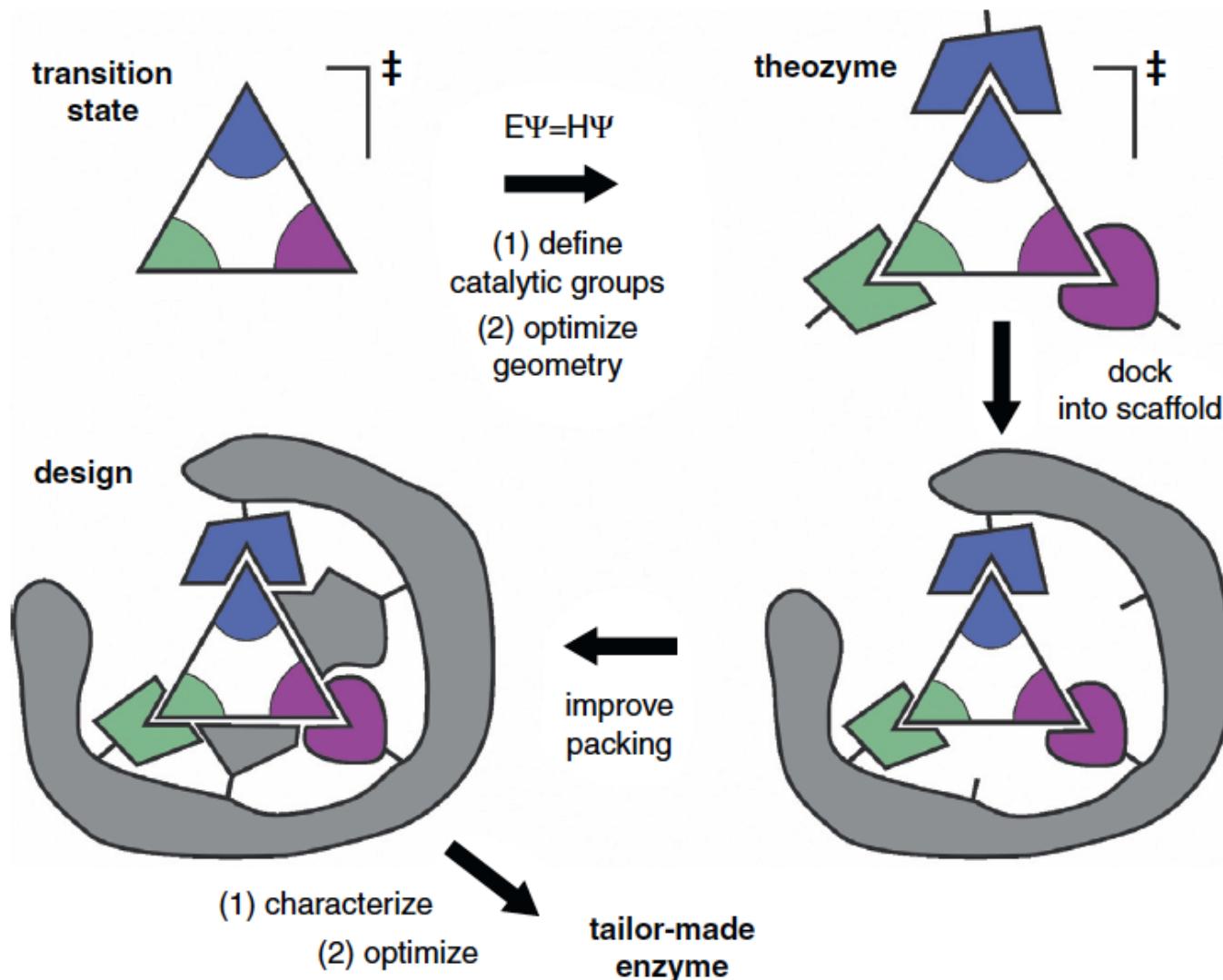
Algorithms

- ✓ *Algorithms with mathematical guarantees*
 - 0 Dead-end elimination (DEE)
 - 0 Branch-and-bound algorithms
 - 0 Optimization as an integer linear program
 - 0 Message-passing based approximating to the linear programming dual
- ✓ *Optimization algorithms without guarantees*
 - 0 Monte Carlo and simulated annealing
 - 0 Faster
 - 0 Belief propagation for protein design

Flowchart of De novo design

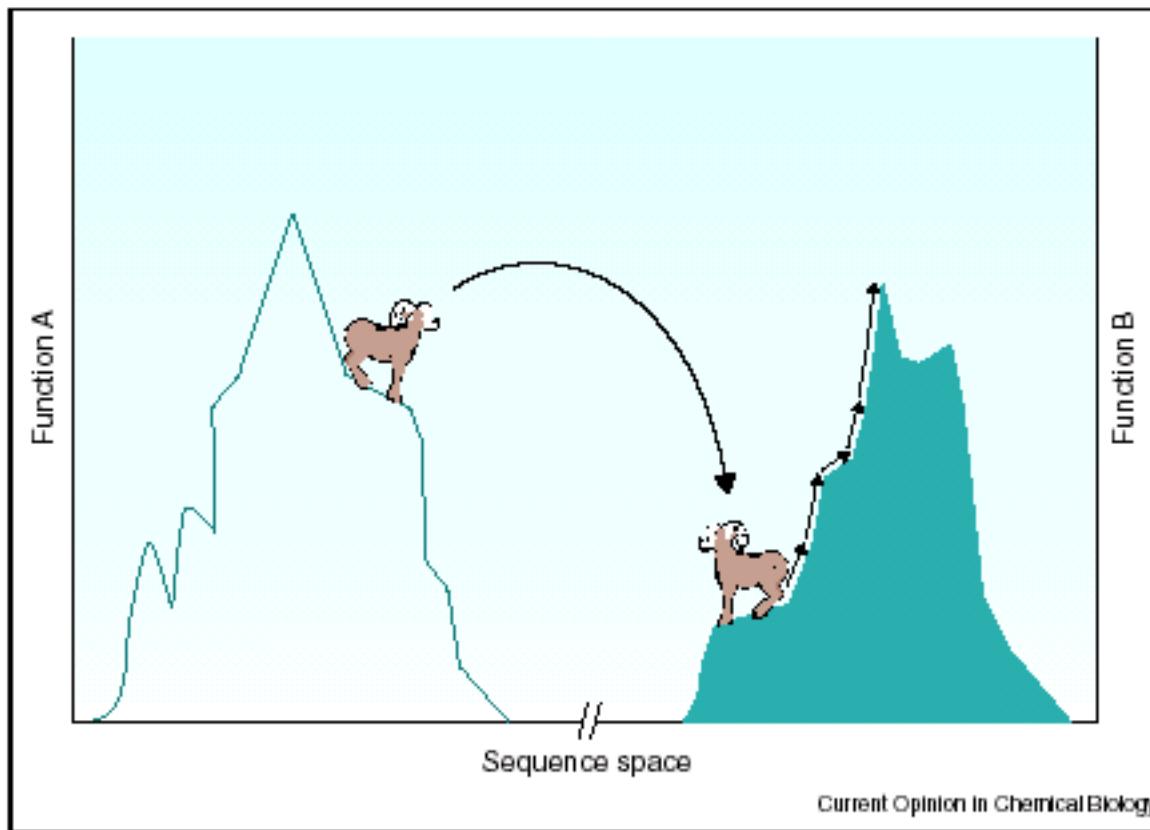


De novo design of enzymes



De novo design

- To identify sequences and geometries of amino acids that are optimal for stabilizing the backbone geometry.
- To develop powerful search algorithms to find optimal solutions



Applications of protein design

What can be engineered in Proteins ?

-> Folding (+Structure):

1. *Thermodynamic Stability*
(Equilibrium between: Native \leftrightarrow Unfolded state)
2. *Thermal and Environmental Stability (Temperature, pH, Solvent, Detergents, Salt)*

Applications of protein design

What can be engineered in Proteins ?

-> Function:

1. *Binding (Interaction of a protein with its surroundings)*

How many points are required to bind a molecule with high affinity?

2. *Catalysis (a different form of binding – binding the transition state of a chemical reaction)*

Increased binding to the transition state \Rightarrow increased catalytic rates and enzyme specificity!!!

Requires: Knowledge of the Catalytic Mechanism !!!

-> engineer K_{cat} and K_m

Stability improvement

Factors which contribute to stability:

- 1 *Hydrophobicity (hydrophobic core)*
- 2 *Electrostatic Interactions:*
 - > *Salt Bridges*
 - > *Hydrogen Bonds*
 - > *Dipole Interactions*
- 3 *Disulfide Bridges*
- 4 *Metal Binding (Metal chelating site)*
- 5 *Reduction of the unfolded state entropy with
 $X \rightarrow P$ mutations*

Stability improvement

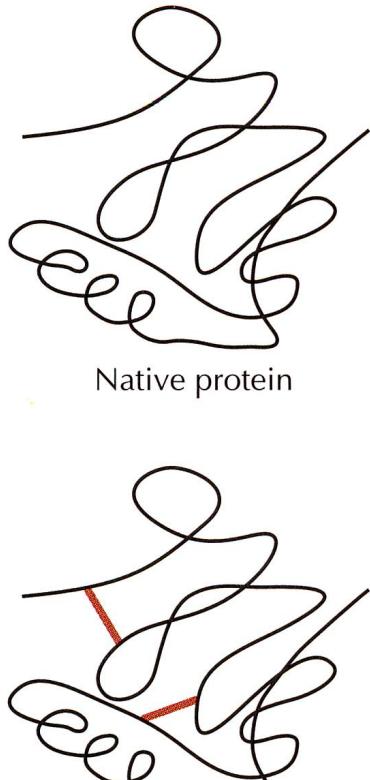
Design of Thermal and Environmental stability:

1. *Stabilization of α -Helix Macro dipoles*
2. *Engineer Structural Motifs (like Helix N-Caps)*
3. *Introduction of salt bridges*
4. *Introduction of residues with higher intrinsic properties for their conformational state (e.g. Ala replacement within a α -Helix)*
5. *Introduction of disulfide bridges*
6. *Reduction of the unfolded state entropy with $X \rightarrow Pro$ mutations*

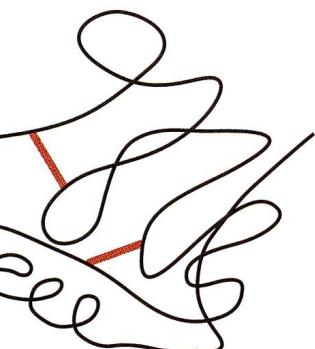
Stability improvement-1

Engineering Stability of Enzymes – T4 lysozyme

-> S-S bonds introduction



Native protein



Engineered protein

Table 8.2 Properties of T4 lysozyme and six engineered variants

Enzyme	Amino acid at position:							No. of -S-S-	% Activity	T_m (°C)
	3	9	21	54	97	142	164			
wt	Ile	Ile	Thr	Cys	Cys	Thr	Leu	0	100	41.9
pwt	Ile	Ile	Thr	Thr	Ala	Thr	Leu	0	100	41.9
A	Cys	Ile	Thr	Thr	Cys	Thr	Leu	1	96	46.7
B	Ile	Cys	Thr	Thr	Ala	Thr	Cys	1	106	48.3
C	Ile	Ile	Cys	Thr	Ala	Cys	Leu	1	0	52.9
D	Cys	Cys	Thr	Thr	Cys	Thr	Cys	2	95	57.6
E	Ile	Cys	Cys	Thr	Ala	Cys	Cys	2	0	58.9
F	Cys	Cys	Cys	Thr	Cys	Cys	Cys	3	0	65.5

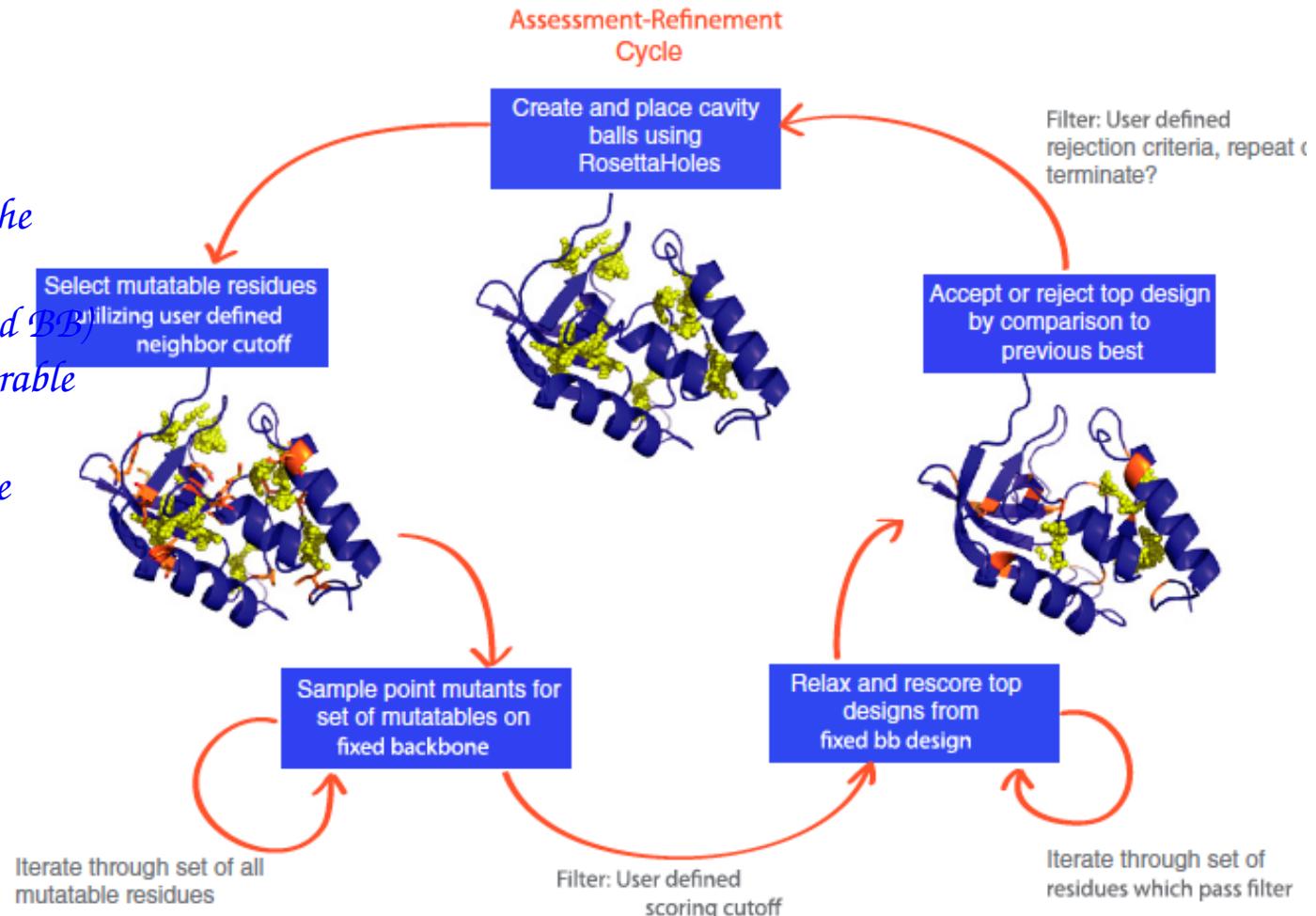
Adapted from Matsumura et al., *Nature* 342:291–293, 1989.

wt, wild-type T4 lysozyme; pwt, pseudo-wild-type enzyme; A through F, six engineered cysteine variants; -S-S-, disulfide bonds; T_m , “melting” temperature (a measure of thermostability).

Stability improvement-2

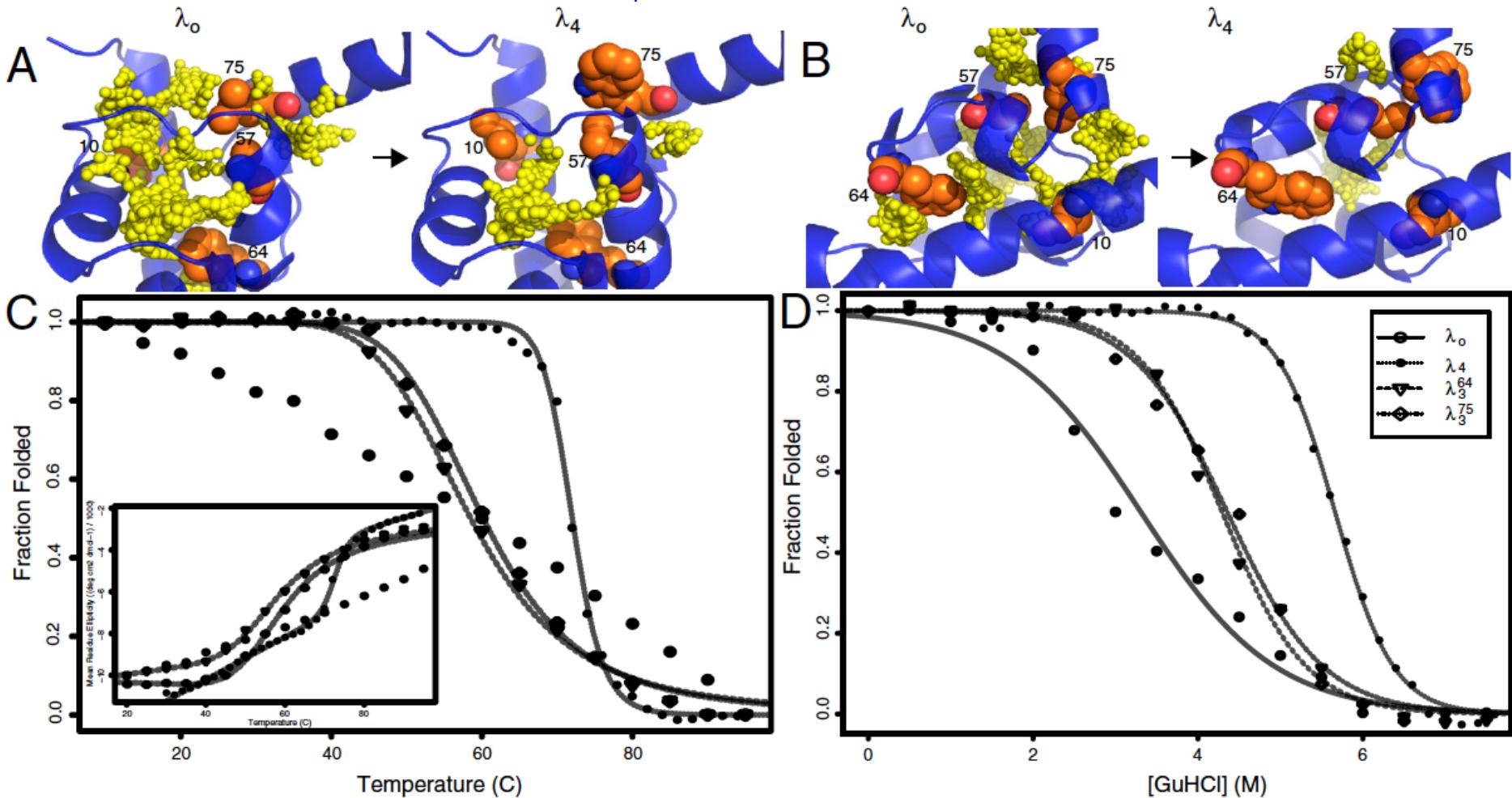
Overview of the Rosetta_{vip} protocol

1. Void identification
2. Mutations to fill the void
3. Optimization (fixed BB)
4. Relaxation of favorable mutations
5. Select and continue



Stability improvement-2

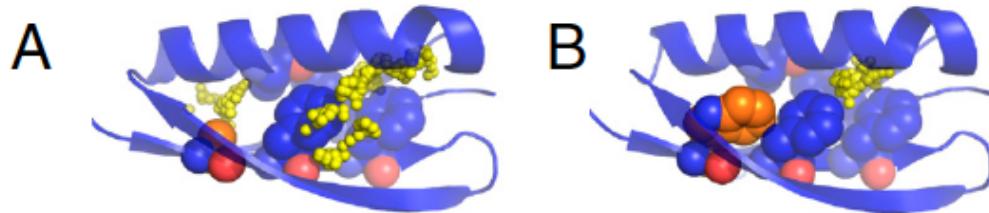
Designed λ repressor by Rosetta_{vip}



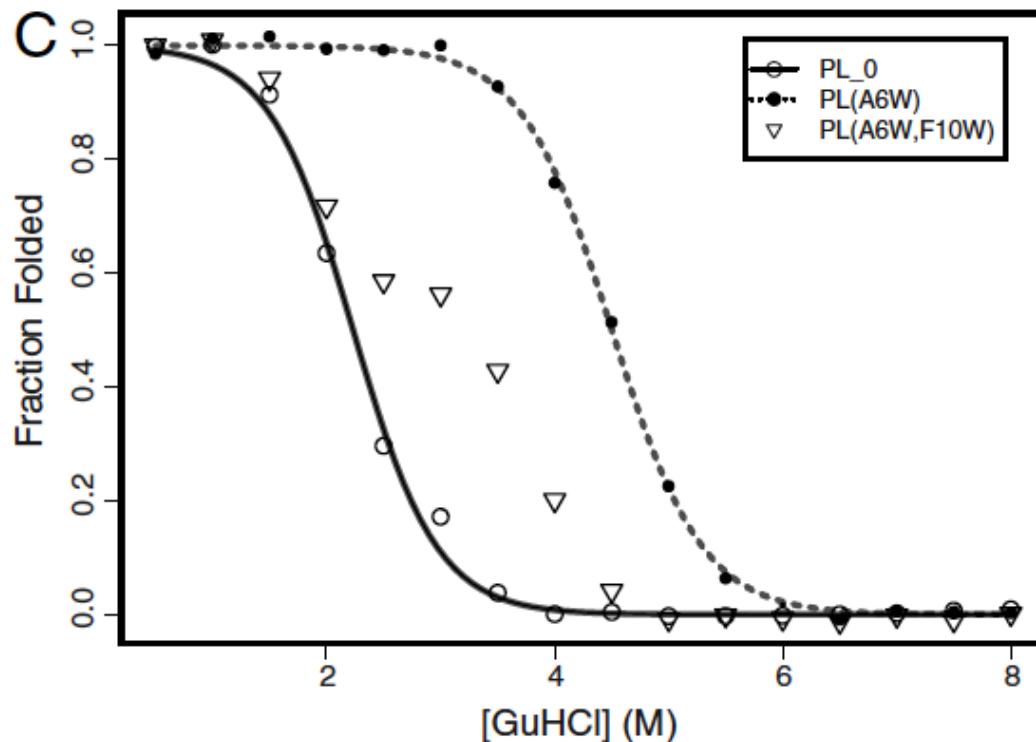
4 mutations, reduction of $>240\text{\AA}^3$, increased stability (thermal and chemical)

Stability improvement-2

Designed Protein L by Rosetta_{vip}

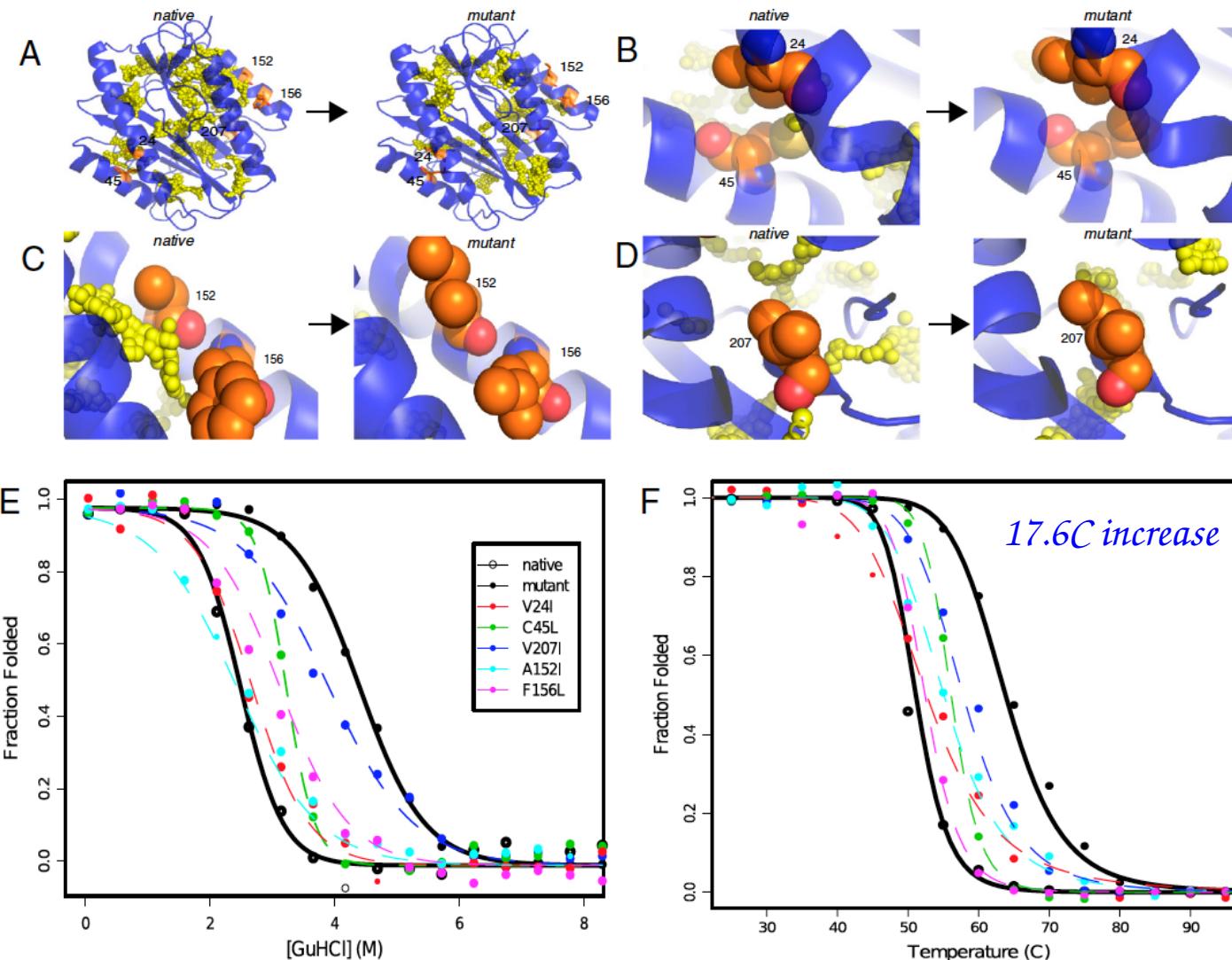


- Two voids identified
- Two mutations
- Transition between folded and unfolded
- False positive
- Single mutation



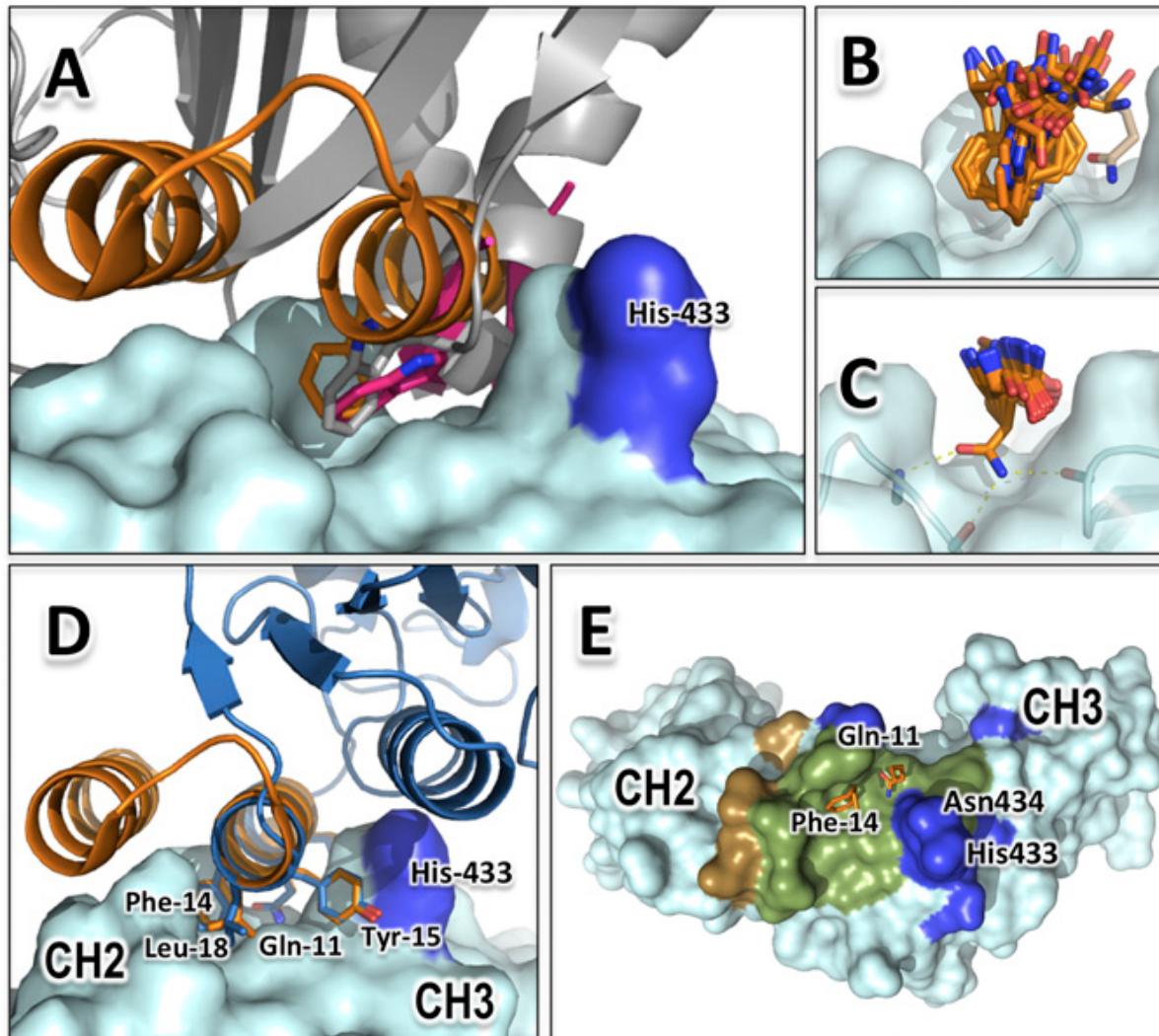
Stability improvement-2

Natural protein eMAP by Rosetta_{vip}



Designing and engineering environmental sensitivity

Computational design of a pH-sensitive IgG binding protein



Computational design of a pH-sensitive IgG binding protein.

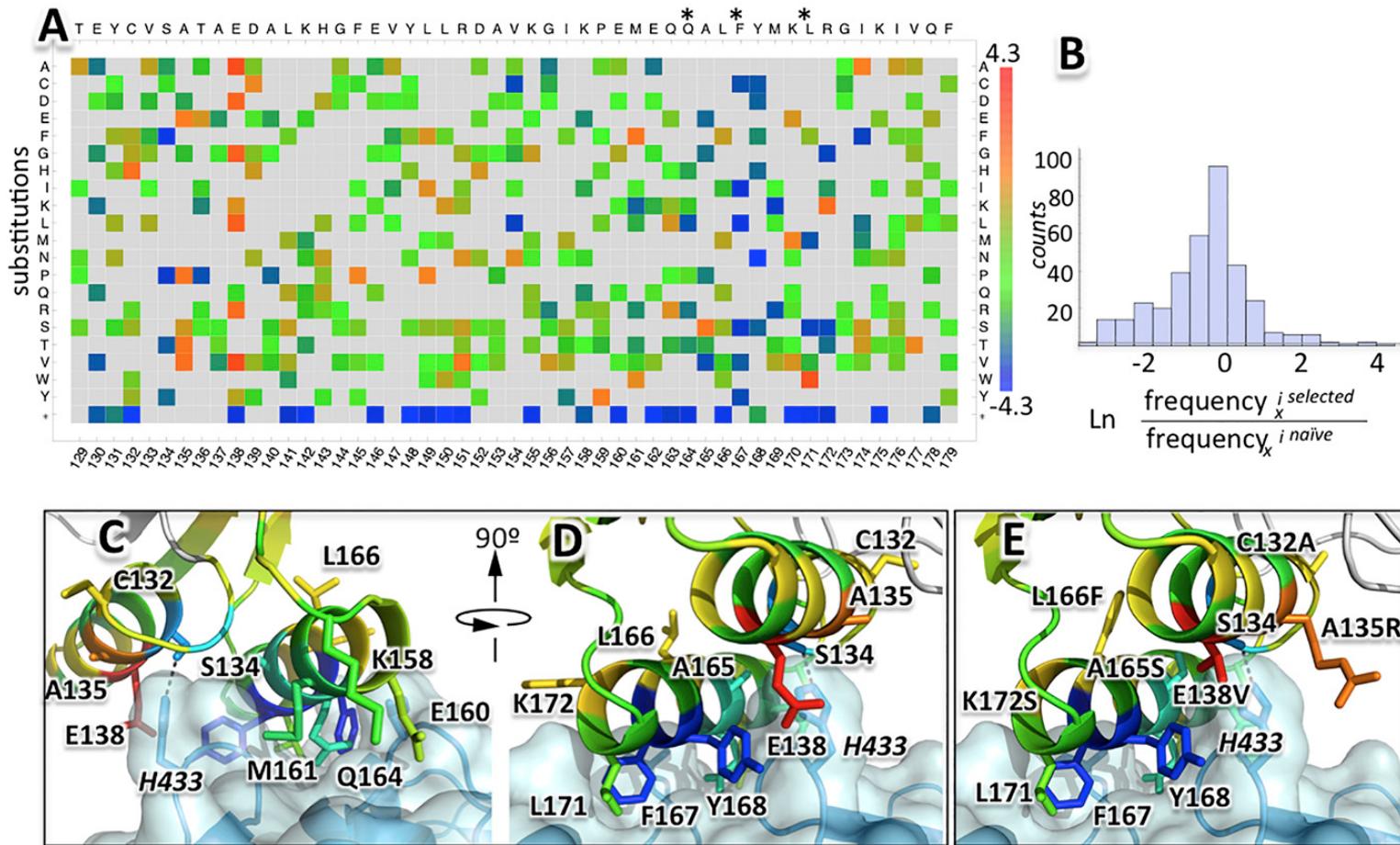
Strauch EM, Fleishman SJ, Baker D.

Proc Natl Acad Sci U S A. 2014 Jan 14;111(2):675-80. doi: 10.1073/pnas.1313605111. Epub 2013 Dec 31.

Designing and engineering environmental sensitivity

Computational design of a pH-sensitive IgG binding protein

Mutagenesis
FACS
NGS



Computational design of a pH-sensitive IgG binding protein.

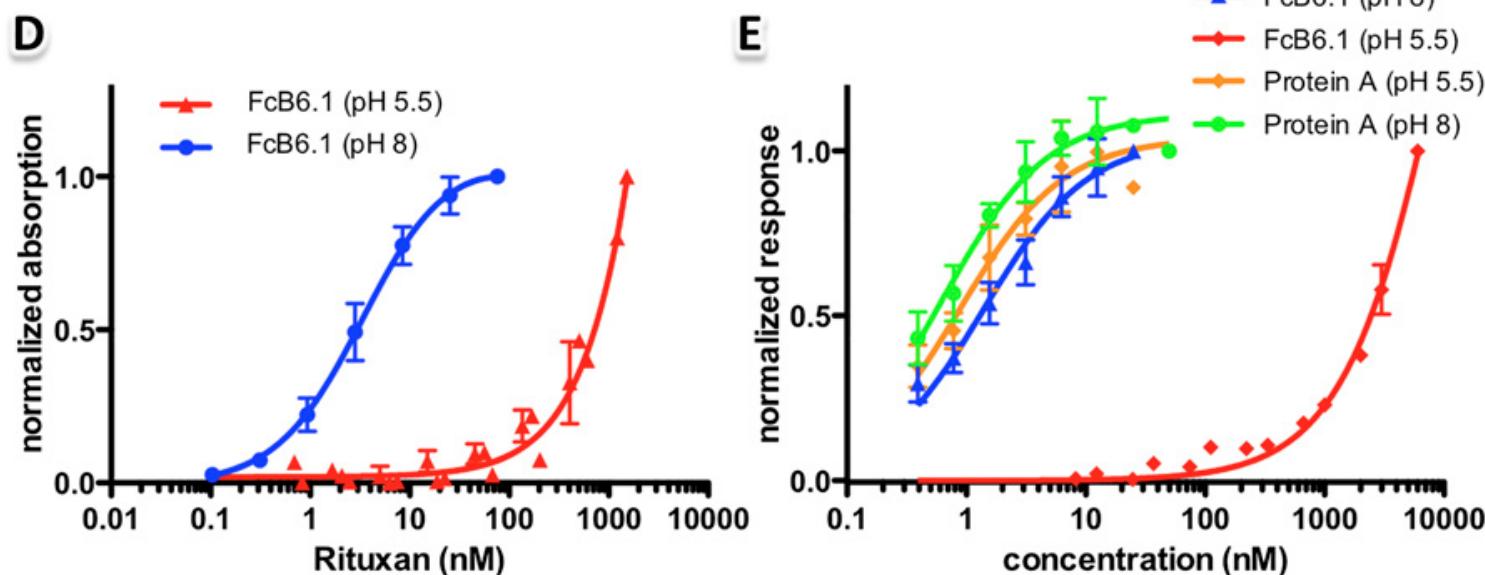
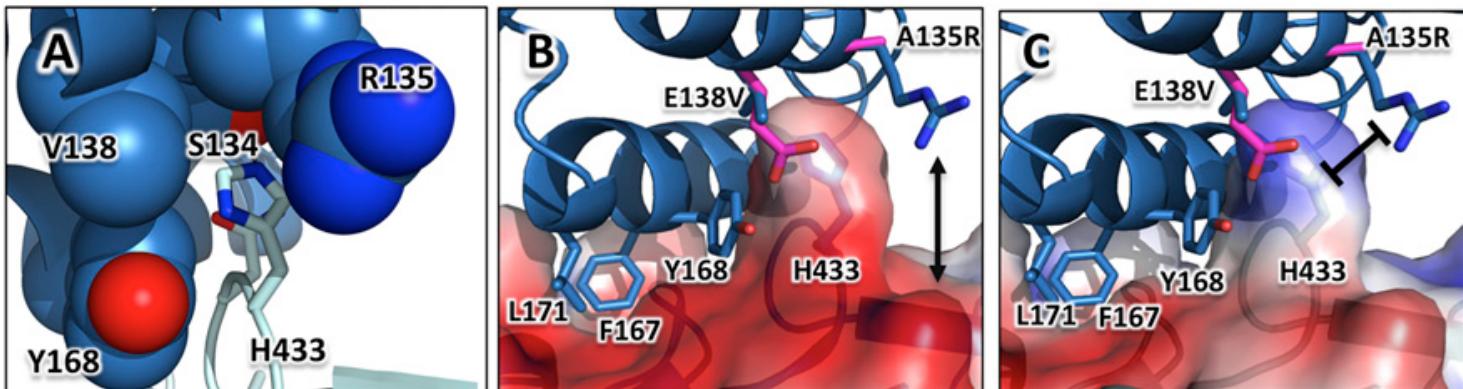
Strauch EM, Fleishman SJ, Baker D.

Proc Natl Acad Sci U S A. 2014 Jan 14;111(2):675-80. doi: 10.1073/pnas.1313605111. Epub 2013 Dec 31.

Designing and engineering environmental sensitivity

Computational design of a pH-sensitive IgG binding protein

1. pH sensitive ($\text{pH} 8.2$ vs $\text{pH} 5.5$, 500 fold)
2. Thermal stability (80°C)
3. Chemical stability (3M Urea, 1.5M Guanidine)
4. Binding specificity ($h\text{IgG}2$, $h\text{IgG}4$, $h\text{IgG}1$, $m\text{IgG}1$, $m\text{IgG}2a$)



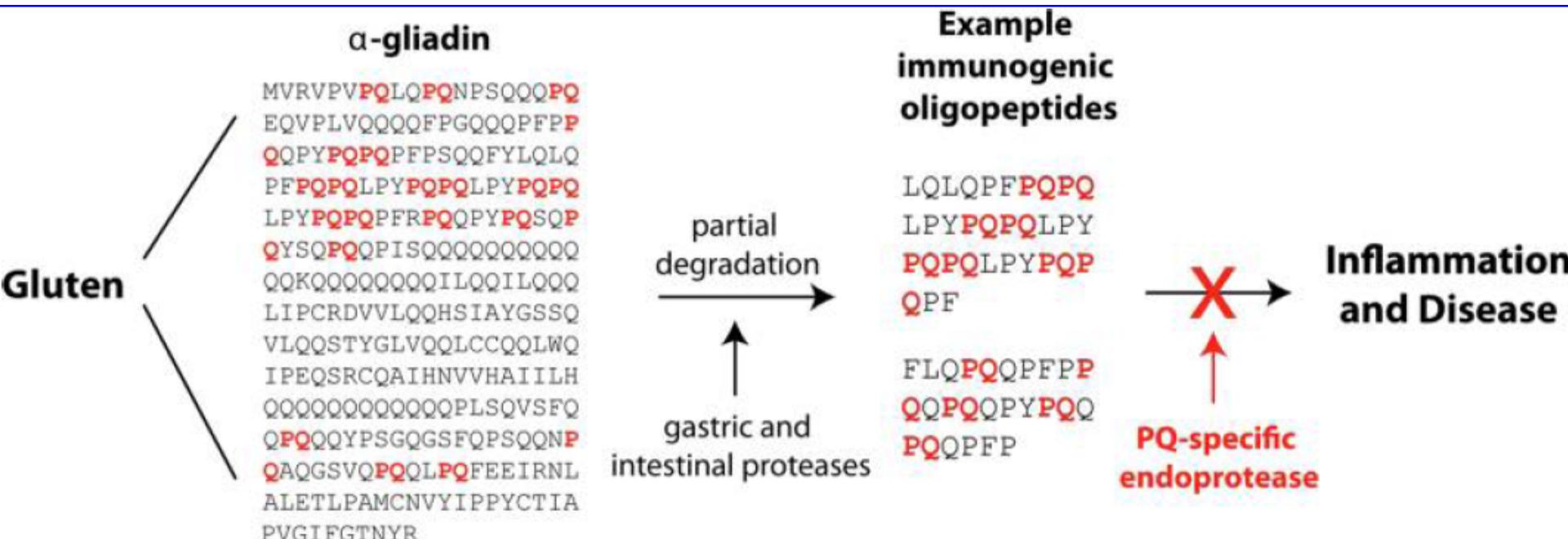
Computational design of a pH-sensitive IgG binding protein.

Strauch EM, Fleishman SJ, Baker D.

Proc Natl Acad Sci U S A. 2014 Jan 14;111(2):675-80. doi: 10.1073/pnas.1313605111. Epub 2013 Dec 31.

Designing and engineering enzyme activity-1

Computational Design of an α -Gliadin Peptidase



Ideal oral enzyme therapeutic:

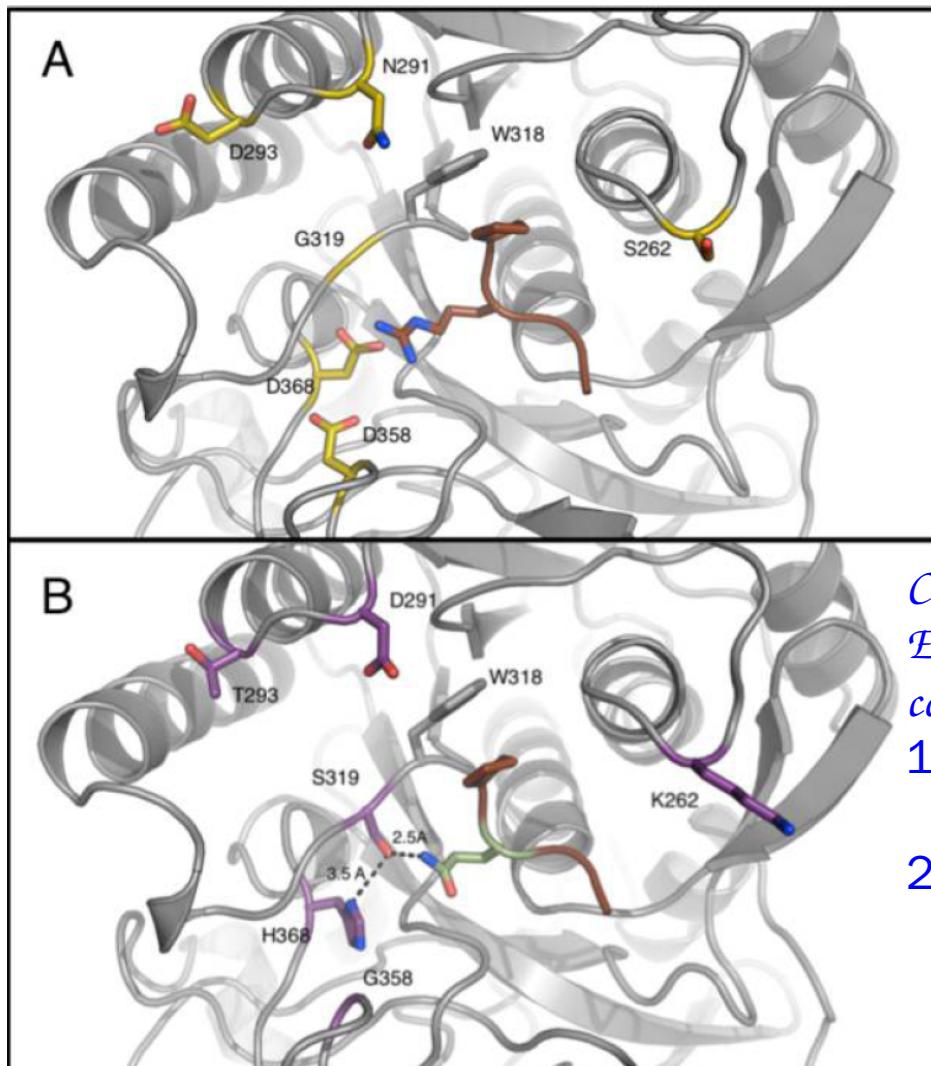
1. Optimal activity at low pH (2~4)
2. Resistance to common digestive proteases
3. Facile recombinant production in soluble form
4. Specificity for the common PQ motif

Designing and engineering enzyme activity-1

Computational Design of an α -Gliadin Peptidase

KumaWT

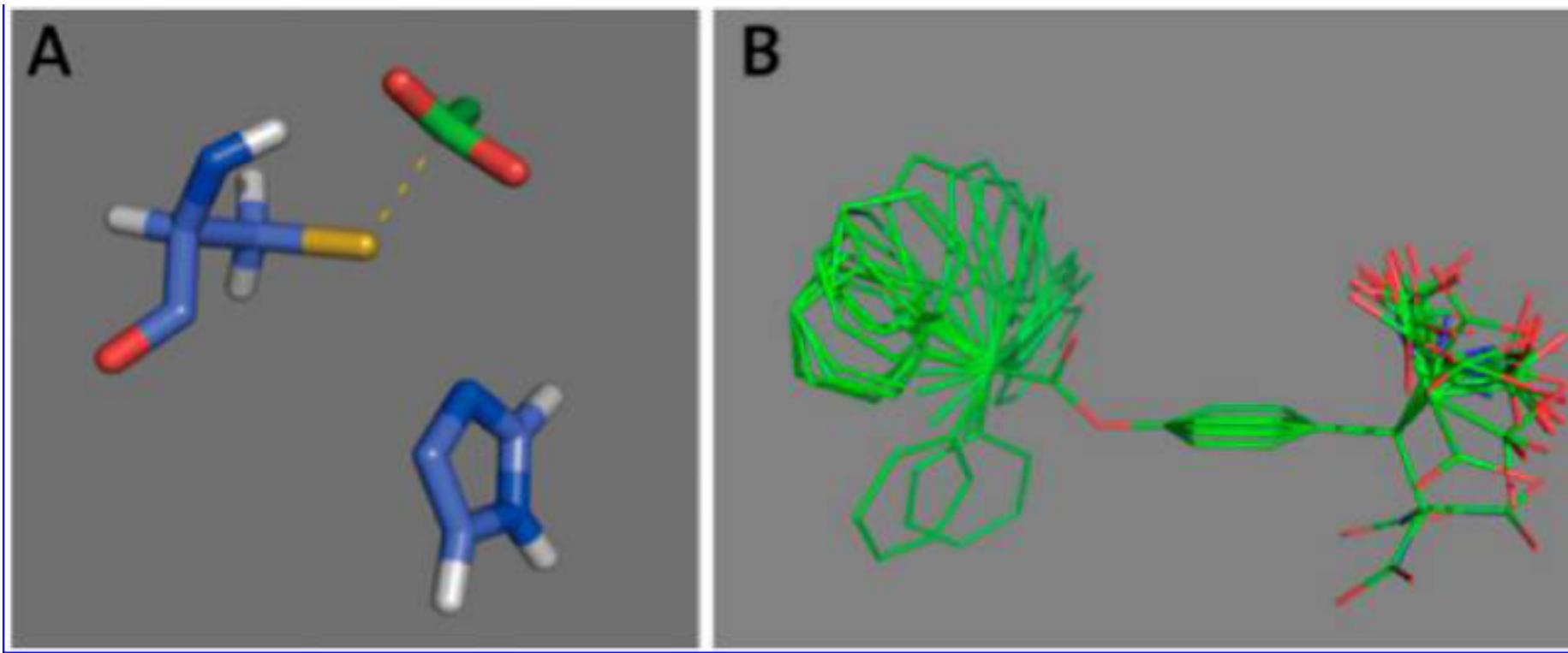
- 1 Optimal activity at low pH (2~4)
- 2 Resistance to common digestive proteases
- 3 Facile recombinant production in soluble form
- 4 Specificity for the common PR/K motif



- Computation design
Energy calculation and comparison
- 1 116-fold activity increase
 - 2 >800 fold switch in substrate specificity

Designing and engineering enzyme activity-2

Catalytic Dyads and Oxyanion Holes for Ester Hydrolysis

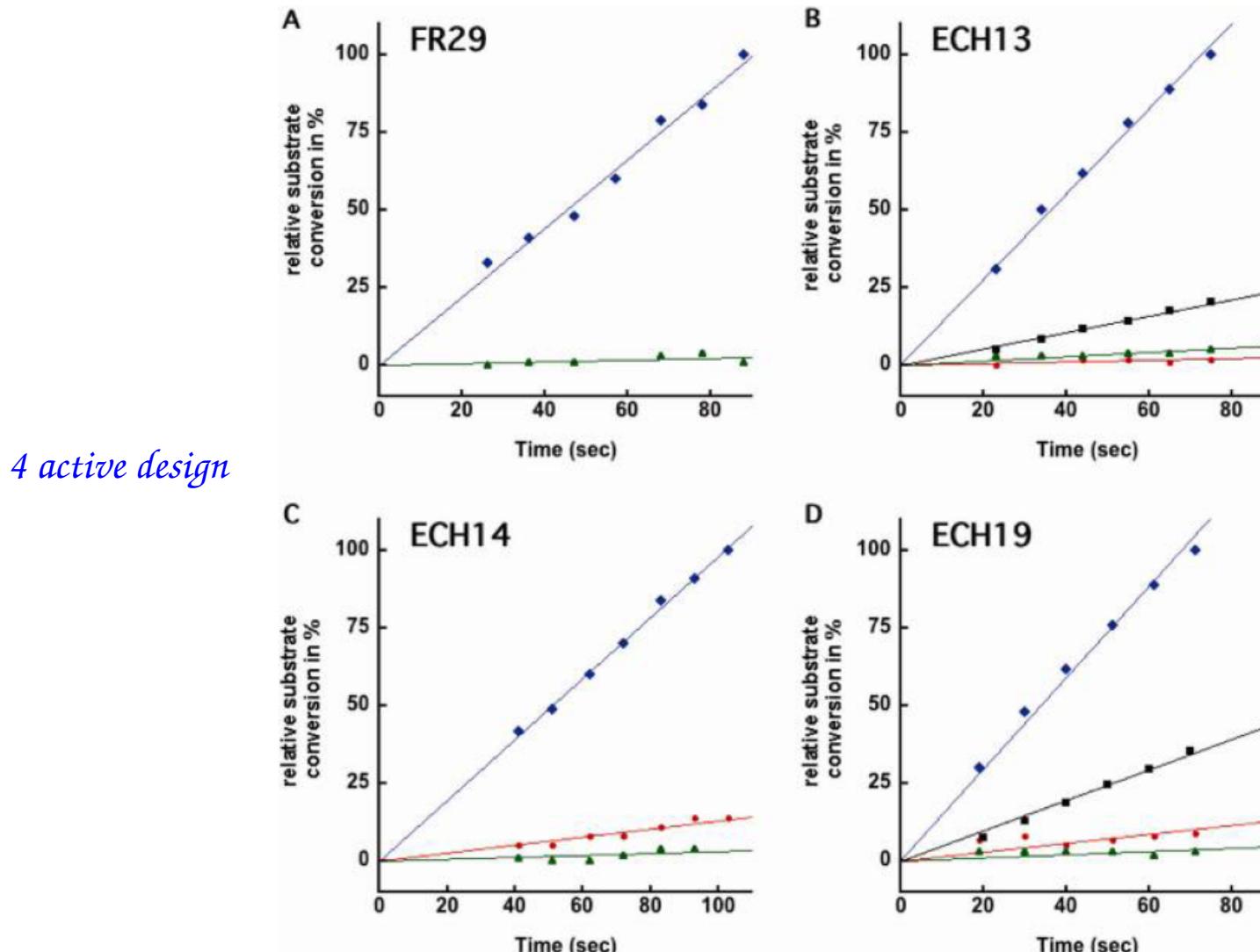


Cys-His dyad

Substrate

Designing and engineering enzyme activity-2

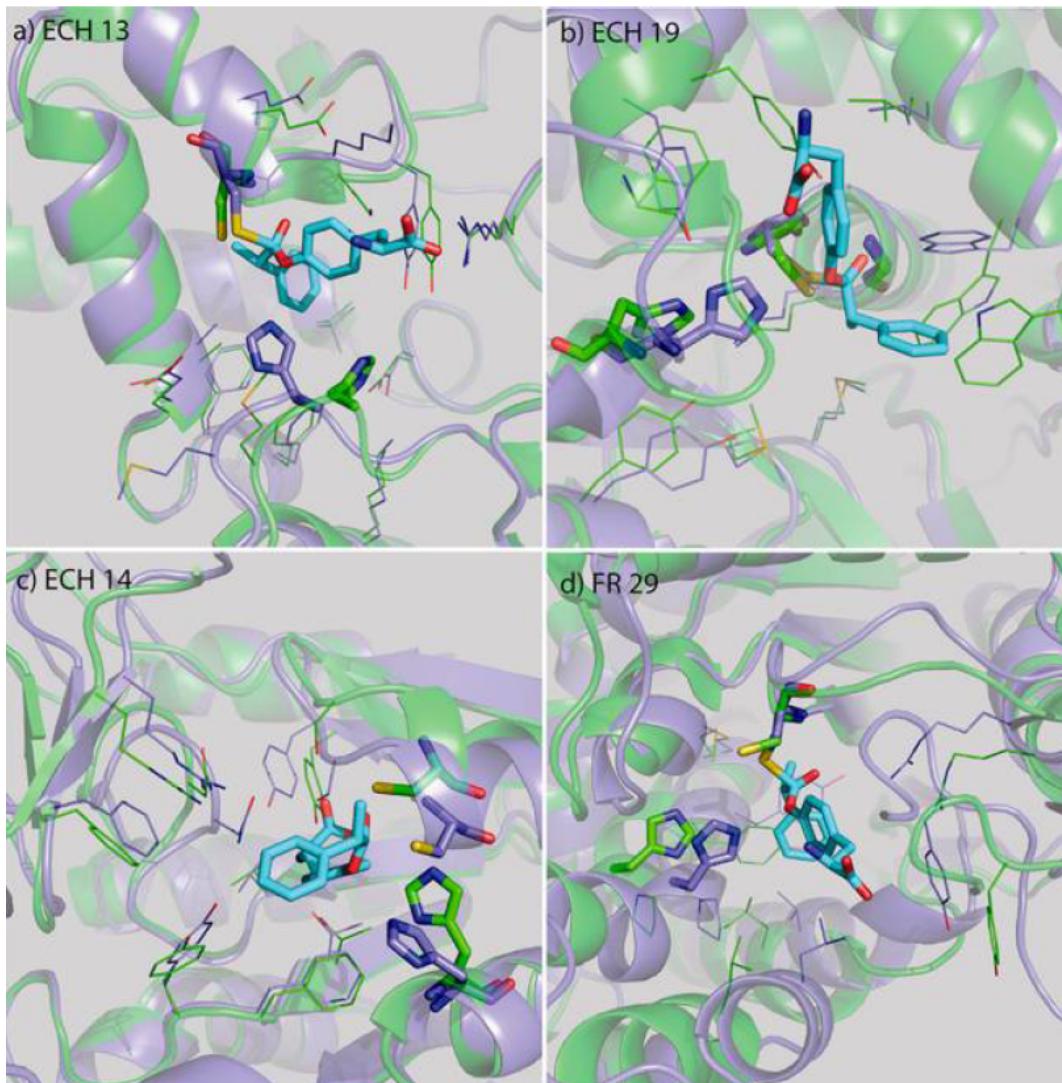
Catalytic Dyads and Oxyanion Holes for Ester Hydrolysis



Designing and engineering enzyme activity-2

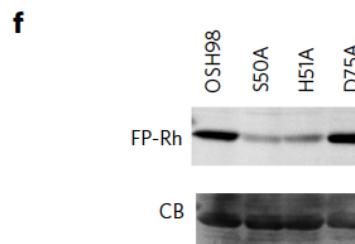
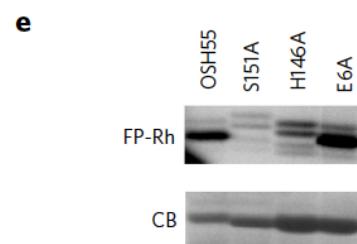
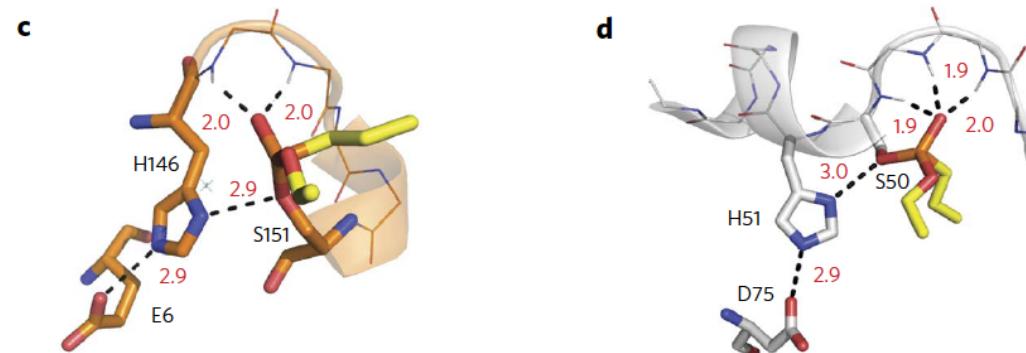
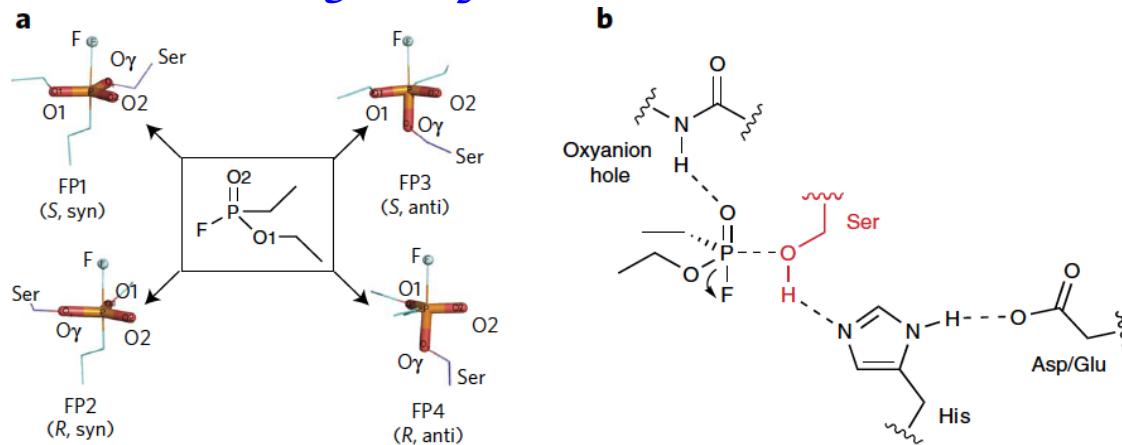
Catalytic Dyads and Oxyanion Holes for Ester Hydrolysis

*His conformation
changed
Backbone shift*



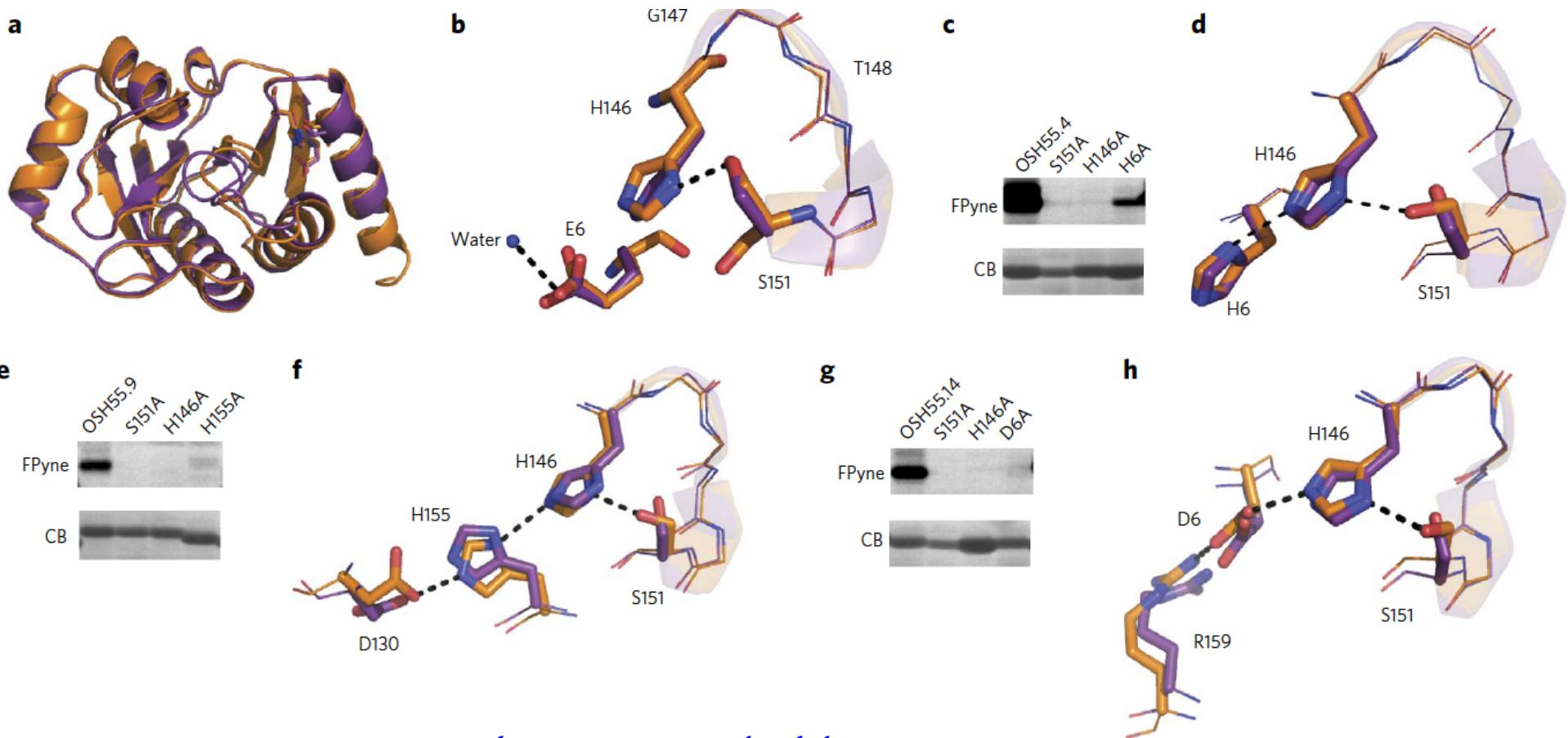
Designing and engineering enzyme activity-3

Design of activated serine-containing catalytic triads with atomic-level accuracy



Designing and engineering enzyme activity-3

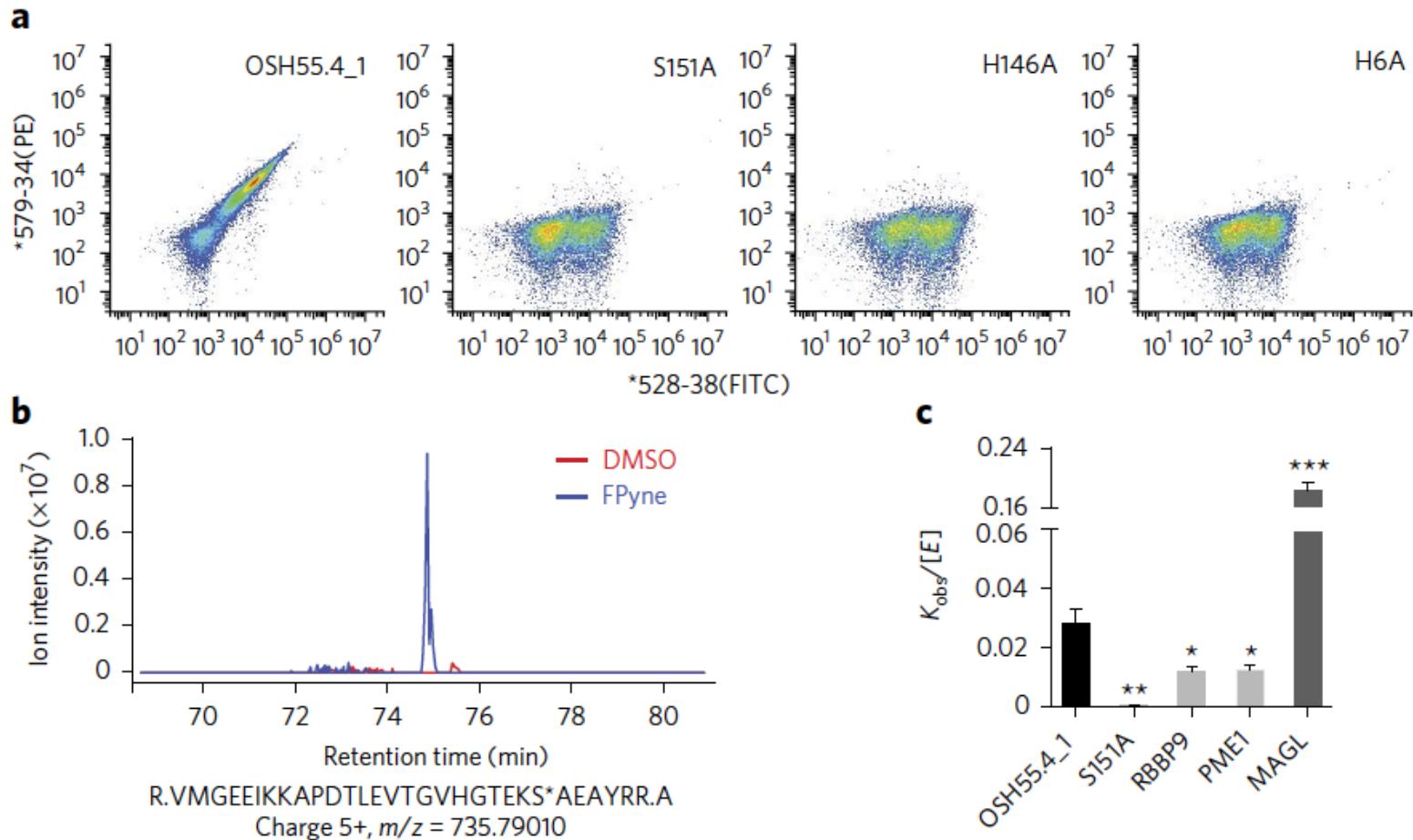
Design of activated serine-containing catalytic triads with atomic-level accuracy



Computational optimization and validation

Designing and engineering enzyme activity-3

Design of activated serine-containing catalytic triads with atomic-level accuracy

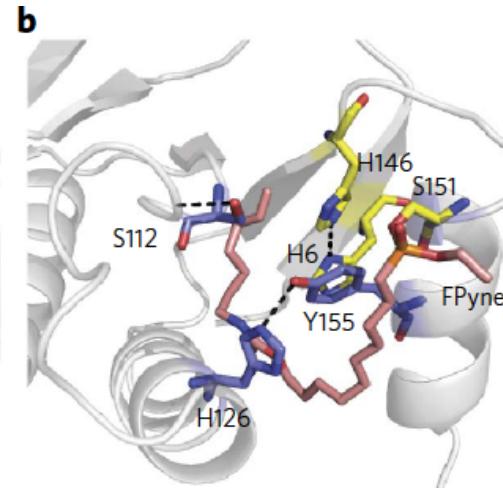
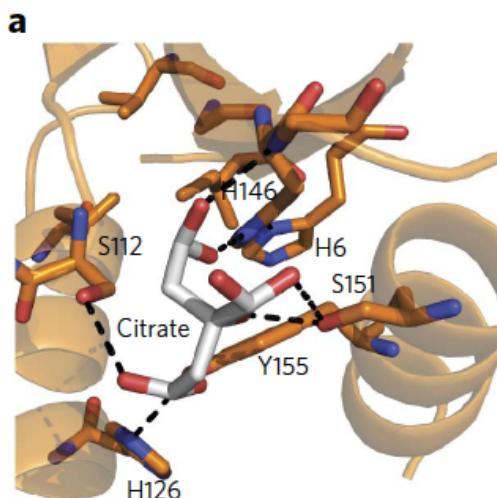


Evolutional optimization and validation

Designing and engineering enzyme activity-3

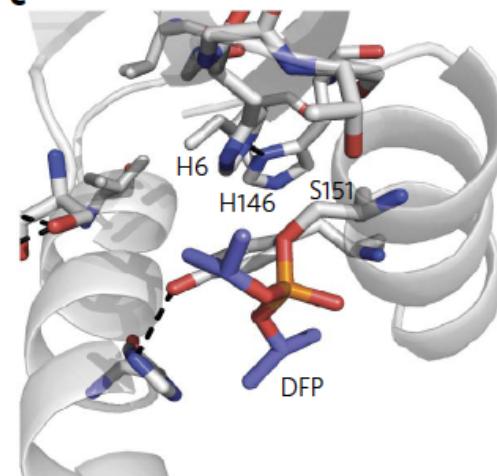
Design of activated serine-containing catalytic triads with atomic-level accuracy

Apo-Citrate

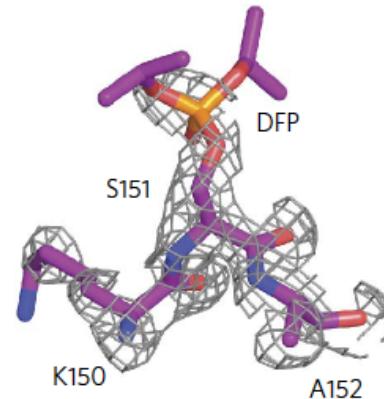


Probe-bound

DFP-bound



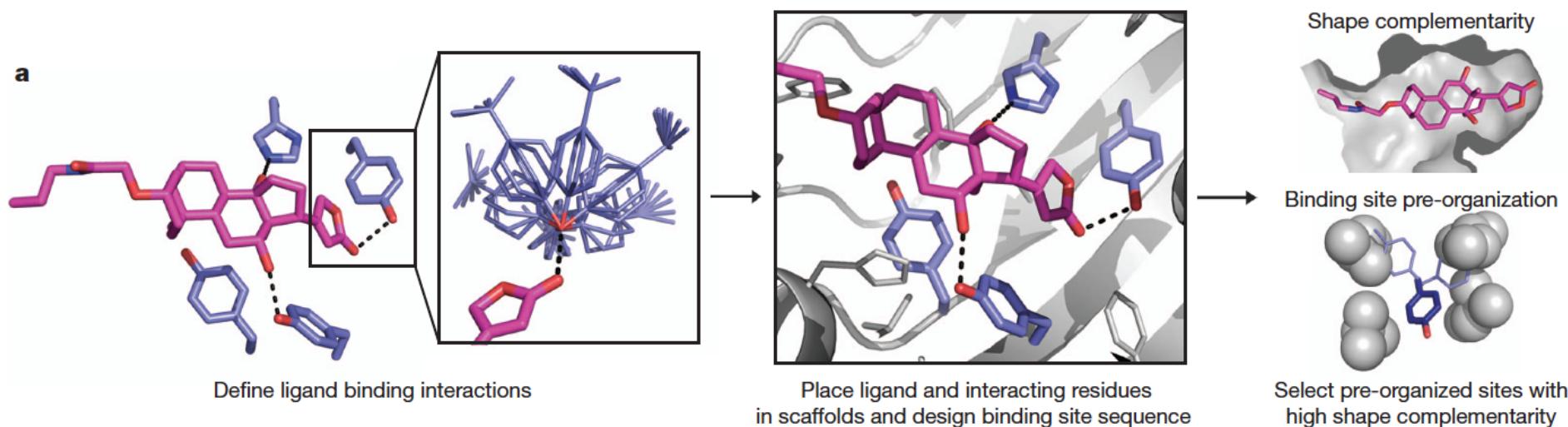
d



Electron-density

Designing and engineering ligand binding affinity and selectivity

Computational design methodology of DIG-binding protein

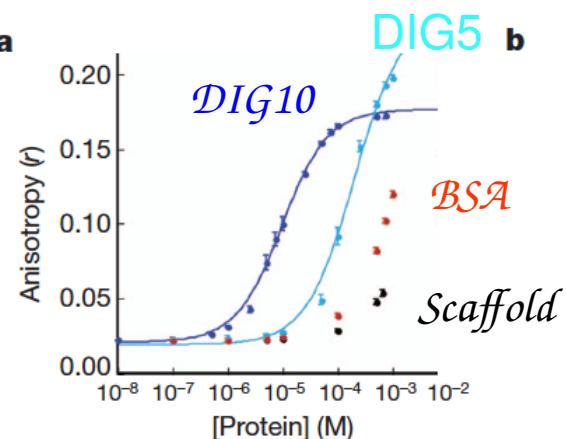


Computational design of ligand-binding proteins with high affinity and selectivity.

Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, Jankowski W, Kalodimos CG, Johnsson K, Stoddard BL, Baker D. Nature. 2013 Sep 12;501(7466):212-6. doi: 10.1038/nature12443. Epub 2013 Sep 4.

Designing and engineering ligand binding affinity and selectivity

Computational design combined with experimental mutagenesis



Variant	K_d for DIG-PEG ₃ -Alexa488 (FP)	K_d for DIG (ITC)
DIG10	$8.9 \pm 1.3 \mu\text{M}$	$12.2 \pm 3.1 \mu\text{M}$
DIG5	$205 \pm 28 \mu\text{M}$	ND
1Z1S	mM (nonspecific)	ND
DIG10.1	$119 \pm 15 \text{nM}$	$196 \pm 25 \text{nM}$
DIG10.2	$8.9 \pm 2.3 \text{nM}$	$168 \pm 59 \text{nM}$
DIG10.3	$541 \pm 193 \text{pM}$	< 6 nM

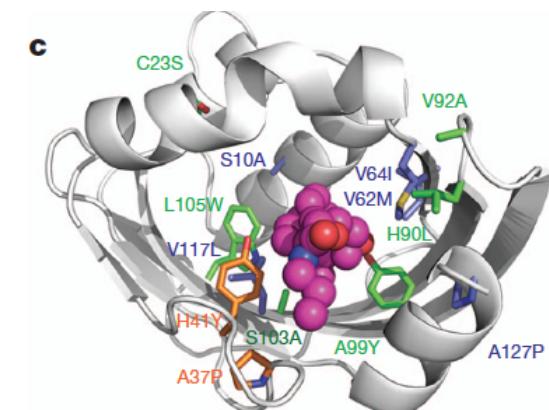
Dissociation constants of designs

Equilibrium fluorescence anisotropy

Saturated mutagenesis

Yeast surface display & FACS

Doped oligonucleotide mutagenesis



Mutations identified during maturation (Blue for 10.1, orange for 10.2 and green for 10.3)

Computational design of ligand-binding proteins with high affinity and selectivity.

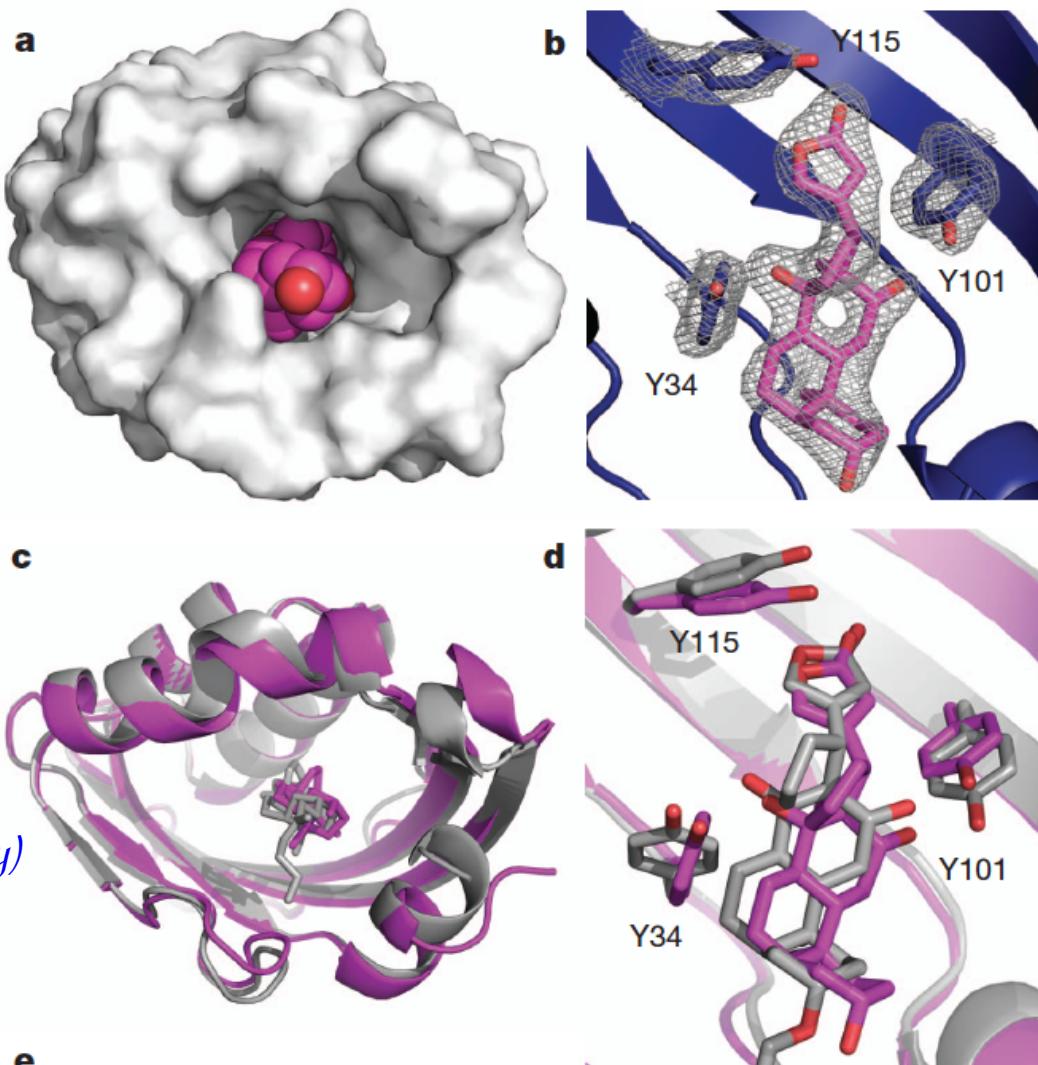
Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, Jankowski W, Kalodimos CG, Johnsson K, Stoddard BL, Baker D. Nature. 2013 Sep 12;501(7466):212-6. doi: 10.1038/nature12443. Epub 2013 Sep 4.

Designing and engineering ligand binding affinity and selectivity

DIG10.2-DIG complex

Structure validation

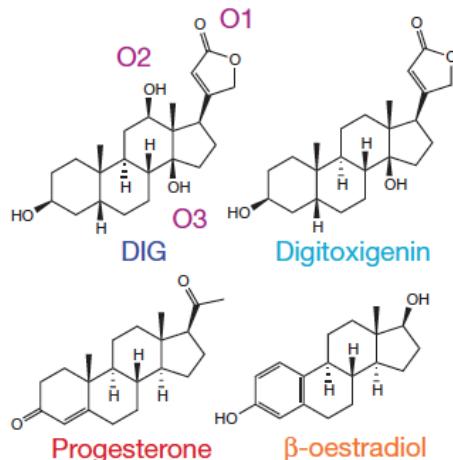
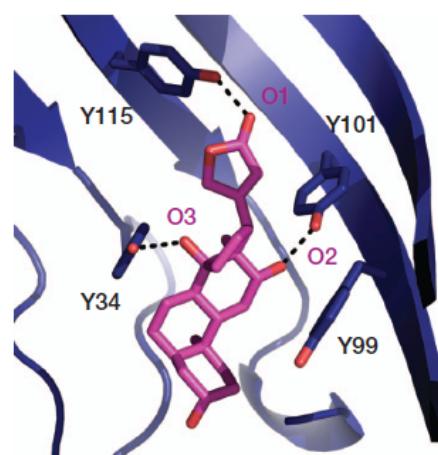
Comparison with model (grey)



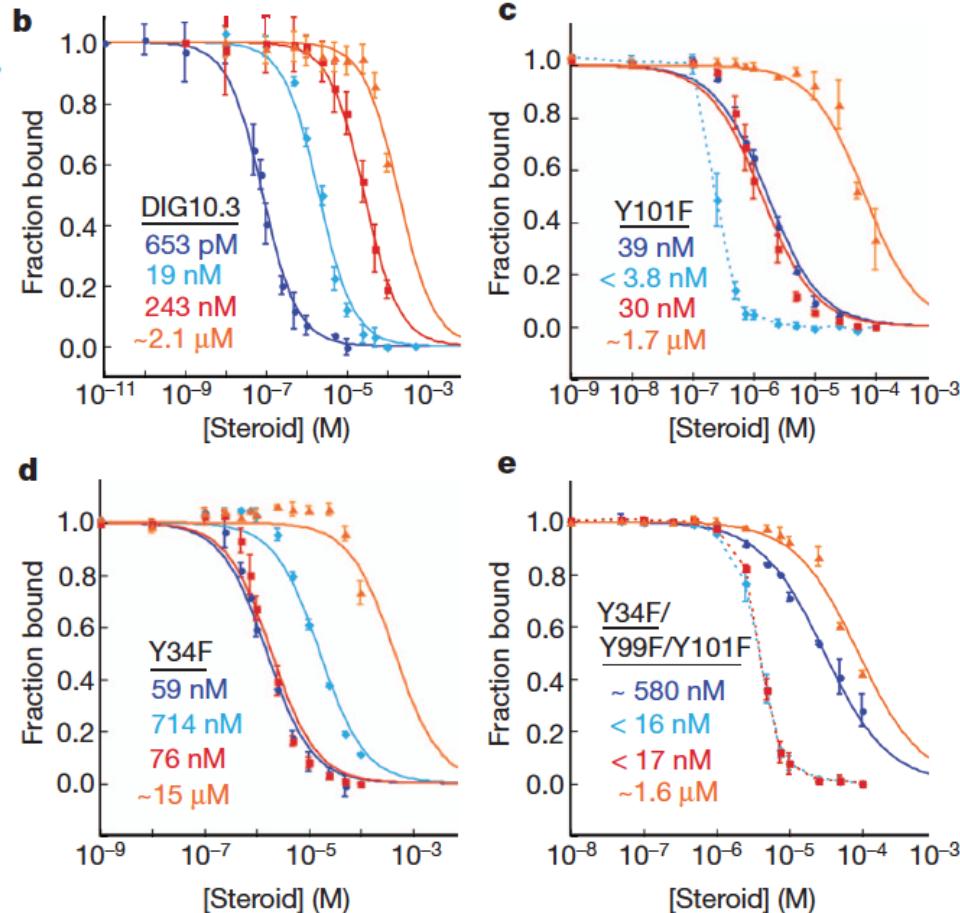
Computational design of ligand-binding proteins with high affinity and selectivity.

Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, Jankowski W, Kalodimos CG, Johnsson K, Stoddard BL, Baker D. Nature. 2013 Sep 12;501(7466):212-6. doi: 10.1038/nature12443. Epub 2013 Sep 4.

Designing and engineering ligand binding affinity and selectivity



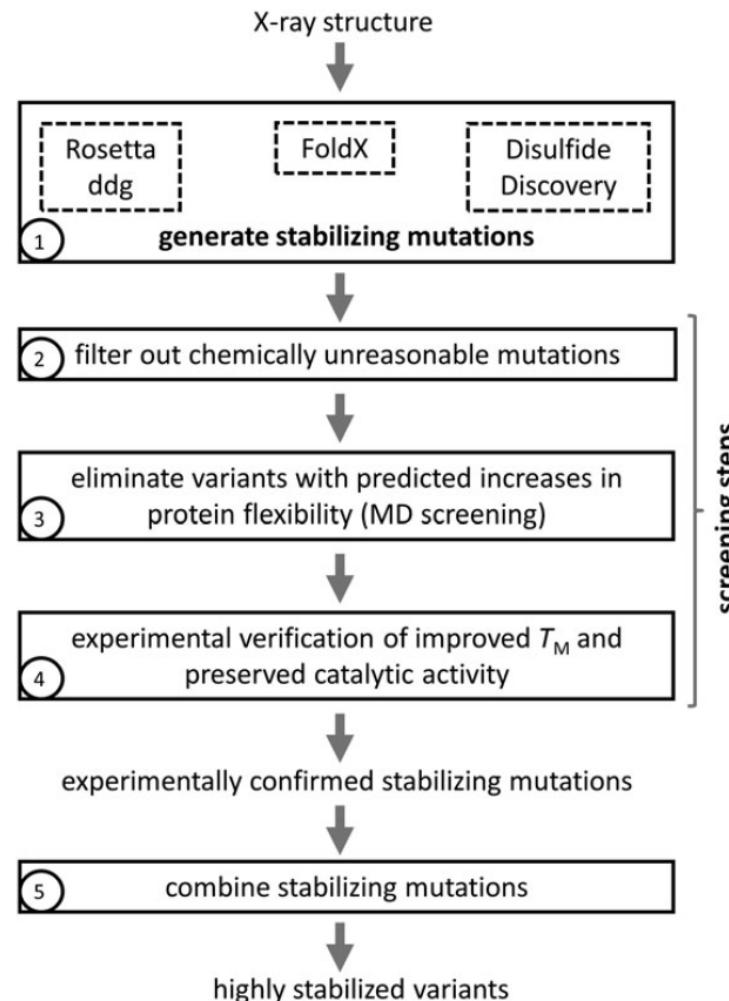
Binding selectivity among different ligands



Designing and engineering enzyme stability and selectivity

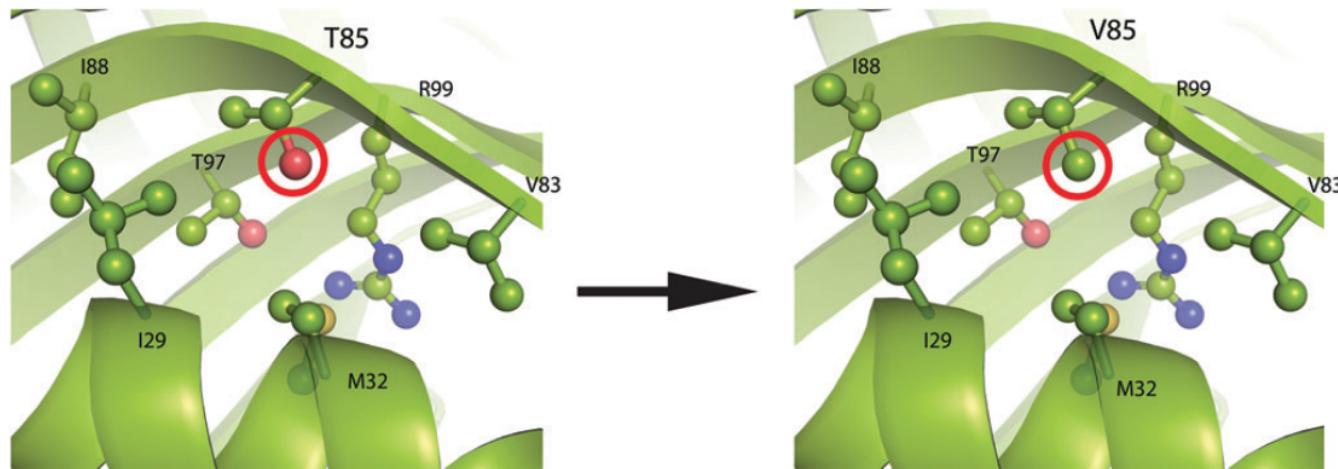
Computationally designed libraries for rapid enzyme stabilization

Framework for Rapid Enzyme Stabilization by Computational libraries (FRESCO)

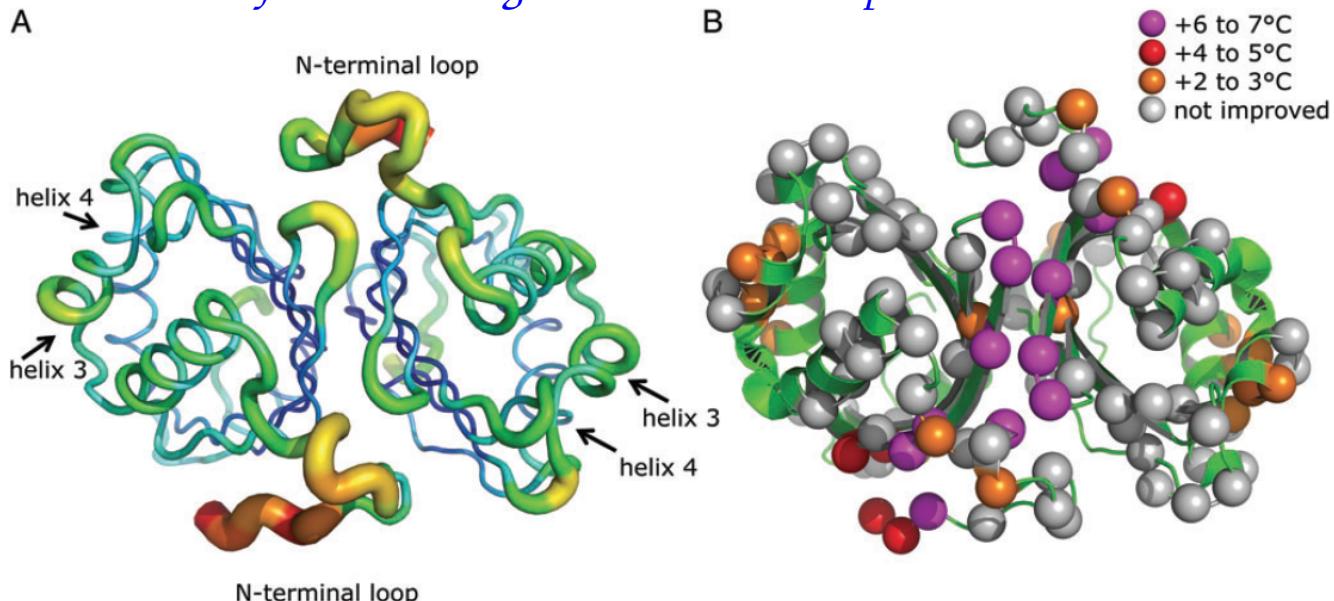


Designing and engineering enzyme stability and selectivity

Examples of the predicted structure for a stabilization mutation

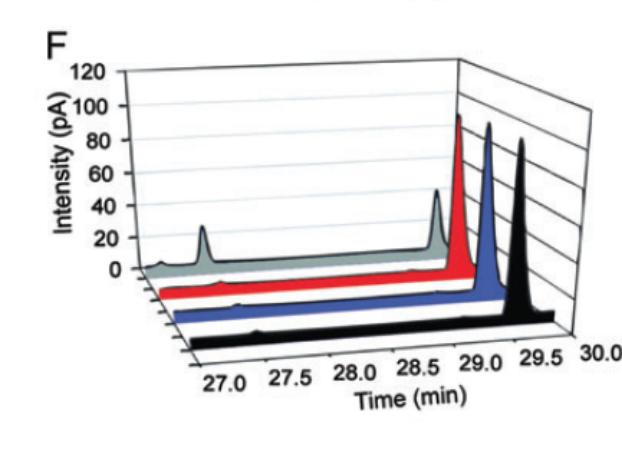
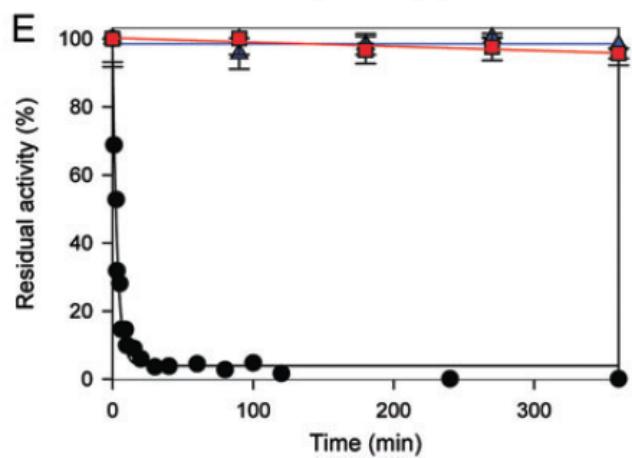
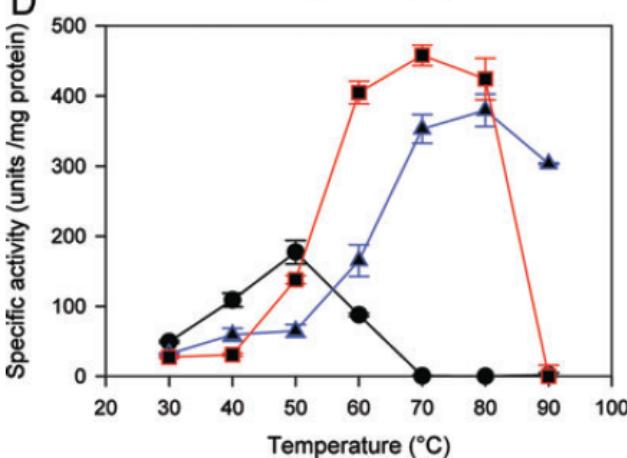
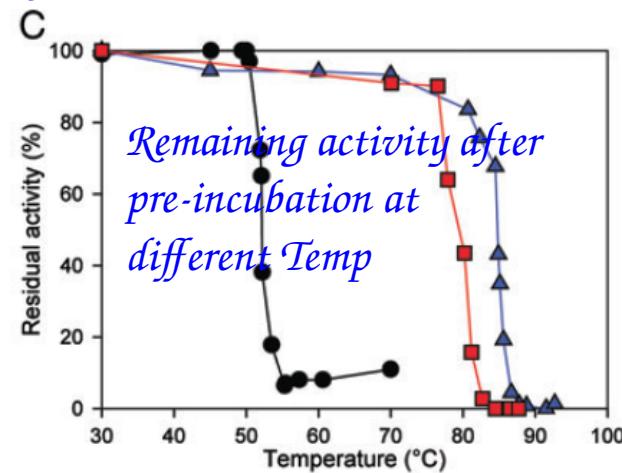
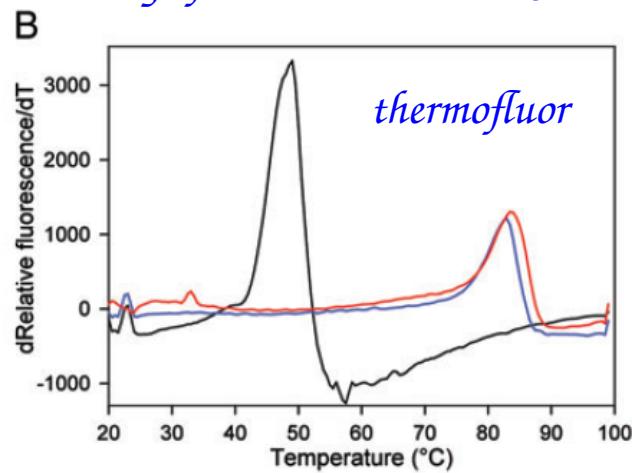
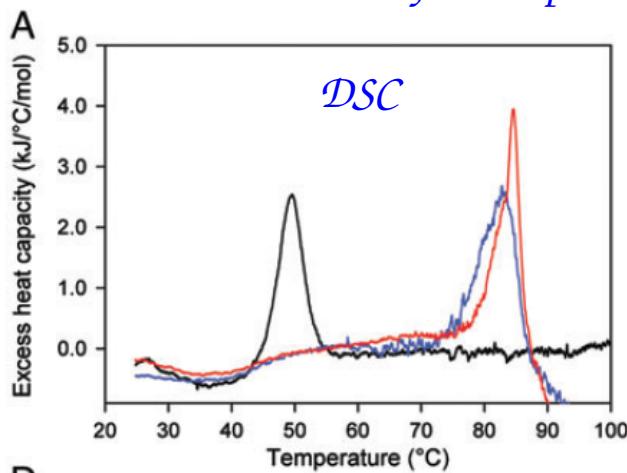


Distribution of the stabilizing mutations over the protein



Designing and engineering enzyme stability and selectivity

Overview of the improved stability of the *LEH* mutants F1 and F2



Designing and engineering enzyme stability and selectivity

Table II. Catalytic parameters of WT LEH and variants F1 and F2

Variant	WT	F1	F2
Temperature (°C)	30	50	30
$k_{\text{cat}} (\text{s}^{-1})$	13.9 ± 0.8	63 ± 4	8.9 ± 0.4
$K_M (\text{mM})$	0.3 ± 0.1	0.6 ± 0.1	$<0.25^{\text{a}}$
$k_{\text{cat}}/K_M (\text{s}^{-1} \text{ M}^{-1})$	4.6×10^4	1.0×10^5	$>3.6 \times 10^4$
			80
			30
			70
			8.2 ± 0.3
			160 ± 7
			0.3 ± 0.1
			4.9×10^5

- 1 *T_m increase from 50°C to 85°C*
- 2 *Activity increase*
- 3 *Half-life >250 fold longer*