COMPUTATIONAL DESIGN OF PROTEINS AND ENZYMES

VAITEA OPUU

Laboratoire de Biologie Structurale de la Cellule,
École Polytechnique

October 29, 2020

Jury Members

Anne-Claude Camproux	Université Paris-Diderot	Rapporteur
Marc Delarue	Institut Pasteur	Rapporteur
Martin Weigt	Sorbonne Université	Examinateur
Sebastian Will	École Polytechnique	Examinateur
Thomas Simonson	École Polytechnique	Directeur de thèse

Design biomolecular systems

Whole protein design

- Complete redesign of a PDZ domain involved in protein-protein interactions

Enzyme design

- Design of MetRS binding site to increase activity for unnatural amino acids

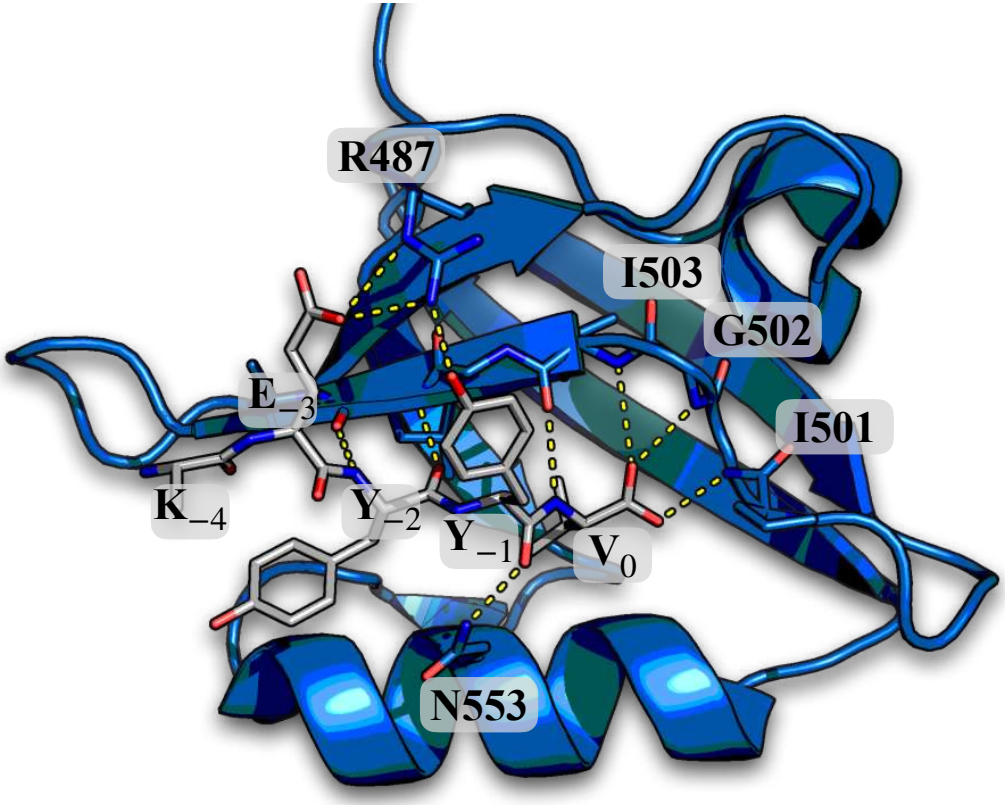
Explore the benefits/limits of the physics-based approach

- Molecular mechanics model
- Boltzmann sampling

Design PDZ pairs with overlapping genes

- Algorithm development
- PDZ application

PDZ DOMAINS



PDZ STRUCTURE

- ~ 90 Amino acids
- 6 β sheets
- 2 α helices (α_1 and α_2)

BIOLOGICAL CONTEXT

- Protein-Protein recognition
- Proteins activity modulation

DESIGN GOAL

Search for sequences that adopt this fold

$\Longrightarrow$ **Solve inverse folding problem**

Main ingredients

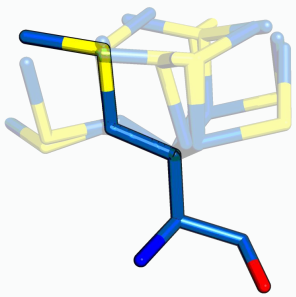
Molecular Mechanics + Continuum Electrostatics

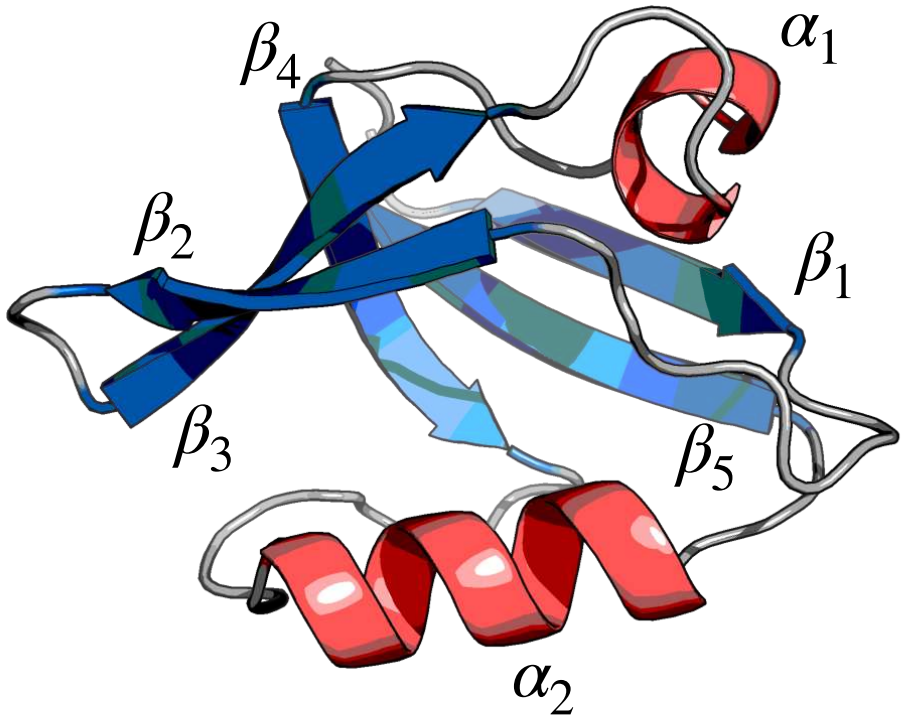
$$E = E^{MM} + E^{GB} + E^{NONPOLAR}$$

Empirical unfolded state

$$\sum_{aa} E^{uf}_{aa}(type_{aa})$$

Discrete conformations





- Experimental geometry for the backbone
- Fixed backbone during the exploration
- Discrete rotamers

Numerical procedure

Design software:

Proteus (Mignon, Druart, Michael, Opuu, Polydorides, Villa *et al*, JPCA, 2021)

Energy function:

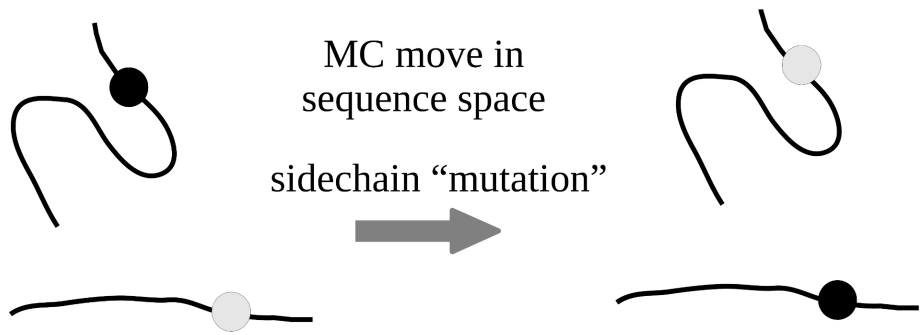
- Amber ff99SB
- pre-calculation in a lookup table
- Unfolded state parametrized empirically

Sequence space:

- Proline & glycine not allowed to mutate
- 13 positions involved in peptide binding are not allowed to mutate
- All 61 others mutate freely: 10^{76} possible sequences

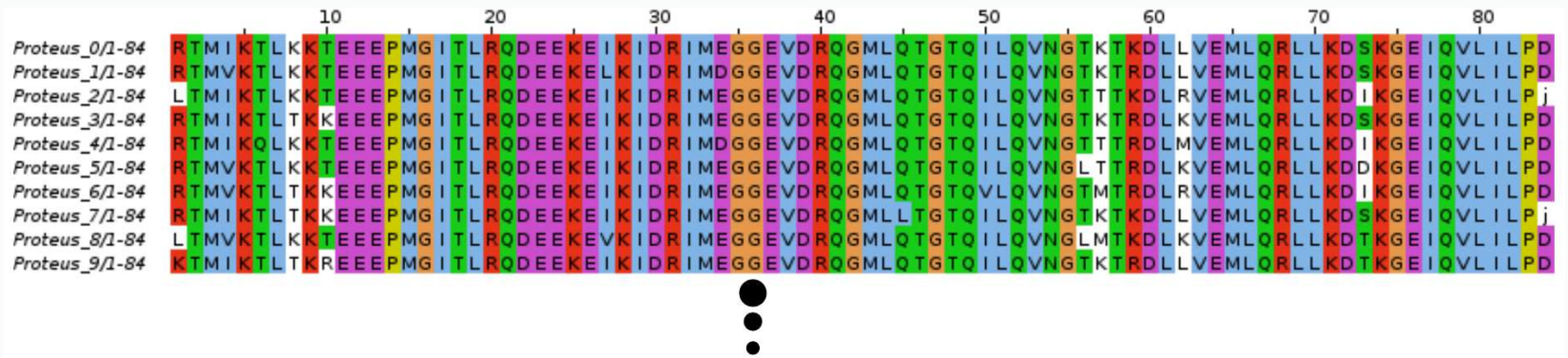
Sampling method:

- Replica exchange MC
- 10^9 Monte Carlo steps per replica

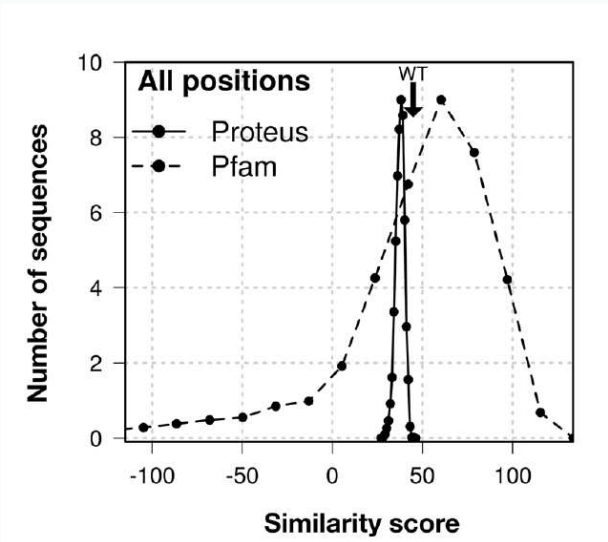


Designed sequences resemble natural sequences

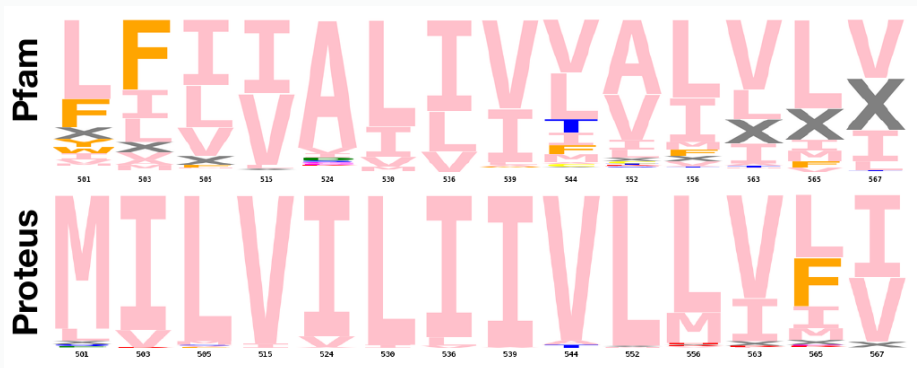
$\simeq 10^5$ designed sequences



37% mean identity

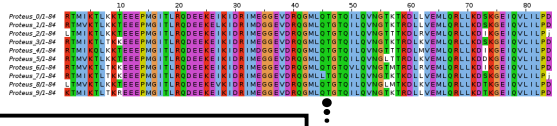


Core positions: Proteus vs Pfam



SELECTING SEQUENCES FOR EXPERIMENTAL TESTING

Top 2 000 sequences



Proteus energy<1.5 kcal/mol

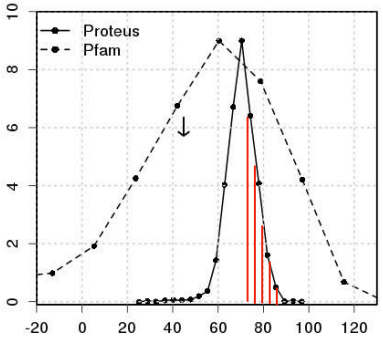
Solubility → **Isoelectric point**

1268 sequences

Negative design → **Superfamily score above average**

692 sequences

Ad-hoc → **Similarity**

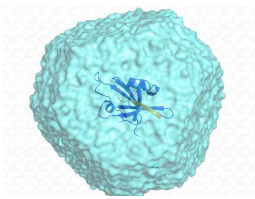


215 sequences

Net charge, ionic mutations

15 sequences

Cluster



6 sequences

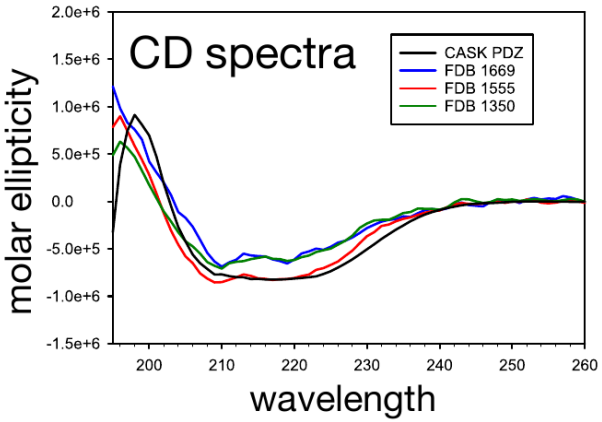
Molecular dynamics

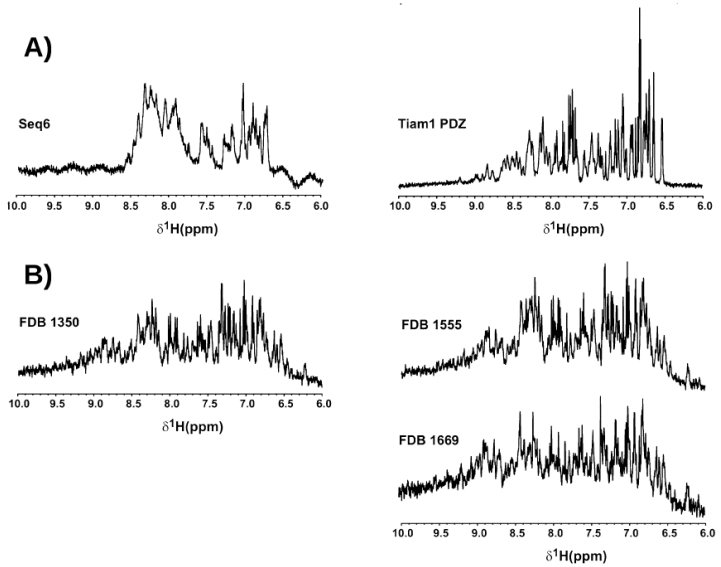
3 for experiments

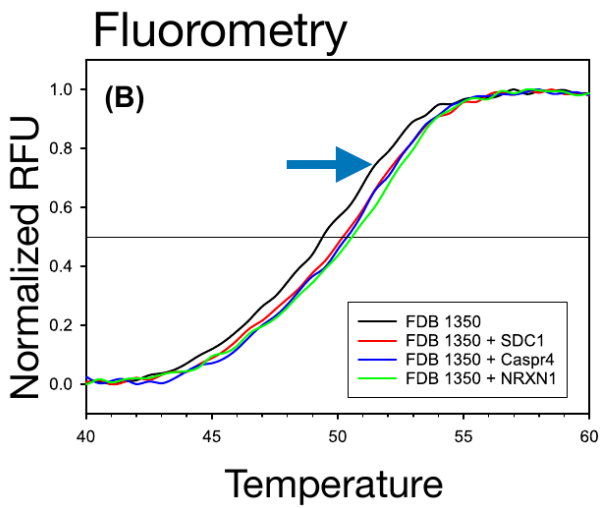
52~53 mutations out of ~90 positions in tested sequences!

EXPERIMENTS CONFIRM ALL 3 DESIGNS

(E. Fuentes, U. Iowa)







RESULTS

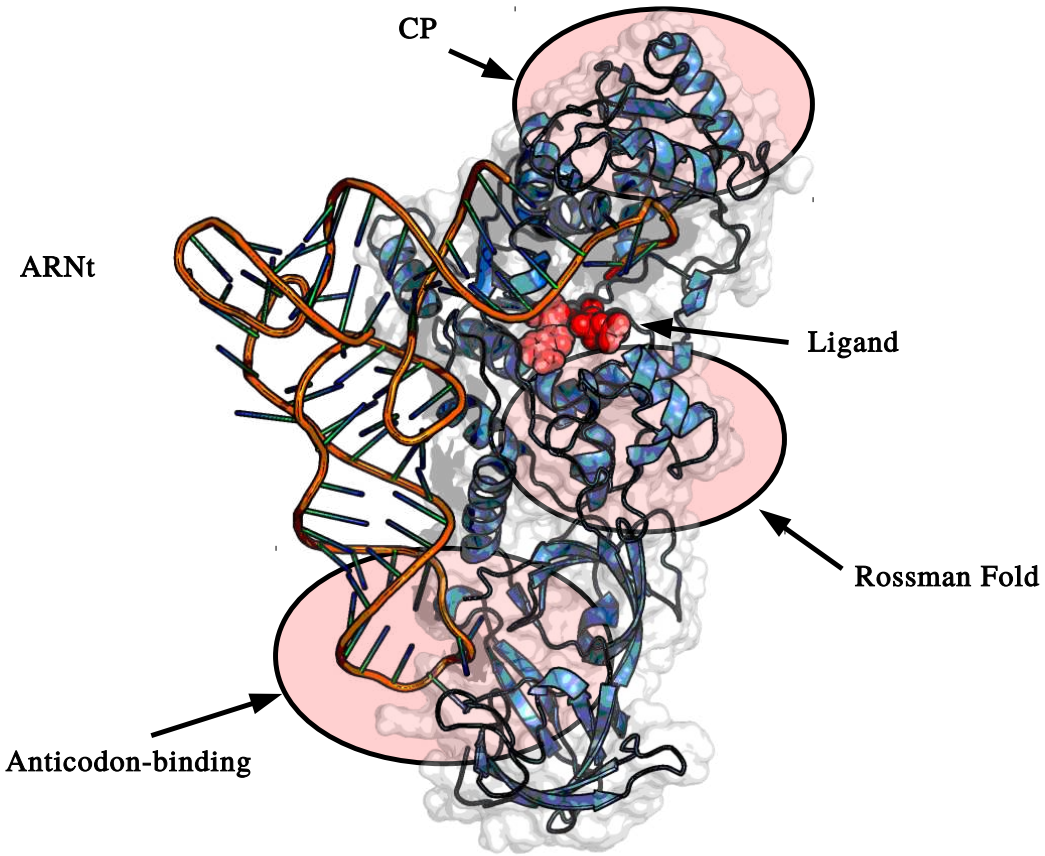
- PDZ-like secondary structure content
- 1D-NMR spectra typical of folded protein
- Thermal denaturation upshifted by peptide binding

$\Longrightarrow$ Designed sequences adopt PDZ fold

First successful protein redesign with a physics based energy function

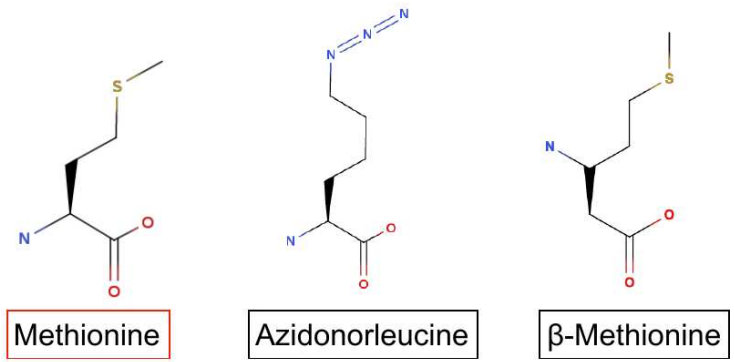
Opuu *et al* (2020) Sci. Rep.

Engineering MetRS for ligand:substrate binding and catalytic power



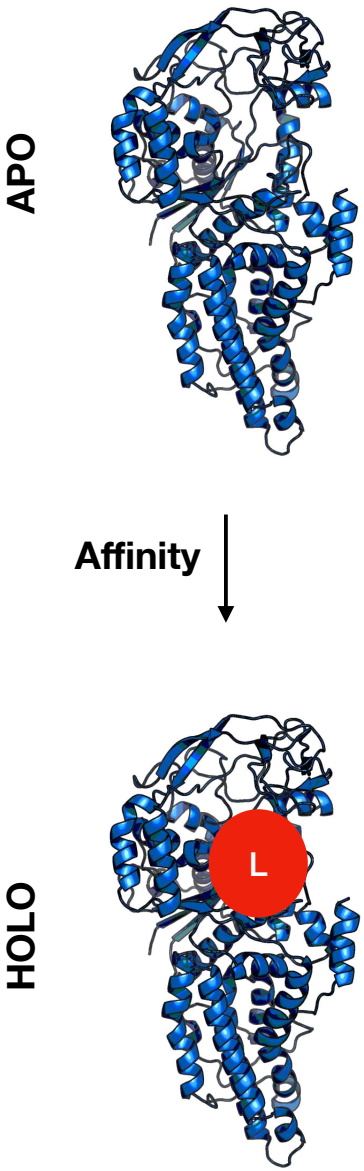
$$Met + ATP \rightleftharpoons MetAMP + PPi$$

$$MetAMP + tRNA \rightleftharpoons Met\text{-}tRNA + AMP$$



UAAs

Designing for affinity or catalysis is very difficult



Calculation strategies

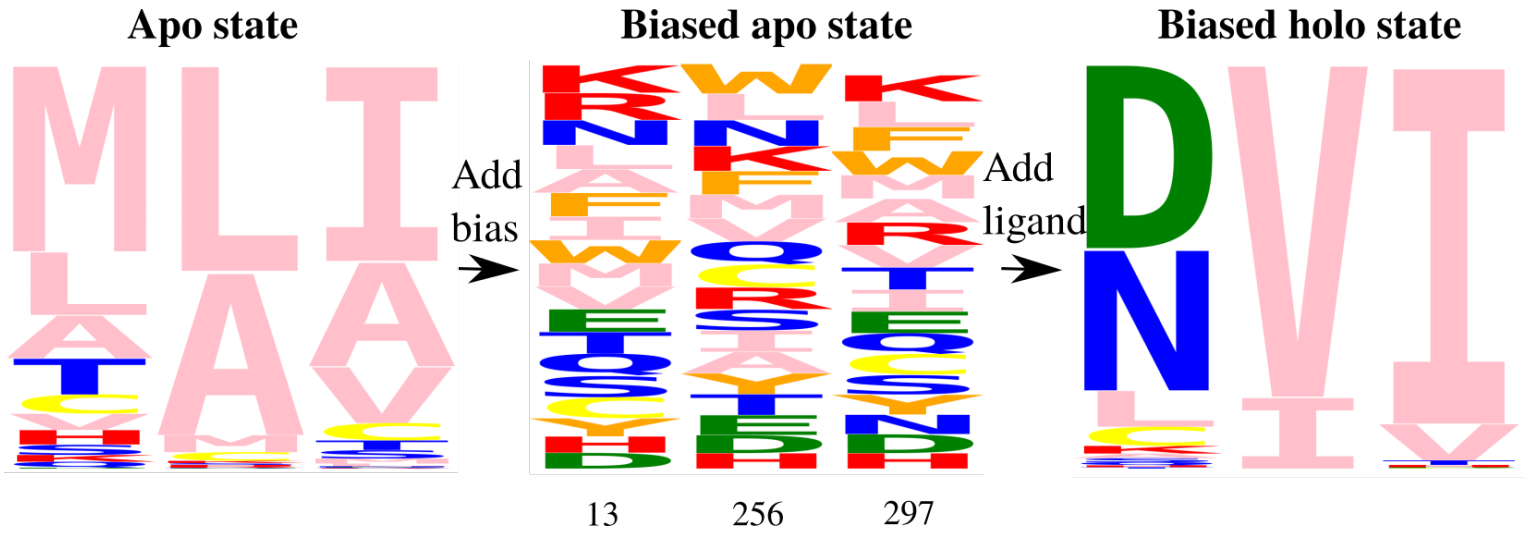
- Large combinatorial space
- Simultaneous optimization directions: stability, affinity, catalysis
- Positive design for the bound state and negative design for the unbound state
- Existing methods are heuristic or very expensive
  - optimize the bound state energy (Rosetta)
  - exhaustive enumeration of states (Osprey)

Physical description

- Transition state is difficult to characterize
- Electrostatics is an important component of the affinity

A RIGOROUS METHOD TO DESIGN FOR AFFINITY: ADAPTIVE LANDSCAPE FLATTENING

Villa, Panel, Chen, Simonson 2018; Bhattacherjee Walling 2013

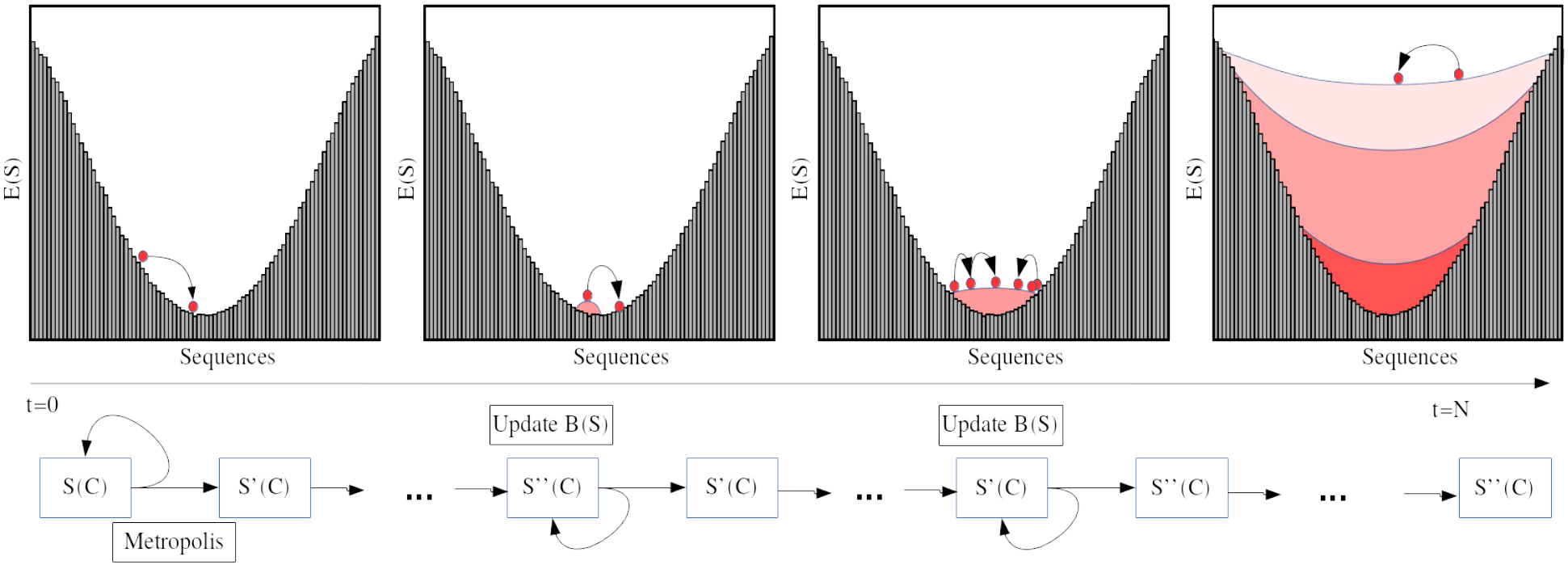


SIMULATE HOLO STATE WITH THE SAME BIAS:

The bias **substracts out** the apo free energy

Sequences populated according to their binding free energy

THE ADAPTIVE LANDSCAPE METHOD APPLIED TO SEQUENCES



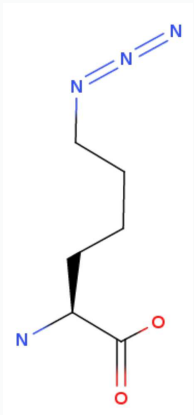
- Update scheme borrowed from WellTempered Metadynamic

$$\mathrm{inc}(B(S_i(t);t)) = e_0 \times \exp\left(-\frac{E_i^B(S_i(t);t)}{E_0}\right)$$

$$\mathrm{inc}(B(S_i(t),S_j(t);t)) = e_0 \times \exp\left(-\frac{E_i^B(S_i(t),S_j(t);t)}{E_0}\right)$$

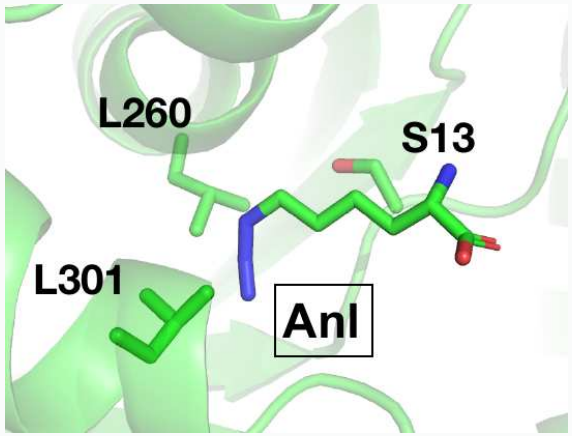
First test: Redesign MetRS for Azidonorleucine binding

Azidonorleucine (Anl)



- Cell localization
- Cell identification
- Protein labeling

Experimental data



(Tanirkulu *et al*, 2009)

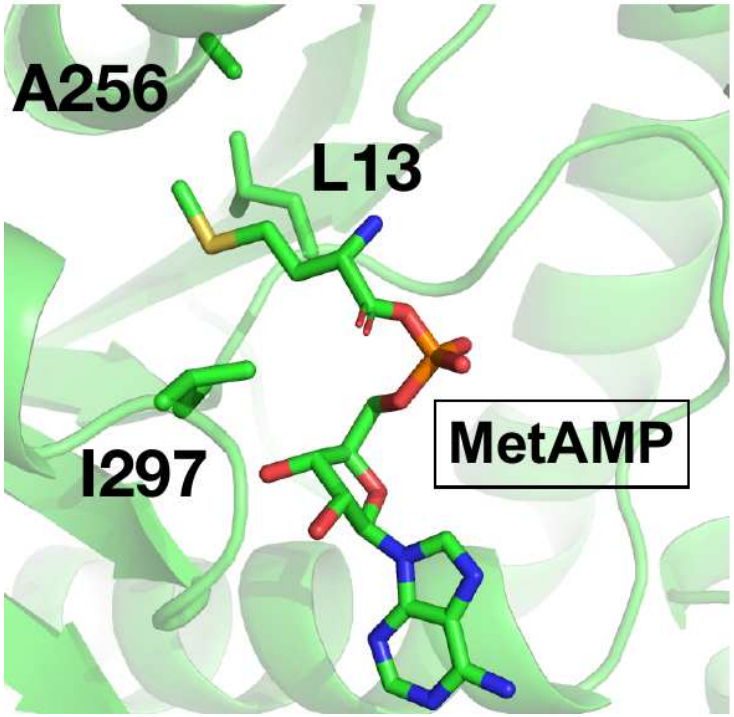
- 21 experimental variants
- 3 variable positions
- 2744 possible variants considered

Main results:

5 of 6 most active variants are among the top 100 predictions

2 most active variants are among the top 10 specificity predictions

2ND TEST: REDESIGN METRS FOR MET BINDING



SYSTEM DATA

- 3 variable positions
- KMSKS loop in its "inactive" conformation

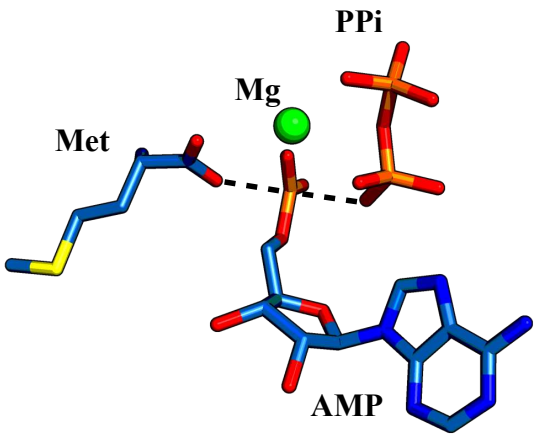
- 528 variants sampled during simulations out of 5832 allowed
- 20 experimental variants tested, based on the variants predicted
- Among top 40 predictions, 5 are active
- Computed affinities in good agreement with experiment:

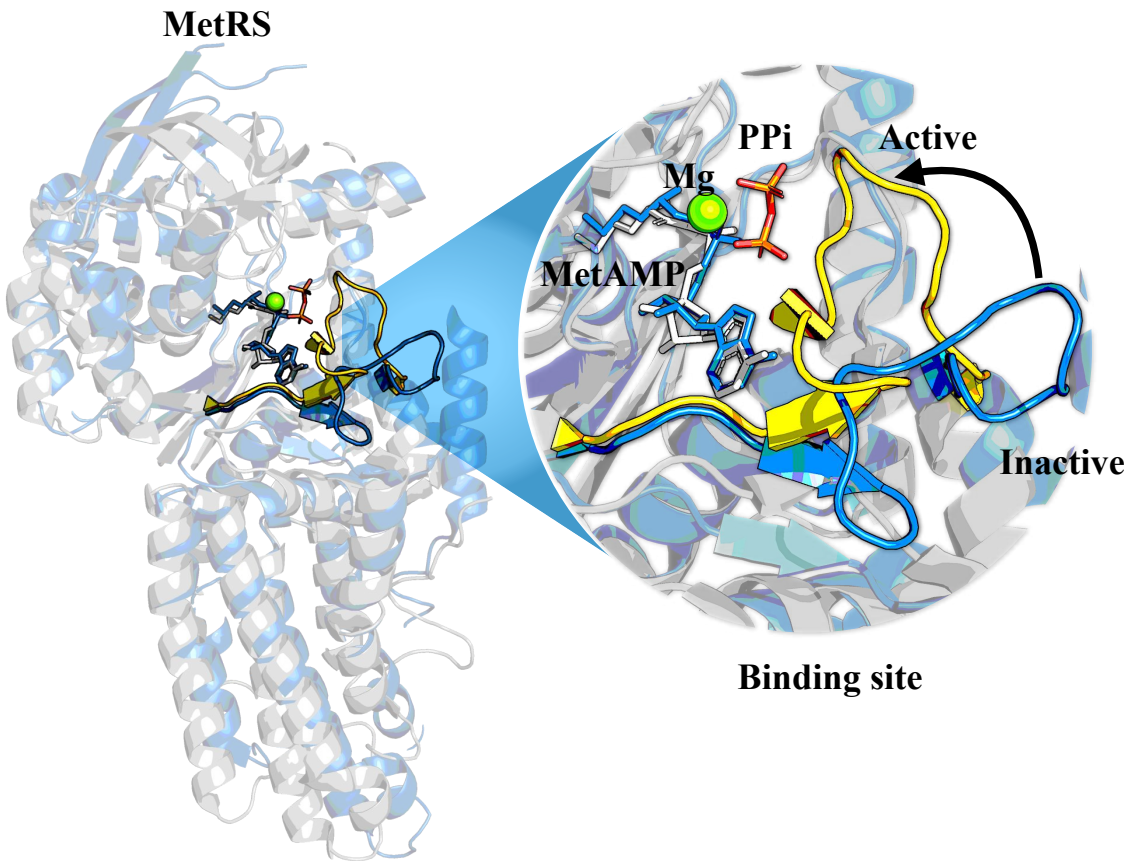
0.9 kcal/mol rms error, 0.75 correlation

Opuu *et al*, Plos Comp. Bio. 2020

Adaptive landscape flattening allows us to design for catalysis

Transition state



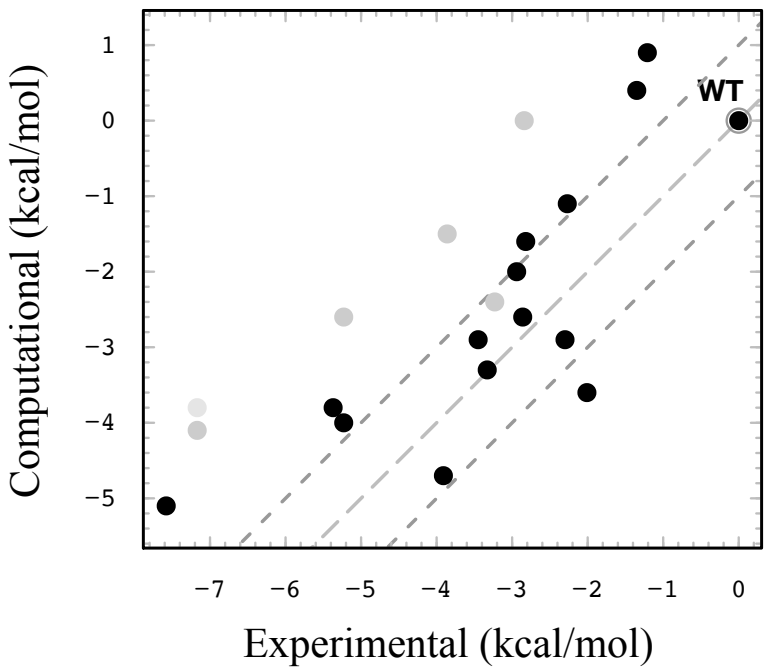


- Mg and ATP are **rigid**
- **"APO"** state: MetRS:ATP
- **HOLO** state: MetRS:[Met:ATP]‡

Variants can now be selected by their catalytic efficiency k_{cat}/K_M

Catalytic efficiency = k_{cat}/K_M

$$\Delta G^{\ddagger} \stackrel{def}{=} kT \times ln\left(\frac{k_{cat}}{K_M}\right)$$

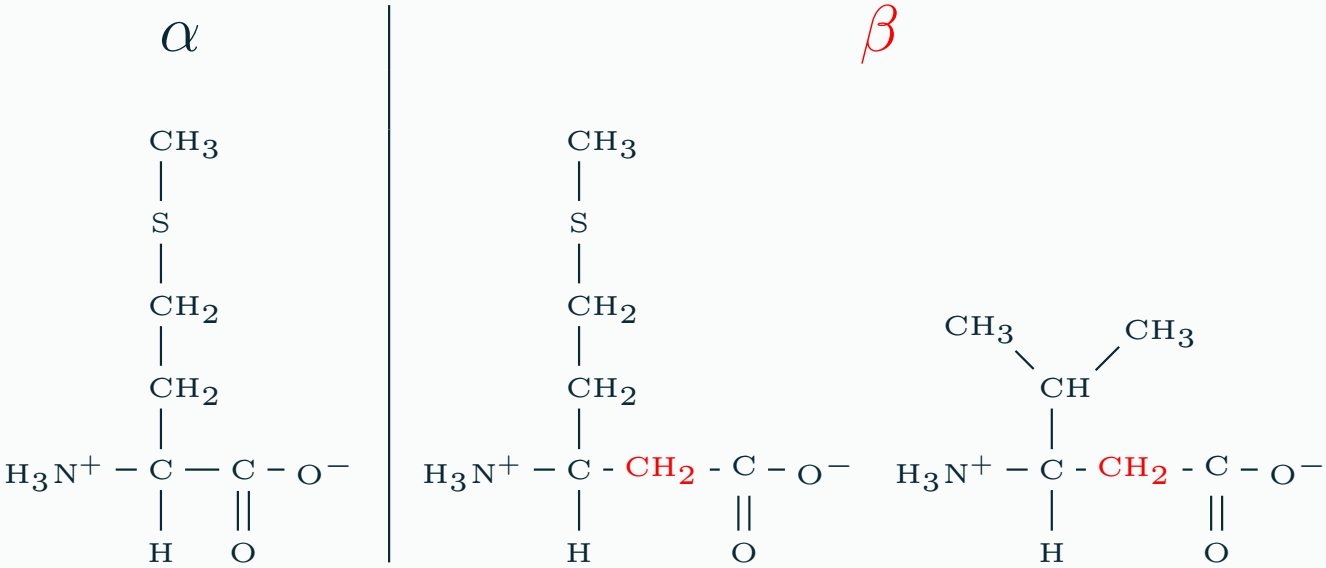


- 20 experimental values (above)
- 13 among top predictions
- Good agreement for k_{cat}/K_M
- 0.8 correlation
- 1.1 kcal/mol rms error

Structural models + physical description capture the important effects

Design MetRS for β amino acid activity

β amino acids



β polypeptide

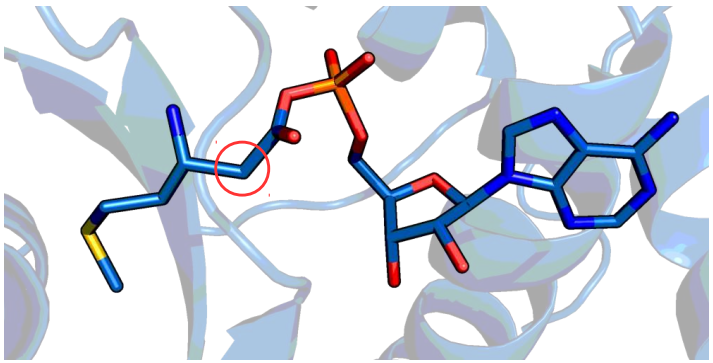
- Allow new backbone geometries
- Reduce peptide degradation

Challenges

- Structural models
- Force field parameters
- Which positions to mutate

1ST ROUND: DESIGN 3 POSITIONS FOR BETA-METAMP AFFINITY

β-METAMP



SYSTEM DATA

- ligand: **β-MetAMP**
- 3 mutable positions: 13, 256, 297
- APO state: MetRS
- HOLO state: MetRS:β-MetAMP

- 206 variants sampled by simulations
- **20** variants experimentally tested, 11 from MC (blind tests)
- Fair agreement for k_{cat}/K_M
- **5 out of 11 have a weak but measurable activity**

☺ **3 variants have decreased α-Met selectivity (factors 2-8)**

☹ **Activity lower than WT β-Met activity**

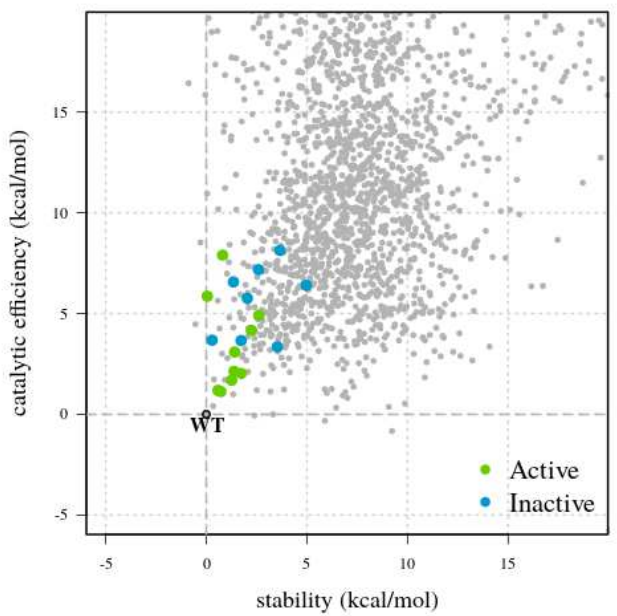
2ND ROUND: DESIGN 3 POSITIONS FOR CATALYTIC EFFICIENCY

SYSTEM DATA

- Same positions can mutate
- Apo state: MetRS:ATP
- Holo state: MetRS:β-Met‡

$\Delta G^{\ddagger}(\beta\text{-Met})$

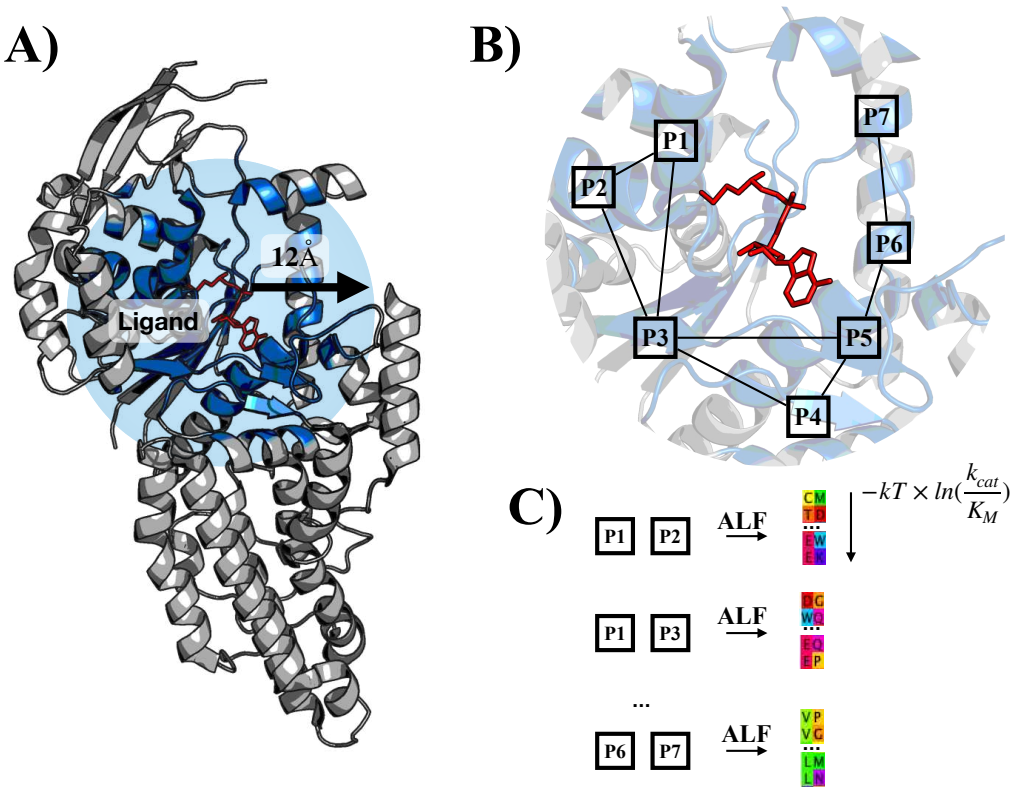
variants	predicted	experiments
CAC	4.5	2.2
CAI	1.0	1.6
CAV	2.4	3.6
LAC	3.2	2.0
LAI	0.0	0.0
LAT	2.3	3.1
LAV	1.4	3.1
MAC	8.1	2.5
MAV	6.1	3.1
SAC	5.3	2.9
SAI	2.0	1.9



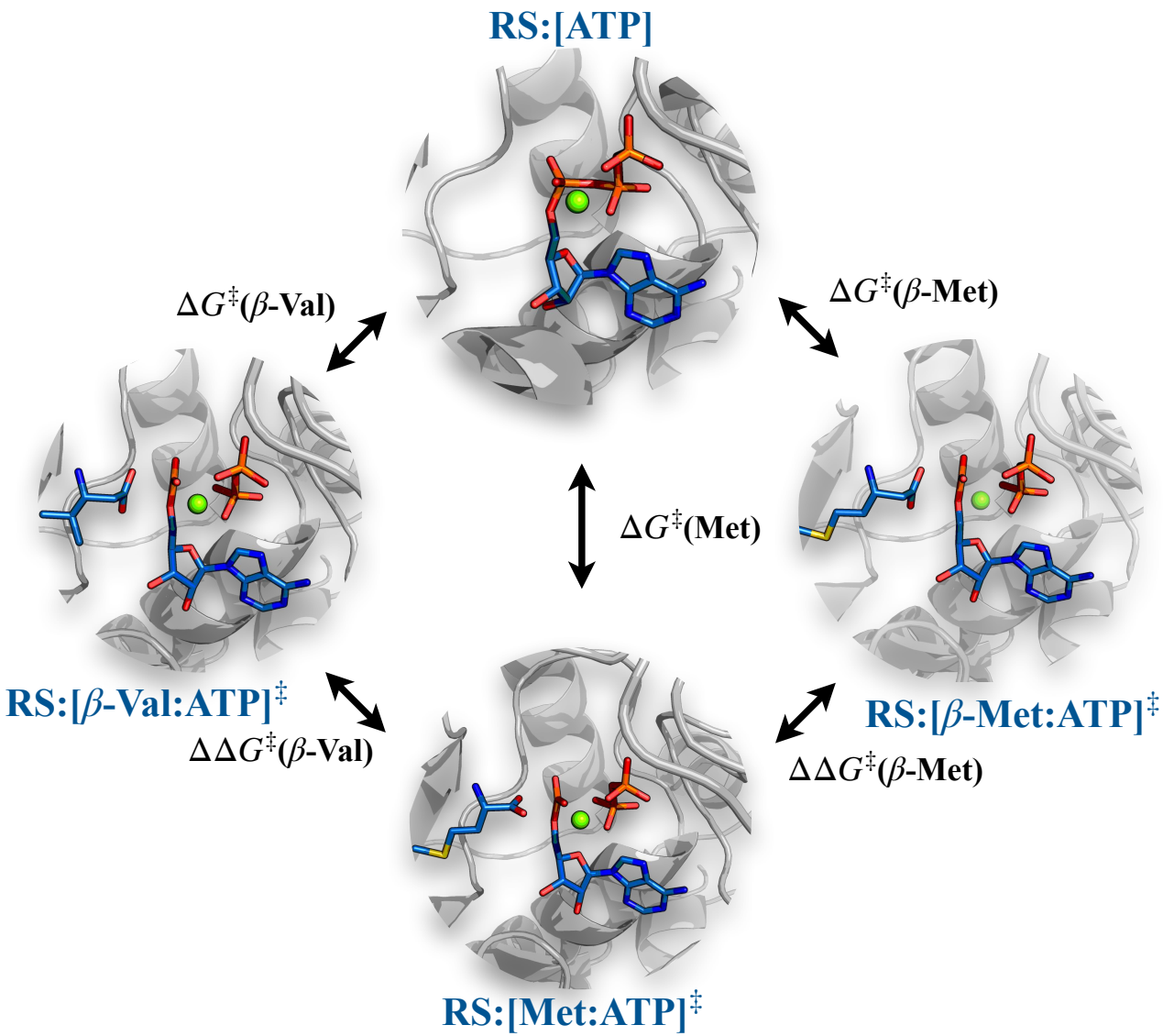
- Fair agreement for k_{cat}/K_M
- Still no variants with improved activity

3RD ROUND: SCAN THE WHOLE ACTIVE SITE

- Select all the positions in the active site (19 positions)
- Design pairs of positions
- Compute a "pair-activity" score
- Select positions with high scores, to form a few quartets
- Redesign quartets



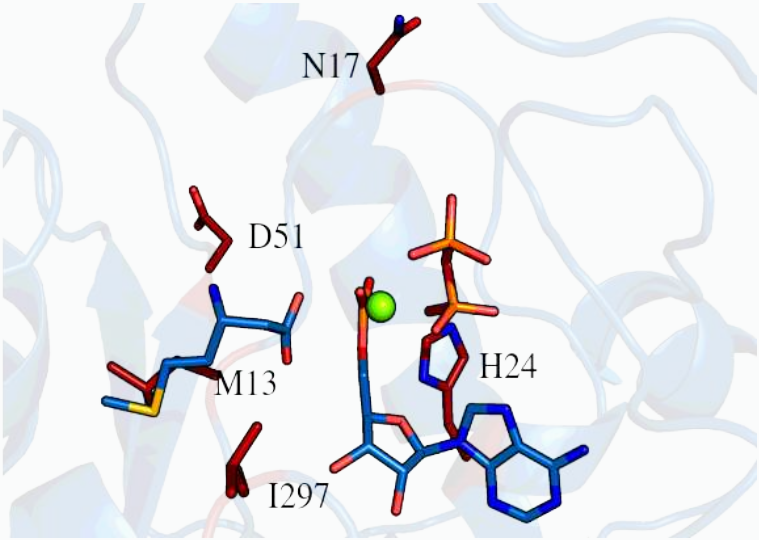
SYSTEMS CONSIDERED



Design four positions at once

QUARTET SELECTIONS

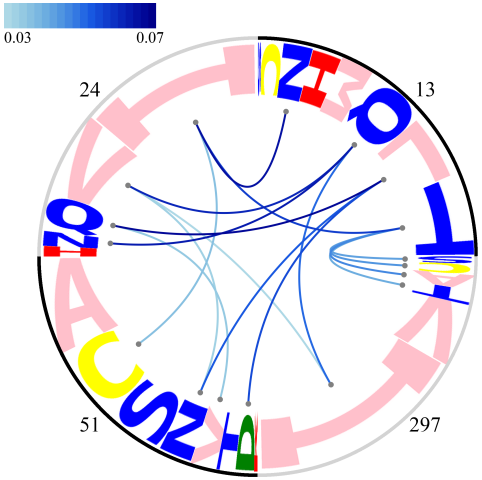
- Positions are in the first and second layer of the binding site
- 3 quartets of 4 positions considered



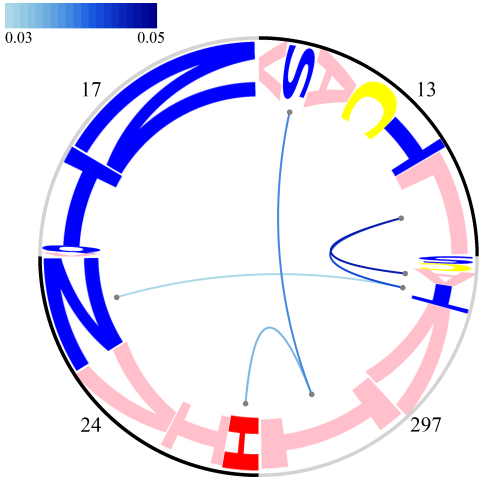
SELECTED VARIANTS

- 2970 + 3168 + 720 variants sampled by MC
- 89 + 81 + 37 variants satisfy various thresholds (stability, catalytic efficiency and selectivity)

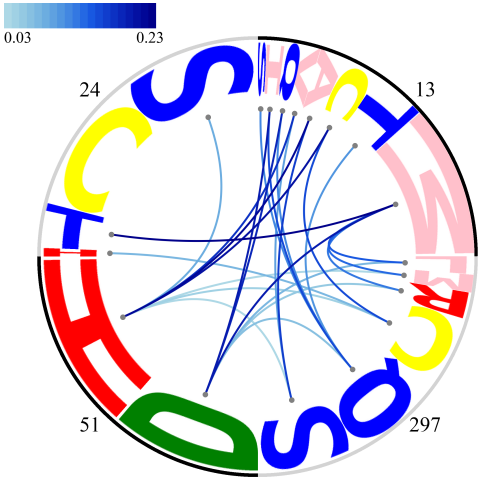
β-MET DESIGNS



Quartet 1



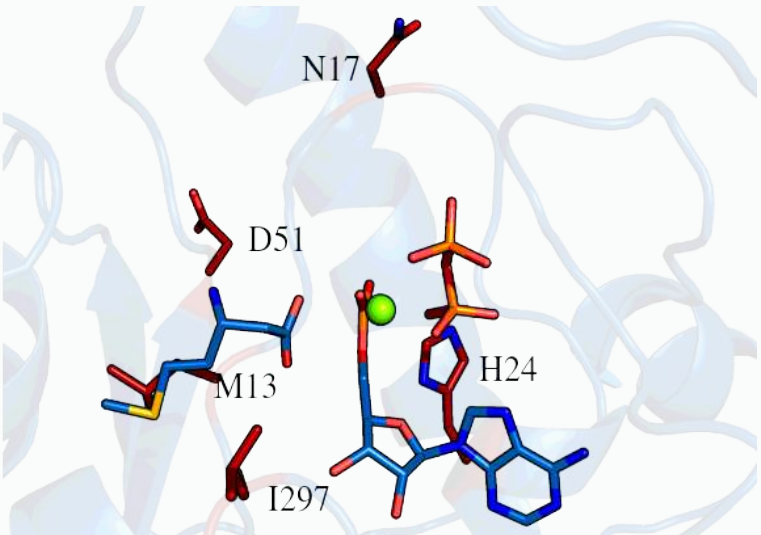
Quartet 2



Quartet 3

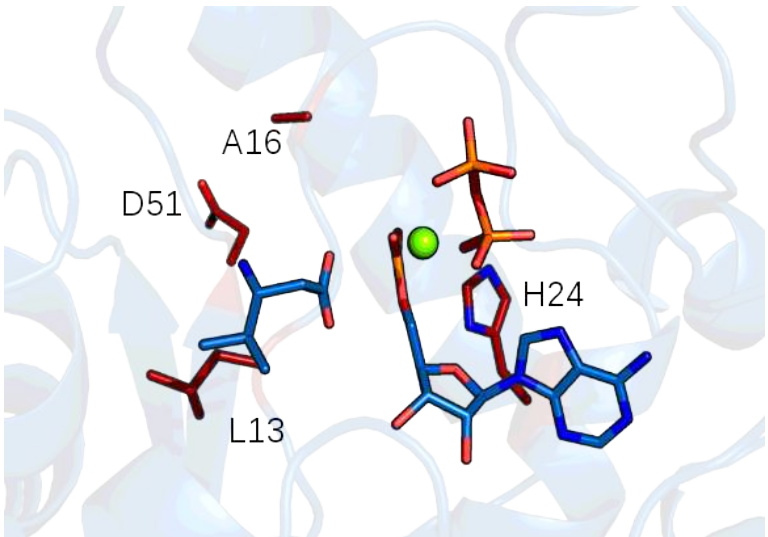
DESIGNS

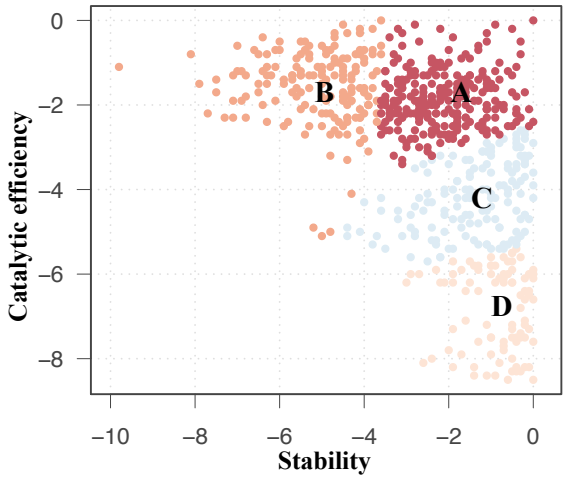
- 2 positions are proposed to be mutated into smaller and more polar types
- ASP 51 is proposed to be mutated into HIS
- 2 positions are maintained in their native types

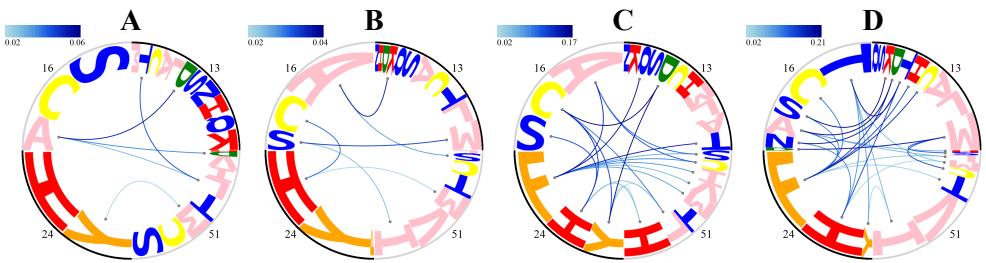


SELECTION OF POSITIONS

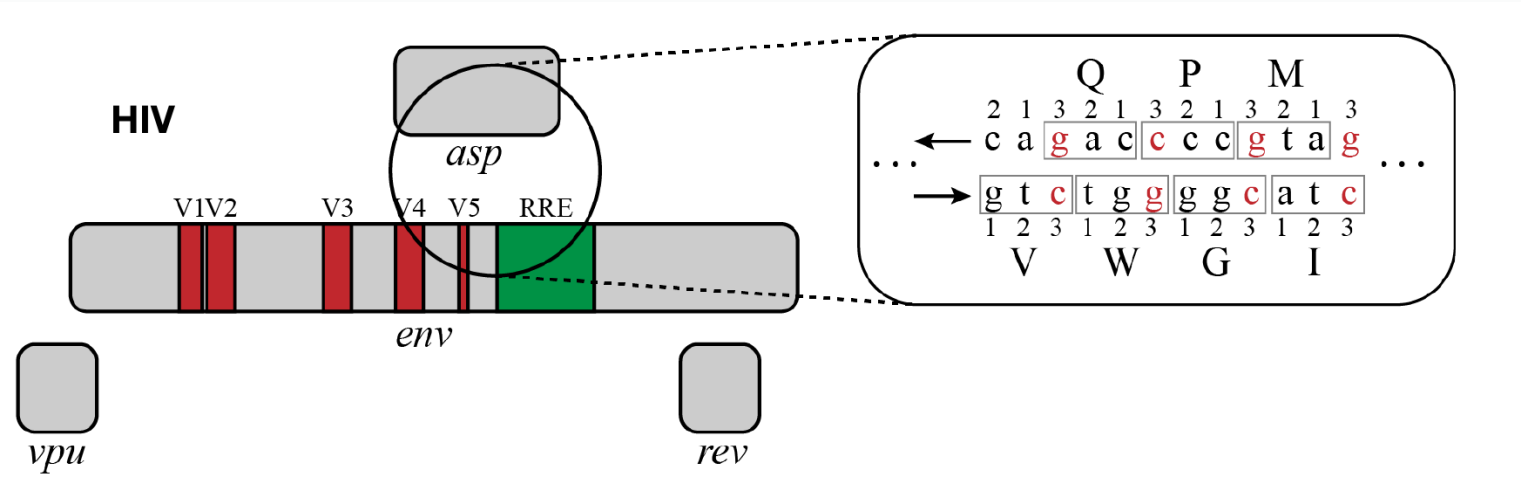
- 1 quartet of 4 positions considered
- 661 variants predicted
- H24F is predicted as a good choice for β-Val (part of HIGH)







HIV (CASSAN *et al* 2016)



VERY RARE OUTSIDE VIRUSES

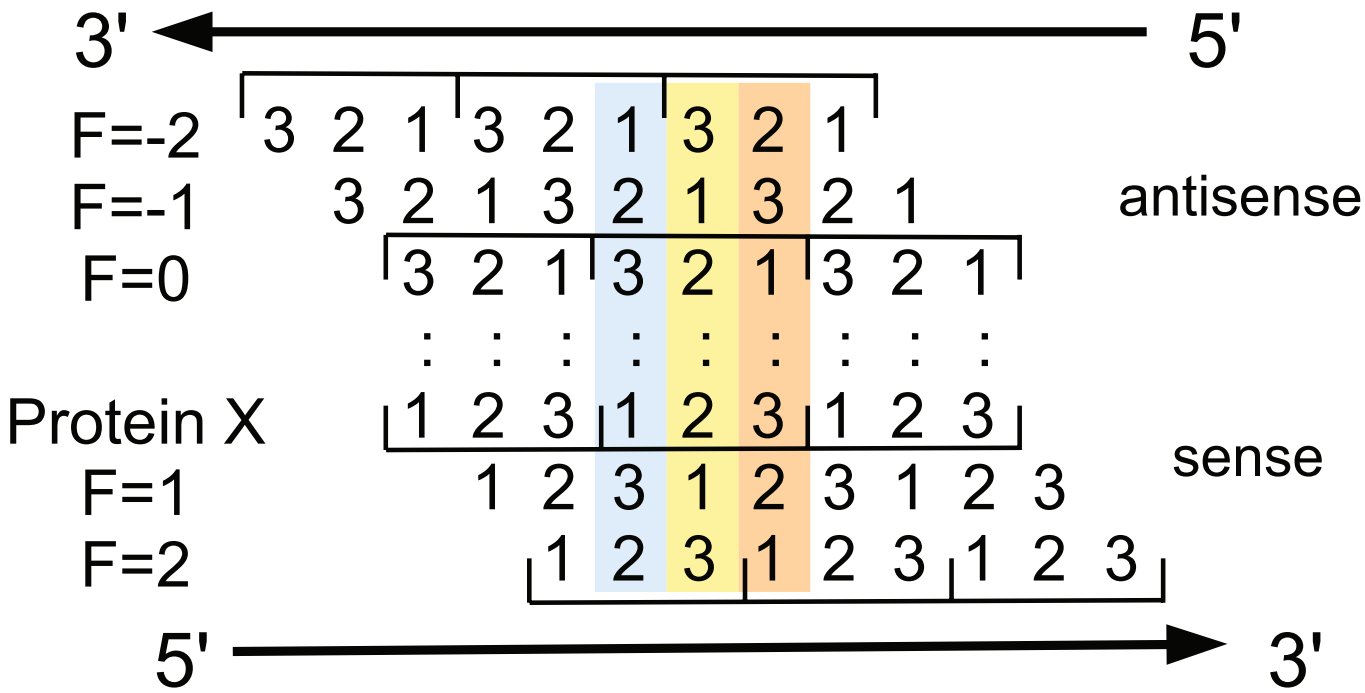
- Several in SARS-COV 2
- Cancer related genes: INK4a and ARF
- Autophagy related genes

...

WHY DESIGN OVERLAPPING GENES

- Limit genetic drift
- bio-confinement of GMOs
- Smaller genomes
- Genetic regulations
- Viral evolution

Design overlapping genes: the problem statement



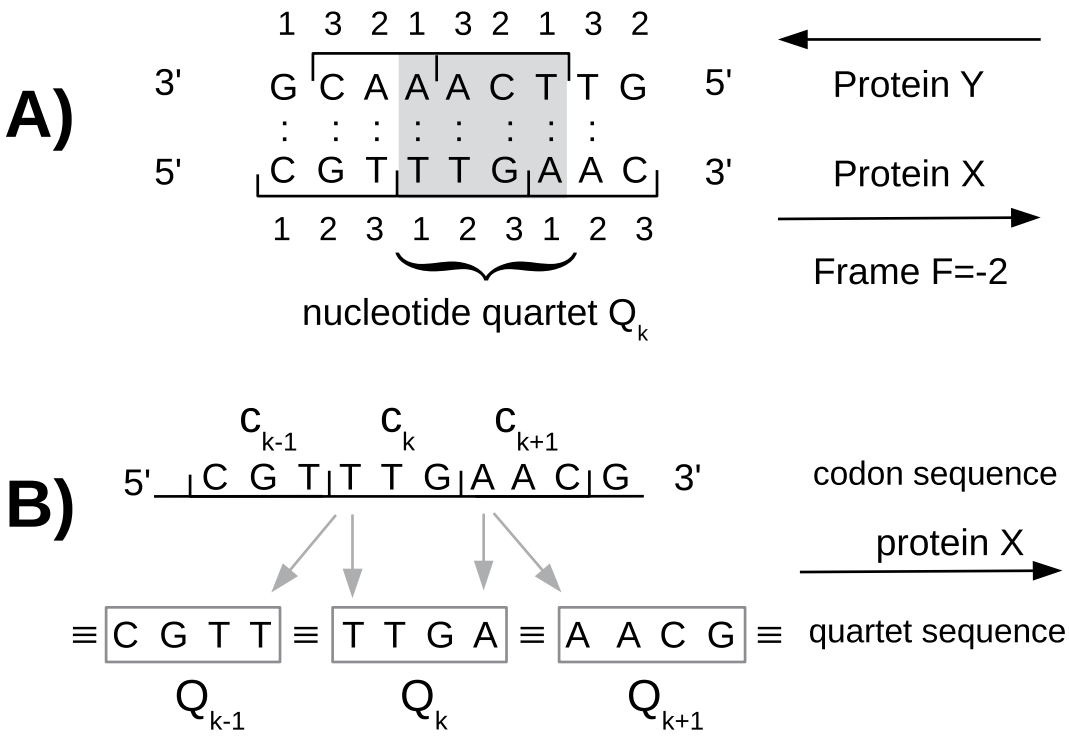
Codon overlap + base pairing $\Longrightarrow$ Strong constraints on sequences

For arbitrary pair of proteins, **overlapping coding rarely exists**!

Based on two arbitrary sequences, find the closest homologs that have an overlapping coding scheme.

EXACT SOLUTION BASED ON LINKED QUARTETS

Opuu *et al*, Sci. Rep, 2017



QUARTET DEFINITION

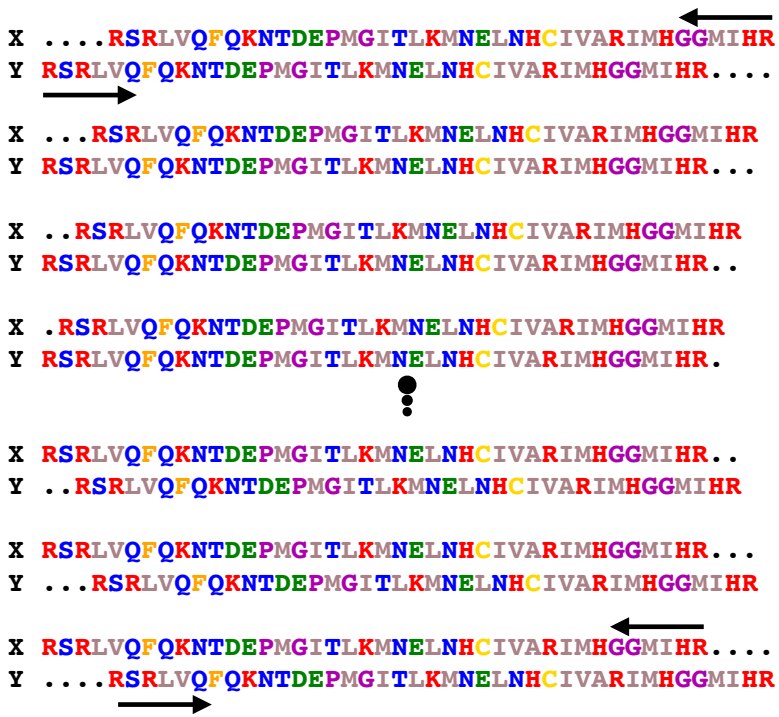
- Redefine the DNA sequence as linked list of quartets
- Similarity based score for each quartet

Reduction to the problem of finding a best path by **dynamic programming**

PDZ APPLICATION

DESIGNS

- 5 PDZ with known structures
- All 5 phases, different offsets
- 1715 designs of overlapping PDZ pairs



CHARACTERIZATION

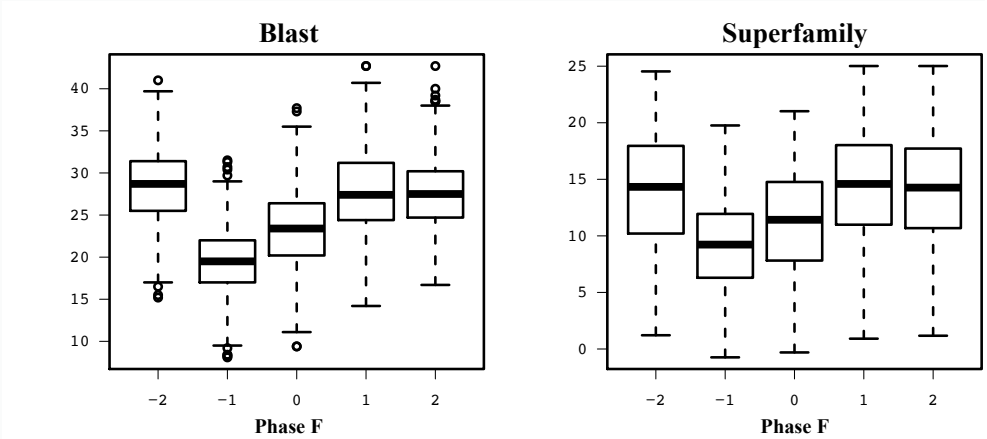
- Similarities to Pfam PDZ database
- Superfamily recognition
- Pi, Charge
- Disorder measures (Iupred)
- Cavity detection (McVol)

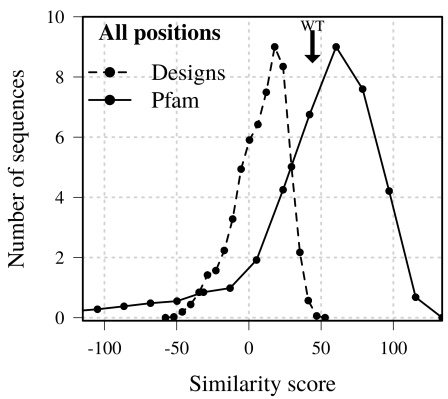
VALIDATION

- Molecular dynamics in explicit solvent 0.5-3 μ s
- Experimental tests (underway)

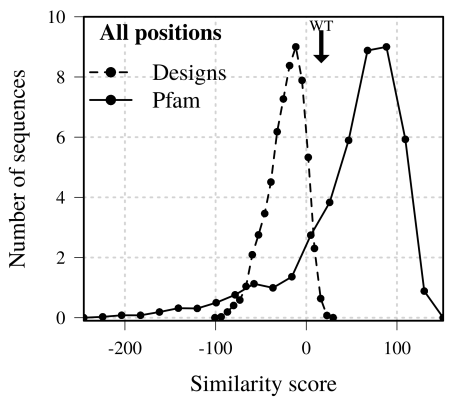
RESULTS

- Over-represented types: LEU, ARG, SER
- Under-represented types: ALA, GLY, VAL, ASP
- Phase F = -2 is the best
- Phase F = {-1, 0} are the poorest





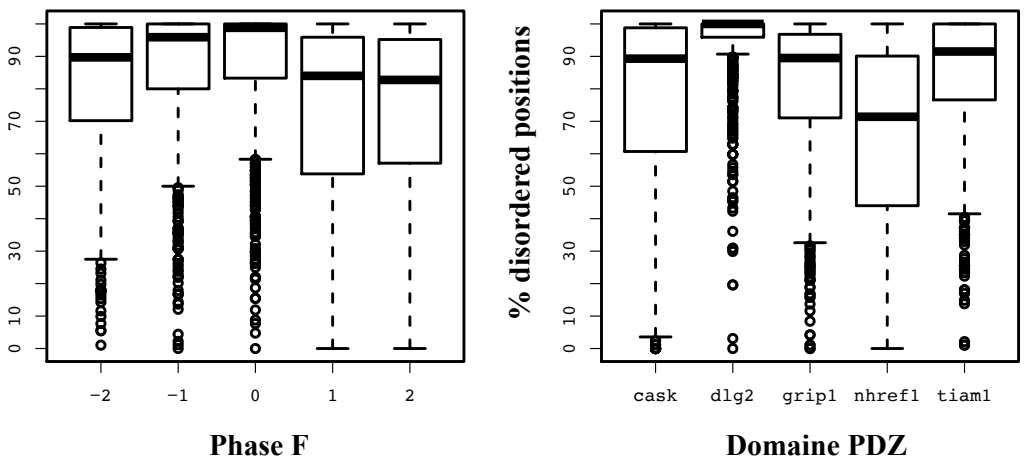
Cask



Nhrf1

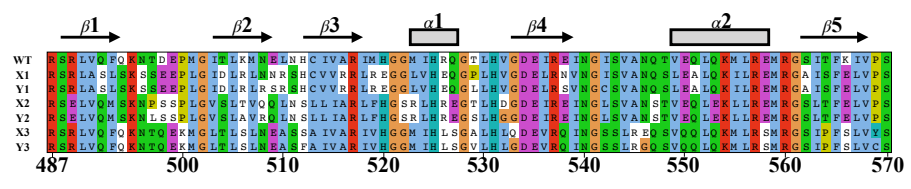
CAVITY AND DISORDER PROPENSITIES

- 3D structure reconstructed with the experimental template structures
  - 715/1715 (41.7 %) pairs have no cavities
- Disorder propensities are computed using the sequence only
  - With the threshold of 80% of non-disorder positions: 65% pairs are non-disordered



MD VALIDATION

CHOSEN SEQUENCES

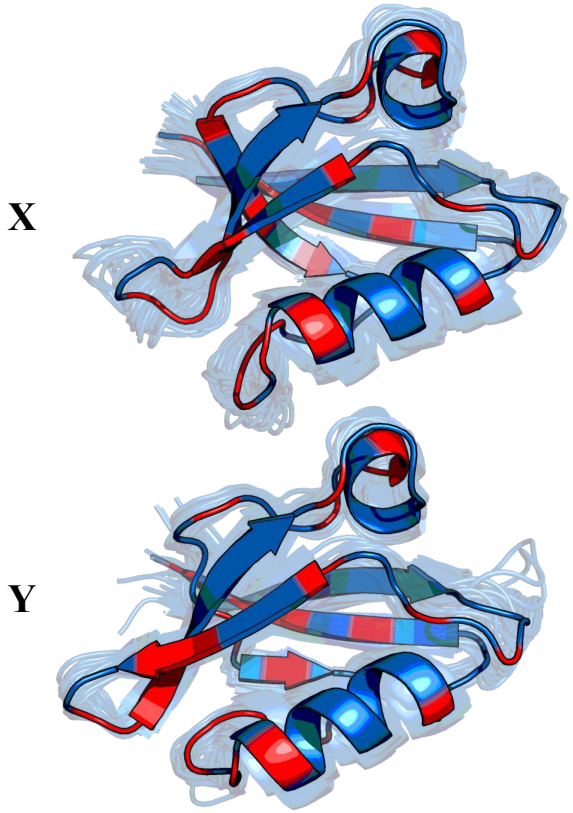


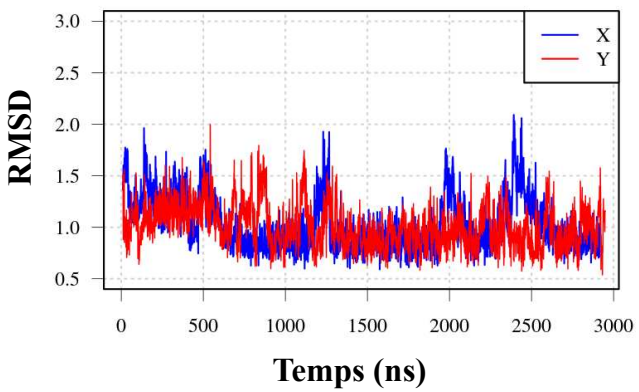
- 3 pairs of Cask sequences selected
- Phase is F=-2

MD STABILITY ASSESSMENT

- 1 pair is stable for 3 μ s (27-28 mutations)
- 1 pair stable for 0.5 μ s (34-36 mutations)
- 1 pair showed some instabilities (31-32 mutations)

Experiments underway (G. Travé)





Overall conclusions

Complete redesign of a PDZ domain

- 1st successful redesign of a PDZ domain with a physics-based approach

Enzyme redesign for affinity and catalytic efficiency

- Affinity based designs can discover WT ligand binders
- First example of selection explicitly for TS binding

Enzyme redesign for the activation of UAAs

- Discovery of active variants for β-Met
- 3 variants with decreased α-Met selectivity
- Systematic scan of active site (experiments underway)

Overlapping coding designs

- Sequence based design algorithm combined with post-filtering
- 1 pair of overlapping PDZ designs very stable in MD

Computational group, experimental groups

Computational group

Thomas Gaillard
David Mignon
Nicolas Panel
Francesco Villa
Alexandrine Daniel
Thomas Simonson

MetRS collaborators

Giuliano Nigro
Christine Lazennec-Schurdevin
Yves Mechulam
Emmanuelle Schmitt

Physics-based PDZ redesign

Young Joo Sun
Titus Hou
Ernesto J. Fuentes

Overlapping PDZ redesign

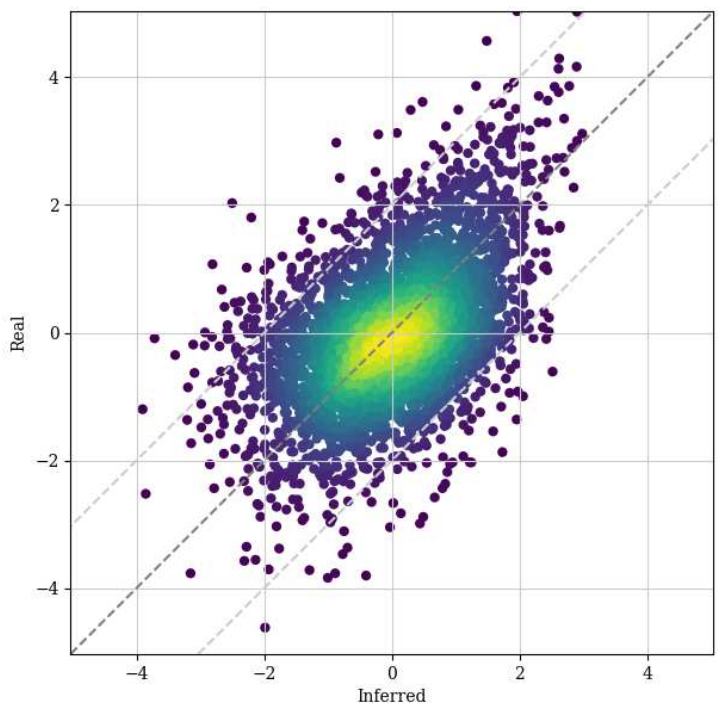
Gilles Travé

MSA BOOTSTRAP + METROPOLIS TEST

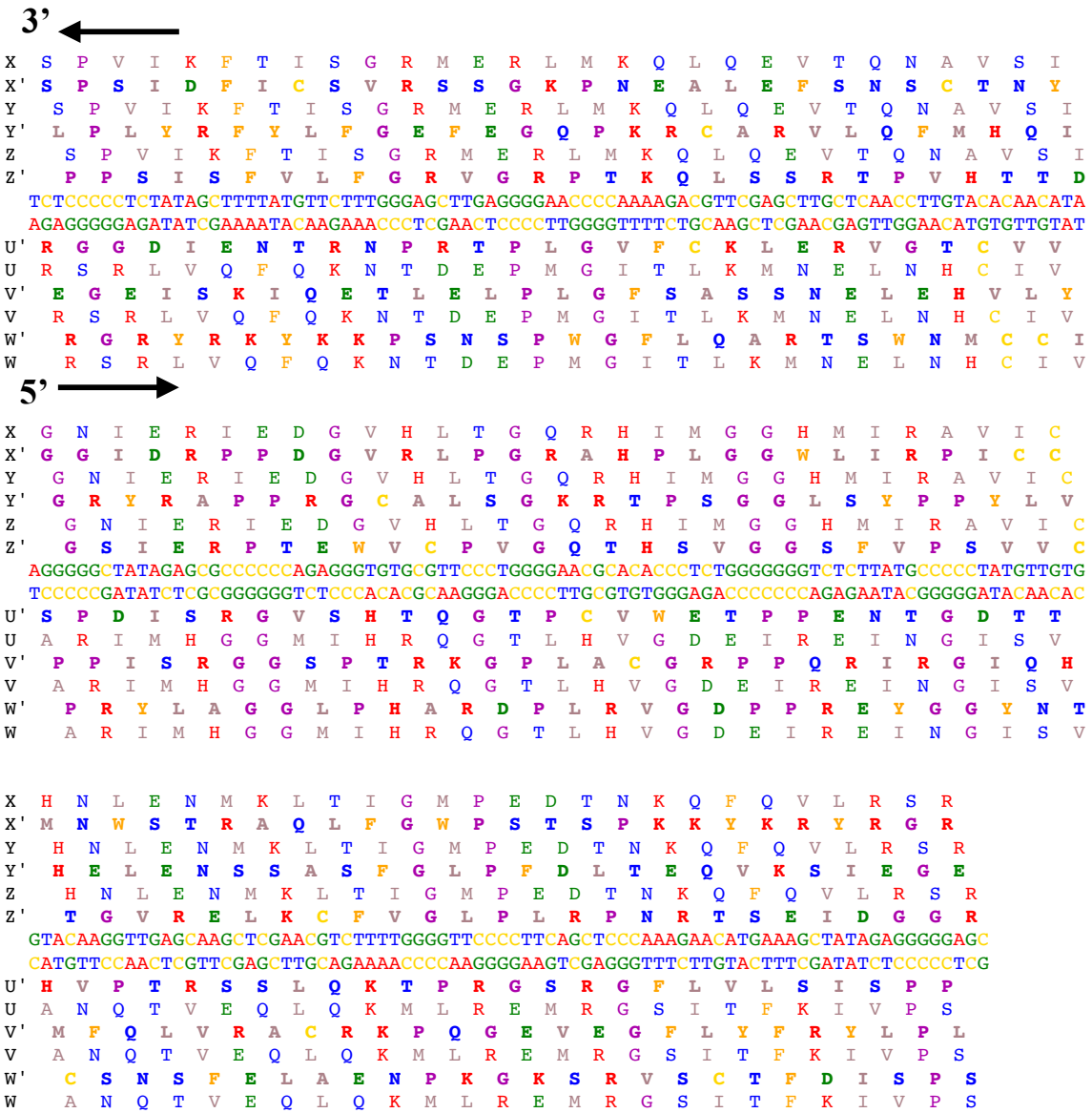
- Sequences are modelled with the Bias only:
  $E(S) = \sum_i B(S_i) + \sum_{i<j} B(S_{ij})$
- Sequences are sampled randomly from the MSA

TOY SYSTEM TESTINGS

- 5 positions: 21^5 ($10^{6.6}$) sequences
- 5 × 21 + 10 × 441 (4515) parameters
- Correlation = 0.5, mean error ~ 2 kcal/mol



Design of 6 overlapping PDZ sequences



- All **6** designs are based on Cask PDZ
- DNA sequence formulated as linked **quintuplets**
- 29.5% of identity
- 103 similarity scores