

# **Structure-based Identification of Catalytic Residues**

Chen Keasar  
BGU

# Lecture outline

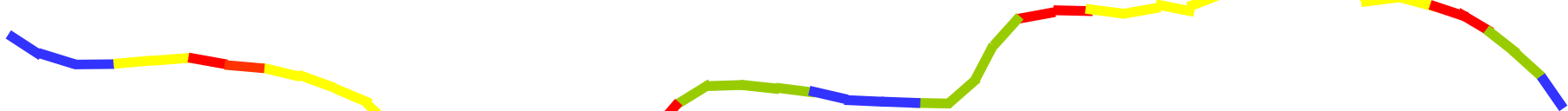
- Proteins & enzymes
- Catalytic residues
  - Who are they?
  - why is it important to identify them?
  - Why structure based identification?
- Our structural features
- SVN based prediction

Enzymes  $\subset$  Proteins

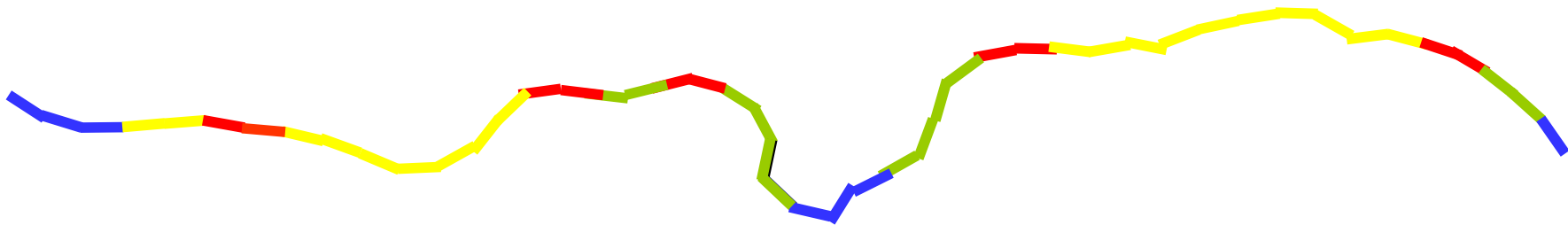
Proteins are chains.





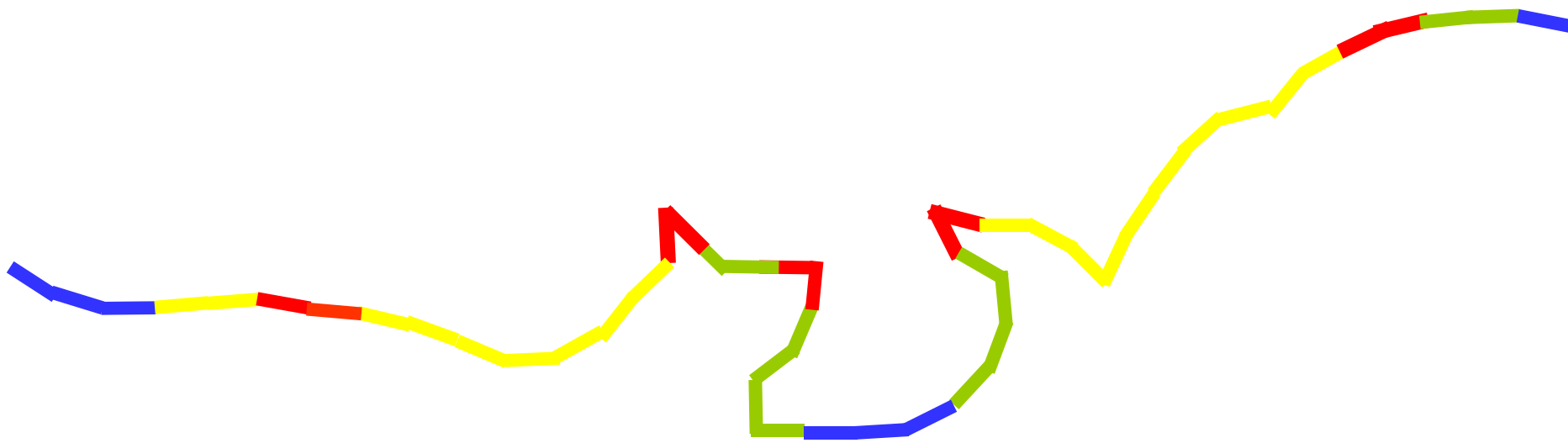


Chen Keasar, BGU

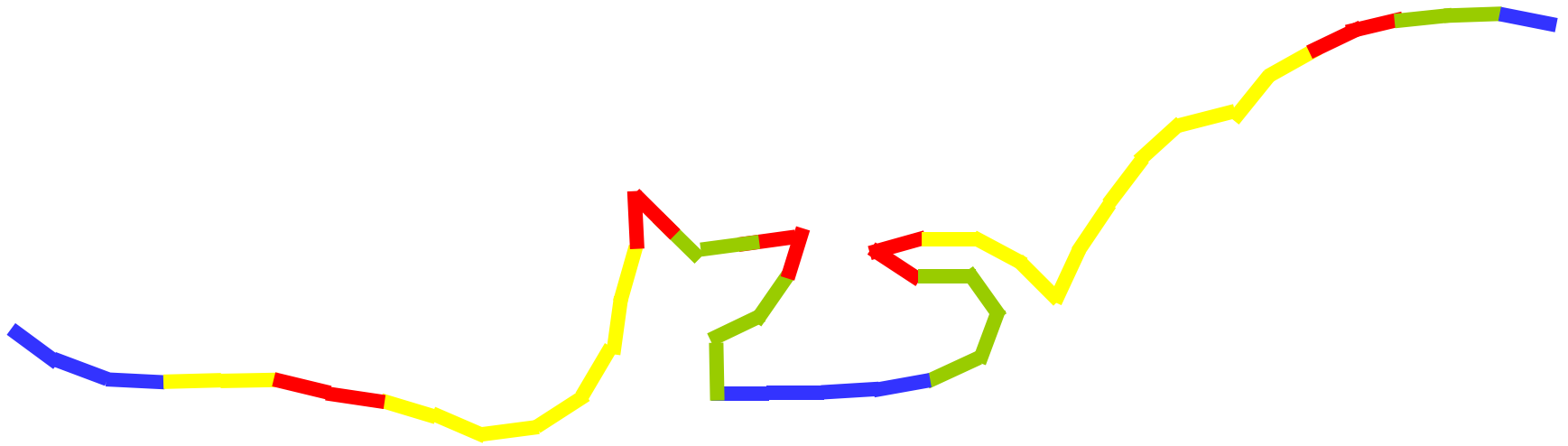


Chen Keasar, BGU

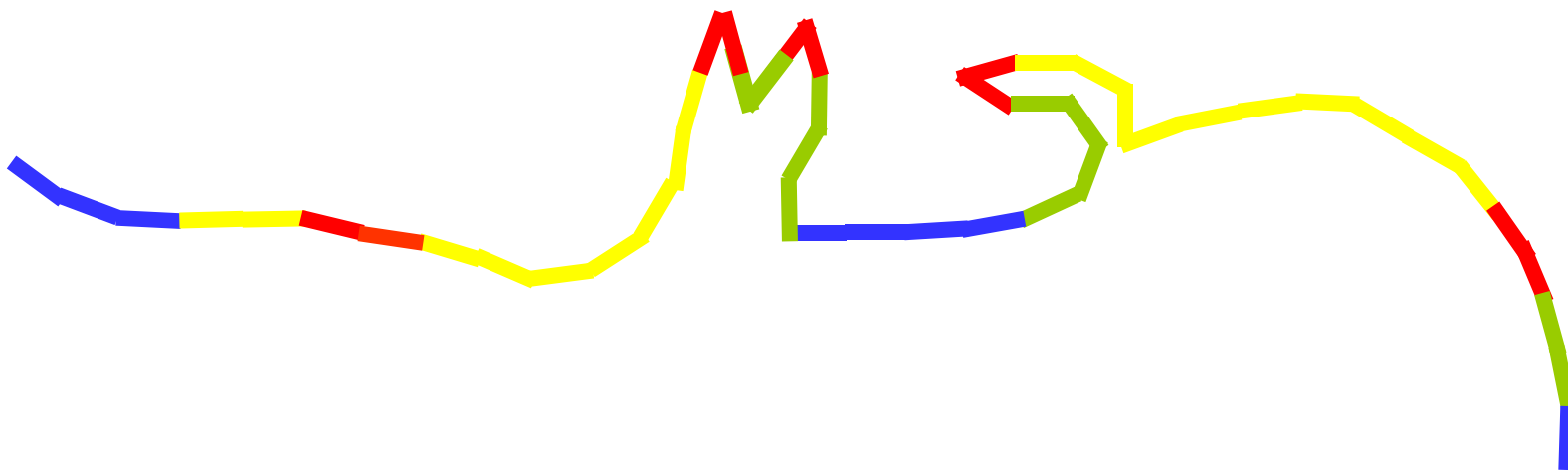




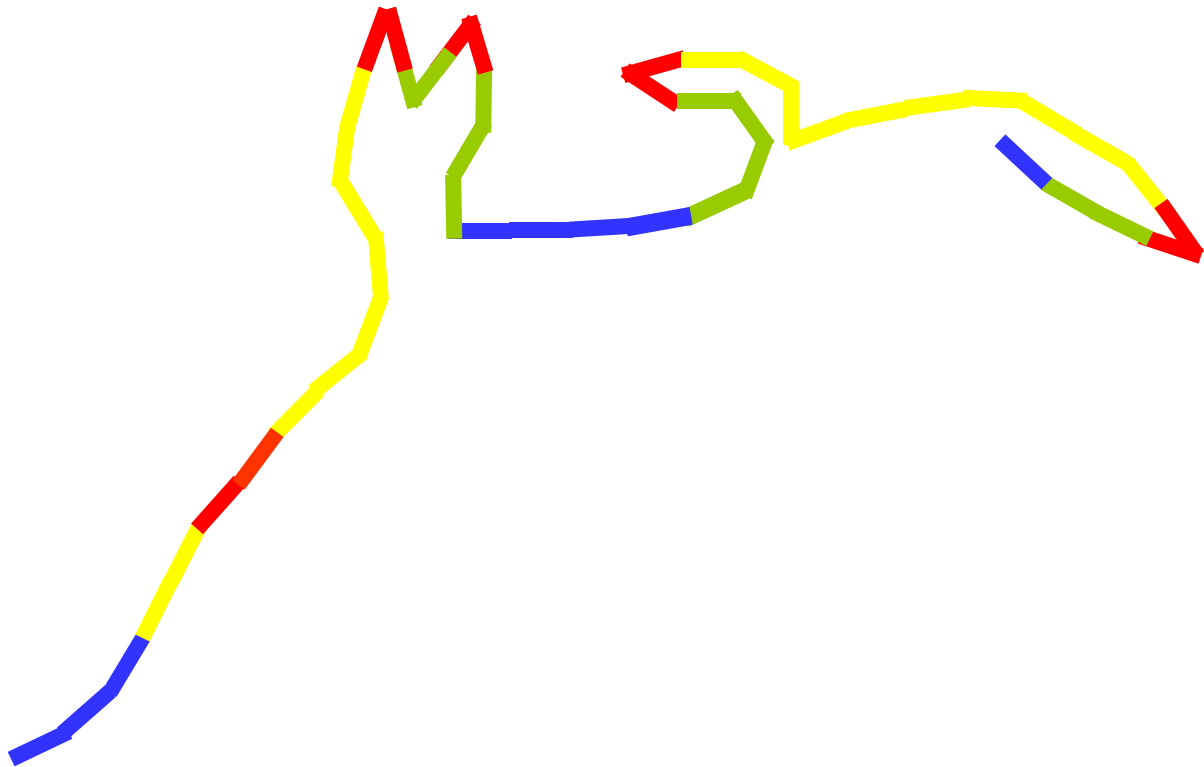
Chen Keasar, BGU



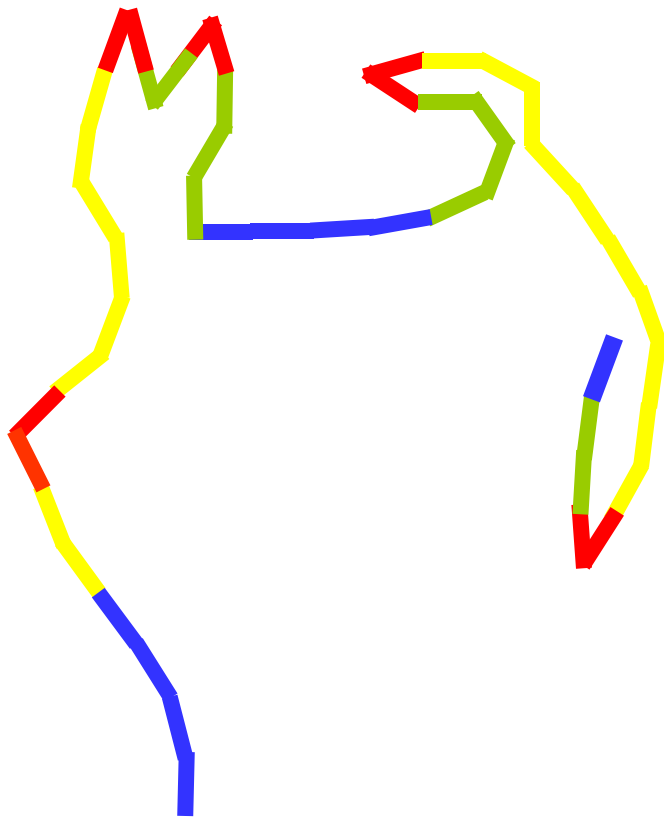
Chen Keasar, BGU



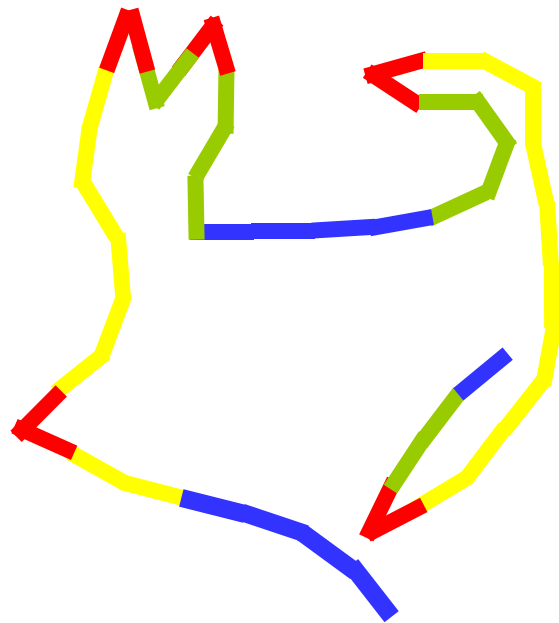
Chen Keasar, BGU



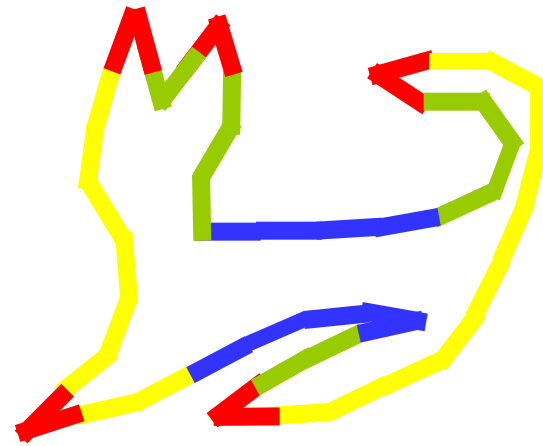
Chen Keasar, BGU



Chen Keasar, BGU

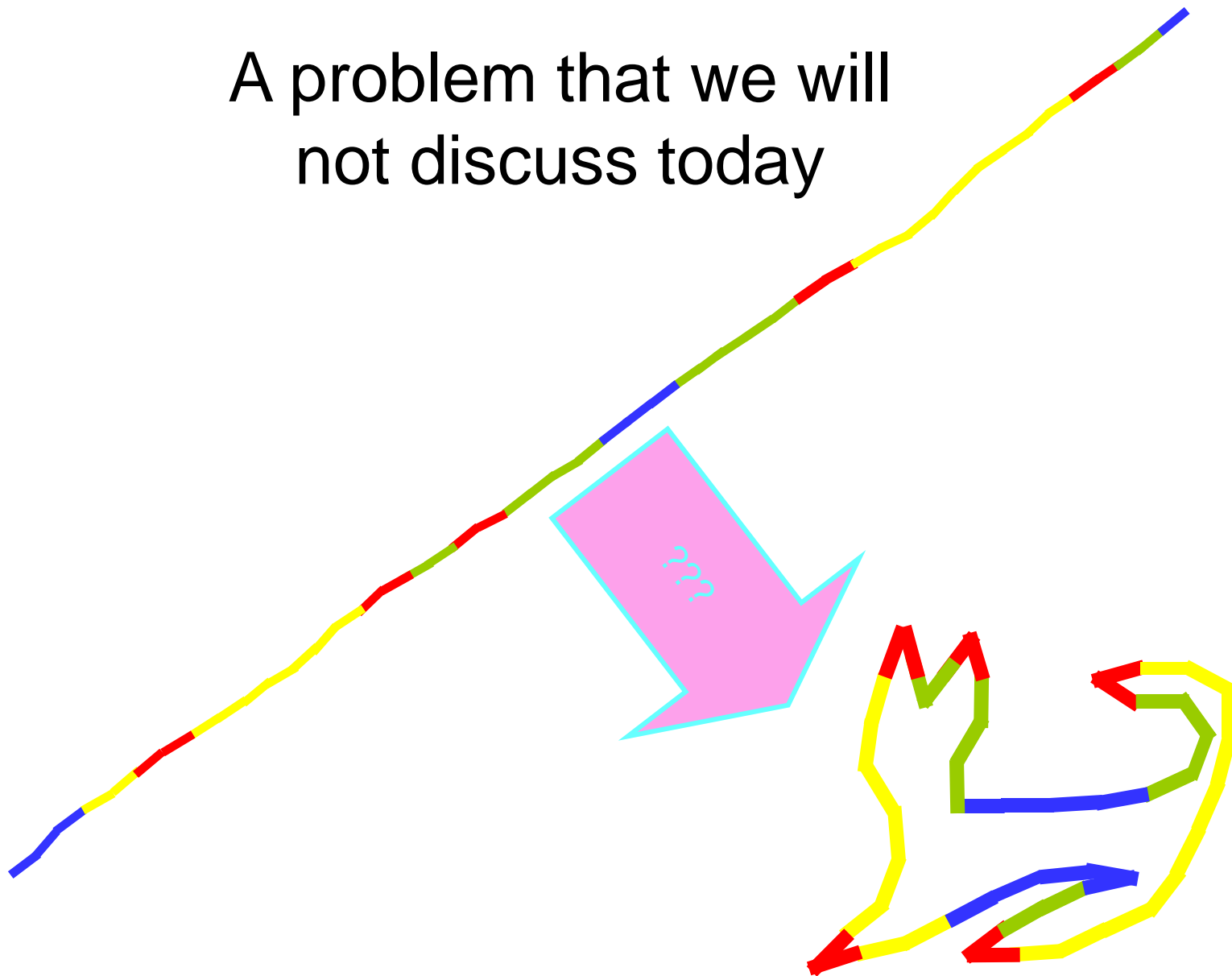


Chen Keasar, BGU

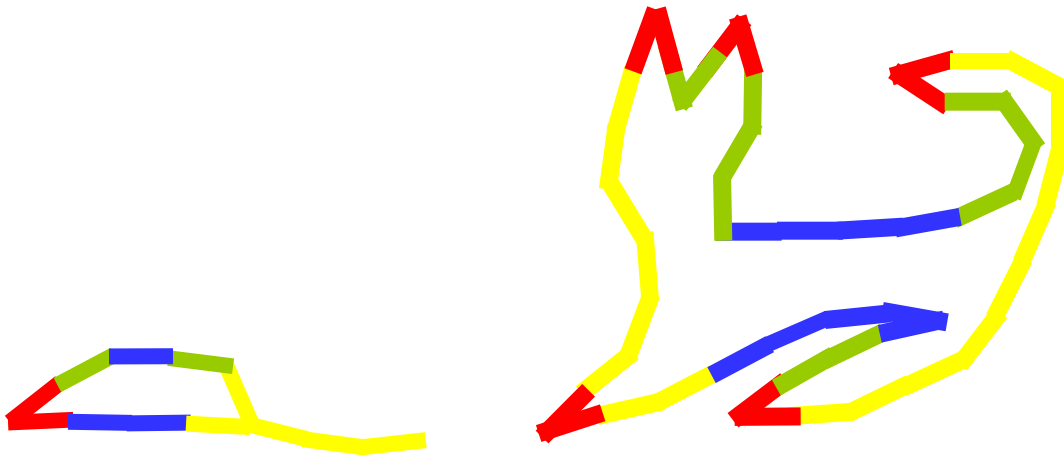


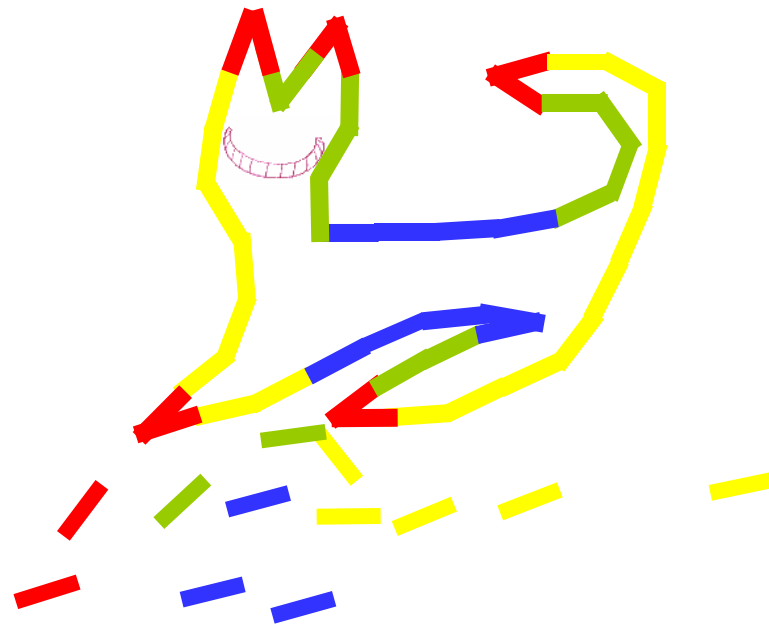
Chen Keasar, BGU

A problem that we will  
not discuss today



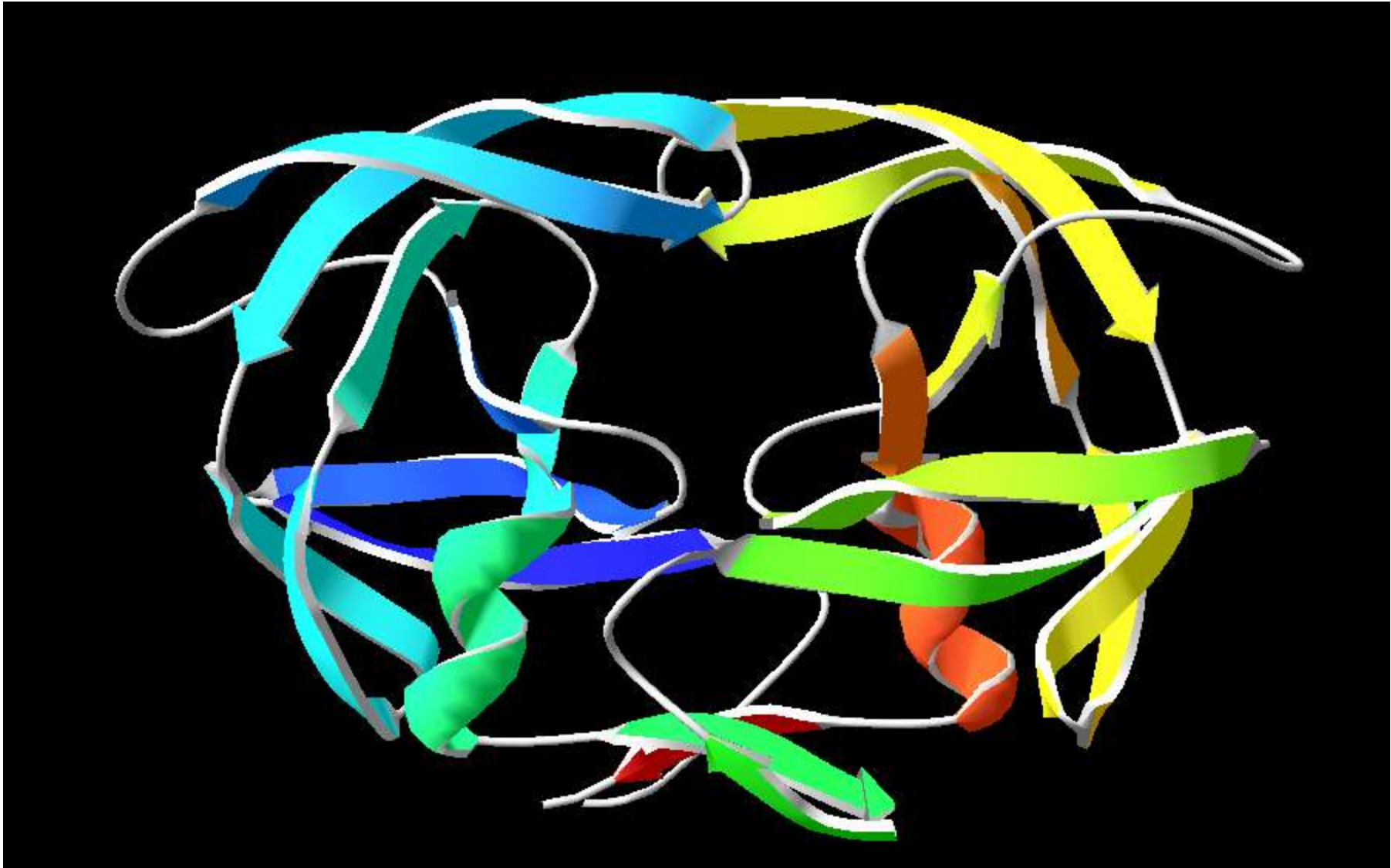
Enzymes are proteins that break and build other molecules



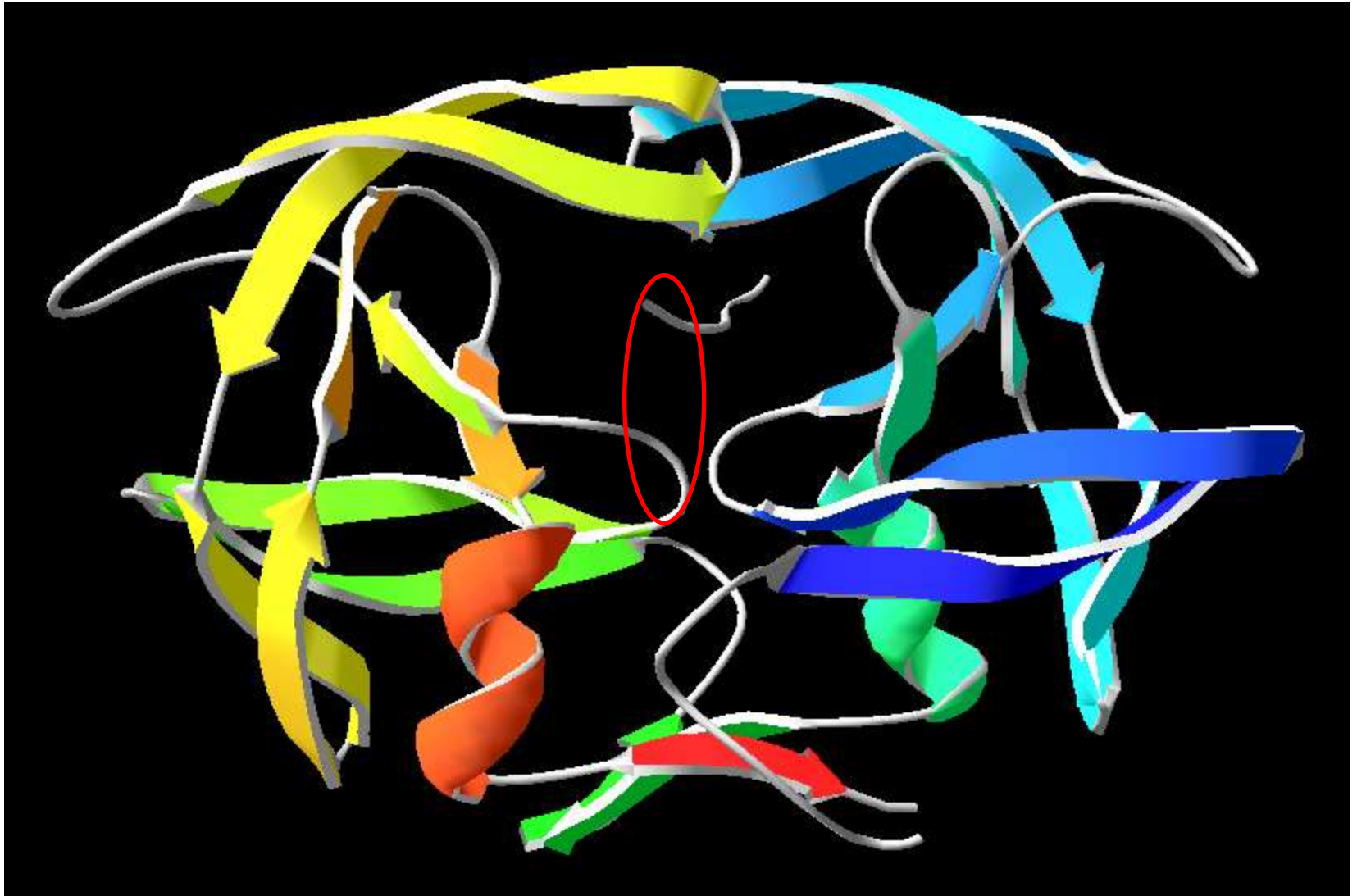


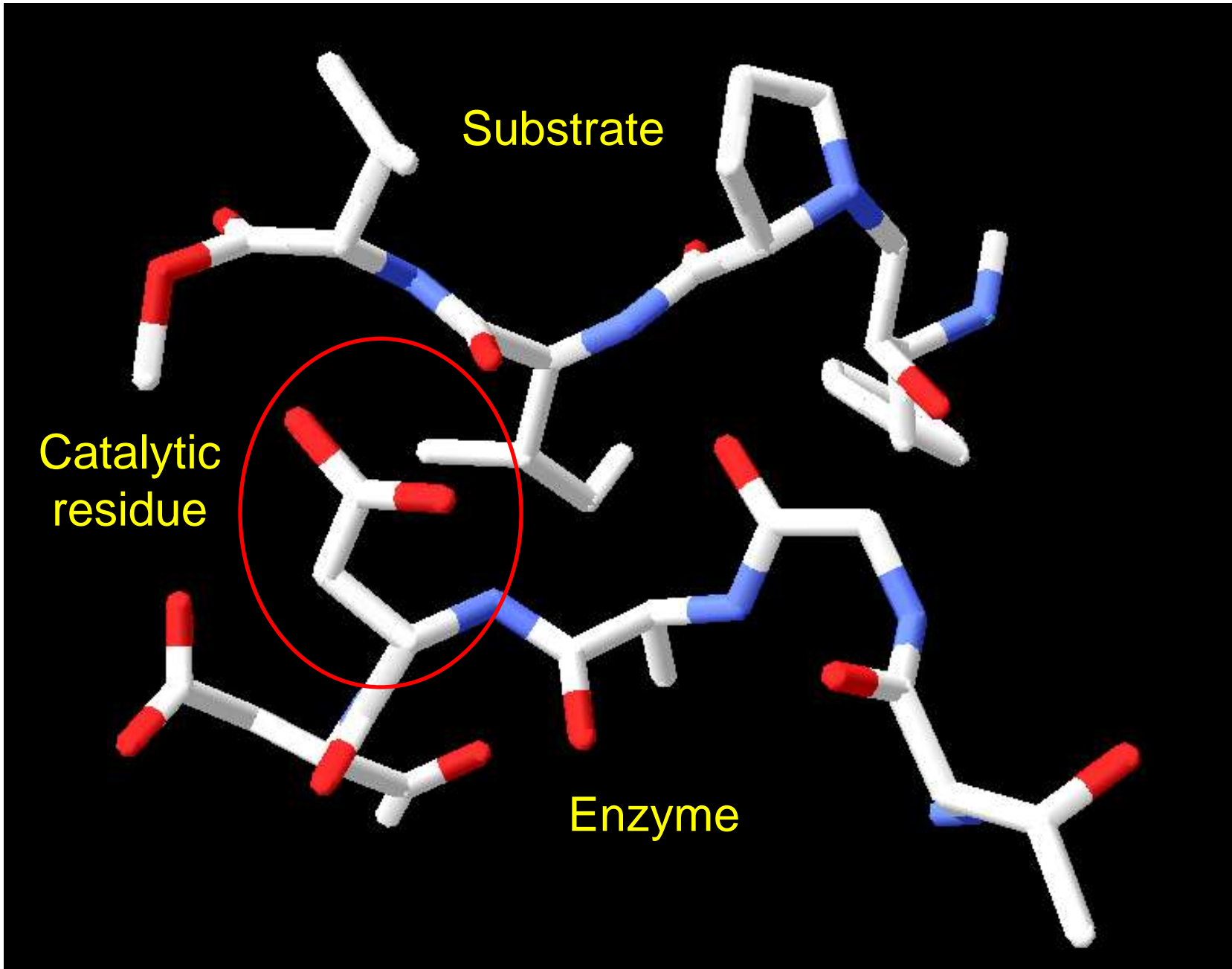
Chen Keasar, BGU

## A Real example – HIV protease

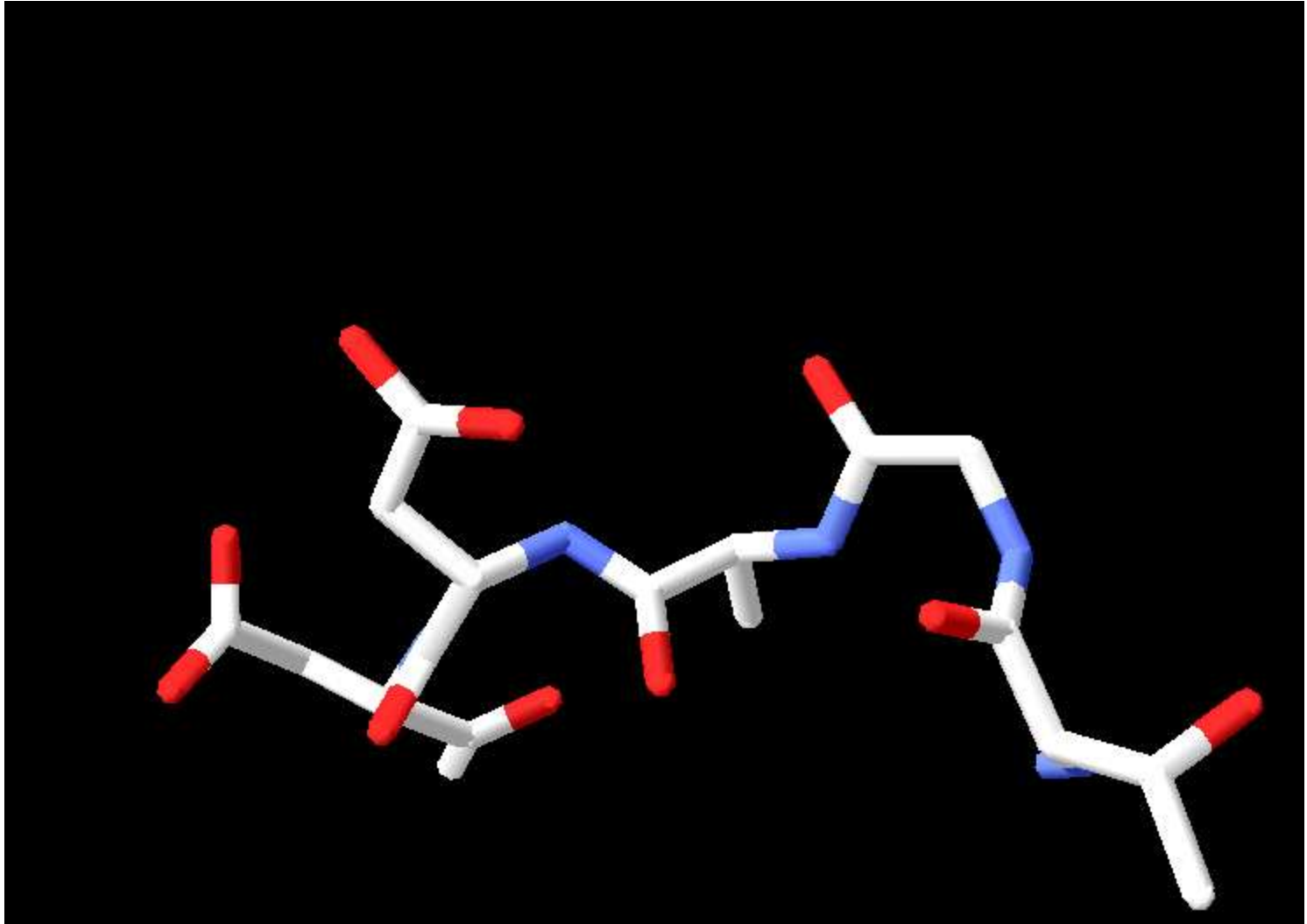


Now with a substrate

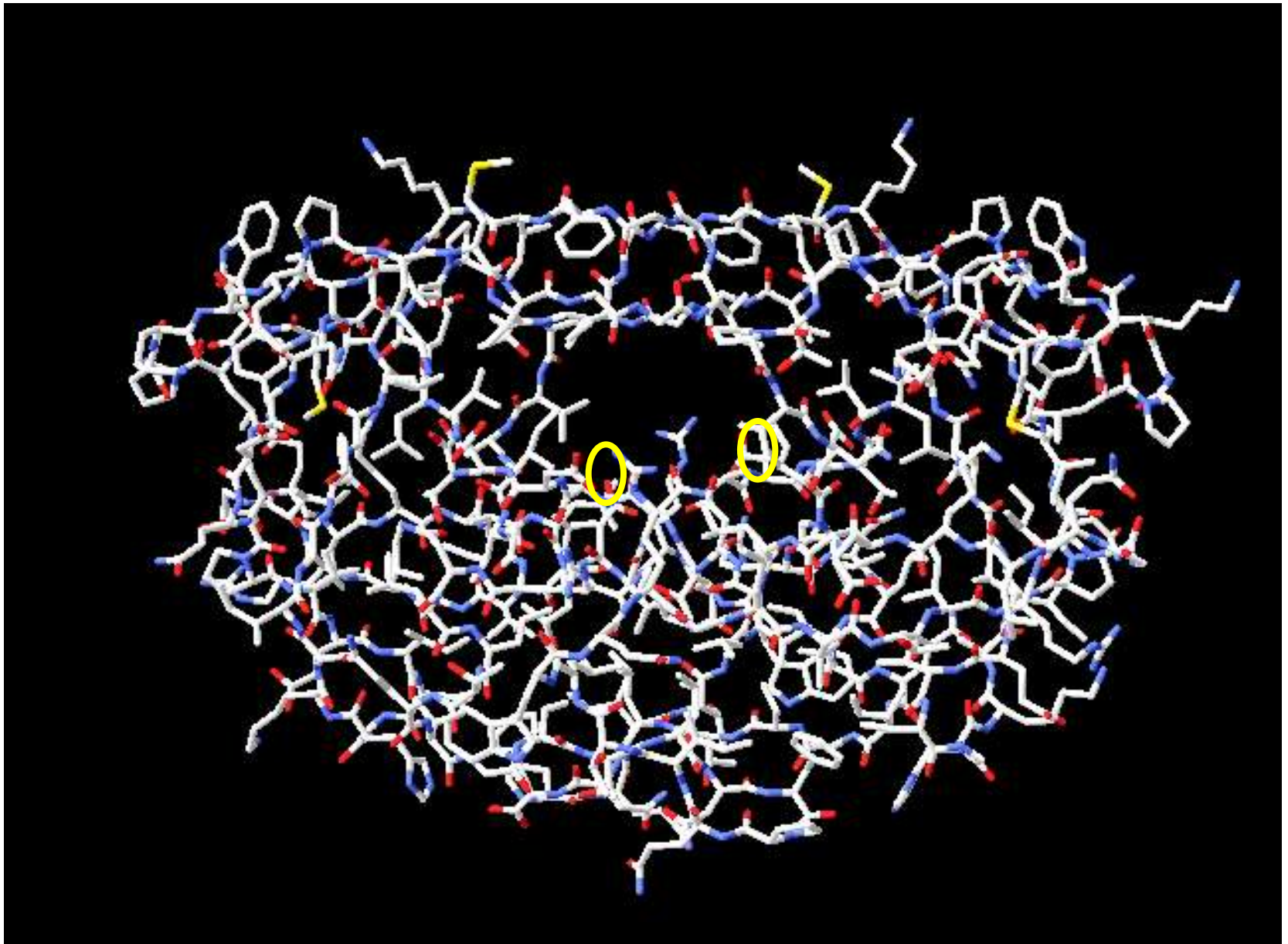




The problem – identify the catalytic residues in the absence of the substrate



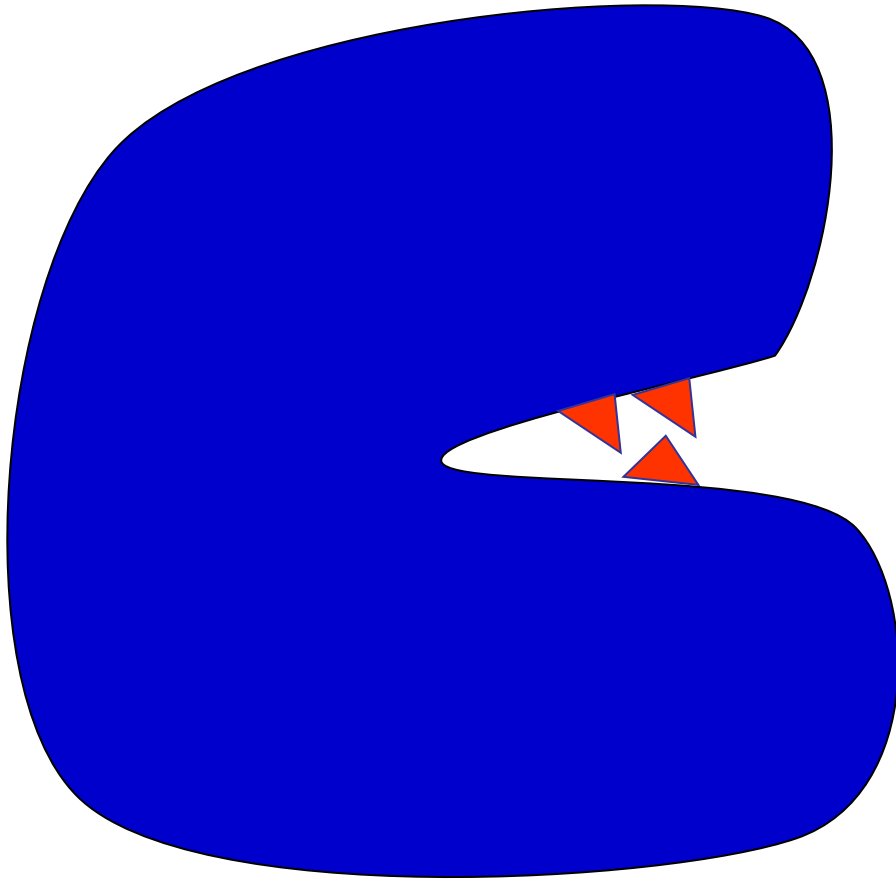
2 catalytic residues vs. 98 non-catalytic residues



## Why is it important?

1. A special case of functional residues identification
2. Interpretation of genomic data
3. Drug design

**Structural clue:**  Catalytic residues are strained

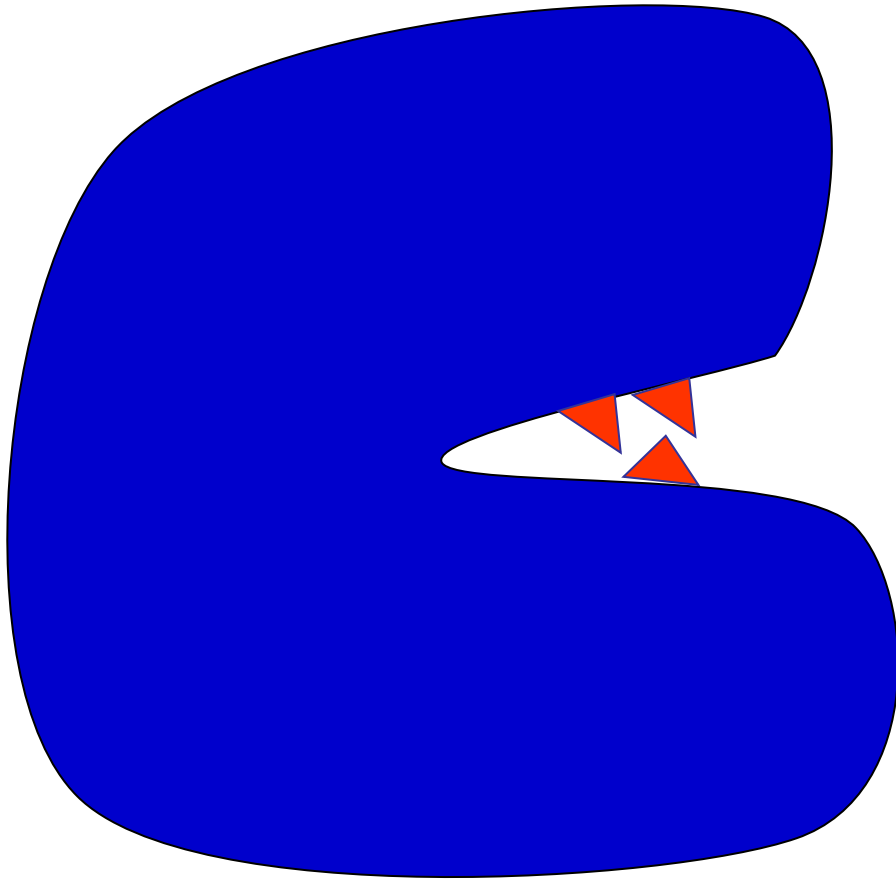


Most of the residues serve as a scaffold.

Catalytic residues have a specific mission.

- **Electrostatic strain:**
  - Warshel, 1978
  - Warshel, et al. 1988
- **Backbone strain**
  - Herzberg & Moult 1991
- **Side-chain strain**
  - Heringa & Argos 1999


Structural clue:  Catalytic residues are strained



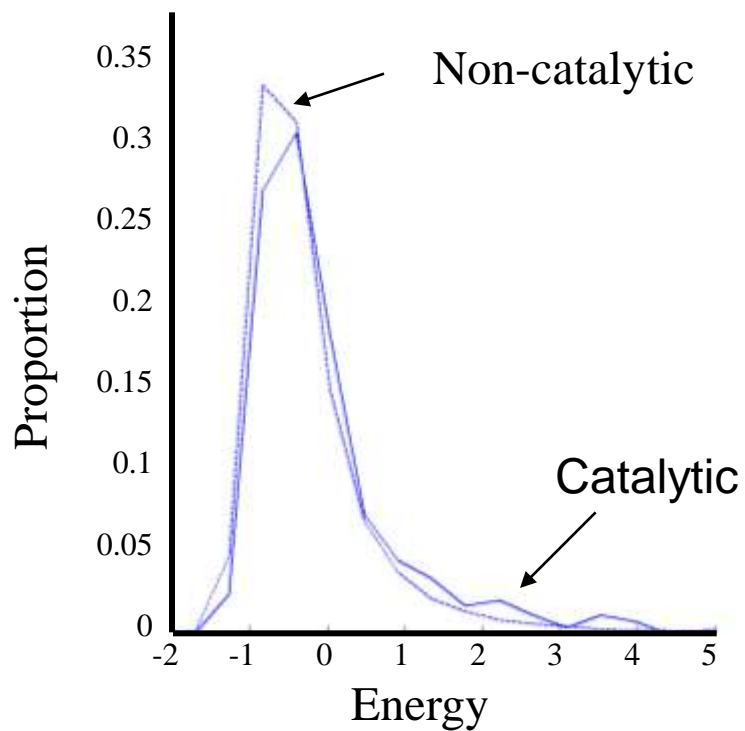
This property has already been used for prediction

- **Conformational strain**
  - Petock *et al.*, 2003
- **Electrostatic strain**
  - Bate and Warwicker, 2004
  - Elcock, 2001
  - Ondrechen *et al.*, 2001
- **Combination of energy terms**
  - Dessailly *et al.*, 2007

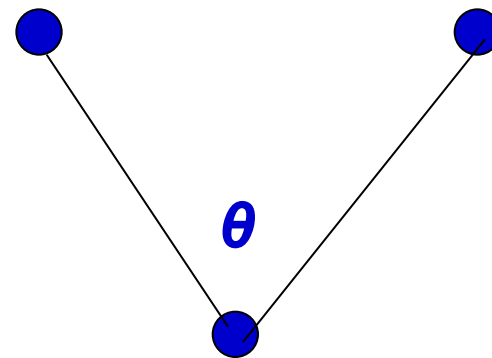
Structural clue:  Catalytic residues are strained

 Strain may be quantified by energy functions

Example: angle energy



$$E = K(\theta - \theta_0)^2$$

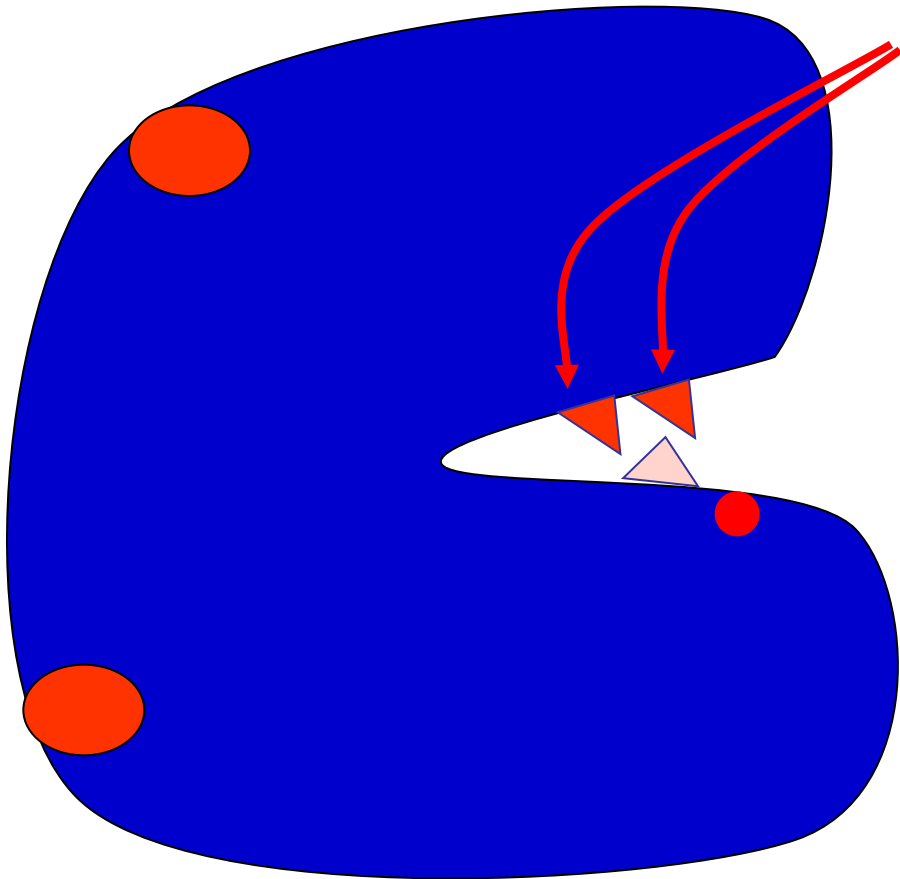


Structural clue:  Catalytic residues are strained

The good news



Indeed many “hot” residues are catalytic

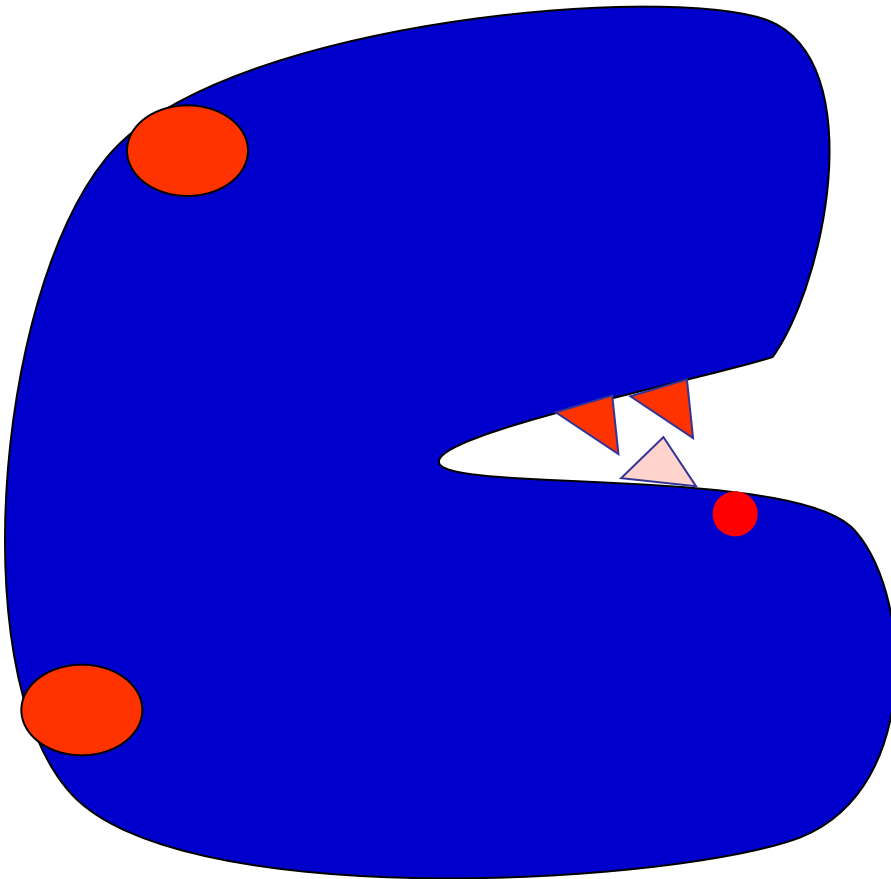


Structural clue:  Catalytic residues are strained

The good news



Most “cold” residues are non-catalytic

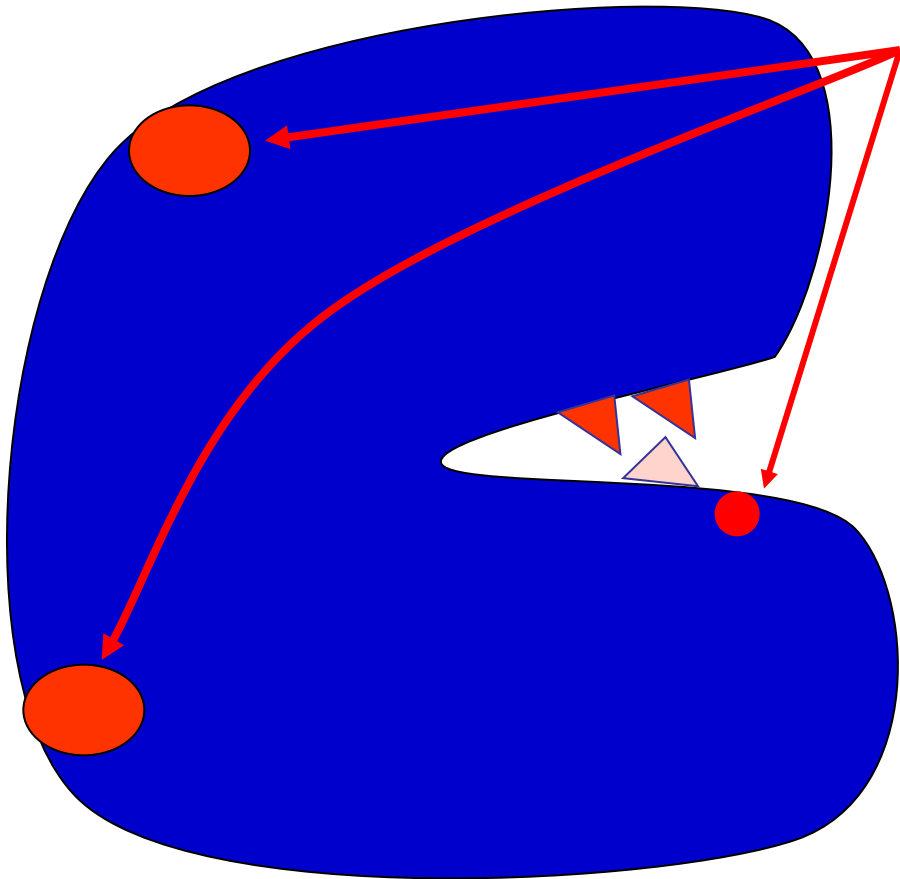


Structural clue:  Catalytic residues are strained

The bad news

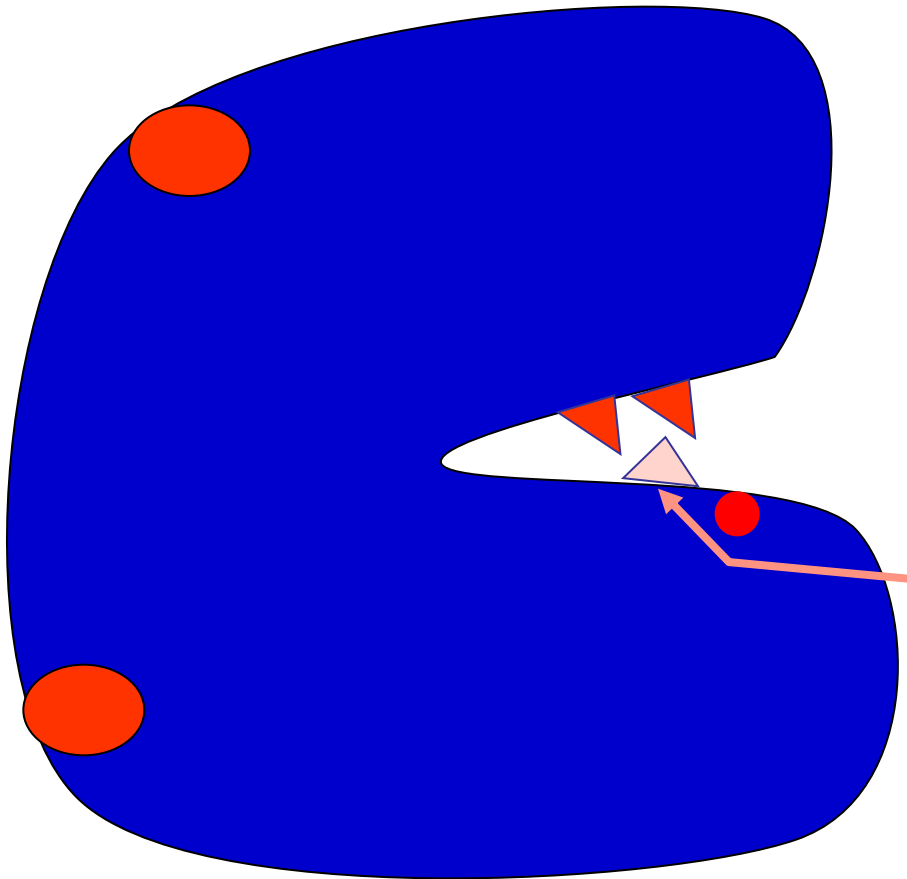


Too many non-catalytic residues are “hot”.



Structural clue:  Catalytic residues are strained

The bad news



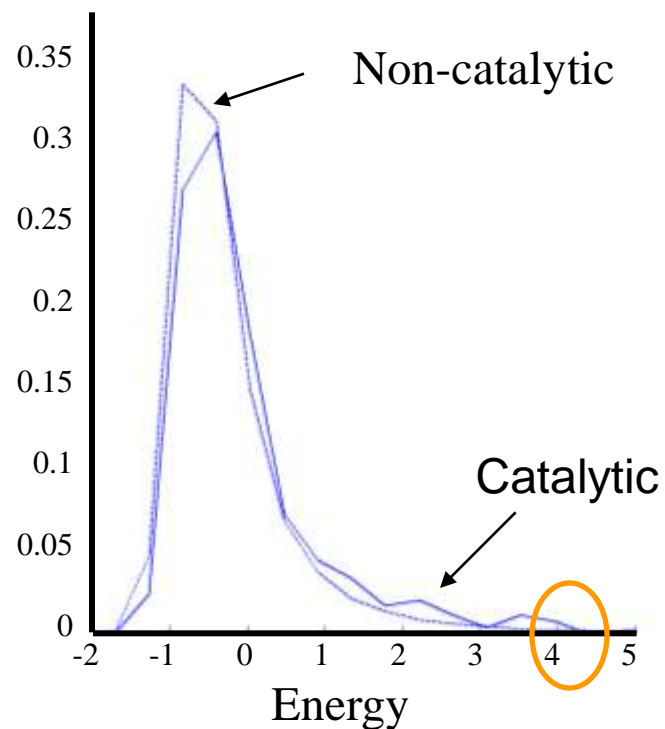
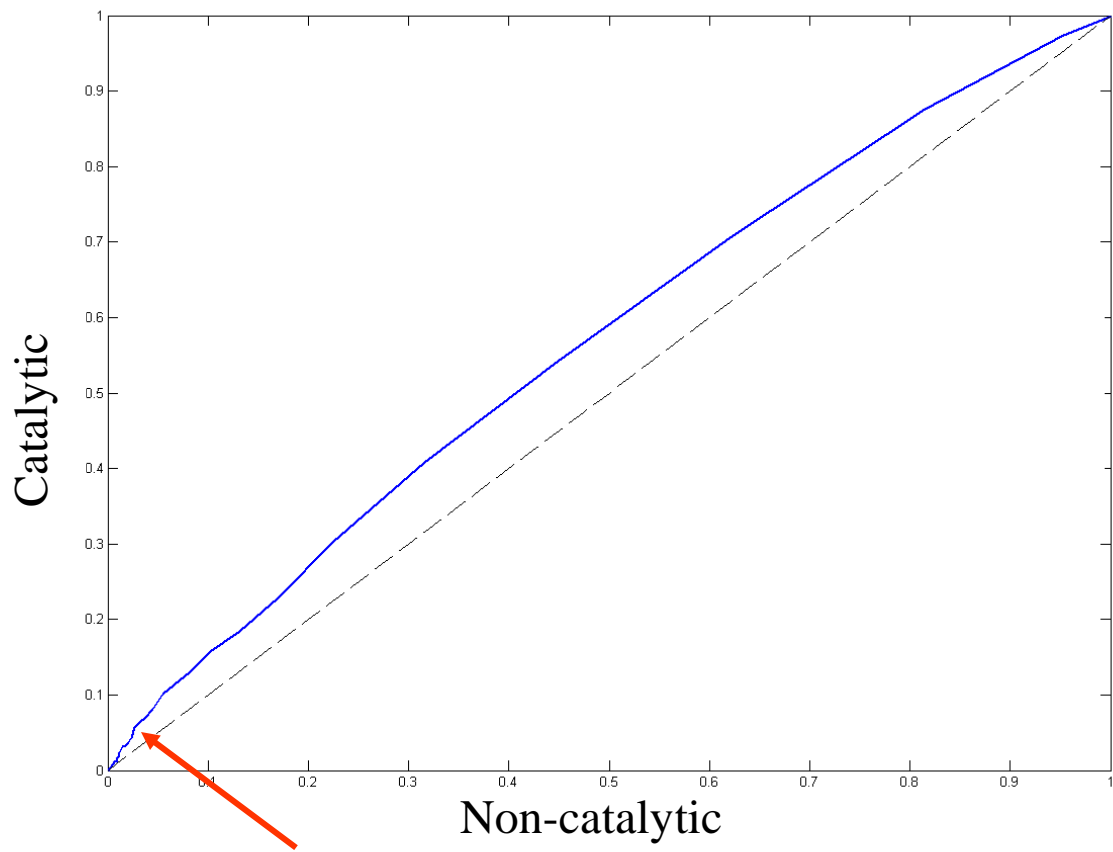
Some catalytic residues are not that “hot”.

Structural clue:  Catalytic residues are strained

The bad news



## ROC -curve

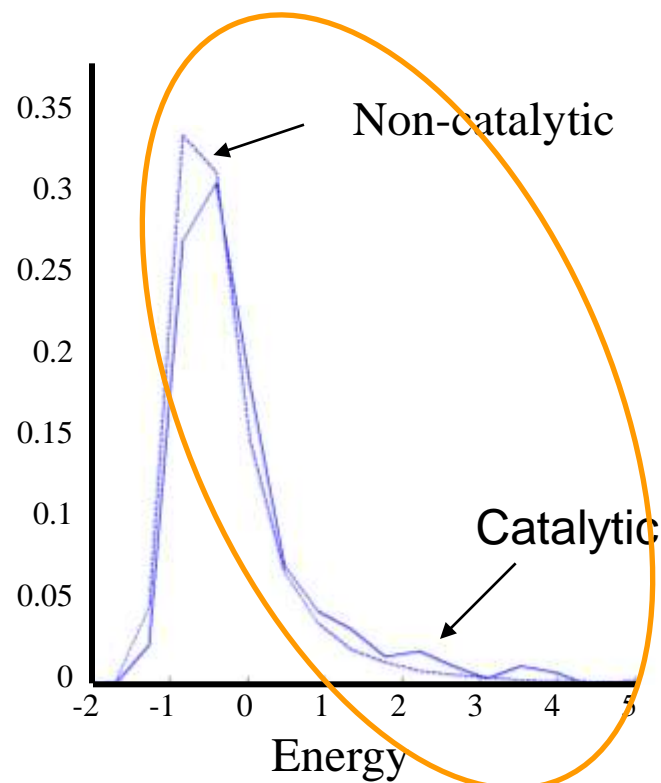
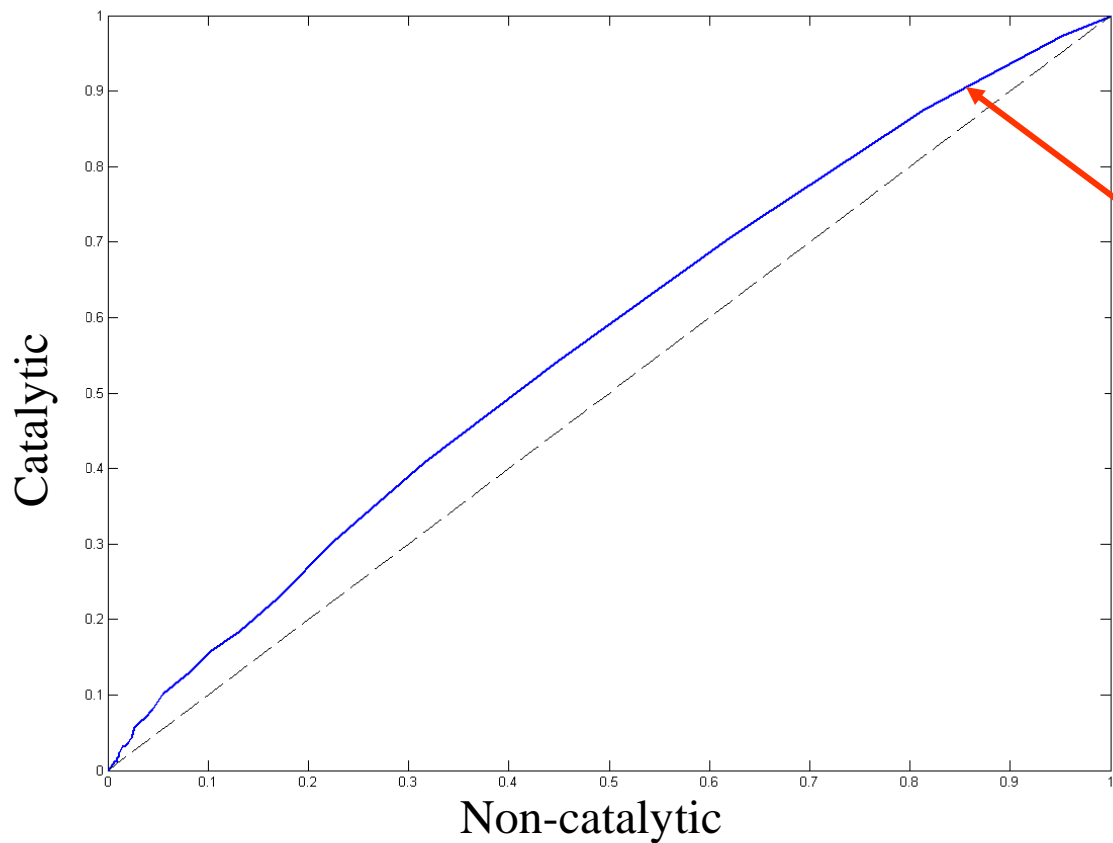


Structural clue:  Catalytic residues are strained

The bad news



ROC -curve

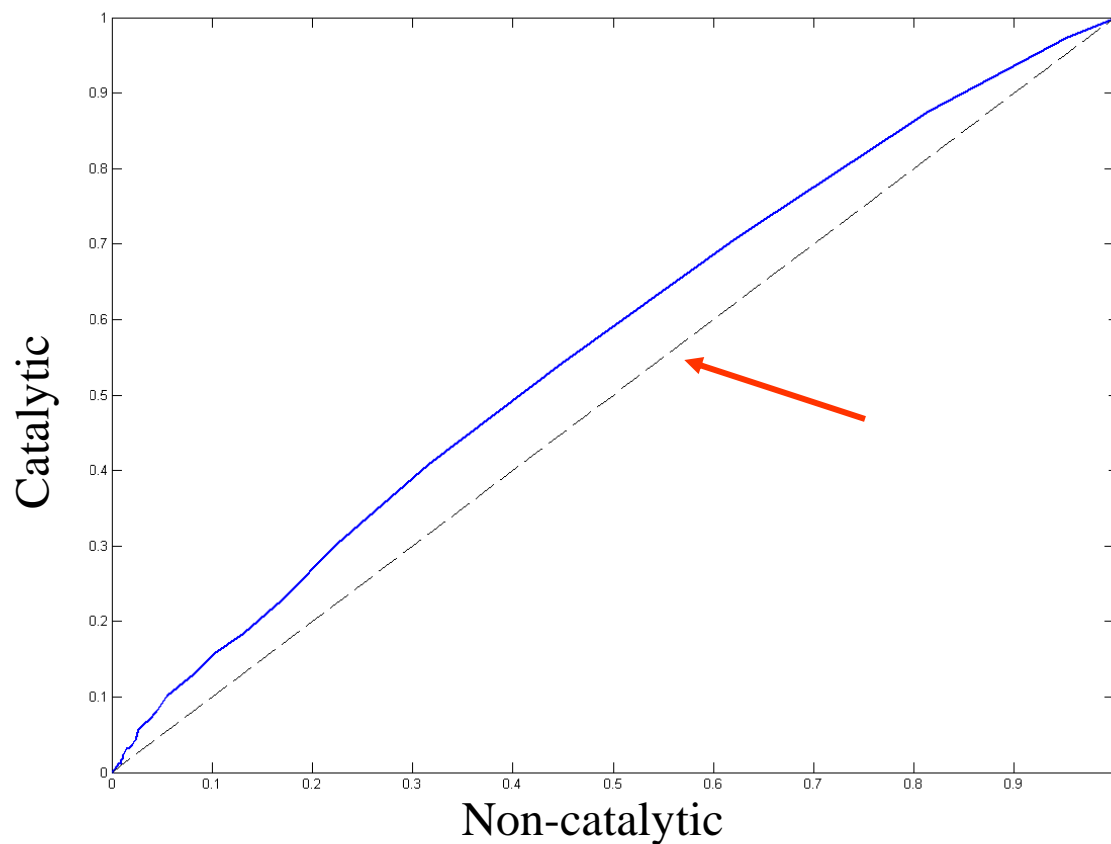


Structural clue:  Catalytic residues are strained

The bad news



ROC -curve

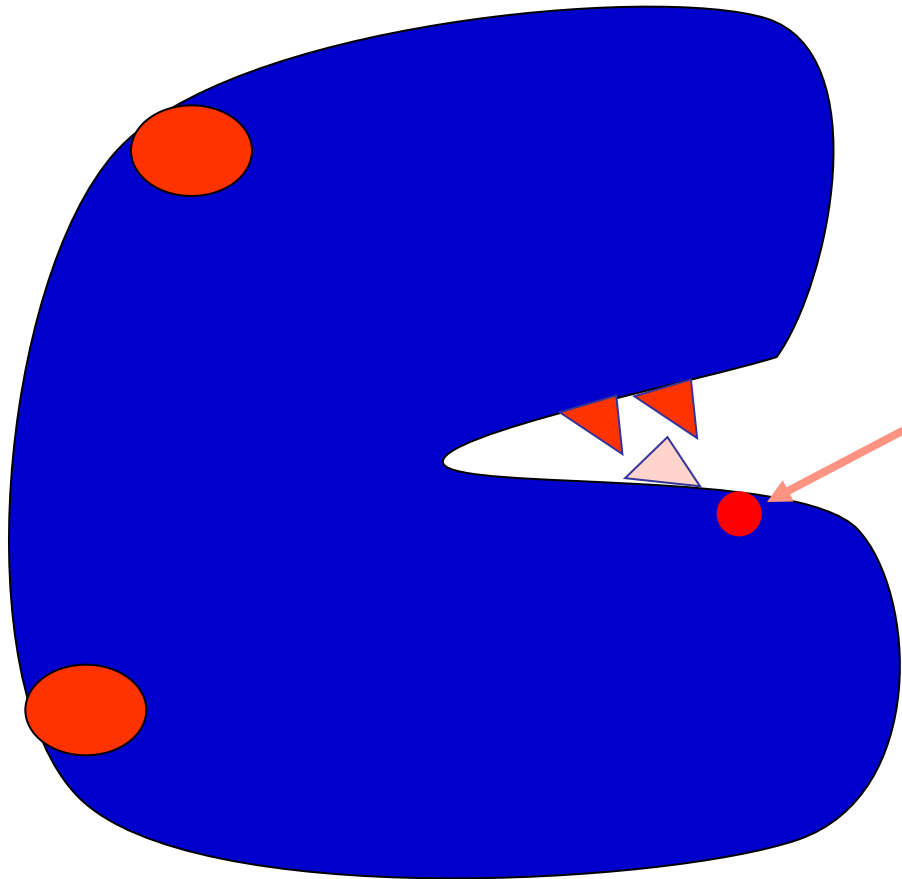





Better than random

BUT not good enough.

Structural clue:  Catalytic residues are strained

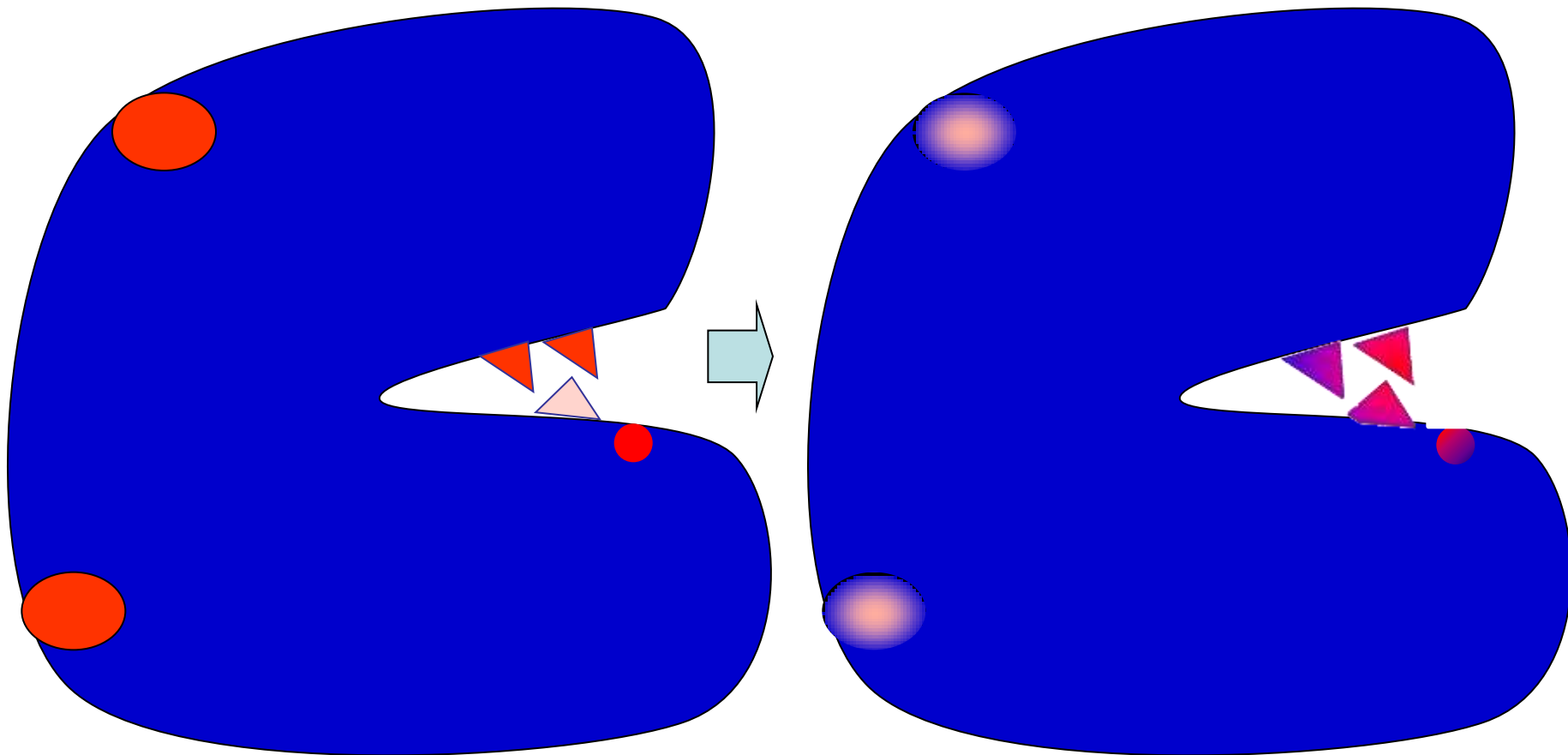
Why?



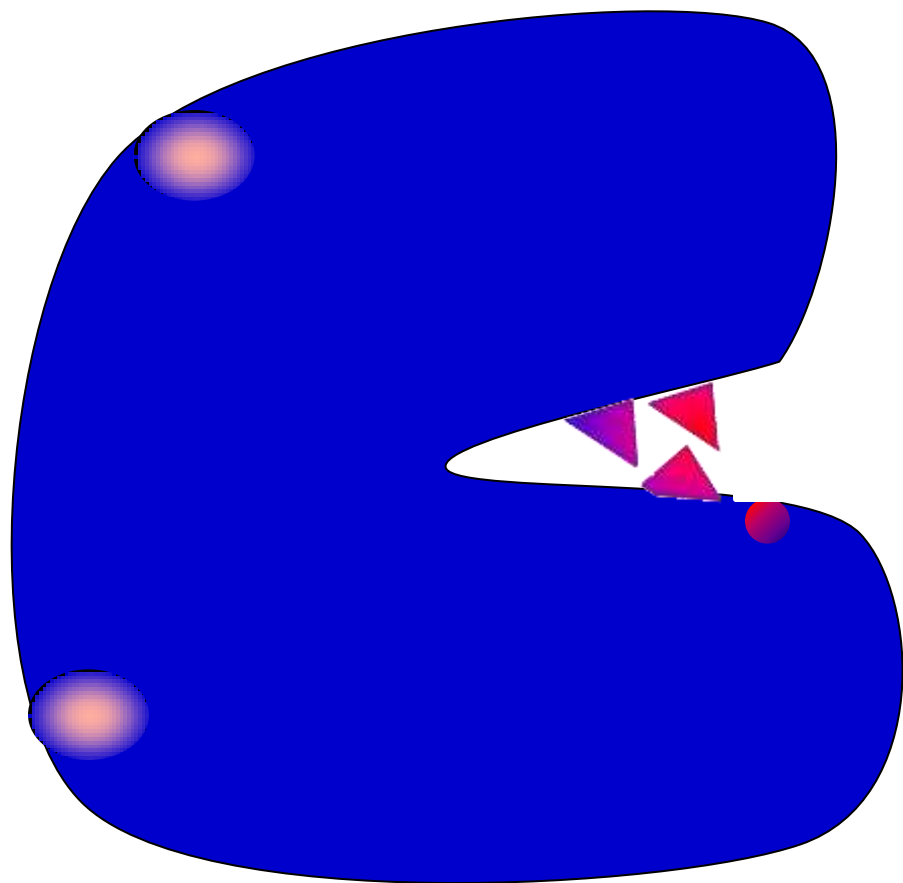
-  Our energy terms are not perfect
-  Proteins are not “designed” for optimal stability
-  Other functional residues (e.g., binding) may also be “hot”

Structural clue:

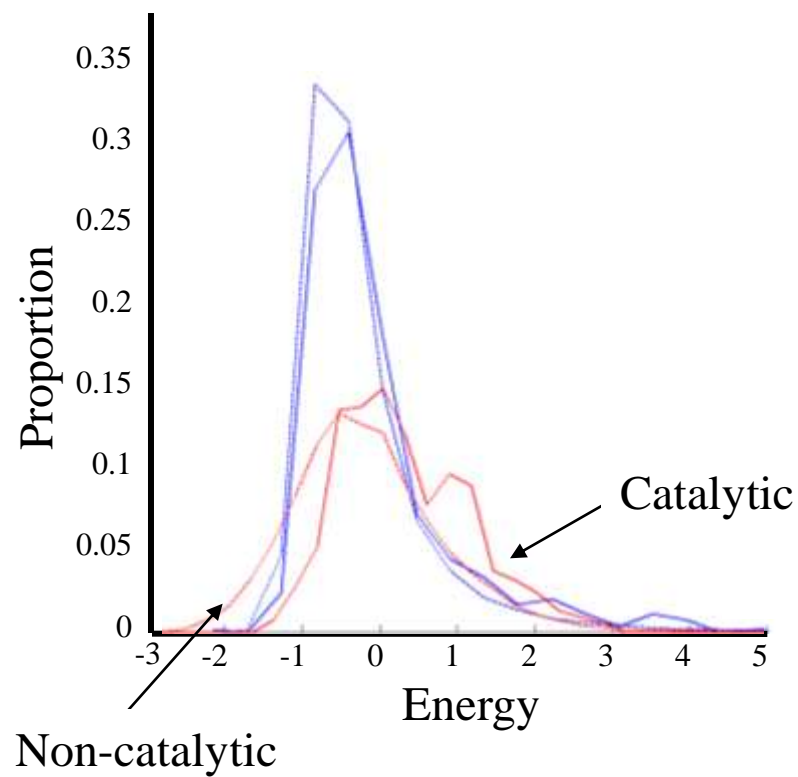
- Catalytic residues are strained
  - Catalytic residues are clustered
- => use spatial averaging



- Structural clue:
- Catalytic residues are strained
  - Spatial averaging



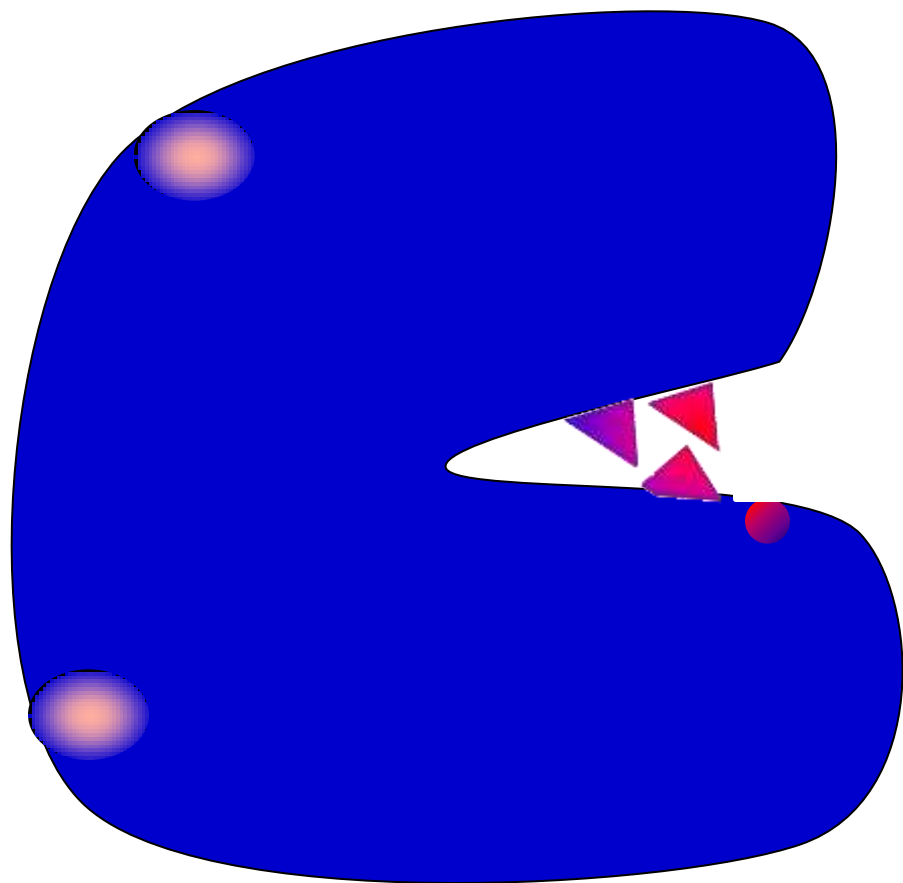
Example: angle energy



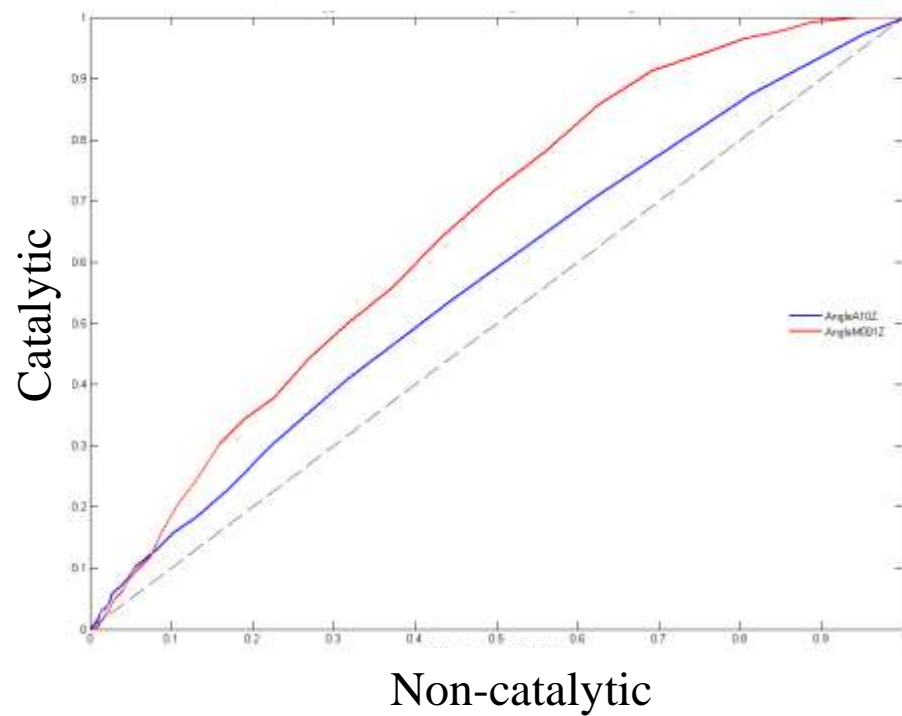
Structural clue:

 Catalytic residues are strained

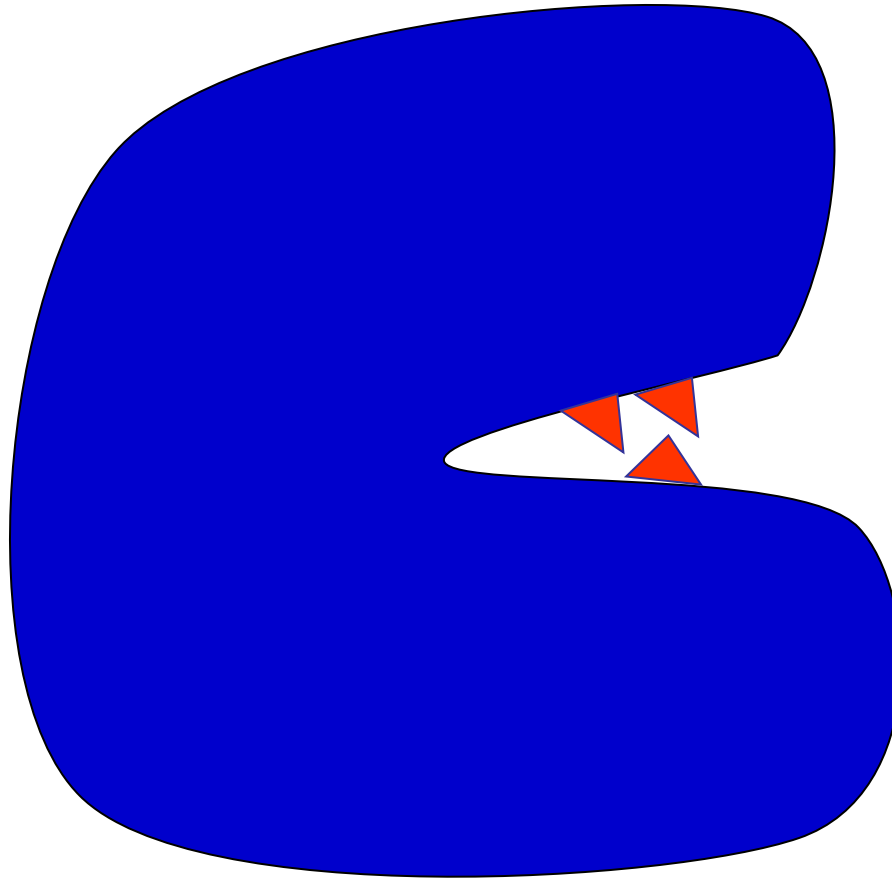
 Spatial averaging



ROC -curve



**Structural clue:**  Catalytic residues tend to be rather buried



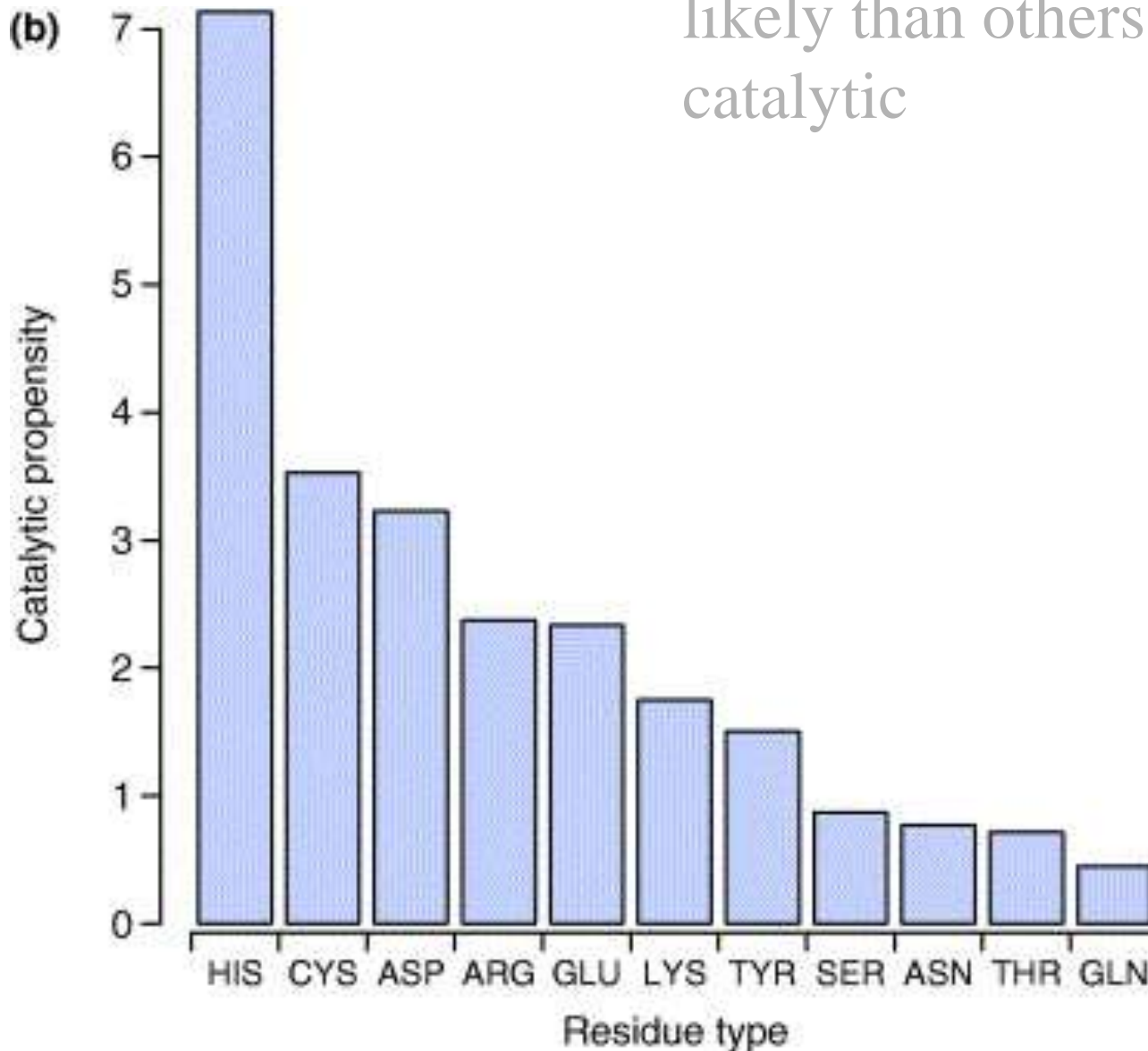
**Common knowledge:** 🌐 Some residue types are more likely than others to be catalytic

## Round Up The Usual Suspects



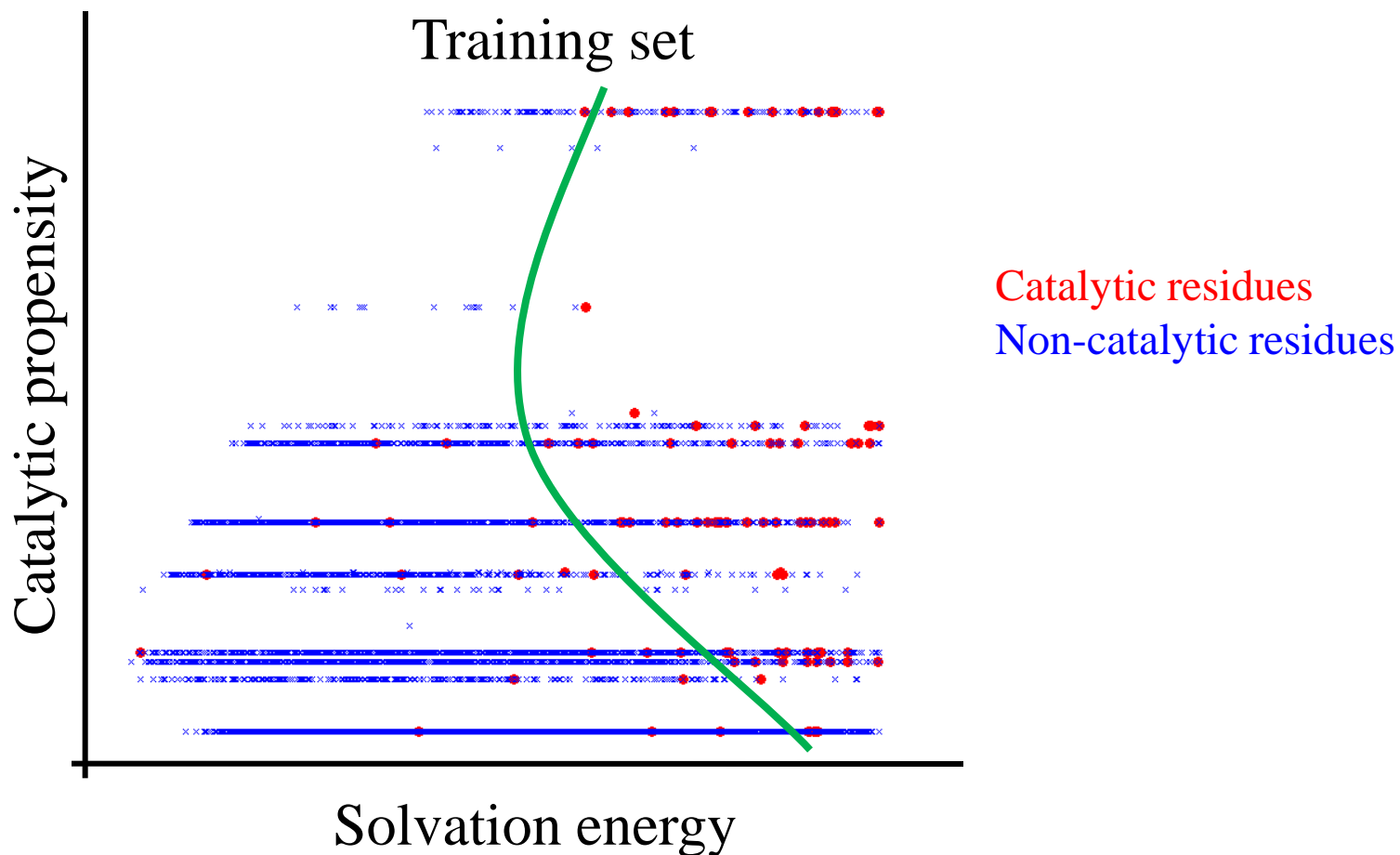
Common knowledge:

Some residue types are more likely than others to be catalytic



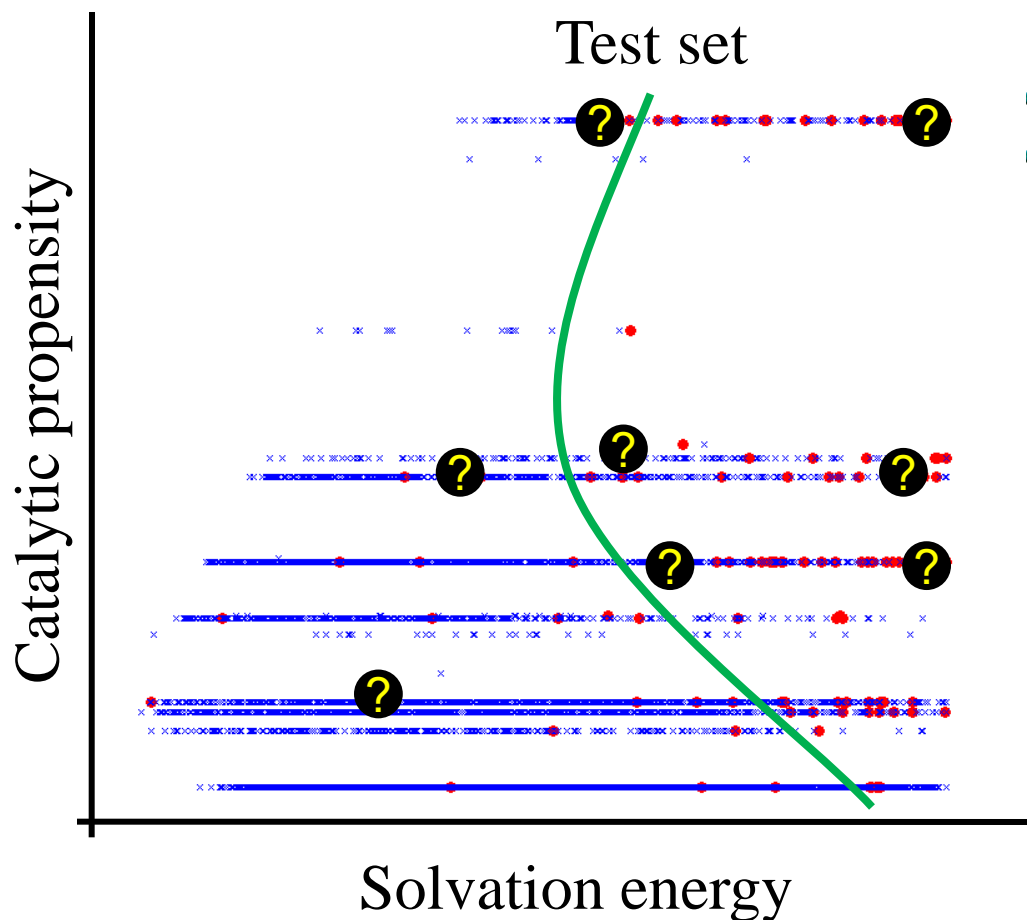
**Our approach:** Use Support Vector Machine (SVM) to integrate various clues.

A two dimensional example



Our approach:  Use Support Vector Machine (SVM) to integrate various clues.

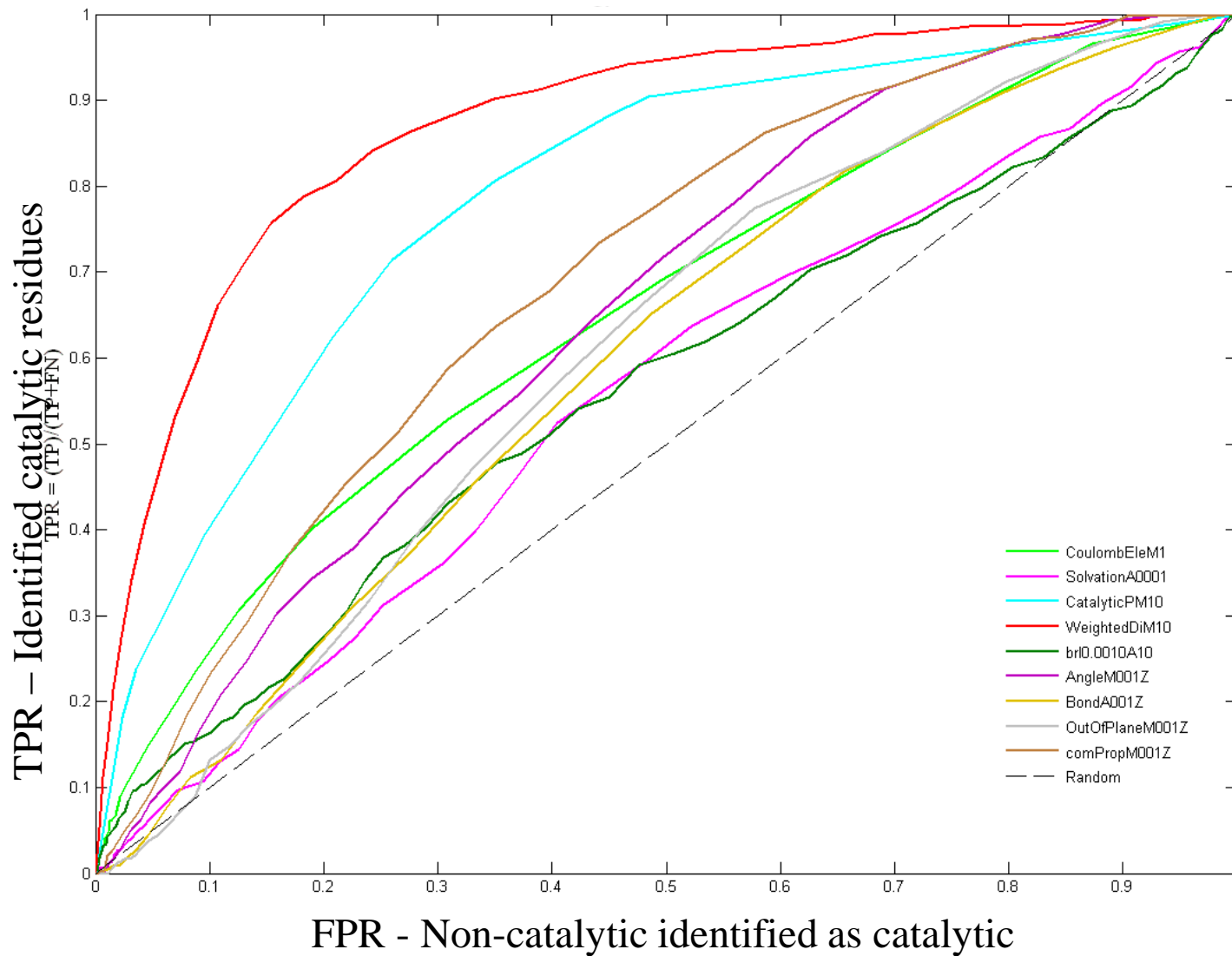
## A two dimensional example



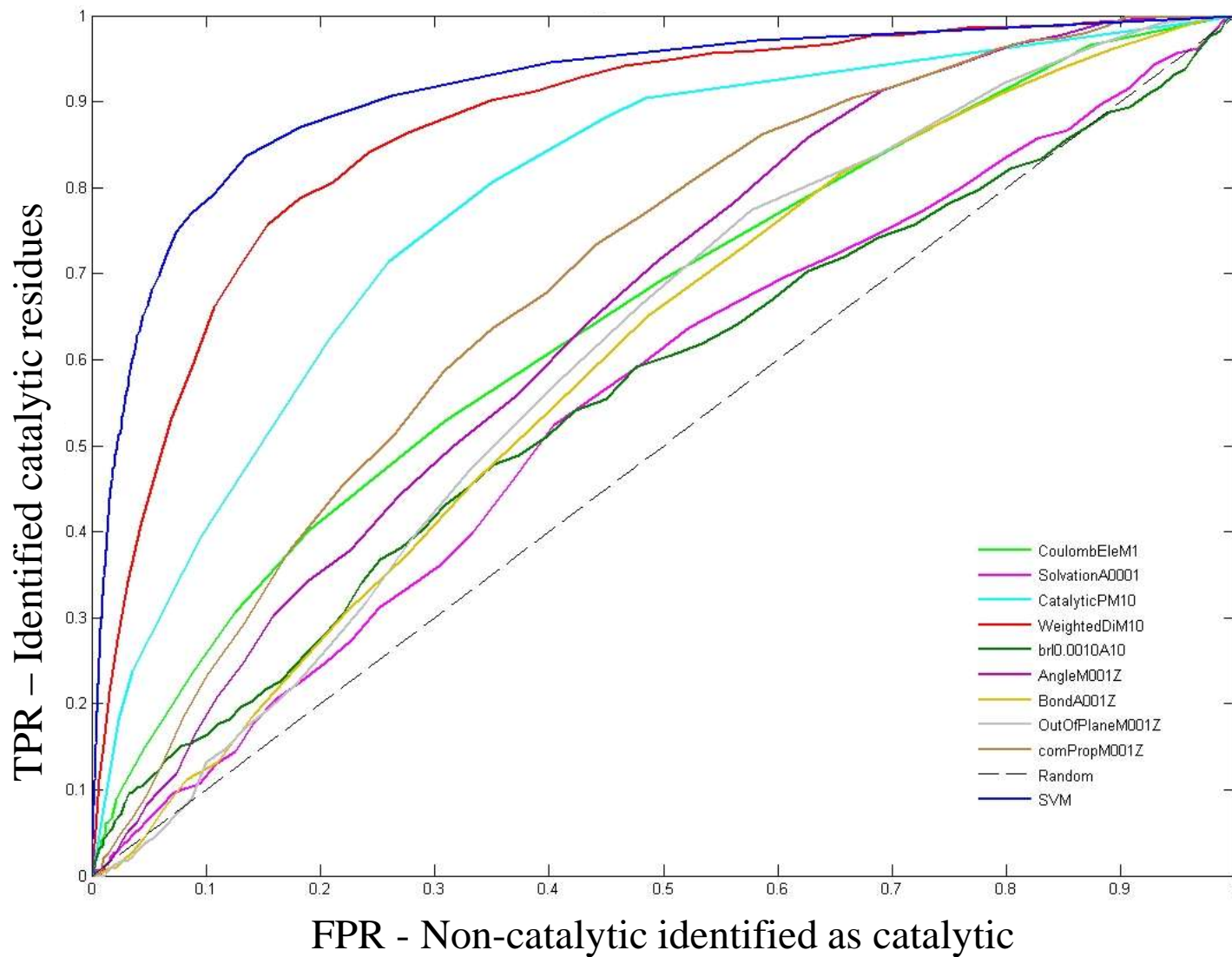
## Problems

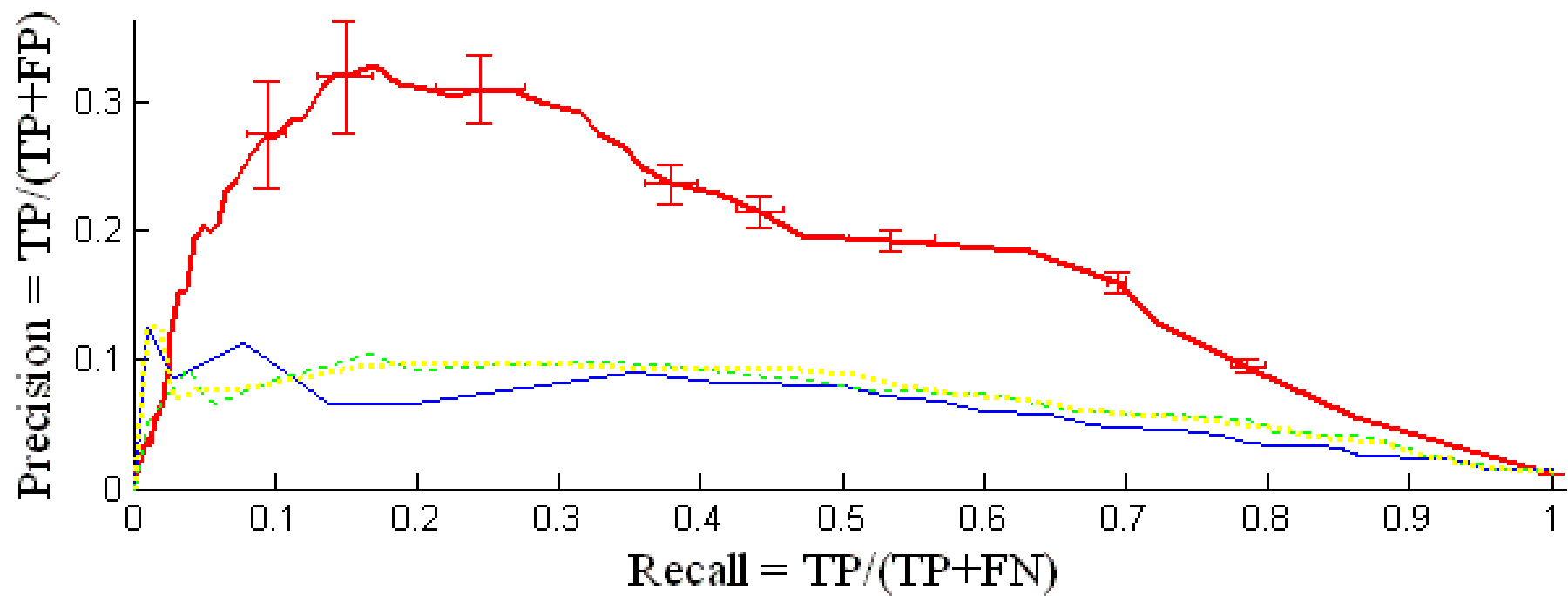
-  Unbalanced dataset
-  Small dataset

# Results: ROC curves

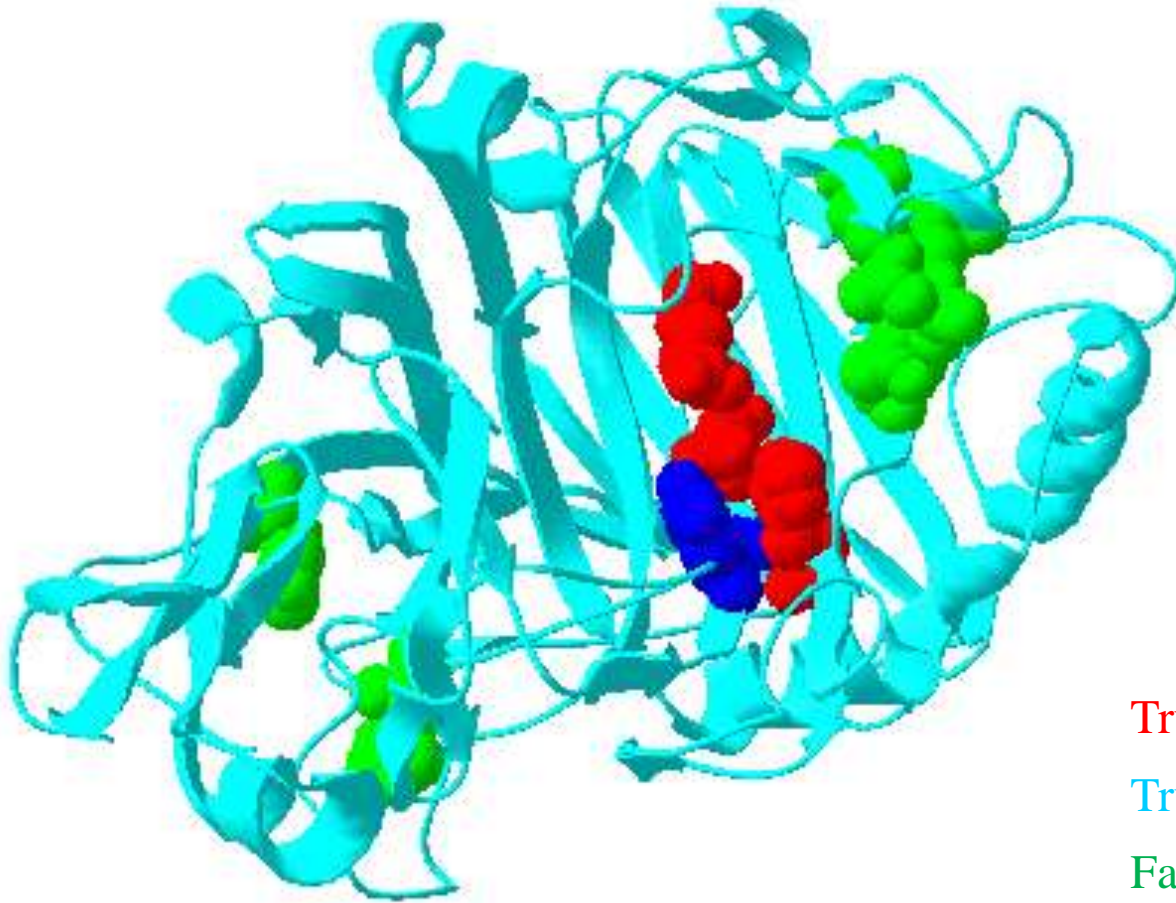


# Results: ROC curves





Results:  Example – 1GPI



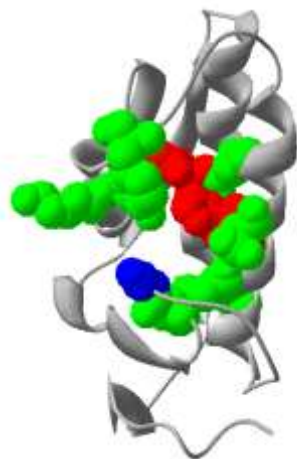
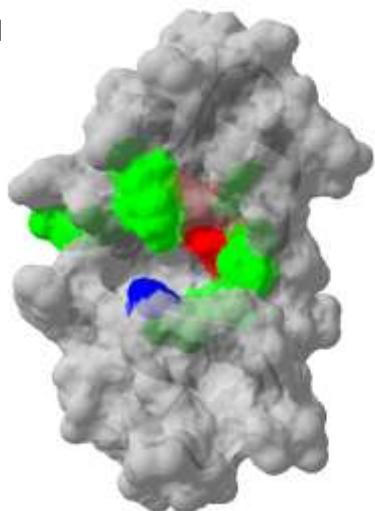
True Positive

True Negative

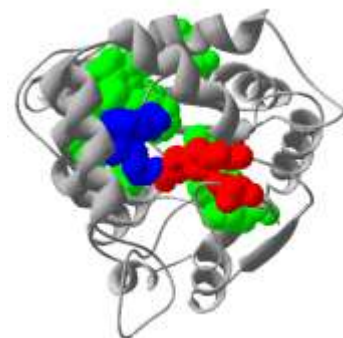
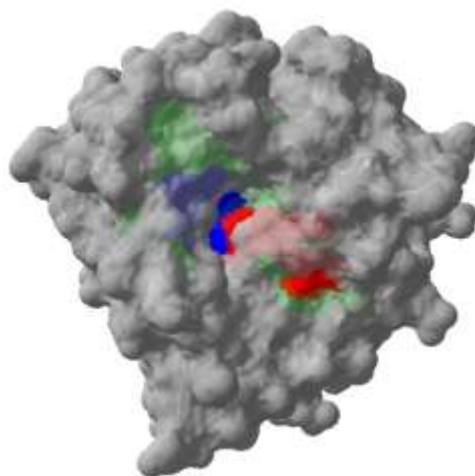
False Positive

False Negative

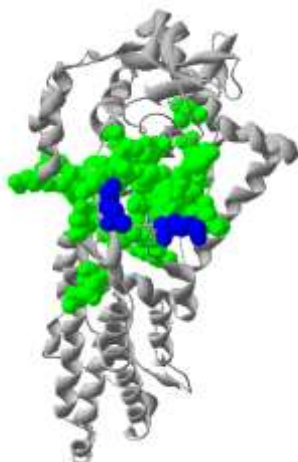
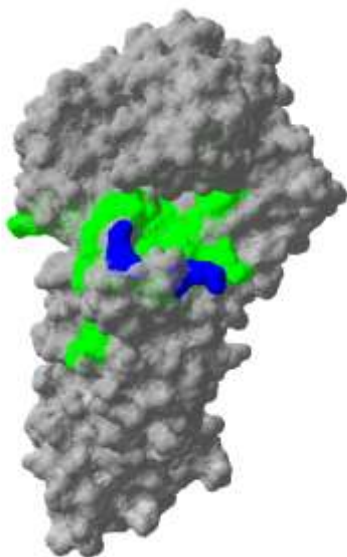
1a3d



1a8q



1a8h



# Thanks

- Dan Reshef  
● Ran Yahalom Did the work
- Boaz Lerner, DIEM Machine learning
- Nir Kalisman  
● El-ad Amir Energy functions
- Tetyana Maximova Software infrastructure

GeneFun & ISF funding