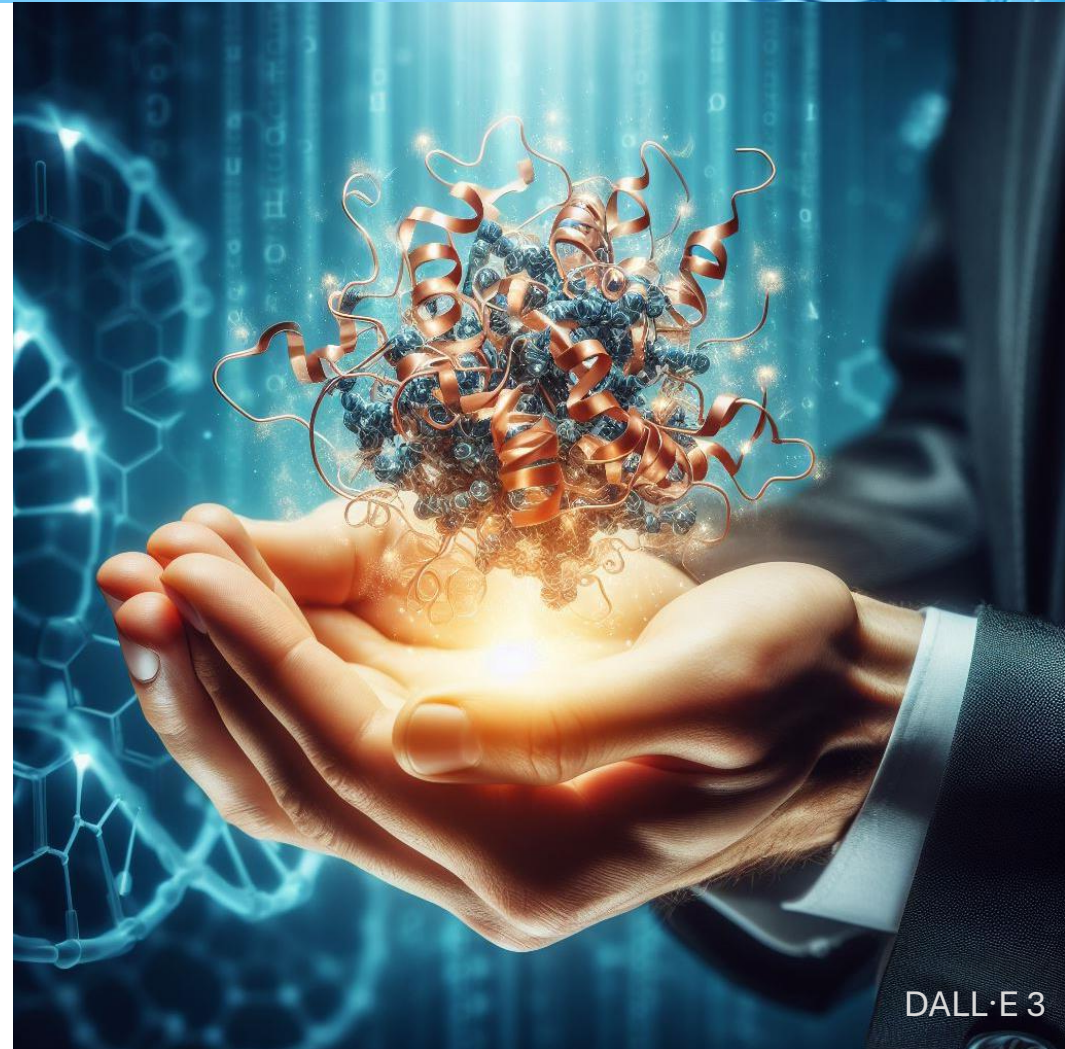# Protein Engineering

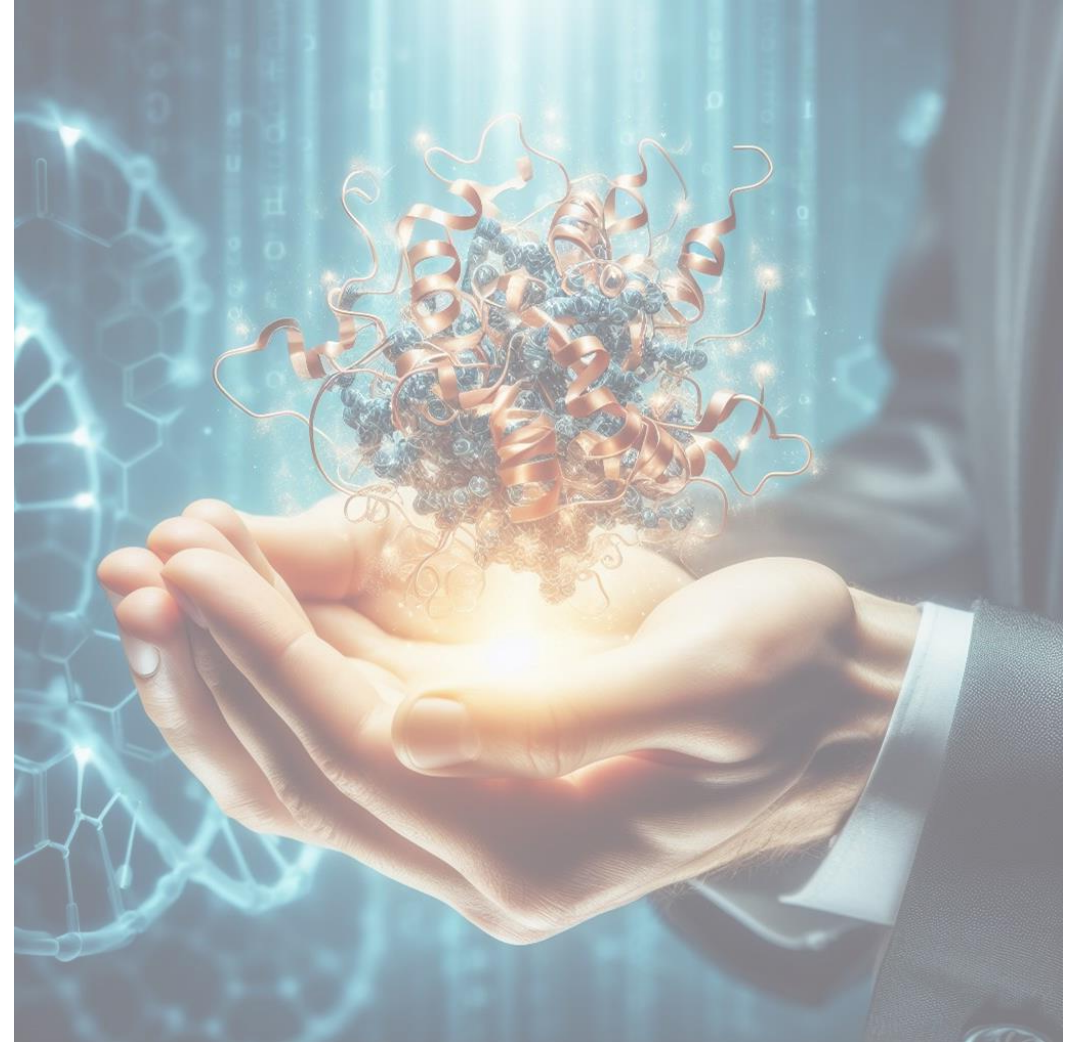Molecular Biotechnology Lecture #4

Michal Vašina

16/10/2024

Outline

1. Proteins in biotechnology

2. Aims of protein engineering

3. Main strategies

   ►**Directed evolution**

   ►**Rational design**

   ►**Machine learning**
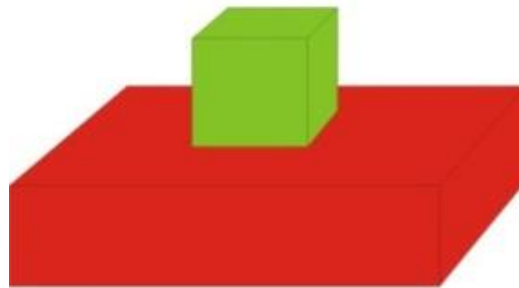
   ►**Semi-rational design**

Proteins in Biotechnology

▶ key problem -availability of optimal protein for specific process

▶ traditional biotechnology - adapt process

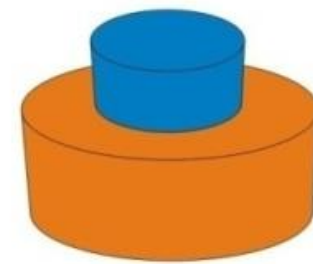▶ modern biotechnology - adapt protein

Available catalyst

Dream Process

Available catalyst

Nightmare Process

Adapted catalyst

Dream Process

▶Applied Molecular Biotechnology **Lectures 8-12**

LOSCHMIDT LABORATORIES

How to get new protein?

Classical Screening
- ► screening culture collections
- ► polluted and extreme environment

Environmental Gene Libraries
- ► metagenomic DNA

Database mining
- ► gene databases
- ► (meta)genome sequencing projects
- ► numerous uncharacterized proteins

Query Sequence
MSAIKMHECOSNLGSEPOFRSGDETSQKARFERLPPADTRYVISETKDKL

Selection Table

Similarity Network

Hon et al. *Nucleic Acids Research* 2020, link

https://loschmidt.chemi.muni.cz/enzymeminer/

How to get new protein?

Classical Screening

▶ screening culture collections
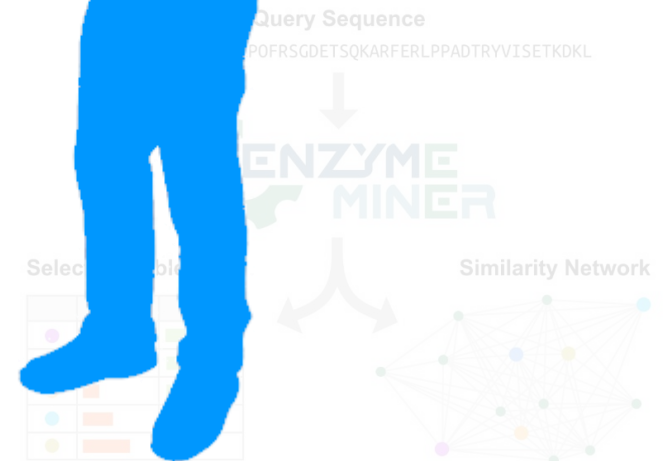
▶ polluted

If suitable protein does not exist in nature?

Environmental Gene Libraries

▶ metagenomic DNA

▶ Protein Engineering

Database mining

▶ gene databases

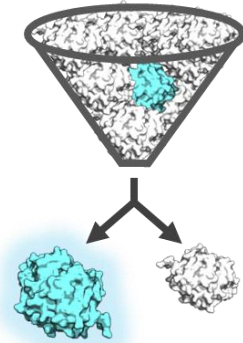▶ (meta)genome sequencing projects

▶ numerous uncharacterized proteins

Hon et al. *Nucleic Acids Research* 2020 (link)
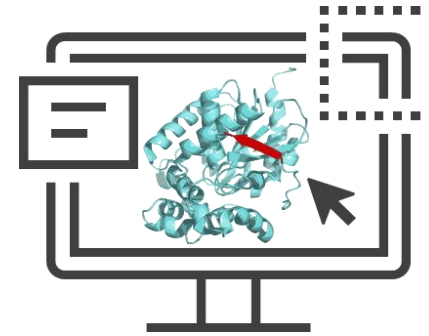
https://loschmidt.chemi.muni.cz/enzymeminer/

LOSCHMIDT LABORATORIES

How to get new protein?

Protein Discovery

Directed Evolution

Rational Design

Machine Learning

LOSCHMIDT LABORATORIES
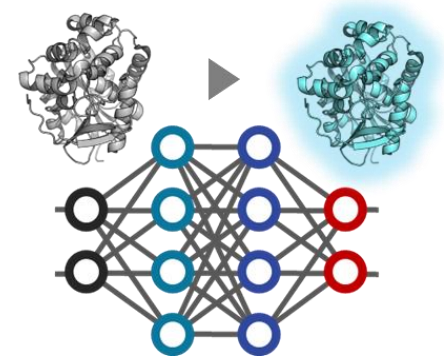
Protein Engineering at a glance

► use of genetic manipulations to alter the coding sequence of a gene and thus **modify the properties of the protein**

► General engineering cycle: **Design-build-test-learn**

AIMS AND APPLICATIONS

► **technological** - optimization of the protein to be suitable in particular technology process

► **scientific** - desire to understand what elements of proteins contribute to folding, stability and function

LOSCHMIDT LABORATORIES

What shall we improve?

structural properties of proteins
- ▶ **stability** (temperature, solvents)
- ▶ tolerance to pH, salt
- ▶ Resolve the **atomic structure** (to understand function)

functional properties of proteins
- ▶ **substrate specificity** and selectivity
- ▶ **kinetic properties** (e.g., K_m, k_{cat}, K_i)
- ▶ Inhibition by small molecules (drugs)
- ▶ protein-protein or protein-DNA interactions

Target properties

Stability
Thermostability
Folding
% Unfolded / Temperature

Structure
X-ray crystallography
NMR, Cryo-EM SAXS

Activity
Optimal T/pH
In vivo / In vitro
Activity / Temperature
Activity / Substrates

Kinetics
[S] / [I]
Steady-state
Transient-state
Rate / [Substrate]
Rate / [Inhibitor]

LOSCHMIDT LABORATORIES

Overview of strategies

Enzyme Discovery

Directed Evolution

Rational Design

Machine Learning

Hit selection

Biochemical Characterization

Stability	Structure	Activity	Kinetics

[S] / [I]

LOSCHMIDT LABORATORIES
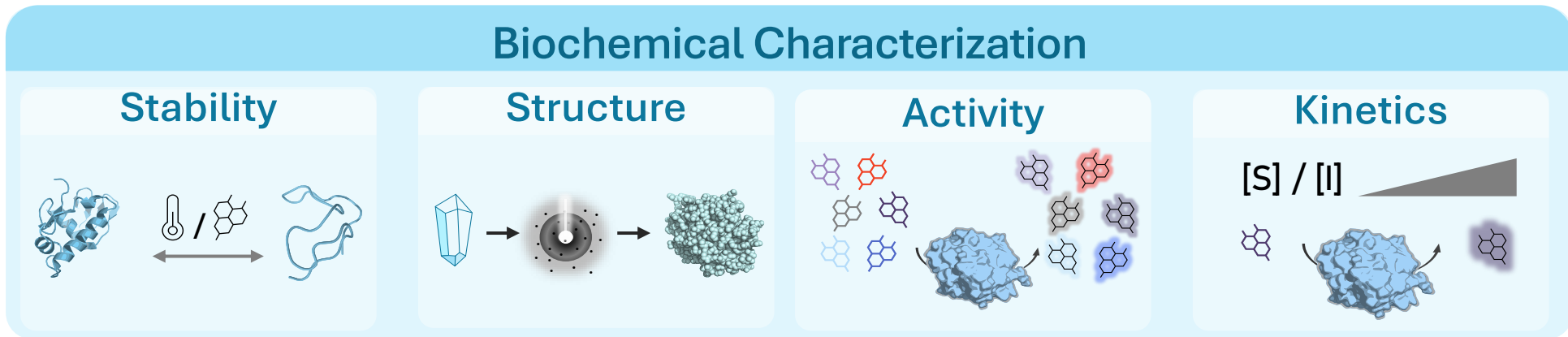
Main strategies

RATIONAL DESIGN

1. Computer aided design

2. Site-directed mutagenesis

Individual mutated gene

3. Transformation

4. Protein expression

5. Protein purification

6. *not applied*

Constructed mutant enzyme

IMPROVED ENZYME

7. Biochemical testing
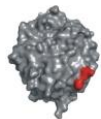
DIRECTED EVOLUTION

1. *not applied*

2. Random mutagenesis
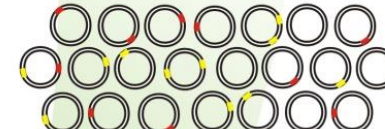
Library of mutated genes
(>10,000 clones)

3. Transformation

4. Protein expression

5. *not applied*

6. Screening and selection
- stability
- selectivity
- affinity
- activity

Selected mutant enzymes

LOSCHMIDT
LABORATORIES
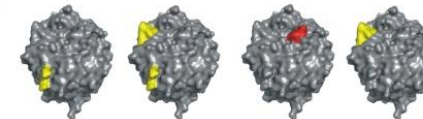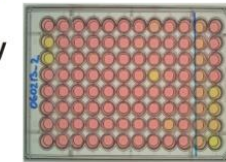
Main strategies

RATIONAL DESIGN

DIRECTED EVOLUTION

GO SMART!

GO HARD!

IMPROVED ENZYME

7. Biochemical testing
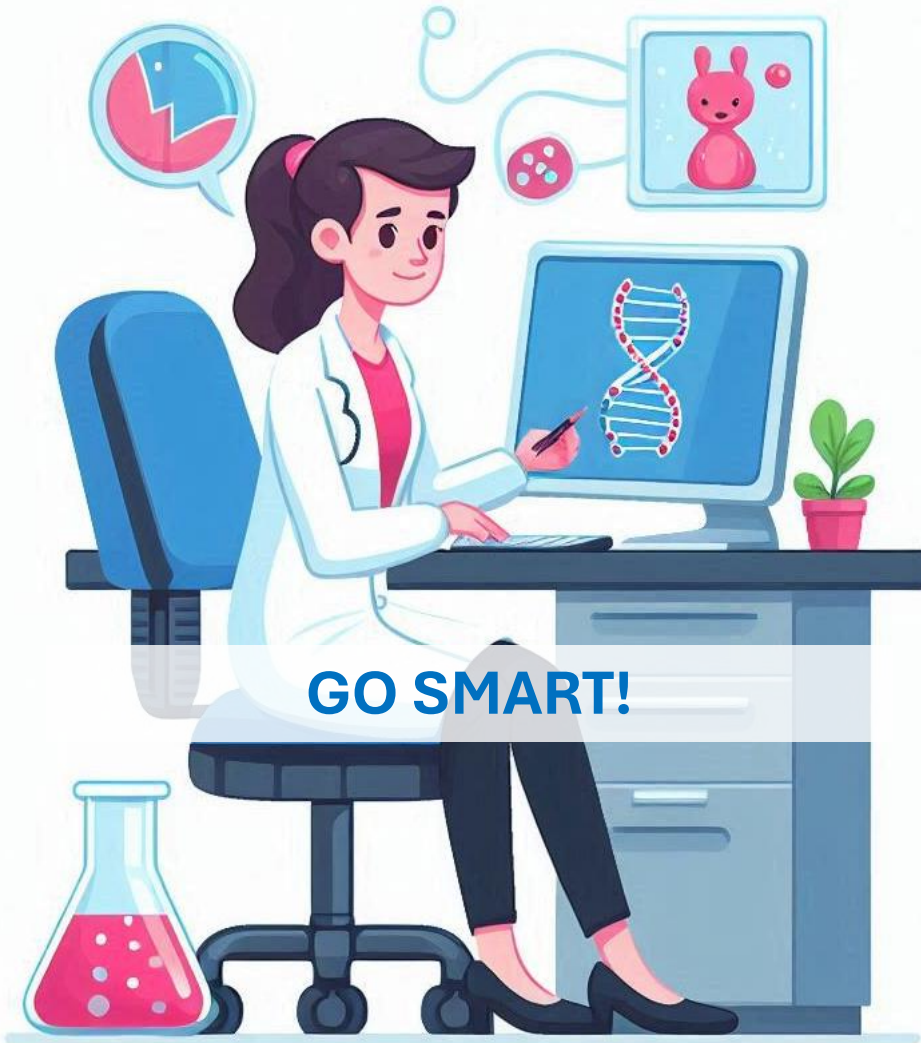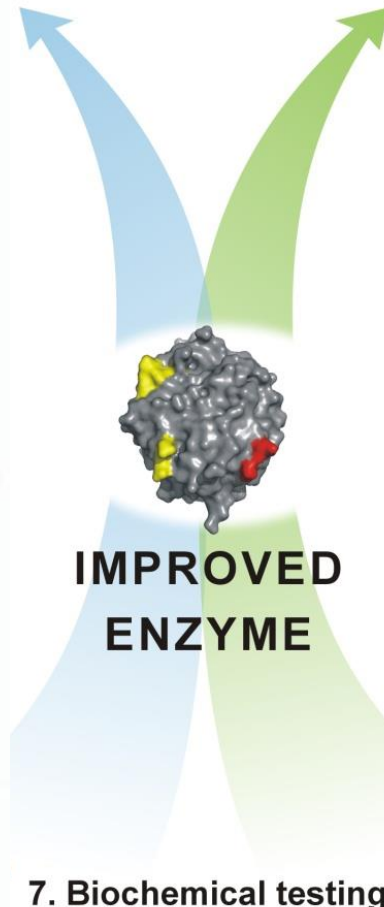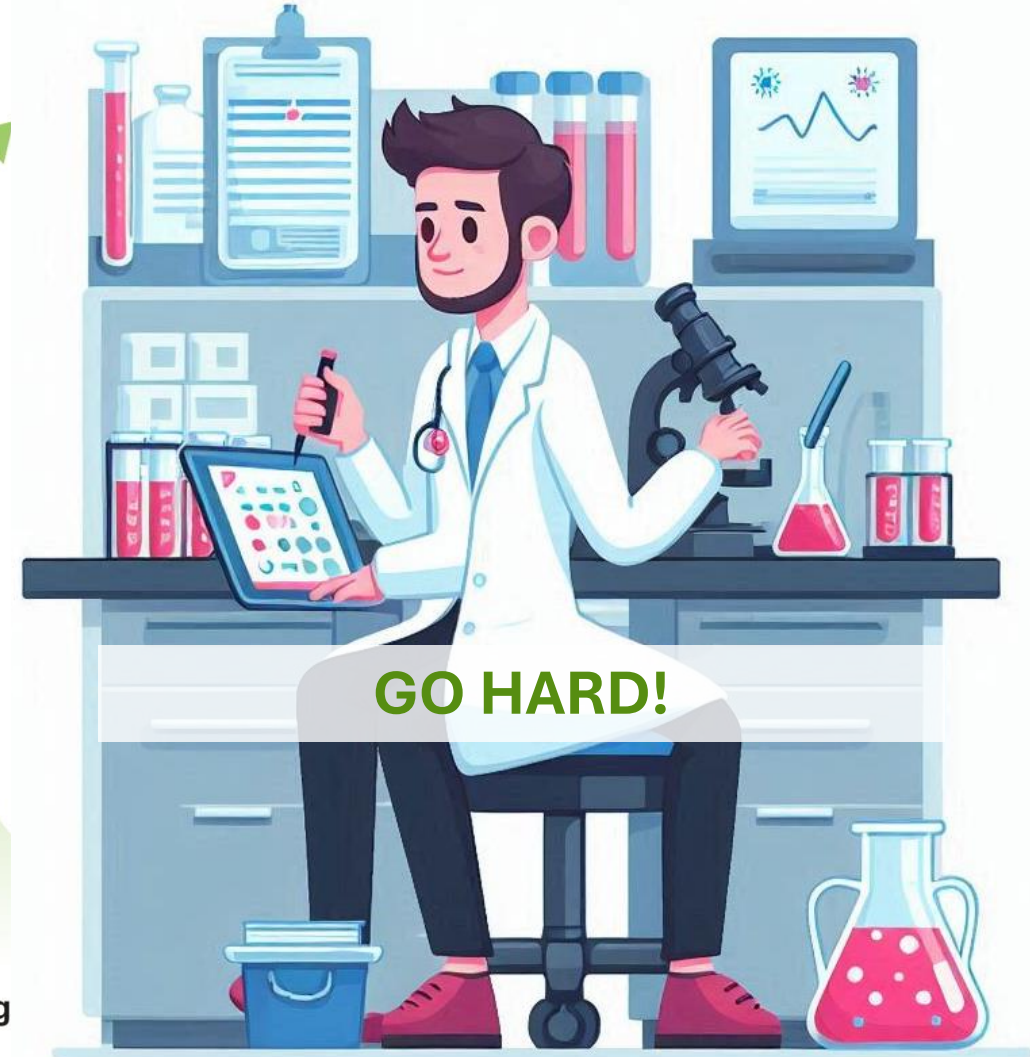
Directed Evolution

► emerged during mid-1990s

► inspired by natural evolution

► "laboratory evolution"

  ► requires outside intelligence, not blind chance

  ► does not take millions of years, but happens rapidly

The Nobel Prize in Chemistry 2018

Frances H. Arnold

Prize share: 1/2

The Nobel Prize in Chemistry 2018 was divided, one half awarded to Frances H. Arnold "for the directed evolution of enzymes",

https://www.nobelprize.org/prizes/chemistry/2018/summary/

LOSCHMIDT LABORATORIES

Directed Evolution

►**evolution in test tube** comprises two steps

    ►**random mutagenesis**
    building mutant library (diversity)

    ►**screening and selection**
    identification of desired biocatalyst

►**prerequisites** for directed evolution

    ►gene encoding protein of interest

    ►method to create mutant library

    ►suitable expression system

    ►screening or selection system

1. *not applied*

2. Random mutagenesis

Library of mutated genes
(>10,000 clones)

3. Transformation

4. Protein expression

5. *not applied*

6. Screening and selection
- stability
- selectivity
- affinity
- activity

Selected mutant enzymes

7. Biochemical testing

IMPROVED ENZYME

LOSCHMIDT LABORATORIES
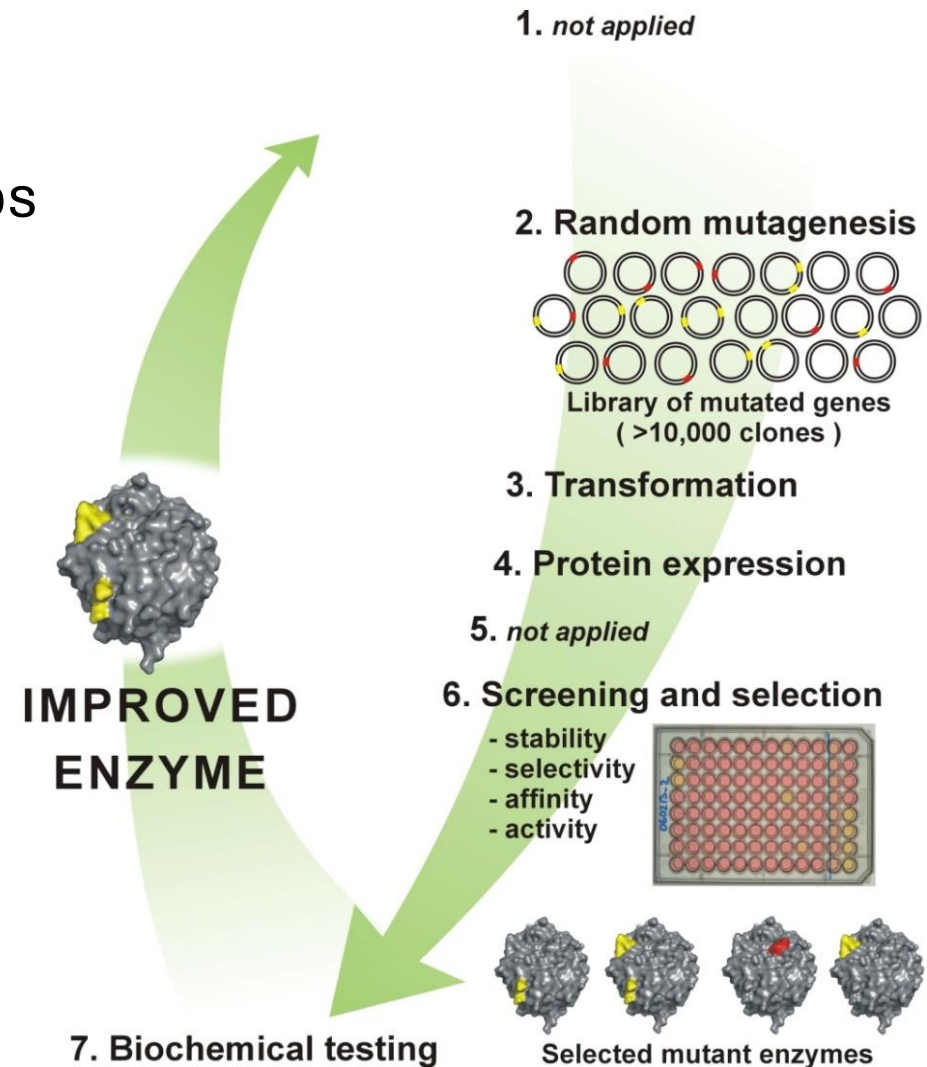
Methods to create mutant libraries

►technology to generate large diversity

►**Non-recombining**
one parent gene -> variants with point mutations
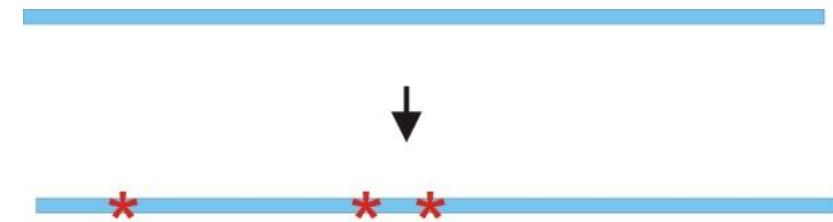
►**Recombining** (also „sexual mutagenesis")
several parental homologous genes -> chimeras
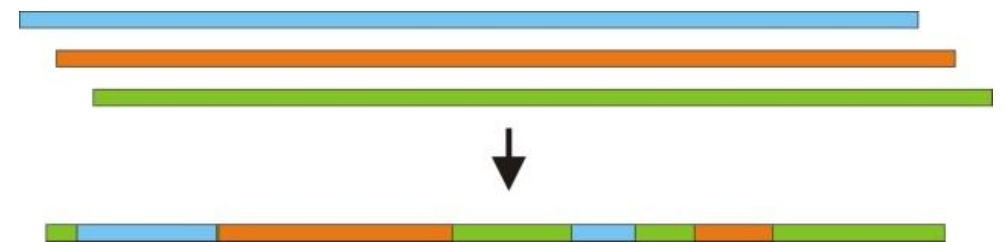
Non-recombining mutagenesis

▶ UV irradiation or chemical mutagens (traditional)

▶ **mutator strains** - lacks DNA repair mechanism

mutations during replication (e.g., *Epicurian coli* XL1-Red)

▶ **error-prone polymerase chain reaction** (ep-PCR)

    ▶ gene amplified in imperfect copying process
    (e.g., unbalanced deoxyribonucleotides concentrations,
    high Mg^{2+} concentration, Mn^{2+}, low annealing temperatures)

    ▶ 1 to 20 mutations per 1,000 base pairs

▶ **site-saturation mutagenesis**

    ▶ randomization of single or multiple codons

    ▶ degenerate primers (NNN for complete randomization)

▶ **other methods**

    ▶ insertion/deletions (InDel)

    ▶ cassette mutagenesis (region mutagenesis)

Error-prone mutagenesis

Site saturation mutagenesis
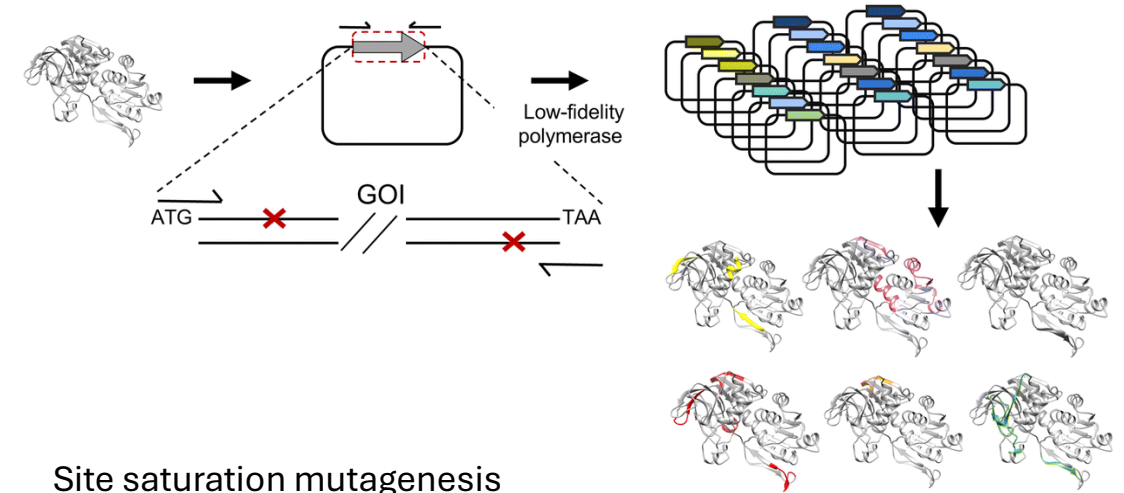
LOSCHMIDT LABORATORIES
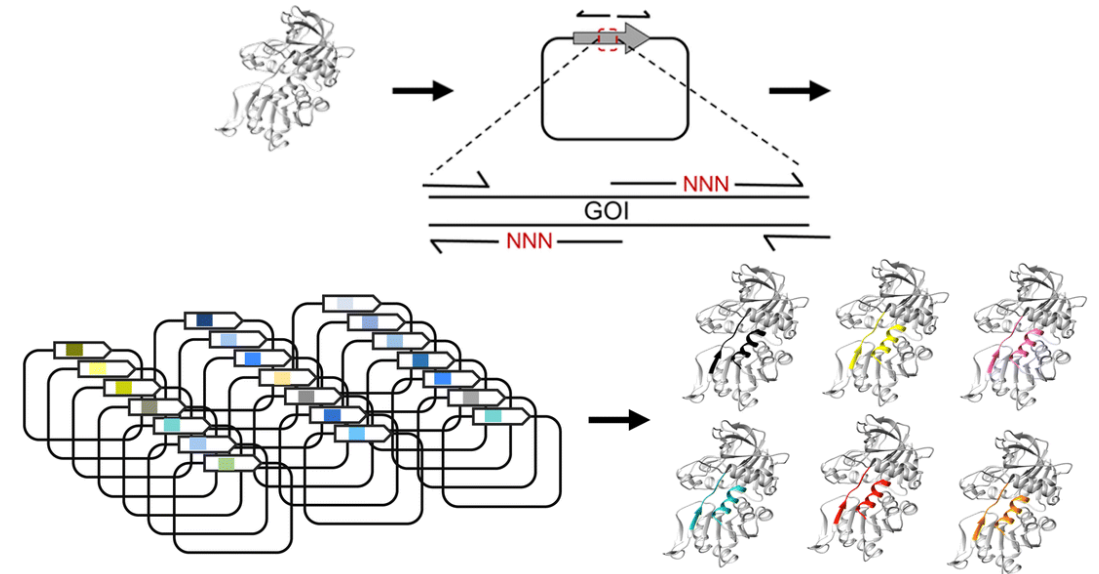
Recombining mutagenesis

▶**DNA shuffling**

  ▶fragmentation step

  ▶random reassembly of segments

▶**StEP - staggered extension process**

  ▶simpler then shuffling, no fragmentation

  ▶random reannealing combined with limited primer extension

▶**other methods**

shuffling of genes with lower homology down to 70% (e.g., RACHITT, ITCHY, SCRATCHY)

DNA-shuffling

GOIs   DNase I   Annealing   Extension   Shuffled gene

StEP

Primers bind denatured template   Short fragments are synthesized   Fragments randomly prime the templates and are extended further   Full-length genes

⬤ = Brief polymerase-catalysed extension
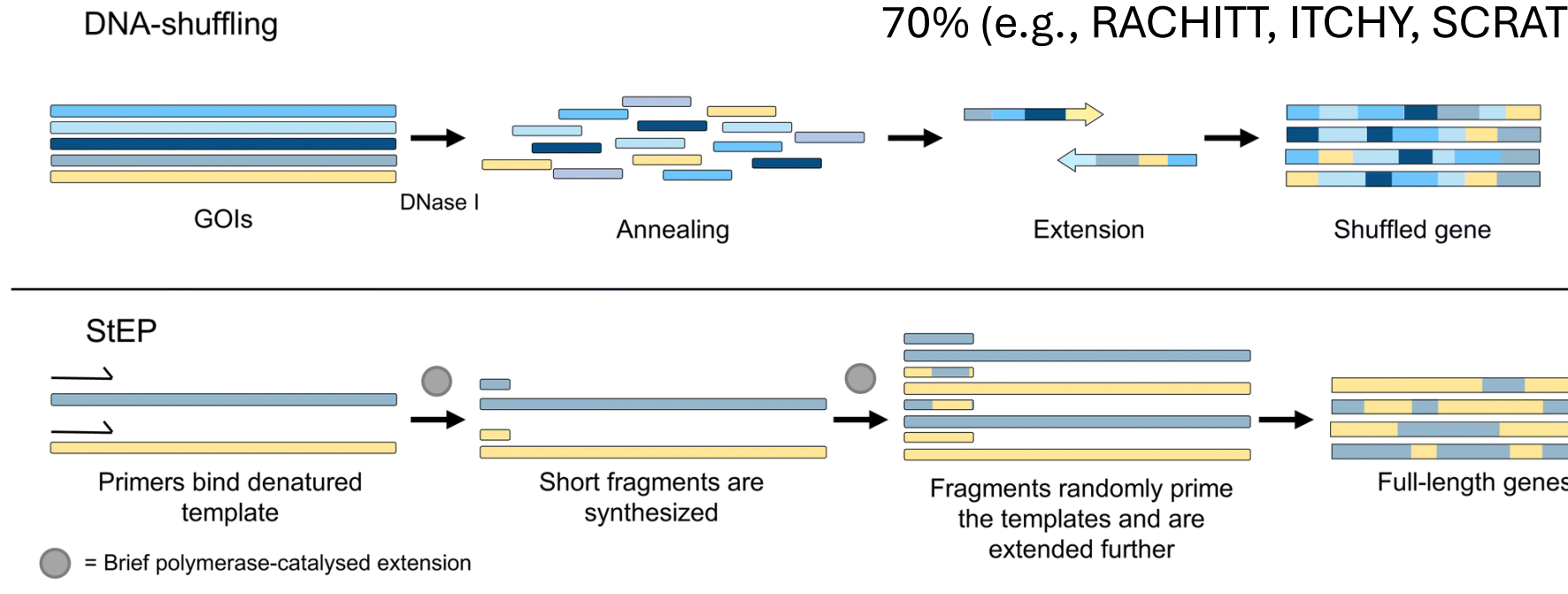
Screening and selection

► most critical step of direct evolution

► isolation of positive mutants hiding in library

► genotype to phenotype linkage is crucial

► **High-throughput screening**
experimental testing of variants one by one

► **Direct selection**
applying selective pressure to the library
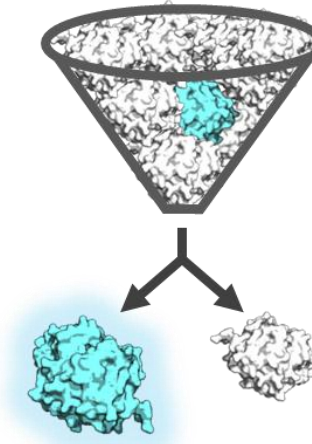
(Ultra)-High throughput screening

►Golden rule: *„You get what you screen for!"*

►agar plate (pre)screening

►microtiter plates screening

  ►96-, 384- or 1536-well formats

  ►robot assistance (colony picker, liquid handler)

  ►10^4 libraries

  ►volume 10 – 100 µL

►microfluidic systems

  ►water in oil emulsions (up to 10 kHz)

  ►FACS sorting (10^8 events/hour)

  ►10^9 libraries

  ►volume 1 – 10 pL
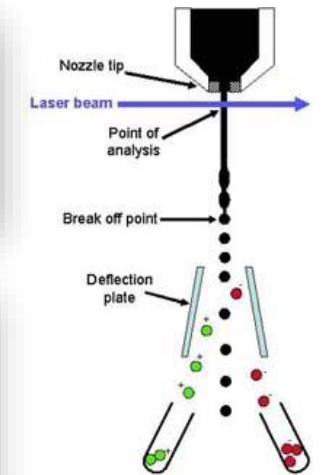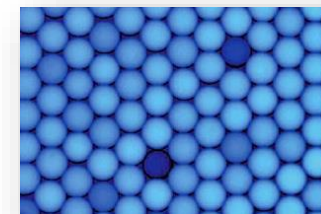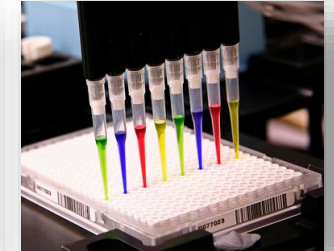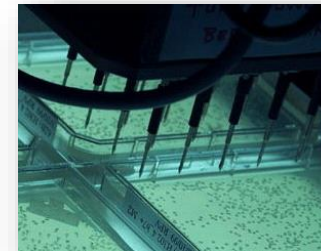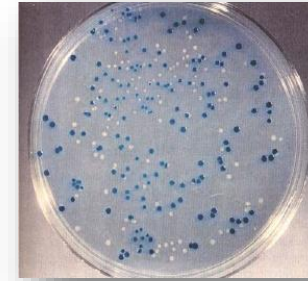
Experimental throughput is critical

STANDARD DESIGN

▶ Random mutagenesis (2-3 positions)

▶ Library of **10^4 clones**

volume: 100´ µL
assays/day: 10^3
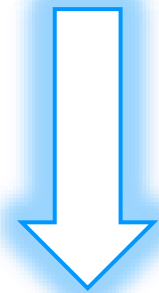
ADVANCED DESIGN

▶ Random mutagenesis (5-7 positions)

▶ Library of **> 10^6 clones**

volume: 10´ pL
assays/day: 10^7

▶ Microfluidics **Lecture 7**

Direct selection

► not generally applicable
(mutant libraries $>10^6$ variants)

► link between genotype and phenotype

► **display technologies**
  ► ribosome, phage display
  ► yeast, bacteria display

► **life-or-death assay**
  ► auxotrophic strain
  ► toxicity based selection
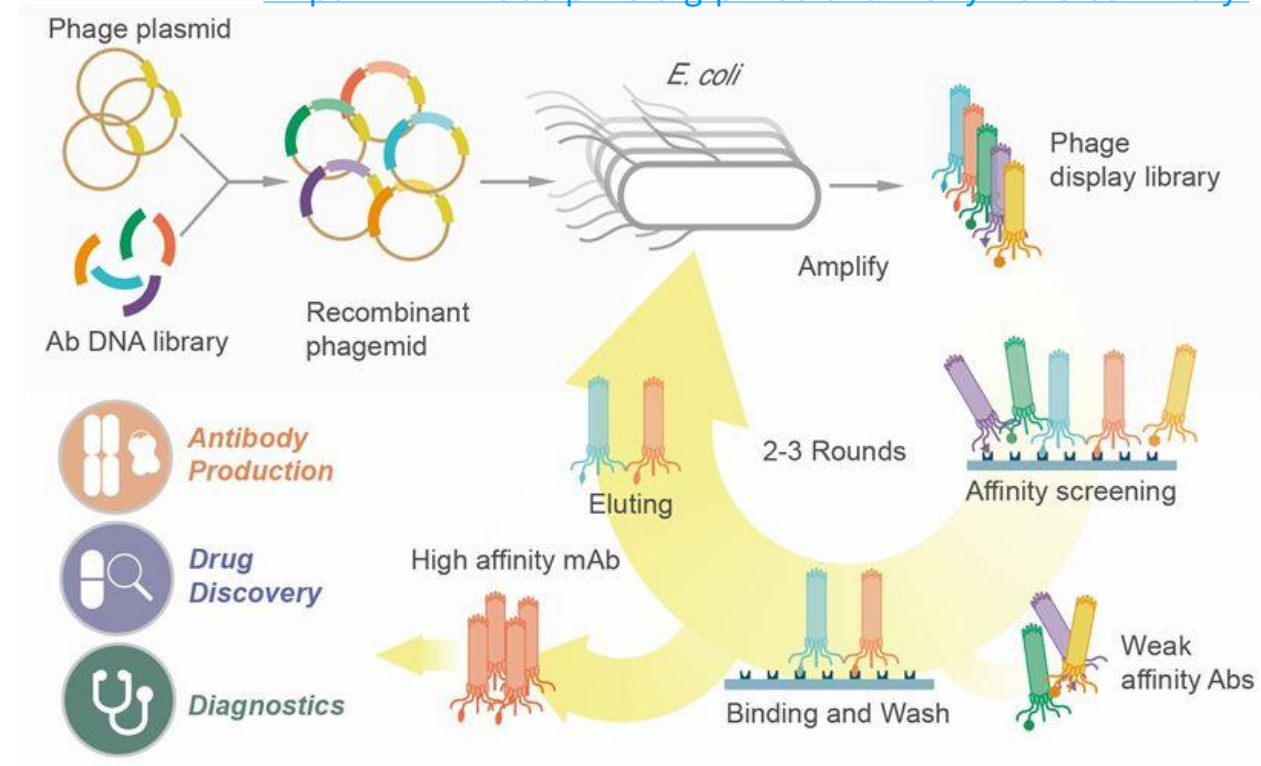
Success story #1

directed evolution of enantioselectivity

►lipase from *P. aeruginosa* (E-value improved from 1.1 into 51)

►spectrophotometric screening of (R)- and (S)-nitrophenyl esters

►40,000 variants screened

►the best mutant contains six amino acid substitutions
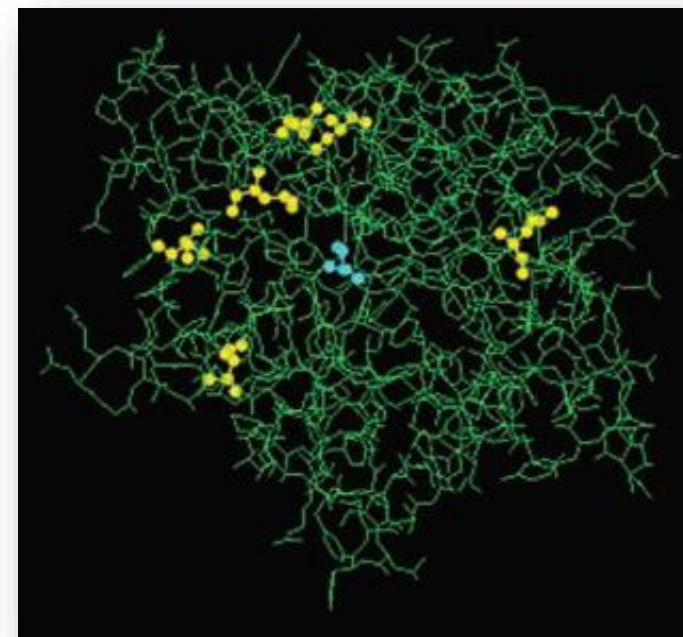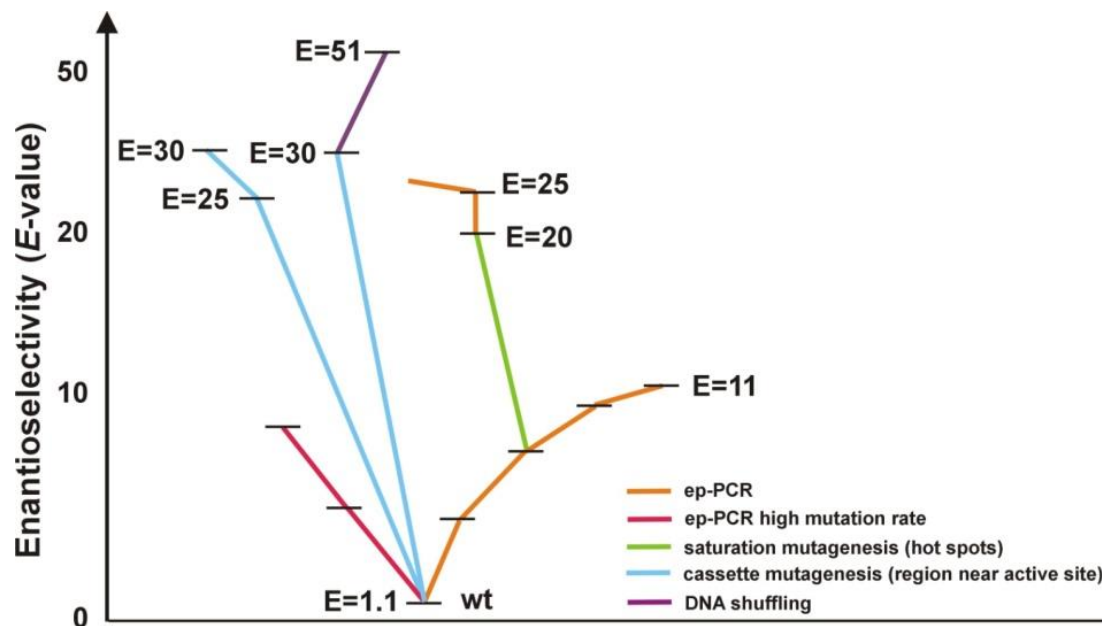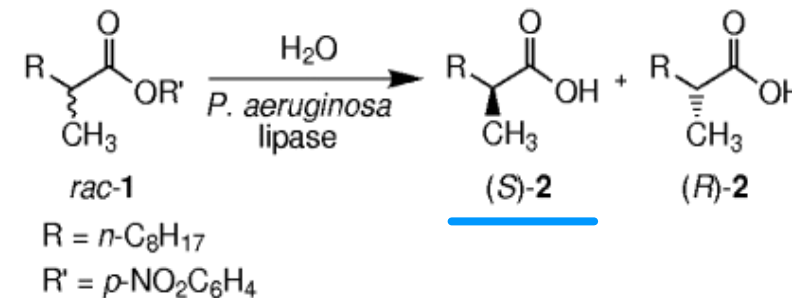
Reetz, et al. *Angew Chem* 2001, link   21

Main strategies

Main strategies
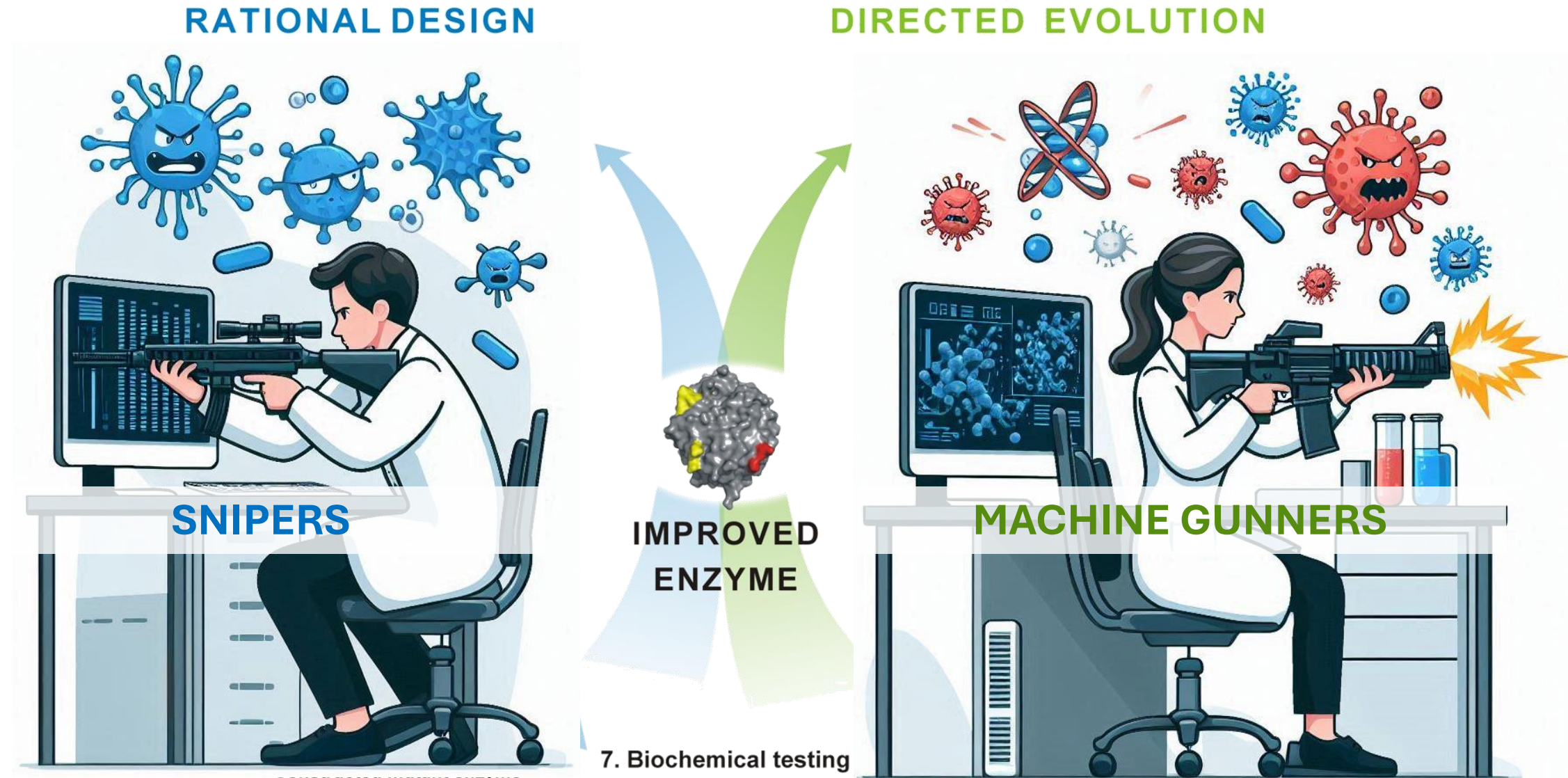
Rational design introduction

1. Computer aided design
2. Site-directed mutagenesis
   Individual mutated gene
3. Transformation
4. Protein expression
5. Protein purification
6. *not applied*
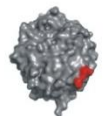
IMPROVED ENZYME

Constructed mutant enzyme
7. Biochemical testing

▶ emerged around 1980s as the original protein engineering approach

▶ knowledge based - combining theory and experiment

▶ protein engineering cycle: „learn-design-build-test-learn"

▶ difficulty in prediction of mutation effects on protein property

▶ *de novo* design most challenging

Principals of rational design

1. Computer aided design
2. Site-directed mutagenesis
   Individual mutated gene
3. Transformation
4. Protein expression
5. Protein purification
6. *not applied*
7. Biochemical testing

Constructed mutant enzyme

IMPROVED ENZYME

► rational design comprises:
- ► **design** - understanding of protein functionality
- ► **experiment** - construction and testing of mutants

► **prerequisites** for rational design:
- ► **gene** encoding protein of interest
- ► **3D structure** (e.g., X-ray, NMR) or sequence alignment
- ► computational methods and capacity
- ► site-directed mutagenesis techniques
- ► efficient expression system
- ► biochemical assay to test mutants

LOSCHMIDT LABORATORIES

Bioinformatics-based design

►**Sequence homology approach**

- ►homologous wild-type sequences alignment
- ►identifying amino acid residues responsible for differences
- ►design - combination of positive mutation from all parental proteins

►**Ancestral reconstruction**

- ►construction of phylogenetic tree
- ►design - nods prediction by consensus approach

Bioinformatics-based design

► **Sequence homology approach**

  ► homologous wild-type sequences alignment

  ► identifying amino acid residues responsible for differences

  ► design - combination of positive mutation from all parental proteins

► **Ancestral reconstruction**

  ► construction of phylogenetic tree

  ► design - nods prediction by consensus approach
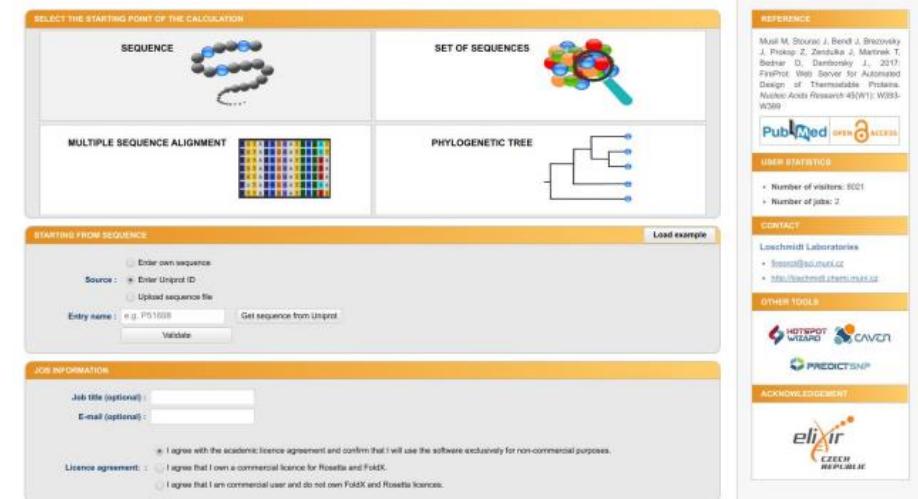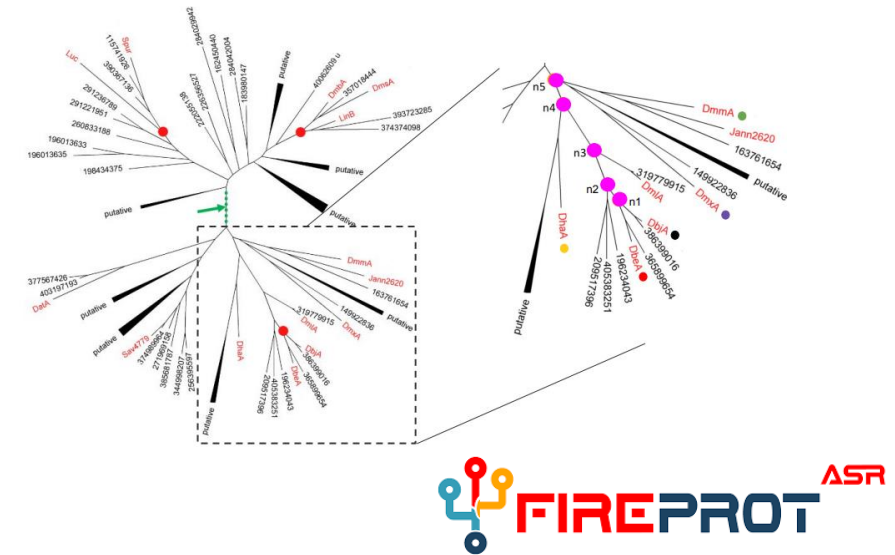
Musil et al. *Brief Bioinform* 2020, link

https://loschmidt.chemi.muni.cz/fireprotasr/

LOSCHMIDT LABORATORIES

Bioinformatika Bi5000/Bi5000c

►Období: podzim

►Rozsah: přednáška 2 hodiny/týden, cvičení 2 hodiny/týden

►Vyučující: prof. Mgr. Jiří Damborský, Dr., prof. RNDr. Roman Pantůček, Ph.D.,

►Osnova:

  ►bioinformatické databáze a jejich prohledávání

  ►analýza nukleotidových a proteinových sekvencí

  ►hledání a identifikace genů

  ►analýza a předpověď struktury proteinů
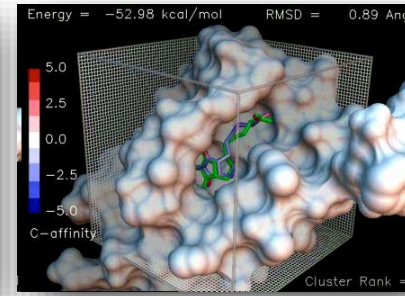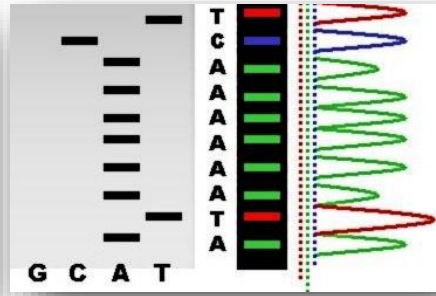
Structure-based design

▶ **prediction** of enzyme function from structure alone is challenging

▶ protein structure: experimental (X-ray crystallography, NMR),
computational (AlfaFold models, homology models!)

▶ molecular modelling

- ▶ molecular docking
- ▶ molecular dynamics
- ▶ quantum mechanics/molecular mechanics (QM/MM)
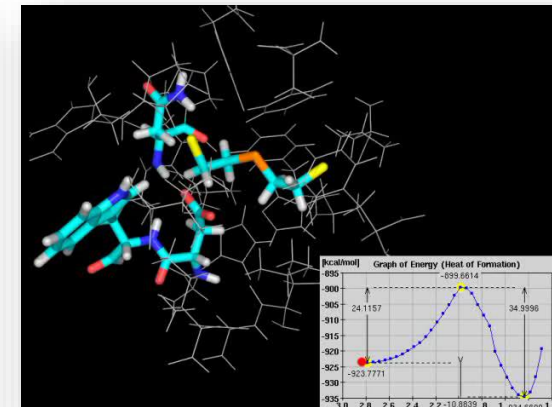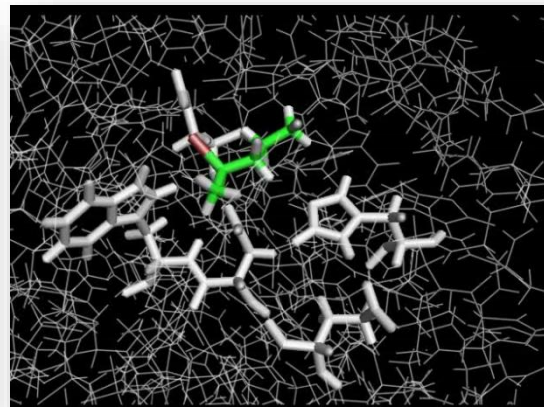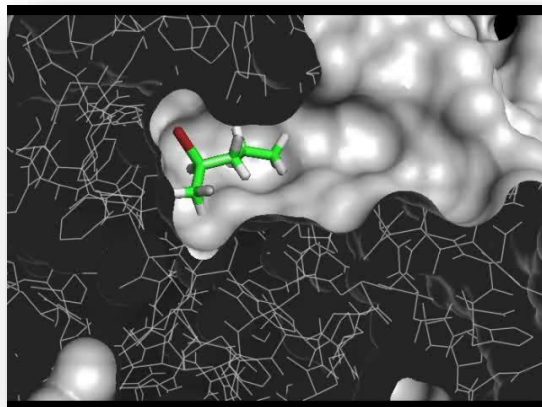
Strukturní biologie Bi9410/Bi9410c

► Období: podzim

► Rozsah: přednáška 2 hodiny/týden, cvičení 2 hodiny/týden

► Vyučující: doc. Mgr. David Bednář, Ph.D.

► Osnova:

    ► struktura, stabilita a dynamika biologických makromolekul

    ► makromolekulární interakce a komplexy

    ► stanovení a předpověď struktury, identifikace důležitých oblastí

    ► stanovení vlivu mutace na strukturu a funkci proteinu

    ► aplikace v biologickém výzkumu, návrhu léčiv a biokatalyzátorů
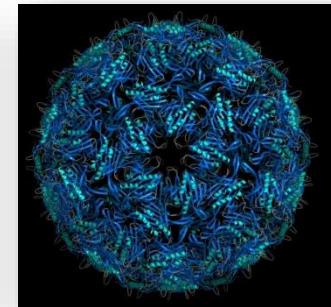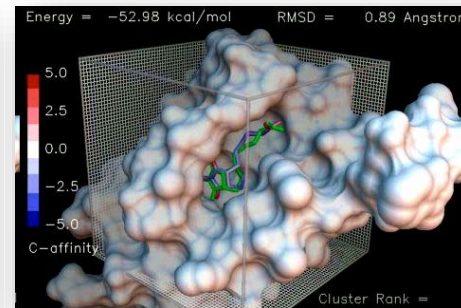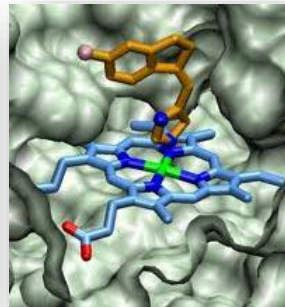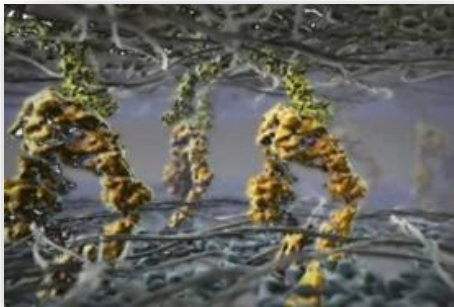
Gene of interest construction

► site-directed mutagenesis

  ► introducing point mutations

► multi site-directed mutagenesis

► gene synthesis

  ► commercial service

  ► codon optimization
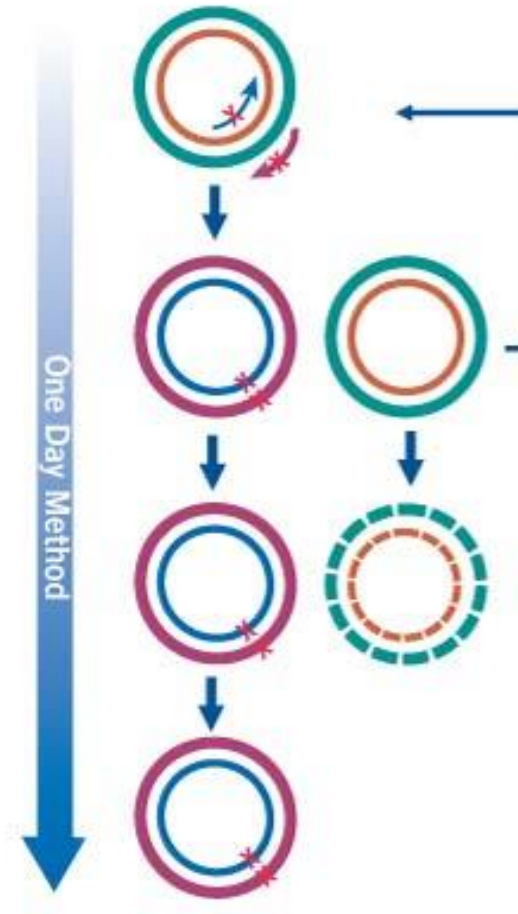
Rational design targets
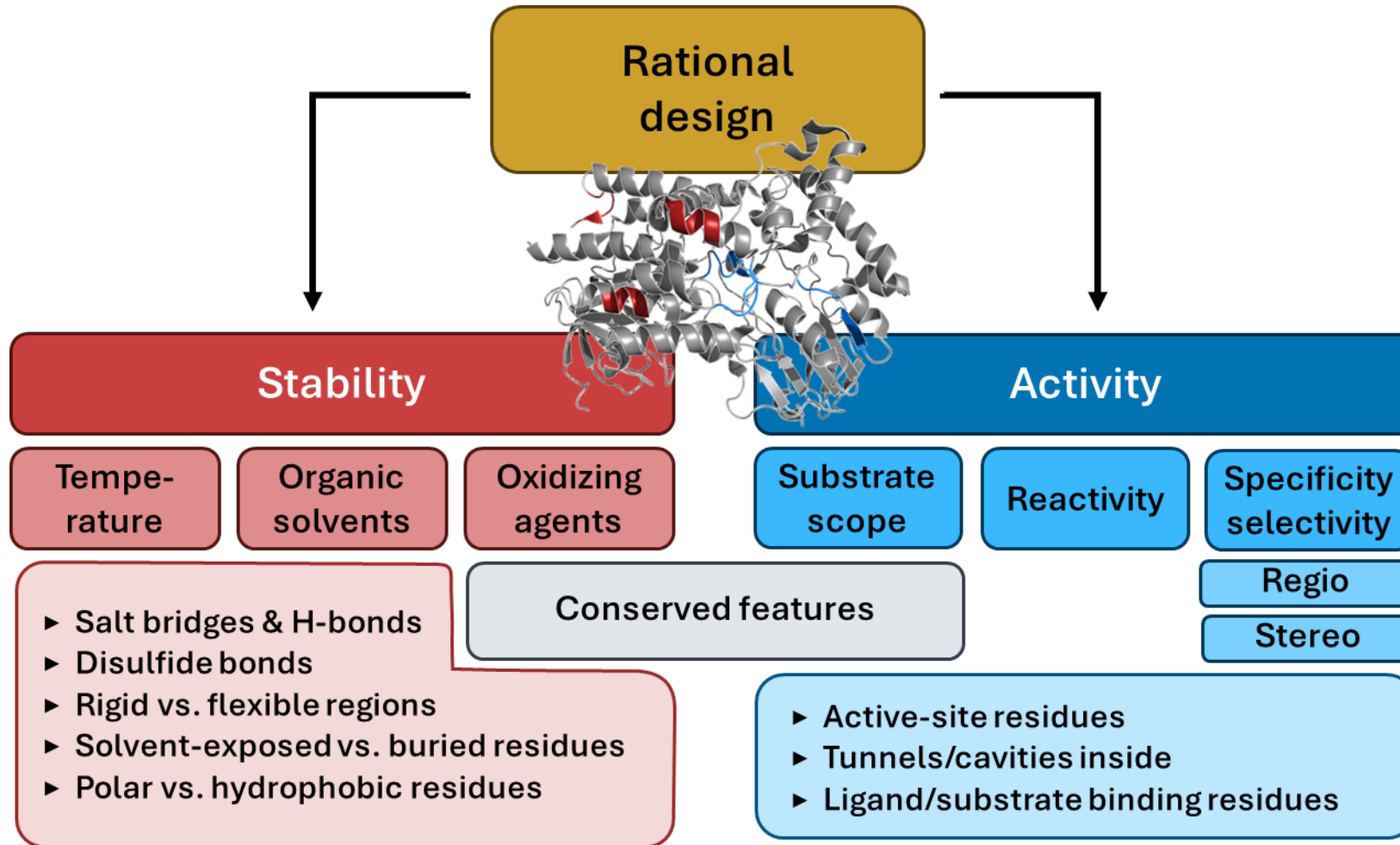
Success story #2

Stabilizing already stabilized enzyme

DhaA115: T_m = 73.3 °C (previously by **FireProt**)

1. Introduction of disulfide bridges
   ► no increase in T_m

2. Automated platforms **FireProt** and **PROSS**
   ► **FireProt** : best T_m = 77.0 °C
   ► **PROSS**   : best T_m = 78.4 °C

3. Further stability increase by **manual curation**
   ► **FireProt** : best T_m = 79.3 °C
   ► **PROSS**   : best T_m = 80.9 °C

4. Automated curation by **machine learning**
   ► **MutCompute** : T_m = 81.7 °C

Kunka, et al. *ACS Catal* 2023, link   35
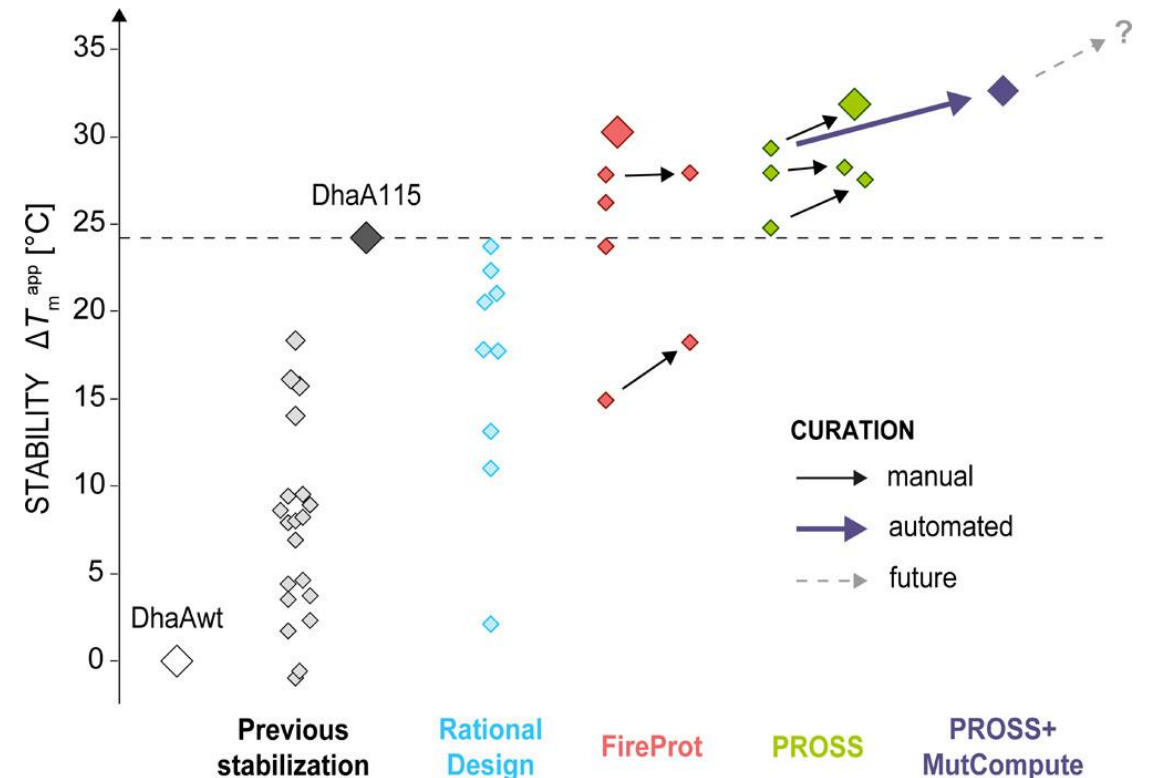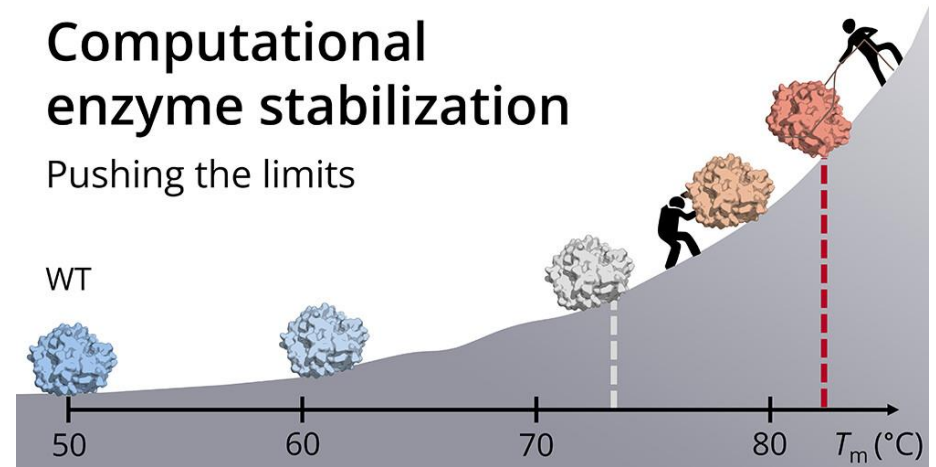
Targets for Machine Learning

Designing molecules

►Design mutations/protein variants

►Design drugs/ligands to bind proteins

Predictions

►Structure prediction (AlphaFold2, …)

►Sequence from structure
(find binding proteins, RF Diffusion)

►Function from sequence

►AI in Life Sciences **Lecture 6**

Design mutations
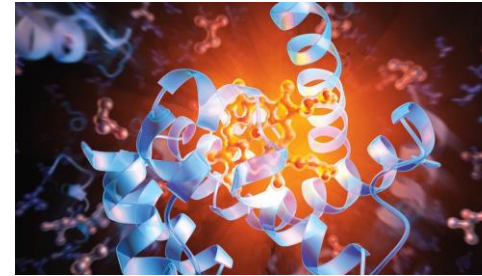
Design drugs/ligands
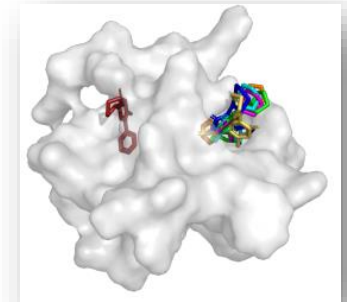
Structure from Sequence
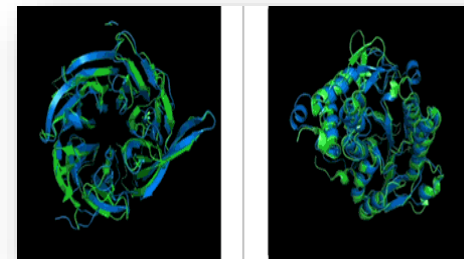
Sequence from Structure

Predict function

LOSCHMIDT LABORATORIES
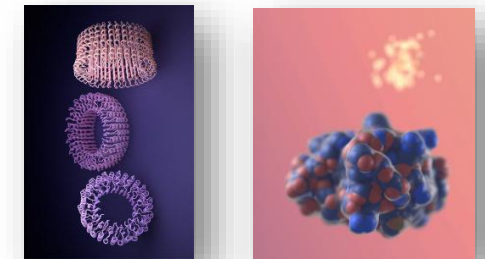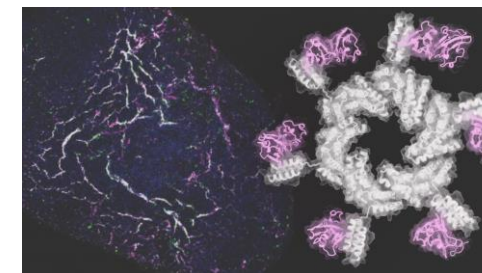
AI in Biology, Chemistry, and Bioengineering Bi9680En

►Období: podzim

►Rozsah: přednáška 2 hodiny/týden

►Vyučující: Dr. Stanislav Mazurenko

►Osnova:

▶ modern bio-challenges: drug design, DNA interpretation, protein engineering

▶ types of AI algorithms and workflow for designing predictors

▶ clustering algorithms, random forests, artificial neural networks

▶ features, databases, and predictors used in applications

Advertisement

Nobel Prize in Chemistry 2024

THE NOBEL PRIZE IN CHEMISTRY 2024

David Baker
"for computational protein design"

Demis Hassabis

John M. Jumper
"for protein structure prediction"
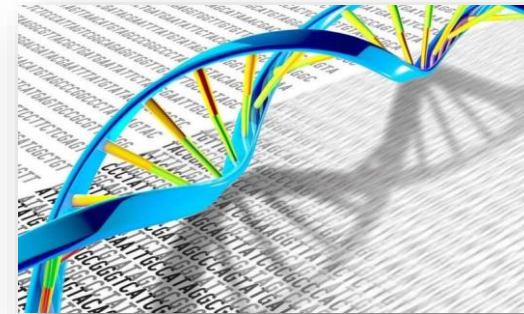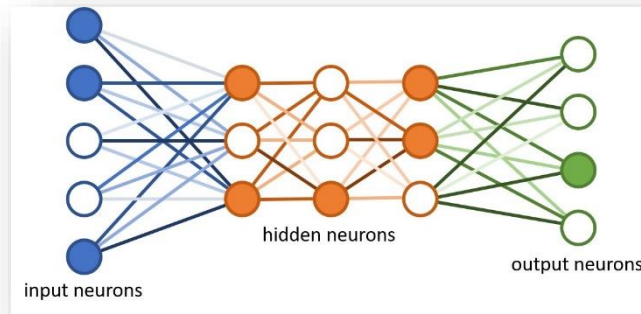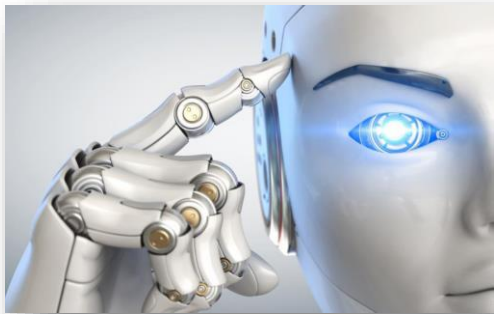
THE ROYAL SWEDISH ACADEMY OF SCIENCES

Illustrations: Niklas Elmehed
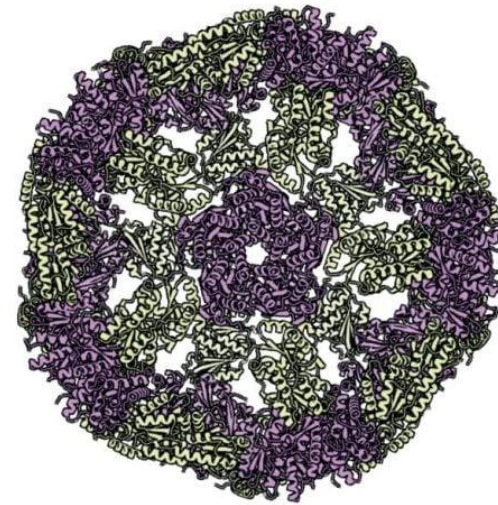
2016: New nanomaterials where up to 120 proteins spontaneously link together.
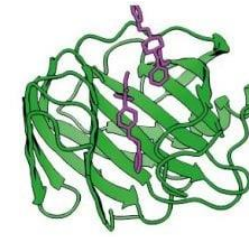
2017: Proteins that bind to an opioid called fentanyl (purple). These could be used to detect fentanyl in the environment.
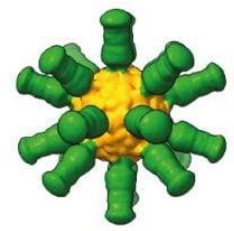
2021: Nanoparticles (yellow) with proteins imitating influenza virus on the surface (green) that can be used as a vaccine for influenza. Successful in animal models.
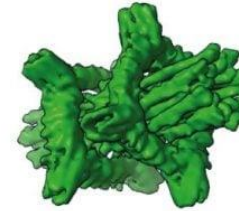
2022: Proteins that function as a type of molecular rotor.
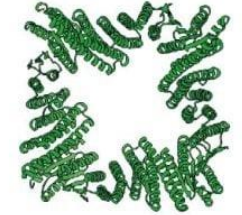
2024: Geometrically shaped proteins that can change their shape due to external influences. Could be used for producing tiny sensors.

DeepMind

AlphaFold

A score above 90 is considered roughly equivalent to the experimentally determined structure

AlphaFold 2

AlphaFold

Global distance test (GDT_TS; average)

Year: 2006 2008 2010 2012 2014 2016 2018 2020
CASP: 7 8 9 10 11 12 13 14
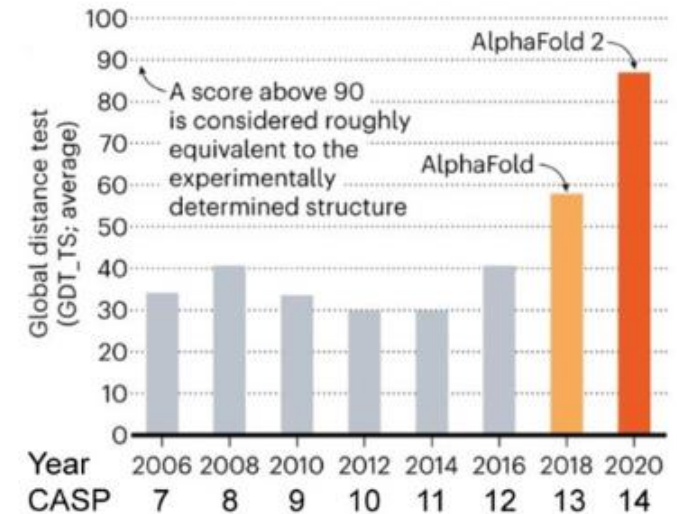
Main strategies
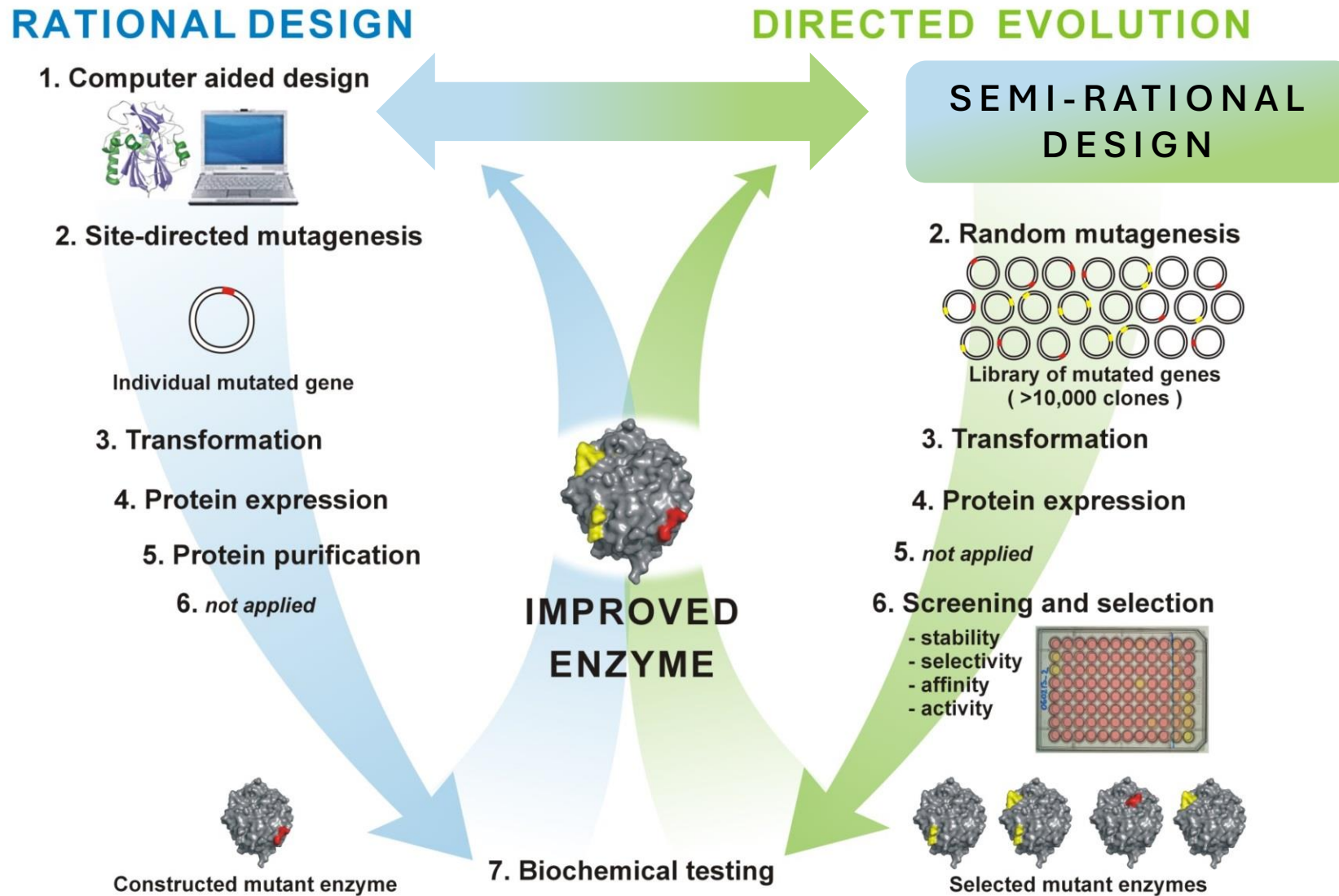
Success story #3: Degrading a toxic pollutant

► conversion of 1,2,3-trichloropropane
    by DhaA from *Rhodococcus erythropolis* Y2
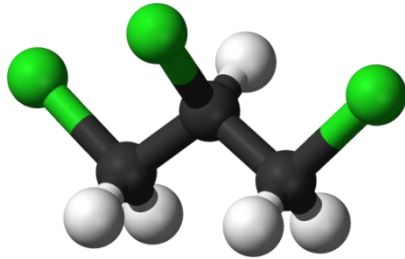
First round: Directed evolution

► conversion of 1,2,3-trichloropropane
   by DhaA from *Rhodococcus erythropolis* Y2

► **Directed Evolution** - importance of access pathways
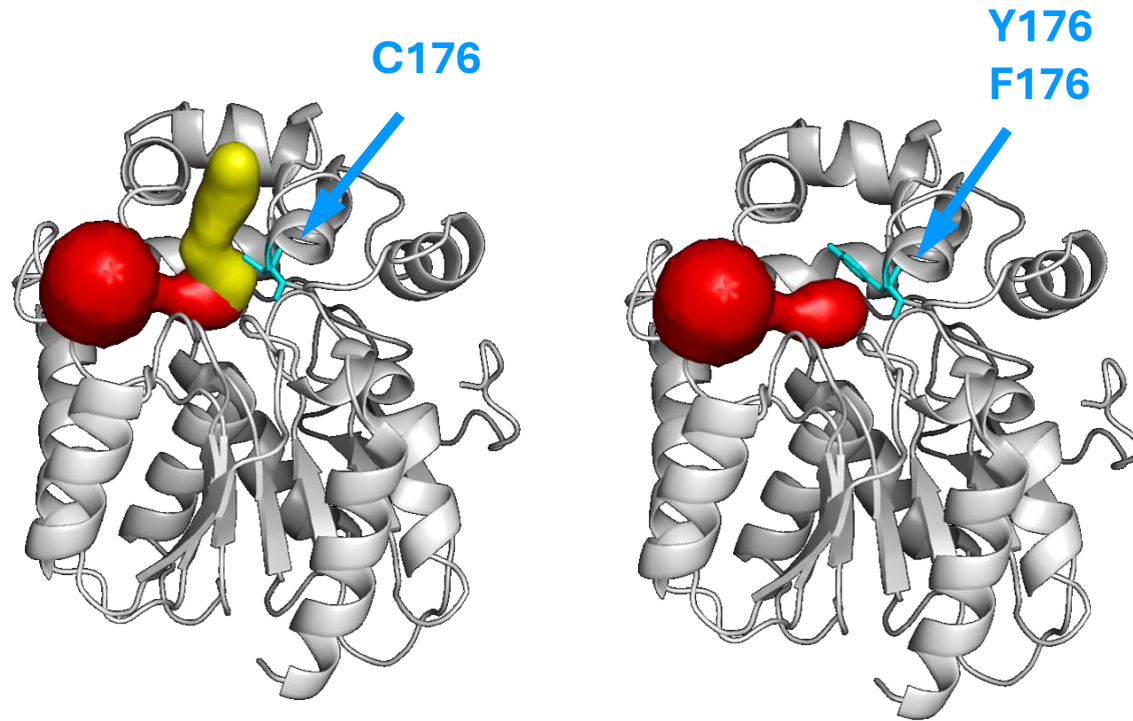
Variant	k_{cat} (s^{-1})
wt	0.08
C176Y+Y273F[1]	0.28
G3D+**C176F**[2]	0.32

[1] Bosma, et al. ***AEM*** 2002, link,
[2] Gray et al. ***Adv. Synth. Catal.*** 2001, link

Second round guided by structural insights

▶ conversion of 1,2,3-trichloropropane
   by DhaA from *Rhodococcus erythropolis* Y2

▶ **Directed Evolution** - importance of access pathways

▶ **Semi-rational Design** - hot spots in access tunnels

▶ library of **5,300** clones screened
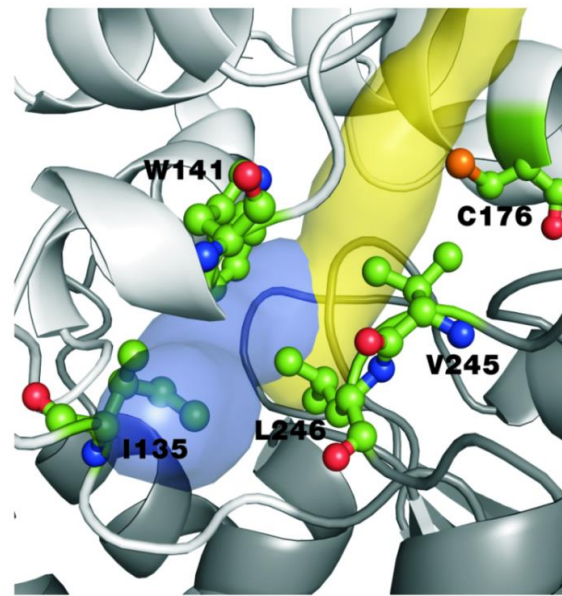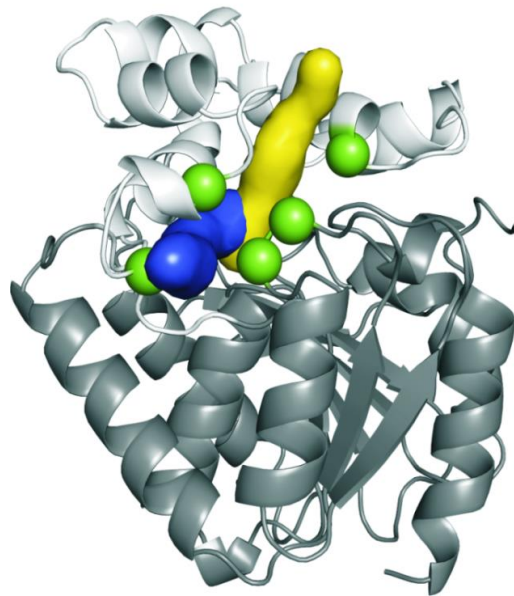
Pavlova, et al. ***Nat Chem Biol*** 2009, link

Results
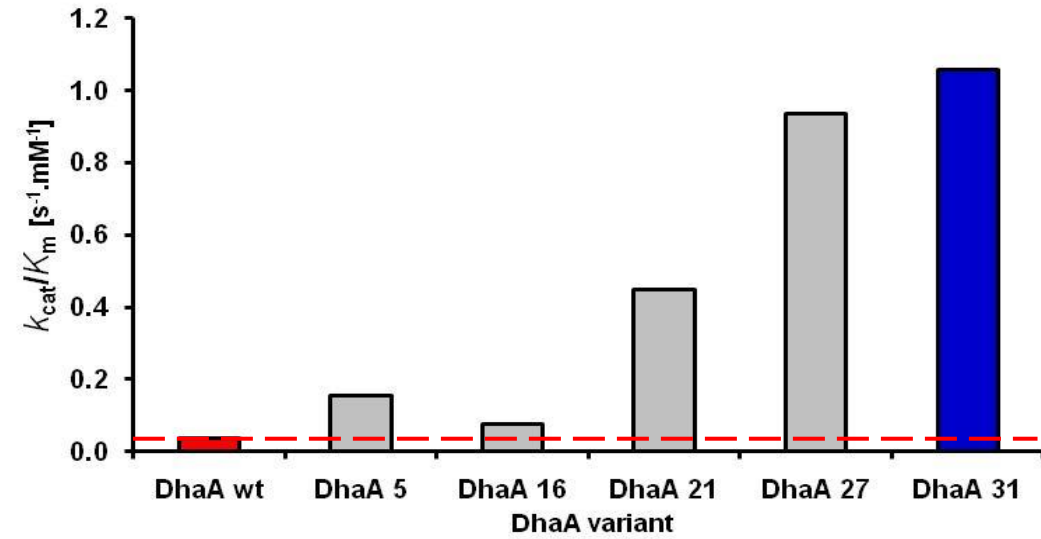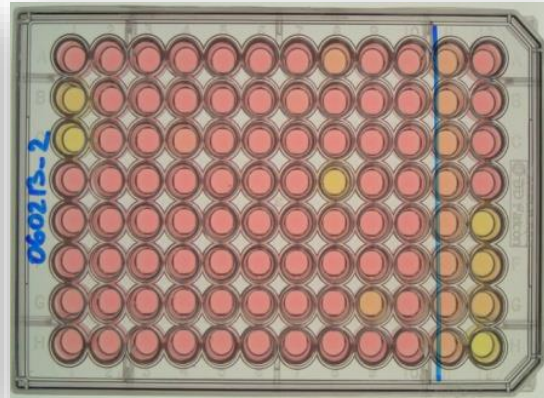
Accessible solvent
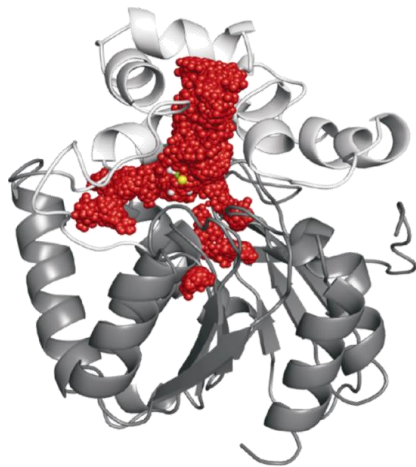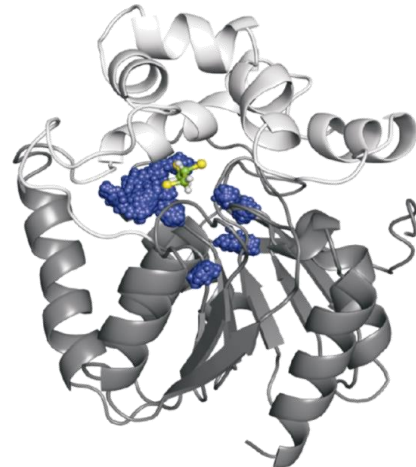
DhaA WT   **DhaA 31**

Active site
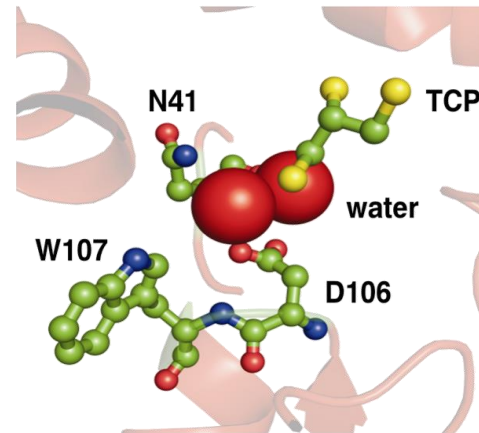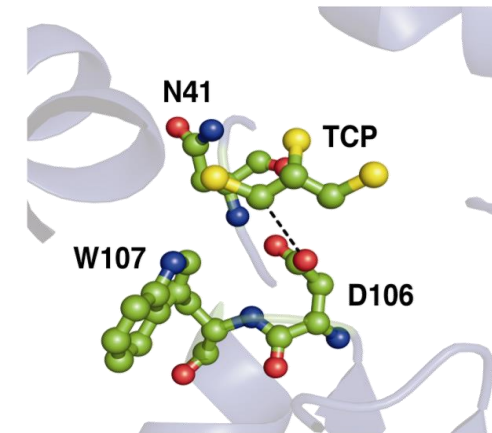
DhaA WT   **DhaA 31**

Pavlova, et al. ***Nat Chem Biol*** 2009, link

Proteinové inženýrství Bi7410

► Období: jaro

► Rozsah: přednáška 2 hodiny/týden

► Vyučující: Mgr. Michal Vašina, Ph.D. ,
doc. Mgr. David Bednář, Ph.D.

► **Osnova:**

  ► strukturně-funkční vztahy proteinů

  ► metody exprese a purifikace rekombinantních proteinů

  ► metody strukturní a funkční analýzy proteinů

  ► racionální design, semi-racionální design a řízená evoluce

  ► příklady využití proteinového inženýrství
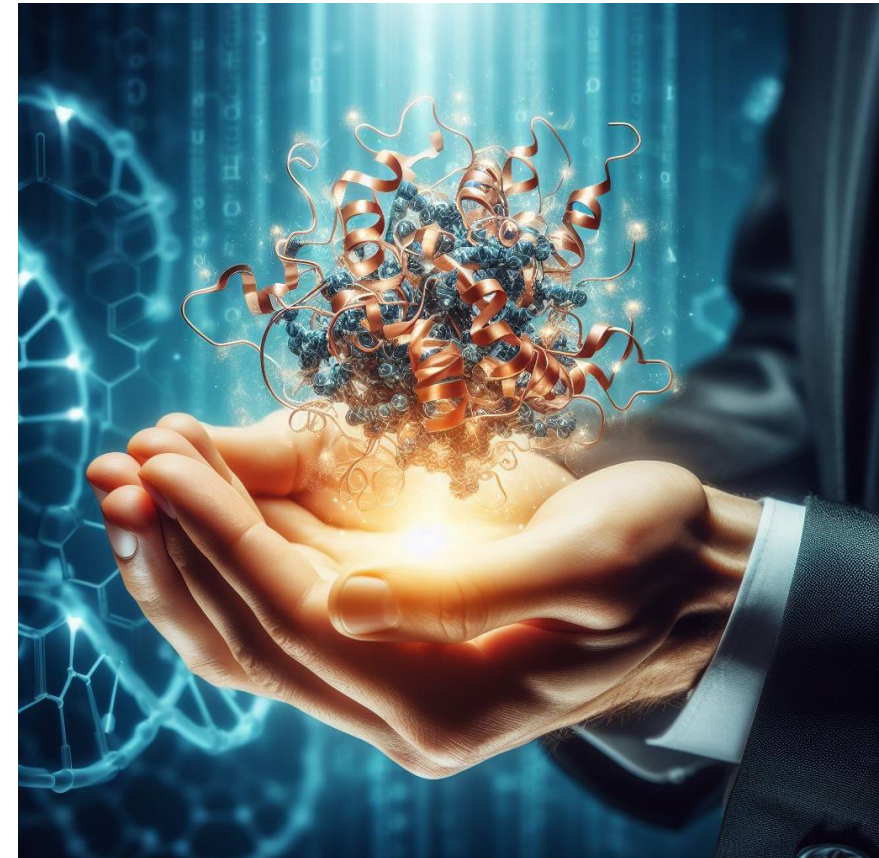
Multidisciplinary in protein research
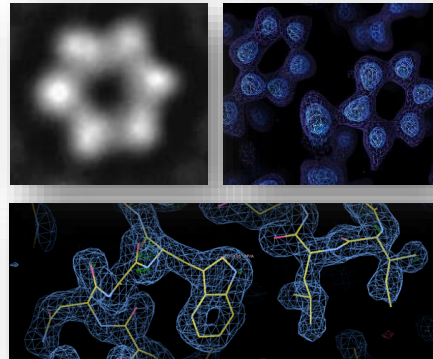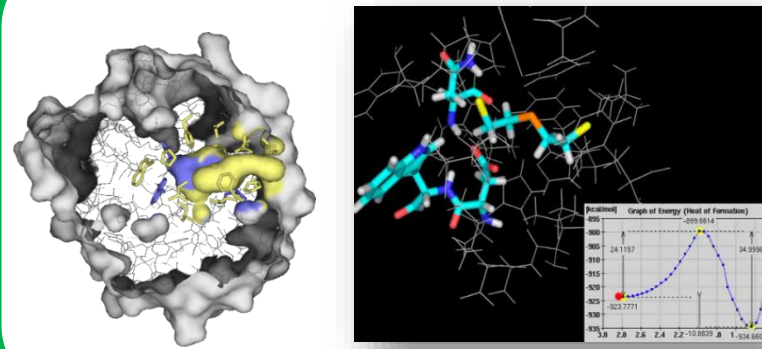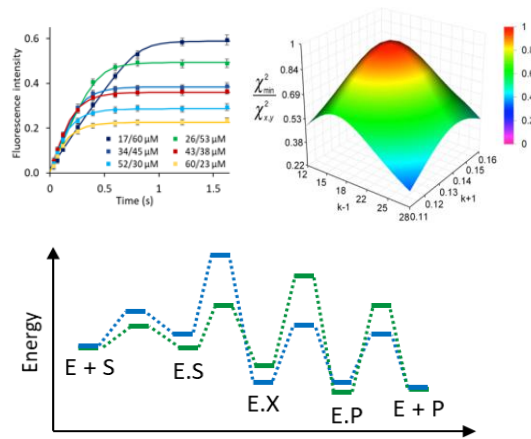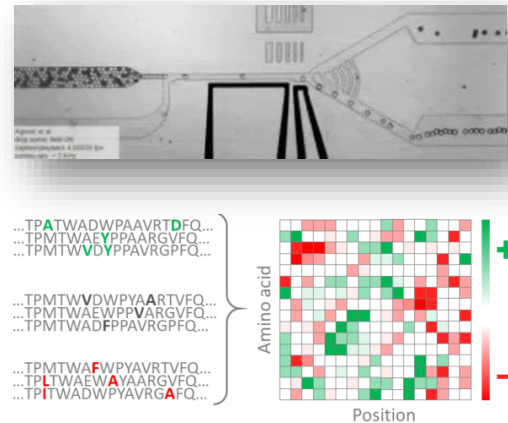
Combine multiple strategies

Protein Discovery
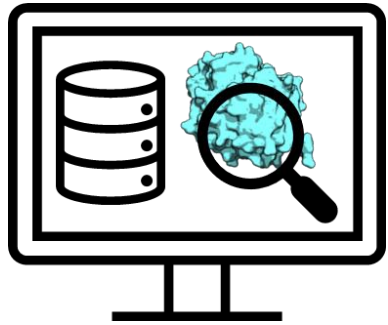
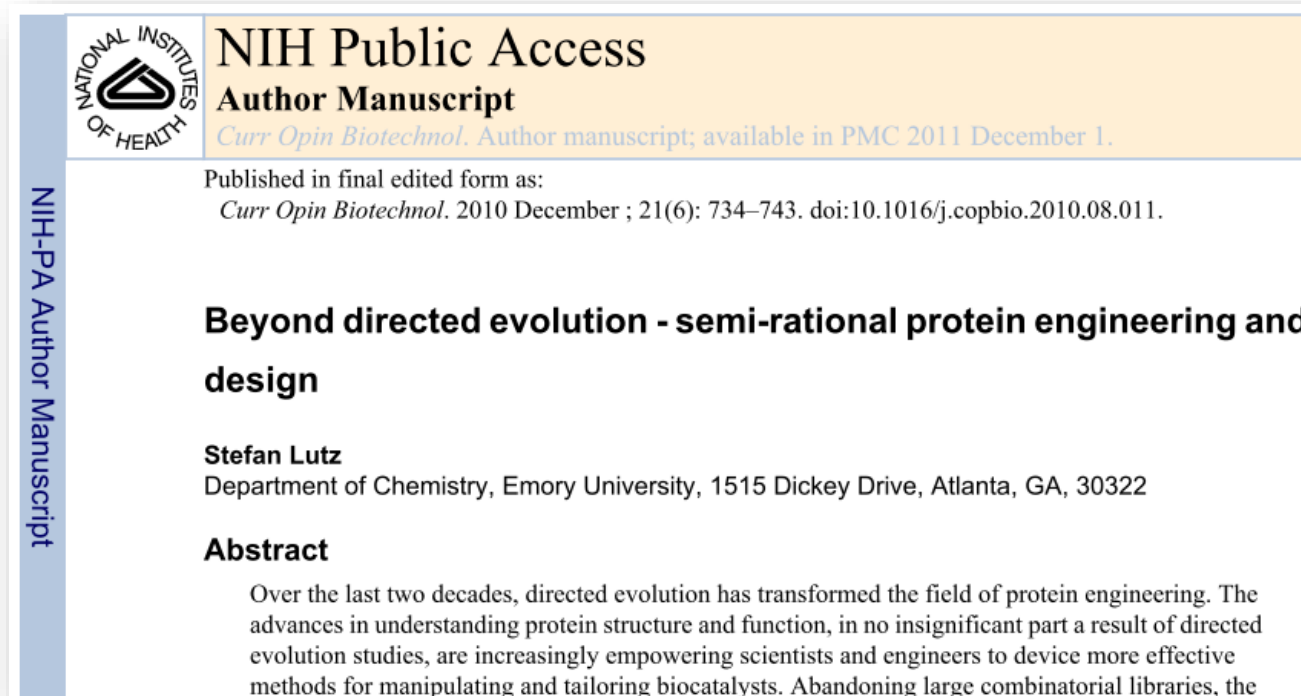**Directed Evolution**
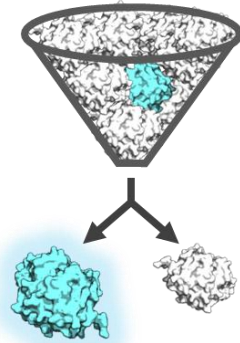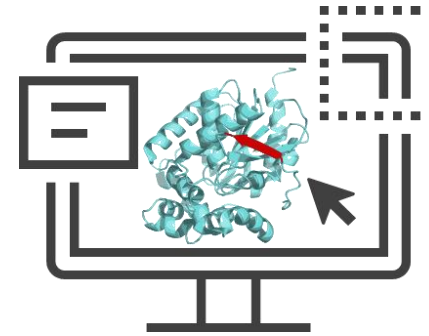
Rational Design

Machine Learning
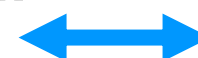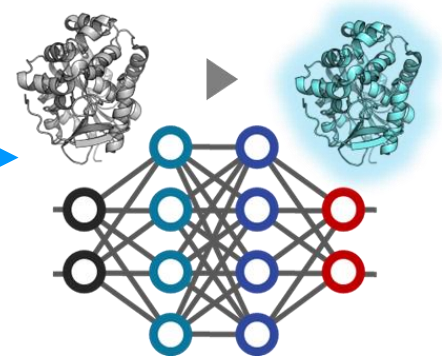
Reading

►Lutz, S. 2010: Beyond directed evolution - semi-rational protein engineering and design. *Curr Opin Biotechnol*. 21(6): 734–743 (link)

►Computational enzyme redesign and Computational de novo enzyme design (page 5-7)