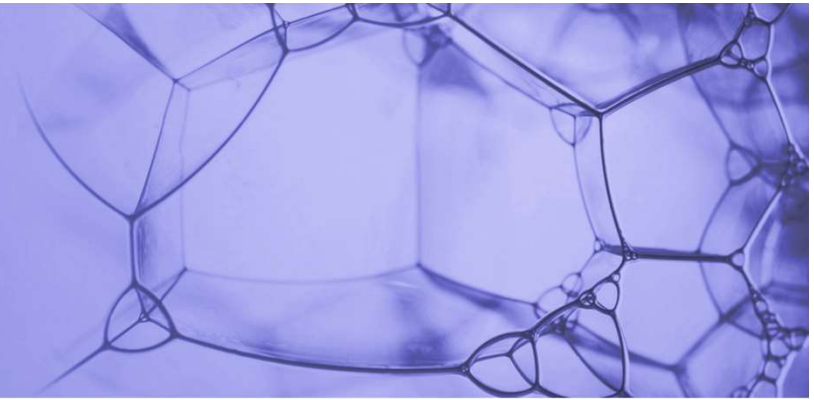


LOSCHMIDT
LABORATORIES



Synthetic protein biology I.

Introduction to protein engineering & computational (*de novo*) protein design

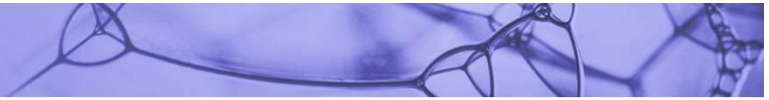
Modifying protein structure and function for biotechnology,
biomedicine and basic research

Dr. Martin Marek
Loschmidt Laboratories
Faculty of Science, MUNI
Kamenice 5, bld. A13, room 332
martin.marek@recetox.muni.cz



What will we talk about

- **Introduction to protein engineering and design**
 - definition, goals and applications
- **Rational protein design (knowledge-based strategies)**
 - concepts, methodology, limitations, success stories
- **Directed evolution (lab-based brute force engineering)**
 - strategies, methodology, disadvantages, success stories
- **Integrative (combined) approaches**
 - the best of both approaches, beneficial synergy, examples
- **Selection and screening technologies**
 - classical versus emerging technologies, unmet challenges



Introduction to protein engineering

Concepts
Methods
Applications

What protein engineering is

- Protein engineering



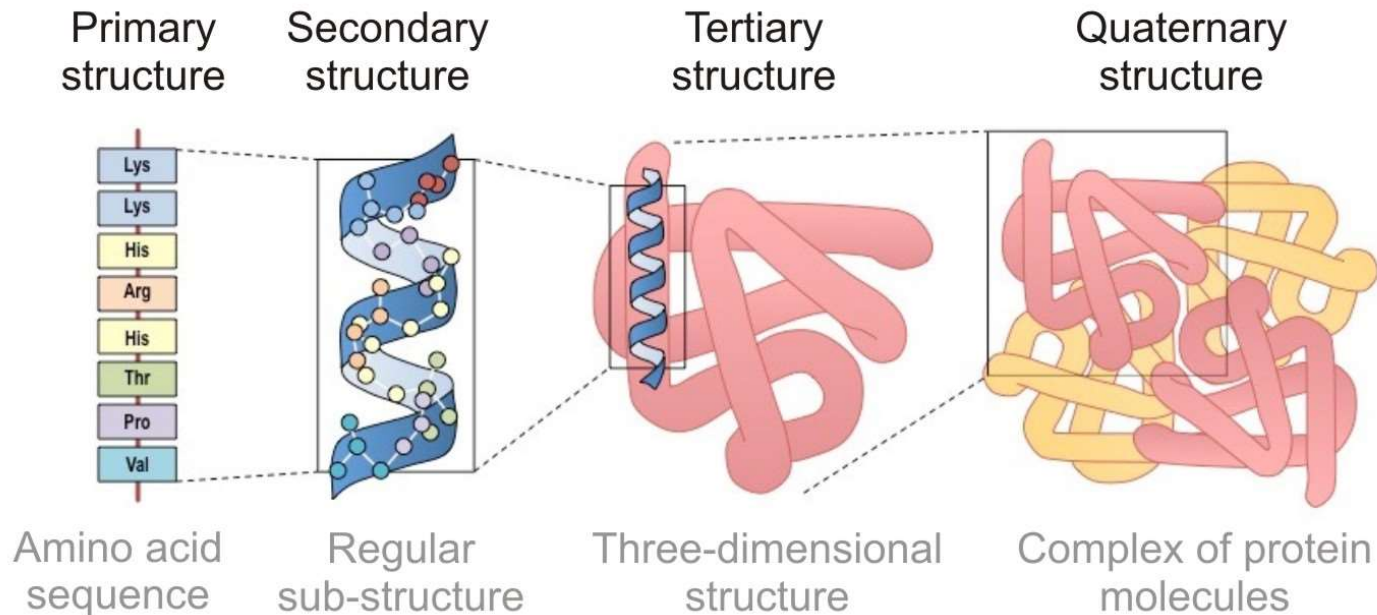
The activities to modify the structure and molecular function of a target protein so that it acquires new specific properties (stability, substrate specificity, enantioselectivity etc.)

- Genetic engineering

The alteration of the genome of a target organism by laboratory techniques

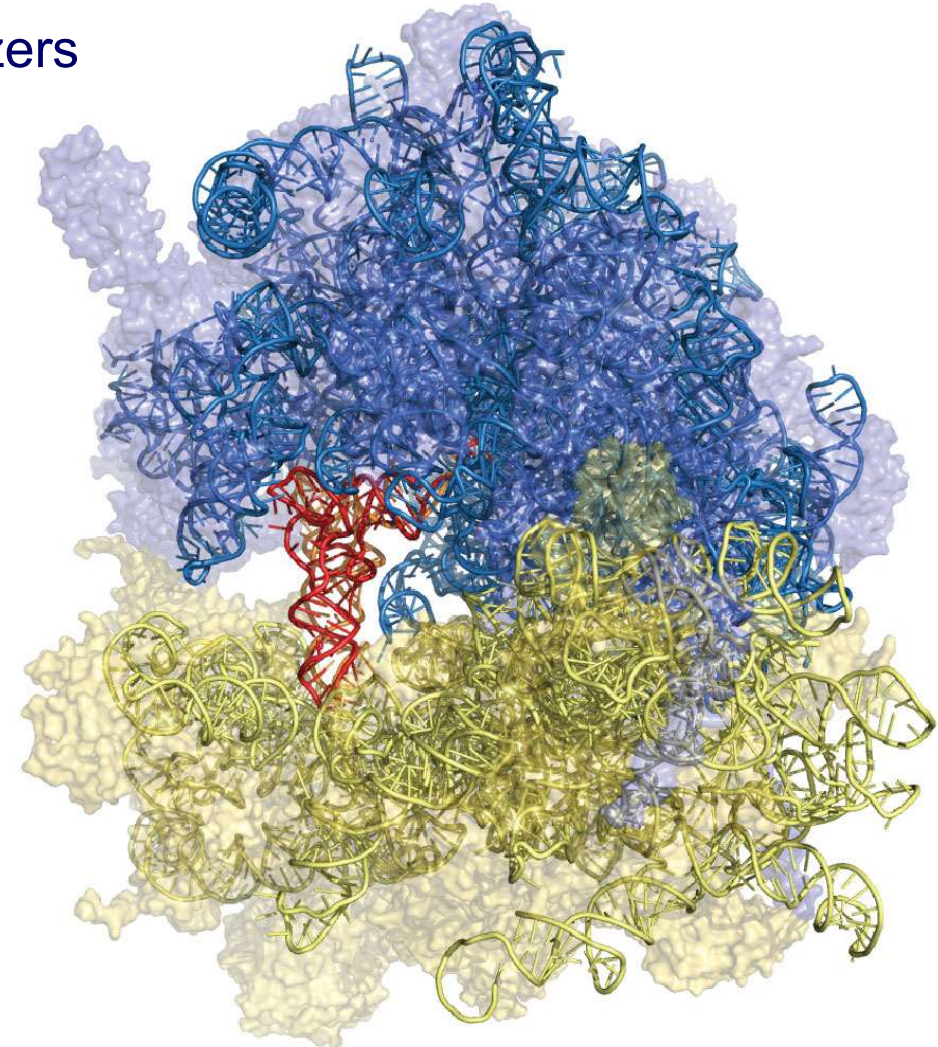
The anatomy of protein structure

- Proteins are an important class of biological macromolecules which are polymers of amino acids
- Biochemists have distinguished several levels of structural organization of proteins



Protein synthesis: the ribosome

- The ribosomes are protein synthesizers of the cell
- Made up of rRNAs and distinct ribosomal proteins
- Arranged into two pieces:
 - Small ribosomal subunit
 - Large ribosomal subunit



Science

[Contents](#) ▾
 [News](#) ▾
 [Careers](#) ▾
 [Journals](#) ▾

SHARE

RESEARCH ARTICLE



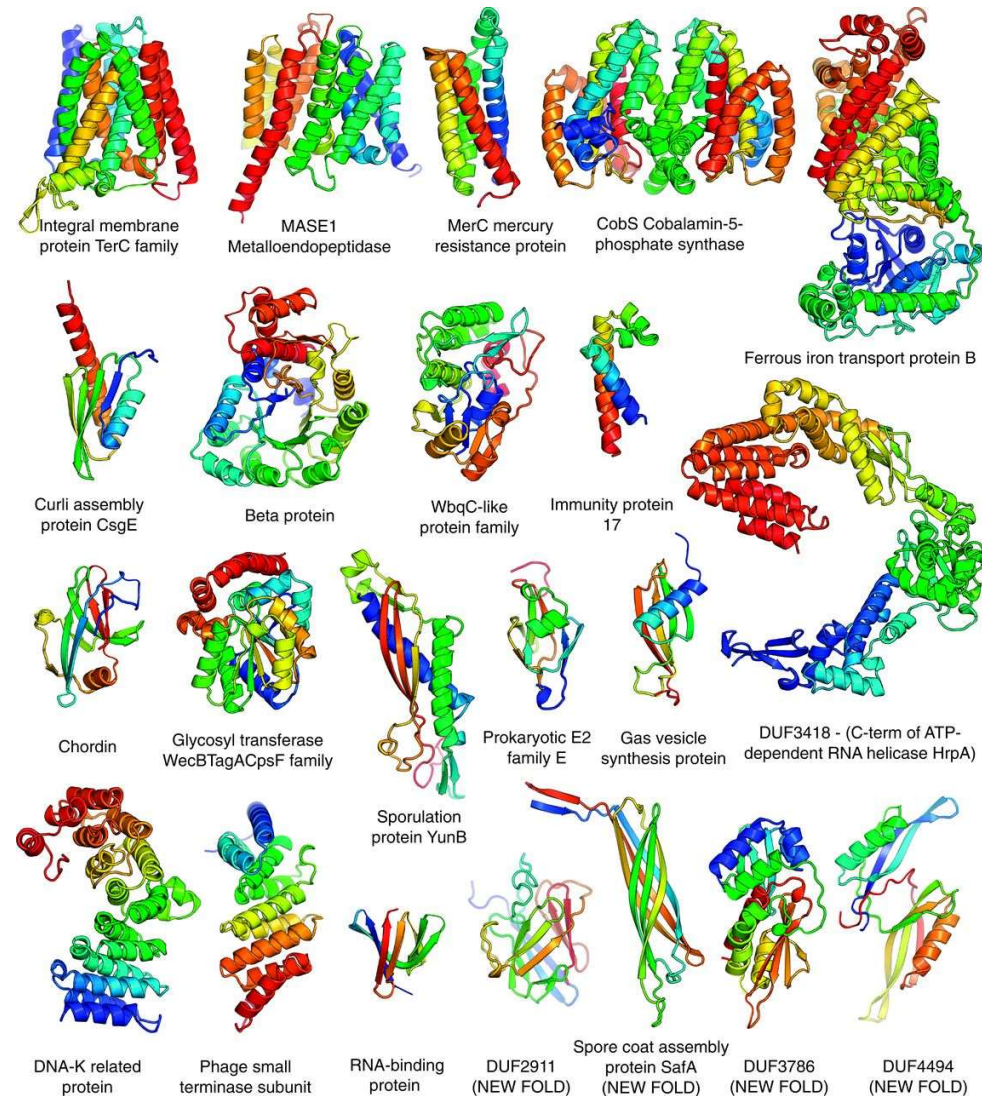
The Structure of the Eukaryotic Ribosome at 3.0 Å Resolution

Adam Ben-Shem^{*,†}, Nicolas Garreau de Loubresse^{*}, Sergey Melnikov^{*}, Lasse Jenner, Gulnara Yusupova, Marat Yusupov[†]

[†] See all authors and affiliations

Protein variety and functional diversity

- Single-domain proteins
- Multi-domain proteins
- Multi-subunit protein complexes
- DNA-binding proteins
- Protein-RNA complexes
- Sugar-binding protein
- Light-emitting proteins
- Integral membrane proteins
- Intrinsically disordered proteins
- Etc.



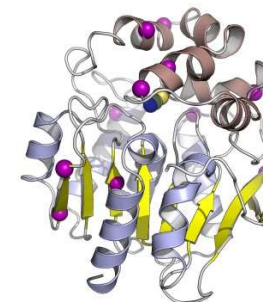
The basic concepts of protein engineering

- The amino acid sequence of a protein affects both its structure and its function
- Thus, the ability to modify the sequence, and hence the structure and activity, of individual proteins in a systematic way, opens up many opportunities, both scientifically and biotechnologically
- Today, large DNA sequences can be synthesised *de novo*, allowing an unprecedented ability to engineer proteins with novel functions
- However, the number of possible proteins (**protein sequence space**) is far too large to test individually, so we need 'navigating system' to find desirable molecular activities and other properties



```

MSSASSNARDEVIAAIIHEEADWVDRTVYPFES
RCIGLSSGAVHYIDEGPDDGGRETLMLHGPNP
TWSFLYRHLVRDLRDEYRCVALDYLGFGLSER
PTDFSYPEDHADVVVEEFIDELGLEDVVLVGH
DWGGPIGFSYAIDHPENVGGLVVMNTWMWPVS
  
```



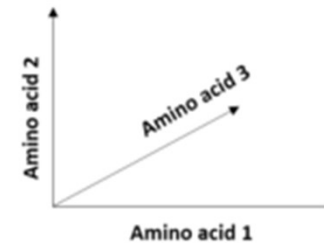
Protein sequence space

All dipeptides in 2D space = 400

Ala	AA	RA	NA	DA	CA	EA	QA	GA	HA	SA	LA	KA	MA	FA	PA	SA	TA	WA	YA	VA
Arg	AR	RR	NR	DR	CR	ER	QR	GR	HR	SR	LR	KR	MR	FR	PR	SR	TR	WR	YR	VR
Asn	AN	RN	NN	DN	CN	EN	QN	GN	HN	SN	LN	KN	MN	FN	PN	SN	TN	WN	YN	VN
Asp	AD	RD	ND	DD	CD	ED	QD	GD	HD	SD	LD	KD	MD	FD	PD	SD	TD	WD	YD	VD
Cys	AC	RC	NC	DC	CC	EC	QC	GC	HC	SC	LC	KC	MC	FC	PC	SC	TC	WC	YC	VC
Glu	AE	RE	NE	DE	CE	EE	QE	GE	HE	SE	LE	KE	ME	FE	PE	SE	TE	WE	YE	VE
Gln	AQ	RQ	NQ	DQ	CQ	EQ	QQ	GQ	HQ	SQ	LQ	KQ	MQ	FQ	PQ	SQ	TQ	WQ	YQ	VQ
Gly	AG	RG	NG	DG	CG	EG	QG	GG	HG	SG	LG	KG	MG	FG	PG	SG	TG	WG	YG	VG
His	AH	RH	NH	DH	CH	EH	QH	GH	HH	SH	LH	KH	MH	FH	PH	SH	TH	WH	YH	VH
Ile	AI	RI	NI	DI	CI	EI	QI	GI	HI	SI	LI	KI	MI	FI	PI	SI	TI	WI	YI	VI
Leu	AL	RL	NL	DL	CL	EL	QL	GL	HL	SL	LL	KL	ML	FL	PL	SL	TL	WL	YL	VL
Lys	AK	RK	NK	DK	CK	EK	QK	GK	HK	SK	LK	KK	MK	FK	PK	SK	TK	WK	YK	VK
Met	AM	RM	NM	DM	CM	EM	QM	GM	HM	SM	LM	KM	MM	FM	PM	SM	TM	WM	YM	VM
Phe	AF	RF	NF	DF	CF	EF	QF	GF	HF	SF	LF	KF	MF	FF	PF	SF	TF	WF	YF	VF
Pro	AP	RP	NP	DP	CP	EP	QP	GP	HP	SP	LP	KP	MP	FP	PP	SP	TP	WP	YP	VP
Ser	AS	RS	NS	DS	CS	ES	QS	GS	HS	SS	LS	KS	MS	FS	PS	SS	TS	WS	YS	VS
Thr	AT	RT	NT	DT	CT	ET	QT	GT	HT	ST	LT	KT	MT	FT	PT	ST	TT	WT	YT	VT
Trp	AW	RW	NW	DW	CW	EW	QW	GW	HW	SW	LW	KW	MW	FW	PW	SW	TW	WW	YW	VW
Tyr	AY	RY	NY	DY	CY	EY	QY	GY	HY	SY	LY	KY	MY	FY	PY	SY	TY	WY	YV	VY
Val	AV	RV	NV	DV	CV	EV	QV	GV	HV	SV	LV	KV	MV	FV	PV	SV	TV	WV	YV	VV

Ala Arg Asn Asp Cys Glu Gln Gly His Ile Leu Lys Met Phe Pro Ser Thr Trp Tyr Val

All tripeptides in 3D space = 8000



All 10 a.a proteins in 10D space = 10^{13}

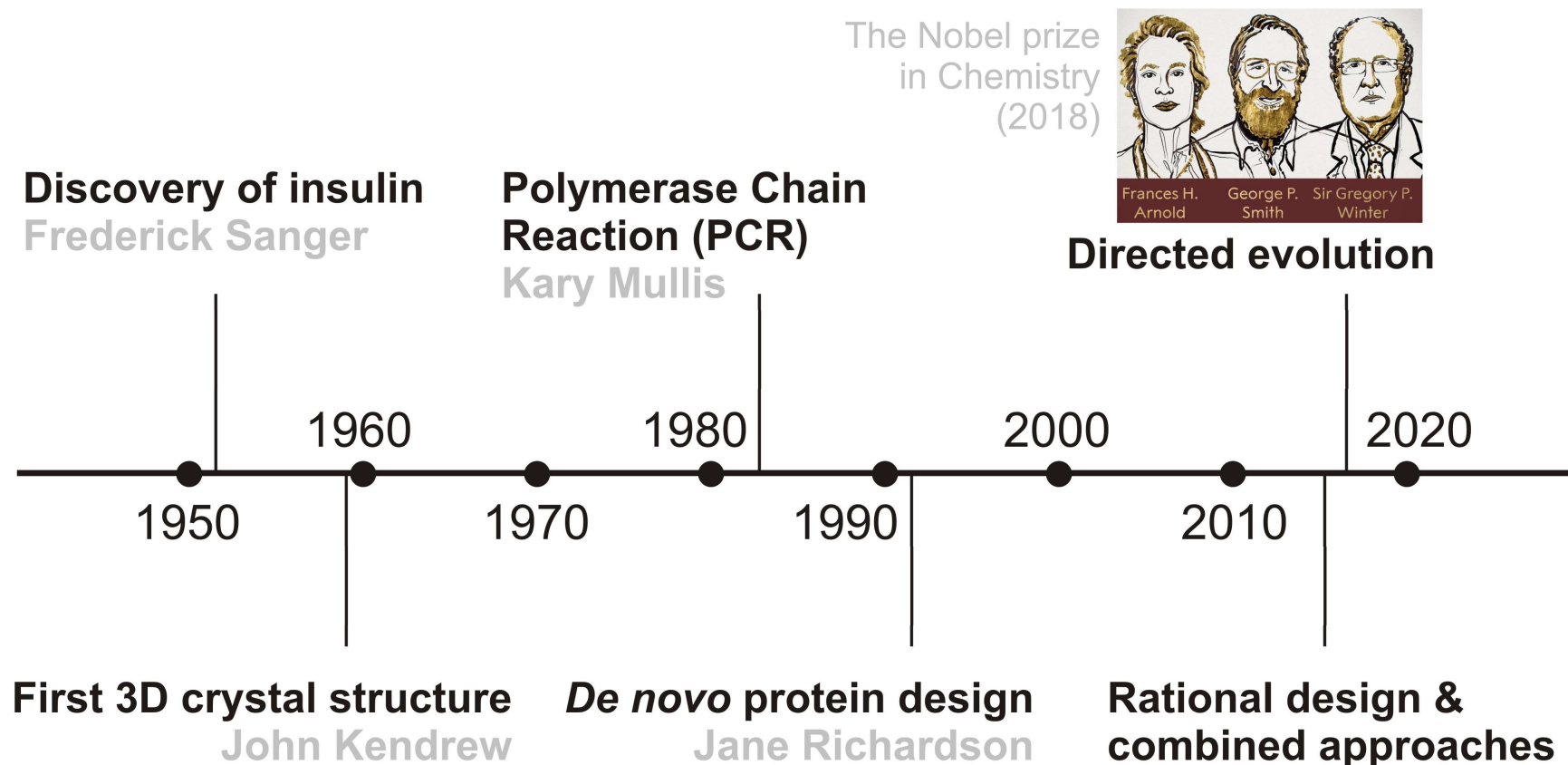
All 50 a.a proteins in 50D space = 10^{65}

All 100 a.a proteins in 100D space = 10^{130}

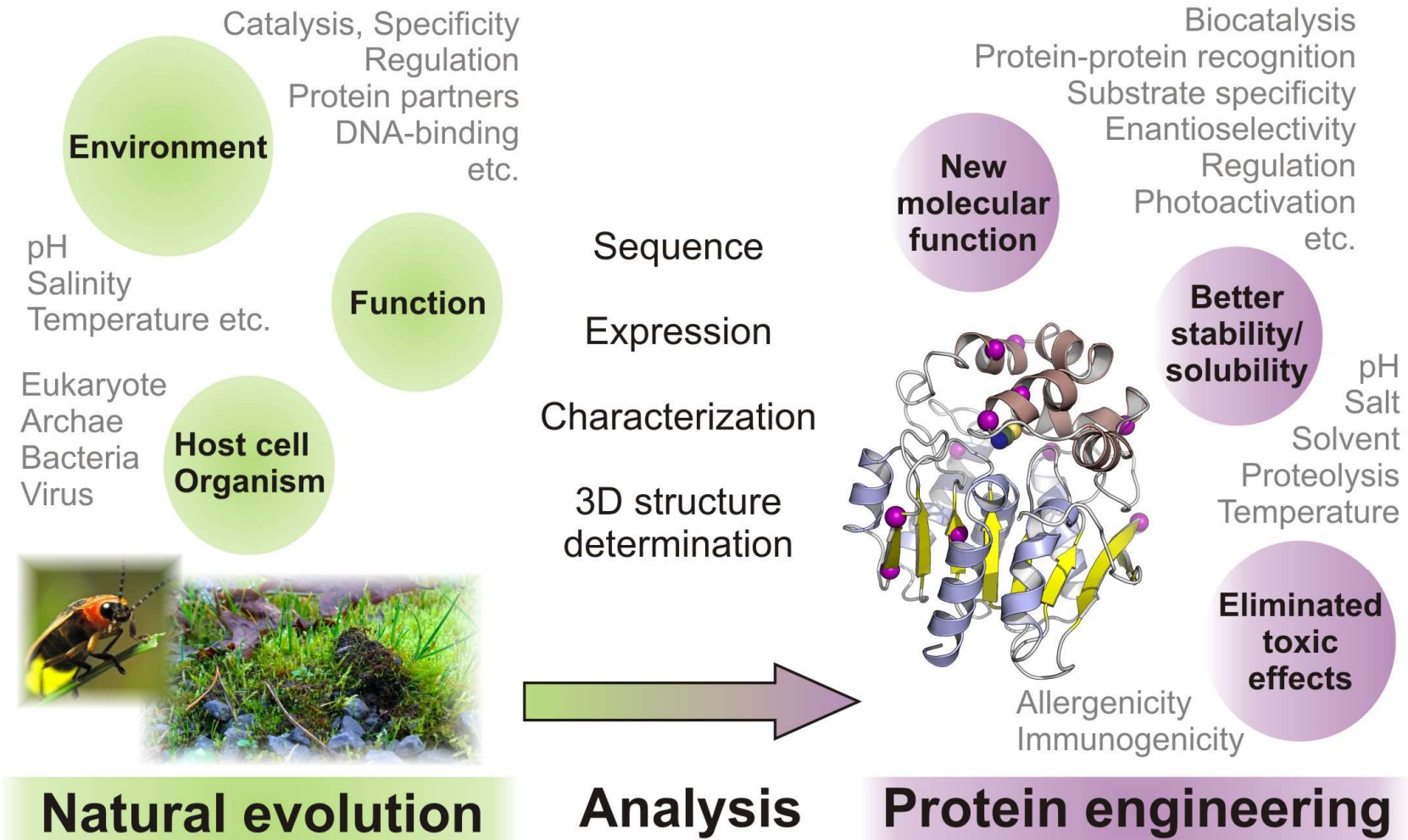
All 300 a.a proteins in 300D space = 10^{325}

- There are 400 possible dipeptides arranged in a 20x20 space but that expands to **10^{130} for even a small protein of 100 amino acids**
- Most sequences in sequence space have no function, leaving relatively small regions that are populated by naturally occurring proteins

Major milestones in protein engineering

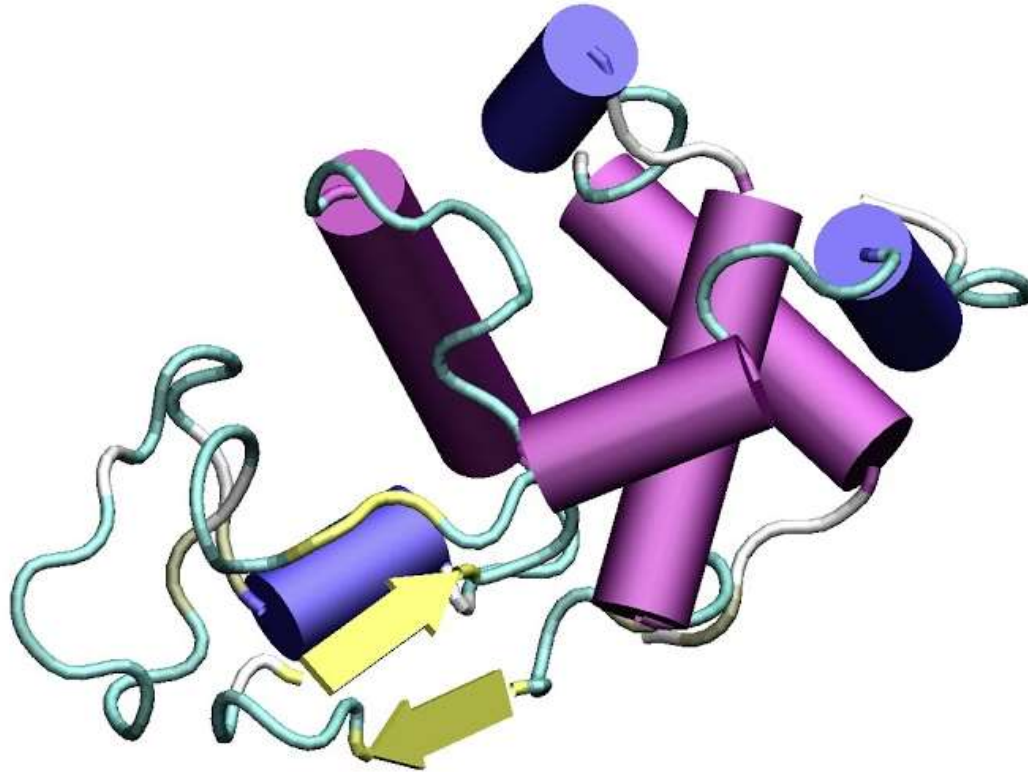


Protein engineering: goals



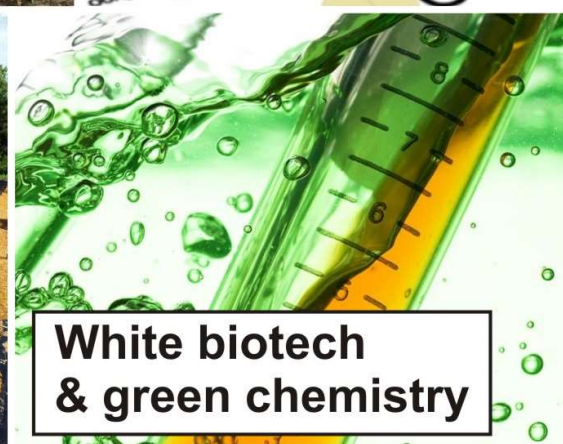
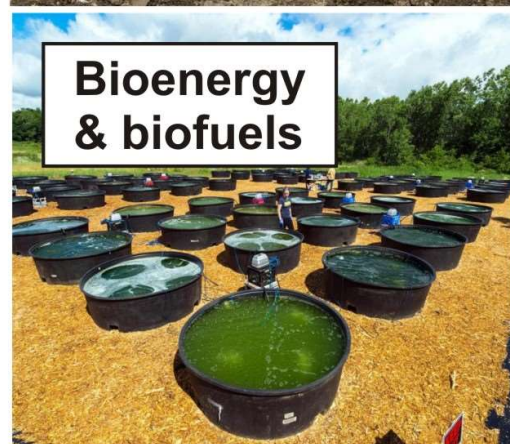
Why is protein engineering important for synthetic biology?

- Protein engineering is creating building blocks for synthetic biology applications

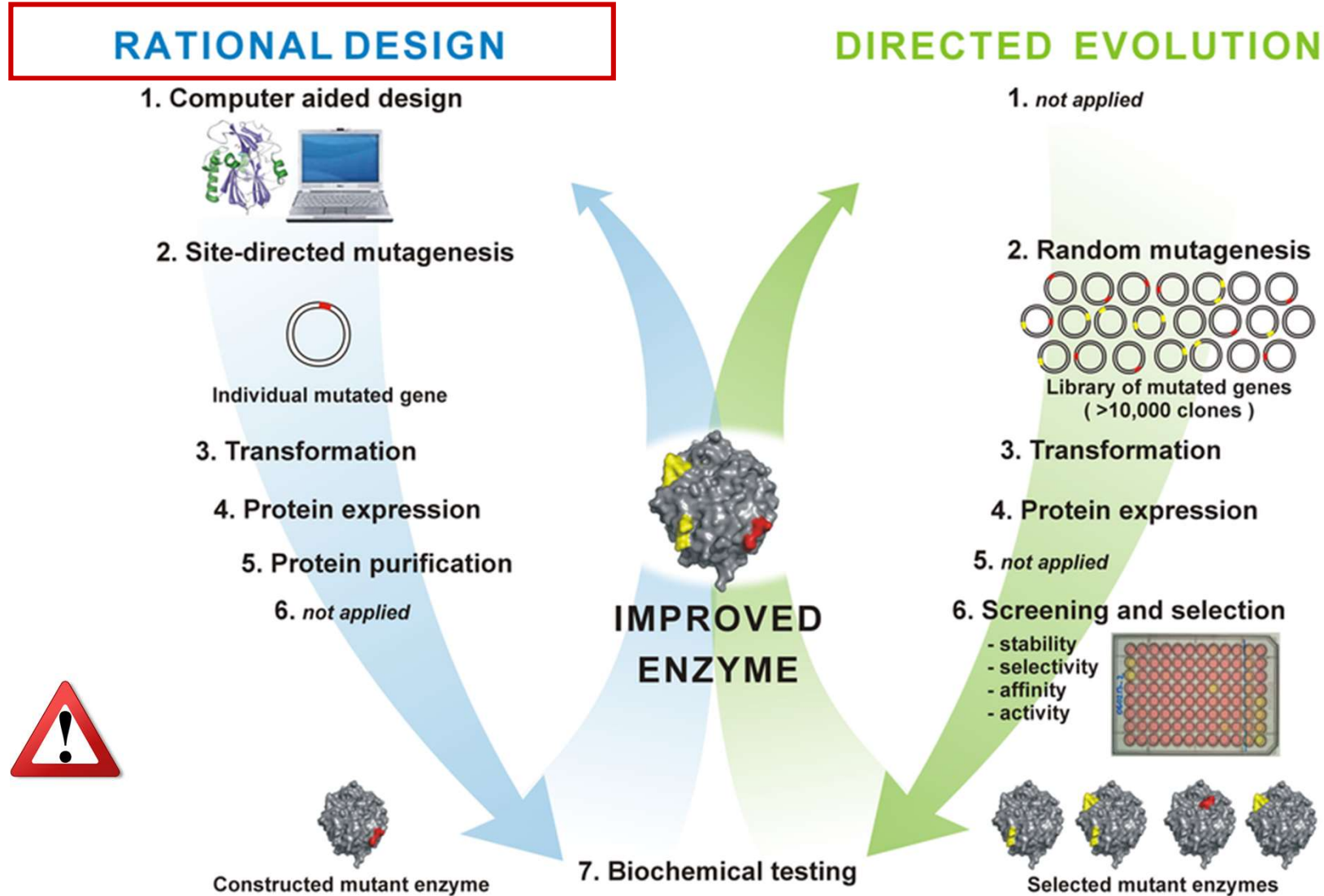


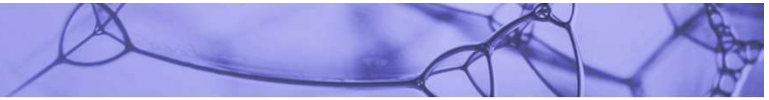
Practical applications of protein engineering

Why should we bother to think about protein engineering?



The two major strategies of protein engineering





Rational protein design

Concepts
Methods
Applications

Rational protein design workflow

- Based on protein knowledge

Evolutionary history, natural variation

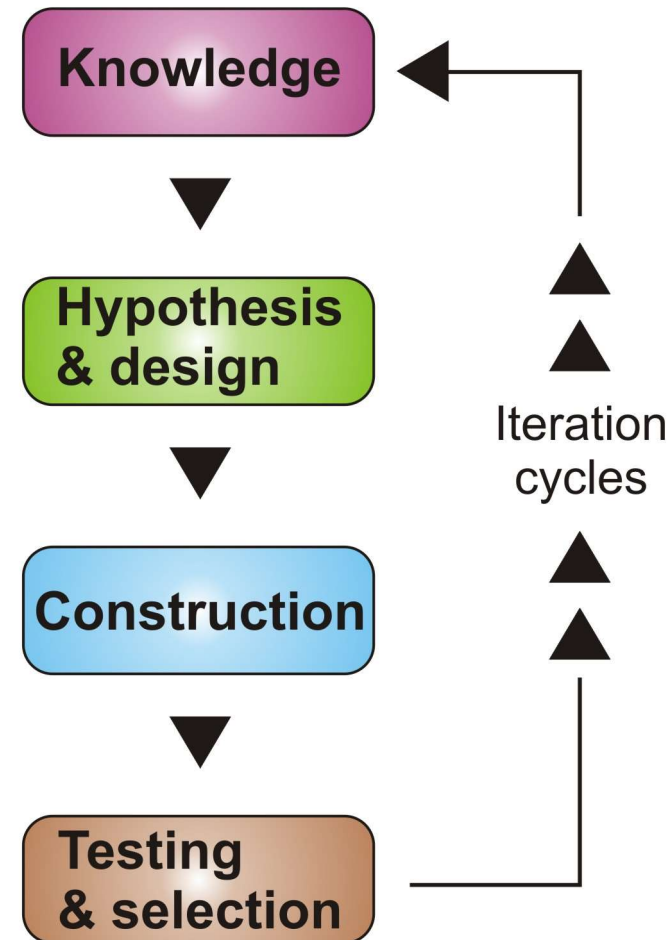
Bioinformatics (MSA, ASR etc.)

Biochemistry & biophysics

Structure & dynamics

Mode of action, molecular mechanism

- Similar to mechanical engineering





Rational protein design

BENEFITS

- Range of available techniques
- Controlled outcome
- Intellectually satisfying
- Increasing computational power

LIMITATIONS

- Deep understanding is essential
 - Evolutionary history, natural variation
 - Structure & dynamics
 - Mode of action, molecular mechanism
- Algorithms are not perfect
- High failure rate
 - Unsuccessful stories are not reported

Rational protein design

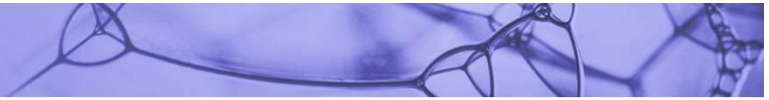
“Low-resolution“ design

- Engineered fusion proteins
- Split proteins
- Chemical modifications
- Antibody-drug conjugates
- Cyclisation
- Disulphides

“High-resolution“ design

- Manipulating existing proteins
- *De novo* protein design
- Computational modelling



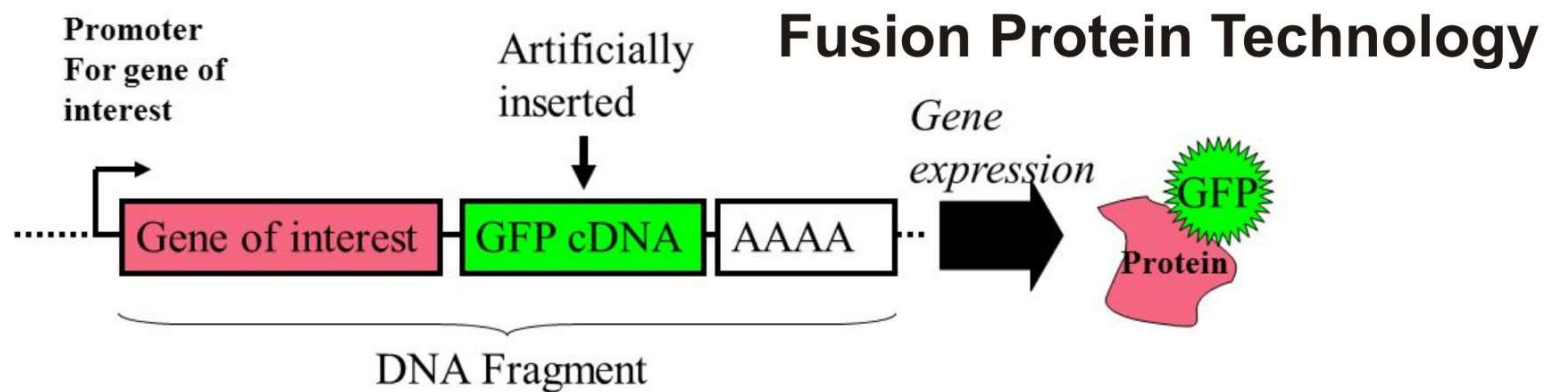


“Low-resolution“ protein design

Concepts
Methods
Applications

Fusion protein technologies

- The fundamental idea is to combine protein units of defined function (domains) to engineer a fusion protein with novel functionality
- Examples include biosensors, chimeric antibodies, signal transduction components, DNA-binding transcription factors, cell biology application, structural biology application, therapeutics etc.

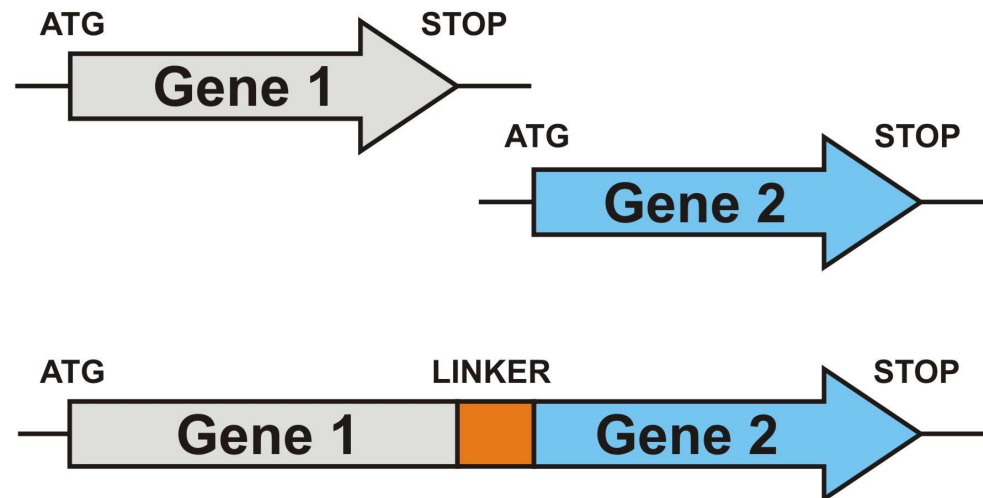


Produces the protein of interest with a GFP unit attached!

If the GFP can be verified (by other means) to not affect the behavior of the protein of interest, we have a way of fluorescently tagging the protein of interest!

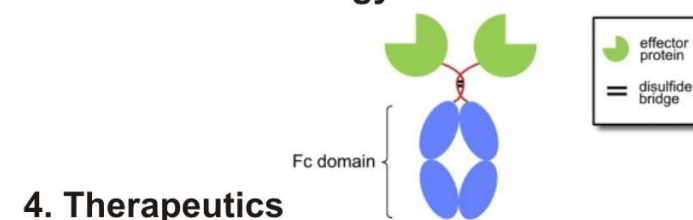
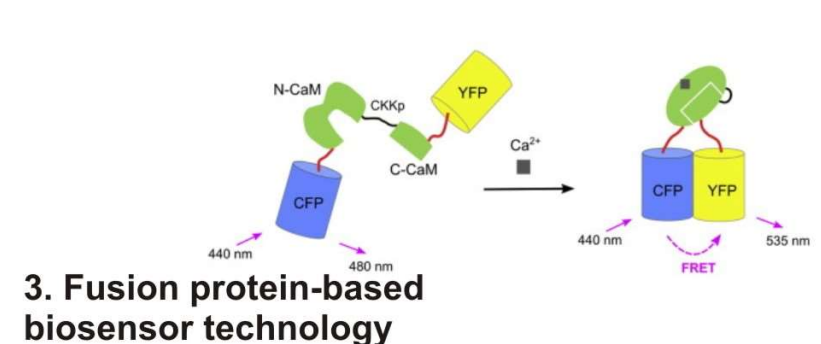
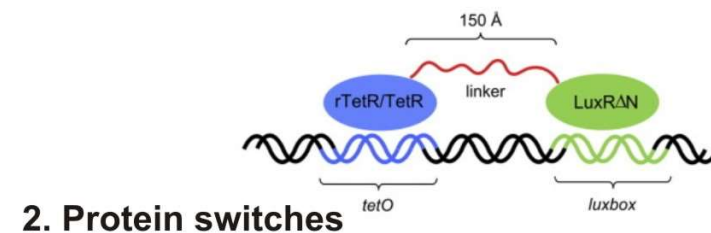
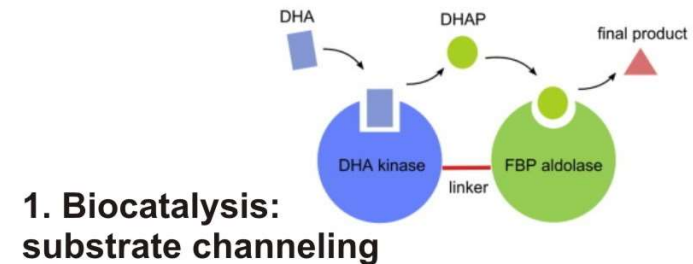
Fusion proteins: considerations

- Remove stop codon of a first gene
- Ligate (fuse) genes together in a frame
- Include linker codons
 - Linker length and flexibility
 - Distance between protein components
 - Ability for proteins rotate relatively to each other
 - Protease resilience
 - Ability for domains to fold



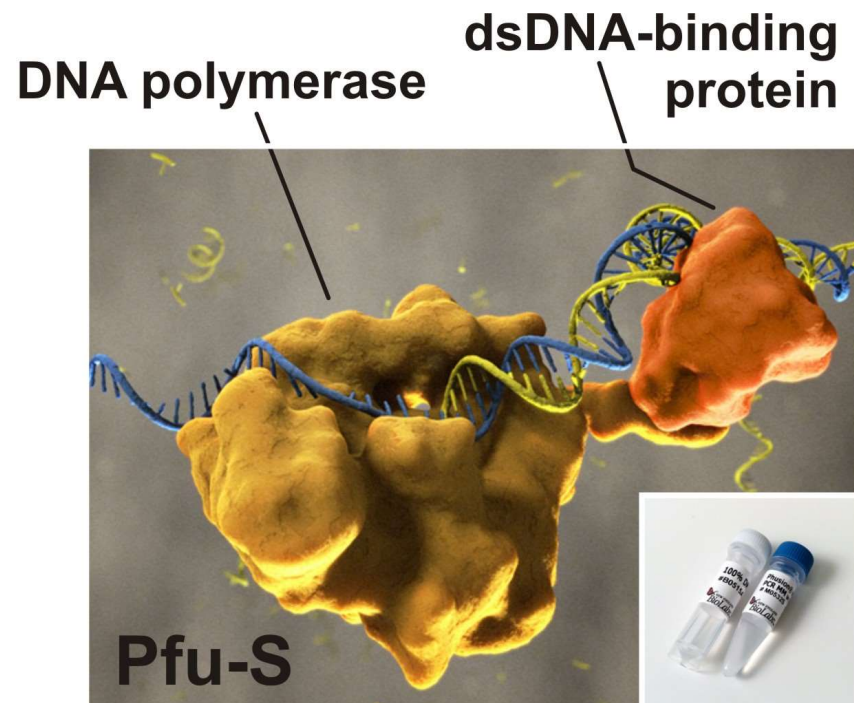
Examples of synthetic fusion protein applications

1. Substrate channeling was achieved by fusing dihydroxyacetone (DHA) kinase with fructose-1,6-biphosphate (FBP) aldolase using a linker peptide QGQGQ.
2. A protein On and Off switch was built by linking the tetracycline repressor protein (TetR or rTetR) and the transcription activator (LuxR Δ N) together. Depending on the presence of anhydrotetracycline, TetR/rTetR undergoes a conformational change and binds to *tetO*, which allows the fused LuxR Δ N to bind the *luxbox* sequence, thereby controlling downstream gene expression.
3. A fusion protein-based biosensor was created for Ca²⁺ detection. The system consisted of a tandem fusion of the cyan fluorescent protein (CFP), N-terminal fragment of calmodulin (CaM), CaM-binding peptide from CaM-dependent kinase (CKKp), C-terminal fragment of CaM and yellow fluorescent protein (YFP). CaM is able to bind Ca²⁺ and thus wraps around the fused CKKp, which places the two fluorophores in close proximity to enhance the FRET efficiency.
4. The dimeric structure of a typical Fc fusion protein. An effector protein is covalently attached to an immunoglobulin Fc domain that provides immune functions and extends half-life.

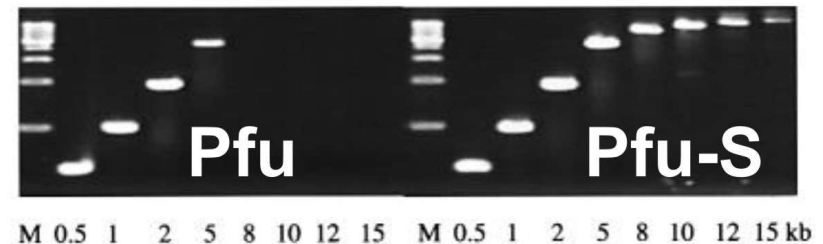


Enhanced DNA polymerase processivity via protein fusion

- The fusion of a heterologous dsDNA binding protein to a polymerase can increase processivity without compromising catalytic activity and enzyme stability.
- Second, polymerase processivity is limiting for the efficiency of PCR, such that the fusion enzymes exhibit profound advantages over unmodified enzymes in PCR applications.
- This technology improved the performance of nucleic acid modifying enzymes.



Wang et al., *Nucleic Acid Res.* 32: 1197-1207 (2004)

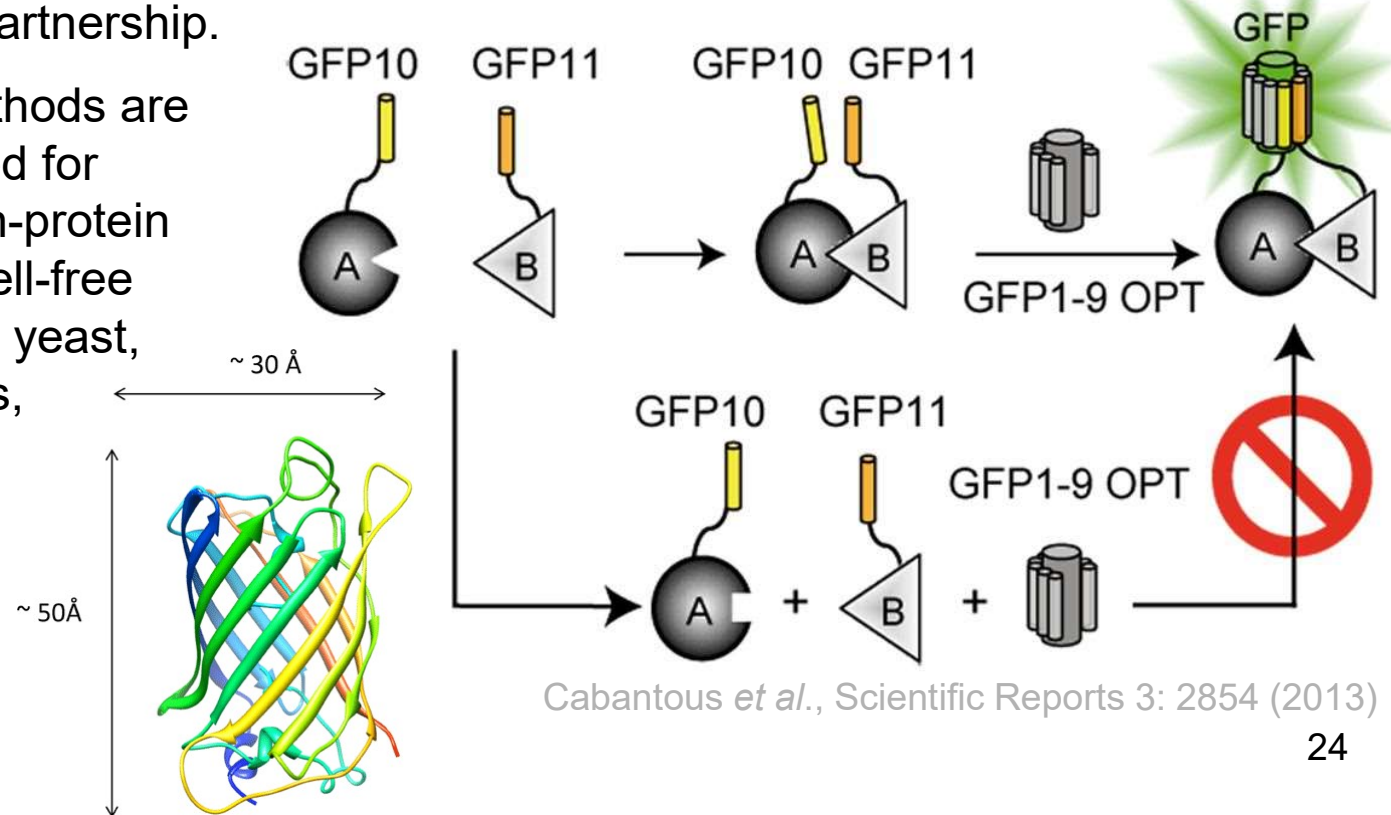


- **10-fold increase in processivity**
- **Improved salt tolerance**
- **Amplification >15 kb templates**

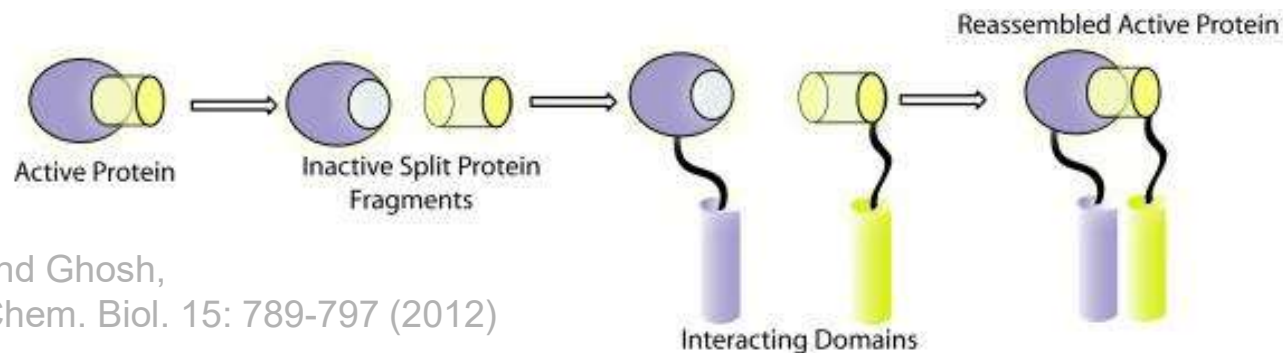
Split protein technology

- A promising approach for deconvoluting the role of macromolecular partnerships is split-protein reassembly, also called protein fragment complementation.
- This approach relies on the appropriate fragmentation of protein reporters, such as the green fluorescent protein or firefly luciferase, which when attached to possible interacting partners can reassemble and regain function, thereby confirming the partnership.

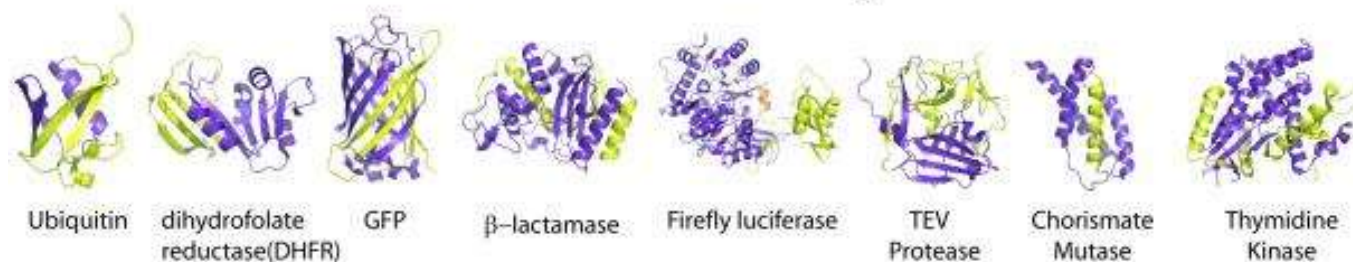
- Split-protein methods are effectively utilized for detecting protein-protein interactions in cell-free systems, *E. coli*, yeast, mammalian cells, plants and live animals.



Split protein technology: basic concept



Shekhavat and Ghosh,
Curr. Opin. Chem. Biol. 15: 789-797 (2012)



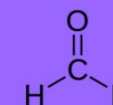
- A generic split-protein system is shown where a functional protein is dissected into two inactive fragments, purple and yellow.
- The attachment of two interacting proteins or protein domains brings the inactive fragments into close proximity and overcomes the entropic cost of fragmentation.
- This leads to the reassembly or complementation of the fragments thus providing a **direct readout for the partnership between the interacting domains**.
- Crystal structures of representative proteins which have been shown to be amenable to interaction dependent reassembly.

Chemical modification of proteins

- **Formaldehyde**

Extensive modification → inactivated protein toxins

Formaldehyde



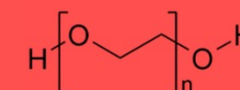
- **Poly-ethylene glycol**

Flexible hydrophilic coat → solubility

Reduced accessibility → Protease resistance, non-antigenicity

Increased size → Serum half-life

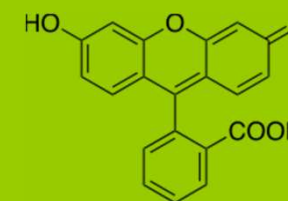
Poly-ethylene glycol



- **Fluorescent probes (Fluorescein, Cy5 etc.)**

Labelling → tracking location in cell, dynamics

Fluorescein



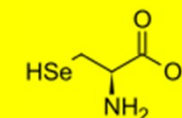
- **Prosthetic catalytic groups**

Modified reactivity → new catalytic properties

- **Antibody-drug conjugates (ADCs)**

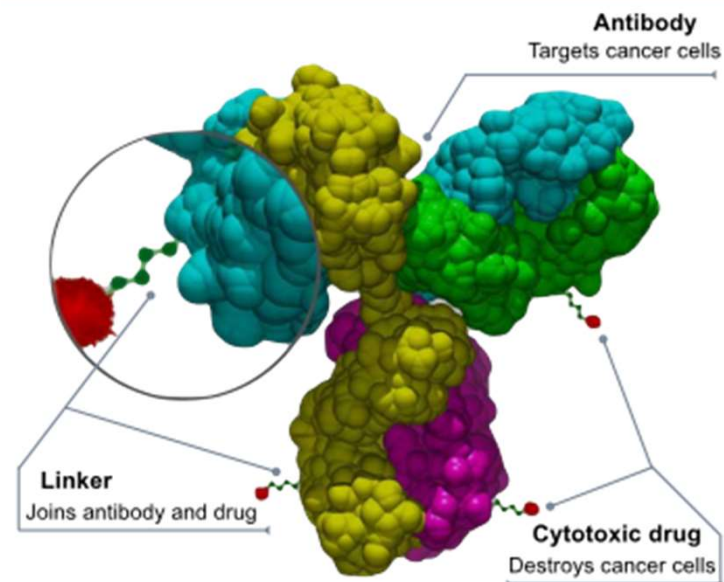
Covalent linking of antibodies with small-molecule drugs

Selenocysteine

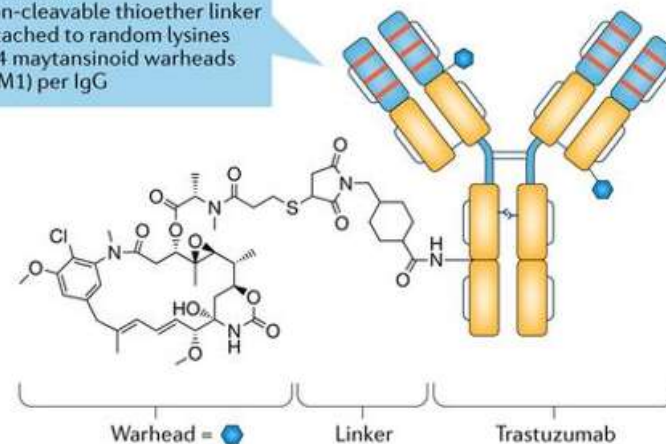


Antibody-drug conjugates (ADCs)

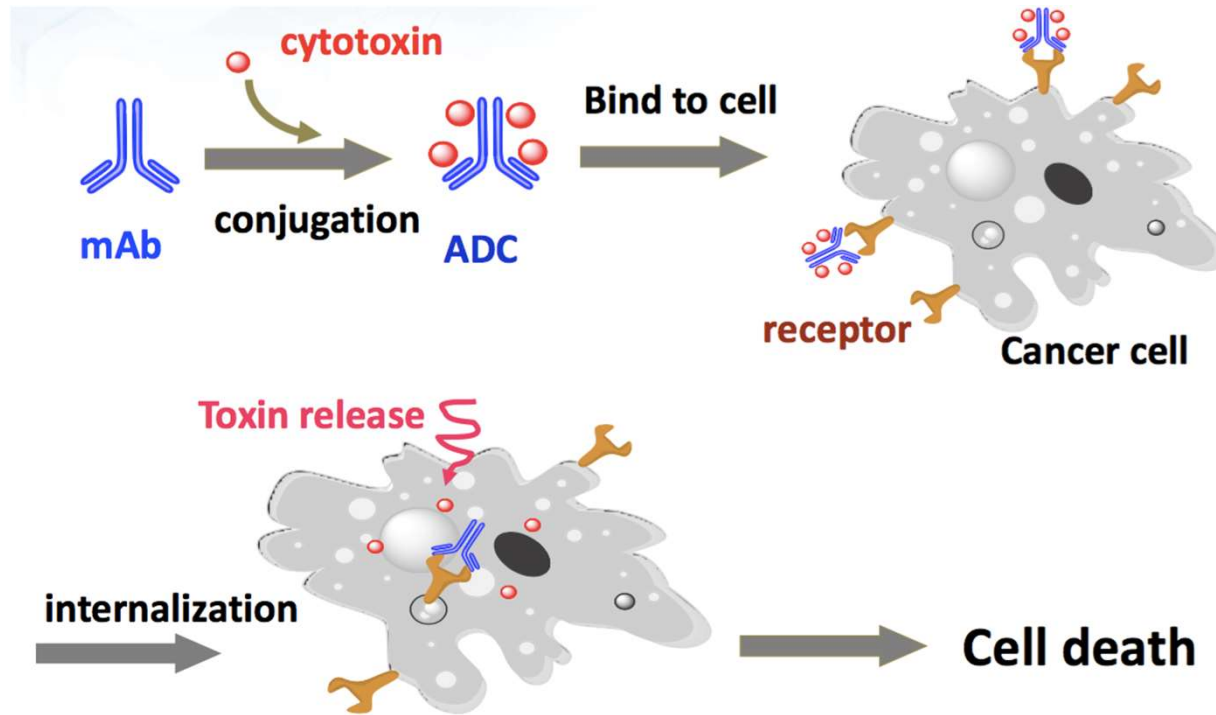
- Antibody–drug conjugates (ADCs) are one of the fastest growing classes of oncology therapeutics
- ADCs consist of recombinant monoclonal antibodies (mAbs) that are covalently bound to cytotoxic chemicals (known as warheads) via synthetic linkers
- Such immunoconjugates combine the antitumour potency of highly cytotoxic small-molecule drugs (300–1,000 Da, with subnanomolar half-maximal inhibitory concentration (IC₅₀) values) with the high selectivity, stability and favourable pharmacokinetic profile of mAbs
- Alternative formats to mAbs, such as protein scaffolds (designed ankyrin-repeat proteins (**DARPin**s), nanobodies, single-chain variable fragments (scFvs) and peptide–drug conjugates), dual-labelled ADCs and biparatopic drug conjugates, present new research avenues



Second generation
 • IgG1 mAb
 • Non-cleavable thioether linker attached to random lysines
 • 3–4 maytansinoid warheads (DM1) per IgG



Antibody-drug conjugates: the mode of action



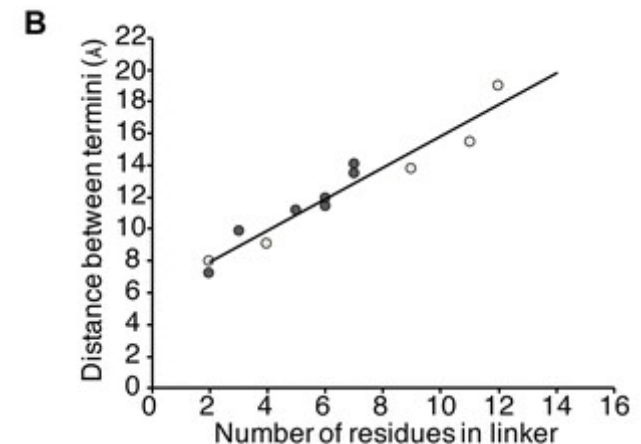
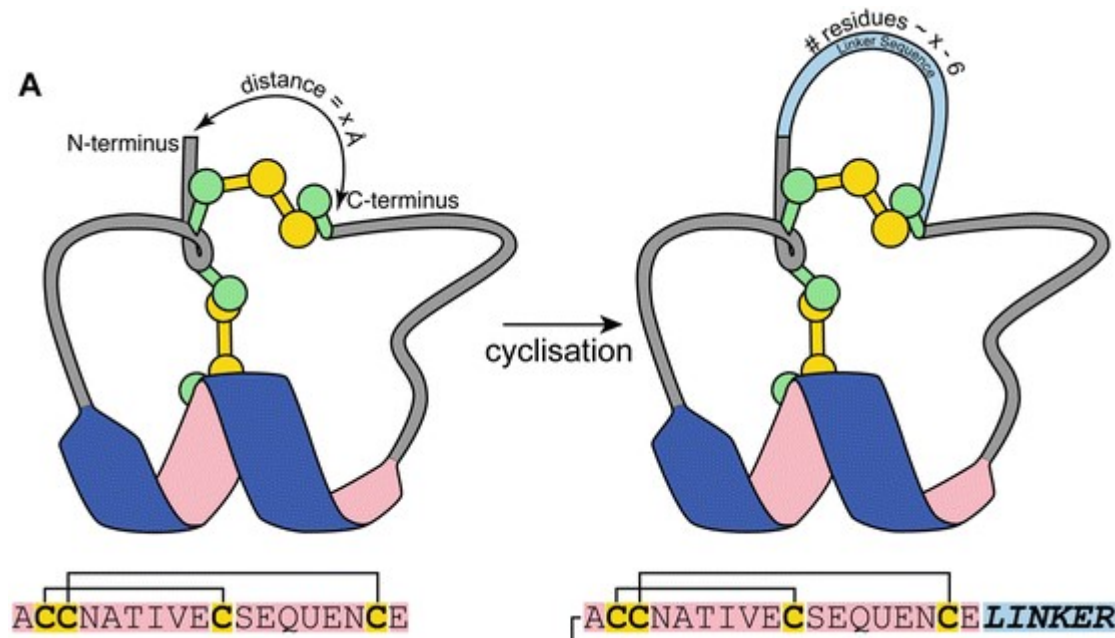
Anti-epidermal growth factor receptor 2 (HER2) monoclonal antibody **trastuzumab** conjugated to the maytansinoid DM1 via a nonreducible thioether linkage (MCC) with potential antineoplastic activity



Antibody-drug conjugate composed of a CD30-directed monoclonal antibody that is covalently linked to the antimicrotubule agent monomethyl auristatin E (MMAE)

Protein cyclisation

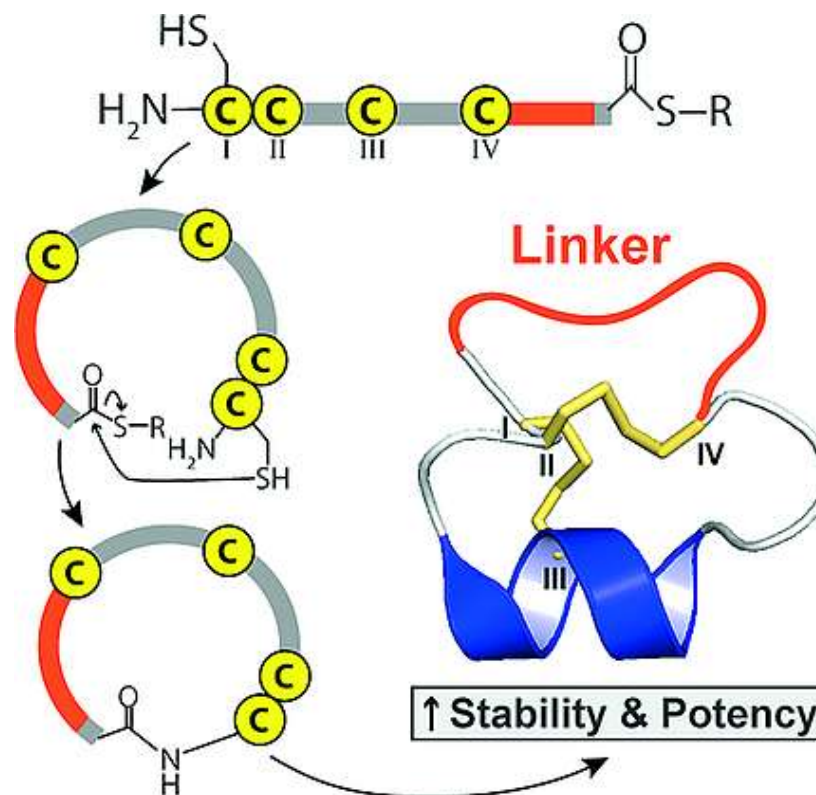
- Termini of most proteins happen to be close together
- Cyclisation via head-to-tail linkage of the termini of a peptide chain occurs in only a small percentage of proteins, but engenders the resultant **cyclic proteins with exceptional stability**
- Engineering efforts to cyclise peptides or proteins to gain superior stability and/or protease resistance



Daniel and Clark, Adv. Exp. Med. Biol., vol. 1030, pp 229-225, Springer, (2017)

An example of successful protein cyclisation

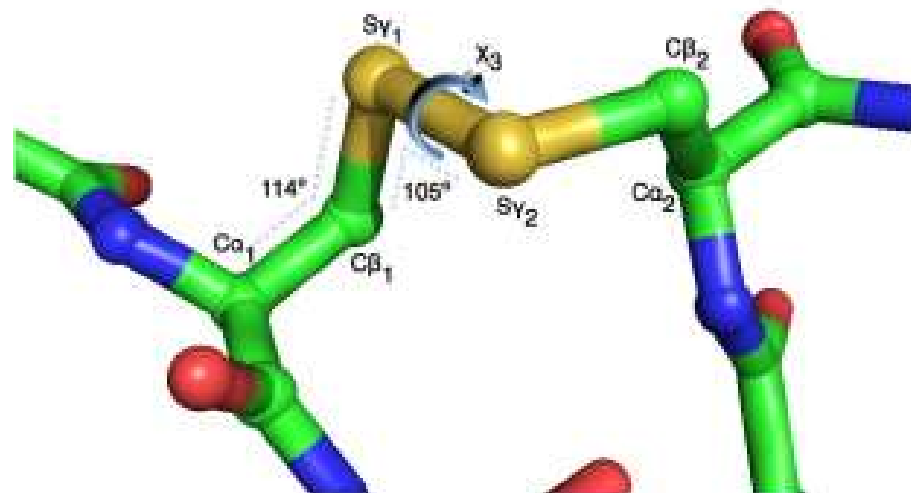
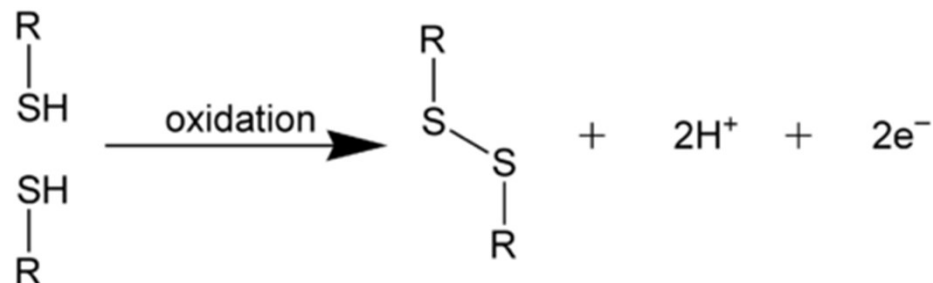
- **Conotoxins** are disulfide-rich peptides found in the venoms of marine snails (*Conus*)
- Pain killer activity by specific binding to ion channels
- They have attracted great attention from the pharmaceutical industry because of their potential uses as drug leads, but like most peptides, conotoxins are susceptible to proteolysis and typically are not orally bioavailable
- Multiple approaches have been used to stabilise conotoxins to improve their potential pharmaceutical use
- Specifically, the use of **backbone cyclisation dramatically improved their stability** in biological fluids and protease resistance

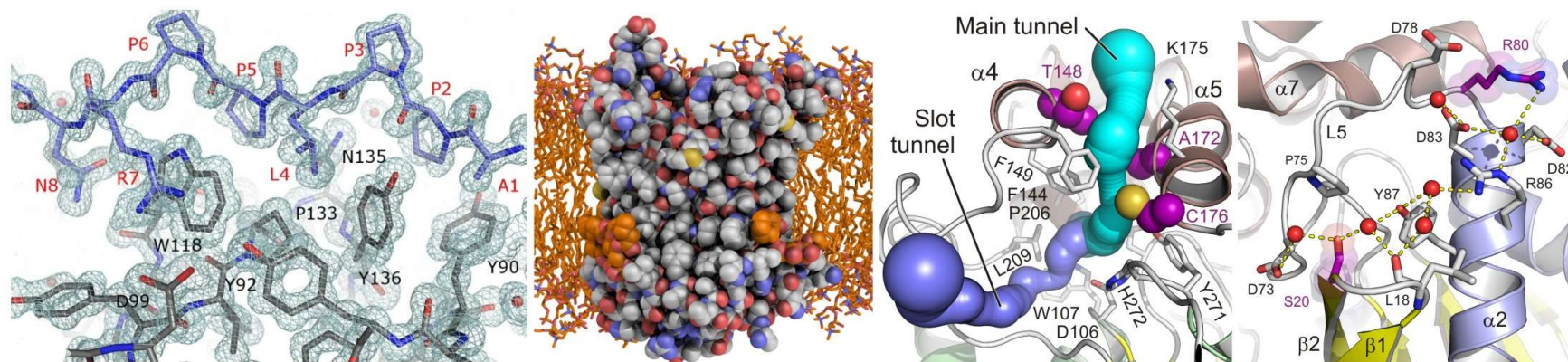


Wu *et al.*, *Eur. J. Organ. Chem.*
 3462-3472 (2016)

Disulfide bond engineering to enhance protein stability

- Two cysteine residues in close proximity will form a covalent bond
- Disulphide bond, disulphide bridge
- **Significantly stabilizes protein tertiary structure**
- Engineering efforts to mutate two codons into cysteines → **creation of disulphide bond in oxidising environment**
- Considerations: inter-cysteine distance and inter-cysteine orientation

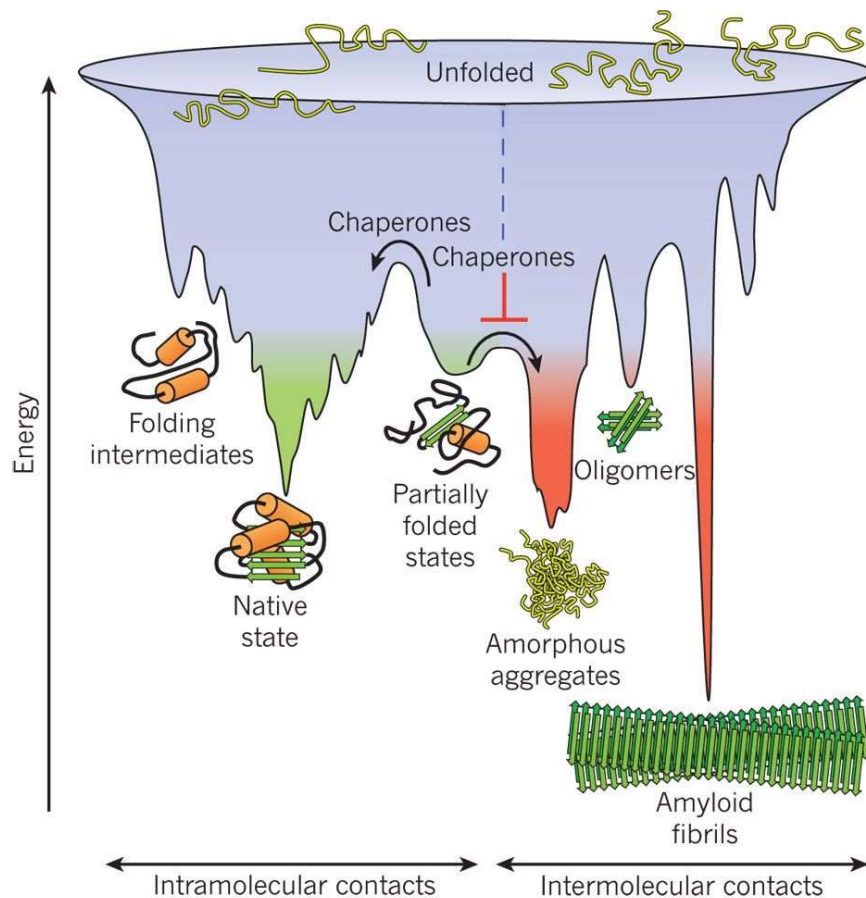




“High-resolution“ protein design

Concepts
Methods
Applications

Design of new protein functions



- **Proteins fold to their lowest free energy states**
- For designing new proteins, we must be able to calculate energies reasonably accurately, and sample protein conformations sufficiently to **find global minimum**
- To design proteins with new functions, we need hypotheses about configurations of atoms necessary to achieve desired function
- Finally, we have to test experimentally all designed proteins



“High-resolution“ protein design: requirements

What knowledge is required for “high-resolution” protein engineering:

- determination of 3D structure, for mutagenesis-based engineering
- knowledge of protein folding rules for *de novo* engineering
- computational modelling techniques usually required

Computational methods important for protein engineering:

- modelling & visualization
- energy/thermodynamic calculations
- searching conformation and sequence spaces
- comparison with known protein structures/sequences

The basis of more automated analysis of structural perturbations than our own “inspect and try” approach involves use of **an energy function** to evaluate plausibility of candidate structures:

- $E_{tot} = E_{bond} + E_{angl} + E_{dihe} + E_{impr} + E_{VDW} + E_{elec} + E_{Hbond} + \dots$
- This may be evaluated using a force field (e.g. CHARMM, Amber) and atomic coordinates available from simulation or modified PDB file



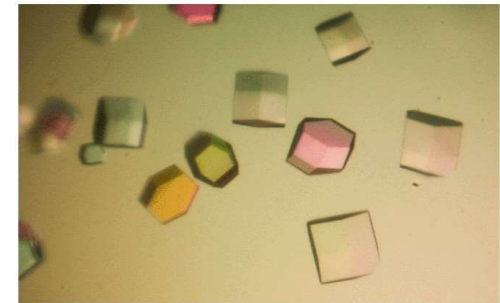
Protein structure determination techniques

- Determining 3D structure of proteins help protein engineers to understand molecular mechanism of protein action and its biological function.

Structural biology methods

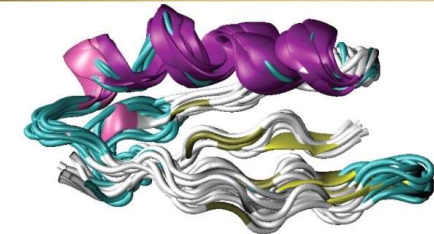
- X-ray crystallography

Crystallization required, no size limits, challenging for highly flexible proteins



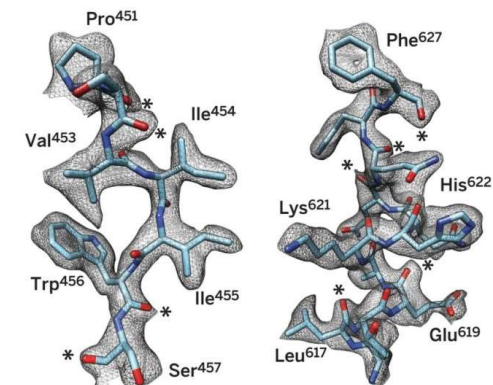
- Nuclear magnetic resonance (NMR) spectroscopy

Labelling required, suitable for smaller proteins, capturing protein motions



- Cryo-electron microscopy

Automation, direct electron detectors, image processing suitable for large protein complexes



Retrieval of atomic coordinates – Protein Data Bank (PDB)

The **Protein Data Bank (PDB)** is a database for the three-dimensional structural data of large biological molecules (proteins, nucleic acids), and determined by X-ray crystallography, NMR spectroscopy and cryo-electron microscopy.



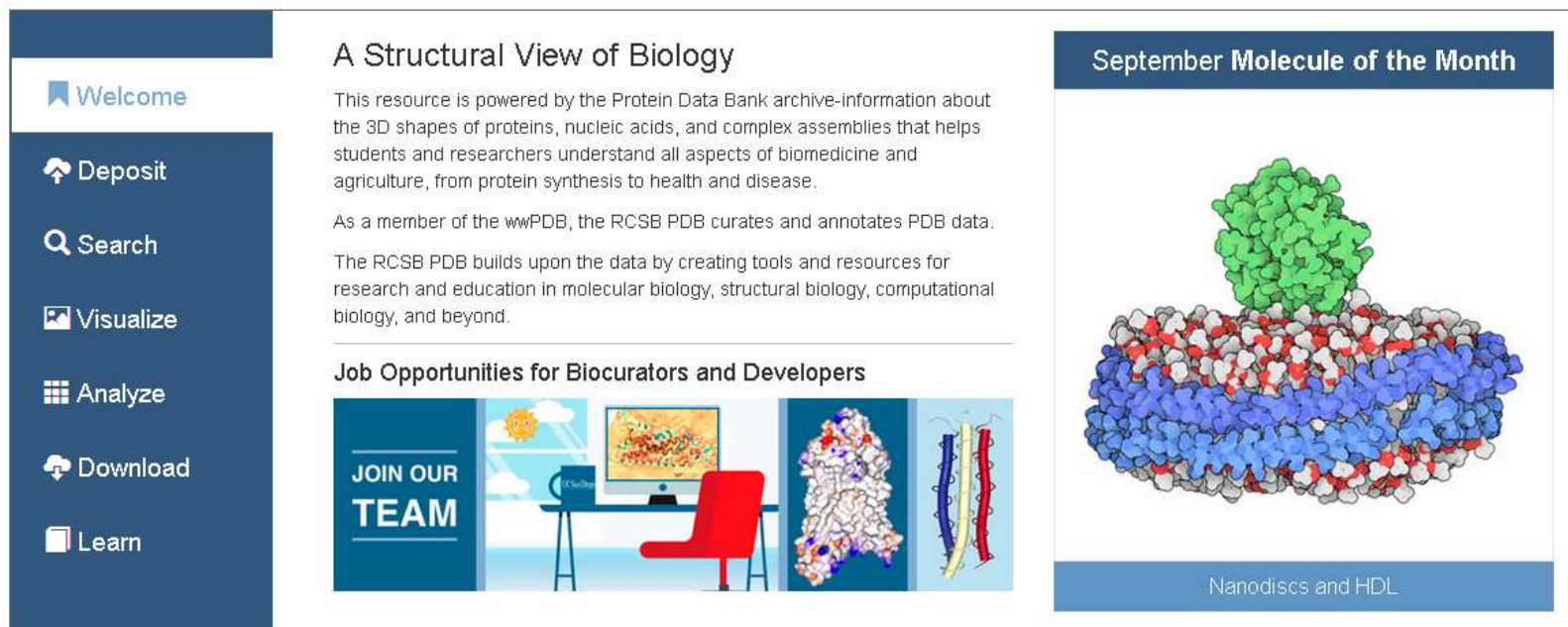
RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB

RCSB PDB 155407 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Search by PDB ID, author, macromolecule, sequence, or ligands Go

[Advanced Search](#) | [Browse by Annotations](#)

PDB-101 | Worldwide Protein Data Bank | EMDatabank | Nucleic Acid Database | Worldwide Protein Data Bank Foundation



Welcome

- Deposit
- Search
- Visualize
- Analyze
- Download
- Learn

A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

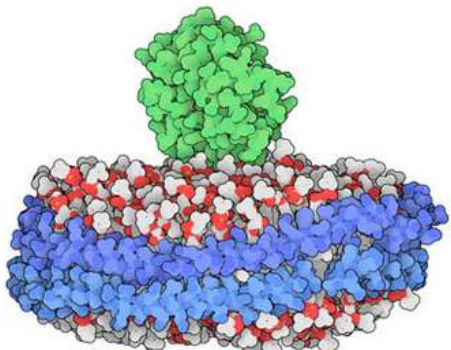
As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

Job Opportunities for Biocurators and Developers

JOIN OUR TEAM

September Molecule of the Month

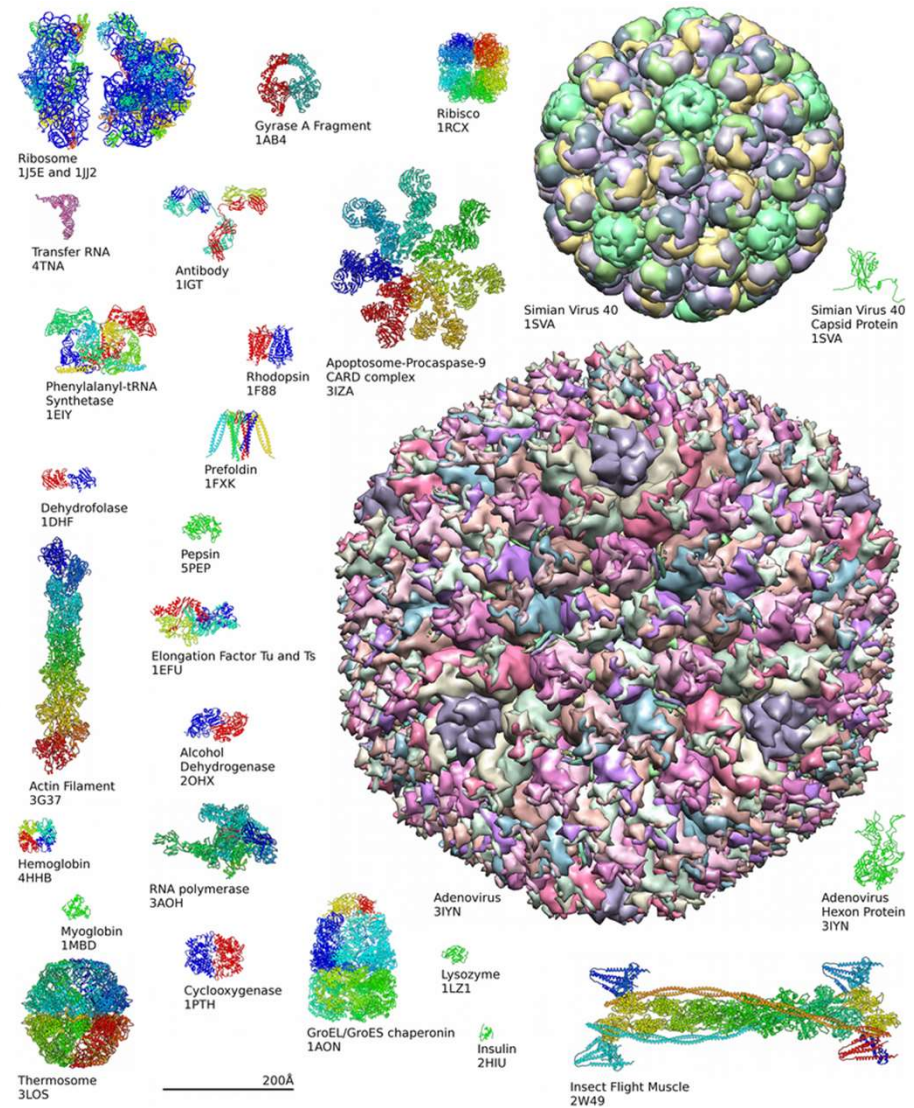
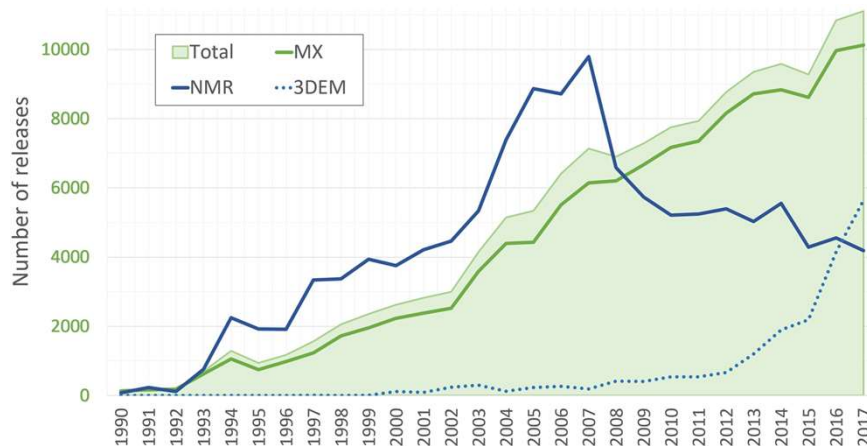


Nanodiscs and HDL



Protein Data Bank (PDB): overview

- The PDB is a key in areas of structural biology. Many other databases use protein structures deposited in the PDB. For example, SCOP and CATH
- Each structure published in PDB receives a four-character alphanumeric identifier, its PDB ID
- The file format initially used by the PDB was called the PDB file format
- From 2019, mmCIF is the master format for the PDB archive



PDB (mmCIF) file format

- **PDB file format:** The Protein Data Bank (**PDB**) format provides a standard representation for macromolecular structure data derived from X-ray diffraction, NMR and cryo-EM studies.

A **textfile** that includes **atomic coordinates**, observed sidechain rotamers, secondary structure assignments, atomic connectivity, ...



record type	atom number	atom	amino acid	chain ID	residue number	coordinates			occupancy	temperature factor	element name
						x	y	z			
ATOM	1	N	MET	D	1	14.322	20.430	-2.337	1.00	17.78	N
ATOM	2	CA	MET	D	1	14.423	20.285	-0.855	1.00	18.66	C
ATOM	3	C	MET	D	1	15.153	21.479	-0.242	1.00	18.46	C
ATOM	4	O	MET	D	1	15.811	22.241	-0.941	1.00	18.84	O
ATOM	5	CB	MET	D	1	15.068	18.970	-0.457	1.00	20.20	C
ATOM	6	CG	MET	D	1	16.569	18.895	-0.674	1.00	20.60	C
ATOM	7	SD	MET	D	1	17.240	17.319	-0.103	1.00	22.81	S
ATOM	8	CE	MET	D	1	16.378	16.194	-1.196	1.00	13.23	C
ATOM	9	N	LEU	D	2	14.983	21.653	1.071	1.00	18.40	N
ATOM	10	CA	LEU	D	2	15.568	22.825	1.718	1.00	19.14	C
ATOM	11	C	LEU	D	2	17.093	22.722	1.765	1.00	18.53	C
ATOM	12	O	LEU	D	2	17.655	21.647	1.945	1.00	19.07	O
ATOM	13	CB	LEU	D	2	15.025	23.078	3.121	1.00	21.35	C
ATOM	14	CG	LEU	D	2	15.438	24.404	3.773	1.00	22.45	C
ATOM	15	CD1	LEU	D	2	14.856	25.606	3.049	1.00	23.53	C
ATOM	16	CD2	LEU	D	2	15.042	24.430	5.244	1.00	23.83	C

Retrieval of coordinate files from the PDB: considerations

- PDB (mmCIF) file format – 3D model
- Structure factors file (CIF) – Data file



2PSD

Crystal Structures of the Luciferase and Green Fluorescent Protein

DOI: 10.2210/pdb2PSD/pdb

Classification: [OXIDOREDUCTASE](#)

Organism(s): [Renilla reniformis](#)

Expression System: [Escherichia coli](#)

Mutation(s): 8

Deposited: 2007-05-06 Released: 2007-06-05

Deposition Author(s): [Loening, A.M.](#), [Fenn, T.D.](#), [Gambhir, S.S.](#)

Experimental Data Snapshot

Method: X-RAY DIFFRACTION

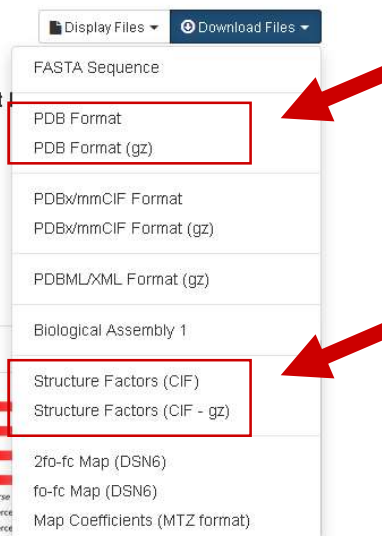
Resolution: 1.4 Å

R-Value Free: 0.183

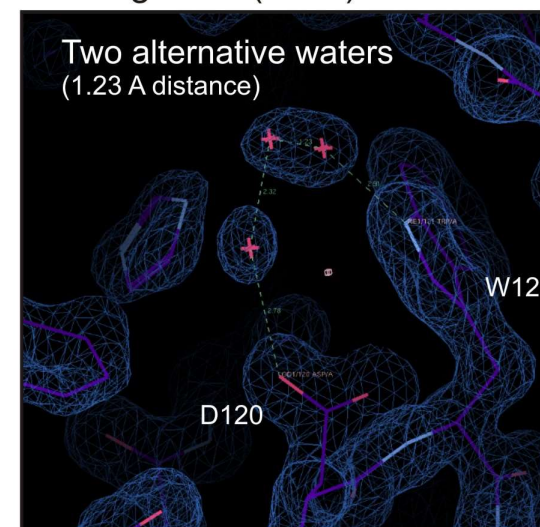
R-Value Work: 0.165

wwPDB Validation

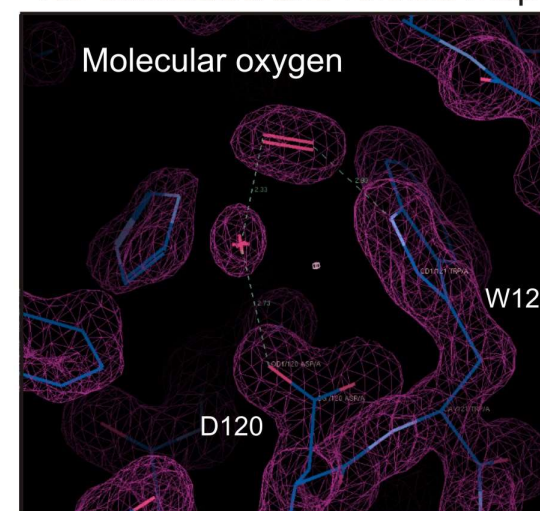
Metric	Value
Rfree	0.183
Clashscore	0.183
Ramachandran outliers	2fo-fc Map (DSN6)
Sidechain outliers	fo-fc Map (DSN6)
	Map Coefficients (MTZ format)



Loening *et al.* (2007) J. Mol. Biol.

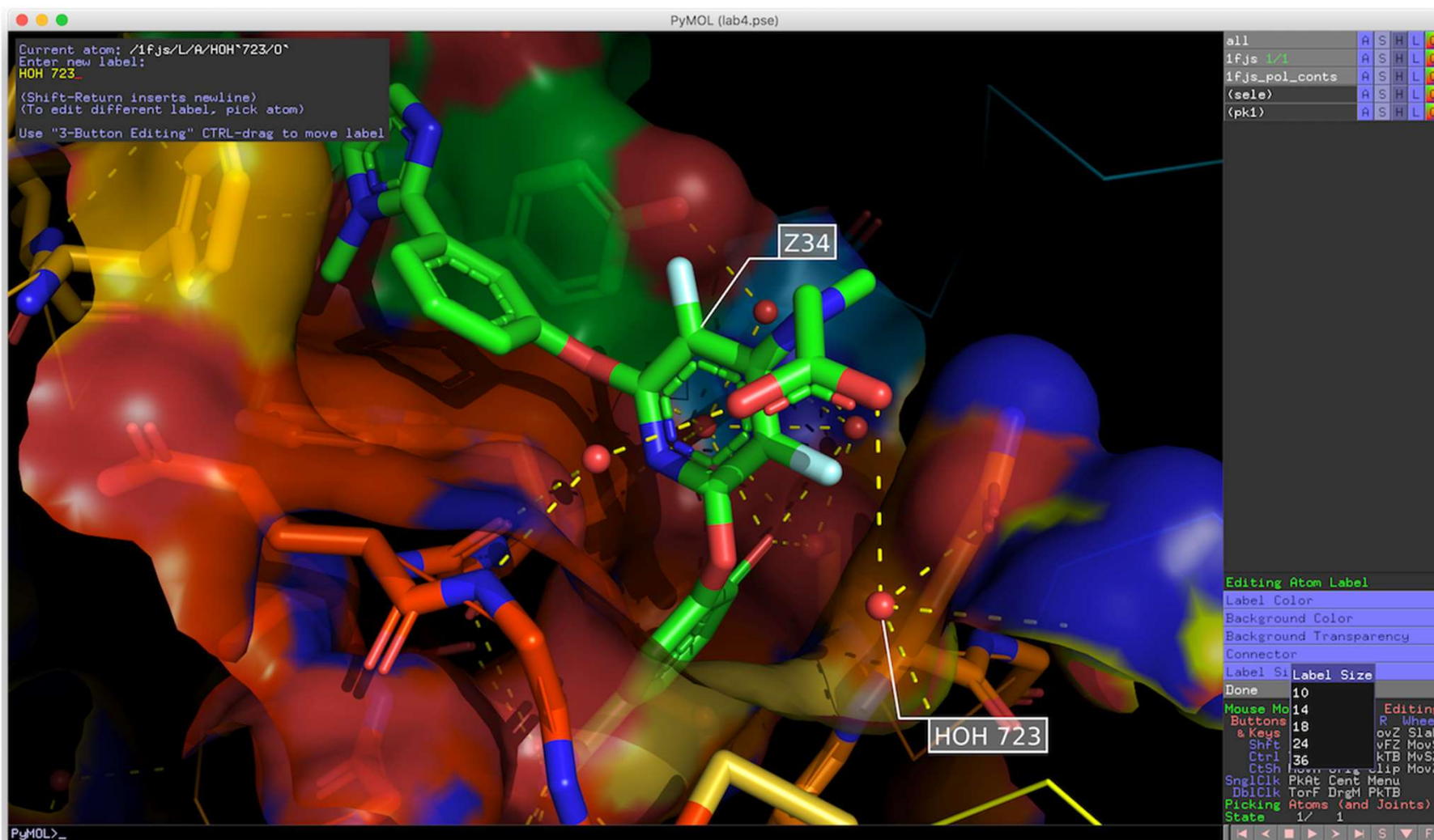


Re-calculated and refined map



Visualisation, modelling and computational tools

https://www.rcsb.org/pages/thirdparty/molecular_graphics





Selected visualisation and modelling software tools

- **PyMOL**
A free and open-source molecular graphics system for visualization, animation, editing, and publication-quality imagery. PyMOL is scriptable and can be extended using the Python language. Supports Windows, Mac OSX, Unix, and Linux
- **Chimera**
Interactive molecular modeling system, free to academic/non-profit; displays multiple sequence alignments and associated structures, atom-type and H-bond identification, molecular dynamics trajectories (AMBER format), and offers ligand-screening interface (DOCK), filter by number/position of H-bonds, and extensibility to create custom modules - for Windows, Linux, Mac OS X, IRIX, and Tru64 Unix
- **Swiss PDB viewer**
A 3D graphics and molecular modeling program for the simultaneous analysis of multiple models and for model-building into electron density maps. The software is available for Mac (OSX or PPC), Windows, Linux, or SGI
- **YASARA**
A complete molecular graphics and modeling program, including interactive molecular dynamics simulations, structure determination, analysis and prediction, docking, movies and eLearning for Windows, Linux and MacOSX
- **VMD**
VMD (Visual Molecular Dynamics) runs on many platforms including MacOS X, and several versions of Unix and Windows. VMD provides visualization, analysis, and Tcl/Python scripting features, and has recently added sequence browsing and volumetric rendering features. VMD is distributed free of charge
- **Foldit**
Foldit is a crowdsourcing computer game based on protein modelling

De novo protein design: problem definition

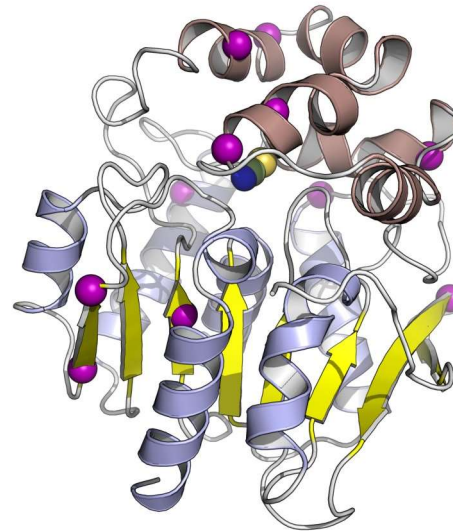
```

>dhmeA-NC
MSSASSNARDEVIAAI
HEEADWVDRTVYPFES
RCIGLSSGAVHYIDEG
PDDGGRETLMLHGNP
TWSFLYRHLVRDLRDE
YRCVALDYLGFLSER
PTDFSYRPEDHADVVE
EFIDELGLEDVVLVGH
DWGGPIGFSYAIDHPE
NVGGLVVMNTWMPVS
DDKHFSRFSKLLGGRI
GRELCERYDLFTRVIM
PMGFADRSRFTESARE
QYRAANRGDRTGTGIF
  
```

Protein folding



Protein design



- Given the desired three dimensional structure of a protein, design an amino acid sequence that will assume that structure
- Of course, a precise set of atomic coordinates would determine sequence. Usually we start with an **approximate desired structure**
- Alternatively, we may want to design for a particular function (e.g. ability to bind a particular ligand)
- **Protein design is the inverse of the protein folding problem!**

Protein design workflow

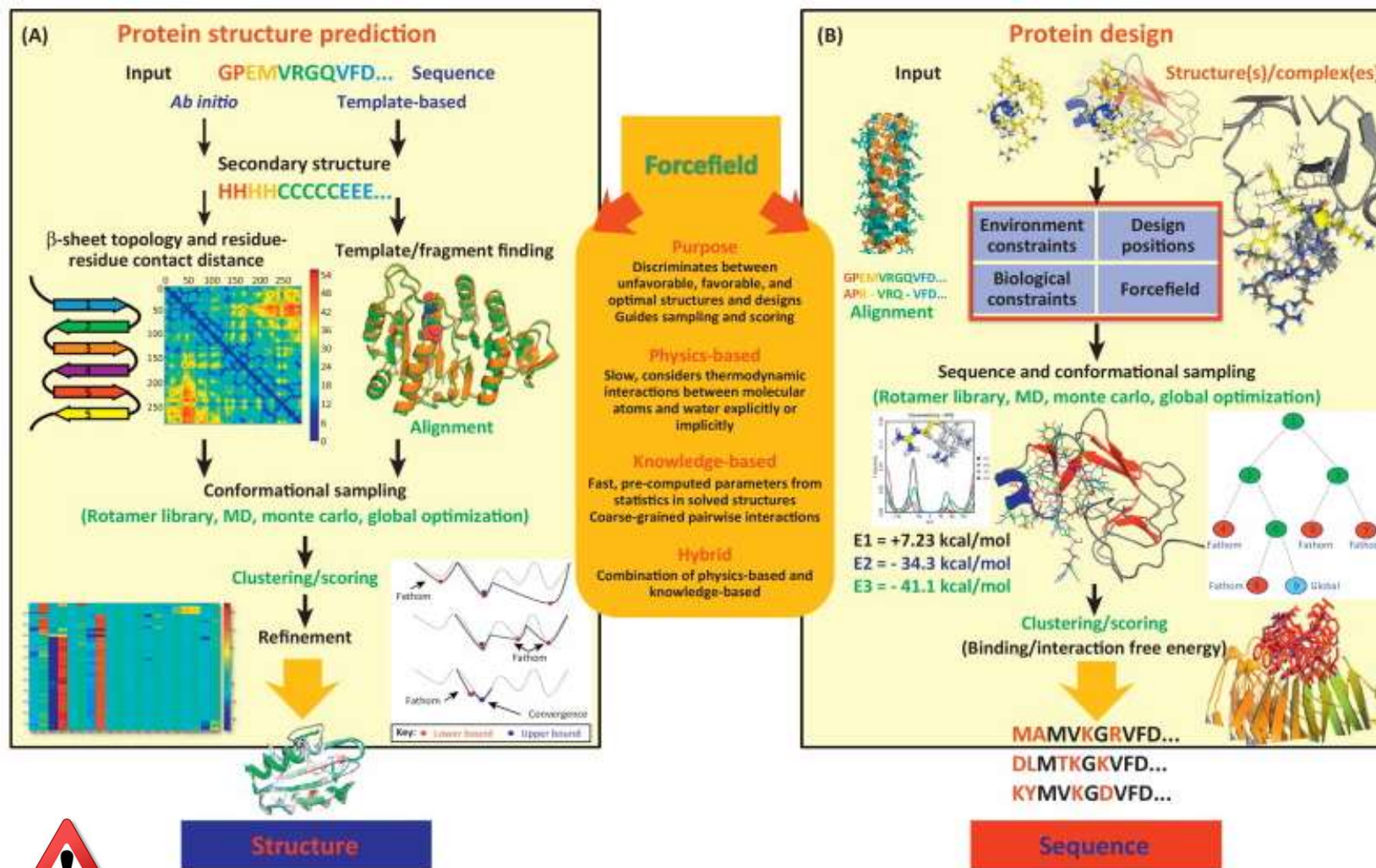


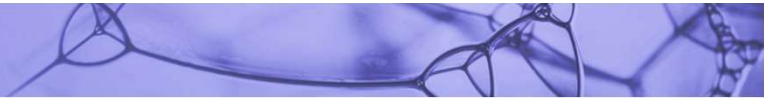
- Computer calculation of optimal sequence for desired structure and function
- Read off amino acid sequence of designed protein
- Back translate to DNA sequence, and make a gene
- Make a protein and assay



Computational protein design

Protein folding and design are two sides of the same coin





The key calculation: total energy of the protein

- Design along the backbone or scaffold
- Rotamer/backbone and rotamer/rotamer interaction energies tabulated
- Given a target backbone geometry, we aim to select side-chain rotamers at each position to minimize total protein energy:

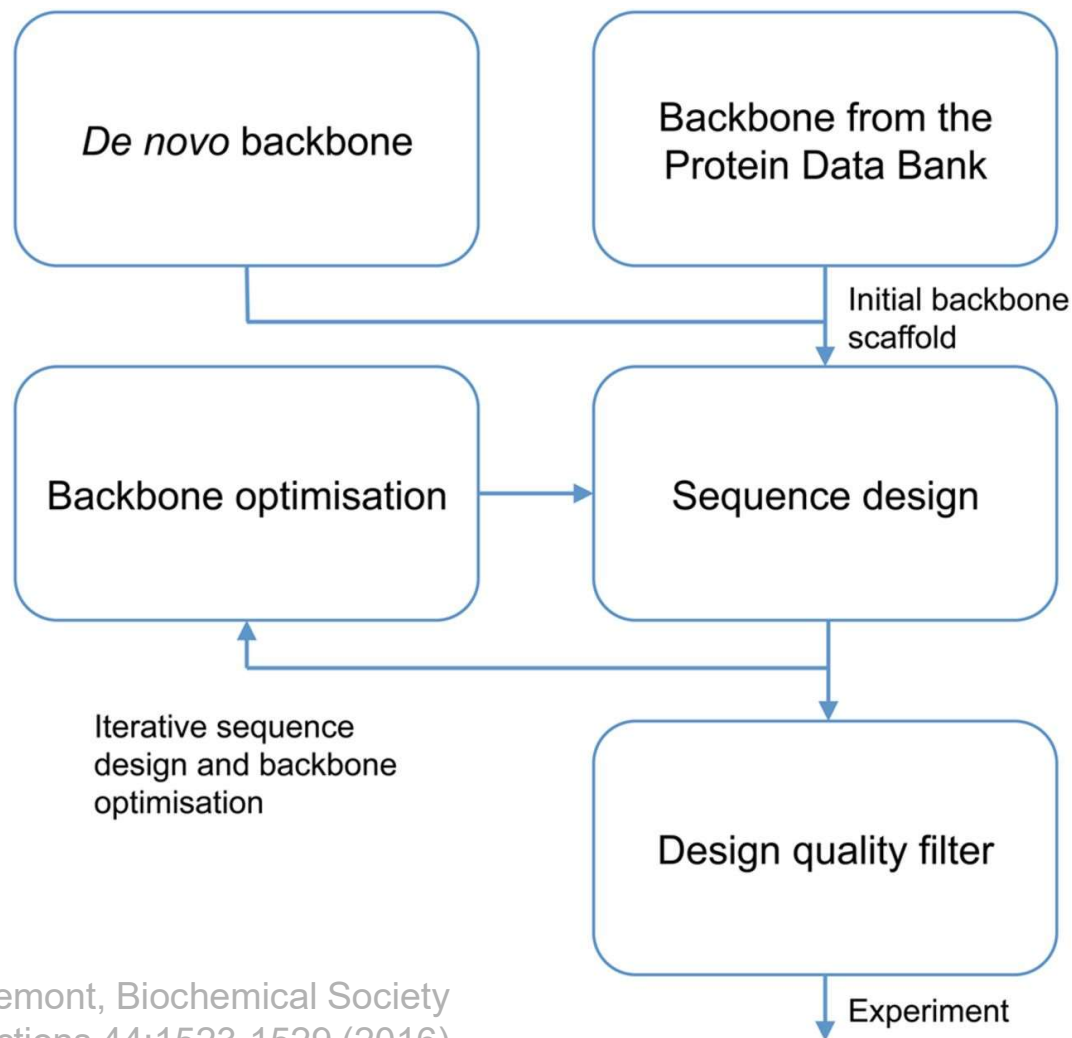
$$E_T = \sum_i \left[E_i(r_i) + \sum_{i \neq j} E_{ij}(r_i, r_j) \right]$$

where:

r_i	specifies both the amino acid at position i and its side-chain geometry
E_i	the self-energy of a particular rotamer r_i
E_{ij}	the pair energy of rotamers r_i, r_j

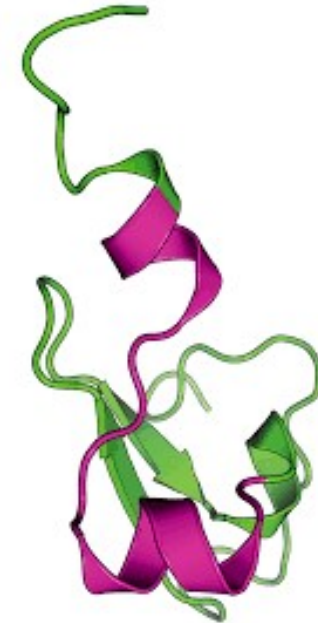
A typical computational protein design workflow

- Initial backbone structures can be either generated *de novo* or taken from solved protein structures
- Sequences that stabilise the designed backbone structure are then computationally designed, and the backbone may be permitted to move as part of an iterative design cycle
- Finally, promising designs are selected for experimental characterisation



The protein backbone design and selection

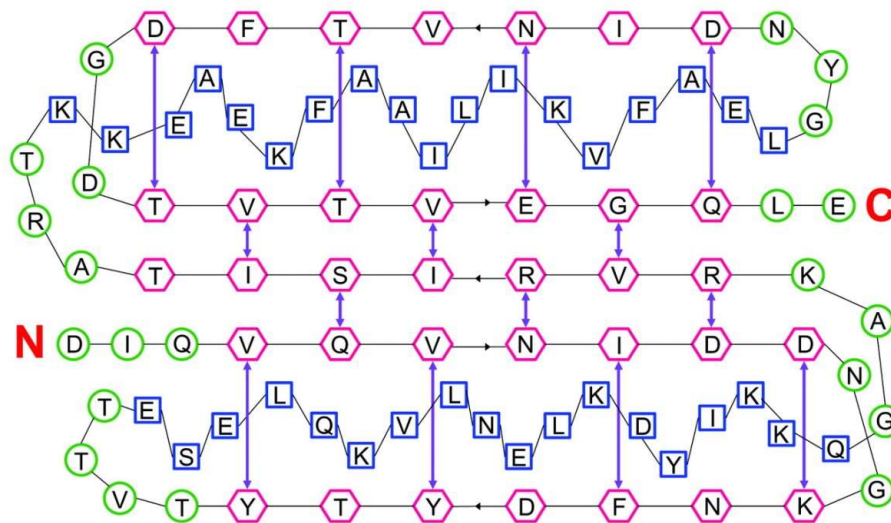
- First step in *de novo* protein design is selection of target backbone structures:
 - This is as much art as science
 - Often multiple target structures are selected, some won't work
 - Proteins can only adopt a limited set of backbone structures
 - But we do not have a perfect description of that set
- Approaches used towards this goal:
 - Use an experimentally determined backbone structure
 - Assemble secondary structural elements by hand
 - Use a fragment assembly program like Rosetta, and select fragment combinations that fit approximate desired scaffolds



Example of successful backbone design

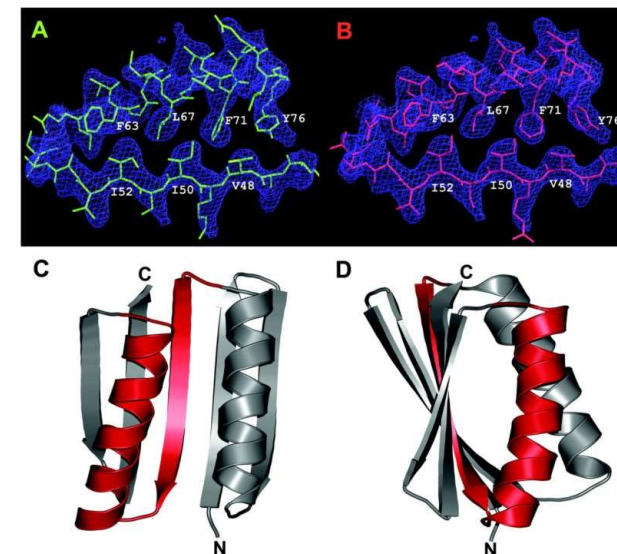
- Towards design of “Top7,” a protein with a novel fold
- Started with a 2D schematic, then used Rosetta fragment assembly
- Ended up with 172 backbone models that fit initial scaffold

Two-dimensional schematic of the target fold



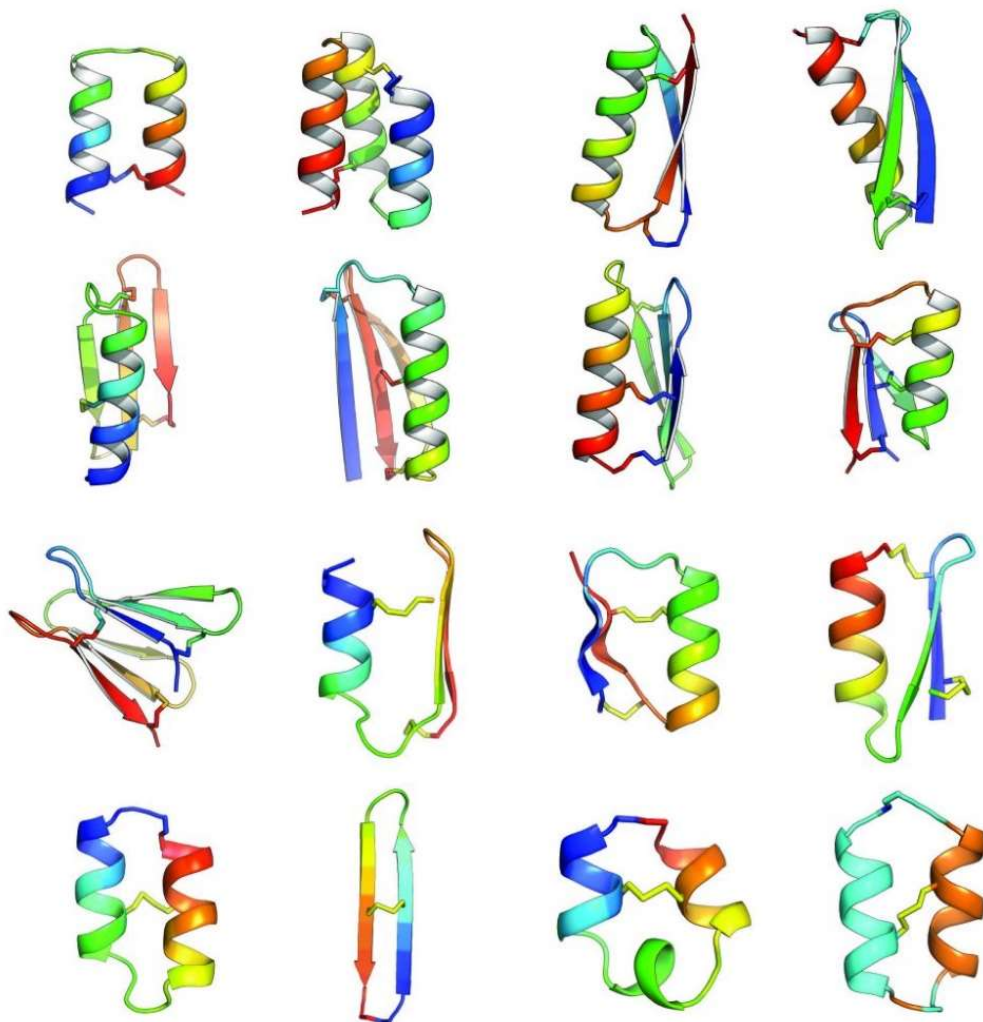
Initial schematic of target scaffold: hexagons; β sheet, squares; α helix, arrows; hydrogen bonds. Letters indicate amino acids in final designed Top7 sequence

Crystal structure of final Top7 protein



New topologies of *de novo* designed hyperstable peptides

Bhardwaj *et al.*, Nature (2016): doi: 10.1038/nature19791

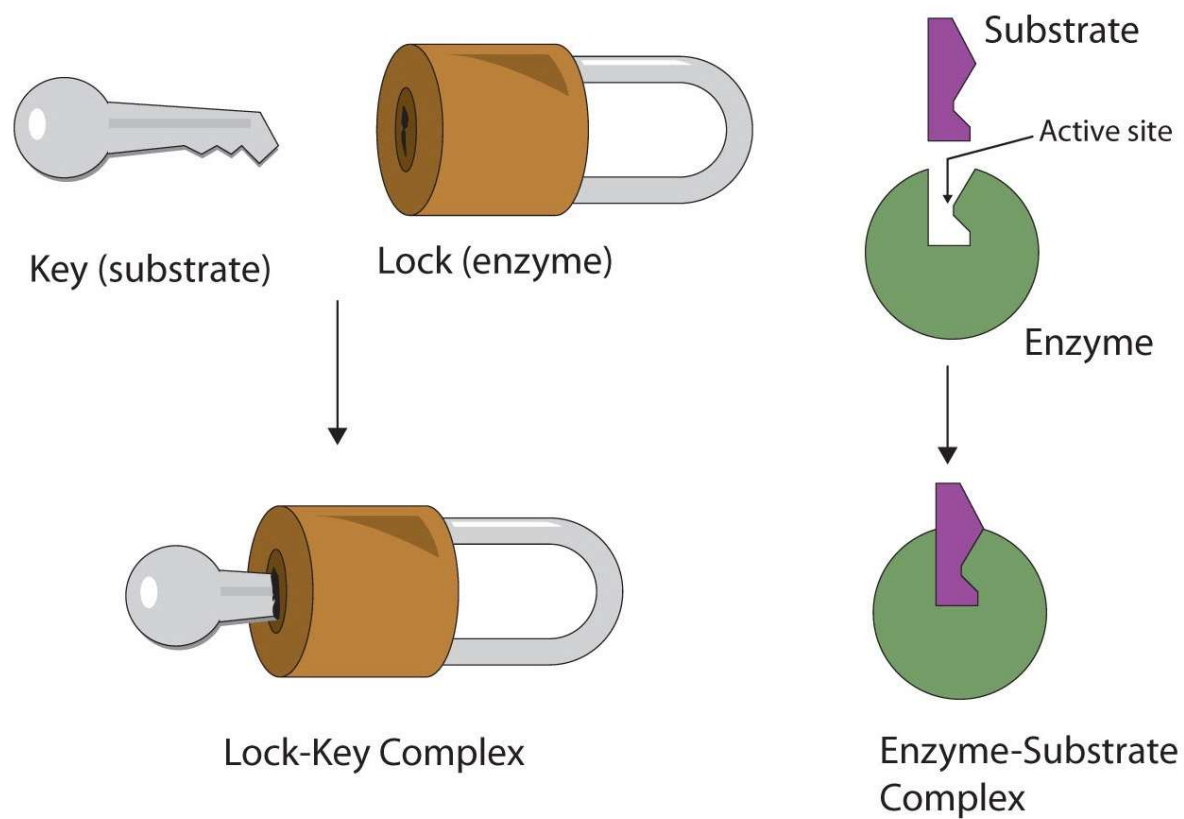


- The development of computational methods for accurate *de novo* design of conformationally restricted peptides, and the use of these methods to design 18-47 residue, disulfide-crosslinked peptides, a subset of which are heterochiral and/or N-C backbone-cyclized.
- Both genetically encodable and non-canonical peptides are exceptionally stable to thermal and chemical denaturation, and 12 experimentally determined X-ray and NMR structures are nearly identical to the computational design models.
- The computational design methods and stable scaffolds presented here provide the basis for development of a new generation of peptide-based drugs.

Computational design of a new enzyme



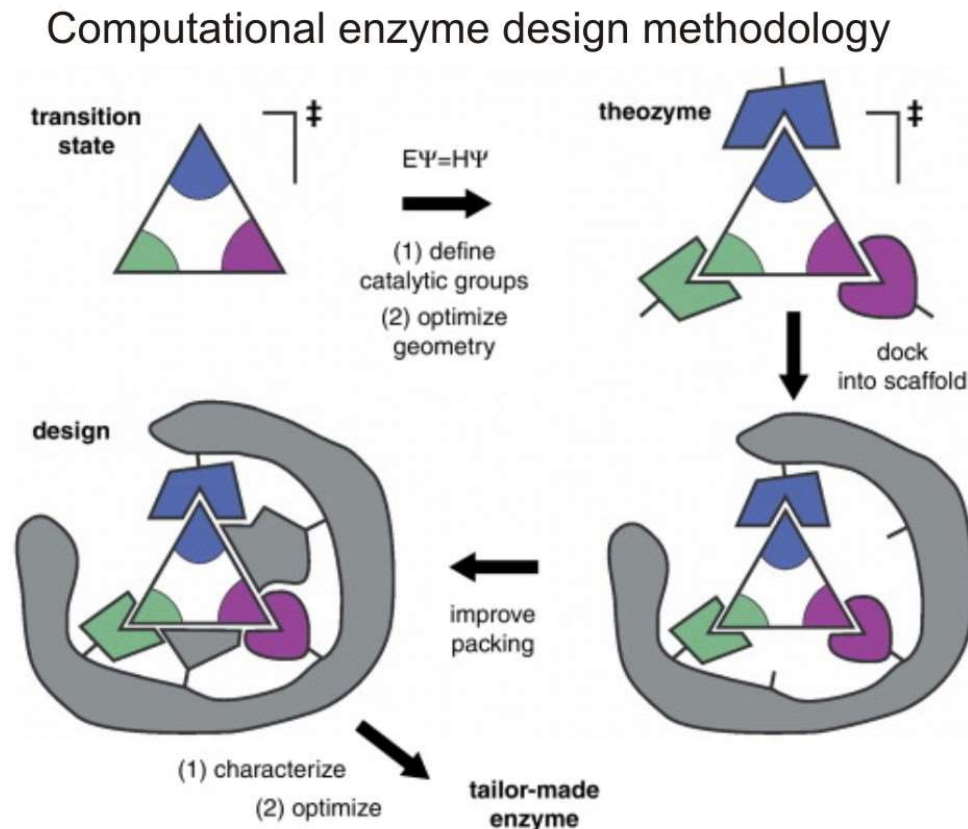
Lock and Key analogy: lock = enzyme; key = substrate



Computational design of a new enzyme



1. Disembodied residues are placed to stabilise reaction transition state - **THEOZYME**
2. PDB database is then searched for protein structures (backbones) with correct orientations (geometries)
3. Remaining residues in active site are optimized for proper packing and elimination clashes
4. Final adjustments of overall structure of tailor-made enzyme



Kries *et al.*, *Curr. Opin. in Chem. Biol.*
 17:221-8 (2013)

The hub for Rosetta modelling software



<https://www.rosettacommons.org/>

 Search

Home

Software

Documentation
& Support

Developer
Resources

About

RosettaCON

A banner for Rosetta Software with a blue background. On the left, there is a 3D ribbon diagram of a protein structure in various colors (red, orange, yellow, green, blue). On the right, the text "Rosetta Software" is written in a bold, sans-serif font. Below it, a paragraph describes the software as a dynamic and evolving macromolecular modeling suite. Further down, the text "Powered by the Commons" is written in a bold, sans-serif font, followed by a paragraph stating that RosettaCommons members enable notable scientific advances in computational biology.

Rosetta Software
A dynamic and evolving macromolecular modeling suite addressing biomolecular structure prediction and design.

Powered by the Commons
RosettaCommons members enable notable scientific advances in computational biology

- The Rosetta software suite includes **algorithms for computational modelling and analysis of protein structures**. It has enabled notable scientific advances in computational biology, including *de novo* protein design, enzyme design, ligand docking, and structure prediction of biological macromolecules and macromolecular complexes.

RosettaDesign

<http://rosettadesign.med.unc.edu/>

rosetta commons

high resolution structure prediction and design software

- login ×
- documentation ×
- register ×
- rosetta pipeline ×
- contact us ×

free.servers

- × [Rosetta Dock](#)
- × [Robetta Structure Prediction](#)
- × [Rosetta Antibody/Homology Modeling](#)

swiftlib

rosetta.design^{3.5}
computational protein design software

welcome.

Rosetta design can be used to identify sequences compatible with a given protein backbone. Some of Rosetta design's successes include the design of a novel protein fold, redesign of an existing protein for greater stability, increased binding affinity between two proteins, and the design of novel enzymes.

If you would like to use Rosetta.design, please [register for an account](#). If you already have an account, you can login below.

- Kuhlman Lab

login.

User Name:

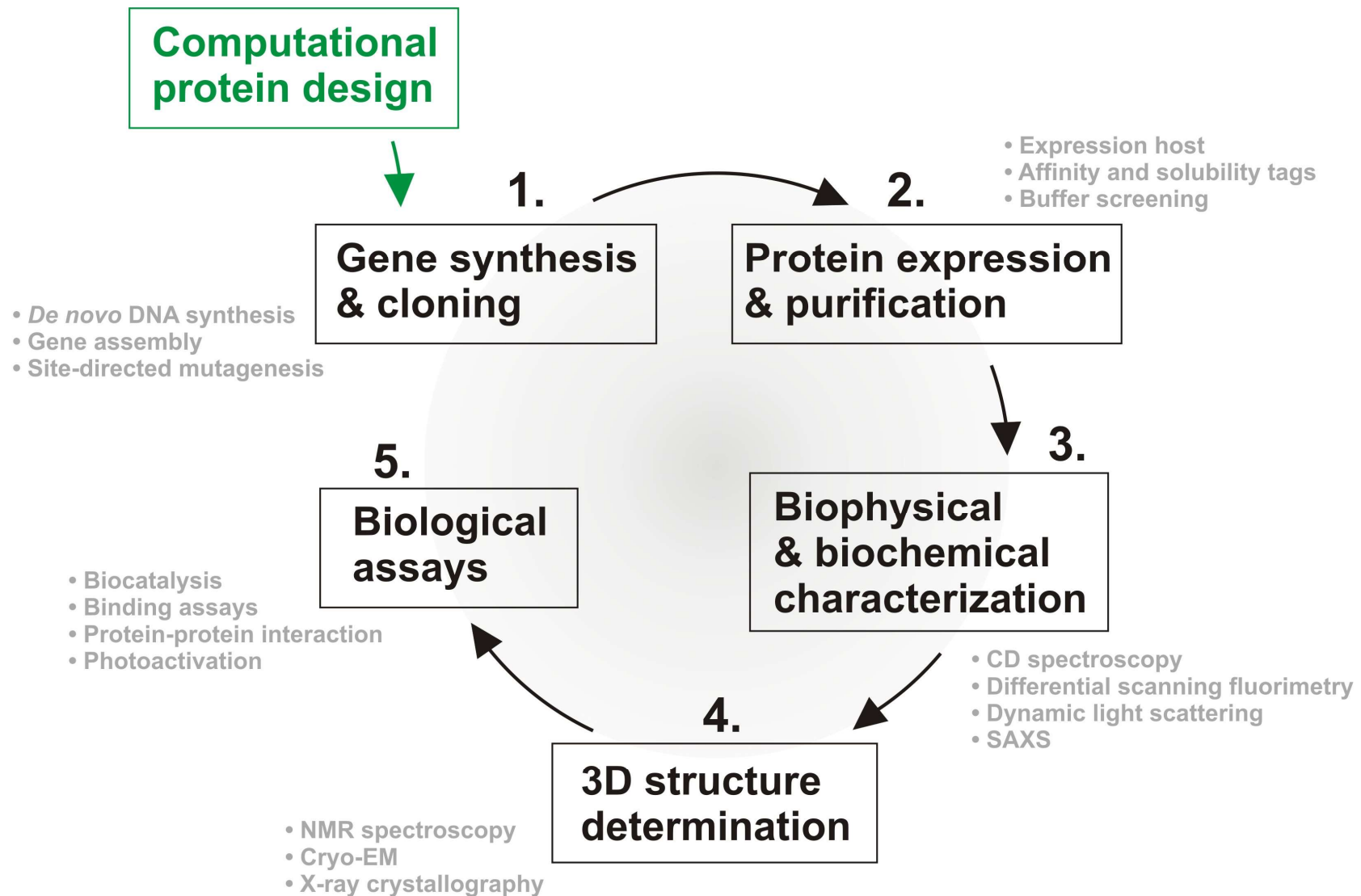
Password:



- **Rosetta design** can be used to identify sequences compatible with a given protein backbone. Some of Rosetta design's successes include the design of a novel protein fold, redesign of an existing protein for greater stability, increased binding affinity between two proteins, and the design of novel enzymes.

What to do after the design is finished ...

... Experimental validation

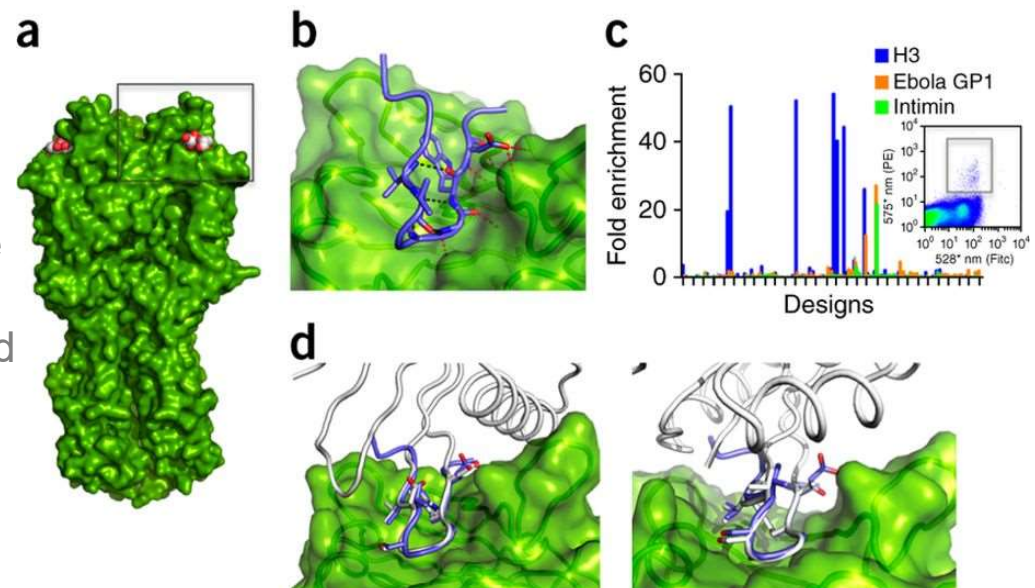


De novo design of trimeric influenza-neutralizing proteins

An example of computational protein design I.

- Many viral surface glycoproteins and cell surface receptors are homo-oligomers and thus can potentially be targeted by geometrically matched homo-oligomers that engage all subunits simultaneously to attain high avidity and/or lock subunits together
- A general strategy for the computational design of homo-oligomeric protein assemblies with binding functionality precisely matched to homo-oligomeric target sites
- **In the first step, a small protein is designed that binds a single site on the target. In the second step, the designed protein is assembled into a homo-oligomer such that the designed binding sites are aligned with the target sites**
- This approach was used to design high-avidity trimeric proteins that bind influenza A hemagglutinin (HA) at its conserved receptor binding site
- The designed trimers can both capture and detect HA in a diagnostic format, neutralizes influenza in cell culture, and completely protects mice

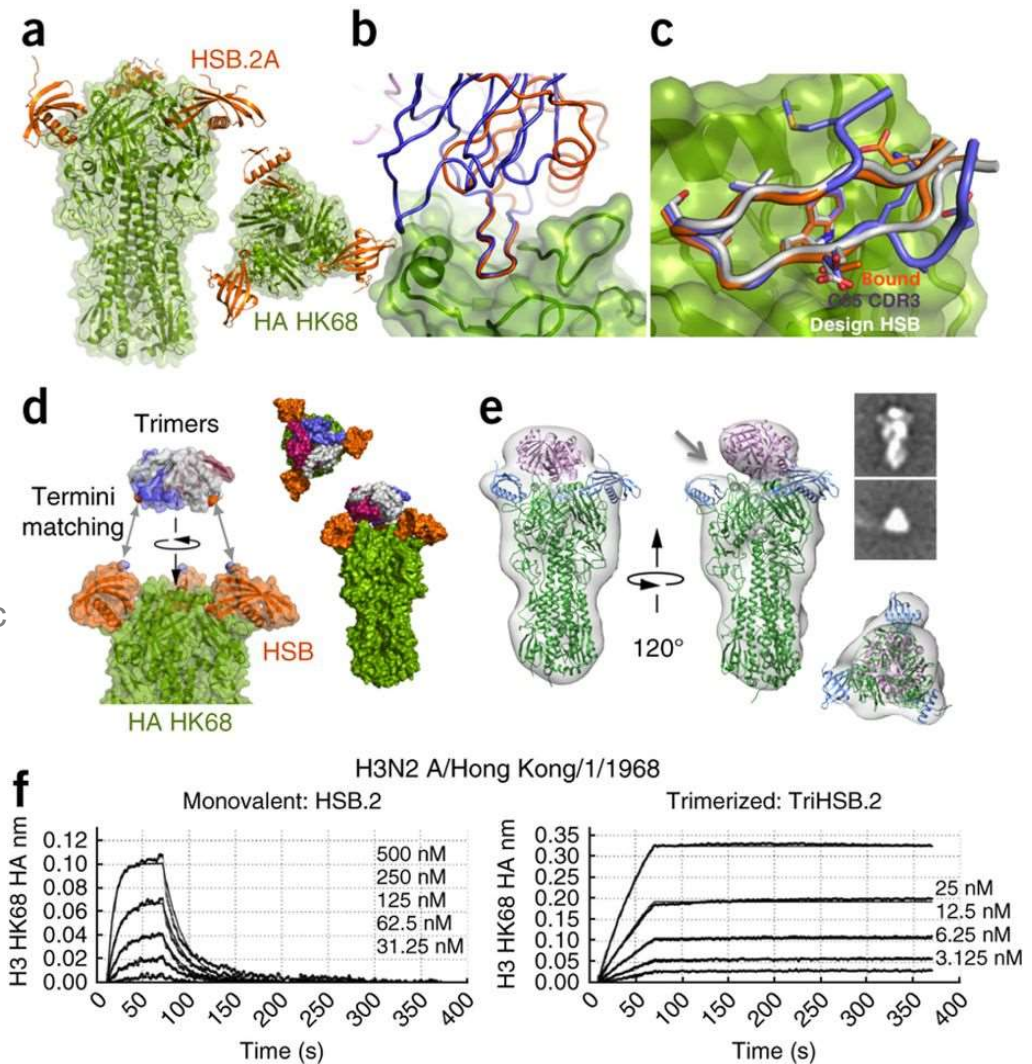
Strauch *et al.*, Nature Biotechnology,
 35: 667-671 (2017)



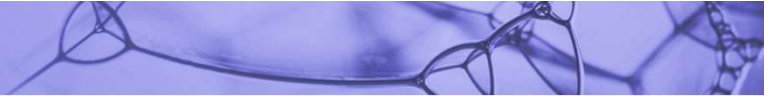
De novo design of trimeric influenza-neutralizing proteins

An example of computational protein design I.

- (a) A co-crystal structure of HSB.2A bound to HK68 HA shows that the design binds to the RBS as designed
- (b) Superposition of antibody C05 and the Tri-HSB.2A-HA crystal structure
- (c) Close agreement is found for the contacts with HA between the tip of HCDR3 of C05 (blue), the HSB designed model (gray) and the bound crystal structure HSB.2A (orange)
- (d) Translational and rotational sampling of trimeric protein scaffolds (gray, magenta, blue) to identify trimers that connect the termini of three HSB (orange) molecules bound to trimeric HA (green)
- (e) EM reconstruction of Tri-HSB.1C bound to HK68 HA
- (f) BLI titrations of H3 HK68 HA binding to monomeric HSB.2 and trimer Tri-HSB.2



Strauch *et al.*, Nature Biotechnology,
 35: 667-671 (2017)

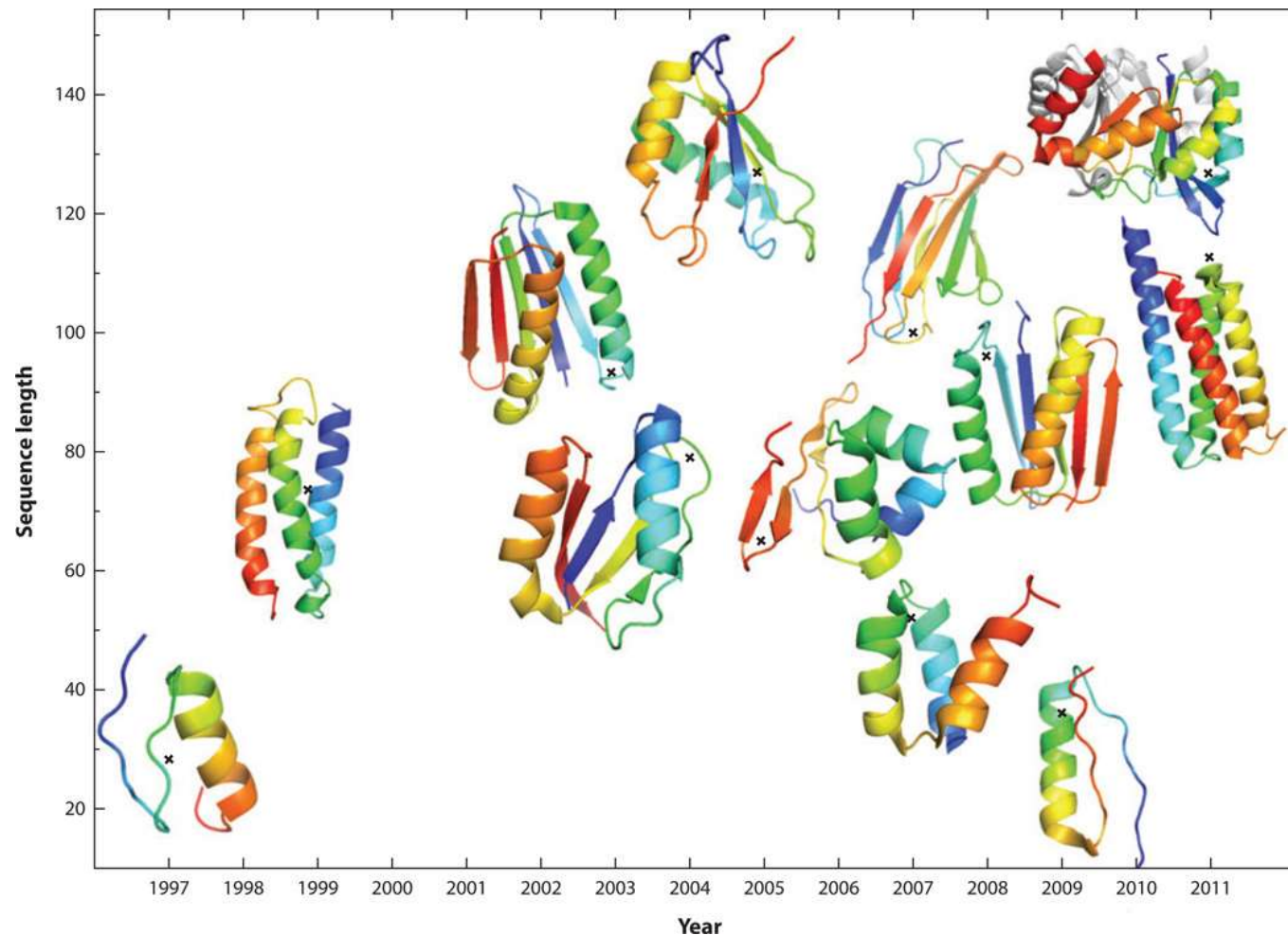


How well does *de novo* protein design work?

- Impressive recent successes
- But, keep in mind that:
 - successful protein design projects often involve creating and experimental testing dozens of candidate proteins to find one
 - unsuccessful projects are not reported
 - design of membrane proteins is still challenging

Computational protein design: the reality

- An increase in the sequence lengths of computationally designed and structurally validated proteins



Great future for protein designers

MENU ▾

nature
International journal of science

Subscribe


Search


Login

TECHNOLOGY FEATURE • 23 JULY 2019

The computational protein designers

A new breed of protein engineers is finding that the best way to create a molecule is to build it from scratch.



Summary

Today

RATIONAL DESIGN

1. Computer aided design



2. Site-directed mutagenesis



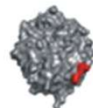
Individual mutated gene

3. Transformation

4. Protein expression

5. Protein purification

6. *not applied*



Constructed mutant enzyme

Next lecture

DIRECTED EVOLUTION

1. *not applied*

2. Random mutagenesis



Library of mutated genes
(>10,000 clones)

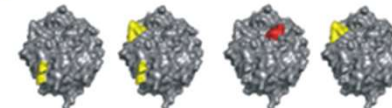
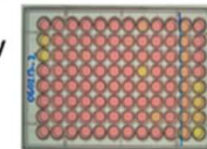
3. Transformation

4. Protein expression

5. *not applied*

6. Screening and selection

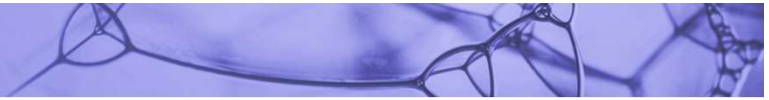
- stability
- selectivity
- affinity
- activity



Selected mutant enzymes

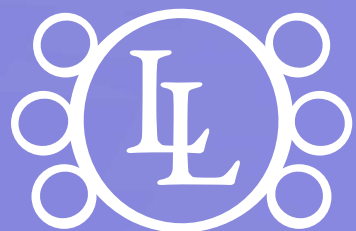
**IMPROVED
ENZYME**

7. Biochemical testing

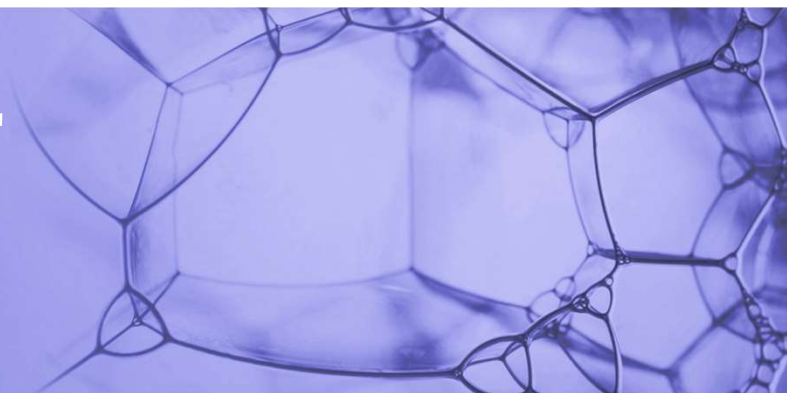


Questions

Dr. Martin Marek
Loschmidt Laboratories
Faculty of Science, MUNI
Kamenice 5, bld. A13, room 332
martin.marek@recetox.muni.cz



LOSCHMIDT
LABORATORIES





Supplementary materials

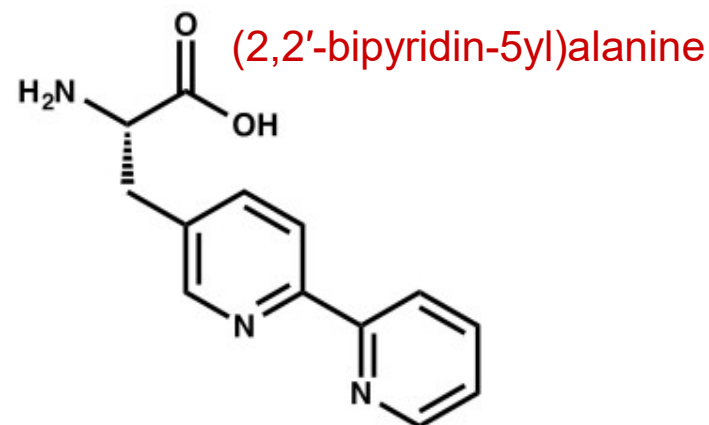
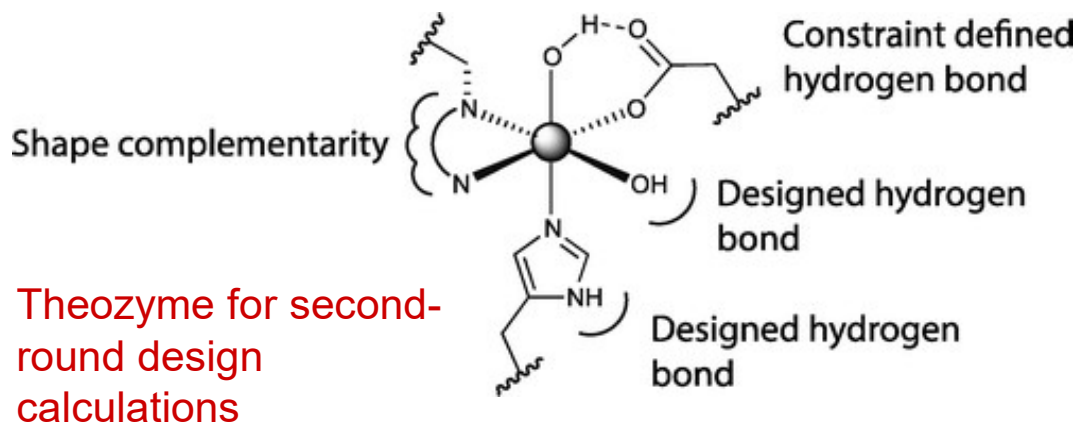
Computational protein design: the glossary

- *De novo* protein design: computationally designed proteins that can fold into a target structure with a desired function
- Intrinsic disorder proteins or regions in a protein that do not have a unique three-dimensional structure as a monomer at physiological conditions
- Knowledge-based energy function an energy function derived from statistical or statistical mechanical analysis of known protein structures
- Physical-based energy function an energy function derived by the laws of physics that is composed of many approximate terms
- Energy function the scoring function that is minimized during iterative protein design
- Local interaction the interaction between amino acid residues that are sequence neighbours
- Nonlocal interaction the interaction between amino acid residues that are located close to each other in three-dimensional space but far from each other in their sequence positions

A designed metalloprotein using an unnatural amino acid

An example of computational protein design II.

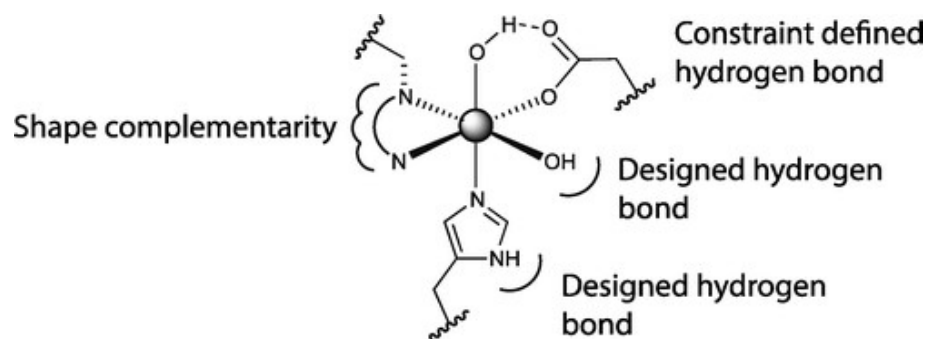
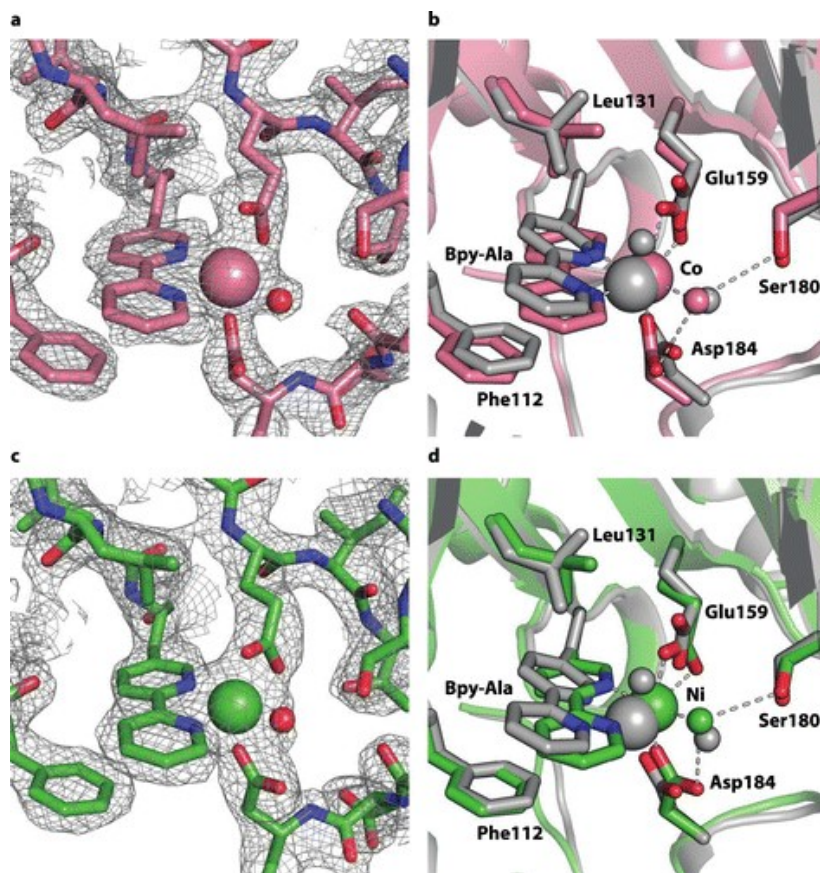
- Genetically encoded unnatural amino acids could facilitate the design of proteins and enzymes of novel function
- The Rosetta design was used to design metalloproteins in which the amino acid **(2,2'-bipyridin-5yl)alanine** (Bpy-Ala) is a primary ligand of a bound metal ion
- A buried metal binding site with octahedral coordination geometry consisting of Bpy-Ala, two protein-based metal ligands, and two metal-bound water molecules
- Experimental characterization revealed a Bpy-Ala-mediated metalloprotein with the ability to bind divalent cations including Co^{2+} , Zn^{2+} , Fe^{2+} , and Ni^{2+} , with a K_d for Zn^{2+} of ~ 40 pM



A designed metalloprotein using an unnatural amino acid

An example of computational protein design II.

- X-ray crystal structures of the designed protein bound to Co^{2+} and Ni^{2+} have RMSDs to the design model of 0.9 and 1.0 Å respectively over all atoms in the binding site.

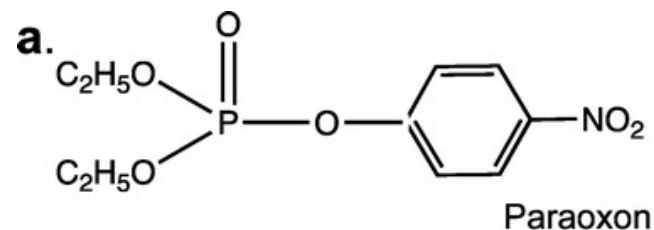


X-ray crystallographic analysis of MB_07 bound to Co^{2+} and Ni^{2+} . Electron density from a $2\text{Fo}-\text{Fc}$ map in the vicinity of the Bpy-Ala bound to Co^{2+} (a) and Ni^{2+} (c) contoured at 1.0σ . Density for Bpy-Ala, Co^{2+} or Ni^{2+} (pink and green spheres, respectively), D184, E159, and a metal-bound water molecule (red spheres) is visible.

Design to improve detoxification rates of nerve agents

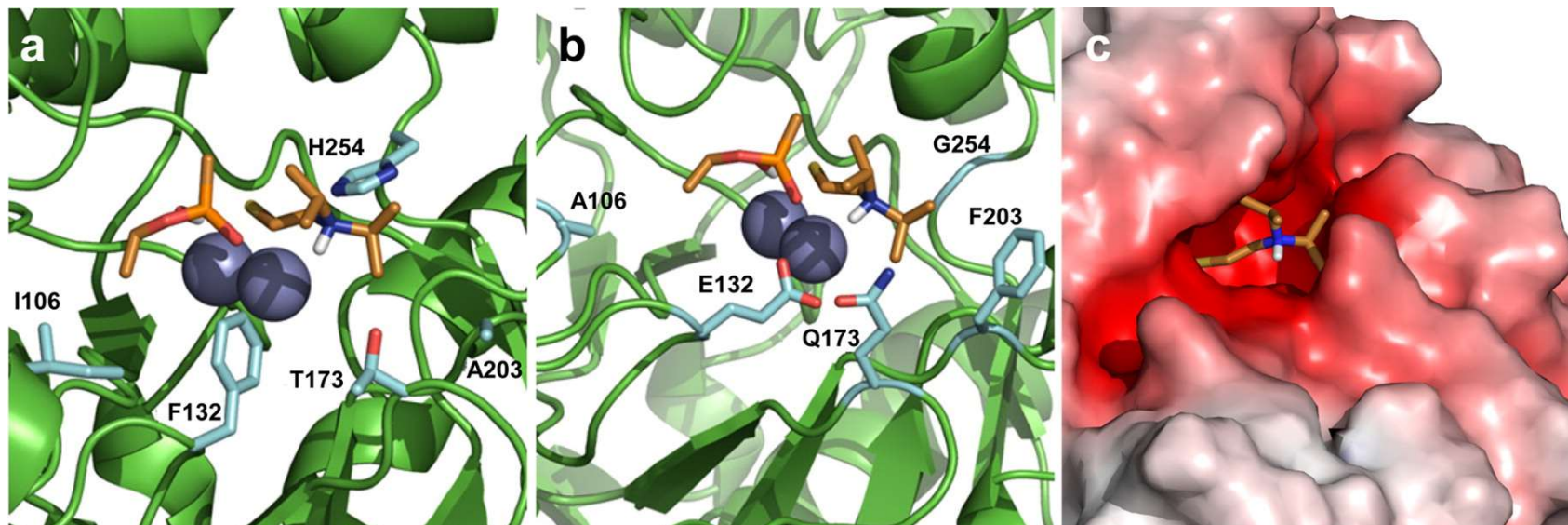
An example of computational protein design III.

- Organophosphate pesticides rapidly inactivate acetylcholinesterase and are the most toxic stockpile nerve agents
- An integrated computational and experimental approach was applied to increase *Brevundimonas diminuta* phosphotriesterase's (PTE) detoxification rate of V-agents by 5000-fold
- Computational models were built of the complex between PTE and V-agents. On the basis of these models, the active site was redesigned to be complementary in shape to VX and RVX and to include favorable electrostatic interactions with their choline-like leaving group
- five rounds of iterating between experiment and model refinement led to variants that hydrolyze the toxic SP isomers of all three V-agents with k_{cat}/K_M values of up to $5 \times 10^6 \text{ M}^{-1} \text{ min}^{-1}$ and also efficiently detoxify G-agents
- These new catalysts provide the basis for broad spectrum nerve agent detoxification



Design to improve detoxification rates of nerve agents

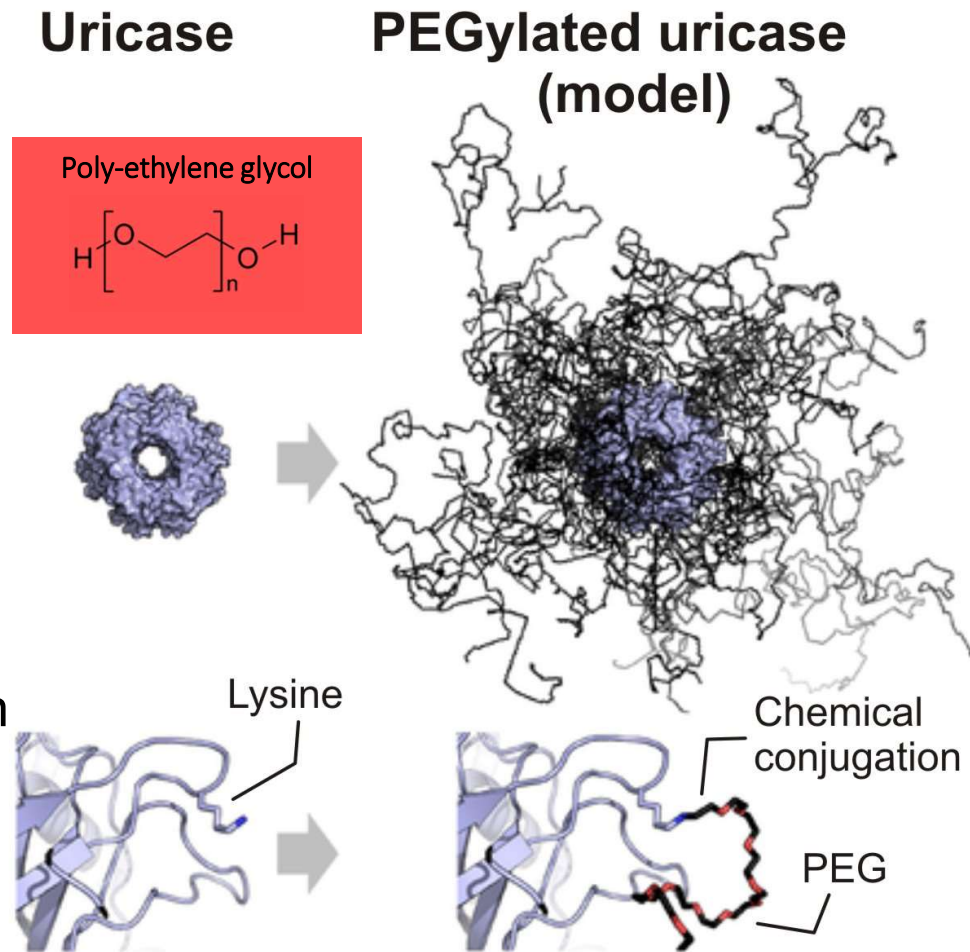
An example of computational protein design III.



- Computational models of wild-type PTE (a) and the 5th generation variant A53 (b) with the bound substrate model (the SP isomer of a VX-RVX hybrid). (c) The designed pocket of the A53 variant is complementary to VX's leaving group, including charge complementarity to the choline-like moiety

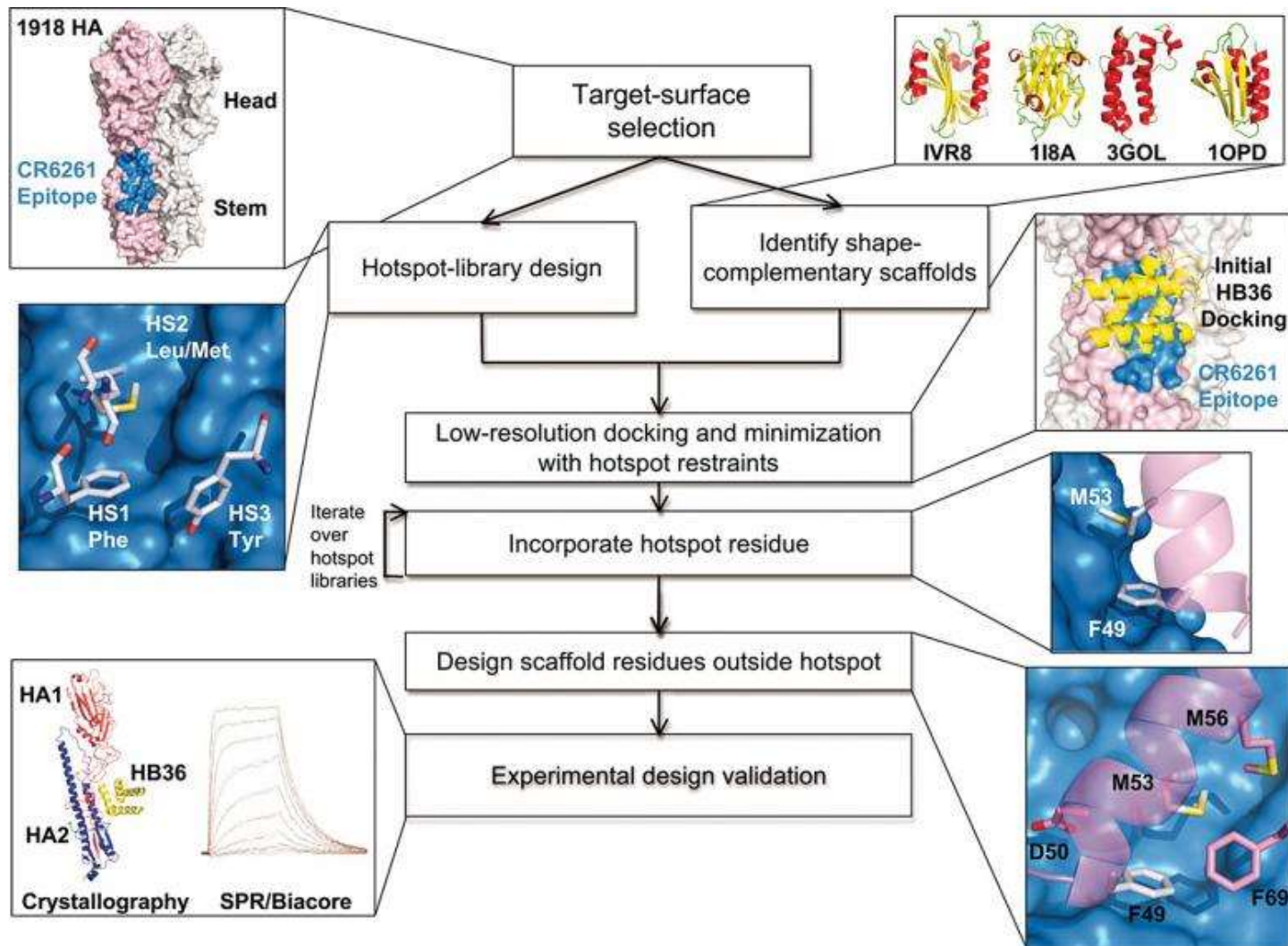
Uricase PEGylation to reduce immunogenicity

- Uricase is used for gout treatment, but its high immunogenicity is a limitation
- Attachment of PEG polymers via lysine coupling to reduce side-effects
- PEG number optimization
10-kDa optimal size
- 1000x reduced antigenicity upon PEGylation
- Improved solubility and increased serum half-life

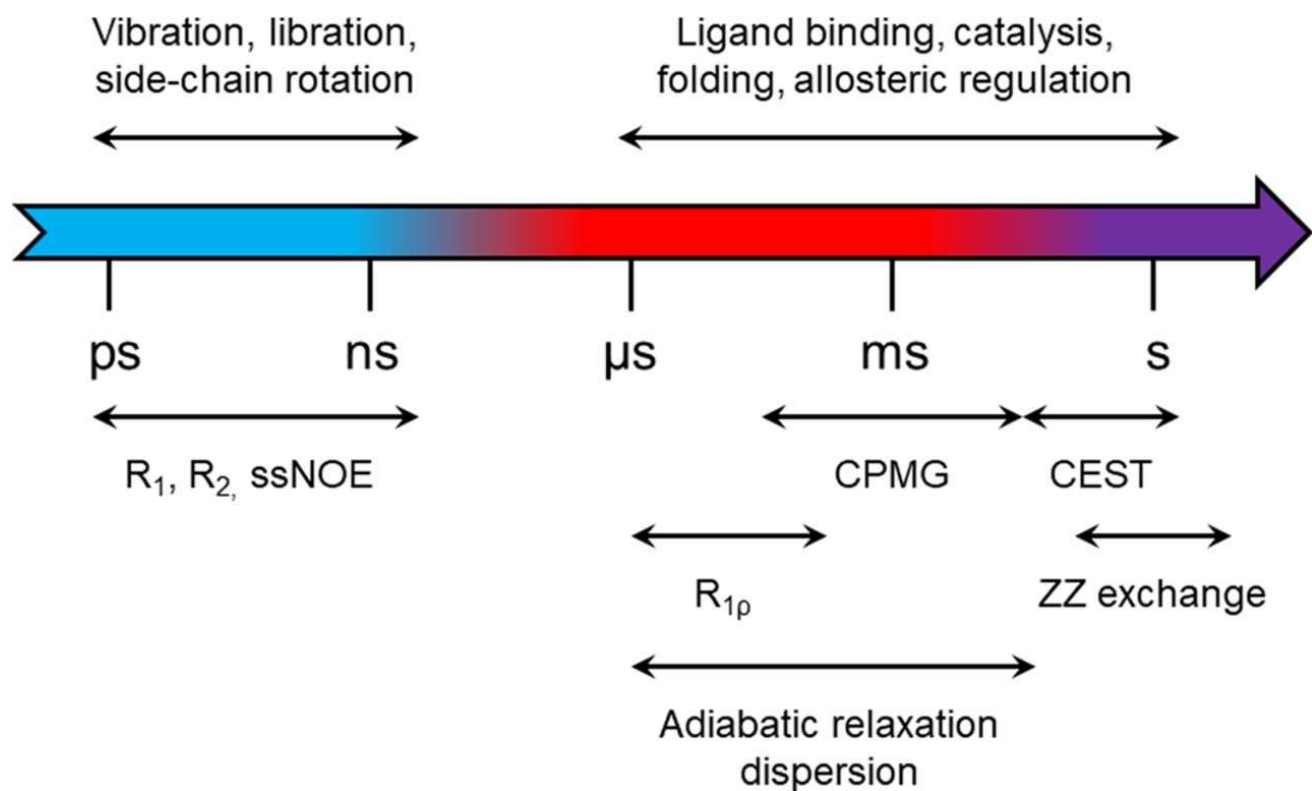


Sherman *et al.*, *Adv. Drug Deliv. Rev.* 60:59-68 (2008)

Flow chart of the key steps in the design of novel proteins



Protein dynamics

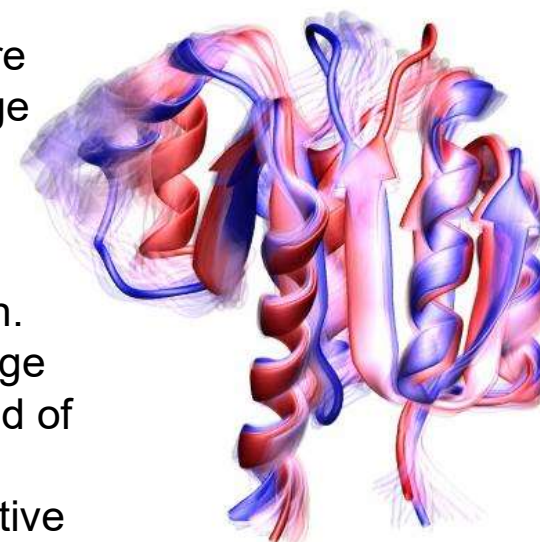


Chao and Byrd, *Em. Top. Life Sci.*, 2:93-105 (2018)

Design of proteins that exchange on functional timescales

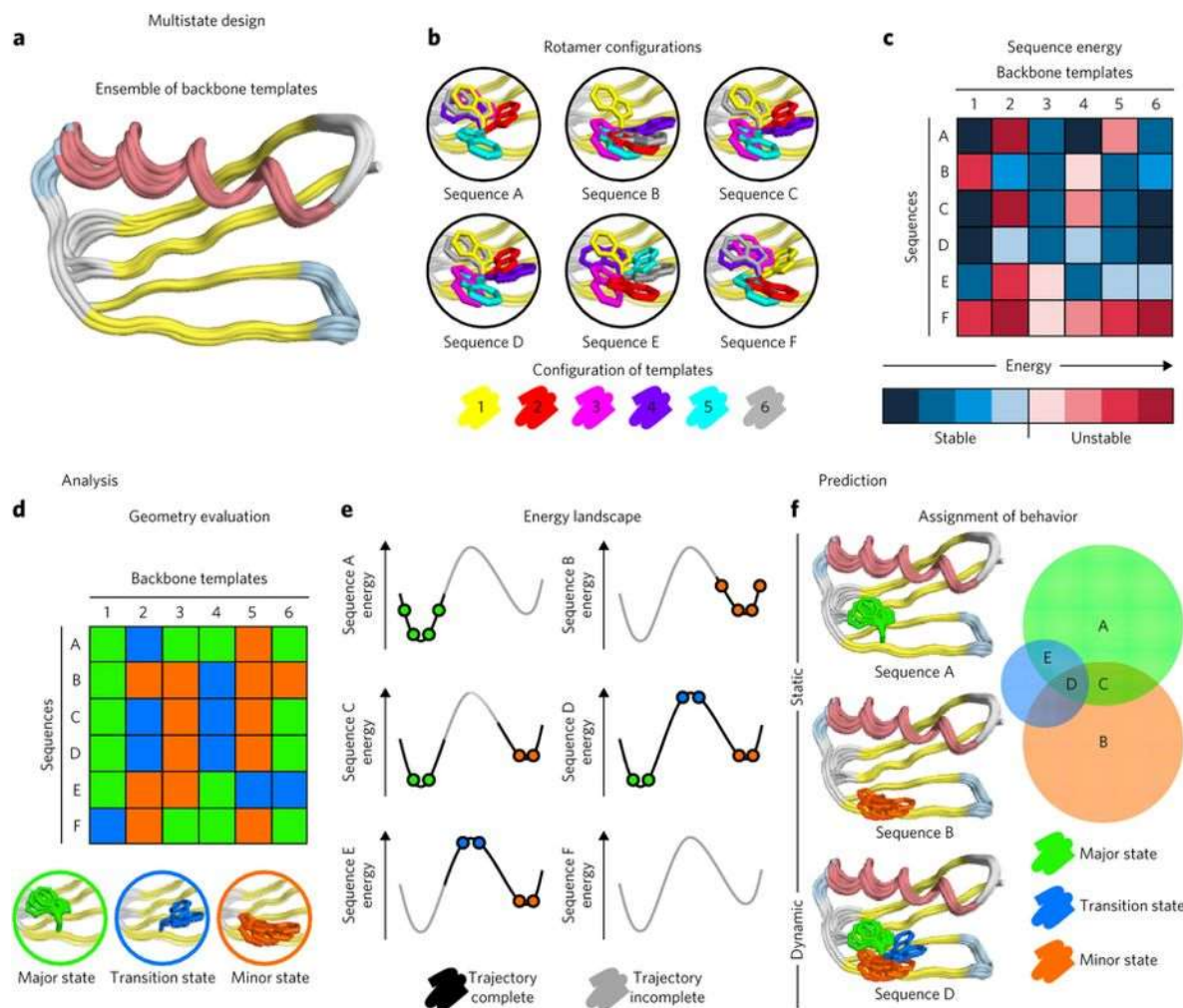
Davey et al., Nat. Chem. Biol. 13, 1280–1285 (2017)

- Proteins are intrinsically dynamic molecules that can exchange between multiple conformational states, enabling them to carry out complex molecular processes with extreme precision and efficiency.
- Attempts to design novel proteins with tailored functions have mostly failed to yield efficiencies matching those found in nature because standard methods do not allow the design of exchange between necessary conformational states on a functionally relevant timescale.
- A broadly applicable computational method was developed to engineer protein dynamics that we term *meta*-multistate design.
- This methodology was capable to design spontaneous exchange between two novel conformations introduced into the global fold of Streptococcal protein G domain β 1.
- The designed proteins, named DANCERs, for dynamic and native conformational exchangers, are stably folded and switch between predicted conformational states on the millisecond timescale.
- The successful introduction of defined dynamics on functional timescales opens the door to new applications requiring a protein to spontaneously access multiple conformational states.



The *meta*-multistate design framework for design of conformational exchange

- (a,b) Multistate design (MSD) with an ensemble of backbone templates approximating the conformational landscape for dynamic exchange between targeted states
- (c) MSD also returned an energy value for each microstate that reflects its predicted stability.
- (d,e) Geometry-based analysis of the rotamer-optimized microstates (d) allowed assignment of each microstate to major, minor, or transition state regions of the energy landscape (e).
- (f) Prediction of conformational dynamics was done based on an evaluation of the relative energies of these states. For *meta*-MSD to predict a sequence as dynamic, all three states must be stable and have an energy profile that is compatible with exchange (for example, sequence D).



Examples of computationally designed enzymes

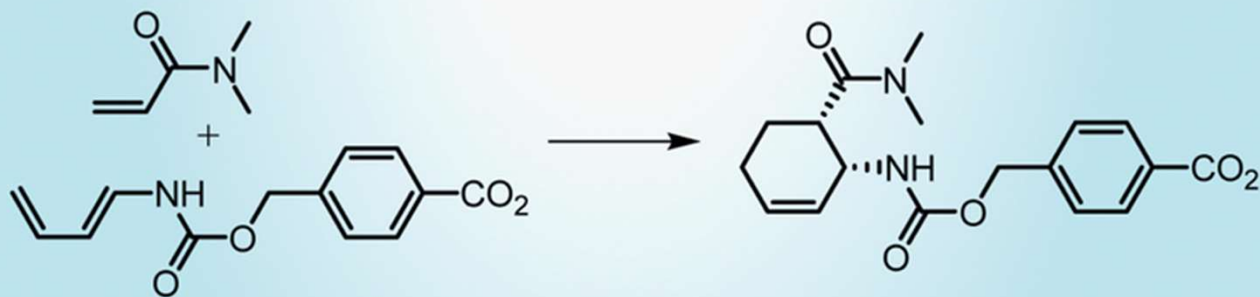
Kemp Elimination Enzyme



Retro-Aldolase



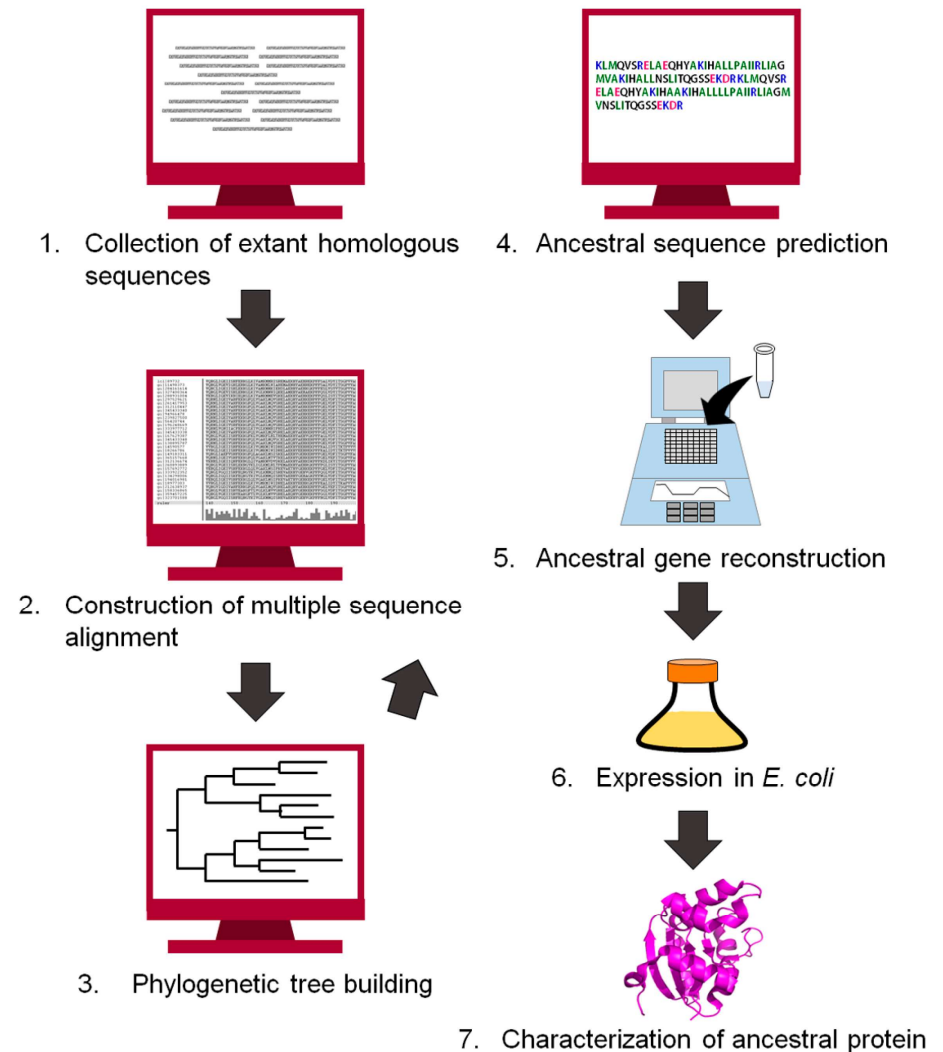
Diels-Alderase



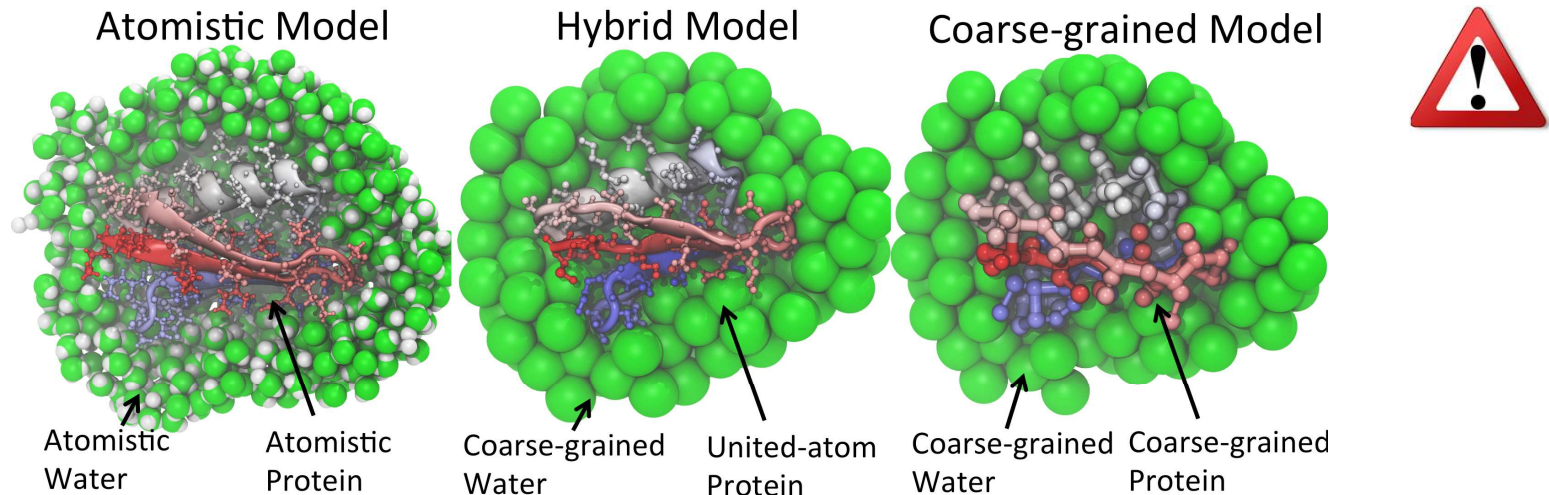
Design of ancestral proteins



1. The hypothesis suggests that ancestral proteins were able to withstand the harsh conditions prevalent on earth at that time
2. In addition to their **thermostability**, ancestral enzymes may have been **promiscuous with respect to substrates**. The evolution theory of proteins holds that current proteins evolved from low-specificity ancestral proteins
3. Because of their low specificity, the ancestral proteins evolved to become more efficient at using specific substrates
4. Thus, the reconstruction of ancestral sequences from multiple sequence alignments and phylogenetic trees may provide the **opportunity to change enzyme specificity**



Computational challenge in protein simulations



- Computer simulations can provide atomistic details of the processes that are hardly observed through experimental measurements. **Biological processes typically microsecond or even millisecond long**, need to be followed in computer simulations femtosecond by femtosecond, simulating such processes in every atomistic detail is computationally challenging.
- Despite ever-growing computing power, computer simulations cannot be used for the modelling of large biomolecular systems over time scales long enough to be of biological interest.
- To overcome the challenge, **coarse-grained models**, in which multiple atomistic sites are grouped into one site, have been developed. The models significantly reduce computational cost and, thereby, enhance the speed of simulation

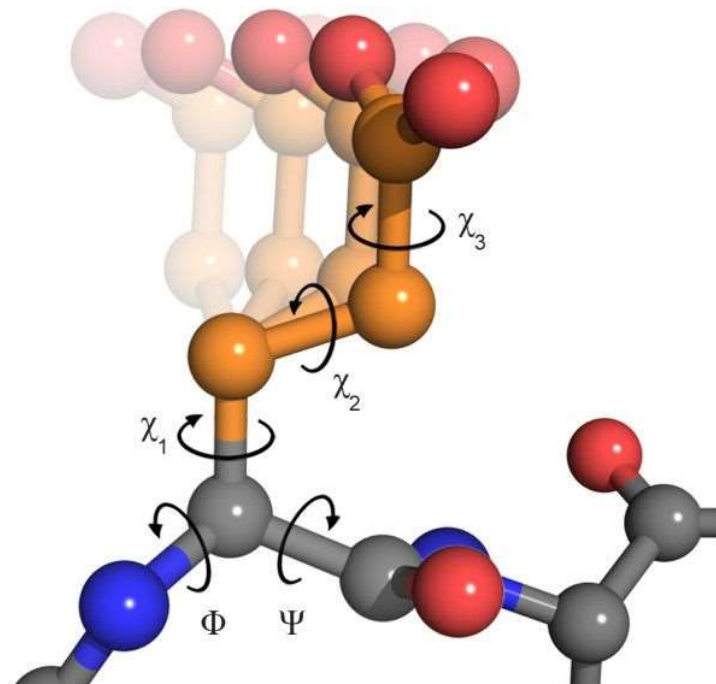
“High-resolution“ protein design workflow



- Protein design begins with a structure or complex and produces new sequences
- Design positions are chosen to be mutated. Next, the sequence may be aligned to other homologous sequences to produce biological constraints on the sequence space
- Sequence design is then performed and can be done using a single state or multiple states. In this step, the structure being designed can remain fixed, with only side-chain rotamers changing, or may be completely flexible
- The algorithms for sampling come from the same classes of techniques used in protein folding
- Designed sequences may then be clustered and evaluated with a more detailed scoring function
- Design produces one or many sequences that are predicted to fold into the input structure, often with enhanced biophysical characteristics

Rotamers

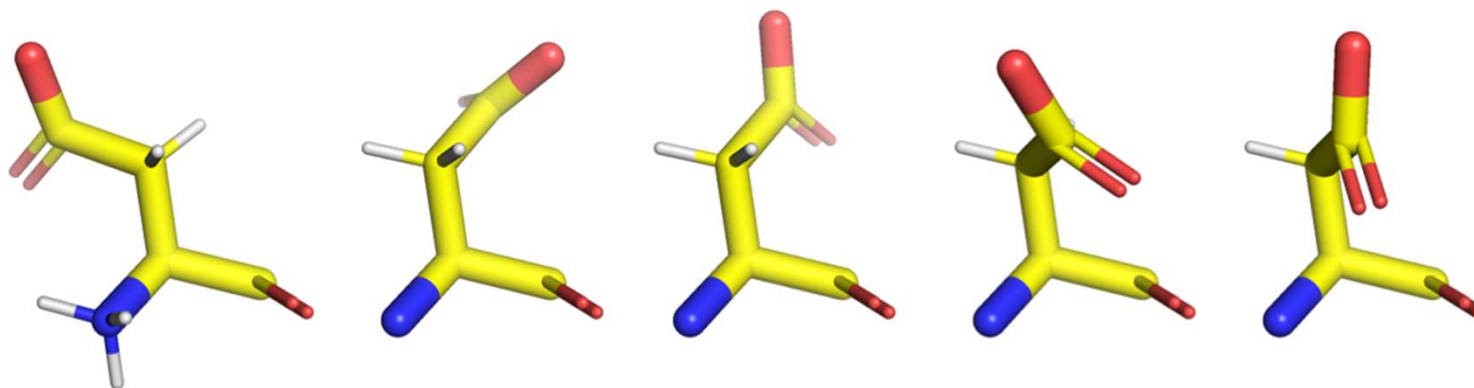
- Protein side chain may have many different conformations
- They are mostly defined by the dihedral angles (bond length and bond angle is relative fixed)
- Side chains are only permitted to adopt a discrete set of statistically preferred conformations: **rotamers**
- The figure shows dihedral angles in glutamate: dihedral angles are the main degrees of freedom for the backbone (ϕ and ψ angles) and the side chain (χ angles) of an amino acid. The number of χ angles varies between zero and four for the 20 standard amino acids. The figure shows a ball-and-stick representation of glutamate, which has three χ angles. The fading conformations in the background illustrate a rotation around χ_1 .



Rotamer libraries

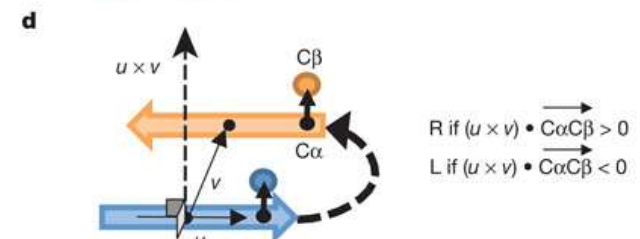
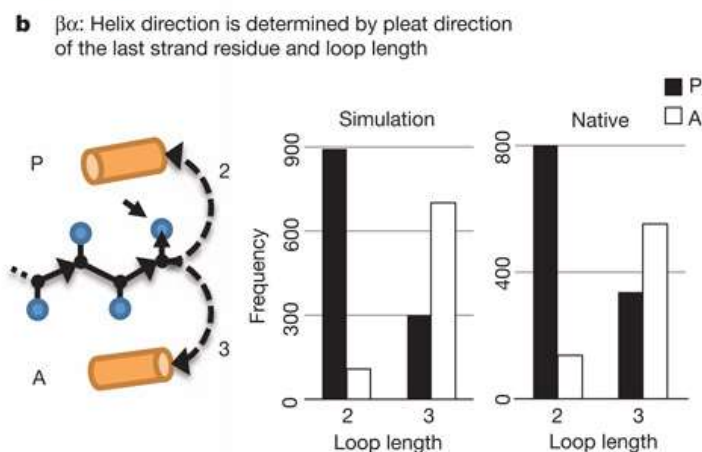
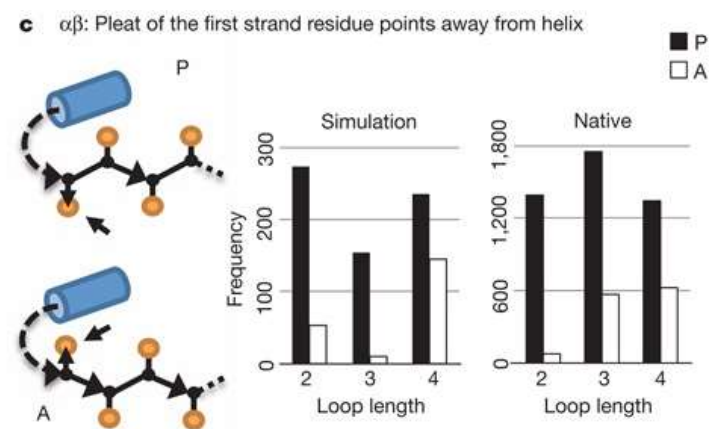
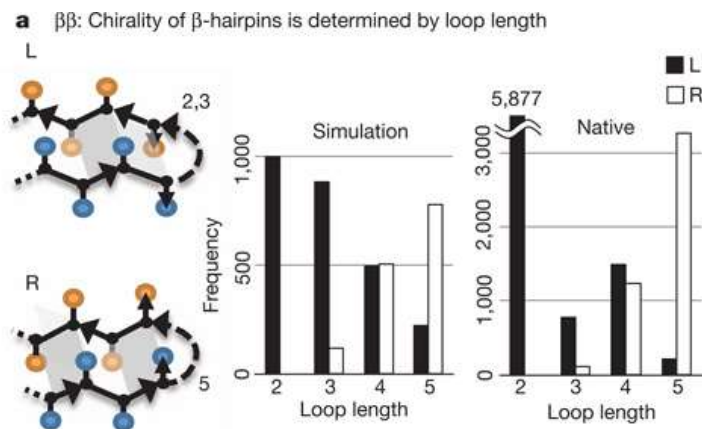
- Rotamer libraries were compiled by clustering the side chains of each amino acid over the whole database. Each cluster is a representative conformation (rotamer), and is represented in the library by the best side-chain angles

Five Rotamers of Aspartate



Principles for designing ideal protein structures

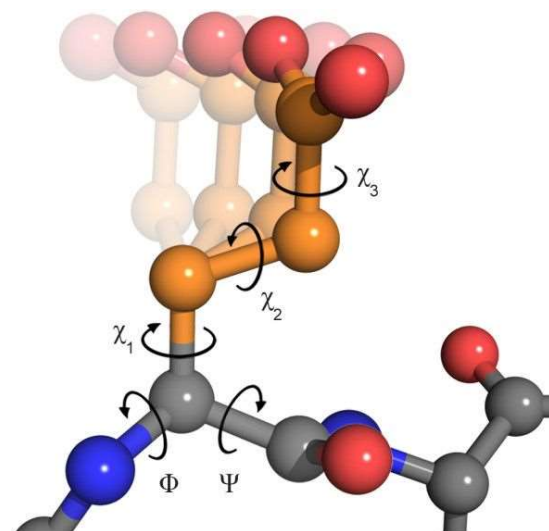
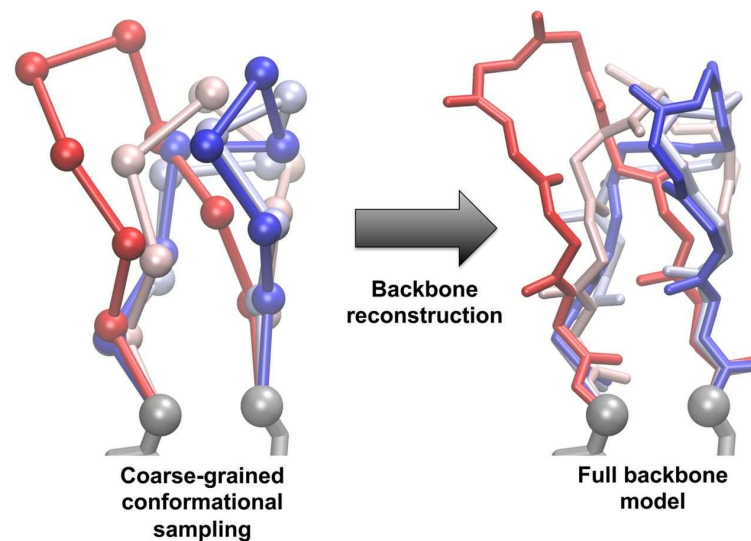
- Fundamental rules:** (a), $\beta\beta$ -rule. L (left-handed) and R (right-handed) $\beta\beta$ -units are illustrated, (b), $\beta\alpha$ -rule. P (parallel) and A (antiparallel) $\beta\alpha$ -units are illustrated. (c), $\alpha\beta$ -rule. (d), Chirality (L versus R) of a $\beta\beta$ -unit. The chirality is defined on the basis of the orientation of the $C\alpha$ -to- $C\beta$ vector.



Koga *et al.*, Nature 491: 222-227 (2012)

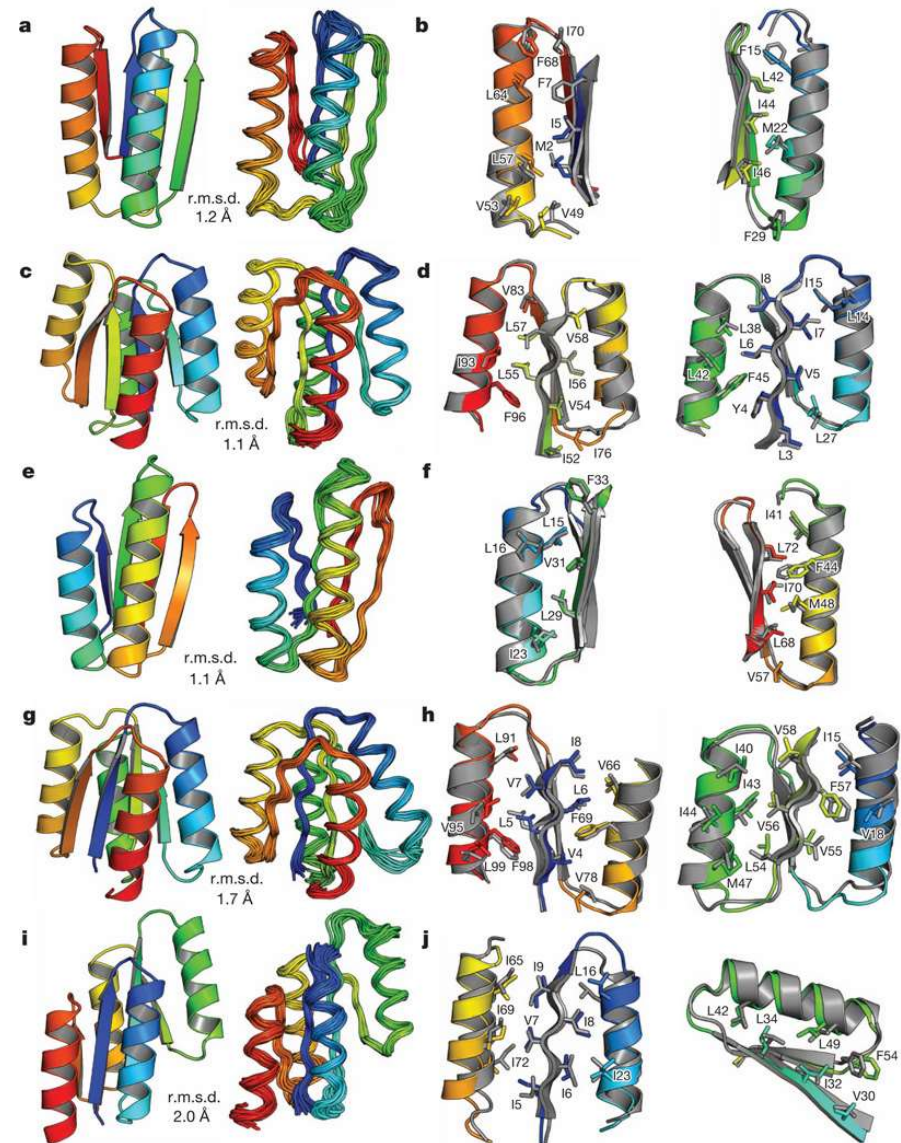
Energy functions and molecular force fields

- In structure-based computational protein design, folds are represented by the spatial coordinates of the backbone atoms or design scaffold
- Protein design is done by amino acid side chain along the scaffold
- Side chains are only permitted to adopt a discrete set of statistically preferred conformations: rotamers
- Rotamer/backbone and rotamer/rotamer interaction energies are tabulated
- These potential energies can then be approximated by using any of the standard force fields: CHARMM, AMBER, GROMOS



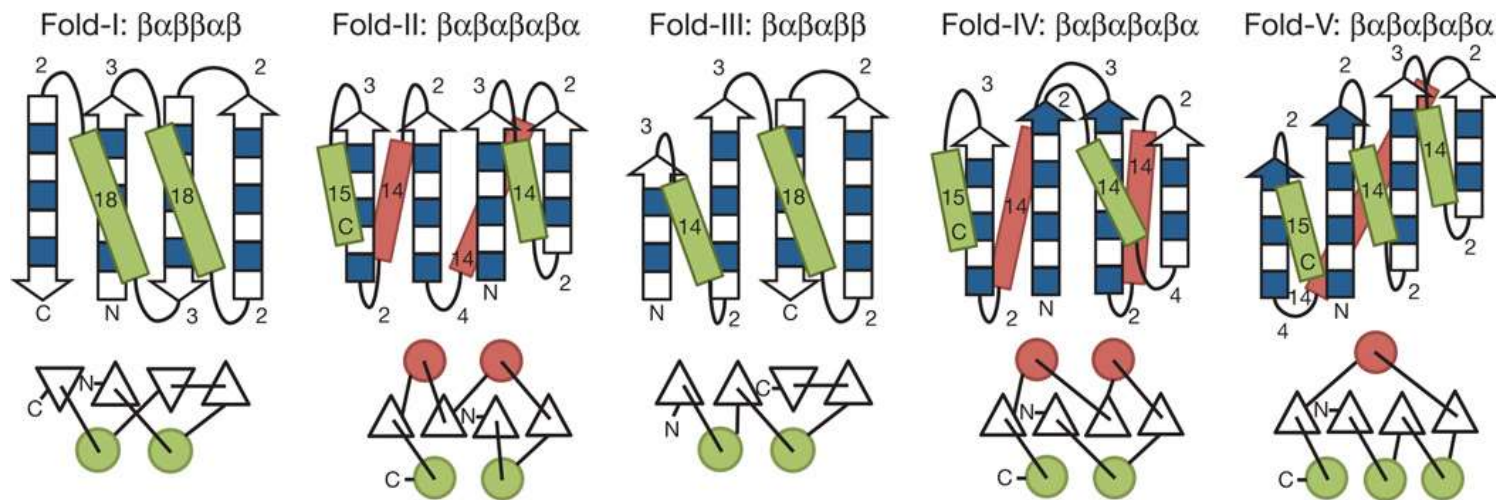
Comparison of models with experimental structures

- Experimental validation by NMR
- Comparison of overall topology. Design models (left) and NMR structures (right); the C α root mean squared deviation (r.m.s.d.) between them is indicated
- Comparison of core side-chain packing in superpositions of design models (rainbow) and NMR structures (grey)
- These results illuminate how the folding funnels of natural proteins arise and provide the foundation for engineering a new generation of functional proteins free from natural evolution



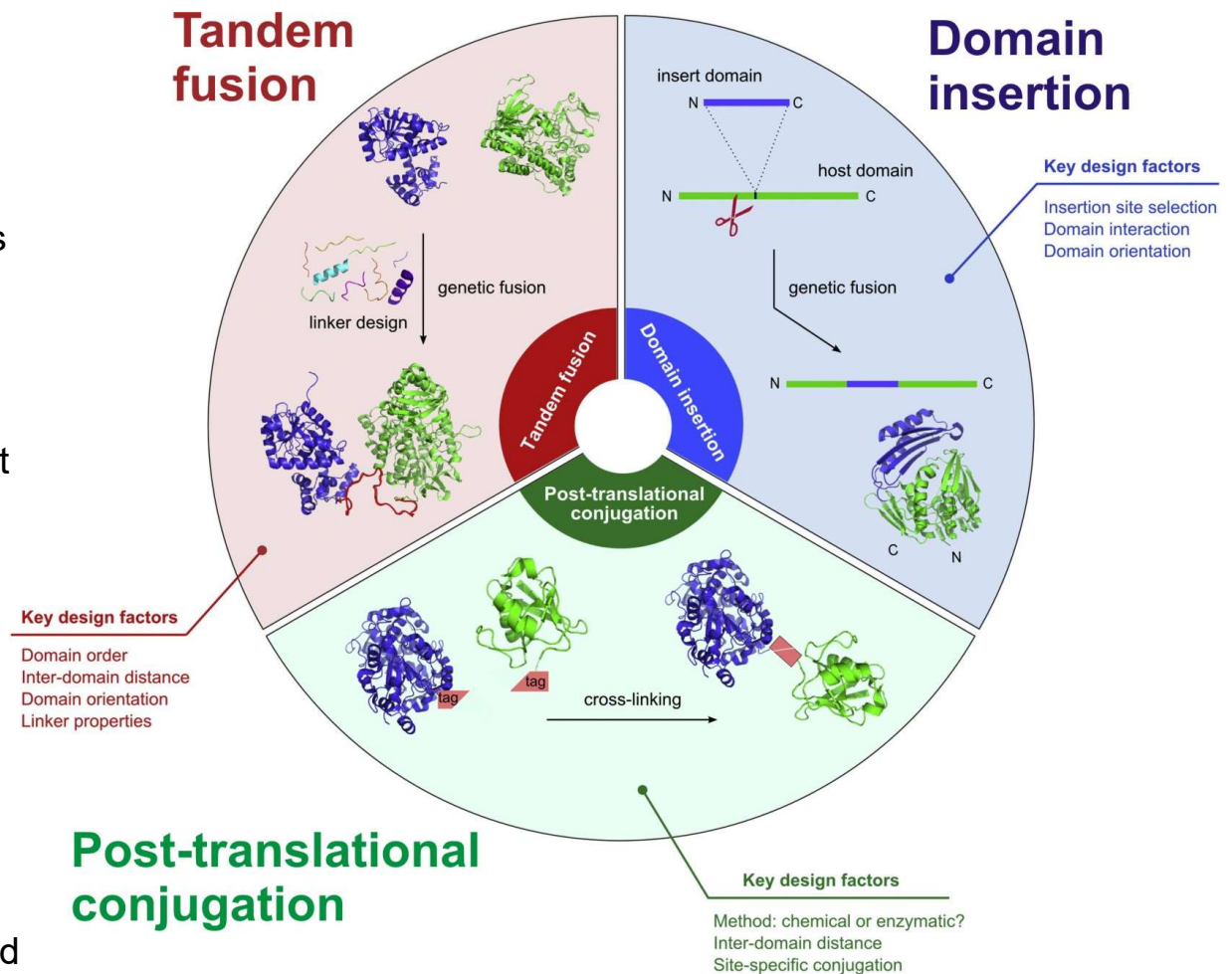
Principles for designing ideal protein structures

- A new approach based on a set of rules relating secondary structure patterns to protein tertiary motifs, which make possible the design of funnel-shaped protein folding energy landscapes leading into the target folded state
- Guided by these rules, they designed sequences predicted to fold into ideal protein structures consisting of α -helices, β -strands and minimal loops
- Designs for five different topologies were found to be monomeric and very stable and to adopt structures in solution nearly identical to the computational models



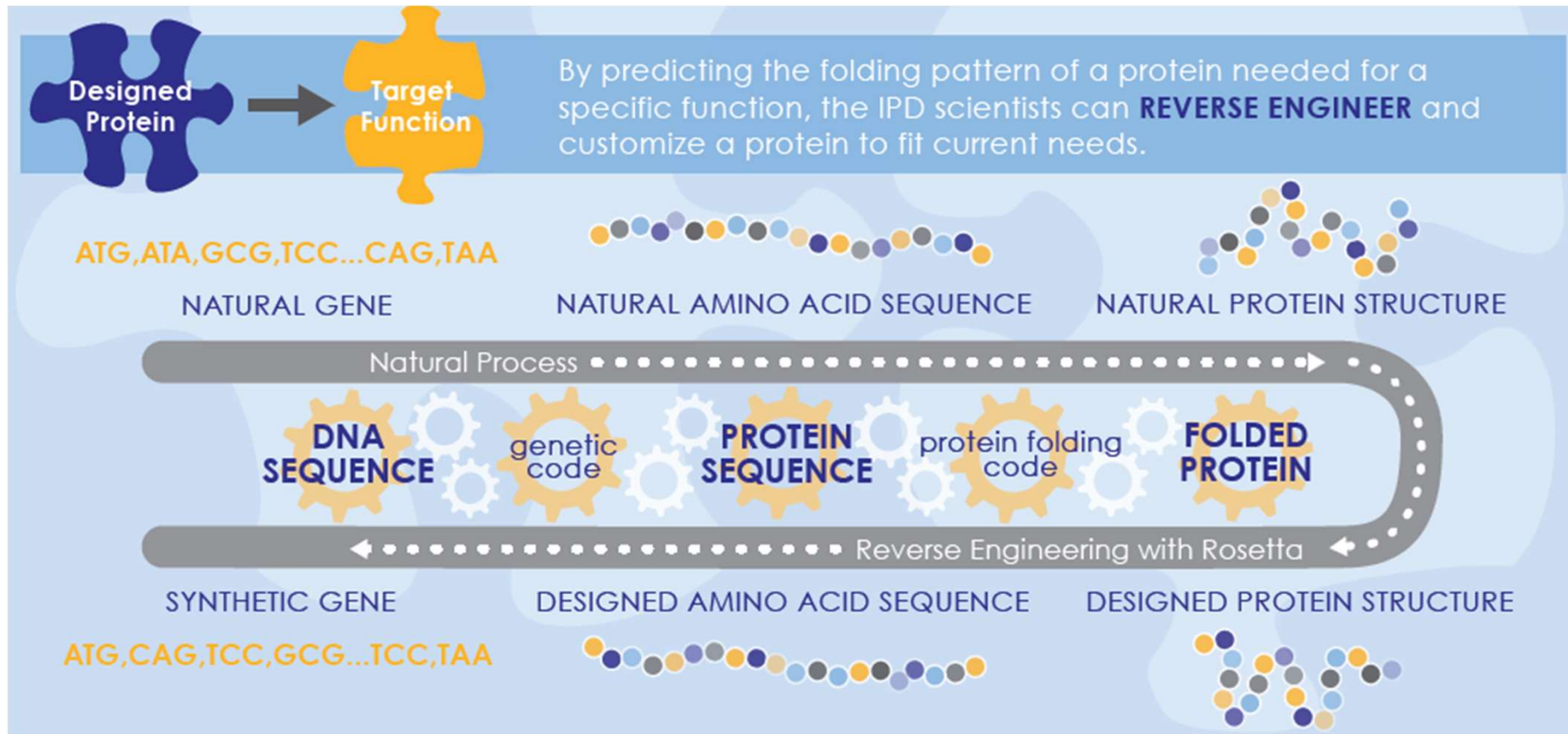
Strategies for the construction of synthetic fusion proteins

1. Linker-mediated tandem fusion is achieved by joining two proteins in a head-to-tail manner with a linker peptide in between, which can be selected from natural linker reservoir or artificially designed to separate fusion partners as well as to maintain favorable interactions between them.
2. Domain insertion is implemented by inserting one domain into a host domain through carefully selected recombination sites.
3. Post-translational conjugations used to combine separately expressed proteins to form a branched architecture, in which chemical reagents or enzymes such as transglutaminase and sortase A that recognize specific amino acids or sequences are used to cross-link these tagged proteins.



Computational protein design: what is it?

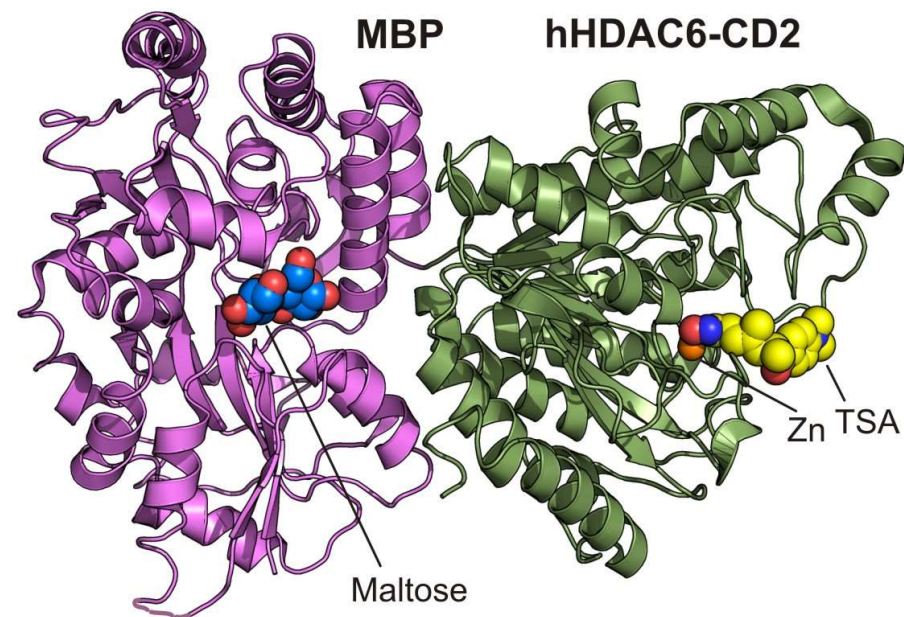
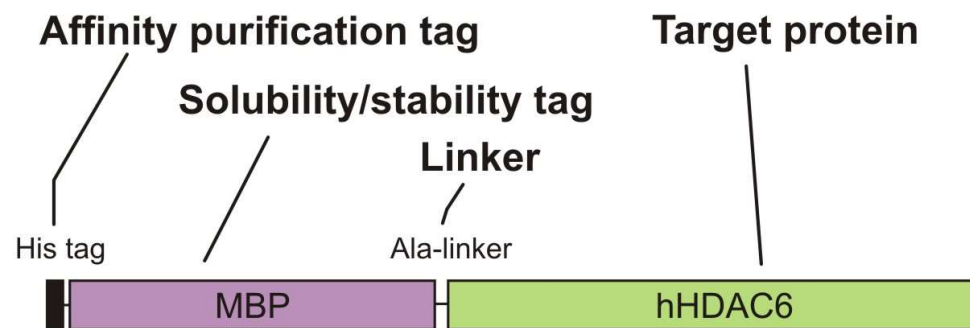
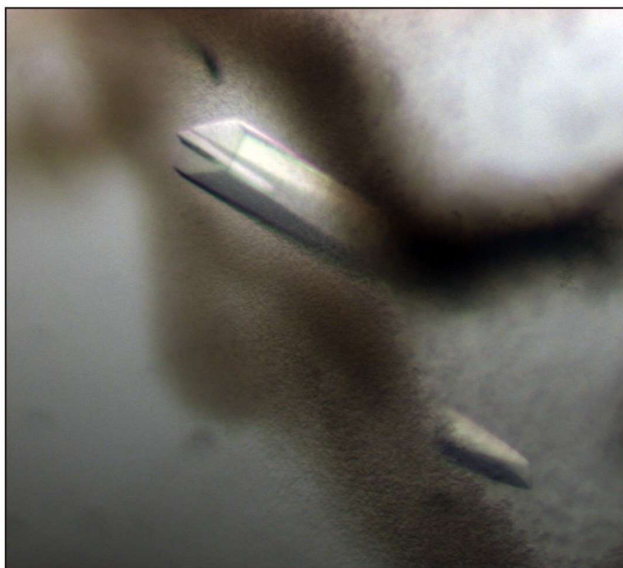
- Dissecting the rules that govern protein structures
- Implementation of these rules into a computer program, **Rosetta, Robetta**
- Cracking the protein folding code enables to model protein structures and design new proteins with desired properties



Engineered fusion proteins help protein crystallographers

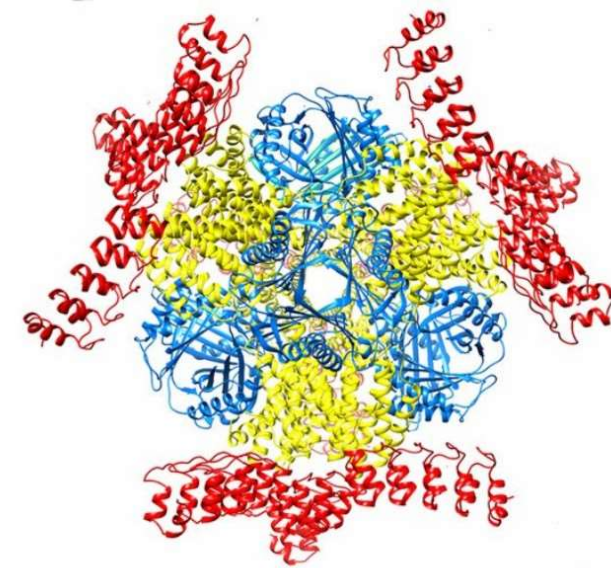
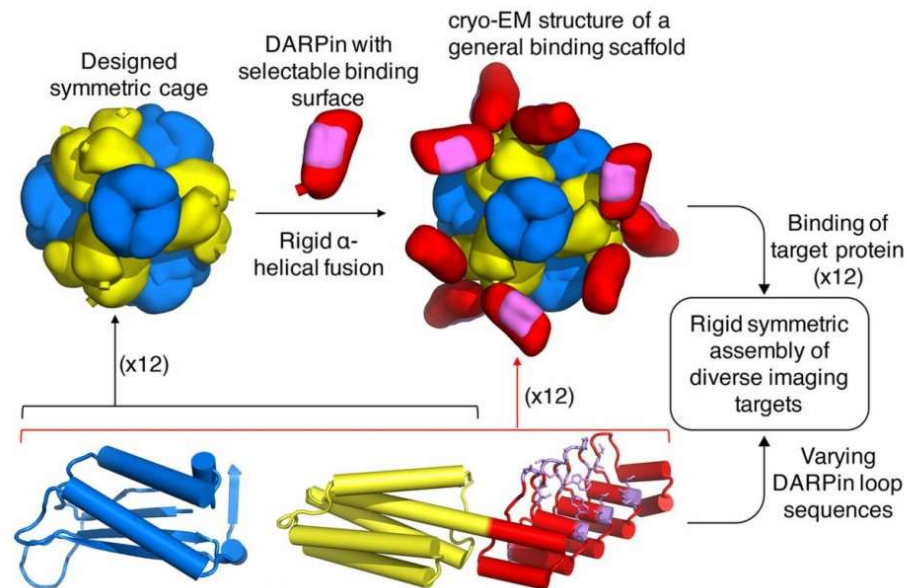
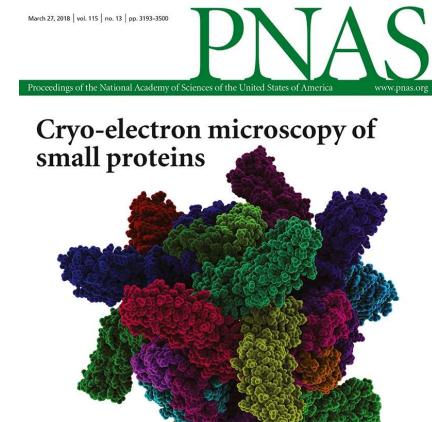
Fusion tags used:

- Thioredoxin (Thx)
- Maltose-binding protein (MBP)
- Glutathione S-transferase (GST)
- Small ubiquitin-like modifier (SUMO)
- Polyhistidenes (6xHis, 12xHis)



Imaging of small proteins displayed on protein scaffolds

- New electron microscopy (EM) methods are making it possible to view the structures of large proteins and nucleic acid complexes at atomic detail, but the methods are difficult to apply to molecules smaller than approximately 50 kDa.
- This limit can be successfully visualized when it is attached to a large protein scaffold designed to hold 12 copies of the attached protein in symmetric and rigidly defined orientations.





The Rosetta modelling software: overview

- The Rosetta software suite includes **algorithms for computational modelling and analysis of protein structures**. It has enabled notable scientific advances in computational biology, including *de novo* protein design, enzyme design, ligand docking, and structure prediction of biological macromolecules and macromolecular complexes.
- Rosetta is available to all non-commercial users for free and to commercial users for a fee.
- Rosetta development began in the laboratory of Dr. David Baker at the University of Washington as a structure prediction tool but since then has been adapted to solve common computational macromolecular problems.
- Development of Rosetta has moved beyond the University of Washington into the members of RosettaCommons, which include government laboratories, institutes, research centers, and partner corporations.
- The Rosetta community has many goals for the software, such as:
 - Understanding macromolecular interactions
 - Designing custom molecules
 - Developing efficient ways to search conformation and sequence space
 - Finding a broadly useful energy functions for various biomolecular representations