how can machines learn protein engineering ?

– one possible approach of machine learning concept to predict and improve enzyme activities.

winterseminar-Bled 2016-02-19 mark doerr university greifswald

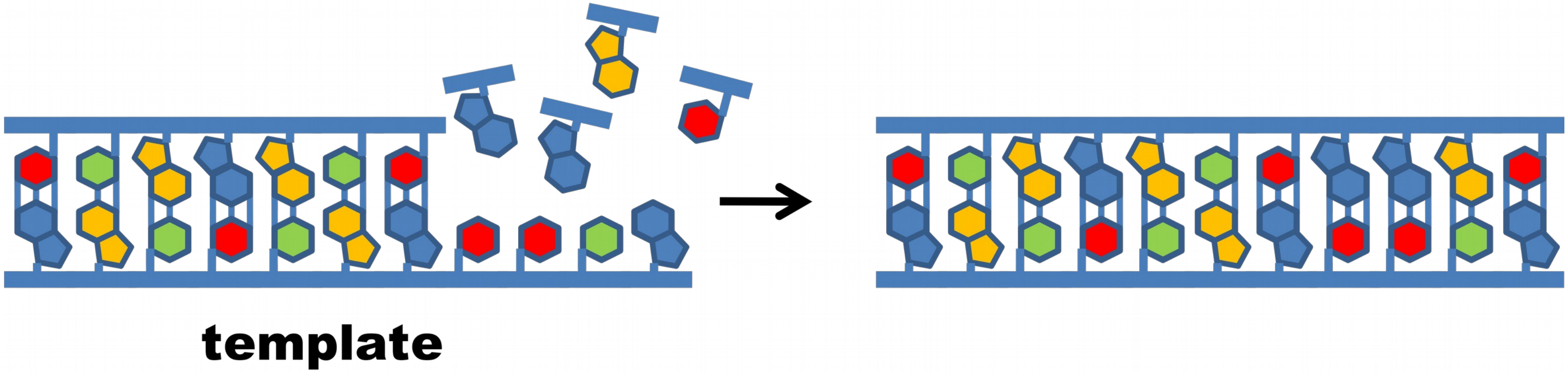
before we start talking about enzymes,
we should talk about RNA

RNA condensation and template directed extension reactions

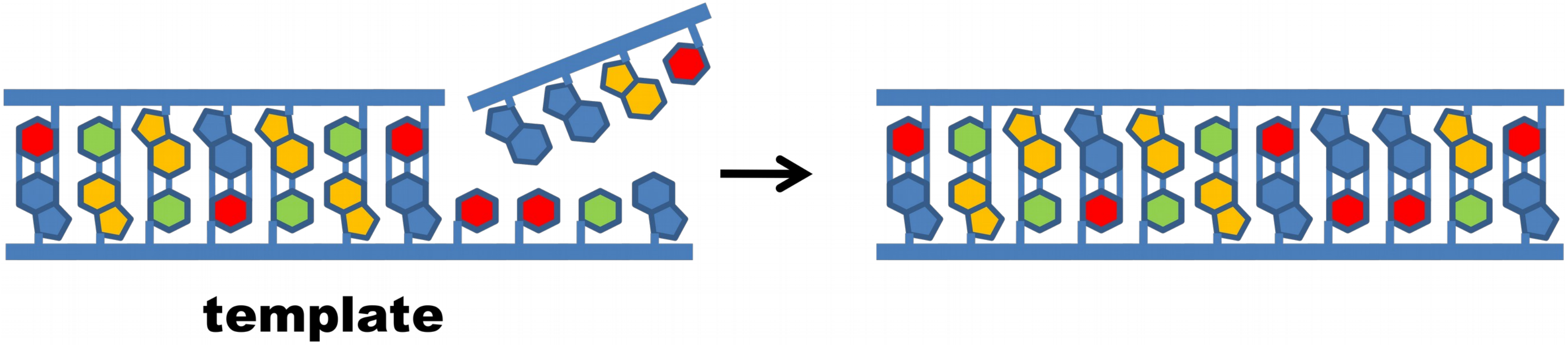
work performed at University of Southern Denmark Odense (SDU)
in **Steen Rasmussen**'s and **Pierre-Alain Monnard**'s group
together with **Philipp M.G. Löffler** (main experimenter)

template-directed nucleotide condensation/ligation

(a) Primer extension with monomers



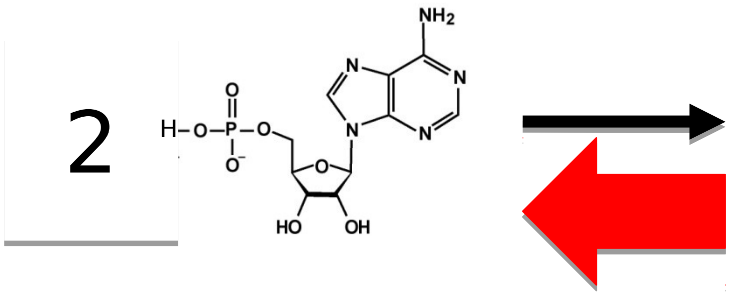
(b) Template directed ligation



condensation of RNA monomers

$$\mathbf{M + M \rightarrow M\text{-}M + M \xrightarrow{\text{n steps}} (M)_{2+n} + n\ H_2O}$$

In aqueous solution condensation reactions are not favored,
BUT the reverse reaction, DECOMPOSITION, is favored

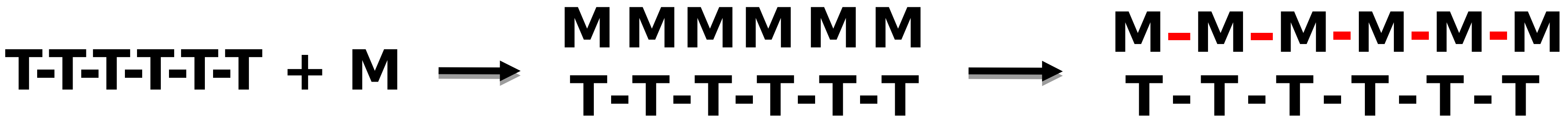


pApA + H_2O

Activation of monomers, compartmentalization and/or CATALYST (e.g. METAL ions, surface,….) are needed !

RNA self-replication in HOMOGENEOUS aqueous phase

Monomers template-directed polymerization



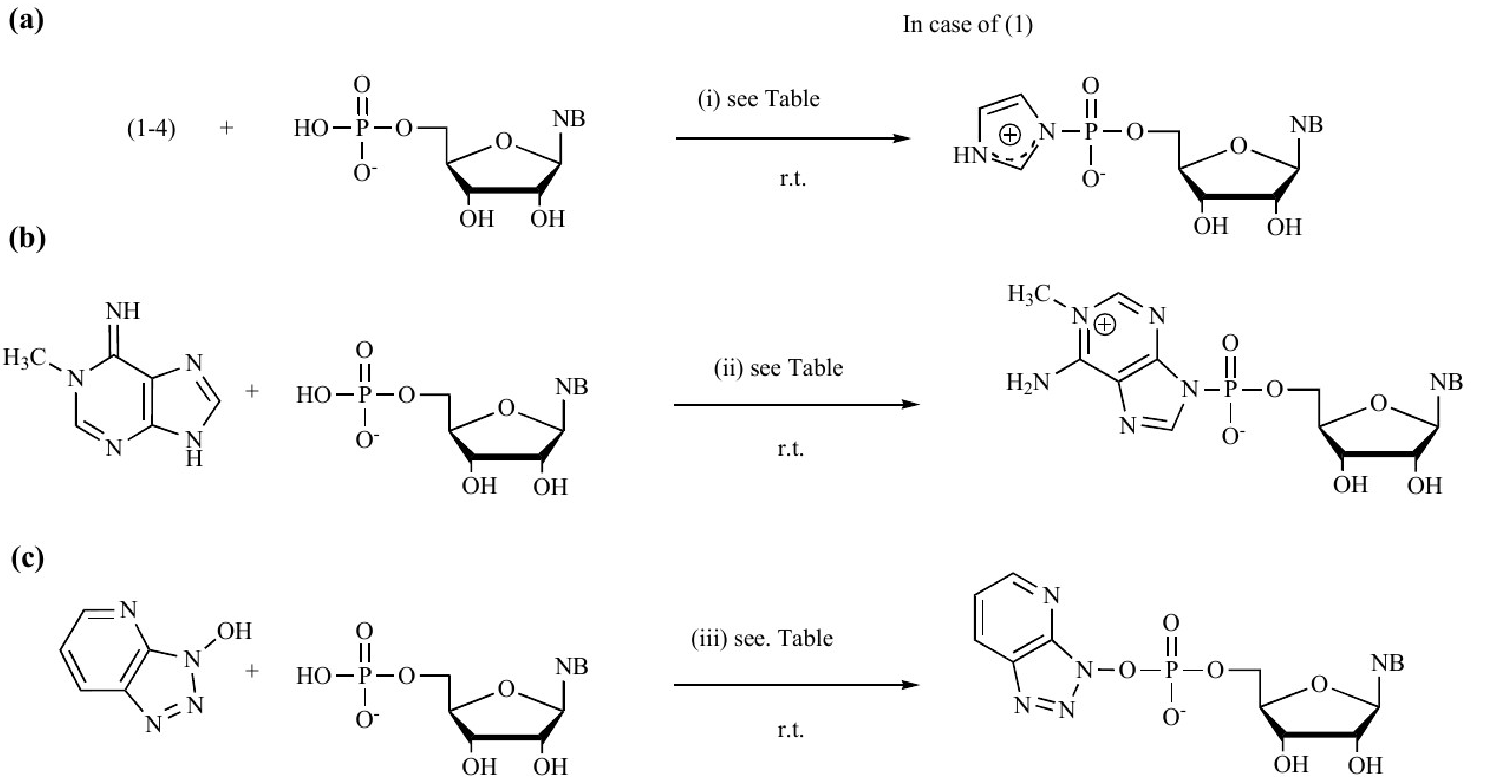
T-T-T-T-T-T + M → M M M M M M / T-T-T-T-T-T → M-M-M-M-M-M / T-T-T-T-T-T

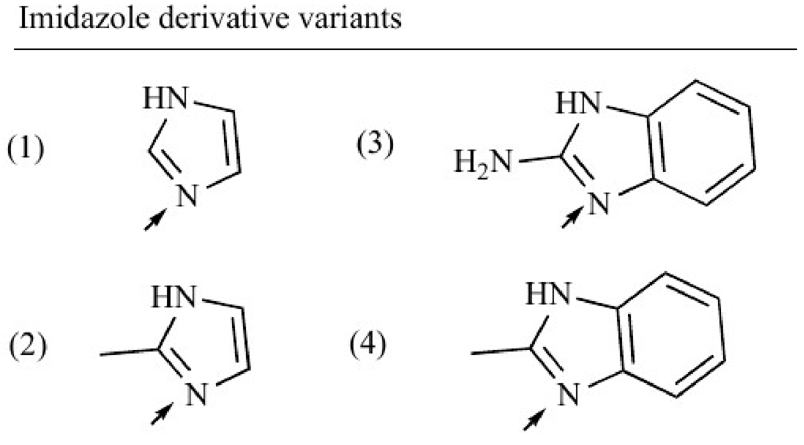
Orgel et al. (1980-1997)

Template with at least 60 % C ⇨ Poly G efficiently formed

Consecutive AA (UU to be polymerized) ⇨ Block (U too hydrophilic)

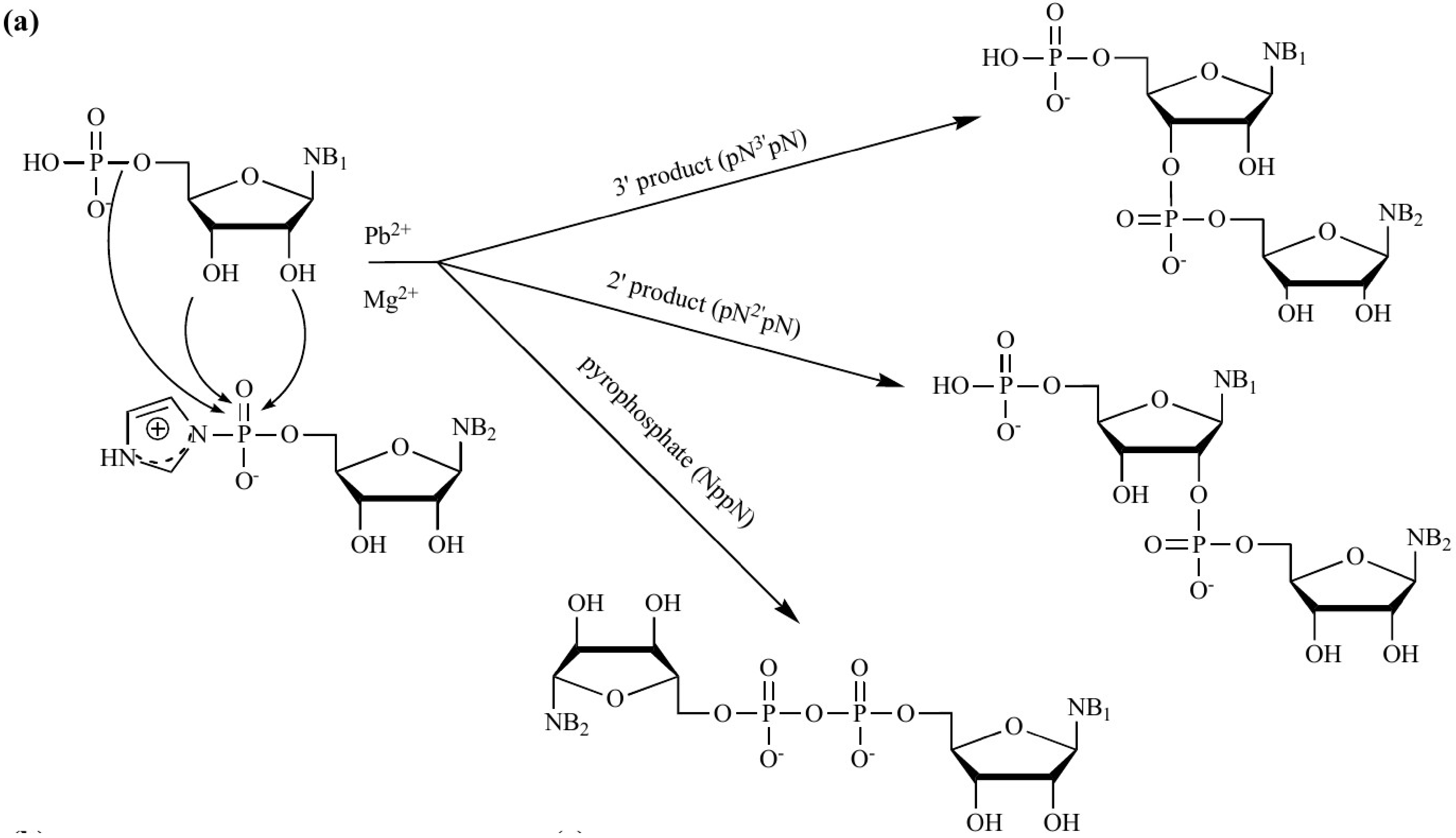
⇨ **NO AMPLIFICATION POSSIBLE WITH MONOMERS**





Coupling agents	Conditions
(i) $(PyS)_2$	DMF/DMSO, $P(Ph)_3$, $(Et)_3N$, Ar, dry
(ii) EDC	H_2O, NaOH, pH 5
(iii) HATU	DMF, DIEA, Ar, dry

Doerr, M.; M.G. Loffler, P.; Monnard, P.-A. Current Organic Synthesis 2012, 9 (6), 735–763.



Doerr, M.; M.G. Loffler, P.; Monnard, P.-A. Current Organic Synthesis 2012, 9 (6), 735–763.

Condensation of nucleotides (ImpU) on a template

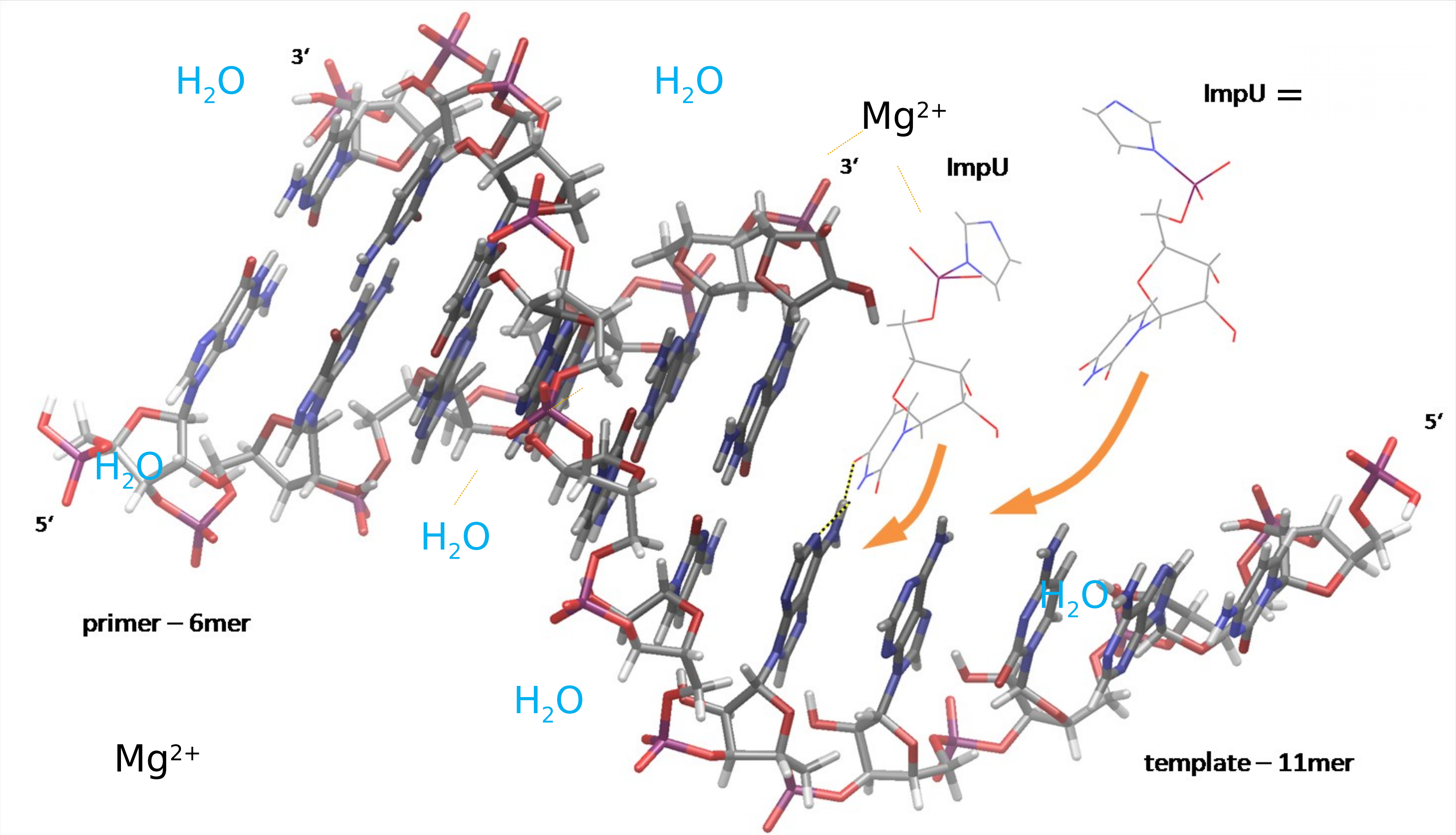
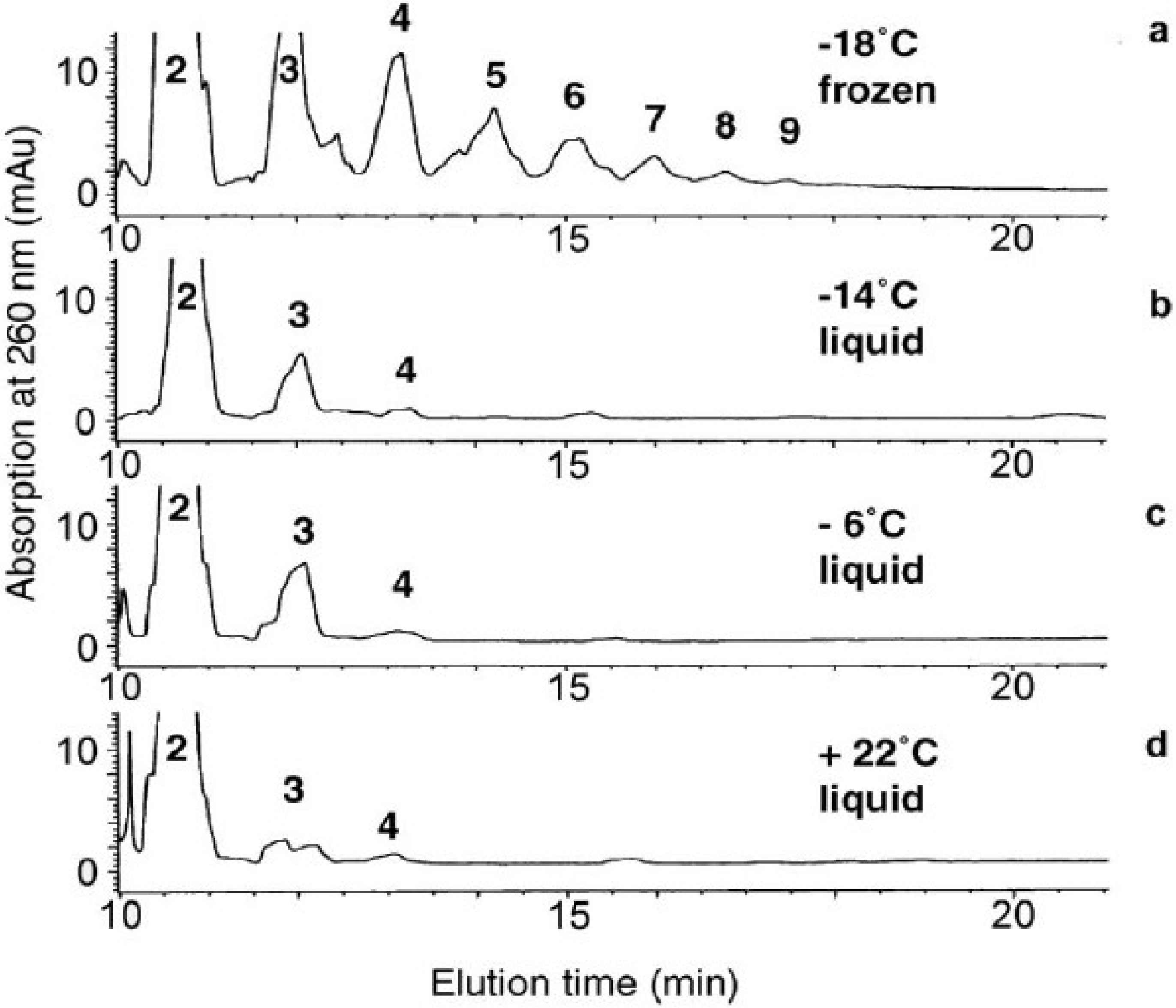


Illustration of possible conformations Simple MM2 simulation

Doerr, M.; M.G. Loffler, P.; Monnard, P.-A. Current Organic Synthesis 2012, 9 (6), 735–763.

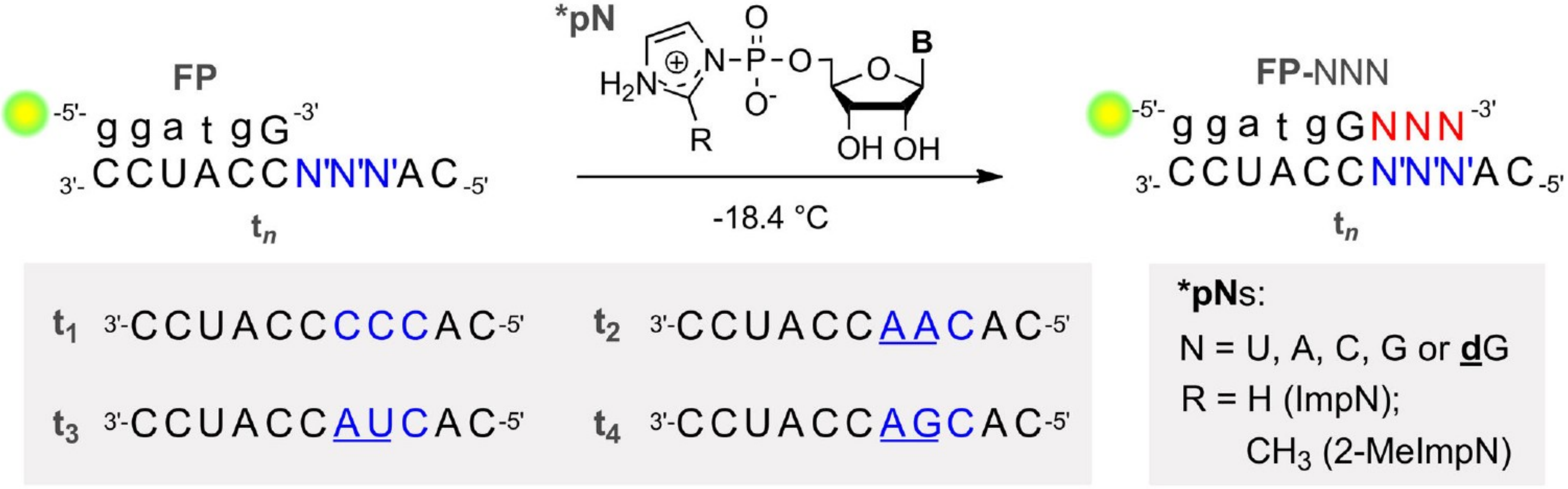
temperature dependence of oligomerization



Doerr, M.; M.G. Loffler, P.; Monnard, P.-A. Current Organic Synthesis 2012, 9 (6), 735–763.

eutectic ice

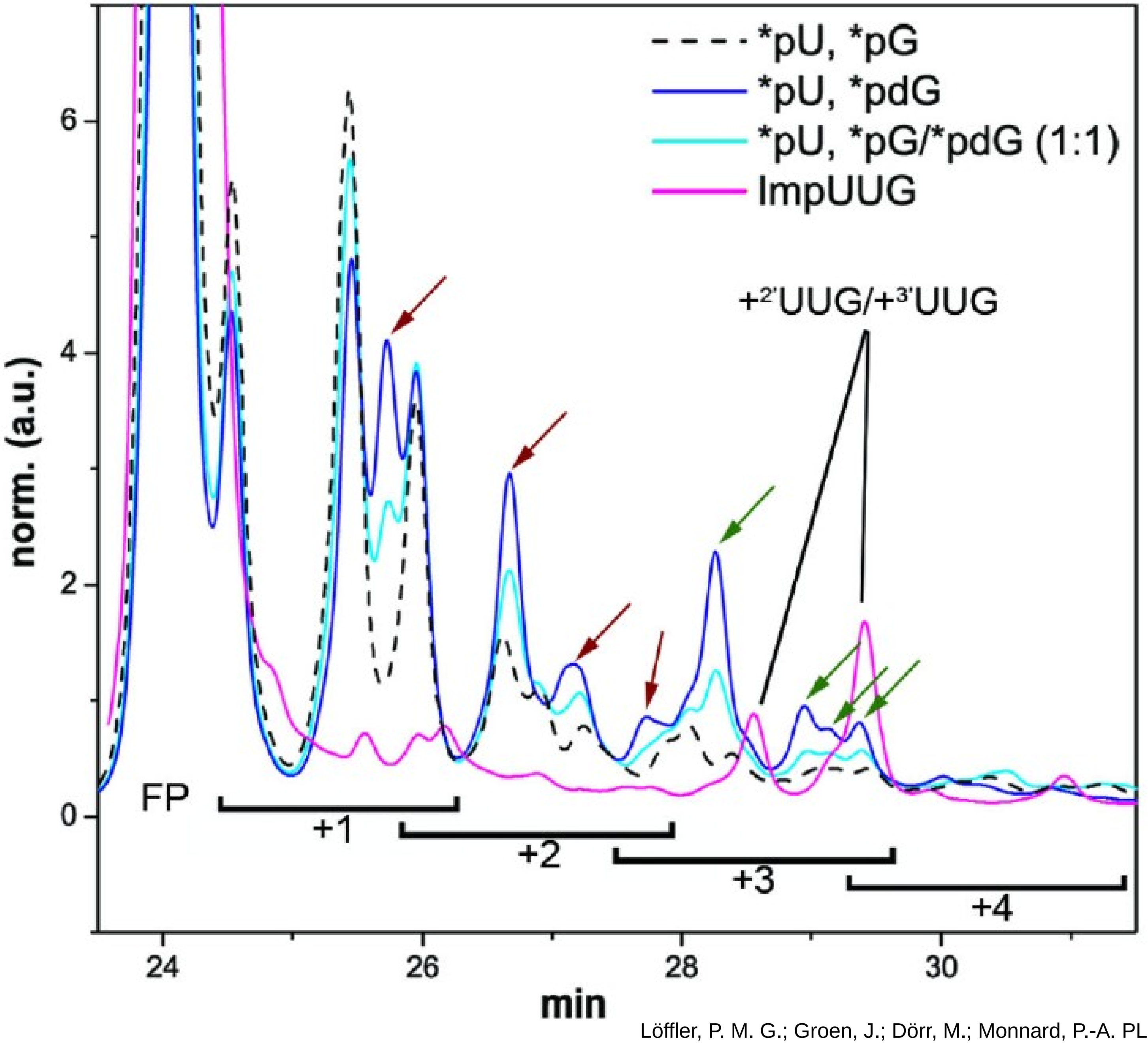
primer extension experiment designs


*pN
O
N
P
O
O
B
H2N
O-
R
OH OH
FP
-5'- g g a t g G -3'
3'- C C U A C C N'N'N' A C -5'
tn
-18.4 °C
FP-NNN
-5'- g g a t g G N N N -3'
3'- C C U A C C N'N'N' A C -5'
tn
t1 3'-C C U A C C C C C A C-5'
t2 3'-C C U A C C A A C A C-5'
t3 3'-C C U A C C A U C A C-5'
t4 3'-C C U A C C A G C A C-5'
*pNs:
N = U, A, C, G or dG
R = H (ImpN);
CH3 (2-MeImpN)

Löffler, P. M. G.; Groen, J.; Dörr, M.; Monnard, P.-A. PLoS ONE 2013, 8 (9), e75617.

template directed primer elongation

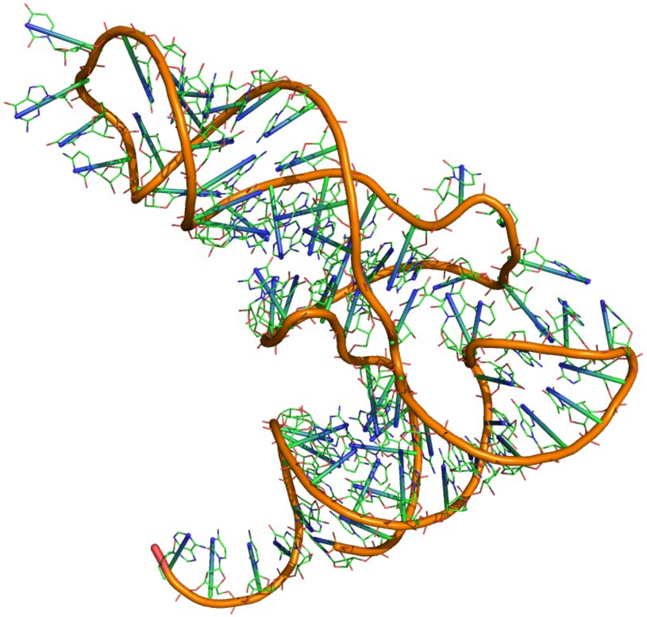
t_2 3'-CCUACCAACAC-5'



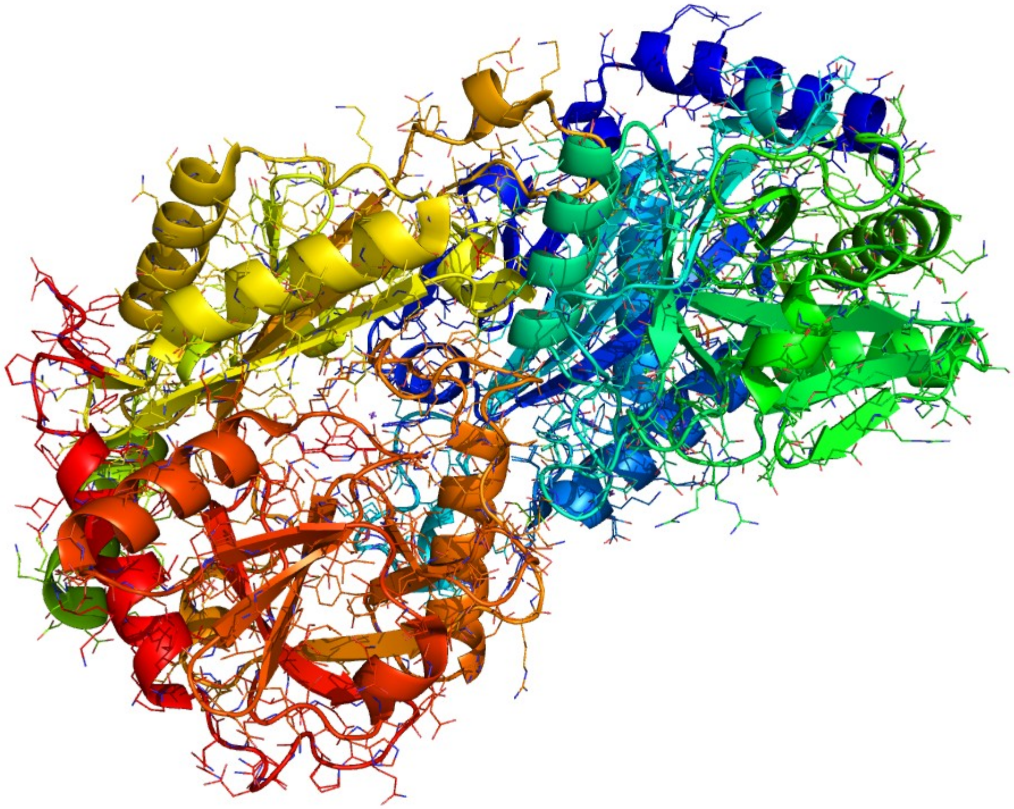
Löffler, P. M. G.; Groen, J.; Dörr, M.; Monnard, P.-A. PLoS ONE 2013, 8 (9), e75617

RNA condensation and folding is just the beginning

let's increase complexity further



translation



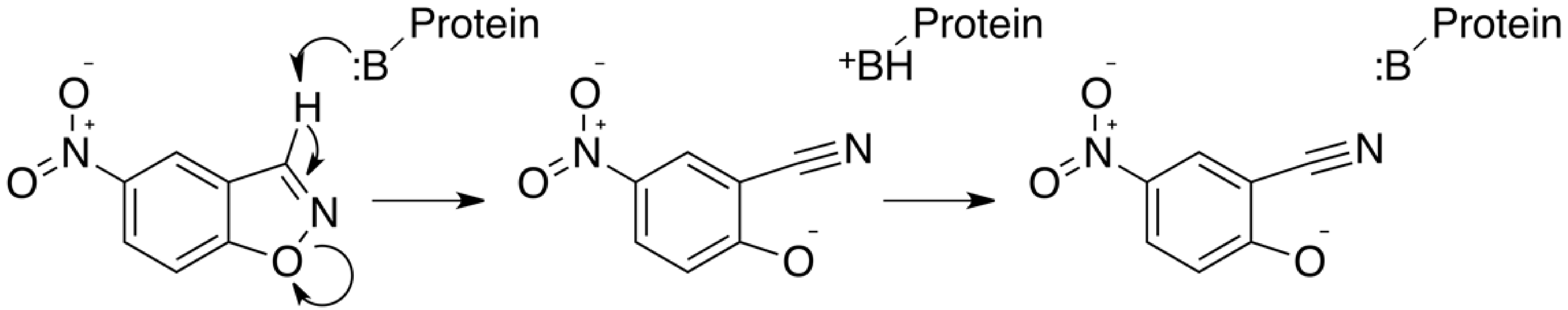
RNA

ASP-tRNA
(PDB:1ASY)

Proteins

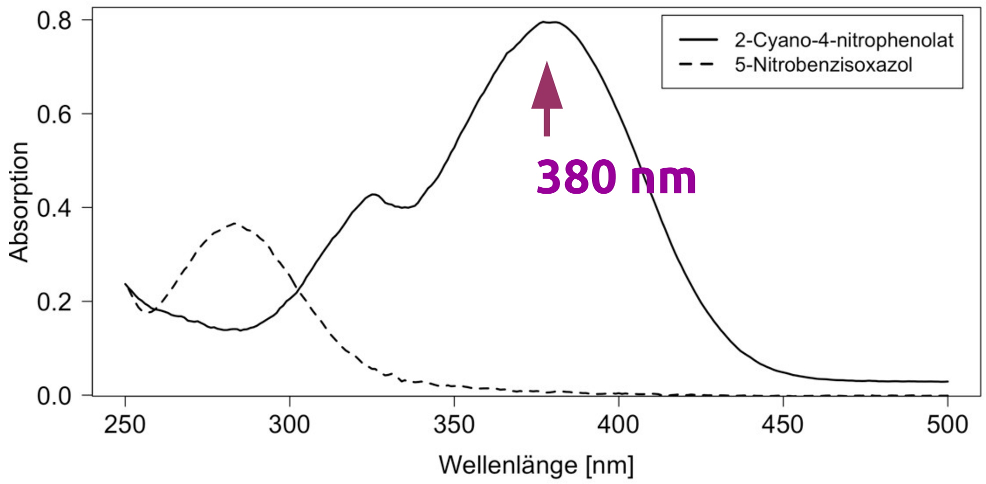
transaminase
(PDB:4CHI)

KEMP-eliminase reaction



5-Nitrobenzisoxazol

2-Cyano-4-nitrophenolat



scheme after Moroz et al., *Angew. Chem. Int. Ed. Engl.* **2013**, *52*, 6246-6249. spectrum: courtesy of Moritz Voß

challenges in protein design/engineering

- increased transformation rate of the substrate
- altered substrate spectrum (e.g. bigger/bulkier substrates)
- enhanced or altered stereo selectivity of the substrate/product

- educt / product tolerance

- usage cheaper cofactors

- higher stability to environmental conditions
- temperature tolerance
- pH tolerance
- organic solvents

- enzyme cascades / pathways to more advanced products
  (related: regulation of
  the protein expression level and or intermediate transformation rate)

how to engineer ?

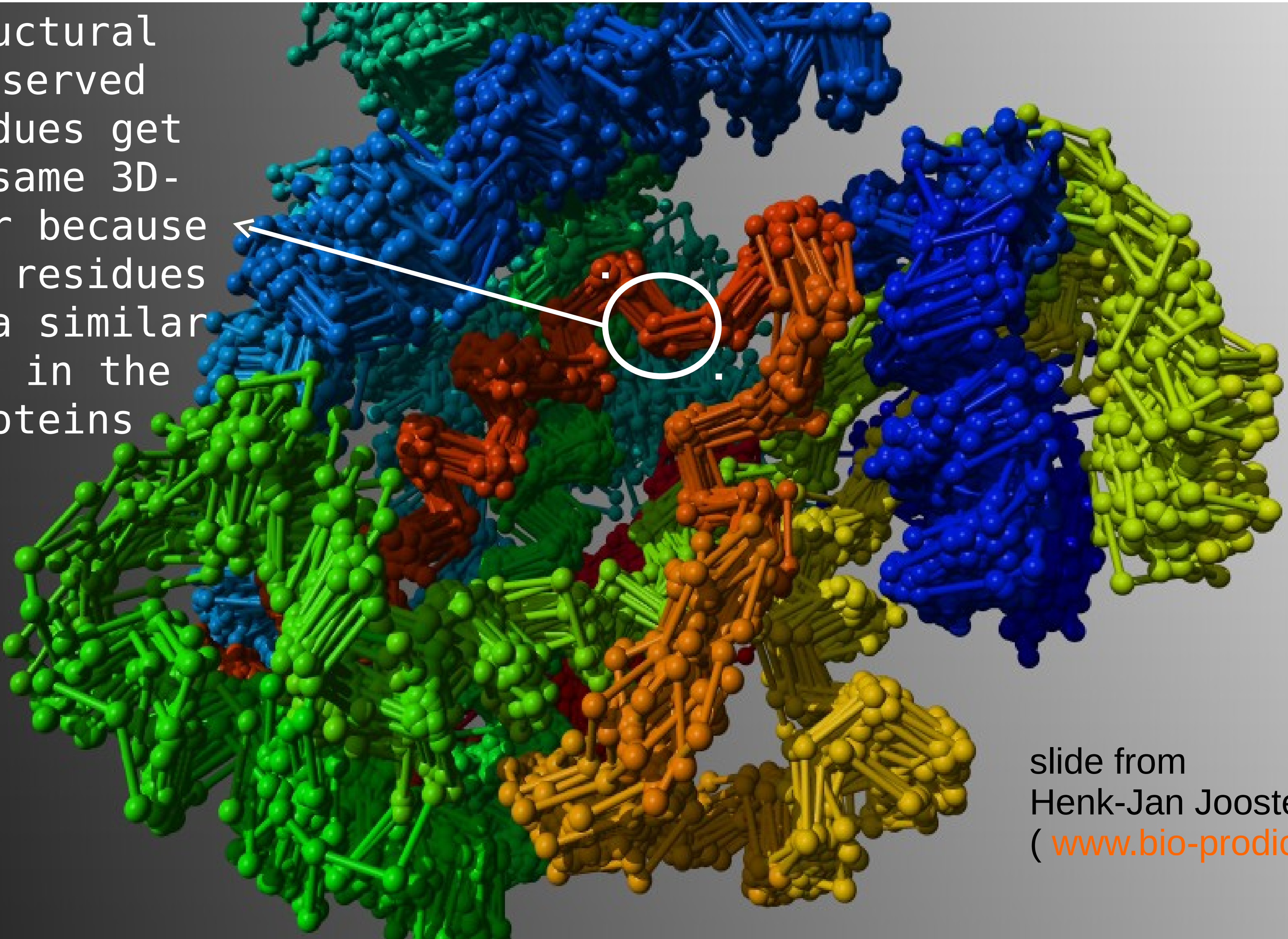
- rational engineering (if structure is known)
  molecular modeling, 3D structure alignments

- (random) mutagenesis strategies
  (error prone PCR, NNK libraries / CASTing)

- semi-rational approaches

Structure based alignments

Structural conserved residues get the same 3D-number because these residues have a similar role in the proteins

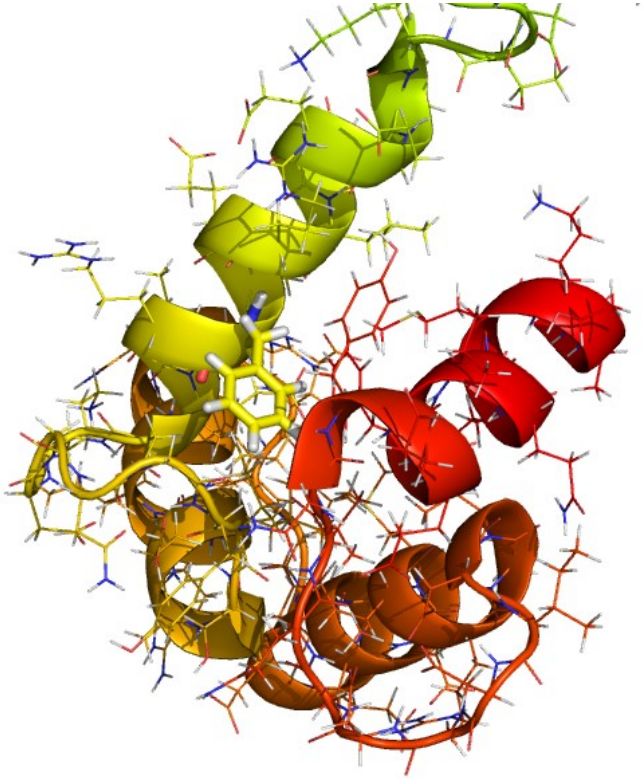


slide from
Henk-Jan Joosten
(www.bio-prodict.nl)

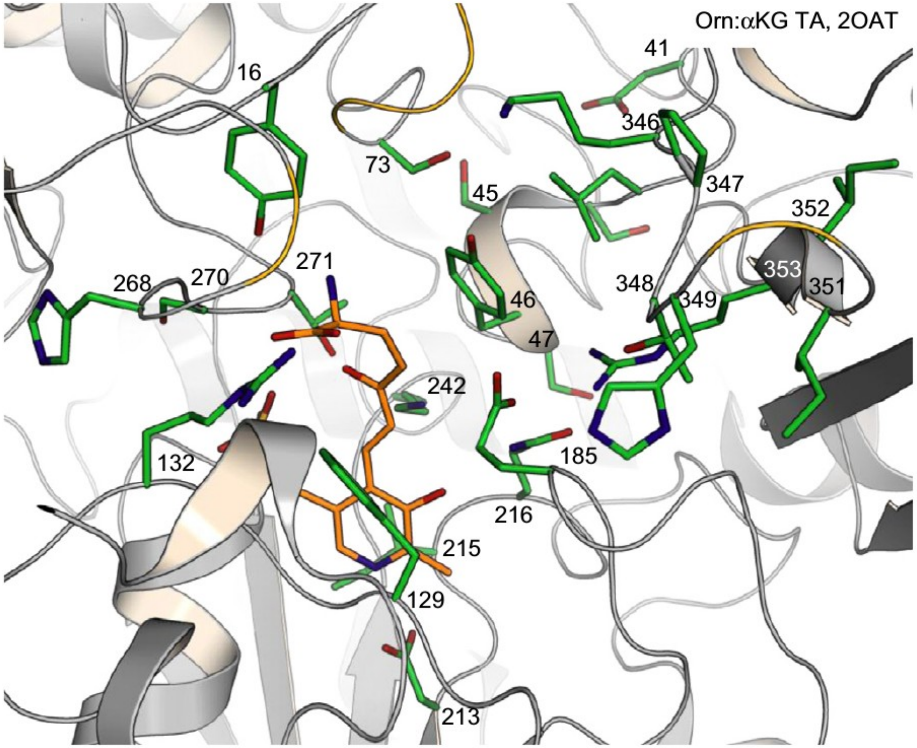
rational design

requirements:

- knowledge about the protein structure (NMR / X-ray)
- certain rigidity of the structure



“simple” active site
(AlleyCatK, PDB 1UPS)



complex active site
(Orn:αKG TA (PDB: 2OAT)

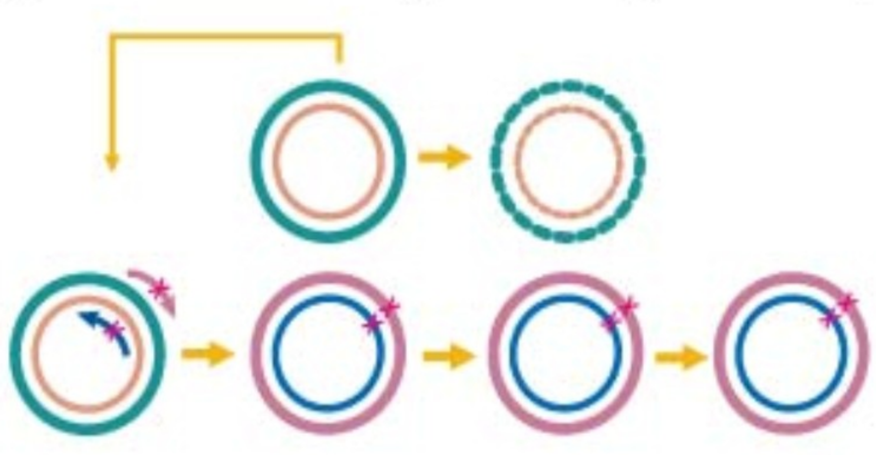
active site modifications

- changes in residue size (e.g. Phe → Ala)
- residue polarity (e.g. Glu → Gln)
- removing loops pointing into the active site
- exchanging loops inside the active site, or in outer spheres

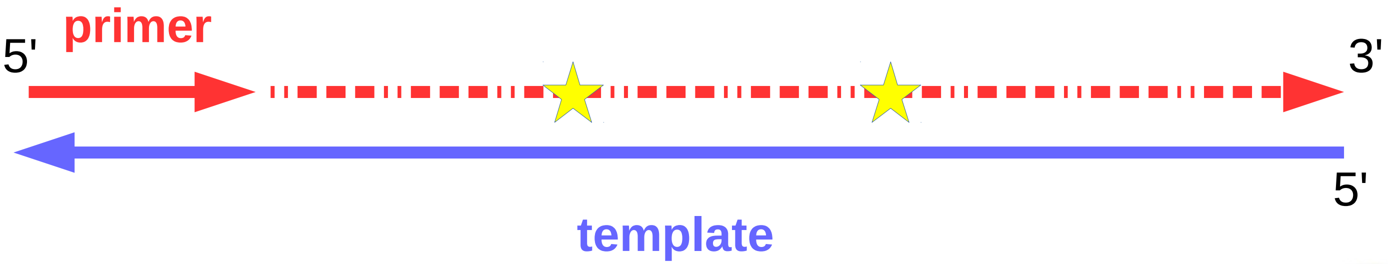
(random) mutagenesis strategies

- QuikChange

(www.agilent.com)

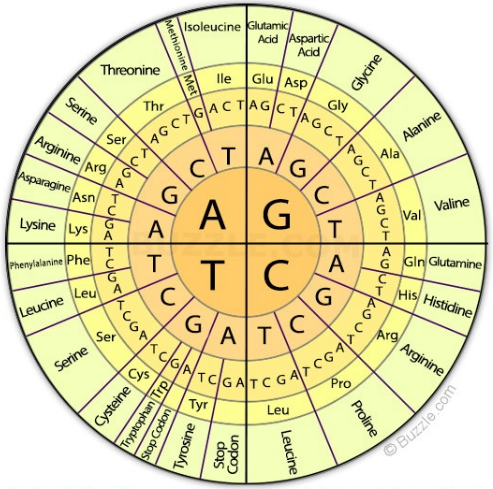


- error prone PCR



- degenerated codons, (e.g.NNK) at a certain position in the gene (where, e.g., is N = A/C/G/T and K = G/T)

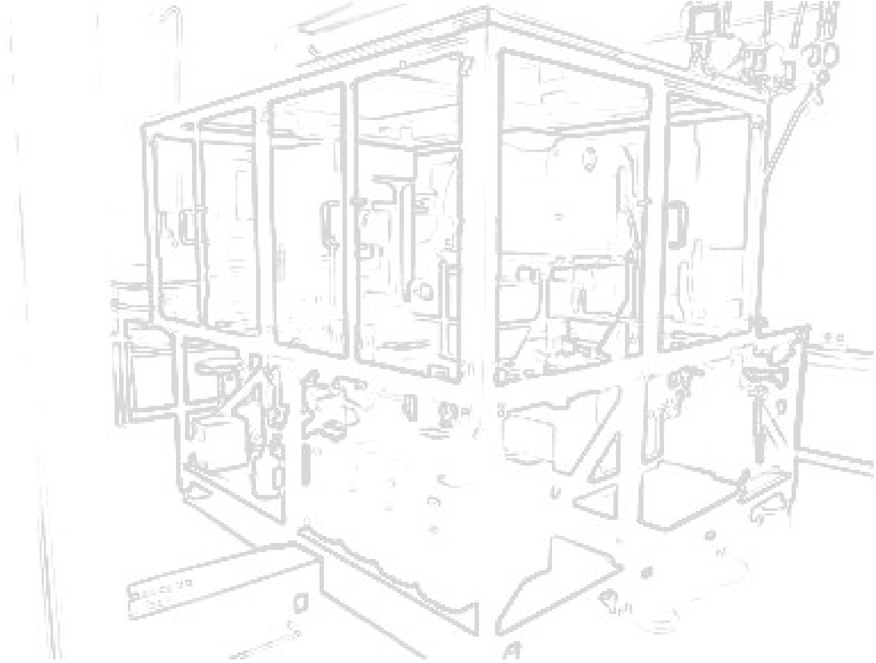
⇒ good coverage of all 20 amino acids at a single position



To decode the codon, move from the center circle towards the periphery.

how do we screen ?

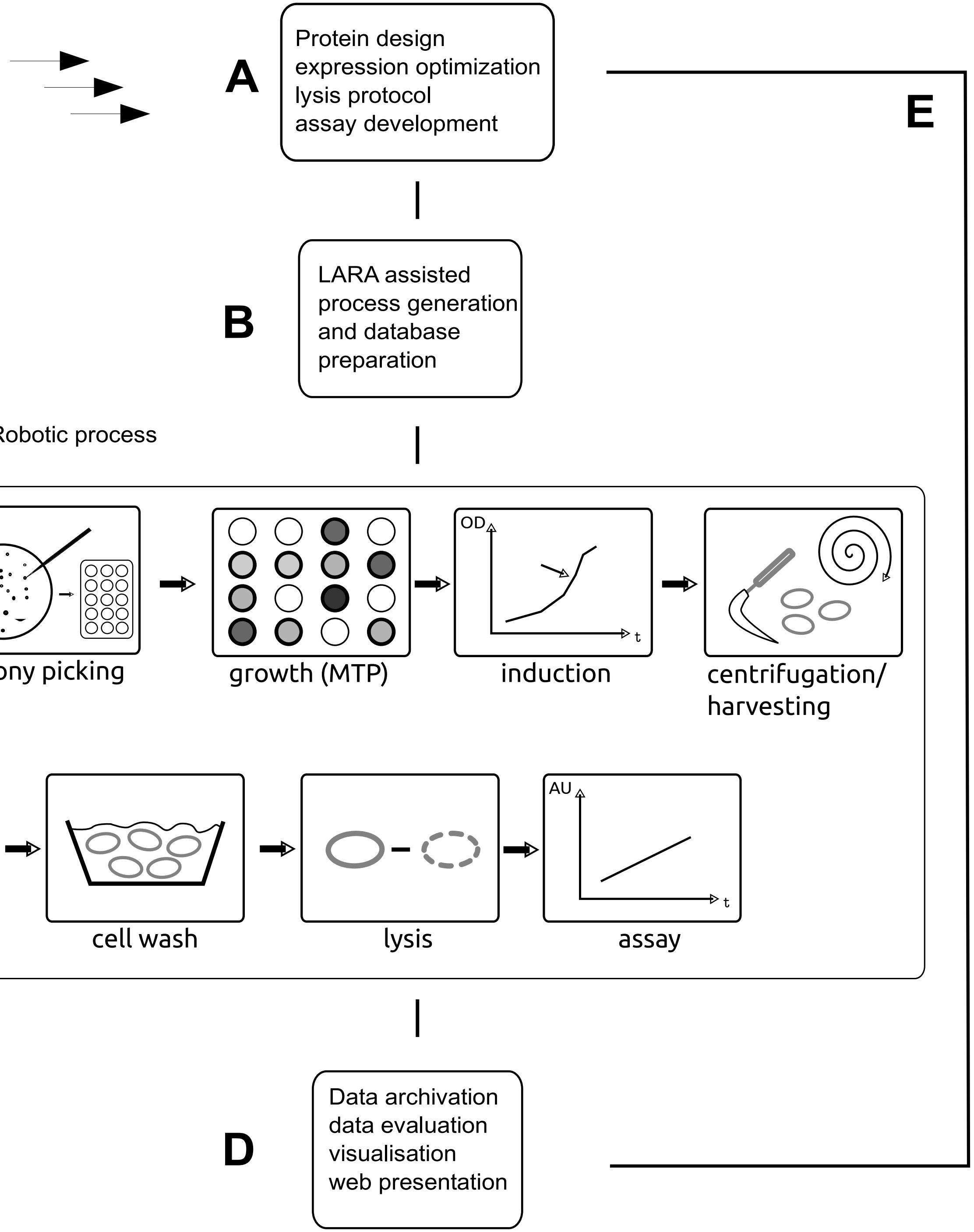
LARA robotic platform



Laboratory Automation
Robotic Assistant

the LARA movie

Our robotic process



A

Protein design
expression optimization
lysis protocol
assay development

B

LARA assisted
process generation
and database
preparation

C Robotic process

colony picking

growth (MTP)

induction

centrifugation/
harvesting

cell wash

lysis

assay

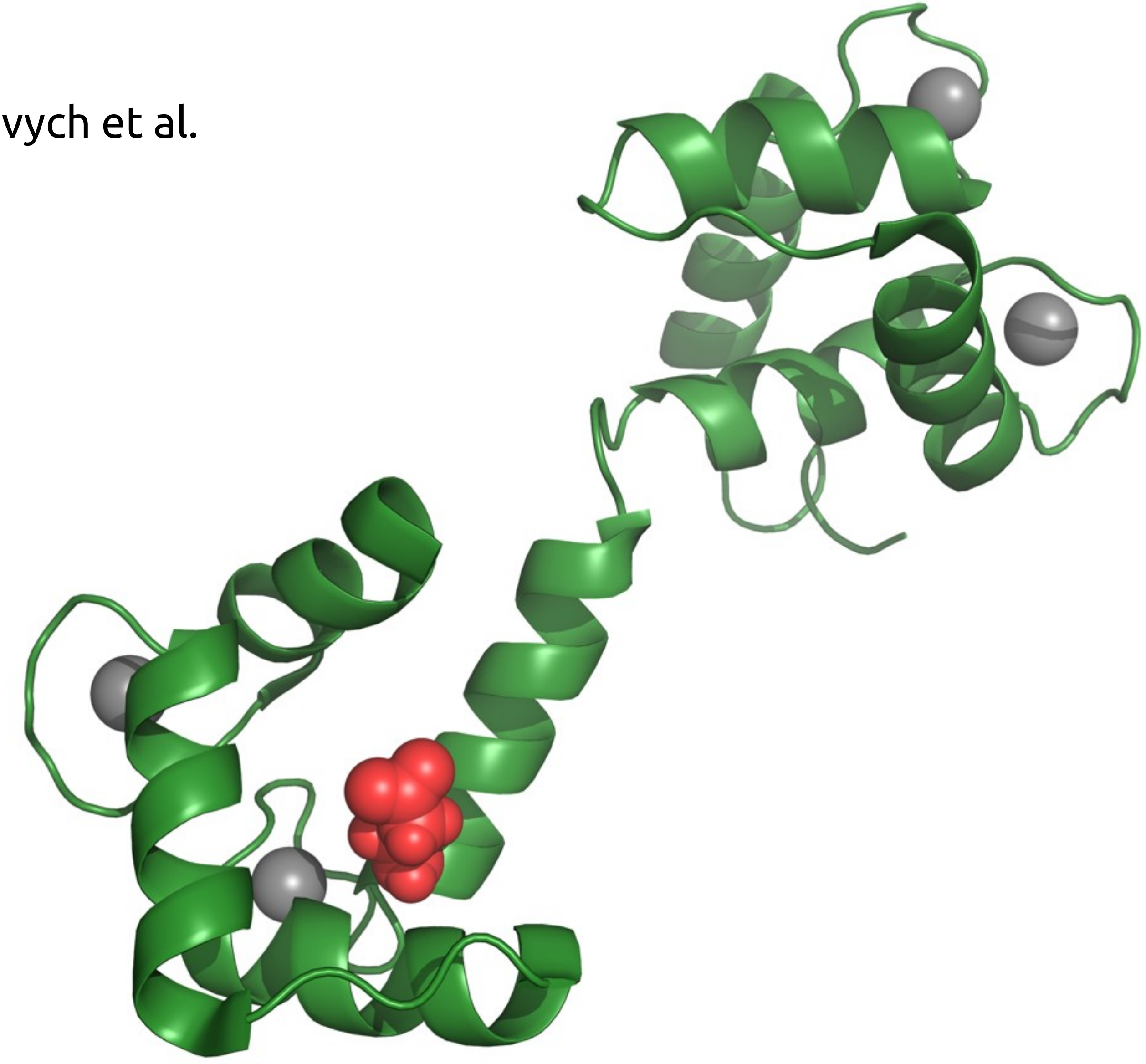
D

Data archivation
data evaluation
visualisation
web presentation

E

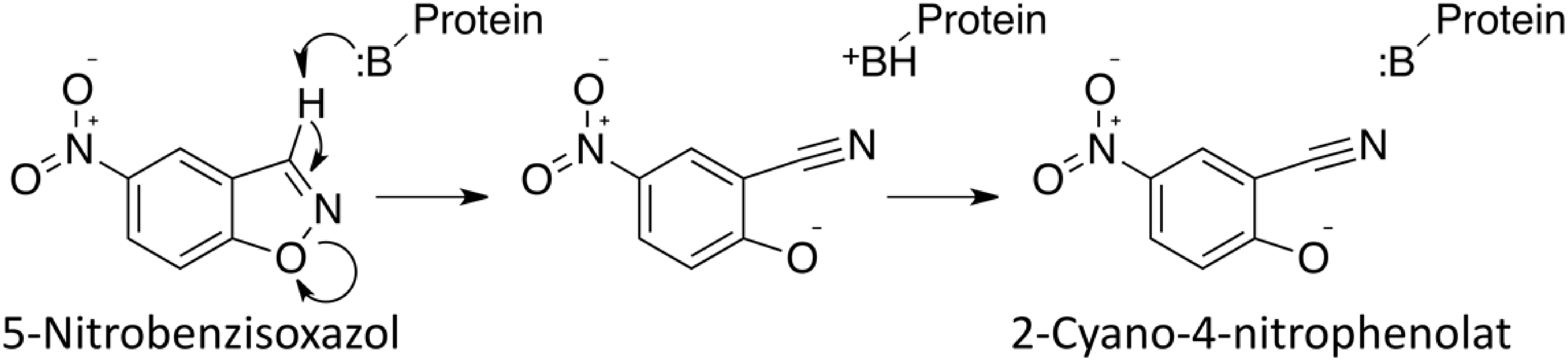
evolution of the KEMP-eliminase

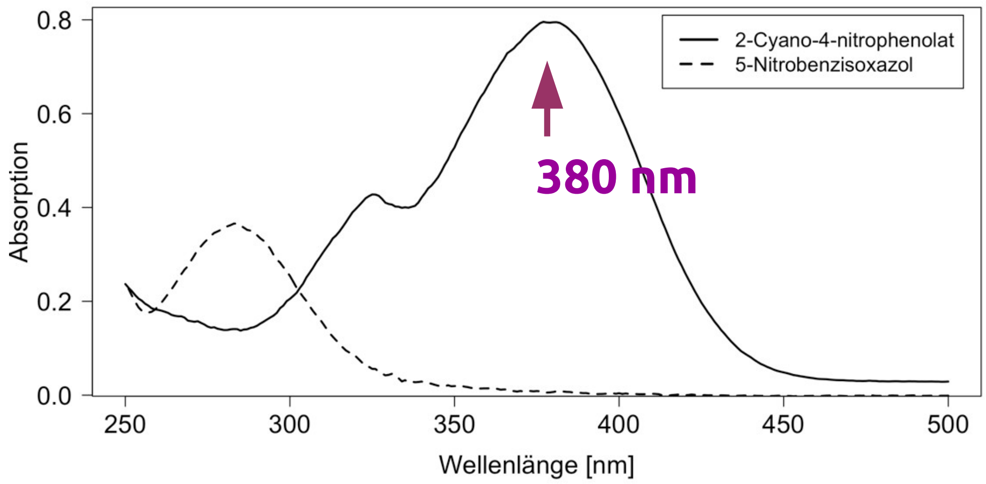
Ivan Korendovych et al.



PDB 2M3S: Moroz et al., *Angew. Chem. Int. Ed. Engl.* **2013**, *52*, 6246-6249, courtesy of Moritz Voss

KEMP-eliminase reaction





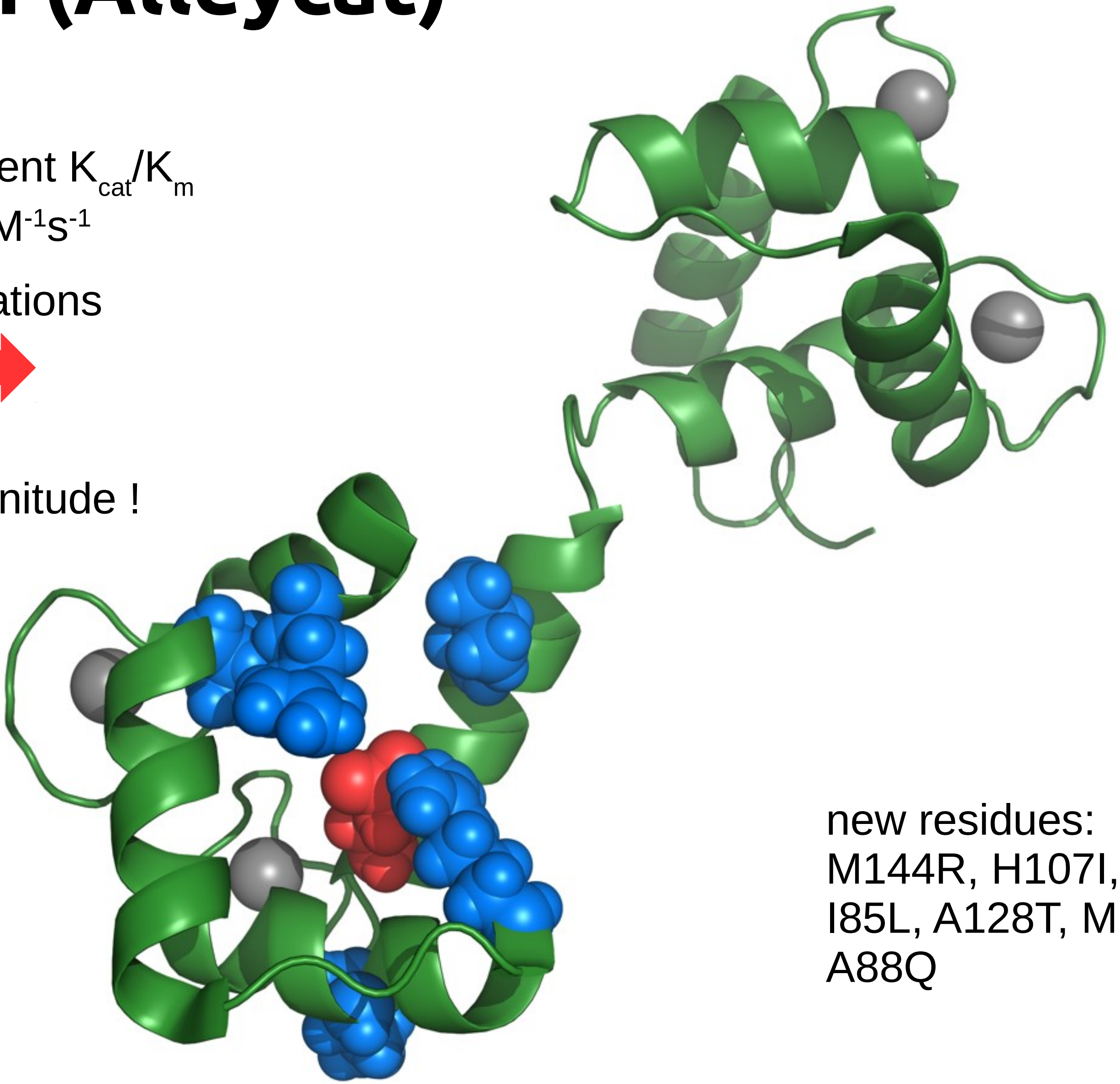
scheme after Moroz et al., *Angew. Chem. Int. Ed. Engl.* **2013**, *52*, 6246-6249. spectrum: courtesy of Moritz Voß

KEMP-eliminase after 7 rounds of evolution (Alleycat)

activity improvement K_{cat}/K_m
6 $M^{-1}s^{-1}$ → 814 $M^{-1}s^{-1}$

theroretical estimations
10^6 $M^{-1}s^{-1}$

activity gap of ca.
2-3 orders of magnitude !



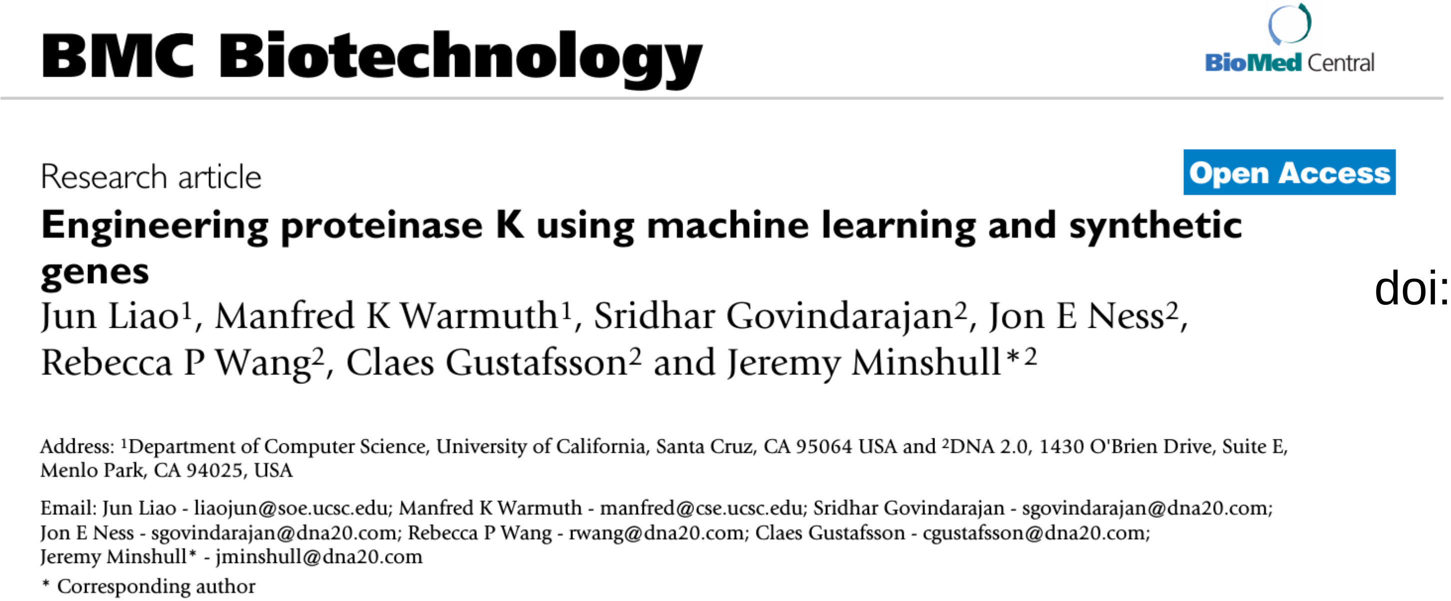
new residues:
M144R, H107I, L112R, I85L, A128T, M124L, A88Q

PDB 2M3S: Moroz et al., *Angew. Chem. Int. Ed. Engl.* **2013**, *52*, 6246-6249, courtesy of Moritz Voss

examples of questions for machine learning

1) beneficial mutations ?
2) correlated / interfering mutations ?
3) improvement of expression / folding ?
4) combination of two properties in one enzyme (e.g. selectivity and stability)

basis of current project:



BMC Biotechnology

BioMed Central

Research article

Open Access

Engineering proteinase K using machine learning and synthetic genes

Jun Liao[1], Manfred K Warmuth[1], Sridhar Govindarajan[2], Jon E Ness[2], Rebecca P Wang[2], Claes Gustafsson[2] and Jeremy Minshull*[2]

Address: [1]Department of Computer Science, University of California, Santa Cruz, CA 95064 USA and [2]DNA 2.0, 1430 O'Brien Drive, Suite E, Menlo Park, CA 94025, USA

Email: Jun Liao - liaojun@soe.ucsc.edu; Manfred K Warmuth - manfred@cse.ucsc.edu; Sridhar Govindarajan - sgovindarajan@dna20.com; Jon E Ness - sgovindarajan@dna20.com; Rebecca P Wang - rwang@dna20.com; Claes Gustafsson - cgustafsson@dna20.com; Jeremy Minshull* - jminshull@dna20.com

* Corresponding author

doi:10.1186/1472-6750-7-16

Basic machine learning tools: linear regressions like:

- ridge regression least absolute shrinkage and selection operator (Lasso)
- partial least square regression (PLSR)
- support vector machine regression (SVMR)
- linear programming support vector machine regression (LPSVMR)
- linear programming boosting regression (LPBoostR)
- matching loss regression (MR)
- one-norm regularization matching-loss regression (ORMR)

explorable sequence space (limitations)

- microtiter plate based screening

  10^3 - 10^4 variants per week

  100 – 300 sequences

- µ-fluidics

  ca. 10^6 - 10^8 per week, # sequences in the same range, but limited information on enzyme activity (a better yes-now answer)

future developments

µ-fluidics and chip based technologies
will allow to combine **100 000 000** activities
in the near future (ca. 3-5 years)
several concepts are under development

➡ **BIG DATA is awaiting you !**

acknowledgements

RNA polymerization crew

Pierre-Alain Monnard (SDU Odense)
Philipp Löffler (SDU Odense)

machine Learning crew
Jan Oldenburg (Bachelor student, Uni Greifswald)
Marc Hellmuth (Uni Greifswald)
Stefan Born (TU Berlin)

de-novo protein crew

Ivan Korendovych (Syracuse, NY)
Moritz Voß (PhD student)
Caroline Nolten (Bachelor student)

group leader

Uwe Bornscheuer

financial support:
DFG and federal state of Mecklenburg-Vorpommern and DFG

Thank you for your attention !

lara.uni-greifswald.de

cartoon: http://xkcd.com/

tRNA and condon usage

- tRNA gene clustering
- tRNA concentration / abundance
- tRNA codon fidelity
- mRNA folding during translation
- tRNA modifications (“stress answer” ?)

nice review on that topic:
Novoa, E. M.; Ribas de Pouplana, L. Trends in Genetics **2012**, 28 (11), 574–581.

dx.doi.org/10.1016/j.tig.2012.07.006