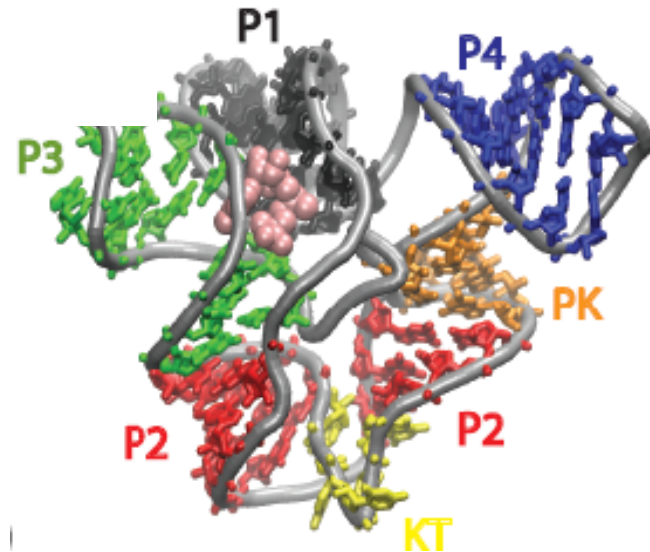
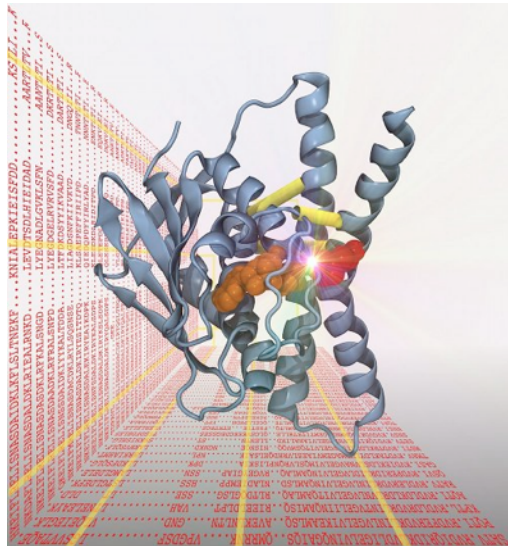


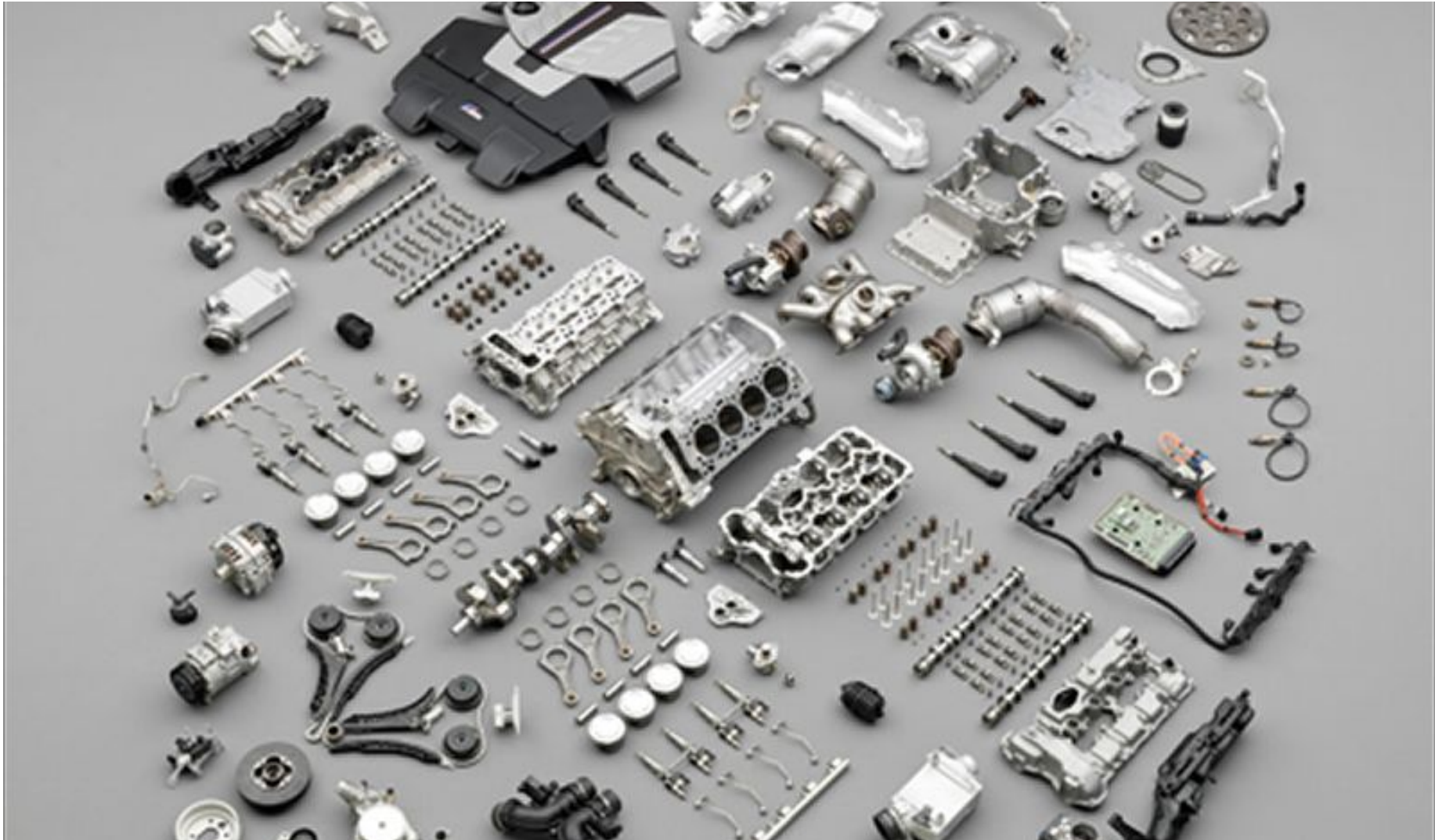
Big Data Analytics in Biology: Biomolecular Structure Prediction and Beyond by Tracing Residue Co-Evolution

GridKA School, September 2016
Alexander Schug
schug@kit.edu

Steinbuch Centre for Computing (SCC) & Department of Physics



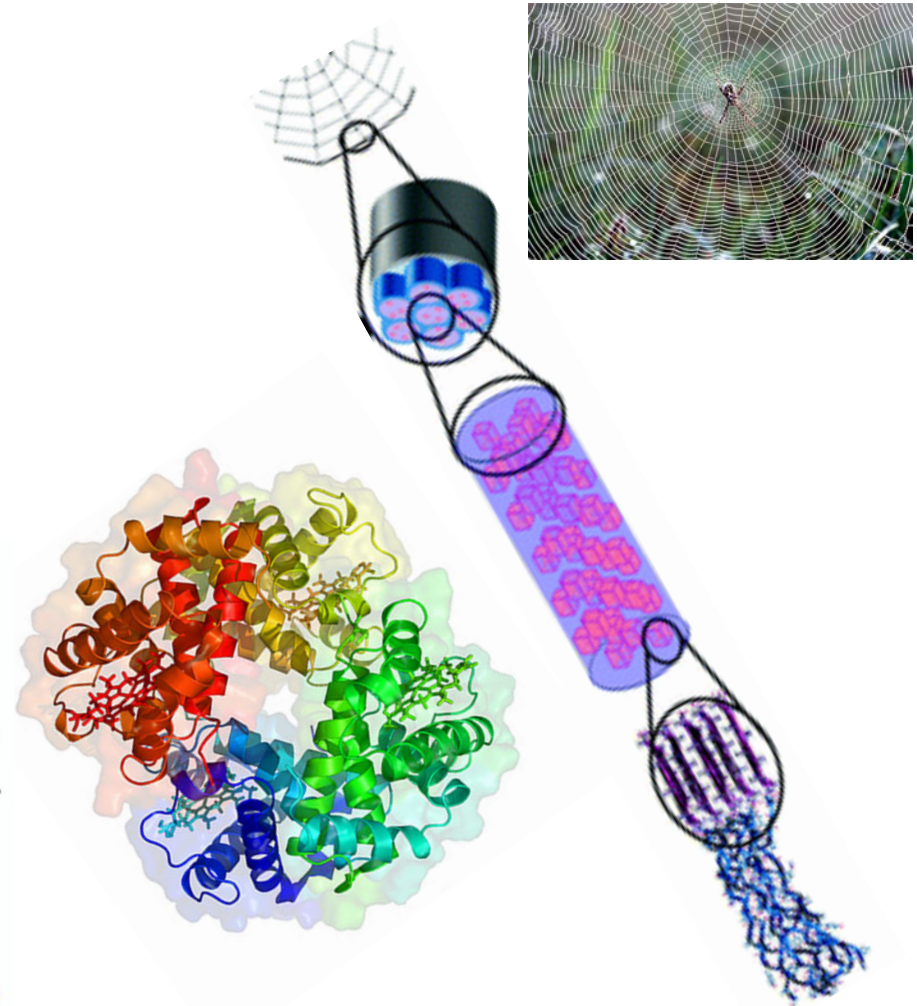
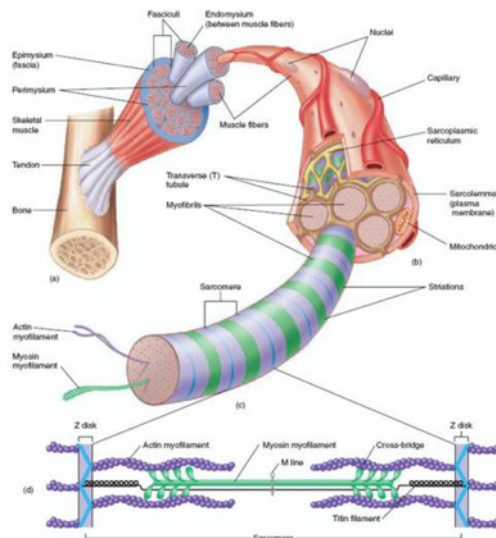
Molecular Machines



Proteins

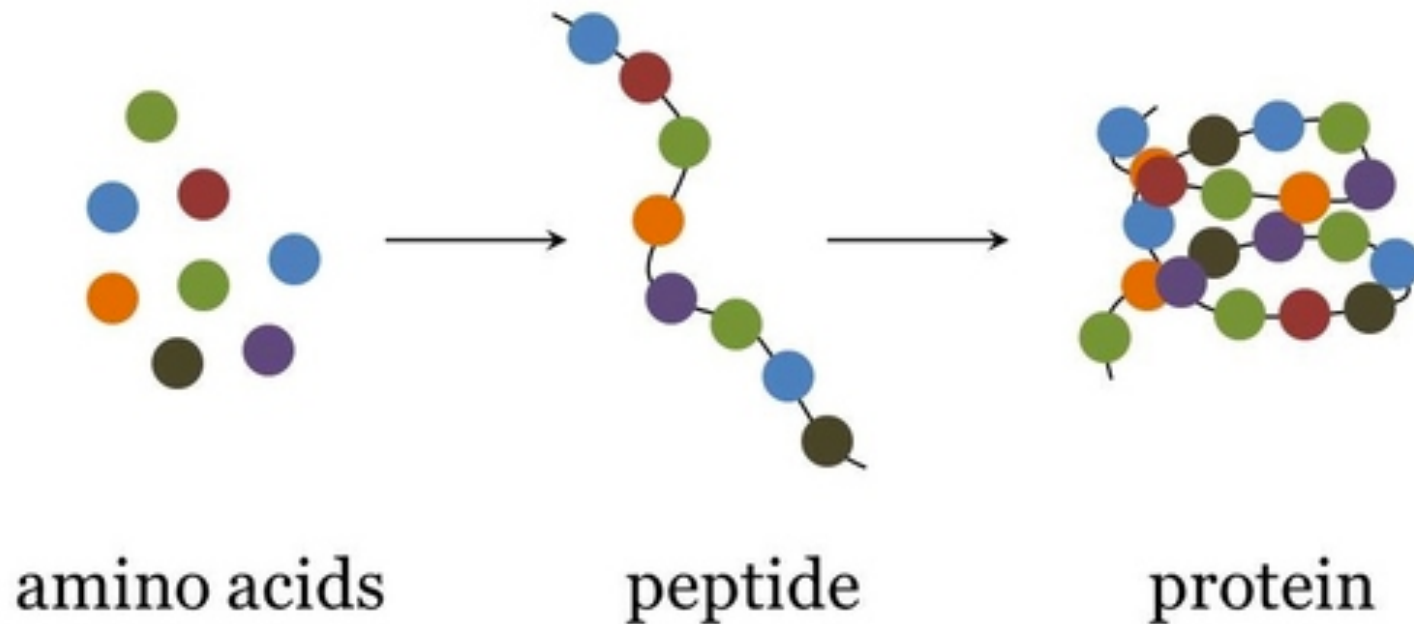
■ Molecular Swiss Knife

- Different Tasks
- Muscles
- Structural Proteins
- Oxygen
- Enzymes
- ...



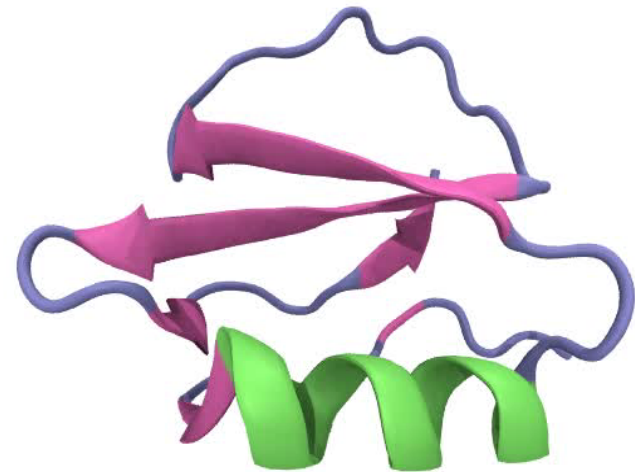
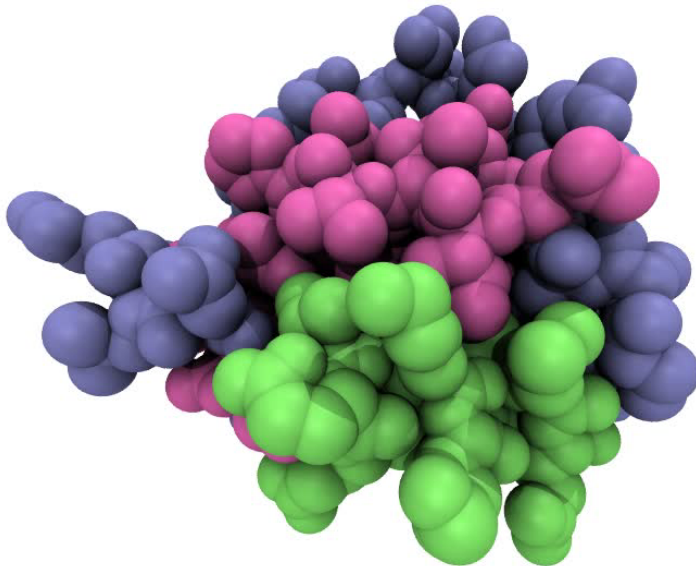
Protein: Sequence of Amino Acids

- Type of protein determines by number of order of amino acids (only 20)
- Folding into unique 3d structures



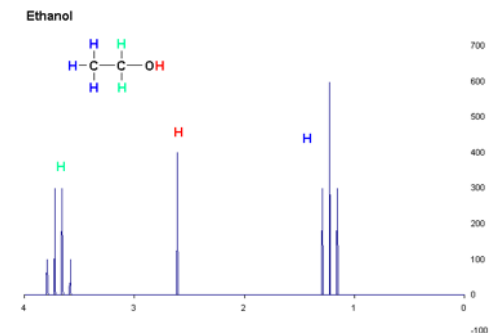
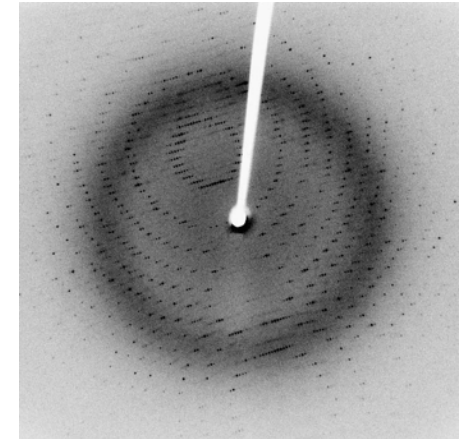
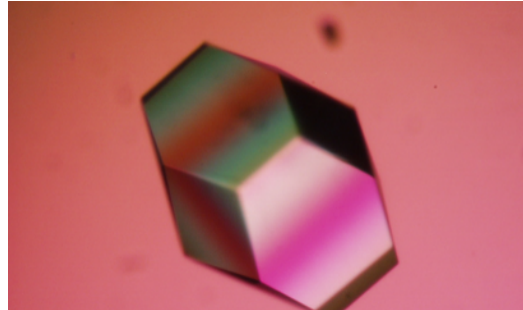
In 3d

- Alpha-helix (green)
- Beta-sheet (purple)



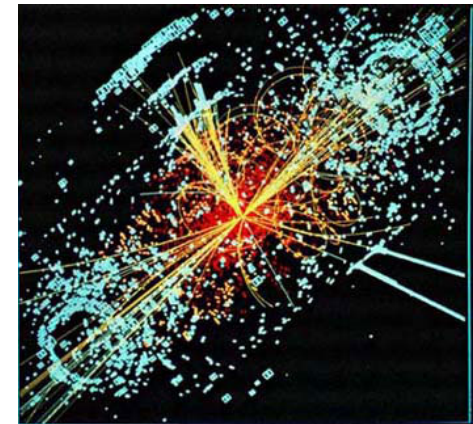
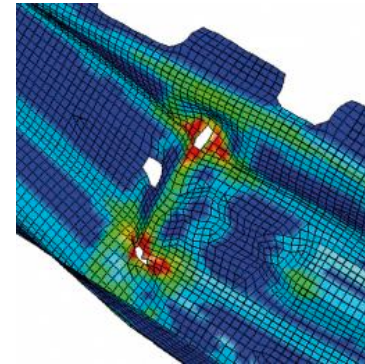
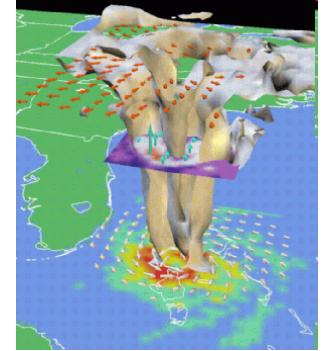
Experiments for Structure Determination?

- Xray
- NMR
- Neutron Scattering
- ...



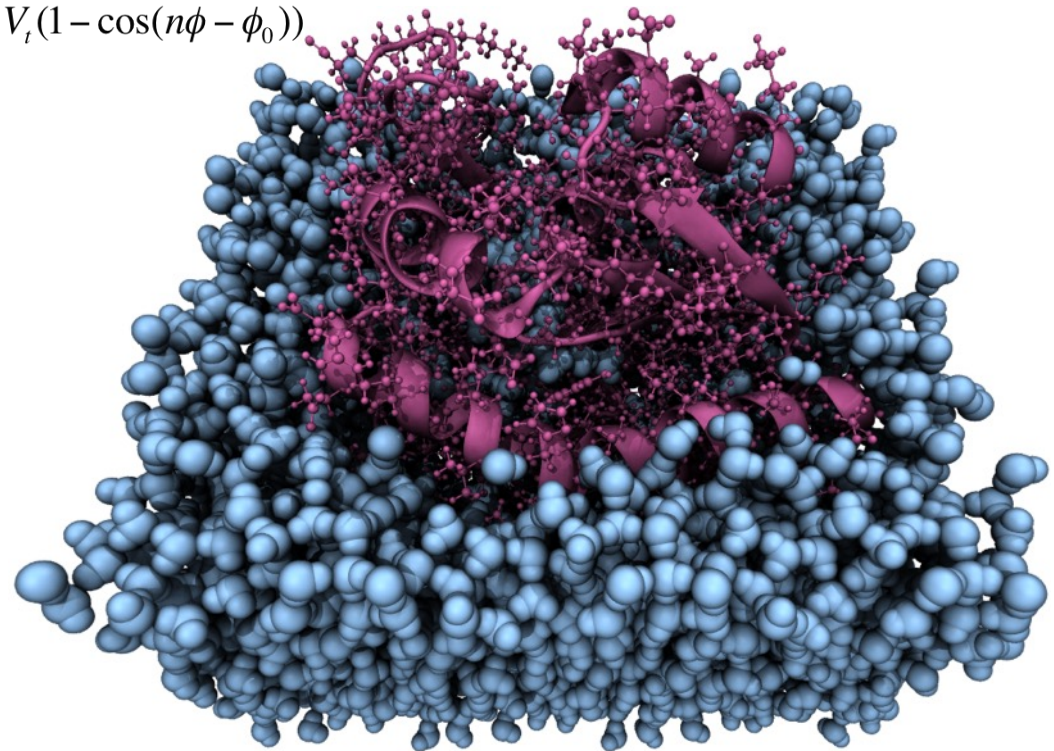
HPC Simulations as „atomic Mikroskopie“

- Simulations as „3rd column“
 - Experiment
 - Theory
 - Simulation
- Molecular Level:
 - *Ab-initio*, i.e. quantum-mechanic Calculation very expensive
 - Classical Force fields
- NP Chemistry 2013 "for the development of multiscale models for complex chemical systems"



Molecular Dynamics

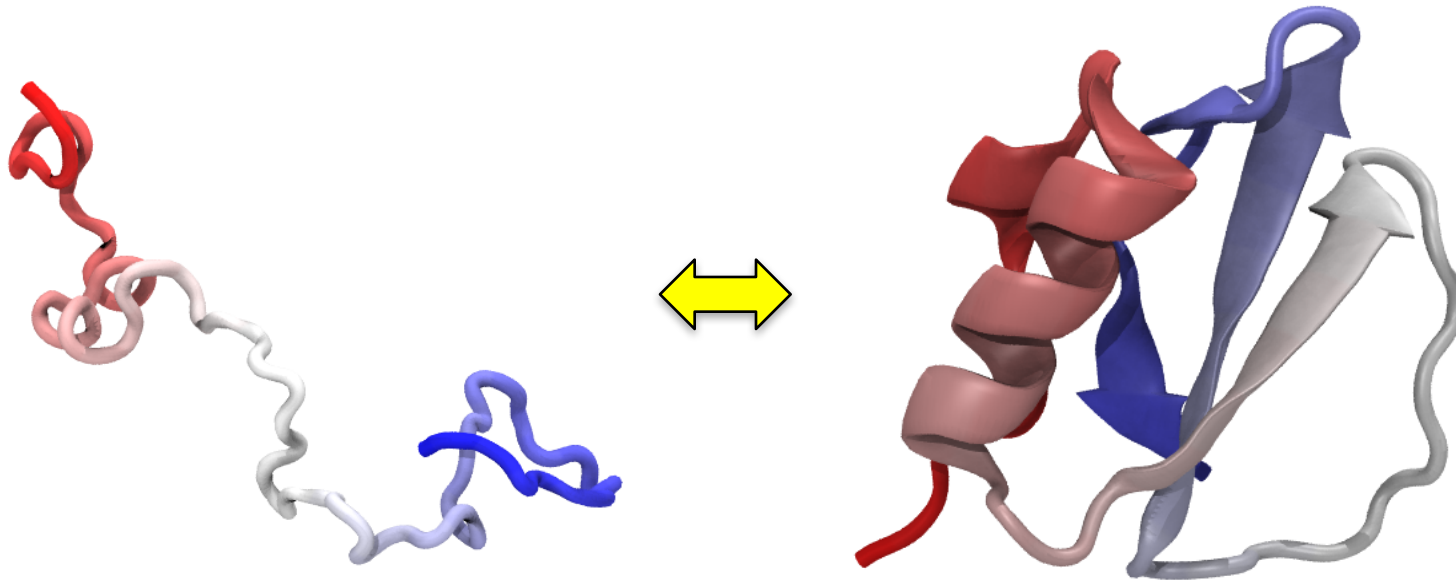
$$\begin{aligned}
 V(\vec{r}) = & \sum_{bonds} K_b (x - x_0)^2 + \sum_{angles} K_a (\theta - \theta_0)^2 + \sum_{torsions} V_t (1 - \cos(n\phi - \phi_0)) \\
 & + \sum_{atoms} V_{ij} \left[\left(\frac{\epsilon_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\epsilon_{ij}}{r_{ij}} \right)^6 \right] + \sum_{atoms} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_R r_{ij}}
 \end{aligned}$$



$$F_i = m_i a_i = m \frac{dv}{dt} = m \frac{d^2 r}{dt^2}$$

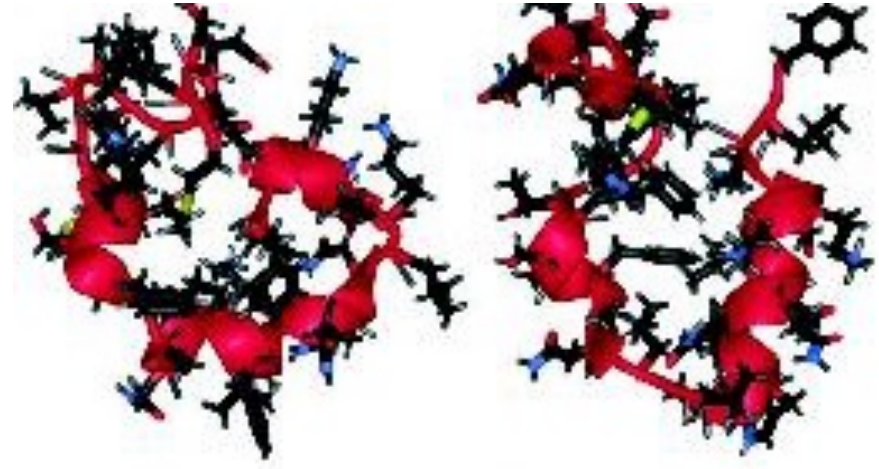
Protein Folding

- Amino acid sequence determines folding



Simulation protein folding 20 yrs back

- Limited size
- Simulated time ($\leq \mu\text{s}$) relevant times (ms-s)
- Limited accuracy of force fields



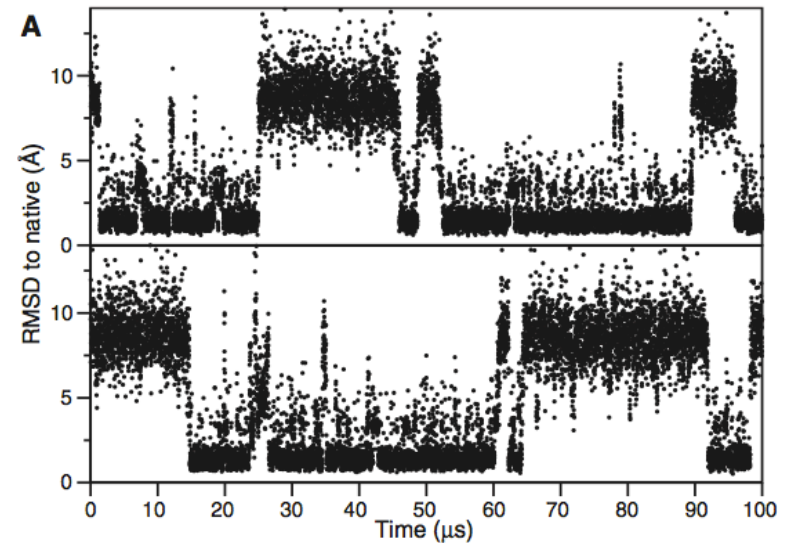
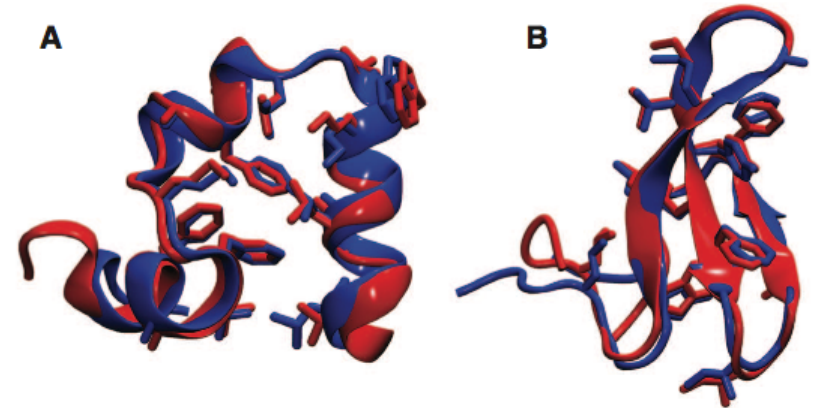
256 nodes CRAY T3E = 85 CPU Years

1vii, 36 AS, explicit solvent, $T = 300\text{ K}$

Y. Duan, P Kollman: Science **23**, 740 (1998)

Supercomputer Anton

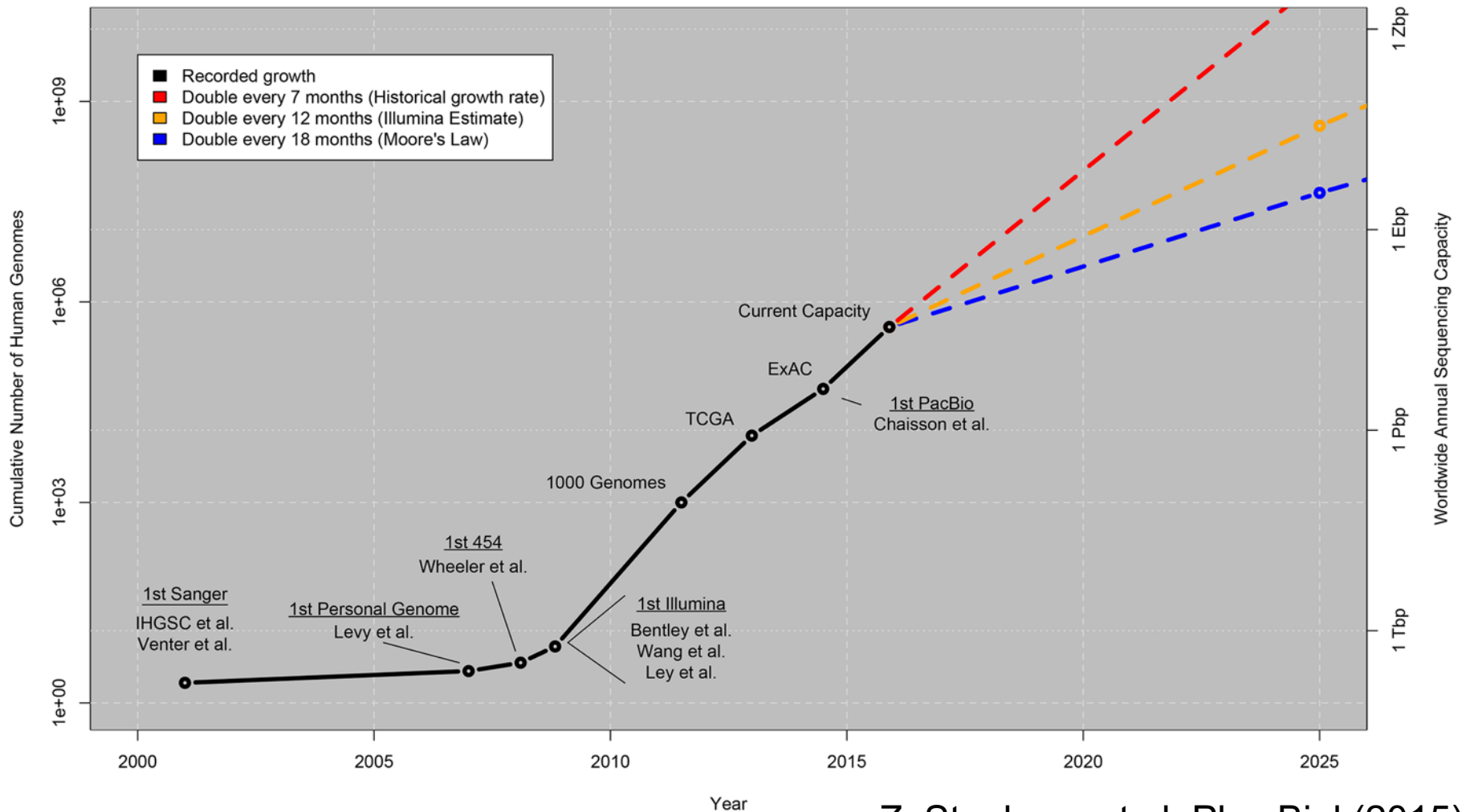
- fs $\rightarrow$ ms ?
- GPU's, Cell's etc.
- ANTON, 0.1-1 *ms* Simulation



D.E. Shaw et al., Science 2010

BIG data?

Growth of DNA Sequencing



Z. Stephens et al., Plos Biol (2015)

Big Data Analytics: Is there information in

```

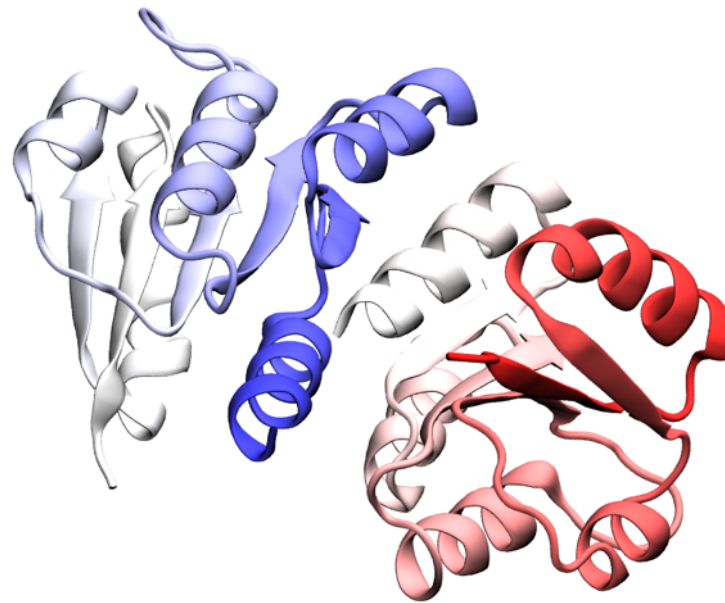
LIIV.EDSPLLMRLQDELK....E.AGY..NNIA.SFENGKEAYEYIMNLAENETLSKQIDMIITD.....IEMPK..M...DGHRLTKLL.KENPKSSD...VPVMIFS.SLIT.DDLRHRGE.VVG.ADEQISKP..EISDLIKKVD
IFCI.DENPIVLNNIKNFLD....D.QIF..AVIG.VTDSLKALMEIL...CTK.....PDIILIN.....VDMPD..L...DGYELCSLL.RKHSYFKN...TPVIMVT.EKAG.LVDRARAK.IVR.ASGHLTKPF.NQGDLLKVIF
ILVA.DDDDVQRKILNNLLK....K.FHY...NVT.LVPNGEIAWEYIN..KGQQK....YDLVLTLD.....VMMPH..I...TGFDLLQRI.NDHPVHRH...IPVILMSGTAVD.YKYANDTI.KIG.GQDFLTKEI.AKELLKKKID
VLLV.EDIELNVIVARSVLE....K.LGN...SVD.VAMTGKAALEMF...KPGE....YDLVLLD.....IQLPD..M...TGLDISRELTCKYPREDL...PPLVALT.ANV..LKDKQEYL.NAG.MDDVLSKPL.SVPALTAMIK
VLLV.EDDAVSIQLCSKFLR....K.YGC...TVQ.VVSDGLSAISTL...EKYR....YDLVLMD.....IVMPN..L...DGATATSIV.RSFD..NE...TPIIAMT.GNIM.NQDLITYL.QHG.MNDILAKPF.TRDDLHSILI
ILVV.EDNHVNQEVIKRMLN....L.EGI..ENIE.LACDGQEAQDKV...KELTSKGENYNNMIFMD.....VQMPK..V...DGLLSTKMIRRDG..YT...SPIVALT.AFAD.DSNIKECL.ESG.MNGFLSKPI.KRPKLKTILT
VLLV.EDNHTNAFILKAFQ....K.YQM...SVE.WVQDGTQALEKL...KEHA....FDLILMD.....NQLPK..M...GGIATREIRETLK..LG...TPIYACT.ADAQ.ESTKQEFL.SAG.ANRVIVKPI.KEQELHDELL
VMAV.DDNPANLKLIGALLE....D.MVQ...HVE.LCDSGHQAVERA...KQMP....FDLILMD.....IQMPD..M...DGIRACELI.HQLPHQQQ...TPVIAVT.AHAM.AGQKEKLL.GAG.MSDYLAKPI.EEERLHNLLL
VLVV.DDHKPNMLLRQQLD....Y.LGQ...RVV.AADSGEAAALW...HEHA....FDVVITD.....CNMPG..I...NGYELARRI.RAAEAAPGYGRTRCILFGFTASAQMDEAQRCA.AAG.MDDCLFKPI.GVDALRQRLN
FLVV.DDFSTMRRIVRNLLK....E.LGF..NNVE.EAEDGVDALNKL...QAGG....FGFIISD.....WNMPN..M...DGLELLKTI.RADSAMSA...LPVLMVT.AEAK.KENIIAAA.QAG.ASGYVVKPF.TAATLEEKLN
ILIA.DDHPTNRLLLKRLN....L.LGY...DVD.EATDGVQALHKV...SMQH....YDLLITD.....VNMPN..M...DGFELTRKL.REQN..SS...LPWGLT.ANAQ.ANEREKGL.SCG.MNLCLFKPL.TLDVLKTHLS
ILIV.DDDQEILKLMKTALE....M.RNY...EVA.THQNISLPLDIT...DFQG....FDLILLD.....IMMSN..I...EGTEICKRI.RREI...S...TPIIFVS.AKDT.EEDIIINGL.GIG.GDDYITKPF.SLKQLVAKVE
LLLI.EDDDAIRTALELSLT....R.QGH...RVA.TAASGEDGLKLL...REQR....PDLIVLD.....VMLPG..I...DGFEVCRR.I.RRTD...Q...LPILLT.ARND.DIDVVVGL.ESG.ADDYVVVVPV.QGRVLDARIR
IVIV.EDEPVTQARLQSYFT....Q.EGY...TVS.VTASGAGLREIM...QNQS....VDLILLD.....INLPD..E...NGLMLTRAL.RERS...T...VGIIIVT.GRSD.RIDRIVGL.EMG.ADDYVTKPL.ELRELVVRVK
VLVI.DDDVAMRHILVEYLT....I.HAF...KVT.AVADSKQFNRLV...CSET....VDVVVVD.....LNLGR..E...DGLEIVRSL.ATKS...D...VPIIIISGARLE.EADKVIAL.ELG.ATDFIAKPF.GTREFLARIR
VLVV.EDEQLADAVATGLR....R.EAM...AVD.VVYDGAALERI...GVND....YDVVVLD.....RDLPL..V...HGDDVCRKI.VELG..MP...TRVMLT.ASGD.VSDRVEGL.EIG.ADDYLPKPF.AFSELIARVR

```

?

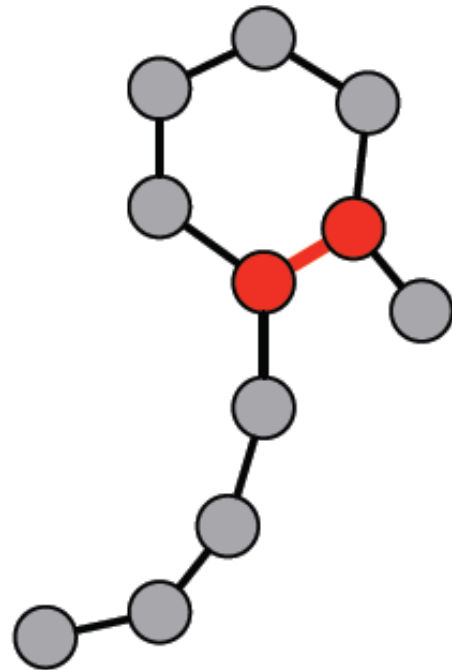
Data Analytics: Is there information in

LIIV.EDSPLLMRLQDELK....E.AGY..NNIA.SFENGKEAYEYIMNLAENETLSKQIDMIITD.....IEMPK..M...DGHRLTKLL.KENPKSSD...VPVMIFS.SLIT.DDLRHRGE.VVG.ADEQISKP..EISDLIKKVD
IFCI.DENPIVLNNIKNFLD....D.QIF..AVIG.VTDSLKALMEIL....CTK.....PDIILIN.....VDMPP..L...DGYELCSLL.RKHSYFKN...TPVIMVT.EKAG.LVDRARAK.IVR.ASGHLTKPF.NQGDLLKVIF
ILVA.DDDDVQRKILNNLLK....K.FHY...NVT.LVPNGEIAWEYIN..KGQQK....YDLVLT...VMMPH..I...TGFDDLQRI.NDHPVHRH...IPVILMSGTAVD.YKYANDTI.KIG.GQDFLTKEI.AKELLKKKID
VLLV.EDIENLVIVARSVLE....K.LGN...SVD.VAMTGKAALEMF...KPGE....YDLVLLD.....IQLPD..M...TGLDISRELTAKYPREDL...PPLVALT.ANV..LKDKQEYL.NAG.MDDVLSKPL.SVPALTAMIK
VLLV.EDDAVSIQLCSKFLR....K.YGC...TVQ.VVSDGLSAISTL...EKYR....YDLVLLD.....IVMPN..L...DGATATSIV.RSFD..NE...TPIIAMT.GNIM.NQDLITYL.QHG.MNDILAKPF.TRDDLHSILI
ILVV.EDNHVNQEVIKRMLN....L.EGI..ENIE.LACDGQEAQDKV...KELTSKGENYNNMIFMD.....VQMPK..V...DGLLSTKMIRRDG..YT...SPIVALT.AFAD.DSNIKECL.ESG.MNGFLSKPI.KRPKLKTILT
VLLV.EDNHTNAFILKAFQ...K.YQM...SVE.WVQDGTQALEKL...KEHA....FDLILMD.....NQLPK..M...GGIATREIRETLK..LG...TPIYACT.ADAQ.ESKQEFEL.SAG.ANRVIVKPI.KEQELHDELL
VMAV.DDNPANLKLIGALLE....D.MVQ...HVE.LCDSGHQAVERA...KQMP....FDLILMD.....IQMPD..M...DGIRACELI.HQLPHQQQ...TPVIAVT.AHAM.AGQKEKLL.GAG.MSDYLAKPI.EEERLHNLLL
VLVV.DDHKPNMLLRQQLD....Y.LGQ...RVV.AADSGEAAALW...HEHA....FDVVITD.....CNMPG..I...NGYELARRI.RAAEAAPGYGRTRCILFGFTASAQMDEAQRCA.AAG.MDDCLFKPI.GVDALRQRLN
FLVV.DDFSTMRRIVRNLLK....E.LGF..NNVE.EAEDGVDALNKL...QAGG....FGFIISD.....WNMPN..M...DGLELLKTI.RADSAMSA...LPVLMVT.AEAK.KENIIAAA.QAG.ASGYVVKPF.TAATLEEKLN
ILIA.DDHPTNRLLLKRLN....L.LGY...DVD.EATDGVQALHKV...SMQH....YDLITD.....VNMPN..M...DGFEELTKL.REQN..SS...LPWGLT.ANAQ.ANEREKGL.SCG.MNLCLFKPL.TLDVLKTHLS
ILIV.DDDQEILKLMKTALE....M.RNY...EVA.THQNISLPLDIT...DFQG....FDLILLD.....IMMSN..I...EGTEICKRI.RREI...S...TPIIFVS.AKDT.EEDIIINGL.GIG.GDDYITKPF.SLKQLVAKVE
LLLI.EDDDAIRTALELSLT....R.QGH...RVA.TAASGEDGLKLL...REQR....PDLIVLD.....VMLPG..I...DGFEVCRRI.RRTD...Q...LPIILLT.ARND.DIDVVVGL.ESG.ADDYVVKPV.QGRVLDARIR
IVIV.EDEPVTQARLQSYFT....Q.EGY...TVS.VTASGAGLREIM...QNQS....VDLILLD.....INLPD..E...NGLMLTRAL.RERS...T...VGIIIVT.GRSD.RIDRIVGL.EMG.ADDYVTKPL.ELRELVRVK
VLVI.DDDVAMRHVIVEYLT....I.HAF...KVT.AVADSKQFNRLV...CSET....VDVVVVD.....LNLRG..E...DGLEIVRSL.ATKS...D...VPIIIISGARLE.EADKVIAL.ELG.ATDFIAKPF.GTREFLARIR
VLVV.EDEQLADAVATGLR....R.EAM...AVD.VVYDGAALERI...GVND....YDVVVLD.....RDLPL..V...HGDDVCRKI.VELG..MP...TRVLMIT.ASGD.VSDRVEGL.EIG.ADDYLPKPF.AFSELIARVR



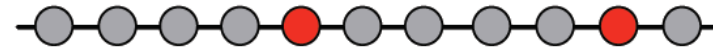
?

Co-Evolution of Amino Acids in Contacts



Contact in 3D

Co-Evolution
→



R	I	D	H	R	L	K	N	T	D	H
F	L	N	G	R	L	R	D	T	D	H
H	E	R	Q	E	T	G	E	L	K	H
K	Y	R	T	R	L	T	D	L	D	H
R	R	A	M	E	V	G	N	L	K	H
T	Q	K	E	E	L	A	N	L	K	H
K	Q	Q	S	E	V	E	N	A	K	H
R	L	N	Q	R	A	D	D	L	D	H

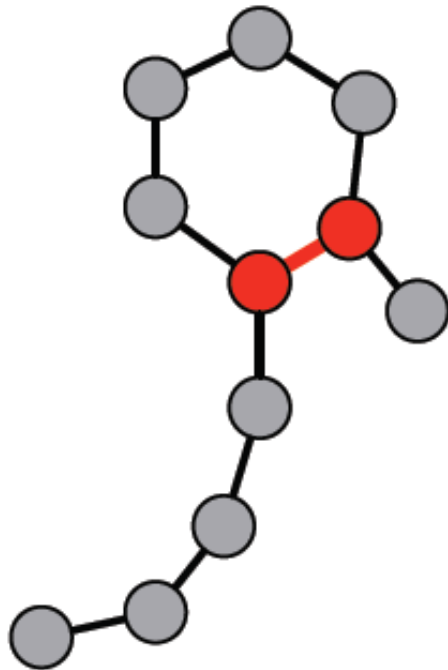
Correlation



Are sequence correlations indicative for residue-residue contacts?

[Göbel et al. 1994, Neher 1994, Ranganathan et al. 1999]

Statistical Inference by Mutual Information?

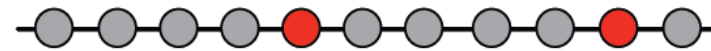


Contact in 3D

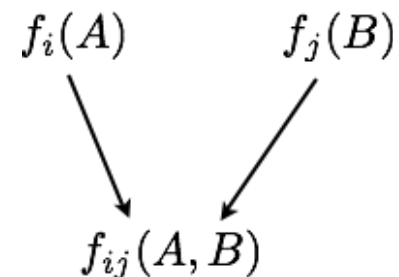
Co-Evolution



Statistical
Inference



R	I	D	H	R	L	K	N	T	D	H
F	L	N	G	R	L	R	D	T	D	H
H	E	R	Q	E	T	G	E	L	K	H
K	Y	R	T	R	L	T	D	L	D	H
R	R	A	M	E	V	G	N	L	K	H
T	Q	K	E	E	L	A	N	L	K	H
K	Q	Q	S	E	V	E	N	A	K	H
R	L	N	Q	R	A	D	D	L	D	H



$$MI_{ij} = \sum_{A,B} f_{ij}(A, B) \ln \frac{f_{ij}(A, B)}{f_i(A) f_j(B)}$$

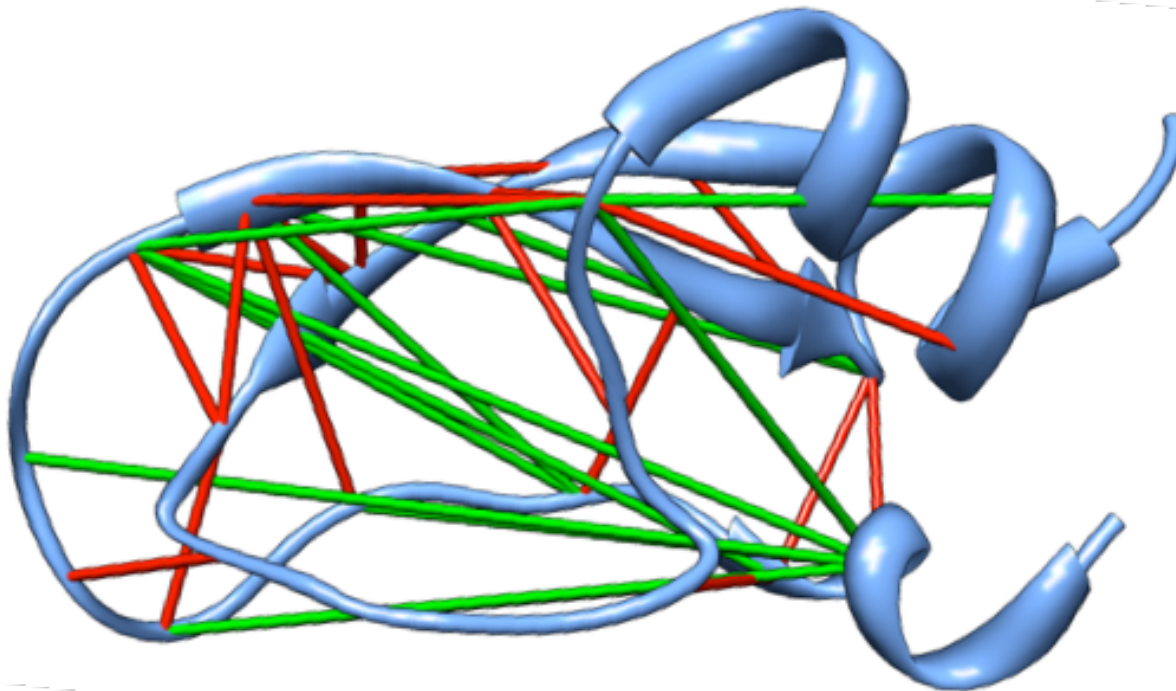
Strong Correlation is not a contact

Trypsin inhibitor

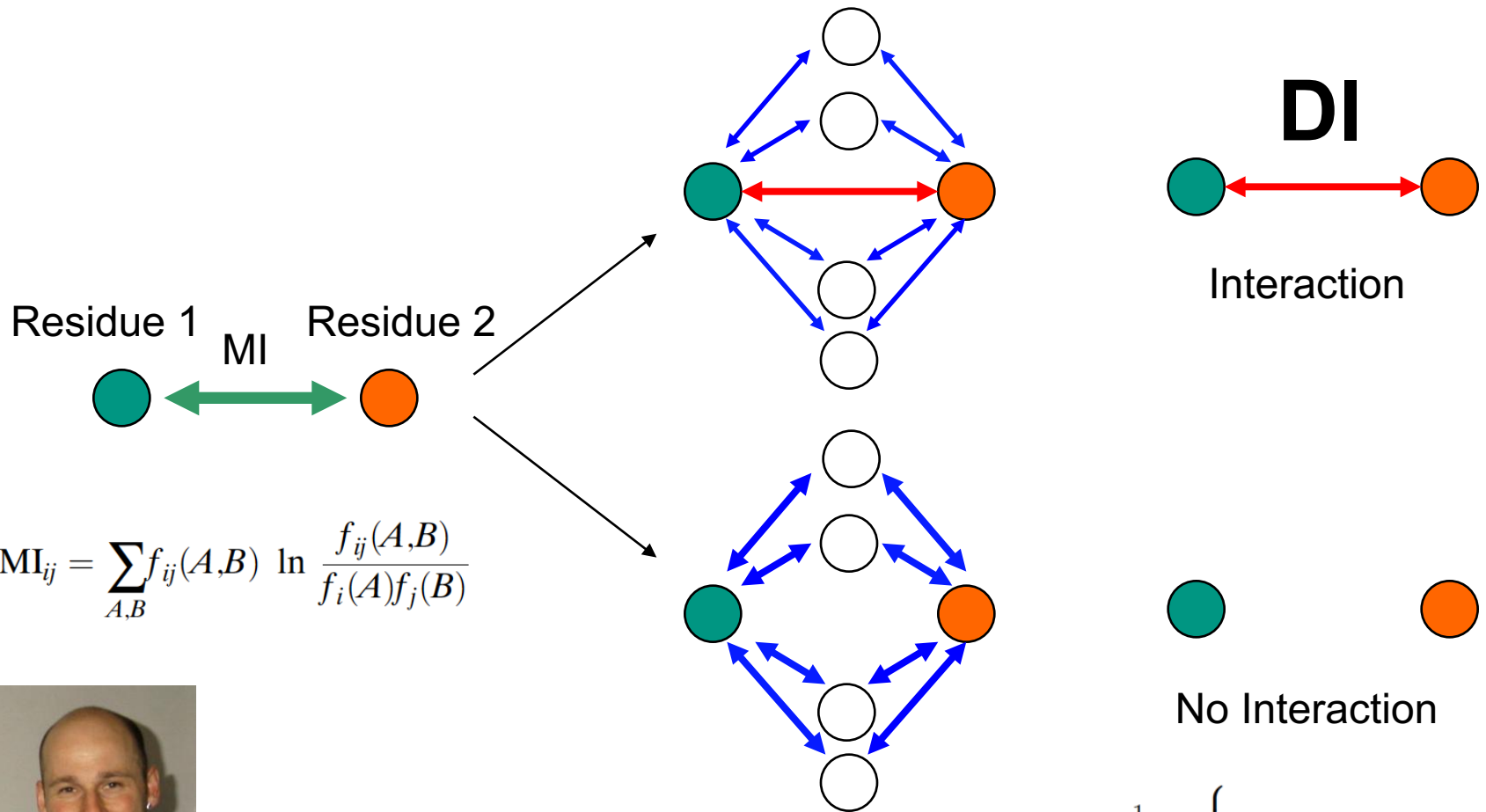
30 strongest MI correlations

Red: contact

Green: no contact



DCA(Direct Coupling Analysis) disentangles direct and indirect correlations



$$MI_{ij} = \sum_{A,B} f_{ij}(A,B) \ln \frac{f_{ij}(A,B)}{f_i(A)f_j(B)}$$



Message Passing: M. Weigt et al., PNAS 2009

Mean Field: F. Morcos et al., PNAS 2011

$$P(A_1, \dots, A_L) = \frac{1}{Z} \exp \left\{ \sum_{i < j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i) \right\}$$

Direct Coupling Analysis

Maximum entropy modeling $P(A_1, \dots, A_L)$

i) Coherence with data

model generates empirical correlations

$$P_{ij}(A_i, A_j) = \sum_{\{A_k | k \neq i, j\}} P(A_1, \dots, A_L) \stackrel{!}{=} f_{ij}(A_i, A_j)$$

ii) Minimal constrained model $P(A_1, \dots, A_L)$

Maximum Entropy

$$- \sum_{\{A_i\}} P(A_1, \dots, A_L) \ln P(A_1, \dots, A_L) \rightarrow \max$$

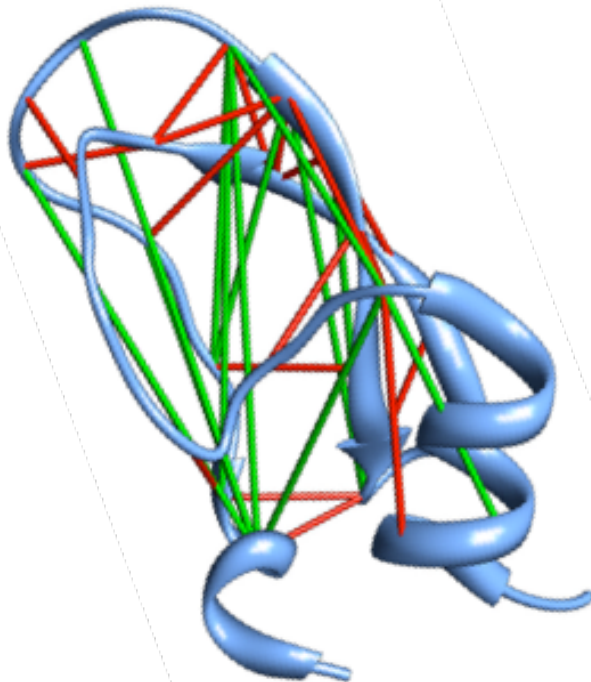
=> (inverse) Potts model

$$P(A_1, \dots, A_L) \sim \exp \left\{ + \sum_{i < j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i) \right\}$$

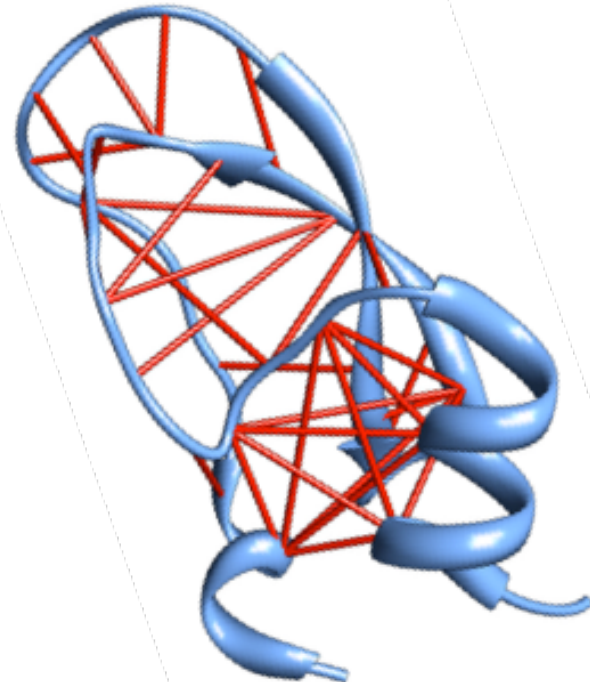
Strong improvement of contact prediction

- Trypsin inhibitor
- Red: contact
- Green: no contact

30 strongest MI correlations

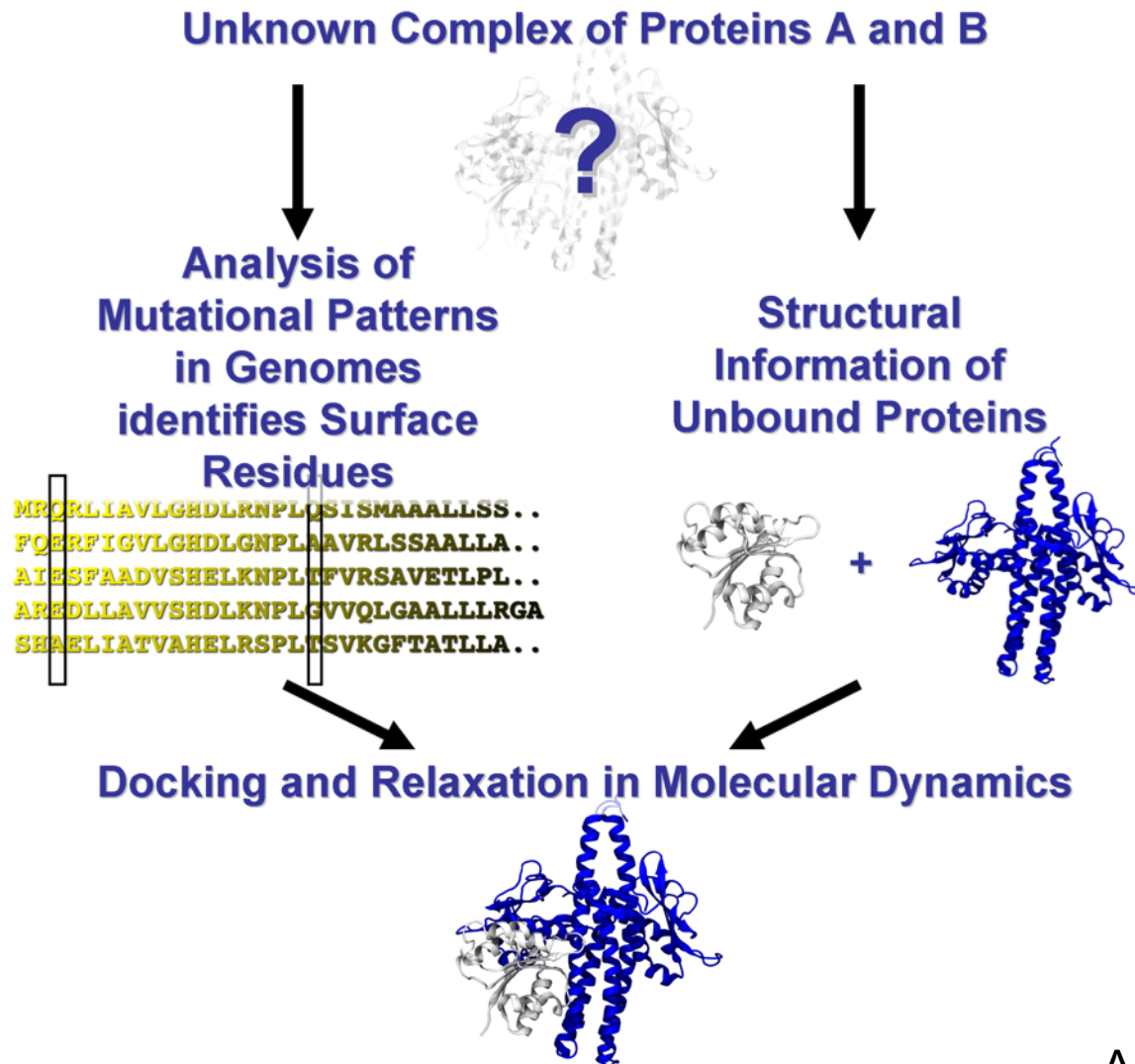


30 strongest DCA couplings



How do proteins talk to each other?

Protein complex prediction



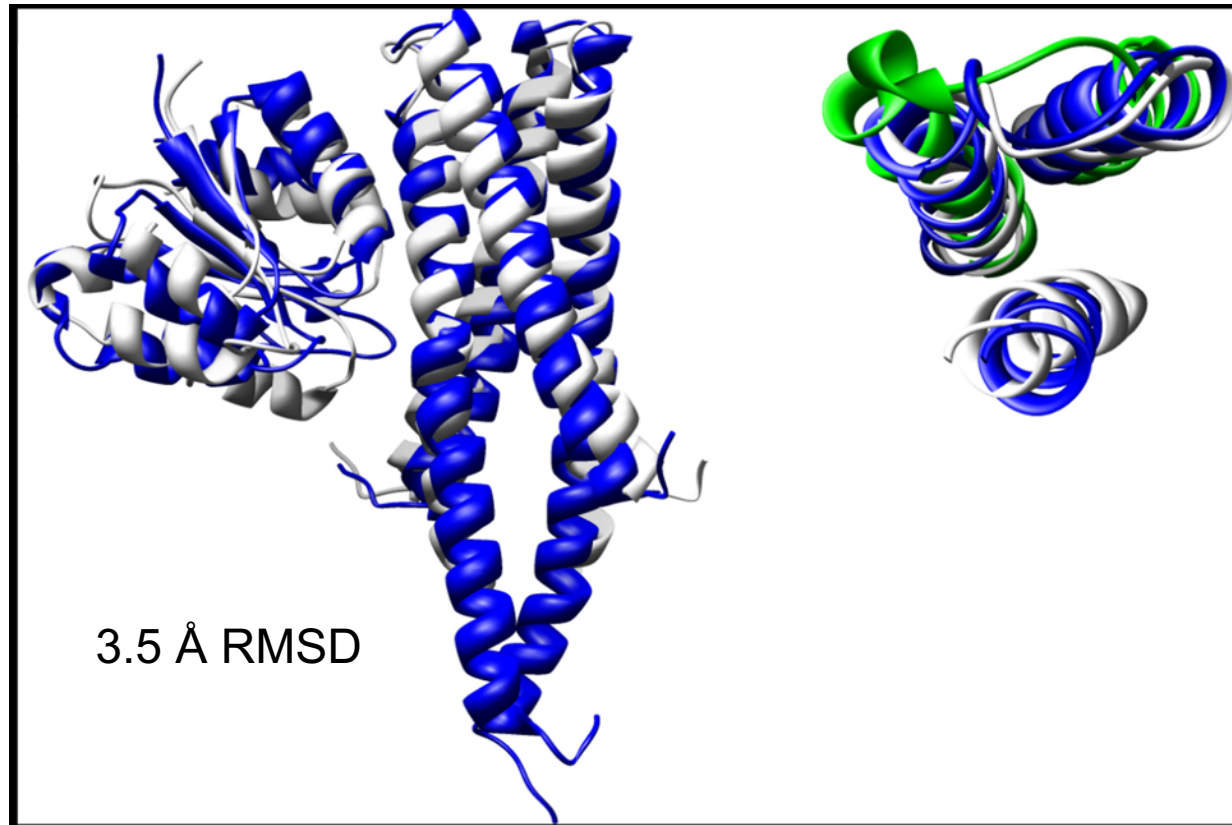
AS et al., PNAS 2009

Docking via predicted Interface Contacts



AS et al., PNAS 2009

Blind Prediction: Experimental Complex of Two Component Signal Transduction System



Blind Prediction: AS et al., PNAS 2009

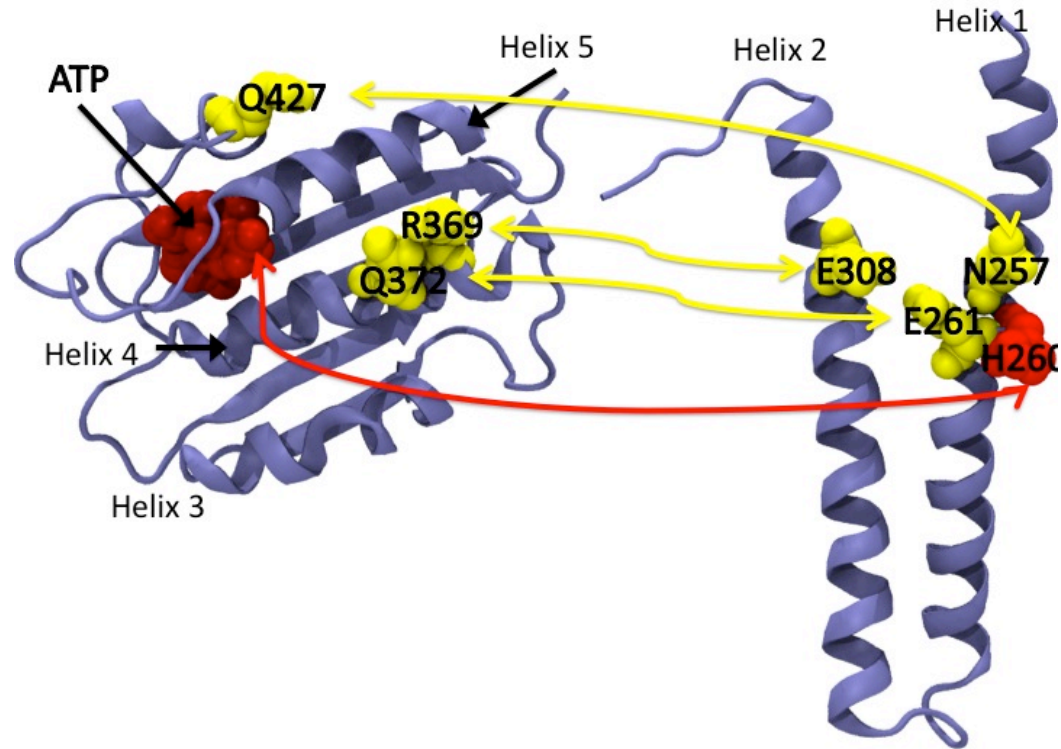
Independent X-Ray: P. Casino et al., Cell 2009

How do proteins move?

Prediction of Conformational Transitions

■ Interdomain contacts Domain 2 (HATPase_c)

Rank	Res 1	Res 2	d/Å	Domain
1	388	392	4.6	22
2	268	272	3.2	11
3	268	298	3.2	11
4	365	456	3.7	22
5	385	392	3.9	22
6	310	311	1.3	11
7	311	312	1.3	11
8	303	307	3.0	11
9	261	372	14.5	12
10	420	421	1.3	22
11	272	298	6.9	11
12	369	372	2.9	22
13	375	379	2.7	22
14	310	312	3.2	11
15	429	431	3.9	22
16	251	255	2.9	11
17	257	272	20.5	11
18	379	383	2.8	22
19	420	429	3.7	22
20	431	432	1.3	22
21	385	388	6.4	22
22	251	252	1.3	11
23	250	251	1.3	11
24	308	369	8.0	12
25	298	310	14.8	11
26	369	455	7.0	22
27	383	384	1.3	22
28	426	429	3.1	22
29	420	431	3.8	22
30	451	455	2.9	22
31	251	268	23.6	11
32	315	451	3.6	12
33	257	427	12.7	12
34	372	375	3.4	22
35	369	456	4.7	22
36	311	372	3.3	12

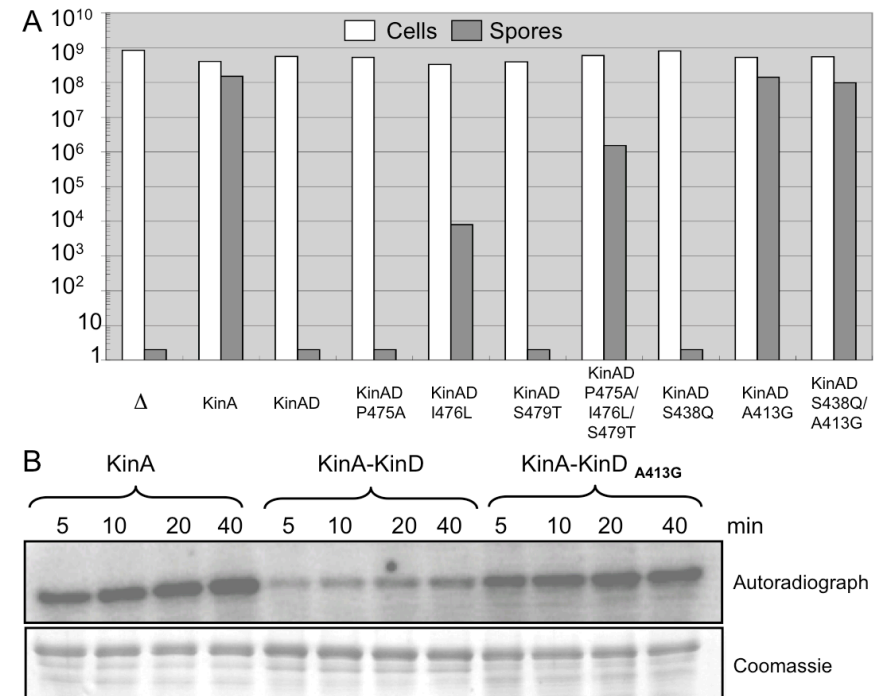


Domain 1 (HisKA)

Dago et al., PNAS (2012)

Verification of active conformation by *in-vivo* mutagenesis

- Biochemical data on hybrid kinases: mutations in helix 3 can fully destroy/restore function

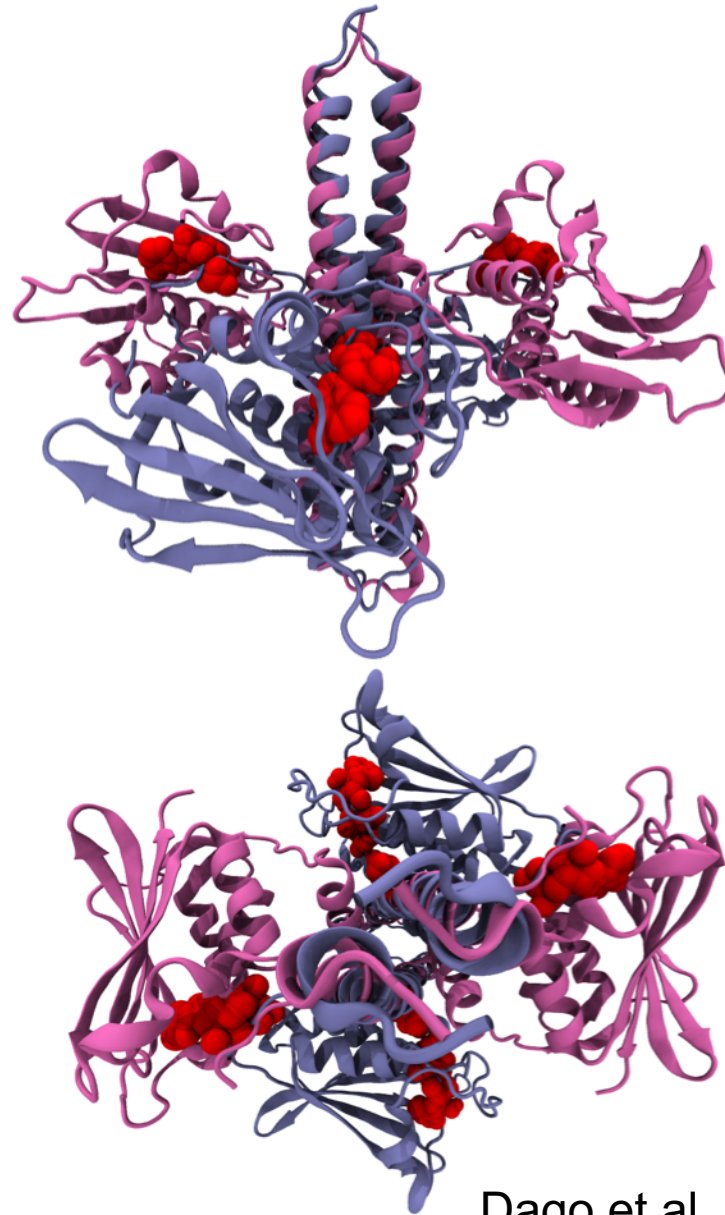


work by H. Szurmant, TSRI
Dago et al., PNAS (2012)



Active Conformation

- Our blind prediction 2012:
 - Contacts from statistical analysis
 - Molecular dynamics
 - Biochemical mutagenesis
- Confirmed
 - Wang et al., Plos Biol (2013)
 - Casino et al., Nat Comm (2014)

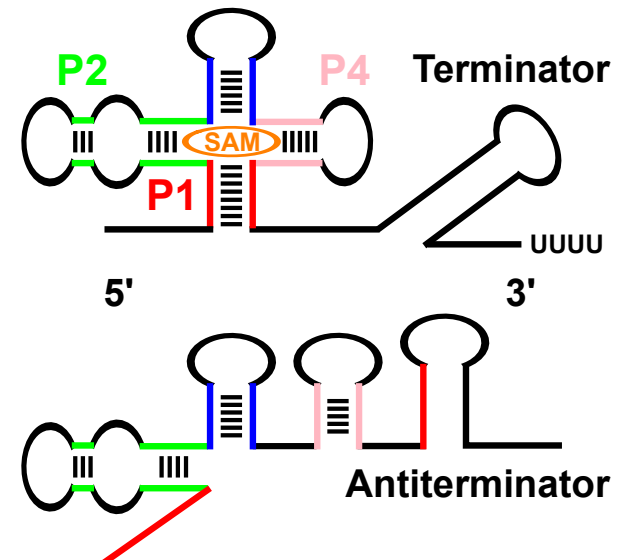
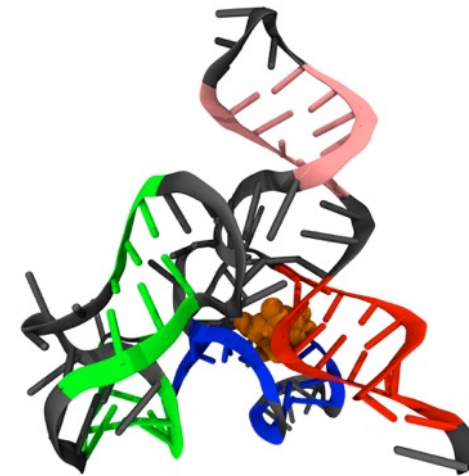


Dago et al., PNAS (2012)

Ok- it works for proteins. What else?

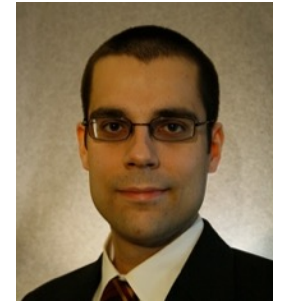
Riboswitches as Example for structured ncRNA

- Reminder: DNA (“storage”), RNA (“blueprint”), Protein (“tool”)
- Riboswitches at start of mRNA
 - Sense specific chemicals to switch between states
 - Genetic regulation (inhibition of translation, transcription termination...)



RNA Structure Prediction?

- 4+1-letter code vs. 20+1-letter code
- Dominance of 2-structure (“short range”)
 - 3-structure contacts (“long range”) for reliable prediction important
- ncRNA: Riboswitches “hard case”



- Select “small gold-standard” for RS prediction from RFAM

- high number sequences
 - high resolution structure
 - Sequence < 100

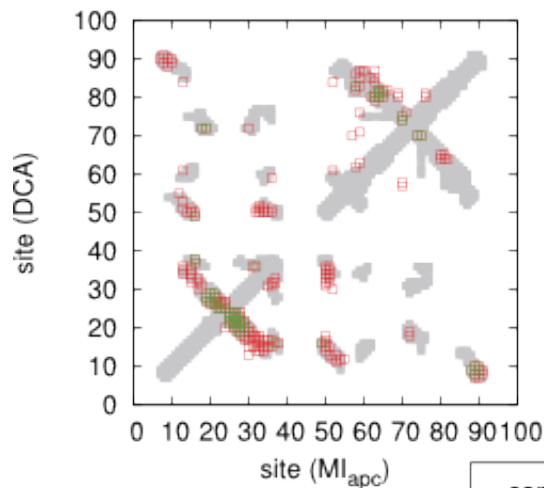
Name of Riboswitch	PDB	Resolution PDB (Å)	Rfam Code	(effective) Number sequences in Rfam	PDB sequence length	Rfam sequence length
Adenine	1y26	2.1	RF00167	2427 (588.88)	71	102
TPP	2gdi	2.05	RF00059	11197 (3347.90)	80	104
SAM-I	2gis	2.9	RF00162	4757 (1165.56)	94	107
c-di-GMP	3irw	2.7	RF01051	1990 (983.21)	98	87
Glycine	3owi	2.85	RF00504	6875 (1828.54)	88	90
Fluoride	3vrs	2.6	R01734	1267 (532.03)	52	64

De Leonardis et al., NAR 2015

Contact Maps

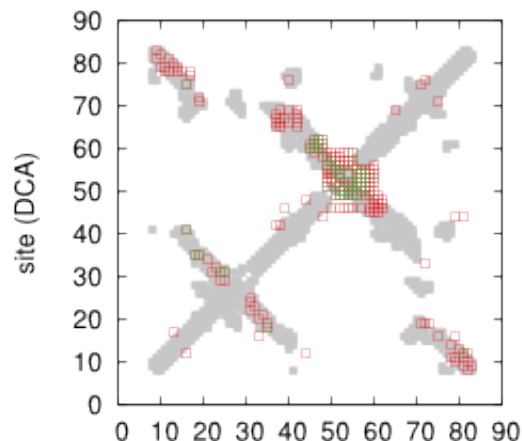
TPP

RF00059 – 2gdi



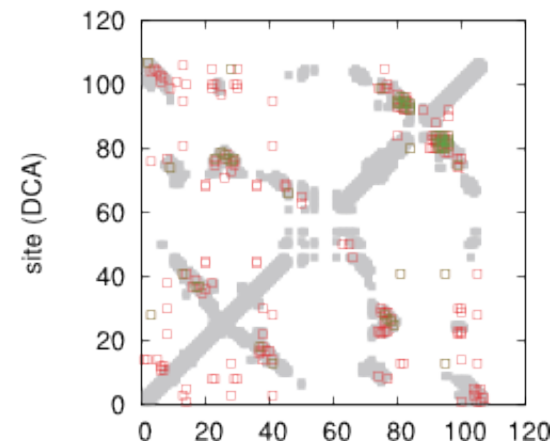
Glycine

RF00504 – 3owi



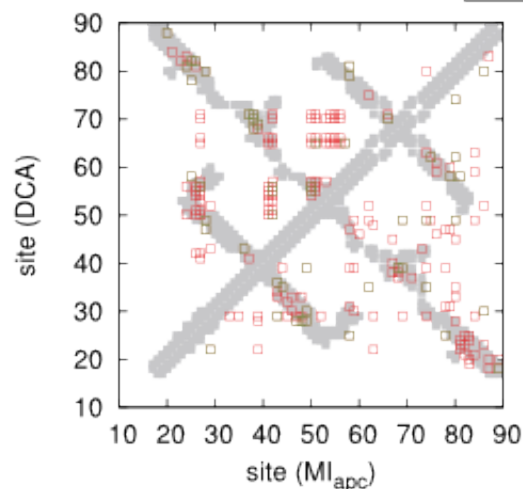
SAM I

RF00162 – 2gis



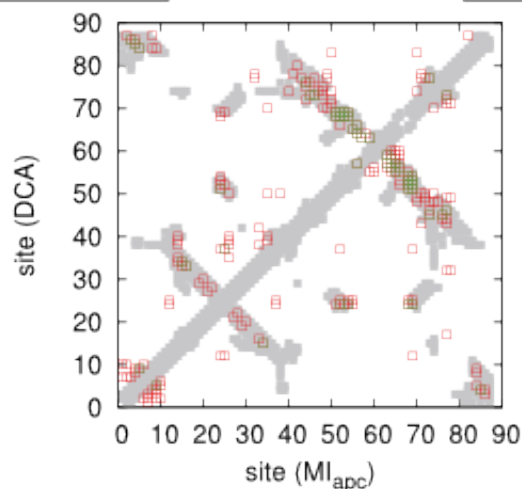
Adenine

RF00167 – 1y26



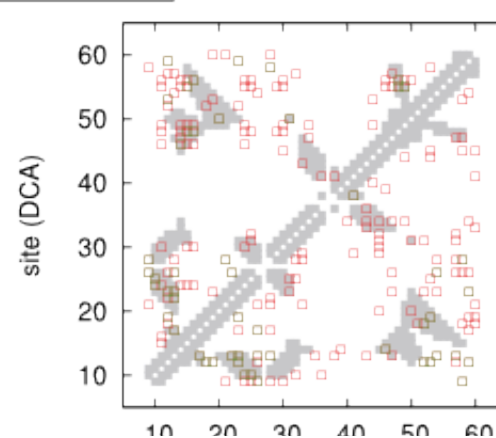
c-di GMP

RF01051 – 3irw



Fluoride

RF01734 – 3vrs

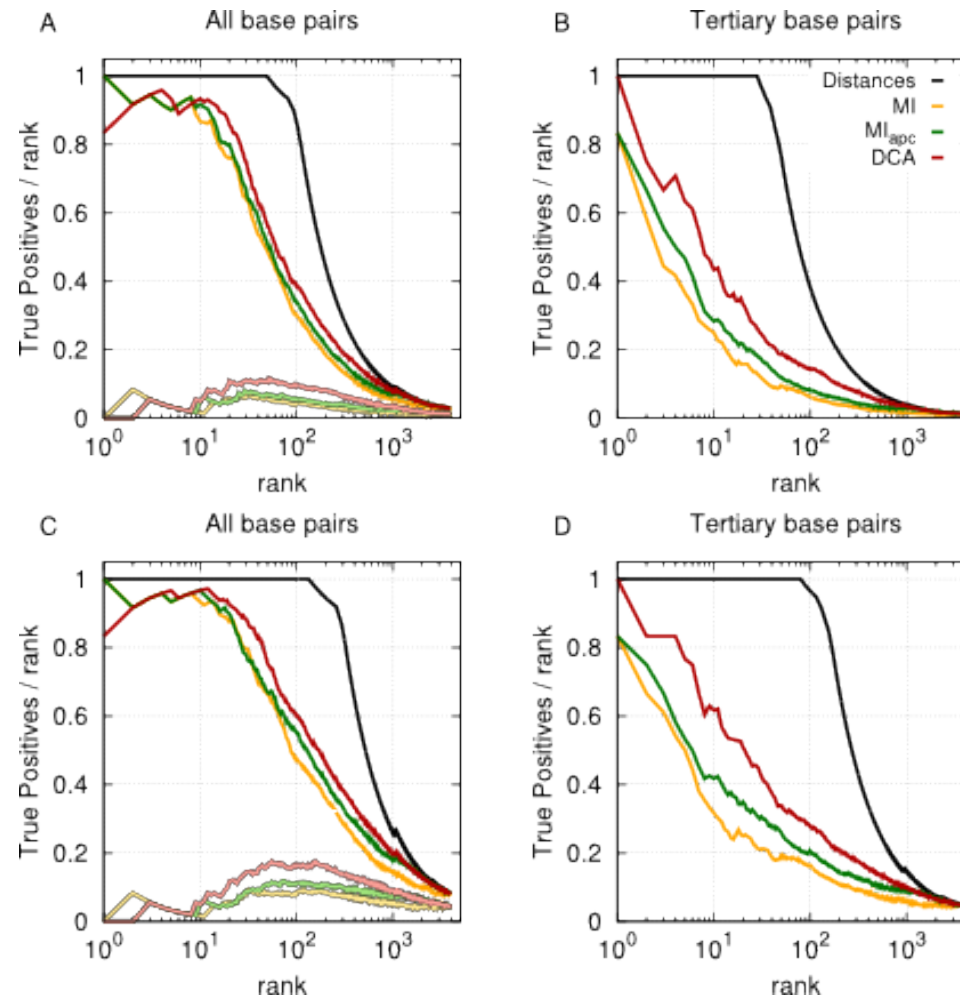


contacts 8Å
rank 26–100
rank 1–25

contacts 8Å
rank 26–100
rank 1–25

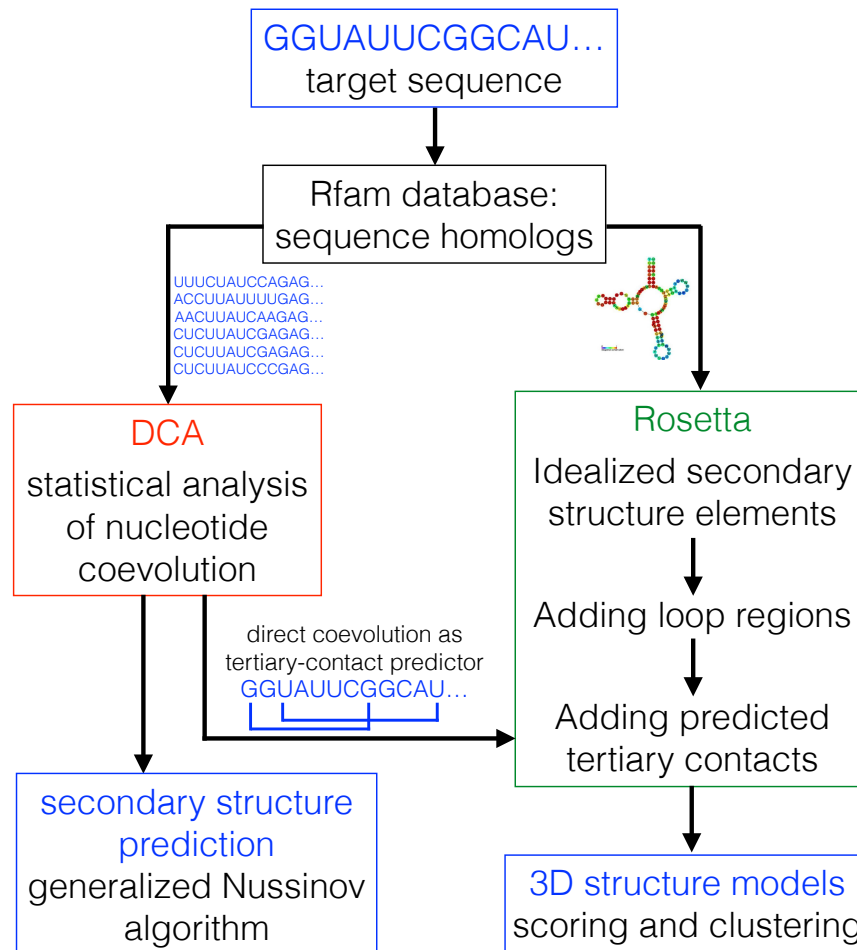
De Leonardis et al., NAR 2015

Precision of Contacts



De Leonardis et al., NAR 2015

Prediction Workflow

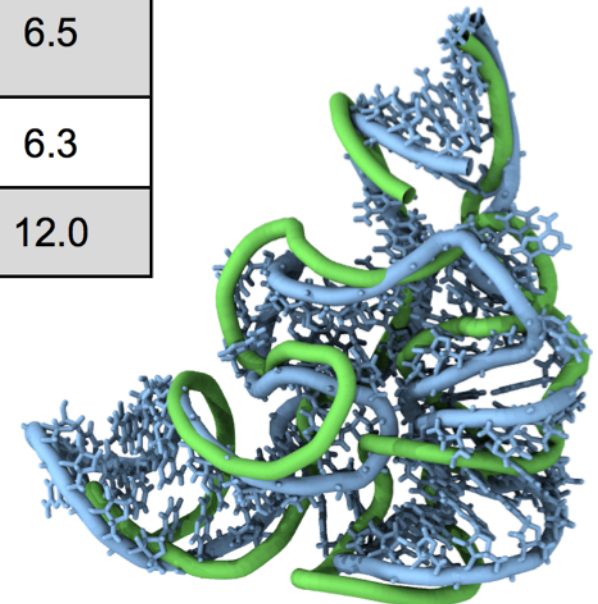


De Leonardis et al., NAR 2015

Prediction Results for 6 Riboswitches (best of 10)

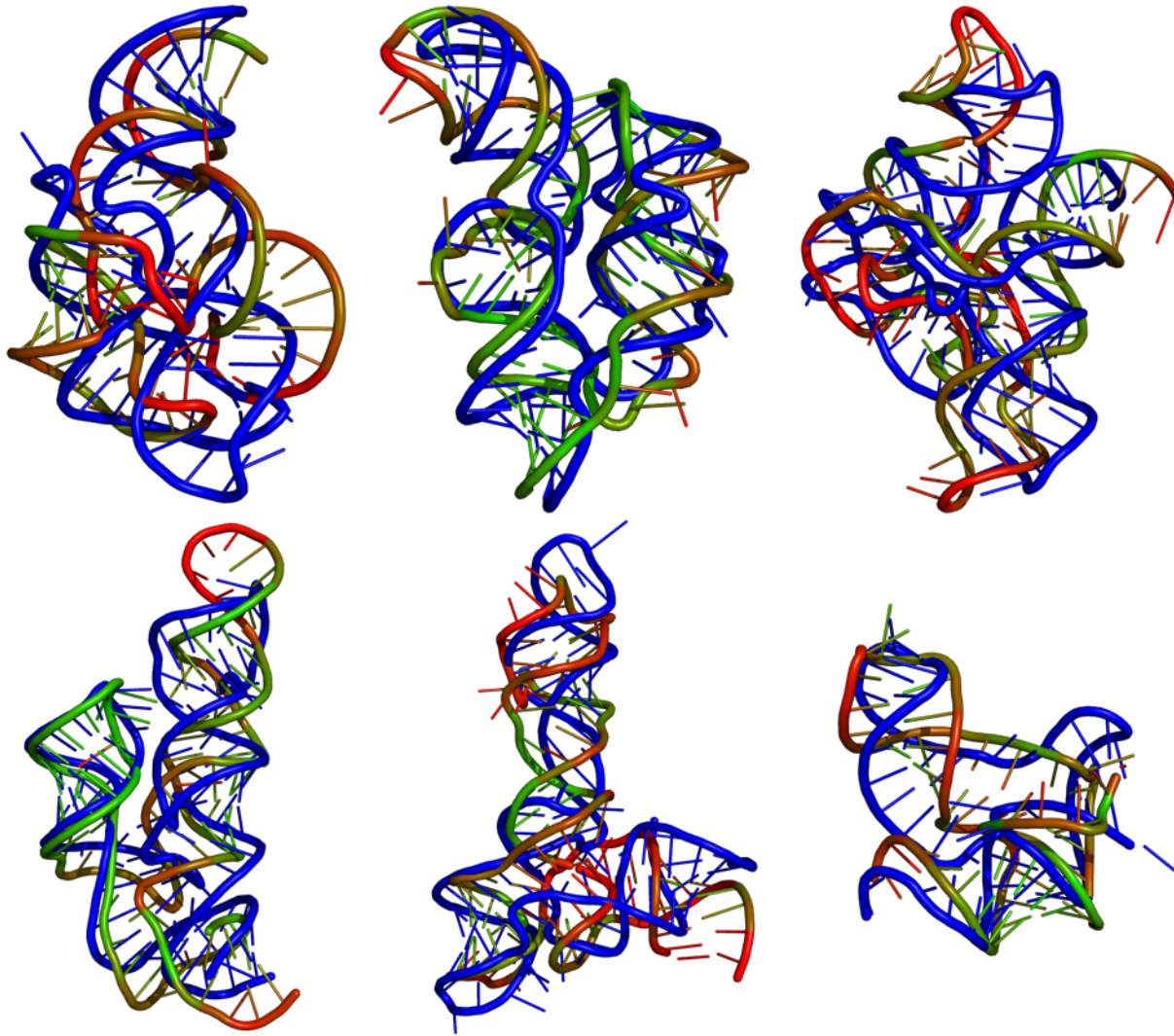
PDB Rfam	Secondary only	+ 100 Mlapc	+ 100 DCA	+ All
1y26 RF00167	12.6	10.2	11.3	7.4
2gdi RF00059	16.1	17.6	6.7	6.3
2gis RF00162	17.4	7.3	10.6	5.8
3irw RF01051	10.2	7.6	7.8	6.5
3owi RF00504	13.5	14.4	11.6	6.3
3vrs RF01734	12.5	13.4	8.8	12.0

TPP-RS (2gdi)
RMSD 6.7 Å



De Leonardis et al., NAR 2015

3d Predictions

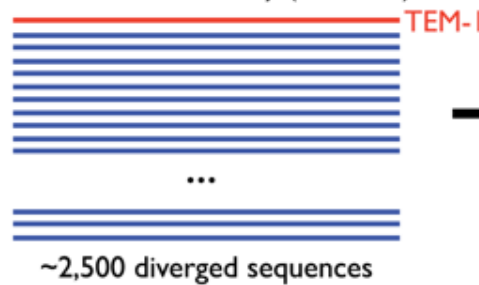


De Leonardis et al., NAR 2015

New Directions: Function

- Infer of protein mutational landscapes to predict phenotypic effect of mutations
- beta-lactamase TEM-1
- Providing resistance against beta-lactam antibiotics

Beta-lactamase2 family (PF13354)



Statistical landscape inference (DCA)

$$\phi(a_1, \dots, a_L)$$

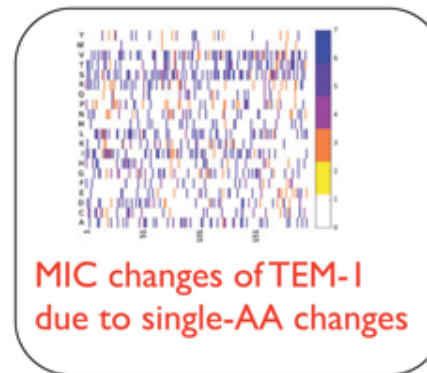
$$= \sum_i \varphi_i(a_i) + \sum_{i,j} \varphi_{ij}(a_i, a_j)$$



Score for mutant AA sequences

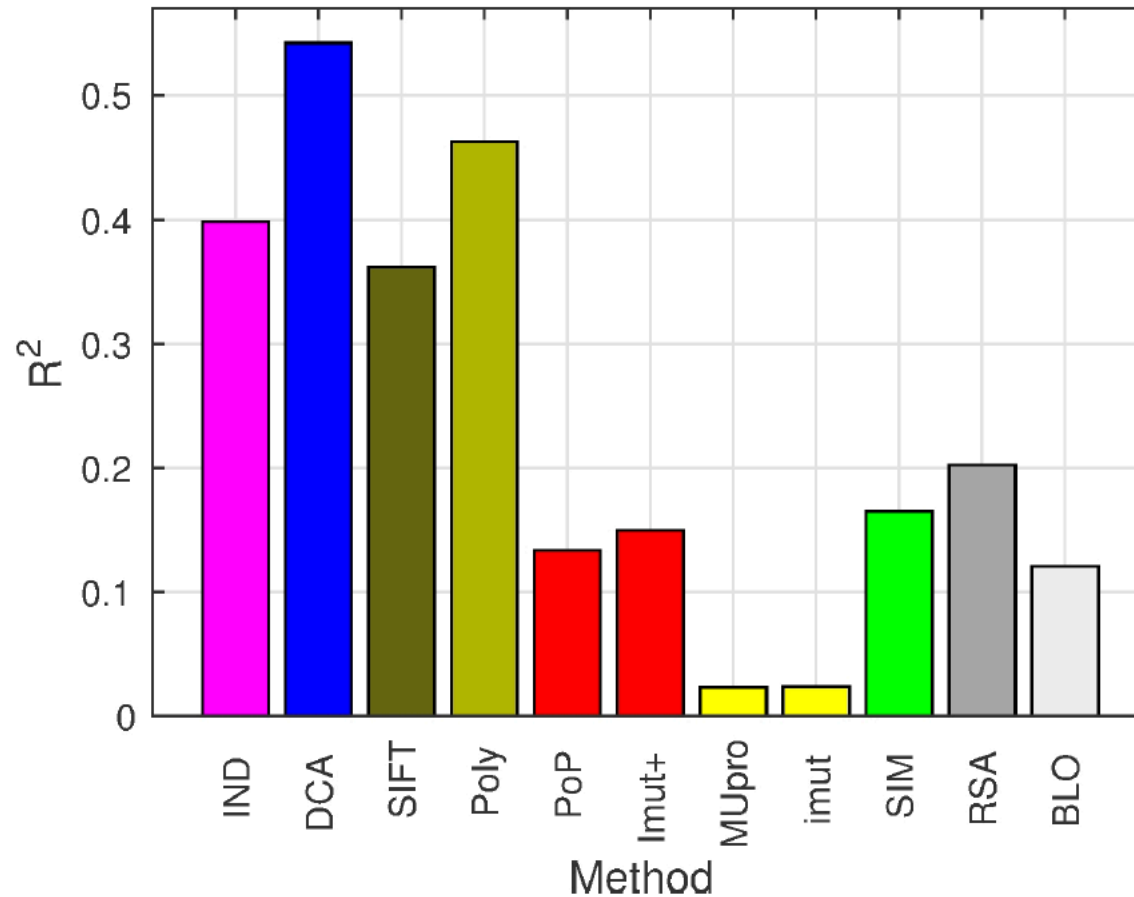
$$\Delta\phi = \phi(\text{mutant}) - \phi(\text{wildtype})$$

Evolutionary constraints
across diverged homologs

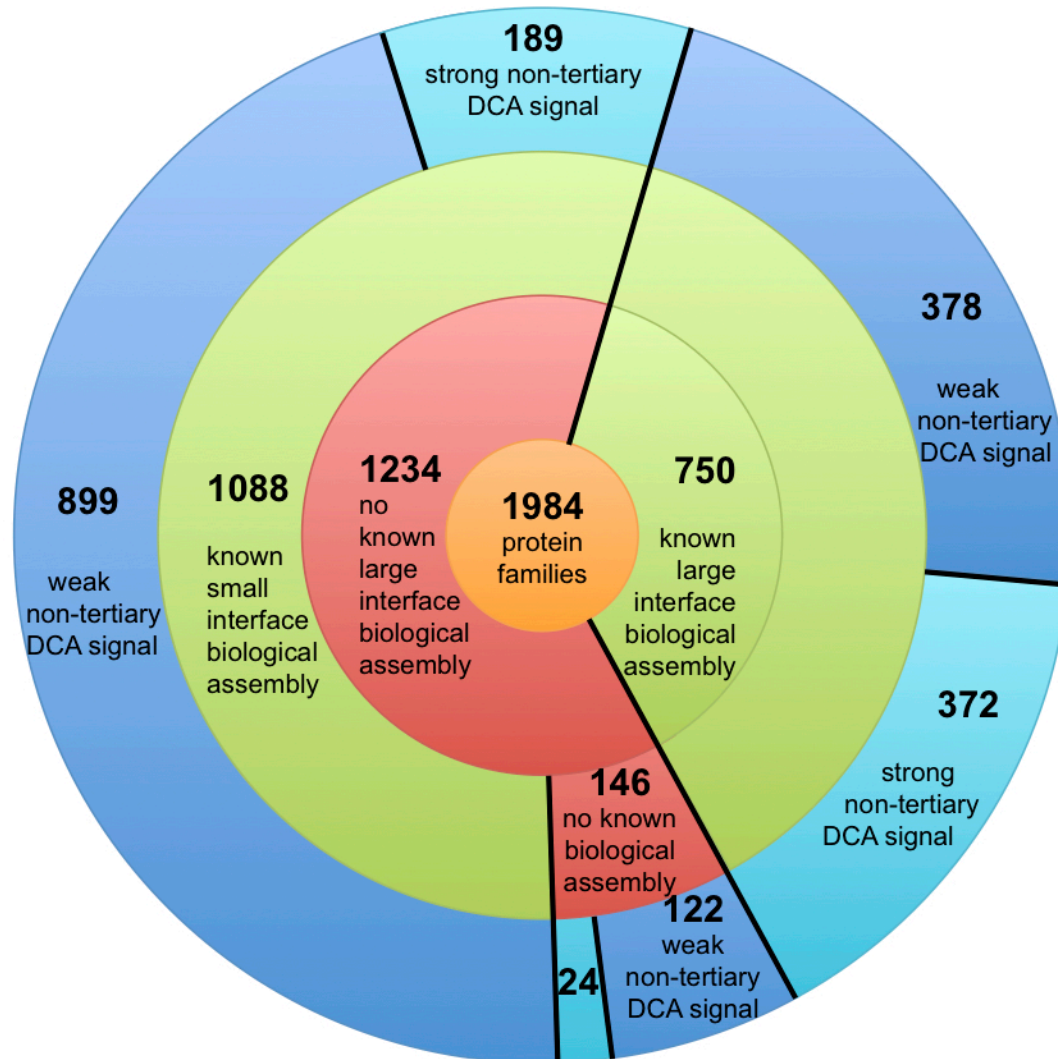


M. Figliuzzi et al., MBE 2016

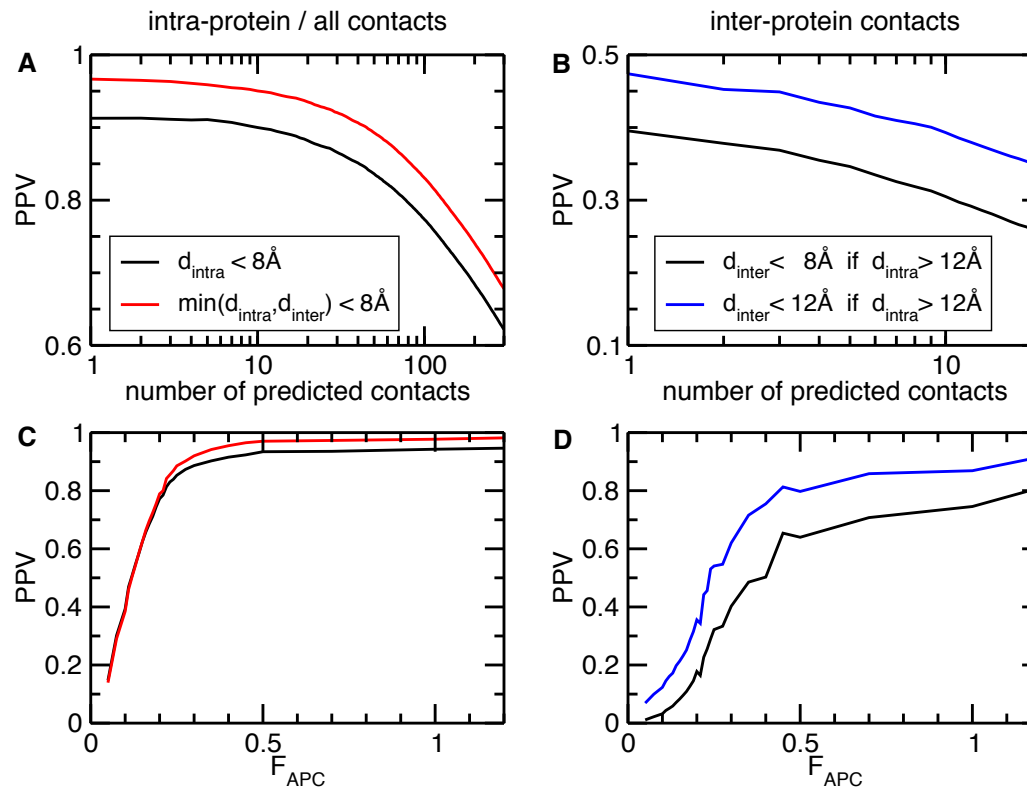
New Directions: Function



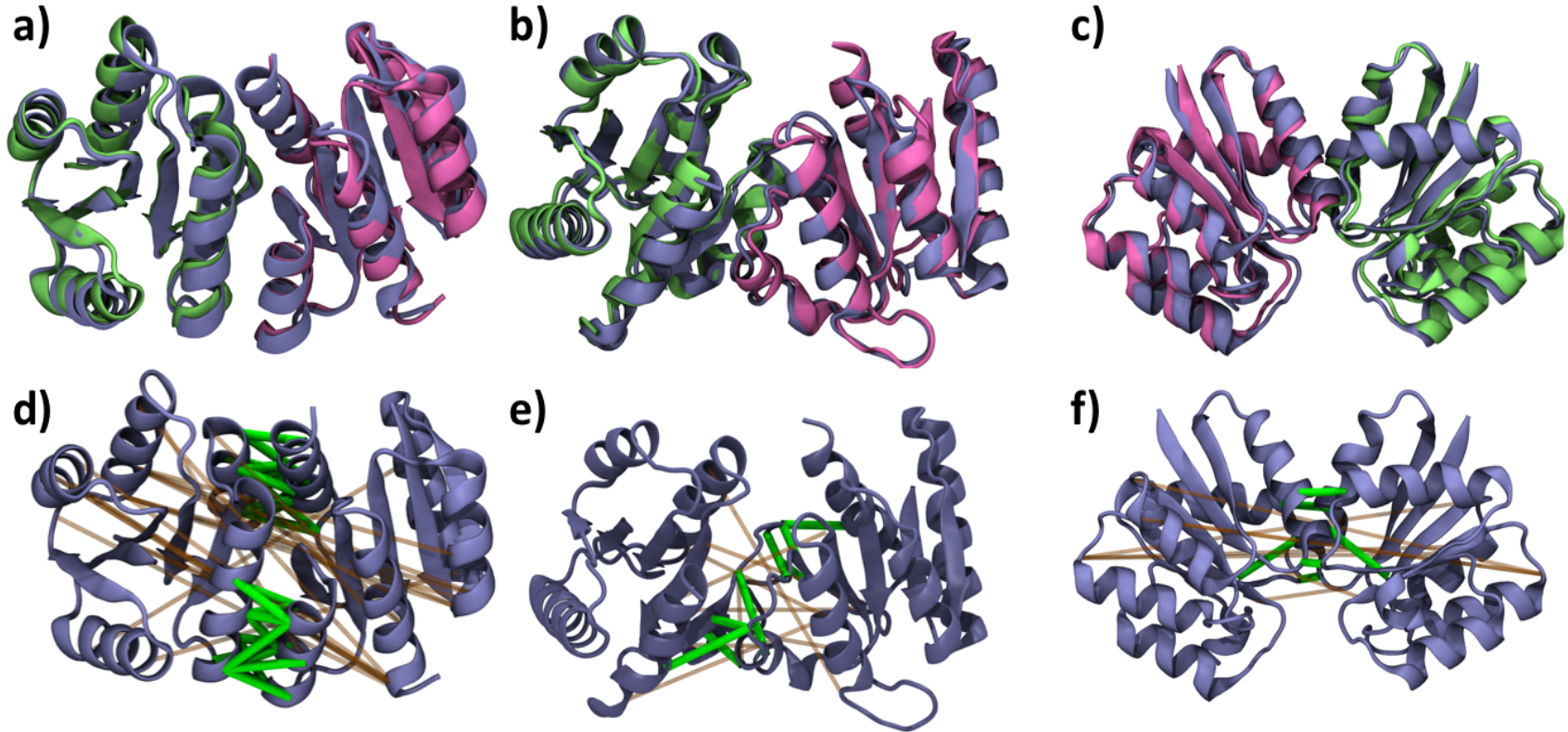
New directions: Large Scale (unpublished results)



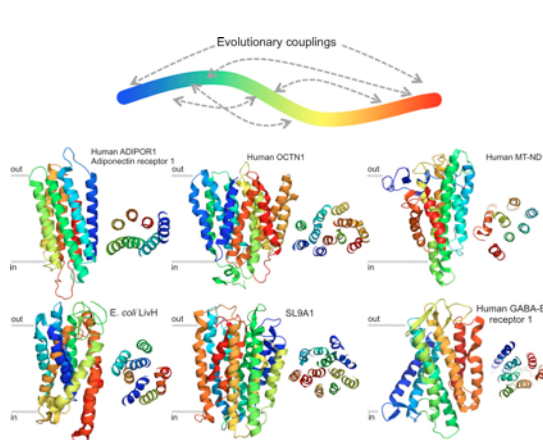
Large Scale (unpublished results)



Large Scale (unpublished results)



Protein & RNA Structure Prediction by Integrating Genomic Information into Molecular Simulation



Complex Formation:

AS et al., PNAS (2009)
Ovchinnikov et al., eLife (2014)
Hopf et al., eLife (2014)

Active Conformation:

Dago et al., PNAS (2012)
Morcos et al., PNAS (2013)

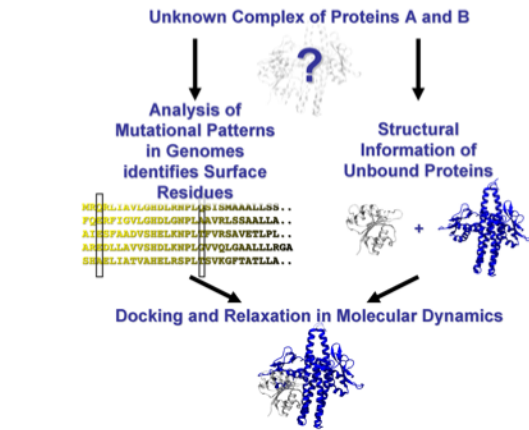
Protein Structure Prediction:

Sulkowska et al., PNAS (2012)

Transmembrane Regions:

Hopf et al., Cell (2012)

[...]

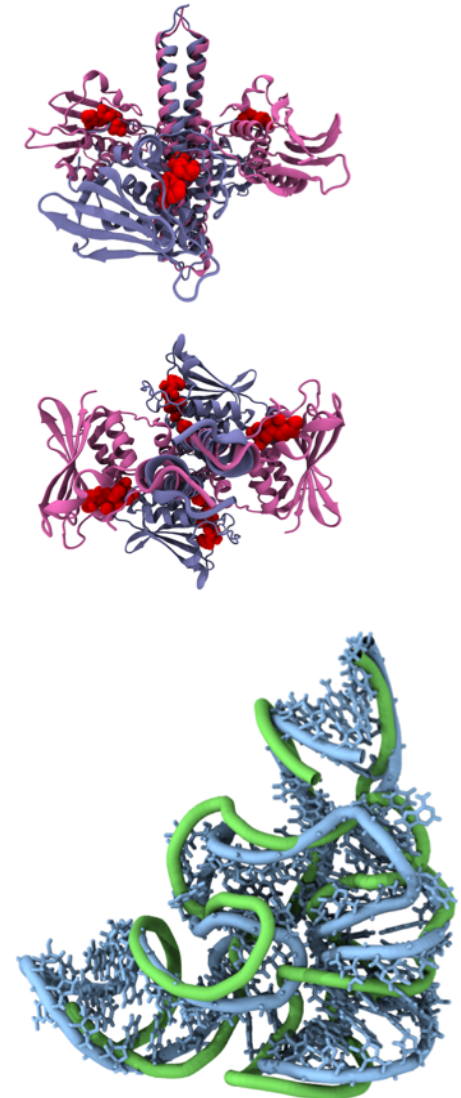
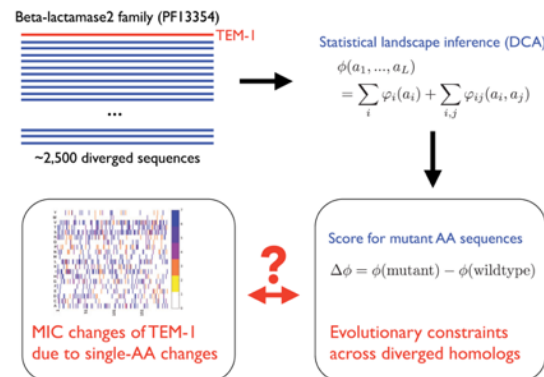


RNA:

De Leonardis et al. NAR (2015)
Weinreb et al., Cell (2016)

Landscape Inference:

Figliuzzi et al., Mol. Biol. Evol. (2016)



Group

■ Shalini John Lovis

■ Benjamin Lutz

■ Fabian Nagel

■ Claude Sinner

■ Sebastian Ratz

■ Ines Reinartz

■ Abhinav Verma

■ Ilaria Mereu

■ Martin Weigt (U Pierre et Marie Curie, Paris)

■ Remi Monasson (U Pierre et Marie Curie, Paris)

■ Simona Cocco (U Pierre et Marie Curie, Paris)

■ Hendrik Szurmant (Western University of Health Sciences, LA)

■ Ben Schuler (U Zürich)

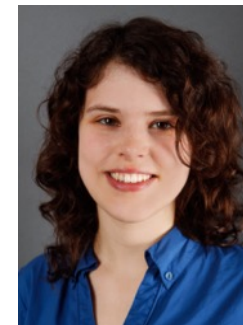
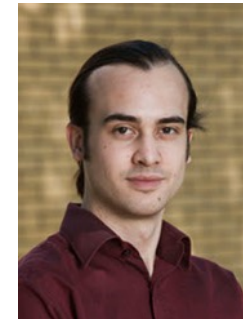
■ Ulrich Nienhaus (KIT)

■ Steffen Scholpp (KIT)

■ Stefan Klumpp (U Potsdam)

■ Jose Onuchic (Rice University)

■ Magnus Wolf-Watz (Umeå U)



Funding

