On the Importance of Physically Correct Models for Protein-Ligand Binding

Martin Lepšík, A. Imberty, P. Hobza, J. Řezáč

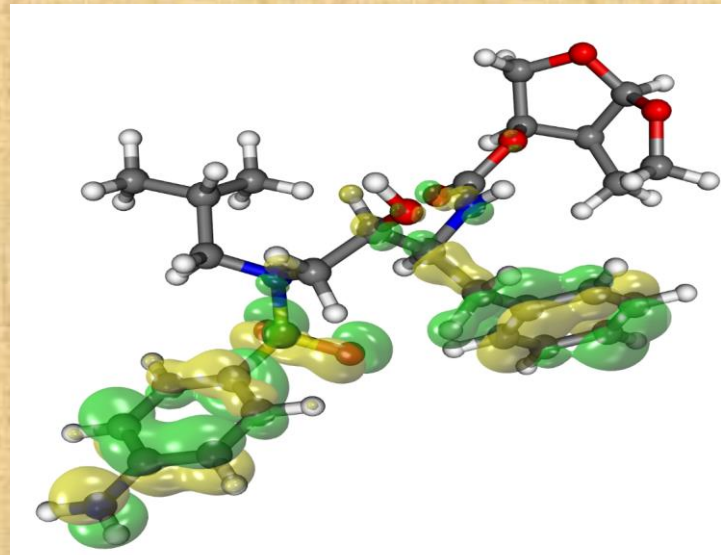
lepsik@uochb.cas.cz

Němčice, ENBIK 06/2022

Ligand Design via Quantum Chemical Scoring

M. Lepšík, J. Fanfrlík, A.Pecina, S. Eyrilmez, C. Köprülüoglu, P. Hobza, J. Řezáč

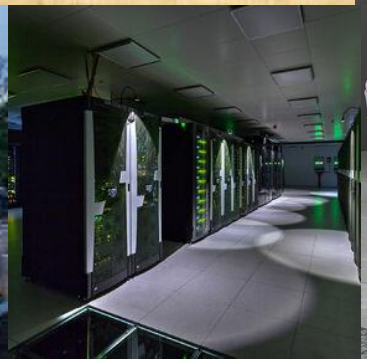
IOCB Prague, Czech Academy of Sciences

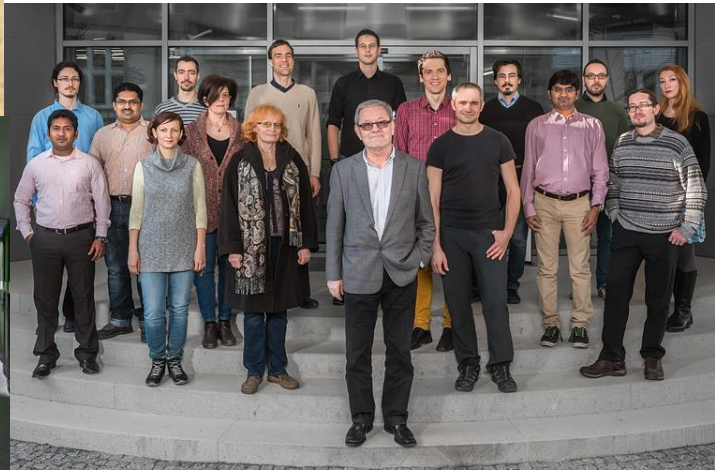


ÚOCHB AV ČR IOCB PRAGUE

IT4I: 20+ MCPU hours

IOCB Tech





Computer-Aided Drug Design

Structure-based approach

- Target/Receptor (protein)
- Ligand (small molecule / drug)

- 3D structures (X-ray crystallography, NMR, cryo-EM)

- non-covalent interactions governing the affinity

Aims

- prioritize compounds for synthesis
- exclude non-binders

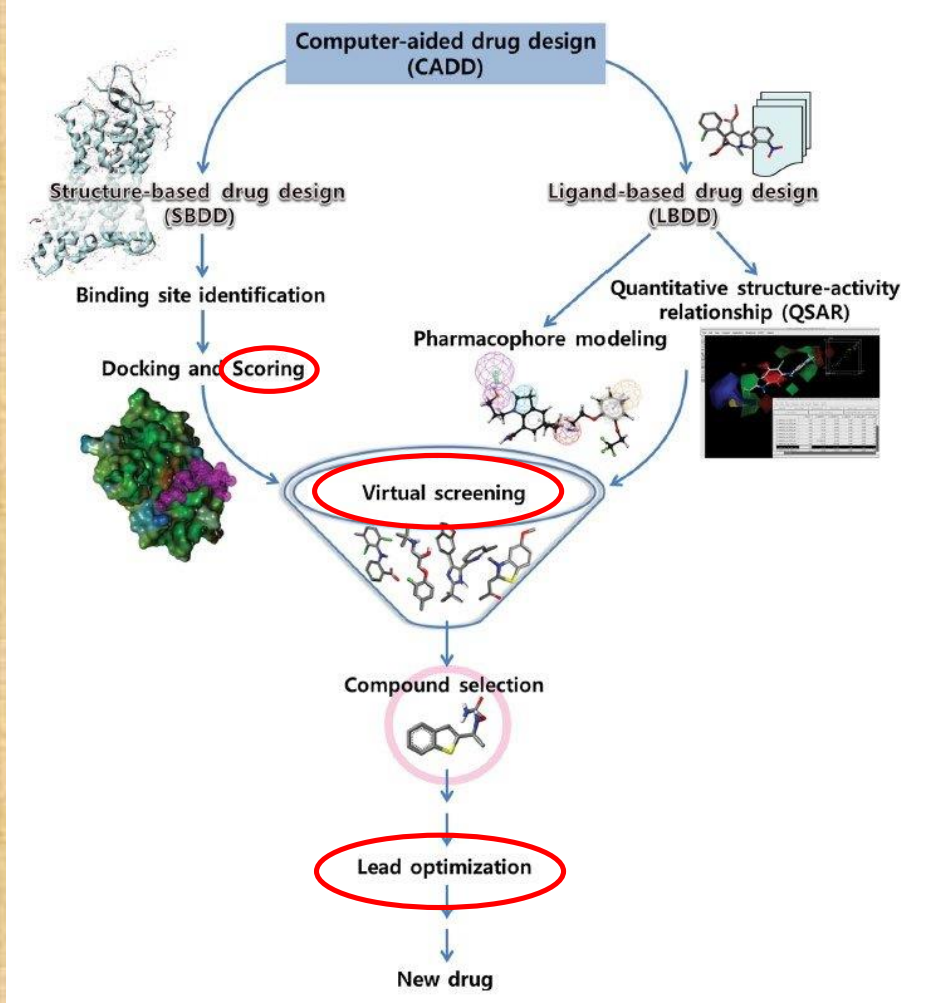


Image: Prajapat P, Agarwal S, Talesara GL, *J. Med Org Chem* **2017**

What is scoring?

Predicting the strength of protein-ligand interaction from structure

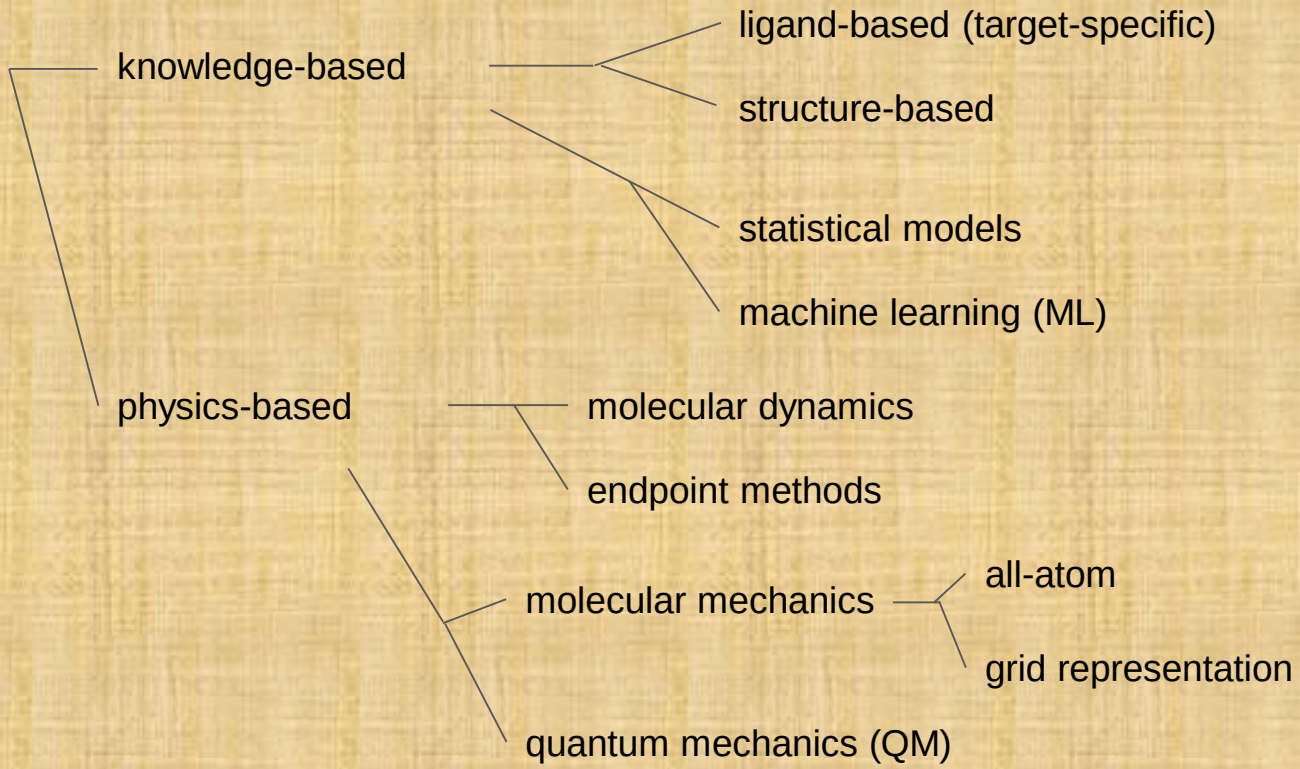
ΔG_{bind} = -12.456 kcal/mol

What is scoring?

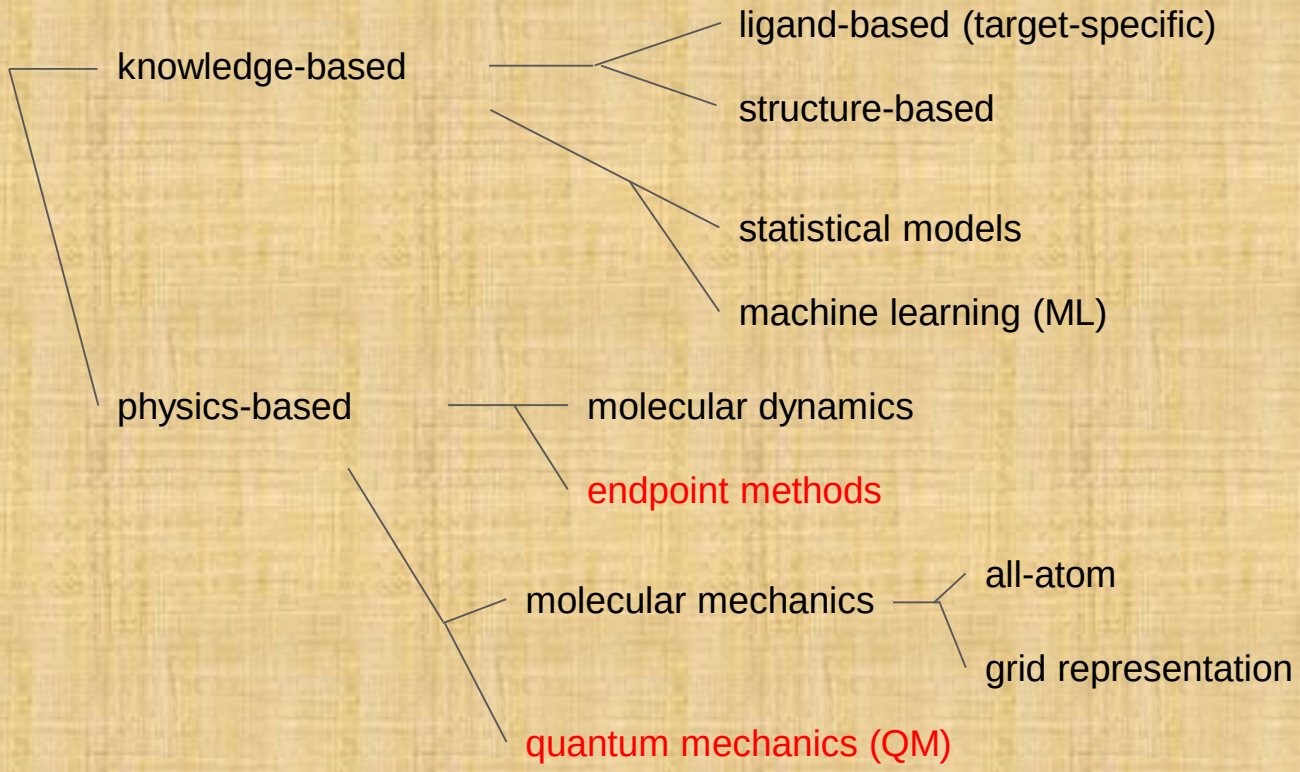
Predicting the strength of protein-ligand interaction from structure

Score = -42.01 units

Scoring function “taxonomy”



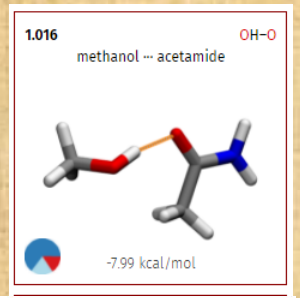
Scoring function “taxonomy”

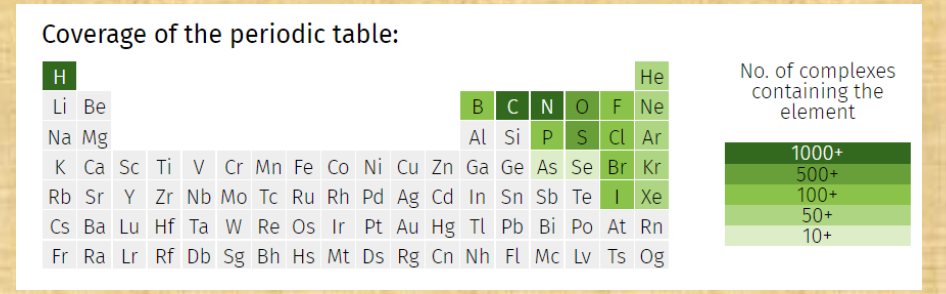


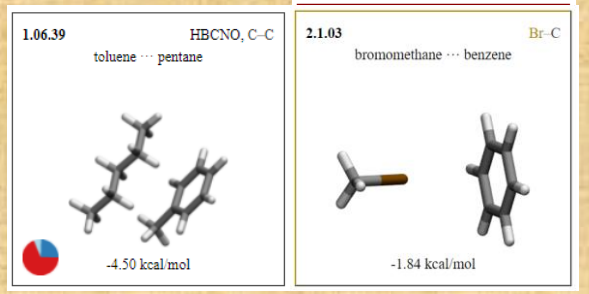
Universal Reliable Scoring Function

QM Methods for Non-covalent Interactions

- Small models, accurate calculations (CCSD(T)/CBS) in vacuum
- www.nciatlas.org (~ 20,000 data points)
- hydrogen bonding, dispersion, sigma-hole interactions,…..



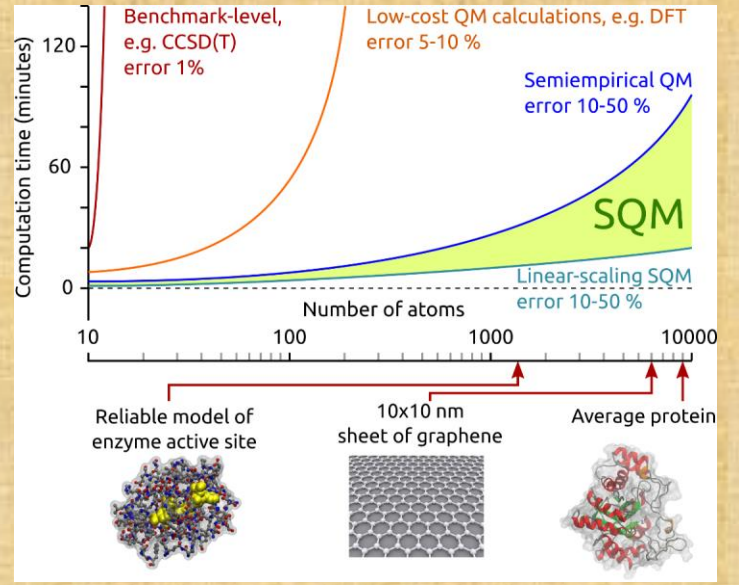




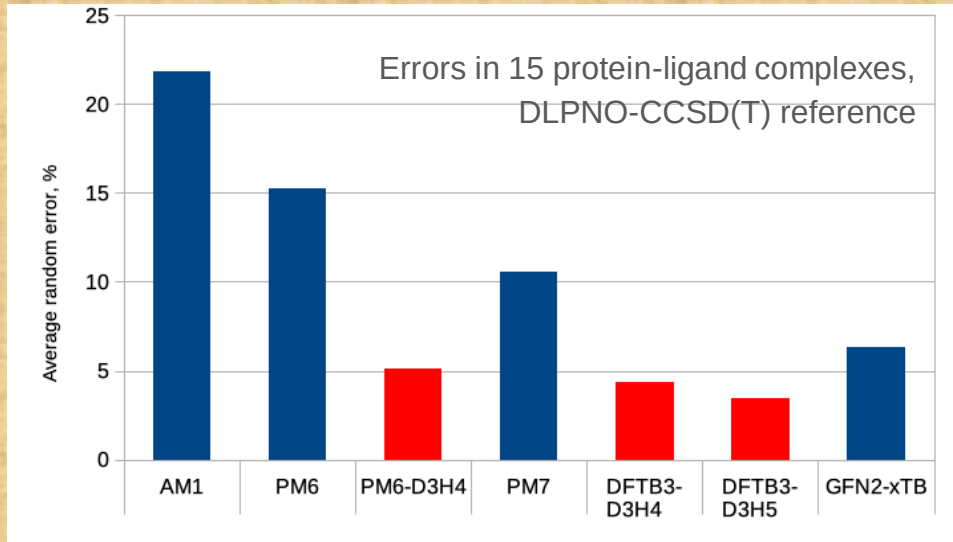
- Development of semiempirical QM methods - corrections for non-covalent interactions
  chemical accuracy (1 kcal/mol) in small dimers

Semiempirical QM methods

- Fast calculation
- Easy preparation
  (no system-specific parameters)
- Accuracy?



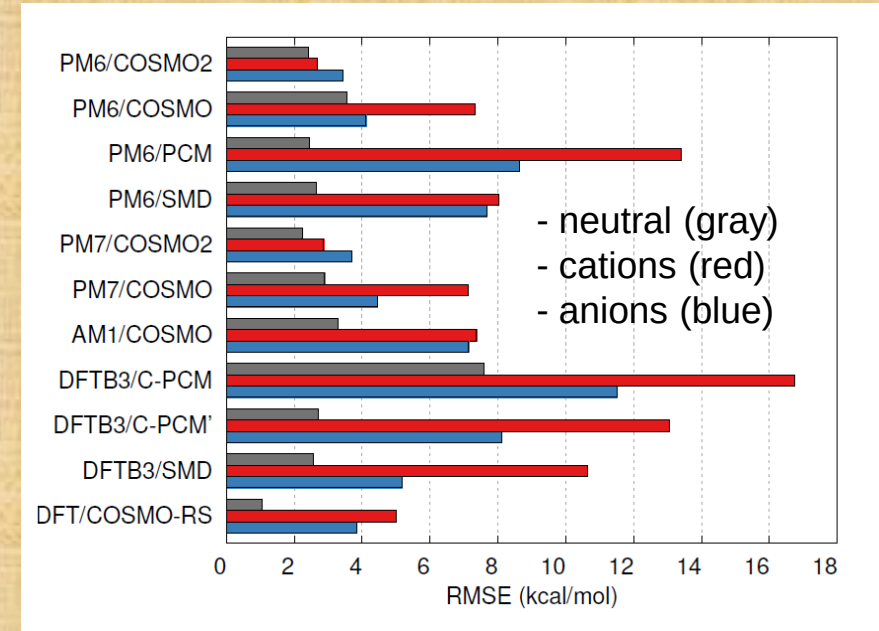
correcting SQM methods
for non-covalent interactions[1-3]



[1] Řezáč et al.; *J. Chem. Theory Comput.* **2009**, 5, 1749
[2] Řezáč and Hobza.; *J. Chem. Theory Comput.* **2012**, 8,141
[3] Řezáč; *J. Chem. Theory Comput.* **2017**, 13, 4804

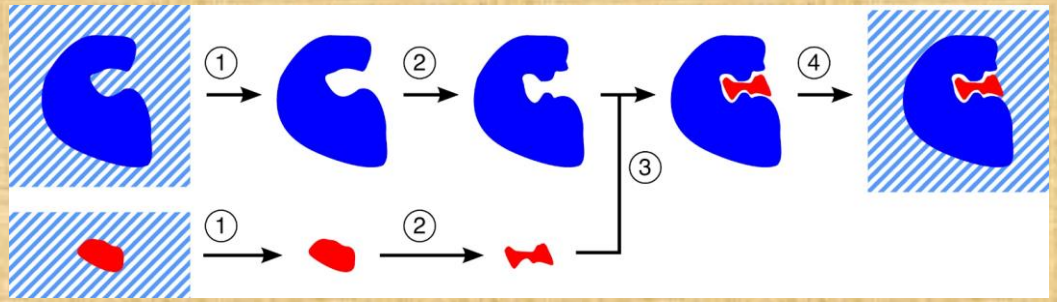
Implicit Solvation Models

- reparametrisation of COSMO → COSMO2



Kříž, Řezáč; *J. Chem. Inf. Model.* **2019**, 59, 229

SQM-based Scoring function



Modular physics-based approach:
components can be replaced if better alternative exists

Score = ΔE_{int} ← PM6-D3H4X + further corrections

+ $\Delta\Delta G_{solv}$ ← PM6/COSMO2

+ $\Delta G_{conf,w}(L)$ ← Optimized free molecule / optional conformation search

+ $\Delta G_{conf,w}(P)$

– $T\Delta S$ ← LM5 model fitted to QM data

Fanfrlík et al.; *J. Phys. Chem. B* **2010**, *114*, 12666
Lepšík et al.; *ChemPlusChem* **2013**, *78*, 921
Pecina et al.; *ChemPlusChem* **2020**, *85*, 2362

QM/MM Setup

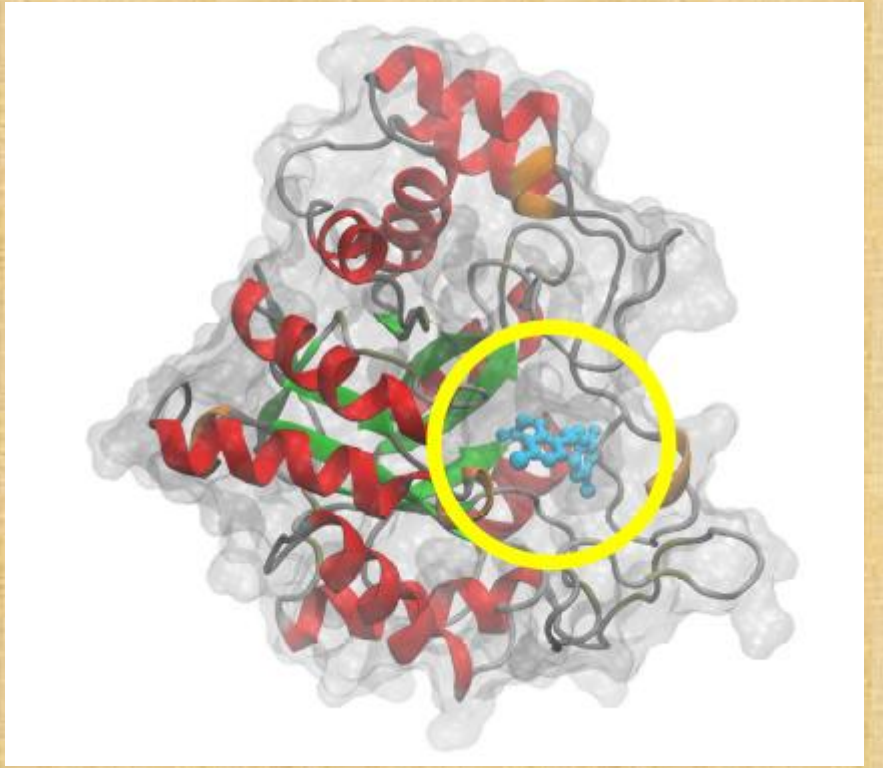
- Ligand ~ 10 to 100 atoms
- Protein ~ 10 000 atoms

- We consider model of the active site with ~1500 atoms
  (10 Å sphere around ligand)
- Proven to converge to the results obtained in the whole protein

- One protein conformation

for a series of ligands

- QM/MM geometry optimization
  + many more steps



Questions

Is SQM-score generally applicable?

How does it compare to commonly used scoring functions (SF) in academia/industry?

Verification

Evaluation against experimental “truth” in multiple diverse data sets

- Input: Experimental structures or a reliable model
- Comparison with experimental affinities

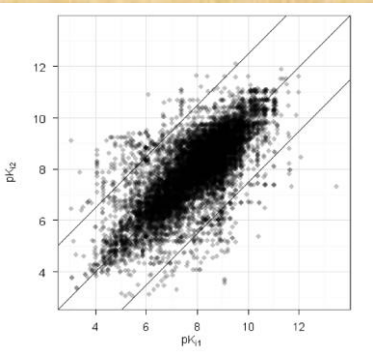
Experimental data

All data

PDBBind database ~20,000 systems

Reproducibility from multiple independent measurements - $R^2 = 0.8$

No time to prepare each system carefully



DOI: 10.1021/jm300131x

Reliable data

Reliable structures, preferably crystal

Measurements from one lab

Only tens of target / ligand series

Careful preparation of each protein

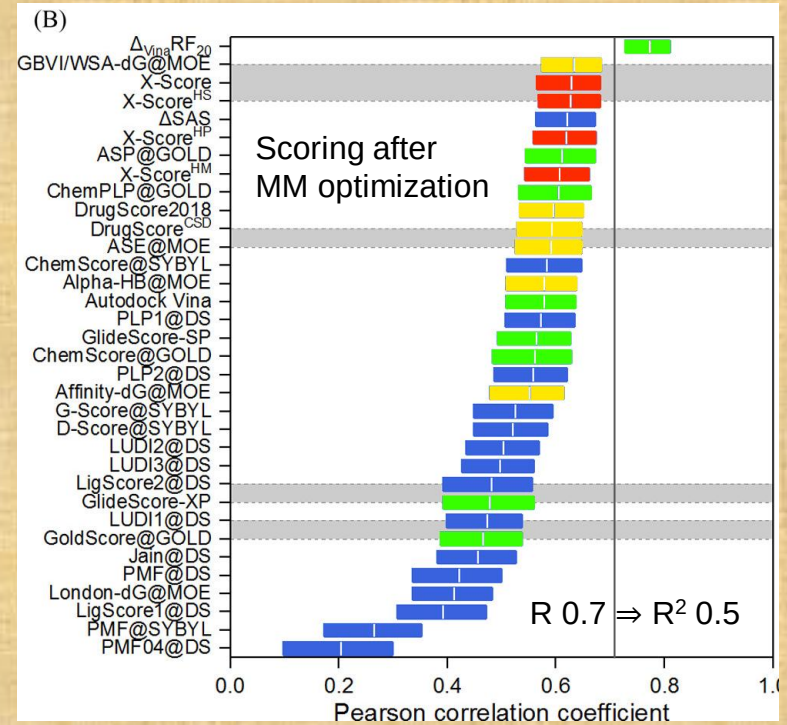
Current Status

- Best SFs in the CASF2016[1]
-
- Structure-based machine learning

× MD-based methods (FEP)

Timing:

- Empirical SFs <= seconds
- SQM-score ~ 30 minutes

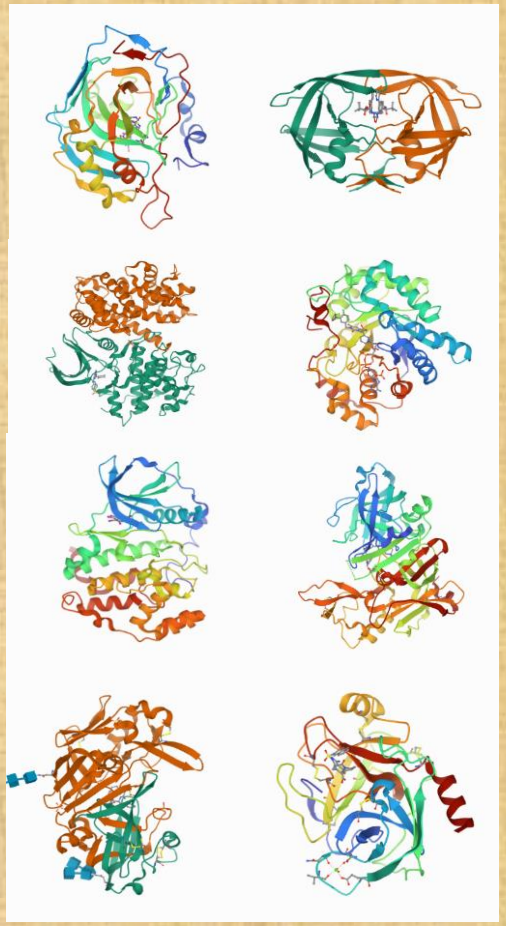


[1] M. Su, Q. Yang, Y. Du, G. Feng, Z. Liu, Y. Li, R. Wang, *J. Chem. Inf. Model.*, **2018**.

Diverse protein-ligand datasets

- consistent inhibition constants, IC_{50}
- reliable crystal structures

		Ligands	Similarity	Crystals	Expt.
001	Carbonic anhydrase 2	10	0.32	10	Ki
002	HIV protease	22	0.43	15+	Ki
003	CDK2	12	0.38	15	Ki
004	Casein kinase 2	9+	0.47	9+	Ki
006	Aldose reductase	14	0.47	14	Ki + IC50
010	CDK2	21	0.88	1	IC50
011	Cathepsin D	10	0.71	3	IC50
032	BACE1	16	0.48	20	IC50
038	JAK1	10	0.58	12	Ki
043	Trypsin	10	0.71	5	Ki



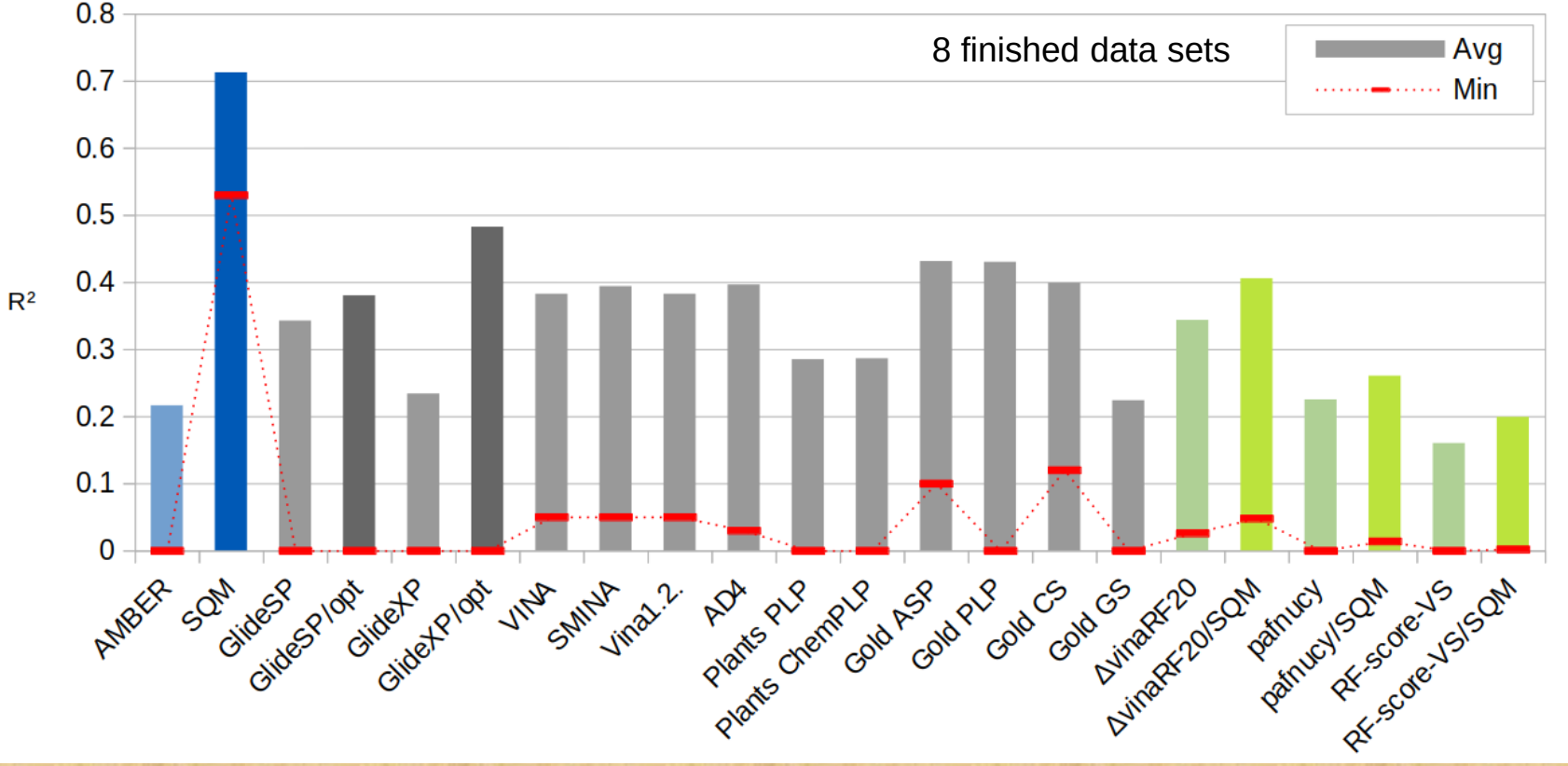
SQM-score Performance

$$\Delta G_{bind} = RT \ln K_i$$

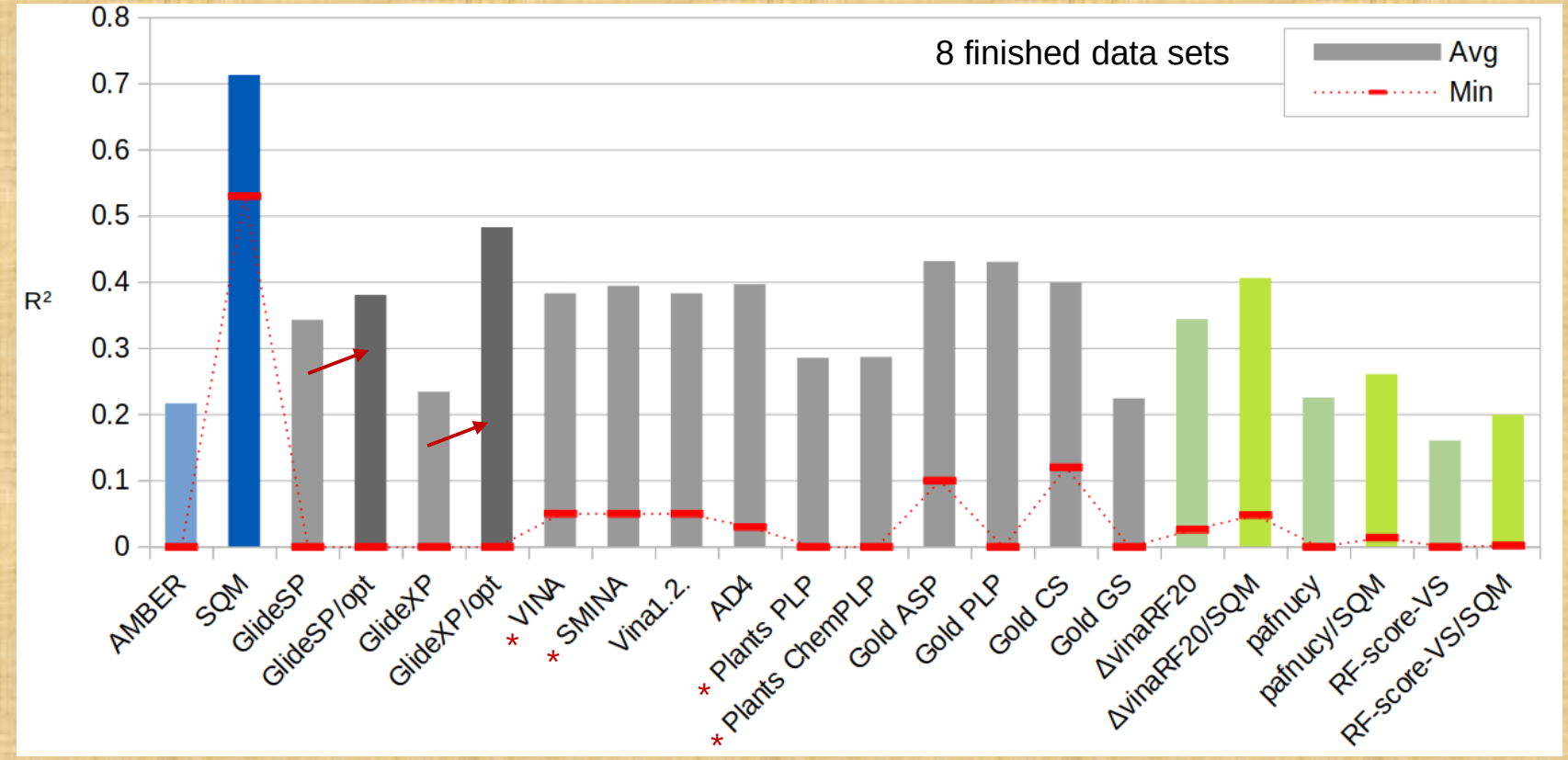
R^2 to experiment			
	Ligands	v 1.02	v 2.19
001-CA2	10	0.21	0.66
002-HIV-PR	12	0.41	0.55
003-CDK2	12	0.53	0.83
004-CK2_Dobes	9	0.64	0.65
006+037-AR	14	0.61	0.78
010-CDK2-biphenyls	21	0.55	0.84
011-Cath-D	10	0.62	0.75
032-BACE1-challenge	16	0.01	0.56
38-JAK1	10	0.04	0.83
AVERAGE		**0.40**	**0.72**

- **multi-step optimization protocol,** focus on fixing H-bond networks
- tight control of SQM calculation
- **reparametrized H4 and X corrections**
- additional corrections for sulfur
- halogen bonding correction in MM

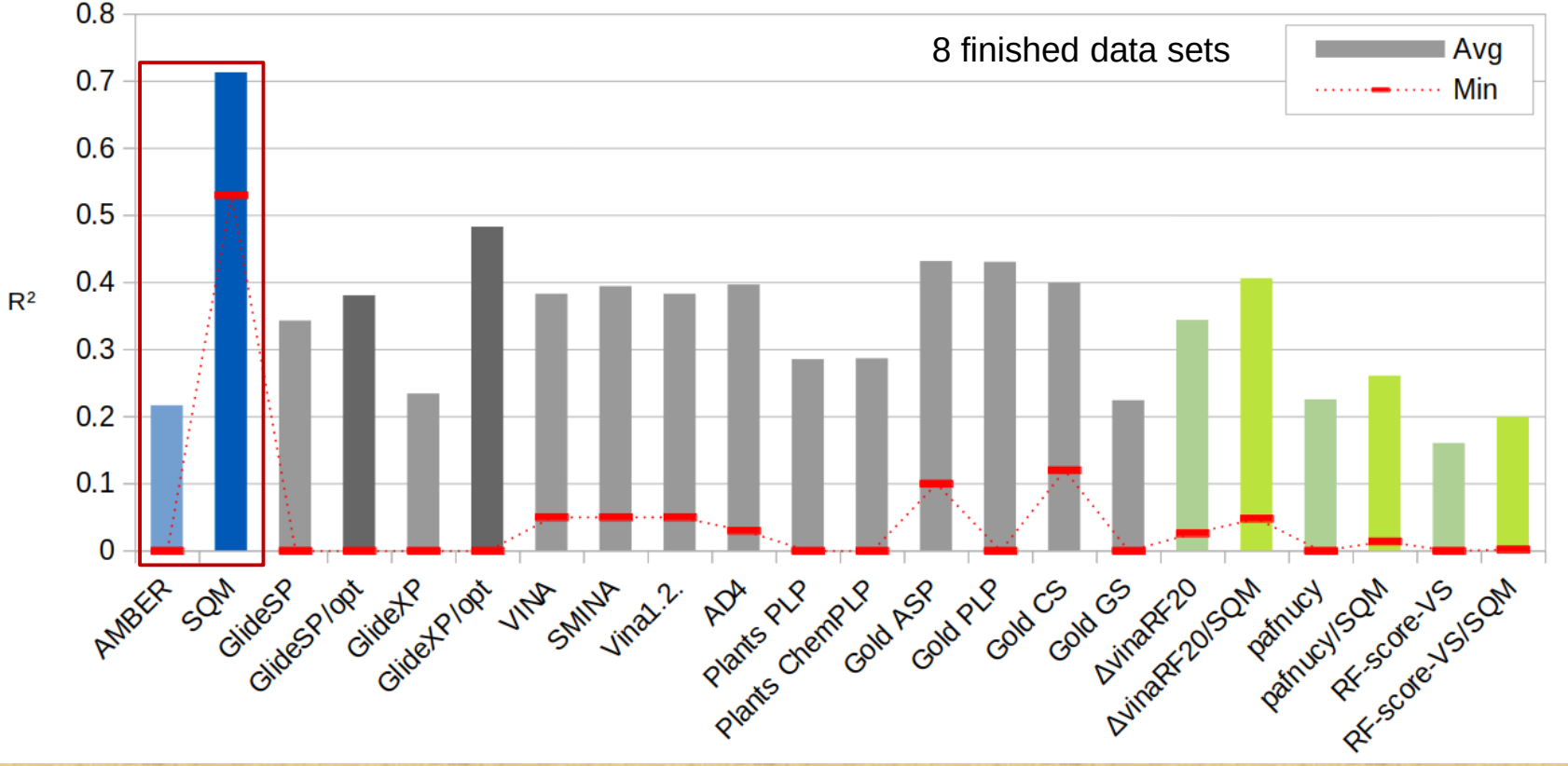
Scoring Function Performance



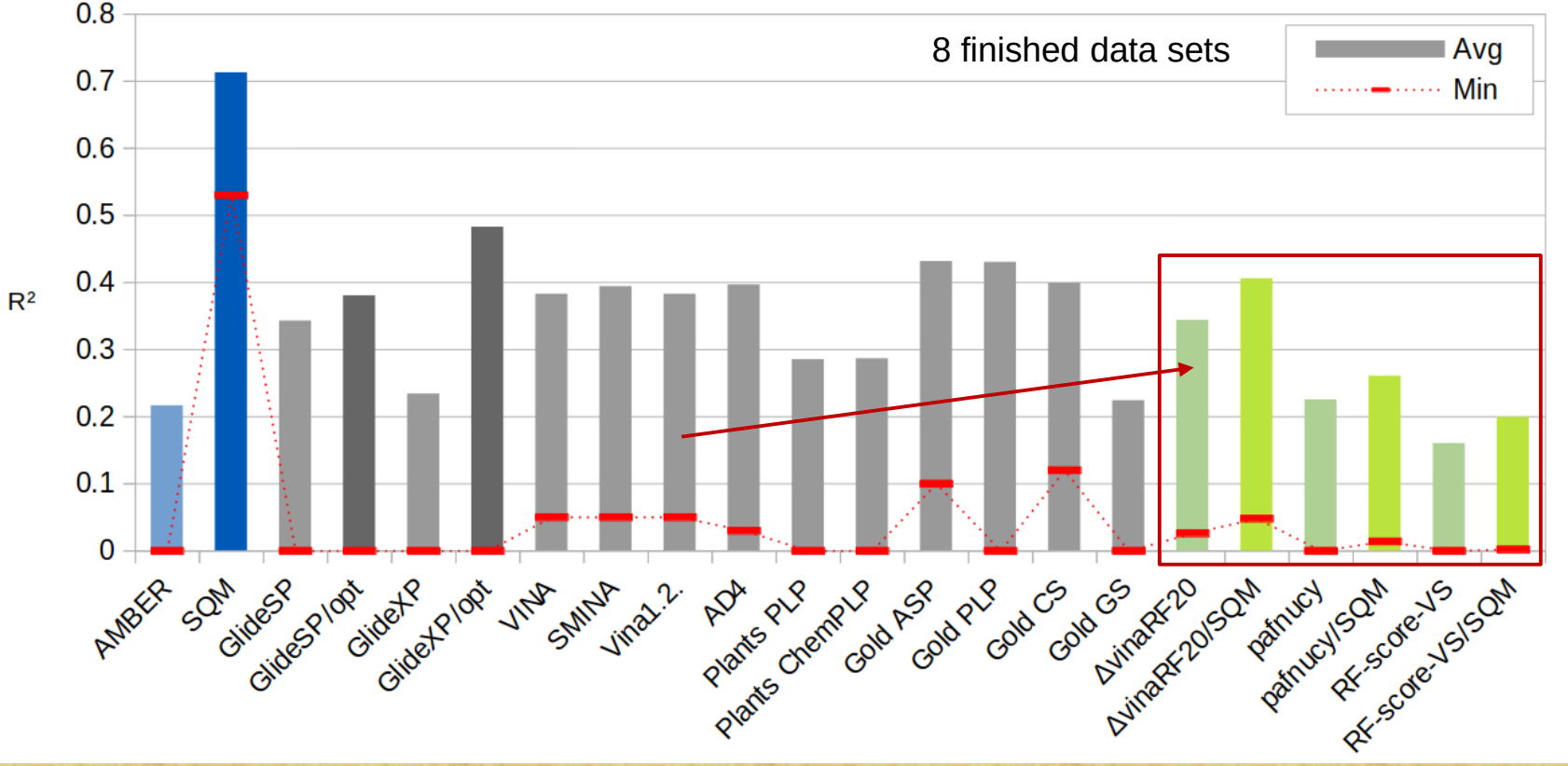
Scoring Function Performance



Scoring Function Performance

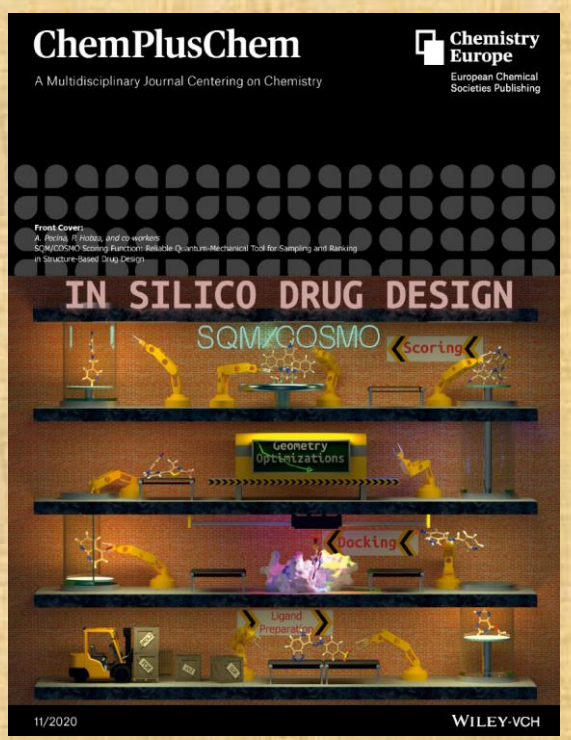


Scoring Function Performance



Summary

- SQM-score outperforms classical and ML scoring functions
- reasonable computational cost (20 min/1 CPU)
- prototype software licensed to US-based pharma company
- Heading towards marketable, stand-alone implementation



Acknowledgements

- IOCB internal grant
- *European Regional Development Fund; OP RDE; Project: "*ChemBioDrug*" (No. CZ.02.1.01/0.0/0.0/16_019/0000729).*

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research, Development and Education

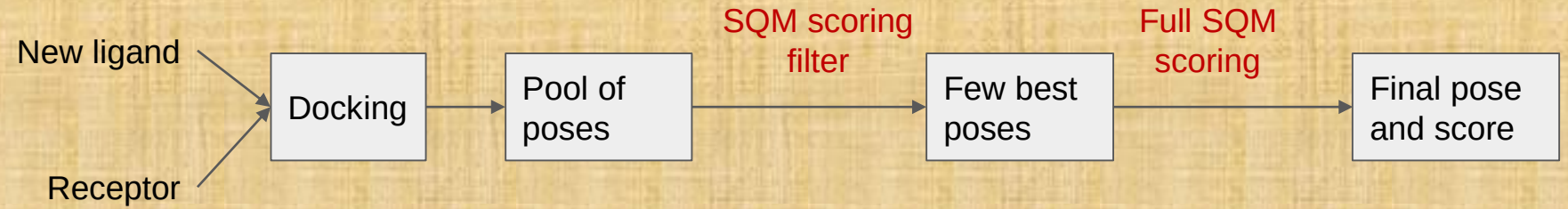
MINISTRY OF EDUCATION, YOUTH AND SPORTS

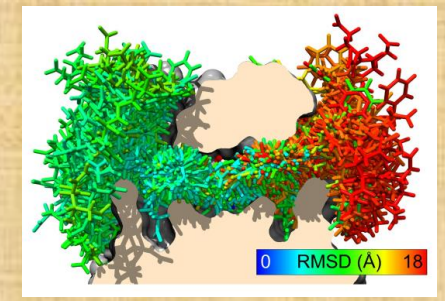
Reviews:
Lepšík et al.; *ChemPlusChem* **2013**, *78*, 921
Pecina et al.; *ChemPlusChem* **2020**, *85*, 2362

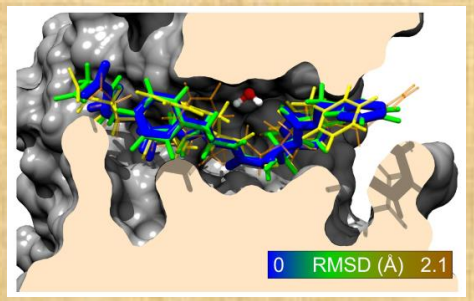
Towards realistic use case

Unknown structure of the complex - ability to calculate new molecules

- faster protocol for selecting best geometries (poses) from docking[1]



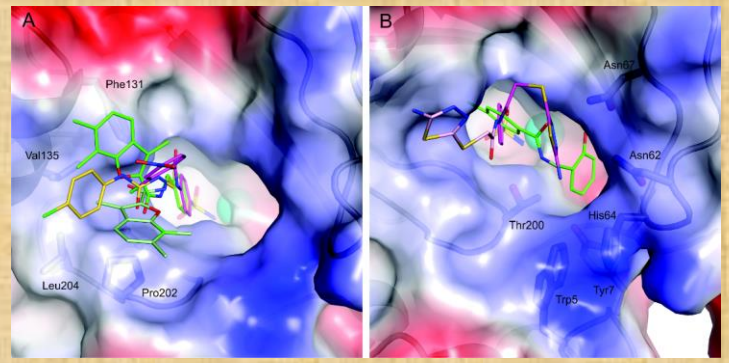


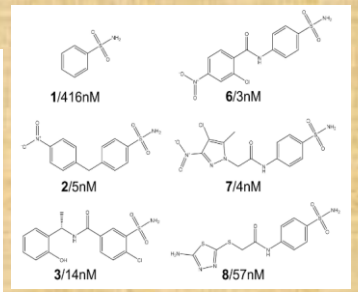


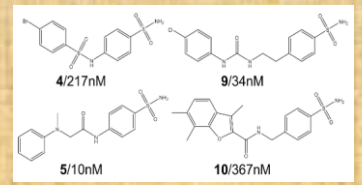
[1] A. Pecina, R. Meier, J. Fanfrlík, M. Lepšík, J. Řezáč, P. Hobza, C. Baldauf; Chem. Commun. 2016

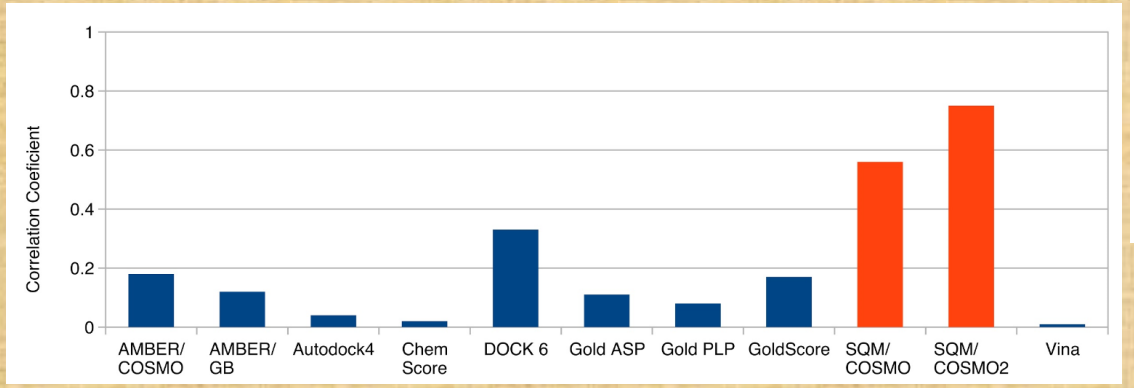
Case Study 1: Ranking - Carbonic anhydrase II

- Set of 10 inhibitors
  binding to carbonic anhydrase II through Zn^{2+}
- 10 high-resolution (1.1–1.4 Å) crystal structures
- Consistent inhibitory constant (K_i) values
  measured at IOCB
- Score vs. $\Delta G_{bind} = RT \ln K_i$





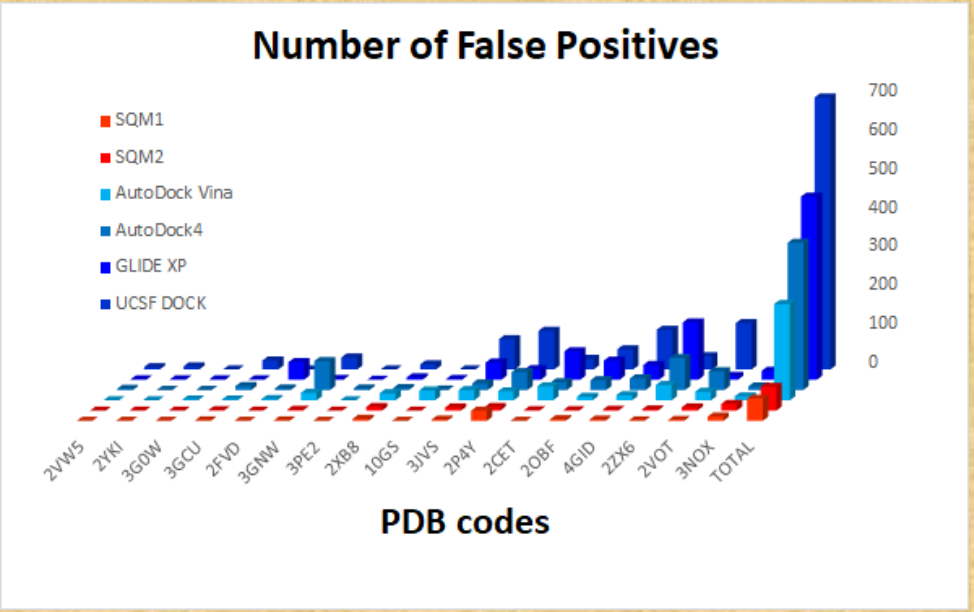




Pecina et al.; *ChemPhysChem* **2018**, 19, 873

Case Study 2: Sampling (Native Pose Identification)

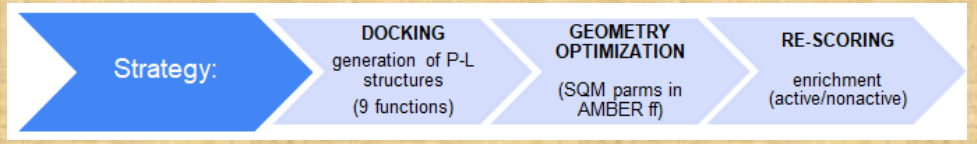
- diverse set of 21 protein-ligand systems (17 shown here)
- crystal structures available in PDB
- two SQM methods: SQM1=DFTB-D3H4X/COSMO
  SQM2=PM6-D3H4X/COSMO
- compared to 4 standard scoring functions
- evaluation:
  false positive = a pose with better score than crystal
  (ideal: zero false positives)
- SQM has 4-12-times less FPs than the standard SFs

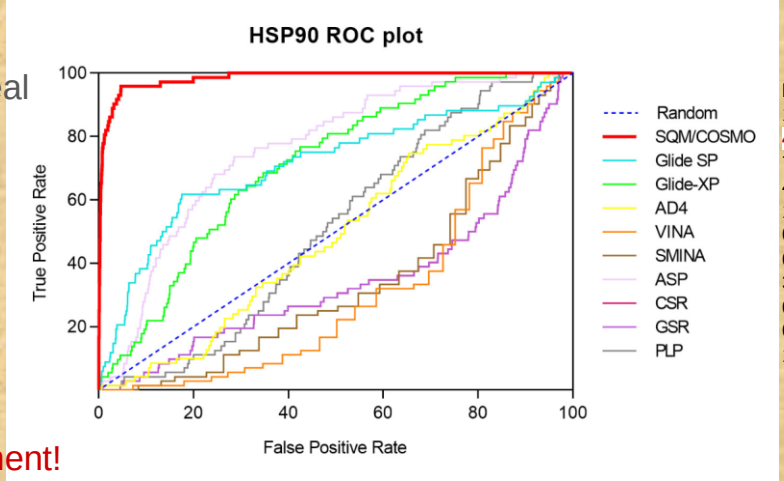


Pecina et al.; *Chem. Commun.* **2016**, 52, 3312
Pecina et al.; *J. Chem. Inf. Model.* **2017**, 57, 127
Ajani et al.; *ACS Omega* **2017**, 2, 4022

Case Study 3: Virtual screening (Library enrichment)

- Heat shock protein (HSP90); important for cancer and immunity
- 72 biologically active compounds + 4469 structurally similar compounds (DUD-E decoys)
- Enrichment factor (EF_1) and ROC curves (AUC%), where random is (1, 50%) and ideal (63, 100%)
- Standard docking provides good poses but standard SFs fail in their correct ranking
- Rescoring by SQM increases enrichment significantly
- Combination of SQM geometries and SQM/COSMO SF leads to the best enrichment!





EF_1	AUC%
1	50
47	98
7	75
4	71
1	49
0	30
0	34
3	76
0	34
0	60
1	51

Eyrilmez et al.; ChemPhysChem **2019**, 20, 2759

Polarisation in Classical Molecular Dynamics (MD) of Protein/Ion/Ligand Complexes

M. Lepšík, S. Kuhaudomlarp, P. Jungwirth, A. Imberty

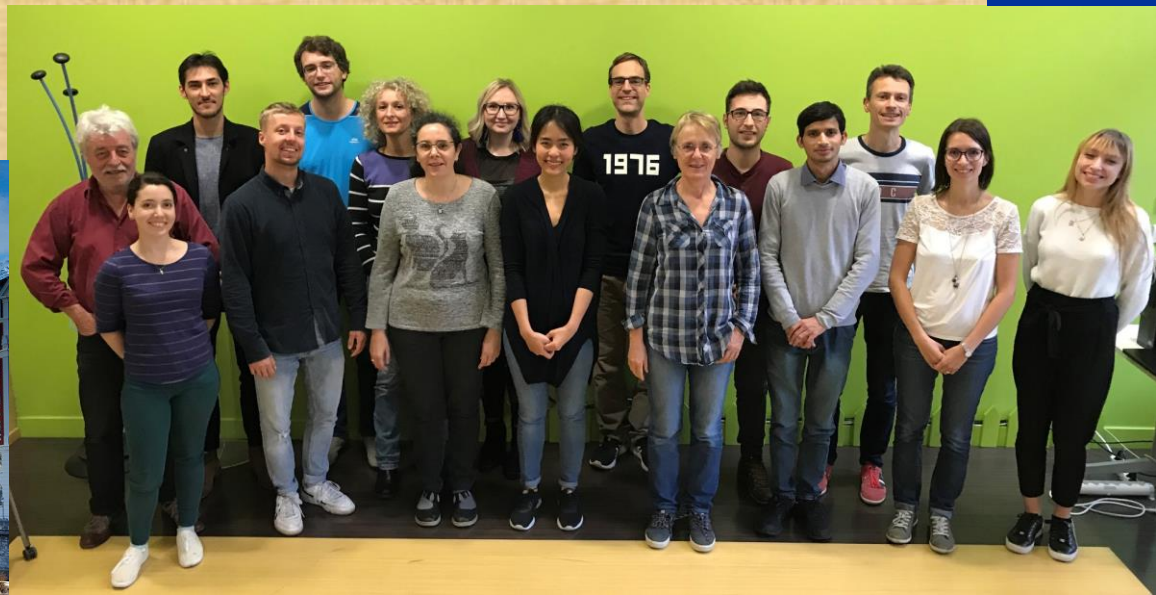
CERMAV, CNRS, Grenoble, France

Marie Sklodowska-Curie

Individual Fellowship

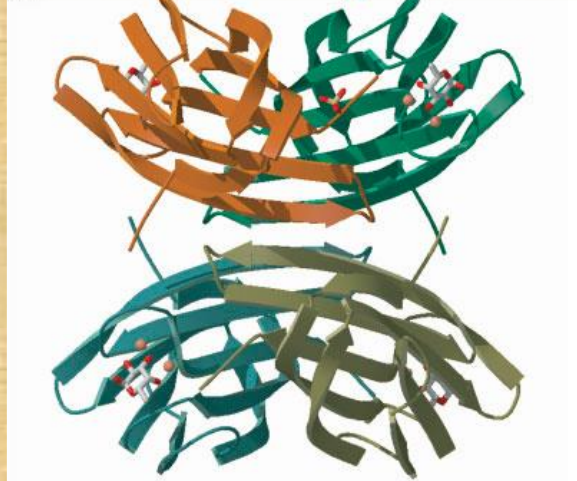
“CaLecLig”

2018-2020

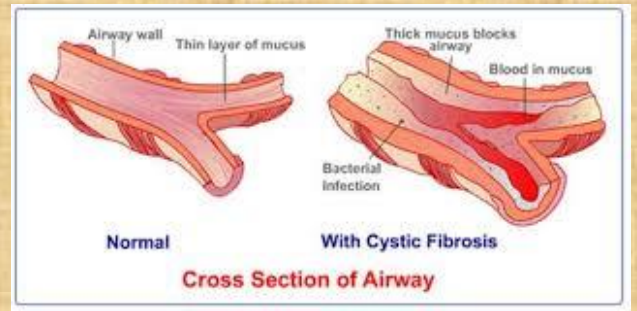


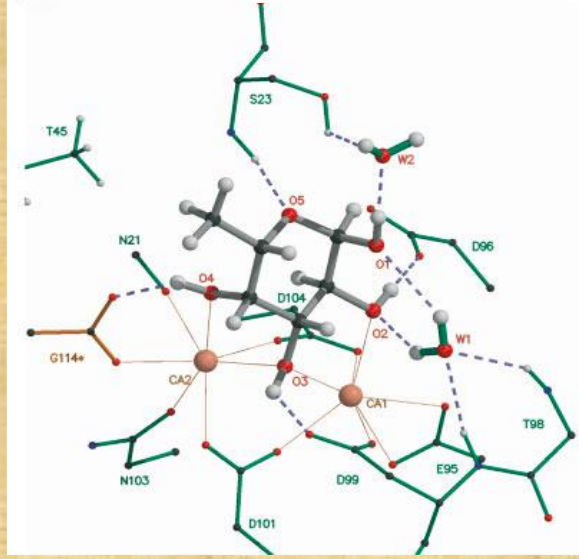
Bacterial Infection: Virulence Lectins

- Biofilm, virulence factors
- LecA, LecB lectins
- tetramers
- Ca^{2+} in binding site
- Bind human oligosaccharides



- *Pseudomonas aeruginosa*
- cystic fibrosis





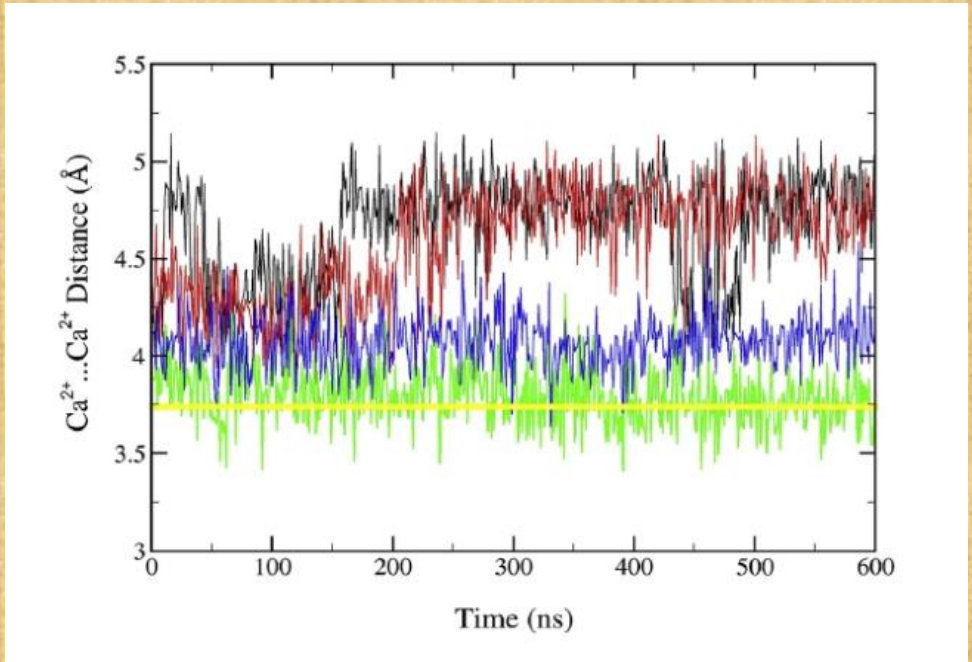
E. Mitchell, et al., *Nat. Struct. Biol.,* 2002, 9, 918

Electronic Continuum Correction for MD

- classical fixed-charge non-polarizable force fields miss electronic polarization, screening of charges

- charge scaling by inverse of square-root of water permittivity at high frequency

$$q_r = \frac{q}{\sqrt{\epsilon_r}}$$



(2+ charge Ca^{2+} - black, red; scaled Ca^{2+} parameters – blue, green), crystallographic value (yellow)

B.J. Kirby & P. Jungwirth, *J. Phys. Chem. Lett.*, **10**, (2019), 7531

M. Lepsik, et al., *Eur. J. Med. Chem.*, **177**, (2019), 212