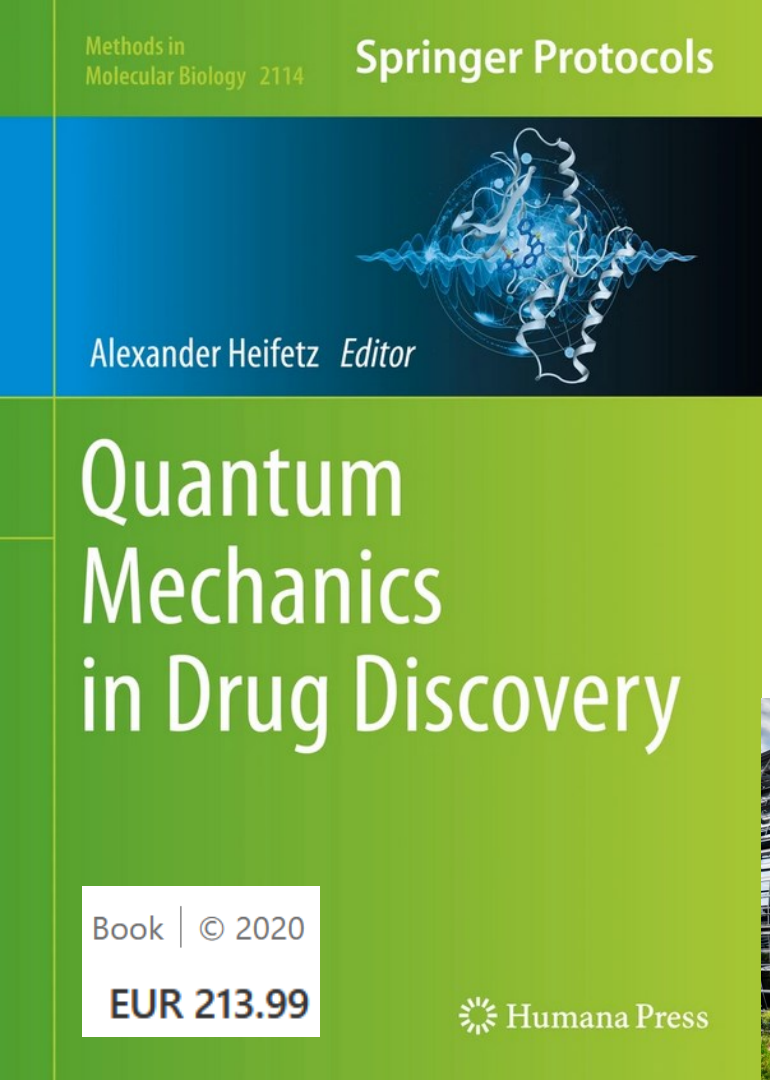


7ADD, Olomouc, Jan 2024

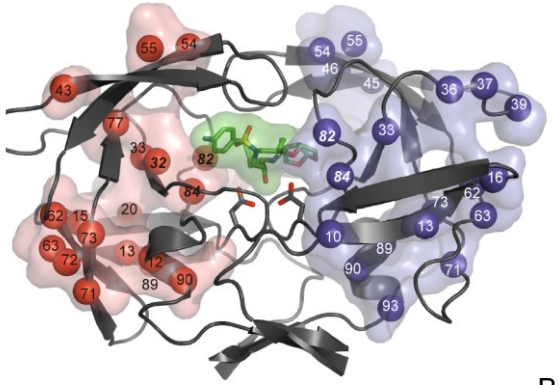
Martin Lepšík

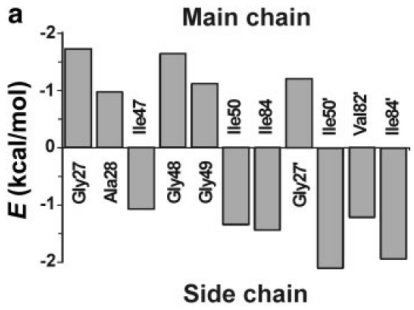
LEPSIK@UOCHB.CAS.CZ

ÚOCHB AV ČR
IOCB PRAGUE

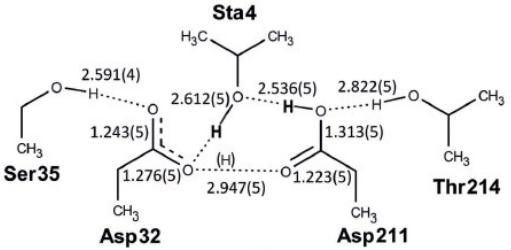
My Scientific Interests I

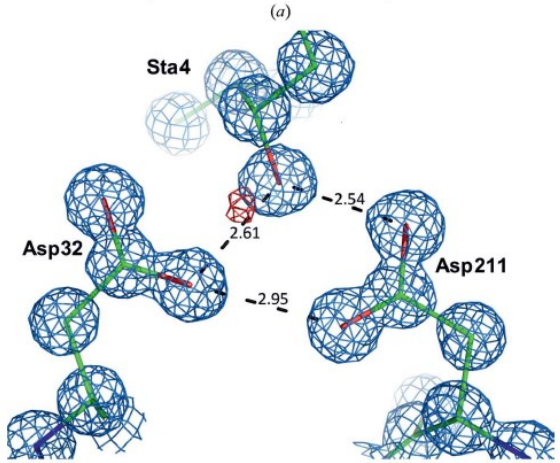
1) Protein-ligand binding at atomistic details



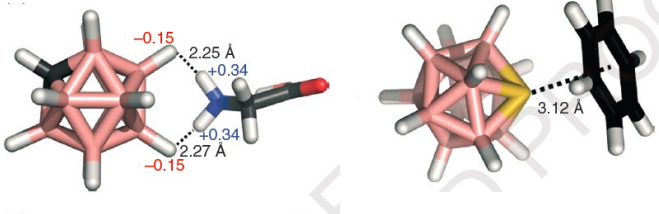


Brynda et al., J Med Chem, **2004,**8, 2030

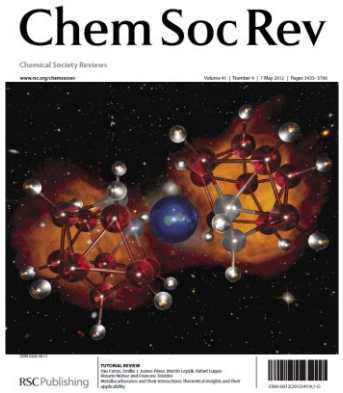




2) Non-classical non-covalent interactions



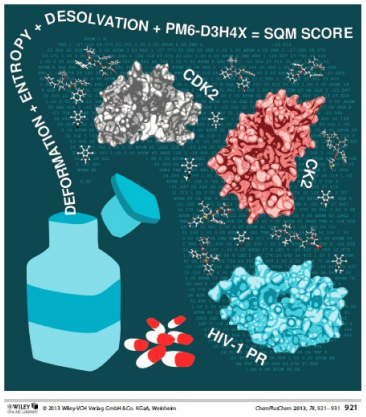
Fanfrlik, J. et al. In: Boron-Based Compounds: Potential and Emerging Applications in Medicine, Wiley, 2018.



Dostal et al., Acta Cryst D, **2015**, 71, 2494

My Scientific Interests II

3) Semiempirical Quantum Mechanical (SQM) Scoring

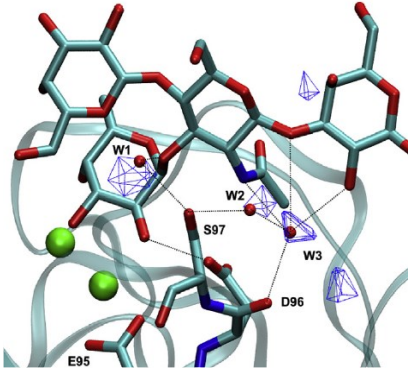


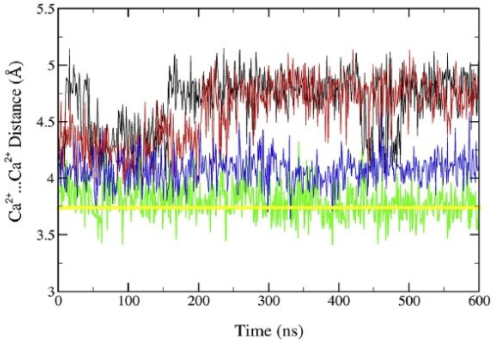
Reviews:

Lepšík et al.; *ChemPlusChem* **2013**, *78*, 921

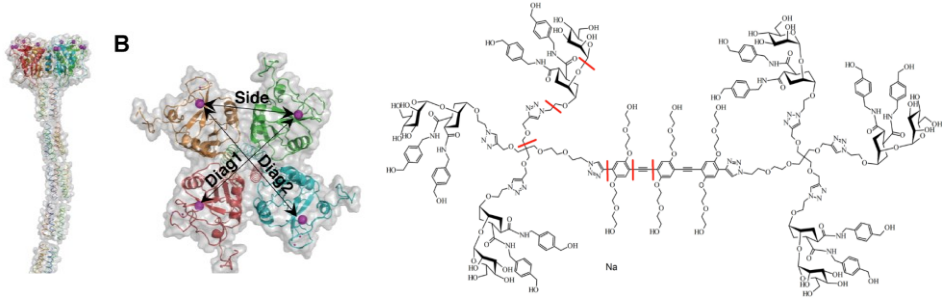
Pecina et al.; *ChemPlusChem* **2020**, *85*, 2362

4) Electronic Continuum Correction in Classical Molecular Dynamics





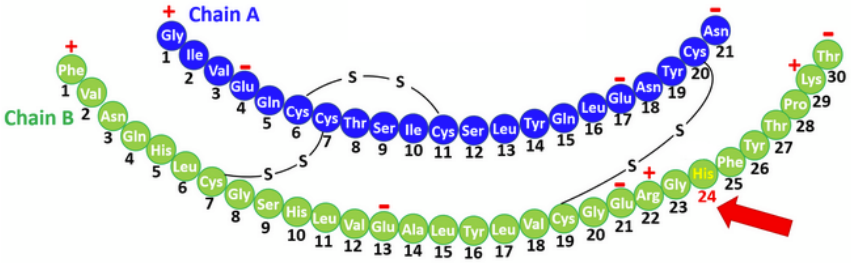
Lepšík et al.; *Eur J Med Chem* **2019**, *177*, 212.



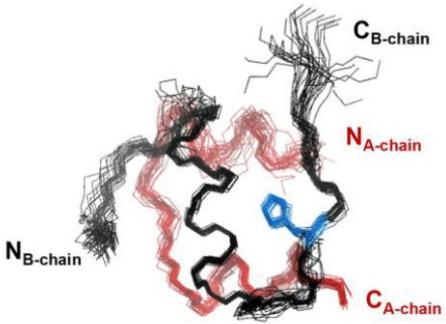
Porkolab, Lepšík et al.; *ACS Cent Sci* **2023**, *9, 4, 709*.

My Scientific Interests III

5) Insulin analogue/Insulin receptor binding



[L-HisB24]-insulin analogue sequence



Posters

Yevgen Yurenko et al

Quantification of Non-covalent Interactions at Protein-Protein Interface

Jiri Zak and Martin Lepsik

Molecular Dynamics of Insulin analogue/Insulin receptor Complexes

Outline

1. Use of QM in Drug Design
2. Advantages and Limitations of QM
3. SQM method development
4. SQM-based Scoring Function

Where in Drug Design can QM help?

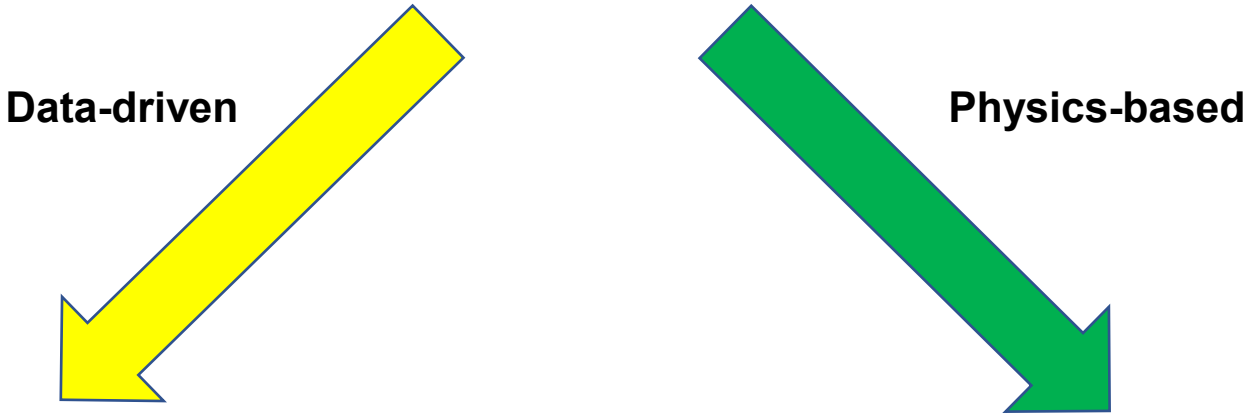
Structure-based DD

- X-ray crystallographic refinement
- Hit Identification (Virtual Screening, Docking, Scoring)
- Hit-to-Lead
- Docking
- Scoring

Ligand-based DD

- Partial charges
- Bioactive conformations
- pKa predictions

Structure-based Affinity Prediction



Standard Scoring Functions (SFs)
- ultrafast (seconds per compound)
- lack both accuracy and reliability

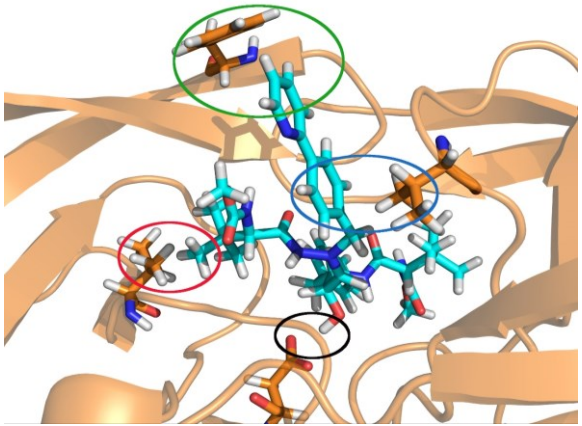
Machine-Learning (M-L)
- ultrafast (seconds per compound)
- ? training data/accuracy
- ? applicability domain

Free Energy Methods (FEP)
- variable accuracy, force-field dependent
- relative vs. absolute; slow on GPU (days)

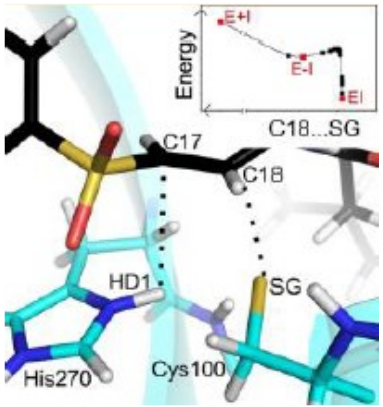
Quantum Mechanics (DFT)
- accurate but slow on 10s CPU (days)
- not applicable to large biomolecules (proteins)

Why Quantum Mechanics?

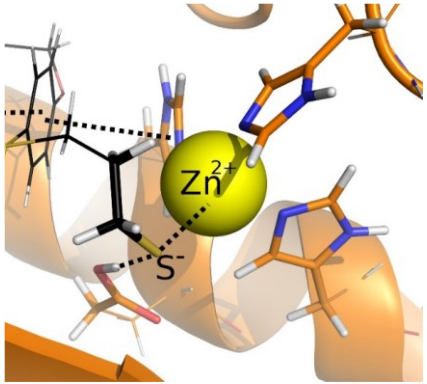
- quantitative: all types of non-covalent interactions
- dispersion, H-bonding, halogen bonding, etc.
- quantitative description
- metal interactions
- polarization, charge transfer
- covalent binding
- no parametrization of ligands



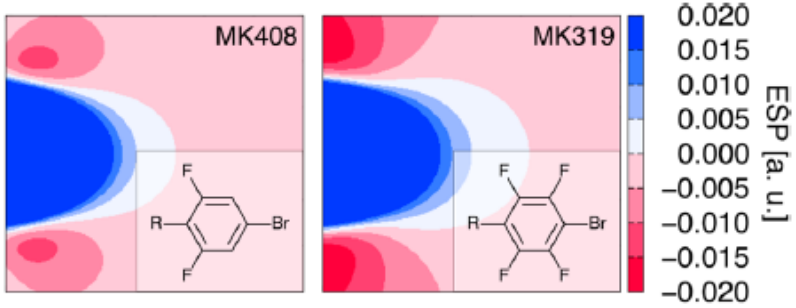
J. Phys. Chem. B, 2010, 114, 12666



J. Phys. Chem. B 2013,117, 14973

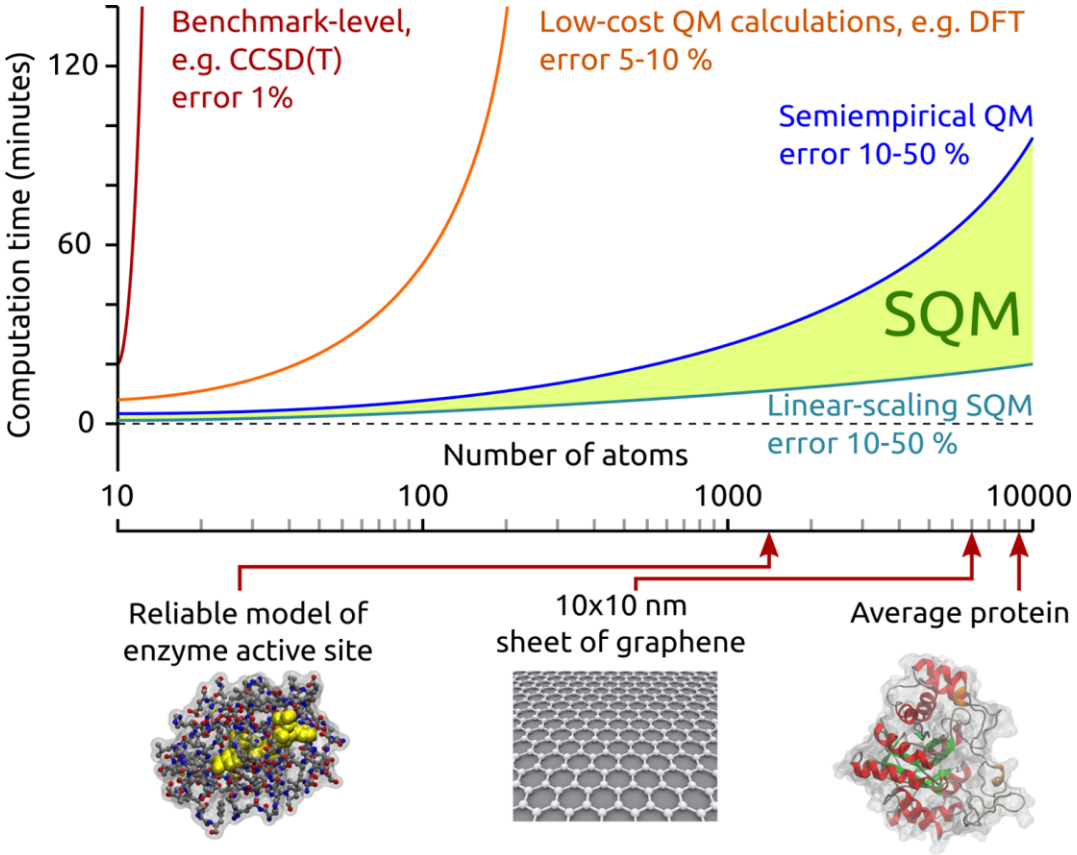


J. Chem. Inf. Model. 2017, 57, 127



ACS Chem. Biol. 2013, 8, 2484

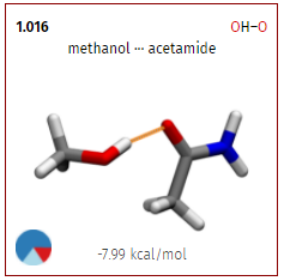
Which QM method?



- Fast
- linear-scaling with system size
- general (periodic table)

QM Methods for Non-covalent Interactions

- Small models, accurate calculations (CCSD(T)/CBS) in vacuum
- S66 benchmark dataset: H-bonding, dispersion
- www.nciatlas.org (~ 20,000 data points)
- hydrogen bonding, dispersion, sigma-hole interactions, repulsion

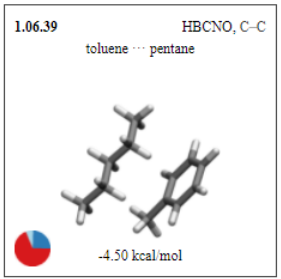


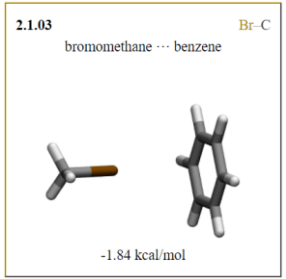
Coverage of the periodic table:

H																	He
Li	Be											B	C	N	O	F	Ne
Na	Mg											Al	Si	P	S	Cl	Ar
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba	Lu	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra	Lr	Rf	Db	Sg	Bh	Hs	Mt	Ds	Rg	Cn	Nh	Fl	Mc	Lv	Ts	Og

No. of complexes containing the element

1000+
500+
100+
50+
10+

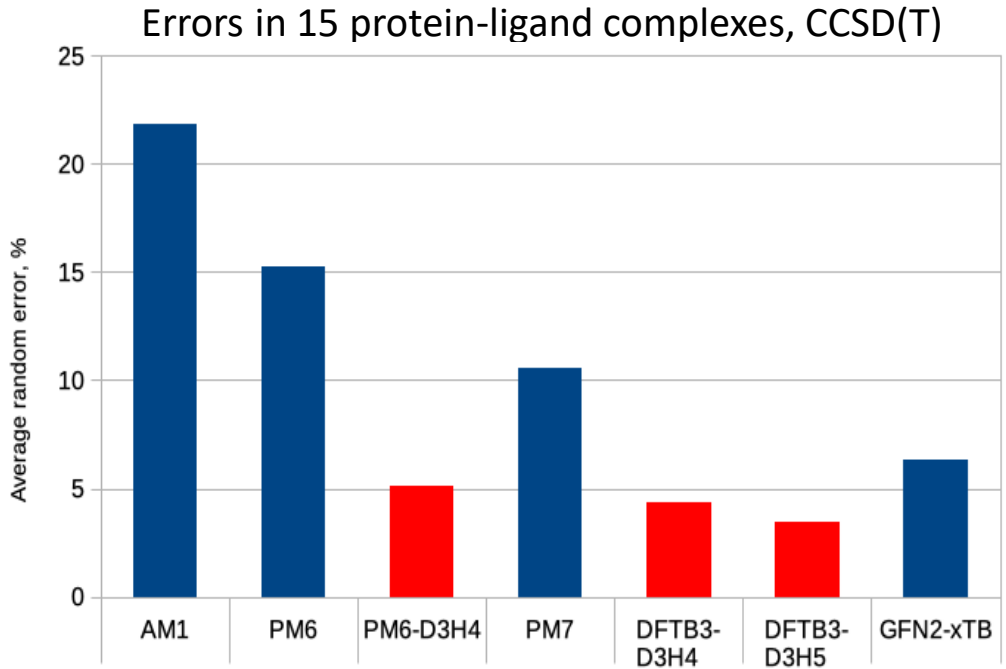




- Development of semiempirical QM methods - corrections for non-covalent interactions
- chemical accuracy (1 kcal/mol) in small dimers

Řezáč, J., Hobza P. Chem. Rev. 2016, 116, 9, 5038

Corrected Semiempirical QM



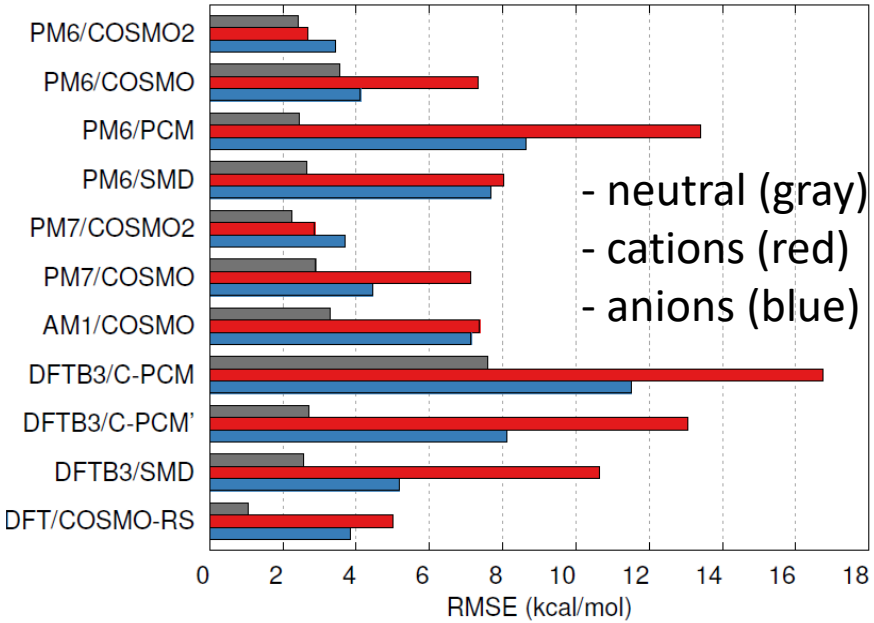
- Fast calculation
- Easy preparation
  (no system-specific parameters)
- Accuracy?

PM6-D3H4X

[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

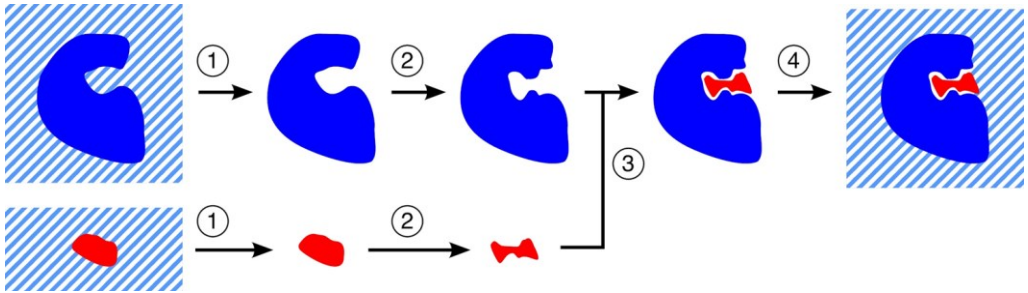
COSMO2 Implicit Solvation Model

- reparametrisation of COSMO
- adding non-polar solvation term



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

SQM2.20 Scoring Function



Modular physics-based approach:

- MM/GBSA-like
- components can be replaced if better alternatives exist

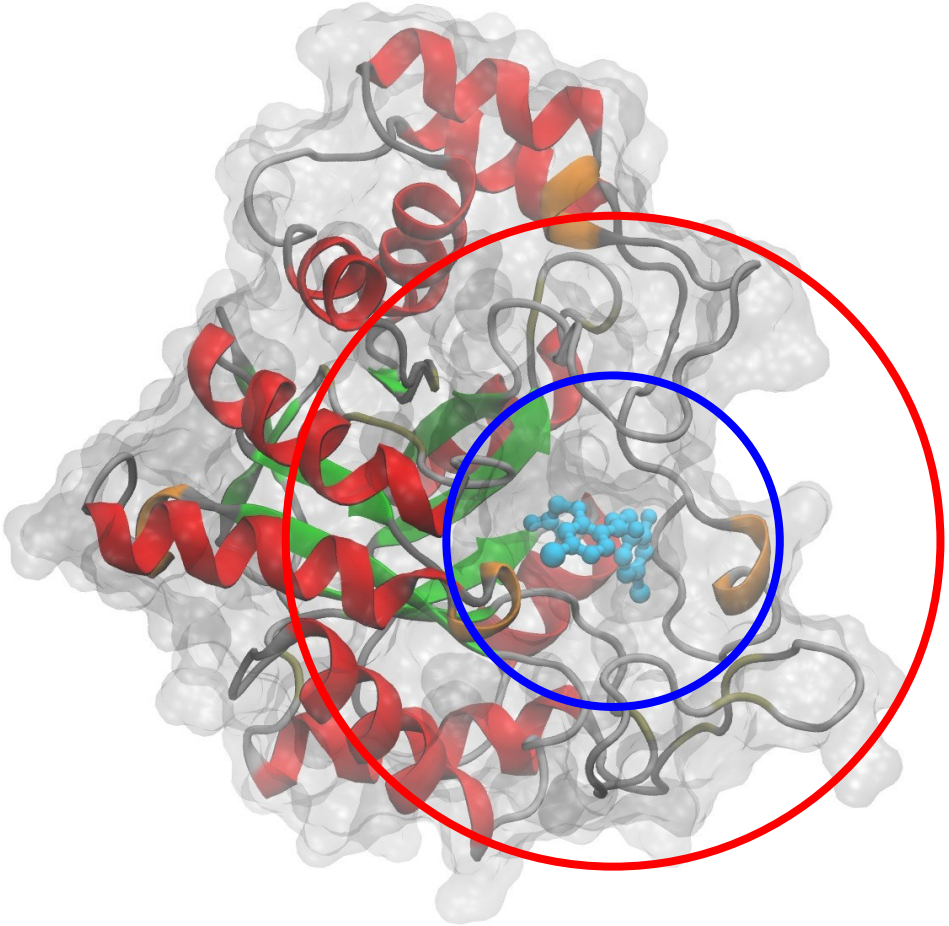
SQM2.20 =

ΔE_{int}	**PM6-D3H4X** + further corrections
$+ \Delta\Delta G_{solv}$	PM6/**COSMO2**
$+ \Delta G_{conf,w}(L)$	PM6-D3H4X/COSMO2 optimization / optional conf search
$+ \Delta G_{H+}$	PM6-D3H4X/COSMO2 difference
$- T\Delta S$	LM5 model fitted to QM data

Fanfrlík et al.; *J. Phys. Chem. B* **2010**, *114*, 12666

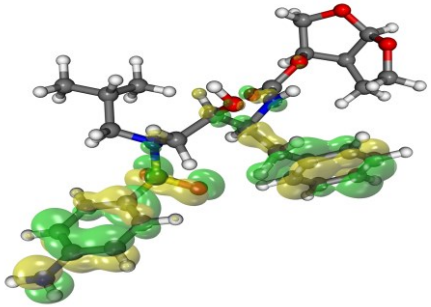
QM/MM Setup

- Internal moving QM part
- Intermediate QM static part
- Outside fixed



Quest for Universal Reliable Scoring Function

Quantum Mechanical Scoring in Structure-based Drug Design

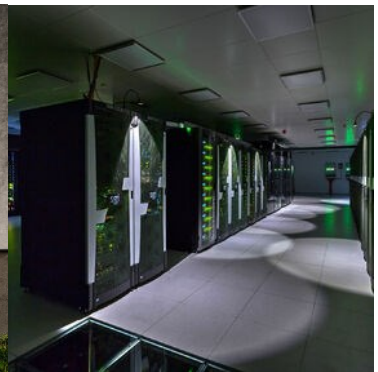


M. Lepšík, J. Fanfrlík,
A. Pecina, J. Řezáč

ÚOCHB AV ČR
IOCB PRAGUE

IOCB Tech

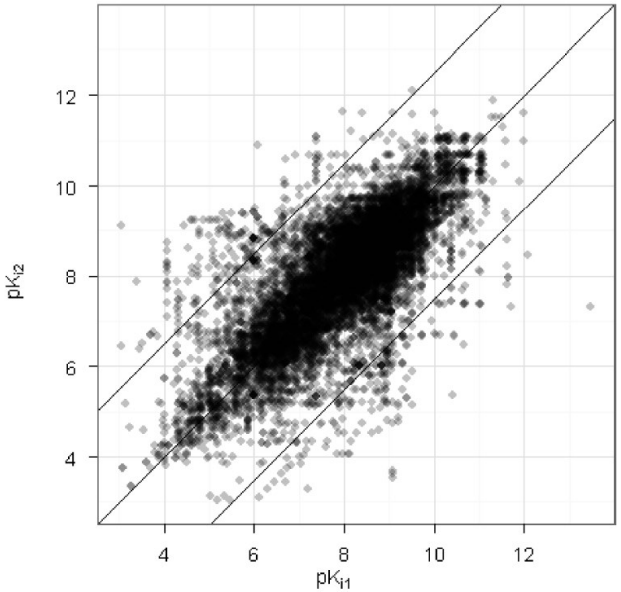
P. Hobza and past members



Is the Scoring Function Universal and Reliable?

Comparison to the “experimental “truth” in multiple diverse data sets

- Input: Experimental structures or a high-quality model
- Comparison with **RELIABLE** experimental affinities
- Reproducibility from multiple independent measurements: $R^2 = 0.8$)



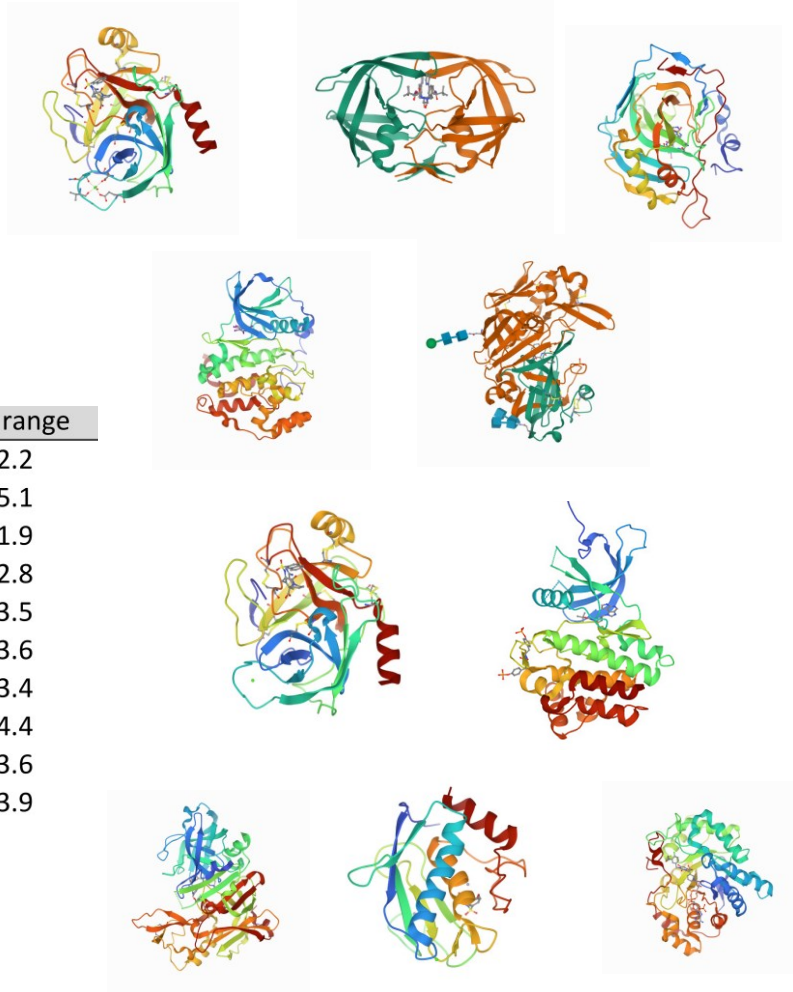
Kramer et al. *J. Med. Chem.* **2012**, *55*, 5165–5173.

PL-REX dataset

Protein-**L**igand / **R**eliable **Ex**periment data set,
understanding of the system and meticulous preparation

- reliable structures, preferably crystal
- measurements from one lab (K_i, IC_{50})
- careful preparation of each protein

Target	Ligands	Crystals	Similarity	Experiment	pKi range
Carbonic anhydrase	10	10	0.32	Ki	2.2
HIV Protease	22	12	0.51	Ki	5.1
Casein kinase 2	16	16	0.35	Ki	1.9
Aldose reductase	14	14	0.47	Ki	2.8
Cathepsin D	10	3	0.71	IC50	3.5
Beta-secretase 1	16	16	0.48	IC50	3.6
Janus kinase 1	12	12	0.55	Ki	3.4
Trypsin	15	15	0.46	Ki	4.4
CDK2	31	31	0.69	IC50	3.6
Matrix metallopeptidase 12	18	18	0.47	Ki	3.9

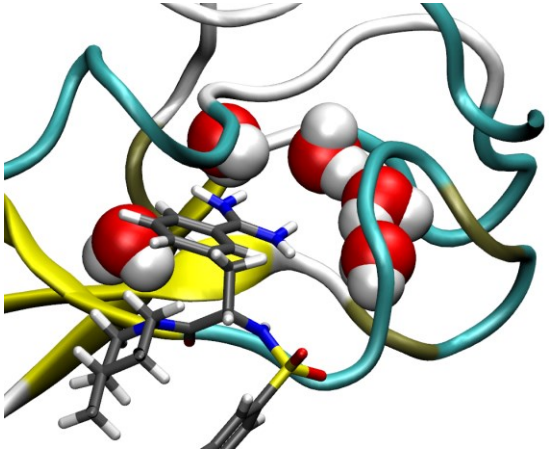


164 data points

Dataset available: https://github.com/Honza-R/PL-REX

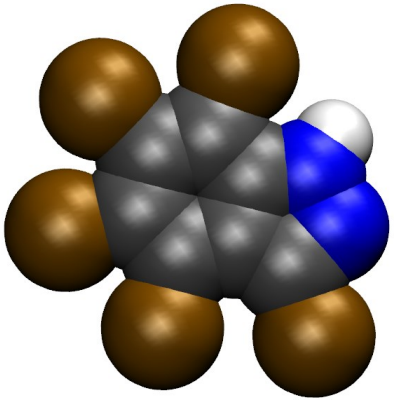
Preprint: https://dx.doi.org/10.26434/chemrxiv-2023-zh03k

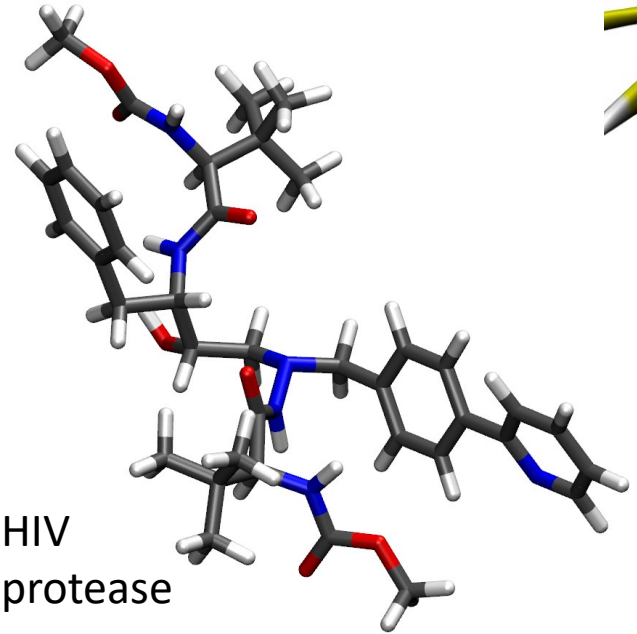
PL-REX dataset



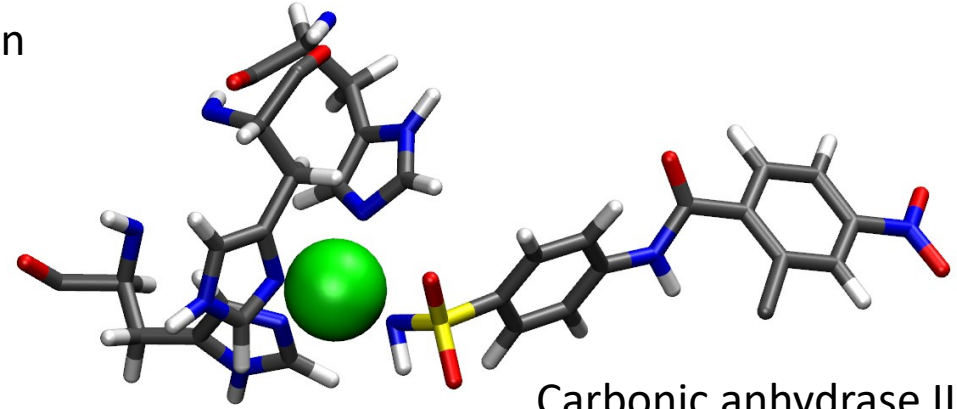
Trypsin

Casein kinase 2





HIV protease



Carbonic anhydrase II

Systems Rejected from PL-REX

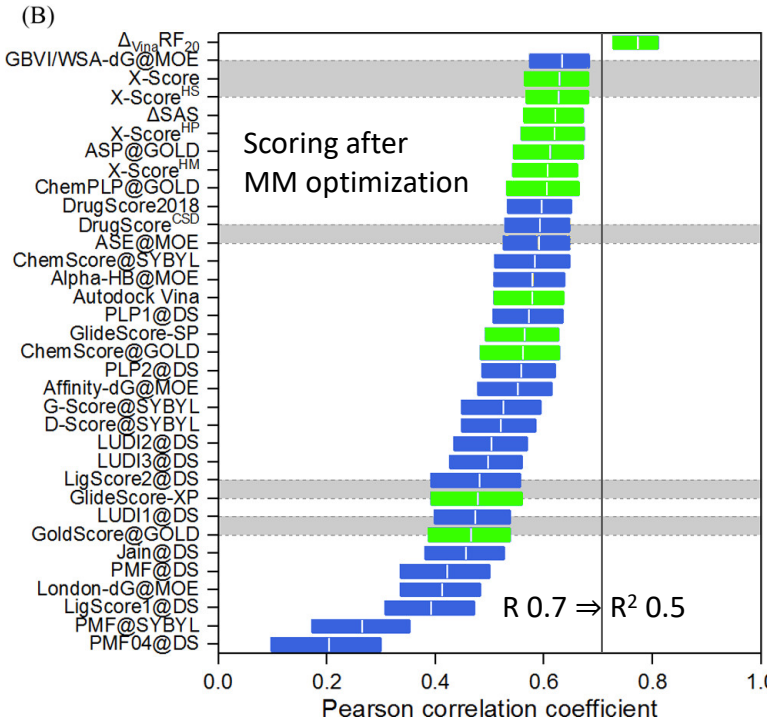
	Source	Ligands	Crystals	Resolution	pKi span	Notes
Cathepsin S	D3R challenge	19	24	1.7 - 3.0	1.3	Wrong ligand conformations and maps in X-rays
Beta-Secretase 1 BACE-1	D3R challenge	13	13	1.7 - 2.3	0.8	
Bromodomain of BRD4	10.1021/acs.jcim.1c01229	14	1	1.6	3.6	single X-ray, docked ligands (worked for FEP)
M. tuberculosis Malate synthase	10.1021/acs.jcim.8b00417	20	20	1.4 - 2.6	3.4	Ligands replace some waters
Receptor tyrosine kinase EPHA2	10.1002/cmdc.201700217	14	18	1.2 - 1.9	3.7	Ligand modifications are in solvent area

Comparison with Scoring Functions

- Best SFs in the CASF2016 set[1]
- Few more used previously in the group
- Structure-based machine learning

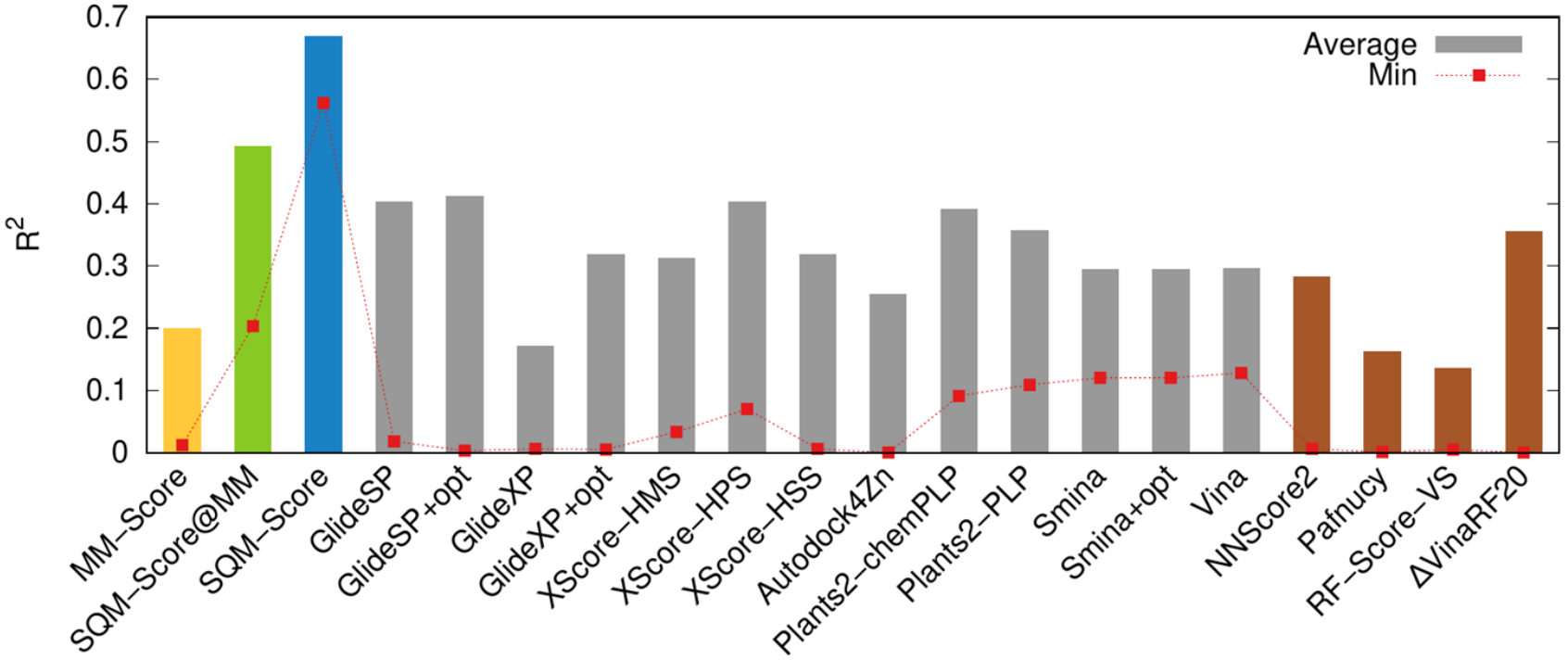
Timing:

- Empirical SFs <= seconds
- SQM-score ~ 20 minutes



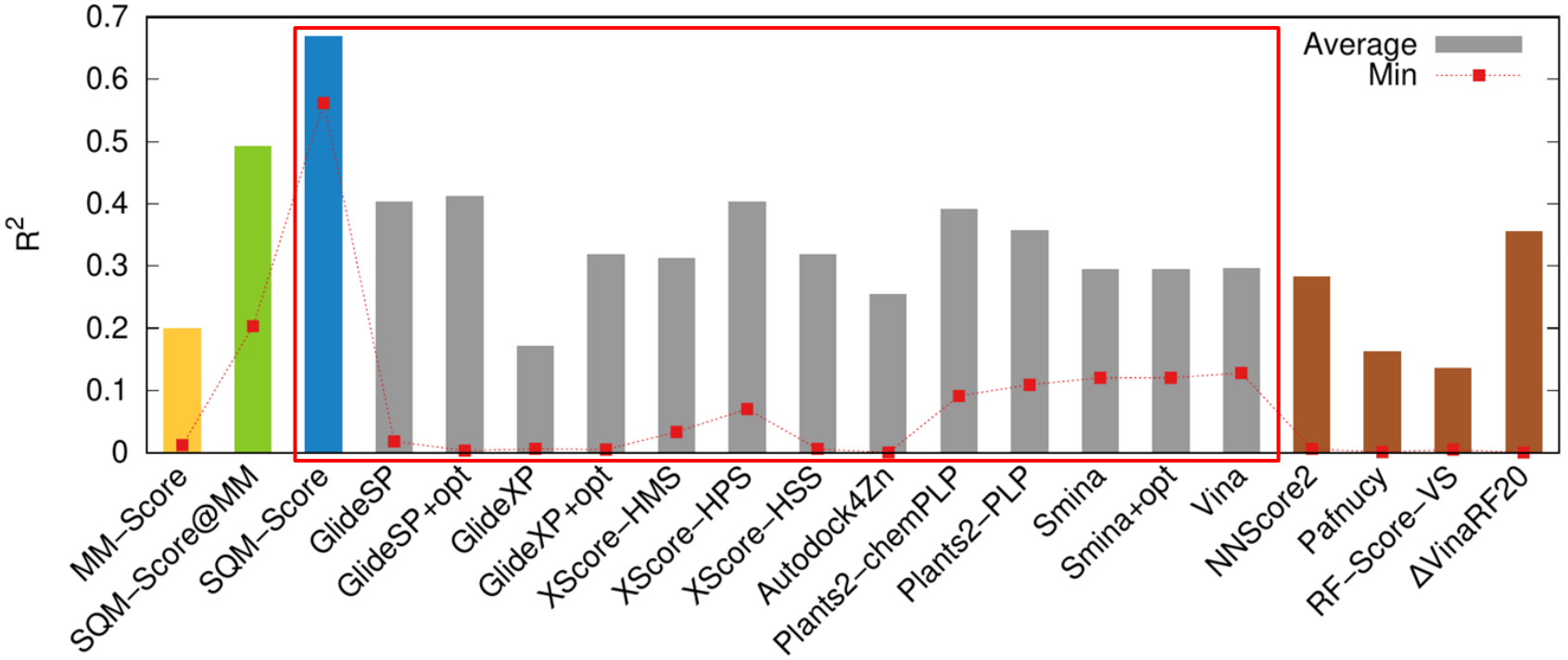
Su, M. et al., *J. Chem. Inf. Model.*, **2019,** 59, 895.

Comparison with Scoring Functions



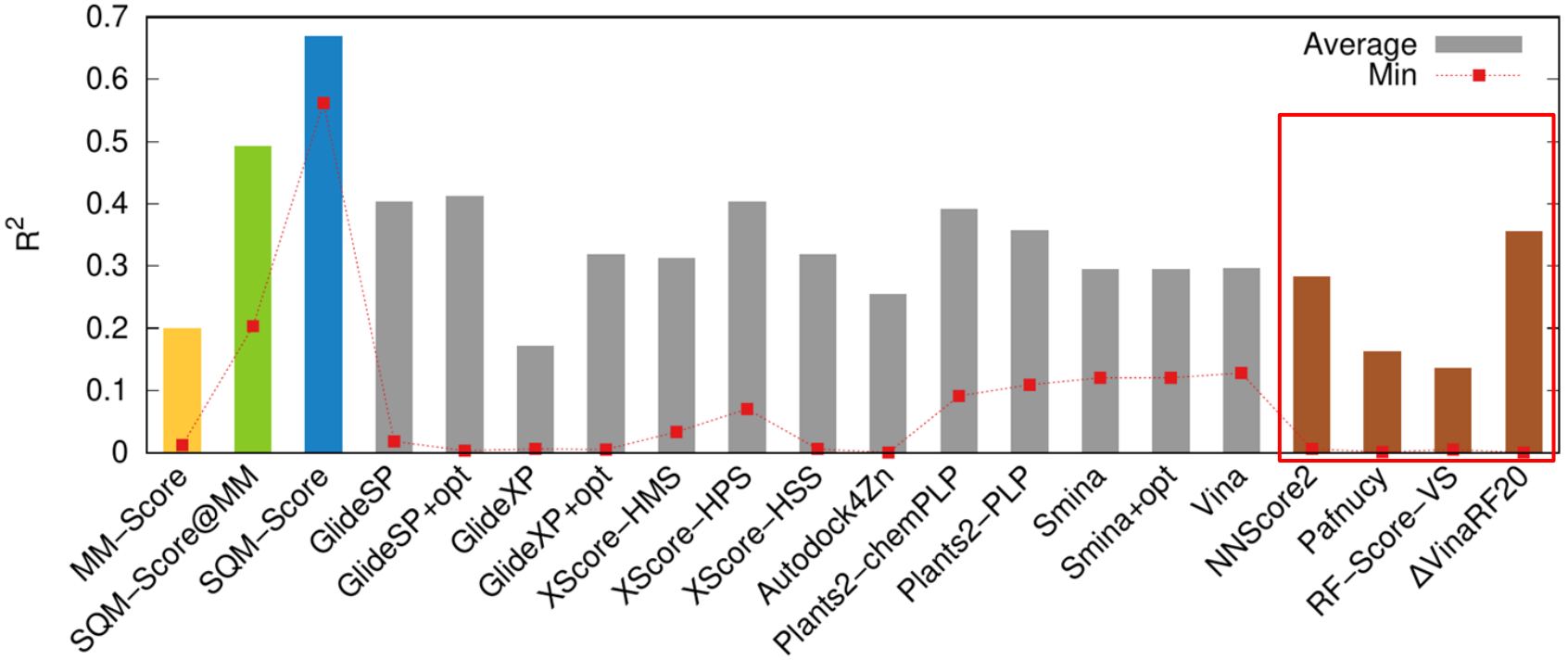
Correlation with experiment, averaged over 10 targets

Comparison with Scoring Functions



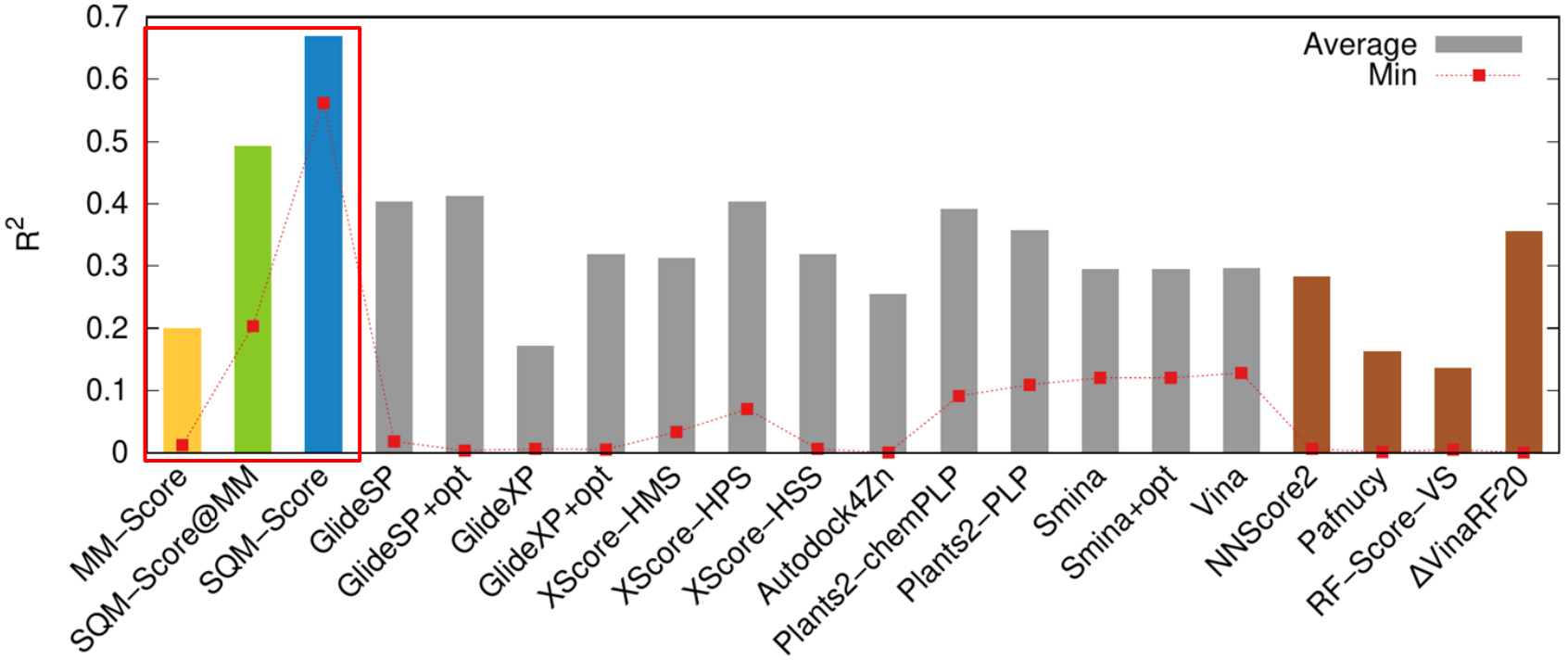
Correlation with experiment, averaged over 10 targets

Comparison with Scoring Functions



Correlation with experiment, averaged over 10 targets

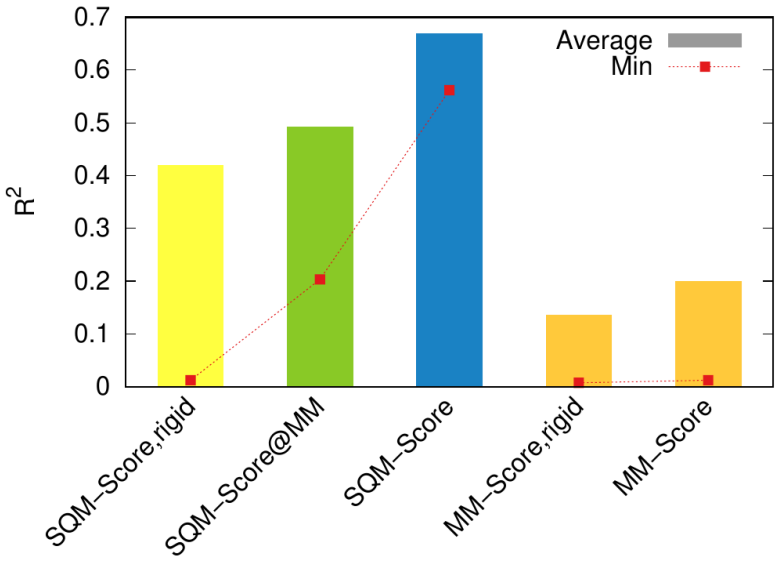
Comparison with Scoring Functions



Correlation with experiment, averaged over 10 targets

P-L complex geometry

- determines the quality of scoring
- The same SQM score computed on increasingly more refined geometries



Comparison with MM and DFT

	Default Model (~2,000 atoms)				Trimmed Model (~1,000 atoms)	
Dataset	**SQM2.20**	SQM2.20 //AMBER	AMBER		SQM2.20	DFT score
01-CA2	0.67	0.36	0.28		0.63	0.85
02-HIV-PR	0.75	0.70	0.33		0.71	0.61
03-CK2	0.81	0.70	0.40		0.79	0.53
04-AR	0.70	0.56	0.01		0.60	N.D.
05-Cath-D	0.66	0.22	0.23		0.70	0.66
06-BACE1	0.63	0.57	0.37		0.37	0.25
07-JAK1	0.56	0.57	0.03		0.59	0.49
08-Trypsin	0.75	0.73	0.54		0.61	0.79
09-CDK2	0.61	0.20	0.07		0.56	0.50
10-MMP12	0.74	0.62	0.03		0.81	0.69
Average	**0.69**	**0.52**	**0.23**		**0.62 (0.67*)**	**0.64***

- SQM: universal performance across targets
- AMBER geometries deteriorate SQM2.20 scoring in some targets
- AMBER scoring: low performance
- SQM2.20 comparable to DFT (ΔE_{int} replaced by ωB97X-D3BJ/DZVP) **BUT**
- SQM2.20 is fast (20 min/system on 1CPU) vs. DFT with ~10^3 CPU-hours / system)

- **DFT brings no statistically significant improvement**
- SQM with corrections very good
- Gas-phase DFT susceptible to errors

Preprint:

https://dx.doi.org/10.26434/chemrxiv-2023-zh03k

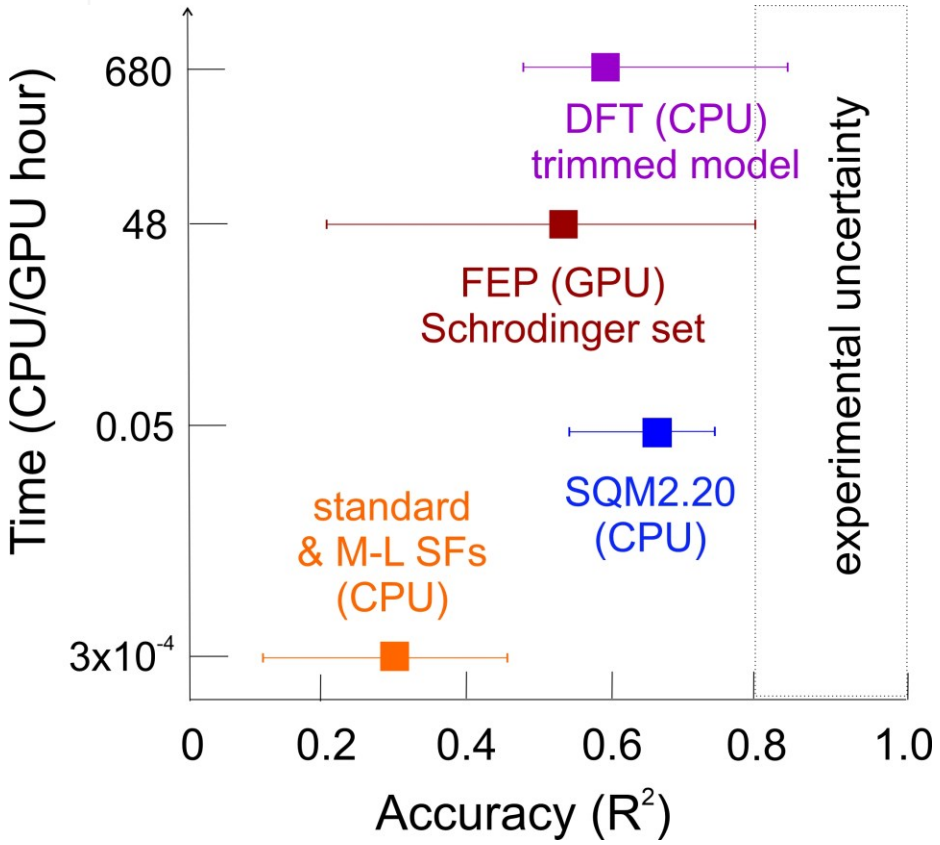
Affinity Prediction: Timing

End-point Methods

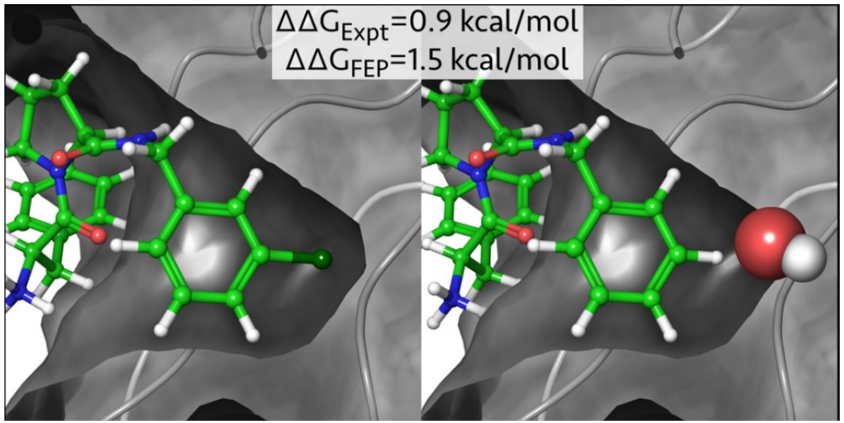
- scoring (seconds, 1CPU)
- SQM2.20 (minutes, 1CPU)
- DFT (hours/days, multi CPU/GPU)

Ensemble Methods

- FEP (hours/days, multi CPU/GPU)

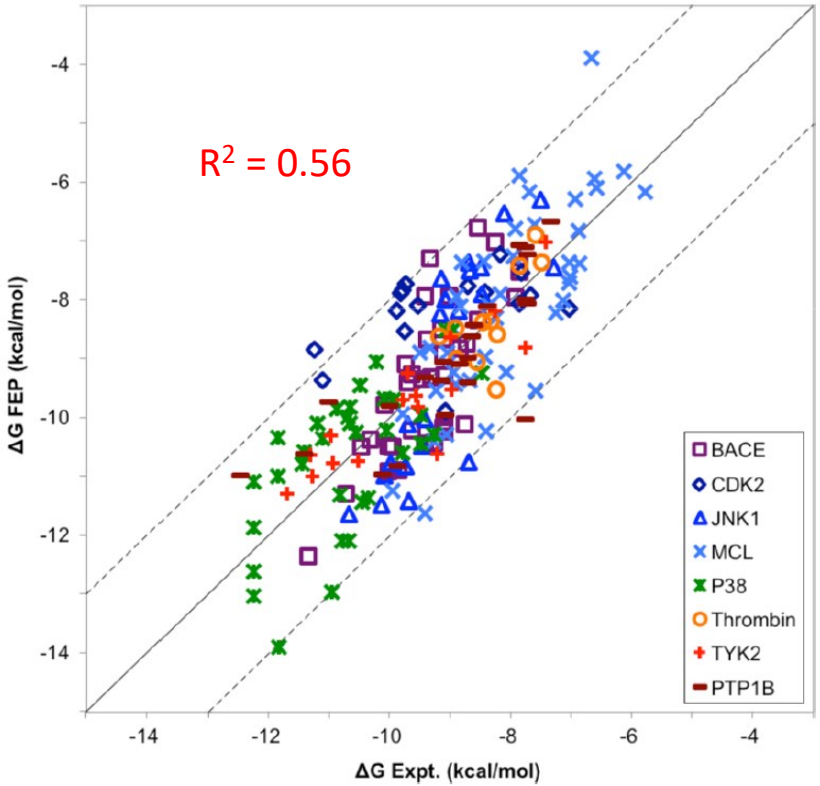


Schrodinger FEP+ Dataset



Schrodinger FEP+

- 8 targets, 10-40 ligands each, similar
- Automatic preparation
- Free-Energy Perturbation
- OPLS 2.1 force field
- REST enhanced sampling
- GPU



Wang L et al., *J. Am. Chem. Soc*. **2015**, *137*, 2695−2703

Comparison with FEP+ on PL-REX

Work in progress

Target	num. of ligands	avg. Tanimoto	charge	**SQM2.20**	**FEP+**
01-CA2	10	0.32	-1	0.67	0.55
02-HIV-PR	22	0.51	0, 1	0.75	0.04
03-CK2	16	0.35	-1	0.81	0.54
04-AR	14	0.47	-1	0.70	0.00
05-Cath-D	10	0.71	0	0.66	0.75
06-BACE1	16	0.48	0, 1	0.63	N.D.
07-JAK1	12	0.55	0, 1	0.56	0.34
08-Trypsin	15	0.46	0, 1, 2	0.75	0.46
09-CDK2	31	0.69	-1	0.61	0.56
10-MMP12	18	0.47	0	0.74	0.42
AVERAGE	**17**	**0.52**		**0.69**	**0.40**

- PL-REX challenging for FEP+
- different ligand charges
- dissimilar ligands

Comparison with FEP+ on Schrodinger Dataset

Work in progress

Target	num. of ligands	avg. Tanimoto	FEP+	SQM2.20	SQM2.20/fixed
BACE	36	0.71	0.61	0.00	0.23
CDK2	16	0.84	0.23	0.29	0.56
JNK1	21	0.85	0.72	0.16	0.19
MCL1	42	0.67	0.59	0.58	0.58
p38	34	0.77	0.42	0.25	0.36
PTP1B	23	0.79	0.64	0.55	0.55
thrombin	11	0.84	0.50	0.63	0.66
Tyk2	16	0.84	0.79	0.58	0.62
AVERAGE	**25**	**0.79**	**0.56**	**0.38**	**0.47**

- SQM2.20 limited by lack of reliable initial structures (severe clashes from docking/modeling)
- simple fixes improve correlations
- further improvements expected after complex refinement of structures

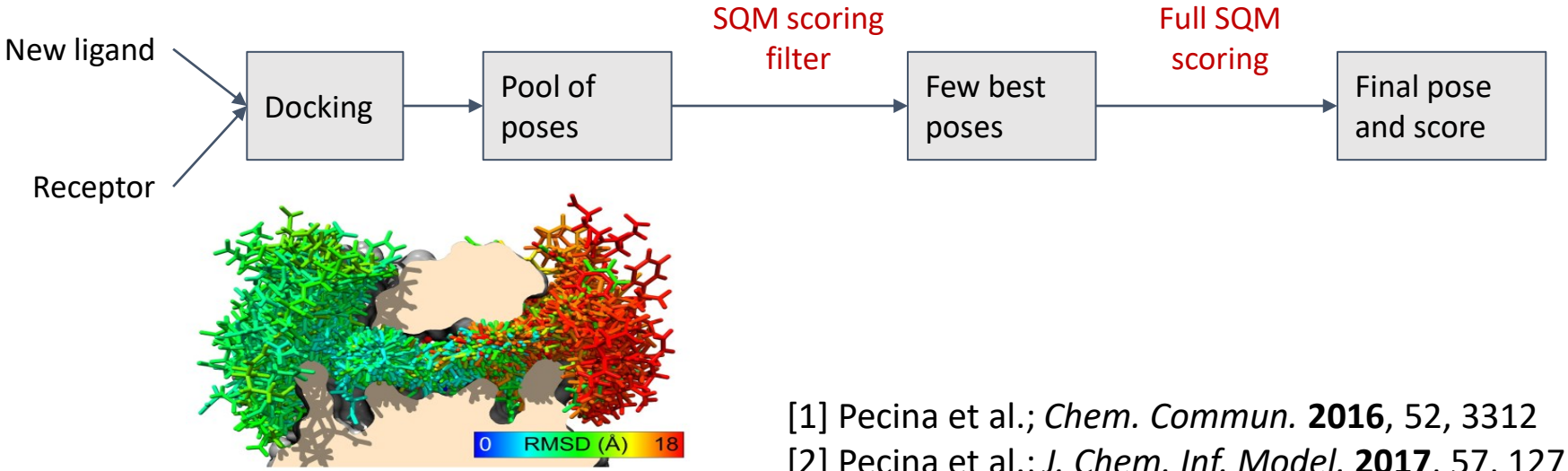
J. Am. Chem. Soc. **2015**, 137, 7, 2695; *J. Chem. Inf. Model.* **2023**, 63, 8, 2438

Integrating SQM2.20 with Docking

Work in progress

solution for complexes with unknown structures?

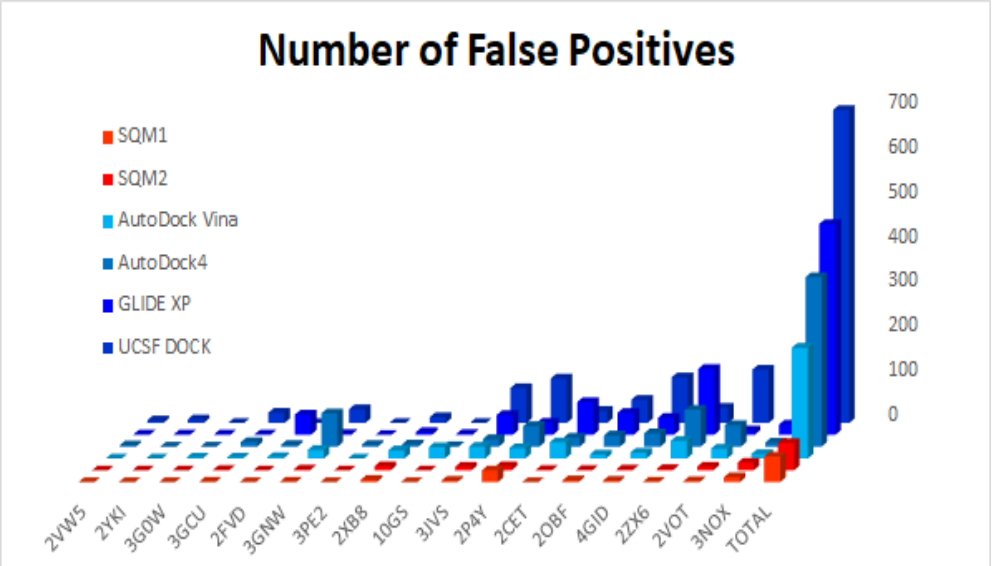
- selecting native-like poses from docking
- previous SQM versions identified the native pose reliably[1,2]



[1] Pecina et al.; *Chem. Commun.* **2016**, 52, 3312
[2] Pecina et al.; *J. Chem. Inf. Model.* **2017**, 57, 127

Native Pose Identification

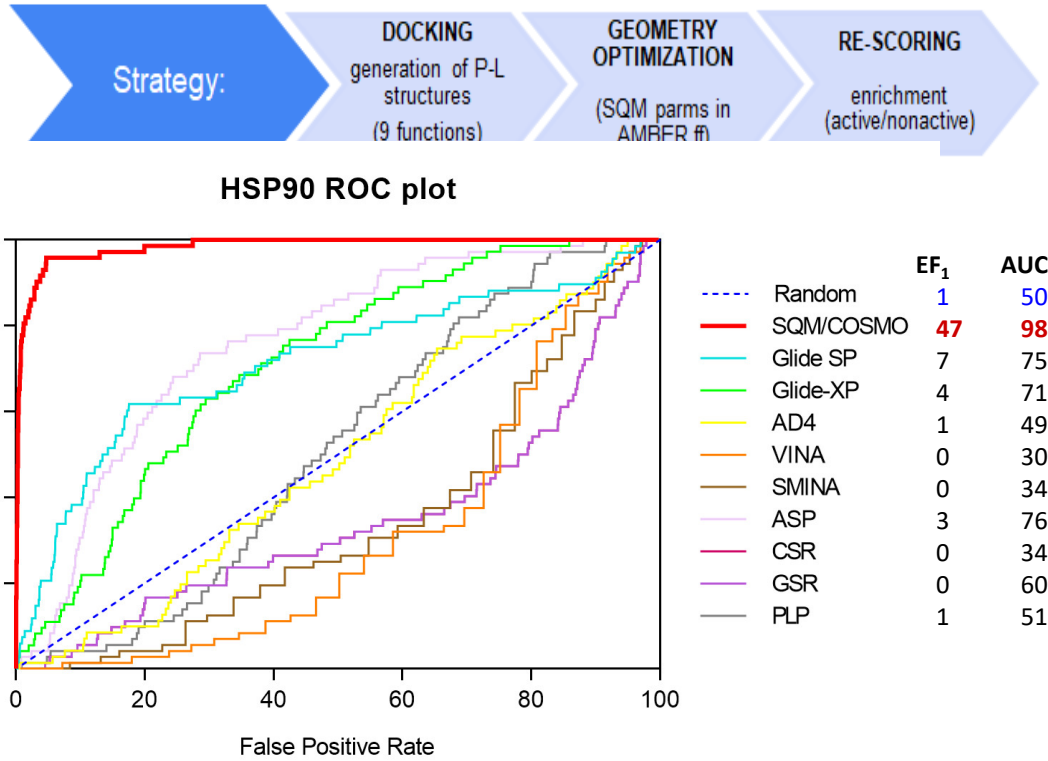
- diverse set of 17 protein-ligand systems
- compared to 8 standard scoring functions
- false positive = a pose with better score than crystal (ideal: zero false positives)
- SQM has 4-12-times less FPs than the standard SFs



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

Towards Virtual Screening

- Heat shock protein (HSP90); cancer and immunity
- 72 biologically active compounds + 4469 structurally similar compounds (DUD-E decoys)
- Enrichment factor (EF1) and ROC curves (AUC%),
  where random is (1, 50%)
  and ideal (63, 100%)



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

SQM2.20: Universal Physics-based Quantum Mechanical Scoring

- **Reliable affinity predictions** (“DFT accuracy”)
- **Reasonable computational cost** (20min/1CPU/compound)
- **Insightful details** of P-L binding (SQM geometries + energetics)
- Tested on **diverse set of curated data**
- publicly available **PL-REX**: 10 proteins, >150 ligands, structures, affinities
- Superior to quick approaches to ranking (MM, standard SFs and M-L)
- Comparable to FEP+ (preliminary results)

SQM 2.20 preprint

DOI:10.26434/chemrxiv-2023-zh03k

Open to collaborations

- Interested in datasets where conventional methods fail
- Comparison to other methods
- Extending data set coverage / application domain
- Trial license for the software

Contact: Lepsik@uochb.cas.cz
LinkedIn: Martin Lepsik
Twitter/X: @Lepsik_science

Thank you for your attention