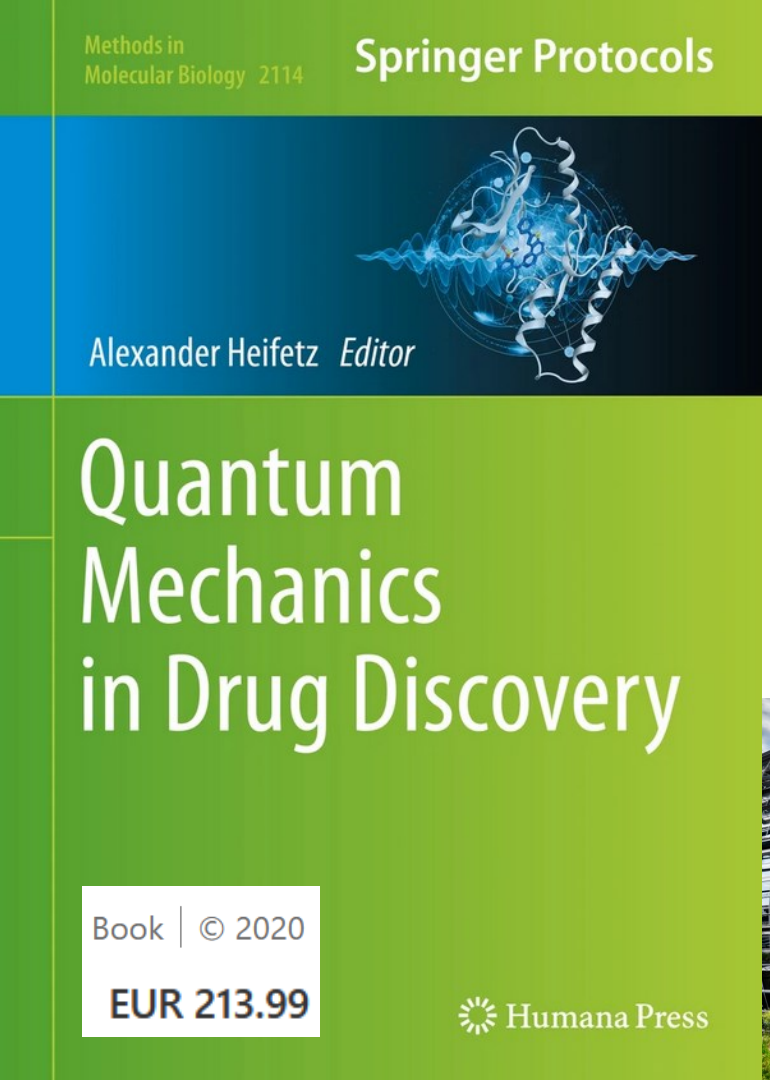


8ADD, Olomouc, Jan 2025

Martin Lepšík

Computational Chemistry for Drug Design

Group leader: Jan Řezáč

ÚOCHB AV ČR
IOCB PRAGUE

Outline

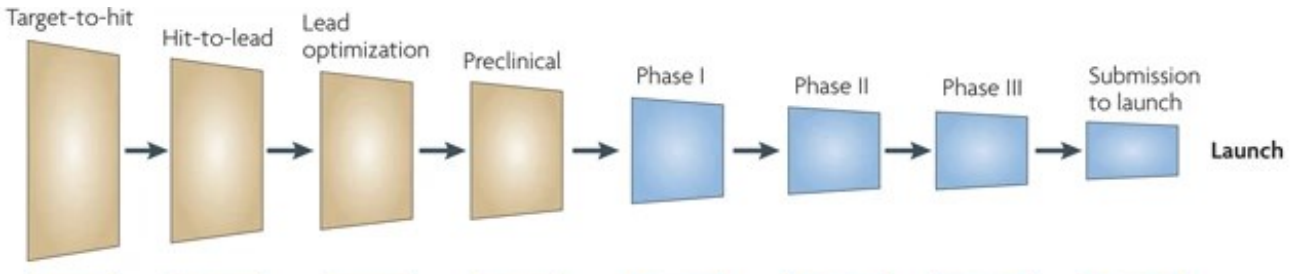
1. Computer-aided drug design
2. SQM Scoring
3. Experimental datasets (structures and affinities)
4. Extensions of SQM (docking, VGS)
5. Insulin – Receptor Case Study

Outline

1. **Computer-aided drug design**
2. SQM Scoring
3. Experimental datasets (structures and affinities)
4. Extensions of SQM (docking, VGS)
5. Insulin – Receptor Case Study

Drug Development: A Lengthy, Costly, Risky Bussiness

– 13+ years
– 1.8 mil. USD
– 4% success rate



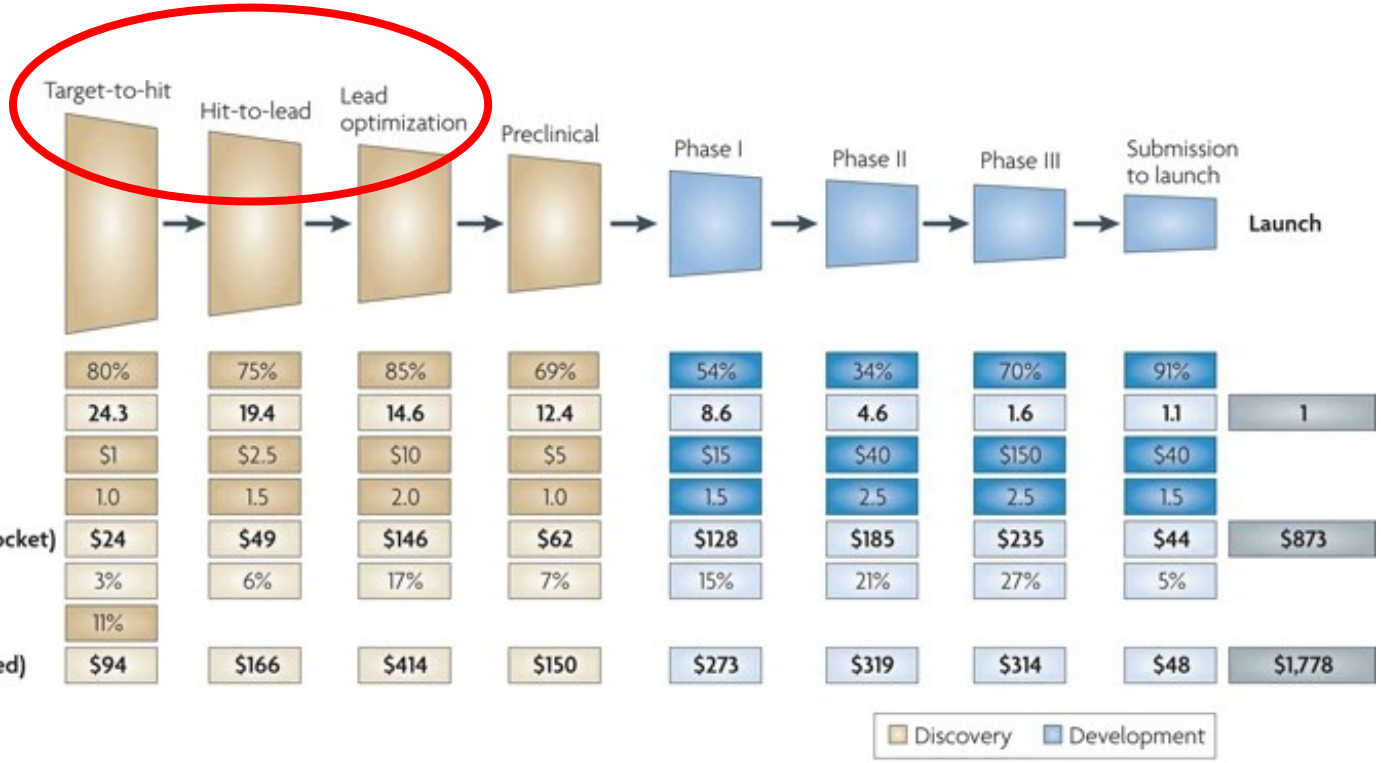
p(TS)	80%	75%	85%	69%	54%	34%	70%	91%	
***WIP* needed for 1 launch**	**24.3**	**19.4**	**14.6**	**12.4**	**8.6**	**4.6**	**1.6**	**1.1**	**1**
Cost per WIP per Phase	$1	$2.5	$10	$5	$15	$40	$150	$40	
Cycle time (years)	1.0	1.5	2.0	1.0	1.5	2.5	2.5	1.5	
Cost per launch (out of pocket)	**$24**	**$49**	**$146**	**$62**	**$128**	**$185**	**$235**	**$44**	**$873**
% Total cost per NME	3%	6%	17%	7%	15%	21%	27%	5%	
Cost of capital	11%								
Cost per launch (capitalized)	**$94**	**$166**	**$414**	**$150**	**$273**	**$319**	**$314**	**$48**	**$1,778**

Discovery Development

Nat Rev Drug Discov **9**, 203–214 (2010)

Drug Development: A Lengthy, Costly, Risky Bussiness

– 13+ years
– 1.8 mil. USD
– 4% success rate

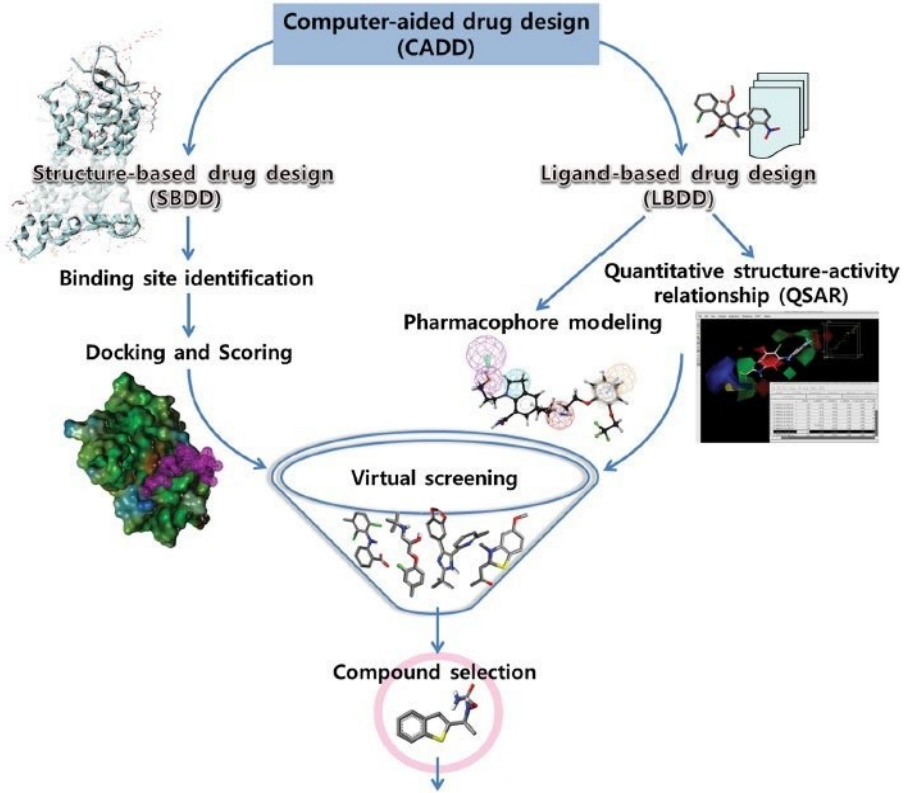


Nat Rev Drug Discov **9**, 203–214 (2010)

Computations in Drug Discovery

– **SBDD**: 3D Structure of Target
– **LBDD**: 2D of Actives/Inactives

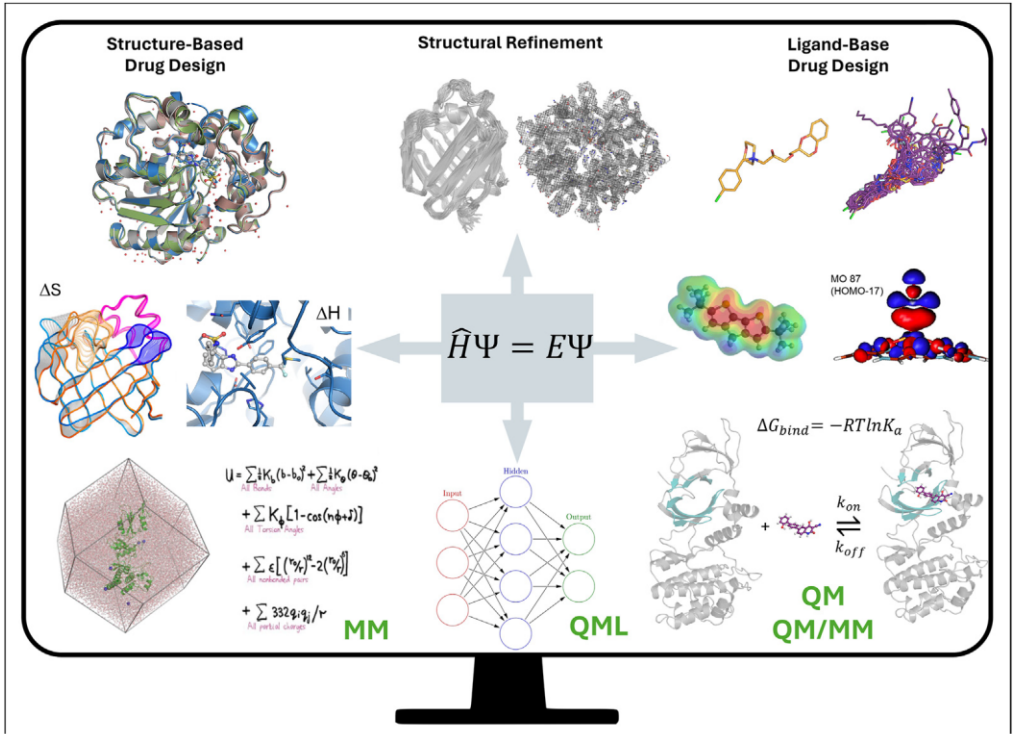
Save Time & Money



Types of Computations

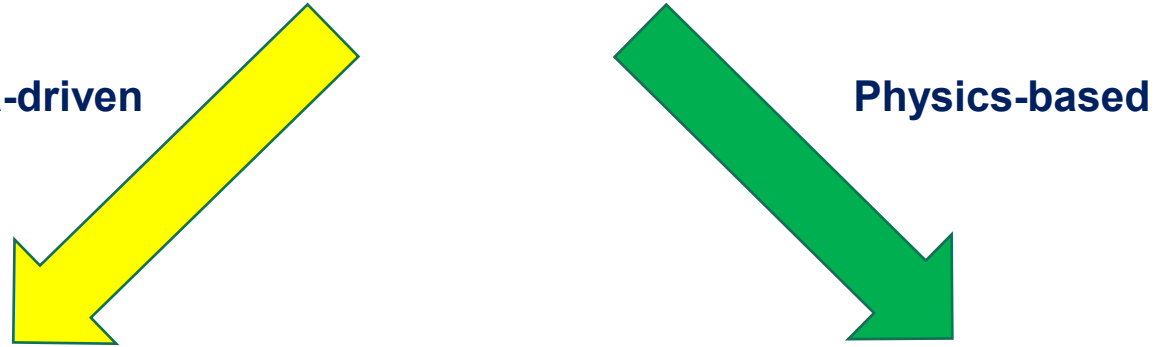
SBDD
– X-ray crystallographic refinement
– Hit Identification (Virtual Screening)
– Docking
– Scoring

LBDD
– Partial charges
– Bioactive conformations
– pK_a predictions



Current Opinion in Structural Biology 2024, 87:102870

Structure-based Affinity Prediction



Standard Scoring Functions (SFs)
- ultrafast (seconds)
- low accuracy

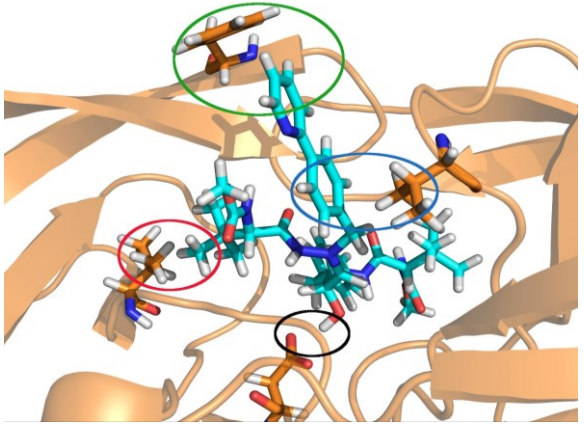
Machine-Learning (M-L)
- ultrafast (seconds)
- ? training data/accuracy,
- ? applicability domain

Free Energy Methods (FEP)
- slow on GPU (days)
- variable accuracy

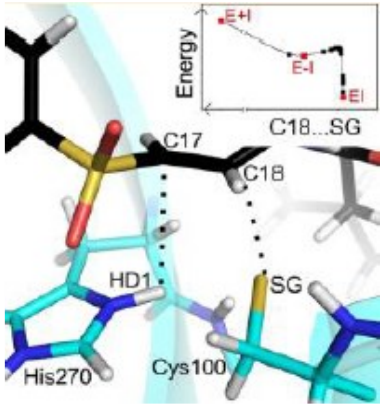
Standard Quantum Mechanics (DFT)
- slow on 10s CPU (days)
- accurate
- ? applicable to biomolecules

Why Quantum Mechanics?

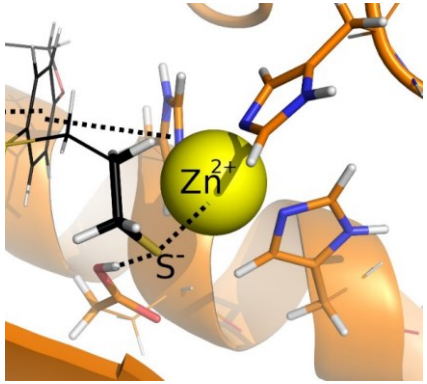
- all types of non-covalent interactions
- dispersion, H-bonding, halogen bonding, etc.
- quantitative description
- metal interactions
- polarization, charge transfer
- covalent binding
- no parametrization of ligands



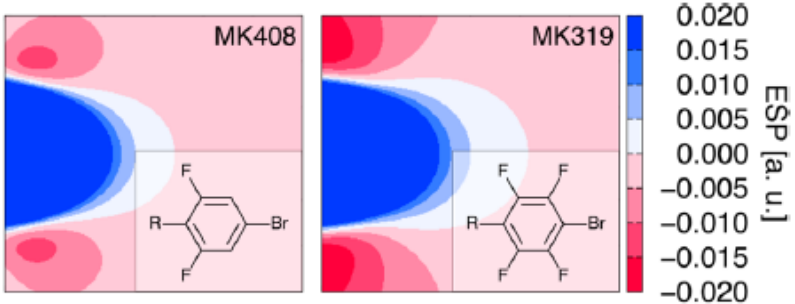
J. Phys. Chem. B, 2010, 114, 12666



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



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



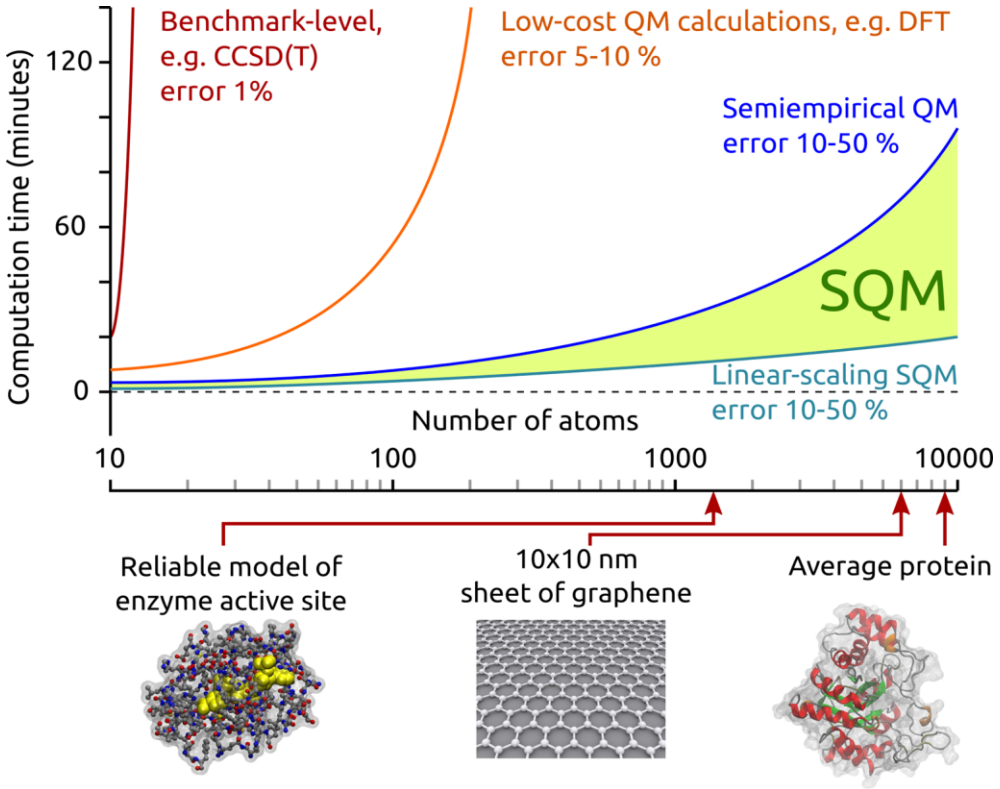
ACS Chem. Biol. 2013, 8, 2484

Which QM method?

CCSD(T)
– slow (weeks; 100s CPUs)
– N^4-N^7 scaling with system size

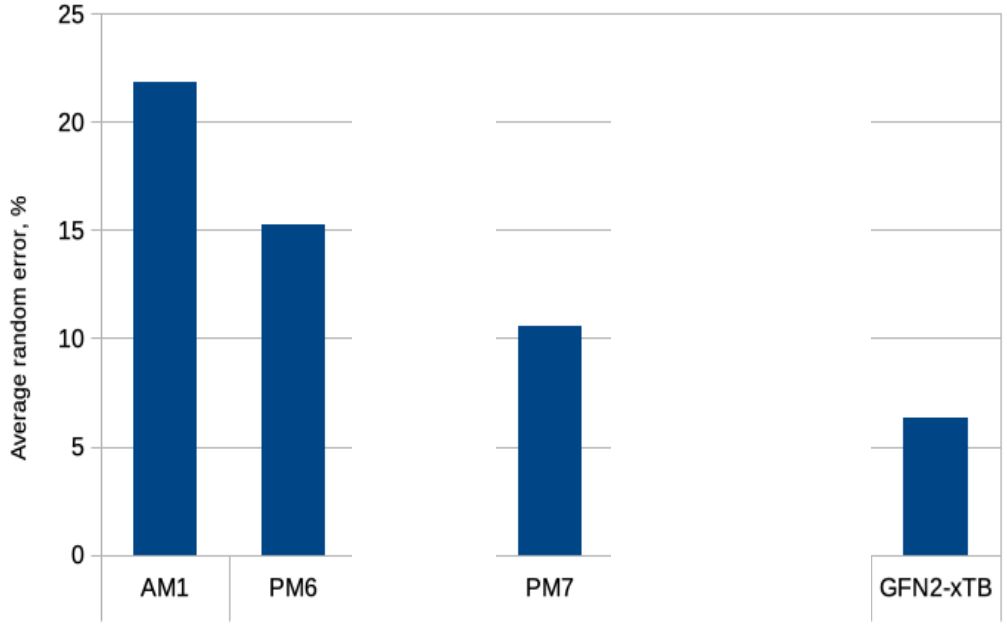
DFT-D
– moderate (days; 10s CPUs)
– biomolecular (100s atoms)

SQM
– fast (minutes; 1 CPU)
– linear-scaling
– 1000s atoms



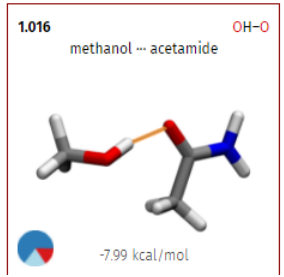
Non-covalent Interactions by SQM

Large errors in 15 protein-ligand complexes
CCSD(T) reference



CCSD(T) Interaction Energies in Small Models

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

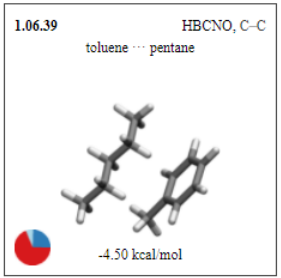


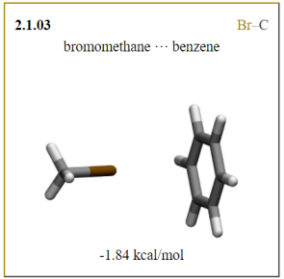
Coverage of the periodic table:

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+





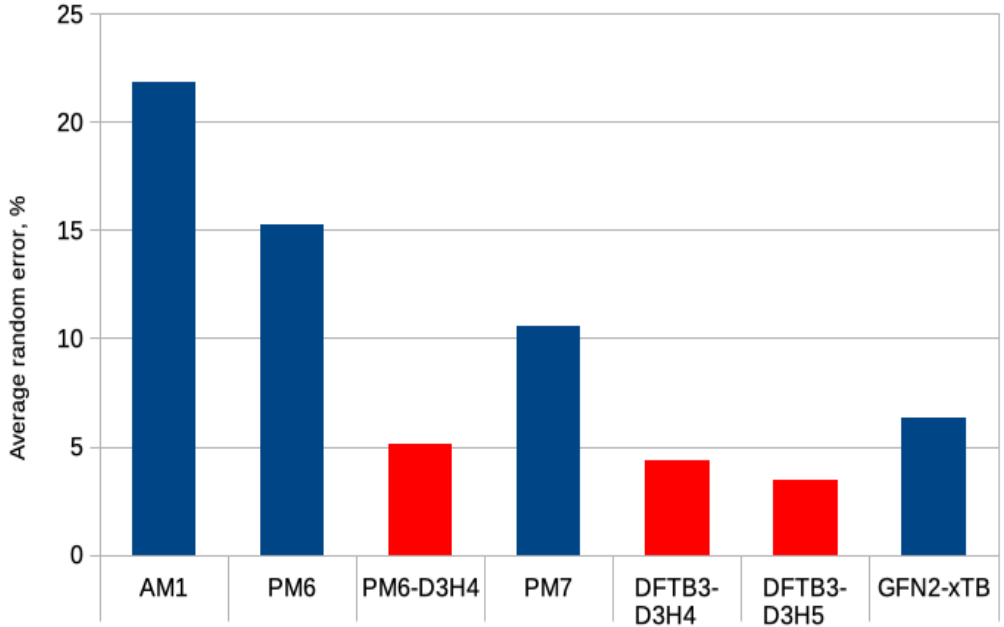
- Development of semiempirical QM methods - corrections for non-covalent interactions
- chemical accuracy (1 kcal/mol) in small dimers

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

Corrected SQM Methods

Errors in 15 protein-ligand complexes,

CCSD(T) reference



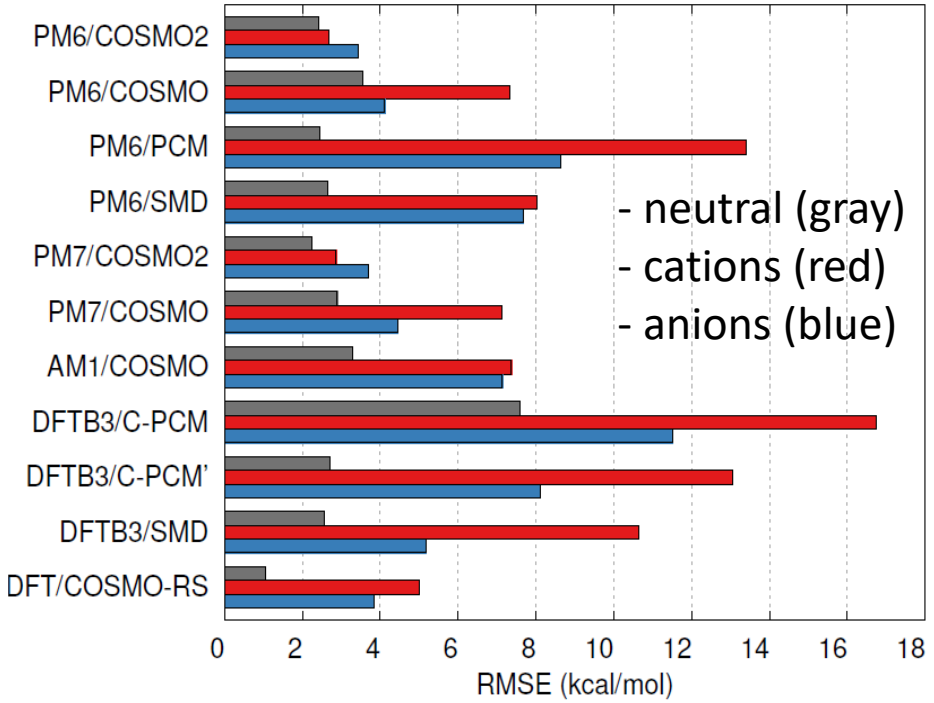
- Fast calculation
- Easy preparation
  (no system-specific parameters)
- Accuracy?

PM6-D3H4X

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

COSMO2 Implicit Solvation Model

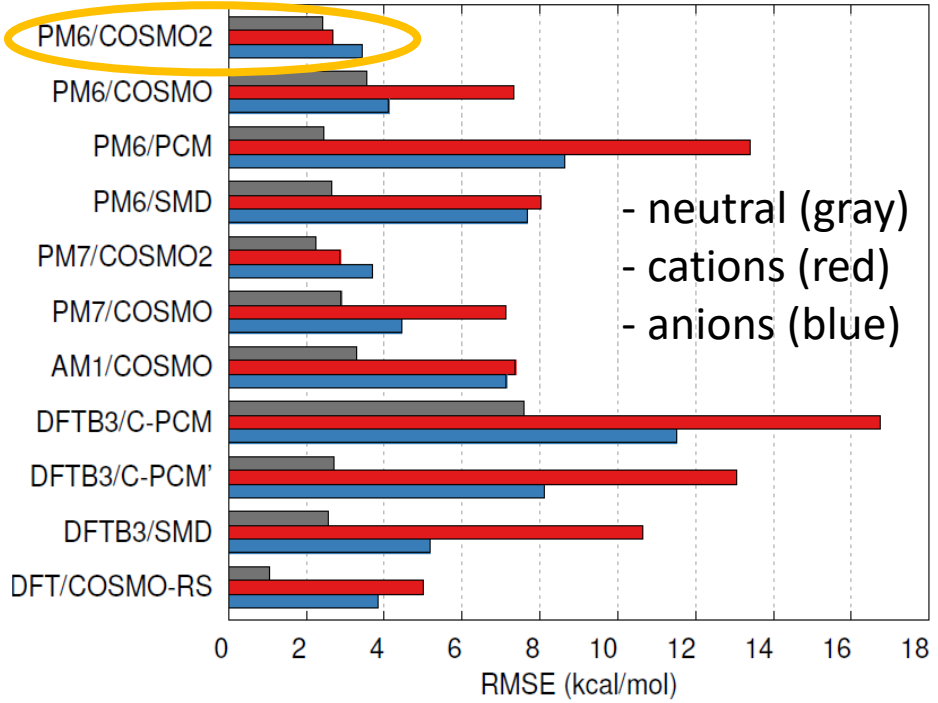
- reparametrisation of COSMO
- adding non-polar solvation
- single-point energies only



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

COSMO2 Implicit Solvation Model

- reparametrisation of COSMO
- adding non-polar solvation
- single-point energies only

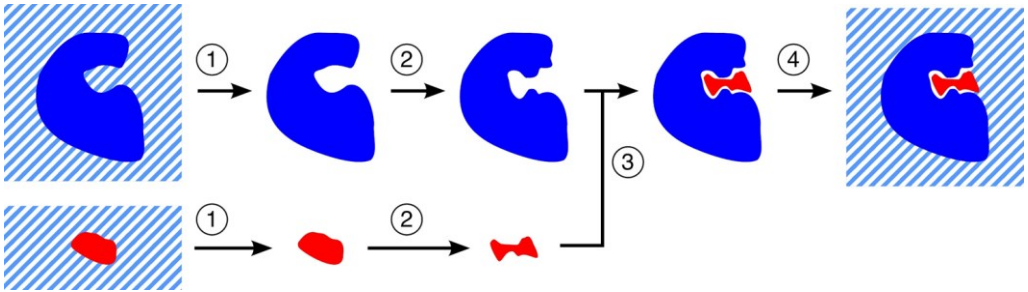


Kříž, K. & Řezáč, J. *J. Chem. Inf. Model.* **2019**, 59, 229

Outline

1. Computer-aided drug design
2. **SQM Scoring**
3. Experimental datasets (structures and affinities)
4. Extensions of SQM (docking, VGS)
5. Insulin – Receptor Case Study

SQM-based 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
$+ \Delta G_{H+}$	PM6-D3H4X/COSMO2 difference
$- T\Delta S$	LM5 model fitted to QM data

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

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

Application of SQM-based Scoring

Ranking, interaction analysis

22 publications (since 2010)

- Cyclin-Dependent Kinases (Cancer) - *5*
- Carbonic Anhydrases (Cancer) - *3*
- Cathepsins (Schistosomiasis) - 4
- Serine Racemase (neuropathologies) - 1
- Aldose-Reductases (Diabetes) - 3
- Insulin Analogues (Diabetes) - 1
- HIV Protease (AIDS) - 2
- Trypsin/Chymotrypsin (cancer) -1
- Polymerases (Influenza) - 2

Sampling and virtual screening

4 publications (since 2016)

- 2 sampling studies on 4 proteins
  - Acetylcholine esterase (Alzheimer's disease)
  - TACE/ADAM17 (inflammation)
  - Aldose-Reductases (Diabetes)
  - HIV Protease (AIDS)
- + 17 proteins in wider sampling study
  - incl. Hepatitis C RNA polymerase, Glutathione S-transferase (cancer resistance)
- Virt. screening - Heat shock protein (cancer)

Reviews: *ChemPlusChem* **2013**, *78*, 921; *ChemPlusChem* **2020**, *85*, 2362

Is SQM-based Scoring Universal?

Outline

1. Computer-aided drug design
2. SQM Scoring
3. **Experimental datasets (structures and affinities)**
4. Extensions of SQM (docking, VGS)
5. Insulin – Receptor Case Study

Free energy calculations https://doi.org/10.1038/s41570-023-00552-4

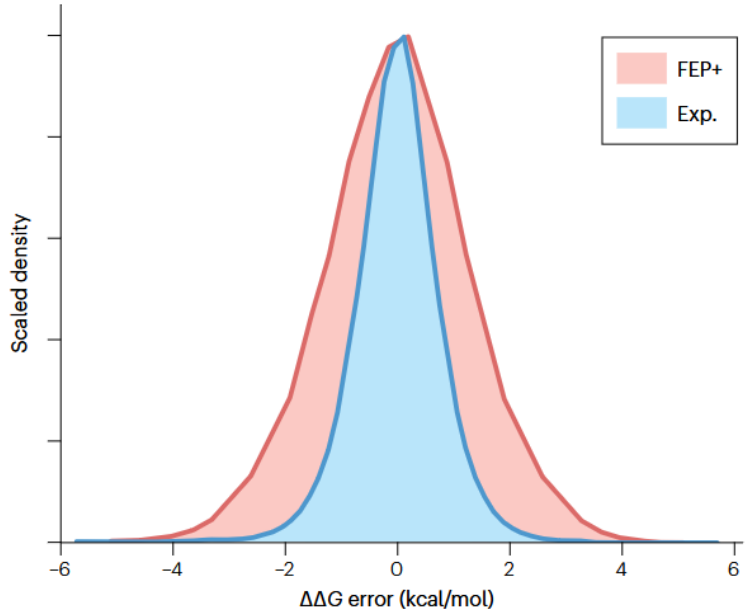
The experiment is the limit

Christina E. M. Schindler, Daniel Kuhn & Ingo V. Hartung

Check for updates

nature reviews chemistry Volume 7 | November 2023 | 752–753 | 752

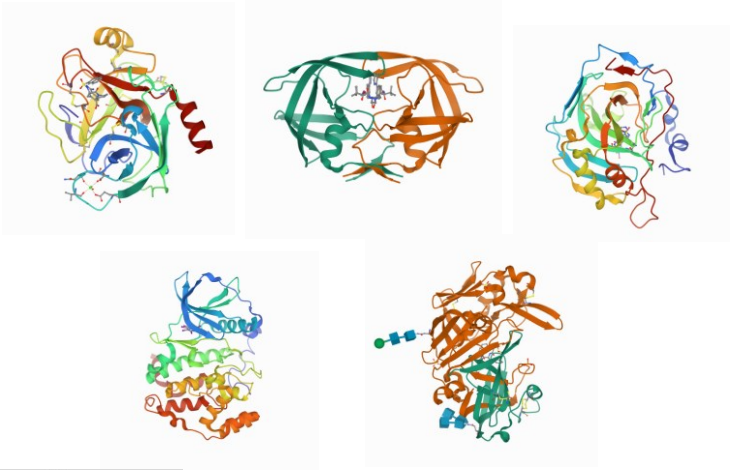
- Reproducibility from multiple independent measurements: $R^2 = 0.8$
- Datasets for Free Energy Perturbation (FEP)



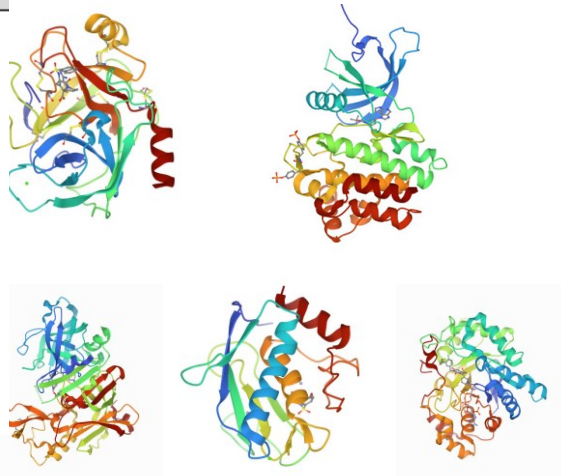
Building high-quality dataset

PL-REX: Protein-**L**igand / **R**eliable **Ex**periment data set

- 10 targets: 10+ ligands per each
- High-resolution crystal structures
- Affinities measured in one lab (K_i preferred over 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

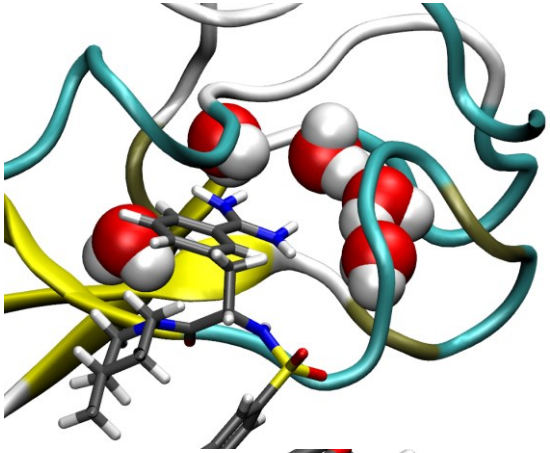


Nat. Commun. 2024, 15, 1127; https://github.com/Honza-R/PL-REX

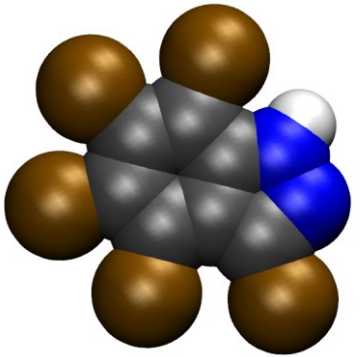
Challenging Cases in PL-REX dataset

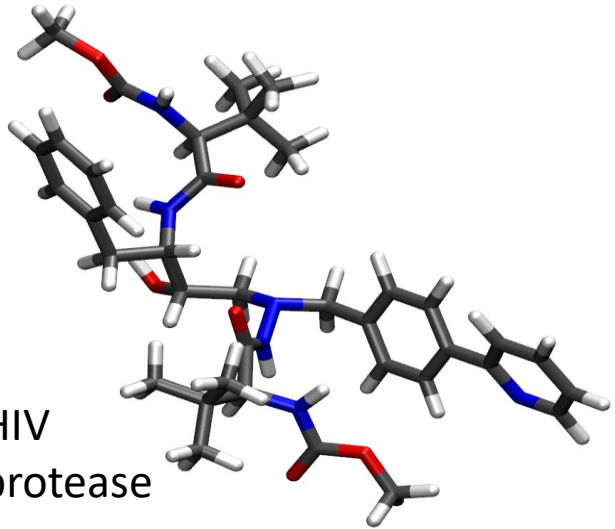
- Large flexible ligands
- Halogen bonding
- Binding via metal
- Protonation upon binding
- Water bridging protein and ligand

Trypsin

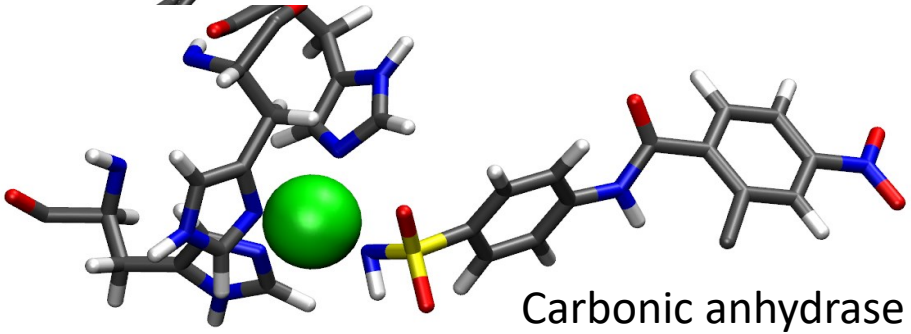


Casein kinase 2





HIV protease



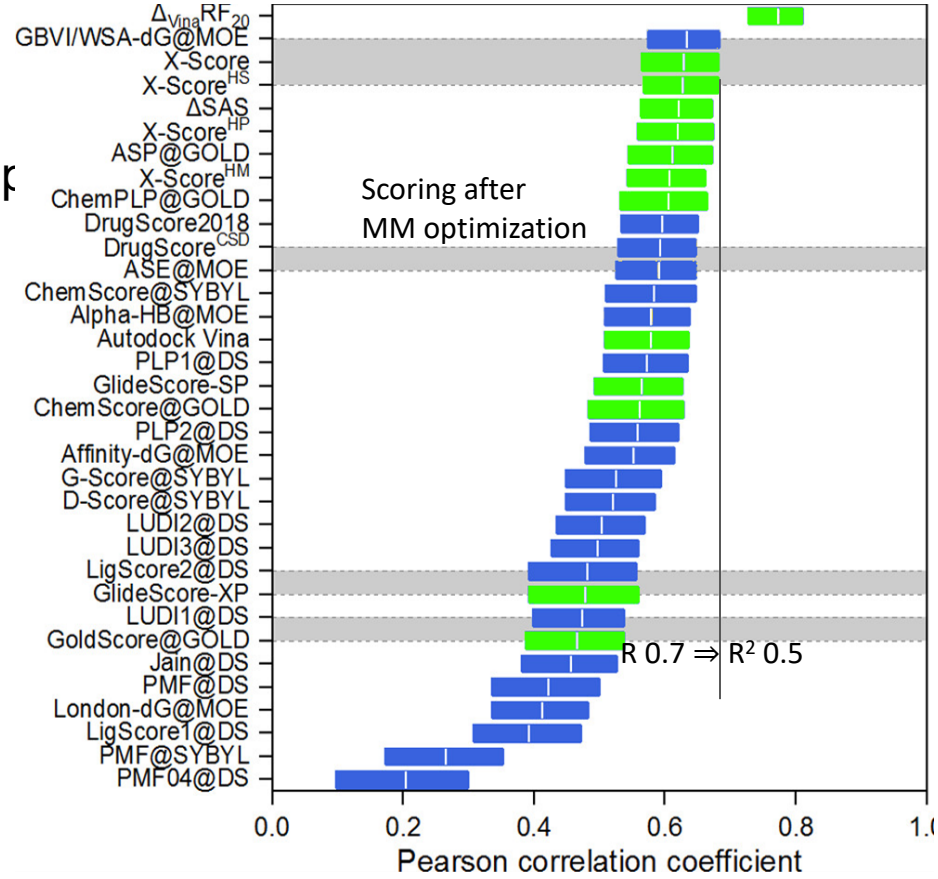
Carbonic anhydrase II

Standard Scoring Functions

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

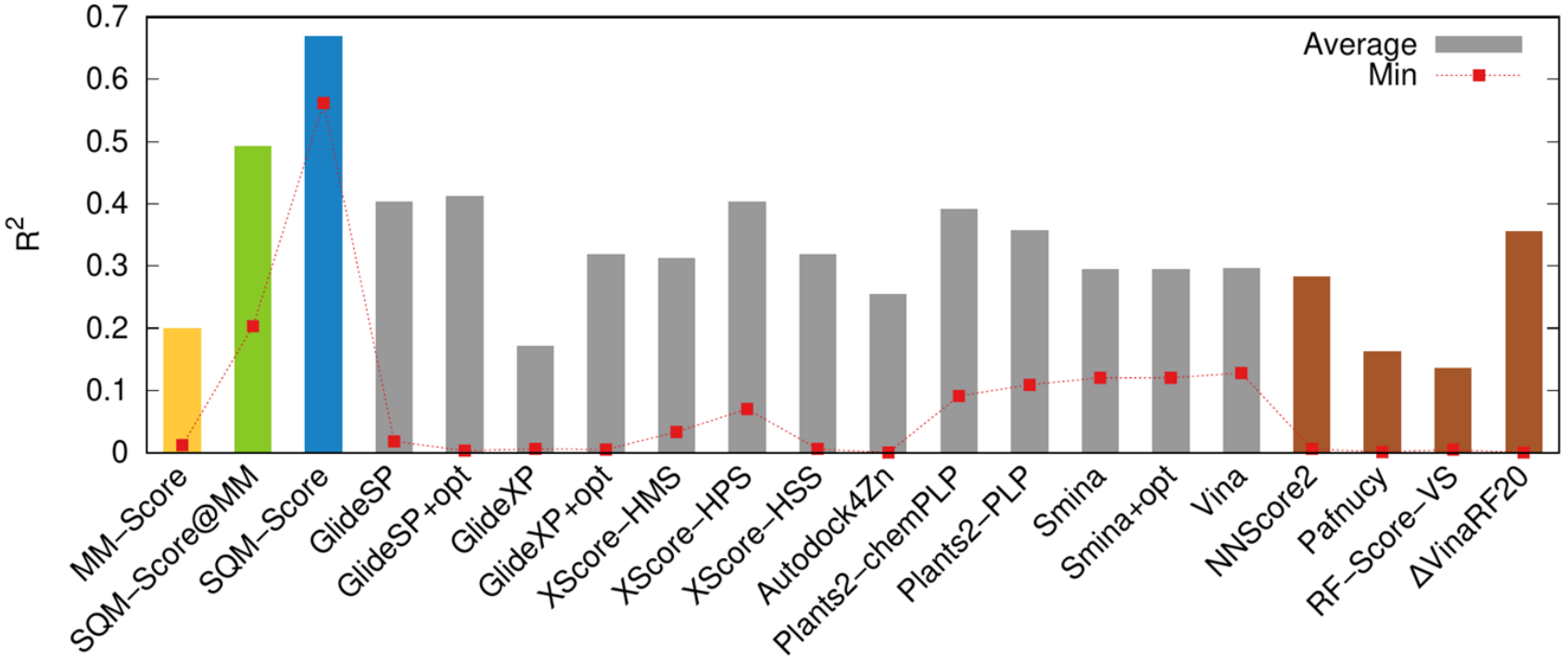
Timing:

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



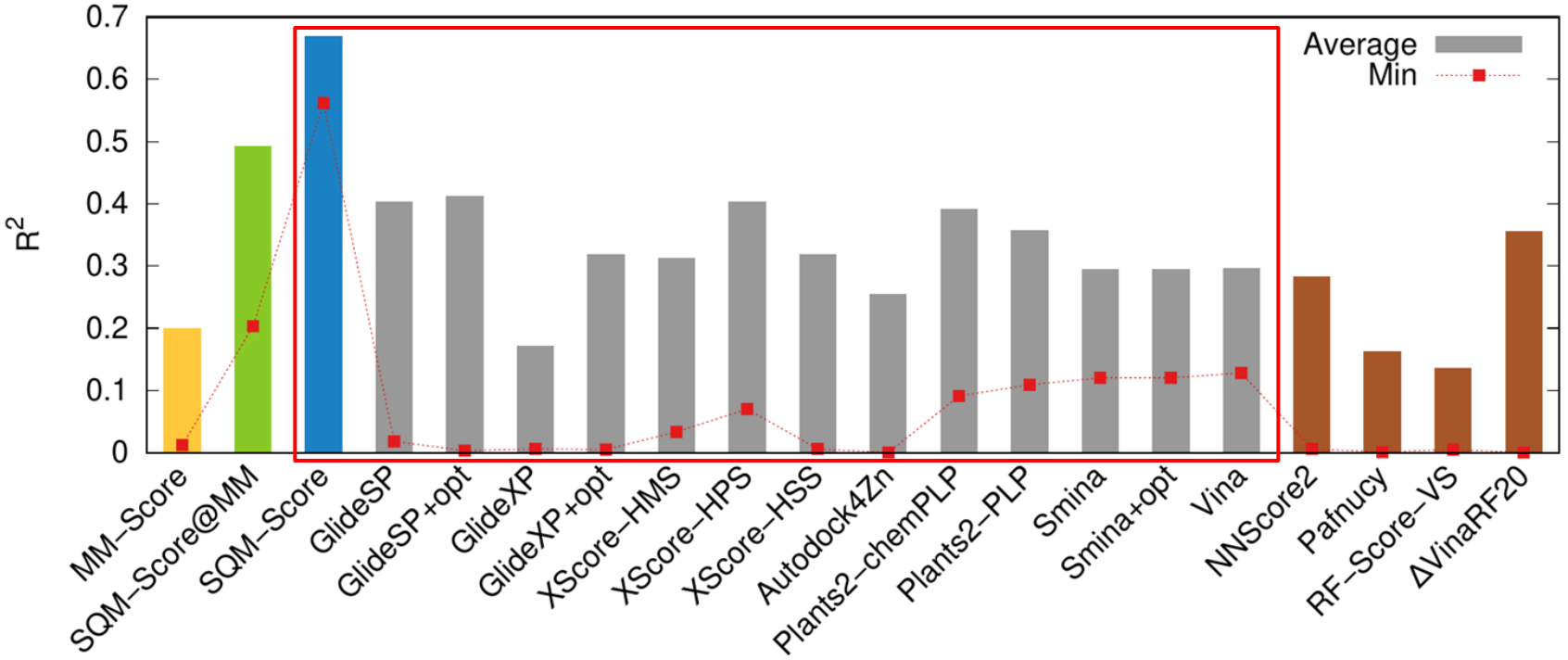
Su, M. et al., *J. Chem. Inf. Model.*, **2019,** 59, 895.

Scoring on PL-REX



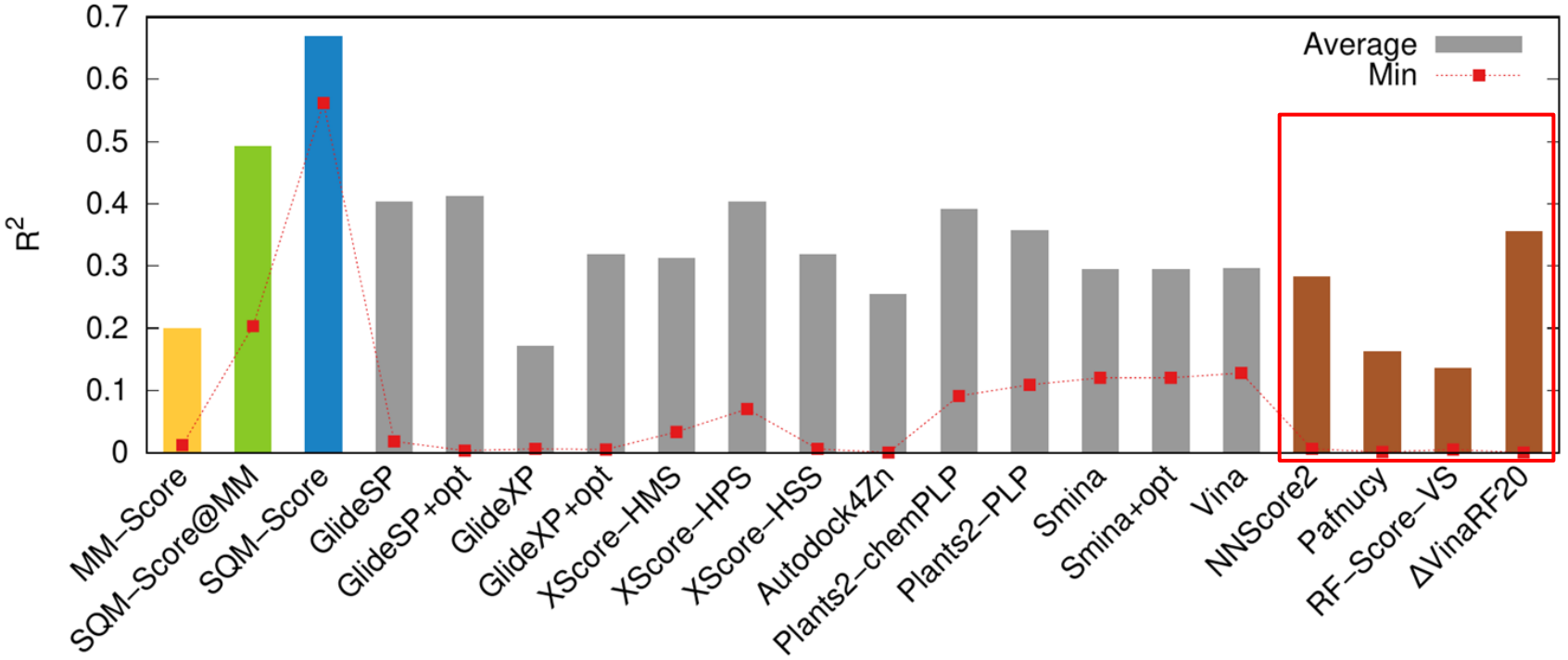
Correlation with experimental affinities, averaged over 10 targets

Scoring on PL-REX



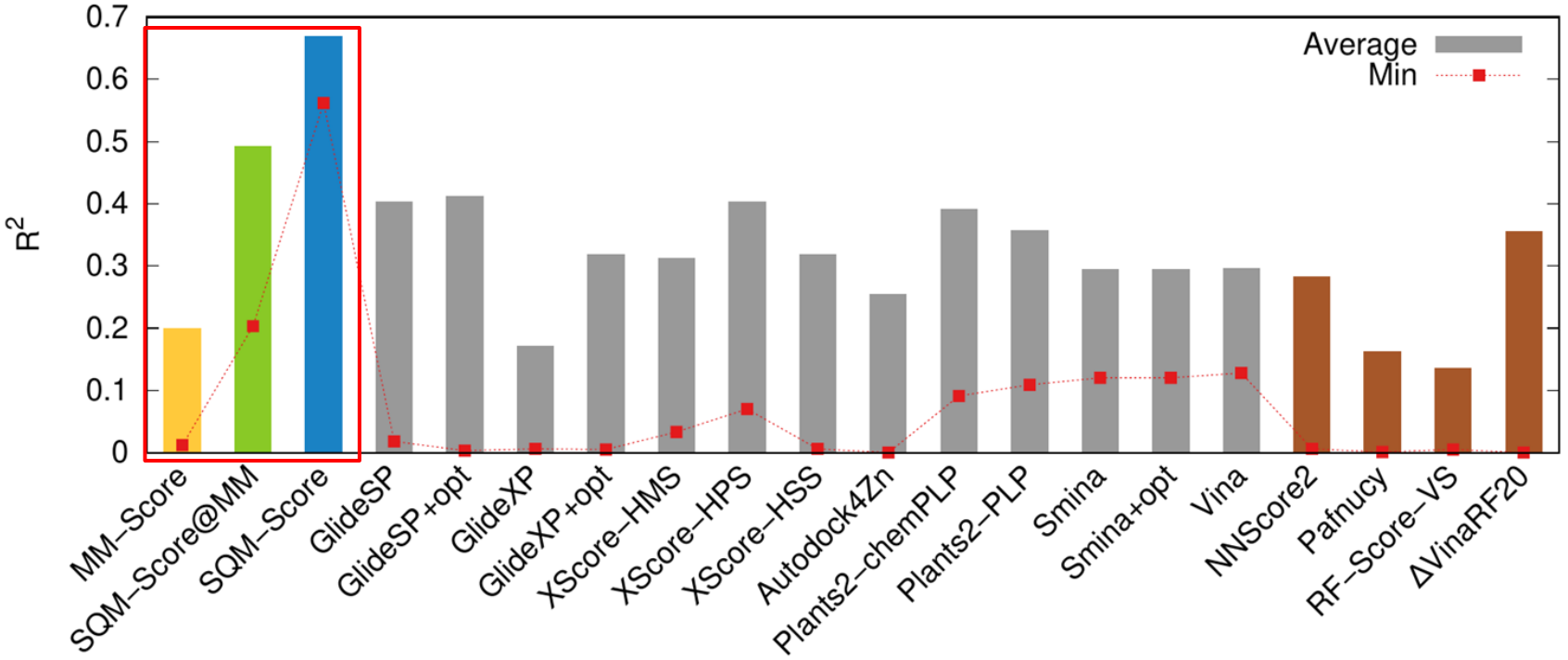
Correlation with experiment, averaged over 10 targets

Comparison with Scoring Functions



Correlation with experiment, averaged over 10 targets

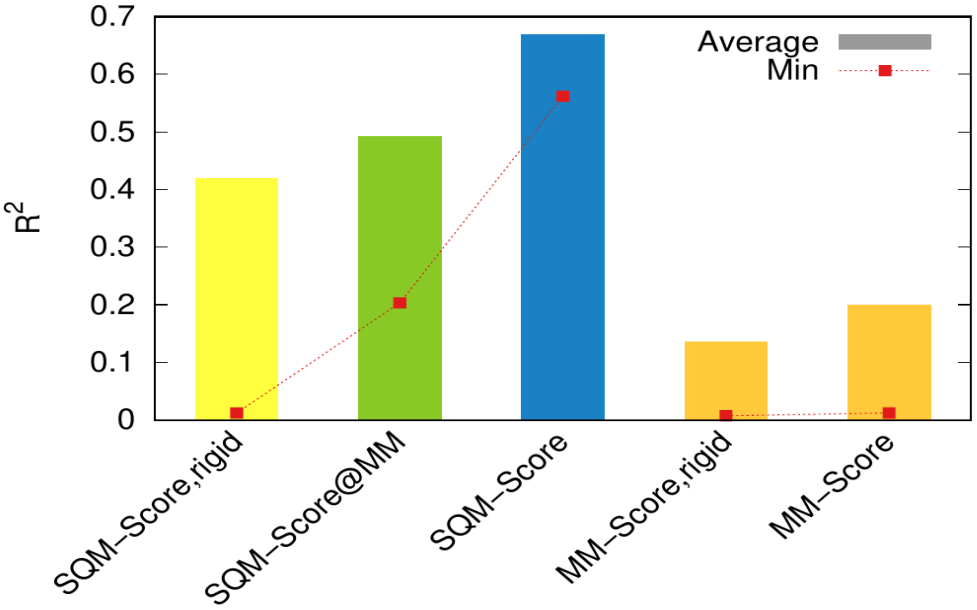
Comparison with Scoring Functions



Correlation with experiment, averaged over 10 targets

P-L complex geometry

- determines the quality of scoring
- SQM score on different geometries



nature communications

Article

https://doi.org/10.1038/s41467-024-45431-8

SQM2.20: Semiempirical quantum-mechanical scoring function yields DFT-quality protein–ligand binding affinity predictions in minutes

Received: 20 July 2023

Accepted: 24 January 2024

Adam Pecina[1,2], Jindřich Fanfrlík[1,2], Martin Lepšík[1] & Jan Řezáč[1]

Nat. Commun. 2024, 15, 1127

SQM2.20 vs. MM or DFT

- SQM: universal performance across targets
- AMBER geometries deteriorate SQM2.20 scoring in some targets
- AMBER scoring: low performance
- SQM2.20 comparable to DFT
- SQM2.20 is fast (20 min/system on 1CPU) vs. DFT with ~10^3 CPU-hours / system)

	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***

Nat. Commun. 2024, 15, 1127

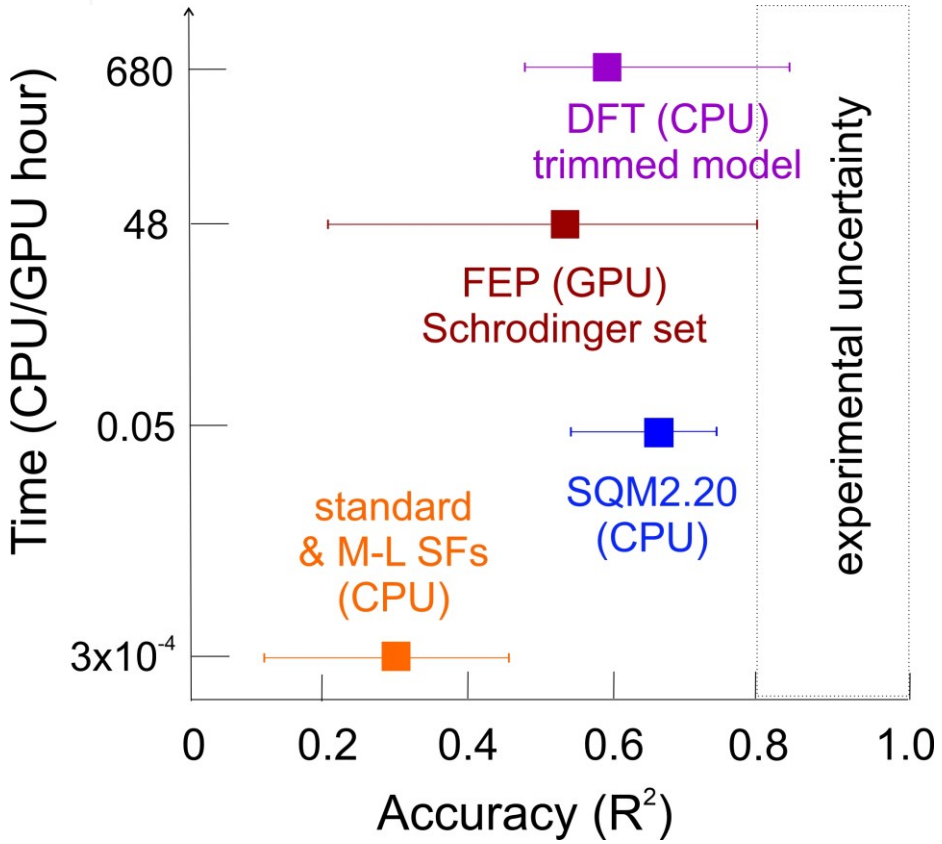
Affinity Prediction: Timing

End-point Methods

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

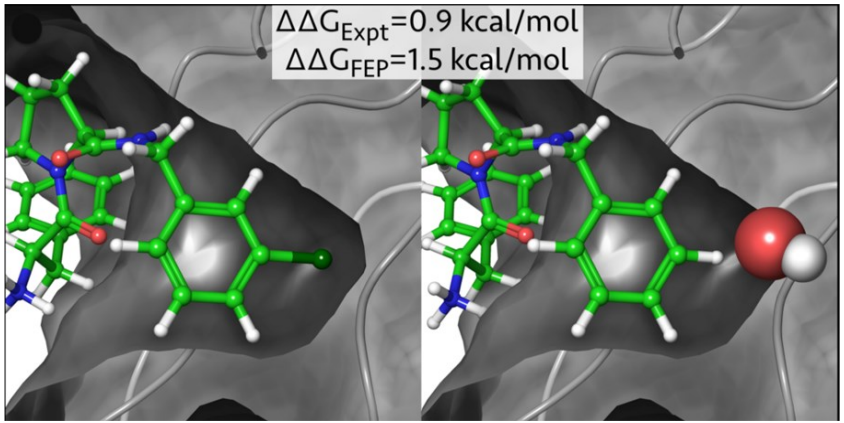
Ensemble Methods

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



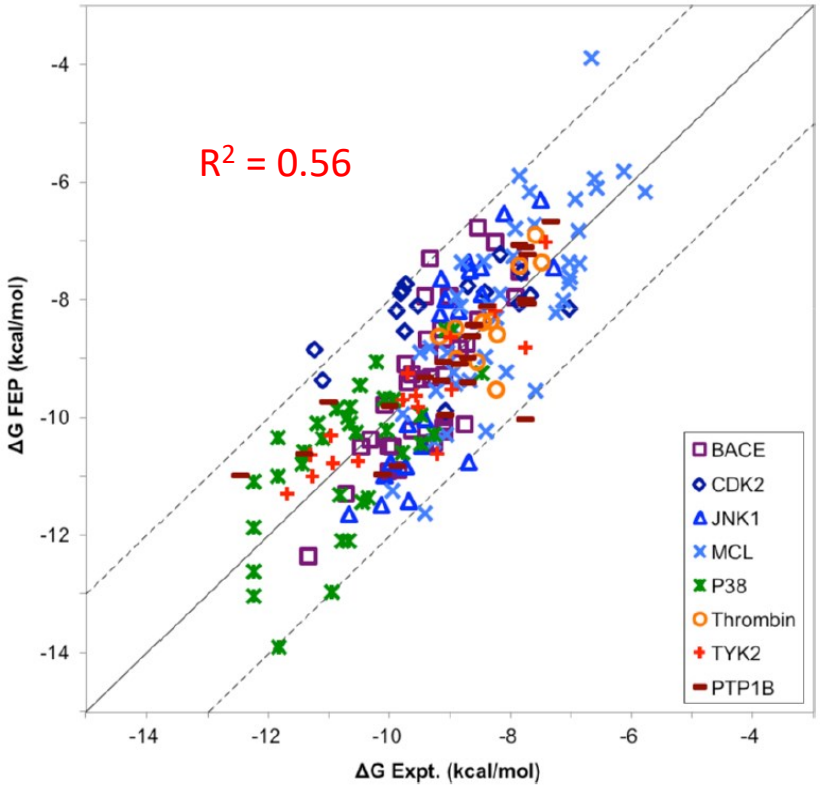
Comparison of SQM2.20 to FEP+

Wang Dataset for Free Energy Perturbation



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

SQM2.20 vs. FEP+ on Wang Dataset

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**

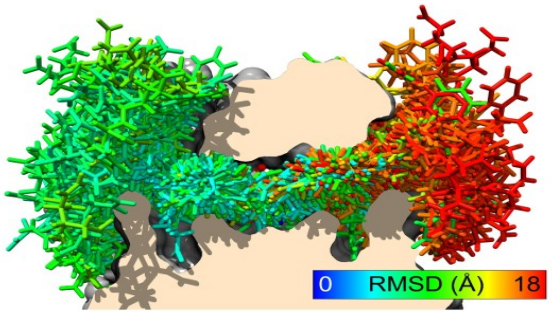
- SQM2.20 limited by lack of reliable initial structures (severe clashes from docking/modeling)
- simple fixes improve correlations
- further improvements expected after complex refinement of structures

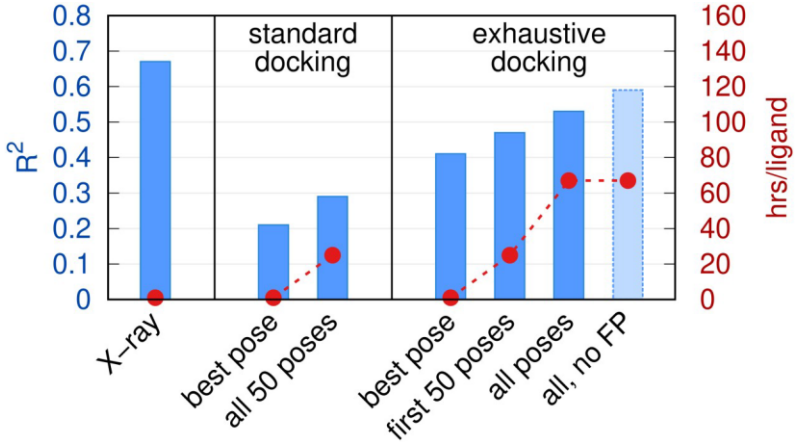
Outline

1. Computer-aided drug design
2. SQM Scoring
3. Experimental datasets (structures and affinities)
4. **Extensions of SQM (docking, VGS)**
5. Insulin – Receptor Case Study

Integrating SQM Scoring with Docking

- automatic protocol for selecting best poses from docking
- SQM identifies the native pose reliably

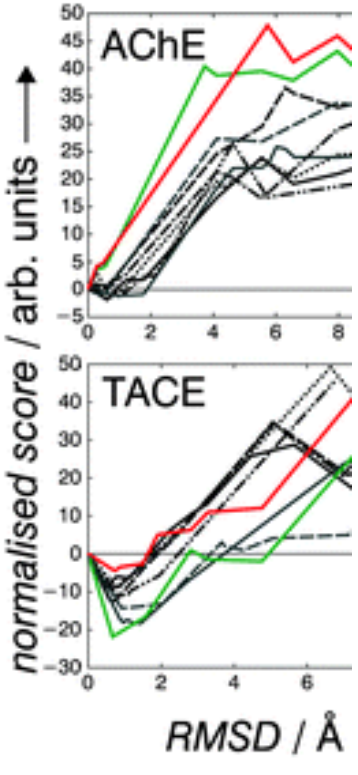


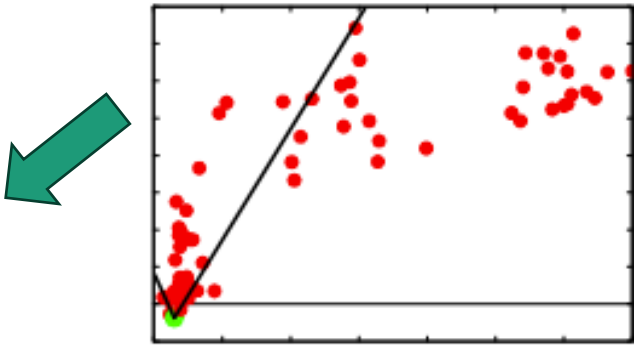


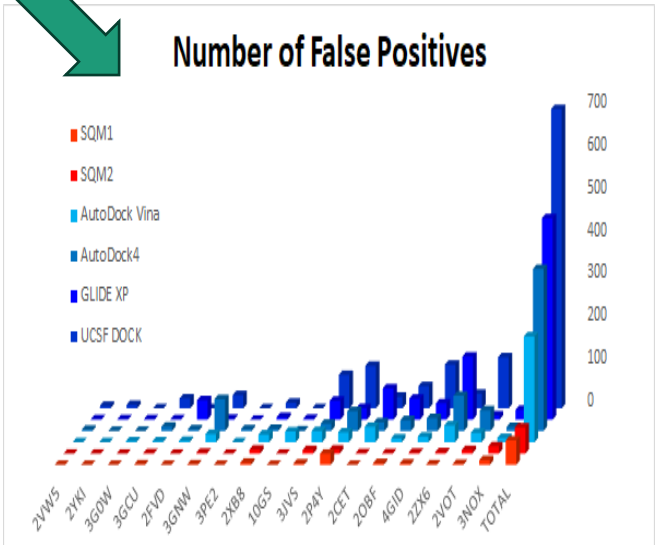
Pecina et al.; *Chem. Commun.* **2016**, 52, 3312
Pecina et al.; *J. Chem. Inf. Model.* **2017**, 57, 127

Native Pose Identification

- diverse set of 17 protein-ligand systems
- SQM and 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







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

Outline

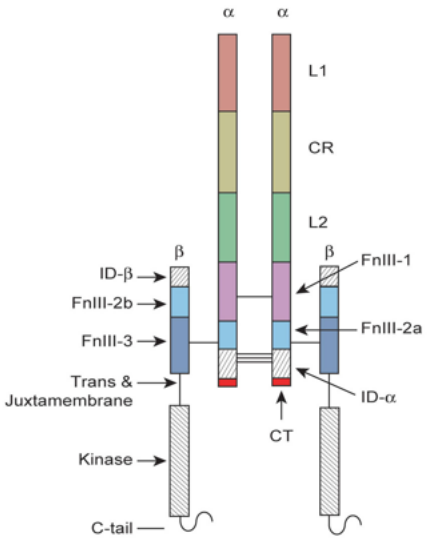
1. Computer-aided drug design
2. SQM Scoring
3. Experimental datasets (structures and affinities)
4. Extensions of SQM (docking, VGS)
5. **Insulin – Receptor Case Study**

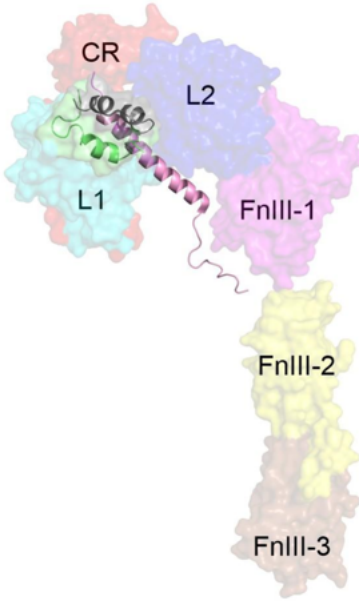
Insulin Receptor (IR)

ectodomain

transmembrane

signaling kinase

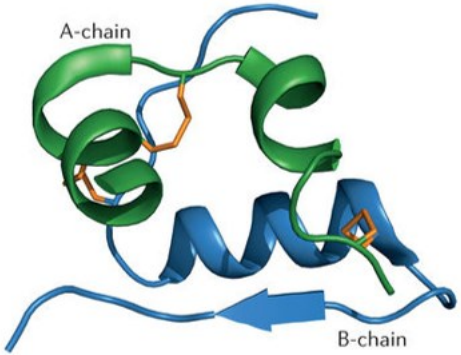


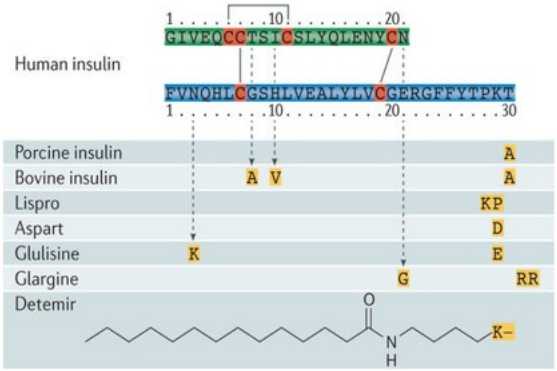


homodimeric (αβ/α'β')
glycoprotein

X-ray structure
resolution 3.8 Å

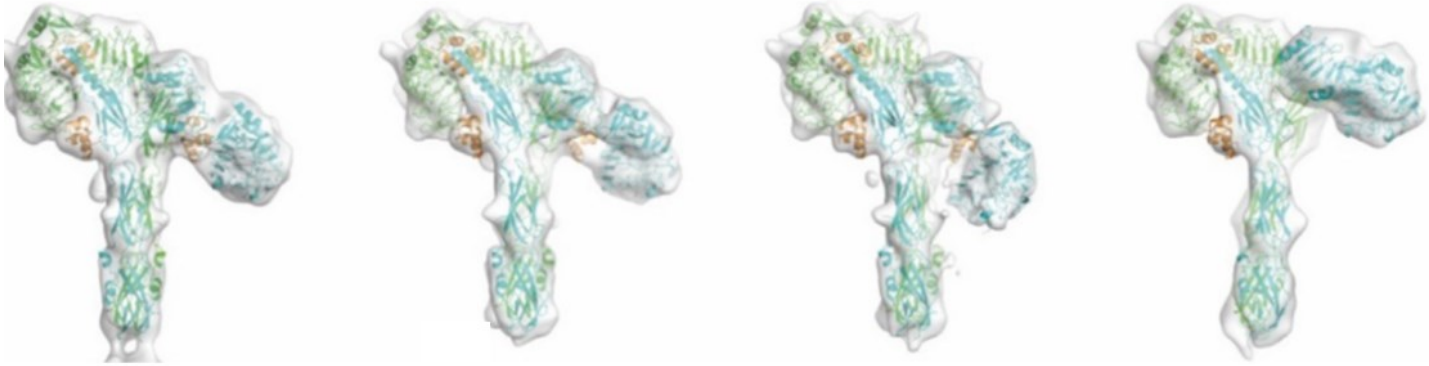
Insulin Analogs



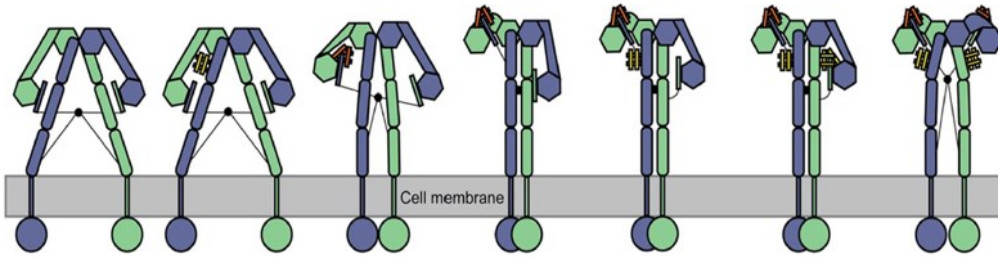


McKern, N., Lawrence, M., Streltsov, V. *et al. Nature* **443**, 218–221 (2006)

Cryo-EM Conformation Continuum

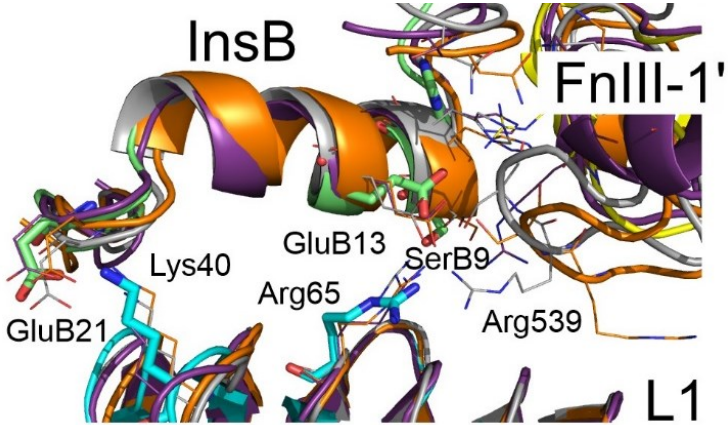


– activation pathway
– 0-4 insulins bound
– resolution 3-9 Å



J. Nielsen, J. Brandt, T. Boesen, et al. J. Mol. Biol. 434 (2022) 167458

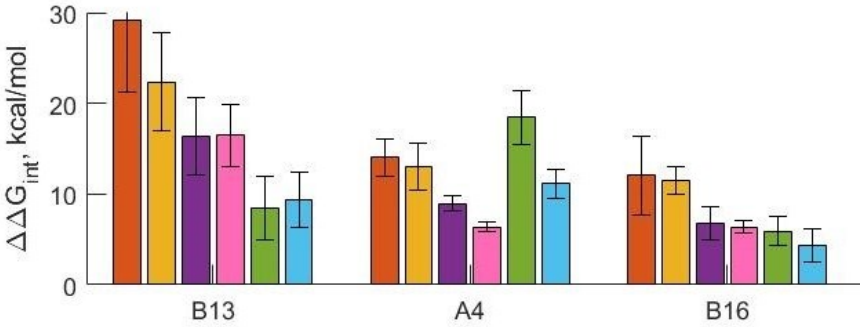
Local Sampling via Molecular Dynamics

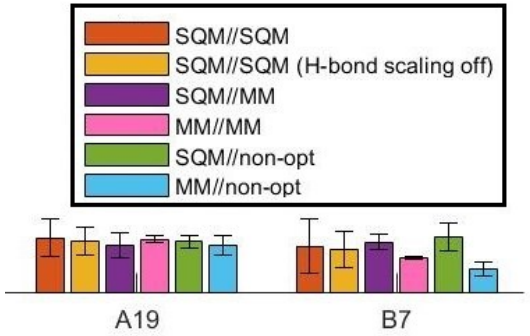


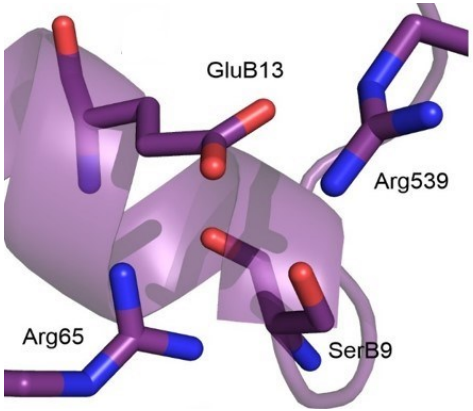
– occupancies of H-bonds and non-polar contacts throughout MD

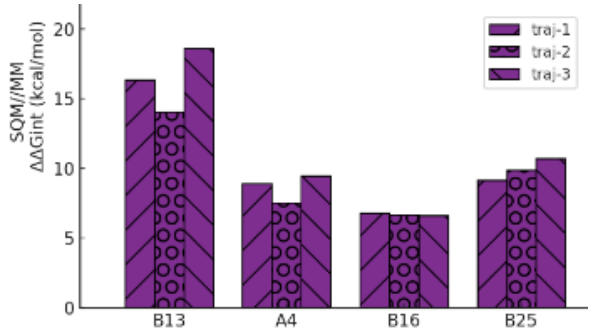
		traj-1	traj-2	traj-3
Ile A2	Phe 714	90	50	62.9
Ile A2	His 710	90	90	55
Val A3	Asp 707	94	77.1	78.6
Tyr A19	Phe 714	86.7	92.5	63.3
Tyr A19	Val 715	70	100	70
Tyr A19	Pro 716	45.3	62.7	60
Gly B8	Glu 706	100	100	83.3
Val B12	Leu 37	90	100	90
Val B12	Phe 64	55	55	36.7
Val B12	Arg 65	80	63.3	60
Val B12	Phe 714	60	70	37.5
Leu B15	Phe 714	93.3	84.3	91.7
Tyr B16	Phe 39	13.7	25.2	13.1
Gly B23	Asn 15	100	93.3	96.7
Phe B24	Leu 37	86	88	2.5
Phe B24	Phe 714	58.6	76.3	66.7
Phe B25	Pro 716	60	100	62.5
Phe B25	Arg 717	84.5	55.3	65
Phe B25	Pro 718	76.7	95	80
Tyr B26	Asp 12	74.3	64.3	88

Virtual Glycine Scan of Insulin - Receptor









Eur. J. Org. Chem. 2018, 5203–5211

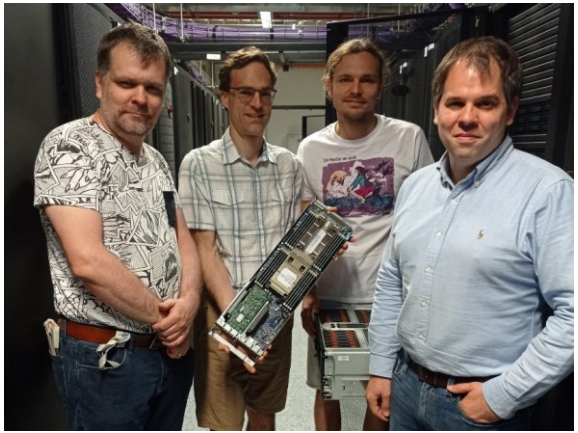
Miloš Halda, poster

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+

Acknowledgements

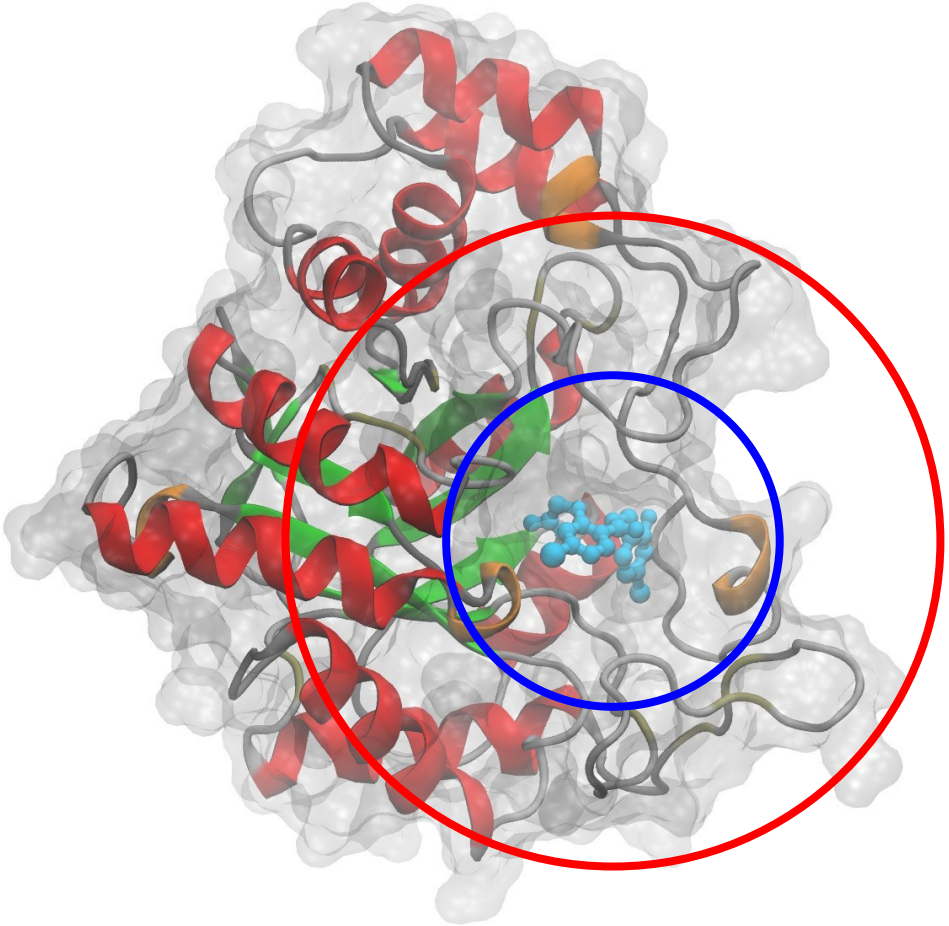
– P. Hobza and his team members
– HPCg team
– IOCB tech
– GA CR



Thank you for your attention

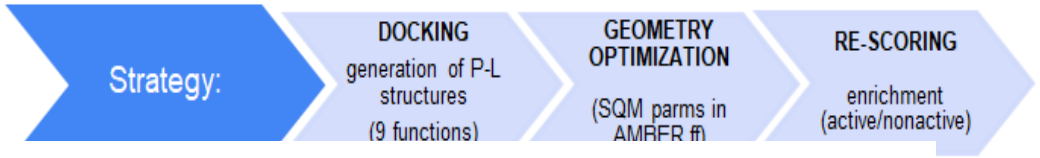
QM/MM Setup

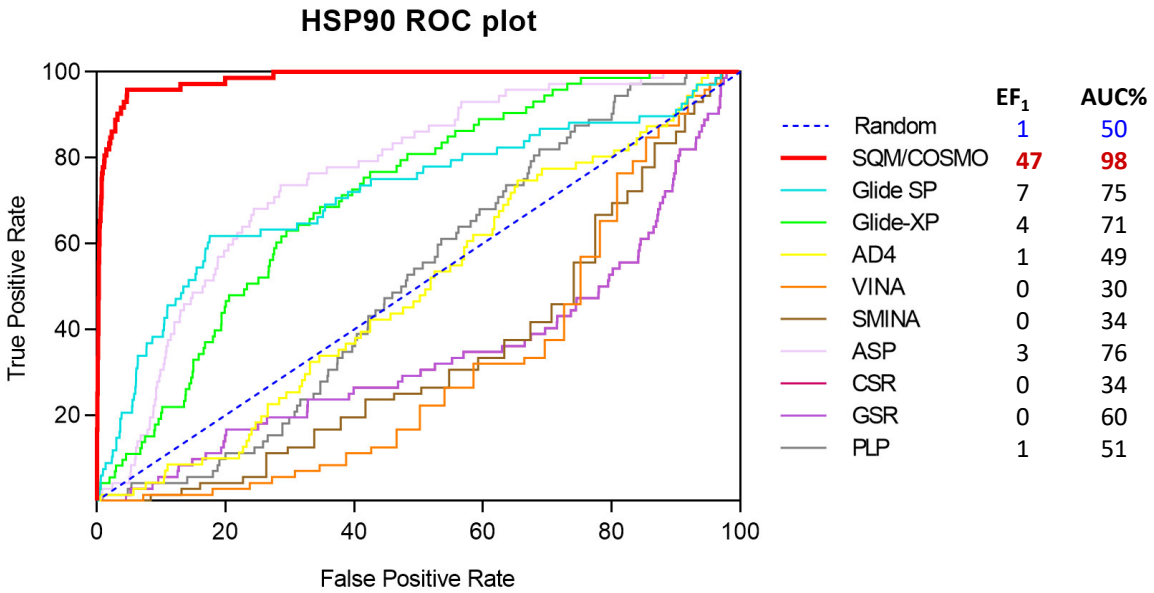
- Internal moving QM part
- Intermediate QM static part
- Outside fixed



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