Electronic Supplementary Material (ESI) for ChemComm.

SUPPLEMENTARY MATERIAL

The SQM/COSMO Filter: Reliable Native Pose Identification Based on the Quantum-Mechanical Description of Protein–Ligand Interactions and Implicit COSMO Solvation

Adam Pecina[1+], René Meier[2+], Jindřich Fanfrlík[1], Martin Lepšík[1], Jan Řezáč[1],

Pavel Hobza[1,3]* and Carsten Baldauf[4]*

[1] *Institute of Organic Chemistry and Biochemistry (IOCB) and Gilead Sciences and IOCB Research Center, Flemingovo nám. 2, 16610 Prague 6 (Czech Republic).*

[2] *Institut für Biochemie, Fakultät für Biowissenschaften, Pharmazie und Psychologie Universität Leipzig, Brüderstrasse 34, D-04109 Leipzig (Germany)*

[3] *Regional Centre of Advanced Technologies and Materials, Department of Physical Chemistry, Palacký University, 77146 Olomouc (Czech Republic)*

[4] *Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, D-14195 Berlin (Germany)*

[+] These authors have contributed equally to this work

E-mail: hobza@uochb.cas.cz, baldauf@fhi-berlin.mpg.de

Abbreviations

AChE … Acetylcholine esterase
AR … Aldose reductase
ASP … Astex Statistical Potential
ChemPLP … GOLD ChemPLP score
COSMO ... Conductor-like Screening model
CS ... Chemscore
FIRE … Fast Inertial Relaxation Engine
GAFF … general AMBER force field
GB … generalised Born implicit solvent model
GlideXP ... GlideScore Extra Precision
GS … Goldscore
HIV PR ... HIV-1 protease
IQR … interquartile range
MAD … mean absolute deviation
MM … molecular mechanics
PB … Poisson-Boltzmann implicit solvent model
PLP ... PLANTS PLP score
P-L … protein–ligand
QM … quantum mechanical
Q1 and Q3 ... the first and the third quartile
RMSD … Root-mean-square deviation
$RMSD^{max}$ … maximal root-mean-square deviation
SD … Steepest descent
SF … scoring function
SMD ... Solvation Model based on Density
SQM… semiempirical quantum mechanical
TACE … TNF-α converting enzyme
TDOF... torsional degrees of freedom
Vina ... AutoDock Vina
vdW … van der Waals

1. Methods

1.1. Protein-ligand complexes

Four unrelated protein-ligand complexes that feature difficult-to-handle noncovalent interactions were chosen for this study. These were resolved by X-ray crystallography at reasonable resolution (Table S1) and the ligand electron density was well distinguishable. The ligands are shown in Figure S1.

Table S1. Protein-ligand complexes used in this study

PDB	Reference	Resolution	Protein	Ligand	Features
1E66	[1]	2.10 Å	**AChE**	Huprine X	Two binding pockets, halogenated ligand
2IKJ	[2]	1.55 Å	**AR**	IDD393	Cofactor, halogenated ligand
3B92	[3]	2.00 Å	**TACE**	440	Metallo-protein, Zn^{2+} cation coordinated by S^-, three water molecules in binding site
1NH0	[4]	1.03 Å	**HIV PR**	KI2	Large, flexible and charged ligand, structural water molecule in binding site

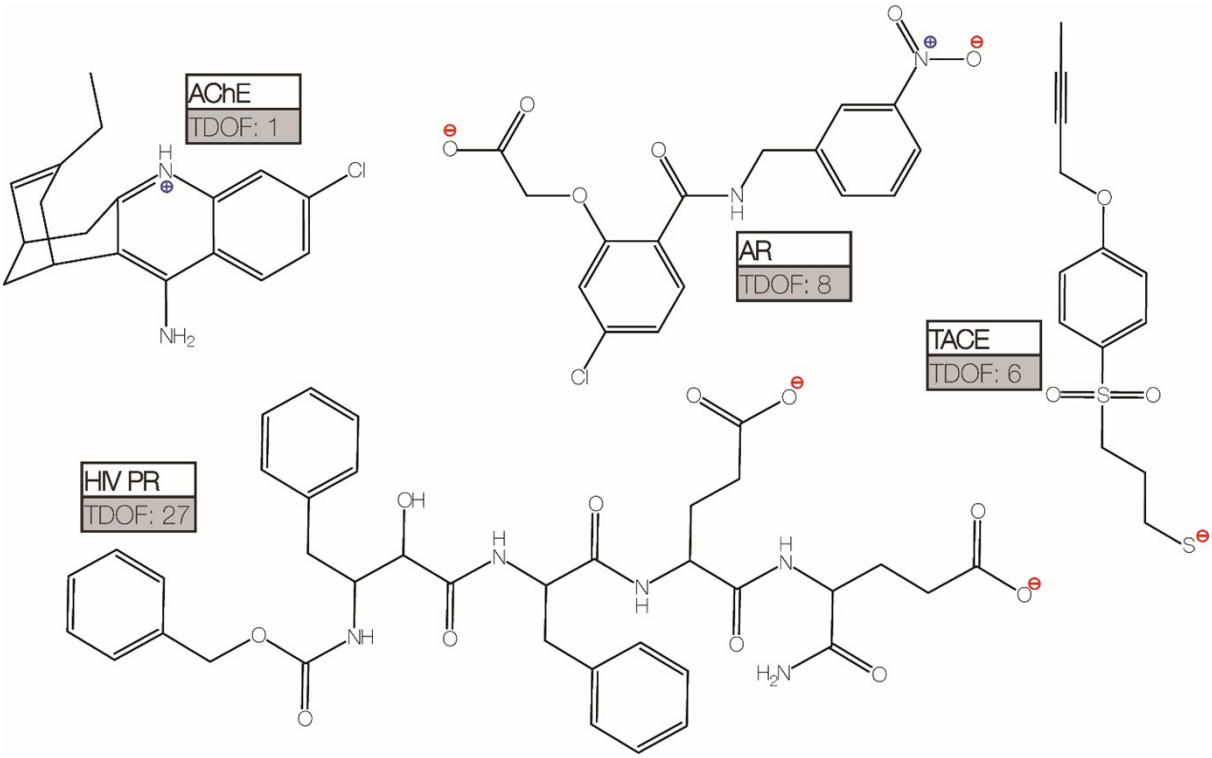


Figure S1. 2D structures of the studied ligands (labelled by their target protein, see table S1) with their charges and the numbers of torsional degrees of freedom (TDOF).

1.2. Generation of Protein-Ligand Poses via Docking

Four different docking programs with overall 7 different scoring functions (Table S2) were used to generate protein-ligand poses, the workflow is summarized in Figure S2. The individual docking runs were started from the structure of the ligand in the respective X-ray structure and in addition from up to 10 randomized ligand conformations. These starting conformations were created with the conformation search in MOE[5] with at least 2 Å RMSD between the conformations and an energy window of 7 kcal/mol using the Amber 10[6]+EHT force field.[7] For each docking run, up to 100 receptor-ligand poses were generated by each of the 7 docking setups. If the docking program supports removal of redundant results, this option was used. The hypothetical maximal number of 7,700 decoys per receptor-ligand pair was reduced by clustering with a cut-off of 0.5 Å for decoys up to 2 Å RMSD to the crystal structure and a cut-off of 2 Å for all other decoys in order to avoid redundant conformations. This yielded more than 2,800 ligand-receptor poses; exact numbers are given in Table S2.

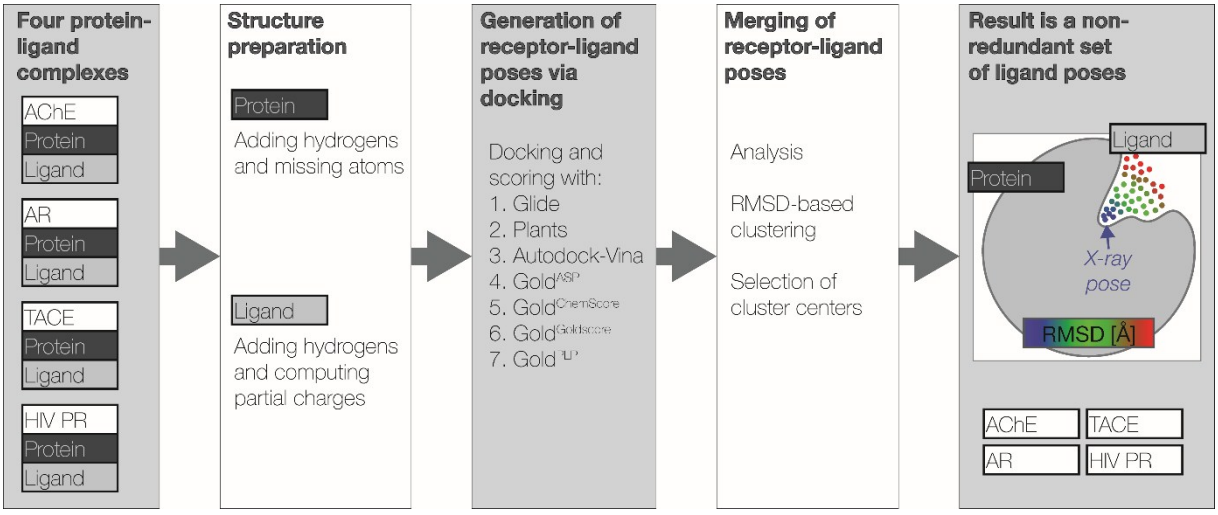


Figure S2. Schematic representation of the workflow that was used to generate sets of alternative and non-redundant binding poses of protein-ligand complexes.

Table S2. Docking protocols and numbers of generated decoy poses.

Setup	Software	Energy function	Number of generated poses			
			AChE	AR	TACE	HIV PR
1	Glide	GlideScore XP	4	19	27	38
2	PLANTS	PLANTS PLP	200	1,100	1,100	700
3	Autodock Vina	Vina	2	168	220	140
4	GOLD	ASP	200	1,100	1,100	700
5		Chemscore	200	1,100	1,100	700
6		Goldscore	200	1,100	1,100	700
7		ChemPLP	200	1,100	1,100	700
Poses after clustering		sum = 2,865	67	163	734	1901

1.3. Physics-based scoring

1.3.1. Structure preparation

Careful preparation of the protein-ligand structures was carried out as physics-based methods (AMBER/GB and SQM/COSMO) are sensitive to molecular details, e.g. protonation states and geometrical clashes generated by the docking procedures.

Ligands were prepared by adding hydrogen atoms with UCSF Chimera.[8] Force-field parameters for the ligands were taken from GAFF[9] and partial charges were derived from RESP fitting of the electrostatic potential (ESP) calculated at the AM1-BCC level.[10]

The protein structures were prepared using the Reduce[11] and LEaP programs[12] that are part of the AMBER 10 package[6]. The protonation states of histidine side chains were manually assigned based on the hydrogen-bonding patterns and pH of the crystallization conditions.

Acetylcholine esterase (AChE). For the 1E66 X-ray structure (Table S1), the carbohydrate modifications of the enzyme were not considered. Based on the experimental pH of crystallization of 5.6,[1] His471 is modeled as doubly protonated. The ligand Huprine X is protonated (charge +1, Figure S1) and forms a hydrogen bond with the backbone carbonyl of His440.

Aldose reductase (AR). The structure 2IKJ (Table S1) features the NADP cofactor (charge -3), singly protonated histidines, and a ligand with charge -1. The O1 of the inhibitor carbonyl group forms a hydrogen-bond with Nε1 of Trp111 and the O2 binds to the side-chain of His110 and Tyr48.[2] The nitrophenyl group of the inhibitor is placed in the specificity pocket of the enzyme where it forms an interaction to Leu300 NH via the nitro oxygen and a face-to-face oriented π…π stacking with the side-chain of Trp111.

TNF-α converting enzyme (TACE). It is a metallo-protease whose structure (PDB code 3B92, Table S1) features a Zn^{2+} cation that is coordinated with the inhibitor thiol moiety and the three histidine side-chains of the protein. The thiol group was modeled as thiolate (S^-) in analogy with deprotonated sulfonamide (SO_2NH^-) group that we studied earlier.[13] Three structural water molecules from the crystal structure

were considered throughout this study. Three water molecules (W524, W538, W676) were required to achieve sensible docking results. The first water molecule is bound by Ala439 and the sulfonyl group of the inhibitor, the second is bound by Glu398 and Val440 and the third is bound by Tyr436 and Ile438.

HIV-1 Protease (HIV PR). This homodimeric enzyme (Table S1) features a structural water molecule in the flap region of the active site that was considered in all the calculations. The Asp25/25' dyad is considered doubly protonated based on the crystallographic findings.[4] The Asp30 side chain is protonated on Oδ2 according to the QM calculations of protein-ligand stabilities and proton transfer barriers.[14]

1.3.2. Geometry Optimization

Hydrogen positions were subjected to steepest-descent optimization (SD) and simulated annealing with the SANDER module of the AMBER package.[6] In the *protein-ligand complexes*, the positions of the hydrogen atoms within 6 Å around the ligand position were optimized in three steps: (i) 50 optimization steps using SD, (ii) simulated annealing for this part of the protein/ligand complex, (iii) optimization of hydrogen positions with the FIRE algorithm. For poses with close contacts between ligand and protein below 1.5 Å, 50 SD optimization steps of the ligand embedded in the fixed protein were performed.

1.3.3. Scoring

In the Pavel Hobza's group, we have been developing an SQM scoring function[15] which correctly describes all types of noncovalent interactions, viz. dispersion, hydrogen and halogen-bonding. We have demonstrated its applicability for various protein-ligand systems, such as protein kinases, aldose reductase, HIV-1 protease and carbonic anhydrase.[13-16] As a special case, we have also extended it to treat covalent inhibitor binding.[17] Recently, there have been several attempts to make QM methods applicable in virtual screening, especially by their acceleration and simplification.[18]

1.3.4 SQM region

To make the calculations faster, we defined a sphere of 8 to 12 Å (roughly 2,000 atoms) around the aligned ligand poses as a region representing the binding site.

This region was treated by SQM and was the same for all the poses. These truncated systems (SQM/COSMO filter) were compared with full-sized systems (full SQM/COSMO) and shown that they behaved nearly identically (see later, Figure S4).

1.4. Score Normalisation

The calculated scores are on different scales and thus are not straightforwardly comparable. In order to generate comparable numbers, they were converted to a normalised score. For each data set, i.e. all poses of a protein-ligand complex ranked by a scoring/energy function, the first quartile (Q_1) and the third quartile (Q_3) were calculated. The interquartile range (IQR) is defined as:

$$IQR = Q_3 - Q_1$$

All poses with energies greater than $Q_3 + 1.5\,IQR$ were considered as high energy outliers and were removed from the dataset. Finally, the relative energies of poses with respect to the energy of the X-ray pose were scaled with a factor F:

$$F = \frac{100}{(Q_3 + 1.5\,IQR) - (Q_1 - 1.5\,IQR)}$$

The resulting normalised scores are comparable between the different energy functions.

1.5. RMSD Measurements

For all the ligand poses generated, the distances in Cartesian space (root-mean square deviation, RMSD) from the X-ray structure were determined. The RMSD values were calculated without considering hydrogens. The algorithm takes full molecule symmetry into account, based on a graph depth-first-search[5] and atom priorities following the Cahn-Ingold-Prelog rules. All RMSD values were calculated without superposition so that the resulting values truly express a distance in the multi-dimensional energy landscape.

1.6. Normalised Scores vs. RMSD

The energy values of every pose were plotted *vs.* the RMSD value to the crystal structure. The cloud of points (see Appendix of the SI) was further simplified to a single graph by showing only the lower boundary of all energies with respect to RMSD from the X-ray structure (Figure S3). The graph was constructed by removing

all data points above a point if the connecting line would have a |slope| > 12 starting with the lowest energy data point. This was repeated on all points in the order of increasing energy until the whole data set was processed. The remaining points were connected with lines.

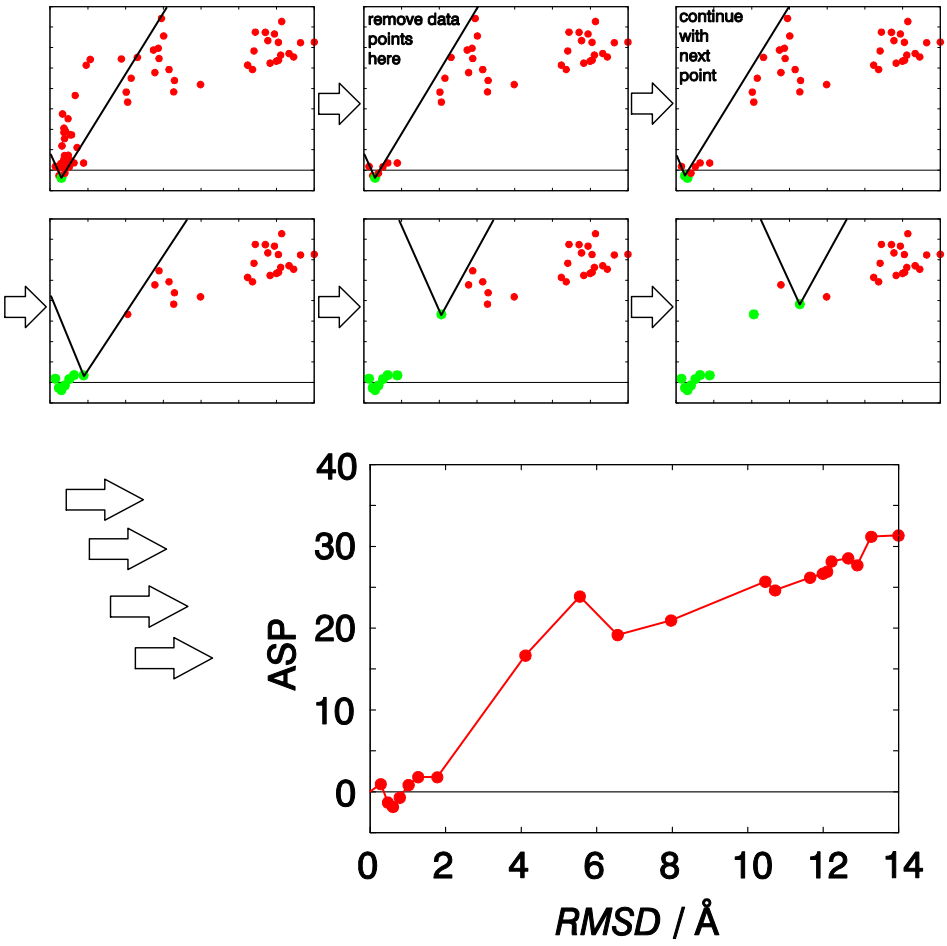


Figure S3. Scheme of the algorithm to create the lower boundary from the whole data set. An iterative process reduces the large amount of data points to the most important points for this study.

2. Results

2.1. Convergence of SQM region size

In all four systems we compared the influence of applying the truncation scheme to covering the full protein-ligand complex in a SQM calculation. Table S3 shows gives the mean absolute deviation (MAD). The MAD values of up to 4 kcal/mol are, however, not visible in the overall shape of the lower-bound representation of the binding energy landscape (see Figure 2). The results of SQM/COSMO filter and full SQM/COSMO are in good agreement (Figure S4). The use of SQM/COSMO filter can thus be recommended for use due its speed.

Table S3. Mean absolute deviations (MAD/kcal.mol^{-1}) between SQM and full-SQM energy approaches.

Protein	AChE	AR	TACE	HIV PR
Overall atoms	8,388	5,160	4,064	3,230
Atoms in SQM region	1,843	1,960	1,489	2,200
MAD / kcal/mol	2.8±1.6	2.5±0.8	0.5±0.9	3.9±0.8

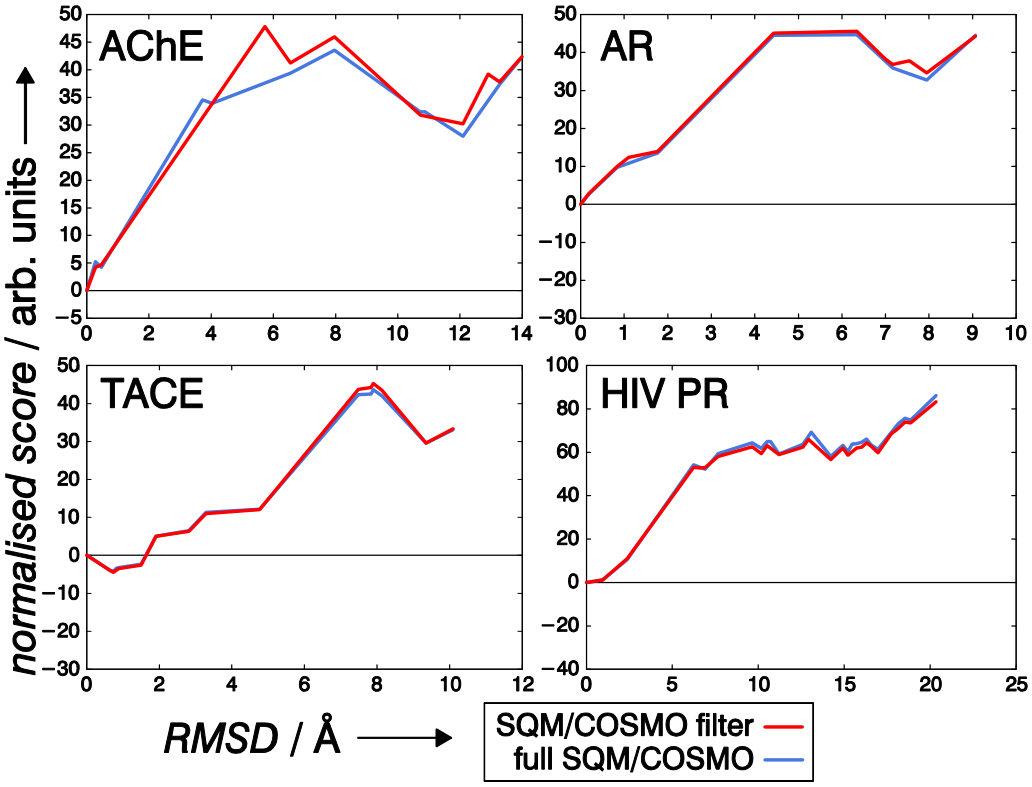


Figure S4. Comparison of the full-size SQM/COSMO vs. SQM/COSMO filter plots

2. 2. Quality Criterion – RMSD^{max}

Here, were present the results of of second criterion, RMSD^{max}, for larger score windows of 10 and 20 (Table S4). In the former, 3 scoring functions (SQM/COSMO, AMBER/GB and Gold CS) recognized the correct binding pose (RMSD < 2 Å). The SQM/COSMO showed the lowest RMSD^{max} (1.32 Å). In the score window of 20, no scoring function met the limit of 2 Å. However, AMBER/GB and SQM/COSMO were close (RMSD^{max} of 2.04 and 2.49 Å).

Table S4. Behaviour of the scoring function within normalised scores up to 10 and 20

	Scoring function								
			Glide	PLANTS	AutoDock			Gold	
	SQM/COSMO	AMBER/GB	XP	PLP	Vina	ASP	CS	GS	ChemPLP
	Maximal RMSD within a window of 10 of the normalised Score								
AchE	0.63	1.01	2.13	2.13	1.01	1.78	1.78	1.14	1.01
AR	0.84	0.19	7.54	3.47	3.54	2.59	1.77	7.66	1.81
TACE	2.81	4.76	3.13	2.91	8.06	2.86	2.63	2.44	2.73
HIV PR	1.01	0.94	17.74	13.13	11.62	1.00	1.08	14.20	12.64
Average	**1.32**	**1.62**	**7.64**	**5.41**	**6.06**	**2.06**	**1.81**	**6.36**	**4.55**
	Maximal RMSD within a window of 20 of the normalised Score								
AChE	1.06	1.14	11.99	4.11	19.85	7.97	6.58	5.55	1.43
AR	1.77	1.16	9.06	7.79	9.75	3.90	2.32	8.18	3.54
TACE	2.37	1.10	18.22	16.51	12.60	1.94	1.93	16.90	14.20
HIV PR	4.76	4.76	3.13	2.91	9.59	7.41	2.63	6.98	7.13
Average	**2.49**	**2.04**	**10.60**	**7.83**	**12.95**	**5.31**	**3.37**	**9.40**	**6.58**

3. References

[1] H. Dvir, D. M. Wong, M. Harel, X. Barril, M. Orozco, F. J. Luque, D. Munoz-Torrero, P. Camps, T. L. Rosenberry, I. Silman et al., *Biochemistry* **2002**, *41*, 2970–2981.

[2] H. Steuber, A. Heine, G. Klebe, *J. Mol. Biol.* **2007**, *368*, 618–638.

[3] U. K. Bandarage, T. Wang, J. H. Come, E. Perola, Y. Wei, B. G. Rao, *Bioorg. Med. Chem. Lett.* **2008**, *18*, 44–48

[4] J. Brynda, P. Řezáčová, M. Fábry, M. Hořejší, R. Štouračová, J. Sedláček, M. Souček, M. Hradílek, M. Lepšík, J. Konvalinka, *J. Med. Chem.* **2004**, *47*, 2030–2036.

[5] Chemical Computing Group Inc., Molecular Operating Environment (MOE) 2013.08, 1010 Sherbooke St. West, Suite #910, Montreal, QC(Canada), **2015**.

[6] D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, M. Crowley, R.C. Walker, W. Zhang, K.M. Merz, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D.H. Mathews, M.G. Seetin, C. Sagui, V. Babin, and P.A. Kollman, AMBER 10, University of California, San Francisco, **2008**.

[7] P. R. Gerber, K. Muller, *J. Comput.Aided Mol. Des.* **1995**, *9*, 251–268.

[8] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, *J. Comput. Chem.*, **2004**, *25*, 1605–1612.

[9] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.*, **2004**, *25*, 1157–1174.

[10] a) A. Jakalian, B. L. Bush, D. B. Jack, C. I. Bayly, *J. Comput. Chem.* **2000**, *21*, 132–146; b) A. Jakalian, D. B. Jack, C. I. Bayly, *J. Comput. Chem.* **2002**, *23*, 1623–1641.

[11] J. M. Word, S. C. Lovell, J. S. Richardson, D. C. Richardson, *J. Mol. Biol.* **1999**, *285*, 1735–1747.

[12] C. E. A. F. Schafmeister, W. S. Ross, V. Romanovski, LEAP, University of California, San Francisco, **1995**.

[13] A. Pecina, M. Lepšík, J. Řezáč, J. Brynda, P. Mader, P. Řezáčová, P. Hobza, J. Fanfrlík, *J. Phys. Chem. B* **2013**, *117*, 16096-16104

[14] A. Pecina, O. Přenosil, J. Fanfrlík, J. Řezáč, J. Granatier, P. Hobza, M. Lepšík, *Collect. Czech. Chem. Commun.* **2011**, *76*, 457-479.

[15] a) J. Fanfrlík, A. K. Bronowska, J. Řezáč, O. Přenosil, J. Konvalinka, P. Hobza, *J. Phys. Chem. B* **2010**, *114*, 12666-12678, b) M. Lepšík, J. Řezáč, M. Kolář, A. Pecina, P. Hobza, J. Fanfrlík, *ChemPlusChem* **2013**, *78*, 921-931.

[16] a) P. Dobeš, J. Řezáč, J. Fanfrlík, M. Otyepka, P. Hobza, *J. Phys. Chem. B* **2011**, *115*, 8581-8589; b) J. Fanfrlík, M. Kolář, M. Kamlar, D. Hurn, F. X. Ruiz, A. Cousido-Siah, A. Mitschler, J. Řezáč, E. Munusamy, E.; M. Lepšík, et al., *ACS Chem. Biol.* **2013**, *8*, 2484-2492; c) J. Fanfrlík, F. X. Ruiz, A. Kadlčíková, J. Řezáč, A. Cousido-Siah, A. Mitschler, S. Haldar, M. Lepšík, M. H. Kolář, P. Majer, et al., *ACS Chem. Biol.* **2015**, *10*, 1637-1642.

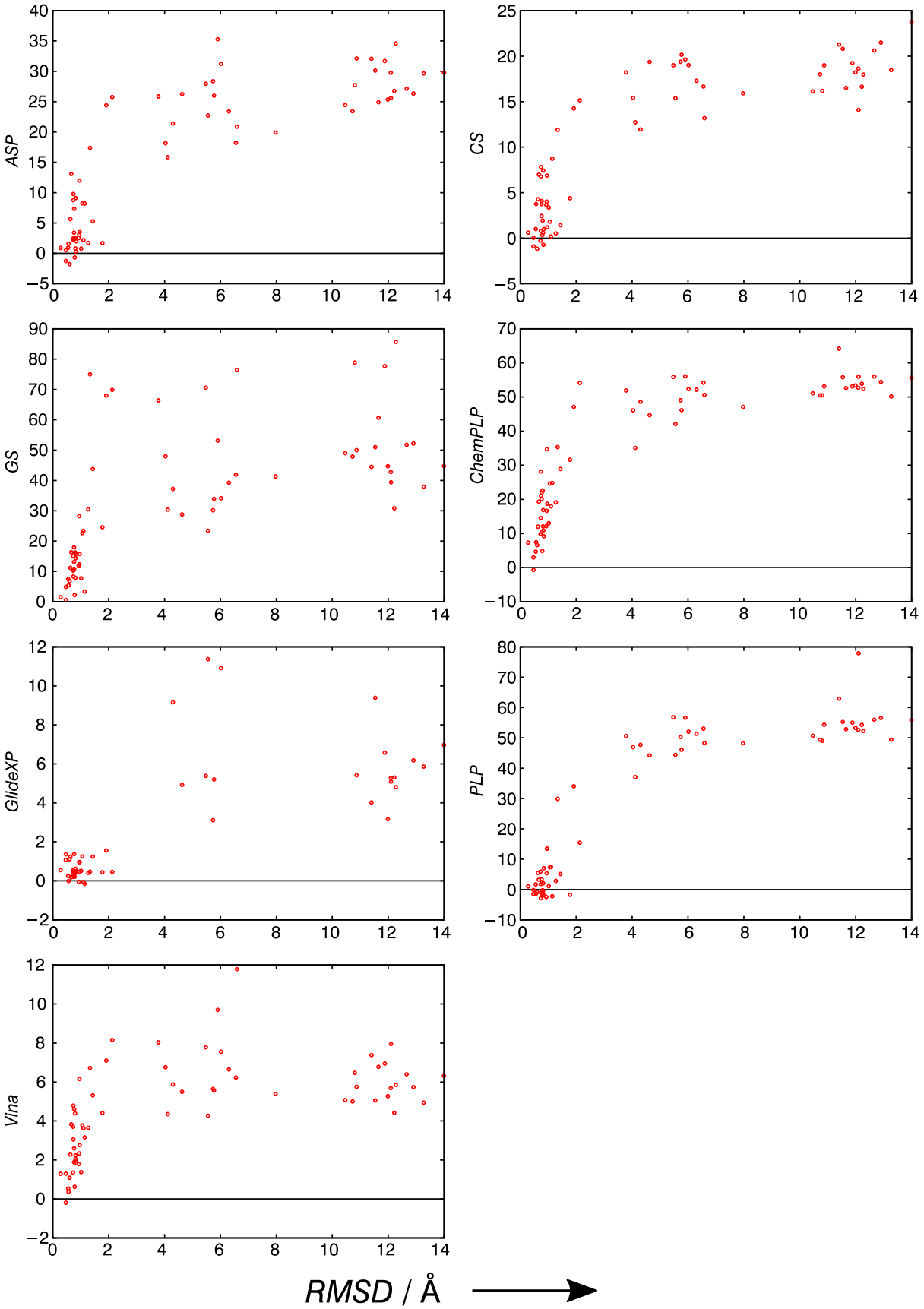
[17] J. Fanfrlík, P. S. Brahmkshatriya, J. Řezáč, A. Jílková, M. Horn, M. Mareš, P. Hobza, M. Lepšík, *J. Phys. Chem. B* **2013**, *117*, 14973-14982.

[18] a) K. Wichapong, A. Rohe, C. Platzer, I. Slynko, F. Erdmann, M. Schmidt, W. Sippl, *J. Chem. Inf. Model.* **2014**, *54*, 881–893; b) P. Chaskar, V. Zoete, U. F. Röhrig, *J. Chem. Inf. Model.* **2014**, *54*, 3137–3152; c) S. K. Burger, D. C. Thompson, P. W. Ayers, *J. Chem. Inf. Model* **2011**, *51*, 93–101.

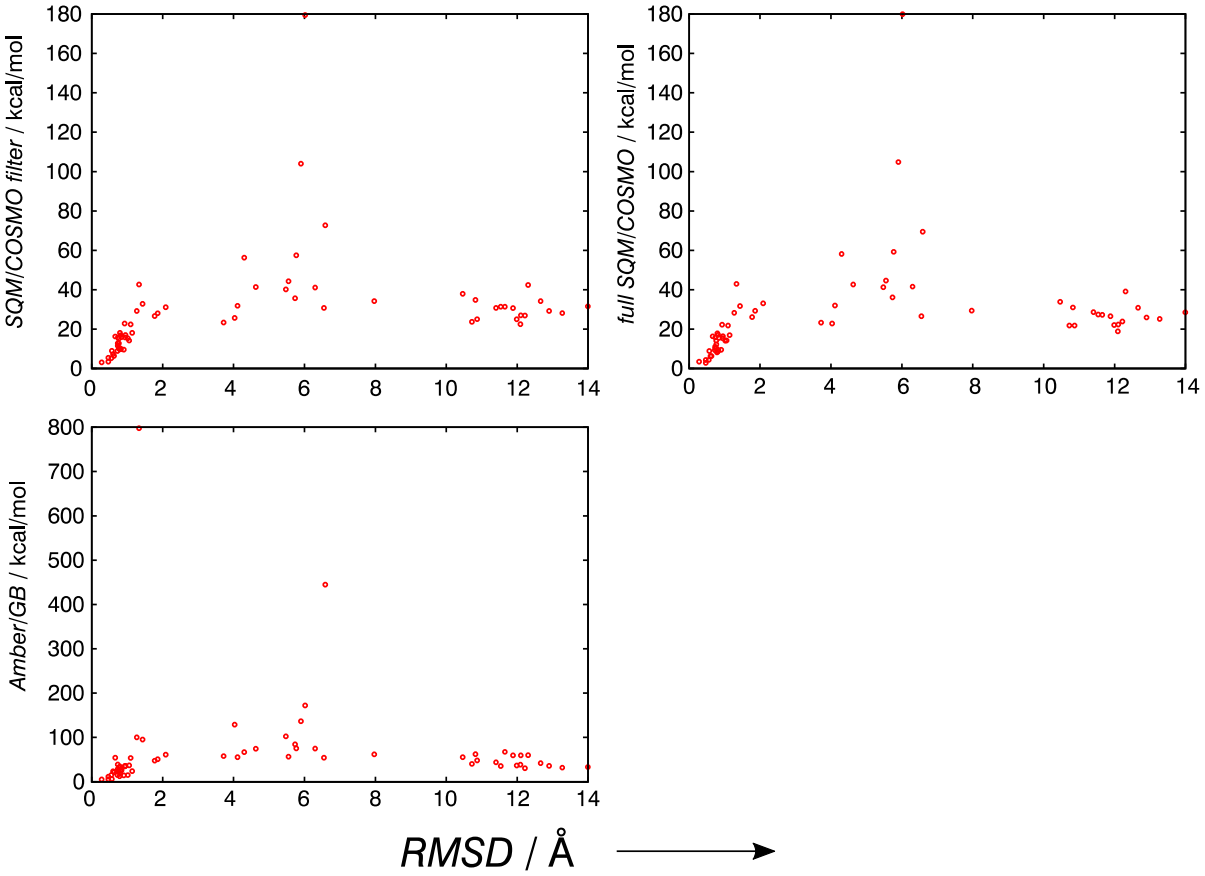
4. Appendix

Raw energy and score values plotted against RMSD values for the tested scoring functions for all poses of AChE, AR, TACE and HIV PR.

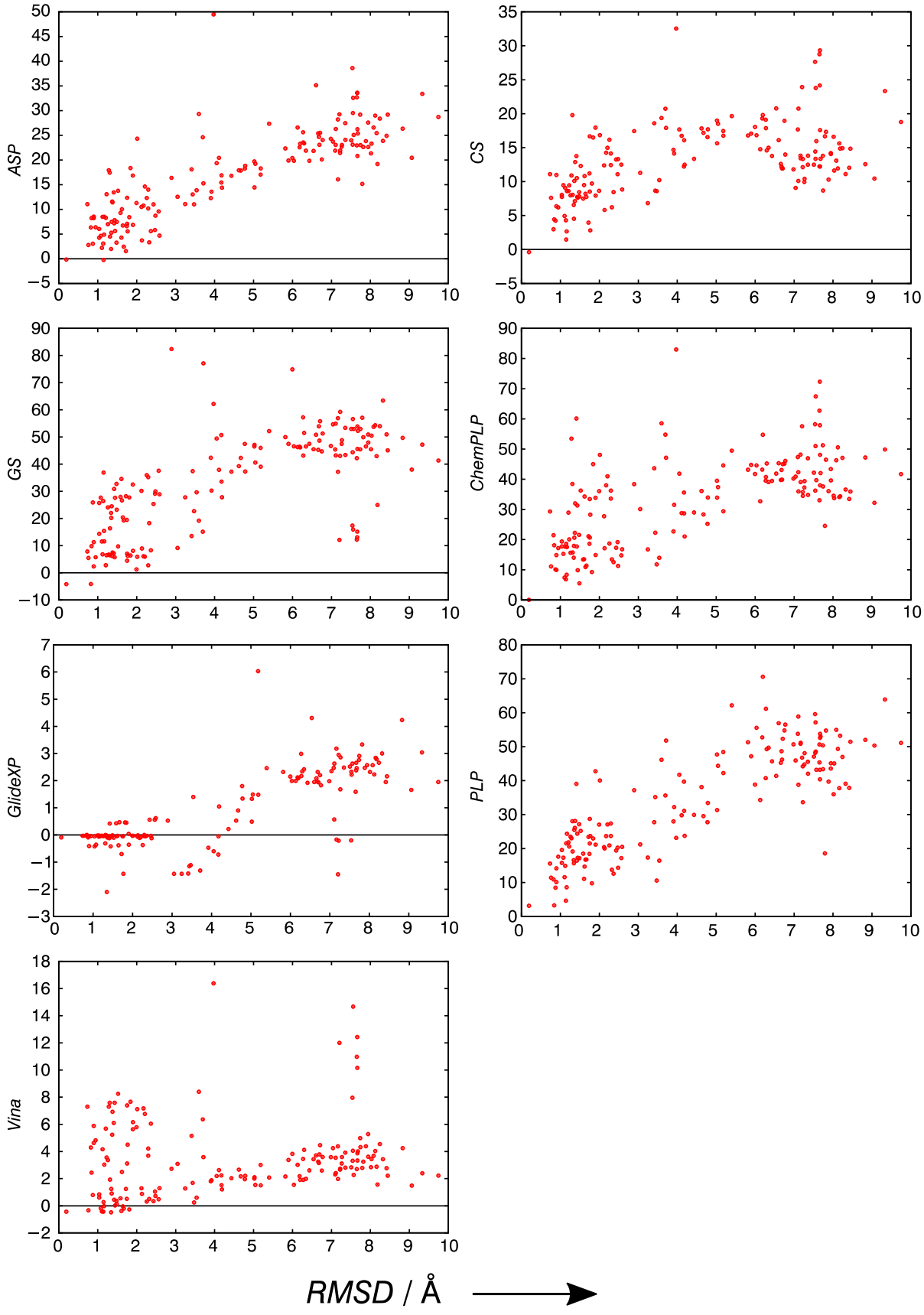
A1: Raw scores and energies for **AChE**.



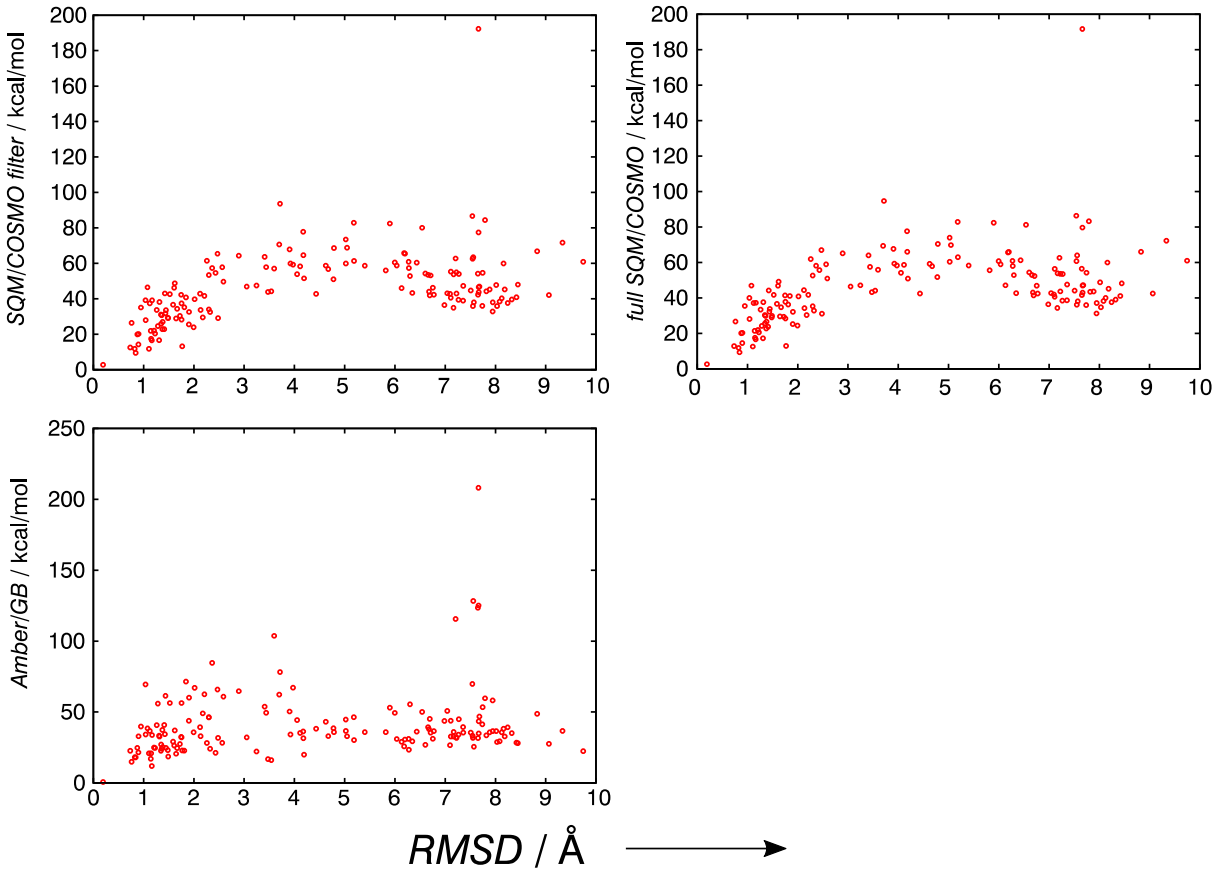
A1 continued: Raw scores and energies for **AChE**.



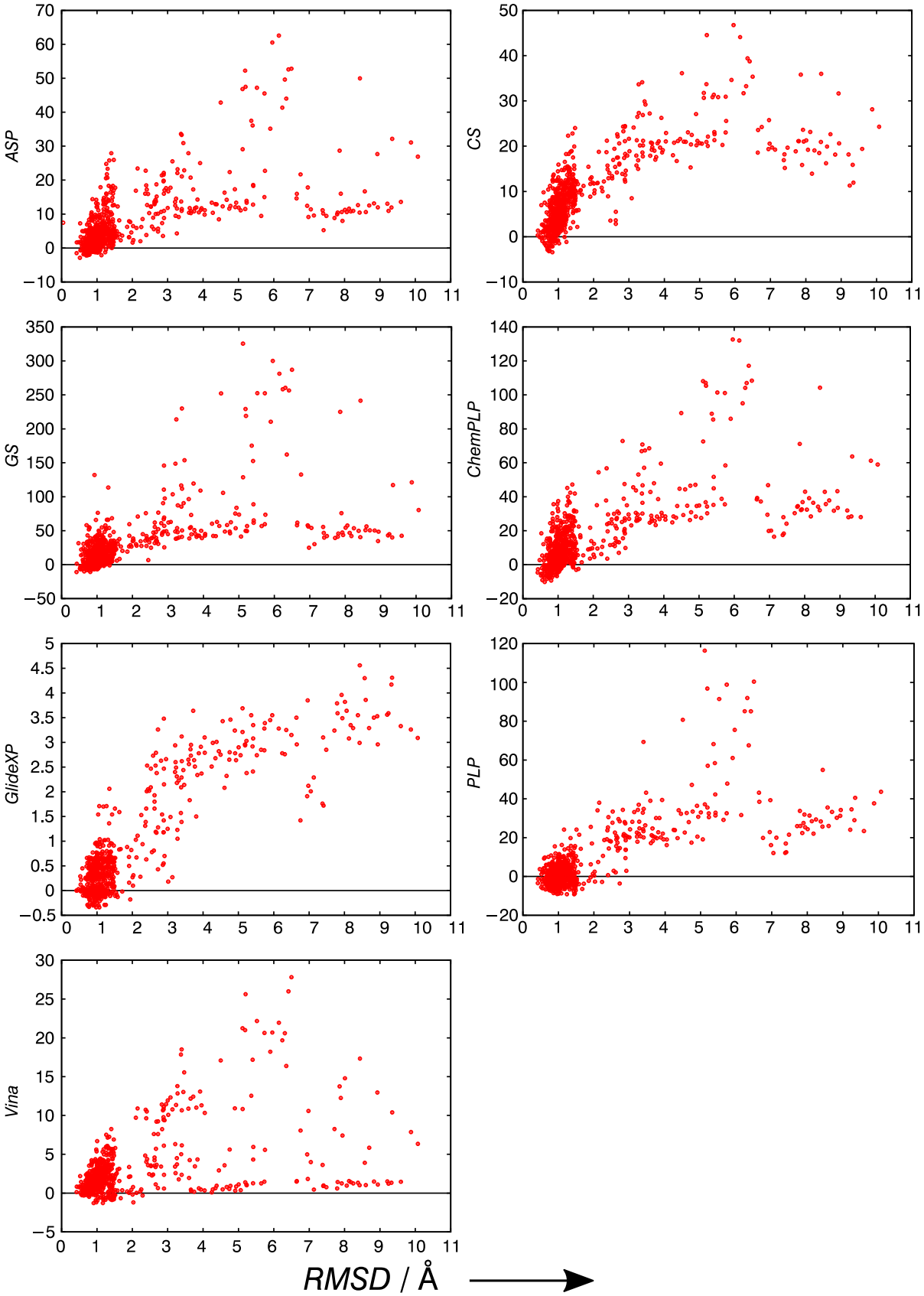
A2: Raw scores for **AR**.



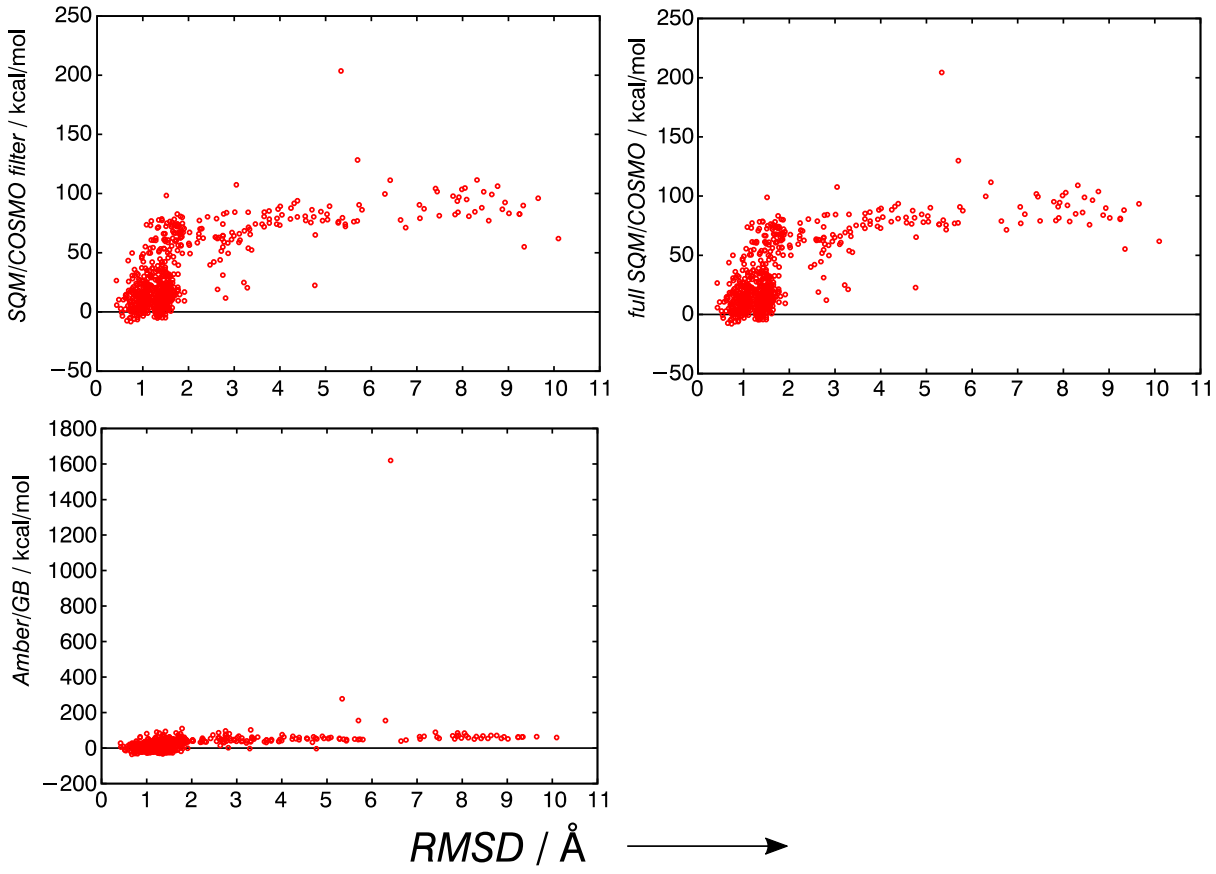
A2 continued: Raw scores and energies for **AR** continued.



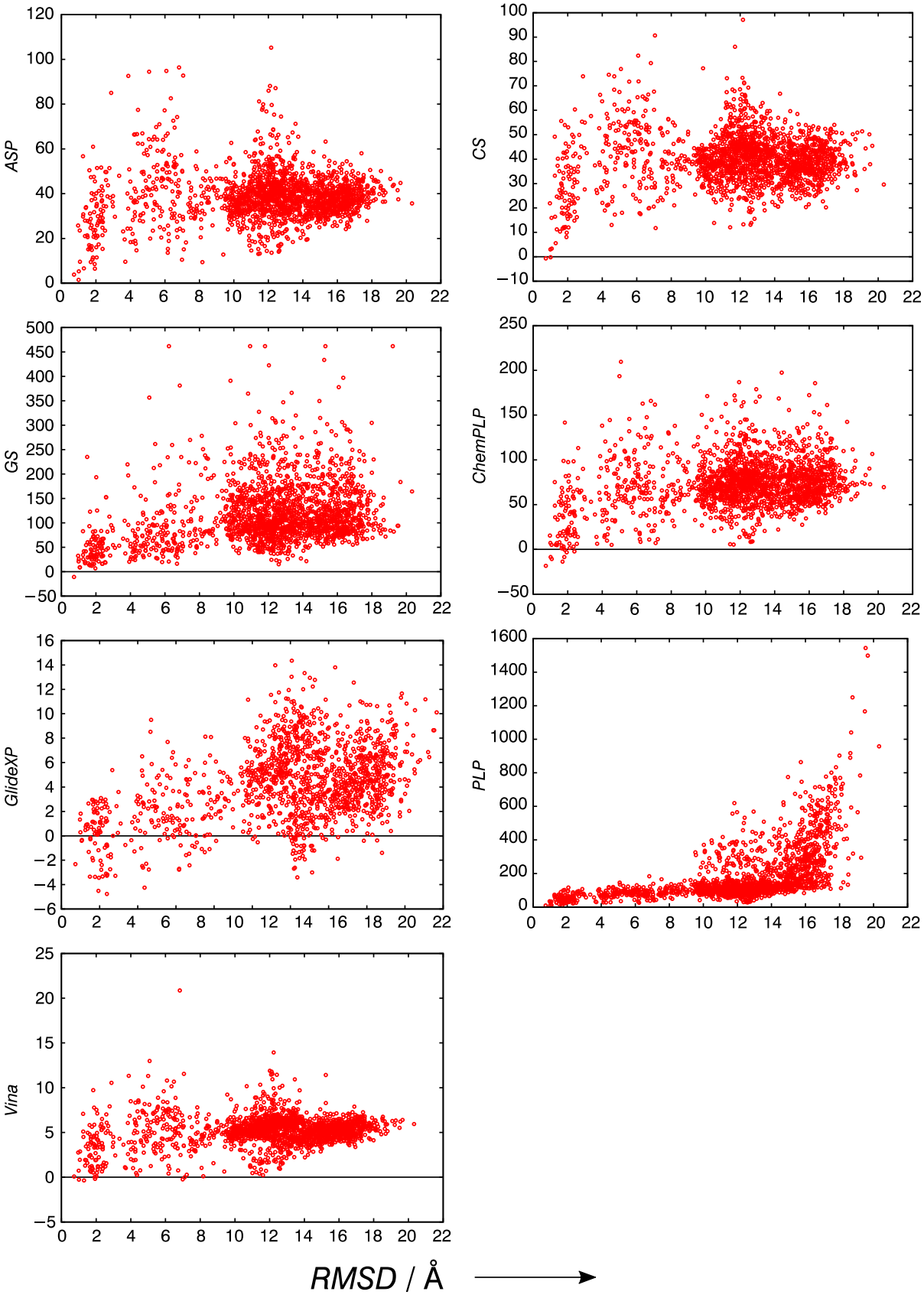
A3: Raw scores and energies for **TACE**.



A3 continued: Raw scores and energies for **TACE** continued.



A4: Raw scores for **HIV PR**.



A4 continued: Raw scores and energies for **HIV PR** continued.