Univerzita Karlova

Přírodovědecká fakulta

Studijní program: Fyzikální chemie

Mgr. Kristian Kříž

Optimalizace semiempirických kvantově mechanických metod pro návrh léčiv *in silico*

Optimization of Semiempirical Quantum Mechanical Methods for *in Silico* Drug Design

Disertační práce

Školitel: doc. RNDr. Jan Řezáč, Ph.D.

Konzultant: RNDr. Martin Lepšík, Ph.D.

Praha, 2020

Charles University

Faculty of Science

Study programme: Physical Chemistry

Mgr. Kristian Kříž

Optimization of Semiempirical Quantum Mechanical Methods for *in Silico* Drug Design

Optimalizace semiempirických kvantově mechanických metod pro návrh léčiv *in silico*

Doctoral Thesis

Supervisor: doc. RNDr. Jan Řezáč, Ph.D.

Konzultant: RNDr. Martin Lepšík, Ph.D.

Prague, 2020

Declaration of the Author

I declare that I have worked out this thesis by myself using the cited references. Neither the thesis nor its parts were used previously for obtaining any academic degree.

Prague, September 24, 2020

Contents

1 Introduction

1.1 Quantum Chemistry

The quantum physical theory has made the quantitative understanding of processes on submicroscopic level possible. This understanding in turn enhanced greatly possibilities of rational approaches to design materials for technology and drugs for medicine. Quantum chemistry is the application of quantum mechanics (QM) on the system of interest in chemistry [1], and also in biology, medicine or material sciences [2]. Quantum chemistry seeks the energy of a system either by solving Schrödinger equation, that is applying a Hamilton operator to a wave function of a system – as in case of wave-function based methods (such as Hartree-Fock, MP2, CCSD(T) – section 3.1), or by using electron probability density instead of the wave function – as in case of density functional theory (DFT) methods (section 3.2). The first quantum chemical methods formulated, wave-function methods based on purely theoretical grounds are called *ab initio* methods [1]. DFT is usually also considered an *ab initio* method, given the computational demands and accuracy. These methods require progressively more time as the system being solved grows larger. For example, Hartree-Fock method, a cornerstone of quantum chemistry, scales with the fourth power of the system size [3]. Even though powerful computers are utilized nowadays, the evaluation of properties of larger molecular complexes is still time consuming [4].

1.2 Semiempirical Quantum Mechanical Methods (SQM)

SQM methods are approximations to QM (quantum mechanical) methods [3]. By invoking simplifications or neglects of equations in places where the most time can be saved for the least accuracy detriment, they offer a compromise

solution for larger systems. While less accurate, they retain their quantum character and as such they can still provide some description of phenomena such as charge transfer [3, 5], often important for binding (of protein and ligand, for example). Although they contain parameters, once they are determined, the method works universally (given all parameters for elements present in the system studied are provided). This is in contrast with molecular mechanics (MM) methods, which require an independent parametrization for each specific system [6, 7]. These properties along with the computational time savings make SQM methods suitable for use in *in silico* drug design [8], as will be discussed in section 1.6.

Two SQM approaches most prominently used today are those based on molecular orbital theory – Hartree-Fock (HF) derived methods, such as PM6 and recent PM7, and those based on density functional theory (DFT) – DFTB density-functional tight-binding methods [3]. These approaches are described in detail in sections 3.1 and 3.2 respectively.

1.3 Corrections for Non-covalent Interactions

To adjust for the loss of accuracy of SQM methods caused by the simplifications they apply (or to improve the performance of less advanced QM methods), corrections can be added to adjust some observed error trends of a method in question. Corrections for dispersion are illustrative examples. It has been found out that neither the HF nor DFT account correctly for the dispersion interaction [9]. Empirical function based on London‘s formula has been used to reproduce dissociation curves of noble gas complexes, where dispersion is important. It has terms inversely proportional to the sixth, eight, tenth, etc. power of the separation of interacting centers [10, 11]. General form of this kind of dispersion correction to the energy, which can be written as

$$\sum_{AB} \sum_{n=6,8,..} s_n f_{R_{AB}}^{damp} \frac{C_{AB,n}}{R_{AB}^n} \qquad \text{(EQ.1)}$$

adds an additive dispersion correction contribution from a multipole n to the calculated energy for each pair of atoms A and B separated by distance R^{AB}. f_{damp} is a damping function which turns the correction off for small distances and s is a global scaling parameter for a given method (functional) [12]. Among the most used is the Grimme‘s correction proposed originally for DFT [13], D3. DFT-D3 includes terms for n=6, 8 with parameters in f^{damp} and coordination number dependent C coefficients determined by higher-level methods (time dependent DFT) [14]. In the most recent version D4, the atomic pairwise C coefficients were made dependent on atomic partial charges by means of a function multiplying atomic polarizabilities (used for the determination of C in these models). Also, three-body dispersion term addressing nonadditivity is utilized by default unlike in the model‘s predecessors [15]. Thus a corrective formula is generally built on a rational basis (London dispersion in this case) but contains parameters that ensure the best results (compared to a reference) for a method being corrected [13, 14]. The widespread use of these dispersion-corrected DFT functionals proves the general usefulness of empirical corrections, even for higher-level computational methods.

SQM methods, being simplified QM methods, also share the drawbacks of their respective QM templates, and so they also require adjusting for the dispersion interaction. Empirical dispersion correction containing the term decaying with the sixth power of the atomic separation and with a damping function as in equation EQ.1 was proposed first for DFTB [16]. The C_{AB} were calculated from empirically determined atomic polarizabilities and numbers of valence electrons. Likewise for wave-function-based MNDO – AM1 and PM3 methods, the polarizability (calculated by these methods) dependent dispersion correction has been

supplied [17]. For PM6, the dispersion correction based on the work of Jurečka et. al. [18] parametrized together with the hydrogen bond correction (see in the further text) PM6-DH has been developed in our group. The parameters C_{AB} were originally taken from ref. [13] and the damping function modified (so that the s_n scaling factor scales van der Waals radii in the damping function rather than the whole function). Recent versions of the these corrections parametrized for use with PM6 and DFTB3 methods (section 3.3.1) are widely used.

It has been found out that in addition to the missing dispersion the MNDO methods failed at the description of hydrogen bonds [19]. Although hydrogen atom was supplied by additional Gaussian functions to its core-core term in AM1 [20] and the term was modified also in PM6 [19], the methods still yielded inaccurate results for hydrogen bonded complexes [21].

The introduction of an additional relatively simple hydrogen correcting empirical term to the SQM method can enhance its accuracy, so it rivals QM methods [22, 23].

The first of hydrogen bond corrections for SQM methods developed in our group has been parametrized for the PM6 method along with the D correction (PM6-DH) [21]. It added an independent energy correction to the several kinds of interactions involving hydrogen and partners Y (X-H $\cdots$ Y)

$$a\left[\frac{q_H q_Y}{r_{HY}^2} cos(\theta_{XHY}) + bA^{-r_{HY}}\right] \qquad \text{(EQ.2)}$$

with a,b and A an interaction specific parameters. The first term is an electrostatic component based on partial atomic charges q, scaled by an angular part reproducing hydrogen bond directionality. The second term is a corrective repulsive exponential. Subsequent H2 correction paired with D correction adjusted the D for a SQM method and addressed several interaction-specific issues [24]. The third variant of the correction H+ abandoned partial charges,

for which it would be demanding to calculate derivatives and has a form

$$\frac{C_X + C_Y}{2r_{XY}^2} \times f_{angular} \times f_{bond}(r_{ZH}) \times f_{damp}(r_{XY}) \qquad \text{(EQ.3)}$$

with atom specific parameters C, $f_{angular}$ containing multiple relevant angles, and two more empirical functions of the interacting atoms separation with Z being the closer atom of X and Y to the hydrogen. Most recent versions of these corrections are H4 (D3H4) and H5 (D3H5), with the latter being specific for DFTB3. These are discussed in *Methods* in sections 3.3.1 and 3.3.3, respectively.

Halogen bond is a noncovalent interaction, in which an electrophilic (positive) region on a halogen (X) interacts with a nucleophilic (negative) region on an interacting partner (Y) in an arrangement Z-X $\cdots$ Y. The anisotropy in charge distribution of the halogen giving rise to the positive region is caused by an influence of a Z atom, to which it is covalently bonded.

Halogen bond can be important for protein-ligand binding, since common ligands are often halogenated [7, 25]. SQM methods tend to overestimate this interaction [26]. For instance, halogen bond corrections X have been developed for PM6 (for example in PM6-D3H4X), supplying a separation dependent exponential repulsion to the energy [27].

1.4 Solvation Modelling Approaches for SQM Methods

If the SQM methods are to reliably describe biological systems (such as protein-ligand complexes – relevant for drug design purposes), the complex environment of the system has to be considered. The first and likely the most important step to model this complexity is the inclusion of solvent – water.

The inclusion of water molecules to be calculated explicitly along with a protein and a ligand would become a bottleneck for *in silico* drug design which requires fast evaluation of binding properties – by SQM or still more approx-

imate methods (section 1.6). Rather than representing solvent as individual molecules, solvation models treating the solvent as a continuum are employed with SQM methods for effective description of larger systems. The solvation (free) energy ΔG_{sol} is then the change in energy that accompanies the transfer of a compound from vacuum to water, the difference of energies in vacuum and in solvent. The energy in vacuum can be provided by electronic structure method and the solvation energy is supplied by the solvation model, which is used in conjunction with the electronic structure method. The influence of the solvent can be separated into polar and nonpolar terms. The latter traditionally features dispersion, cavitation and repulsion energy. There are many approaches for expressing each of these contributions [28]. An example of a simple expression for a combined cavitation and dispersion term is

$$G_{cav} + G_{vdW} = \sum_a \sigma_a S_a \qquad \text{(EQ.4)}$$

the sum of solvent-accessible surface of each atom type multiplied by an atom-specific constant. Among approaches broadly used today accounting for solvent polar influence are apparent surface charge (ASC) methods, multipole expansion and generalized Born methods (GB). In generalized Born methods, solute is represented by point charges i and j at the nuclei and the electrostatic contribution is based on

$$G_{el} = \left(1 - \frac{1}{\epsilon}\right) \sum_i \sum_j \frac{q_i q_j}{f_{GB}} \qquad \text{(EQ.5)}$$

where the f_{GB} is an effective function of a separation between charges and atomic radii. The nonpolar contribution is traditionally expressed as EQ.4 in GB model [29]. Multipole expansion models extend GB approach from point charges (monopoles) to higher-order (multipoles).

In ASC, the potential due to the solute polarization of the solvent is a potential at the point r given by a number of charges Q placed at the surface of cavity formed by the solute

$$V_\sigma(r) = \sum_a \frac{Q_a}{|r - r_a|} \tag{EQ.6}$$

Where the apparent surface charge Q is calculated from the overall potential in the cavity as a sum of a solute potential V_{solute} given by the electronic structure of a solute calculated by a method of choice and a potential of polarized solvent in the cavity $V_{inside,\sigma}$

$$Q_a = \frac{\epsilon - 1}{4\pi\epsilon} S_a \nabla (V_{solute} + V_{inside,\sigma}) \tag{EQ.7}$$

Where S_a is a surface area ascribed to the charge a, ϵ is the permittivity of the medium. As the potential depends on its own value in the cavity, the evaluation is iterative [1, 28]. Among most broadly used models to use with SQM are PCM, COSMO and SMD.

1.5 Parametrization of SQM Methods

As was alluded to in previous sections, parameters in the empirical corrections can be derived by fitting the proposed mathematical form, correction function, to reference data (benchmark). The set of data for fitting a quantity, such as energy, consists of series of molecular geometries with corresponding benchmark quantity (such as energy) values and is thus called benchmark dataset. Benchmark dataset used for this purpose of obtaining parameters for a method or its corrections is a "training dataset" (method/correction is being trained). Prior to the parametrization, data sets can be used to study systematic deviations among methods, so that the proper form of a correcting term can be designed,

or to test the performance of methods.

Obviously, having the best possible reference benchmark for the parametrization of correction forms is highly advantageous. In case when the experimental data is unavailable or unreliable, high-level computations are often resorted to. An interesting aspect of the parametrization against a reference is that parameters absorb all the difference between the reference data and the output of method being corrected, including high-level energy corrections, irrespective of a physical basis of the difference.

The benchmark dataset should be large and diverse, so that various trends in deviations of methods can be studied, captured and corrected for. If the dataset consisted only of a limited variety of system kinds, not all the subtle physical phenomena would be traceable and the outcomes may not necessarily be transferable to other systems. In practice, it is necessary only to cover in the dataset the kind of systems, that the methods will be used on. For use of SQM or QM methods in *in silico* drug design, those are systems featuring proteins, their ligands, nucleic acids or other biomolecules but most often protein-ligand complexes. The vast majority of medically relevant proteins consist of 20 amino acids, most often built from only 5 elements (H, C, N, O, S). However, due to their polymer nature, their 3-dimensional structure is very diverse. Although low-molecular-weight drugs that bind to them are chemically more variable, there are general patterns of their composition, too.

1.6 *In Silico* Drug Design

In the field of drug design, the aim is to find and optimize a compound (drug) that interacts with the target biomolecule, thus curing or at least treating a disease. *In silico* drug design uses computers to achieve this goal. There is a great diversity of approaches and they can be combined in various ways.

Because of this diversity, only a brief overview is presented in this section. The approaches fall into ligand- or structure-based methods, depending on whether a 3D structure of the target protein is absent or present, respectively.

In ligand-based methods, a drug candidate may be devised based on the descriptors of other compounds that bind the target by using quantitative structure-activity relationship (QSAR) or pharmacophore models [30].

Structure-based approaches attempt at finding compounds which would bind to a target protein of known structure from a database of small molecules, putative ligands. This process is called virtual (i.e. computer-aided) screening and consists of two key steps: docking and scoring. Docking algorithms fit the compound to the target protein and scoring functions than estimate how strongly does the docked compound bind. The strongly binding compound found by this approach is then presumably a strong ligand of the target protein – a drug.

In structure-based approaches, the target structures are obtained from experiments (X-ray crystallography, nuclear magnetic resonance) or homology modelling. Protein Data Bank (PDB) is a freely accessible repository containing hundreds of thousands of biomolecular structures. PDB was the source of structures for the construction of datasets in the Dataset part of this work (section 5).

Scoring functions can be classified as those using statistical thermodynamics, as empirical, knowledge-based or physics-based. Physics-based describe individual interactions either by force fields (Newtonian forces, as in molecular dynamics) or by quantum chemical means. The latter can be SQM or *ab initio* QM, in most advanced approaches [6, 30, 31].

In empirical scoring functions a few of additive descriptors of a system (such as number of hydrogen bonds) can be used to produce a quantity that reflects

the ligand affinity to a protein (score). The terms in the function are proposed and then the weights of the term contributions have to be determined by fitting to the reference values of complexes with known affinities – the training datasets. Knowledge-based scoring functions work in a similar way, but in this case, from larger training sets more general principles can be derived. Frequency of occurrence of some of many possible properties or descriptors (such as distance of two specific atom types) among strong binders can be used to derive potential function. In that sense knowledge-based scoring functions are more empirical than above mentioned empirical scoring functions [30]. Recently, artificial intelligence and machine learning algorithms are utilized for the identification of drug candidates, for example in scoring function development [31].

Scoring functions mentioned in the latter paragraph do, however, not provide good estimates of active ligands consistently for different proteins. The quantum-mechanical based (SQM) scoring functions outperform these other scoring functions substantially [32]. SQM scoring functions are described in greater detail in the section 1.7.

Testing constituents of entire databases, which may contain more than 10^7 [33] molecules, by means of virtual screening requires great number of computations (as many as 10^{13} [34]). Thus, when taking the size of the system (few thousands of atoms) into account, the use of *ab initio* QM methods is impractical in this case [35, 36]. Nevertheless, several QM-based scoring functions have been developed [37] and can be used for smaller-scale applications, with a fragmentation scheme or within a QM/MM setup, for instance (next paragraph).

Even though QM *ab initio* methods are usually too demanding to be used on large systems (such as whole proteins) in *in silico* drug design, SQM methods are a viable alternative. Fragmentation approaches can be used to further increase the efficiency of SQM or QM methods for large systems. In such ap-

proaches a large molecule is fragmented into smaller subsystems and the energy of a molecule is taken as a sum of energies of individual fragments [4, 38]. This approach neglects nonlinear effects (such as polarization among fragments), but it is possible to include these missing effects in some approximate way – for example, on a lower level of theory [39]. Another approach related to the fragmentation strategy can be used in case the system of interest is too large for a given higher-lever method: the higher-level method is applied on the ligand and its immediate environment and a lower-level method on the reminder of the molecule. An *ab initio* QM method can be a higher-lever method and a SQM a lower-level one [37] or molecular mechanics or SQM can be a lower-level method for a QM/MM, SQM/MM setups, respectively [40, 41]

The strength of the binding can be expressed as an interaction energy, the difference of energies of a complex (in *in silico* drug design most often ligand-protein complex) and the energies of its constituents (Fig. 1.6). The interaction energy can be used as a score or a score component [42]. Taking the interaction energy alone as a score would neglect the (presumably) subtle structural rearrangements upon binding – relaxation energy, which has to be accounted for in case of highly mobile structures. The force-field (molecular mechanics) or electronic structure SQM methods can provide an energy of a system in a single calculation. A similar important quantity is the change of solvation free energy upon the complexation $\Delta\Delta G_{solv}$, the difference between the solvation energies of a complex and its isolated constituents (protein and ligand). It corresponds to the change of free energy upon binding due to the solvent and it is also used in SQM scoring functions and will be referred to as "interaction solvation energy" further.

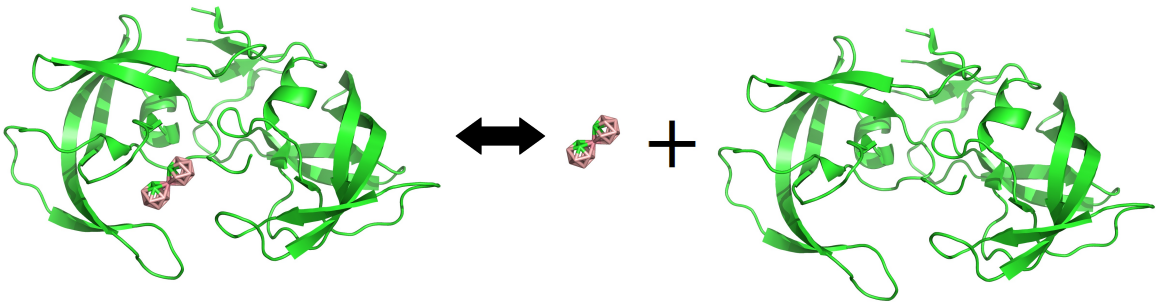

Figure 1: An interaction energy as a difference of energies of a complex (protein-ligand) and energies of individual protein and ligand respectively. Here, HIV protease and its ligand of Protein Data Bank code 1ztz [43] has been used for an illustration. The isolated monomers are extracted from the complex for the interaction energy calculation without a relaxation of geometry. Missing deformation contributions are neglected or considered independently (section 1.7)

1.7 Applications of SQM Methods to *in Silico* Drug Design

For the statistical thermodynamical approach mentioned in section 1.6, QM [44] or SQM [45] methods have been used in QM/MM setups. QM/MM (usually meaning SQM in this context, as was used in the reference [41]) have been used to enhance the quality of a docked complex and DFT has been used in the context of QM/MM for proteins, too [40, 46]. AM1 has been used alongside with COSMO with MOZYME linear scaling algorithm (section 3.3.4) to evaluate the correlation of calculated interaction energies with the experimental affinities [47]. The authors argued that flawed accuracy is largely due to COSMO not being parametrized for biomolecules (RNA-ligand complexes were investigated in this study). The first SQM scoring function *per se* introduced a physically sound score for a protein ligand complex of a form

$$score = H_{bind} + E_{disp} + \Delta\Delta G_{solv} - T\Delta S \qquad \text{(EQ.8)}$$

With the change of solvation energy $\Delta\Delta G_{solv}$ being estimated at the Poisson-Boltzmann model (a generalization to the Generalized Born models) [48]. The H_{bind} interaction energy was calculated at the AM1 level and was augmented by an attractive part of a Lennard-Jones potential (the E_{disp} term decaying with the sixth power of atom-atom separation – as in equation EQ.1). An entropy term was introduced to this score with ΔS being the sum of a linear function of rotatable bonds immobilised upon binding and a term corresponding to the change of a solvent-exposed surface by a complex formation. This SQM scoring function performed better than 11 other non-QM-based scoring functions tested in the study on metalloprotein complexes. As quantum effects are important in active sites of such complexes (charge transfer involving metalloprotein enzyme and its metal cofactor), the use of QM-based methods was appropriate. Molecular-orbital-based SQM methods (section 3.1) used in several studies of protein ligand complexes included PM6-DH2, PM6-D3, PM6-D+, and also PM7 with the implication of their usage in virtual screening [49–51]. Scoring functions featuring DFTB (section 3.2) energy have also been used in this context, to optimize the docked structures [52].

The SQM scoring function developed in our group before has a similar general form as in equation EQ.8

$$score = E_{bind} + \Delta\Delta G_{solv} + G_{conf} - T\Delta S \qquad \text{(EQ.9)}$$

but with the interaction energy E_{int} calculated on the PM6-D3H4X level, and with the solvation contribution $\Delta\Delta G_{solv}$ calculated with the SMD or COSMO models. The conformational energy G_{conf} is the energy change associated with

the change of conformation of a protein and a ligand upon the binding. To capture this effect, energy minimisation of both interacting partners isolated from the geometry of the complex as a starting point has to be done. The last term of equation EQ.9 can be obtained from the number of rotatable bonds of complex and constituents or from their calculated vibrational energy levels [53]. This scoring function was then simplified to include only interaction and solvation interaction energies. Together with some other adjustments in the preparation of the systems, the procedure ran much faster, while still being efficient for systems tested [54]. Later, another variant of this scoring function was introduced, in which the PM6 D3H4X in the interaction energy term was swapped for the DFTB3 D3H4 as an SQM method of choice. This was done to address the difficulties encountered when treating metalloproteins [55]. Recently, COSMO2 [56] has been implemented into this scoring functions in place of solvation model [42].

2 Aim of the Thesis

The aim of this thesis is to optimize the efficiency of semiempirical quantum mechanical (SQM) methods by means of empirical corrections for their utilization in computer-aided drug design.

In the first part of the thesis we describe the reparametrization of the implicit solvation model COSMO along with the addition of a term corresponding to nonpolar contributions. The resultant optimized model is designated COSMO2 (published as reference [56], provided as appendix **A)**). This optimized model is to become a part of scoring function developed in our group to improve the performance of a solvation model as a scoring function accuracy limiting factor.

The subject of the second part of the thesis is the construction of protein-ligand derived datasets of noncovalent interactions for testing the accuracy and

development of SQM methods, PLF547 and PLA15. The PLF547 dataset consists of fragments of protein interacting with ligands and the PLA15 is a dataset of protein active sites complexed with ligands. By constructing of these datasets, we are addressing the lack of quality reference data suitable specifically for the drug design applications (benchmark data on protein-ligand complexes). The accuracy of selected methods, including COSMO2, is tested on the newly developed PLF547 and PLA15 datasets. The results of this part are published as reference [23] and are provided as appendix **B)**.

3 Methods

3.1 Molecular-Orbital-Based SQM Methods

Molecular-orbital-based (wave-function-based) semiempirical methods are akin to HF method in the basic formulation of the problem and can be regarded as its simplification. In this basic formulation a molecular orbital is expressed as a linear combination of atomic orbitals χ (i.e. functions of space centered on atom). The analytically unsolvable two-electron part of Hamiltonian is regarded as an interaction of each electron with a potential formed by a "static" electronic distribution of all the other electrons residing in their respective χ. The composition of molecular orbital from χ of various weights is guessed so it can be substituted for the electron distribution which appears in a two-electron Hamiltonian in addition to being a function on which an operator operates, yielding the energy and improved composition of molecular orbital. This new composition (a wave function) is then used again in an iterative procedure until a state is reached in which a substituted function yields (within a specific threshold) itself – a self consistency is reached. The method is also called self-consistent field (SCF). Each successive step represents a reaction of electrons to

each other happening, as though, at once. Hartree-Fock electronic energy can be expressed as

$$E^{HF} = \sum_{r}^{b}\sum_{s}^{b}\sum_{i}^{n/2} c_{ri}c_{si}\Big[2\int \chi_r^*(1)\big(-\frac{1}{2}\Delta_1 - \sum_A \frac{Z_A}{r_{1A}}\big)\chi_s(1)d\nu_1 + \quad \text{(EQ.10)}$$

$$\sum_{t}^{b}\sum_{u}^{b}\sum_{j}^{n/2} c_{tj}c_{uj}\int\int 2\frac{\chi_r^*(1)\chi_s(1)\chi_t(2)^*\chi_u(2)}{r_{12}} - \frac{\chi_r^*(1)\chi_u(1)\chi_t^*(2)\chi_s(2)}{r_{12}}d\nu_1 d\nu_2\Big]$$

$$+\sum_{A}^{N}\sum_{A\neq B}^{N}\frac{Z_A Z_B}{r_{AB}}$$

Where the first one-electron integral represents a kinetic energy of an electron and an energy of its interaction with nuclei A. Second integral over coordinates of pair of electrons represents Coulomb and exchange interactions between electrons. The sums r, s, t and u goes over b basis functions χ, the sum i and j goes over n/2 spatial molecular orbitals. The last term stands for the pairwise interaction of N nuclei of charges Z.

HF derived molecular-orbital-based SQM methods aside other approximations neglect some of two-electron integrals, which leads to a considerable computational time savings. The types of omitted integrals gives the basic definition of the method. Various extensions of this basic definition can vary in the different parameter sets or in empirical functions used.

Historically, among the most important molecular-orbital SQM methods were complete neglect of differential overlap (CNDO), intermediate neglect of differential overlap (INDO) alleviating the neglect of some integrals, neglect of diatomic differential overlap (NDDO) further increasing the range of integrals and modified neglect of differential overlap (MNDO). A further extension of MNDO leads to Austin model 1 (AM1), Parametrized model 3 (PM3), 6 (PM6)

and latest PM7. For all the methods, only the valence electrons are treated, while the rest is included in the potential due to the nucleus, forming an effective core potential which is invariant under the SCF procedure.

The starting point for SQM methods was the CNDO. CNDO neglected two-electron integrals (both integrals over ν_1 and ν_2) other than those with both basis functions χ over one electron being equal ($\chi_r(1)$ equals $\chi_s(1)$), and likewise for the second electron. INDO included also integrals in which all the basis functions χ were located on the same atom [3].

NDDO and derived more recent methods include all two-electron integrals with both basis functions of first electron (such as $\chi_r(1)$ and $\chi_s(1)$ in the first integral) on the same atom, and both basis functions of second electron likewise are on the same atom. These remaining not neglected integrals are replaced by approximate expressions which include parameters that are fitted to reproduce benchmark results or conform with experimental data. For instance, NDDO remaining two-electron integrals are approximated as γ_{rs}

$$\gamma_{rs} = \frac{1}{R_{rs} + 2(\gamma_{rr} + \gamma_{ss})^{-1}} \qquad \text{(EQ.11)}$$

$$\gamma_{rr} = \int\int \frac{\chi_r^*(1)\chi_r(1)\chi_r^*(2)\chi_r(2)}{r_{12}} d\nu_1 d\nu_2 \qquad \text{(EQ.12)}$$

And likewise for γ_{ss}. The integral γ_{rr} or γ_{ss} can be taken as the difference between the ionization potential and the electron affinity of element on which the basis function χ_r is localized. [57, 58]

The one-electron part of the energy, integral over ν_1, is approximated in NDDO as

$$\int \chi_r^*(1)\big(-\frac{1}{2}\Delta_1 - \frac{Z_A}{r_{1A}}\big)\chi_s(1)d\nu_1 - \sum_{B\neq A} Z_B \int \frac{\chi_r(1)^*\chi_s(1)S_B^*S_B}{r_{1B}} d\nu_1 \quad \text{(EQ.13)}$$

for both basis functions centered on atom A. The first integral is neglected for χ_r not being equal χ_s and is often parametrized as an average of ionization potential and electron affinity of an atom A otherwise [59, 60]. S_B is a spherical core density on B of electrons that have been left out by the valence treatment.

And for basis functions centered on different atoms the one-electron (resonance) integrals are approximated as

$$\beta_{ArBs} = \frac{(\beta_A + \beta_B)}{2} \int \chi_r(1)^* \chi_s(1) d\nu_1 \qquad \text{(EQ.14)}$$

Where β_{ArBs} is approximated as an average of β parameters for different types of atoms and is scaled by an overlap integral [59]. Above described neglections of integrals employed by NDDO and derived methods reduce the number of integrals to the second power of the basis. The diagonalization of the matrix elements to calculate the energy (property) then scales with the third power of the system size (basis functions), which is thus the bottleneck of the method regarding its speed [3, 61].

The interaction of nuclei is augmented with the inclusion of core electrons as

$$\sum_A \sum_{B>A} C_A C_B \int \frac{S_A(1)^* S_A(1) S_B^*(2) S_B(2)}{r_{AB}} d\nu_1 \nu_2 + f_{AB} \qquad \text{(EQ.15)}$$

Where C are charges on nuclei A and B modified by the core electrons on those nuclei. The f_{AB} is the empirical function of interatomic distance containing atom-specific parameter.

MNDO represents a further step in the SQM molecular orbital method development. The integrals of MNDO represent same interactions as in NDDO, except that the two-electronic part is substituted by the multipole expansion:

$$\sum_{l1} \sum_{l2} \sum_m [M^A_{l1;m}, M^B_{l2;m}] \qquad \text{(EQ.16)}$$

Where m and l are respective quantum numbers

$$[M^A_{l1,m} M^B_{l2,m}] = \frac{1}{2^{l_1+l_2}} \sum_i \sum_j [R^2_{ij} + (\rho^A_{l1} + \rho^B_{l2})^2]^{-1/2} \qquad \text{(EQ.17)}$$

Where i and j are point charges of respective multipoles separated by R, ρ^{A} and ρ^{B} are density-like additive terms characteristic for each atom type A and B of a given l. These with the distances of a point charge of a multipole from a nucleus center are atom parameters obtained from other quantities accessible from the calculation. MNDO is the basis of currently used molecular orbital SQM methods in that these subsequent methods retain this MNDO form of integrals [62].

For the AM1 and PM methods, MNDO parameters are reoptimized and the form of f_{AB} is readjusted. PM6 also extends a basis set to d orbitals for selected heavier elements [63]. PM7 augmented integrals with empirical functions mainly correcting description of solids and added empirical dispersion and hydrogen bond corrections [64]. PM6 and PM7 will be further elaborated in the sections 3.3.1 and 3.3.2.

3.2 Density-Functional Tight-Binding Methods

The second major family of SQM methods is density-functional-theory (DFT) based. DFT uses electron probability density rather than a wave function, as is the case in HF, to derive desired properties of a system. In present practice, (Kohn-Sham) orbitals are reintroduced to construct a density. The orbitals are used to calculate the interaction of electron with nuclei, electronic kinetic energy and the Coulombic interelectronic repulsion much the same way as in HF.

$$E^{DFT} = -\sum_A Z_A \int \sum_i^n \frac{\theta_i^2(1)}{r_{1A}} d\nu_1 - \frac{1}{2} \sum_i^n \int \theta_i(1) \Delta_i \theta_i(1) d\nu_1$$

$$+\frac{1}{2}\int\int\sum_{i}^{n}\sum_{j}^{n}\frac{\theta_i^2(1)\theta_j^2(2)}{r_{12}}d\nu_1 d\nu_2 + E_{xc}(\rho) \qquad \text{(EQ.18)}$$

The molecular orbitals θ are then expanded as a linear combination of atomic orbitals, yielding comparable terms as in EQ.10. The quantity of sum over squares of molecular orbitals θ represents an electron density, which is used in a E_{XC} exchange correlation functional. Its correlation part is designed to correct the error due to the approximation of "static" electronic distribution creating a potential for each individual electron in each successive SCF step. The exchange part is meant to describe the Fermi effects of like-spin fermions. This last term effectively covers all the rest to the energy, except for relatively easily attainable effects described by the first three terms [1].

SQM methods derived from DFT are density functional tight-binding methods (DFTB). DFTB is formulated around the Taylor expansion of density around the reference one, which is the superposition of neutral isolated atomic densities [2, 3].

$$E^{DFTB3} = \frac{1}{2}\sum_{AB}V_{AB}^{rep} + \sum_{iAB}\sum_{\mu}\sum_{\nu}n_i c_{i\mu}c_{i\nu}H_{\nu\mu}^{0}$$

$$+\frac{1}{2}\sum_{AB}\Delta q_A \Delta q_B \gamma_{AB} + \frac{1}{3}\sum_{AB}\Delta q_A^2 \Delta q_B \Gamma_{AB} \qquad \text{(EQ.19)}$$

Where the first term, sometimes also zeroth order correction to the reference, represents a repulsion potential. This potential between reference densities of atoms A and B is a function fitted to adjust for the energy difference between the DFTB and a reliable high-level computational method [65]. Thereby it also absorbs some portion of an error introduced by other terms. The second term in the equation contains Hamiltonian matrix of basis functions μ and ν of superposition of densities belonging to neutral free atoms. Integrals for this

term are precalculated [66]. The expansion coefficients c of these neutral ith molecular orbitals μ are on atom A and ν on atom B (basis set is minimal). The Δq in the third term of the expression EQ.19 represents the interaction of excess charges on atoms A and B obtained from Mulliken population analysis of the molecular orbitals. The γ function substitutes for the integral of excess density over volumes of respective atoms and has the form

$$\frac{1}{R_{AB}} - S(R_{AB}; U_A; U_B) \qquad \text{(EQ.20)}$$

with S being an exponential function where Hubbard parameters U for atoms are related to the difference of ionization potential and electron affinity of the respective atom. The γ converges to this difference for no separation and converges to its first term for large separation R_{AB} [2, 67, 68]

The fourth term, third order correction (thus DFTB3), captures the deviation from the reference density due to the interaction of charged species. Γ_{AB} features a dependence of γ_{AB} function on the excess charge on A [66].

3.3 Computational SQM and QM Methods Tested

DFTB3 and PM6 are SQM methods used in scoring functions developed in our group [42, 55]. Their significant advantage over some other SQM is an available implementation with linear scaling algorithms, reducing the speed of the calculations, so that they can be used efficiently in *in silico* drug design [42]. Previously the focus in our group has been on the optimization of these methods and corrections to dispersion, hydrogen and halogen bonds have been developed, parametrized for them [18, 22]. COSMO is our preferred solvation model, as it is also available with a linear scaling algorithm, while being sufficiently accurate.

We augment and reparametrize COSMO for the use with PM6 and PM7 in the COSMO2 part of the thesis. Along with COSMO with original settings and

newly optimized COSMO2, we test SMD and PCM solvation models in various computational setups on the reference datasets of solvation energies. We include higher-level DFT/COSMO-RS results for the comparison with SQM coupled solvation models.

In the second part of the thesis devoted to the construction of datasets, we compare the performance of SQM or QM methods potentially suited for *in silico* drug design. Aside from DFTB4-D3H4 and PM6-D3H4, also DFTB3-D3H5, PM7, GFN2-xTB, HF-3C (a QM method) are tested on the datasets constructed for this part of the thesis. We include older AM1 and PM6 without corrections, as well as QM DFT with selected functionals and MP2 for comparison. As a benchmark for the two datasets constructed, we use composite schemes based on high-level *ab initio* methods – CCSD(T), MP2 and DFT. We also use COSMO-RS as a reference for solvation energies, since no experimental data is available for PLF547 dataset. For PLA15, we use a sum of COSMO-RS solvation energies as an approximation to the solvation energy of a respective protein active site.

3.3.1 PM6-D3H4 and DFTB3-D3H4

The PM6 is a wave-function-based SQM method derived from MNDO as mentioned in section 3.1. The f_{AB} of equation EQ.15 between two atoms A and B has the form

$$x_{AB} \times exp\big[-\alpha_{AB} \times (r_{AB} + 0.0003 r_{AB}^6)\big] \times \big[C_A C_B \int \frac{S_A(1)^* S_A(1) S_B^*(2) S_B(2)}{r_{AB}} d\nu_1 \nu_2\big] \tag{EQ.21}$$

with x and α being the pairwise atomic parameters. The function was slightly augmented for H-O, H-N, C-C and Si-O interactions. A repulsion function was added to the core-core potential

$$f_{AB}^{rep} = 10^{-8} \times \left(\frac{C_A^{1/3} + C_B^{1/3}}{R_{AB}}\right)^{12} \tag{EQ.22}$$

Later, the two-electron integral of equation EQ.11 was modified slightly to reproduce properties of solids more accurately [69]. Altogether, PM6 contains parameters for the first terms introduced in equation EQ.13, β of equation EQ.14, parameters for two-center electronic interaction of equation EQ.11, orbital exponents of basis functions and parameters in core-core functions (some of which are pairwise). In PM6 the parameters were optimized to reproduce heats of formation of about 9000 compounds [19, 62].

D3 and H4 corrections in PM6-D3H4 stand for the dispersion correction and hydrogen bond correction respectively, parametrized for the method, both adding a post-SCF term to the energy. The dispersion correction D3 is built upon the D correction (section 1.3). The coefficients C_{AB} of eq. EQ.1 are calculated for each atom pair and are valence state-specific. The dispersion term has a form as in EQ.1, only without higher than n=6 terms and it includes a corrective function for overly attractive hydrogen atoms (including hydrogen atoms not involved in hydrogen bonds) [22].

The hydrogen bond correction H4 is a product of five functions, which contain geometry arguments and empirical parameters, accounting for: a radial dependence (polynomial of a donor-acceptor distance), an angular dependence (donor-hydrogen-acceptor angle), an occasional proton transfer, the charge of interacting moieties and a function correcting for hydrogen bonds involving water which behaves slightly anomalously in PM6 method. The parameters for the H4 correction along with the parameters of the dispersion correction EQ.1 have been attained by the fit to the reference data of S66 benchmark dataset (a dataset of small organic molecules interacting with one another in several orientations and separations) [70, 71].

The D3H4 corrections for DFTB3 are derived in the same way and on the same benchmark as for PM6, only with the DFTB3 method-characteristic pa-

rameters.The correction has been parametrized for several other methods [22]. For the DFTB3, we use the 3OB parameters set [65].

3.3.2 PM7

PM7 is the successor of the PM6 method. The authors have, as in case of PM6, introduced modifications for the integral over two electrons in order to fix the description of large solid systems. PM7 uses in-built post-SCF dispersion and hydrogen bond corrections. The dispersion correction has the same general form as in equation EQ.1, limiting n to 6 [72] and with a slight modification for solids. Its hydrogen bond correction has a form

$$\frac{C_{AB}}{R_{AB}^2} \times f_{damp} \times f_{geom} \qquad \text{(EQ.23)}$$

where f_{geom} is a function involving goniometric dependences on dihedral angles and f_{damp} dampens the correction for small and large separations. Unlike PM6-D3H4, both dispersion and hydrogen bond corrections have been parametrized along with the rest of PM7 on multiple kinds of reference data which besides noncovalent interactions also included heats of formation and geometries. This, according to authors of PM7 publication (ref. [64]), might have been the reason why PM7 underperformed when tested for noncovalent interactions compared to variants of dispersion and hydrogen bond corrected PM6, predecessors to PM6-D3H4 – which were parametrized on datasets for noncovalent interactions only.

3.3.3 DFTB-D3H5

The hydrogen (H5) part of D3H5 correction to the DFTB3 correcting the hydrogen bonds X-H $\cdots$ Y has a form

$$\gamma_{AB=HY}^{H5} = \gamma_{AB}\big(1 + k \times exp[-\frac{(r_{HY} - r_0)^2}{2w^2}]\big) \qquad \text{(EQ.24)}$$

where k is an atom-specific parameter, r_0 corresponds to the equilibrium bond length and together with w are based on van der Waals radii.

This correction multiplies the γ_{AB} in the third term of the function EQ.19 for the interactions classed as hydrogen bonds by a Gaussian function centered at the equilibrium distance for a given hydrogen bond. This enhances the γ around the distances relevant for the hydrogen bond. The modification of the γ function directly involves the correction in the self-consistent iterative procedure. DFTB3-D3H5 has also been parametrized on the S66 dataset [71].

3.3.4 Linear Scaling Algorithms, MOZYME

Standard solving (diagonalization) of the Fock matrix has to deal with integrals where the first electron is in one orbital and the second electron in one each of all the other orbitals regardless of distance between their respective atom centers. Resulting molecular orbitals are delocalized. Alternatively, localized molecular orbitals can be used, limiting the integrals to centers that are close. Scaling can be made linear with the system size. By MOZYME, interaction of higher-order multipoles were considered only up to certain separation, lower-order multipoles were included up to greater separation and the interaction of monopoles was retained [61].

In DFTB3, similar Divide and Conquer algorithm [73] is available in which the electron density is partitioned into smaller regions of space calculated independently – a procedure similar to the localization of orbitals employed by MOZYME.

3.3.5 GFN-xTB, GFN2-xTB

GFN-xTB (geometry, frequency, noncovalent interactions, extended tight binding) is a method akin to DFTB3. While being semiempirical, it employs additional basis function on hydrogen (double zeta H) and d orbital basis to each of selected heavier elements. The method uses D3 dispersion correction (section 1.3) and a Lennard-Jones potential resembling form of halogen bond correction multiplied by an angular damping function [74].

Its successor, GFN2-xTB has no need for an additional basis function on hydrogen or an empirical halogen correction. Instead, GFN2-xTB extends DFTB-level of description by including multipoles (monopole-dipole, dipole-dipole and monopole-quadrupole interactions) to the charge fluctuation term (third term of EQ.19 plus corresponding exchange-correlation term) [75]. The method uses more recent D4 dispersion correction (section 1.3).

3.3.6 HF-3C

HF-3C is a HF method, so it qualifies as a QM method, but it has been proposed as an alternative for SQM when feasible. The method operates with an individualized small basis set, minimal for H and first row elements other than Li, Be. Along with D3, it contains other two empirical corrections to the energy – a basis set superposition correction and a short-range correction. Basis set superposition error arises from the different basis set sizes of a complex and its constituents, and gets smaller with increasing basis set. The correction for this error in HF-3C is a sum over all the atom pairs A-B of basis set dependent energy differences of atom A multiplied by an exponential function of distance between A and B scaled by a function of number of virtual (unoccupied) orbitals on atom B. The other correction is a core-core augmenting term, also an exponential of an interatomic separation. The latter corrects specific bond

lengths [76].

3.4 Quantum Mechanical Benchmark Methods

In the Dataset part we were in a need of reliable benchmark energies for protein-ligand fragments and for large protein active site complexes. MP2 – second order of Møller-Plesset perturbation theory and CCSD(T) – coupled clusters singles, doubles, perturbative triples are high-level QM wave function methods which incorporate excited Slater determinants to capture the correlation energy missed out by the SCF treatment. The full CCSD(T) method would have been too demanding when used on systems of either PLA15 or PLF547 datasets. The DLPNO-CCSD(T) scheme, which we used instead, allows CCSD(T) to be applied to larger systems by neglecting contributions from weakly interacting electron pairs (such as distant electrons) to the correlation energy [4]. It has been shown, that the DLPNO-CCSD(T) in cc-pVTZ with the "tight" energy threshold for the integral neglection set, the methods reaches about 0.06 kcal/mol standard deviation with errors below 0.25 kcal/mol [4] for the S66 benchmark dataset [71] as compared to the full CCSD(T) in TZ basis. Similarly good agreement has been found between the results of canonical CCSD(T) and DLPNO in DZ basis on the large GMTKN55 set of molecules (with above mentioned threshold setup) [77]. For the benchmark energies of PLF547 dataset we have used

$$E_{Frag}^{Benchmark} = MP2_{F12}^{CBS} + \Delta CCSD(T)^{aug-cc-pVDZ}$$

$$\Delta CCSD(T)^{aug-cc-pVDZ} = DLPNO\ CCSD(T) - MP2 \qquad \text{(EQ.25)}$$

with $MP2_{F12}$ being an explicitly correlated method providing a good estimate of MP2 complete basis set (CBS) energy [78] and ΔCCSDT providing the

correlation energy above the MP2 level (up to DLPNO-CCSD(T)). We have used aug-cc-pVDZ basis set with the "tight" threshold for DLPNO-CCSD(T), cc-pVDZ-F12 for MP2-F12 and aug-cc-pVDZ for MP2.

For PLF547, we also provide results by a few QM methods for the comparison to the SQM performances: B3LYP/def2-QZVP, BLYP/def2-QZVP and BLYP/DZVP-DFT, all with D3 correction, and also MP2 in setup used for the benchmark calculation.

As benchmark solvation energies, we have used COSMO-RS (further section) and we included also SMD at HF/6-31G* level as an alternative QM-based solvation model for comparison. In the COSMO2 part, where the experimental solvation energies were accessible, we used COSMO-RS for comparison.

For the PLA15 dataset of large active site and ligand complexes such an approach would not have been computationally feasible, still. Instead, we calculated the complex on the DFT (B3LYP-D3/DZVP-DFT) level and added a correction to the DFT result. The correction was the difference between the DFT interaction energy of each ith fragment of PLF547 constituting the active site of PLA15 complex in the same B3LYP-D3/DZVP-DFT setup and the above mentioned benchmark for the respective PLF547 fragment.

$$\Delta E_{activesite}^{Benchmark} = \Delta E_{activesite}^{DFT-D3} + \sum_i (\Delta E_i^{Benchmark} - \Delta E_i^{DFT-D3}). \quad \text{(EQ.26)}$$

This correlation represents an additive, "unpolarizable" part of a correlation energy (up to the MP2-F12 – DLPNO-CCSD(T) level) to the DFT, while the nonlinear effects are covered on the lower DFT level. The non-additive portion of the correlation energy has been neglected.

For the PLA15 dataset, reference solvation interaction energies of each active site were approximated by a sum of fragment COSMO-RS energies, since active-

site COSMO-RS is likewise not feasible. This is presumably a crude measure which neglects all nonlinear effects.

3.5 Solvation Models

3.5.1 COSMO

COSMO (conductor-like screening model) is an apparent surface charge (ASC) solvent model (section 1.4). The permittivity constant ϵ is set to ∞ in COSMO, mimicking a conductor. The potential of a solute is then scaled by

$$f = \frac{\epsilon - 1}{\epsilon + k} \qquad \text{(EQ.27)}$$

restituting a dielectric behaviour, where the k is 0.5 for neutral compounds, 0 for ions [79]. Up to present, COSMO imlpementation for SQM methods has been lacking a non-polar term, a term covering effects other than those of electrostatic nature described by ASC formulation.

An essential part of an ASC solvation model is a construction of a cavity around a solute on which the apparent surface charges are placed. In the original implementation, the cavity was constructed as an superposition of overlapping atom-centered spheres of radii close to van der Waals radii (the technical construction was done in a numerical algorithm by projecting surface segments on the spheres). The cusps resulting from overlapping spheres were then eliminated by projecting of sphere segments on a larger sphere (extended by the solvent radius), discarding internal segments and projecting back, leaving holes in a regions around would-be cusps. In later implementations, holes are filled by interlocked triangles in various algorithms [80]. The models described further on use similar cavity construction schemes designed to represent closely molecular shape [79]. The original COSMO implementation uses van der Waals radii

for the cavity construction, [81] later implementations, however, use optimized radii, as COSMO-RS does, as well as more elaborate algorithm of cavity construction [80, 82]. The way how the problematic parts of the cavity are handled (cusps or holes) together with the selection atomic radii are crucial, because the shape of the cavity impacts the accuracy and robustness of the method.

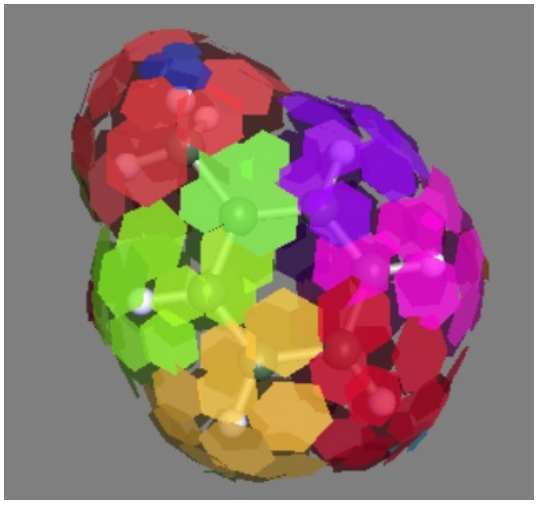

Figure 2: A cavity built from surface segments by COSMO in the recent MOPAC implementation around a toluene molecule. Provided by A. Klamt.

3.5.2 COSMO-RS

The continuum solvation model is an approximation, less adequate for a polar solvent of anisotropic electron density. Important specific interaction like hydrogen bonds are also not addressed by the basic model. COSMO-RS (COSMO for real solvents) builds on the COSMO with DFT calculation. It adds a dispersion term proportional to the exposed surface area of each atom constituting a solute multiplied by an atom-specific constant. Similarly, the model takes into account solute-solute and hydrogen bond interactions by using surface segment charge densities and the contact areas. These contributions are included by means of

statistical thermodynamics [82, 83].

We used B-P/def2-TZVPD DFT to provide the electronic structure of the solute for COSMO-RS. The COSMO-RS has been used for the benchmark solvation interaction energy of PLF547 and PLA15 datasets and as a reference higher-level method to the results of other solvation models when testing COSMO2.

3.5.3 PCM

PCM (Polarizable Continuum Model) was historically the first implementation of ASC solvation models as D- (dielectric) PCM. In a conductor-like PCM variant C-PCM, infinite ϵ is used, using the same form of a scaling function as in EQ.27, so that the electrostatic contribution is the same as in COSMO. In C-PCM, the k is chosen to be 0 [84], as has been recommended by COSMO authors for ions [82].

Additionally to electrostatic term, PCM contains cavitation, repulsion and dispersion nonpolar terms. The cavitation term is

$$\sum_i \frac{S_i}{4\pi R_i^2} \times f \qquad \text{(EQ.28)}$$

with S_i being the area of a surface fragment i from which the cavity is constructed, R_i is a radius of a sphere belonging to the fragment i and f is a function of R_i, solvent radius, solvent density number, pressure and temperature [85]. Similarly, dispersion energy is

$$\rho \sum_i \sum_M \sum_N -\frac{C_{NM}}{R_{Mi}^6} \times S_i \frac{R_i^2 + R_{Mi}^2 - R_{Ml}^2}{2R_i} \qquad \text{(EQ.29)}$$

Here, ρ is the solvent number density, R_{Mi} a distance between solute atom M and surface fragment i, R_{Ml} between solute atom and a center of sphere associated with surface fragment i. A repulsion energy is represented by a

similar, but positive, term featuring exponential function of R_{Mi} and expression of first, second and third power of R_{Mi} rather than sixth.

3.5.4 SMD

SMD (Solvation Model Density), unlike to its Generalized-Born-based predecessors, also belongs to the ASC approaches [86] and its formulation is similar to that of PCM methods [79, 86]. It features a nonpolar term of a form

$$(\xi_k + \xi^M) \times \sum_{k}^{atoms} A_k \qquad \text{(EQ.30)}$$

with ξ_{k} being the atomic surface tension dependent on an atomic number and solvent characteristics, ξ^{M} being the solvent characteristics dependent molecular surface tension and A being the solvent accessible surface area of an atom constructed from different radii than those used for the electrostatic contribution [86].

3.6 Software

PM6 and PM7 methods were used in this work as available in MOPAC 2016 [87], except for the PM6/PCM and PM6/SMD in the COSMO2 part, which were calculated in GAUSSIAN 09 [88]. HF/6-31G* paired with SMD was used also in GAUSSIAN in the Dataset part for comparison with SQM-paired solvent models. PM6 predecessor AM1, tested in COSMO2 part for the comparison with newer methods, was also used with MOPAC. In the Dataset contribution, DFTB3 calculations (including DFTB3-D3H5, in the further sections) were performed in DFTB+ code [89]. In the COSMO2 part, the DFTB3 was used in conjunction with PCM (for which we included a setup with nonpolar term excluded – denoted C-PCM' further) and SMD solvation models in GAMESS [90]. GFN-xTB methods use their own code written by the authors of the methods

– xtb code [74, 75]. We used HF-3C as was implemented with default settings in ORCA code [91]. All the DFT and MP2 calculations were performed in TURBOMOLE 7.3 [92]. DLPNO-CCSD(T) calculations were done in ORCA. COSMO-RS was used in COSMOTHERM-v17 [93].

3.7 Cuby4

Cuby in an integrative platform for computational chemistry [94]. It can be used for running calculations in various codes and putting together constituents of interaction energy and similar quantities. Aside from this functionality, it has available options for geometry manipulations, which were used in fragmentation of complexes into PLF547 dataset in Dataset part, and options for parameter optimization, used for the COSMO2 development. All the calculations, fragmentation into constituents of PLF547 dataset, addition of hydrogen atoms to truncated residues and optimization of parameters with respect to RMSE (Root Mean Square Error, a measure of deviation) were done in cuby4 framework.

4 Reparametrized COSMO Solvation Model with a Nonpolar Term – COSMO2

This part of the thesis, COSMO2 part, is dedicated to the optimization of a COSMO solvation model applied along with the PM6 or PM7 methods. The motivation behind this undertaking is the development of an SQM scoring function utilizing a solvation model (COSMO) for *in silico* drug design [42, 53, 95]. As will be elaborated in dedicated sections, the error in interaction energy produced by SQM methods is smaller in magnitude than the solvation energy error of COSMO (solvent models in general). Thereby, when using SQM together with solvent model to estimate a score of a compound, it is advantageous to fo-

cus preferentially on the optimization of the solvent model – accuracy limiting part of the score.

4.1 Strategy of COSMO2 Development

In the work we present here, we have included a nonpolar term to the solvation energy expression of COSMO model as a total cavity surface A scaled by a surface tension coefficient ξ

$$E_{np} = \xi A \qquad \text{(EQ.31)}$$

This nonpolar term should cover in average way other than electrostatic contributions to the solvation energy (dispersion, repulsion and cavitation energy).

Additionally, we have optimized the parameters for atomic radii used for the COSMO cavity construction. The optimization of parameters and surface tension parameter was performed for PM6 and PM7 SQM methods with respect to a reference dataset obtained with permission from Truhlar and coworkers [96]. We tested the newly optimized methods along with other solvation models on independent datasets as well as on actual protein ligand complexes (as a part of a scoring function).

4.2 Training and Validation Sets for COSMO2 Parametrization

The training dataset prepared from the data of Truhlar et. al., Minnesota Solvation Database [96], contains geometries and experimental solvation energies of mostly small molecules in various solvents from which we used water. We excluded 7 compounds for which the calculations did not converge by some of tested/benchmark methods. Further, we excluded 57 compounds that were part of one of the testing data sets (SAMPL1, see below). The resultant training

data set prepared from Minnesota Solvation Database (MNSol) containing 331 neutral and 138 charged compounds will be denoted MNSol*. The authors of the original COSMO model used only neutral compounds, when the parameters were optimized in the later implementations [82, 83]. Detailed list of compounds in training data set has been provided in supplementary material of ref. [56]. For the testing of methods we prepared three non-overlapping datasets of neutral compounds – datasets prepared from data of two SAMPL challenges designated SAMPL1 [97] and SAMPL4 [98] in this work, a dataset derived from data collected by Mobley et. al. [99] (Mobley266) and one small set of charged species from Lee et. al. [100] (C10). The SAMPL challenge 1 and 4 provides experimental solvation data for compounds which we had to build and optimize with the B3LYP-D3/def2-QZVP and with COSMO solvation model. The SAMPL compounds include herbicides, pesticides or other bioactive small compounds and are thus drug design relevant. We excluded compounds (10 from SAMPL1 challenge, 6 from SAMPL4) for which experimentally supplied and COSMO-RS calculated solvation energy differed significantly. This has been done to eliminate possible errors due to selection of conformers. We removed few compounds of SAMPL4 which were in the training dataset (SAMPL1 has been removed from the training dataset, so there was no longer any overlap). SAMPL1 and SAMPL4 derived datasets ended up with 53 and 42 molecules respectively. The C10 compounds were prepared and optimized as in case of SAMPL. The data adapted from Mobley et. al. included geometries. After excluding compounds shared with the training set, 266 small organic molecules remained. Moreover, we tested the newly parametrized method as a constituent of a SQM scoring function on a series of complexes of carbonic anhydrase II with 10 compounds, for which the experimental affinities and structures of complexes were determined [95]. We then assessed the efficacy of the model on a basis of correlation

between computed score and an experimental affinity.

4.3 Optimization of COSMO2 Parameters

For the optimization of element radii and the nonpolar term, we used iterative Broyden–Fletcher–Goldfarb–Shanno numerical gradient-based algorithm. The starting values for element radii in our considerations were the values found for the original COSMO implementation in MOPAC (section 3.5.1, Tab. 1). The starting value for the surface tension parameter was chosen to be 0.05 kcal/mol/Å^2. Many runs with different initial parameters were performed to confirm the minimum found was not only the local one. Firstly, we parametrized only radii for H, C, N, O elements with the surface tension parameter of nonpolar term by excluding compounds with other elements. Then, we included all the elements – compounds with S, P, F, Cl, Br, I, but kept H, C, N, O radii fixed. We focused on these elements for drug design application in mind, since they are overwhelmingly most common constituents of proteins and their ligands (pharmaceutics). In the COSMO2 publication (reference [56]), radii from the original COSMO implementation in MOPAC, which were used also as the starting parameters for our reparametrization, were accidentally listed as for COSMO, rather than those of the latest version (parametrized ones, which were actually used in the calculations).

The radius of the hydrogen was significantly reduced by the optimization, while other radii became larger (Tab. 1). The PM6/COSMO2 optimization without the nonpolar term yielded about 0.8 kcal/mol greater RMSE result (compared to the 2.65 kcal/mol, Tab. 2), so its inclusion is significant.

Table 1: The atomic radii (in Ångstroms) used in the COSMO implementation in MOPAC, the starting parameters for our reparametrization and the newly optimized COSMO2 parameters for PM6 and PM7. The effective surface tension parameter ξ (kcal/mol/Å^2).

	COSMO	Startpoint	PM6/COSMO2	PM7/COSMO2
H	1.30	1.08	0.828	0.929
C	2.00	1.53	1.821	1.699
N	1.83	1.48	1.904	1.913
O	1.72	1.36	1.682	1.686
P	2.13	1.75	2.118	2.242
S	2.16	1.70	2.369	2.346
F	1.72	1.3	1.602	1.528
Cl	2.05	1.65	1.911	1.901
Br	2.16	1.8	2.178	2.211
I	2.32	2.05	2.276	2.062
ξ	—	0.05	0.046	0.042

Table 2: Errors in solvation free energies (as RMSE relative to the experimental values, in kcal/mol) in the training set (MNSol*) and four validation sets. C-PCM' denotes the C-PCM model excluding the default cavitation and dispersion terms.

Methods	MNSol*	Mobley266	SAMPL1	SAMPL4	C10
PM6/COSMO2	2.65	2.81	5.08	2.44	2.18
PM6/COSMO	4.31	3.72	9.07	4.01	2.88
PM6/PCM	6.35	2.24	3.93	1.92	8.38
PM6/SMD	4.80	3.03	4.91	2.59	8.39
PM7/COSMO2	2.62	2.54	3.73	1.92	2.28
PM7/COSMO	3.96	3.44	6.07	3.21	2.87
AM1/COSMO	4.80	2.26	8.89	2.26	4.68
DFTB3/C-PCM	9.92	6.40	12.87	8.96	12.55
DFTB3/C-PCM'	6.16	2.31	3.66	2.84	8.01
DFTB3/SMD	4.86	2.65	4.85	3.15	5.57
DFT/COSMO-RS	2.54	1.08	1.88	1.59	3.57

4.4 Results

4.4.1 Performance on Datasets of Small Molecules

We included a higher-level DFT-D3/COSMO-RS for the comparison to solvation models used with SQM methods. Interestingly, the COSMO-RS was outperformed by some of the simpler solvation models for charged species (Tab. 2, the C10 is composed of ions and the training dataset MNSol* includes ions – Fig. 3). The optimization of atomic radii parameters with the newly included nonpolar term for PM6 and PM7 alike did consistently increase the accuracy of the COSMO(2) model across the datasets (Tab. 2). PM7/COSMO had been slightly more accurate than PM6/COSMO before the COSMO2 optimization and PM7/COSMO2 was still more accurate than PM6/COSMO2 after the optimization. PM7/COSMO2 was among the most accurate method/model combinations tested overall. Some SQM methods delivered slightly better estimates than PM7/COSMO2, but not consistently throughout the datasets – especially for datasets with charged species does PM7/COSMO2 (followed by PM6/COSMO2) best of tested methods. This is understandable, since ions compose a significant portion of the training dataset. The error reduction upon COSMO2 optimization is largest for cations (Fig. 3). Interestingly, the improvement of the COSMO2 optimization was the most prominent for the SAMPL1, neutral dataset. Several compounds of SAMPL1 which produced a very large deviation (overly negative energies) for most models tested, often contained phosphorodithioic or other thio-phosphate containing groups. Errors for these compounds were significantly reduced by the COSMO2 optimization, too.

PM6/PCM was surprisingly accurate for datasets without charged species. Together with the COSMO2 parametrization for the PM7, it was the best performing SQM-paired solvation model for those datasets, despite the nonpolar terms are not included in this implementation. As it has been argued in the

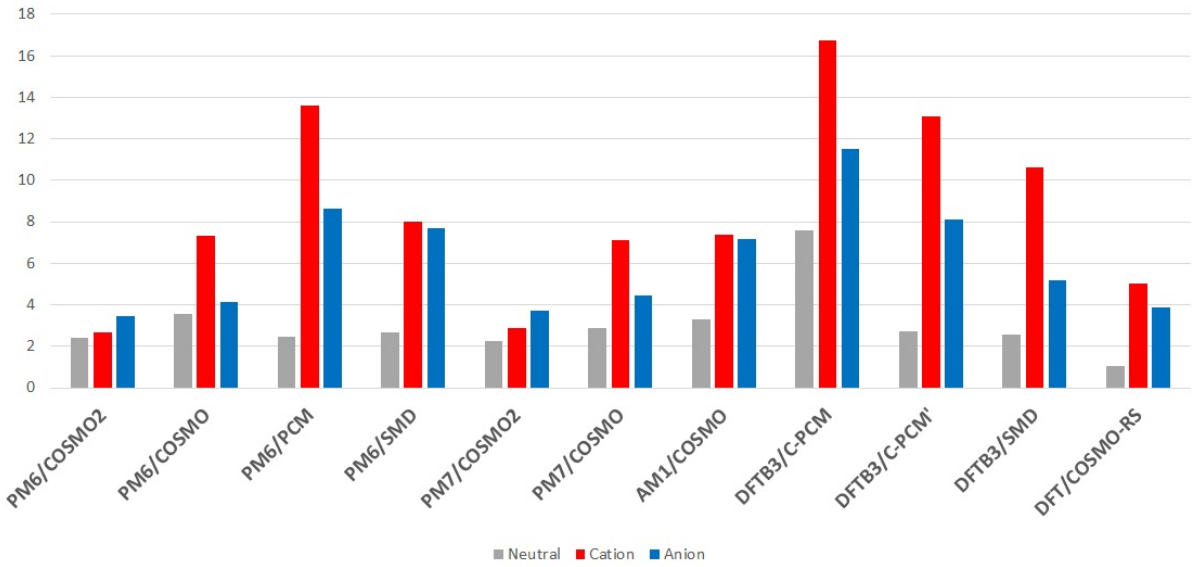

Figure 3: The accuracy of SQM methods combined with solvation models (and QM DFT paired with COSMO-RS) tested on the COSMO2 training dataset MNSol* (in RMSE, kcal/mol), the dataset is separated into groups by the compound charge.

COSMO2 publication (reference [56]), the reason for this could be the error compensation between missing polarization in the SQM method (i.e. missing stabilization of a system) and missing cavitation and repulsion positive terms in the solvation model (i.e. missing destabilization of a system). To further this argument, C-PCM without nonpolar terms (C-PCM') with DFTB3 in GAMESS delivered more accurate estimates than with those terms included (2) - the difference was most significant for neutral compounds. In these cases, reduction of mean signed error (a measure of systematic deviation) from overly positive values, goes indeed hand in hand with reduction of RMSE upon turning nonpolar terms off.

SMD model is comparable to PCM, but is not originally developed for SQM calculations. Despite of the nonpolar term inclusion, PM6/SMD (GAUSSIAN) performed worse than PM6/PCM (GAUSSIAN) except for MNSol*,

which nearly coincided with the training set for the SMD development, and C10 with ions. Similarly, DFTB3/SMD in GAMESS did worse than C-PCM' without nonpolar terms in some of tested datasets. Although the authors of the SMD model highlighted the portability of the model among electronic structure methods [86], clearly, this can result in a loss of accuracy. We recommend method-specific parametrization and thus from this point of view, high-quality benchmark datasets are of crucial importance.

4.4.2 Performance on Large Protein-Ligand Complexes

We ultimately aim for the application in drug design, on complexes of proteins and pharmaceutics. Thus we tested the newly parametrized method embedded in a scoring function on 10 complexes of a carbonic anhydrase II with a sulfoamide-containing analogues [95]. Since the experimental affinities of the complexes are know, we correlated these with the calculated interaction energies in the solvent. These were expressed by the scoring function as a sum of two terms – a change of energy upon the protein-ligand complex formation in the vacuum (calculated by a SQM method) and a change of solvation energy upon the complex formation (achieved with the PM6 or PM7 SQM method paired with a solvation model).

$$score = \Delta E_{int} + \Delta\Delta G_{solv}. \qquad \text{(EQ.32)}$$

For these particular high-resolution structures, position of water molecules are known. As those may somehow interfere with the implicit solvent models and are usually not attainable, we provide corresponding results for structures with water molecules removed (Fig. 5).

Although PM7/COSMO2 in this experimental setup did not consistently outperform PM6/COSMO2 (PM6-D3H4/COSMO2 is slightly more accurate

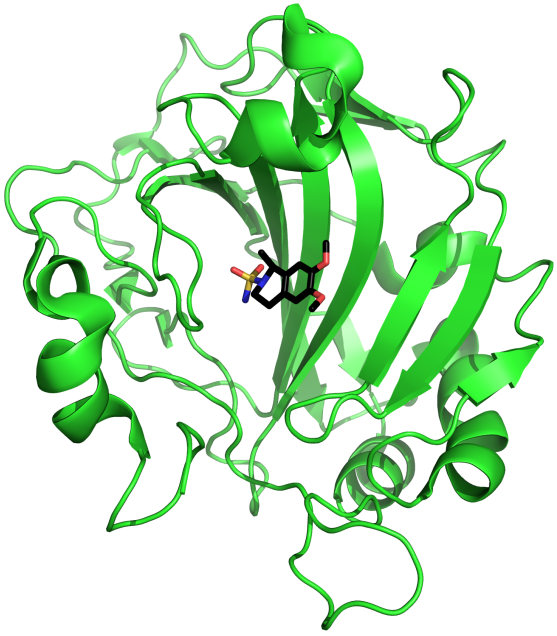

Figure 4: Carbonic anhydrase II (PDB code: 3PO6) [101] with a sulfonamide containing ligand.

than PM7/COSMO2 for complexes without explicit water molecules), COSMO2 reparametrization either way increased the correlation between the score and the experiment, more so for the complexes without the explicit water molecules (Fig. 5).

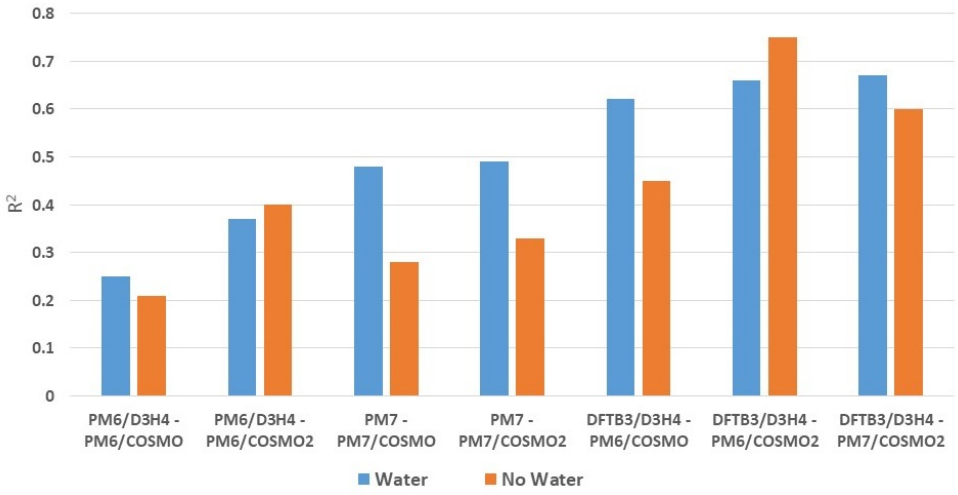

Figure 5: Coefficient of determination (R^2) in the series of carbonic anhydrase II inhibitors for scoring functions combining different methods for the calculation of gas-phase interaction energy (ΔE_{int}) and the solvation interaction energy ($\Delta\Delta G_{solv}$). Results obtained with and without explicit structural waters are listed.

5 The Development of Protein-Ligand Derived Benchmark Datasets

This "Dataset" part presents the development of PLF547 (Protein Ligand Fragment, 547 complexes) and PLA15 (Protein Ligand Active site, 15 complexes) datasets with benchmark reference interaction energies and solvation energies computed on high computational levels.

Surprisingly, not many databases or datasets are devoted to protein-ligand systems. There are several dataset of interacting small organic molecules [71, 102] and there is a protein-ligand derived dataset [103], which is, however, limited in its size and variability (one ligand and its fragmented protein environment). Recently a large GMTKN55 [104] multi-purpose dataset containing also PCONF21 [105] dataset with high-level QM conformational energies of amino-

acid oligomers has been published. A MPCONF196 dataset includes also conformers of several larger macrocycles (representing common ligands, but without sulfur, phosphorus or halogen atoms) [78]. An argument was made in the COSMO2 part for the need of good-quality reference data serving the development of SQM methods (or solvation models).

Here, we pursue this motivation further by constructing large datasets for noncovalent interactions devoted specifically to protein-ligand complexes and derived systems with the aim of testing and development of SQM methods for *in silico* drug design. Any semiempirical parameters or corrections for SQM methods obtained on these datasets should then be well suited for the description of systems relevant to *in silico* drug design, biochemistry and related applications.

5.1 The Complexes Constituting the Datasets

The complexes of the two datasets devised were based on structures which we obtained from reference [32]. These are the structures of 17 medically relevant protein-ligand complexes of several protein families with a resolution below 2.5 Å taken originally from the Protein Data Bank. The preparation of structures done earlier (ref. [32]) included removal of crystal waters, protonation (addition of hydrogen atoms) and disulfide bond assignment. Ligands are up to 100 atoms in size, with charge of 0, +1 or -1, all contain aromatic rings, some of them sulfur and some are halogenated (Fig. 6).

The PLA15 consists of 15 ligands interacting with its protein surroundings, amino acids up to 4 Å apart (Fig. 7). Two of 17 complexes were excluded because reliable benchmark data were not achieved due to the self interaction error (see below) as a DFT is used as a basis of the reference benchmark (sections 3.4, 3.5.2). Additional two amino acid residues were excluded from the selection due to intermolecular clashes (tyrosine 139 of 2YKI and lysine 98 of 2VW5).

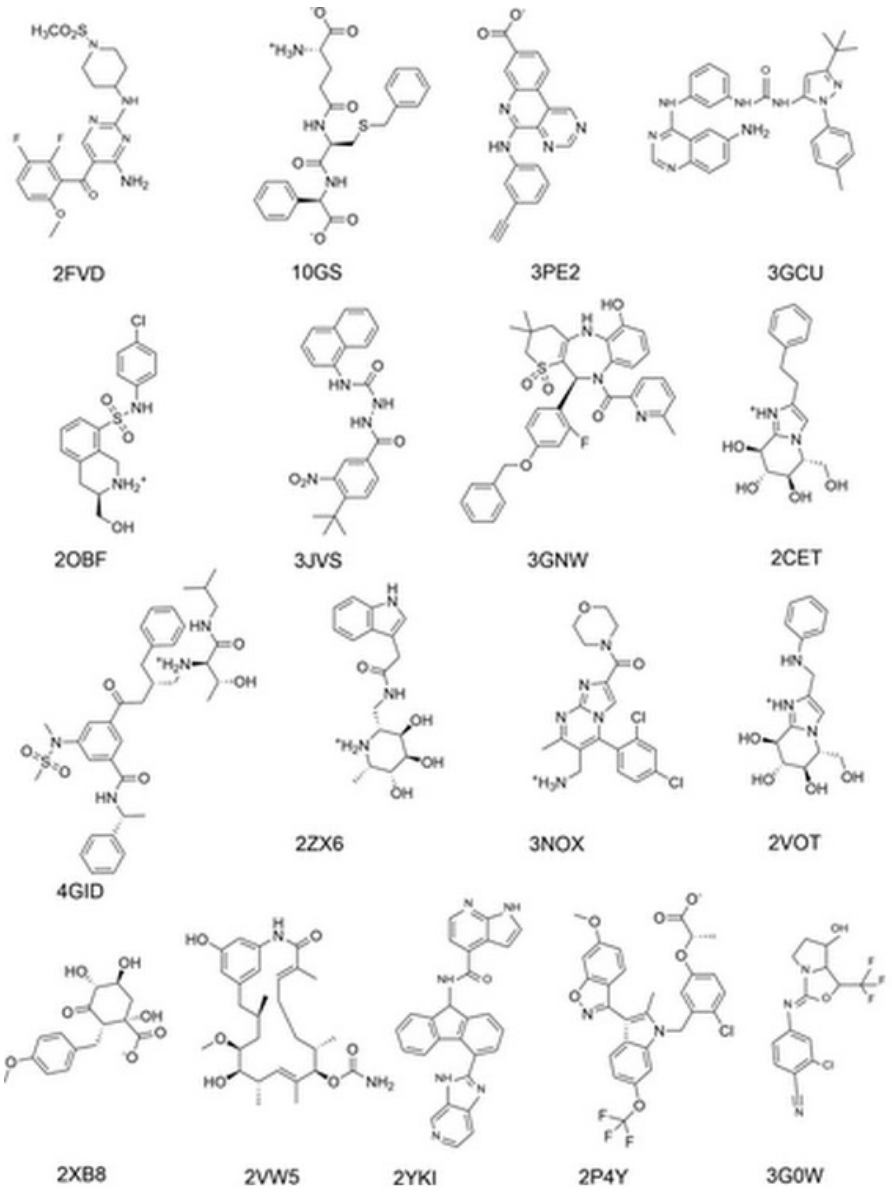

Figure 6: Ligands of protein structures from which the datasets were prepared. PDB codes are listed. 3GCU and 3JVS were excluded from the PLA15 dataset due to large self-interaction error (as discussed in section 5.2.1). Figure taken from ref. [32].

PLF547 consists of complexes of the same 17 ligands interacting with each protein fragment (backbone or amino acid side chain) within 4 Å of the respective ligand at once (Fig. 8). Two fragments corresponding to residues removed from PLA15 due to steric clashes were also excluded from PLF547. In other words, the protein part from PLA15 (and also from the two of original 17 active site-ligand complexes not included in PLA15) have been fragmented to individual amino acids or backbone parts, each one forming a complex with its respective ligand. Bonds between carbonyl and C_α as well as between C_α and C_β (except for proline, glycine) carbons atoms were truncated and terminated with hydrogen atoms, yielding N-methylformamide (backbone model) or side chain models, respectively. This fragmentation ultimately yielded 547 complexes of protein fragments interacting with their respective ligands. The benchmark interaction energy values have been calculated by composite schemes of QM methods. For benchmark solvation energies, we used DFT/COSMO-RS. Out of the remaining complexes, 8 additional fragments most affected by the self interaction error were excluded from the analysis of solvation interaction energies, since our benchmark to those is DFT based (DFT/COSMO-RS). The datasets were provided as supplementary material of publication in reference [23]. The performance of SQM methods on the newly devised dataset hinted at avenues of future improvement or cases when caution is due.

Trends in the accuracy of methods tested on the PLF547 dataset have been analyzed by separating its constituents into 4 subsets by the distance between interacting partners (below 90 % of the van der Waals radii sum of closest atoms – short, 90-110 – equilibrium, 110-130 – long and above 130 % – distant), 3 subsets by the total charges of interacting partners (both charged, one neutral, both neutral) and into 6 subsets by the type of protein residue fragments (backbone

– N-methylformamide; aromatic – His, Trp, Phe, Tyr; other nonpolar – Ala, Val, Ile, Leu, Met; polar – Ser, Thr, Asn, Gln, Cys, Pro; anion – Glu, Asp and cation – Arg, Lys, protonated His).

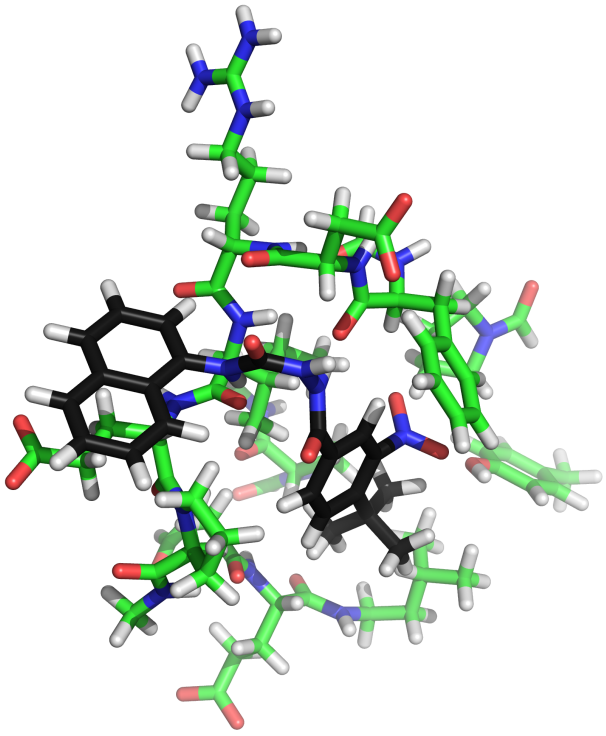

Figure 7: A complex derived from 3JVS Protein Data Bank structure, one of the PLA15 active site complexes. Carbon atoms of a protein part in green, of a ligand in black.

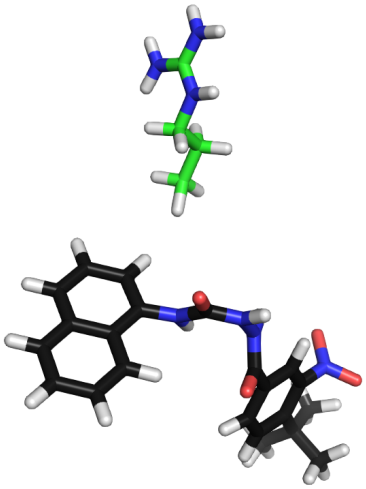

Figure 8: One of the PLF547 fragment complexes derived from 3JVS.

5.2 Results

5.2.1 PLF547, Interaction Energies

Here we present the accuracy in interaction energy estimates of selected SQM and QM methods on the PLF547 dataset grouped by the distance between interacting partners (Fig. 9), by the type of a protein fragment (Tab. 3) and by the total charge of interacting moieties and overall (Fig. 10).

Trends in Accuracy for the Complexes Grouped by the Distance of Interacting Partners We have found out that DFT and DFTB methods provided large errors (interaction overestimation) for several systems that contained a charged protein fragments further away from the ligand (Fig. 9, equilibrium, long and distant separations). This is contrary to expectation that the error should become smaller with the increasing separations, since the absolute magnitude of interaction is smaller. We have attributed this coounterintuitive results to the self-interaction error (SIE) of the DFT approach, when the density

interacts with itself unphysically.

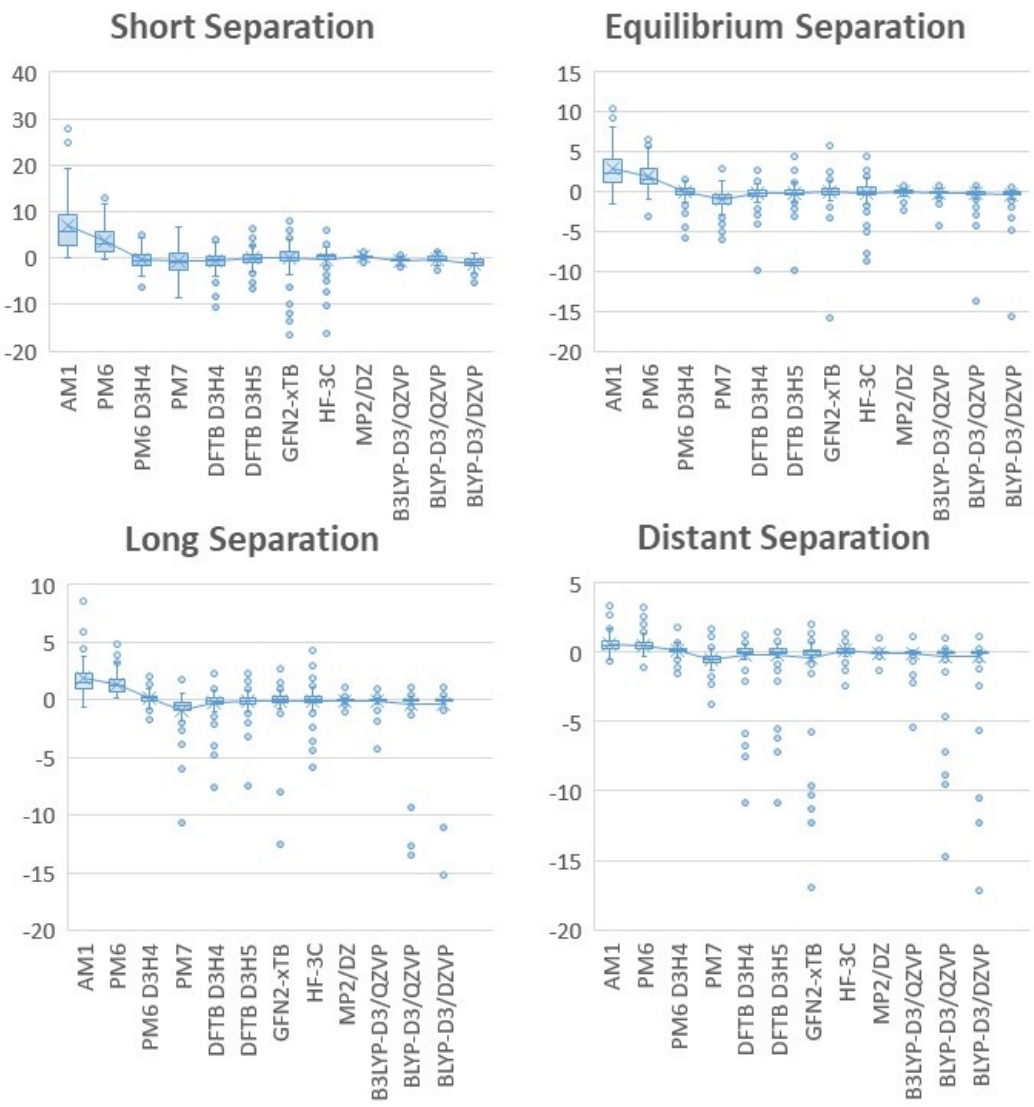

Figure 9: The error of methods on PLF547 complexes divided into groups by interacting partner separation. Even though the overall error decreases with the separation, several outliers with large errors emerge. Notably, this is true for the DFT and DFTB methods. Even AM1 – the least accurate of SQM methods presented here (Tab. 10), is not encumbered by such outliers on greater separations.

Table 3: The RMSE of the methods tested on the PLF547 dataset, in kcal/mol. Complexes sorted by the type of the protein fragment.

Methods	Backbone	Aromatic	Nonpolar	Polar	Anion	Cation
MP2/aug-cc-pVDZ	0.27	0.65	0.27	0.48	1.05	0.59
B3LYP-D3/QZVP	0.26	0.45	0.22	0.35	1.68	0.80
BLYP-D3/QZVP	0.30	0.50	0.24	0.36	4.97	3.31
BLYP-D3/DZVP	0.41	0.72	0.25	0.58	6.31	1.56
AM1	2.61	5.72	2.64	4.50	12.30	4.96
PM6	1.44	3.47	1.92	2.37	6.38	4.18
PM6-D3H4	0.73	1.33	0.40	1.47	2.13	1.89
PM7	1.41	2.58	0.90	1.88	3.00	3.65
DFTB3-D3H4	0.66	1.15	0.39	1.45	4.89	2.86
DFTB3-D3H5	0.61	1.12	0.40	1.05	4.10	2.55
GFT2-xTB	0.63	1.31	0.46	0.93	6.88	3.63
HF-3C	0.84	1.69	1.99	2.32	7.40	2.44

Trends in Accuracy for the Complexes Grouped by the Protein Fragment Type Although less apparent than the self-interaction error of DFT-based methods, a few trends seem noticeable in the error distribution in systems separated by the kinds of protein fragments. Firstly, aromatic systems seem to have rather big error, comparable to that of ions, even though absolute interaction energies among ions are much larger (Tab. 3). For example, for AM1, interaction involving aromatic protein residue have even slightly bigger RMSE than those involving cations. Additionally, HF-3C seems to overestimate interaction involving sulfur and likely also interactions with halogenated ligands. For anion-fragment complexes, HF-3C (GFN2-xTB and some other methods to lesser extent) provide larger error caused by a slight overestimation of interaction of several systems, while the cationic complexes are described better. Relatively larger error of PM7 for "nonpolar" and "backbone" complexes may be due to the inaccurate description of hydrogen bonds.

The Accuracy for the Complexes Grouped by the Net Charges and Overall Out of QM methods, MP2 results (also used for the benchmark) and

B3LYP DFT are very accurate with RMSE 1 kcal/mol for the interaction between charged systems and about 0.25 kcal/mol for neutral ones (Fig. 10). The error of the non-hybrid BLYP DFT functional (with two basis sets) is significantly higher for other than neutral systems - higher than the error obtained with some of SQM methods. This error of non-hybrid functionals is caused by the SIE (which is apparent from Fig. 9 for larger separations). Observed dependencies on the functional and on separation can be found in the literature [106, 107]. Overall, PM6-D3H4 provided the most accurate results for SQM methods,

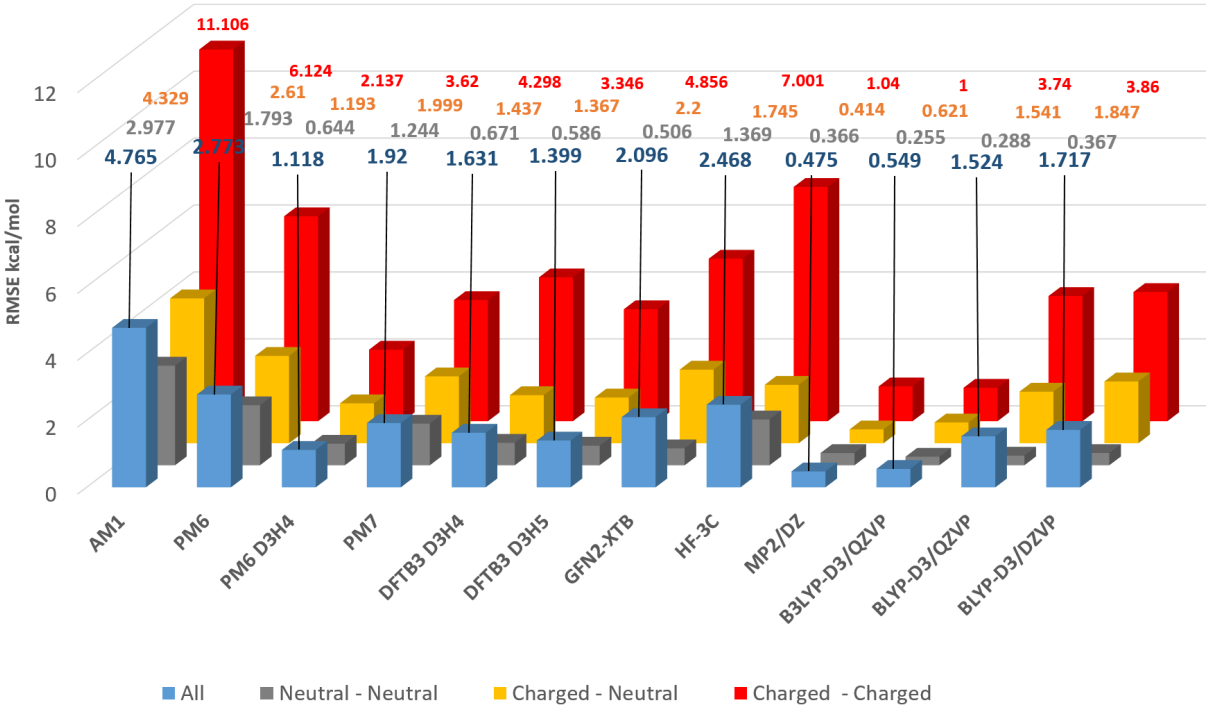

Figure 10: The root mean square error (RMSE) in the interaction energy estimates given by the tested methods on the PLF547 fragment data set relative to the benchmark. Four columns represent complexes divided into groups where both interaction partners are ions (red), one neutral, one ion (yellow) or both neutral (gray). Blue column is an overall RMSE for all the 547 fragments. MP2 had been calculated in aug-cc-pVDZ basis. In the DFT calculations, QZPV stands for def2-QZVP and DZVP for DZVP-DFT basis sets, respectively.

judging by its RMSE, with DFTB3-D3H4 and DFTB3-D3H5 following (Fig. 10, blue columns). Without the 8 systems with the large SIE (which were removed from the solvation interaction energy analysis, section 5.1), all the DFT-based SQM methods yield better results and DFTB-D3H5 becomes the most accurate, slightly more so than PM6-D3H4 (both methods yielding about 1.1 kcal/mol RMSE in this setup). PM7, GFN2-xTB and HF-3C were the least accurate, however without the 8 SIE outlier systems, GFN2-xTB performs better (about 1.35 kcal/mol RMSE as compared to 2.1, Fig. 10). GFN2-xTB is also the most accurate SQM method for neutral-neutral type complexes (which were also less affected by the SIE error, Fig. 10). PM7 overestimates the interaction slightly (Fig. 9), likely due to overly attractive dispersion term [108].

5.2.2 PLF547, Solvation Interaction Energies

Here we describe the performance of the selected solvation models used with SQM methods in terms of solvation interaction energy accuracy compared to DFT/COSMO-RS. The 8 self-interaction error systems have been removed, since the reference benchmark is DFT-based. We include another QM-based solvation model HF/6-31G*/SMD. The important notion is, that COSMO-RS solvation model is not as reliable benchmark for solvation energy as the correlated QM methods for interaction energy are. It is thereby necessary to be cautious about the results presented in this section. The same applies for PLA15 solvation energy references composed of these COSMO-RS benchmark values presented further on (section 5.2.4). For example, QM-based HF/SMD gave comparable results to SQM methods. But given the estimated benchmark error is comparable to errors provided by the tested methods, the results may be much different when compared to the "true" solvation energy, which is inaccessible. As we have shown in COSMO2 part, COSMO-RS gives about 1 kcal/mol RMSE with respect to experimental solvation energies for neutral compounds

Table 4: The RMSE of the energy change upon the formation of the fragment–ligand complex in kcal/mol on the PLF547 complexes for the combinations of SQM or QM methods and solvation models tested. The results listed overall and grouped by the net charge of interacting partners.

Methods	Ion-Ion	Ion-Neutral	Neutral-Neutral	All
HF/6-31G*/SMD	4.67	2.29	1.25	2.21
PM6/COSMO	6.06	2.11	1.26	2.35
PM6/COSMO2	4.18	1.83	1.24	1.89
PM7/COSMO	5.83	2.03	1.26	2.27
PM7/COSMO2	4.31	1.82	1.25	1.91
DFTB3/SMD	4.29	2.02	1.08	1.97
DFTB3/PCM	8.0	3.43	3.27	3.87

(3) – already comparable with the error SQM and HF/SMD methods provide here, with respect to COSMO-RS for neutral compounds (Tab. 4).

With the above noted precaution in mind, all these methods give similarly accurate estimates, aside from DFTB/C-PCM, which was off the mark for either neutral and charged systems, as well. This is consistent with the results in the COSMO2 part, where the method performed poorly, too. The effect of exaggerated "repulsion" is indicated by a positive and significant mean signed error (2.3 as opposed to below 0.5 kcal/mol from other methods). Without the nonpolar terms, the methods improve the accuracy by about 1 kcal/mol RMSE (from 3.8 to 2.8), while mean signed error drops to -0.8 (now, in an opposite direction – slight overestimation). There seems to be an imbalance in nonpolar terms of the PCM implementations in GAMESS. COSMO2 seems to be an improvement over COSMO in this setup, too. Again, the difference is apparent for charged systems.

5.2.3 PLA15, Interaction Energies

In this section, the errors of interaction energies on ligands complexed with their respective protein surroundings (PLA15) obtained by selected SQM methods are provided. The PLA15 complexes are designed to represent faithfully the

protein-ligand interaction. These results are thus of immediate significance for drug design.

In the presented comparison we exclude HF-3C and most of other QM methods due to size of the systems. The QM BLYP/DZVP-DFT interaction energies, as have been used for the calculation of the PLA15 benchmark (*Methods*, section. EQ.26), are shown for the comparison with SQM methods. We discuss the errors in % of kcal/mol relative to the benchmark values (Tab. 5, "benchmark" column).

AM1 and PM6 methods without semiempirical corrections underestimated the interaction for all the complexes, with the average error 55% of the benchmark interaction energy (Tab. 5). Clearly, errors of this magnitude are unacceptable. PM7 showed an improvement over AM1 and PM6 (for PM7 the interaction is overestimated, AM1 and PM6 underestimate it), but was outperformed by all the other SQM methods. Other SQM methods – PM6-D3H4, DFTB2-D3H4, DFTB3-D3H5 and GFN2-xTB – the most accurate of them, are roughly comparable, with the error about 10% of a total interaction energy. The performance of these SQM methods is comparable to that of BLYP/DZVP-DFT. Interestingly, out of these methods, only GFN2-xTB underestimated the interaction. A systematic shift in error is apparent for the methods. Such a behaviour can be easily corrected for. Moreover, in *in silico* drug design often only the relative values are of importance, since systematic deviation cancels out when comparing predictions given by a single method for multiple systems. When the average deviation is subtracted, we obtain a random, nonsystematic part of an error (Tab. 5, absolute random error). For this setup, DFTB3-D3H5 yielded the most accurate results of SQM methods.

Table 5: The errors for the interaction energies of the 15 active site complexes in percentage of the benchmark interaction energies (listed in the last column), the average % error for a method, the total RMSE for a method in kcal/mol and the absolute random % error of a method. The latter are presented averaged for all the systems and they represent a random part of a deviation without a systematic error ("absolute" meaning positive values). Aside from BLYP/DZVP-DFT, semiempirical methods only are shown. Net ligand charges are in parentheses.

Methods, interaction energy error, %								
Complexes	AM1	PM6	PM6-D3H4	PM7	DFTB3-D3H4	DFTB3-D3H5	GFN2-xTB	DFT
10GS (-1)	34.3	19.2	-7.2	-8.8	-8.1	-4.1	11.6	-5.8
2CET (+1)	34.5	14.0	-10.9	-16.2	-6.8	-7.0	2.5	-4.6
2FVD (0)	67.2	38.2	-13.7	-34.8	-15.8	-11.6	0.2	-12.2
2OBF (+1)	34.9	18.1	-10.2	-23.7	-12.3	-9.1	2.2	-7.1
2P4Y (-1)	49.2	30.7	-2.3	-18.1	-2.9	-3.2	14.3	-6.8
2VOT (+1)	39.3	14.1	-12.6	-21.3	-8.7	-3.8	2.1	-6.2
2VW5 (0)	69.2	40.3	-15.1	-28.3	-12.1	-8.0	14.5	-10.6
2XB8 (-1)	52.5	28.0	-3.1	-8.2	-6.2	-6.6	8.9	-6.6
2YKI (0)	90.5	54.2	-7.2	-37.8	-7.0	-5.2	21.4	-6.2
2ZX6 (+1)	36.7	16.6	-13.5	-16.5	-17.9	-9.7	0.5	-5.9
3G0W (0)	98.4	66.8	2.8	-29.6	-8.4	-10.4	23.7	-5.1
3GNW (0)	99.8	62.5	-9.8	-46.6	-14.6	-13.7	8.7	-13.8
3NOX (+1)	20.7	8.2	-13.5	-21.8	-12.7	-5.6	-0.7	-4.7
3PE2 (-1)	54.5	33.2	-2.8	-10.0	-12.2	-8.5	3.9	-3.2
4GID (+1)	56.8	28.0	-21.5	-39.6	-20.8	-19.2	7.2	-8.7
Average %	55.9	31.5	-9.4	-24.1	-11.1	-8.4	8.1	-7.2
Methods, RMSE (kcal/mol)								
Total	79.1	42.1	21.0	38.8	21.4	15.2	13.4	11.9
Methods, absolute random error, %								
Average %	19.5	14.2	4.9	9.6	3.9	3.1	6.2	2.2

5.2.4 PLA15, Solvation Interaction Energies

we show the error (in % of kcal/mol) of selected solvation models relative to the benchmark on PLA15 dataset. We took the reference interaction energies in the solvent as a sum of COSMO-RS energies for the fragments constituting the respective active site. This approach neglects any nonlinear (polarization related) effect altogether, however, any better solvation (interaction) energy reference for large protein complexes would be hard to achieve. All the methods tested were calculated in this setup.

HF/SMD overestimated the interaction in solvent significantly and systematically – for all the complexes, compared to the COSMO-RS reference. Most notably, DFTB3/C-PCM produced very large errors, especially for the neutral ligands (Tab. 6). Upon closer inspection, this error is generated by the summing up of interaction energies of fragments that are systematically shifted to positive values (interaction underestimation). Since the approximate benchmark for PLA15 solvation is a mere sum of COSMO-RS solvation energies of fragments, the systematic underestimation is more apparent in this setup. For charged ligands (at least one of the partners is an ion), the effect is less apparent. Here, COSMO2 seems to be slightly more accurate than respective COSMO variants, too (Tab. 6, RMSE).

Table 6: The errors of solvation energy change upon protein-ligand complexation of chosen SQM methods (and HF/6-31G*/SMD) for the 15 active site complexes expressed as a percentage of a reference benchmark (listed in the last column), the average % error for a method, the total RMSE for a method in kcal/mol. The solvation energy changes of active site complexes are approximated by summing up all the respective fragment energies. Net ligand charges are in parentheses.

	Methods, interaction energy error, % of benchmark energy							benchmark
Complexes (lig. charge)	PM6/ COSMO	PM6/ COSMO2	PM7/ COSMO	PM7/ COSMO2	DFTB3/ SMD	DFTB3/ PCM	HF/ SMD	abs. energy kcal/mol
10GS (-1)	-10.3	-6.5	-8.7	-6.3	-11.0	1.3	-23.3	288.2
2CET (+1)	-6.9	-0.8	-6.8	-1.4	-6.5	5.4	-13.6	297.5
2FVD (0)	30.7	21.6	33.7	30.3	9.2	103.1	-27.6	90.1
2OBF (+1)	13.4	11.6	12.8	10.3	4.0	36.9	-8.2	214.9
2P4Y (-1)	24.1	15.8	24.7	19.3	13.9	96.9	-16.4	135.8
2VOT (+1)	-1.3	6.7	-0.8	6.5	-3.3	7.2	-10.2	243.9
2VW5 (0)	4.3	2.6	5.1	6.7	0.2	77.4	-37.1	109.4
2XB8 (-1)	3.4	4.0	7.1	9.2	6.8	47.0	-22.1	162.5
2YKI (0)	4.7	15.5	5.9	15.3	-8.7	146.2	-49.1	73.5
2ZX6 (+1)	3.3	5.4	4.1	6.8	-0.8	12.8	-10.7	284.4
3G0W (0)	27.2	27.4	32.8	31.8	0.2	199.1	-57.0	42.0
3GNW (0)	26.8	21.1	22.6	17.4	-1.3	144.7	-54.4	71.4
3NOX (+1)	8.7	6.3	8.0	5.6	-0.7	17.7	-11.1	226.6
3PE2 (-1)	-8.3	-13.9	-5.9	-10.0	-13.7	49.1	-38.5	126.0
4GID (+1)	-10.4	-10.8	-11.4	-11.4	-21.6	25.2	-48.6	297.5
Average %	7.3	7.1	8.2	8.7	-2.2	66.7	-28.5	
	Methods, RMSE (kcal/mol)							
Total	20.1	17.1	20.2	18.5	20.9	76.3	51.2	

6 Conclusions

6.1 COSMO2

In the COSMO2 part, we have shown that one can achieve a reasonable accuracy improvement of a solvation model by means of relatively simple optimization. Changes to the underlying COSMO included reparametrization of parameters for element radii and addition of a function containing an extra parameter. In this way, the method can still resolve large systems very quickly – COSMO2 can readily be used along with MOZYME in MOPAC, and in this way, the energy of a 1000 atom system takes about 2 minutes to calculate). Solvation energies are larger for ionic species, and solvation energy estimates of solvation methods are correspondingly large for ionic species. However, when training dataset includes charged moieties, the parametrization can absorb a substantial amount of this error, as we have seen for the COSMO2 datasets. Newly parametrized methods outperformed even more elaborate COSMO-RS/DFT for charged species.

The accuracy of solvation models depends strongly on how and with which SQM method the model is implemented. We have attained better results for DFTB3 without the default nonpolar term than when we included it (likely due to error cancellation). When using solvation models, it appears that the efficacy of various parametrizations for concrete SQM methods may not be fully transferable among various implementations.

These results stress the significance of a robust and encompassing datasets.

6.2 PLF547/PLA15 Datasets

For the development of datasets for noncovalent interactions, PLF547 and PLA15, our task was simpler in a way that for (interaction) energies, there are high-level QM approaches which are trusted enough to provide energy estimates close to

true experimental energies. On the other hand, this is not the case for solvation models. In this work, we used MP2 and CCSD(T)-DLPNO to produce accurate reference data of interaction energies for small 547 systems. For a dataset of whole 15 active site models, we introduced an approximation, to get as close to reliable energy estimates as possible. Our subsequent study of SQM methods on the new datasets showed that both PM6 and DFTB3 with D3H4 (or D3H5) corrections yielded very accurate interaction energy estimates, advocating the effectiveness of empirical corrections. We show that SQM methods with empirical corrections can provide results of accuracy approaching those of DFT methods on the complexes studied. Furthermore, we have discovered some error trends specific to a SQM method for specific system types. For instance, we have encountered a density self-interaction error produced by DFT and DFT-based SQM methods when applied to several complexes of developed datasets. The error of these systems had a significant effect on a statistical analysis of method accuracy. Although these seem to be functional-dependent (in case of DFT) regarding their magnitude, the methods reproduce the error consistently for these specific systems (for which other methods provide relatively accurate estimates). Some other weaker trends were also noticeable for some of SQM methods.

With a better knowledge of weak sides of concrete SQM methods, caution is due when using them on high-risk systems until semiempirical corrections are devised to cover them.

When considering the results of both parts of this work together, the absolute errors in estimates of solvation energies with solvation models currently in use are larger that the errors of interaction energies of comparable SQM methods (comparing COSMO2 results to interaction energy accuracy of SQM methods

tested in the Dataset part). For the purpose of *in silico* drug design, it is thereby advantageous to focus preferentially on the accuracy of solvation energy estimation, since it has the same impact on the overall score, ligand binding affinity prediction, as the interaction energy does. In the Dataset part, where we used DFT/COSMO-RS as a reference for a solvation energy part of the study, COSMO2 seems to outperform standard COSMO for charged species, confirming the observations made in the COSMO2 part.

7 Future Perspectives

Regarding COSMO2, preliminary results indicate, that it is possible to improve on the model further by implementing parameters for an atom in a specific environment (group). It seems like some molecular groups produce larger solvation energy error than others, as indicated by a large error spread in SAMPL1 dataset. The dependence on valence state has been observed for parameters for dispersion correction [9]. Arguably, implementing different atom types for solvation models would be computationally inexpensive way to improve the accuracy. The nonpolar term can likewise be improved by using atom-type specific surface, as in case of SMD nonpolar term. The inclusion of the COSMO2 in the scoring function developed in our group is being tested.

The work on additional dataset of small organic moieties is already in progress designed to introduce corrections to SQM methods for specific interactions. The correction functions for atom-pair-specific repulsion potential have been proposed.

References

(1) Levine, I. N., *Quantum Chemistry*, 5th ed.; Prentice Hall, Inc.: Upper Saddle River, New Jersey, 2000.

(2) Koskinen, P.; Mäkinen, V. *Comput. Mater. Sci.* **2009**, *47*, 237–253.

(3) Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M. *Chem. Rev.* **2016**, *116*, 5301–5337.

(4) Liakos, D. G.; Sparta, M.; Kesharwani, M. K.; Martin, J. M. L.; Neese, F. *J. Chem. Theory Comput.* **2015**, *11*, 1525–1539.

(5) Raha, K.; Peters, M. B.; Wang, B.; Yu, N.; Wollacott, A. M.; Westerhoff, L. M.; Merz, K. M. *Drug Discov Today* **2007**, *12*, 725–731.

(6) Ciancetta, A.; Genheden, S.; Ryde, U. *J. Comput. Aided Mol. Des.* **2011**, *25*, 729–742.

(7) Fanfrlík, J.; Brahmkshatriya, P. S.; Řezáč, J.; Jílková, A.; Horn, M.; Mareš, M.; Hobza, P.; Lepšík, M. *J. Phys. Chem. B* **2013**, *117*, 14973–14982.

(8) Kotev, M.; Sarrat, L.; Gonzalez, C. D. *Methods Mol. Biol.* **2020**, *2114*, 231–255.

(9) Wu, Q.; Yang, W. *J. Chem. Phys.* **2001**, *116*, 515–524.

(10) Ahlrichs, R.; Penco, R.; Scoles, G. *Chem. Phys.* **1977**, *19*, 119–130.

(11) Toennies, J. P. *Chem. Phys. Lett.* **1973**, *20*, 238–241.

(12) Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(13) Grimme, S. *J. Comput Chem.* **2004**, *25*, 1463–1473.

(14) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104.

(15) Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. *J. Chem. Phys.* **2019**, *150*, 154122.

(16) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149–5155.

(17) Martin, B.; Clark, T. *Int. J. Quantum Chem.* **2006**, *106*, 1208–1216.

(18) Jurečka, P.; Černý, J.; Hobza, P.; Salahub, D. *J. Comput. Chem.* **2007**, *28*, 555–569.

(19) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173–1213.

(20) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

(21) Řezáč, J.; Fanfrlík, J.; Salahub, D.; Hobza, P. *J. Chem. Theory Comput.* **2009**, *5*, 1749–1760.

(22) Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2012**, *8*, 141–151.

(23) Kříž, K.; Řezáč, J. *J. Chem. Inf. Model.* **2020**, *60*, 1453–1460.

(24) Korth, M.; Pitoňák, M.; Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 344–352.

(25) Kolář, M. H.; Hobza, P. *Chem. Rev.* **2016**, *116*, 5155–5187.

(26) Řezáč, J. *J. Comput Chem.* **2019**, *40*, 1633–1642.

(27) Řezáč, J.; Hobza, P. *Chem. Phys. Lett.* **2011**, *506*, 286–289.

(28) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3094.

(29) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(30) Gohlke, H.; Klebe, G. *Angew. Chem.* **2002**, *41*, 2644–2676.

(31) Yang, X.; Wang, Y.; Byrne, R.; Schneider, G.; Yang, S. *Chem. Rev.* **2019**, *119*, 10520–10594.

(32) Ajani, H.; Pecina, A.; Eyrilmez, S. M.; Fanfrlík, J.; Haldar, S.; Řezáč, J.; Hobza, P.; Lepšík, M. *ACS Omega* **2017**, *2*, 4022–4029.

(33) Irwin, J. J.; Shoichet, B. K. *J. Med. Chem.* **2016**, *59*, 4103–4120.

(34) Anderson, A. C. *Chem. Biol.* **2003**, *10*, 787–797.

(35) Klebe, G. *Drug Discov Today* **2006**, *11*, 580–594.

(36) Young, D. C., *Computational Drug Design: A Guide for Computational and Medicinal Chemists*; John Wiley & Sons, Inc.: 2009.

(37) Rao, L.; Zhang, I. Y.; Guo, W.; Feng, L.; Meggers, E.; Xu, X. *J. Comput Chem.* **2013**, *34*, 1636–1646.

(38) Zhou, T.; Caflisch, A. *ChemMedChem* **2010**, *5*, 1007–1014.

(39) Thapa, B.; Beckett, D.; Erickson, J.; Raghavachari, K. *J. Chem. Theory Comput.* **2018**, *14*, 5143–5155.

(40) Sinnecker, S.; Neese, F. *J. Comput Chem.* **2006**, *27*, 1463–1475.

(41) Beierlein, F.; Lanig, H.; Schurer, G.; Horn, A. H. C.; Clark, T. *Mol. Phys.* **2003**, *101*, 2469–2480.

(42) Pecina, A.; Eyrilmez, S. M.; Köprülüoğlu, C.; Miriyala, V. M.; Lepšík, M.; Fanfrlík, J.; Řezáč, J.; Hobza, P. *ChemPlusChem* **2020**.

(43) Cígler, P.; Kožíšek, M.; Řezáčová, P.; Brynda, J.; Otwinowski, Z.; Pokorná, J.; Plešek, J.; Grüner, B.; Dolečková-Marešová, L.; Máša, M.; Sedláček, J.; Bodem, J.; Kräusslich, H.-G.; Král, V.; Konvalinka, J. *PNAS* **2005**, *102*, 15394–15399.

(44) Reddy, M. R.; Singh, U. C.; Erion, M. D. *J. Comput Chem.* **2007**, *28*, 491–494.

(45) Olsson, M. A.; Söderhjelm, P.; Ryde, U. *J. Comput Chem.* **2016**, *37*, 1589–1600.

(46) Ryde, U.; Söderhjelm, P. *Chem. Rev.* **2016**, *116*, 5520–5566.

(47) Vasilyev, V.; Bliznyuk, A. *Theor. Chem. Acc.* **2004**, *112*, 313–317.

(48) Raha, K.; Merz, K. M. *J. Am. Chem. Soc.* **2004**, *126*, 1020–1021.

(49) Nagy, G.; Gyurcsik, B.; Hoffmann, E. A.; Körtvélyesi, T. *J. Mol. Graph. Model.* **2011**, *29*, 928–934.

(50) Kamel, K.; Kolinski, A. *Acta Biochim. Pol.* **2012**, *59*.

(51) Urquiza-Carvalho, G. A.; Fragoso, W. D.; Rocha, G. B. *J. Comput. Chem.* **2016**, *37*, 1962–1972.

(52) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Phys. Chem. B* **2001**, *105*, 569–585.

(53) Lepšík, M.; Řezáč, J.; Kolář, M.; Pecina, A.; Hobza, P.; Fanfrlík, J. *ChemPlusChem* **2013**, *78*, 921–931.

(54) Pecina, A.; Meier, R.; Fanfrlík, J.; Lepšík, M.; Řezáč, J.; Hobza, P.; Baldauf, C. *Chem. Commun.* **2016**, *52*, 3312–3315.

(55) Pecina, A.; Haldar, S.; Fanfrlík, J.; Meier, R.; Řezáč, J.; Lepšík, M.; Hobza, P. *J. Chem. Inf. Model.* **2017**, *57*, 127–132.

(56) Kříž, K.; Řezáč, J. *J. Chem. Inf. Model.* **2019**, *59*, 229–235.

(57) Hasanein, A. A.; Evans, M. W., *Computational Methods in Quantum Chemistry*; World Scientific: 1996; 264 pp.

(58) Roberts, G.; Warren, K. D. *Theoret. Chim. Acta* **1969**, *13*, 353–354.

(59) Pople, J. A.; Beveridge, D., *Approximate molecular orbital theory*, Approximate molecular orbital theory; McGraw-Hill, Inc.: New York, 1970.

(60) Oleari, L.; Sipio, L. D.; Michelis, G. D. *Mol. Phys.* **1966**, *10*, 97–109.

(61) Stewart, J. J. P. *Int. J. Quantum Chem.* **1996**, *58*, 133–146.

(62) Dewar, M. J. S.; Thiel, W. *Theoret. Chim. Acta* **1977**, *46*, 89–104.

(63) Thiel, W.; Voityuk, A. A. *J. Phys. Chem.* **1996**, *100*, 616–626.

(64) Stewart, J. J. P. *J. Mol. Model.* **2013**, *19*, 1–32.

(65) Gaus, M.; Goez, A.; Elstner, M. *J. Chem. Theory Comput.* **2013**, *9*, 338–354.

(66) Gaus, M.; Cui, Q.; Elstner, M. *J. Chem. Theory Comput.* **2011**, *7*, 931–948.

(67) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260–7268.

(68) Řezáč, J. *J. Chem. Theory Comput.* **2017**, *13*, 4804–4817.

(69) Stewart, J. J. P. *J. Mol. Model.* **2008**, *14*, 499–535.

(70) Řezáč, J.; Riley, K. E.; Hobza, P. *J. Chem. Theory Comput.* **2011**, *7*, 3466–3470.

(71) Řezáč, J.; Riley, K. E.; Hobza, P. *J. Chem. Theory Comput.* **2011**, *7*, 2427–2438.

(72) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787–1799.

(73) Yang, W.; Lee, T. *J. Chem. Phys.* **1995**, *103*, 5674–5678.

(74) Grimme, S.; Bannwarth, C.; Shushkov, P. *J. Chem. Theory Comput.* **2017**, *13*, 1989–2009.

(75) Bannwarth, C.; Ehlert, S.; Grimme, S. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(76) Sure, R.; Grimme, S. *J. Comput. Chem.* **2013**, *34*, 1672–1685.

(77) Liakos, D. G.; Guo, Y.; Neese, F. *J. Phys. Chem. A* **2020**, *124*, 90–100.

(78) Řezáč, J.; Bím, D.; Gutten, O.; Rulíšek, L. *J. Chem. Theory Comput.* **2018**, *14*, 1254–1266.

(79) Mennucci, B. *WIREs Comput. Mol. Sci.* **2012**, *2*, 386–404.

(80) Klamt, A.; Diedenhofen, M. *J. Comput Chem.* **2018**, *39*, 1648–1655.

(81) Klamt, A.; Schüürmann, G. *J. Chem. Soc., Perkin Trans. 2,* **1993**, *0*, 799–805.

(82) Klamt, A. *WIREs Comput. Mol. Sci.* **2018**, *8*, e1338.

(83) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. W. *J. Phys. Chem. A* **1998**, *102*, 5074–5085.

(84) Barone, V.; Cossi, M. *J. Phys. Chem. A* **1998**, *102*, 1995–2001.

(85) Wang, Y.; Li, H. *J. Chem. Phys.* **2009**, *131*, 206101.

(86) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

(87) Stewart, J. J. P., MOPAC 2016, Stewart Computational Chemistry, Colorado Springs, CO, USA.

(88) Frisch, M. J. et al., *Gaussian09 Revision D.01*, Gaussian09 Revision D.01; Gaussian Inc. Wallingford CT: 2009.

(89) Aradi, B.; Hourahine, B.; Frauenheim, T. *J. Phys. Chem. A* **2007**, *111*, 5678–5684.

(90) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347–1363.

(91) Neese, F. *WIREs Comput. Mol. Sci.* **2012**, *2*, 73–78.

(92) Furche, F.; Ahlrichs, R.; Hättig, C.; Klopper, W.; Sierka, M.; Weigend, F. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 91–100.

(93) Klamt, A.; Eckert, F. **2001**, COSMOtherm, Version C1.1-Revision01.01, COSMOlogic GmbH & Co. KG, Leverkusen, German.

(94) Řezáč, J. *J. Comput. Chem.* **2016**, *37*, 1230–1237.

(95) Pecina, A.; Brynda, J.; Vrzal, L.; Gnanasekaran, R.; Hořejší, M.; Eyrilmez, S. M.; Řezáč, J.; Lepšík, M.; Řezáčová, P.; Hobza, P.; Majer, P.; Veverka, V.; Fanfrlík, J. *ChemPhysChem* **2018**, *19*, 873–879.

(96) Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. **2012**, Minnesota Solvation Database – Version 2012.

(97) Guthrie, J. P. *J. Phys. Chem. B* **2009**, *113*, 4501–4507.

(98) Guthrie, J. P. *J. Comput Aided Mol. Des.* **2014**, *28*, 151–168.

(99) Mobley, D. L.; Dill, K. A.; Chodera, J. D. *J. Phys. Chem. B* **2008**, *112*, 938–946.

(100) Lee, S.; Cho, K.-H.; Lee, C. J.; Kim, G. E.; Na, C. H.; In, Y.; No, K. T. *J. Chem. Inf. Model.* **2011**, *51*, 105–114.

(101) Mader, P.; Brynda, J.; Gitto, R.; Agnello, S.; Pachl, P.; Supuran, C. T.; Chimirri, A.; Řezáčová, P. *J. Med. Chem.* **2011**, *54*, 2522–2526.

(102) Řezáč, J.; Hobza, P. *Chem. Rev.* **2016**, *116*, 5038–5071.

(103) Faver, J. C.; Benson, M. L.; He, X.; Roberts, B. P.; Wang, B.; Marshall, M. S.; Kennedy, M. R.; Sherrill, C. D.; Merz, K. M. *J. Chem. Theory Comput.* **2011**, *7*, 790–797.

(104) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184–32215.

(105) Goerigk, L.; Karton, A.; Martin, J. M. L.; Radom, L. *Phys. Chem. Chem. Phys.* **2013**, *15*, 7028–7031.

(106) Lundberg, M.; Siegbahn, P. E. M. *J. Chem. Phys.* **2005**, *122*, 224103.

(107) Polo, V.; Kraka, E.; Cremer, D. *Mol. Phys.* **2002**, *100*, 1771–1790.

(108) Hostaš, J.; Řezáč, J.; Hobza, P. *Chem. Phys. Lett.* **2013**, *568-569*, 161–166.