ulm university universität ulm

Computational Screening of Energy- and Bio-materials

Dissertation
zur Erlangung des Doktorgrades Dr. rer. nat.
der Fakultät für Naturwissenschaften der
Universität Ulm

vorgelegt von
Nusret Duygu Yılmazer

Supervisor: Jun. Prof. Martin Korth

2015

Amtierender Dekan: Prof. Dr. Peter Dürre

1. Gutachter: Herr Jun. Prof. Dr. Martin Korth
2. Gutachter: Herr Prof. Dr. Gerhard Taubmann

Tag der Promotion: 24. November 2015

Abstract

Computing accurate protein-ligand interaction energies is essential for virtual drug design. For this purpose, in this work, the opportunities of applying fast, enhanced SQM methods are explored. The PDBbind set is used as the experimental reference for protein and ligand structures and binding affinities.

As a part of this work, an algorithm is developed to prepare the structures for the computations. For the calculations, Gilson's minima mining approach is followed, which reduces the problem of obtaining binding energies into computing optimized energies and vibrational frequencies for a large number of binding modes.

Protein-ligand interaction energies are computed by various computational methods: i.e. semi empirical quantum mechanical (SQM) methods, Molecular mechanical (MM) methods, Density functional theory (DFT) methods and Wave function theory (WFT) methods. The accuracy of the results are compared with each other and experimental binding energies.

In total, three benchmarking studies are conducted. In all of them, PM6-DH+ performs as almost as accurate as DFT and WFT methods for realistic model systems, while being fast enough to be used for real protein-ligand systems within the minima mining approach.

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

List of Publications

Parts of this work were published in:

Computational Screening of Biomaterials

1. Yilmazer, N.D. and Korth, M., “Enhanced semiempirical quantum-mechanical methods for biomolecular interactions”: Invited mini-review for *Comp. Struct. Biotech. J. (Elsevier)*, **2015**, 13, 169.

2. Yilmazer, N.D., Heitel, P., Schwabe, T. and Korth, M., “Benchmark of electronic structure methods for protein–ligand interactions based on high-level reference data*”, J. Theor. Comput. Chem.,* **2015**, 14, 1540001.

3. Yilmazer, N.D. and Korth M., “Comparison of molecular mechanics, semi-empirical quantum mechanical, and density functional theory methods for scoring protein-ligand interactions.” *J Phys Chem B,* **2013**, 117, 8075. (Computational Chemistry Highlight, November 2013)

Computational Screening of Energy Materials

4. Husch, T., Yilmazer, N. D., Balducci A. and Korth, M., "Large-scale virtual high-throughput screening for the identification of new battery electrolyte solvents: computing infrastructure and collective properties", *Phys Chem Chem Phys.,* **2015**, 17, 3394. Computational chemistry highlight March 2015, top-scoring Altmetrics article in PCCP March 2015.

5. Yilmazer N. D. and Korth, M., “Computational approaches for the prediction of solid-electrolyte interface formation”, *Bunsen Magazin,* **2013**, 6, 294. (Invited Article)

Table of Contents

Tables

Chapter 2.1.1

Chapter 2.1.3

Chapter 2.3

Chapter 3.1.1 (Research Stage I)

Chapter 3.1.2 (Research Stage II)

Chapter 3.1.3 (Research Stage III)

Chapter 3.2

Figures

Chapter 2.1.1

Chapter 2.2

Chapter 2.3

Chapter 3.1.1 (Research Stage I)

Chapter 3.1.3 (Research Stage III)

Graphs

Chapter 3.1.1 (Research Stage I)

Lists

Chapter 2.1.3

Chapter 3.1.1 (Research Stage I)

Schemes

Chapter 2.1.1

Chapter 2.2

Chapter 3.1.3 (Research Stage III)

Acronyms

DFT	Density Functional Theory
FF	Force Field
HF	Hartree Fock
MM	Molecular Mechanics
RRHO	Rigid Rotor Harmonic Oscillator
SQM	Semi Empirical Quantum Mechanical
WFT	Wave Function Theory

Chapter 1

INTRODUCTION

This thesis consists of two parts based on two distinctive research topics.

The first part is regarded as the main part of the work, with a title of Computational Screening of Bio Materials (Section 3.1). It has three subsections, titled with Research Stage I, II and III which are integrated to each other. The overall goal is to have a performance comparison between the various computational methods with a focus on SQM-DH for scoring protein-ligand interactions, and then by analysing the outcomes, to have an investigation of the biomolecular interaction balances among the scoring function terms.

The second (minor) part has the title Computational screening of battery electrolyte materials (Section 3.2). The study is about screening molecular electrolyte components and then ranking them with respect to their collective properties. Collective properties which are evaluated at lower-level methods are compared with the higher level estimator results. A short summary on this research will be given at the end of thesis.

Chapter 2

LITERATURE SURVEY

2.1 Overview of Computational Chemistry

A large part of the theoretical chemistry can be regarded as the mathematical description of chemistry. Computational chemistry mainly involves the implementation and automation of mathematical methods and models for applications to problems in chemistry [1].

Research with computational chemistry uses advanced tools, theories and models. This involves tasks like: developing algorithms, solving equations, sorting, encoding or visualizing a large number of data to be used in the models. The need for more CPU cycles, bigger memory and disk space increases dramatically, as the size of the examined systems increase. Therefore, even though computational chemistry is considered as a cost and energy efficient approach for chemical problems compared to the experimental studies, it is not without an expense. As a consequence, one of the main research goals in computational chemistry is to improve the accuracy of the results while minimizing the computational cost [2].

Computational chemistry continues to serve as an essential tool for wide range of applications in various disciplines amongst natural sciences and engineering.

2.1.1 Theoretical Background & Computational chemistry methods

Schrödinger Equation [3]

Newton's second law describes the dynamics of a system in classical mechanics. However, electrons are too small particles and they have both wave and particle characteristics. Hence, they cannot be described by classical mechanics (like Newton's law), which is the reason why molecular systems are treated by quantum mechanics.

In order to have a better definition for electrons' special behaviours, terms that are attributed to their wave characteristics have to be considered and introduced.

With the following terms,

H : being the Hamilton operator,
Ψ : denoting the wave function,
ħ : Planck's constant divided by 2π
t : time

The time dependent general Schrödinger equation can be given as in the following:

$$H\psi = i\hbar \frac{\partial \psi}{\partial t} \qquad \text{(Eqn. 2.1.1-1)}$$

Based on the time and position dependencies, this equation, can be written as follows:

$$H(r,t)\psi(r,t) = i\,\hbar \frac{\partial \psi(r,t)}{\partial t} \qquad \text{(Eqn. 2.1.1-2)}$$

For an N-particle system, Hamilton operator, H, contains both kinetic and potential energy terms:

$$H=T+V \qquad \text{(Eqn. 2.1.1-3)}$$

$$H(r,t) = T(r,t) + V(r,t) \qquad \text{(Eqn. 2.1.1-4)}$$

For bound systems, the potential energy is considered to be independent of time:

$$V(r,t) \rightarrow V(r) \qquad \text{(Eqn. 2.1.1-5)}$$

This, in return, affects the Hamiltonian as well, so that it also becomes time-independent:

$$H(r,t) \rightarrow H(r) = T(r) + V(r) \qquad \text{(Eqn. 2.1.1-6)}$$

These all eventually lead to the point, where, the space variables of the wave function can be separated. Solving the first order differential equations with respect to time, we obtain the time dependence factor in the form of a simple factor, $e^{-iEt/\hbar}$, multiplied by the spatial wave function ψ.

$$\psi(r,t) = \psi(r)e^{-iEt/\hbar} \qquad \text{(Eqn. 2.1.1-7)}$$

For time independent problems, this phase factor is omitted, and the other terms are considered as a starting point. Therefore, with the following assignment,

$$\psi(r,t) \rightarrow \psi(r) \qquad \text{(Eqn. 2.1.1-8)}$$

the **time-independent general Schrödinger** equation is obtained:

$$\boldsymbol{H(r)\psi(r) = E(r)\psi(r)} \qquad \text{(Eqn. 2.1.1-9)}$$

After this point, with the description of the kinetic and potential energy terms, the nuclear and electronic variables can be separated further.

For an N-particle system, Hamilton operator can be expanded.

$$H=T+V \qquad \text{(Eqn. 2.1.1-3)}$$

Kinetic energy, T, can be defined as follows:

$$T = \sum_{i=1}^{N} T_i = -\sum_{i=1}^{N} \frac{\hbar^2}{2m_i} \nabla_i^2 \qquad \text{(Eqn. 2.1.1-10)}$$

m :particle mass

Where the inner terms here can be given in the following:

$$\nabla_i^2 = \left(\frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2}\right) \quad \text{(Eqn. 2.1.1-11)}$$

The potential energy, V, in other words the Coulomb potential (which consists of Nuclear–Electron Attraction, Electron–Electron Repulsion and Nuclear–Nuclear Repulsion) can be given as:

$$V = \sum_{i=1}^{N} \sum_{i>j}^{N} V_{ij} \quad \text{(Eqn. 2.1.1-12)}$$

Born-Oppenheimer Approximations

Energy terms can be analysed more in detail with respect to electrons and nuclei,

Nuclei are heavier than electrons, therefore their velocities are much smaller These allows for making the assumption that electronic motion and the nuclear motion in molecules can be regarded as independent from each other. This is called "the Born-Oppenheimer Approximation (BOA)". Based on this approximation, a molecular wave function is analysed in two approaches:

1. The nuclear motion is regarded as much slower than electron motion so that the nuclear geometry is considered to be fixed. Therefore, the electronic wave function only depends on the nuclear positions (in other words, parametrically on the nuclear coordinates), but not on their velocities or momentum. The drawback of this separation approach is that the coupling between nuclear and electronic velocities is neglected. This leads to standard (electronic) QM methods.
2. The perspective of the first approach above is changed the other way around here with the second approach. This time, the energy from the electronic wave function is regarded as the potential energy part, whereas the nuclear wave function is regarded as the part in motion (e.g., rotation, vibration) i.e. that it sees a smeared out potential from the speedy electrons. The electronic wave function provides a Potential Energy Surface (PES), which, with a good approximation shows where the nuclei move. This ultimately leads to classical MM methods.

If nuclei (n) coordinates are denoted as R, and electron (e) coordinates are denoted as r, then together with the two main ideas behind the Born-Oppenheimer approximation, the Schrödinger equation can be written as follows:

$$H_{tot}\psi_{tot}(R,r) = E_{tot}\psi_{tot}(R,r) \quad \text{(Eqn. 2.1.1-13)}$$

$$H_{tot} = H_e + T_n \quad \text{(Eqn. 2.1.1-14)}$$

$$H_e = T_e + V_{ne} + V_{ee} + V_{nn} \quad \text{(Eqn. 2.1.1-15)}$$

$$\psi_{tot}(R,r) = \psi_n(R)\psi_e(R,r) \quad \text{(Eqn. 2.1.1-16)}$$

$$H_e\psi_e(R,r) = E_e(R)\psi_e(R,r) \quad \text{(Eqn. 2.1.1-17)}$$

$$\big(T_n + E_e(R)\big)\psi_n(R) = E_{tot}\psi_n(R) \quad \text{(Eqn. 2.1.1-18)}$$

The Born-Oppenheimer approximation is usually regarded as a very useful approximation with an efficiency getting better as the nuclei gets heavier for a system.

When the “electronic” Schrödinger equation is solved for a large number of nuclear geometries, then the Potential Energy Surface (PES) can be obtained, and the “nuclear” part of the Schrödinger equation can be solved as well.

The electron–electron repulsion term (which is included the Coulomb potential, V, as stated above), prevents the direct solution to the electronic structure.

Then the solution is usually obtained with a convergence method for the electronic structure. This is achieved with an iterative scheme, which is known as the self-consistent field approximation [3].

Potential Energy Surfaces

The potential energy surface (PES) is the mathematical or graphical relationship between the energy of a molecule (or a group of molecules) and its nuclear coordinates.

The Born–Oppenheimer approximation simplifies the application of the Schrödinger equation so that it also explains the concept of molecular shape (geometry) which makes the concept of a PES possible [4].

PES, is a function for the relevant nuclear degrees of freedom. Assuming that a system has N number of nuclei, then N number of atoms can move in three dimensions. This can be defined by 3N coordinates x, y, z for each atom giving 3N degrees of freedom. However, 6 of those, which describe three translations – in x, y, z directions, and three rotations –along x, y and z axes of the molecule are removed. This results in 3N-6 number of independent coordinates at the end.

For N=2, two nuclei are assumed to be on a line with having only two rotational degrees of freedom. Therefore, only for N=2 (a linear system), potential energy is a function of 3N −5 coordinates (linear system) and this leads to a potential energy curve.

For other N ≥3 values (nonlinear system), energy is a function of 3N −6 coordinates, and this leads to a potential energy hypersurface. [2, 5].

A PES graph can be given as follows:

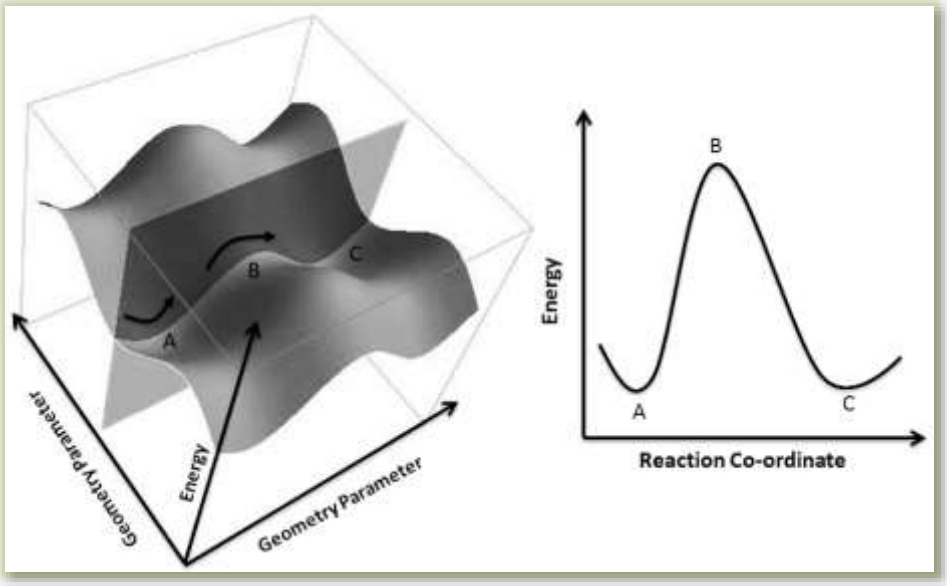


Figure 2.1.1-1*: Potential Energy Surface and Corresponding Reaction Coordinate Diagram* [6]

Self-Consistent Field Theory [8]

This approximation replaces the previously mentioned individual electron–electron repulsion terms by repulsive interactions between individual electrons and the mean electronic field of other electrons that is expressed with the spatially dependent electron density. The disadvantage is that, the electron density depends on the each electron interacting with it, while the electron interaction with the field depends on the density.

Hence, an iterative approach is required up until the first assigned input density converges to the density which is used to calculate the electron-field interaction. This is called Self Consistent Field (SCF) approach. SCF avoids difficult multicentre integrals which are describing the electron-electron interactions, and it diminishes a 3n variable problem into n single electron function problem with three variables for each.

Solving for n molecular orbitals within a self-consistent field is known as the Hartree–Fock solution which will be explained more in detail soon. In order to obtain approximate solutions, Pauli Principle and Variational Principle are also followed so that the wave functions can be constructed via Slater Determinants.

Pauli Principle [3]

There are four electronic quantum numbers defined [7]: the principal quantum number (n), the orbital angular momentum quantum number (l), the magnetic quantum number (m_l), and the electron spin quantum number (m_s). According to the Pauli principle, two electrons cannot have all the four electronic quantum numbers equal.

Electrons are fermions with a half integer spin of ½ for which Fermi-Dirac statistics and Pauli Exclusion Principle applies.

Two possible spin states and the spin functions which obey the orthonormality conditions are given as follows:

$$\langle\alpha|\alpha\rangle = \langle\beta|\beta\rangle = 1 \qquad \text{(Eqn. 2.1.1-18)}$$

$$\langle\alpha|\beta\rangle = \langle\beta|\alpha\rangle = 0 \qquad \text{(Eqn. 2.1.1-19)}$$

The corresponding spin functions are denoted α and β, and in in Dirac Notation.

Orthonormality

In linear algebra, supposing there are two vectors v_1 and v_2, if these vectors are

- Unit vectors (vectors of length, 1), and,
- Orthogonal (perpendicular) to eachother,

Then they these vectors are pronounced as “orthonormal”.

The wave functions of electronic systems are required to be completely antisymmetric. In case the coordinates and spins of any two particle are exchanged, then the wave function changes its sign.

Molecular orbitals (Molecular one electron functions) [3]

Molecular orbitals are a product of spatial orbital and spin orbitals (or spin functions denoted as α or β, which can be taken as orthonormal).
The shape of a given molecular orbital includes attraction to all nuclei, and average repulsion to all other electrons as well.
The molecular orbital picture gives a clue about the probability of finding an electron. An electron is described by its respective orbital, and hence the total wave function is defined as a product of these orbitals. This condition cannot be hold true for the real molecular systems, the wave function cannot be separated into distinct parts for each electron.

Variational Principle [3]

The variational principle states that approximate wave function energy is always less than or equal to the exact energy, and this is only valid if the wave function is exact.

If the exact Schrödinger solution is:

$$H\psi_i = E_i\psi_i \qquad \text{(Eqn. 2.1.1-19)}$$

$$\text{i=0,1,2,....}\infty$$

This means there are infinitely many solutions, and E_0 can be labelled as the lowest. If solutions are orthonormal, then,

$$\langle\psi_i \mid \psi_j\rangle = \delta_{ij} \qquad \text{(Eqn. 2.1.1-20)}$$

$$\psi = \sum_{i=0}^{\infty} a_i\psi_i \qquad \text{(Eqn. 2.1.1-21)}$$

The energy of an approximate wave function can be obtained with dividing the expectation value of the Hamilton operator by the norm of the wave function. Then the equation becomes :

$$E_{wave\ function} = \frac{\langle\psi|H|\psi\rangle}{\langle\psi|\psi\rangle} \qquad \text{(Eqn. 2.1.1-22)}$$

For a normalized wave function, $\langle\psi \mid \psi\rangle$ equals to 1, therefore energy of the approximate wave function becomes:

$$E_{wave\ function} = \langle\psi|H|\psi\rangle \qquad \text{(Eqn. 2.1.1-23)}$$

Slater determinant (SD) [9]

Slater determinants are used to describe the wave functions for a multi-fermionic system that satisfies (i.e. Pauli principle by changing sign upon exchange of two electrons or other fermions) anti-symmetry requirements.

Separation of Variables [3]

For N number of particles, the Hamilton operator H, can be written in independent terms:

$$H = \sum_i h_i \quad \text{(Eqn. 2.1.1-24)}$$

As also mentioned above with a similar summation notation (Eqn. 2.1.1-21), wave functions can also be expanded via one-electron functions. These are called molecular orbitals:

$$\psi = \prod_i \phi_i \quad \text{(Eqn. 2.1.1-25)}$$

and

$$E = \sum_i \varepsilon_i \quad \text{(Eqn. 2.1.1-26)}$$

Then, one variable Schrödinger Equation is:

$$h_i \phi_i = h_i \varepsilon_i \quad \text{(Eqn. 2.1.1-27)}$$

Similarly, the solution to the two-particle problem can then be obtained from the solutions of one variable Schrödinger equations.

$$H = h_1 + h_2 \quad \text{(Eqn. 2.1.1-28)}$$

$$\psi = \phi_1 \phi_2 \quad \text{(Eqn. 2.1.1-29)}$$

and

$$E = \varepsilon_1 + \varepsilon_2 \quad \text{(Eqn. 2.1.1-30)}$$

Wave function being antisymmetric can be written as follows:

$$\psi = -\psi \quad \text{(Eqn. 2.1.1-31)}$$

For N-electrons and N spin orbitals, Slater Determinant can be given as follows [3, 8]:

$$\phi(1,2,\ldots.N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(1) & \phi_2(2) & \ldots. & \phi_n(N) \\ \phi_1(1) & \phi_2(2) & \ldots. & \phi_n(N) \\ \ldots & \ldots & & \ldots \\ \phi(1) & \phi_2(2) & \ldots. & \phi_n(N) \end{vmatrix} \quad \text{(Eqn. 2.1.1-32)}$$

Together with the orthonormality condition (like the Eqn. 2.1.1-20 above),

$$\left\langle \phi_i \middle| \phi_j \right\rangle = \delta_{ij} \quad \text{(Eqn. 2.1.1-33)}$$

In a Slater determinant, columns denote single-electron wave functions (molecular orbitals), whereas rows denote electron coordinates.

Hartree-Fock Approximation

An exact solution of the Schrödinger equation requires the full treatment of electron correlation, therefore more advanced electronic structure methods are used. These methods focus on the electrons. The systems are described by the fundamental forces acting upon the electrons, which requires a multi-determinant wave function.

Multi-determinant wave function methods can generate results that systematically approach the exact solution of the Schrödinger equation. However, for a many-electron system, the dynamics of the system is very complicated and the exact solution of the Schrödinger equation cannot be attained. Hence, apart from the expansion of the wave function in slater determinants, an expression of the molecular orbitals in basis functions is required.

Basis Set Expansion-Approximation [3]

The definition of the molecular orbitals in terms of a series of basis functions is known as Basis Set Approximation.

$$\phi_i(r) = \sum_{\alpha=1}^{M\ basis} C_{\alpha i}\chi_\alpha(r) \qquad \text{(Eqn. 2.1.1-34)}$$

M_{basis} : set of basis functions located on the nuclei,
ϕ : each molecular orbital
χ : basis functions,
C : coefficient which relates the atomic orbital α to molecular orbital i.

For molecular based systems, the mentioned basis functions are mostly chosen as atomic orbitals.

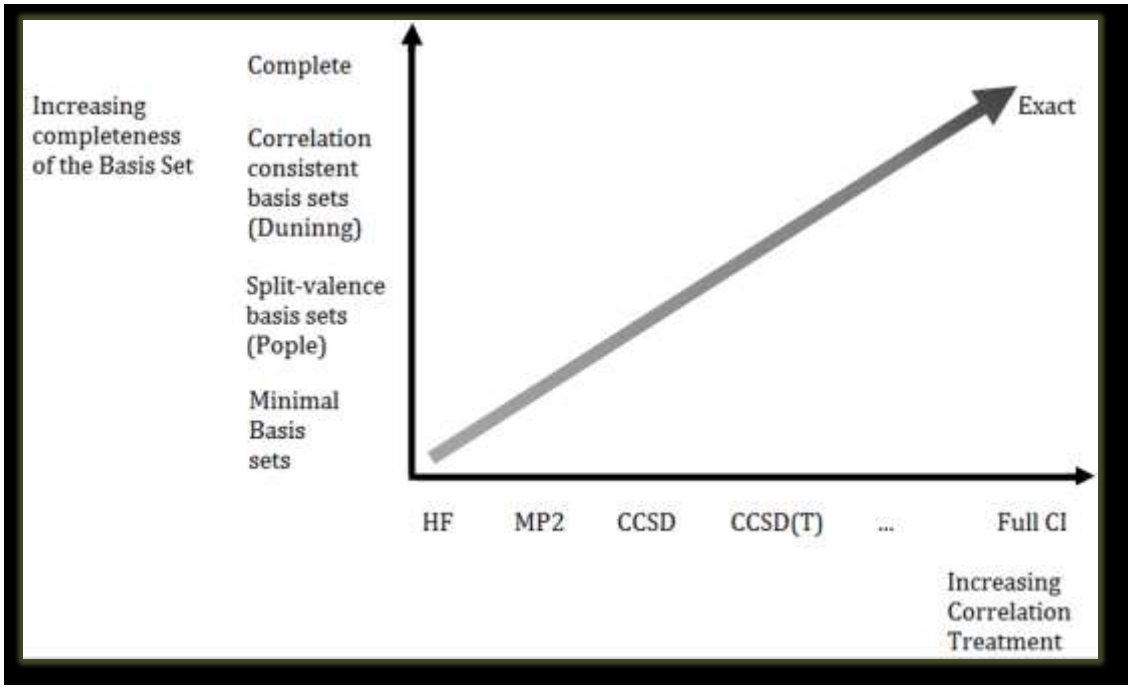


Figure 2.1.1-2: *Comparison of models and their improvement in accuracy based on the correlation treatments and the completeness of the basis set (based on ref.* [8]).

As, with the SCF approach mentioned above, N-electron Schrödinger equation can be viewed as an "n single-electron" case, and the solution of this "n single-electron" is known as the Hartree–Fock (HF) solution.

The disadvantage here is, HF solution does not treat the interactions between the electrons, namely, the electron correlations, which in return, leads to some errors.

There are, however, general approaches which have been developed to treat electron correlations. These can be listed as follows and will be reviewed shortly in the following sections:

- Many Body Perturbation Theory,
- Configurational Interaction (CI),
- Coupled Cluster (CC) theory.

Electron Correlation [8]

In a system, electrons try to avoid eachother and their motion is correlated accordingly. In a fixed orbital state, description of the electrons are missing at this point.

From this perspective, electron distances predicted by HF methods have to be further away from each other in reality, and Electron correlation is defined as the difference in between exact electronic energy and a HF solution.

Energy of a Slater Determinant & Derivation of Hartree-Fock Equation [3]

Single determinant trial wave function, together with the variational principle can be used to obtain Hartree-Fock equations.

Denoting " $\prod_i \phi_i$ " as "**Π**", together with the following notations,

1	: identity operator
A	: antisymmetrizing operator
P_{ij}	: generates all possible permutations of two electron coordinates
P_{ijk}	: generates all possible permutations of three electron coordinates
H	: Hamilton operator,

The diagonal of the determinant can be written as follows:

$$\psi = A[\phi_1(1)\phi_2(2) \ldots . \phi_N(N)] = A\mathbf{\Pi} \qquad \text{(Eqn. 2.1.1-35)}$$

$$A = \frac{1}{\sqrt{N!}} \sum_{p=0}^{N-1} (-1)^p \, P \qquad \text{(Eqn. 2.1.1-36)}$$

$$A = \frac{1}{\sqrt{N!}} \sum_{p=0}^{N-1} [\mathbf{1} - \sum_{ij} P_{ij} + \sum_{ijk} P_{ijk} - \ldots \,] \qquad \text{(Eqn. 2.1.1-37)}$$

$$AH=HA \qquad \text{(Eqn. 2.1.1-38)}$$

$$AA=\sqrt{N!}\, A \qquad \text{(Eqn. 2.1.1-39)}$$

Starting with the equation Eqn. 2.1.1-15,

$$H_e = T_e + V_{ne} + V_{ee} + V_{nn} \quad \text{(Eqn. 2.1.1-15)}$$

Based on the number of electron indices,

h_i : motion of electron i in the field of all nuclei,
g_{ij} : two electron operator for electron-electron repulsion

The following can be presented:

$$T_e = -\sum_i^N \frac{1}{2}\nabla_i^2 \quad \text{(Eqn. 2.1.1-40)}$$

$$V_{ne} = -\sum_i^N \sum_a \frac{Z_a}{|R_a - r_i|} \quad \text{(Eqn. 2.1.1-41)}$$

$$V_{ee} = \sum_i^N \sum_{j>i} \frac{1}{|r_i - r_j|} \quad \text{(Eqn. 2.1.1-42)}$$

$$V_{ne} = \sum_a \sum_{b>a} \frac{Z_a Z_b}{|R_a - R_b|} \quad \text{(Eqn. 2.1.1-43)}$$

Particles are at rest at zero point energy (ZPE) have the following conditions,

$$T_e=0 \ \& \ V_{ne}=V_{ee}=V_{nn}=0$$

Then, the following equations are obtained:

$$h_i = -\frac{1}{2}\nabla_i^2 - \sum_a \frac{Z_a}{|R_a - r_i|} \quad \text{(Eqn. 2.1.1-44)}$$

$$g_{ij} = \frac{1}{|r_i - r_j|} \quad \text{(Eqn. 2.1.1-45)}$$

$$H_e = \sum_{i=1}^N h_i + \sum_{i=1}^N \sum_{j>i}^N g_{ij} + V_{nn} \quad \text{(Eqn. 2.1.1-46)}$$

Starting from the Eqn. 2.1.1-23,

$$E = \langle \psi | H | \psi \rangle \quad \text{(Eqn. 2.1.1-23)}$$

together with the descriptions introduced above,

$$E = \langle A\mathbf{\Pi} | H | A\mathbf{\Pi} \rangle \quad \text{(Eqn. 2.1.1-47)}$$

$$E = \sqrt{N!}\, \langle \mathbf{\Pi} | H | A\mathbf{\Pi} \rangle \quad \text{(Eqn. 2.1.1-48)}$$

$$E = \sum_p (-1)^p \langle \mathbf{\Pi} | H | A\mathbf{\Pi} \rangle \quad \text{(Eqn. 2.1.1-49)}$$

for the nuclear repulsion, the operator can be turned into a constant:

$$\langle \psi | V_{nn} | \psi \rangle = V_{nn} \langle \psi | \psi \rangle = V_{nn} \quad \text{(Eqn. 2.1.1-50)}$$

For one electron operator, when all molecular orbitals are normalized,

$$\langle \mathbf{\Pi} | h_1 | \mathbf{\Pi} \rangle = \langle \phi_1(1)\phi_2(2) \ldots . \phi_N(N) | h_1 | \phi_1(1)\phi_2(2) \ldots . \phi_N(N) \rangle \quad \text{(Eqn. 2.1.1-51)}$$

$$= \langle \phi_1(1) | h_1 | \phi_1(1) \rangle \langle \phi_2(2) | \phi_2(2) \rangle \ldots . \langle \phi_N(N) | \phi_N(N) \rangle \quad \text{(Eqn. 2.1.1-52)}$$

$$= \langle \phi_1(1) | h_1 | \phi_1(1) \rangle = h_1 \quad \text{(Eqn. 2.1.1-53)}$$

all the matrix elements involving a permutation operator, P, give zero.

$$\langle \mathbf{\Pi} | h_1 | P_{12}\mathbf{\Pi} \rangle = \langle \phi_1(1)\phi_2(2) \ldots . \phi_N(N) | h_1 | \phi_2(1)\phi_1(2) \ldots . \phi_N(N) \rangle \quad \text{(Eqn. 2.1.1-54)}$$

$$= \langle \phi_1(1) | h_1 | \phi_2(1) \rangle \langle \phi_2(2) | \phi_1(2) \rangle \ldots . \langle \phi_N(N) | \phi_N(N) \rangle \quad \text{(Eqn. 2.1.1-55)}$$

This also equates to zero based on the integral over 2 being an overlap of two different orthogonal molecular orbitals.

In case of a two electron integral, only the identity operator, **1**, and the permutation, P_{ij}, operators lead to a non-zero result. On the other hand, the three electron permutation operator, P_{ijk}, is still going to give a zero from at least one of the overlap integrals between two different molecular orbitals.

This leads to:

$$\langle \mathbf{\Pi} | g_{12} | \mathbf{\Pi} \rangle = \langle \phi_1(1)\phi_2(2) \ldots . \phi_N(N) | g_{12} | \phi_1(1)\phi_2(2) \ldots . \phi_N(N) \rangle \quad \text{(Eqn. 2.1.1-56)}$$

$$= \langle \phi_1(1)\phi_2(2) | g_{12} | \phi_1(1)\phi_2(2) \rangle \ldots . \langle \phi_N(N) | \phi_N(N) \rangle \quad \text{(Eqn. 2.1.1-57)}$$

$$= \langle \phi_1(1)\phi_2(2) | g_{12} | \phi_1(1)\phi_2(2) \rangle = J_{12} \quad \text{(Eqn. 2.1.1-58)}$$

The Coulomb integral, J, explains the classical repulsion between two charge distributions, and,

$$\langle \mathbf{\Pi} | g_{12} | P_{12}\mathbf{\Pi} \rangle = \langle \phi_1(1)\phi_2(2) \ldots . \phi_N(N) | g_{12} | \phi_2(1)\phi_1(2) \ldots . \phi_N(N) \rangle \quad \text{(Eqn. 2.1.1-59)}$$

$$= \langle \phi_1(1)\phi_2(2) | g_{12} | \phi_2(1)\phi_1(2) \rangle \ldots . \langle \phi_N(N) | \phi_N(N) \rangle \quad \text{(Eqn. 2.1.1-60)}$$

$$= \langle \phi_1(1)\phi_2(2) | g_{12} | \phi_2(1)\phi_1(2) \rangle = K_{12} \quad \text{(Eqn. 2.1.1-61)}$$

the exchange integral, K.

Based on J_{12} and K_{12}, the overall energy becomes:

$$E = \sum_{i=1}^{N} h_i + \sum_{i=1}^{N} \sum_{j>i}^{N} (J_{ij} - K_{ij}) + V_{nn} \quad \text{(Eqn. 2.1.1-62)}$$

$$E = \sum_{i=1}^{N} h_i + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (J_{ij} - K_{ij}) + V_{nn} \quad \text{(Eqn. 2.1.1-63)}$$

Factor ½ is for the double sum over all electrons.

$$E = \sum_{i}^{N} \langle \phi_i | h_i | \phi_i \rangle + \frac{1}{2} \sum_{ij}^{N} \left(\langle \phi_i | J_i | \phi_i \rangle - \langle \phi_i | K_i | \phi_i \rangle \right) + V_{nn} \quad \text{(Eqn. 2.1.1-64)}$$

$$J_i | \phi_j(2) \rangle = \langle \phi_i(1) | g_{12} | \phi_i(1) \rangle | \phi_j(2) \rangle \quad \text{(Eqn. 2.1.1-65)}$$

$$K_i | \phi_j(2) \rangle = \langle \phi_i(1) | g_{12} | \phi_j(1) \rangle | \phi_i(2) \rangle \quad \text{(Eqn. 2.1.1-66)}$$

Keeping the molecular orbitals orthonormal, a set of molecular orbitals which gives the minimum energy will be obtained. This is classified as a constrained optimization problem that will be handled via Lagrange multipliers.

Lagrange Multipliers [3, 10-13]

Lagrange method is one of the techniques that is used for the constrained optimization problems.

This involves the following assumption:

- There is a function, f(x1,x2,x3.., xn), which needs
  - to be optimized or
  - its (local or global) maximum or minimum points to be found,
- This function f(x1,x2,x3.., xn) is subject to another function, g, as a constraint: g(x1,x2,x3…,xn) =0,
- The continuous first partial derivatives of the functions f(x1,x2,x3.., xn) and g(x1,x2,x3.., xn) are on the open set including the curve g(x1,x2,x3.., xn)=0 and that $\nabla g \neq 0$ at any point on the curve.

$$L = E - \sum_{ij}^{N} \lambda_{ij}(\langle \phi_i | \phi_j \rangle - \delta_{ij}) \qquad \text{(Eqn. 2.1.1-67)}$$

$$\delta L = \delta E - \sum_{ij}^{N} \lambda_{ij}\left(\langle \delta\phi_i | \phi_j \rangle + \langle \phi_i | \delta\phi_j \rangle\right) = 0 \qquad \text{(Eqn. 2.1.1-68)}$$

$$\delta E = \sum_{i}^{N}(\langle \delta\phi_i | h_i | \phi_i \rangle + \langle \phi_i | h_i | \delta\phi_i \rangle) + \frac{1}{2}\sum_{ij}^{N}\left(\langle \delta\phi_i | J_i - K_j | \phi_i \rangle + \langle \phi_i | J_j - K_j | \delta\phi_i \rangle + \langle \delta\phi_j | J_i - K_i | \phi_j \rangle + \langle \phi_j | J_i - K_i | \delta\phi_j \rangle\right) \qquad \text{(Eqn. 2.1.1-69)}$$

$$\delta E = \sum_{i}^{N}(\langle \delta\phi_i | h_i | \phi_i \rangle + \langle \phi_i | h_i | \delta\phi_i \rangle) + \sum_{ij}^{N}\left(\langle \delta\phi_i | J_j - K_j | \phi_i \rangle + \langle \phi_i | J_j - K_j | \delta\phi_i \rangle\right) \qquad \text{(Eqn. 2.1.1-70)}$$

Here, **Fock operator** is defined, F:

$$\boldsymbol{F_i = h_i + \sum_j^N (J_j - K_j)} \qquad \text{(Eqn. 2.1.1-71)}$$

Then equation becomes,

$$\delta E = \sum_{i}^{N}(\langle \delta\phi_i | F_i | \phi_i \rangle + \langle \phi_i | F_i | \delta\phi_i \rangle) \qquad \text{(Eqn. 2.1.1-72)}$$

Then, the Lagrange function can be written as,

$$\delta L = \sum_{i}^{N}(\langle \delta\phi_i | F_i | \phi_i \rangle + \langle \phi_i | F_i | \delta\phi_i \rangle) - \sum_{ij}^{N} \lambda_{ij}\left(\langle \delta\phi_i | \phi_j \rangle + \langle \phi_i | \delta\phi_j \rangle\right) \qquad \text{(Eqn. 2.1.1-73)}$$

For a stationary state with orthonormal orbitals, δL=0, is desired.

Then, with the following equivalence:

$$\langle\phi|\delta\phi\rangle = \langle\delta\phi|\phi\rangle^* \quad \text{(Eqn. 2.1.1-74)}$$

$$\langle\phi|F|\delta\phi\rangle = \langle\delta\phi|F|\phi\rangle^* \quad \text{(Eqn. 2.1.1-75)}$$

This becomes,

$$\delta L = \sum_i^N \langle\delta\phi_i|F_i|\phi_i\rangle - \sum_{ij}^N \lambda_{ij} \left\langle\delta\phi_i\middle|\phi_j\right\rangle + \sum_i^N \langle\delta\phi_i|F_i|\phi_i\rangle^* - \sum_{ij}^N \lambda_{ij} \left\langle\delta\phi_i\middle|\phi_j\right\rangle^* = 0 \quad \text{(Eqn. 2.1.1-76)}$$

$$\sum_{ij}^{\lambda}(\lambda_{ij} - \lambda_{ji}^*) \left\langle\delta\phi_i\middle|\phi_j\right\rangle = 0 \quad \text{(Eqn. 2.1.1-77)}$$

Lagrange multipliers being the elements of a Hermitian Matrix, thus, the following can be written:

$$\lambda_{ij} = \lambda_{ji}^* \quad \text{(Eqn. 2.1.1-78)}$$

Hermitian Matrix [14, 15]

If a matrix is self-adjoint, which means, if

$A=(a_{ij})$ is a defined matrix, then it is called Hermitian matrix in case it holds the following condition:

$A=A^H$; where A^H denotes the conjugate transpose.

Then finally, **Hartree-Fock** equation is obtained:

$$\boldsymbol{F_i\phi_i = \sum_i^N \lambda_{ij}\phi_j} \quad \text{(Eqn. 2.1.1-79)}$$

Equation can be simplified further via taking the Lagrange multipliers diagonal, thus, $\lambda_{ij}\rightarrow 0$ and $\lambda_{ii}\rightarrow\varepsilon$

Special set of molecular orbitals, called **canonical orbitals**, are obtained, hence Hartree-Fock equation can also be shown as:

$$F_i\phi_i' = \varepsilon_i\phi_i' \quad \text{(Eqn. 2.1.1-80)}$$

So from this point on, if we also apply the basis set approximation by expanding the molecular orbitals,

$$\phi_i = \sum_\alpha^M c_{\alpha i}\chi_\alpha \quad \text{(Eqn. 2.1.1-81)}$$

then the Hartree-Fock equation becomes

$$F_i \sum_\alpha^M c_{\alpha i}\chi_\alpha = \varepsilon_i \sum_\alpha^M c_{\alpha i}\chi_\alpha \quad \text{(Eqn. 2.1.1-82)}$$

For a closed shell system, multiplying this equation with a specific basis function and then integrating leads to the **Roothan-Hall equation**:

With:

F: **Fock matrix** elements,

$$F_{\alpha\beta} = \langle \chi_\alpha | F | \chi_\beta \rangle \quad \text{(Eqn. 2.1.1-83)}$$

S: matrix containing overlap elements between basis functions:

$$S_{\alpha\beta} = \langle \chi_\alpha | \chi_\beta \rangle \quad \text{(Eqn. 2.1.1-84)}$$

The result is the **Fock equation** in the atomic orbital basis:

$$\boldsymbol{FC = SC\varepsilon} \quad \text{(Eqn. 2.1.1-85)}$$

To expand further,

$$\langle \chi_\alpha | F | \chi_\beta \rangle = \langle \chi_\alpha | h | \chi_\beta \rangle + \sum_j^{occ.MO} \langle \chi_\alpha | J_j - K_j | \chi_\beta \rangle \quad \text{(Eqn. 2.1.1-86)}$$

$$= \langle \chi_\alpha | h | \chi_\beta \rangle + \sum_j^{occ.MO} \left(\langle \chi_\alpha \phi_j | g | \chi_\beta \phi_j \rangle - \langle \chi_\alpha \phi_j | g | \phi_j \chi_\beta \rangle \right) \quad \text{(Eqn. 2.1.1-87)}$$

$$= \langle \chi_\alpha | h | \chi_\beta \rangle + \sum_j^{occ.MO} \sum_\gamma^{AO} \sum_\delta^{AO} c_{\gamma j} c_{\delta j} \left(\langle \chi_\alpha \chi_\gamma | g | \chi_\beta \chi_\delta \rangle - \langle \chi_\alpha \chi_\gamma | g | \chi_\delta \chi_\beta \rangle \right) \quad \text{(Eqn. 2.1.1-88)}$$

$$= \langle \chi_\alpha | h | \chi_\beta \rangle + \sum_\gamma^{AO} \sum_\delta^{AO} D_{\gamma\delta} \left(\langle \chi_\alpha \chi_\gamma | g | \chi_\beta \chi_\delta \rangle - \langle \chi_\alpha \chi_\gamma | g | \chi_\delta \chi_\beta \rangle \right) \quad \text{(Eqn. 2.1.1-89)}$$

$$D_{\gamma\delta} = \sum_j^{occ.MO} c_{\gamma j} c_{\delta j} \quad \text{(Eqn. 2.1.1-90)}$$

$$F_{\alpha\beta} = h_{\alpha\beta} + \sum_{\gamma\delta} G_{\alpha\beta\gamma\delta} D_{\gamma\delta} \quad \text{(Eqn. 2.1.1-91)}$$

$$F = h + G.D \quad \text{(Eqn. 2.1.1-92)}$$

G.D= contraction of the D matrix with four dimensional G tensor

Total energy is:

$$E = \sum_i^N \langle \phi_i | h_i | \phi_i \rangle + \frac{1}{2} \sum_{ij}^N \left(\langle \phi_i \phi_j | g | \phi_i \phi_j \rangle - \langle \phi_i \phi_j | g | \phi_j \phi_i \rangle \right) + V_{nn} \quad \text{(Eqn. 2.1.1-93)}$$

$$E = \sum_i^N \sum_{\alpha\beta}^M c_{\alpha i} c_{\beta i} \langle \chi_\alpha | h_i | \chi_\beta \rangle + \frac{1}{2} \sum_{ij}^N \sum_{\alpha\beta\gamma\delta}^M c_{\alpha i} c_{\gamma j} c_{\beta i} c_{\delta j} \left(\langle \chi_\alpha \chi_\gamma | g | \chi_\beta \chi_\delta \rangle - \langle \chi_\alpha \chi_\gamma | g | \chi_\delta \chi_\beta \rangle \right) + V_{nn} \quad \text{(Eqn. 2.1.1-94)}$$

$$E = \sum_{\alpha\beta}^M D_{\alpha\beta} h_{\alpha\beta} + \frac{1}{2} \sum_{\alpha\beta\gamma\delta}^M D_{\alpha\beta} D_{\gamma\delta} \left(\langle \chi_\alpha \chi_\gamma | g | \chi_\beta \chi_\delta \rangle - \langle \chi_\alpha \chi_\gamma | g | \chi_\delta \chi_\beta \rangle \right) + V_{nn} \quad \text{(Eqn. 2.1.1-95)}$$

$$E = \sum_{\alpha\beta}^M D_{\alpha\beta} h_{\alpha\beta} + \frac{1}{2} \sum_{\alpha\beta\gamma\delta}^M \left(D_{\alpha\beta} D_{\gamma\delta} - D_{\alpha\delta} D_{\delta\beta} \right) \langle \chi_\alpha \chi_\gamma | g | \chi_\beta \chi_\delta \rangle + V_{nn} \quad \text{(Eqn. 2.1.1-96)}$$

$$\langle \chi_\alpha | h | \chi_\beta \rangle = \int \chi_\alpha(1) \left(-\frac{1}{2} \nabla^2 \right) \chi_\beta(1) dr_1 + \sum_a \int \chi_\alpha(1) \frac{Z_a}{|R_a - r_1|} \chi_\beta(1) dr_1 \qquad \text{(Eqn. 2.1.1-97)}$$

$$\langle \chi_\alpha \chi_\gamma | g | \chi_\beta \chi_\delta \rangle = \int \chi_\alpha(1) \chi_\gamma(2) \frac{1}{|r_1 - r_2|} \chi_\beta(1) \chi_\delta(2) dr_1 dr_2 \qquad \text{(Eqn. 2.1.1-98)}$$

This equation can also be written as:

$$\int \chi_\alpha(1) \chi_\gamma(2) \frac{1}{|r_1 - r_2|} \chi_\beta(1) \chi_\delta(2) dr_1 dr_2 = \left(\chi_\alpha \chi_\gamma | \chi_\beta \chi_\delta \right) \qquad \text{(Eqn. 2.1.1-99)}$$

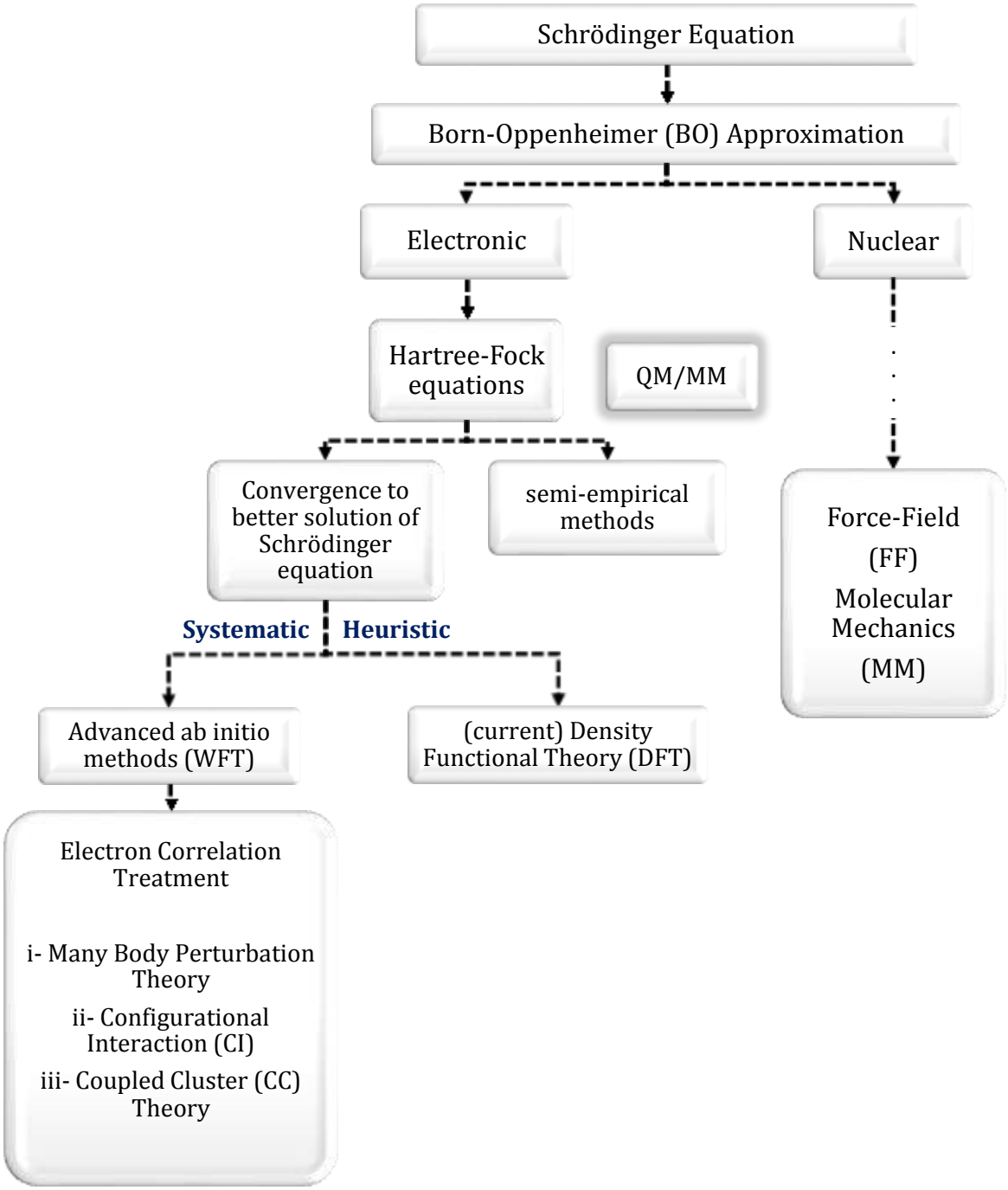
Ordering of the function shown in electron indices is called physicist's notation.

In a similar fashion, this time gathering the electrons at different parts, it is called chemist's (Mulliken) notation.

$$\int \chi_\alpha(1) \chi_\beta(1) \frac{1}{|r_1 - r_2|} \chi_\gamma(2) \chi_\delta(2) dr_1 dr_2 = \langle \chi_\alpha \chi_\beta | g | \chi_\gamma \chi_\delta \rangle \qquad \text{(Eqn. 2.1.1-100)}$$

Concluding Remarks

From the methodology point of view, the following Scheme 2.1.1-1 can be followed.



Scheme 2.1.1-1*: Overview of Computational Chemistry Methods*

BO approximation eventually leads either to Quantum Mechanical (QM) or Force-Field (FF) methods. Quantum mechanical methods are also varied within eachother after the Hartree-Fock (HF) equations.

HF model is considered as an important stage which puts the electronic structure methods into two main categories:

- the models which are aimed to converge to the exact solution of the Schrödinger equation (ab initio methods), and,
- the semi empirical models.

Ab initio methods also have two related sub categories:

- advanced ab initio methods (in other words, Wave Function Theory (WFT) methods), and,
- Density Functional Theory (DFT) methods.

Wave function theory (WFT) makes use of HF solutions by including electron correlations. As it is mentioned previously, electron correlation treatments are present and they can be listed as follows: Many Body Perturbation Theory, Configurational Interaction (CI) and Coupled Cluster (CC) Theory.

Density Functional Theory (DFT) on the other hand, is also derived from first principles, but it can be considered as an improvement on HF theory, since DFT is based on the electron density of the N-particle system instead of wave functions.

Practically today's Kohn-Sham DFT methods work with orbitals like HF does. DFT is an independent-particle model, and it is computationally comparable to HF, but it provides much better results.

It is shown by Hohenberg and Kohn [3, 8, 16] that, for DFT, the energy of a system is a unique functional of its electron density. On the other hand, the accuracy of the DFT methods are based on the quality of the exchange-correlation functionals. One important disadvantage of DFT is that, there is no systematic approach for them in order to improve their results towards the exact solution of the Schrödinger equation.

Semi-empirical methods are derived from the HF model by neglecting all the integrals which involve more than two nuclei during the construction of the Fock matrix. The success of this method relies on the ability of turning the remaining integrals into the parameters, so that these parameters try to fit the molecular energies and geometries to the experimental data. Therefore, SQM is computationally much more efficient than the ab initio HF methods, while it being limited to systems, for which the parameters exist.

In general, with a good computing system, QM methods are known to be practical with their computational costs for the systems which have maximum few hundred number of atoms.

There are also hybrid QM/MM and methods which combine a Quantum mechanical approach with another lower level theory approach (i.e. like MM: Molecular Mechanics), and these type of methods are also often used for large systems nowadays.

Force-Field Methods [8]

A nuclear based description of a system is more advantageous especially when the electron transfer is not critical for a system, but instead the simulation of a structural property or a dynamic response of a system is more of a concern. In addition, with this approach, larger systems can be simulated for longer time scales as well.

Force-fields methods are convenient for these purposes. They can track both:

- the forces in between the individual atoms,
- intra and intermolecular forces in between molecules.

The potential energy of the Force-Fields contains both intra and intermolecular forces:

Intra molecular Forces:

The contributions of the intra–molecular interactions to the potential energy are the result of the changes in the bond length, bond angle and torsion angle from their 'standard' positions.

Following terms take place in the equations:

- calculated bond length: r_i & a universal bond length for specific type of bond r_0
- calculated bond angle θ_i & universal bond angle θ_0
- calculated torsion angle φ_i & universal torsion angle φ_0

There are also empirical coefficients for the fit between experimental bond lengths, as K_B, the bond angles, as K_θ and for torsion angles, as C_i.

The bond length potential is often a parabolic equation based on Hooke's law.

Bond Length:

$$V_r = \sum_{i=1}^{N_m-1} \frac{1}{2} K_B (r_i - r_o)^2 \qquad \text{(Eqn. 2.1.1-101)}$$

Bond angle:

$$V_\theta = \sum_{i=1}^{N_m-2} \frac{1}{2} K_\theta (\theta_i - \theta_o)^2 \qquad \text{(Eqn. 2.1.1-102)}$$

Torsion angle:

$$V_\phi = \sum_{i=1}^{N_m-3} \sum_{j=0}^{p} C_i (cos\phi_i)^j \qquad \text{(Eqn. 2.1.1-103)}$$

Intermolecular Forces:

These include electrostatic dispersive and repulsive forces. The van der Waals interactions are been modelled via Lennard–Jones [8, 17-22], Morse and Buckingham type potentials [8, 23].

Coulombic:

$$V_C = \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{q_i q_j}{4\pi\varepsilon_o r_{ij}} \quad \text{(Eqn. 2.1.1-104)}$$

Non-polar:

$$V_{\mathrm{LJ}} = \frac{A_o}{r^{12}} - \frac{B_o}{r^6} \text{; Lennard-Jones} \quad \text{(Eqn. 2.1.1-105)}$$

$$V_{\mathrm{exp}} = Ae^{-Br} \text{ ; Exponential} \quad \text{(Eqn. 2.1.1-106)}$$

$$V_{\mathrm{B}} = Ae^{-Br} - \frac{C_6}{r^6} \text{; Buckingham} \quad \text{(Eqn. 2.1.1-107)}$$

q_i	:Charges on the atom i
ε	: the dielectric constant of the medium
A_0, B_0, A_1 and C	: Fitting coefficients

Then, the total potential energy V_T of the system (addition of intra and intermolecular forces) can be given as follows:

$$V_T = V_r + V_\theta + V_\varphi + V_C + V_{LJ} \quad \text{(Eqn. 2.1.1-108)}$$

Force-field methods provide

- the potential energy which is used to carry out energy minimization to identify the most stable structures,
- Monte Carlo simulations to determine the properties of equilibrated systems, and,
- Molecular Dynamics simulations to follow the dynamics of the system.

It needs to be emphasized that the accuracy of the most atomistic simulations is highly dependent on the accuracy and applicability of the force field that has been developed, and there are very good results achieved with this approach in computational chemistry.

Advanced Ab Initio Methods [3, 8]

As stated above, HF solutions are approximate solutions to the exact Schrödinger equation. Instead of a direct electron-electron interaction solution, HF replaces the terms with a mean field approach (which is without a correlation). Therefore, there is a difference in between the exact solution and the approximate solution of the Schrödinger equation. This difference is named as the correlation energy. There are multi-reference methods and standard correlation methods. Standard correlation methods are explained in the following:

Advanced Ab Initio Methods – Electron Correlation Treatments [3, 8]

I. Many Body Perturbation Theory

This approach treats the configurational interactions as small perturbations to the HF Hamiltonian. This is expressed as:

$$H = H_o + \lambda H' \qquad \text{(Eqn. 2.1.1-109)}$$

H_o : reference HF Hamiltonian
λ : variable describing the relative degree of perturbation
H' : small perturbation

The perturbation is a consequence of the constructs of true Hamiltonian and this is equivalent to the nuclear attraction and electron repulsion terms. In a Taylor Series expansion, energy and wave function can be expressed as follows:

$$E = \lambda^0 E_0 + \lambda^1 E_1 + \lambda^2 E_2 + \lambda^3 E_3 + \cdots \qquad \text{(Eqn. 2.1.1-110)}$$

$$\psi = \lambda^0 \psi_0 + \lambda^1 \psi_1 + \lambda^2 \psi_2 + \lambda^3 \psi_3 + \cdots \qquad \text{(Eqn. 2.1.1-111)}$$

ψ_0 : unperturbed solution of the wave function-from the H_0 Hamiltonian
E_0 : unperturbed solution of the energy-from the H_0 Hamiltonian
E_1, E_2, E_3 : higher order corrections to the energy
ψ_1, ψ_2, ψ_3 : higher order corrections to the wave function

Examples include:

- Møller–Plesset (MP) perturbation theory: The reference of the unperturbed Hamiltonian operator is obtained by the sum over Fock operators.
- MP2: a low perturbation order. MP2 mostly recovers 80–90% of the correlation energy.

II. Configurational Interaction (CI)

CI methods treat the trial wave function, ψ, as a linear combination of the Hartree Fock wave function, ψ_o, and the virtually excited wave functions, ψ_1, ψ_2, ψ_3, etc. Hence, based on the variational principle, which enables the optimization of the coefficients, the general solution is expressed as:

$$\psi = C_o \psi_o + C_1 \psi_1 + C_2 \psi_2 + C_3 \psi_3 + \ldots \qquad \text{(Eqn. 2.1.1-112)}$$

Trial wave function, ψ, might include the exchange of one or more electrons from the occupied into unoccupied orbitals:

- Single (S)
- Double (D)
- Triple (T) excitations, respectively.

Then the full equation becomes:

$$\psi_{CI} = C_o\psi_{SCF} + \sum_S C_S\psi_S + \sum_D C_D\psi_D + \sum_T C_T\psi_T + \cdots \qquad \text{(Eqn. 2.1.1-113)}$$

C_S, C_D and C_T: coefficients for singly, doubly, triply excited states.

With these additions, the wave function, ψ, now has the parts both from the HF determinant, ψ_o , and other possible determinants.

Coefficients for each state (like: C_S, C_D, etc …) can be variationally optimized by minimizing the energy. In principle, solutions should reach to an accurate solution as the number of excitations do increase. In such a case, within the limit of the basis set expansion, full-CI provides exact solutions beyond Hartree-Fock. However, full-CI calculations are only possible for very small systems because of the computational costs, and truncated CI is not size consistent.

III. Coupled Cluster (CC) theory

A size consistent way of adding determinants can be achieved by coupled cluster theory.

Since single excitations (S) do not extend the HF solution for the energy, the truncation level is usually at the double (D) excitations. The addition of the singles to the doubles improve the solution, and this is named as CCSD (Coupled Cluster Singles and Doubles).

The expansion until the fourth degree (quadruple) excitations are performed only for very small systems (3 to 5 atoms) because of the demanding computational costs.

In comparison to the full CI, the correlation treatment levels based on their accuracy can be given as follows:

$$\text{HF} < \text{MP2} < \text{CCSD} < \text{CCSD(T)} < \text{CCSDT} << \text{Full CI}$$

Concluding Remarks

Accurate predictions of the properties are mostly pronounced to be provided by higher level CC calculations. With a high CPU cost, it is reported that, the accuracy is reached within:

- 1% for structure determination
- 1 kcal/mol for reaction energies and enthalpies,
- to 2 kcal/mol for free energies, and,
- less than 2 pKa for acid strengths.

Scaling of the computational time [24]

If,

N : is the number of orbitals or number of electrons

x : is an exponent which usually equals to or larger than 3.

Then, for most quantum mechanical methods, the computational time t_{CPU} is proportional to a certain power of the system size; N^x, and thiscan be denoted as : $O(N^x)$

When the different WFT methods are compared, their scaling can be given as follows:

Name	Scaling
HF	$O(N^4)$
MP^n	$O(N^{n+3})$
CCSD	$O(N^6)$
CCSD(T)	$O(N^7)$
CCSDT	$O(N^8)$
CCSDTQ	$O(N^{10})$

Table 2.1.1-1: *Scaling of CI methods [3]*

Then the order can be listed as:

$$HF << MP2 < CISD < MP4\ (SDQ) \sim CCSD < MP4 < CCSD\ (T)$$

Density Functional Theory [3, 8]

Density functional theory is often counted as an "ab initio" type of method since it is derived from the first-principles and that it does not require any adjustable parameters at all.

- Hohenberg and Kohn proved that the ground-state electronic energy for a system is a unique functional of its electron density [16]. In other words, it is shown that, there is one-to-one correspondence between the electronic density of a system and the energy. The exact correspondence is unfortunately not known.

- Kohn-Sham theory provides the kinetic energy with an assumption of non-interacting electrons. The difference between the exact kinetic energy and the Kohn-Sham theory kinetic energy is small. This remaining kinetic energy difference is expressed as the the exchange-correlation term (E_{xc} [ρ]).

As a result, general DFT equation can be given as follows:

$$E_{DFT}[\rho] = T_S[\rho] + E_{ne}[\rho] + J[\rho] + E_{xc}[\rho] \qquad \text{(Eqn. 2.1.1-114)}$$

DFT method has usually scaling as $O(N^3)$ to $O(N^4)$, but lower scaling algorithms are available. DFT accuracy is not as high as the higher level ab initio wave function methods, but it is much better than HF.

Hohenberg-Kohn (H-K) Theorem Proof by Contradiction [3]

For N nuclei and electrons, electronic Hamiltonian is:

$$H_e = -\sum_{i=1}^{N_{elec}} \frac{1}{2}\nabla_i^2 - \sum_{i=1}^{N_{elec}} \sum_{A=1}^{N_{nuclei}} \frac{Z_A}{|R_A - r_i|} + \sum_{i=1}^{N_{elec}} \sum_{j>1}^{N_{elec}} \frac{1}{|r_i - r_j|} + \sum_{A=1}^{N_{nuclei}} \sum_{B=A}^{N_{nuclei}} \frac{Z_A Z_B}{|R_A - R_B|}$$

(Eqn. 2.1.1-115)

H-K theorem can be proved by contradiction based on the following (wrong) assumption:

Assuming, if there are two external potentials(for example from nucleis) , V_{ext} and V_{ext}', resulting in the same electron density (ρ), then these two external potentials lead to two different Hamiltonians, H and H', where two lowest wave functions can be demonstrated as ψ and ψ' too. Due to the variational principle, and also first taking ψ' as an approximate wave function for H, and then ψ as for H', the following is obtained:

$$\langle \psi' | H | \psi' \rangle > E_0 \qquad \text{(Eqn. 2.1.1-116)}$$

$$\langle \psi' | H | \psi' \rangle + \langle \psi' | H - H | \psi' \rangle > E_0 \qquad \text{(Eqn. 2.1.1-117)}$$

$$E_0' + \langle \psi' | V_{ext} - V_{ext}' | \psi' \rangle > E_0 \qquad \text{(Eqn. 2.1.1-118)}$$

$$E_0' + \int \rho(r) \, (V_{ext} - V_{ext}') dr > E_0 \qquad \text{(Eqn. 2.1.1-119)}$$

$$E_0 - \int \rho(r) \, (V_{ext} - V_{ext}') dr > E_0' \qquad \text{(Eqn. 2.1.1-120)}$$

Adding the inequalities, one gets:

$$E_0' + E_0 > E_0' + E_0 \qquad \text{(Eqn. 2.1.1-121)}$$

which contradicts the assumption.

This is explained as having a one to one correspondence in between electron density and nuclear potential, and also with the Hamiltonian and the energy, so energy is a unique functional of electron density E[ρ].

Methods for Larger Systems [8, 25]

In general, for the molecules with sizes around 1000 atoms, molecular orbital calculations are easily managed with SQM types. However, for larger molecules the only option is to use the MM methods. At this point some alternatives were developed, which can be listed as follows:

I. Hybrid QM/MM approaches (to combine QM and MM methods)
II. Linear scaling algorithms.

I. QM/MM Methods

With this hybrid approach, reaction environment is divided into two different sections that are treated with different level of theories.

For the atoms that are in the actual reaction zone (this zone can be named as "core"), the electronic processes are assumed to be localized in small regions and they are treated with QM methods (i.e. DFT, ab initio or semiempirical methods based on the accuracy needed), meanwhile, the rest of the atoms that are far away can be treated with lower level methods (mostly with the classical force-field, MM method).
Creating a link between these two different approaches is difficult.

However, for a general description, when the following is assigned,

- MM; being the denotation which refers to either Molecular Mechanics or another lower level method,
- QM; Quantum Mechanical method,
- Core; stands for the atoms that are in the actual reaction zone
- System; stands for the full system,

Then the energy for a system can be described as follows:

$$E_{QM}(System) = E_{QM}(core) + E_{MM}(system) - E_{MM}(core) \qquad \text{(Eqn. 2.1.1-122)}$$

This combination is titled as "QM/MM" methods, which creates a fast and a powerful method allowing the simulation of the large systems.
This method can also be used to increase the accuracy of a particular calculation by using high level calculations to describe the central QM core region.

II. Linear Scaling

With these type of approaches, the goal is, to obtain a linear scaling of the computational effort with system size by exploiting the local character of most relevant interactions.

Semi-Empirical Methods [3, 25]

The cost of Hartree-Fock methods is mainly based on the number of two-electron integrals in the Fock-matrix. Semi-empirical methods reduce the number of these integrals and therefore reduce the computational costs. The large majority of semi-empirical methods consider only s and p functions.

Zero Differential Overlap (ZDO) approximation, is the main assumption of most semi-empirical methods. This approximation ignores all products of basis functions that depend on the same electron coordinates when located on different atoms.

Assuming there are two atoms, A and B, and denoting the atomic orbitals, as μ_A on centre A, and as ν_B on centre B, then, ZDO approximation is set equal to zero, $\mu_A \nu_B = 0$. This is the product of functions on different atoms which are set equal to zero, instead of having an integral result. [3]

In order to compensate the approximation that this condition brings, the remaining integrals are regarded as parameters, and these parameters are assigned based on computational or experimental data. Due to this, semi-empirical method types mostly vary based on these parameters.

While these integral approximations and parameterizations limit the accuracy of the semi-empirical methods, at the same time, they make the methods efficient enough to model large molecules in a successful manner.

Starting with the previous Eqn. 2.1.1-90 above,

$$D_{\gamma\delta} = \sum_j^{occ.MO} c_{\gamma j} c_{\delta j} \qquad \text{(Eqn. 2.1.1-90)}$$

$$F_{\alpha\beta} = h_{\alpha\beta} + \sum_{\gamma\delta} G_{\alpha\beta\gamma\delta} D_{\gamma\delta} \qquad \text{(Eqn. 2.1.1-123)}$$

denoting the two electron integrals as,

$$\langle \mu\nu | \lambda\sigma \rangle \qquad \text{(Eqn. 2.1.1-124)}$$

Then, the Fock matrix for semiempirical can be generally written as follows:

$$F_{\mu\nu} = h_{\mu\nu} + \sum_{\lambda\alpha}^{AO} D_{\lambda\alpha} [\langle \mu\nu | \lambda\sigma \rangle - \langle \mu\lambda | \nu\sigma \rangle] \qquad \text{(Eqn. 2.1.1-125)}$$

$$h_{\mu\nu} = \langle \mu | h | \nu \rangle \qquad \text{(Eqn. 2.1.1-126)}$$

From this point on, there are approximations for one and two electron parts (i.e. NDDO, INDO, CNDO types), and some additional information on semi-empirical methods will be given in the following sections.

Overview

To sum up, when MM, DFT and SQM are compared [25]:

- There is about six orders magnitudes [25, 26] of difference in between Molecular Mechanics (MM) and Density Functional Theory (DFT) methods (DFT; being a part of fully quantum mechanical (QM) methods).

- SQM methods are the simplest electronic structure theory methods. They are less robust and usually less accurate than DFT methods, but they are at least approximately 1000 times faster in computations.

- When compared with MM methods, SQM methods are about 1000 times slower than MM methods, but unlike the classical force fields, they can treat electronic effects (i.e. such as polarization effects, chemical reactions and electronic excitations).

Therefore, it can be stated that, SQM methods have a place in between DFT and MM methods with their realistic electronic structure calculations, together with an acceptably good accuracy and capability to treat large systems. This leads to some consequences:

- For example, even though lately DFT methods are also using linear scaling approaches which made them applicable for large biomolecular systems, in case of a DFT research (i.e. which involves fast pre-optimization of geometries, or high – throughput screening based in silico optimization studies), SQM methods can be used as the initial step for these type of calculations. Then, the rest of the research calculations with can be continued with DFT methods.

- SQM methods also use linear scaling type of algorithms (i.e. like MOZYME in MOPAC), however, the biggest advantage of the SQM methods is their speed compared to the DFT and their accuracy compared to MM methods [27].

The following Table 2.1.1-2 and Figure 2.1.1-3 can be an illustration for this comparison:

Method Type	General Features	Advantages	Disadvantages
Molecular Mechanics	• Uses classical physics • Relies on force-field with embedded empirical parameters	• Computationally least intensive – fast and useful with limited computer resources • Can be used for molecules as large as enzymes	• Particular force field applicable only for a limited class of molecules • Does not calculate electronic properties • Requires experimental data (or data from *ab initio* methods) for parameters • Inaccurate treatment of polarizatıon • Usually no bond-breaking possible, which means no reactivity
Semi-Empirical	• Uses approximation extensively • Uses experimentally or theoretically derived empirical parameters • Uses quantum physics	• Less demanding computationally than *ab initio* methods • Capable of calculating transition states and excited states • Includes polarization effects • Can model chemical reactivity	• Require experimental data (or data from *ab initio*) for parameters • Less rigorous than *ab initio*) methods • Fail for unusual compounds
DFT	• uses quantum physics • mathematically rigorous • few parameters	• Useful for a broad range of systems • Does not depend on experimental or theoretical data • Capable of calculating transition states and excited states • includes polarization effects • Can model chemical reactivity	• Computationally more expensive
WFT	• Very accurate	• Useful for a broad range of systems • Does not depend on experimental or theoretical data • Capable of calculating transition states and excited states • includes polarization effects • Can model chemical reactivity	• More expensive

Table 2.1.1-2: *Comparison of MM, DFT, WFT and semi-empirical methods (based on & inspired from, ref. [28])*

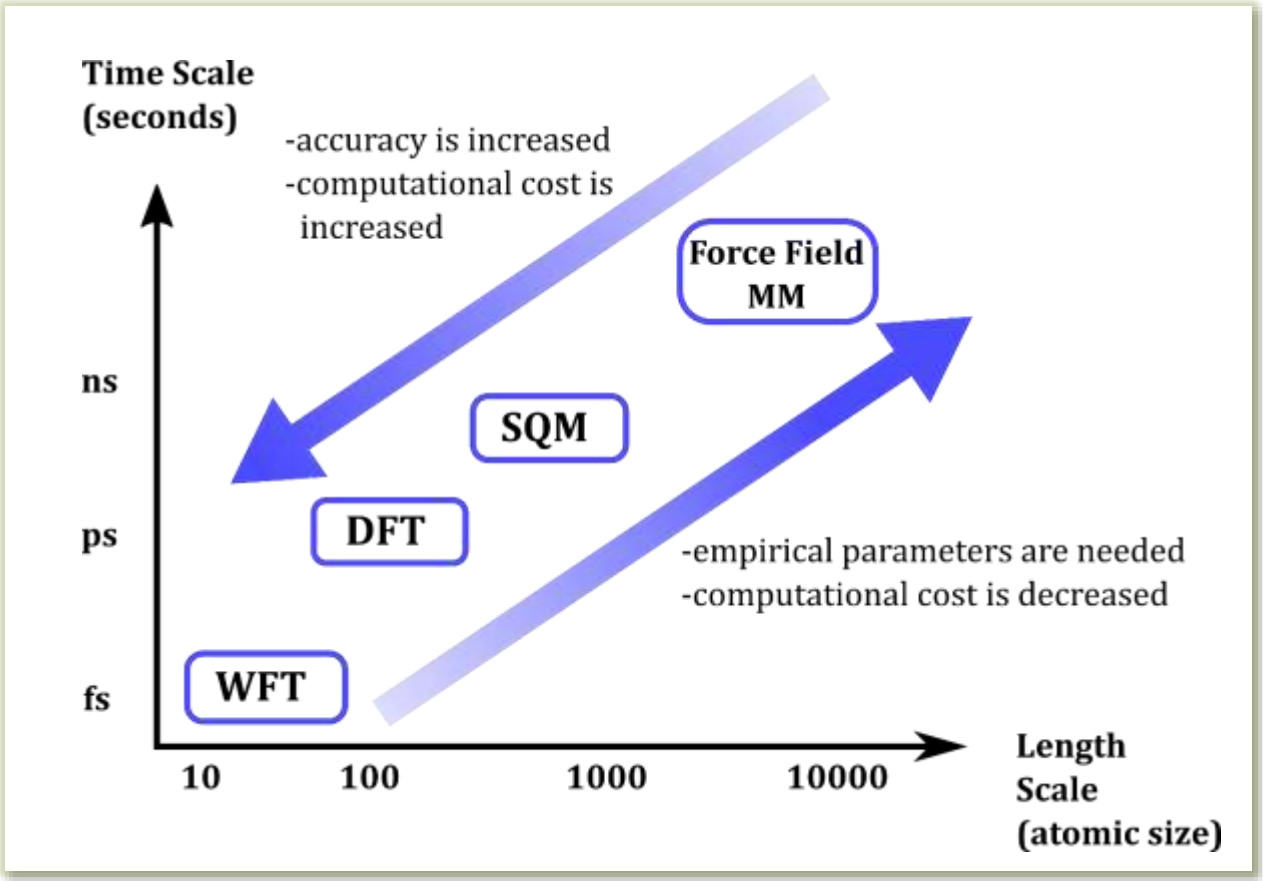


Figure 2.1.1-3: *Comparison of WFT (wave function theory), DFT (density functional theory), SQM (semi-empirical) and MM (molecular mechanics) methods (based on & inspired from ref. [29])*

2. 1. 2 Types of Semi Empirical Quantum Mechanical (SQM) Methods [3,25]

As it is previously mentioned, from the methodology point of view, semi-empirical methods simplify the Hartree-Fock Self Consistent Field-Molecular Orbital (HF-SCF) equations with their integral approximations [25].

There are three approximations made for one and two electron parts. These are,

- NDDO : Neglect of Diatomic Overlap Approximation
- INDO : Intermediate Neglect of Differential Overlap Approximation
- CNDO : Complete Neglect of Differential Overlap Approximation

Since the ZDO (Zero Differential Overlap) approximation already decreases the accuracy, the direct application of these NDDO, INDO and CNDO approximations are not resulting well either. In order to improve these approximations, some parameters and methods are introduced [3].

In Jensen's book [3], for improving these NDDO, INDO and CNDO approximations, there are three approaches which are nicely described for this purpose. Due to these:

I. Remaining integrals can be calculated from the functional form atomic data.
II. Remaining integrals can be made into parameters, which are assigned values based on a few (usually atomic) experimental data.
III. Remaining integrals can be made into parameters, which are assigned values based on fitting to many (usually molecular) experimental data.

As a result of these approaches, associatively, some methods were developed:
CNDO/1, CNDO/2, CNDO/S, CNDO/FK, CNDO/BW, INDO/1, INDO/2, INDO/S, SINDO1, and MSINDO.

Especially after M. J. S. Dewar combining the approaches II and III, a new class of "modified" models have been obtained and these are commonly used afterwards. Parameters for these methods are obtained by fitting, and the molecular data used for these parametrizations can be listed as: geometries, heats of formation, dipole moments and ionization potentials [25, 30-33].

The names of these modified models can be briefly listed below:

- MINDO: Modified Intermediate Neglect of Differential Overlap
  (MINDO/1, MINDO/2, MINDO/3)

- Modified NDDO models:
  - MNDO: Modified Neglect of Diatomic Overlap
    (MNDO, MNDOC)
  - AM1 : Austin Model 1
    - a straight re-parametrization of the AM1 method with a much larger set of reference data afforded the general-purpose RM1 variant with improved results
  - MNDO-PM3 (PM3 in short): Modified Neglect of Diatomic Overlap, Parametric Method Number 3
    - With a modified empirical core repulsion function using Pairwise Distance Directed Gaussians (PDDG), lead to PDDG/MNDO and PDDG/PM3.
  - MNDO/d: MNDO which also includes d orbitals
    - Contributed to the development of PM6 and PM7
  - SAM1 and SAM1D: Semi-ab initio Method 1, again based on NDDO.
  - OMx: orthogonalization models
    (OM1, OM2, and OM3): include orthogonalization corrections in the one-electron terms of the NDDO Fock matrix to correctly account for the effects of Pauli exchange repulsion [25, 34-37].

Concluding Remarks

MNDO models getting upgraded with d orbitals, and becoming MNDO/d lead to the development of PM6 and PM7 methods. The advantage of these PM6 and PM7 methods also arises from their capability to cover the whole periodic table elements, and therefore enabling them to compute molecular and solid-state properties [25, 38, 39].

In addition to these above listed traditional semi-empirical methods, which are in other words simplified ab initio MO treatments, there are also other methods available and popular among biochemical and materials science studies. These are: Semiempirical tight-binding (TB) versions of DFT methods (DFTB approach), and, the self-consistent charge (SCC) DFTB methods [25, 40-43].

2. 1. 3 Enhanced SQM Methods

For biomolecular interactions, SQM methods have an advantage over Force Field (FF) methods by taking the charge transfer and polarization effects into account, while these effects are not accurately described with FF methods [44, 45].

Nonetheless, previously, SQM methods had their own problems especially with hydrogen bond and dispersion interactions. These interactions are improved with the help of various method development studies leading to the "enhanced SQM" methods (SQM-DH, D: stands for dispersion, H: stands for hydrogen bonds) [46-49]. The relevant studies are briefly listed in Table 2.1.3-1.

SQM-DH methods

Dispersion Correction

Dispersion interactions are improved for SQM-DH methods with the inclusion of the FF terms to the empirical hydrogen bonding correction, so that it gets closer to the level of DFT-D methods [50]. Similar to the DFT-D methods, SQM-DH methods also have either system independent D2-type [51] or system dependent D3-type [52] of coefficients [27, 53]. The following equations can be presented [55]:

$$E_{dispersion} = -S_6 \sum_i^N \sum_j^N \frac{C_6^{ij}}{R_{ij}^6} . f_{damping} \qquad \text{(Eqn. 2.1.3-1)}$$

R_{ij} : the interatomic distance between atoms i and j.
C^{ij}_6 : the dispersion coefficient for the pair of atoms i and j (calculated from the atomic C_6 coefficients)
s_6 : scaling factor
$f_{damping}$:damping function

$$f_{damping}(R_{ij}) = \frac{1}{1+e^{-\alpha\left(\frac{R_{ij}}{R_0-1}\right)}} \qquad \text{(Eqn. 2.1.3-2)}$$

R_0 : the sum of the atomic van der Waals radii
α : a parameter determining the steepness of the damping function.

Once the correction part is defined, it is added to the main energy.

$$E_{SQM\text{-}D} = E_{SQM} + E_D \qquad \text{(Eqn. 2.1.3-3)}$$

Hydrogen Bond Correction

Different than the dispersion correction approaches, Hydrogen bond interaction was subjected to several suggestions because of its complicated nature. Hobza and co-workers [55] approached to the problem in a similar way to the dispersion problem case. They introduced a correction that can be added to an unmodified semiempirical calculation.

Zeroth-, first-, second- and third-generation H terms were introduced in detail [27, 56], and if the overall effect is described as:

$$E_{hydrogen-bond} = g_{distance} \cdot h_{orientation} \quad \text{(Eqn. 2.1.3-4)}$$

$E_{hydrogen\text{-}bond}$: a function that serves for the sterical arrangement of the two fragments relative to eachother and positioning of an H atom in between them.
$g_{distance}$: a function based on distance
$h_{orientation}$: a function based on orientation,

Then, the description of these 0th 1st 2nd and 3rd generation terms are expressed as follows:

$$g_{distance}^{\text{0th,1st and 2nd generation}} = f(r_{HA}) \quad \text{(Eqn. 2.1.3-5)}$$

$$g_{distance}^{\text{3rd generation}} = f(r_{DA}) \quad \text{(Eqn. 2.1.3-6)}$$

$$h_{orientation}^{\text{0th generation}} = 1 \quad \text{(Eqn. 2.1.3-7)}$$

$$h_{orientation}^{\text{1st generation}} = f(\Theta) \quad \text{(Eqn. 2.1.3-8)}$$

$$h_{orientation}^{\text{2nd generation}} = f(\Theta, \phi, \psi) \quad \text{(Eqn. 2.1.3-9)}$$

$$h_{orientation}^{\text{3rd generation}} = f(\Theta, \phi, \phi 2, \psi, \psi 2, r_{HX}) \quad \text{(Eqn. 2.1.3-10)}$$

r_{HA} : hydrogen acceptor distance
r_{DA} : donor acceptor distance
r_{HX} : hydrogen-electronegative atom distance

0th-generation approach:

- Non-directional terms are added.

1st-generation approach:

- Directional terms that depend only on the acceptor–hydrogen distance, and,
- The main (donor–hydrogen–acceptor) angle are introduced.

2nd-generation approach:

- A secondary (base–donor/acceptor–hydrogen) angle, and ,
- Torsional (base–donor/acceptor–hydrogen) angle are introduced.

3rd-generation approach:

- Hydrogen-bonds are taken as the interaction between two electronegative atoms (X and Y), which is smoothly switched on by the favourable placement of one (or more) hydrogen atom(s) in between them.

REMARKS [27, 56]

1st generation terms were and still are used by some FF methods, but for large systems this type is not regarded as advantageous.

The main advantage of 2nd generation terms is the exclusion of unphysical interaction contributions which arise for 1st generation terms [53, 57] and was leading to substantial problems with geometry optimizations.

In case of the 3rd generation types, the most important part is regarded as the change from the use of hydrogen-acceptor distance into a core-core interaction namely by using the donor-acceptor distance instead. This approach is therefore more robust, for example when it is about a proton transfer.

DH2 hydrogen-bond correction is basically considered as charge-independent between two atoms that are regarded as Donor and Acceptor. It weighs this term with a function which shows the sterical arrangement of these two relative to eachother and relative to the placement of an H atom in between them.

In order to describe the geometric information to model the behaviour (directionality) of hydrogen-bonds, the following indications are used in this model:

- A possible hydrogen donor (D)
- Hydrogen (H)
- A possible hydrogen acceptor (A)

As also shown in Figure 2.1.3-1, main coordinates are described as [27, 53, 56]:

- Distance between the Hydrogen and the Acceptor,
  H···A
- Main hydrogen bond angle: involving Donor, Hydrogen, Acceptor and their angle in between (Θ),
  D–H···A, and Θ
- Secondary angle: involves Hydrogen, Acceptor, Acceptor based Atom I, and their angle in between (ϕ),
  H···A–R_2, angle ϕ; R_2 is a donor "base atom"
- Torsion angle: involving Hydrogen, Acceptor, Acceptor based Atoms I, and the angle (ψ)
  H···A–R_1–R_2, angle ψ

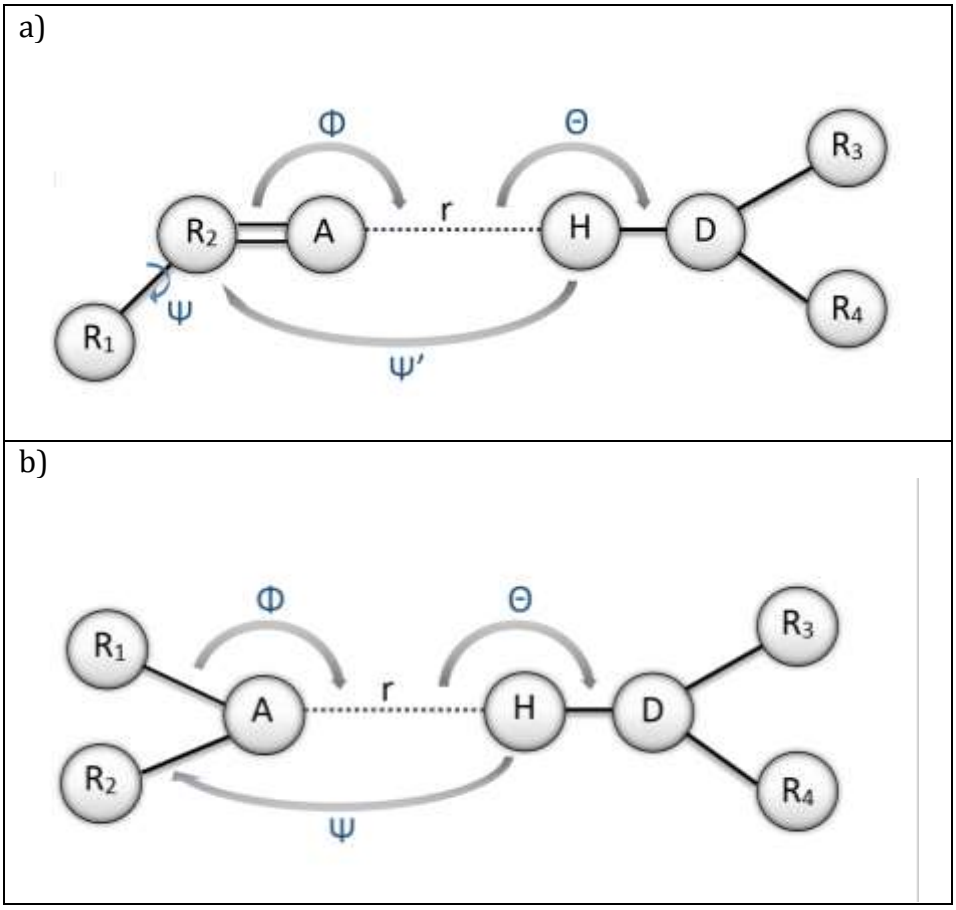


Figure 2.1.3-1: *Graphical representation for the geometric features of hydrogen bonding. Illustrates the H-bond distance r and the angles: Θ, Φ, and Ψ for two different cases. Figure a) shows sp2 oxygen-type acceptor atom, whereas Figure b) shows sp2 nitrogen or general sp3-type acceptor atoms that require a different choice of atoms for the definition of the torsion angle coordinate. The out-of-plane "movement" in case a (Ψ′) is actually realized by a combined change of the two internal coordinates: Φ and Ψ. (Based on a figure from ref. [53])*

In DH+ the following approach is applied.

Similar to the descriptions above,

A and B	: two donor/acceptor atoms,
C_A and C_B	: two element-wise correction parameters respectively.
ϕ, ϕ_2, ψ, ψ_2	: Angles can also be used for both of these atoms in a symmetric way.
r_{AB}	: the interatomic distance between atoms A and B.

The obtained term is also corrected with a damping function, f_{damp}. These can be demonstrated as below [56]:

$$E_{H-bond} = \frac{C_{AB}}{r_{AB}^2} h_{orientation} f_{damp} \quad \text{(Eqn. 2.1.3-11)}$$

$$h_{orientation} = \cos(\Theta_A)^2 \cos(\phi_A)^2 \cos(\psi_A)^2 \cos(\phi_B)^2 \cos(\psi_B)^2 f_{bond} \quad \text{(Eqn. 2.1.3-12)}$$

$$f_{bond} = 1 - \frac{1}{1+e^{\left[-60\left(\frac{r_{XH}}{1.2}-1\right)\right]}} \quad \text{(Eqn. 2.1.3-13)}$$

$$f_{damp} = \left\{\frac{1}{1+e^{\left[-100\left(\frac{r_{AB}}{2.4}-1\right)\right]}}\right\} \cdot \left\{1 - \frac{1}{1+e^{\left[-10\left(\frac{r_{AB}}{7.0}-1\right)\right]}}\right\} \quad \text{(Eqn. 2.1.3-14)}$$

$$C_{AB} = \frac{C_A + C_B}{2} \quad \text{(Eqn. 2.1.3-15)}$$

Damping function can be selected so that no fitting will be necessary for it. Also a long-range cutoff can be regarded as a fit parameter.

The following explanation is directly taken from the main reference [50],
For f_{damp}, it is:

- "Smoothly switched on between a donor-acceptor distance of 2.3 and 2.5 Å (safe choice for the assumption of no H bonds below 2.5 Å), and,
- Smoothly switched off between 3.5 and 10.5 Å (safe choice for the assumption of full H-bond strength up to 3.5 Å and no strength anymore at three times this distance)."

And meanwhile for f_{bond},

- "It brings the correction to zero if the hydrogen wanders away too far from both electronegative atoms (with r_{XH} being the smaller one of the two distances r_{AH} and r_{BH})
- Smoothly switched off between 1.15 and 1.25 Å (safe choice for the assumption of a maximum distance of 1.15 for a covalent hydrogen bond)."

C_A and C_B values can be found in the following lists (List 2.1.3-1 and List 2.1.3-2) for semi empirical and for force-field methods respectively [50].

Element	Method			
	OM3	PM6	AM1	DFTB
N	-0.05	-0.16	-0.29	-0.21
O	-0.07	-0.12	-0.29	-0.08

List 2.1.3-1 *Hydrogen Bonding Correction Parameters for Semi Empirical Methods*

Element	Method					
	MM2	MM3	AMBER	OPLS	OPLSAA	MMFF94
N	-0.64	-0.63	-0.21	-0.24	-0.25	-0.21
O	-0.08	-0.17	-0.03	-0.00	-0.00	-0.05

List 2.1.3-2 *Hydrogen Bonding Correction Parameters for Force Field Methods*

Our Reference Method: PM6-DH+ [27, 50, 56]

PM6-DH+ is a third generation SQM-DH model developed by Korth. It is currently the only method where the distinction between the acceptor and donor atoms are skipped to avoid many conceptual problems. As a result, there is only one parameter per electronegative element left to fit in order to reach a high accuracy level.

Its geometric factor takes into account all available—that is, angular and torsional—information. Its core/core-term like definition, as well as its element wise parameterization (with the corresponding low number of parameters to fit), makes it to be generally applicable and well transferable.

PM6-DH+ is implemented in MOPAC2009 [58].

Within our research, the validity checks for PM6-DH+ method are made with experimental data set references, and its performance is compared with other possible candidate computational methods.

The names of the commonly used enhanced SQM (D/H) methods are given in the following Table 2.1.3-1 together with their specifications.

Name of the Method	Developed By	Year	Additional Specification
AM1-D	Collignon and co-workers [59]	2006	• First, mostly unrecognized version of a SQM-D method
PM3-D & independently AM1-D	McNamara and Hillier [54]	2007	• Independent reimplementation of the SQM-D approach
OMx(-D)	Tuttle and Thiel [60]	2008	• Extension of the established OMx methods with D corrections
PM6-DH	Rezac and co-workers [55]	2009	• PM6-DH is the first dispersion and hydrogen-bond corrected SQM method. It never was publicly available due to severe technical problems.
PM6-DH2	Korth, Hobza and co-workers (under supervision of Hobza)[53]	2009	• PM6-DH2 is the first robust SQM-DH method readily available in MOPAC [58] from MOPAC2009 on and now used by many groups around the world for problems in life and materials science [27, 61]
Approach	Wang and Bryce [63]	2009	• Approach to add MM dedicated hydrogen-bond terms as a QM/MM interface term (in addition to the usual dispersions and repulsion terms) was pursued.
AM1-FS1	Foster and Sohlberg [64]	2010	• Also makes use of both hydrogen-bond and dispersion terms. • Performance is roughly similar to PM6-DH2
PM6-DH+	Korth [50]	2010	• PM6-DH+ allows for proton transfer reactions and uses only 2 fit parameters while not losing accuracy in comparison to PM6-DH2. • PM6-DH+ is readily available in MOPAC [58] from MOPAC2009 on, and is now used by many groups all over the world for problems in life and materials science [27, 62]
PM6-DH2X	Rezac and Hobza [65]	2011	• PM6-DH2X includes halogen-bond (X) correction terms, halogen-bond (X) terms, which work analogously to hydrogen-terms, but with an opposite sign, as standard SQM methods underestimate repulsion in halogen bonds and thus deliver distances too short and interaction energies too high [65, 66]. • Similar corrections were later on also applied to MM methods [67, 68].
Higher-level theory approaches	Laikov [69]	2011	• A systematic derivation of SQM parameters from higher level coupled cluster data is made.
PM6-D3H4	Rezac and Hobza [70]	2012	• PM6-D3H4 includes both improved dispersion corrections terms of D3-type and a new hydrogen-bond correction scheme (nevertheless neglecting important terms), and it is claimed to have an increased robustness for geometry optimizations and molecular dynamics simulations.
Post-SCF correction	Foster and Sohlberg [71]	2012	• Self-consistent, atomic charge dependent hydrogen bond correction terms. They are usually added as post-SCF correction and thus lead to non-variational methods if partial charges are used – which is a problem most of all for PM6-DH2 [64].

PM3-D and AM1-D models	Anikin et al. [72]	2012	• Independent re-implementation of the older models
corresponding GPU-enabled algorithms	Carvalho, Maia, And co-workers [73]	2012	• This algorithm allowed for a very impressive illustration of the capability of PM6-DH+ to identify native protein structures out of large sets of decoy conformations.
PM7	Stewart [39]	2013	• PM7 includes dispersion and hydrogen-bond correction terms of mixed PM6-DH2/PM6-DH+ type directly into the SQM fitting process. • Performance for non-covalent interactions is roughly similar to PM6-DH2/DH+ • Available in MOPAC [58]
Polarized molecular orbital (PMO)	Truhlar and co-workers [74-76]	2011 - 2014	• PMO with damped dispersion and orbitals on H atoms for—also hydrogen–bond type–polarization effects, currently parameterized for H, C, N, O, and S. • Benchmark data in comparison to SQM-DH methods is not available.
Development of minimal QM models, especially HF-3c	Sure and Grimme[77]	2013	• Comparably slower but supposedly more robust than SQM-DH methods and thus they are capable of filling the gap between SQM and DFT methods in terms of cost and accuracy. • Performance for non-covalent interaction roughly similar to PM6-DH2/DH+.
PM6-D3H+	Korth,Jensen and co-workers [78]	2014	• PM6-D3H+, is an updated PM6-DH+ model with an improved dispersion corrections of D3-type and a more robust (third-generation) H+ term, for an improved performance in geometry optimizations and molecular dynamics. • Source code freely available on GITHUB [79].
MSINDO-D3H+	Grimme, Bredow and co-workers [80]	2014	• MSINDO-D3H+ uses more recent D3-type terms and the above mentioned improved implementation of the H+ term by Korth/Jensen for enhancing Bredow's MSINDO approach.
AM1/d-CB1	Govender, Naidoo and co-workers [81, 82]	2014	• Includes some d-orbitals and new core-core repulsion terms. • Benchmark data in comparison to SQM-DH is not yet available.

Table 2.1.3-1: *List of Enhanced Semi Empirical Methods*

Concluding Remarks

Based on the latest developments and updates in literature, the following parts from the Table 2.1.3-1 can be highlighted as a summary of this section:

- OMx methods are probably the best choice for non-covalent interactions, but as a disadvantage, they are limited to first row elements for now. This limited applicability unfortunately prevents this method from being available for many systems. This results in the preference of several other commonly used methods.

- In that regard, SQM-DH methods are currently the most common approaches for many applications, especially where the non-covalent interactions and computational costs are the essential concerns.

The next section will list the applications of several enhanced semi empirical methods for various fields.

Applications of the Enhanced Semi Empirical Methods

I. Benchmarking – small biomolecular model systems

Latest significant studies are listed below in Table 2.1.3-2.

Date	Studied By	Content
2011	Prenosil and co-workers [83]	Study of hydrogen-bond cooperativity effects with PM6-DH2 in comparison to MM, DFT and WFT methods. **Outcome:** Unlike MM methods, PM6-DH2 performs reasonably accurate in comparison to high-level Coupled Cluster data.
2012	Rezac and co-workers [66]	Investigation of the performance of their newer PM6-DH2X and PM6-D3H4X models for a large number of halogenated systems.
2013	Hostas and co-workers [84]	Usage of OMx-D and PM7 to benchmark the performance of all these methods for non-covalent interactions including conformational changes.
2013	Sedlak and co-workers [85]	Study on large dispersion-dominated biomolecular systems. **Outcome:** An excellent price/performance ratio for enhanced SQM methods is found.
2014	Li and co-workers [86]	A benchmark study with Hobza's BEGDB benchmark database using several SQM, SQM-DH, DFT, and symmetry-adapted perturbation theory (SAPT) methods. **Outcome:** SQM-based methods are dramatically faster than DFT and SAPT methods (and thus readily available for large systems), but also that they are somewhat less accurate. None of the tested SQM-DH methods was found to be clearly superior with respect to the achieved accuracy; the authors do not discuss the conceptual advantages and disadvantages of the individual methods.
2014	Barberot and co-workers [87]	Benchmark study using their AlgoGen genetic algorithm approach for the extensive sampling of the conformational space of typical benchmark set systems in order to investigate the performance of the PM6-DH+ model. **Outcome:** Excellent price/performance ratio, but, several spurious minima are found.

Table 2.1.3-2: *List of benchmarking studies on small molecular model systems with enhanced semiempirical methods.*

II. Benchmarking – large biomolecular model systems

Large biomolecular systems require more computational power or faster computational methods. From this perspective, the amount of benchmark studies on large systems are comparably far less than the benchmark studies on small systems, and some of these are tabulated below.

Our first two publications with large biomolecular model systems, which will be further explained in Sections 3.1.1 and 3.1.2 (as Research Stage I [89] and II [91]), also belong to this category of benchmarking.

Date	Studied By	Content
2012	Mikulskis and co-workers [88]	Investigation of the performance of MM and SQM methods including (their own versions of) dispersion and hydrogen-bond corrections for protein–ligand interactions. **Outcome:** The importance of empirical corrections for SQM methods is highlighted and AM1-DH version is suggested as a competitive alternative to MM/GBSA calculations.
2013	Yilmazer and Korth [89]	PM6-DH+ method's performance is studied in comparison to DFT methods for several hundred systematically generated protein–ligand model systems from the PDBbind2007 [90] benchmark set. **Outcome:** An excellent performance of the SQM-DH method in comparison to DFT-D data is found.
2015	Yilmazer, Korth and co-workers [91]	PM6-DH+ and DFT methods are compared to high-level WFT reference data for the smallest complexes (from the large biomolecules list of their previous study). **Outcome:** SQM is performing as well as DFT, which in turn performs as well as WFT method.

Table 2.1.3-3: *List of benchmarking studies on large molecular model systems with enhanced semiempirical methods.*

III. Benchmarking – interactions in water

Interactions in water or basically solvation effects, and also the proton transfer phenomena are important concerns. Proton transfer is possible only with SQM methods but not with MM methods. The relevant SQM-DH studies can be listed as follows:

Date	Studied By	Content	
2013	Bulo and co-workers [92]	Focusing on the investigation of QM/MM setup parameters, a benchmark of solute–water interactions using PM6-DH+ and DFTB in a QM/MM setup is published.	
2014	Marion and co-workers [93]	The performance of OMx-D, PM6-DH2, PM6-DH+, PM6-D3H4 and PM7 methods are benchmarked amongst eachother for the interaction of water with hydrophobic groups. **Outcome**: A new model is developed, PM3-PIF3 to describe systems in aqueous solutions.	
2013	Wu and co-workers [94]	Special models were constructed for OMn, and proton-transfer in water.	**Outcome:** These models outperform 'pure' SQM methods, but data for a comparison with SQM-DH methods is still missing.
2014	Wang and co-workers [95]	Special models were constructed by AM1-W and AM1PGG-W, for proton-transfer in water.	

Table 2.1.3-4: *List of benchmarking studies on water interaction model systems with enhanced semiempirical methods.*

IV. Benchmarking – various

There are also other type of benchmarking studies which are worth mentioning.

Date	Studied By	Content
2010- 2012	Merz and co-workers [96-99]	Propagation of systematic and random errors in the computation of protein–ligand interaction energies and protein-folding including MM, SQM-DH and DFT data. **Outcome:** DFT was found to be the best choice, but interestingly some MM potentials outperformed some (non-enhanced) SQM methods, illustrating the need for empirical corrections.
2011-2014	Truhlar and co-workers [74-76]	Benchmark study of PM7 amongst other methods, and the development of their polarized molecular orbital (PMO) method. **Outcome:** PMO2 outperforms PM7 on their set of atmospherically relevant compounds.
2010-2013	Raju, Leverentz and co-workers [100-101]	Benchmark study involving AM1-D and PM3-D. **Outcome:** A sub-optimal performance of SQM-D methods for hydrogen bonded systems is found.

Table 2.1.3-5: *List of other type of benchmarking studies with enhanced semiempirical methods.*

V. (Pre-)optimization, dynamical studies, structure refinement, conformational searches

There are also other application areas where SQM-DH are used due to their advantageous calculation cost and speed. Type of the studies can generally be listed as follows:

- Fast optimization [102-103] and pre-optimization of biomolecular systems with SQM-DH, prior to higher level computations [104-105] ,
- Studying SQM-DH as an intermediate level in hybrid systems (as an example, within DFT-D, SQM-DH and MM) [106],
- In dynamical QM/MM calculations, SQM-DH as the QM method [107],
- SQM-DH based X-ray structure refinement [108]
    - α-helical structures [109]
- Non-local optimization of molecular structures with SQM-DH, for screening:
    - the conformational space of the FGG tripeptide [110]
    - DNA quadruplex/molecule complexes [111] .
- Helping the analysis of the experimental data [112-115]

VI. Host/guest systems

In addition to pre-optimization, especially for biomolecular interactions, SQM-DH methods are valuable tools. When compared to high level method results (like gold standard: CCSD (T)/CBS references), their accuracy is good enough together with their advantageous computational costs. Then, the arising question is whether SQM-DH methods can replace DFT-D methods for this problem.

Several studies on this host/guest category include research on:

- Complexes [115-120] ,
- Molecular tweezers [121] ,
- Non-covalent complexation [122-126],
- Supramolecular chemistry [127-128] .

Studies of Hari S. Muddana and Michael K. Gilson so far can be pronounced as the main contributions to this class of applications. Some of their studies were about:

- 29 CB7 host-guest systems with PM6-DH+/COSMO based on their 'minima mining' (M2) approach for predicting binding affinities and found good agreement with experiment [129],
- Blind prediction of 14 CB7 binding affinities within the SAMPL4 challenge [130].

VII. Materials science

SQM-DH methods are also used for investigating:

- Nano systems with including graphite, graphene, fullerenes, nano tubes, DNA bases and combinations [131-144],
- Molecular self-assemblies [145-148]
- Molecular switches [149]
- Screening of thousands of compounds for methane storage [150-152] and hydrogen storage [153],
- Adsorption energies on graphene by PM6-DH2 method [154], (as an outcome, the accuracy was comparable with DFT-D method and also was in a good agreement with experimental studies),
- Calculation of small molecule adsorption energies on graphene in comparison to FF, other SQM and DFT approaches, (as an outcome, PM6-DH+ is found to be the most efficient method. Chemical accuracy for PM6-DH+ was comparable to temperature-programmed desorption experiments. First tests for the rational design of improved graphene surfactants were presented [155].),
- Other studies where SQM-DH performance had similar results with the DFT-D performance [114, 109, 156,157, 158, 138, 139, 148, 159- 164],
    - Sometimes high accuracy was obtained with SQM-DH (PM6-DH2) methods as in the interaction of DNA bases with Li@C60 [165].
    - Sometimes low accuracy was obtained, but usually with older, only dispersion corrected AM1-D or PM3-D methods [166-168].

VIII. Other Applications

Due to latest developments, there are also studies where SQM-DH methods are uncommonly used and promising results were obtained. These are the studies, where,

- Chiral discrimination [169] and piezoelectric effects of applied electric fields on hydrogen-bond interactions [170] are investigated,
- PM6-DH+ is used for computing protonation sites and proton affinities of amino acids with the goal of solving mass spectrometry problems [171],
- PM6-DH+ derived descriptors are used to estimate the glass transition temperature for 209 molecular liquids [172].
- A heuristic approach is developed to estimate kinetic effects in complex chemical reaction networks [173], by using PM7 (including DH terms) for structure optimization and reaction thermodynamics.

IX. Virtual Drug Design

The advantages of SQM-DH methods have made them interesting for the field of rational drug design, and there are several applications for the virtual drug design field.

This field is also closely related to our research [89, 91], therefore in Section 2.2 the relevant concepts and terminology will be explained.

Applications and studies in this category are tabulated in Table 2.1.3-6.

Date	Studied By	Content
2009	McNamara, Hillier and co-workers [174]	QM/MM-based scoring is studied with SQM-D.
2010	Hobza and co-workers [175]	Scoring of 22 HIV-1 ligands with PM6-DH2/SMD (and MM-based entropy terms) **Outcome:** A substantial improvement over the conventional DOCK procedure. The authors emphasize that their scheme is free of system-specific parameters and thus readily available also for other protein/ligand systems.
2011	Hobza and co-workers [176]	Scoring 15 structurally diverse CDK2 inhibitors with PM6-DH2/COSMO **Outcome:** Very good agreement with experiment is found.
2011, 2013	Hobza and co-workers [177,158]	Investigation of SmCB1 inhibitors. **Outcome:** Results are in close agreement with DFT-D data.
2011	Nagy and co-workers [178]	Computation of DNA/zinc–finger–protein interactions.
2011	Kamel and Kolinski [179]	QM-based scoring is studied with SQM-DH.
2012	Avila Salas and co-workers [180]	Computation of the interaction energies of 4 drug molecules with 8 polyamidoamine dendrimer fragments from overall 320 million configurations of about 30 to 170 heavy atoms with PM6-DH+. **Outcome:** An excellent correlation with experiment is found (R^2=0.9), especially in comparison to MM data- experimental correlation (R^2= 0.75).
2012	Benson and co-workers [181]	Prediction of the enthalpic part of the SAMPL3 challenge [182] trypsin/fragment binding affinities with PM6-DH2 to in combination with different solvation models [181].

		Outcome: The importance of (multiple) docking poses and the influence of the solvation model are emphasized, so that, at least in their approach, SQM-DH methods offer no benefits over purely empirical scoring functions.
2012	Quevedo and co-workers [183]	The interactions of organic molecules in the solid state and in solution are investigated.
2012	Kamel and Kolinski [184]	QM-based scoring is studied with SQM-DH.
2012	Stigliani and co-workers [185]	Docking based on Autodock Vina structures which are re-ranked with PM6-DH2/COSMO, is studied. **Outcome:** This version is found to improve the docking results.
2013	Pan and co-workers [186]	QM/MM-based scoring is studied with SQM-DH.
2013	Ahmed and co-workers [187]	Silico design of biologically active compounds is investigated.
2014	Ucisik and co-workers [188]	Prediction of protein–ligand binding affinities is studied by using Monte Carlo estimates of the configuration integrals based on SQM-DH microstate energies using different implicit solvation models **Outcome:** A good correlation with experiment is found for PM6-DH2-based estimates.
2014	Temelso and co-workers [189]	Screening of peptides with anti-breast-cancer properties at different theoretical levels.
2014	Pavlicek and co-workers [190]	QM/MM-based scoring is studied. **Outcome:** With SQM-DH, a good agreement with DFT is found.
2014	Kruse and co-workers [191]	Complete nucleic acids building blocks are investigated.
2015	Hobza and co-workers [192]	Malonate-based inhibitors of mammalian serine racemase are studied, a new repulsion correction to their PM6-D3H4X/COSMO approach is added.

Table 2.1.3-6: *List of virtual drug design related studies with enhanced semiempirical methods.*

From the studies listed above, it can be concluded that, SQM-DH methods are indeed good candidates for many different scientific fields, like:

- (bio-) organic/inorganic hybrid materials,
- bio-nano structures,
- de novo design & optimization of functional bio-macromolecules.

In the following, there will be two separate sections based on our two different (one major, one minor) projects.

Our main research is related to “Computer aided drug design”, therefore in Section 2.2 some fundamental concepts will be given.

The minor research is related to “Computational screening of battery electrolyte materials”, which will be briefly introduced in Section 2.3.

2.2 Overview of Computer Aided Drug Design

Some lexical definitions of the commonly used terms are introduced here.

Drug:
A chemical compound and most commonly an organic small molecule that activates or inhibits the function of a biomolecule (i.e. protein) which is used for a diagnosis, cure, treatment, mitigation or prevention of a disease. [193, 194].

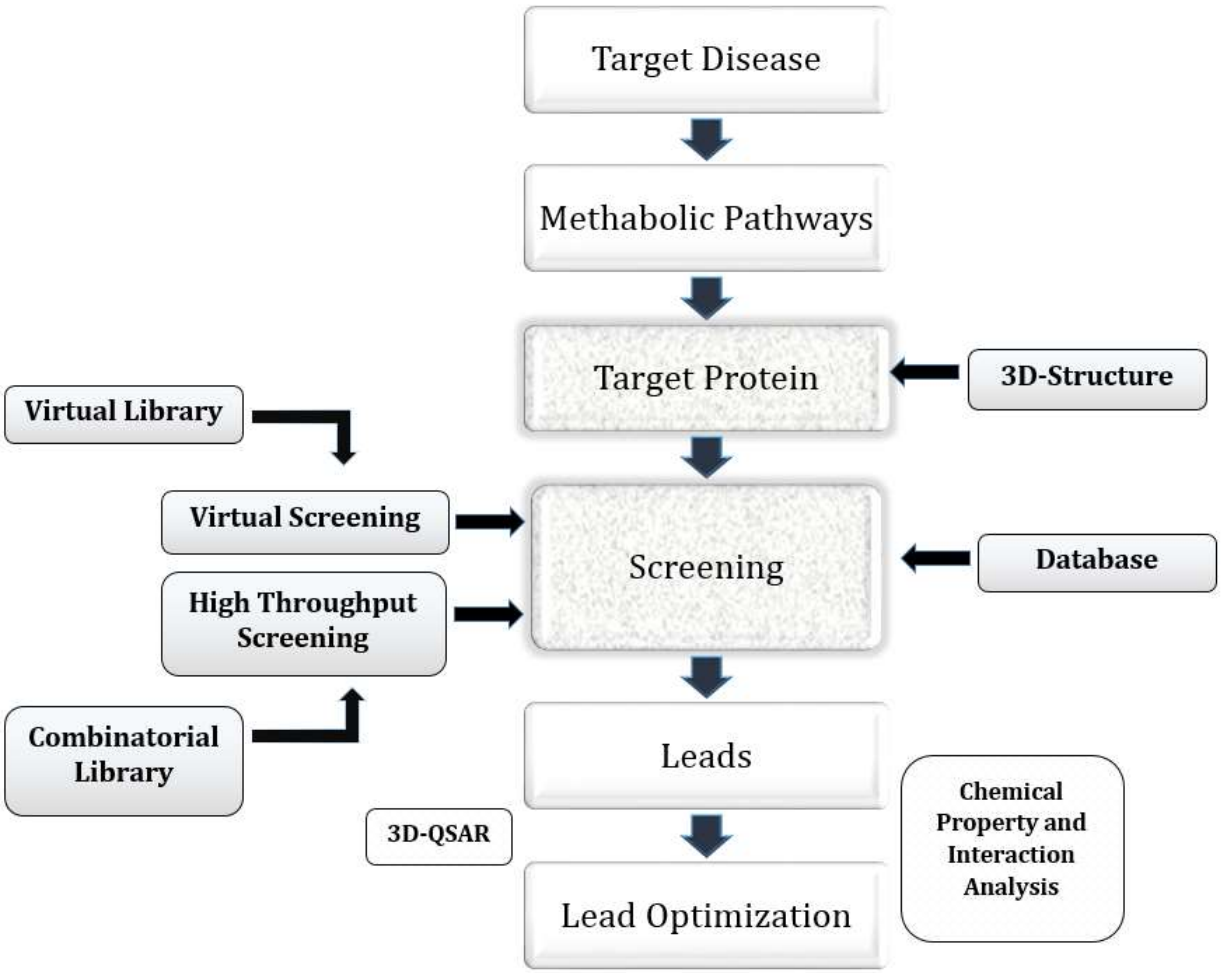
Protein:
A complex organic compound group, which mainly contains carbon, hydrogen, oxygen, nitrogen and sulphur and occasionally some other elements. It essentially consists of combinations of amino acids [193].

Ligand:
An organic molecule (which can be an antibody, hormone or a drug) that binds to a receptor (protein) [195].

Ligand /Protein binding site:
A location on a protein structure where the chemical interaction (the binding) with a ligand takes place. [196]

Drug Design:
The inventive process of designing or finding new medications out of small molecules (i.e. ligands) which are complementary in shape and charge to a biomolecular target (i.e. protein) so that they will interact with and bind to the target [197, 198].

The overall workflow for a drug design can be given in the following Scheme 2.2-1



***Scheme 2.2-1**: Simplified Drug Design Workflow based on ref.* [199].

Briefly, [200] drug design aims to:

I. Predict the binding mode of a known active ligand
II. Identify new ligands
III. Predict the binding affinities of related compounds from known active series

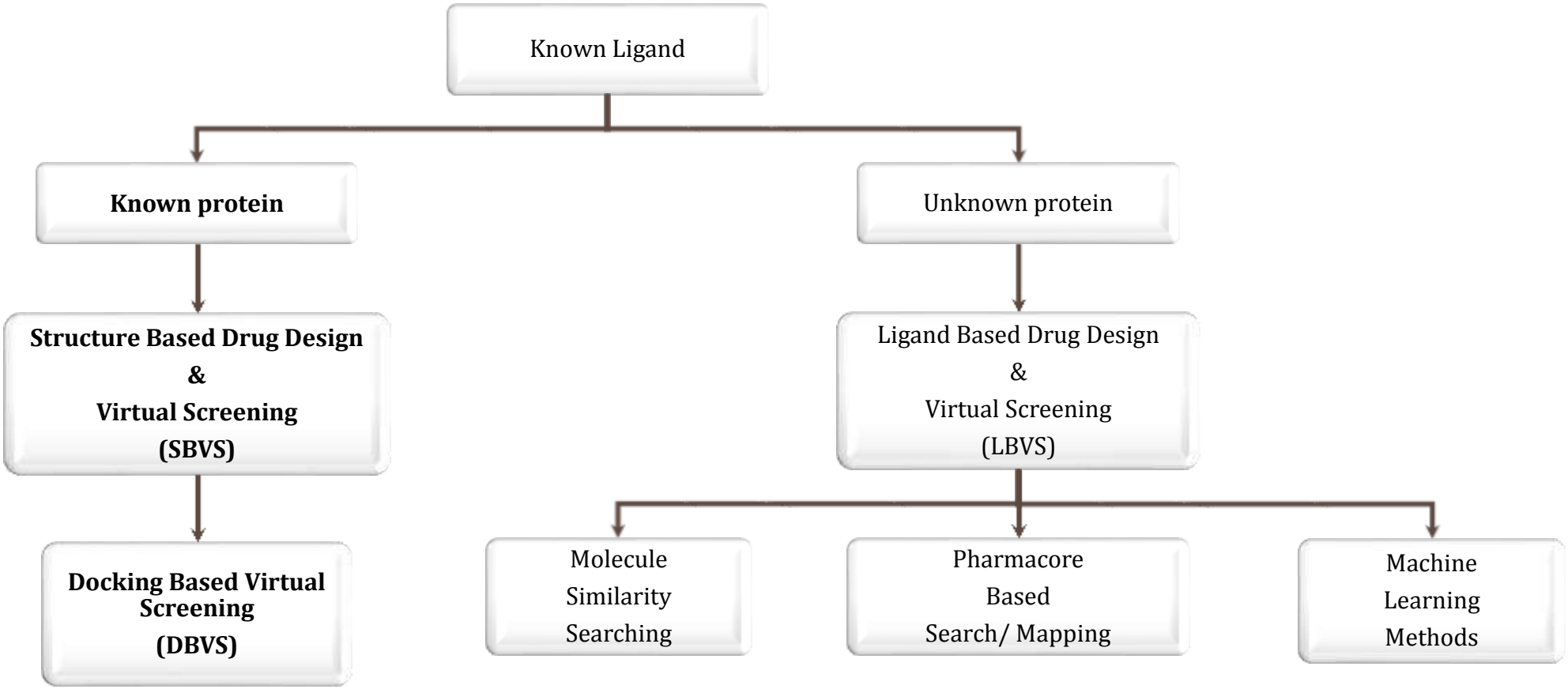
Virtual screening, is an automated computational technique to evaluate very large libraries of compounds [201] in order to diminish the needed experimental effort and to increase the hit rate in the selection of new drug candidates. This approach is a common procedure for now for many pharmaceutical companies [202, 203].

Virtual screening can be classified in two main categories [199]:

- Ligand-based virtual screening (LBVS),
- Structure-based virtual screening (SBVS).

Ligand based virtual screening (LBVS) involves a large number of molecules being evaluated based on the similarity of "already known ligands".

Structure-based virtual screening (SBVS), on the other hand, involves a number of molecules being evaluated for their "specific binding to the active sites of target proteins".



Scheme 2.2-2: *Illustration of the Structure and Ligand Based Drug Design Processes based on ref.* [199]

Advances in the structural biology, together with the X-ray crystallography and the nuclear magnetic resonance (NMR) helped protein and ligand structures to be known better.

Due to these advances, **Structure-Based Virtual Screening (SBVS)** has become quite common and preferable among the drug design processes. Following this, especially, **Docking-Based virtual screening (DBVS)** is reported as the most widely applied approach in practice [204].

DOCKING

In a simple definition, docking means identifying the most important binding poses (or modes) of the drug candidate molecules (namely, ligands) on the receptors (mostly the target proteins) [200].

With the known protein structures and a database of known ligand structures (potential candidates), the goal by docking is to have a range of protein-ligand complex conformations so that based on their stabilities and binding energies they can be sorted.

A basic illustration for docking can be given in Figure 2.2-1 as follows:

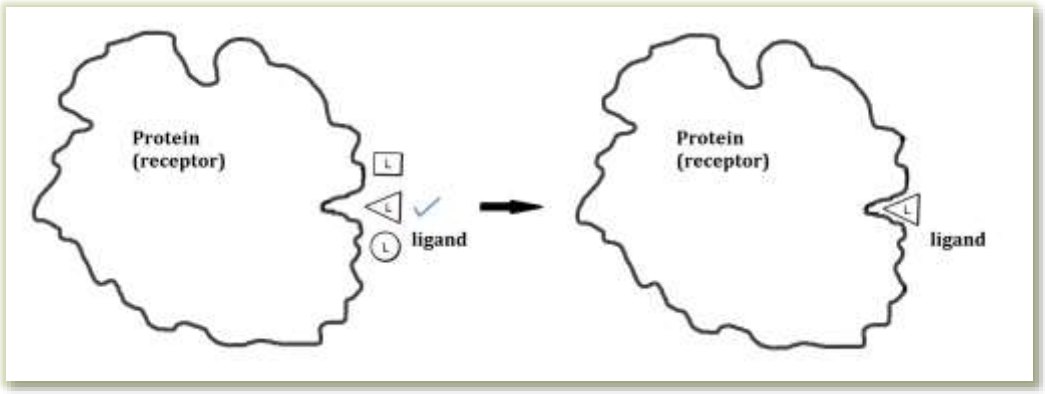


Figure 2.2-1: *Basic Illustration of Docking. Depending on the protein structure, the binding site and the ligand structure, the most suitable ligand candidate (with a predicted pose) is chosen from the database.*

This is usually done by a docking software with the help of accurate structural modelling and a correct prediction of [204, 205]:

- Ligand's conformation, and,
- Ligand's orientation (or posing) within a targeted protein's binding site.

Docking softwares have two main parts: A search algorithm (which generates a large number of poses of a molecule in the binding site), and a scoring function (which calculates a score or binding affinity for a particular pose).

At this point, there is an important concern as Peishoff and his co-workers emphasize [200]: the docking results are mostly judged by the enrichment of the hits while the correct ranking is not being the central focus of the docking procedure anymore.

This is also the reason why, when it is only about "docking", the underlying scoring functions are accepted to have a good success [206, 207]. On the other hand, when it is indeed about scoring, then these same functions are not regarded to have as much as good performance [199, 208].

SCORING/ RESCORING

Scoring is also known or pronounced as "rescoring" sometimes, because during the "docking" stage, scoring algorithms are already being used.

Scoring as a stage, briefly refers to the correct ranking of the docked structures in terms of their overall free energy of binding [200]. In other words, the strength of binding, or namely, the evaluation of the ligand-receptor interactions are highly concerned at this stage. The goal is, to be able to distinguish some of the experimentally observed modes from the others, so that, only the most promising candidates can be subjected to the tests in the in following experiments.

To continue from Figure 2.2-1 above, the next basic illustration is given for the scoring (rescoring) stage in Figure 2.2-2. Here, the already docked ligands are now subjected to the scoring (rescoring), so that the best candidate specimen amongst them can be chosen.

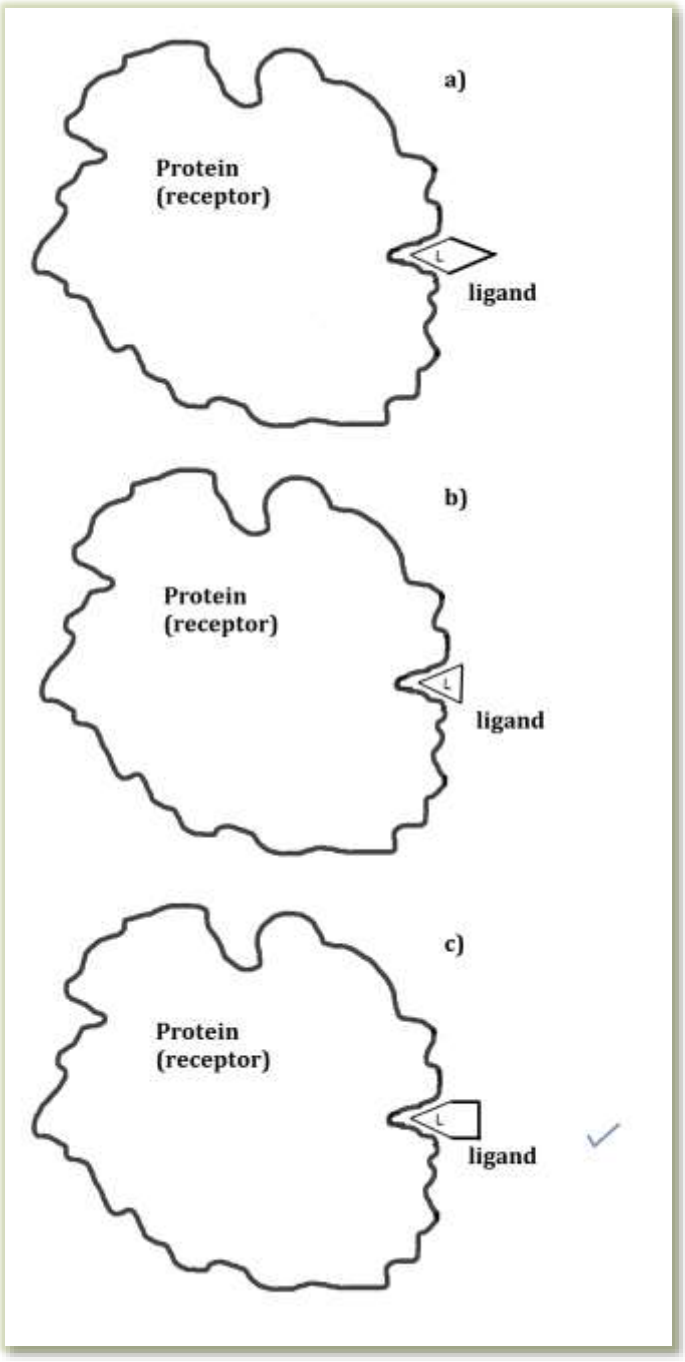


Figure 2.2-2: *Basic Illustration of Scoring. Based on the binding affinities, the most suitable candidate is chosen from the database of already docked ligands.*

To carry out these type of calculations, at first, free energy simulation techniques have been developed for the quantitative modelling of protein-ligand interactions and for predicting the binding affinity. However, initial attempts were not suitable due to their computational costs. Then, in order to improve the results, the scoring functions (which take place both in the docking and scoring stages) were critically under consideration once again [205].

As it is mentioned previously, despite the fact that these scoring functions have an acceptably good success for docking [206, 207], they were not satisfactory enough for the scoring (rescoring) part.

Accurate ranking of binding affinities is still a very difficult task for the commonly used scoring functions [209, 210]. Common scoring functions have a low theoretical level of oversimplifying assumptions for the treatment of the protein-ligand interactions.

For example, ligand binding in nature occurs in a condensed phase with many degrees of freedom [200], which indicates that, these calculations are closely related with a combination of enthalpic and entropic effects. Though, most of the scoring functions are focused on the energetic effects rather than entropic effects [205], and thus, this also complicates the situation.

Binding Energy

In our calculations, the experimental binding energies are taken from equilibrium constants:

$$\Delta G = -RT\ lnK \quad \text{(Eqn. 2.2-1)}$$

Our calculations are based on the free energy of interaction formula, [211]

$$\Delta G = \Delta H - T\Delta S \quad \text{(Eqn. 2.2-2)}$$

ΔG: change in Gibbs free energy
ΔH: change in enthalpy
T: temperature
ΔS: change in entropy

and on Gilson's [212-216] Minima-Mining approach, where the following formulas are considered:

$$H = \sum_i p_i H_i \quad \text{(Eqn. 2.2-3)}$$

$$S = \sum_i p_i S_i - k_B \sum_i p_i \ln p_i \quad \text{(Eqn. 2.2-4)}$$

k_B: Boltzmann's constant

Minima-Mining approach, thus the equation (Eqn. 2.2-4) above suggests that [212]:

The total configurational entropy is the summation of two terms:

- first term, as the weighted average of the entropies of the individual wells, S_i, and
- the second term, which is similar to the entropy of mixing and which is also the entropy associated with the distribution of the system across the energy wells, i.

Then the following can be represented:

$$\Delta G = \sum_i p_i \Delta H - T(\sum_i p_i \Delta S_i - R \sum_i p_i \ln p_i) \quad \text{(Eqn. 2.2-5)}$$

$$\Delta G = \sum_i p_i \Delta H - T \sum_i p_i \Delta S_i + RT \sum_i p_i \ln p_i \quad \text{(Eqn. 2.2-6)}$$

$$\Delta G = \sum_i p_i \Delta G_i + RT \sum_i p_i \ln p_i \quad \text{(Eqn. 2.2-7)}$$

R : gas constant
i : energy well
p_i : probability of finding the system in energy well, i

$$p_i = \frac{e^{-\Delta G_i/RT}}{\sum_i e^{-\Delta G_i/RT}} \quad \text{(Eqn. 2.2-8)}$$

Where, this will be computed in the following with the RRHO approximation.

$$\Delta G = \Delta E_{SQM} + \Delta E_{solvation} + \Delta H_{0-298}^{RRHO} - T\Delta S_{298}^{RRHO} \quad \text{(Eqn. 2.2-9)}$$

Further Explanations

Gilson's Minima Mining Approach *[212 -216]*

As there is a need to improve the methods for ranking the protein-ligand binding energies, there have been approaches/methods developed for this purpose. These can be categorized based on their complexity as follows [213]:

Simple approaches

- **Docking methods/ scoring functions**

These methods try to find the single most stable conformation of a protein-ligand complex. This is usually based on a sum of free energy contributions.

Advantage: They are fast calculations.

Disadvantage: They oversimplify the important free energy contributions (i.e. enthalpy, entropy)

Complex approaches

- **Free Energy Pathway Methods**

Examples can be listed as Monte Carlo or Molecular Dynamics method, which are the free energy perturbation and thermodynamics methods.

These methods usually use explicit solvent models to calculate absolute or relative work of binding energies of ligands.

Advantage: They provide more accurate results.

Disadvantage: They are computationally too demanding to be practical for automation.

In between these two types of approaches/methods, a third type exists, which is called **"end-point" free energy methods**, and **Minima-Mining** is from this category. End point free energy methods:

- Provide better detail than the simple approaches because they account for the bound and unbound states of the protein-ligand complexes.
- They are computationally advantageous when compared to the complex approaches.
    - They use a smaller set of conformations (local energy minima) of the free and bound proteins and ligands.
    - Each local energy minimum linked free energy is calculated based on the molecule's (protein's or ligand's) energy well's depth and width.
    - The contributions of the energy wells are combined and used to approximate the overall free energy.
- With this approach, conformational search or basically the identification of the stable conformations can be done without a need of crossing energy barriers with a thermal motion. The number of explicit degrees of freedom can be further limited by using an implicit solvent model and the number of conformations can be limited by treating parts of the system as restrained or rigid especially in case of large receptors.

Accordingly, these methods are used together with QM energy models.

Thermodynamics of each energy well are obtained with Rigid Rotor Harmonic Oscillator (RRHO) model, hence, the configurational entropy of the local energy well, S_j, is expressed in three terms as: translational, rotational and vibrational contributions.

The following is presented based on Atkin's descriptions [217] for more information on finding the system in an energy well using the probability function.

Quantum mechanics focuses on individual molecules whereas, statistical thermodynamics focuses on the average behaviour of the large numbers of molecules, and relates the microscopic properties of a matter to its bulk properties.

Being a part of statistical thermodynamics, Boltzmann distribution is used for the prediction of a system's populations of states when there is a thermal equilibrium. This concept benefits from the **partition functions** and their relevancy to thermodynamics.

If a closed system is taken as a basis, with N number of molecules and with a constant energy E, then there is still a question of how the energy is distributed amongst the dynamic molecules within this system.

One possibility is to come up with the population of states so that the distributions of the average number of molecules can be described.

The following are assigned:

n_i: average number of molecules (i.e. n_0, n_1, n_2,…)

in,

ε_i: state of energy (i.e. ε_0, ε_1, ε_2, …)

It is mentioned that, with every collision the precise identities of the molecules in each state may change, but the population of the states almost remain constant.

To calculate the populations of states at any temperature, for any type of molecule, in any mode of motion, the only constraint is reported to be the independency of the molecules, so that, when their individual energies add up, it will give the total energy of the system. Also, with this approach, populations of the states will only depend on temperature.

Configurations and weights

At any instant, it is assumed that, n_0 molecules will have the energy state of ε_0, and n_1 will have ε_1, and likewise. The reference state is assumed to be the lowest state, namely the zero of energy ($\varepsilon_0 = 0$). In order to measure all energies, this reference state is the starting point, and all others are calculated relative to this.

The instantaneous configuration of the system is defined as: {n_0, n_1, n_2...}. A general configuration {n_0, n_1, n_2...} can be achieved in "W" different ways, where, W is for the "weight of the configuration". With a general formula, it is given as follows:

$$W = \frac{N!}{n_0!n_1!n_2!\ldots} \qquad \text{(Eqn. 2.2-10)}$$

With a logarithmic expression, this can be rewritten as,

$$\ln W = ln\frac{N!}{n_0!n_1!n_2!\ldots} \qquad \text{(Eqn. 2.2-11)}$$

Which eventually becomes,

$$\ln W = \ln N! - \sum_i \ln n_i\,! \qquad \text{(Eqn. 2.2-12)}$$

with Stirling's approximation to simplify,

$$\ln x! \approx x \ln x - x \qquad \text{(Eqn. 2.2-13)}$$

Overall it results in the approximate expression for the weight:

$$\ln W = (N \ln N - N) - \sum_i (n_i \ln n_i - n_i) = N \ln N - \sum_i n_i \ln n_i \qquad \text{(Eqn. 2.2-14)}$$

Boltzmann Distribution

From the equations above, it can be estimated that the configuration, which weight W gets its maximum, will mostly likely be the dominating configuration and it will be the part where the system will be mostly found in. The maximum value of W can be found via taking its derivative and equating it to zero (namely where dW=0 is satisfied) or via finding the maximum value of "ln W" (in case the logarithmic functions are to be used).

However two restrictions are mentioned:

i. Only some of the configurations are allowed to be included, and these are the ones, where the total energy of the system is constant. Therefore, a configuration has to satisfy the condition of constant total energy:

$$\sum_i n_i \varepsilon_i = E \qquad \text{(Eqn. 2.2-15)}$$

ii. Since N (the total number of molecules) is already defined, populations have to be arranged accordingly in the sense that if there will be some increase, then there has to be a decrease for a compensation. Thus, accordingly, W has to be limited based on the constant total number of the molecules too. The condition to keep is given as follows:

$$\sum_i n_i = N \qquad \text{(Eqn. 2.2-16)}$$

Together with these restrictions, the populations in the greatest weight depend on the energy of the state according to the Boltzmann distribution. This is given as:

$$\frac{n_i}{N} = \frac{e^{-\beta\varepsilon_i}}{\sum_i e^{-\beta\varepsilon_i}} \qquad \text{(Eqn. 2.2-17)}$$

with the following conditions:

$$\varepsilon_0 \leq \varepsilon_1 \leq \varepsilon_2 \qquad \text{(Eqn. 2.2-18)}$$

$$\beta = \frac{1}{kT} \qquad \text{(Eqn. 2.2-19)}$$

k: Boltzmann's constant
T: temperature

Then, the Boltzmann distribution is given as:

$$p_i = \frac{e^{-\beta\varepsilon_i}}{q} \qquad \text{(Eqn. 2.2-20)}$$

p_i: fraction of the molecules in the state i,

$$p_i = \frac{n_i}{N} \qquad \text{(Eqn. 2.2-21)}$$

q: molecular partition function

$$q = \sum_i e^{-\beta\varepsilon_i} \qquad \text{(Eqn. 2.2-22)}$$

Rigid Rotor, Harmonic Oscillator (RRHO) Approximation *[212]*

Rigid rotor, harmonic oscillator approximation treats the molecules as essentially "rigid", so that their internal motions are assumed to include only vibrations of small amplitude. This approach enables uncoupled kinetic and potential energy approximations of translational, rotational and vibrational motions. The partition function can be factorized as follows:

$$Q = Q_t Q_r Q_{vib} \qquad \text{(Eqn. 2.2-23)}$$

Q : full partition function
Q_t : overall translational motion of the molecule
Q_r : overall rotational motion of the molecule
Q_t : overall vibrational motion of the molecule

The overall translation of the molecule has the kinetic energy contribution of overall translation, and it is given in the following expression:

$$Q_t = V\left(\frac{2\pi m}{\beta h^2}\right)^{3/2} \qquad \text{(Eqn. 2.2-24)}$$

m : molecular mass
V : a factor
β : $(k_B T)^{-1}$
h : Planck's constant
k_B : Boltzmann's constant

Since molecules are assumed rigid, then this assumption results in approximating the moments of inertia as constants, therefore, the rotational contribution can be written as follows:

$$Q_r = 8\pi^2\left(\frac{2\pi}{\beta h^2}\right)^{3/2}(I_1 I_2 I_3)^{1/2} \qquad \text{(Eqn. 2.2-25)}$$

I_1, I_2, and I_3: the molecule's three principal moments of inertia.

Then, vibrational contribution can be introduced as,

$$Q_{vib} = e^{-\beta E_o}\prod_i \frac{e^{-\beta h\omega_i/4\pi}}{1-e^{-\beta h\omega_i/2\pi}} \qquad \text{(Eqn. 2.2-26)}$$

ω_i : frequencies of the vibrations

E_o : ground state energy

Overview

As it is mentioned above, scoring functions in principle have two duties [199].

- To differentiate between various poses of a single ligand in the receptor (protein)-binding site. (Docking-pose determination task)

- After docking is finished, to estimate binding affinities of different receptor–ligand complexes within the database and to rank order the compounds. (Scoring task)

There are no QM scoring functions established as a standard in the pharmaceutical industry. Our goal is to develop scoring functions, where basic interactions are more accurately treated. For this, we rely on the PM6-DH+ method. Starting from the Section 3.1.1, our results are presented.

2.3 Overview computational screening of battery electrolyte materials [218]

With the worldwide increasing energy demand, working on the renewable energy sources became crucial. Both being able to harvest energy from an energy source, and moreover, being able to store this harvested energy, are so challenging.

When it is about storage methods, there are different categories, like: biological, mechanical, electrical, chemical, thermal or electrochemical storage methods. These methods are also the main concerns of today's industry too.

Automobile industry can be listed amongst one of the biggest industries which can highly benefit from the developments in this field. In that sense, for the automotive industry, the electrochemical storage methods (i.e. batteries, fuel cells) are more of a concern.

Electrochemical storage, mainly involves the research topics closely related to [219-221]:

- Basic principles of the catalysis at electrochemical interfaces (i.e. structure of electrochemical solid-liquid interfaces, proton-electron transfer reactions at interfaces…),

- Transferring methods from surface science to electrochemistry (for example in order to understand structure-reactivity relationships for nanostructured electrodes, and likewise).

Advanced batteries, currently consist of interconnected electro-chemical cells with lithium-ion based cell chemistries. Lithium-ion ones have higher density compared to other alternatives. They are combining graphite as the anode, with a lithiated transition metal oxide as the cathode [222].

On the first charge, solution species are reduced from a passivating film on the anode. The solid-electrolyte interface (SEI) prevents further irreversible processes, thus, the choice of the electrolyte plays an important role.

Properties of especially high interest for electrolyte solutions include
- electrochemical stability windows,
- melting, boiling and flash points,
- dielectric constants, and,
- viscosity [223].

There are also transportation concerns that affect the design of the batteries [224], like:
- safety,
- cost,
- gravimetric energy density, and,
- thermal stability of the organic electrolyte solvent component.

Cells are produced from high voltage transition metal cathodes or nanocomposite anodes and these type of batteries need electrolytes. There are also “superbatteries”, which are based on Lithium-Sulphur or Lithium-Air design. Amongst the others, these type have a higher performance potential, however they are also limited by [225]:

- Interfacing of electrodes and electrolytes
- Danger of shortcuts through the dendrite formation on Lithium-metal anodes.

Usually being based on the cyclic and linear carbonate like mixture formulations, the commonly used electrolytes are [223]:

- Ethylene carbonate (EC)
- Dimethyl carbonate (DMC),

These are usually linked together with: Lithium hexafluorophosphate (LiPF6) as a salt.

Battery working principle
(Li-ion Battery for demonstration)

Discharging	Charging
Lithium ions between the layers in the negative electrode material (anode) pass through the separator and into positive electrode material (cathode), resulting in a discharging current flow [226].	Lithium ions positive electrode material (cathode) pass through the separator and into the layers in the negative electrode material (anode), resulting in a charging current flow [226].

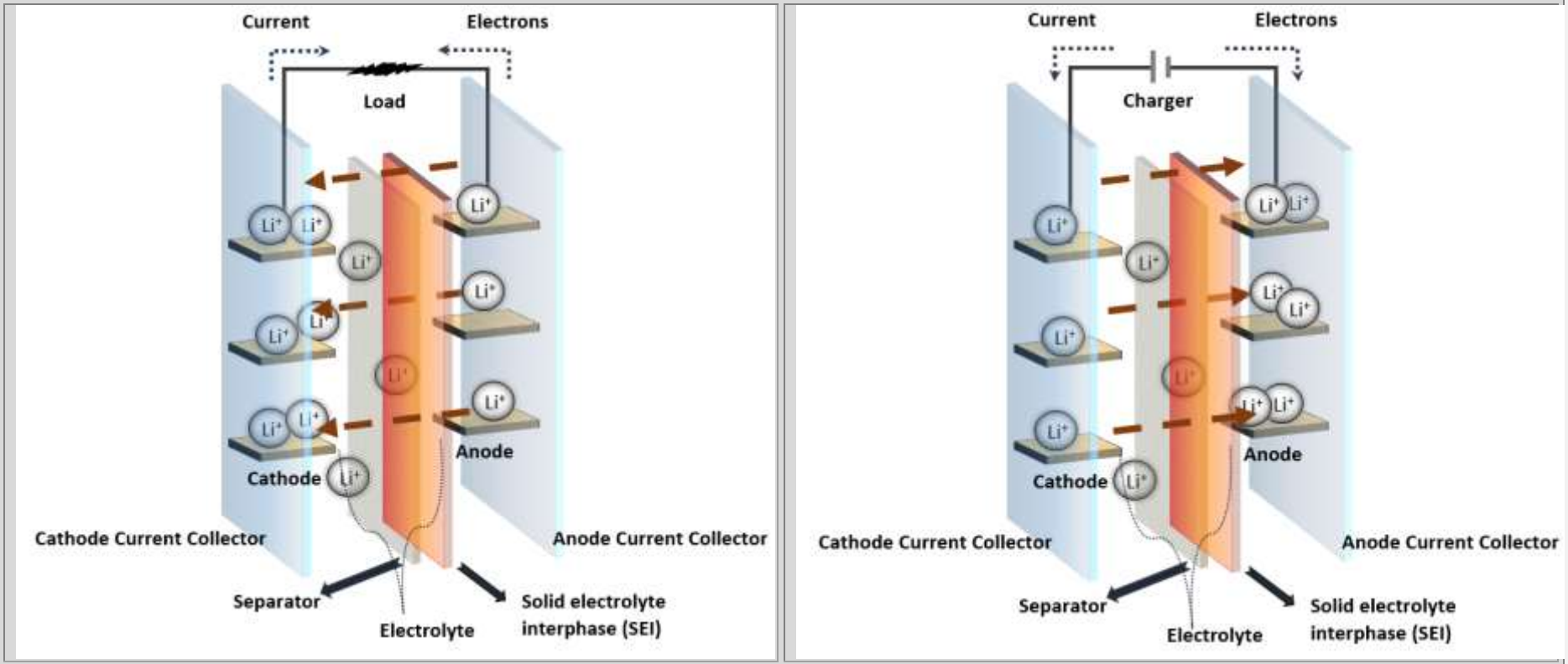


Figure 2.3-1 *Li-ion Battery Working Principle*

Due to their importance for the energy density, cathode materials were the focus of the scientific interest in the past, but then, electrolytes have become important for the studies for further developments [227-229].

There are several contributing studies to this field, and again, computational methods are aiding to the research and development of the technologies.

Table 2.3-1 shows some of the theoretical studies on battery materials.

Selected virtual screening work on battery materials		
Year	**Studied By**	**Content**
2008	Tarascon [230]	Ionic liquids and 'green' electrode materials from biomass
2010	Halls and Tasaki [231]	Exploratory study on screening for electrolytes
2011	Ceder and co-workers [232, 233]	A study to screen for new electrode materials which focuses on solid state physics derived electronic structure methods (i.e., periodic density functional theory calculations). It is suitable for the development of conventional electrode materials, but not efficient with molecular organic materials.
2011	Park and co-workers [234]	Exploratory study on screening for electrolytes

Table 2.3-1: *Theoretical work on battery materials*

Collective' properties (like melting points etc.) can only be treated at comparably low theoretical levels for electrolyte components; for instance with statistical models based on quantum chemistry calculations or completely empirical qualitative structure property relationship (QSPR) approaches.

Application of semiempirical models is a rather straightforward procedure, a very challenging problem arises from the fact that chemical reactivity plays an essential role for the performance of electrolyte solutions.

The accurate modelling of electrochemical systems is difficult due to the complexity of the liquid phase and the existence of varying electrode potentials [235], so that, even by experimental means, atomistic information on the composition of SEI layers is rarely available.

However, the reactive formation of interface structures and the atomic scale processes of SEI formation in current Lithium-ion batteries can be investigated with computational methods. Several studies exist, and they are tabulated in Table 2.3-2.

Selected Studies of Atomistic modelling of SEI formation		
Year	**Studied By**	**Content**
2009	Xing and co-workers [236]	Investigation of the oxidative stability and decomposition of carbonate electrolyte solvents with quantum chemistry methods (at B3LYP/6-311++G(d,p) level) takes place. **Outcome:** aldehydes and oligomers of alkyl carbonates as main products are found in good agreement with experimental results.
2010	Leung and coworkers [237]	Methodologically most advanced theoretical work (as a number of *ab initio* molecular dynamics (AIMD) studies) on SEI formation so far, is performed. It involves the investigation of the initial stages of SEI formation. **Outcome:** The important role of carbon edge terminations for EC decomposition was revealed.
2011	Leung and coworkers [238]	A Hybrid DFT MD study on excess electrons in liquid ethylene carbonate is performed. **Outcome:** the excess electrons were found to be localized on single EC molecules in all cases.
2011	Leung and coworkers [239]	Both experimentally and with AIMD simulations, the use of atomic layer deposition to hinder solvent decomposition is investigated. **Outcome:** Computational predictions were confirmed by experimental results. On bare Lithium metal electrodes, EC decomposes within picoseconds, while with oxide coating electron transfer to EC is slowed down.
2011	Xing and co-workers [240]	The role of anions for the decomposition processes, that is, a reduction of the oxidation stability and a change of the importance of different decomposition paths, is investigated.
2011	Kim and co-workers [241]	The structure and formation of SEI layers are theoretically investigated with classical molecular dynamics (MD) simulations. **Outcome:** Distributions of different SEI components were obtained as a function of the distance from the surface for EC, DMC and mixed electrolytes with anodes of different Lithium surface densities.

		1. LEDC is indeed the main SEI component for electrolytes with EC at low Lithium surface densities (comparably to graphite anodes), but quickly decomposes to inorganic salts for higher Lithium surface densities. 2. A multilayer structure with more inorganic salts is found out to be closer to the anode and more organic based salts are found close to the bulk electrolyte.
2012	Xu and co-workers [242]	A detailed review on experimental and theoretical details concerning SEI chemistries and formation mechanism is presented.
2012	Von Wald Cresce and coworkers [243]	A direct link between the Li+ solvation sheath structure and SEI formation for mixtures of EC and propylene carbonate (PC) electrolytes is found. **Outcome:** A reversed preference of PC over EC by Li+ is found with quantum chemistry methods and molecular dynamics simulations. SEI formation strongly depends not only on the electrolyte composition but also on the anode material.
2012	Owejan and co-workers [244]	In situ neutron reflectometry measurements of the SEI layer in a working lithium half-cell is made. **Outcome:** Thicknesses of less than 10nm and a uniform mixing of SEI components for their cell setup are found. It is also known that not only anodes are covered with SEI layers, but also cathodes.
2012	Takamatsu and co-workers [245]	The solid-liquid interface on the cathode side with in situ Total-Reflection X-ray Absorption Spectroscopy, is investigated. **Outcome:** An initial irreversible reduction of cathode transition metal ions at the electrode/electrolyte interface is found.
2012	Xing and co-workers [246]	Systematic differences between EC and sulfolane-based electrolytes are studied with molecular dynamics (MD).

2012	Bedrov and co-workers [247]	Another MD study is performed. **Outcome:** a rather long life time for singly-reduced EC species is found, allowing these compounds to react with other singly-reduced species. At low concentrations of reduced species, LEDC is indeed (one) major product, while at higher concentrations diradical compounds form and in turn react with other radicals to form oligomer species.
2012	Leung [248]	Organic solvent decomposition on cathode materials are investigated. **Outcome:** Proton transfer is predicted to follow after EC oxidation with both products weakening the cathode ionic bonding network
2012	Ganesh and co-workers [249]	An advanced study of SEI formation for several electrolytes and graphite electrodes is published. **Outcome:** Orientational ordering of the electrolyte molecules near the interface precedes reduction and that the reduced species depend strongly on surface functionalization and the presence of salts.
2013	Nie and co-workers [250]	The surface reactions of electrolytes with graphite anodes are investigated by using transmission electron microscopy with energy dispersive X-ray spectroscopy (and complementary methods). **Outcome:** For electrolytes containing EC, SEI layers of about 50nm thickness consisting mainly of dilithium ethylene dicarbonate (LEDC) and LiF, with a strong dependence of the thickness and composition on the initial composition of the electrolyte solution, is found.
2013	Borodin and coworkers [251]	In comparison to G4MP2 reference data, quantum chemistry methods (the M05-2X, LC- ωPBE density functionals with implicit solvation) are used, where it is possible to investigate the oxidative stability and decomposition reactions of carbonate, but also sulfone, sulfonate and alkyl phosphate based electrolyte solvents and a number of different lithium salts. **Outcome:** Spontaneous hydrogen-transfer to salt anions decreases the solvent stability and that the presence of anions but also other solvent molecules can significantly reduce barriers for oxidation-induced decomposition reactions and therefore product composition.
2013	Leung [252]	Publication of a review of AIMD studies on SEI formation.

Table 2.3-2: *Selected Literature for Atomistic modelling of SEI formation*

Chapter 3

Our Projects

Our main goal was to develop large scale screening strategies for molecular compounds and we needed methods to describe all relevant properties of these compounds. In that regard, this thesis includes two different projects.

- One of them is our main research topic and corresponds to the “Computational Screening of Biomaterials” section of the thesis title. This part is linked with the literature Section “2.2: Overview of Computer Aided Drug Design“, and details of the project will be further explained in Section 3.1.

- The second, minor part of my work is about the introductory research for “Computational Screening of Energy Materials” which is related with the literature section “2.3: Overview computational screening of battery electrolyte materials”. It will be introduced briefly in Section 3.2 up until the point at which I was not involved anymore. Then the rest of this work was carried out by our workgroup colleagues.

3.1 Computational Screening of Bio Materials

The first project is given in three main "Research Stage" steps, which can be abstracted as follows:

- In Research Stage I (Section 3.1.1), we compare the performance of various quantum chemical approaches for tackling this previously mentioned "scoring" problem, since the correct ranking protein–ligand interactions with respect to overall free energy of binding is a grand challenge for virtual drug design.
  Generating systematic benchmark sets of PDBbind based large protein/ligand model complexes by using our cutting algorithm, we show that the performance first of all depends on the general level of theory.
  Then, comparing classical molecular mechanics (MM), semiempirical quantum mechanical (SQM), and density functional theory (DFT) based methods, we find that enhanced SQM approaches perform very similar to DFT methods and substantially different from MM potentials.

- In Research Stage II (Section 3.1.2), we benchmark several wave function theory (WFT), density functional theory (DFT) and semiempirical quantum mechanical (SQM) approaches against high-level theoretical references for realistic test cases. Based on our findings, we can recommend SCS-MP2 and B2-PLYP-D3 as reference methods for WFT, moreover, TPSS-D3+Dabc/def-TZVPP as the best DFT approach, and finally, PM6-DH+ as a fast and accurate alternative SQM method to full ab initio treatments, especially for systematically generated model systems of real protein/ligand complexes from the PDBbind database.

- Research Stage III (Section 3.1.3) is about the further investigation of the problem of scoring protein/ligand interactions, by analysing the delicate balance of biomolecular interactions with quantum mechanical (QM), semi-empirical QM (SQM) and molecular mechanics (MM) methods.
  The biggest differences between QM and SQM or MM methods are found for the treatment of enthalpic/entropic effects.
  It is also observed that, trying to improve the description of only one sub-balance, for example with a QM-level description of energetic protein/ligand interaction, is not likely to improve the accuracy of scoring functions.

3.1.1 Research Stage I [89]

I.1 Introduction

Accurate and fast modelling of protein-ligand systems is essential for various scientific fields. Biomolecular interactions are highly important for *in silico* drug design [253].

For a virtual drug design, the preferred approach is the pre-selection of most promising candidates, so that the following main computational procedures can be applied more efficiently. We hereby refer to these main procedures as docking and as scoring. Often the docking processes are assumed to have an acceptable accuracy to identify the most important binding poses [206, 207]. On the other hand, the scoring stage is still regarded as a challenge due to the low accuracy of the binding affinity values [209, 210]. In scoring studies, the ligands and proteins are usually being treated at a low theoretical level. Therefore, a treatment with higher level methods creates a potential for an improvement. There are some systematic works on high-level methods [88, 254, 255], and some among them involve small host-guest systems [120, 129]. Still, the number of relevant studies in this field is insufficient. Improvement possibilities for the scoring functions are seen in a better treatment of polarization, solvation and entropic effects [206, 207], where entropic effects are the major complications since they are linked with the dynamical complexity of the problem.

For the improvement of polarization and solvation effects, SQM methods are very promising candidates to focus on. In addition to the Stewart's linear scaling approaches [38], there are many other studies on SQM method developments as well [88, 256, 257].

Korth contributed to the development of empirical corrections for non-covalent interactions [56]. For protein-ligand systems, these non-covalent interactions are significant, but often they are not treated well enough at semi-empirical levels. Korth workgroup additionally studied SQM methods augmented with empirical DH corrections, and have found out that the accuracy of SQM-DH methods can reach to that of DFT-D methods for a large number of cases, while SQM methods are about three orders of magnitude faster [50, 53].

In this work, we wanted to assess the performances of different computational methods (MM, SQM and DFT) for the scoring of protein/ligand interactions. This aim was motivated by

- method performance studies [42, 258]
- studies supporting the possibility of the scoring methods development via experimental data sets [259], and,
- the availability of the Protein-Ligand Complex Structure Data sets (in our case PDBbind2007 as our initial choice [260, 261]),

Once it is about a theoretical study, it is important to acknowledge that the molecular models are vital for the quality of the simulations. We are aware that even the smallest

protein in the Protein-Ligand Data sets is much larger than the largest ligand. Given that we have the 'docked' poses, in other words, the 'fixed' protein and ligand couple structures in the PDBbind sets, the interaction zone between the ligand and the protein is also fixed and it is confined to a certain region.

The following sections will explain our model preparations and hence how our calculations are performed. In order to check our approach and its validity, the prepared models will be subjected to several comparison tests as well.

It is highly important here to note that a fixed protein-ligand complex is not a perfect model, since in reality they are flexible and in motion. We will later on (Section 3.1.3), take this flexibility into account by averaging overall relevant binding modes.

I.2 Choice and Generation of Model System Benchmark Sets

The PDBbind Database Sets

As a first step, PDBbind database is used which is based on (Protein Data Bank) PDB database set.

This set contain a collection of experimentally measured binding affinities and detailed structures for biomolecular complexes [90], therefore they provide a good opportunity to compare computational results with experimental values. A compilation of the specifications of these PDBbind sets can be found in the main references [209, 260-263]. Based on the information contained therein, there are various PDBbind set categories, named as: "general set", "refined set" and "core set".

To obtain these sets, experimentally determined PDB structures are screened at the PDB database and after some validity checks, structures are admitted to the sets as described below [209, 260-263]:

Once the primary reference of each complex was examined to collect experimentally determined binding affinity data (K_d, K_i, or IC_{50}) of the given complex, the set is called as general set having a large number of entries. Afterwards, this general set is going through a filtering step regarding binding data, crystal structures, as well as the nature of the complexes, and thus the refined set is obtained. Finally, a high-quality version of these sets with a smaller number of complexes, the core set is meant to serve as a high-quality benchmark for evaluating scoring methods. Core set is obtained by clustering the refined set by protein sequence similarities with a cutoff of 90% and only selecting sets with at least five members. From each cluster, the one with the highest binding constant, the one with the lowest binding constant, and the one with a medium binding constant are selected as the representatives of this cluster.

A reasonably large set was needed for our study. Based on the requirements, the refined set of the PDBbind2007 database set was chosen. The set has around 1300 protein-ligand complexes [260, 261].

Later on newer PDBbind versions are released, however, in the beginning of our study, PDBbind 2007 was the only set with separate protein and ligand structures available at that time, and therefore it became our choice.

It also has to be noted that different sets can sometimes have different experimental conditions or properties, therefore, we also tested the other (more recent) PDBbind sets when they became available too, but their quality was rather similar. Due to this similarity, in order to stay consistent, this study will mainly focus on our initial set selection: PDBbind 2007 set.

Generation of Benchmark Model Systems

The size of the large complexes (complex: protein + ligand) within the PDBbind set ranges from 800 to 90000 atoms (in average, roughly 7000 atoms). To speed up the computations, the initial task of this study was to prepare reasonably good model systems of smaller sizes, which resemble the original structures sufficiently.

This had to be done only for the proteins, because computational cost depends on the number of atoms in a molecule and ligands are several orders of magnitude smaller than proteins (the relevant numerical data about sizes is given in Table I.2-1 below). Therefore, a protein cutting (down-sizing) algorithm is written. With this cutting algorithm, main goal was to create smaller models while changing the original protein structures as little as possible, so that a large enough sub-set of protein atoms from the original structure could be kept to reflect the original structure's characteristics. After obtaining the model cuts, also to find out an optimum model size, a comparison test amongst these model cuts was needed in order to select the optimum "cut-off" distance.

The proposed cutting algorithm first looks at the defined/assigned cut-off distance, then detects the individual atom-centres inside this cut-off distance, and then finds out to which residue these atoms belong. No matter how little a given residue intersects, the function will always include the full residue into the smaller model structure that is to be kept– that we name as "pocket" (see Figure I.2-1). This condition of "keeping the whole residue in the pocket", holds true and applied for, even when only a single atom of a residue is inside the cutoff line.

The process of keeping the residues can be explained similar to an illustration of taking out some certain branches from a tree (Figure I.2-2). In case even only a little part of the branch is in the interested zone, then the complete branch segment is selected.

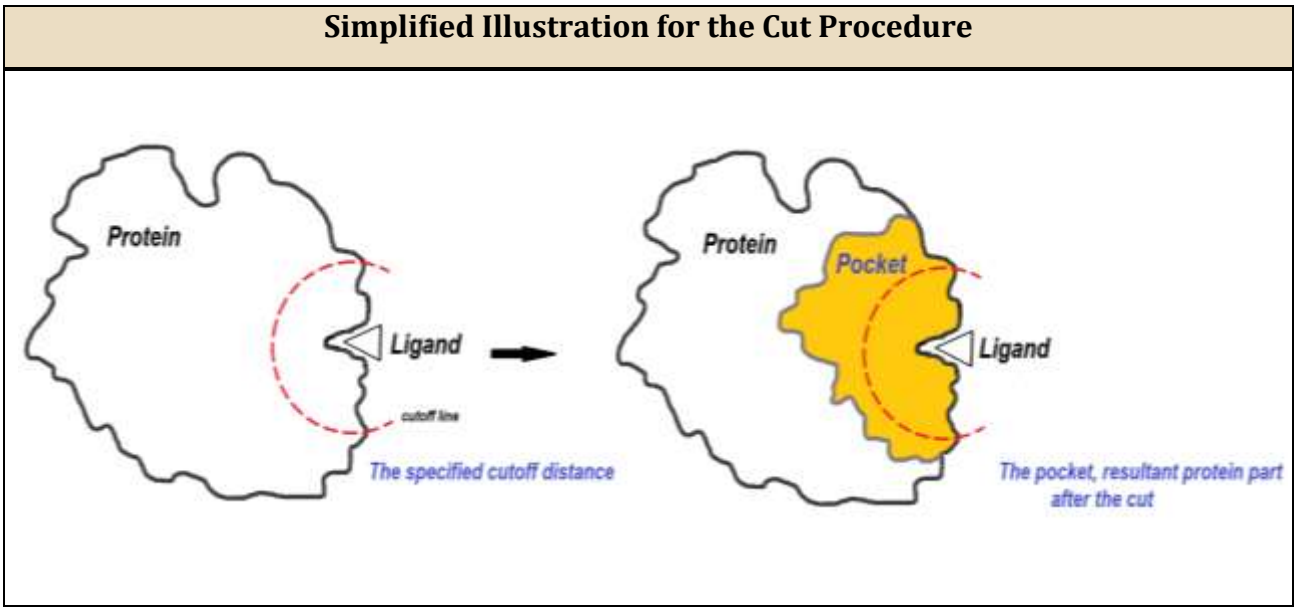


Figure I.2-1: *Basic Illustration for the Cutting Algorithm. Shown is a schematic representation of a protein "P" and a ligand "L" bound to it. The cutting algorithm selects all the relevant residues and pocket model (yellow part) is obtained.*

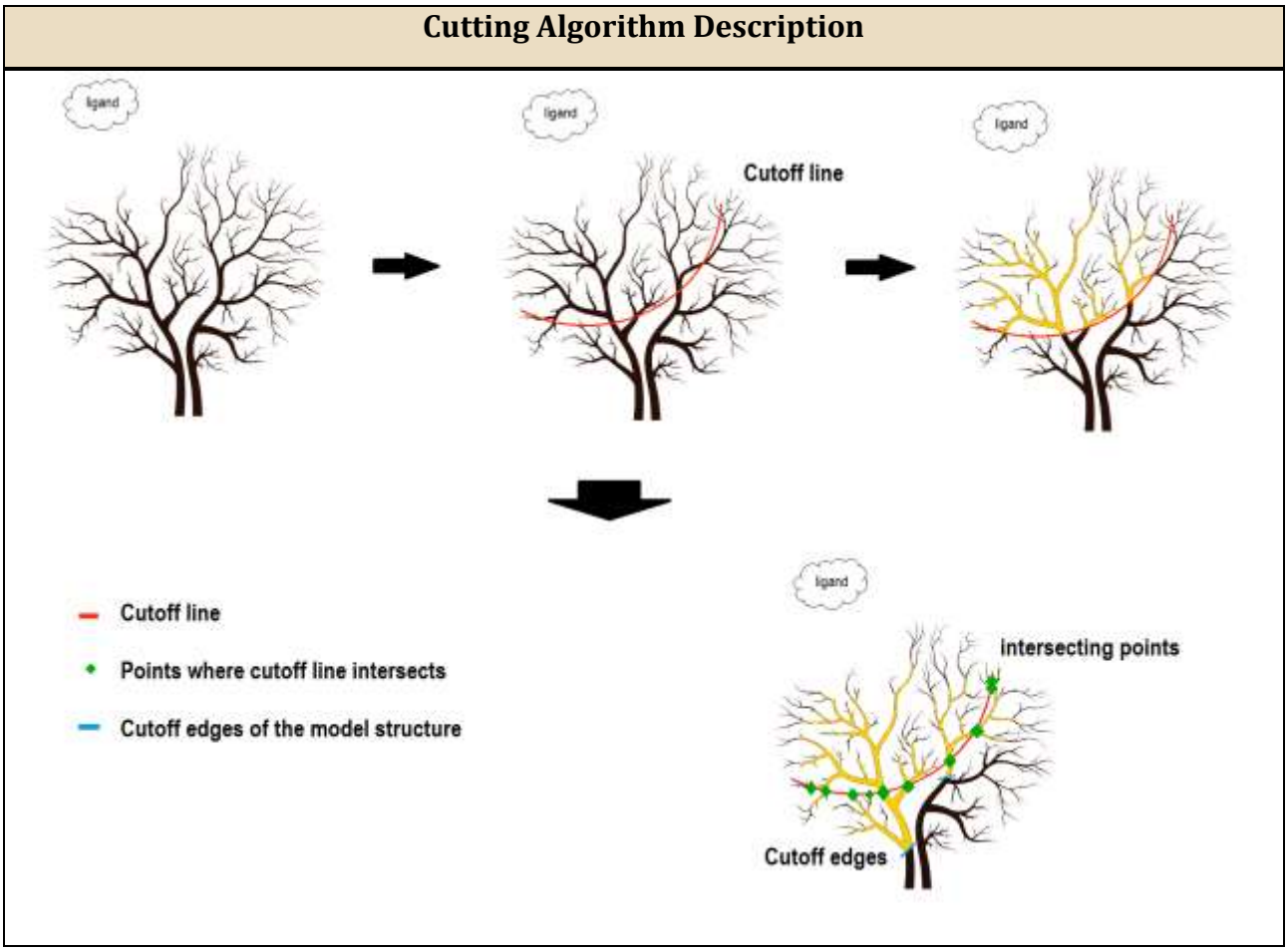


Figure I.2-2: *Descriptive Illustration for the Cutting Algorithm. Green indication points are the points where the cutoff distance intersects with the branched structures. This is used to define the branches to be tracked along, selected and kept afterwards (shown in yellow). This yellow part corresponds to the pocket model. Blue line indicates the terminal points of the cut structures. These are chosen as the central stops to be able to keep the structures in a uniform manner. These are namely the end points of the pockets.*

As an additional demonstration, the following Figure I.2-3 makes use of structures directly from PDB files.

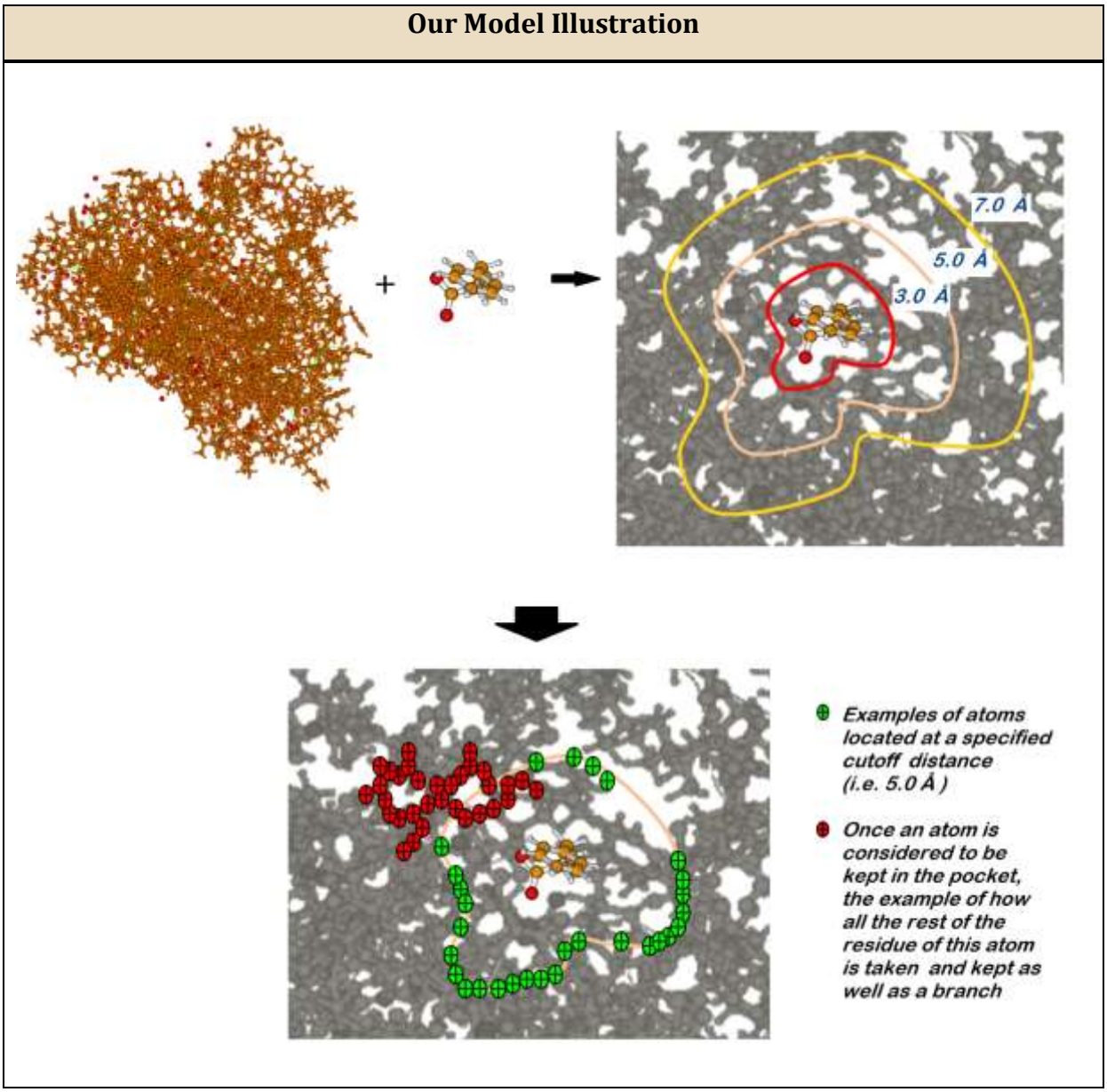


Figure I.2-3: *Model Illustration for the Cutting Algorithm. Green points are drawn to illustrate the atoms which are located at an example cutoff distance. Red points indicate how a residue can be tracked starting from one of the intersecting (one of the green) atoms. By this way, when all greens are found out and following that when all red residue ones are identified, then pocket is formed.*

With the help of our algorithm, protein cuts were made around the docked ligands with the following distances: 3.0, 5.0, 7.0, 10.0 and 20.0 Å. While investigating the structures, there were some problems encountered in the original PDBbind files, mostly about hemiacetals, phosphors, contact points, bonding and charges.

Therefore, together with these concerns, structures for the binding-affinity calculations were prepared in the following way:

1. Smaller model protein structures (the "pockets") were created by the cutting algorithm.
2. A number of structures, whose errors were inside the pocket that could not be fixed easily, were excluded from our calculations to keep our main data set close to the original structures.
3. The resultant pockets were capped with hydrogen atoms at the cut edges.
4. Due to keeping the whole residue even for one single atom intersecting the cutoff line, the overall sizes of the pockets were always systematically larger (by about 3-5 Å) than the specified cut-off distances. For example, a cutoff distance of 3 Å would yield a final pocket size of around 6-8 Å.
5. In case of HIS amino-acid residues, protonation was necessary [264] where they are assumed to be neutral.
6. Water molecules within the pockets were discarded [265].
7. Overall charges were assigned according to automatic Lewis-structure analysis and double-checked with the assignment by MOPAC [58].

Table I.2-1 below shows the minimum and maximum sizes of the model complexes after the cutting procedure. The sizes of pockets ranged from minimum 70 to maximum 700 atoms for 3.0 Å, whereas it reached up to about minimum 900 to 7000 atoms for 20.0 Å.

Cutoff Distances (Ångström, Å)	Size of the pockets (Number of Atoms) min – max	Size of the model complexes (Number of Atoms) min – max	Size of the ligands (Number of Atoms) min – max	Size of the original (full) protein (Number of Atoms) min – max
3.0	70 – 686	84 – 821	6 – 148	884 – 92514 (ave=7286) Diameter ranging from 35 Å to 1774 Å
5.0	148 – 940	171 – 1081		
7.0	271 – 1722	309 – 1857		
10.0	456 – 2624	489 – 2759		
20.0	906- 6989	1030-7046		

Table I.2-1: *Atomic sizes of model complexes (ligand + pocket), ligands and pockets*

After obtaining the pocket models, the goal is to calculate the statistical correlation of the binding energies with respect to different cutoff distances to find an optimal pocket model compromising the computational cost and accuracy.

The binding (interaction) energy can be calculated as follows:

$$E_{interaction} = E_{complex} - (E_{pocket} + E_{ligand}) \qquad \text{(Eqn. 3.1.1-1)}$$

For the Research Stage I, the binding energies are computed by MM (FF), SQM and DFT methods. The limitation of finite computation time results in some of the computational methods only being available for a small number of atoms.

As can be seen in Table I.2-1, the 10.0 Å and 20.0 Å pocket model benchmark set consist of large number of atoms per protein, that's why for both of these, SQM is the only easily available method due its computational costs. Similarly, when all methods are concerned and need to be compared with eachother, then, only small benchmark sets, mainly the 3.0 Å and for some cases the 5.0 Å cutoff pocket models are available for all type of methods.

Our research involve:

- Performance comparisons of the pocket models with various cutoff pocket models,
- Performance comparisons in between different computational methods, based on a certain cutoff pocket model.

The following statistical tools are used for these comparisons:

Pearson Product-Moment Correlation Coefficient:

The Pearson product-moment correlation coefficient, R is commonly used in statistics to measure the degree of dependence between two linearly related variables. The Pearson R values can change in between +1 and −1 inclusive, where 1 indicates a total positive correlation, 0 indicates no correlation, and −1 indicates total negative correlation. In our case, these linearly related variables will be pocket models with different cutoff distances [266].

If there are two datasets, where one of them as an X set with the following variables: $\{x_1,...x_n\}$, and another dataset Y, having the following variables, $\{y_1,...y_n\}$, and that both of them are containing n total values,

Then, Pearson R is given as:

$$R = R_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \quad \text{(Eqn. 3.1.1-2)}$$

Kendall Rank Correlation Coefficient

As a second tool to consider, the Kendall tau rank correlation coefficient, τ is a non-parametric test that measures the strength of dependence between two variables. Here again the range is in between +1 and −1 inclusive, +1 indicates that there is a perfect agreement between the two sets of ranks, whereas −1 indicates that there is a complete disagreement between the two sets of ranks (as the rank of one variable increases the other one decreases [267].

Once again continuing with the same notation, having datasets of: $\{x_1,...x_n\}$ and another as: $\{y_1,...y_n\}$ with containing n total values, then, the following are expressed:

- (x_i, y_i) and (x_j, y_j) are called "concordant", in case:

$$\frac{x_i - x_j}{y_i - y_j} > 0 \quad \text{(Eqn. 3.1.1-3)}$$

  So, this corresponds to: both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$.

Likewise,

- (x_i, y_i) and (x_j, y_j) are called "discordant", if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$.

$$\frac{x_i - x_j}{y_i - y_j} < 0 \quad \text{(Eqn. 3.1.1-4)}$$

  Which comes with the condition as well that when, $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant.

The Kendall τ coefficient is then defined as [268, 269]:

$$\tau = \frac{(number\ of\ concordant\ pairs) - (number\ of\ discordant\ pairs)}{Number\ of\ pairs} \quad \text{(Eqn. 3.1.1-5)}$$

- τ =1: when all the pairs are concordant, and it means variables are in exactly the same order.
- τ =-1: when all the pairs are discordant, which means the variables are in exactly the opposite order.
- τ =0: equal numbers of concordant and discordant pairs, and this is labelled as "there's no relationship between the variables."

Since we are interested in ranking, it is advantageous to examine both Pearson's R and Kendall's τ parameter. The interpretation of the R and τ value depends on the field of studies. In our case, correlations are used as follows:

R or τ values	Meanings
+0.90 to +1.00	Very strong positive relationship
+0.70 to +0.90	Strong positive relationship
+0.50 to +0.70	Moderate positive relationship
+0.00 to +0.50	No, negligible or weak positive relationship
-0.00 to -0.50	No, negligible or weak negative relationship
-0.50 to -0.70	Moderate negative relationship
-0.70 to -0.90	Strong negative relationship
-0.90 to -1.00	Very strong negative relationship

List I.2-1 *Pearson R and Kendall τ value interpretations*

Computational methods that are used for this section are tabulated below in Table I.2-2:

Computational Methods		
Method Type	**Name**	**Computational Details**
MM-FF	MMFF94[270]	OpenBabel [271]
	Amber ff99sb [272] and GAFF [273]	Amber11 [274]
DFT	BP86 [275,276] PBE [277] TPSS [278] with empirical dispersion corrections of D2 [279] D3 [52] and D3 plus three-body-dispersion (named D33 in the following) Turbomole [280, 281]	Calculations are done with Turbomole 6.4 [280, 281] using TZVP and TZVPP Gaussian AO basis sets [282] the RI approximation for two-electron integrals [283, 284], and with COSMO as well as COSMO-RS via COSMOtherm [285]
Semi empirical	AM1[33] AM1-D\cite [53,54] PM6 [38] PM6-D [53] PM6-DH2 [53] PM6-DH2X [65] versus PM6-DH+ [50]	MOPAC2012 [58] with MOZYME linear scaling algorithm [286] and COSMO solvation models [285].

Table I.2-2: *List of computational methods used for Research Stage I*

The reasoning behind some of these selections can be briefly mentioned:

- MMFF94 (implemented in Open Babel) is chosen as an example Force Field (FF) method also because of its parameterizations being available. Also, Amber ff99sb/GAFF approach (with Amber 11) was also added for further comparisons amongst FF methods, and Amber tools were able to do automatized input preparations for 352 entries.
- AM1, AM1-D, PM6, PM6-D, PM6-DH2, PM6-DH2X, and PM6-DH+ methods were included as semi-empirical methods. The most interesting method apart from these are OMx methods among other types of SQM methods. Unfortunately OMx methods are not parametrized for S and P elements, and since we have these elements in proteins structures OMx was not an available option to be included in our tests.
- Chosen DFT methods were with BP86, PBE, and TPSS functionals. We wanted to test and compare the effect of GGA and meta-GGA as well. Therefore, BP86 and TPSS were used for this comparison, but we found that their performances were similar. The reason why BP86 was included, was to test COSMO-RS which is fitted for BP86. The dispersion corrections were additionally involved because DFT does not include these interactions on its own.
- Mopac2012 was the software selected for the semi-empirical calculations because SQM-DH type of methods are available in this software.
- MOZYME was developed to enable very large organic compounds to be easily calculated. This "MOZYME keyword replaces the standard SCF procedure with a localized molecular orbital (LMO) method", therefore with this linear-scaling algorithm, a shorter computation time is aimed.
- COSMO and COSMO-RS were tested because in order to compare their performances in terms of ranking.
- Turbomole 6.4 was the convenient choice because it is often reported to be faster in these type of benchmarking systems [287].
- We haven't studied VASP ONETEP since using periodic codes for biomolecular systems is still in development [288].

At SQM and MM (MMFF94) level, calculations on all generated model systems were possible. We have arrived at the following number of binding energy results at the respective cutoff distances:

Cutoff Distances (in Ångström, Å)	**Number of Binding Energies (SQM and MM : MMFF94)**
3.0	736
5.0	733
7.0	725
10.0	714
20.0	623

Table I.2-3: *Number of Binding Energy Data Points calculated via SQM and MM (MMFF94) methods, with indicated cutoff distances.*

At DFT level, not all generated model systems could be successfully treated due to the software, RAM limitations and SCF convergence problems.

Since we are comparably limited with the computational powers for DFT calculations, the following points were able to be obtained:

Cutoff Distances (in Ångström, Å)	**Number of Binding Energy Data Points (DFT)**
3.0	695 BP86/TZVP
	487 BP86/TZVPP
5.0	539 BP86/TZVP
	539 PBE/TZVP
	513 TPSS/TZVP

Table I.2-4: *Number of Binding Energy Data Points calculated via DFT methods, with indicated (3.0, 5.0 Å) Cutoff Distances.*

These data points demonstrate the number of pockets and ligands that we can calculate binding energies from. After this data points are obtained for different cutoff distances, the correlation in between their binding energy results are given in the following sections.

While presenting the correlation results, in case there were some systematic errors that are encountered, then there was a procedure that we have applied to the data plots.

When there is a systematic error, it means there is an introduction of an error/inaccuracy within the system so that it shifts or scales all the data values in a consistent way. As a result, it is possible to have a systematic change/correction on these data values, and this can be done as in the following:

First, assuming that we obtain a plot of binding energies out of many complexes like in example Figure I.2-4, where each circle represents a complex's binding energy and axes represent different computational methods to compare with eachother.

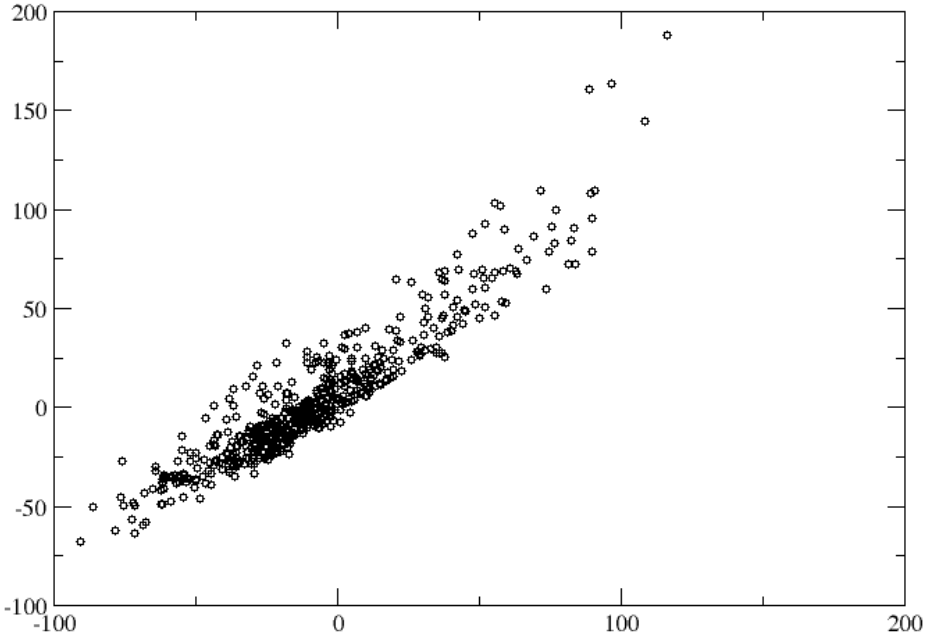


Figure I.2-4: *Scaling and Shifting Stage-I*

To be able to know the overall ranking tendency, we first would like to see the linear regression, namely, the trendline for this data collection as shown in Figure I.2-5:

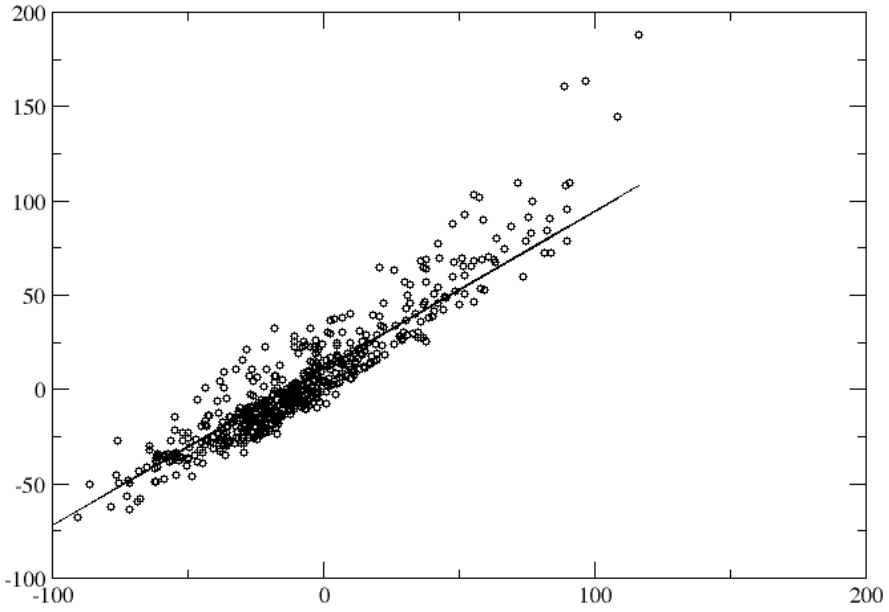


Figure I.2-5: *Scaling and Shifting Stage-II*

Trendline is a visual tool to have a graphical representation of the outputs, and in this study they are obtained via xmgrace program in Linux [289].

For a trendline (linear regression), the equation is:

$$Y = a + bX, \qquad \text{(Eqn. 3.1.1-6)}$$

X: is the explanatory variable,
Y: is the dependent variable,
b: the slope of the line,
a: is the intercept (the value of y when x = 0)

Then, the systematic error correction is done by:

- Scaling: Dividing the trendline equation by the numerical value of the trendline slope (the numerical value of b, as in Eqn. 3.1.1-6), and then,
- Shifting: Subtracting the numerical value of the intercept (the numerical value of a, as in Eqn. 3.1.1-6) from this.

What we obtain afterwards, can be illustrated in the following Figure I.2-6 below. Here, the new, corrected version of the data points and the new trendline are represented. This is basically the red coloured version of the previous data set and it is labelled as the "Scaled and Shifted" version.

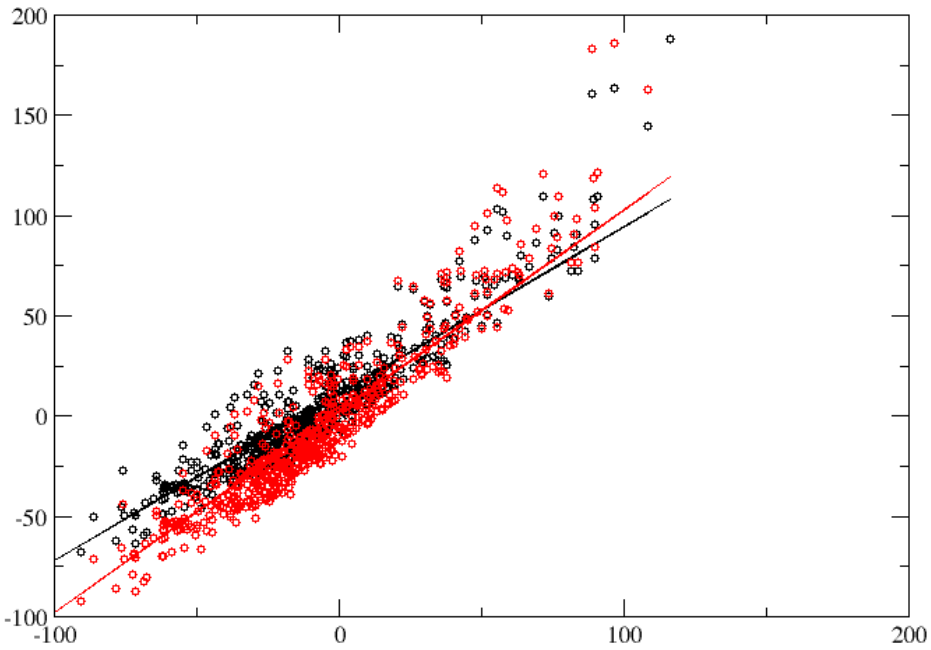


Figure I.2-6: *Scaling and Shifting Stage-III*

Therefore, from this stage on, when we mention about "scaling and shifting", this is the procedure what we are basically referring to and likewise applying to the plots to have better comparisons.

There are also some error criteria definitions to be introduced here, and they will be used quite frequently for the comparisons. Two among them require special attention: the "mean deviation; MD" and "mean absolute deviation; MAD". In general, MD is obtained to see the systematic errors, while MAD is obtained to have an idea on the average errors. The equations in their simplest form can be given as follows:

Mean Deviation is:

$$MD = \frac{\sum(x - \mu)}{N} \qquad \text{(Eqn. 3.1.1-7)}$$

Mean Absolute Deviation is:

$$MAD = \frac{\sum|(x - \mu)|}{N} \qquad \text{(Eqn. 3.1.1-8)}$$

Where, μ is the mean value, x is each value, and N is the total number of values.

Related to the Scaling and Shifting procedure that is just described, there are also two additional MD and MAD values in order to present new scaled and shifted versions of them. They are labelled as MD* and MAD* (denoted with a star), and mostly presented in addition to the MD and MAD values:

MD* : "mean deviation value- after scaling and shifting"

MAD* : "mean absolute deviation value- after scaling and shifting".

I.3 Results and Discussion

The binding energy results of the benchmark sets obtained by means of the MM (FF), SQM and DFT methods (depending on the computational availabilities with cutoff distances: 3.0, 5.0, 7.0, 10.0 and 20.0 Å), are presented in this section. The Statistical parameters- correlation values- are also reported for the evaluation of these findings. The error values which will be given here will be including: MD and MAD values and MD* and MAD* values respectively where they are needed. Further details will be given below.

Comparison of Pocket Model Sizes Based on Cutoff Distances

The aim is to be able to have a reasonable cutoff distance for the pocket model, so that this model can be chosen as a basis.

Focusing only on one method for now, on a method which can computationally handle all the pocket sizes, semiempirical PM6-DH+ was our first chosen method as a reference.

Figure I.3-1 compares PM6-DH+ binding energies for the benchmark sets with increasing distances of 3.0, 5.0, 7.0, 10.0 and 20.0 Å cutoffs.

20.0 Å can be regarded as sufficiently large for our testing purposes because of the typical ranges of the intermolecular interactions involved. It should nevertheless be kept in mind that for the real protein structures the diameters ranges from minimum 35 Å to maximum 1774 Å (average 102 Å).

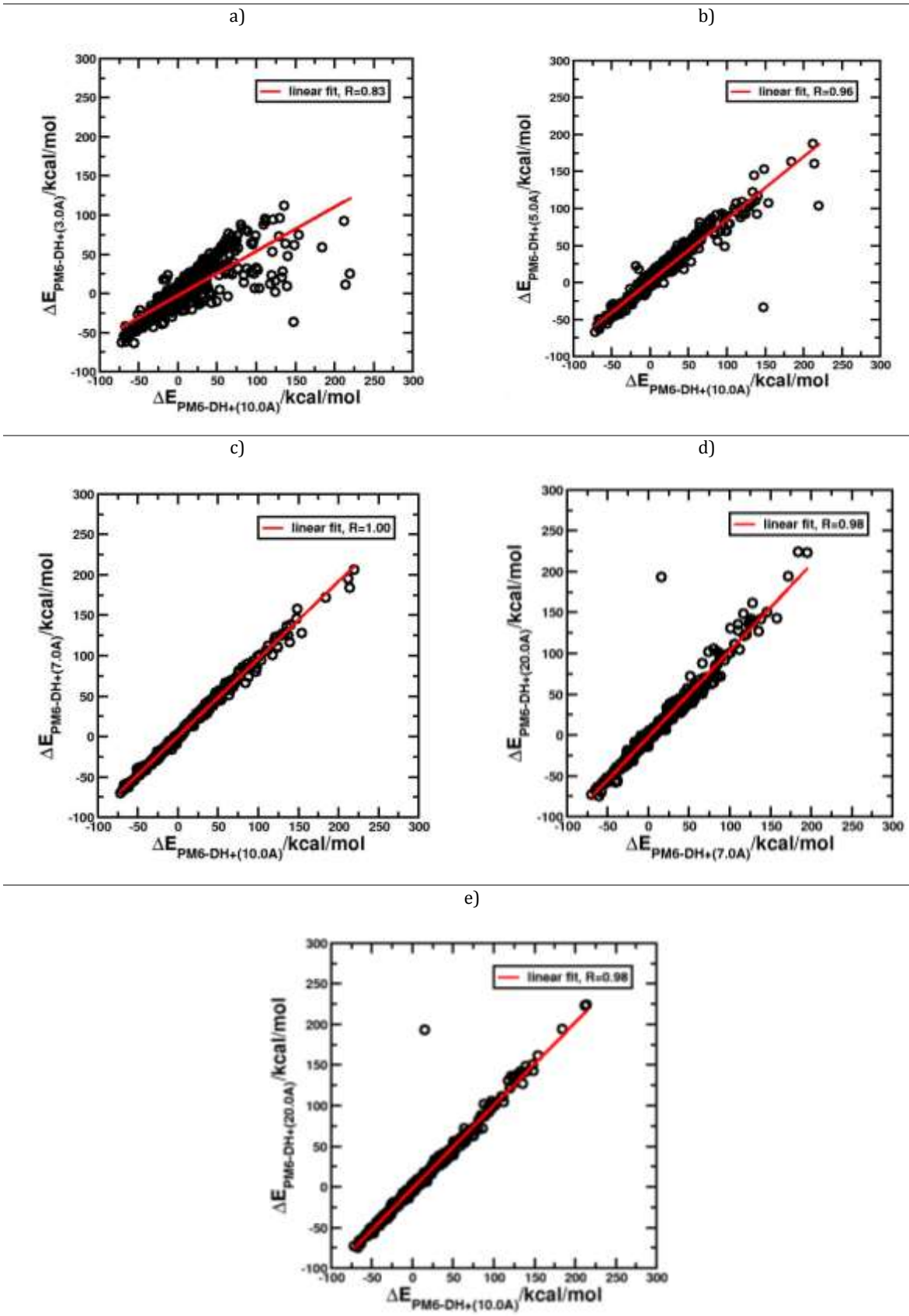


Figure I.3-1: *Correlation between PM6-DH+ data for benchmark sets generated with 3.0, 5.0, 7.0, 10.0 Å cutoff distances: a) 3.0 vs. 10.0 Å, b) 5.0 vs 10.0 Å, c) 7.0 vs 10.0 Å cutoff distances d) 7.0 vs 20.0 Å e) 10.0 vs 20.0 Å. All computations with COSMO solvation model.*

As it can be seen from the plots, we observe a quick convergence as the cutoff distance increases, and this means, as cutoff distance is increased, pockets behave similarly after a certain point. This can be reported respectively follows:

- As for the cutoff 3.0 Å, there is a correlation of Pearson, R=0.83 with the 10.0 Å pocket,
- For 5.0 Å it is R=0.96 with the 10.0 Å pocket,
- For 7.0 Å it is R=1.00 with the 10.0 Å pocket,
- In between 7.0 Å and 20.0 Å, it is R=0.98,
- Finally in between 10.0 Å and 20.0 Å, this value is R=0.98.

Within the plots, Pearson R is shown, however, Kendall (τ) values also show a similar increasing trend with the same models. The statistical parameters for this section can be tabulated as follows:

Entry (all with COSMO solvation models)	Figure I.3-1	Pearson (R) correlation	Kendall (τ) correlation
PM6-DH+ 3.0 vs 10.0 Å	a	0.83	0.76
PM6-DH+ 5.0 vs 10.0 Å	b	0.96	0.90
PM6-DH+ 7.0 vs 10.0 Å	c	1.00	0.96
PM6-DH+ 7.0 vs 20.0 Å	d	0.98	0.93
PM6-DH+ 10.0 vs 20.0 Å	e	0.98	0.95

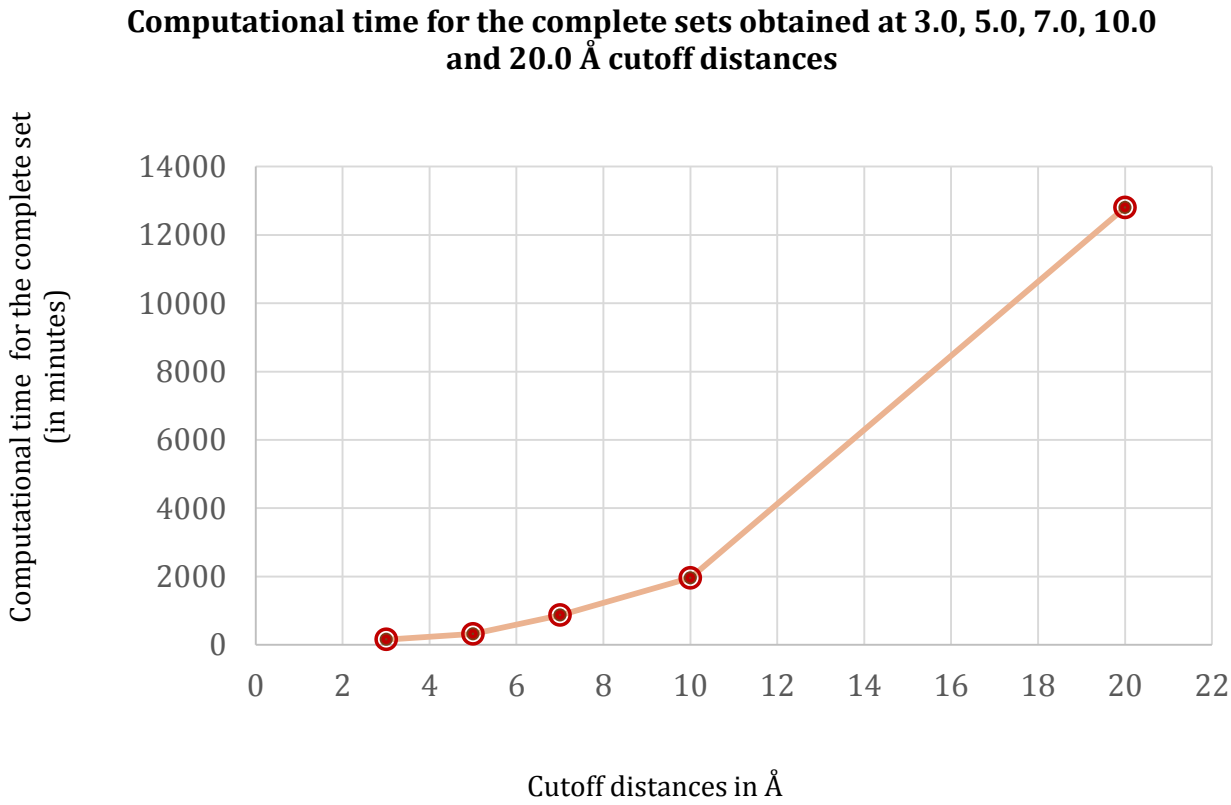
Table I.3-1: *Pearson and Kendall values for the data presented in Figure I.3-1.*

It is clear from the general trends that, as we increase the cutoff size, which means increasing the model sizes of our pockets, it converges to an effective size limit, and this size limit which can be regarded as the efficient pocket size limit which is enough to represent the overall protein in an acceptable way.

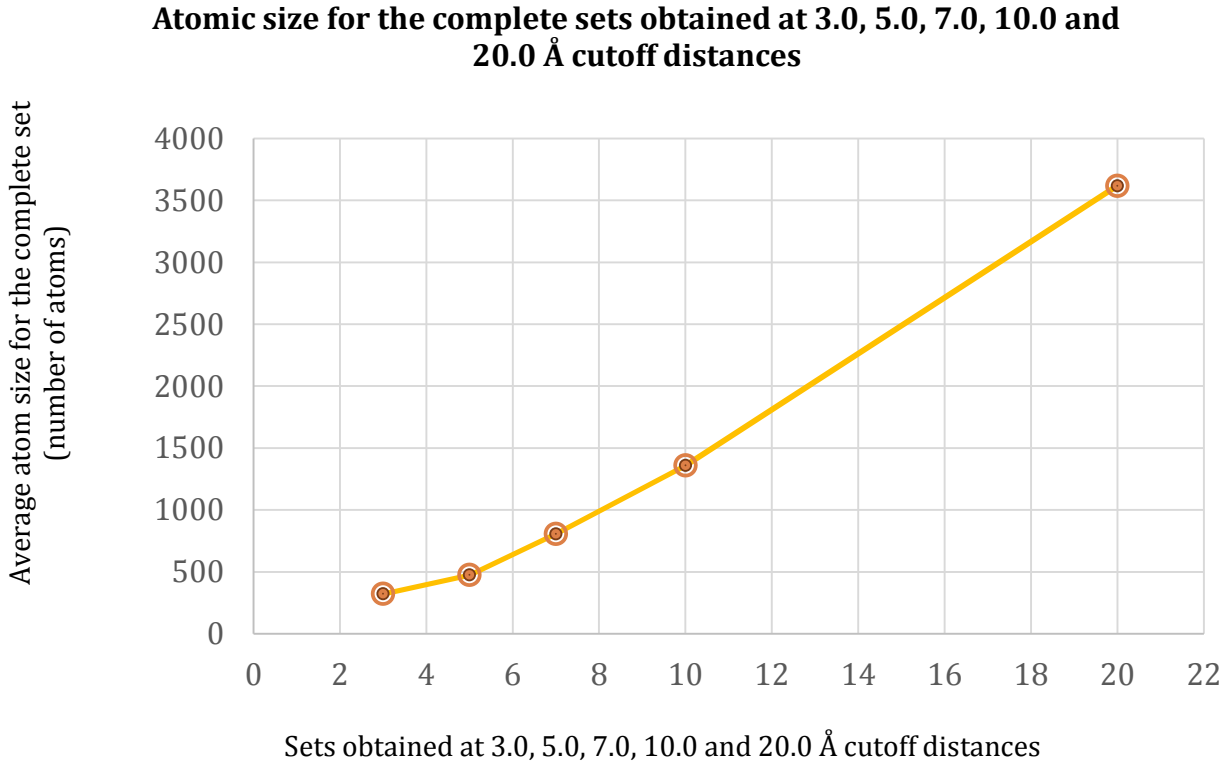
We can state that 3.0 Å pocket is a bit different than the 10.0 Å pocket based on the correlation values (R=0.83 or τ=0.76). As we increase the cutoff to 5.0 Å, and compare with 10.0 Å again, a better correlation is observed (R=0.96 or τ=0.70), which means the 5.0 Å pocket is a better representative for the 10.0 Å pocket model. Then, as the cutoff size is increased further and when 7.0 Å pocket is examined, it is seen that this model is almost in one to one model resemblance with the 10.0 Å pocket with a correlation of R=1.00 or τ=0.96. This means that 7.0 Å leads to a reasonably good model of our protein.

Still to investigate further, 7.0 vs 20.0 Å has a correlation of R=0.98 (including only one outlier data) and R=0.99 (without any outlier data), whereas, 10.0 vs 20.0 Å has a correlation of R=0.98 (including only one outlier data) and R=1.00 (without any outlier data). This means going beyond 7.0 Å will not change the results critically, and in general, the smaller the model, the faster the calculations. Therefore, 7.0 Å can be considered as the optimum cutoff distance.

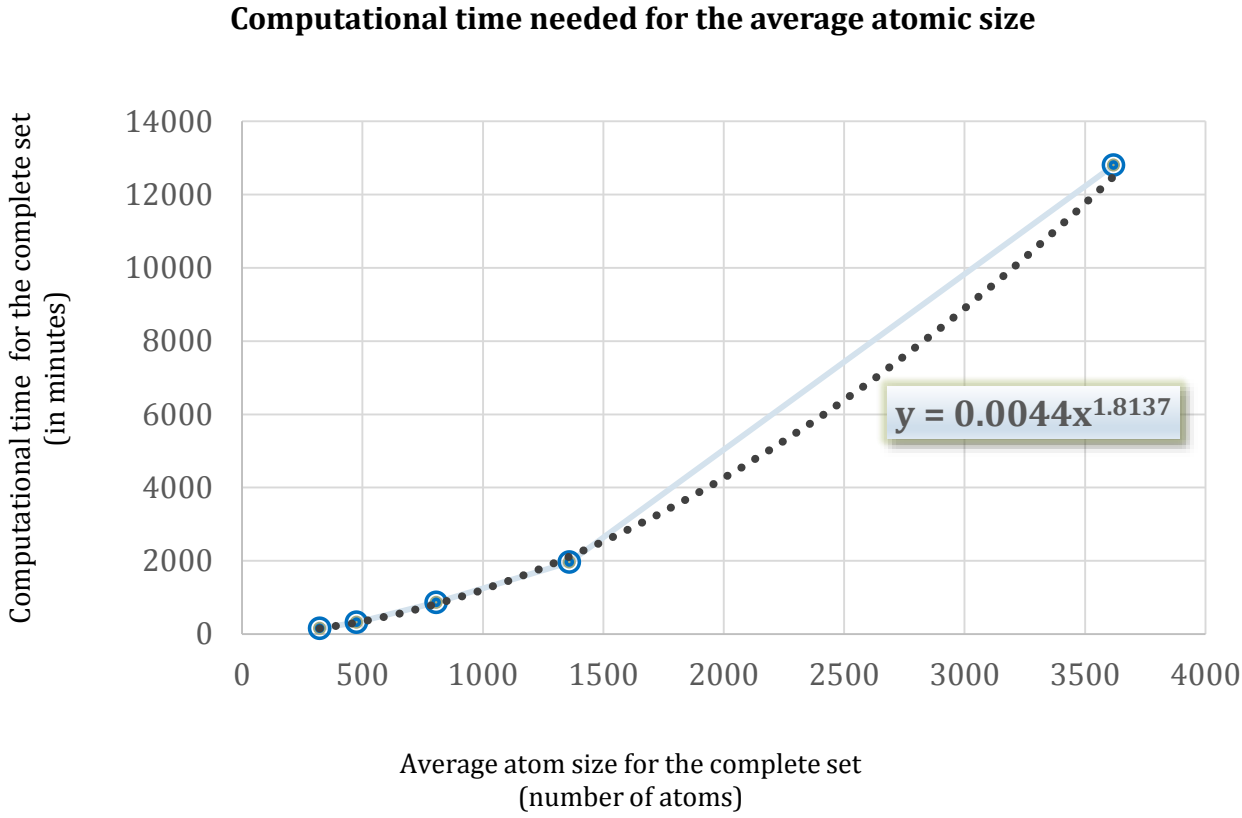
Computational cost comparison for these calculations are also given in Graph I.3-1. Then, this is followed with the atomic size comparison in Graph I.3-2. Finally, the computational time and atomic size comparison is given in Graph I.3-3.



Graph I.3-1: *Computational time (in minutes) to calculate the full sets at given cutoff distances.*
154 minutes (2.6 hours) for 3.0 Å,
317minutes (5.3 hours) for 5.0 Å,
863 minutes (14.4 hours) for 7.0 Å,
1958 minutes (32.6 hours or 1.4 days) for 10.0 Å,
12801 minutes (213 hours or 8.9 days) for 20.0 Å cutoff sets.



Graph I.3-2: *Average atom size (number of atoms) in the full sets at given cutoff distances.*
322 number of atoms in average within 3.0 Å set,
475 number of atoms in average within 5.0 Å set,
807 number of atoms in average within 7.0 Å set,
1358 number of atoms in average within 10.0 Å set,
3618 number of atoms in average within 20.0 Å set.



Graph I.3-3: *Computational time needed for the average atom size (number of atoms).*
322 number of atoms in average within 154 minutes (2.6 hours),
475 number of atoms in average within 317minutes (5.3 hours),
807 number of atoms in average within 863 minutes (14.4 hours),
1358 number of atoms in average within1958 minutes (32.6 hours or 1.4 days),
3618 number of atoms in average within 12801 minutes (213 hours or 8.9 days).

The scaling is found out to be as $N^{1.8}$, which is close to quadratic scaling N^2, thus it is rather good for a QM method which usually scale as N^{3-4}.

Due to our cutting algorithm, 7.0 Å cutoff has an actual corresponding effective distance of about 10-12 Å. This result is motivating since it is also found out to be in a good agreement with studies on small model systems [257].

In that sense, we would have liked to continue with this optimum cutoff 7.0 Å for all the comparisons with different methods, but unfortunately, unlike the SQM methods giving us this opportunity to handle large systems, the computational costs are limited for many other methods, like for example DFT methods.

As DFT is also available for them, 3.0 Å and 5.0 Å cutoff pocket models are the chosen for a detailed SQM and DFT comparisons.

Moreover, for the overall method comparison purposes, which will be including WFT, DFT, SQM and MM methods, 3.0 Å cutoff level models will be the only ones to focus on due to the computational availabilities.

For a real scoring application, a larger cutoff – as much as computations do allow – is desirable, but that seems only possible with SQM methods for now.

The next section aims to test and compare various SQM methods in order to have a valid SQM reference method to begin with.

Comparison of SQM methods

The goal with this comparison here is to see the best suited SQM methods for our study and test the reliability of our own method (Korth's PM6-DH+ method [50]) for this research as well.

As it is mentioned in the previous sections, PM6-DH+ method is amongst the most reliable enhanced semi empirical methods which was also preferred in many different application studies [27]. Its performance is tested and compared with several other SQM methods as well, which are: AM1, PM6, PM6-D, PM6-DH2, and PM6-DH2X methods.

Results are given in Figure I.3-2 and in Table I.3-2. Data points are plotted in comparison to PM6-DH+ values with the correlation values indicated. All computations are done with COSMO solvation models.

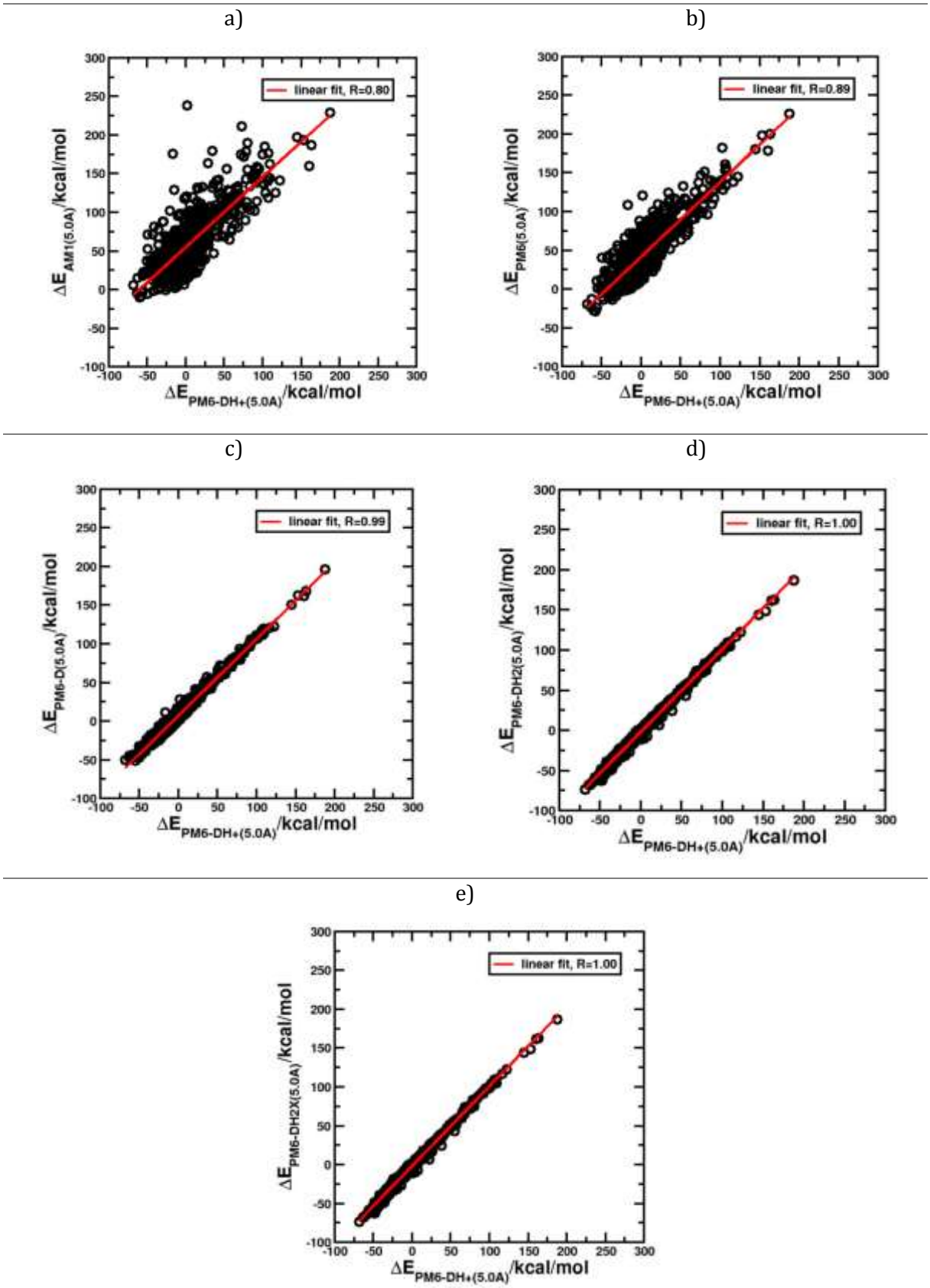


Figure I.3-2: *Correlations of different SQM approaches a) AM1 b) PM6 c) PM6-D d) PM6-D2 e) PM6-DH2X, each method compared against PM6-DH+. All computations are performed at 5.0 Å cutoff distances and involve COSMO solvation models.*

List of SQM Methods compared with PM6-DH+ method at 5.0 Å cutoff (all with COSMO solvation models)	Figure I.3-2	Pearson (R) correlation	Kendall (τ) correlation
AM1 (5.0 Å)	a	0.80	0.56
PM6 (5.0 Å)	b	0.89	0.67
PM6-D (5.0 Å)	c	0.99	0.90
PM6-DH2 (5.0 Å)	d	1.00	0.94
PM6-DH2X (5.0 Å)	e	1.00	0.94

Table I.3-2: *Pearson and Kendall values for the data presented in Figure I.3-2.*

The results overall confirm our choice of PM6-DH+ as the reference method.

AM1 shows a bigger deviation in the light of small benchmark studies, it seems likely less accurate [27]. Therefore, after this first comparison, AM1 methods were excluded from further consideration.

It is also observed that there is a significant importance of dispersion corrections, as there is a difference in between PM6 and PM6-D methods.

However, looking at the correlation values of PM6-D, PM6-DH2 and PM6-DH2X, only a negligible contribution/change is observed in between these which is resulting from the empirical corrections for hydrogen- and halogen-bonding. This finding is also in good agreement with Ryde and his coworkers study [88].

More specifically, PM6-DH2 and PM6-D2HX are both showing the same performance as PM6-DH+ with a correlation of R=1.00. Hence, a correlation of R=1.00 can be expressed as having similar results as PM6-DH+.

Additionally, related to this test, Table I.3-3 shows the error values MD, MAD, MD*, MAD* in comparison to the PM6-DH+ method.

List of SQM Methods compared with PM6-DH+ method at 5.0 Å cutoff (all with COSMO solvation models)	Figure I.3-2	MD	MAD	MD*	MAD*
		(kcal/mol)			
AM1 (5.0 Å)	a	-54	54	-5	20
PM6 (5.0 Å)	b	-41	41	-2	14
PM6-D (5.0 Å)	c	-7	7	0	4
PM6-DH2 (5.0 Å)	d	2	2	0	2
PM6-DH2X (5.0 Å)	e	2	2	0	2

Table I.3-3: *MD, MD*, MAD, MAD* values for SQM method comparisons*

Based on Table I.3-3, MD and MAD values for the "non-enhanced" SQM methods (basically the methods without a Hydrogen (H) or a Dispersion (D) correction: the AM1 and PM6 methods) against an "enhanced" SQM method (our enhanced PM6-DH+ method in this case) are found to be very large. These are:

- MD: -54 and MAD: 54 kcal/mol for the case of, AM1 vs PM6-DH+
- MD: -41 and MAD: 41 kcal/mol for the case of, PM6 vs PM6-DH+

This shows that there is a huge difference in between these non-enhanced and enhanced methods. Even when these are corrected for the systematic shifts, these values are still high.

AM1, when it is scaled and shifted according to the following equation,

$y = 54.643 + 0.90931 * x$, then we obtain,

- MD*: -5 and MAD*: 20 kcal/mol for the case of, AM1 vs PM6-DH+.

For the PM6, with

$y = 41.382 + 0.96407 * x$, then the values become:

- MD*: -2 and MAD*: 14 kcal/mol for the case of, PM6 vs PM6-DH+.

For the comparison of "enhanced SQM methods", which include the following methods: PM6-D, PM6-DH2 and PM6-DH2X, against our reference: "enhanced SQM", PM6-DH+ method, now that we observe that the MD and MAD values are much smaller:

- MD: -7 and MAD: 7 kcal/mol, for PM6-D vs PM6-DH+
- MD: 2 and MAD: 2 kcal/mol both for PM6-DH2 and PM6-DH2X vs PM6-DH+

They are corrected again with scaling and shifting procedure. Then,

For PM6-D, with:

$y = 6.9079 + 0.99185 * x$, the values became:

- MD*: 0 and MAD*: 4 kcal/mol for PM6-D vs PM6-DH+.

For PM6-DH2 and PM6-DH2X, with,

$y = -2.0686 + 1.0287 * x$, the values became:

- MD*: 0 and MAD*: 2 kcal/mol for both PM6-DH2 and PM6-DH2X vs PM6-DH+.

Comparison of DFT methods

Next, DFT methods are compared with different functionals and basis sets. To begin with, the reference method here is taken as BP86-D2/TZVP method. The results are given in Figure I.3-3. Here, all the computations are done with COSMO solvation. Afterwards, based on a certain DFT method (BP86-D2/TZVP once again), in order to see the solvation effects, COSMO and COSMO-RS were compared in Figure I.3-4. Finally, in Figure I.3-5 different dispersion correction schemes are compared.

Comparison of Basis sets and Functionals

The following Figure I.3-3 and Table I.3-4 are obtained:

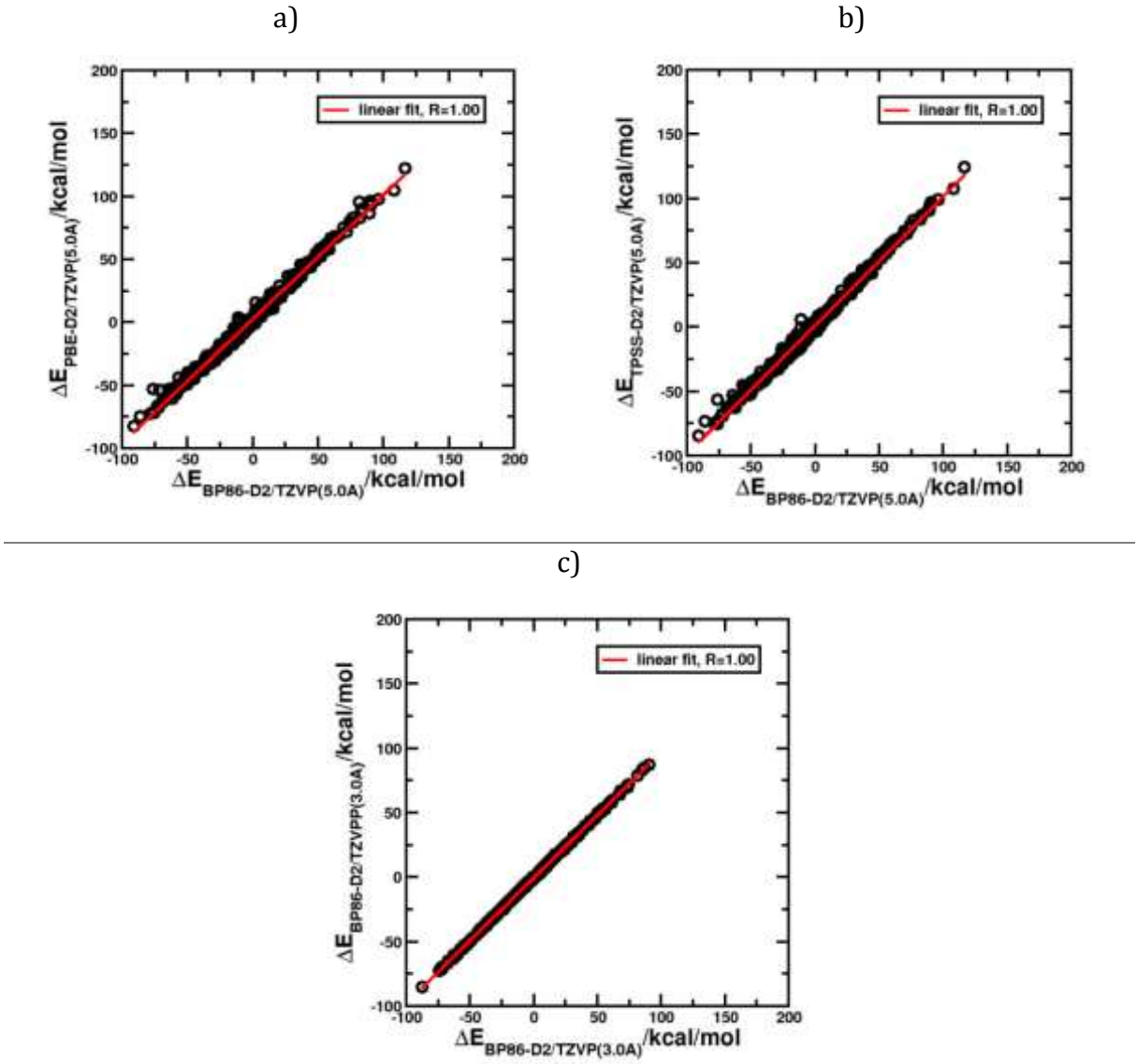


Figure I.3-3: *Correlation between different DFT functionals and basis sets: a) PBE-D2/TZVP b) TPSS-D2/TZVP c) BP86-D2/TZVPP, each of them plotted against the reference method BP86-D2/TZVP. Computations a) and b) are performed at 5.0 Å, c) is performed at 3.0 Å cutoff distances. All cases involve COSMO solvation models.*

List of DFT Methods compared with BP86-D2/TZVP method at 5.0 Å and 3.0 Å cutoff (all with COSMO solvation models)	Figure I.3-3	Pearson (R) correlation	Kendall (τ) correlation
PBE-D2/TZVP (5.0 Å) vs BP86-D2/TZVP (5.0 Å)	a	1.00	0.94
TPSS-D2/TZVP (5.0 Å) vs BP86-D2/TZVP (5.0 Å)	b	1.00	0.94
BP86-D2/TZVPP (3.0 Å) vs BP86-D2/TZVP (3.0 Å)	c	1.00	0.99

Table I.3-4: *Pearson and Kendall values for the data presented in Figure I.3-3*

From Figure I.3-3 and Table I.3-4, it can be concluded that, the DFT functionals (using PBE-D2 or TPSS-D2 instead of using BP86-D2) give very similar results. Also for the basis set comparisons, increasing the level beyond TZVP (and having TZVPP instead) only makes a little change. The interaction energies are rather uniform with a correlation of Pearson R=1.00 for the cases we have investigated. Additionally, MD, MD*, MAD and MAD* values are also given in Table I.3-5 as follows:

List of DFT Methods compared with BP86-D2/TZVP method at 5.0 Å and 3.0 Å cutoff (all with COSMO solvation models)	Figure I.3-3	MD	MAD	MD*	MAD*
		(kcal/mol)			
PBE-D2/TZVP (5.0 Å) vs BP86-D2/TZVP (5.0 Å)	a	-2.6	3.0	0.0	2.3
TPSS-D2/TZVP (5.0 Å) vs BP86-D2/TZVP (5.0 Å)	b	-1.2	2.4	0.0	2.3
BP86-D2/TZVPP (3.0 Å) vs BP86-D2/TZVP (3.0 Å)	c	0.5	0.8	0.0	0.4

Table I.3-5: *MD, MD*, MAD, MAD* values for DFT functionals and basis set comparisons*

As can be seen from the Table I.3-5, all of them have resulted in similar values. The details can be listed as follows:

Before shifting and scaling the values were:

- PBE-D2/TZVP (5.0 Å) vs BP86-D2/TZVP (5.0 Å) ; MD: -2.6 and MAD: 3.0 kcal/mol
- TPSS-D2/TZVP (5.0 Å) vs BP86-D2/TZVP (5.0 Å) ; MD: -1.2 and MAD: 2.4 kcal/mol,

Whereas, after shifting and scaling with

y = 2.4491 + 0.98718 * x, and y = 1.2083 + 1.0019 * x respectively, these became:

- PBE-D2/TZVP (5.0 Å) vs BP86-D2/TZVP (5.0 Å) ; MD*: 0.0 and MAD* 2.3 kcal/mol
- TPSS-D2/TZVP (5.0 Å) vs BP86-D2/TZVP (5.0 Å) ; MD*: 0.0 and MAD*: 2.3 kcal/mol

For the comparison of TZVPP with TZVP, before shifting and scaling, it was:

- BP86-D2/TZVPP (3.0 Å) vs BP86-D2/TZVP (3.0 Å); MD: 0.5 and MAD: 0.8 kcal/mol

And, after the adjustment with y = -0.71515 + 0.97313 * x, this became:

- BP86-D2/TZVPP (3.0 Å) vs BP86-D2/TZVP (3.0 Å); MD*: 0.0 and MAD*: 0.4 kcal/mol.

Comparison of COSMO and COSMO-RS

The differences between COSMO and COSMO-RS models are investigated, and the result is presented in Figure I.3-4, Table I.3-6 and Table I.3-7.

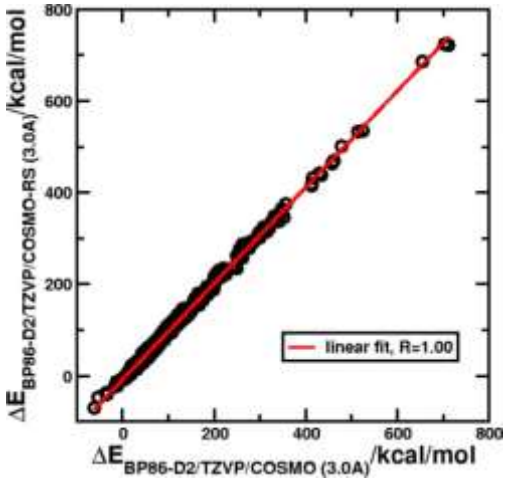


Figure I.3-4: *Correlation between the solvation contributions of COSMO and COSMO-RS for BP86/TZVP calculations.*

Entry	Figure	Pearson (R) correlation	Kendall (τ) correlation
BP86-D2/TZVP/COSMO-RS (3.0 Å) vs BP86-D2/TZVP/COSMO (3.0 Å)	I.3-4	1.00	0.95

Table I.3-6: *Pearson and Kendall values for the data presented in Figure I.3-4.*

Entry	Figure	MD	MAD	MD*	MAD*
		(kcal/mol)			
BP86-D2/TZVP/COSMO-RS (3.0 Å) vs BP86-D2/TZVP/COSMO (3.0 Å)	I.3-4	2.9	6.8	-0.4	4.4

Table I.3-7: *MD, MD*, MAD, MAD* values for solvation models correlations*

Before scaling and shifting, the values were MD: 2.9 and MAD: 6.8kcal/mol, and after scaling and shifting with the following equation,

$y = -8.3776 + 1.0511 * x$, they became:

- MD*: -0.4 and MAD*: 4.4 kcal/mol.

Comparison of Dispersion Corrections

Investigation of the different dispersion corrections are presented in Figure I.3-5 and Tables I.3-8 and I.3-9 below.

It is again shown that, only for the case when rankings are compared, differences in dispersion corrections does not play a decisive role either.

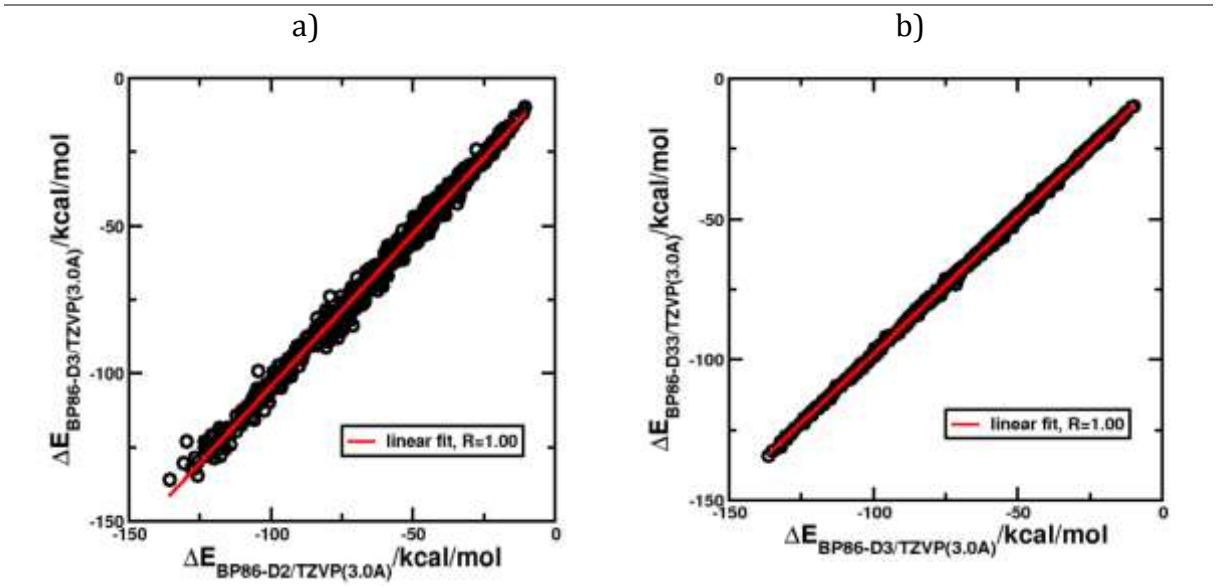


Figure I.3-5: *Correlation between different dispersion schemes for DFT, BP86/TZVP methods: a) D3 against D2, b) D33 against D3 is plotted. All calculations are done with COSMO solvation models and at 3.0 Å cutoff distances.*

Entry (all with COSMO solvation models)	Figure I.3-5	Pearson (R) correlation	Kendall (τ) correlation
BP86-D3/TZVP (3.0 Å) vs BP86-D2/TZVP (3.0 Å)	a	1.00	0.94
BP86-D33/TZVP (3.0 Å) vs BP86-D3/TZVP (3.0 Å)	b	1.00	0.99

Table I.3-8: *Pearson and Kendall values for the data presented in Figure I.3-5.*

Entry (all with COSMO solvation models)	Figure I.3-5	MD	MAD	MD*	MAD*
		(kcal/mol)			
BP86-D3/TZVP (3.0 Å) vs BP86-D2/TZVP (3.0 Å)	a	3.2	3.4	0.0	1.9
BP86-D33/TZVP (3.0 Å) vs BP86-D3/TZVP (3.0 Å)	b	-1.2	1.2	0.0	0.4

Table I.3-9: *MD, MD*, MAD, MAD* values for solvation models correlations*

Before scaling and shifting, the values were:

- MD: 3.2 and MAD: 3.4 kcal/mol for BP86-D3/TZVP (3.0 Å) vs BP86-D2/TZVP (3.0 Å), and ,
- MD: -1.2 and MAD: 1.2 kcal/mol for BP86-D33/TZVP (3.0 Å) vs BP86-D3/TZVP (3.0 Å).

After scaling and shifting with: $y = -1.1542 + 1.0332 * x$ for the first case, values became:

- MD*: 0.0 and MAD*: 1.9 kcal/mol for BP86-D3/TZVP (3.0 Å) vs BP86-D2/TZVP (3.0 Å),

and with $y = -0.10118 + 0.97991 * x$, for the second case, the values became:

- MD*: 0.0 and MAD*: 0.4 kcal/mol for BP86-D33/TZVP (3.0 Å) vs BP86-D3/TZVP (3.0 Å).

To sum up, based on the overall performance results so far, the following methods will be used as references for further comparisons.

- For the SQM methods: The PM6-DH+ method,
- For the DFT methods: BP86-D2/TZVP method.

Hereby, it is important to emphasize once more that, 'ranking' is our main consideration overall. That's why our findings and statements are not counter claims for the importance of differences among the basis sets, functionals, dispersion correction schemes or the usage of COSMO versions for many systems and applications, especially where the numerical data values are important rather than the overall tendency (ranking). Instead, on the contrary, when the absolute data values are required, the differences in these parameters are highly crucial and distinctive to take into considerations.

Comparison between different computational methods

After the SQM methods and DFT methods are tested and compared within their own classes, a reference method for SQM and DFT are assigned as a result. This time, these reference SQM and DFT methods are compared with each other and preferably with the 5.0 Å pocket models, since DFT methods are computationally available at this range.

Comparison between SQM and DFT methods at 5.0 Å cutoff

The correlation between the SQM (PM6-DH+) and DFT (BP86-D2/TZVP) methods at the 5.0 Å cutoff range is investigated and given in Figure I.3-6 and in Table I.3-10 below.

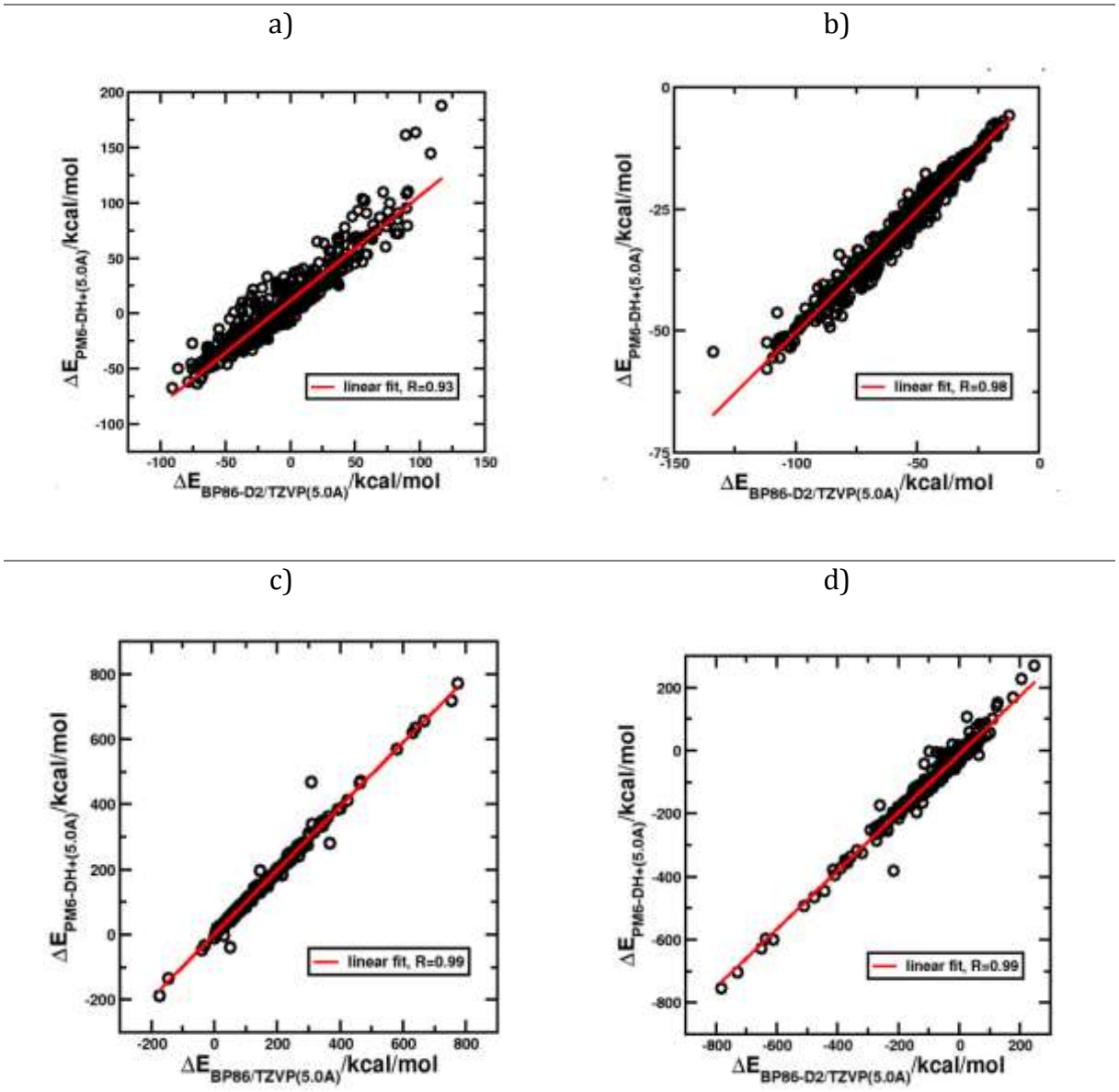


Figure I.3-6: *Detailed comparison between SQM (PM6-DH+) and DFT (BP86-D2/TZVP) methods: a) Overall Interaction Energy b) dispersion contribution c) solvation contribution d) electronic contribution. All computations are with COSMO solvation models and at 5.0 Å cutoffs.*

Entry	Figure I.3-6	Pearson (R) correlation (5.0 Å)	Kendall (τ) correlation (5.0 Å)
Overall interaction energy (with all contributions)	a	0.93	0.77
Only dispersion contribution	b	0.98	0.89
Only solvation contribution	c	0.99	0.95
Only Electronic contribution	d	0.99	0.84

Table I.3-10: *Pearson and Kendall values for the data presented in Figure I.3-6.*

It is observed that, correlation increases from 0.93 to 0.98 (parts a & b) when there is the dispersion contribution. Correlation again increases from 0.93 to 0.99 (parts a & c) when there is only the solvation contribution. A similar increasing tendency is observed for the polarization effects as well (electronic contributions), where the correlation increases from 0.93 to 0.99 for the (parts a & d).

These all indicate that SQM methods still need some improvement and that the balance of the intermolecular interactions are not obtained well enough with SQM methods. Dispersion contribution on its own is a bit problematic too, because SQM methods already model noncovalent effects but without taking care of a proper long range behaviour. In conclusion, SQM methods can be improved in several ways [69, 36, 74-76].

Comparison in between MM methods and best suited SQM and DFT methods

This part includes an overall comparison in between MM, SQM and DFT methods this time. We better remark on the selection of the MMFF94 method here, since it is indicated that interaction energies from most MM methods are highly correlated [290],we are initially assuming that the qualitative results would not be so different for the other MM methods as well. Amber ff99sb/GAFF is in any case evaluated to test this assumption.

Figure I.3-7 and Table I.3-11 below, compares the correlations in between SQM: PM6-DH+, DFT: BP86-D2/TZVP and MM: MMFF94 methods. As the next step, two MM methods (MMFF94 and ff99sb/GAFF) are compared with eachother as well in Table I.3-11. All computations are done at 3.0 Å cutoff level.

a) b)

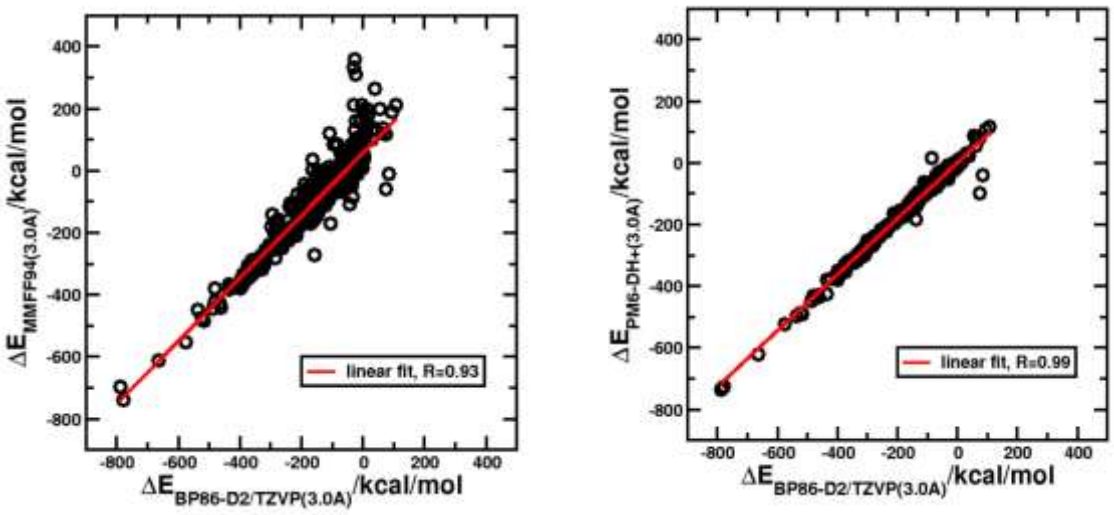


Figure I.3-7: *Detailed comparison between SQM (PM6-DH+), MM (MMFF94) and DFT (BP86-D2/TZVP) methods: a) MM (MMFF94) against DFT (BP86-D2/TZVP), b) SQM (PM6-DH+) against DFT (BP86-D2/TZVP). All calculations are done with 3.0 Å cutoff.*

It is clear from the plots that, SQM results are closer to the DFT results when compared with the MM results. Therefore, SQM seems to be an improvement over MM methods in case of rankings.

Entry		Pearson (R) correlation	Kendall (τ) correlation
I	ff99sb/GAFF (3.0 Å) vs MMFF94 (3.0 Å)	0.95	0.75
II	ff99sb/GAFF (3.0 Å) vs BP86-D2/TZVP (3.0 Å)	0.96	0.73
III	ff99sb/GAFF/GBSA (3.0 Å) vs BP86-D2/TZVP/**COSMO**(3.0 Å)	0.60	0.51
IV	MMFF94 (3.0 Å) vs BP86-D2/TZVP (3.0 Å) **Figure I.3-7a**	0.93	0.76
V	PM6-DH+ (3.0 Å) vs BP86-D2/TZVP (3.0 Å) **Figure I.3-7b**	0.99	0.93
VI	PM6-DH+/**COSMO** (3.0 Å) vs BP86-D2/TZVP/**COSMO**(3.0 Å)	0.92	0.74

Table I.3-11: *Pearson and Kendall values for the data presented in Figure I.3-7 and some additional method comparison tests.*

Due to the findings, the MMFF94 and ff99sb/GAFF similarity assumption is validated in comparison with the DFT method (second entry: ff99sb/GAFF vs BP86-D2/TZVP has a correlation of R= 0.96 and τ= 0.73; whereas, fourth entry: MMFF94 vs BP86-D2/TZVP has a correlation of R= 0.93 and τ= 0.76).

The comparison in between the second and third entries, as well as, the comparison in between the fifth with sixth entries show the performance results of the same entries only by being with or without COSMO solvation models. It can be observed that the solvation effects are not improving the ranking correlation in between SQM and DFT methods.

Among the ones which are having solvation effects, it is also observed that, the third entry ff99sb/GAFF/GBSA (3.0 Å) vs BP86-D2/TZVP/COSMO (3.0 Å), has a correlation of R= 0.60 and τ= 0.51, while, the sixth entry PM6-DH+ (3.0 Å) COSMO vs BP86-D2/TZVP (3.0 Å) COSMO has a correlation of R= 0.92 and τ= 0.74. These result are favouring the SQM methods in this category.

Using more sophisticated charge and solvation models might improve these results for the case of MM methods, however due to the already observed slight changes in between MM methods with the previous tests, we somehow do not expect a difference in qualification in that sense. Therefore, from the findings here, it can be reported that, SQM methods are an improvement for the MM methods.

MD, MD*, MAD and MAD*values are also tabulated in the following Table I.3-12 to compare our SQM, DFT and MM references.

Entry	MD	MAD	MD*	MAD*
	(kcal/mol)			
PM6-DH+ (3.0 Å) vs BP86-D2/TZVP (3.0 Å)	-12	14	-0.1	7.1
MMFF94 (3.0 Å) vs BP86-D2/TZVP (3.0 Å)	-55	57	0.4	30.9

Table I.3-12: *MD, MD*, MAD, MAD* values for solvation models correlations*

The differences in between the numerical values once more suggest the reference SQM method over the reference MM method.

Due to the numerical values, the mean absolute deviations from the DFT method are roughly at the size of 1% (14/1098*100) and 5% (57/1098*100) for the SQM and MM methods respectively in percentages and these are on the electronic scale with an energy range of about 1098kcal/mol (this 1000kcal/mol around range can be seen on the axes of the relevant plots as well).

Once the solvation is considered, then it is a resulting energy range of about 250kcal/mol, then these values become 6% (14/250*100) for the SQM and 23% (57/250*100) for the MM methods.

It is already known that, any type of a possible calculation method right now is unfortunately still far away from the desired accuracy level (there is still a big gap between the experimental values and the computed values). However, even this is still being the case, at least, reaching to the best and computationally the most cost effective alternative method, is still regarded as an improvement.

From our first work (Research Stage I, Section 3.1.1), by evaluating the performances of different computational methods, it is clear that the SQM methods are doing perfectly well for the ranking purposes and with an accuracy similar to DFT.

It should however be emphasized that, there is still a need and a room for better scoring functions, and hence better model approaches for SQM methods.

Among the reasons of why SQM is far away from the experimental values, we can list the following: First of all, we are using simple model systems, and these model systems rely only on a single binding mode (docked, fixed, ligand-protein pairs). Second, during our calculations, we do not ignore the ones with some possible problems from our complex sets (i.e. like the problematic cases with protonation, cavity water and relaxation effects so on). Therefore, improving these conditions might improve the results too. In addition, according to our estimations, especially neglecting the entropic effects is an important drawback in our approaches. On this occasion, we will have further tests and analysis on entropy, enthalpy effects in our Research Stage III, Section 3.1.3.

I.4 Conclusion

To summarize our findings:

- Quantum mechanical calculations can be restricted to smaller model system without losing their predictive capability as we have also shown in our tests with different sized model systems.
- The constructed model systems are well within the reach of SQM methods, but still challenging for DFT approaches with their sizes.
- Different SQM methods have different features (as it is mentioned in our literature Section 2.1.3 before). With our findings, it can be concluded that our reference SQM method, PM6-DH+ is reliable enough to be recommended.
- Even for the various and large scaled protein-ligand sets, in case when the ranking is the only concern, there is almost no significant difference between using or not using the dispersion corrections and implicit solvent models. This is only being said for the ranking, but for any other calculations, it is still recommend to use DFT-D3/TZVP/COSMO (-RS) methods.
- For the comparison in between the different computational models that are having different theoretical levels, like DFT vs SQM, we have found out that, as an enhanced SQM approach, PM6-DH+ performs very similar to DFT-D and this shows a substantial improvement upon classical potentials.
- When we have tests on the smaller energy scales, by looking at the differences, it can be stated that SQM has a deviation of 5% from DFT whereas MM (FF) has a deviation of 15% from DFT.
- We have gone through different stages for model preparations in our study and based on this experience, we can state that, an automatic way of a model system preparation needs further adjustments and especially the neglected entropic (and/or enthalpic) effects should be included or revised.

After these remarks, our results bring us to the second stage of our research (Section 3.1.2), where, this time our focus will be on the comparison of several wave function theory (WFT), density functional theory (DFT) and semiempirical quantum mechanical (SQM) approaches against high-level theoretical references, and again by making use of the realistic protein-ligand (pocket) model structures.

3.1.2 Research Stage II [91]

II.1 Introduction

As it is previously mentioned in our literature Section 2.2, for *in silico* drug design, scoring functions mostly perform well but they still need to be improved for a higher accuracy. When it is about "scoring", there are some QM-level studies on computational methods [89, 88, 254, 255, 129, 120, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300]. To extend our research, this time our focus will be on benchmarking several wave function theory (WFT), density functional theory (DFT) and semiempirical quantum mechanical (SQM) approaches against high-level theoretical references for realistic test cases.

II.2 Computational Details & Generation of PLI10 model systems

Our previous study (Section 3.1.1) was covering PDBbind 2007 data set and the pocket models prepared at the cutoff distances of: 3.0, to 5.0, 7.0, 10.0 and 20.0 Å. For the preparation of the pocket models our own cutting algorithm was used and the resultant structures were always again 3-5 Å larger than the specified cutoff-distance labels based on the algorithm. The same PDBbind 2007 set is used again for this study and therefore previously mentioned concerns are also valid. For example,

- After obtaining pockets, the terminal structures are capped with hydrogen atoms.
- Histidine residues were assumed to be neutral (with protonation at the sterically less crowded place if necessary), because automatic pKa prediction was technically not possible for all systems.
- Overall charges were assigned according to automatic Lewis-structure analysis. This assignment is based on the number of additional or missing bonds found via comparison to sums of van-der-Waals radii, and then these were double-checked with the automatic assignment implemented in MOPAC. Also, MOPAC values are taken in case of disagreement.
- Irregular ligand files, for example files with atoms having an unreasonable number of bonds, or files having atoms missing from amino acids, or files where some atoms were too close to each other (this was judged by van-der-Waals radii) were discarded.

Amongst the obtained pocket models, the sizes of ranged from minimum 70 to maximum 700 atoms for 3.0 Å, whereas it reached up to about minimum 900 to 7000 atoms for 20.0 Å. Albeit this time, in order to be able to do WFT computations, we have picked out only the 10 smallest complexes from the pocket models at the cutoff 3.0 Å set. This set is named as "PLI10 set".

This PLI10 set involves the following complexes:

Abbreviation	Name of the Complex
1AVN	Human carbonic anhydrase II complexed with histamine [301]
1D7J	FK506 binding protein (FKBP) complexed with 4-hydroxy-2-butanone [302]
1E4H	Human transthyretin complexed with two bromophenols [303]
1JYS	Nucleosidase from E. Coli complexed with MTA/AdoHcy [304]
1QPB	Pyruvate decarboxylase from yeast complexed with pyruvamide [305] (this structure supersedes the now removed PDB entry 1ypd)
1WEI	The catalytic domain of muty from E. Coli K20A complexed with adenine [306]
2BJM	SPE7 complexed with anthrone [307]
2F8I	Human transthyretin (TTR) complexed with benzoxazole [308]
2HDQ	AmpC beta-lactamase from E. Coli K12 complexed with 2-carboxythiophene [309]
2HDR	AmpC beta-lactamase from E. Coli K12 complexed with 4-Amino-3-hydroxybenzoic acid [309]

***Table II.2-1**: The names and descriptions of the complexes in PLI10 set*

Among these, 2BJM has a positive binding energy, which is most probably an indication of a wrong geometry setup in PDBbind2007, however this structure is still a valid point on the Potential Energy Hyper Surface, therefore we keep it still for the benchmarking.

The names of the methods and programs that are used for this research stage II are listed in Table II.2-2 as follows:

Computational Methods		
Method Type	**Name**	**Computational Details**
DFT	B3-LYP [310, 311], B-LYP [275, 312] BP86 [275, 276], B97-D [279], PBE [277], and TPSS [278], with empirical dispersion corrections of: D2 [279], D3 [52] and D3 plus three-body-dispersion (named Dabc the following) types [280, 281].	The calculations are done with Turbomole 6.4 [280,281], using SVP, TZVP and TZVPP [282] Gaussian AO basis sets, RI approximation for two-electron integrals [283,284, 313] and COSMO as well as COSMO-RS (via COSMOtherm) solvation models [285].
	M06 and M06-2X [314]	Calculations are done via NWChem [315]
WFT	Perturbation methods of second order, (SCS)-MP2 [316], B2-PLYP [317] and explicitly correlated MP2-F12 [318].	Calculations are done also with TURBOMOLE [281] by using def2-TZVPP and aug-cc-pVDZ [319, 320, 321] basis sets, along with the RI approximation applied to the Coulomb and exchange integrals in the Hartree--Fock part [322] with corresponding auxiliary basis sets [323] as well as to the MP2 part with corresponding basis sets [313]
WFT	pCCSD-1a [324, 325] within the local pair natural orbital approximation [326, 327]	def2-TZVPP basis set was used and calculations are done in ORCA, version 2.9 [328]. RI approximation was used in the Hartree--Fock and correlation part and the same auxiliary basis sets as before.
Semi empirical	PM6-DH+[50]	MOPAC2012 [58] was used, with MOZYME linear scaling algorithm [286] and COSMO solvation models [285].

***Table II.2-2**: List of computational methods used for Research Stage II*

Interaction energies are computed again with the previously mentioned equation,

$$E_{interaction} = E_{complex} - (E_{pocket} + E_{ligand}) \quad \text{(Eqn. 3.1.1-1)}$$

II.3 Results and Discussion

Non-covalent bonds, which are important for biomolecular interactions, require a high-level theory method for their computations if there isn't any empirical input available [258, 329]. The CCSD (T) (i.e, coupled-cluster theory with single and double excitations and perturbative triples) is known as the "gold standard" method of Quantum Chemistry [330]. However, this method is not applicable for the large sized systems, therefore it is common to refer to the extrapolation schemes to produce highly-accurate reference values (Complete Basis Set or CBS) [331, 332].

CBS values being the reference data, Table II.3-1 involves the results of the PLI10 set data values for various WFT methods. These WFT calculations were done by Tobias Schwabe; co-author of our article [91].

Once these WFT data is obtained, then the deviations with respect to this reference method are also calculated for a variety of DFT and SQM methods and these are all presented as MAD and MD error statistics in Table II.3-2.

WFT calculations

For the CCSD (T) gold standard method, as for the extrapolations to be considered, LPNO-pCCSD (local pair natural orbital parameterized coupled-cluster theory with single and double excitations) is a good comparable candidate [333].

Extrapolation was done as follows:

- First, a correction term:

$$\left(E_{LPNO-pCCSD}^{def2-TZVPP} - E_{MP2}^{def2-TZVPP}\right) \quad \text{(Eqn. 3.1.2-1)}$$

  from, local pair natural orbital parameterized coupled-cluster theory with single and double excitations (LPNO-pCCSD), was added to

- the counter-poise corrected (CPC) explicitly correlated second-order Moller–Plesset perturbation theory (MP2-F12) data:

$$E_{MP2-F12(CPC)}^{aug-cc-pVDZ} \quad \text{(Eqn. 3.1.2-2)}$$

  close to the basis set limit.

This F12 approach with aug-cc-pVDZ basis sets are comparable to aug-cc-pVQZ.

Then, the reference energies are obtained as follows:

$$E_{reference} = E_{MP2-F12(CPC)}^{aug-cc-pVDZ} + \left(E_{LPNO-pCCSD}^{def2-TZVPP} - E_{MP2}^{def2-TZVPP}\right) \qquad \text{(Eqn. 3.1.2-3)}$$

Results are presented in the Table II.3-1 below. This also includes the plain LPNO-pCCSD/def2-TZVPP values, and, MP2 data that is with/out CPC or F12 basis sets, as well as having them both or not at all.

		I	II	III	IV	V	VI
PLI10 set	**Reference CBS values**	**ΔLPNO-pCCSD/ def2-TZVPP**	**ΔMP2-F12(CPC) aug-cc-pVDZ**	**ΔMP2(CPC) aug-cc-pVDZ**	**ΔMP2-F12 aug-cc-pVDZ**	**ΔMP2 aug-cc-pVDZ**	**ΔMP2 def2-TZVPP**
1AVN	-15.9	-1.6	-2.4	-1.2	-4.3	-8.4	-4.1
1D7J	-14.5	-1.2	-3.0	-0.5	-10.4	-12.3	-4.2
14EH	-11.9	-0.3	-2.4	0.5	-6.5	-12.6	-2.7
1JYS	-15.6	-1.7	-5.0	-1.1	-16.3	-17.6	-6.7
1QPB	-9.3	-2.3	-2.5	-0.8	-4.8	-9.9	-4.8
1WEI	-30.1	-3.4	-4.7	-1.9	-15.3	-16.0	-8.1
2BJM	10.2	-4.5	-6.8	-4.4	-16.9	-24.7	-11.3
2F8I	-84.3	-1.6	-4.4	1.5	-11.9	-16.8	-6.0
2HDQ	-205.8	-5.5	-4.3	-1.1	-8.3	-16.4	-9.8
2HDR	-0.9	-0.6	-4.4	1.1	-14.0	-12.5	-5.0
	MD	-2.3	-4.0	-0.8	-10.9	-14.7	-6.3
	MAD	2.3	4.0	1.4	10.9	14.7	6.3
	MIN	-5.5	-6.8	-4.4	-16.9	-24.7	-11.3
	MAX	-0.3	-2.4	1.5	-4.3	-8.4	-2.7

Table II.3-1: *Comparison of Wave Function Theory methods against the reference values (reference CBS) together with minimum and maximum values (MIN, MAX) given for the data results as well as error statistics (MD, MAD). CPC is counter poise corrected values, CBS is complete basis set limit extrapolations.*

		VII	VIII	IX	X
PLI10 set	**Reference CBS values**	**ΔSCS-MP2 def2-TZVPP**	**ΔSCS-MP2 CBS**	**ΔB2-PLYP-D3 def2-TZVPP**	**ΔB2-PLYP-D3 CBS**
1AVN	-15.9	-1.2	0.4	-3.1	-1.5
1D7J	-14.5	0.1	1.3	-2.2	-1.0
14EH	-11.9	1.1	1.4	-2.3	-1.9
1JYS	-15.6	-0.5	1.2	-2.6	-0.9
1QPB	-9.3	-1.2	1.1	-3.9	-1.6
1WEI	-30.1	-2.5	0.9	-4.7	-1.3
2BJM	10.2	-3.8	0.7	-5.3	-0.8
2F8I	-84.3	0.5	2.1	-2.6	-1.0
2HDQ	-205.8	-3.8	1.7	-5.7	-0.2
2HDR	-0.9	-0.1	0.8	-2.5	-1.9
	MD	-1.1	1.2	-3.5	-1.2
	MAD	1.5	1.2	3.5	1.2
	MIN	-3.8	0.4	-5.7	-1.9
	MAX	1.1	2.1	-2.2	-0.2

Table II.3-1 cont'd.: *Comparison of Wave Function Theory methods against the reference values (reference CBS) together with minimum and maximum values (MIN, MAX) given for the data results as well as error statistics (MD, MAD). CPC is counter poise corrected values, CBS is complete basis set limit extrapolations.*

From Table II.3-1, the following can be observed:

- The first entry (I), ΔLPNO-pCCSD/def2-TZVPP is without F12 and results in a MAD value of 2.3kcal/mol, (i.e. about 2kcal/mol).

- The second entry (II), ΔMP2-F12(CPC) /aug-cc-pVDZ, which is this time with CPC, is missing higher order correction from pCCSD, and it ends up with a MAD value of 4 kcal/mol.

- The third entry (III), ΔMP2(CPC)/aug-cc-pVDZ, uses CPC, but it also misses a higher order correlation when compared with pCCSD. However, it shows a small MAD value of 1.4 kcal/mol. The reason for this might be due to CPC being large for the small aug-cc-pVDZ basis sets (without F12), and since CPC is known to overshoot [3], it is here compensating for the missing higher order correction. Therefore the small value of this MAD value is obtained for the wrong reasons.

- The fourth entry (IV), ΔMP2-F12/aug-cc-pVDZ, is missing a higher order correlation again when compared with pCCSD, and unlike the third entry, it only uses the F12 approach but not CPC. This results in a MAD value of 10.9 kcal/mol. Accordingly, this creates a big effect when it is compared with the second entry (II) as well, because when the second and fourth entries are compared, they both have the same setting, but the second entry also has CPC as an addition to the F12. This seem to be resulted in having a smaller MAD value of 4kcal/mol with CPC addition when it is compared to not having it (11 kcal/mol- fourth entry). There are smaller overshooting problems as well. This is usually based on the completeness of the basis set. (More complete the basis set, smaller the basis set error, therefore corrections are smaller and overshooting is smaller too.)

- Looking at the fifth entry (V), ΔMP2/aug-cc-pVDZ is lacking a higher order correlation again when compared with pCCSD but which now has neither F12 nor CPC in addition, and it results in a large MAD value which is 14.7kcal/mol. In other words, this version has the worst performance overall.

- The sixth entry (VI), ΔMP2/def2-TZVPP, similar to the fifth entry, is again without any higher order correlation and it is without F12 or CPC, but, instead has a larger basis set than the previous one. This leads to have a MAD value of 6.3 kcal/mol. Accordingly, lower MAD values are obtained here, and the very low value comes from missing basis sets and/or higher order correlations error compensation somehow.

- The seventh entry (VII), ΔSCS-MP2/def2-TZVPP, has the same basis set of def2-TZVPP with the sixth entry, but it also has a SCS (spin-component-scaling) MP2 basis this time. SCS is introduced with an initial aim to remove the double counting of correlation effects that leads to an overestimation of dispersion effects by MP2. (To make a remark here, this effect is being removed via higher order correlations in our references.) This results in a MAD value of 1.5kcal/mol.

- As for the next, eighth entry (VIII), ΔSCS-MP2/CBS, keeping the method same as the previous one but this time having a "complete basis set" (CBS), MAD value of 1.2kcal/mol is obtained. CBS is obtained in the same way for our reference method.

- As one of the suggested methods for the treatment of large systems [334], the ninth entry (IX) is ΔB2-PLYP-D3/def2-TZVPP; and where B2-PLYP-D3 is being a double hybrid density functional theory (DH-DFT) approach, has correlation effects from a MP2-like treatment on top of DFT orbitals in addition to the D3 dispersion correction contributions. This resulted in a MAD value of 3.5kcal/mol.

- Tenth entry (X) is the next trial of the previous entry, ΔB2-PLYP-D3/CBS, and this time B2-PLYP-D3 is being tested with CBS instead of def2-TZVPP. This leads to a MAD value of 1.2kcal/mol.

In addition to these findings from WFT comparisons, we also wanted to compare the performances of DFT and SQM methods. The following Table II.3-2 presents their numerical deviation results from the reference WFT method.

Mean deviation (MD), mean absolute deviation (MAD), root mean square deviation (RMSD) and error span (MIMA), (i.e. difference between minimum and maximum errors) are presented here as well as including Pearson (R) and Kendall (τ) values.

		Method	Basis Set	MD	MAD	RMSD	MIMA	R	τ
DFT	I	BP86	D2/def2-SVP	-10.14	10.14	10.55	11.55	1.00	0.91
		B3-LYP		-12.27	12.27	12.73	12.64	1.00	0.96
		B-LYP		-12.51	12.51	12.81	10.68	1.00	0.96
		PBE		-11.21	11.21	11.55	9.73	1.00	0.96
		B97-D		-9.36	9.36	9.64	8.22	1.00	0.96
		TPSS		-11.15	11.15	11.51	10.79	1.00	0.91
	II	BP86	D2/def2-TZVP	-3.70	3.70	4.15	6.95	1.00	0.96
		B3-LYP		-4.79	4.79	5.00	3.29	1.00	0.96
		B-LYP		-3.78	3.78	4.07	5.10	1.00	1.00
		PBE		-3.96	3.96	4.22	4.69	1.00	1.00
		B97-D		-2.33	2.33	2.72	4.81	1.00	1.00
		TPSS		-4.23	4.23	4.56	6.34	1.00	0.96
	III	BP86	D2/def2-TZVPP	-3.28	3.28	3.69	6.46	1.00	0.96
		B3-LYP		-4.30	4.30	4.47	3.42	1.00	0.96
		B-LYP		-3.27	3.27	3.59	4.63	1.00	1.00
		PBE		-3.58	3.58	3.82	4.61	1.00	1.00
		B97-D		-2.01	2.01	2.41	4.23	1.00	1.00
		TPSS		-3.94	3.94	4.26	5.95	1.00	1.00
	IV	BP86	D3/def2-TZVPP	-4.04	4.04	4.41	6.00	1.00	0.96
		B3-LYP		-3.58	3.58	3.81	3.30	1.00	1.00
		B-LYP		-3.04	3.04	3.37	4.14	1.00	1.00
		PBE		-2.64	2.64	2.90	3.46	1.00	1.00
		B97-D		-2.61	2.61	2.96	5.06	1.00	1.00
		TPSS		-1.99	1.99	2.43	4.17	1.00	1.00
	V	BP86	D3+D_{abc}/def2-TZVPP	-3.64	3.64	4.09	6.16	1.00	0.96
		B3-LYP		-3.18	3.18	3.45	3.94	1.00	1.00
		B-LYP		-2.64	2.64	3.08	5.98	1.00	1.00
		PBE		-2.24	2.24	2.63	5.21	1.00	1.00
		B97-D		-2.21	2.24	2.70	5.11	1.00	1.00
		TPSS		-1.59	1.71	2.21	5.67	1.00	1.00
	VI	M06-2X	def2-TZVP	-0.13	1.70	1.96	6.18	1.00	0.89
		M06		-0.04	1.65	2.04	6.81	1.00	0.83

***Table II.3-2**: Comparison of DFT methods with Pearson R and Kendall τ values correlated against WFT reference method values, as well as with error statistics MD, MAD, RMSD and MIMA.*

	Method	MD	MAD	RMSD	MIMA	R	τ
SQM	**PM6-DH+**	-4.24	4.98	6.41	18.80	1.00	0.87
	PM6-DH+ (+ΔD3)	-2.29	3.38	4.93	17.33	1.00	0.91
	PM6-DH+ (+ΔD3 +ΔD$_{abc}$)	-1.90	3.31	4.88	17.66	1.00	0.91

Table II.3-2 cont'd.: *Comparison of DFT methods with Pearson R and Kendall τ values correlated against WFT reference method values, as well as with error statistics MD, MAD, RMSD and MIMA.*

- The first section, I, with the basis set of D2/def2-SVP, has quite large MD, MAD, RMSD and MIMA values compared to the other parts due to small basis sets. Changing from section II, D2/def2-TZVP, to section III, D2/def2-TZVPP methods do improve the conditions but only makes a little impact on the error values. This also makes us assume the following: In case we want to change from TZVPP into QZVP one, maybe again a small effect will be observed, but it is also likely that the conclusions will not really go through a big change again.

- Looking at the dispersion correction arrangements, it can be observed through the sections IV, V and VI that, when we change D2/def2-TZVPP into D3/def2-TZVPP first, and then into D3+Dabc/def2-TZVPP, there is a gradual drop in the values once more, as an overall tendency in most cases. Then it can be concluded that, among all the variations that are listed above, TPSS/D3+Dabc/def2-TZVPP is the best one overall with the lowest MAD value in its group with 1.7kcal/mol.

  However for some of the cases here, it cannot be labelled as an improvement. For example unlike others, with the method BP86 and B97-D, when dispersion changes into D3 or into D3+D$_{abc}$, then there is an increase in MAD values (2.6 and 2.2kcal/mol respectively) instead when compared to its previous state with D2 dispersion (2.0kcal/mol).

- Section VI, having the M06-2X and M06 method with def2-TZVP the functionals, already describes the dispersion interactions without needing any correction terms for them. They have resultant MAD values around 1.7 kcal/mol which also looks promising. Only that with these methods, 2HDR has a convergence problem (which is most likely due to numerical instabilities [335]), therefore, since it cannot be treated with these methods properly, the relevant data for 2HDR was excluded from the statistical concerns.

- In the SQM part of the table, PM6-DH+ methods and its variations are given. There are two additional variations: with "+ΔD3" and with "+ΔD3+ΔD$_{abc}$". PM6-DH+ (+ΔD3) means that dispersion is changed from D2 to D3 for PM6-DH+ method, and it shows an improvement with MAD values changing from 4.98kcal/mol into 3.38kcal/mol. Likewise, PM6-DH+ (+ΔD3+ΔD$_{abc}$) is the part showing the effect of changing from D2 to D$_{abc}$ (first D2 changing into D3, and then D3 changing into D$_{abc}$). However for this latter case, only a small change is observed with the MAD values (from 3.38kcal/mol to 3.31kcal/mol).

At this point, if we recall our first research stage (3.1.1) again, we were having comparisons in between our reference DFT (BP86) and SQM (PM6-DH+) methods in a detailed manner. Here with this research stage 3.1.2, we have another opportunity for a similar comparison in between SQM and DFT methods once again.

BP86-D3+D_{abc}/def2-TZVPP has a MAD value of 3.64kcal/mol whereas PM6-DH+ (+ΔD3+ΔD_{abc}) has a MAD value of 3.31kcal/mol. Hence, this again shows that PM6-DH+ has a similar (even a slightly better) accuracy, while being roughly three times computationally faster than the DFT methods.

We frequently mention that SQM methods are faster, but in order to show how fast they are, the performance of our reference SQM method is presented in terms of its computational time as well. Table II.3-3 shows a comparison of the average computing time needed for calculating one "interaction energy" in the PLI10 set for different methods. Data will be given in per CPU core, therefore, to find the measured CPU time, this data needs to be divided by number of cores if the calculation was done in a parallel way.

	Method	Basis Set	Core seconds[a]
SQM	PM6-DH+	-	3.2
DFT	HF	def2-SVP	4550.2
	GGA: RI-PBE-D	def2-SVP	756.0
	m-GGA: RI-TPSS-D	def2-SVP	913.1
	Hybrid: B3LYP-D	def2-SVP	4165.9 (1.2 hours)
	HF	def2-TZVP	12718.6
	GGA: RI-PBE-D	def2-TZVP	1179.4
	m-GGA: RI-TPSS-D	def2-TZVP	2299.5
	Hybrid: B3LYP-D	def2-TZVP	19861.4 (5.5 hours)
	HF	def2-TZVPP	71611.5
	GGA: RI-PBE-D	def2-TZVPP	4562.5
	m-GGA: RI-TPSS-D	def2-TZVPP	6935.3
	Hybrid: B3LYP-D	def2-TZVPP	73862.8 (20.5 hours)
WFT	Reference method	CBS	3681085.4 (42.6 days)

Table II.3-3: *Comparison of SQM, DFT and WFT methods based on the computation time they took for completions. Numerical values show the average computation time needed for calculating only one interaction energy in our PLI10 data set. Values are given in core seconds, i.e. adjusted for the number of CPU cores used.*

[a]: Serial SQM and DFT calculations on Intel Core i7-3770, 3.40 GHz, 8 cores, 8MB cache and 16GB RAM; parallel WFT calculations on Intel Xeon, 2.67 GHz, 8 cores, 12MB cache and 65GB RAM.

Looking at the reported values, SQM, PM6-DH+ requires seconds to accomplish the task, while DFT methods can take time up to a day, and the WFT reference needs approximately 40days for the same task.

Overall, the following results can be highlighted from this part of our research:

- Based on Table II.3-1, **SCS-MP2/CBS** (Entry VIII) and **B2-PLYP-D3/CBS** (Entry X) both give the lowest MAD values as 1.2 kcal/mol overall. Therefore, in case of necessity, these can be regarded as a recommendation for more costly computations.
- Again referring to the Table II.3-1, it is observed that **MP2/def2-TZVPP** (Entry VI) has a MAD value around 6kcal/mol, therefore, this cannot be recommended at all at least for the scoring purposes, because there are already other methods available which are more accurate and faster than this one (even some other DFT methods and even SQM methods).
- Based on Table II.3-2, when it is with dispersion corrections and triple-ζ (TZ) basis set types in general, DFT methods give MAD values of around or below 4 kcal/mol. Therefore, the best DFT approach from this list seems to be **TPSS-D3+D_{abc}/def2-TZVPP** with a MAD value of 1.71kcal/mol (section V). Moreover, the others which look promising are: **M06 and M06-2X** with MAD values of around 1.65 kcal/mol and 1.70kcal/mol respectively. Also, amongst the D2 related group, the **B97-D2/def2-TZVPP** can be considered with a low MAD value of 2.01kcal/mol.
- Referring to the Table II.3-2 once again, it is observed that SQM-DH usually give MAD values of around 3-5 kcal/mol, but, once more to mention, unlike the other methods, SQM-DH methods are about three orders of magnitude faster. This becomes a real advantage and makes SQM-DH methods profitable and preferable among others for most of the cases (i.e. especially for instance for the modelling of large systems).

II.4 Conclusion

With PLI10 set which is derived from the PDBbind data set, Wave Function Theory (WFT), Density Functional Theory (DFT) and semiempirical quantum mechanical (SQM) methods were benchmarked against eachother and against experimental references. The WFT methods studied here are the state-of-art theoretical reference data for the benchmarking of lower-level methods.

SCS-MP2 and B2-PLYP-D3 were found to be the most efficient WFT methods, whereas TPSS-D3+D_{abc}/def2-TZVPP could be assigned as the best DFT approach.

Overall, our SQM reference method, PM6-DH+, was found to be a fast and a surprisingly accurate alternative to full *ab initio* treatments in comparison to the theoretical reference values.

After these findings, with our next Research Stage III (Section 3.1.3), the transferability of these performance results into experimental results are investigated.

3.1.3 Research Stage III

III.1 Introduction

As it is mentioned in the previous sections, still some improvement is needed for scoring functions. This is mainly referred to be a scoring problem rather than a docking problem. The performance of the software programs are regarded well enough for docking [206, 253, 200, 202, 207], but on the other hand, even if they include physically accurate terms or involve quantum-mechanical nature of the interactions [256, 88, 254, 255], scoring functions do not provide the required scoring performance.

There is potential for improvements, for instance, about polarization and solvation treatments and entropy related terms [206, 207]. Any improvement in one of them can also help advancing the other term as well. For example, obtaining a better polarization and solvation effect, can also be an important contribution for the correct estimation of entropy values too [120, 129, 89]. Since any development at one side might affect the development of other terms as well, then, it can be concluded that, any interaction in between them better be investigated as well. In addition, while doing so, the compensations in between the main interactions have to be taken into account as well (like in between entropy and enthalpy to speak of). Hence, a careful approach is needed to identify the reasons of the complexity of these biomolecular interactions [336].

For instance, adding chemical functionality is regarded as a good strategy to increase "enthalpic" binding affinity contributions, however, these type of contributions are cancelled out by their "entropic" counterparts, and as a result, optimization of the drug candidates becomes very difficult to handle [337].

Concerning these, there are some studies focusing on the compensation relations using the publicly available protein-ligand complex sets [260, 261, 259, 257] but, unfortunately, the relation in between these interactions or compensations are not investigated well enough.

Within this Section 3.1.3, Research Stage III, we will try to sort out the terms and the interactions in between them. Therefore, this part of our research will be less of an assessment of the various computational methods we have tested so far, and it will be more about analysing the existing connections, rather than presenting a new method or a finding. This topic is quite complicated and even though we cannot either solve the problems or see the full picture here properly, we hope that our attempts might lead to a useful contribution.

Main focus will be on the SQM methods once again, but QM and MM methods will be also included to our tests. Once we will arrive at some general analysis, we also would like to test our implications on the design of scoring functions.

III.2 Delicate balance of biomolecular interactions

Figure III.2-1 below illustrates our approach for these interactions and their balances/compensations.

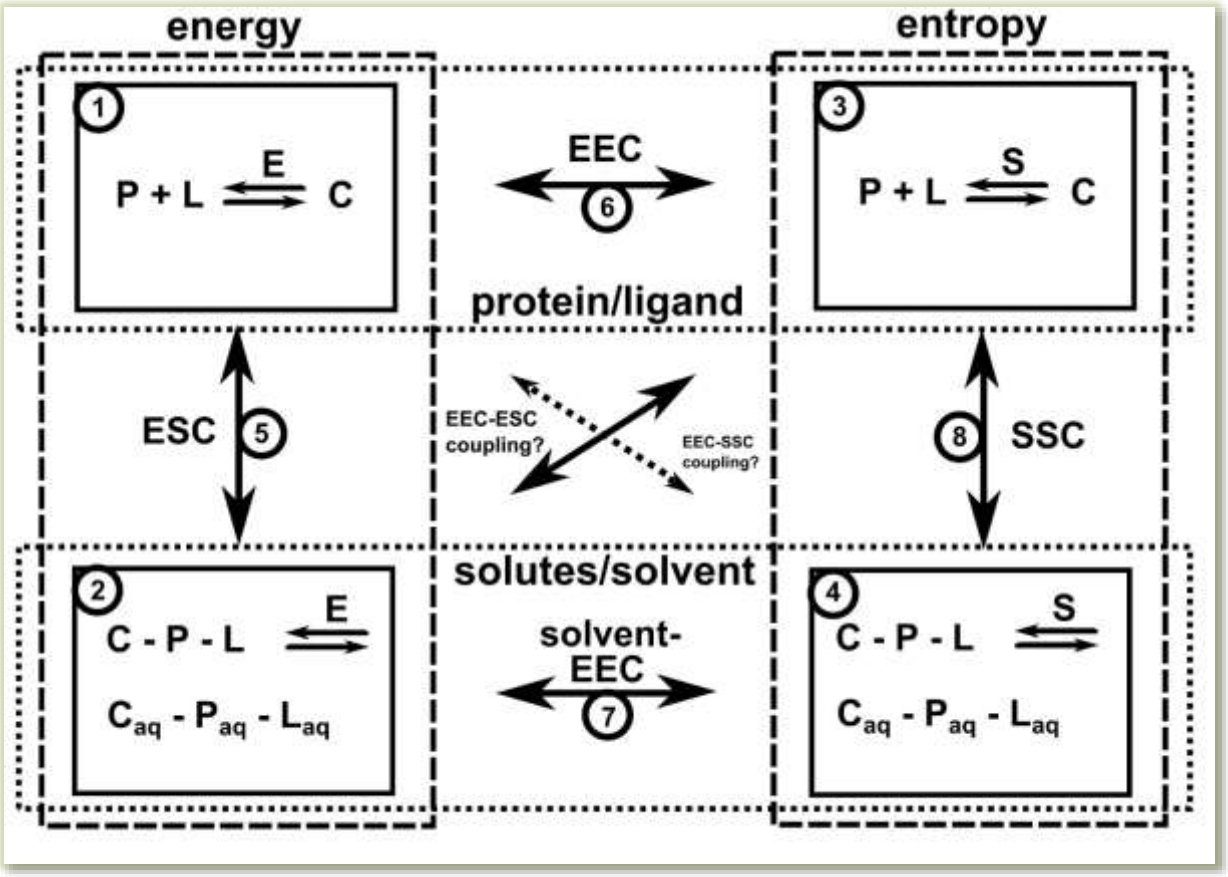


Figure III.2-1: *The representation of the delicate balance of biomolecular interactions.*
Part 1: energetic protein-ligand interactions,
Part 2: energetic solute-solvent interactions,
Part 3: entropic protein-ligand interactions,
Part 4: entropic solute-solvent interactions,
Part 5: energy-solvation compensation (ESC),
Part 6: energy-entropy compensation (EEC),
Part 7: energy-entropy compensation: EEC with solvent: (solvent EEC),
Part 8: entropy-solvation compensation (SSC).

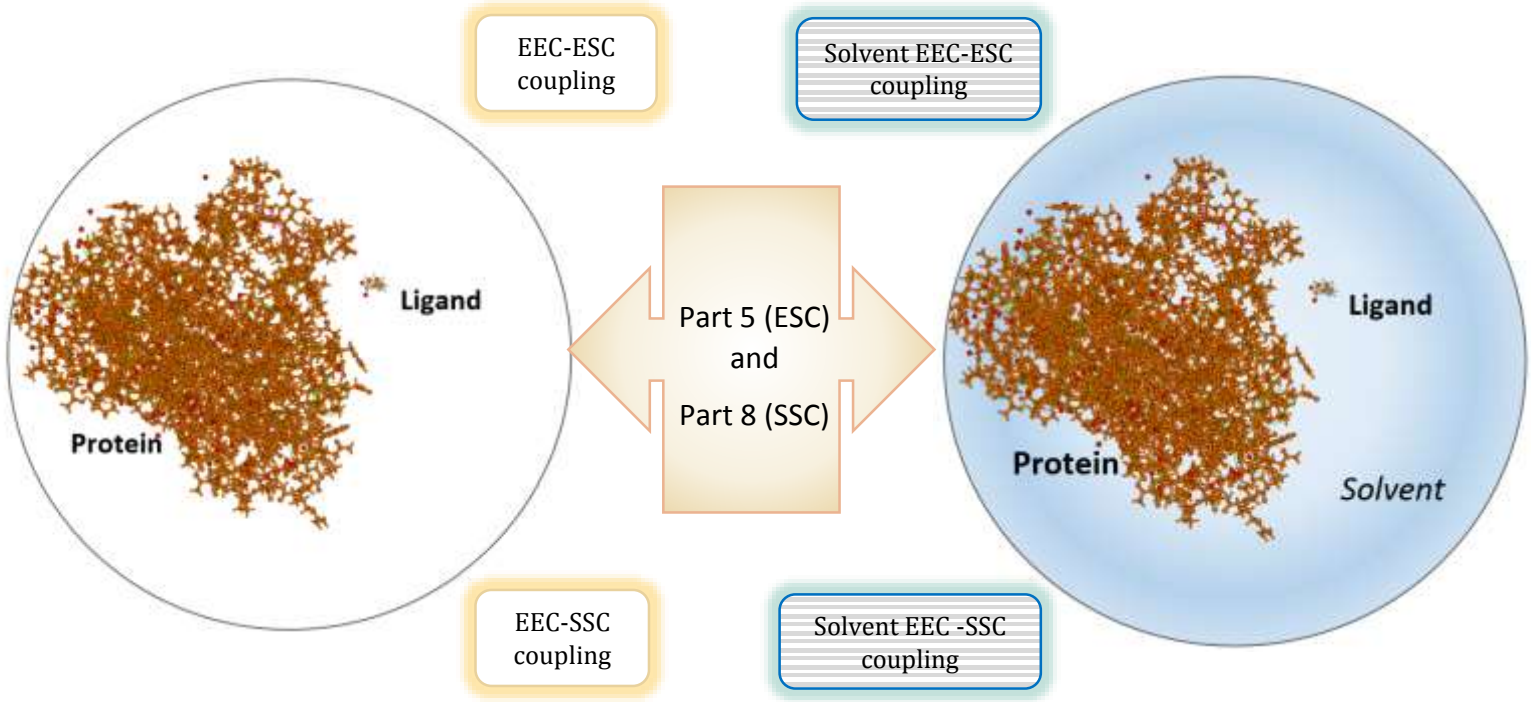
As an additional illustration,

For parts 1, 3 and 6 (EEC):

Only the interactions in between proteins and ligands, energy and entropy matters.

For parts 2, 4 and 7 (solvent EEC):

The interactions of solutes with solvent, energy and entropy matters.



Scheme III.2-1: *Schematic representation of balances in Figure III.2-1.*

- Part 1 demonstrates the energetic interactions only between the protein and ligands.
- Part 2 is about the energetic interactions between the protein-ligand interactions in solutes (complex, protein and ligand) and the solvent.
- Part 3 is about the entropic contributions to protein ligand binding.
- Part 4 demonstrates the entropic contributions from solvation.
- Part 5, denoted as **ESC** (**E**nergy**S**olvation**C**ompensation), is for the compensation of energetic protein-ligand polar interactions with solvation effects.

  Polar interactions between the protein and ligand molecules are mentioned to be partly cancelled [338] due to the possibility of similar polar interactions of the dissociated protein and ligand with water molecules.
- Part 6, denoted as **EEC** (**E**nergy**E**ntropy**C**ompensation), is another compensation in between the enthalpic protein-ligand and entropic protein-ligand interactions without solvation concern, together with and **EEC-ESC coupling** and with **EEC-SSC coupling**.
- Similar approach goes for Part 7, which is denoted as **solvent EEC** (solvent-**E**nergy**E**ntropy**C**ompensation). This time, the compensation is considered for the energetic and entropic part of the solvation.

  For Parts 6 and 7, it can be stated that, a stronger energetic binding results in an increased restriction of molecular flexibility.
- Finally Part 8, denoted as **SSC** (Entropy**S**olvation**C**ompensation), is about the compensation of the entropic protein-ligand interactions and entropic solvation effects.

  As an example, the number of water molecules freed from the solutes' surface can compensate the interaction between the protein and the ligand when there is a strong restriction of a conformational entropy.

Enthalpy-entropy compensation is suggested as a property of non-covalent interactions [339-346], and that a more dominant enthalpic interaction results in a strongly opposing entropic contribution since the conformational freedom is limited by the strong enthalpic effect. There are some literature sources where the enthalpy-entropy compensation (EEC) is being questioned [347], and there are some studies which report non-compensation [348, 349], but so far there is no conclusive theory about EEC [336].

There are some contributing works over host-guest systems or protein-ligand interactions, while the enthalpy-entropy compensation (EEC) effect was being investigated [214, 350, 351, 129, 352]. Korth showed that [338] energetic interactions are not uniformly cancelled by entropic effects, but also that, energetic interactions are dependent on their electronic nature.

Figure III.2-1 tries to separate the intermolecular balance into pieces for an analysis. It includes main parts as well as coupling terms in between them. Part 5 and 8 are the compensation relations, and they already couple the main basic interactions (part 1, 2, 3 and 4), but they might not be independent of eachother.

Before proceeding further, it should be mentioned that the entropic solvent effects could not be calculated due to the technical reasons. In order to study entropic-solvent effects, Molecular Dynamics Simulation are needed. COSMO-RS [285] model has a clever way of guessing these dynamics. However, at the time of our research COSMO-RS was also not available. Hence, due to the unavailabilities, Parts 4 and its relevant connections, Part 7, Part 8, EEC-SSC coupling, solvent EEC-ESC coupling and solvent EEC-SSC coupling were not able to be calculated. Therefore, our study will represent a better picture mainly for Part 1, 2, 3, 5 and 6.

Part 1 and part 2 of Figure III.2-1 include the intermolecular interactions, which is a commonly studied field within the "theory of intermolecular interactions" [212, 353, 258]. This theory divides the energetic (non-covalent) interactions as [354]:

- First order exchange repulsion and electrostatic interactions, and,
- Second order polarization/charge transfer (induction) and dispersion interactions.

And the intermolecular binding is expressed to be a result of a balance of these two mentioned interactions, together with:

- large exchange-repulsion interactions,
- large attractive electrostatic interactions,
- smaller attractive contributions from induction and dispersion,

and that, these add up to a small overall binding effect.

Some interactions, like hydrogen-bonding, π-stacking or the hydrophobic effect, which are called non-basic type of interactions are regarded valuable in the sense of understanding concepts on a higher level. These interactions can be associated with the main balances.

III.3 Computational Details and Selected Data Sets

The number of data sets that we are working on are increased. The following basis sets are used:

- "S22" and "S66" : small biomolecular model basis benchmark sets [331, 332],

  S22 and S66 were selected due to their balanced representation of polar and nonpolar contributions for the biomolecular interactions, as well as providing high-level QM geometries and interaction energies. These sets contain up to around 30 atoms [355].

- "PLI10" set from our previous Research Stage II (Section 3.1.2) is used. It consists of the smallest 10 complexes from the 3.0 Å pocket models that is derived from PDBbind 2007 [356]. This set contains up to about 100 atoms.

- Complexes from the refined PDBbind 2007 set and from PDBbind 2009 set [260, 261] are included. These sets are taken from the references mentioned. Overall, 1297 complexes from the refined set of PDBbind 2007 and 580 complexes from the PDBbind 2009 set are used.

For the computations, MM, SQM and QM (DFT) methods are considered. The details are indicated in the Table III.3-1as follows:

Computational Methods		
Method Type	**Name**	**Computational Details**
MM (FF)	MMFF94 [270]	Calculations are done with Tinker 6.2 [357], and solvation effects are treated with GBSA + RRHO approximation [212]
QM (DFT)	PBE-D2/TZVPP [277, 279, 282]	Geometry optimization and frequency calculations are done with TURBOMOLE 6.4 [280, 281] Using: D2 dispersion corrections [279] the RI approximation for two-electron integrals [283,284] and TZVPP AO basis sets [282] Solvation effects are treated with COSMO [285] + RRHO approximation [212]
SQM	PM6-DH+ [50]	Calculations are done with MOPAC2012 [58] , having D, dispersion and H+, hydrogen-bond corrections [50, 53] by making use of COSMO solvation models [285] + RRHO approximation [212]

Table III.3-1: *List of computational methods used for Research Stage III*

Some additional explanations can be given as follows:

- MMFF94 is chosen since it can be easily applied to all benchmark sets. The performance of MMFF94 is found very similar to AMBER [357].
- PM6-DH+ is again preferred as our reference SQM method.
- As the QM (DFT) approach, PBE-D2/TZVPP is the choice because of its very good accuracy reported for the benchmark sets [331].
- Thermodynamical data is calculated at 298.15 K based on the rigid-rotor harmonic-oscillator (RRHO) approximation [212]

  As it is previously mentioned in Section 2.2, this approximation,

  - Treats molecules as "rigid",
    - This approximation works well for small molecules, but real systems are dynamic. According to Gilson [212], with this approximation, the treated host-guest systems have results which are close to the experimental values.
  - The conformational entropy is missing in this approximation.
- For (QM) DFT and SQM methods solvation effects are treated with COSMO [285]. (COSMO-RS could have been more advantageous, but at the time of our research, program was not available.)

III.4 Results

- Only energetic effects are presented in Table III.4-1 and Figure III.4-1.
- Solvation effects (without any concern of the enthalpy or entropy) are presented in Table III.4-2, Table III.4-3 and Figure III.4-2.
- Then the results, which involve both enthalpic and entropic corrections, but on the other hand exclude the solvation effects, are given in the Table III.4-4, Table III.4-5, Table III.4-6 and in Figure III.4-3.
- Ssolvation, enthalpic and entropic effects are considered altogether and these are investigated in Table III.4-7, Table III.4-8, Table III.4-9 and Figure III.4-4.
- About these tables and figures:
  - As it is mentioned above, due to their system sizes, S22 and S66 sets are available for all methods: QM, SQM and MM. However PLI10, which was actually a derivation from the PDBbind 2007 that is used in our previous Research Stage 3.1.2, includes large protein and ligand models. Thus, PLI10 was mainly only possible for SQM method, and sometimes also for demanding QM calculations.

- Kendall tau (τ) values are not used during this Research Stage III (Section 3.1.3). Pearson R value is explanatory enough to see the main tendencies in between the relations of interaction terms.

Energetic Effects

S22 and S66 sets are compared with three methods, DFT, SQM and MM, for the optimization interaction energies. Table III.4-1 gives the errors statistics, including mean deviation (MD), mean absolute deviation (MAD), root mean square deviation (RMSD) and the maximum error span (MIMA) values, with respect to the high-level (extrapolated CCSD(T)/CBS) reference data [331, 332].

These are the results without any enthalpic, entropic or solvation effects. Results seem to be in a very good correlation with the high level reference data, as values being lower than 1kcal/mol especially for SQM and QM, and also up to 2kcal/mol for MM. Low MAD values, show that a method is fine and satisfying enough for an interaction energy calculation. Since error values are similar to eachother, then, due to the computational speed, SQM, PM6-DH+ method seems to be the most favourable one among these.

Data set	Method	MD	MAD	RMSD	MIMA
S22[d]	**QM[a]**	-0.89	0.90	1.65	5.16
	SQM	-0.23	0.40	0.60	2.17
	MM[b]	1.69	1.70	3.14	10.57
S66[e]	**QM[a]**	-0.43	0.43	1.09	5.78
	SQM	0.30	0.37	0.53	2.84
	MM[c]	2.02	2.02	2.48	7.53

Table III.4-1: *Error statistics for the comparison of optimized interaction energies at QM, SQM and MM methods with respect to the high-level (extrapolated CCSD(T)/CBS) reference data [331,332] presenting MD (mean deviation), MAD (mean absolute deviation), RMSD (root mean square deviation) and MIMA (maximum error span) values.*
[a]: data from reference [338]
[b]: Excluding entry 7 and 14 because of missing parameters.
[c]: Entries 55 and 56 are excluded because of missing FF parameters

Figure III.4-1 is directly linked with Table III.4-1. Here, the interaction energies are represented rather than the MAD error values (which are the differences in between actual values). In order to see the MAD values in Table III.4-1, the differences in between the data points belonging to the same entry in Figure III.4-1, have to be analysed.

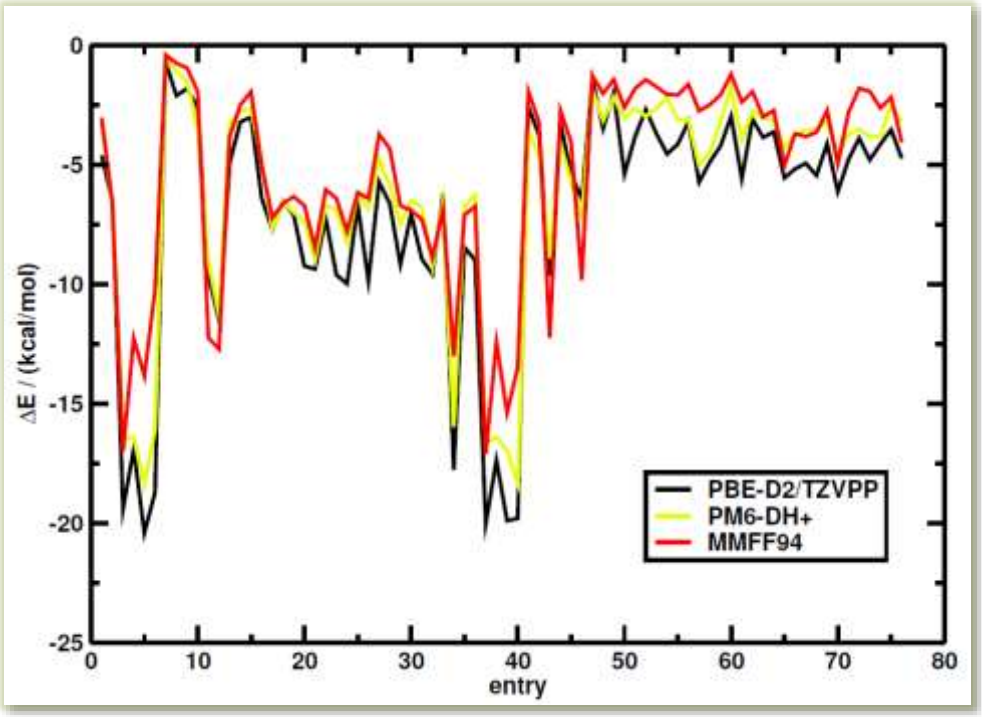


Figure III.4-1: *Both for S22 and S66, graphical representation of the optimized interaction energies (ΔE) for the QM, SQM and MM methods based on Table III.4-1.*
MM methods are excluding entry 7 and 14 because of missing parameters.
MM methods are excluding entry 55 and 56 because of missing FF parameters
QM data from reference [338]

Solvation Effects

The correlation between the solvation interaction energy (ΔE_{solv}) and the polar terms of interaction energy, ΔE, is presented in terms of Pearson, R values in Table III.4-2. All of the sets: S22, S66 and PLI10, are studied. Due to the system size, PLI10 set was only available for SQM and for some demanding QM methods.

Data Set	Pearson (R) correlation for QM	Pearson (R) correlation for SQM	Pearson (R) correlation for MM
S22	-0.93	-0.93	-0.96[a]
S66	-0.83	-0.89	-0.54[b]
PLI10	-0.99	-0.99	-

Table III.4-2: *Pearson R values for the correlation in between the solvation interaction energy contributions ΔE_{solv} and the polar terms of ΔE at QM, SQM and MM levels.*
[a]: Excluding entry 7 and 14 because of missing parameters.
[b]: Entries 55 and 56 are excluded because of missing FF parameters

- It is observed that, there is a high but a negative correlation in between the solvation interaction energy contributions ΔE_{solv}, and the polar part of ΔE for all methods (which is ranging in between -0.83 and -0.99). This finding is in a good agreement with the common knowledge, which states that the favourable polar interactions between proteins and ligands are largely being cancelled by similarly favourable polar interactions of the opposite sign between these two species and water molecules.
- All methods have similar results in general, only that MM method with the data set, S66, is being a little bit exceptional. These similar values also indicate that the Energy Solvation Compensation, ESC, is being reproduced similarly for these methods too (part 5 in Figure III.2-1 & Scheme III.2-1).

  However, this similarity does not mean that solvation effects are "same" for these methods. Table III.4-3 is constructed to have a close-up analysis with error statistics to investigate this matter further.

In Table III.4-3, solvation interaction energies for the SQM and MM results compared with the QM level. Mean deviation (MD), mean absolute deviation (MAD), root mean square deviation (RMSD) and the maximum error span (MIMA) values are given as well as Pearson R values.

		SQM				MM				SQM	MM
		MD	**MAD**	**RMSD**	**MIMA**	**MD**	**MAD**	**RMSD**	**MIMA**	**R**	**R**
S22[a]	**ΔE**	-2.20	2.73	4.13	11.40	-0.61	1.42	4.05	29.4	0.98	0.97
	ΔE_{solv}	1.06	1.07	1.91	7.29	1.06	3.01	3.81	14.73	0.94	0.68
S66[b]	**ΔE**	-1.77	2.63	3.74	14.37	-0.61	1.42	4.05	29.48	0.91	0.77
	ΔE_{solv}	0.51	0.75	1.10	4.44	-0.05	1.88	2.46	14.96	0.95	0.75
PLI10	**ΔE**	-13.62	13.79	16.18	30.90	-	-	-	-	0.99	-
	ΔE_{solv}	1.20	2.71	3.22	10.22	-	-	-	-	1.00	-

Table III.4-3: *Error statistics for SQM and MM methods, compared with QM data for the optimized solvation energy contributions, ΔE_{solv}, with mean deviation MD, mean absolute deviation MAD, root mean square deviation RMSD and Maximum error span MIMA values, in addition to the Pearson R values for the correlations.*
[a]: MM methods are excluding entry 7 and 14 because of missing parameters.
[b]: MM methods are excluding entry 55 and 56 because of missing FF parameters

- It is observed that only for MM level, when the solvation effects are included, the correlation with QM is reduced (especially for the S22 set). Looking at the Pearson, R values, the decline can be seen when the ΔE_{solv} is added.
- For S22 and S66 sets, SQM has a MAD value around 2.5 kcal/mol for the overall energy (ΔE), and around 1 kcal/mol for the solvation contribution (ΔE_{solv}). For MM, meanwhile, MAD is about 1.5 kcal/mol for the overall energy (ΔE), and 2 to 3 kcal/mol for the solvation contribution (ΔE_{solv}).
- As for PLI10, since it was only available to calculate with SQM and QM methods, only the SQM results can be presented here as their correlation with QM. PLI10 set has MAD values of about 14 kcal/mol for the overall interaction energy (ΔE), and 3 kcal/mol for the solvation contribution (ΔE_{solv}). These values for PLI10 are much larger than the ones for S22 and S66 sets, where S22 and S66 sets are comparably smaller-sized sets.
- On the other hand, when it is about the correlation with QM methods, then PLI10 set has a higher correlation compared to the S22 and S66 sets.

Figure III.4-2 below is linked with Table III.4-3, and it also shows the solvation with QM, SQM and MM levels. Looking at the differences in between peaks (interaction energies), it is clear that for ΔE_{solv} contribution, MM behaves a bit different than others with its distinctive peaks.

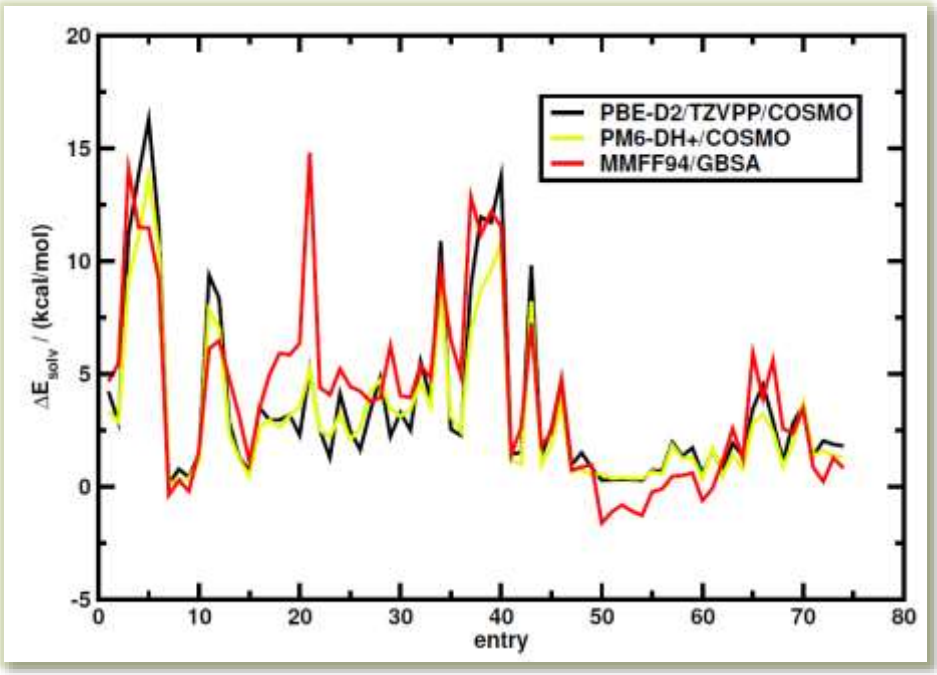


Figure III.4-2: *Both for S22 and S66, solvation interaction energies (ΔE_{solv}), at QM, SQM and MM level.*
MM methods are excluding entry 7 and 14 because of missing parameters.
MM methods are excluding entry 55 and 56 because of missing FF parameters

Enthalpic and Entropic Effects (without Solvation)

As it is given in the previous Section 2.2, keeping the Gibbs free energy (Free Interaction Energy) equation in mind:

$$\Delta G = \Delta H - T\Delta S \qquad \text{(Eqn. 2.2-2)}$$

ΔG: Change in free energy
ΔH: Change in Enthalpy
ΔS: Change in Entropy

The relation between enthalpy and interaction energy within the RHHO approximation and contribution from volume change has the following:

$$\Delta \mathrm{H} = \Delta \mathrm{E} + \Delta \mathrm{H}^{\mathrm{RHHO}}_{0 \text{ to } 298\mathrm{K}} \qquad \text{(Eqn. 3.1.3-1)}$$

$$\Delta E = \Delta E_{SQM} + \Delta E_{solvation} \qquad \text{(Eqn. 3.1.3-2)}$$

All results in:

$$\Delta G = \Delta E_{SQM} + \Delta E_{solvation} + \Delta H^{RRHO}_{0-298} - T\Delta S^{RRHO}_{298} \qquad \text{(Eqn. 3.1.3-3)}$$

Table III.4-4 compares various compensation relations for enthalpy and entropy values. Comparisons in between non-optimized interaction energies, ΔE_o, optimized interaction energies ΔE, interaction enthalpies ΔH and free interaction energies, ΔG, calculated with the QM, SQM and MM methods, take place. Correlations are given by Pearson correlation parameters, R.

		I	II	III	IV	V	VI
		ΔE_o vs ΔE	**ΔE vs ΔH**	**ΔE vs −TΔS**	**ΔE vs ΔG**	**ΔH vs -TΔS**	**ΔH vs ΔG**
S22[d]	**QM[a]**	1.00	1.00	-0.87	0.95	-0.86	0.95
	SQM	1.00	0.97	-0.79	0.71	-0.66	0.85
	MM[b]	0.87	1.00	-0.97	0.67	-0.96	0.70
S66[e]	**QM[a]**	1.00	1.00	-0.59	0.91	-0.59	0.91
	SQM	0.99	0.98	-0.44	0.80	-0.36	0.87
	MM[c]	0.92	1.00	-0.76	0.75	-0.75	0.75

Table III.4-4: *Pearson correlation coefficients R, for the comparison of non-optimized interaction energies ΔE_o, and optimized interaction energies ΔE, interaction enthalpies ΔH, entropic contributions −TΔS and free interaction energies ΔG for QM, SQM and MM methods.*
[a] data from reference [338]
[b] Excluding entry 7 and 14 because of missing parameters
[c]: 55 and 56 are excluded because of missing FF parameters
[d]: 3 linear structures are excluded for S22
[e]: 7 linear structures are excluded for S66

- Entry I mainly has high correlation values in between non-optimized (ΔE_o) and optimized interaction energies (ΔE). Within all results, the highest values are with SQM and QM level methods (QM and SQM has R=1.00 for S22 set, and both of them have 1.00 and 0.99 respectively for S66).

  This finding also indicates that optimization causes almost no change for QM, SQM and MM methods, especially for these tested systems. If to compare between each other, only MM seems behaves a little bit different than the others with a lower R=0.87 for the S22 set, and R=0.92 for the S66 set. Overall, it can be mentioned that, optimizing the interaction energy, comparably have more effect on MM method, and specifically more for the S22 set.

- Entry II shows high correlation values between the optimized interaction energies (ΔE) and enthalpies (ΔH) for all QM, SQM and MM methods (with R values ranging in between 0.97 and 1.00 values).

  For comparison, SQM is the one with a minor numerical deviation here (having R=0.97 or R=0.98) from the "perfect" correlation, whereas, QM and MM have that perfect correlation (R=1.00). Overall, it can be reported that enthalpic corrections have only a little effect on the ranking for all these methods.

- Entry III is about the correlation between the optimized interaction energy (ΔE) and entropic contributions ($-T\Delta S$), and it brings about huge differences. Almost all correlations are low, and even lower for the S66 set than the S22 set.

  The highest correlations amongst them are with MM methods for both S22 and S66 with highest R values. However these high values for MM (even higher values than QM), indicates that, it is because MM might overestimate the compensation.

- Entry IV shows the relation between the optimized interaction energy (ΔE) and the free interaction energy (ΔG).

  SQM and MM results are still far from a good correlation. MM has lower correlation values for both S22 and S66 sets, and it is observed that, SQM is not close to the QM results when entropic effects are involved.

  Therefore, regarding both to Entry III and Entry IV, results indicate that entropy is a special concern to keep in mind, and based on these results, its contribution leads to correlations with lower values.

- Entries V and VI are similar to the Entries III and IV, only this time, instead of optimized interaction energies (ΔE), the enthalpy (ΔH) is the being compared with entropic effects ($-T\Delta S$) and free energies (ΔG). Similar results are observed.

  Previously, Korth showed that [338],

  - non-covalent interactions are compensated by entropic effects not in equal terms but instead by the differences based on their electronic natures, and,

- much higher compensating effects for non-polar interactions (dispersion) are found when compared with the compensating effects for polar (hydrogen-bond) interactions.

Therefore, due to these previous findings, the correlation between the ΔE and –TΔS values as in Table III.4-4 are expected to be different for dispersion or hydrogen-bond dominant systems.

In order to analyse this consideration further, Table III.4-5 is constructed. Here, the S22 and S66 data sets are grouped up in few different categories based on the main characteristic groups of the sets, and so that, EEC, (part 6 in Figure III.2-1) is once more investigated with different perspectives.

In total, three main type of sets are considered and grouped. Aim here is to compare these with the original full set which they are derived from. The newly grouped sets are called: "H set" and "D set" and they are prepared for S22 and S66. The further explanations are given below:

Full set : the full version of the set (either S22 or S66).

H set : dominantly hydrogen-bonded subsets (of either S22 or S66).

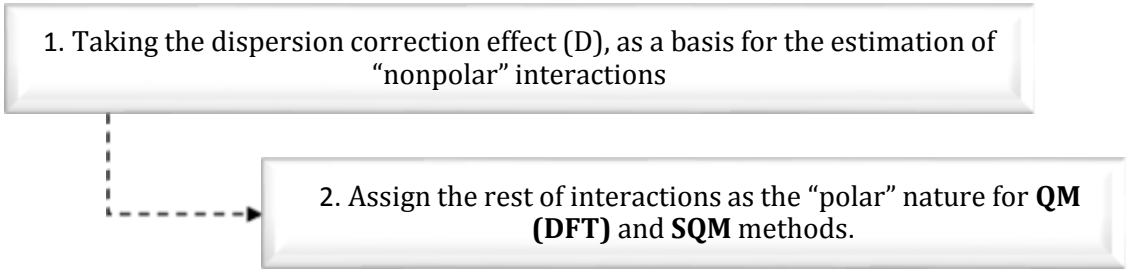
D set : dominantly dispersion-bound subsets (of either S22 or S66).

It is observed that, classifying the characteristics and then grouping the similar ones within the subsets enables a more detailed analysis, and sometimes yields to a higher correlation depending on these specific group features.

When decomposing the interaction energies, the following approach is used:

For QM and SQM methods:

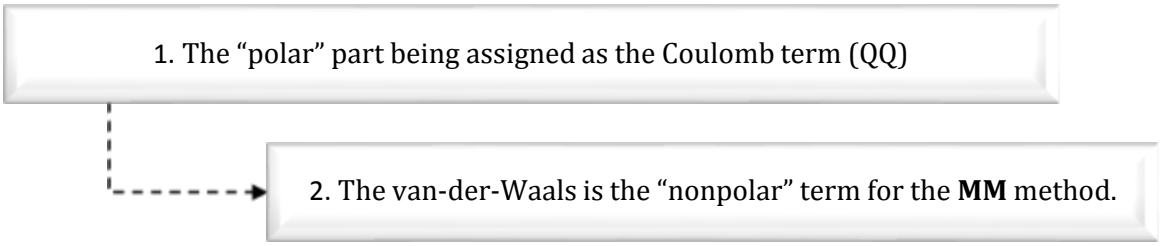
First dispersion correction is taken as the basis for "nonpolar" interaction estimation, and then, the rest is assigned as the "polar" part for these methods.



Scheme III.4-1: *Approach for SQM and QM methods*

For MM methods:

Coulomb terms are assigned as "polar" parts and Van der Waals (vdW) is assigned as the "nonpolar" part.



Scheme III.4-2: *Approach for MM methods*

Basically with this approach, when we look at the "repulsive van der Waals interaction (Pauli repulsion)" terms, they are assigned differently by the methods. They are,

- in the "polar" part of QM(DFT) and SQM methods, but
- in the "nonpolar" part of the MM methods.

The way how MM is treated is more correct, however assignments are done in a different way for the QM and SQM methods. These difference in the approaches is based on the ease of the further applications that are considered. For example, we will later on need extrapolations of the energy terms for enthalpic/entropic and solvation contributions. Hence for QM and SQM methods, the assignments are convenient in this way.

		S22[a,c]			**S66[b,d]**		
		Full set[e]	**H set**	**D set**	**Full set[e]**	**H set**	**D set**
QM	**ΔE vs -TΔS**	-0.87	-0.98	-0.97	-0.59	-0.84	-0.75
	ΔH vs -TΔS	-0.86	-0.97	-0.97	-0.59	-0.84	-0.74
SQM	**ΔE vs -TΔS**	-0.79	-0.97	-0.98	-0.44	-0.76	-0.60
	ΔH vs -TΔS	-0.66	-0.93	-0.87	-0.36	-0.72	-0.65
MM	**ΔE vs -TΔS**	-0.97	-0.95	-0.98	-0.76	-0.84	-0.57
	ΔH vs -TΔS	-0.96	-0.95	-0.98	-0.75	-0.84	-0.56

Table III.4-5: *Pearson correlation coefficients, R, for the comparison of ΔH vs –TΔS and ΔE vs –TΔS values with QM, SQM and MM methods, for three specified sets of S22 and S66.*
[a] MM: excluding entry 7 and 14 because of missing parameters
[b]: 55 and 56 are excluded because of missing FF parameters
[c]: 3 linear structures are excluded for S22
[d]: 7 linear structures are excluded for S66
[e]: data from Table III.4-4

- Generally it is clear from the results that all methods show higher correlation values with their separated H sets and D sets compared to their full set versions.
- However, for the less diverse S22 set, the results show much higher correlations. It can be stated that, in contradiction with QM the methods, there is something different with the polar and non-polar interactions of the full set treatments of SQM and MM methods which leads them to have very high correlation values specifically with these subsets.
- In case of S66 set with the MM method, there is some decrease in the correlation with the D set compared to the full set version of it.

In the following section with Table III.4-6, non-optimized interaction energies ΔE_o, optimized interaction energies ΔE, interaction enthalpies ΔH and free interaction energies ΔG, are listed for SQM and MM methods based on their correlation to QM levels.

		SQM				MM				SQM	MM
		MD	**MAD**	**RMSD**	**MIMA**	**MD**	**MAD**	**RMSD**	**MIMA**	**R**	**R**
S22[a,c]	**ΔE_o**	-1.41	1.64	2.34	7.71	-4.51	4.51	5.98	14.30	0.99	0.91
	ΔE	-0.74	1.01	1.30	4.16	-1.63	2.09	3.08	11.10	0.99	0.92
	ΔH	-3.15	3.21	4.30	9.23	-1.75	2.20	3.24	7.15	0.97	0.92
	$-T\Delta S$	-3.79	3.79	3.94	3.98	2.36	2.36	2.83	4.51	0.93	0.94
	ΔG	-6.93	6.93	7.35	6.91	0.61	2.88	3.45	12.35	0.85	0.70
S66[b,d]	**ΔE_o**	-1.70	1.80	2.46	9.35	-4.10	4.10	4.81	10.48	0.97	0.90
	ΔE	-0.96	1.14	1.39	4.30	-1.65	1.88	2.25	9.71	0.96	0.94
	ΔH	-2.51	2.51	3.12	8.04	-1.68	1.88	2.30	9.83	0.96	0.94
	$-T\Delta S$	-3.49	3.49	3.71	4.96	2.13	2.25	2.62	9.99	0.73	0.77
	ΔG	-6.00	6.00	6.25	5.11	0.45	1.71	2.19	9.59	0.89	0.81

Table III.4-6: *Comparison of non-optimized interaction energies ΔE_o and optimized interaction energies ΔE, interaction enthalpies ΔH, entropic contributions $-T\Delta S$ and free interaction energies ΔG for SQM and MM level with the same magnitudes at QM level.*
[a] MM excluding entry 7 and 14 because of missing parameters
[b]: 55 and 56 are excluded because of missing FF parameters
[c]: 3 linear structures are excluded for S22
[d]: 7 linear structures are excluded for S66

- MAD values for SQM method are around 3kcal/mol for enthalpic and entropic contributions (for ΔH and –TΔS) for both the S22 and S66 sets. Comparably these group of values are lower for the MM method and they are around 2kcal/mol.
- On the other hand, when the correlation for enthalpic and entropic contributions with QM values is concerned, then for ΔG results, it looks like SQM is producing a better correlation with it (R=0.85 and R=0.89 for S22 and S66 sets respectively), whereas MM has a bit lower values (R=0.70 and 0.81 for the S22 and S66 sets respectively). Altogether, it looks like enthalpic/entropic effects somehow affect these methods adversely especially if ranking will be the matter to focus.

Figure III.4-3, is the graphical representation related to Table III.4-6. Once more to indicate, these are actual data values in the graph, and the tabulated MAD values are the differences in between these data points.

This time, a scaled and shifted SQM results are also given here (in blue) because of the systematic error profile as can be seen clearly in the Figure III.4-3.

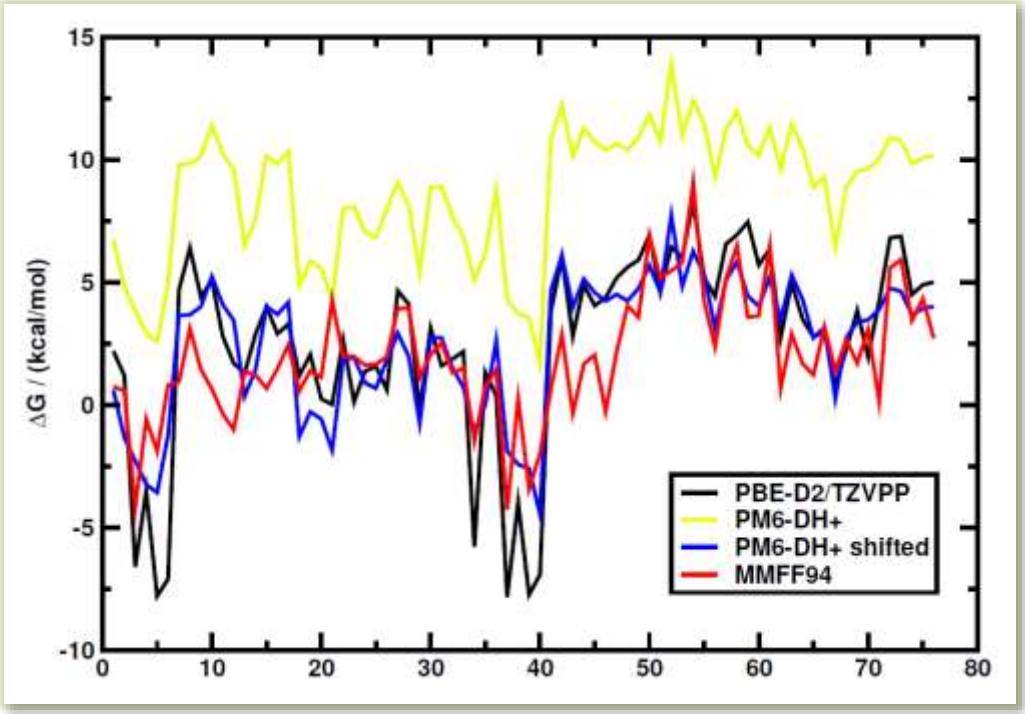


Figure III.4-3: *Both for S22 and S66, actual free interaction energies ΔG (without solvation energies), at QM, SQM, SQM shifted and MM level.*
MM methods are excluding entry 7 and 14 because of missing parameters.
MM methods are excluding entry 55 and 56 because of missing FF parameters
3 linear structures are excluded for S22
7 linear structures are excluded for S66

Energetic Effects (with solvation)

Optimized interaction energies that are compared with the high-level extrapolated CCSD(T)/CBS reference data [331,332], are given in the cases of either or not including the solvation effects. The previously obtained Table III.4-1 values (which are denoted by [x]) show the values "without solvation", and they are appended next to the newly obtained "with solvation" results to have an easy comparison.

		MD	MD[x]	MAD	MAD[x]	RMSD	RMSD[x]	MIMA	MIMA[x]
S22[e]	**QM[a]**	0.08	-0.89	0.15	0.90	0.20	1.65	0.62	5.16
	SQM[b]	0.14	-0.23	0.24	0.40	0.31	0.60	1.08	2.17
	MM[c]	1.65	1.69	1.66	1.70	2.41	3.14	7.30	10.57
S66[f]	**QM**	0.13	-0.43	0.21	0.43	0.29	1.09	1.18	5.78
	SQM	0.31	0.30	0.45	0.37	0.65	0.53	3.09	2.84
	MM[d]	2.22	2.02	2.22	2.02	2.63	2.48	8.11	7.53
PLI10	**SQM**	1.82	–	11.90	–	19.45	–	82.92	–

Table III.4-7: *Mean deviation , MD, mean absolute deviation, MAD, root mean square deviation, RMSD, maximum error span, MIMA values for QM, SQM and MM methods , all including solvation effects, which is compared with high-level (extrapolated CCSD(T)/CBS) reference data from references [331,332] for the original geometries which is excluding solvation effects (MD[x], MAD[x], RMSD[x], MIMA[x]).*

[a] Excluding entry 14 for technical reasons

[b] Excluding entry 10 for technical reasons

[c] Excluding entries 7 and 14 because of missing parameters and, value 12 for technical reasons

[d] Excluding entries 55 and 66 for technical reasons

[x] Table III.4-1values appended (energetic effects without solvation effects)

[e] 3 linear structures are excluded for S22

[f] 7 linear structures are excluded for S66

- From Table III.4-7, it can be clearly observed that, in case of QM methods, error values seem to get better for both S22 and S66 sets when the solvation is included.
- This is artificial in the sense that the reference method is the high-level extrapolated CCSD (T)/CBS, which does not take solvation effects into account. Therefore, this positive effect is thought to be a result of the geometry optimization changes compared to the original structures being smaller within the implicit solvent.
- In case of PLI10, no data is available to compare, but this resultant optimized interaction energies are no surprise for us, because, PLI10 is a set derived from a real protein-ligand systems, and they lack the stabilizing effect of the rest of the protein structure/surrounding.

Enthalpic and Entropic Effects (with Solvation)

Next, the correlation values with solvation effects are given for the comparison of non-optimized interaction energies ΔE_o, optimized interaction energies ΔE, interaction enthalpies ΔH, and free interaction energies ΔG. Data from Table III.4-4 (which is having enthalpic and entropic effects "without solvation") is appended next to the new results which are this time "with solvation".

		I	II	III	IV	V	VI
		ΔE_o vs ΔE	**ΔE vs ΔH**	**ΔE vs $-T\Delta S$**	**ΔE vs ΔG**	**ΔH vs $-T\Delta S$**	**ΔH vs ΔG**
S22[x] (without solvation)	**QM[e]**	1.00	1.00	-0.87	0.95	-0.86	0.95
	SQM	1.00	0.97	-0.79	0.71	-0.66	0.85
	MM[f]	0.87	1.00	-0.97	0.67	-0.96	0.70
S22[g]	**QM[a]**	1.00	1.00	-0.57	0.20	-0.60	0.16
	SQM[b]	0.99	0.35	-0.64	-0.30	0.29	0.71
	MM[c]	0.25	1.00	-0.80	-0.75	-0.77	-0.71
S66[x] (without solvation)	**QM[e]**	1.00	1.00	-0.59	0.91	-0.59	0.91
	SQM	0.99	0.98	-0.44	0.80	-0.36	0.87
	MM	0.92	1.00	-0.76	0.75	-0.75	0.75
S66[h]	**QM**	0.99	0.99	0.00	0.69	-0.05	0.66
	SQM	0.96	0.68	-0.27	0.16	0.19	0.67
	MM[d]	0.99	0.97	-0.03	0.89	0.01	0.93
PLI10 (without solvation)	**SQM**	–					
PLI10	**SQM**	0.06	0.98	-0.89	0.95	-0.85	0.99

Table III.4-8: *Correlation coefficients, Pearson R, for the comparison of non-optimized interaction energies ΔE_o, optimized interaction energies ΔE, interaction enthalpies ΔH, entropic contributions $-T\Delta S$ and free interaction energies ΔG for the QM, SQM and MM methods, with inclusion of solvation effects. (Parts from Table III.4-4, which are the values in case without solvation effects, are appended for S22 and S66 comparison purposes and denoted by [x])*

[a] Excluding entry 14 for technical reasons

[b] Excluding entry 10 for technical reasons

[c] Excluding entries 7 and 14 because of missing parameters and 12 for technical reasons

[d] Excluding entries 55 and 66 for technical reasons

[e] Data from reference38

[f] Excluding entry 7 and 14 because of missing parameters

[x] Data from III.4-4

[g] 3 linear structures are excluded for S22

[h] 7 linear structures are excluded for S66

- From Entry I, it is observed that the solvation effect makes a difference for MM methods in a prominent way and especially for S22 set (when it was R=0.87 without solvation, now it is R=0.25 with the solvation effects).

  QM and SQM stay similar with/without solvation contributions.

  MM, has such a big change, and this might be indicating that MM level has a low quality treatment for the solvation terms. This assumption will nevertheless be studied with the next Table III.4-9.

- For Entry II, there is a major change for SQM methods. The R values were around 0.99 and 1.00 "without solvation" for the sets S22 and S66 before, but this time values drop to 0.35 and 0.68 respectively when it is "with solvation" effects. The numerical drop for the S22 case is even more apparent.

- As a general tendency, it is seen that when the solvation effects are included, then almost all correlation values becomes smaller. To have a comparison amongst the situation of the data sets, it is observed that, when it is with solvation effects, then the S22 set has less correlation than before, but on the other hand there is almost no correlation for S66 set.

- Looking at the other entries III, IV, V and VI, it is observed that, QM has almost no correlation when solvation is added. When it is about Entry VI (which shows the correlation in between enthalpy and free interaction energy), then, a little correlation can be mentioned for SQM with S22 set, but this is not the case with the Entry IV (which shows the correlation in between interaction energy and free interaction energy). In meantime, with these Entries IV and VI and for the set S22, MM has a higher but a negative correlation.

These results overall show that EEC is less systematic when solvation is considered. The reason are thought to be either because of the inaccuracies either in the solvation treatments, or, about the interaction balance.

Table III.4-9 shows the comparison of the SQM and MM level non-optimized interaction energies ΔE_o, optimized interaction energies ΔE, interaction enthalpies ΔH, entropic contributions $-T\Delta S$, free interaction energies ΔG, and reports their correlation with QM level.

		SQM				MM				SQM	MM
		MD	**MAD**	**RMSD**	**MIMA**	**MD**	**MAD**	**RMSD**	**MIMA**	**R**	**R**
S22[a,c]	**ΔE_o**	-1.08	1.41	1.71	5.55	-8.95	8.97	21.30	88.17	0.93	0.13
	ΔE	-1.02	1.32	1.64	4.73	-2.87	2.93	3.86	8.46	0.93	0.64
	ΔH	-7.26	7.26	7.96	6.68	-6.24	6.24	6.93	3.99	0.26	0.54
	-TΔS	2.94	3.60	4.36	12.47	13.30	20.95	27.18	83.68	0.80	0.56
	ΔG	-4.32	4.45	5.54	11.02	7.25	8.64	10.25	26.17	0.44	0.22
S66[b,d]	**ΔE_o**	-1.57	1.71	2.12	6.00	-4.49	6.14	9.91	90.17	0.82	-0.03
	ΔE	-1.39	1.59	1.98	5.50	-2.40	4.14	8.94	86.82	0.83	-0.07
	ΔH	-6.78	6.78	7.13	7.57	-5.58	8.06	12.42	104.82	0.51	-0.08
	-TΔS	4.11	4.13	4.65	9.89	-0.52	9.31	13.17	90.95	0.62	0.33
	ΔG	-2.67	3.13	4.11	12.42	5.46	7.65	13.65	109.42	0.51	-0.15

Table III.4-9: *Comparison of the SQM and MM methods with QM data, non-optimized interaction energies ΔE_o, optimized interaction energies ΔE, interaction enthalpies ΔH, entropic contributions –TΔS and free interaction energies ΔG*

[a] SQM: excluding entries 10 and 14 for technical reasons, MM: excluding entries 7 and 14 because of missing parameters and 12 for technical reasons

[b]: Entries 55 and 56 are excluded for MM because of missing FF parameters

[c]: 3 linear structures are excluded for S22

[d]: 7 linear structures are excluded for S66

- For non-optimized (ΔE_o) and optimized interaction energies (ΔE), SQM has very good (low) MAD values which are around 1.5±0.2 kcal/mol and, also together with high R values being around 0.8 and 0.9 for S22 and S66 sets respectively. This is a better result when compared with the MM findings for these entries.
- However when it is about the enthalpy (ΔH), SQM and MM have similarly poor behaviour, and their MAD values are both around 7±1 kcal/mol. SQM has slightly better results for the S66 set, and MM for the S22 set. R values are not promising for them (for SQM it is R=0.26 and R=0.51 whereas, for MM, it is R=0.54 and R=-0.08 for S22 and S66 respectively).
- In case of the entropic contributions (-TΔS), MAD values are around 3-4 kcal/mol for SQM for both sets, whereas for MM method, they are 20.95 and 9.31 kcal/mol for S22 and S66 respectively. Correlation values are R=0.80 and R=0.62 for SQM and, R=0.56 and R=0.33 for the MM method respectively for S22 and S66 sets.
- About free energies (ΔG), MAD values for SQM are in between 3-4 kcal/mol for both sets, and they have correlations of R=0.44 and R=0.51 for S22 and S66 respectively. On the other hand, for MM, MAD values are around 7.5 to 8.5 kcal/mol for both sets where it resulted in low correlation values of R=0.22 and R=-0.15 for S22 and S66. This basically indicates that there isn't any correlation with QM.

The following Figure III.4-4 is related to the Table III.4-9, which presents the free interaction energies, ΔG, with solvation effects. SQM is given in addition with its scaled and shifted form for a better comparison.

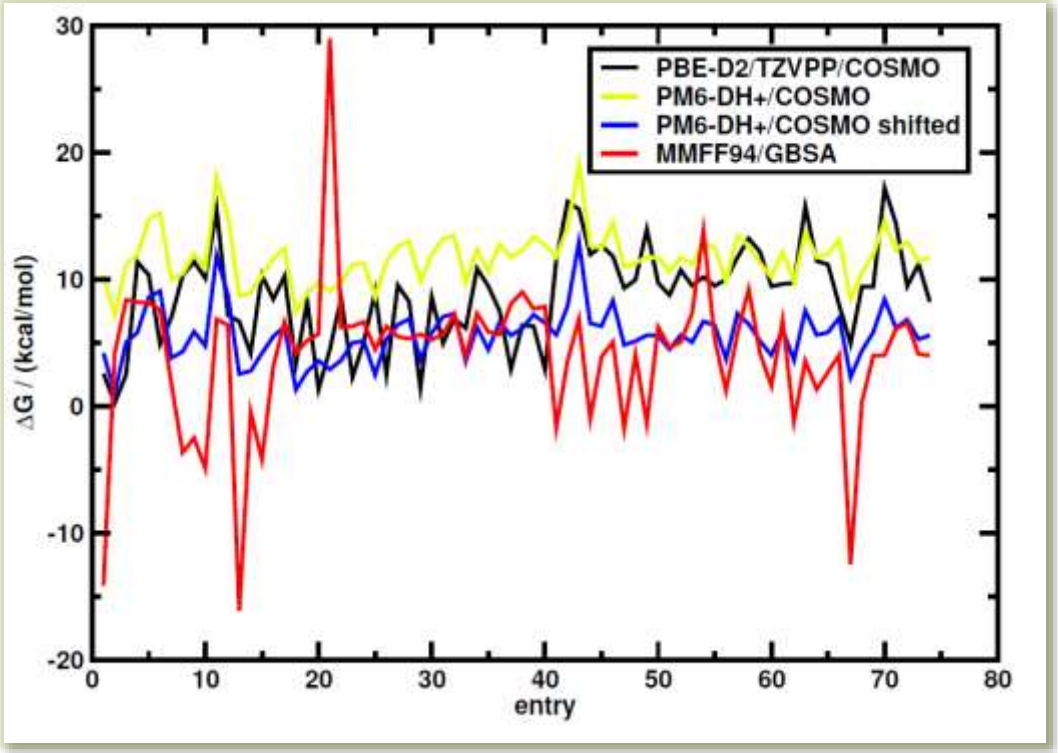


Figure III.4-4: *Both for S22 and S66, actual free interaction energies ΔG (with solvation energies at QM, SQM, SQM-shifted and MM level.*
MM methods are excluding entry 7 and 14 because of missing parameters.
MM methods are excluding entry 55 and 56 because of missing FF parameters
3 linear structures are excluded for S22
7 linear structures are excluded for S66
data from reference [338]

As it can be observed, the scaled and shifted version of SQM shows a closer figure to the QM trend, whereas, MM method clearly seems to have several outliers.

III.5 Discussions

Once more presenting Figure III.2-1,

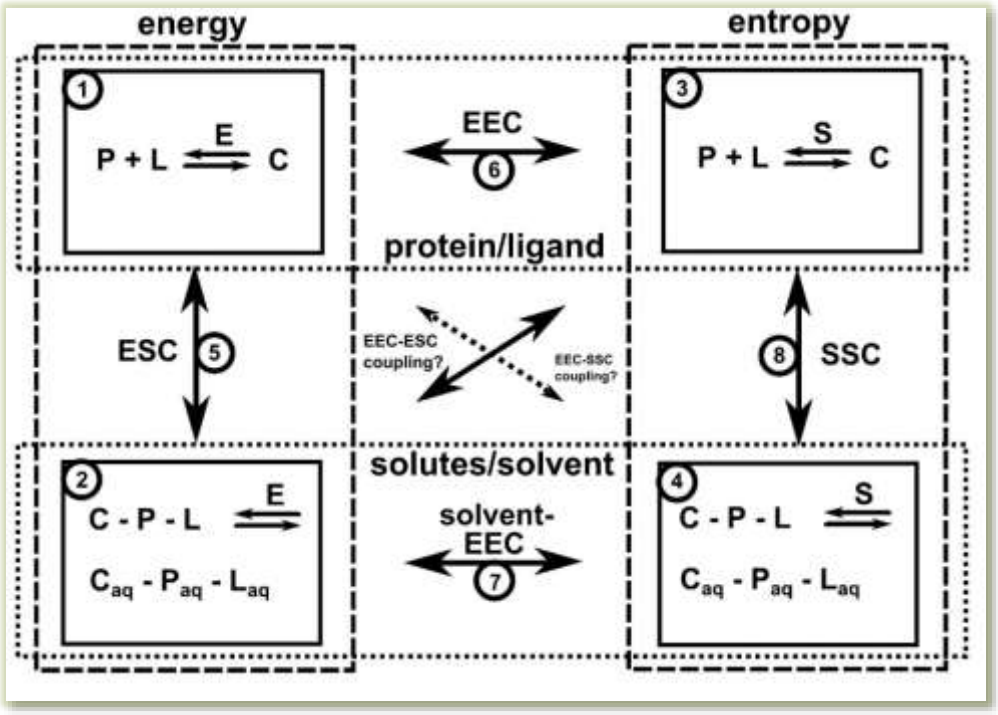


Figure III.2-1: *The representation of the delicate balance of biomolecular interactions.*
Part 1: energetic protein-ligand interactions,
Part 2: energetic solute-solvent interactions,
Part 3: entropic protein-ligand interactions,
Part 4: entropic solute-solvent interactions,
Part 5: energy-solvation compensation (ESC),
Part 6: energy-entropy compensation (EEC),
Part 7: energy-entropy compensation: EEC with solvent: (solvent EEC),
Part 8: entropy-solvation compensation (SSC).

Compared with the QM/COSMO values, the following can be stated:

- For Part 1: SQM and MM methods lead to very good (close to QM) interaction energies, SQM being a bit better in comparison.
- For Part 2: SQM/COSMO solvation is observed to be quite good for energetic solute-solvent interactions, MM/GBSA solvation is rather poor.
- For Part 3: Neither SQM nor MM treats entropic contributions well enough.
- For Part 5: It is found out that, QM, SQM and MM methods are similar to eachother for describing energy solvation compensations (ESC).

- For Part 6: Overall description of the energy-entropy compensation is similar for all methods, however, it is found that SQM underestimates the regularity of enthalpy-entropy compensation (EEC) whereas, MM overestimates.
- Coupling of ESC (energy-solvation compensation) and EEC (energy-entropy compensation) makes EEC less systematic, but this is not the case for ESC. However this observation might be due to the QM/COSMO reference.
- Enthalpic-entropic contributions for the SQM and MM methods get worse with solvation effects included.
- Compensation relations are more systematic for the larger systems (as in PLI10).
- Part 4 (entropic solvation effects), part 7 (solvent EEC) and part 8(SSC) could not be investigated due to technical reasons. Thus, EEC/SSC coupling was not able to be investigated either. It is in principle possible that there will be some compensation resulting from the replacement of the "fixed" water molecules from the surfaces of proteins and ligands. In that sense, entropic solvation effects are expected to be depending on the number of displaced water molecules, as well as depending on the effective size of the ligand, which is related with the conformational flexibility as well.

The most important results so far are the following: After testing the performances of MM and SQM and observing their problems, we find that QM is preferable, especially when it is about "solvation" and more importantly when it is about "enthalpic and entropic effects". On the other hand, when it is about the interaction energies, then, SQM is a good choice due to its computational speed as well.

SQM and MM have problems with enthalpic/entropic/solvation terms, on the other hand, somewhat a systematic compensation relation between the interaction energy terms and these enthalpic/entropic/solvation terms are observed. Then, such a question arises: what would be the result, if we extrapolate these values for SQM and MM methods?

III.5.1 Extrapolation of Energy Terms

The extrapolation results of the solvation interaction energy contributions from the polar interaction energy terms are tabulated in Table III.5.1-1.

The upcoming Table III.5.1-2 includes extrapolation of enthalpic and entropic free energy contributions from non-polar and polar interaction energy terms.

These tables involve both the intermediate and the final extrapolation results for the S22 and S66 benchmark sets. Parameter dependence and the overall robustness of the extrapolation procedures are described in the following parts.

Later on with Table III.5.1-3 the calculated and extrapolated free interaction energies are compared.

Extrapolation of Solvation Interaction Energy Contributions from Polar Terms
(Part 2 of Figure III.2-1)

	Type	a(S22)	New values fitted according to a(S22)				a(S66)	New values fitted according to a(S66)	a(ave)	New values fitted according to a(ave)	
			S22		S66			S66		S22	S66
			MAD	R	MAD	R		MAD		MAD	MAD
QM	**P**	0.70	1.5	0.93	1.7	0.83	0.6	1.7	0.75	1.5	1.6
SQM	**P**	1.0	1.1	0.93	0.7	0.89	0.9	0.7	0.95	1.1	0.7
	P*	1.1	1.1	0.96	1.0	0.87	1.1	1.0	1.10	1.1	1.0
MM[a,b]	**P**	0.7	1.3	0.96	1.5	0.54	0.7	1.5	0.70	1.3	1.5
	P*	0.9	2.0	0.86	-	-	0.6	1.8	0.70	2.2	1.9

Table III.5.1-1: *Extrapolation of solvation interaction energy contributions from polar interaction energy terms. Polar energy term based extrapolation is denoted by P. If the optimization is done, not with the original level of theory, but with respect to the QM data, then it is indicated by P*. a(S22), a(S66) and a(ave) are the definitions of parameters optimizations with for sets S22, S66 and averaged S22-S66 respectively.*
[a]S22; excluding entry 7 and 14 because of missing parameters
[b]MM methods are excluding entry 55 and 56 because of missing FF parameters

First of all, the optimized parameters that used for the sets S22 and S66 are averaged and these new parameters are denoted as a(S22) and a(S66).

The parameters are used as in the following:

$$\Delta E_{solvation} = - E_{polar} * a \qquad \text{(Eqn. 3.1.3-4)}$$

Also, an additional parameter is created from averaging these a(S22) and a(S66) values (averaging the average). It is denoted as a(ave), and obtained as follows:

$$a(ave) = \frac{a(S22)+a(S66)}{2} \qquad \text{(Eqn. 3.1.3-5)}$$

Once an average parameter, a(S22), a(s66) or a(ave) is obtained, then, the relevant computed results with it, are listed right next to it.

When this procedure is done for a(S66), only the S66 set was tested with this parameter, since a(S66) turned out to be very similar to a(S22) in terms of the numerical values.

The extrapolations of solvation interaction energy contributions that are calculated from the polar terms are symbolized with "P". The results that are obtained with P all have very similar numerical values of R and MAD for all parameters: a(S22), a(S66) and a(ave). This indicates that extrapolation from P is very robust.

The part shown as P*, is for the extrapolations which are done with respect to the QM level reference. This QM level based parameter supposedly gives more accurate values. However looking at the values, it is observed that P and P* are similar for this case.

- To have a comparison in between P and P*, values with the a(S22) parameter can be examined further for the set S22. SQM has a correlation of R=0.93 and R=0.96 respectively for similar P and P* values. On the other hand, this situation changes a bit for MM, with the respective correlations of R=0.96 and R=0.86 for P and P*.

Extrapolation of Enthalpic and Entropic Free Energy Contributions from Non-polar (Dispersion) and Polar (Hydrogen Bond) Interaction Energy Terms (Part 3 of Figure III.2-1)

Previously, Korth showed that [338] the compensation effects for non-polar (dispersion) were found to be three times higher than for polar (hydrogen-bond) interactions, and two different extrapolation schemes were tested there. Here, in a similar way, the extrapolation for the enthalpic and entropic free interaction energy contributions from non-polar and polar energy terms are tested and results are tabulated in Table III.5.1-2.

One of the extrapolation is based on full interaction E, and the other one is based on split polar (P) and non-polar (dispersion, D) type of interactions. E* and PD* denotes extrapolations which are done according to the QM level.

	Type	a(S22)[c]	New values fitted to a(S22)				a(S66)[d]	New values fitted to a(S66)	a(ave)	New values fitted to a(ave)	
			S22		S66			S66		S22	S66
			MAD	R	MAD	R		MAD		MAD	MAD
QM	**E**	0.8	3.8	0.94	4.7	0.80	1.2	3.8	1.0	4.1	4.1
	PD	0.6	2.1	0.94	2.1	0.90	0.75	1.6	0.7	2.6	1.7
SQM	**E**	1.4	6.3	0.88	7.1	0.77	2.2	5.8	1.8	6.6	6.1
	PD	1.1	4.0	0.92	4.4	0.80	1.4	3.2	1.3	4.5	3.4
	E*	0.8	4.0	0.73	4.6	0.39	1.25	3.7	1.0	4.1	4.0
	PD*	0.6	2.3	0.90	3.0	0.60	0.8	2.1	0.7	2.5	2.4
MM[a,b]	**E**	1.0	1.4	0.97	2.9	0.77	1.2	2.6	1.1	1.6	2.7
	PD	1.1	1.0	0.96	2.0	0.78	1.4	1.7	1.2	1.0	1.8
	E*	0.7	3.9	0.78	4.8	0.33	1.2	3.7	1.0	4.0	4.0
	PD*	1.1	3.4	0.82	3.0	0.40	1.5	2.3	1.3	3.5	2.5

Table III.5.1-2: *Extrapolation of enthalpic and entropic free interaction energy contributions from polar and non-polar interaction energy terms. Polar and dispersion energy term based extrapolation is denoted by PD, overall energy based extrapolation is denoted by E. If the optimization is done, not with the original level of theory, but with respect to the QM data, then it is indicated by PD* and E* respectively. a(S22), a(S66) and a(ave) are the definitions of parameter optimization for sets S22, S66 and averaged S22-S66 respectively.*
[a] S22; excluding entry 7 and 14 because of missing parameters
[b]: MM methods are excluding entry 55 and 56 because of missing FF parameters
[c]: 3 linear structures are excluded for S22
[d]: 7 linear structures are excluded for S66

The parameters are used as in the following:

$$\Delta G_{enthalpic/entropic} = - \Delta E * a \quad \text{(for E)} \qquad \text{(Eqn. 3.1.3-6)}$$

$$\Delta G_{enthalpic/entropic} = - \Delta E_P * a_P - \Delta E_D * a_D \quad \text{(for PD part)} \qquad \text{(Eqn. 3.1.3-7)}$$

$$a_D = a_P * \text{constant}$$

- This time, the extrapolation is found to be comparably less robust than it was for solvation contribution in Table III.5.1-1.

- For all cases, the version based on the P and D interactions (PD), is always found to more beneficial than the full interaction (E) based extrapolation.

- The optimal ratio of P and D parameters are around 3 for QM and SQM methods, whereas for MM method, this is about 2. Therefore, for all these methods, it seems that, only a very little can be gained with optimizing parameters independently, and that for all methods, only one parameter can fix both of these contributions.

- When a(ave) was considered for SQM and MM, then final parameters with them are P*=0.7 and P*=1.3 respectively. This is with a MAD value of about 2.5 kcal/mol MAD and R=0.90 for SQM and with a MAD value about 3 kcal/mol MAD and R=0.82 for MM in comparison to QM contribution. Here the wrong ratio of polar and non-polar interactions for MM and the high MAD values for both SQM and MM, again emphasize the difficulty of treating enthalpic/entropic effects correctly.

Extrapolation of Free Interaction Energies

Here, the previously calculated values for the free interaction energy, will be compared with the free interaction energy values obtained by extrapolation. These will be presented under the entries as "ΔG-calculated" and "ΔG_{extra}-extrapolated". Results are given in Table III.5.1-3 with the following additional denotations:

ΔG- ΔE :enthalpic/entropic contribution

(combined) :Electronic, solvation, enthalpic/entropic contribution calculated at the same time.

		Extrapolated [c]				Calculated			
		S22		S66		S22[a,d]		S66[b,e]	
		MAD	R	MAD	R	MAD	R	MAD	R
ΔE	**MM**	-	-	-	-	3.0	0.78	1.9	0.94
	SQM	-	-	-	-	1.0	0.99	1.1	0.98
ΔE_{solv}	**MM**	2.0	0.86	1.8	0.66	1.7	0.91	1.5	0.83
	SQM	1.1	0.96	1.0	0.87	1.1	0.94	0.8	0.95
	QM	1.5	0.93	1.6	0.83	-	-	-	-
ΔG- ΔE	**MM**	3.5	0.82	2.5	0.40	3.5	0.70	2.2	0.79
	SQM	2.5	0.90	2.4	0.60	6.2	0.94	5.0	0.73
	QM	2.6	0.94	1.7	0.90	-	-	-	-
ΔE(combined)	**MM**	-	-	-	-	3.0	0.70	4.1	-0.07
	SQM	-	-	-	-	1.3	0.95	1.6	0.83
ΔE_{solv}(combined)	**MM**	2.0	0.86	1.9	0.66	1.6	0.90	1.8	0.77
	SQM	1.1	0.96	1.5	0.93	0.8	0.89	0.7	0.95
	QM	1.7	0.97	2.3	0.90	-	-	-	-
ΔG- ΔE(combined)	**MM**	6.4(4.5)[c]	0.62	7.6(6.8)[c]	0.95	9.6	0.53	8.2	0.27
	SQM	6.3(4.0)[c]	0.57	7.5(3.8)[c]	0.32	4.0	0.51	2.1	0.54
	QM	5.3(3.4)[c]	0.67	5.9(3.1)[c]	0.40	-	-	-	-

Table III.5.1-3: *Comparison of the free interaction energies, ΔG, based on extrapolated enthalpic entropic and solvation contributions, with the free interaction energies, ΔG based on computed enthalpic entropic and solvation contributions [a,b].*
[a] S22:excluding QM entry 14, SQM entry 10, MM entry 12 for technical reasons and MM entries 7 and 14 because of missing parameters
[b] S66: excluding MM entries 55 and 66 for technical reasons
[c] With a parameter optimized including solvation effects: a^{QM}=1.1, a^{SQM}=1.4, a^{MM}=3.2.
[d]: 3 linear structures are excluded for S22
[e]: 7 linear structures are excluded for S66

Upper Half of the Table III.5.1-3 (About ΔE, ΔE_{solv} and ΔG-ΔE)

It is observed that when it is about the solvation contributions (ΔE_{solv}), then, the extrapolated and calculated MAD values are so similar, but looking at the R values, they are slightly less correlated with QM. In case of enthalpic-entropic contributions, (ΔG- ΔE), only for MM method, the extrapolated and calculated values are so similar to eachother. The extrapolated SQM values are more accurate than the calculated ones, but they are also less correlated with the QM data (especially for the more diverse (S66) set). This indicates that the extrapolation becomes less reliable with more mixed-up interactions that are in the S66 set.

Lower Half of the Table III.5.1-3 (About ΔE(combined), ΔE_{solv}(combined) and ΔG-ΔE(combined)

Calculation of combined contributions results in less systematic correlations. Following Figure III.5.1-1 includes the extrapolated data, and this is compared with Figure III.4-4, which has the calculated data instead. For an easier comparison, they are shown next to each other.

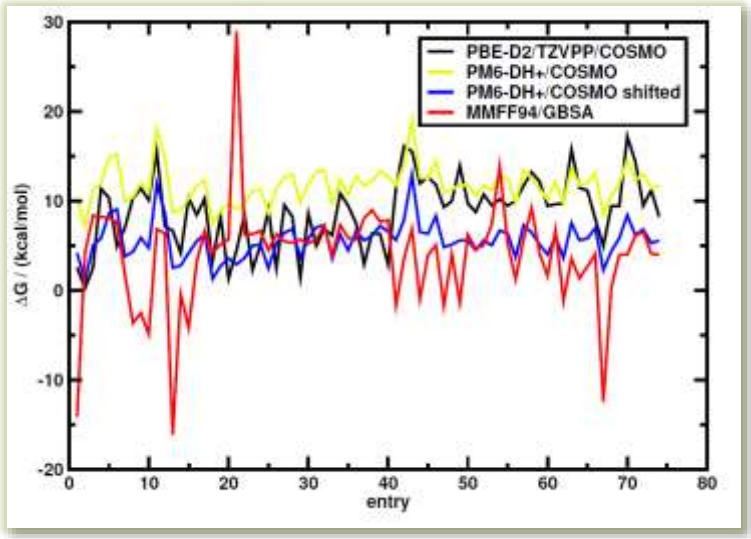


Figure III.4-4: *Both for S22 and S66, free interaction energies ΔG (with solvation) at QM, SQM, SQM-shifted and MM level.*
MM methods are excluding entry 7 and 14 because of missing parameters.
MM methods are excluding entry 55 and 56 because of missing FF parameters
3 linear structures are excluded for S22
7 linear structures are excluded for S66
data from reference [338]

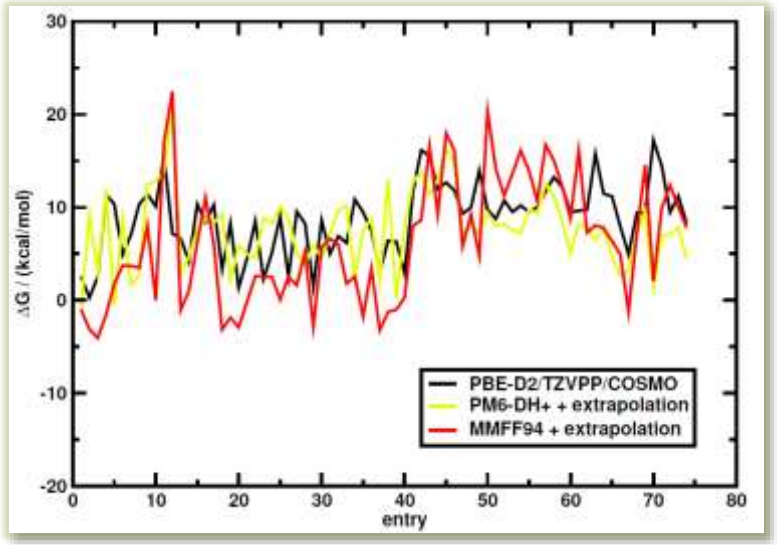


Figure III.5.1-1: *Both for S22 and S66 sets, free interaction energies ΔG (which includes solvation effects) at QM, SQM and MM level with SQM and MM enthalpic, entropic and solvation energy contributions extrapolated from interaction energies (equivalent to Figure III.4-4 only now with extrapolated SQM and MM data from Table III.5.1-3).*
S22:excluding QM entry 14, SQM entry 10, MM entry 12 for technical reasons and MM entries 7 and 14 because of missing parameters
S66: excluding MM entries 55 and 66 for technical reasons with a parameter optimized including solvation effects: aQM=1.1, aSQM=1.4, aMM=3.2.
3 linear structures are excluded for S22
7 linear structures are excluded for S66

It can be clearly seen that, the extrapolated version of MM does a similar good job as the calculated version (when it is compared with the QM data).

Implications for the Design of Scoring Functions

i. Part 1 of Figure III.2-1 (Energetic protein-ligand contributions), will be partly cancelled through EEC and ESC by part 2 (energetic solute/solvent contributions) and part 3 (entropic protein/ligand contributions).

ii. ESC, part 5 is very systematic and cancels by around 75% of all polar energetic contributions for our model systems.

iii. EEC, part 6 is somehow systematic and cancels about 25% of polar and 75% of non-polar energetic contributions for our model systems.

iv. As mentioned above, part 4, part 7 (solvent EEC) and part 8 (SSC) could not be investigated in this research, but the following assumption can be made: if they behave similar to the EEC and ESC, then, there should be a partial cancellation of the interaction terms again.

Once more, it should be emphasized that, in reality these partial balances are dynamically coupled unlike our uncoupled, static analysis. Therefore a rough approximation is made here, but still, our tests can identify/emphasize some important contributions, and these can be listed as follows:

a. The non-polar interaction energy can be counted as the most important one amongst these contributions, because by scaling this energy, the largest part of the effective internal (not conformational) entropy change could also be taken into account.

b. The polar interaction energy is almost completely cancelled. Even though it is not a full cancellation, the resulting term is small in comparison to the other terms.

Both of these reasoning (in a. and b. above) can be seen from the optimized parameters.

$$\Delta G_{enthalpic/entropic} = - \Delta E_p * a_p - \Delta E_D * a_D \quad \text{(for PD part)} \qquad \text{(Eqn. 3.1.3-8)}$$

The term "$\Delta E_p * a_p$" is roughly about 3, whereas, the term "$\Delta E_D * a_D$" is roughly about 1 in ratio.

Also, for the following:

$$\Delta E_{solvation} = - E_{polar} * a \qquad \text{(Eqn. 3.1.3-9)}$$

The term "$E_{polar} * a$" is about 1 for SQM, and about 0.75 for QM.

Due to our findings SQM level solvation can almost be equally treated via scaling polar interactions.

c. Related with the cancellation of the non-polar effects and up to a certain degree the cancellation of the polar effects, the protein/ligand internal entropic effects play a role and, at least at MM level, they can be included via scaling interaction energies.

The effect of improving the interaction energies for protein-ligands, while going from MM to SQM and then to QM level is low. Hence, only a better treatment of the energetic solvation contributions (at least at SQM level) and internal entropic contributions (at least at QM level) can likely make a change.

d. Finally, the entropic solvation contributions and the conformational entropy have likely also some impact.

Deriving conclusions from the remarks listed above, it looks like, even the simplest scoring function:

- Should have at least a scaled estimate of the dispersion (non-polar) protein/ligand interaction energy, which also implicitly gives the reasoning for the main part of the internal entropy changes, and also for the conformational entropy changes up to a point.
- Can be an estimator for conformational entropy effects, but unfortunately there isn't any proper QM way of treating conformational entropy so far, and this results in an option of skipping the relevant term totally. Even if it is not an ideal solution that way, at least, this approach might also eliminate the problematic terms in it.
- In order to implicitly account for the solvation effects, scaled polar interaction terms can be added.

  Due to the low accuracy & poor treatment of hydrogen bonding of MM, separate solvent terms might not be advantageous at this level, but, on the other hand, surface area related non-polar solvent terms can be helpful in the regard that they might indirectly add some flexibility to the explanation of protein-ligand dispersion contributions, even if they cannot directly help describing the solvation effects.
- Any possible terms, which can additionally or indirectly influence the important terms (like dispersion interaction or conformational entropy related terms) would be beneficial to consider.

QM based scoring functions have the advantage of having a proper treatment for the polar interaction energy contributions, solvation effects and internal enthalpic/entropic contributions. Still, even if the effect of all terms are considered together, their overall effect is not as much as:

i. The dispersion interaction terms,

ii. conformational entropy, presumably.

In that regard, conformational entropy is important to investigate.

The contributing terms for commonly used scoring functions are tabulated in Table III.5.1-4 to show that our findings are in line with existing scoring functions.

		contributions							
	Scoring functions	**N[1]**	**P**	**C**	**S**	**H**	**N[2]**	**M**	**other**
FF-based	**Goldscore** [358]	√	√	X	X	Optional	X	X	X
	Autodock [359]	√	√	Optional	Optional	√	X	X	X
	VSGB 2.0 [360]	√	√	√	X	√	√	X	X
Empirical	**Chemscore** [361]	X	X	X	√	√	√	√	X
	Glide [362]	√	√	√	√	√	√	√	P/N* interactions
	XScore [363]	√	X	X	√	√	√	X	X
Knowledge-based	**LigandFit** [364]	X	X	X	X	X	√	X	Shape factor
	eHits [365]	X	X	X	X	(√)	√	(√)	T factor

Table III.5.1-4: *Common scoring functions listed according to the interaction terms included.*
√: indicates that feature is included.
X: indicates that feature is not included.
Optional: feature is optional.
N[1]: Non-polar contribution (dispersion and exchange repulsion contributions)
N[2]: Non-polar equivalent contribution (hydrophobic/lipophilic, special π stacking terms, surface point interactions etc.)
P: polar
C: solvation
S: entropy
H: hydrogen bonding
M: metal interaction

There are different types of scoring functions [366].

FF based scoring functions use scaled force-field terms for all main interactions, and amongst them Autodock or Goldscore can be counted as advanced types. These main interactions can be listed as,

- Polar and non-polar energetic contributions
  - Description of hydrogen bonds terms by these are usually improved by special hydrogen bonding terms.
- Explicit internal entropy and solvation terms can also be represented through scaling of these interaction terms.

Empirical scoring functions, like ChemScore, Glide, or XScore scoring functions, omit the basic interaction terms, and treat the others in a simple way too.

However, as an important note here, this type of scoring functions still

- take care of the dispersion interaction terms (as an example: entropy or dispersion-equivalent ones) in a good way, and,
- They also have additional special hydrogen bond terms.

Knowledge based scoring functions, like LigandFit or eHiTS, store the contact-area related information. “Surface point interactions” and “shape fit” can be listed among the information that is kept. These are relevant to the dispersion interactions and also roughly to conformational entropy loss.

Overall, it is explained that [210] the very best scoring functions from each (FF, empirical or knowledge based) category, perform similar to each other in terms of the accuracy and the correlation. The similar performance of them is attributed to the parametrization that they have, rather than their conceptual differences.

III.5.2 Application Examples

Two application cases are chosen to further investigate the findings of the previous sections.

III.5.2.1 Case Study A

First study is based on a scoring study by Greenidge et al. [367] on 855 protein/ligand complexes of the PDBbind2009 database. The advanced VSGB 2.0 scoring function is used in their work.

Taking this 855 number of (PDBbind 2009 set based) complexes from Greenidge's study as a reference, we have studied a subset with a number of 580 complexes from PDBbind 2007.

Their study resulted in Pearson R value of R=0.79 for 855 complexes, whereas our subset of 580 complexes give about R=0.81. This shows that our PDBbind2007 based smaller subset is not a simplification when compared with the PDBbing2009 based set.

Following Table III.5.2.1-1 shows the correlation in between different energy contributions based on the 580 complexes computed with VSGB 2.0. Only 580 files are selected [367] to have a correspondence with our study that has 580 complexes in common.

The terms are given according to the reference denotations.

ΔE	: overall,
ΔE_{QQ}	: polar interaction term,
ΔE_{GB}	: polar solvation term,
ΔE_{vdw}	: explicit dispersion (and repulsion) interaction term, and,
$\Delta E_{hydrophobic}$	: hydrophobic contributions (i.e. implicit dispersion term)

ΔE	**ΔE_{QQ}**	**ΔE_{GB}**	**ΔE_{vdw}**	**$\Delta E_{hydrophobic}$**
ΔE	-0.03	0.00	0.87	0.89
	ΔE_{QQ}	-0.97	-0.12	-0.15
		ΔE_{GB}	0.00	0.04
			ΔE_{vdw}	0.85

Table III.5.2.1-1: *Pearson R values for the VSGB 2.0 and the energy contributions, ΔE: overall, ΔE_{QQ}: polar interaction term, ΔE_{GB}: polar solvation term, ΔE_{vdw}: explicit dispersion (and repulsion) interaction term and $\Delta E_{hydrophobic}$: hydrophobic contributions. All data is based on PDBbind 2009 data reference [367] selected for 580 protein and ligand complexes.*

One finds that, the polar interaction term ΔE_{QQ} cancels outs almost perfectly with the (polar) solvation term ΔE_{GB} with a correlation of R=-0.97. This corresponds to an average polar interaction contribution (ΔE_{QQ}) of -34 kcal/mol, average solvation contribution (ΔE_{GB}) of +38 kcal/mol.

It can be also observed that, the overall interaction energy (ΔE) has the highest correlations with two terms, and these are:

- ΔE_{vdw}, explicit dispersion (and repulsion) interaction term, and,
- $\Delta E_{hydrophobic}$, for hydrophobic interactions by reinforcing non-polar contacts.

Hydrophobicity is related to energetic and entropic balance of the non-polar and polar interactions between a solute and solvent (in this situation, protein), relative to the balance of the non-polar and polar interactions between a solute and water. Therefore, if there is any term linked with hydrophobicity, then that term is related to the dispersion interaction energy term as well. It can add weight or flexibility to the dispersion interaction term or it can account for solvent and protein/ligand entropic contributions.

Table III.5.2.1-2 compares these mentioned interaction terms and their combinations based on their ability to make predictions with respect to the experimental binding affinities.

Entry	Contribution Type	Pearson, R
I	ΔE vs ΔE_{vdw}	0.87
II	ΔE vs $\Delta E_{vdw} + \Delta E_{QQ} + \Delta E_{GB}$	0.87
III	ΔE vs $\Delta E_{vdw} + \Delta E_{hydrophobic}$	0.91
IV	ΔE vs $\Delta E_{vdw} + \Delta E_{hydrophobic} + \Delta E_{QQ} + \Delta E_{GB}$	0.98
V	pK vs ΔE	-0.81
VI	pK vs ΔE_{vdw}	-0.74
VII	pK vs $\Delta E_{vdw} + \Delta E_{hydrophobic}$	-0.77
VIII	ΔE_{D2} vs ΔE_{vdw}	0.92
IX	ΔE_{D2} vs $\Delta E_{vdw} + \Delta E_{hydrophobic}$	0.90
X	ΔE_{D2} vs ΔE	0.89

Table III.5.2.1-2: *Pearson R values for the correlation between VSGB 2.0 energy contributions [a], with experimental binding affinities pK and with D2 dispersion energy contributions, ΔE_{D2} for 580 protein/ligand complexes from PDBbind 2009 database [367]. ΔE: overall, ΔE_{QQ}: polar interaction term, ΔE_{GB}: polar solvation term, ΔE_{vdw}: explicit dispersion (and repulsion) interaction term, based on PDBbind 2009 data reference (70) selected for 580 protein and ligand complexes.*

The following can be stated:

- ΔE_{vdw} and $\Delta E_{hydrophobic}$ are the most significant (effective) terms.
- Comparing the Entries I, II, III and IV, the addition of the term "$\Delta E_{QQ} + \Delta E_{GB}$" seems to have no effect on the correlation.
- As for the Entries V, VI and VII, ΔE_{vdw} term alone, seems to have the predictive power itself (with R=-0.74 in Entry VI), which is almost similar to the full scoring function effect (with R=0.-81 in Entry V). However, adding $\Delta E_{hydrophobic}$ term to ΔE_{vdw} improves the situation a bit more (with R=-0.77 in Entry VII).
- So, it can be indeed observed that, ΔE_{vdw} and $\Delta E_{hydrophobic}$ terms are the most important ones to have an impact. Also from the Entries VIII, IX and X, it can be noted that ΔE_{D2} is perfectly correlated with ΔE_{vdw} (or with "$\Delta E_{vdw} + \Delta E_{hydrophobic}$") terms even if it is differently computed.

III.5.2.2 Case Study B

The second case study focuses on the scoring function comparison study by Wang et al. [209], where they use PDBbind2007 refined set (with 1297 protein/ligand complexes). Here, the dispersion interaction energy based estimates are compared to the results of Wang's study. The importance of this contribution is indeed observed here once again, especially for diverse protein/ligand systems.

		Pearson, R		
		Experimental Geometry (PDBbind set)[a]	Top score[b]	Best pose[b]
	ΔE_{D2}	0.51 (0.54[c])	-	-
Knowledge-based	LigandFit [364] (ver. 2.3)	-	0.33	0.28
	eHits [365] (ver. 9.0)	-	0.62	0.54
Empirical	FlexX [368] (ver. 2.2.1)	-	0.32	0.30
	Glide [362] (ver. 4.5)	-	0.50	0.48
	Surflex [369] (ver. 2.2)	-	0.57	0.47
FF-based	GOLD [370] (ver. 3.2)	-	0.41	0.25
	AutoDock [359] (ver. 4.2.1)	-	0.50	0.44

Table III.5.2.2-1: *For the 1297 protein-ligand complexes of the refined set of the 2007 PDBbind database, the Pearson R values for the correlation in between the dispersion interaction energy ΔE_{D2} and experimental binding affinities pK for several commonly used scoring functions.*
[a] This work
[b] Data from Plewczynski et al. [210] here as R instead of R2 values
[c] Logarithmic regression

The correlation of TPSS-D2 dispersion interaction energies (ΔE_{D2}), with the experimental binding affinities (pK) of the refined PDBbind2007 set (with 1297 protein-ligand complexes) is given. ΔE_{D2} itself shows a medium degree correlation with R=0.51 or R=0.54 with logarithmic regression.

"Top score" in the Table III.5.2.2-1 means, geometries are chosen according to the best score values. Based on the study of Plewczynski and coworkers [210] the Pearson R correlation values are in the best range of 0.32 to 0.62 for this top scoring set and for different types of the scoring functions. This is in agreement with an earlier study by Cheng and coworkers [209], where they tested the 195 protein-ligand complexes from the "core set" of the PDBbind2007 database.

"Best pose" in the table means that the geometries that are closer to the taken experimental values. Plewczynski and coworkers [210] have the correlation values in the best range of 0.25 to 0.54 for this part. For the top scoring and for the best pose cases, their results are similar or worse compared to the ΔE_{D2} prediction cases. If the DFT-D parameters are investigated, modifying the relevant parameters (s_6 and s_r) does not allow improvements on the R values.

Based on the high correlation (R=0.84) between the ligand size and ΔE_{D2}, ligand size also show that it can be a good indicator for proton-ligand binding affinities (R=0.40). These results support our previous statements that any good empirical dispersion interaction energy estimator should work well enough on its own to predict the protein-ligand binding affinities in the PDBbind set.

This is not in agreement with the literature on the performance of scoring functions for selected problems, which indicates that the quality of the PDBbind data is low, because:

i. The trend which is dominating, is the correlation of binding affinities with the dispersion contribution.

ii. Dispersion interactions are first of all related to the size of the ligand. The binding energy is known to be first of all related to the size of the ligand. The only systematic relationships we find for the PDBbind set is thus the most basic relation at all, and all other relations seem to be mixed up (hydrogen bond, solvation, and enthalpic, entropic terms). The reason behind this is likely because of the differences amongst the experiments behind the referred PDBbind data sets.

III.6 Conclusions

Looking at the delicate balance of biomolecular interactions and its implication for the design of scoring functions, compensation relations were investigated in order to find out whether they can be used to extrapolate enthalpic and entropic contributions to the free energy of single binding modes. Even though the investigated compensation relations are very systematic, they are not systematic enough to be used for this purpose.

3.2 Computational screening of energy (battery electrolyte) materials [371]

A volunteer computing approach is presented for screening a large number of molecular structures with respect to their suitability as new battery electrolyte solvents. Collective properties (i.e. melting, boiling and flash points) are evaluated with COSMOtherm and with quantitative structure–property relationship (QSPR) based methods, while electronic structure theory methods are used for the computation of electrochemical stability window estimators. Computational details are tabulated as follows:

Computational Methods		
Method Type	**Name**	**Computational Details**
DFT	BP86 [275,276]	Geometry optimization and frequency calculations were done with TURBOMOLE 6.4 [280, 281] Using: D2 dispersion corrections [279] the RI approximation for two-electron integrals [283,284] and Solvation effects are treated with COSMO [285] TZVP, TZVPPP and QZVP AO basis sets are employed [282].
		Using: D2 dispersion corrections [279] the RI approximation for two-electron integrals [283,284] LPNO (local pair natural orbital) CEPA (coupled electron pair [372] approximations are done with ORCA 2.8 [328] TZVP, TZVPPP and QZVP AO basis sets are employed [282]
SQM	PM6-DH+ [50]	Calculations were done with MOPAC2012 [58] by making use of COSMO solvation models [285]

***Table 3.2-1**: List of computational methods used for screening of battery electrolyte solvents*

Two application examples are studied:

i. First, the results of a previous large-scale screening test [373] are re-evaluated with respect to the mentioned collective properties.

ii. Second, all reasonable nitrile solvents up to 12 heavy atoms are generated and used to illustrate a suitable filter protocol for picking Pareto-optimal candidates.

As a result, the comparison with experimental references showed the high value of COSMOtherm and QSPR models for estimating collective properties of electrolyte components, especially for ranking compounds with respect to these properties.

SQM-based COSMOtherm estimates were much faster and they are found almost as valuable as DFT-based ones for this purpose.

Comparison of these two application examples showed that a diversity-oriented approach offers more opportunities for balancing thermal stability with ion conductivity.

From the second study, adiponitrile is found as one of the 17 Pareto-optimal candidates, in accordance with recent suggestions from experimental work (as well as several other small di-nitriles previously investigated).

Chapter 4

Discussions and Conclusions

From our research it is found out that, quantum mechanical calculations can be represented with smaller model systems without losing their predictive capability. However that for an automatic model preparation further adjustments are needed, like for instance entropic (and/or enthalpic) effects better be included or revised.

SQM methods are capable of handling large model systems, while it is still challenging for DFT approaches. When ranking is the only concern, it can be concluded that our reference SQM method, PM6-DH+ is reliable enough to be recommended for this purpose. On the other hand, for any other type of calculations, it is still recommend to use DFT-D3/TZVP/COSMO (-RS) methods.

From a methodological perspective, it is observed that an enhanced SQM approach, PM6-DH+ performs very similar to DFT-D and this shows a substantial improvement upon classical potentials. Based on the tests with smaller energy scales, SQM showed a deviation of 5% from DFT whereas MM (FF) had a deviation of 15% from DFT.

After the comparison with higher level methods, SCS-MP2 and B2-PLYP-D3 are found to be the most efficient WFT methods, whereas TPSS-D3+D_{abc}/def2-TZVPP is assigned as the best DFT approach. Our SQM reference PM6-DH+ is a fast and an accurate alternative to full *ab initio* treatments.

Then, main biomolecular interactions and the compensations in between them are studied. Overall, it is found that PM6-DH+ provides the opportunity to calculate the electronic energy part of the protein-ligand interactions for large number of large protein-ligand model systems in an accurate way. The extrapolation of enthalpic and entropic contributions from energies does not seem to be promising, and therefore, these contributions have to be computed using the RRHO approximation.

On the other hand, in combination with the minima mining approach, averaging over free energies for multiple binding modes seems to be a very promising strategy to address the protein-ligand biding problem. However, to evaluate this, it is observed that accurate experimental energies are needed, and it can be stated that the PDBbind database does not seem well suited for this purpose.

Though, fortunately within the CSAR 2014 scoring challenge, more accurate data seems to be available to continue from this point.

Bibliography

[1] Young, D. C., Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems, John Wiley & Sons, Inc., New York, USA, 2002.

[2] Cramer, C. J., Essentials of Computing Chemistry Theories and Models, 2nd Ed., John Wiley & Sons Ltd., West Sussex, England, 2004.

[3] Jensen, F., Introduction to Computational Chemistry, 1st Ed., John Wiley & Sons Ltd., West Sussex, England, 1999.

[4] Lewars, E. G., Computational Chemistry Introduction to the Theory and Applications of Molecular and Quantum Mechanics, 2nd Ed., Springer Netherlands, 2011.

[5] Truhlar, D. G., Potential Energy Surfaces, In Encyclopedia of Physical Science and Technology, 3rd Ed., edited by Meyers RA, Academic Press, New York, 2004; 13, 9-17.

[6] AimNature, Potential Energy Surface and Corresponding 2-D Reaction Coordinate Diagram derived from the plane passing through the minimum energy pathway between A and C and passing through B. http://en.wikipedia.org/wiki/File:Potential_Energy_Surface_and_Corresponding_Reaction_Coordinate_Diagram.png

[7] Silberberg, M. S., Chemistry: The Molecular Nature of Matter and Change, 2nd Ed., Boston: McGraw-Hill, 2000.

[8] van Santen, R. A. and Neurock, M., Molecular Heterogeneous Catalysis: A Conceptual and Computational Approach, Wiley-VCH Verlag GmbH & Co. KGaA, 2006, Print ISBN: 9783527296620, Online ISBN: 9783527610846, DOI: 10.1002/9783527610846

[9] Atkins, P. W., Molecular Quantum Mechanics: an Introduction to Quantum Chemistry. Oxford: Clarendon Press, 1970.

[10] Arfken, G., Lagrange Multipliers. §17.6 in Mathematical Methods for Physicists, 3rd Ed. Orlando, FL: Academic Press, 1985; 945-950.

[11] Lang, S., Calculus of several variables, 2nd Ed., Addison-Wesley Publishing Company, Massachusetts, 1979.

[12] Simmons, G. F., Differential Equations with Applications and Historical Notes, McGraw-Hill Book Company, New York, 1972.

[13] Zwillinger, D., Lagrange multipliers, §5.1.8.1 in CRC Standard Mathematical Tables and Formulae, 31st Ed., Boca Raton, FL: CRC Press, 2003; 389–390.

[14] Arfken, G., Hermitian Matrices, Unitary Matrices. §4.5 in Mathematical Methods for Physicists, 3rd Ed., Orlando, FL: Academic Press, 1985; 209-217.

[15] Ayres, Jr. F., Schaum's Theory & Problems of Matrices Schaum Publishing Co, McGraw-Hill, 1st Ed. 1962; 117-118.

[16] Hohenberg, P. and Kohn, W., Phys. Rev., 1964; B864, 136.

[17] Wimmer, E., J. Comput-Aided Mater. Des., 1993; 1, 215.

[18] Parr, R.G. and Yang, W., Density Functional Theory of Atoms and Molecules, Oxford University Press, New York, 1989.

[19] Zerner, M. C., in Reviews in Computational Chemistry, Lipkowitz, K. B., and Boyd D. B., Eds., VCH Publishers, New York, 1991; 2: 313-366.

[20] Stewart, J. J. P., in Reviews in Computational Chemistry, Lipkowitz, K. B., and Boyd D. B., Eds., VCH Publishers, New York, 1990; 45–82.

[21] Hartree, D.R., Proc. Cambridge Philos. Soc., 1928; 24, 328.

[22] Fock, V. A., Z. Phys., 1930; 15, 126.

[23] Catlow, C. R. A., in Handbook of Heterogeneous Catalysis, Ertl, H.K.G., Weitkamp, J. Eds., Wiley-VCH, New York, 1997; 1149.

[24] Yam, C.Y., Zhang, Q., Wang, F. and Chen, G. H., Chem. Soc. Rev., 2012; 41: 3821-3838

[25] Thiel, W., WIREs Comput. Mol. Sci. 2014; 4: 145–157.

[26] Cui, Q. and Elstner, M., PCCP, 2014; 16: 14368-14377.

[27] Yilmazer, N. D. and Korth, M., Comput. Struct. Biotechnol. J., 2015; 13: 169–175.

[28] http://www.shodor.org/chemviz/overview/ccbasics.html

[29] http://image.slidesharecdn.com/ding-md-for-cnt-graphene-growth-130701191703-phpapp01/95/what-can-we-learn-from-molecular-dynamics-simulations-of-carbon-nanotubeand-graphene-growth-15-638.jpg%3Fcb%3D1372724279

[30] Bingham, R.C., Dewar, M. J. S. and D. H. Lo, J. Am. Chem. Soc., 1975; 97: 1285–1293.

[31] Dewar, M. J. S. and Thiel, W., J. Am. Chem. Soc., 1977; 99: 4899–4907.

[32] Dewar, M. J. S. and Thiel, W., J. Am. Chem. Soc., 1977; 99: 4907–4917.

[33] Dewar, M. J. S., Zoebisch, E., Healy, E. F. and Stewart, J. J. P., J. Am. Chem. Soc., 1985; 107: 3902–3909.

[34] Kolb, M. and Thiel, W., J. Comput. Chem., 1993; 14: 775–789.

[35] Weber, W., Ein neues semiempirisches NDDO Verfahren mit Orthogonalisierungskorrekturen: Entwicklung des Modells, Parametrisierung und Anwendungen. PhD Thesis, Universität Zürich; 1996.

[36] Weber, W. and Thiel, W., Theor. Chem. Acc., 2000; 103: 495–506.

[37] Scholten, M., Semiempirische Verfahren mit Orthogonalisierungskorrekturen: Die OM3 Methode. Ph.D. Thesis, Universität Düsseldorf: Düsseldorf, Germany, 2003.

[38] Stewart J. J. P., J. Mol. Model., 2007; 13: 1173–1213.

[39] Stewart, J. J. P., J. Mol. Model., 2013; 19: 1–32.

[40] Seifert, G., Eschrig, H. and Bieger, W., Z. Phys. Chem., 1986; 267: 529–539.

[41] Seifert, G. and Joswig, J.-O., WIREs Comp. Mol. Sci., 2012; 2: 456–465.

[42] Elstner, M., Porezag, D., Jungnickel, G., Elsner, J., Haugk, M., Frauenheim, T., Suhai, S., and Seifert, G., Phys Rev B, 1998; 58: 7260–7268.

[43] Elstner, M., Gaus, M. and Cui, Q., WIREs Comp. Mol. Sci., 2014; 4: 49–61.

[44] Wollacott, A. M. and Merz, K. M., J. Chem. Theory Comput., 2007; 3: 1609–19.

[45] van der Vaart, A. and Merz, K. M., J. Am. Chem. Soc. 1999; 121: 9182–90.

[46] Clark, T., THEOCHEM, 2000;530:1–10.

[47] Dannenberg, J. J., THEOCHEM, 1997; 401: 279–86.

[48] Csonka, G. I. and Ángyán, J. G., THEOCHEM 1997; 393: 31–8.

[49] Winget, P., Selcuki, C., Horn, A. H. C., Martin, B., Clark, T., Theor. Chem. Acc., 2003; 110: 254–66.

[50] Korth, M., J. Chem. Theory Comput., 2010; 6: 3808–3816.

[51] Grimme, S., J. Comput. Chem., 2004; 25: 1463–73.

[52] Grimme, S., Antony, J., Ehrlich, S. and Krieg, H., J. Chem. Phys., 2010; 132: 154104.

[53] Korth, M., Pitoňák, M., Řezáč, J., and Hobza, P., J. Chem. Theory Comput., 2010; 6: 344–352.

[54] McNamara, J. P. and Hillier, I. H., Phys. Chem. Chem. Phys., 2007; 9: 2362–2370.

[55] Řezáč, J., Fanfrlík, J., Salahub, D. and Hobza, P., J. Chem. Theory Comput., 2009; 5: 1749–1760.

[56] Korth, M., Chemphyschem., 2011; 12: 3131-42.

[57] Vedani, A. and Huhta, D. W., J. Am. Chem. Soc., 1990; 112: 4759–67.

[58] http://www.openmopac.net.

[59] Collignon, B., Hoang, P. N. M, Picaud, S., Liotard, D., Rayez, M. T., et al., THEOCHEM, 2006; 772: 1–12.

[60] Tuttle, T. and Thiel, W., Phys. Chem. Chem. Phys., 2008; 10: 2159–66.

[61] Durán-Lara, E. F., López-Cortés, X. A., Castro, R. I., Avila-Salas, F., González-Nilo, F. D., Laurie V. F. and Santos, L. S., 2015; 168: 464-470.

[62] Gammack Yamagata, A. D., Datta, S., Jackson, K. E., Stegbauer, L., Paton, R. S. and Dixon, D. J., 2015; 54: 4899-4903.

[63] Wang, Q. and Bryce, R. A., J. Chem. Theory Comput., 2009; 5: 2206–11.

[64] Foster, M. E. and Sohlberg, K., J. Chem. Theory Comput., 2010; 6: 2153–66.

[65] Řezáč, J. and Hobza, P., Chem. Phys. Lett., 2011; 506: 286–9.

[66] Řezáč, J., Riley K.E. and Hobza., P., J. Chem. Theory Comput., 2012; 8: 4285–4292.

[67] Ibrahim, M. A. A., J. Chem. Inf. Model., 2011; 51: 2549–2559.

[68] Ibrahim, M. A. A., J. Phys. Chem. B., 2012; 116: 3659–3669.

[69] Laikov, D. N., J. Chem. Phys., 2011; 135: 134120.

[70] Řezáč, J. and Hobza. P., J. Chem. Theory Comput., 2012; 8: 141–151.

[71] Foster, M. E. and Sohlberg, K., Comp. Theor. Chem., 2012; 984: 9–12.

[72] Anikin, N. A., Bugaenko, V. L., Kuzminskii, M. B. and Mendkovich, A. S., Russ. Chem. Bull. 2012; 61: 12–16.

[73] Maia, J. D. C., Carvalho, G. A. U., Mangueira, C. P., Jr., Santana, S. R., Cabral, L. A. F. and Rocha, G. B., J. Chem. Theory Comput., 2012; 8: 3072–3081.

[74] Zhang, P., Fiedler, L., Leverentz, H. R., Truhlar, D. G. and Gao, J., J. Chem. Theory Comput., 2011; 7: 857–867.

[75] Isegawa, M., Fiedler, L., Leverentz, H. R., Wang, Y. and Nachimuthu, S., J. Chem. Theory Comput., 2013; 9: 33–45.

[76] Fiedler, L., Leverentz, H. R., Nachimuthu, S., Friedrich, J. and Truhlar, D.G., J. Chem. Theory Comput., 2014; 10: 3129–3139.

[77] Sure, R. and Grimme, S., J. Comput. Chem., 2013; 34: 1672–1685.

[78] Kromann, J. C., Christensen, A. S., Steinmann, C., Korth, M. and Jensen, J. H., Peer J. 2014; 2: e449.

[79] https://github.com/jensengroup/hydrogen-bond-correction-f3

[80] Brandenburg, J. G., Hochheim, M., Bredow, T. and Grimme, S., J. Phys. Chem. Lett., 2014; 5: 4275–4284.

[81] Govender, K., Gao, J. and Naidoo, K. J., J. Chem. Theory Comput., 2014; 10: 4694–4707.

[82] Govender, K.K. and Naidoo, K. J., J. Chem. Theory Comput., 2014; 10: 4708–4717.

[83] Prenosil, O., Pitonak, M., Sedlak, R., Kabelac, M. and Hobza P., Z. Phys. Chem., 2011; 225: 553–574.

[84] Hostaš, J., Řezáč, J. and Hobza P., Chem. Phys. Lett. 2013; 568: 161–166.

[85] Sedlak, R., Janowski, T., Pitoňák, M., Řezáč, J. and Pulay, P., J. Chem. Theory Comput., 2013; 9: 3364–3374.

[86] Li, A., Muddana, H. S. and Gilson, M. K., J. Chem. Theory Comput., 2014; 10: 1563–1575.

[87] Barberot, C., Boisson, J. C., Gerard, S., Khartabil, H. and Thiriot, E., Comp. Theor. Chem., 2014; 1028: 7–18.

[88] Mikulskis, P., Genheden, S., Wichmann, K. and Ryde, U., J. Comput. Chem., 2012; 33: 1179–1189.

[89] Yilmazer, N. D. and Korth, M., J. Phys. Chem. B., 2013; 117: 8075–8084.

[90] http://www.pdbbind-cn.org/

[91] Yilmazer, N. D., Heitel, P., Schwabe, T. and Korth, M., J. Theor. Comput. Chem., 2015; 14: 1540001.

[92] Bulo, R. E., Michel, C., Fleurat-Lessard, P. and Sautet, P., J. Chem. Theory Comput., 2013; 9: 5567–5577.

[93] Marion, A., Monard, G., Ruiz-Lopez, M. F. and Ingrosso, F., J. Chem. Phys. 2014; 141: 034106.

[94] Wu, X., Thiel, W., Pezeshki, S. and Lin, H., J. Chem. Theory Comput., 2013; 9: 2672–2686.

[95] Wang, S., MacKay, L. and Lamoureux, G., J. Chem. Theory Comput., 2014; 10: 2881–2890.

[96] Merz, K. M., Jr., J. Chem. Theory Comput., 2010; 6: 1769–1776.

[97] Faver, J. C., Benson, M. L., He, X., Roberts, B. P. and Wang, B., J. Chem. Theory Comput., 2011; 7: 790–797.

[98] Faver, J. C., Benson, M. L., He, X., Roberts, B.P. and Wang B., PLoS One, 2011; 6: e18868.

[99] Faver, J. C., Yang, W. and Merz, K. M., J. Chem. Theory Comput., 2012; 8: 3769–3776.

[100] Raju, R. K., Burton, N. A. and Hillier, I. H., Phys. Chem. Chem. Phys., 2010; 12: 7117–7125.

[101] Leverentz, H. R., Qi, H. W. and Truhlar, D. G., J. Chem. Theory Comput., 2013; 9: 995–1006.

[102] Todoroki, K., Nakano, T., Watanabe, H., Min, J. Z. and Inoue, K., Anal. Sci. 2014; 30: 865–870.

[103] Kamachi, T. and Yoshizawa, K., Org. Lett. 2014; 16: 472–475.

[104] Dresselhaus, T., Weikart, N. D., Mootz, H. D. and Waller, M. P., RSC Adv. 2013; 3: 16122–16129.

[105] Enriquez-Victorero, C., Hernandez-Valdes, D., Montero-Alejo, A. L., Durimel, A., Gaspard, S. and Jáuregui-Haza, U., J. Mol. Graph Model., 2014; 51: 137–148.

[106] Fanfrlík, J., Kolář, M., Kamlar, M., Hurny, D. and Ruiz, F.X., Cousido-Siah, A., Mitschler, A., Řezáč, J., Munusamy, E., Lepsik, M., Matejicek, P, Vesely, J., Podjarny, A. and Hobza, P., ACS Chem. Biol., 2013; 8: 2484–2492.

[107] Waller, M.P., Kumbhar, S. and Yang, J., Chem. Phys. Chem., 2014; 15: 3218–3225.

[108] Borbulevych, O. Y., Plumley, J. A., Martin, R. I., Merz, K. M., Jr. and Westerhoff, L. M., Acta Crystallogr. Sect. D: Biol. Crystallogr. 2014; 70: 1233–1247.

[109] Lupan, A., Kun, A. -Z., Carrascoza, F. and Silaghi-Dumitrescu, R., J. Mol. Model. 2013; 19: 193–203.

[110] Dresselhaus, T., Yang, J., Kumbhar, S. and Waller, M. P., J. Chem. Theory Comput. 2013; 9: 2137–2149.

[111] Husby, J., Todd, A. K., Platts, J. A. and Neidle, S., Biopolymers, 2013; 99: 989–1005.

[112] Morales-Toyo, M., Alvarado, Y. J., Restrepo, J., Seijas, L., Atencio, R. and Bruno-Colmenarez, J., J. Chem. Crystallogr. 2013; 43: 544–549.

[113] Nunez-Dallos, N., Reyes, A. and Quevedo, R., Tetrahedron Lett., 2012; 53: 530–534.

[114] Pietrusiewicz, K. M., Flis, A., Ujj, V., Körtvélyesi, T., Drahos, L., Pongrácz, P., Kollár, L. and Keglevich, G., Heteroatom. Chem., 2011; 22: 730–736.

[115] Pistarà, V., Corsaro, A., Rescifina, A., Catelani, G., D'Andrea, F. and Guazzelli, L., J. Org. Chem., 2013; 78: 9444–9449.

[116] Kubota, D., Macedo, O. F., Andrade, G. R., Conegero, L. S., Almeida, L. E., Costa, N. B., Jr., and Gimenez, I., F., Carbohydr. Res., 2011; 346: 2746–2751.

[117] Leclercq, L., Suisse, I., Roussel, P. and Agbossou-Niedercorn, F., J. Mol. Struct., 2012; 1010: 152–157.

[118] Xia, Y., Wang, X., Zhang, Y. and Luo, B., Comp. Theor. Chem., 2011; 967: 213–218.

[119] Gavvala, K., Sengupta, A., and Hazra, P., Chemphyschem., 2013; 14: 532–542.

[120] Grimme, S., Chem. Eur. J., 2012; 18: 9955–9964.

[121] Kessler, J., Jakubek, M., Dolensky, B. and Bour, P., J. Comput. Chem., 2012; 33: 2310–2317.

[122] Miletic, T., Kyriakos, K., Graovac, A. and Ibric, S., Carbohydr. Polym., 2013; 98: 122–131.

[123] Mora, A. J., Belandria, L. M., Ávila, E. E., Seijas, L. E., Delgado, G. E., Miró, A., Almeida, R., Brunelli, M. and Fitch, A.N., Cryst. Growth Des., 2013; 13: 1849–1860.

[124] Nachtigallova, D., Aquino, A. J. A., Horn, S. and Lischka, H., Photochem. Photobiol. Sci. 2013; 12: 1496–1508.

[125] Vuong, Q. V., Siposova, K., Nguyen, T. T., Antosova, A., Balogova, L., Drajna, L., Imrich, J., Li, M. S. and Gazova, Z., Biomacromolecules. 2013; 14: 1035–1043.

[126] Zimnicka, M., Troc, A., Ceborska, M., Jakubczak, M., Kolinski, M. and Danikiewicz W., Anal Chem. 2014;86:4249–4255.

[127] Krieg, E., Weissman, H., Shimoni, E., Bar On (Ustinov) A., Rybtchinski B., J. Am Chem. Soc. 2014; 136: 9443–9452.

[128] Leclercq, L., Lubart, Q., Dewilde, A., Aubry, J.-M. and Nardello-Rataj, V., Eur. J. Pharm. Sci., 2012; 46: 336–345.

[129] Muddana, H. S. and Gilson, M. K., J. Chem. Theory Comput., 2012; 8: 2023–2033.

[130] Muddana, H. S., Fenley, A. T., Mobley, D. L. and Gilson, M. K., J. Comput. Aided Mol. Des., 2014; 28: 305–317.

[131] Lukas, M., Meded, V., Vijayaraghavan A., Song L., Ajayan, P. M., Fink, K., Wenzel, W., and Krupke, R., Nat. Commun. 2013; 4: 1379.

[132] Cousins, B. G., Das, A. K., Sharma, R., Li, Y., McNamara, J. P., Hillier, I. H., Kinloch, I. A., Ulijn, R. V., Small. 2009; 5: 587–590.

[133] McNamara, J. P., Sharma, R., Vincent, M. A., Hillier, I. H. and Morgado, C. A., Phys. Chem. Chem. Phys., 2008; 10: 128–135.

[134] Chamorro-Posada, P., Vazquez-Cabo, J., Sanchez-Arevalo, F. M., Martin-Ramos, P., Martin-Gil, J. J, Navas-Gracia, L. M. and Dante, R. C., Solid State Chem., 2014; 219: 232–241.

[135] Ramraj, A. and Hillier, I. H., J. Chem. Inf. Model., 2010; 50: 585–588.

[136] Robert, P. T., Danneau, R., New. J. Phys., 2014; 16: 013019.

[137] Schrier, J., ACS Appl. Mater. Interfaces., 2011; 3: 4451–4458.

[138] Polynski, M. V. and Ananikov, V. P., Computational Modeling of Graphene Systems Containing Transition Metal Atoms and Clusters. In: Ananikov V. P. Ed., Understanding Organometallic Reaction Mechanisms and Catalysis. Wiley-VCH Verlag GmbH & Co.; KGaA, Weinheim: 2014.

[139] Ramraj, A., Hillier, I. H., Vincent, M. A. and Burton, N. A., Chem. Phys. Lett., 2010; 484: 295–298.

[140] Vijayaraj, R., Raman, S. S., Kumar, R. M. and Subramanian, V., J. Phys. Chem. B., 2010; 114: 16574–16583.

[141] Wang, H., Xu, X., Lee, C., Johnson, C. and Sohlberg, K., J. Phys. Chem. C, 2012; 116: 4442–4448.

[142] Yamada, M., Harada, K., Maeda, Y. and Hasegawa, T., New. J. Chem., 2013; 37: 3762–3769.

[143] Araujo, R. F., Silva, C. J. R., Paiva, M. C., Franco, M. M. and Proenca, M. F., RSC Adv., 2013; 3: 24535–24542.

[144] Foster, M. and Sohlberg, K., Fullerenes Nanotubes Carbon Nanostruct., 2012; 20: 72–84.

[145] Bende, A., Grosu, I. and Turcu, I., J. Phys. Chem. A, 2010; 114: 12479–12489.

[146] Bende, A. and Turcu, I., Int. J. Mol. Sci., 2011; 12: 3102–3116.

[147] Garcia, F., Costa, R. D., Arago, J., Bolink, H. J. and Orti, E., Langmuir, 2014; 30: 5957–5964.

[148] Molla, M. R., Gehrig, D., Roy, L., Kamm, V. and Paul, A., Chem. Eur. J., 2014; 20: 760–771.

[149] Simpson, S., Van Fleet, A. and Zurek, E., J. Chem. Educ., 2013; 90: 1528–1532.

[150] Haranczyk, M., Lin, L. -C., Lee, K., Martin, R. L., Neaton, J. B. and Smit, B., Phys. Chem. Chem. Phys., 2013; 15: 20937–20942.

[151] Martin, R. L., Shahrak, M. N., Swisher, J. A., Simon, C. M., Sculley, J. P., Zhou, H. -C., Smit, B. and Haranczyk, M., J. Phys. Chem. C., 2013; 117: 20037–20042.

[152] Martin, R. L., Simon, C. M., Smit, B. and Haranczyk M., J. Am. Chem. Soc., 2014; 136: 5006–5022.

[153] Sigal, A., Villarreal, M., Rojas, M. I. and Leiva, E. P. M., Int. J. Hydrog. Energy. 2014; 39: 5899–5905.

[154] Vincent, M. A. and Hillier, I. H., J. Chem. Inf. Model., 2014; 54: 2255–2260.

[155] Conti, S. and Ceccini, M., J. Phys. Chem C., 2015; 119: 1867–1879.

[156] Wang, Q., Suzuki, K., Nagashima, U., Tachikawa, M. and Yan, S., Chem. Phys., 2013; 419: 229–236.

[157] Wang, Q., Suzuki, K., Nagashima, U., Tachikawa, M. and Yan, S., Chem. Phys., 2013; 426: 38–47.

[158] Fanfrlík, J., Brahmkshatriya, P. S., Řezáč, J., Jilkova, A., Horn, M., Mares, M., Hobza, P. and Lepsik, M., J. Phys. Chem. B., 2013; 117: 14973–14982.

[159] Foster, M. E. and Sohlberg, K., J. Phys. Chem. A., 2011; 115: 7773–7777.

[160] Gordeev, E. G., Polynski, M. V. and Ananikov, V. P., Phys. Chem. Chem. Phys., 2013; 15: 18815–18821.

[161] Margiotta, N., Marzano, C., Gandin, V., Osella, D., Ravera, M., Gabano, E., Platts, J. A., Petruzzella, E., Hoeschele, J. D. and Natile, G., J. Med. Chem., 2012; 55: 7182–7192.

[162] Raju, R. K., Hillier, I. H., Burton, N. A., Vincent, M. A., Doudou, S. and Bryce, R., A., Phys. Chem. Chem. Phys., 2010; 12: 7959–7967.

[163] Ramraj, A., Raju, R. K., Wang, Q., Hillier, I. H., Bryce, R. A. and Vincent, M. A., J. Mol. Graph Model., 2010; 29: 321–325.

[164] Sharma, R., McNamara, J. P., Raju, R. K., Vincent, M. A., Hillier, I. H. and Morgado, C. A., Phys. Chem. Chem. Phys., 2008; 10: 2767–2774.

[165] Sun, W., Bu, Y. and Wang, Y., J. Comput. Chem., 2012; 33: 490–501.

[166] Morgado, C. A., McNamara, J. P., Hillier, I. H. and Burton, N. A., J. Chem. Theory Comput., 2007; 3: 1656–1664.

[167] Narth, C., Gillet, N., Levy, B., Demachy I. and de la Lande, A., Can. J. Chem., 2013; 91: 628–636.

[168] Strutynski, K., Gomes, J. A. N. F. and Melle-Franco, M., J. Phys. Chem. A, 2014; 118: 9561–9956.

[169] Keunchkarian, S., Franca, C. A., Gagliardi, L. G. and Castells, C. B., J. Chromatogr. A., 2013; 1298: 103–108.

[170] Werling, K. A., Hutchison, G. R. and Lambrecht, D. S., J. Phys. Chem. Lett., 2013; 4: 1365–1370.

[171] Amorim Madeira, P. J., Vaz, P. D., Bettencourt da Silva, R. J. N. and Florêncio, M. H., ChemPlusChem., 2013; 78: 1149–1156.

[172] Preiss, U. P. and Saleh, M. I., J. Pharm. Sci., 2013; 102: 1970–1980.

[173] Rappoport, D., Galvin, C. J., Zubarev, D. Y. and Aspuru-Guzik, A., J. Chem. Theory Comput., 2014; 10: 897–907.

[174] Fong, P., McNamara, J. P., Hillier, I. H. and Bryce, R. A., J. Chem. Inf. Model., 2009; 49: 913–924.

[175] Fanfrlík J., Bronowska, A. K., Řezáč, J., Prenosil, O., Konvalinka, J. and Hobza, P., J. Phys. Chem. B., 2010; 114: 12666–12678.

[176] Dobeš, P., Fanfrlík, J., Řezáč, J., Otyepka M. and Hobza, P., J. Comput. Aided Mol. Des., 2011; 25: 223–235.

[177] Jilkova A., Rezacova, P., Lepsik, M., Horn, M., Vachova, J., Fanfrlík, J., Brynda, J., McKerrow, J. H., Caffrey, C. R. and Mares, M., J. Biol. Chem., 2011; 286: 35770–35781.

[178] Nagy, G., Gyurcsik, B., Hoffmann, E. A. and Körtvélyesi, T., J. Mol. Graph Model., 2011; 29: 928–934.

[179] Kamel, K. and Kolinski, A., Acta Biochim. Pol., 2011; 58: 255–260.

[180] Avila-Salas, F., Sandoval, C., Caballero, J., Guiñez-Molinos, S., Santos, L. S., Cachau, R. E. and González-Nilo, F. D., J. Phys. Chem. B., 2012; 116: 2031–2039.

[181] Benson, M. L., Faver, J. C., Ucisik, M. N., Dashti, D. S., Zheng, Z. and Kenneth, M. M., Jr., J. Comput. Aided Mol. Des., 2012; 26: 647–659.

[182] The SAMPL experiment. http://www.eyesopen.com/SAMPL [Accessed: 2015-05-03]

[183] Quevedo, R., Nunez-Dallos, N., Wurst, K. and Duarte-Ruiz, A., J. Mol. Struct. 2012; 1029: 175–179.

[184] Kamel, K. and Kolinski, A., Acta Biochim. Pol., 2012; 59: 653–660.

[185] Stigliani, J. -L., Bernardes-Genisson, V., Bernadou, J. and Pratviel, G., Org. Biomol. Chem., 2012; 10: 6341–6349.

[186] Pan, X. -L., Liu, W. and Liu, J. -Y., J. Phys. Chem. B., 2013; 117: 484–489.

[187] Ahmed, M., Sadek, M. M., Abouzid, K. A. and Wang, F., J. Mol. Graph Model., 2013; 44: 220–231.

[188] Ucisik, M. N., Zheng, Z., Faver, J. C. and Merz, K. M., J. Chem. Theory Comput., 2014; 10: 1314–1325.

[189] Temelso, B., Alser, K. A., Gauthier, A., Palmer, A. K. and Shields, G. C., J. Phys. Chem. B., 2014; 118: 4514–4526.

[190] Pavlicek, J., Ptacek, J., Cerny, J., Byun, Y., Skultetyova, L., Pomper, M. G., Lubkowski, J. and Barinka, C., Bioorg. Med. Chem. Lett. 2014; 24: 2340–2345.

[191] Kruse, H., Havrila, M. and Sponer, J., J. Chem. Theory Comput., 2014; 10: 2615–2629.

[192] Vorlová, B., Nachtigallová, D., Jirásková-Vaníčková, J., Ajani, H., Jansa, P., Rezáč, J., Fanfrlík, J., Otyepka, M., Hobza, P., Konvalinka, J. and Lepšík, M., Eur. J. Med. Chem., 2015; 89: 189–197.

[193] Merriam-Webster's Collegiate Dictionary. 11th Ed., Springfield, MA: Merriam Webster, 2003.

[194] Cohen, N. C., Guidebook on Molecular Modeling in DrugDesign. Boston: Academic Press. ISBN 0-12-178245-X., 1996.

[195] Random House Webster's College Dictionary with CD-ROM. New York: Random House Reference, 2005.

[196] Oxford English Dictionary, 2nd Ed. 20 Vols. ,Oxford: Oxford University Press, 1989.

[197] Madsen, U., Krogsgaard-Larsen, P. and Liljefors, T., Textbook of Drug Design and Discovery. Washington, DC: Taylor & Francis. ISBN 0-415-28288-8, 2002.

[198] Anderson, A. C., Chemistry & Biology, 2003; 10: 787-797.

[199] Reddy, A. S., Pati, S. P., Kumar, P. P., Pradeep, H. N. and Sastry, G. N., Bentham Science Publishers, 2007; 8: 329-351.

[200] Leach, A. R., Shoichet, B. K. and Peishoff, C. E., J. Med. Chem., 2006; 49: 5851-5855.

[201] Walters, W. P., Stahl, M. T. and Murcko, M. A., Drug Discov. Today, 1998; 4: 160–178.

[202] Klebe, G., Drug Discovery Today, 2006; 11: 580-594.

[203] Giuliatti, S., Computer-Based Methods of Inhibitor Prediction, An Integrated View of the Molecular Recognition and Toxinology - From Analytical Procedures to Biomedical Applications, Prof. Gandhi Radis-Baptista Ed., ISBN: 978-953-51-1151-1, 2013, InTech, DOI: 10.5772/52334.

[204] Cheng, T., Li, Q., Zhou, Z., Wang, Y. and Bryant, S. H., AAPS J., 2012; 14: 133–141.

[205] Kitchen, D. B., Decornez, H., Furr, J. R. and Bajorath, J., Nature Reviews Drug Discovery, 2004; 3: 935-949.

[206] Gilson, M. K. and Zhou, H. -X., Annu. Rev. Biophys. Biomol. Struct., 2007; 36: 21-42.

[207] Gallicchio, E. and Levy, R. M., Adv. Protein Chem. Struct. Biol., 2011; 85: 27-80.

[208] Kroemer, R. T., Current Protein and Peptide Science, 2007; 8: 312-328.

[209] Cheng, T., Li, X., Li, Y., Liu, Z. and Wang, R., J. Chem. Inf. Model., 2009; 49: 1079-1093.

[210] Plewczynski, D., Lazniewski,M., Augustyniak, R. and Ginalski, K., J. Comput. Chem., 2011; 32: 742-755.

[211] Gibbs, J. W., American Journal of Science and Arts, vol. XVI 1878, 441-458.

[212] Zhou, H. –X. and Gilson, M. K., Chem. Rev., 2009; 109: 4092-4107.

[213] Chen, W., Gilson, M. K., Webb, S. P. and Potter, M. J., J. Chem. Theory Comput., 2010; 11: 3540-3557.

[214] Chang, C. -E. A. and Gilson, M. K., J. Am. Chem. Soc., 2004; 126: 13156-13164.

[215] Head, M. S., Given, J. A. and Gilson, M. K., 1997, 101, 1609-1618.

[216] Chen, W., Chang, C. –E. and Gilson, M. K., Biophys. J., 2004; 87: 3035-3049.

[217] Atkins, P. and de Paula, J., Atkins Physical Chemistry 8th Ed., WH Freman, 2006.

[218] Yilmazer, N. D. and Korth, M., Bunsen Magazin, 2013; 6: 294-298.

[219] Frauenhofer ISI, Technologie-Roadmap Lithium-Ionen-Batterie 2030, Karlsruhe 2010.

[220] Santos, E. and Schmickler, W., Catalysis in Electrochemistry: From Fundamentals to Strategies for Fuel Cell Development, Wiley, 2011.

[221] http://www.uni-ulm.de/en/nawi/dfg-research-unit-for-1376-theory-meets-experiment.html

[222] Goodenough, J. B., Acc. Chem. Res., 2013; 46: 1053-1061.

[223] Xu, K., Chem. Rev., 2004; 104: 4303-4418.

[224] Xu, K. and von Cresce, A., J. Mater. Chem., 2011; 21: 9849-9864.

[225] Bruce, P. G., Freunberger, S. A., Hardwick, L. J. and Tarascon, J.-M., Nature, 2012; 11: 19-29.

[226] http://www.shimadzu.com

[227] Scrosati, B., Hassoun, J. and Sun, Y. -K., Energy Environ. Sci., 2011; 4: 3287-3295.

[228] Tarascon, J. -M., Phil. Trans. R. Soc. A, 2010; 368: 3227-3241.

[229] Marom, R., Amalraj, S. F., Leifer, N., Jacob, D. and Aurbach, D., J. Mater. Chem. 2011, 21, 9938-9954.

[230] Tarascon, J. -M., ChemSusChem, 2008; 1: 777-779.

[231] Halls, M. D. and Tasaki, K., J. Power Sources, 2010; 195: 1472-1478.

[232] Hautier, G., Jain, A., Ong, S. P., Kang, B., Moore, C., Doe, R. and Ceder, G., Chem. Mater. 2011, 23: 3495-3508.

[233] G. Hautier, Jain, A., Chen, H., Moore, C., Ong, S. P. and Ceder, G., J. Mater. Chem., 2011; 21: 17147-17153.

[234] Park, M. H., Lee, Y. S., Lee, H. and Han, Y. -K., J. Power Sources, 2011; 196: 5109-5114.

[235] Schnur, S. and Groß, A., Catal. Today, 2011; 165: 129-137.

[236] Xing, L., Li, W., Wang, C., Gu, F., Xu, M., Tan, C. and Yi, J., J. Phys. Chem. B, 2009; 113: 16596-16602.

[237] Leung, K., and Budzien, J. L., Phys. Chem. Chem. Phys., 2010; 12: 6583-6586.

[238] Yu, J., Balbuena, P. B., Budzien, J. and Leung, K., J. Electrochem. Soc., 2011; 158: A400-410.

[239] Leung, K., Qi, Y., Zavadil, K. R., Jung, Y. S., Dillon, A. C., Cavanagh, A. S., Lee, S. -H., George, S. M., J. Am. Chem. Soc., 2011; 133: 14741-14754.

[240] Xing, L., Borodin, O., Smith, G. D. and Li, W., J. Phys. Chem. A, 2011; 115: 13896-13905.

[241] Kim, S.-P., van Duin, A. C. T., Shenoy, V. B., J. Power Source, 2011; 196: 8590-8597.

[242] Xu, K. and von Cresce, A., J. Mater. Res. 2012; 27: 2327-2341.

[243] von Wald Cresce, A., Borodin, O. and Xu, K., J. Phys. Chem. C, 2012; 116: 26111-26117.

[244] Owejan, J. E., Owejan, J. P., DeCaluwe, S. C. and Dura, J. A., Chem. Mater., 2012; 24: 2133-2140.

[245] Takamatsu, D., Koyama, Y., Orikasa, Y., Mori, S., Nakatsutsumi, T., Hirano, T., Tanida, H., Arai, H., Uchimoto, Y. and Ogumi, Z., Angew. Chem. Int. Ed., 2012; 51: 11597-11601.

[246] Xing, L., Vatamanu, J., Borodin, O., Smith, G. D. and Bedrov, D., J. Phys. Chem. C, 2012; 116: 23871-23881.

[247] Bedrov, D., Smith, G. D. and van Duin, A. C.T., J. Phys. Chem. A, 2012; 116: 2978-2985.

[248] Leung, K., J. Phys. Chem. C, 2012; 116: 9852-9861.

[249] Ganesh, P., Kent, P. R. C. and Jiang, D.-E., J. Phys. Chem. C, 2012; 116: 24476- 24481.

[250] Nie, M, Chalasani, D., Abraham, D. P., Chen, Y., Bose, A. and Lucht, B. L., J. Phys. Chem. C, 2013; 117: 1257-1267.

[251] Borodin, O., Behl, W. and Jow, T. R., J. Phys. Chem. C, 2013; 117: 8661-8682.

[252] K. Leung, J. Phys. Chem. C, 2013; 117: 1539-1547.

[253] Jorgensen, W. L., Acc. Chem. Res., 2009; 42: 724–733.

[254] Antony, J., Grimme, S., Liakos, D. G. and Neese, F., J. Phys. Chem. A, 2011; 115: 11210–11220.

[255] Antony, J. and Grimme, S., J. Comput. Chem., 2012; 33: 1730–1739.

[256] Raha, K. and Merz, K. M., J. Med. Chem., 2005; 48: 4558–4575.

[257] Hobza, P., Acc. Chem. Res., 2012; 45: 663–672.

[258] Hobza, P. and Müller-Dethlefs, K., Non-covalent Interactions. Theory and Experiment; RSC Publishing: London, 2010.

[259] Nicola, G., Liu, T. and Gilson, M. K., J. Med. Chem., 2012; 55: 6987-7002.

[260] Wang, R., Fang, X., Lu, Y. and Wang, S., J. Med. Chem., 2004; 47: 2977–2980.

[261] Wang, R., Fang, X., Lu, Y., Yang, C. -Y. and Wang, S., J. Med. Chem., 2005; 48: 4111–4119.

[262] Li Y., Liu Z. H., Li J., Han L., Liu J., Zhao Z. X. and Wang R. X., J. Chem. Inf. Model., 2014; 54: 1700-1716.

[263] Li Y., Han L., Liu Z. H. and Wang R. X., J. Chem. Inf. Model., 2014; 54: 1717-1736.

[264] Li, H., Robertson, A. D. and Jensen, J. H., Proteins: Structure, Function, and Bioinformatics, 2005; 61: 704-721.

[265] Grebner, C., Kästner, J., Thiel, W. and Engels, B., J. Chem. Theory Comput., 2013; 9: 814–821.

[266] http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/

[267] Kendall, G., Biometrika, 1938; 30: 81-93.

[268] Nelsen, R.B., "Kendall tau metric", in Hazewinkel, Michiel, Encyclopedia of Mathematics, Springer, ISBN 978-1-55608-010-4, 2001.

[269] http://stamash.org/calculating-kendalls-taurank-correlation-coefficient/

[270] Halgren, T. A., J. Comput. Chem., 1996; 17: 490–519.

[271] O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T. and Hutchinson, G. R., J. Chem. Inf., 2011; 3: 33–47.

[272] Wang, J., Cieplak, P. and Kollman, P. A, J. Comput. Chem., 2000; 21: 1049–1074.

[273] Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. and Case, D. A., J. Comput. Chem. 2004; 25: 1157–1174.

[274] Case, D. A., Darden, T. A., Cheatham, T. E., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Walker, R. C., Zhang, W., Merz, K. M., Roberts, B., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossvazary, I., Wong, K. F., Paesani, F., Vanicek, J., Wu, X., Brozell, S. R., Steinbrecher, T., Gohlke, H., Cai, Q., Ye, X., Wang, J., Hsieh, M. -J., Cui, G., Roe, D. R., Mathews, D. H., Seetin, M. G., Sagui, C., Babin, V., Luchko, T., Gusarov, S., Kovalenko, A., Kollman, P. A., AMBER 11; University of California: San Francisco, 2010.

[275] Becke, A. D., Phys. Rev. A, 1988; 38: 3098–3100.

[276] Perdew, J. P., Phys. Rev. B, 1986; 33: 8822–8824.

[277] Perdew, J. P., Burke, K. and Ernzerhof, M., Phys. Rev. Lett., 1996; 77: 3865–3868.

[278] Tao, J., Perdew, J. P., Staroverov, V. N. and Scuseria, G. E., Phys. Rev. Lett., 2003; 91: 146401–146405.

[279] Grimme, S., J. Comput. Chem., 2006; 27: 1787–1799.

[280] Ahlrichs, R., Bär, M., Häser, M., Horn, H. and Kölmel, C., Chem. Phys. Lett., 1989; 162: 165–169.

[281] TURBOMOLE V6.4 2012, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989–2007, TURBOMOLE GmbH, since 2007; available from http://www.turbomole.com.

[282] Schäfer, A., Huber, C. and Ahlrichs, R., J. Chem. Phys., 1994; 100: 5829–5835.

[283] Eichkorn, K., Treutler, O., Öhm, H., Häser, M. and Ahlrichs, R., Chem. Phys. Lett., 1995; 242: 652–660.

[284] Eichkorn, K., Weigend, F., Treutler, O. and Ahlrichs, R., Theor. Chem. Acc., 1997; 97: 119–124.

[285] Klamt, A., WIREs Comput. Mol. Sci., 2011; 1: 699–709.

[286] http://openmopac.net/manual/mozyme.html,

[287] http://www.bannedbygaussian.org/

[288] Cole, D. J., Skylaris, C. -K., Rajendra, E., Venkitaraman, A. R. and Payne, M. C., Europhys. Lett., 2010; 91: 37004–37009.

[289] http://plasmagate.weizmann.ac.il/Grace/.

[290] Englebienne, P. and Moitessier, N., J. Chem. Inf. Model., 2009; 49: 2564–2571.

[291] Rao, L., Zhang, I. Y., Guo, W., Feng, L., Meggers, E. and Xu, X., J. Comput. Chem., 2013; 34: 1636-1646.

[292] Gleeson, D., Tehan, B., Gleeson, P. and Limtrakul, J., Org. Biomol. Chem., 2013; 10: 7053-7061.

[293] Kantardjiev, A. A., Nuc. Acids Res., 2012; 40: W415-W422.

[294] Bryce, R. A., Fut. Med. Chem., 2011; 3: 683-698.

[295] Gleeson, M. P., Hannongbua, S. and Gleeson, D., J. Mol. Graph Mod., 2010; 29: 507-517.

[296] Zhou, T. and Caflisch, A., Chem. Med. Chem., 2010; 5: 1007-1014.

[297] Zhang, X., Gibbs, A. C., Reynolds, C. H., Peters, M. B. and Westerhoff L. M., J. Chem. Inf. Mod., 2010; 50: 651-661.

[298] Söderhjelm, P., Kongsted, J., Genheden, S. and Ryde, U., Int. Sci. Comput. Life Sci., 2010; 2: 21-37.

[299] Söderhjelm, P., Kongsted, J. and Ryde, U., J. Chem. Theory Comput., 2010; 6: 1726-1737.

[300] Söderhjelm, P., Aquilante, F. and Ryde, U., J. Phys. Chem. B, 2009; 113: 11085-11094.

[301] Briganti, F., Mangani, S., Orioli, P., Scozzafava, A., Vernaglione, G. and Supuran, C. T., Biochemistry, 1997; 36: 10384-10392.

[302] Burkhard, P., Taylor, P. and Walkinshaw, M. D., J. Mol. Biol., 2000; 295: 953-962.

[303] Ghosh, M., Meerts, I. A. T. M, Cook, A., Bergman, A., Brouwer, A. and Johnson, L. N., Acta. Cryst. D, 2000; 56: 1085-1095.

[304] Lee, J. E., Cornell, K. A., Riscoe, M. K. and Howell, P. L., Structure, 2001; 9: 941-953.

[305] Lu, G., Dobritzsch, D., Baumann, S., Schneider, G. and König, S., Eur. J. Biochem., 2000; 267: 861-868.

[306] Manuel, R. C., Hitomi, K., Arvai, A. S., House, P. G., Kurtz, A. J., Dodson, M. L., McCullough, A. K., Tainer, J. A. and Lloyd, R. S., J. Biol. Chem., 2004; 279: 46930-46939.

[307] James, L. C. and Tawik, D. S., Proc. Natl. Acad. Sci. USA, 2005; 102: 12730-12735.

[308] Razavi, H., Palaninathan, S. K., Powers, E. T., Wiseman, R. L., Purkey, H. E., Mohamedmohaideen, N. N., Deechongkit, S., Chiang, K. P., Dendle, M. T. A., Sacchettini, J. C., Kelly, J. W., Angew Chem. Int. Ed., 2003; 42: 2758-2761.

[309] Babaoglu, K. and Shoichet, B. K., Nat. Chem. Biol., 2006; 2: 720-723.

[310] Becke, A. D., J. Chem. Phys., 1993; 98: 5648-5652.

[311] Stephens, P. J., Devlin, F. J., Chabalowski, C. F. and Frisch, M. J., J. Phys. Chem., 1994; 98: 11623-11627.

[312] Lee, C., Yang, W. and Parr, R. G., Phys. Rev. B, 1998; 37: 785-789.

[313] Weigend, F., Köhn, A., Hättig, C., J. Chem. Phys., 2002; 116: 3175-3183.

[314] Zhao, Y. and Truhlar, D. G., Theor. Chem. Acc., 2008; 120: 215-241.

[315] Valiev, M., Bylaska, E. J., Govind, N., Kowalski, K., Straatsma, T. P., van Dam, H. J. J., Wang, D., Nieplocha, J., Aprà, E., Windus, T. L., De Jong, W. A., Comput. Phys. Commun., 2010; 181: 1477-1489.

[316] Grimme, S., J. Chem. Phys., 2003; 118: 9095-9102.

[317] Grimme, S., J. Chem. Phys., 2006; 124: 034108-034116.

[318] Bachorz, R. A., Bischoff, F. A., Glöß, A., Hättig, C., Höfener, S., Klopper, W., and Tew, D. P., J. Comput. Chem., 2011; 32: 2492-2513.

[319] Dunning, T. H., Jr., J. Chem. Phys., 1989; 90: 1007-1023.

[320] Kendall, R. A., Dunning, T. H., Jr. and Harrison, R. J., J. Chem. Phys., 1992; 96: 6796-6806.

[321] Woon, D. E., Dunning, T. H., Jr., J. Chem. Phys., 1993; 98: 1358-1371.

[322] Weigend, F., Phys. Chem. Chem. Phys., 2002; 4: 4285-4291.

[323] Weigend, F., J. Comput. Chem., 2008; 29: 167-175.

[324] Huntington, L. M. and Nooijen, M., J. Chem. Phys., 2010; 133: 184109.

[325] Huntington, L. M. J., Hansen, A., Neese, F. and Nooijen, M., J. Chem. Phys., 2012; 136: 064101.

[326] Neese, F., Wennmohs, F. and Hansen, A., J. Chem. Phys., 2009; 130: 114108.

[327] Neese, F., Hansen, A. and Liakos, D. G., J. Chem. Phys., 2009; 131: 064103.

[328] Neese, F., WIREs Comput. Mol. Sci., 2012, 2, 73-78.

[329] Korth, M., Grimme, S. and Towler, M. D., J. Phys. Chem. A, 2011; 115: 11734-11739.

[330] Raghavachari, K., Trucks, G. W., Pople, J. A., Head –Gordon, M., Chem. Phys. Lett., 1989; 157: 479-483.

[331] Jurečka, P., Šponer, J., Černý, J. and Hobza, P., Phys. Chem. Chem. Phys., 2006; 8: 1985-1993.

[332] Rezác, J., Riley, K. and Hobza, P., J. Chem. Theory Comput., 2011; 7: 2427-2438.

[333] Schwabe, T., J. Comput. Chem., 2012; 33: 2067-2072.

[334] Schwabe, T. and Grimme, S., Phys. Chem. Chem. Phys., 2007; 9: 3397-3406.

[335] Korth, M. and Grimme, S., J. Chem. Theory Comput., 2009; 5: 993–1003.

[336] Liu, L. and Guo, Q. -X., Chem. Rev., 2001; 101: 673-695.

[337] Lafont, V., Armstrong, A. A., Ohtakta, H., Kiso, Y., Amzel, L. M. and Freire, E., Chem. Biol. Drug Des., 2007; 69: 413-422.

[338] Korth, M., Med. Chem. Commun., 2013; 4: 1025-1033.

[339] Dunitz, J., Chem. Biol., 1995; 2: 709-712.

[340] Searle, M. S., Williams, D. H., J. Am. Chem. Soc., 1992; 114: 10690-10697.

[341] Searle, M. S. and Williams, D. H., Nucleic Acids Research, 1993; 21: 2051-2056.

[342] Westwell, M., Searle, M. S., Klein, J. and Williams, D. H., J. Phys. Chem., 1996; 100: 16000-16001.

[343] Searle, M. S., Westwell, M. S., Williams, D. H., J. Chem. Soc. Perkin Trans., 1995; 2: 141-151.

[344] Williams, D. H. and Westwell, M.S., Chem. Soc. Rev., 1998; 27: 57-63.

[345] Williams, D. H., Stephens, E., O'Brien, D. P. and Zhou, M., Angew. Chem. Int. Ed. Engl., 2004; 43: 6596-6616.

[346] Houk, K. N., Leach, A. G., Kim, S. P. and Zhang, X., Angew. Chem. Int. Ed., 2003; 42: 4872-4897.

[347] Ford, D. M., J. Am. Chem. Soc., 2005; 127: 16167- 16170.

[348] Gallicchio, E., Kubo, M. M., Levy, R. M., J. Am. Chem. Soc., 1998; 120: 4526- 4527.

[349] Graziano, G., J. Chem. Phys., 2004; 120: 4467-4471.

[350] Olsson, T. S. G., Ladbury, J. E., Pitt, W. R. and Williams, M. A., Protein Sci., 2011; 20: 1607-1618.

[351] Moghaddam, S., Inoue, Y. and Gilson, M. K., J. Am. Chem. Soc., 2009; 131: 4012-4021.

[352] Forrey, C., Douglas, J. F. and Gilson, M. K., Soft Matter, 2012; 8: 6385- 6392.

[353] Bissantz, C., Kuhn, B. and Stahl, M., J. Med. Chem., 2010; 53: 5061-5084.

[354] Stone, A. J., The Theory of Intermolecular Forces, Oxford University Press, Oxford, 1997.

[355] http://www.begdb.com/

[356] Yilmazer, N.D., Schwabe, T. and Korth, M., "On the delicate balance of biomolecular interactions and its implication for the design of scoring functions", in preparation.

[357] Paton, R. S. and Goodman, J. M., J. Chem. Inf. Model., 2009; 49: 944-955.

[358] Baxter, C. A., Murray, C. W., Clark, D. E., Westhead, D. R. and Eldridge, M. D., Proteins: Struct. Funct. Genet., 1998; 33: 367-382.

[359] Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R. Hart, W. E., Belew, R. K., Olson, A. J., J. Comput. Chem., 1998; 19: 1639-1662.

[360] Li, J., Abel, R., Zhu, K., Cao, Y., Zhao, S. and Friesner, R. A., Proteins, 2011; 79: 2794-2812. (VSGB2.0 reference)

[361] Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V. and Mee, R. P., J. Comput. Aided Mol. Des., 1997; 11: 425-445.

[362] Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P. and Shenkin, P. S., J. Med. Chem., 2004; 47: 1739-1749.

[363] Wang, R., Lai, L. and Wang, S., J. Comput.-Aided Mol. Des., 2002; 16: 11-26.

[364] Venkatachalam, C. M., Jiang, X., Oldfield, T. and Waldman, M., J. Mol. Graph Model, 2003; 21: 289-307.

[365] Zsoldos, Z., Reid, D., Simon, A., Sadjad, S. B. and Johnson, A. P., J. Mol. Graph Model, 2007; 26: 198-212.

[366] Ferrara, P., Gohlke, H., Price, D. J., Klebe, G. and Brooks III, C. L., J. Med. Chem., 2004; 47: 3032-3047.

[367] Greenidge, P. A., Kramer, C., Mozziconacci, J. –C. and Wolf, R. M., J. Chem. Inf. Model., 2013; 53: 201-209.

[368] Rarey, M., Kramer, B., Lengauer, T. and Klebe, G., J. Mol. Biol., 1996; 261: 470-489.

[369] Jain, A. N., J. Med. Chem., 2003; 46: 499-511.

[370] Jones, G., Willett, P., Glen, R. C., Leach, A. R. and Taylor, R., J. Mol. Biol., 1997; 267: 727-748.

[371] Husch, T., Yilmazer, N. D., Balducci, A. and Korth, M., Phys. Chem. Chem. Phys., 2015; 17: 3394-3401.

[372] Neese, F., Hansen, A., Wennmohs, F. and Grimme, S., Acc. Chem. Res., 2009; 42: 641-648

[373] Korth, M., Phys. Chem. Chem. Phys., 2014; 16: 7919-7926.

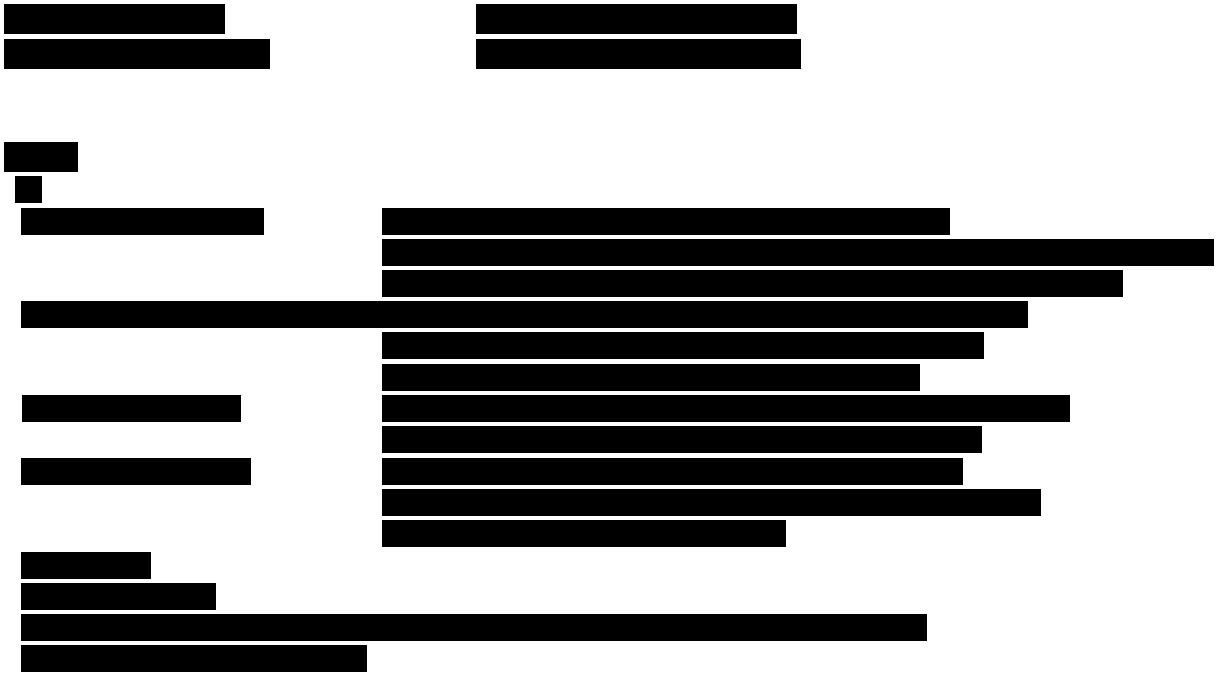
Erklärung

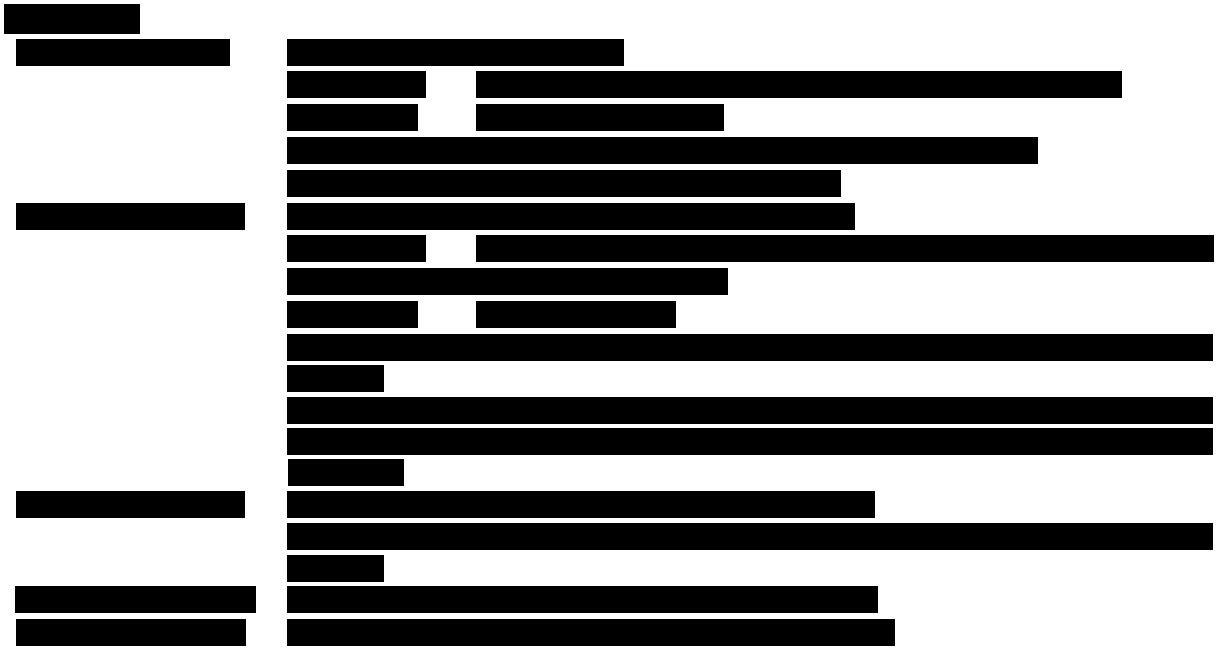
Ich erkläre, dass ich die Arbeit selbständig verfasst und keine anderen als die angegebenen. Quellen und Hilfsmittel verwendet habe.

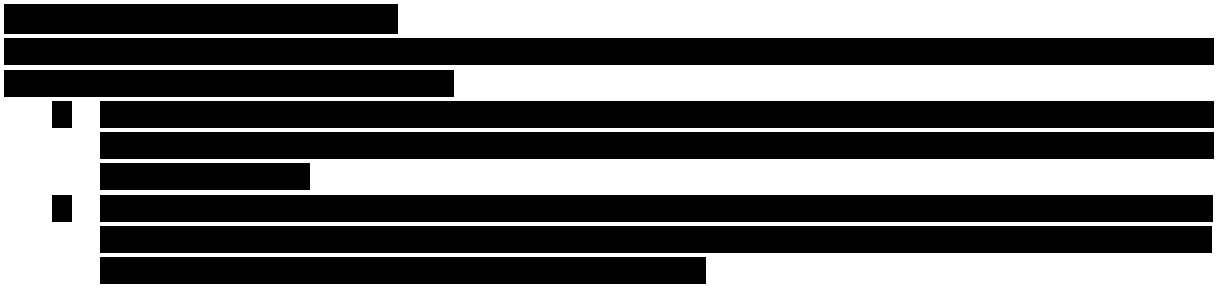
Ulm, den

Nusret Duygu Yilmazer

[redacted]







[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

PUBLICATIONS

1. Yilmazer, N.D. and Korth, M., "Recent progress in treating Protein-Ligand interactions with quantum mechanical methods": Invited review for Int. J. Mol. Sci., *in preparation.*

2. Yilmazer, N.D. and Korth, M., "Semiempirical & molecular mechanics treatment of noncovalent interactions": Invited chapter for *Encyclopedia of Physical Organic Chemistry* (*John Wiley & Sons*), *in preparation.*

3. Yilmazer, N.D., Schwabe, T. and Korth, M., "On the delicate balance of biomolecular interactions and its implication for the design of scoring functions", *submitted.*

4. Yilmazer, N.D. and Korth, M., "Enhanced semiempirical quantum-mechanical methods for biomolecular interactions": Invited mini-review for *Comp. Struct. Biotech. J. (Elsevier)*, ***2015***, 13, 169-175.

5. Husch, T., Yilmazer, N. D., Balducci A. and Korth, M., "Large-scale virtual high-throughput screening for the identification of new battery electrolyte solvents: computing infrastructure and collective properties", *Phys Chem Chem Phys.*, ***2015***, 17, 3394.

6. Yilmazer, N.D., Heitel, P., Schwabe, T. and Korth, M., "Benchmark of electronic structure methods for protein–ligand interactions based on high-level reference data*", J. Theor. Comput. Chem.,* ***2015***, 14, 1540001.

7. Yilmazer N. D. and Korth, M., "Computational approaches for the prediction of solid-electrolyte interface formation", *Bunsen Magazin,* ***2013***, 6, 294. (Invited Article)

8. Yilmazer, N.D. and Korth M., "Comparison of molecular mechanics, semi-empirical quantum mechanical, and density functional theory methods for scoring protein-ligand interactions." *J Phys Chem B,* ***2013***, 117, 8075 – 8084. (Computational Chemistry Highlight, November 2013)

9. Yilmazer, N.D., Fellah M.F. and Onal I., "A DFT Study of Ethylene Hydrogenation Reaction Mechanisms on Ni13 Nanocluster", *Topics in Catalysis,* ***2013***, 56, 789 – 793.

10. Yilmazer, N.D., Fellah M.F. and Onal I., "Ni_{55} Nanocluster: A Density Functional Theory Study of Binding Energy of Nickel and Ethylene Adsorption", *Turk J Chem,* ***2012***, 36, 55 – 67.

11. Yilmazer, N.D., Fellah M.F. and Onal I., "A Density Functional Theory Study of Ethylene Adsorption on Ni10 (111), Ni13 (100) and Ni10 (110) Surface Cluster Models and Ni13 Nanocluster", *Appl. Surf. Sci.,* ***2010***, 256, 5088 – 5093.