Theoretical Prediction of the Interaction between Peptides and Major Histocompatibility Complex II Receptor

Sarah Aldulaijan

A thesis submitted to

Cardiff University

in accordance with the requirements for the degree of

Doctor of Philosophy

School of Chemistry

Cardiff University

May 2012

DECLARATION

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed ....(Candidate) Date: 25-5-2012

STATEMENT 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of PhD

Signed ....(Candidate) Date: 25-5-2012

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed ....(Candidate) Date: 25-5-2012

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ....(Candidate) Date: 25-5-2012

ACKNOWLEDGEMENTS

I would like to thank many people who encouraged me to complete this study. First of all, I would like to gratefully thank my supervisor, Dr James A. Plattes, for his patience, enthusiasm, kindness, outstanding support, and for providing me the opportunity, the knowledge and the guidance to complete this project.

Moreover I would like to thank Prof. Peter Knowles and Dr. Massimo Mella for their advice, support, and encouragement during the project-process meeting. Also, I would like to thank all academic staff of the Theoretical and Computational Chemistry at Cardiff University for their professional academic skills and their friendly personalities during my master and PhD studies which encouraged me to be interesting on this field of science.

Last but not least I would like to thank my family for their endless emotional, moral, and financial support through all my studying years, and for believing in me and giving me the trust to be what I am today.

Publications

“Theoretical prediction of a peptide binding to major histocompatibility complex II”, S. Aldulaijan, J.A. Platts, *Journal of Molecular Graphics and Modelling*, 29, **2010**, 240-245.

“Prediction of Peptide Binding to Major Histocompatibility II Receptors with Molecular Mechanics and Semi-Empirical Quantum Mechanics Methods”, S. Aldulaijan, J.A. Platts, *Journal of molecular Biochemistry*, 1, **2012**, 54-64.

“Quantum Chemical Methods for Calculation of Non-Covalent Interactions in Biological Molecules”, S. Aldulaijan, J.A. Platts, *Current Physical Chemistry*, **2012**, in press.

ABSTRACT

Ab initio, density functional (DFT), semi-empirical and force field methods are used to predict non-covalent interactions between peptides and major histocompatibility complex (MHC) class II receptors. Two *ab initio* methods are shown to be in good agreement for pairwise interaction of amino-acids for myelin basic protein (MBP)-MHC II complex. These data are then used to benchmark more approximate DFT and semi-empirical approaches, which are shown to be significantly in error. However, in some cases significant improvement is apparent on inclusion of an empirical dispersion correction. Most promising among these cases is RM1 with the dispersion correction. This approach is used to predict binding for progressively larger model systems, up to binding of the peptide with the entire MHC receptor, and is then applied to snapshots taken from molecular dynamics simulation. These methods were then compared to literature values of IC_{50} as a benchmark for three datasets, two sets of IC_{50} data for closely structurally related peptides based on hen egg lysozyme (HEL) and myelin basic protein (MBP) and more diverse set of 22 peptides bound to HLA-DR1. The set of 22 peptides bound to HLA-DR1 provides a tougher test of such methods, especially since no crystal structure is available for these peptide-MHC complexes. We therefore use sequence based methods such as SYFPEITHI and SVMHC to generate possible binding poses, using a consensus approach to determine the most likely anchor residues, which are then mapped onto the crystal structure of an unrelated peptide bound to the same receptor. This shows that methods based on molecular mechanics and semi-empirical quantum mechanics can predict binding with reasonable accuracy, as long as a suitable method for estimation of solvation effects is included. The analysis also shows that the MM/GBVI method performs particularly well, as does the AMBER94 forcefield with Born solvation. Indeed, MM/GBVI can be used as an alternative to sequence based methods in generating binding poses, leading to still better accuracy. Finally, we investigated the influence of motion in implicit and explicit solvents for a set of 22 peptides. Binding free energies were calculated by Molecular Mechanics Generalized -Born Surface Area (MM/GBSA) method, but it was found that the results are worse than MM/GBVI on MOE, which show that the MM/GBVI approach can deliver reasonable predictions of peptide-MHC binding in a matter of a few seconds on a desktop computer.

Contents

Table of Contents

Chapter 1:

Introduction:

1. Introduction:

In order to create and develop new drugs, we have to understand the way that a drug interacts with its receptor in order to affect the biological system in the body's cells. One important concept is understanding the chemical interactions between the drug and the receptor [1-3]. Non-covalent interactions have a large influence on many properties of biological molecules [2, 4]. For example, they affect the structure of proteins, DNA and RNA [2, 4], controlling the folding of nucleic acids, molecular recognition and protein-ligand interaction [4, 5]. Proteins, DNA and RNA are crucial for all life, and in order to study their functions, we need to understand their structures [6]. Molecular recognition, “which is one of the most important processes in our life” [4], occurs when a molecule interacts with another at relatively long distances through non-covalent interactions [6]. In addition, understanding the way that a drug or ligand interacts with a protein opens the way to design or develop new drugs [6].

1.1 Non-Covalent interactions:

Non-covalent interactions may be inter- or intra-molecular in nature, and occur when the distance between the subsystems is larger than the typical range for covalent bonds of up to or slightly more than 2 Å [4]. Non-covalent interactions include ionic (or electrostatic) interactions, hydrogen bonds and dispersion-based interactions, which include π-π stacking interactions between aromatic groups. Hydrogen bonding is one of the most important bonds in all chemical cases [7], and was recently re-defined by IUPAC as “an attractive interaction between a hydrogen atom from a molecule or a molecular fragment X–H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule, in which there is evidence of bond formation”[8]. The interaction energies of hydrogen bonds are between -2.4 and -12 kcal/mol [7] (1kcal/mol=4.184

kJ/mol), although interactions such as C—H...O or C—H...π may be rather weaker than that.

Dispersion interactions occur between all compounds, and are particularly significant in those with large polarizability. π-π stacking interactions occur between aromatic groups without overlap of π-orbitals [9, 10]. Hydrogen bonds and stacking are the most important non-covalent interactions in biological complexes such as proteins and nucleic acids [6]. For a long time, hydrogen bonds were considered the most important interactions in such molecules, but with improved theoretical and experimental methods the importance of stacking interactions has clarified, such that stacking interaction is now believed to be equally important as hydrogen bonds [6].

1.2 Biological background:

1.2.1 Major Histocompatibility Complex (MHC):

Major Histocompatibility Complex (MHC) molecules are an important class of receptor in the immune system of all vertebrates: in humans they are termed human leukocyte antigens (HLA). Their role is to bind peptides presented to cell surfaces, hence allowing recognition of self or non-self and stimulating appropriate immune response in the case of non-self. Incorrect recognition of self peptides as being non-self is implicated in a number of auto-immune diseases such as multiple sclerosis and rheumatoid arthritis. The exact mechanism of this is not known but the concept of “molecular mimicry”, in which certain self-peptide sequences are sufficiently similar to non-self sequences to induce immune attack on the body, has been proposed. Prediction of the key binding event between peptide and MHC is therefore desirable, both in understanding the origin of these debilitating diseases and in design of new therapies to treat them.

MHC receptors are generally separated into class I and class II. Both have a single peptide binding site, which in class I is made up of a single amino-acid chain, whereas in class II the active site is located at the junction between two chains, as shown in Figure (1.1)[11, 12]. MHC class I includes heavy chain transmembrane glycoproteins α and β_2-microglobulin ($ß_2$-m), whereas class II consists of two transmembrane glycoprotein chains, α and β Figure (1.2) [12, 13].

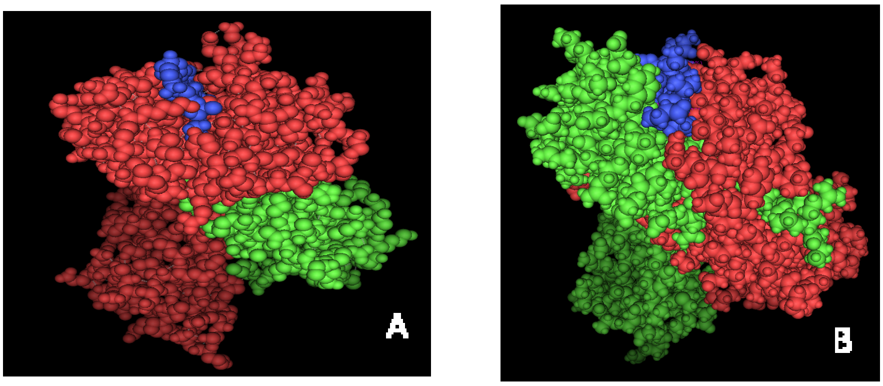


Figure 1.1: MHC class I (A) and class II (B). The bound peptide is shown in blue, the receptor chains in red and green.

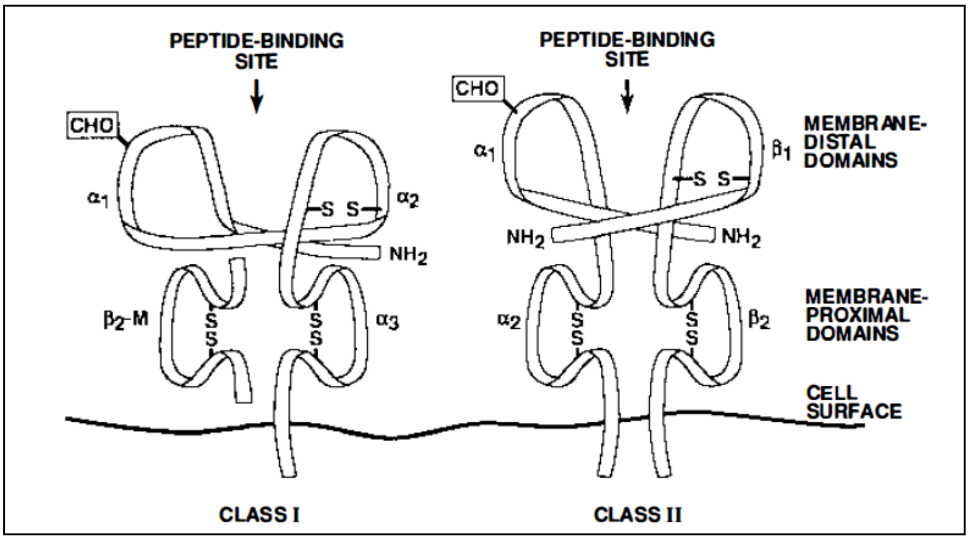


Figure1.2: The MHC classes from reference [13].

The dimensions of the peptide-binding site are around 25Å long, 10Å wide and 11Å deep [13]. Class II can bind with peptides from 13 to 25 residues in length, because the ends of peptide-binding site are open Figure (1.3). In contrast, class I binds with peptides from 8 to 11 residues in length [14], since in class I tryptophan-167 and tyrosine-171 of the A pocket, along with tyrosine-84 in the F pocket, act to "close" the binding site: class II does not include such resides [13]. In addition, peptide binding to the MHC, in both class I and class II, occurs because of particular side chain residues. Many interactions are responsible for MHC-peptide binding, including with peptide side chain as well as between NH_2 and CHO in the main chains of MHC and peptide [13, 14].

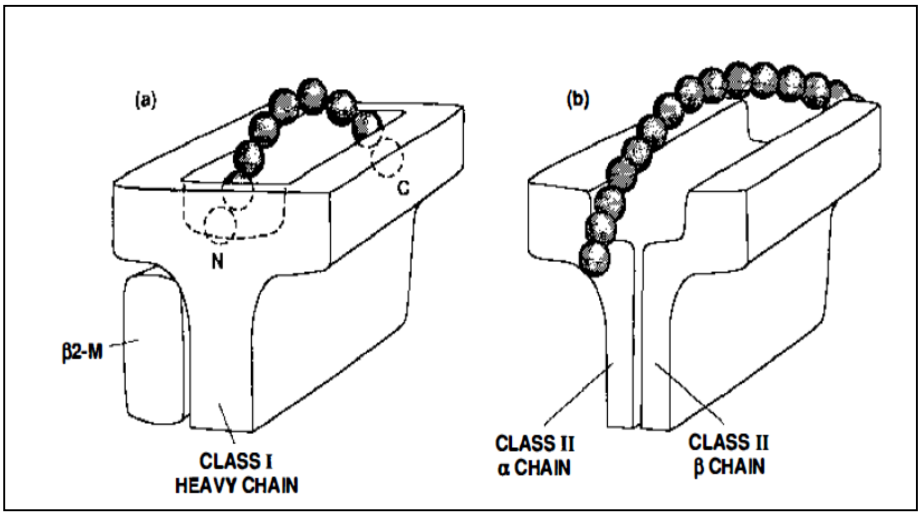


Figure1.3: The MHC classes from reference [13].

The aim of this study is understanding the non-covalent interaction between peptide and MHC class II receptor with the ultimate goal of designing drugs for Multiple Sclerosis (MS) and other auto-immune diseases.

1.3 Theoretical background:

Experimental studies of relatively weak non-covalent interactions can be difficult, so computer simulation is an important and rapidly expanding field.

1.3.1 *Ab initio*, or Wavefunction Methods:

Ab initio methods, also known as wavefunction theory (WFT), attempt to solve from first principles the electronic structure of atoms and molecules. The simplest such method is Hartree-Fock (HF), which averages electron-electron repulsion and is the typical starting point of more advanced methods. The details of electron-electron repulsion that are lost in HF treatment are termed "electron correlation", and most *ab initio* methods seek accurate calculation of correlation energies [4]. Such methods are typically used to calculate interaction energies *via* the supermolecular approach, in which the interaction energy is calculated as the difference between the energy of the complex and the energy of all subsystems in the super-molecular methods. This requires that the method is size consistent, *i.e.* that the energy of two atoms or molecules at infinite separation equals the sum of the energy of each atom or molecule. *Ab initio* methods that are size consistent include many-body perturbation theory, and full configuration interaction (CI), whereas truncated CI is not size consistent and so is not suitable for description of many non-covalent interactions.

The most accurate WFT method in common use for interactions in biological systems is the coupled cluster (CC) ansatz, which is size-consistent. CC methods can incorporate electronic excitations to any level, but these are typically truncated to triples, the so-called CCSDT method. This method , which includes single, double and triple electron excitations, is highly accurate but very computationally expensive and so only applicable to small model systems, especially in calculating dispersion interaction energy where large orbital basis sets are required [15]. The CCSD(T) approach, in which triple excitations are included perturbatively, provides almost identical accuracy to CCSDT method with much reduced computational cost, and hence is widely used for studying hydrogen bonded and stacked complexes at less cost [4, 6, 16, 17]. This method is now widely known as the "gold standard" of single-reference calculations [6]. It has been shown that a CCSD(T) approach with complete basis set (CBS) performs very well in studying all types of non-covalent interactions, including those with importance in biological complexes such as hydrogen bond and dispersion [6]. Unfortunately, this method can be applied to systems containing approximately 24 atoms or less with AO basis sets approaching the basis set limit [4, 6, 17-19], but scaling like N^7 (where N is the number of electrons in the system) means that this approach rapidly becomes unfeasible. This method is therefore used as a benchmark for studying small systems [20] and for testing the performance of other methods [4, 6, 21].

CCSD(T) calculations of extended complexes can be achieved by using "a medium basis set of a DZP-quality", (DZP = double-ζ + polarisation) but unfortunately the cost for such calculation is high [4]. The problem was solved by approximating the CBS CCSD(T) interaction energy by using the difference between CCSD(T) and MP2 interaction energies ($\Delta E^{CCSD(T)}$ - ΔE^{MP2}) eq (1.1), because this difference is less dependent on the basis set than either CCSD(T) or MP2 alone.

$$\Delta E_{CBS}{}^{CCSD(T)} = \Delta E_{CBS}{}^{MP2} + (\Delta E^{CCSD(T)} - \Delta E^{MP2})\,\big|_{\text{medium basis set}} \qquad (1.1)$$

In order to calculate $\Delta E_{CBS}{}^{MP2}$, many extrapolation schemes have been published, such as that by Helgaker et al [22], which is widely used. The extrapolation is often done by using "systematically improved basis sets" such as aug-cc-pVDZ - aug-cc-p-VTZ or aug-cc-pVTZ-aug-cc-pVQZ.

An alternative method for large systems is second order Møller-Plesset perturbation theory (MP2) [4, 15, 20, 23]. MP2 is widely used for studying non-covalent interaction energies [6] because of relatively good accuracy and computational cost, compared with CCSD(T) method. MP2 performs very well in calculating the energy of H-bond, but it significantly overestimates dispersion energy [4, 6, 15, 17, 20, 24, 25], which is one of the major drawbacks of this method. MP2 interaction energies are also strongly basis-set dependent [4, 6]. For example, by using a DZP- quality basis set, this method provides reliable interaction energies for non-covalent interactions, whereas these can be significantly overestimated using larger basis sets [4]. For complexes mainly bound by dispersion energy, it has been shown that MP2 with small basis sets with expanded polarisation functions is successful in calculating dispersion energies, for example MP2/6-31G(0.25)d, the results are close to CCSD(T) as benchmark data [6, 25, 26]. Moreover, MP2 with such small basis sets can be applied to molecules with hundreds of atoms [6]. It has been found that the most accurate method, after CCSD(T), for calculating interaction energies between amino acid residues is MP2/aug-cc-pVTZ [27]. CCSD(T) and MP2 methods "do not contain any parameters, but their performance for non-covalent interaction is limited" [6].

The problem of MP2, overestimated the dispersion binding energy, was solved to some extent by the "MP2.5" approach, in which the correlation energy from MP2, which overestimates dispersion energy, and MP3, which underestimates dispersion energy, is averaged, leading to much improved performance in calculating non-covalent interactions [15, 28, 29]. However, the computational cost of the MP3 step means that MP2.5 is time-consuming when compared with MP2, although it is still much less computationally demanding than CCSD (T) method [6]. Unfortunately, MP2.5 can not be applied for large systems [6, 15].

Werner *et al* have significantly improved the performance of and reduced CPU time required for MP2 in calculating large systems, naming their method density-fitted local MP2 (DF-LMP2) [30]. DF-LMP2 makes use of the local nature of electron correlation to reduce the computational resources required for MP2 calculations. Importantly for the study of non-covalent interactions, this also effectively eliminates basis set superposition error (BSSE), thereby removing the need for potentially expensive counterpoise corrections [31-34]. This method is successful in studying large systems especially when suitable basis sets are used [30].

Spin-component scaled MP2 (SCS-MP2) is a modification of MP2 method introduced by Grimme [35], with the aim of improving the systematic errors in results of MP2. One such systematic error is the overestimation of the dispersion contribution to binding energy, as noted above. This method depends on separation of the correlation energy resulting from anti-parallel and parallel spin pairs of electrons, and assigns two new scaling factors to these contributions. This method successfully improves the performance of MP2 method for dispersion energy in some cases, where the standard MP2 is overestimated [4-6, 35] . Unfortunately, this method leads to worse performance in calculation of hydrogen bond interaction energies [5, 6, 21, 36, 37].

This failure of SCS-MP2 method led to development of other methods, such as SCSN-MP2 and SCS(MI)-MP2, in which the error in interaction energy was reduced by re-parameterization of the scaling factors. Spin-component scaled for nucleobases (SCSN) was developed by optimising the scaling factors against binding energy of stacked nucleobases, and totally neglects the anti-parallel-spin electron pairs contribution and gives the parallel spin contribution a scaling of 1.76 [5]. For a set of 22 complexes containing hydrogen bonded, dispersion bound, and mixed complexes (the widely-used “S22” data set), SCSN gives an error of 0.3 kcal/mol relative to CCSD(T) benchmark [5]. SCS-MP2 for molecular interactions (SCS(MI)-MP2) is an improvement of SCS-MP2 approach [27, 38]. In this method the parallel-spin contribution and anti-parallel spin contribution scaled 1.29 and 0.40, respectively [38]. It has been found that the performance of SCS(MI)-MP2/cc-pV(DT)Z is very close to CCSD(T) results which are the benchmark data to date [38]. The speed and the accuracy of this method make this method suitable to study non-covalent interactions in large systems [27, 38]. SCS(MI)-MP2 performs very well in calculating the interaction energies between amino acid residues [27]. For the S22 data set, this method provides accurate results, better than MP2 and SCS-MP2 methods [6]. The RMSD errors for SCS(MI)-MP2, MP2, and SCS-MP2 are 0.31, 0.99, and 1.45 kcal/mol, respectively [6].

Explicitly correlated MP2 (MP2-F12) is a new method which provides accurate results in calculating the interaction energy for non-covalent interaction [15, 39] . This is achieved by explicitly including a term for electron-electron repulsions in the Hamiltonian, which is a function (F12) or the inter-electron distance r_{12}. For S22 benchmark data, the results of interaction energies calculated by this method with aug-cc-pVDZ basis set are more accurate than by the conventional MP2/AV5Z approach, and MP2-F12/AVTZ and MP2-F12/AVQZ provide similar results to the CBS predictions [39]. The success of MP2-F12 leads to present CCSD(T)-F12a approach [15, 39]. Explicitly correlated couple-cluster method with augmented double-ξ basis sets provides more accurate results – and less cost- than CCSD(T) standard method with same basis set [15, 39-41].

1.3.2 Density Functional Theory:

Density functional theory (DFT) is widely used in studying large systems because of the relatively low cost and overall good performance [4, 42, 43]. The fundamental theorems of

Hohenberg and Kohn [44] show that the ground state total energy of a system can be calculated exactly from knowledge of the electron density alone. However, the exact functional that would allow one to do this is not known, so current implementations of DFT calculate the interaction energy using approximate exchange and correlation functionals. A great many such functionals of varying complexity exist, employing the electron density and its gradient, along with exact (HF) exchange in so-called hybrid functionals. For hydrogen bond energy, DFT shows reliable results comparing with reference data [4]. Unfortunately, most standard methods in DFT fail in calculating non-local dispersion energy [4, 6, 20, 25-27, 45-48]. The importance of dispersion energy in biological complexes means that this method is not suitable in such calculations.

Many attempts have been made in order to include the dispersion energy in HF and DFT calculations. One successful approach has been reached by calculating a dispersion term separately by means of a damped C_6R^{-6} formula, where R is interatomic distance and C_6 is a dispersion coefficient, [49-52] then adding it to HF and DFT calculations eq (1.2) [4, 20, 43].

$$E_{MF\text{-}D}=E_{MF} + E_{disp} \qquad (1.2)$$

Where E_{MF} is mean field energy (HF or DFT), and E_{disp} is dispersion correction. Early studies showed that by adding a dispersion correction to HF energy, calculation of binding energy of rare-gas and larger complexes can be successful [53-56]. This success led to many groups adding a dispersion correction term to DFT energy [49, 57-59].

Grimme “recognized the need for the dispersion to be adjusted for a given functional form and introduced a simple scaling factor optimized for each particular density functional” [20].

This method succeeds in studying non-covalent interactions, especially dispersion and hydrogen-binding energies [20, 27, 52]. For the S22 data set, the root mean square error (RMSE) for DFT (B3LYP) with TZVP basis set, with and without dispersion term, are 0.82 and 3.28 kcal/mol, respectively. [60, 61]. Hobza and co-workers developed an alternative damping function using accurate CCSD(T)/CBS data. [61]. The parameters were modified to recover the dispersion energy in stacking and hydrogen bonded systems.

In addition to the DFT-D approach, several DFT methods show promise for description of at least some classes of non-covalent interactions. Hybrid meta-exchange-correlation functional methods such as M05-2X and M06-2X present accurate calculation for non-covalent interactions, especially dispersion interactions [62-64]. The MUE error for dispersion-bound complexes for M06-2X is 0.2 kcal/mol when compared with best estimated values for peptides. These methods have been recommended to be used in biochemistry calculations [6, 62, 63].

Becke’s Half-and-Half functional (BHandH) “contains an equal mixture of the exact Hartree-Fock and local density approximation for the describing of exchange energy, coupled with Lee, Yang and Parr’s expression for the correlation energy”. This method provides reliable results in calculating stacking energy [6, 9, 24, 65, 66], but overestimates the hydrogen binding energy [6, 9, 24, 65, 66]. In the S22 data set, the mean unsigned error (MUE) is 5.54 kcal/mol for hydrogen bonded complexes, compared to 0.84 kcal/mol for dispersion-bound complexes. Another DFT improvement is PW91 functional [67]. This method provides well for stacking. However, this method, similar to BHandH, fails in describing hydrogen-bond energy [4, 6, 9, 67].

Non-local van der Waals density functional (vdW-DF) was created in order to include the energy of van der Waals interaction in DFT calculations [68, 69]. It has been found that vdW-DF with semilocal functional revPBE exchange performs very well in calculating the van der Waals binding energy [48]. One disadvantage of this method, however, is that with hybrid exchange functionals it results in significant overbinding of many complexes [48], and that this method is rather slow, such that more work is needed in order to improve it [70, 71].

Dispersion-corrected atom-centred potential (DCACP) is another method in the direction to include the dispersion energy in DFT calculations [72], which depends on the use of pseudopotentials [72]. DCACP is an empirical method, where the dispersion attraction between electrons and nuclei occurs by addition of an artificial potential which is optimized to provide the interaction energy. A disadvantage of this method is that it can not be used without pseudopotentials.

DFT normally scales as N^3 or N^4, where N is the number of electrons in the system, such that calculations on large systems such as biological molecules rapidly become unfeasible. Linear scaling of DFT methods would be a highly desirable property, allowing application to much more realistic models of biological systems. Recent developments in the ONETEP (order-*N* electronic total energy package) density functional package bring this goal within reach [73]. Specifically designed for use on parallel computers, and with careful control of accuracy and errors due to approximations in the linear scaling process, [74] ONETEP speeds up DFT calculations to allow calculation of the energy of large systems containing thousands of atoms [73].

1.3.3 Semi-Empirical Methods:

The need for accurate and fast methods to calculate large systems has been realized for more than three decades, when the use of non-empirical *ab initio* methods was prohibitive [6, 75, 76]. Dewar and Thiel introduced their first semi-empirical method (modified neglect of differential overlap, or MNDO) [76, 77], and the Austin Model 1 (AM1) method subsequently developed by "adding a stabilization Gaussian function to the MNDO core-core interaction" [78, 79]. AM1 shows good performance in many chemical studies, and has been widely used since then [80]. It is available in many chemical software packages. In 1989, Stewart improved the techniques of parameterization and published parameterized model 3 (PM3), which gave "lower average errors than AM1, mainly for the enthalpies of formation" [80-82]. PM4 and PM5 are subsequent improvements on PM3, and both present good performance across many areas of chemistry [80]. Semi-empirical methods perform well [83] in calculating covalent interactions of the main group molecules in ground states [6]. However, semi-empirical methods often fail in calculating the non-covalent interactions, especially, dispersion and hydrogen bond interaction energies [42, 80].

At first, it was believed that dispersion energy is more important than hydrogen bonding [6], so all attention was focused on correction of dispersion. Similar to DFT-D method discussed above, a damped dispersion term was added to AM1 and PM3 energies, and some parameters of each semi-empirical method modified [6, 42, 43]. The resulting AM1-D and PM3-D give reliable results in predicting the interaction energy between biological systems with an error average 1 to 1.5 kcal/mol of the results of high-level *ab initio* methods [42, 43]. In addition, another advantage is these calculations can be carried out with same speed as AM1 and PM3, making them useful in calculating large biological systems [42]. Semi-empirical methods with dispersion correction succeed in calculating dispersion energy, but their performance for hydrogen bond energies is limited [42, 43, 80].

OMx (orthogonalization models OM1, OM2, and OM3) semi-empirical method present very accurate results in calculation biological systems, better than standard semi-empirical methods due the use of orthogonalization corrections [6, 84]. With a dispersion term and without modification of the parameters, OMx-D methods are highly successful in calculation biological systems [84]. Table (1.1) shows the mean unsigned error (MUE) for the hydrogen bonded complexes, stacked base pairs, the S22 set, and the JSCH-2005 set (a larger set of peptide and nucleic acid complexes) taken from [84].

	AM1	OM1	OM1-D	OM2	OM2-D	OM3	OM3-D
Hydrogen-bonded complexes	14.78	10.25	6.99	5.67	2.41	5.72	2.54
stacked base pairs	10.67	6.68	1.08	6.36	1.52	6.53	1.26
S22	7.02	5.32	2.52	3.28	1.15	3.81	1.23
JSCH-2005	8.67	6.04	2.35	4.55	1.41	4.77	1.31

Table 1.1: MUE of AM1, OMx and OMx-D for hydrogen-bonded complexes, stacked base pairs, S22 set, and JSCH-2005 in kcal/mol (from ref [84]).

This table shows that OMx methods, even without the dispersion term, perform rather better than AM1. It also shows that OM3 performs better than OM2, which in turn is better than OM1 for all complexes. In addition, it shows that by adding a dispersion term, all methods results improved, especially for stacked base pairs. OM2-D and OM3-D provide similar results. For S22 set, the MUE for OM2-D and OM3-D are 1.15 and 1.23 kcal/mol, respectively. For JSCH-2005, the MUE for OM2-D and OM3-D are 1.41 and 1.31 kcal/mol, respectively.

Self-consistent charge density functional tight-binding (SCC-DFTB) method is a semi-empirical method based on DFT approach [85]. This approach applies a minimal valence basis set, and shows similar computational speed to semi-empirical methods [86]. SCC-DFTB provides good results in calculating many biological molecules such as peptides and nucleic acids [87]. This method, augmented with a dispersion term (SCC-DFTB-D) is a successful method in calculating non-covalent interactions. It can be used to study large systems with several thousands of atoms [6, 88]. It has been found that the root mean square error (RMSE) for hydrogen complexes from S22 benchmark set for this method is 1.53 kcal/mol, for dispersion complexes is 0.82 kcal/mol, and for mixed complexes is 0.86 kcal/mol [89].

In 2006, RM1 (Recife Model 1) was published [78]. This is a new parameterization method of AM1 [78, 80, 83], and can be easily used in any software that contains AM1 [80]. The parameterization set of this method contains 1736 important molecules in biochemistry containing ten atoms (C, H, N, O, P, S, F, Cl, Br, and I) using only the s-p basis set [78, 90], with all parameters re-optimized [80]. RM1 method performs better than related methods, and is therefore widely chosen in modeling organic compounds [78], and is successful in studying large molecules when used with MOZYME (see below). RM1 is not just considered an overall improvement over AM1, but also over PM3 [80, 83, 91], and corrected a problem with nitrogen charge in PM3 [80]. The average error for this method comparing with AM1, PM3, and PM5 for calculating the enthalpies of formation, dipole moment, ionization potential, bond length, and angles for the all 1736 molecules comparing with reference data are shown in Table (1.2). It is clear that the errors of RM1 in all properties (except for the bond angles) is the smallest [80]. However, while RM1 improves the accuracy of calculating non-covalent interactions in biological systems, its results are not perfect.

Properties	AM1	PM3	PM5	RM1	N
Enthalpies of formation (kcal/mol)	11.15	7.98	6.03	5.77	1480
Dipole moment (D)	0.37	0.38	0.5	0.34	127
Ionization potential (eV)	0.6	0.55	0.48	0.45	232
Bond length (A)	0.036	0.029	0.037	0.027	904
Angles (degree)	5.88	6.98	9.83	6.82	910

Table 1.2: MUE (kcal/mol) in properties for 1736 molecules for AM1, PM3, PM5, and RM1 taken from ref [80]. N is the number of quantities used in the comparison.

One drawback of RM1 method has been solved in the RM1-BH method, where BH stands of biological hydrogen bonding. This was achieved by adding one more Gaussian function to the core-core repulsive term in the semi empirical formula for atoms contributing to hydrogen bonds [83]. The prediction of RM1-BH for hydrogen-bond interaction energy is very promising when compared with the methods discussed above [83]. The MUE (kcal/mol) for PM3, RM1, and RM1-BH comparing with the MP2 method as a benchmark for 35 hydrogen bonded base pairs, hydrogen bonding amino acid residues, and hydrogen bonding protein-nucleic acid complexes are shown in Table (1.3).

Complexes	PM3	RM1	RM1-BH
Base-pair dimers	6.4	5.5	1.7
Amino acid residue dimers	5.6	4	1.6
Dimers between a base and an amino acid residue	6.7	6.5	1.9

Table 1.3: MUE (kcal/mol) for PM3, RM1, and RM1-BH for 35 hydrogen bonded base pairs, hydrogen bonding amino acid residues, and hydrogen bonding protein-nucleic acid complexes (taken form ref [83]).

For the base pair dimers, it is clear that PM3 and RM1 underestimate hydrogen bonding energy, with MUE 6.4 and 5.5 kcal/mol, respectively [83], while RM1-BH gives reliable results (MUE 1.7 kcal/mol) comparing with MP2 benchmark data [83]. The same conclusion for the amino acid residues, and protein-nucleic acid complexes, PM3 and RM1 underestimate the hydrogen binding energy for these sets [83]. For amino acid residues set, the MUE for PM3 and RM1 are 5.6 and 4 kcal/mol, respectively [83]. For protein-nucleic acid complexes, the MUE for PM3 and RM1 are 6.7 and 6.5 kcal/mol, respectively [83]. From these results, it is clear that RM1-BH overall improves performance over RM1 and PM3 methods in calculating hydrogen bonding energies for biological molecules [83].

After several incremental improvements, the most recent semi-empirical method is PM6 (parameterized model 6), published by Stewart in 2007 [78], which has parameters for 80 atoms, covering most of the periodic table. PM6 is the most accurate method among the semi-empirical method up to date [6, 92]. It performs better than high level method such as HF or B3LYP DFT with 6-31G(d) basis set in calculating heats of formation [78]. For the HF and B3LYP set, the average unsigned errors for HF, B3LYP, and PM6 are 7.4, 5.2, and 4.4 kcal/mol, respectively [78]. Although PM6 is the most accurate method among semi-empirical methods, its performance in calculating non-covalent interactions, especially dispersion and hydrogen bonding, was not as accurate as might have been expected [6, 78]. Even by adding a dispersion term, as in DFT, its performance for non-covalent interactions does not improve, as shown in Table (1.4) [93].

Three generations of correction for hydrogen bonds have been published: PM6-DH1, PM6-DH2, and PM6-DH+ [92-94]. These methods contain two correction terms, one for dispersion and one for specific hydrogen bond interactions [6]. The first generation correction (PM6-DH1) was successful in calculating dispersion energy, but for hydrogen bond energy the performance was disappointing [6]. Although PM6-DH1 approach shows a large improvement in studying hydrogen bond energy and opened a new path in hydrogen bond corrections[93], it suffers some problems [92-94]. These problems were solved in the second generation of the hydrogen bond correction PM6-DH2 [93, 94]. This method succeeds in calculating hydrogen bond energies with accuracy close to DFT-D approach, but is three orders of magnitude faster [93, 94]. A disadvantage of this method is that it fails when the acceptor atom changes, such as the case in proton transfer [92-94]. PM6-DH+ solves the drawbacks of the first and second generations of the correction with very close accuracy to PM6-DH2, as shown in Table (1.4) [92-94], and also solving the limitations of PM6-DH2 for proton transfer.

Complexes	PM6-D	PM6-DH2	PM6-DH+
H-bonded complexes from S26 set	3.56	0.27	0.88
H-bonded complexes from JSCH2005 set	6.3	2.23	1.59

Table 1.4: RMSE (kcal/mol) for PM6-D, PM6-DH2, and PM6-DH+ for H-bonded complexes from S26 and JSCH2005 sets taken from [93].

The RMSE for H-bonded complexes from S26 and JSCH2005 benchmark data set for PM6-D, PM6-DH2, and PM6-DH+ are shown in table (1.4). From these results, it is clear that adding a dispersion term to PM6 does not improve results for hydrogen-bonding interactions. Improvement for hydrogen-bonding interactions is achieved by adding the hydrogen correction as mentioned above. For the H-bonded complexes from S26, the RMSE errors are 0.27 and 0.88 kcal/mol for PM6-DH2 and PM6-DH+, respectively, *i.e.* both methods perform very well. From these results, it is evident that PM6-DH2 performs better than PM6-DH+. This is because the S22 set was used to parameterize this method, so this accuracy does not reach beyond this set. This is clear in the results for hydrogen-bonding complexes from JSCH2005 set. The RMSE errors are 1.59 kcal/mol for PM6-DH+ and 2.23 kcal/mol for PM6-DH2 approach.

The applicability of semi-empirical methods to large systems is further enhanced by the MOZYME method implemented in current versions of MOPAC [95]. PM6 with MOZYME can be used to calculate large systems [78]. MOZYME uses localized molecular orbital instead of the standard SCF procedure [78, 95].

1.3.4 Basis Set Superposition Error and Perturbation Theory:

All the supermolecular approaches discussed above suffer to some extent from basis set superposition error (BSSE). The interaction energy can be calculated using eq (1.3) only when infinite basis set is used:

$$\Delta E = \Delta E^{R-T} + (E^{R} - E^{T}) \qquad (1.3)$$

With finite basis sets, the supermolecular calculation on the complex uses all basis functions of all subsystems, whereas calculations on constituent monomers do not. This leads to better but non-physical description of the complex than of the monomers, and the error is called BSSE [4]. To solve this error, the counterpoise procedure of Boys and Bernardi may be applied [96], in which each monomer is calculated in the full complex basis set. BSSE becomes smaller by using expanded basis sets, and the error typically becomes negligible by using extended basis set such as cc-pVQZ [4]. It has also been shown [21] that use of augmented, diffuse functions on hydrogen atoms exacerbates BSSE, such that a basis set consisting of aug-cc-pV*n*Z on heavy atoms and cc-pV*n*Z on hydrogen can reduce BSSE with little or no reduction in accuracy. BSSE is generally larger in correlated *ab initio* methods than in Hartree-Fock or DFT approaches, as the former depend on virtual as well as occupied orbitals, which are often much more diffuse. One exception to this general rule is local correlation methods, where restriction to spatially close virtual orbitals reduces BSSE to negligible levels [21]. Semi-empirical methods are not thought to suffer significantly from BSSE, since spatially restricted minimal basis sets are usually employed.

An alternative to the supermolecular approach to calculation of non-covalent interaction energies is symmetry adapted perturbation theory (SAPT) [15, 89, 97]. This is a highly accurate method to calculate the interaction energy directly without the disadvantages of BSSE [89]. The interaction energy is calculated as the sum of first-order (electrostatic and exchange energies), second-order (induction and dispersion energies), and higher-order (charge transfer energy) contributions. While it is highly accurate, the computational cost of a standard SAPT approach is comparable to CCSD(T), and so is only applicable to small model systems. [4, 15, 97].

Further success has been gained by using a combination of SAPT and DFT methods with extended basis sets [89, 98-107]. The method becomes able to calculate molecules such as benzene dimer and DNA base pairs [101, 106]. This (DFT)SAPT method, also known as SAPT-DFT, calculates the interaction energy as eq (1.4)

$$E_{int} = E^1_{pol} + E^1_{ex} + E^2_{ind} + E^2_{ex\text{-}ind} + E^2_{disp} + E^2_{ex\text{-}disp} + \delta\,(HF) \qquad (1.4)$$

$E^1_{pol} + E^1_{ex}$ are the first order contributions, and include electrostatic (polarization) and exchange components. $E^2_{ind} + E^2_{disp}$ are the second order contributions, which contain induction and dispersion components. $E^2_{ex\text{-}ind}$ and $E^2_{ex\text{-}disp}$ are the exchange counterparts of the induction and dispersion. Higher-order contributions are replaced with δ (HF), which allows this method to calculate intra-molecular correlation in DFT level, while the inter-molecular interaction treated in SAPT method [4].

1.3.5 Atomistic force fields:

Atomistic force fields are widely used in simulation of biological systems by reducing the essentials of systems of interest to simple mathematical forms based on Newtonian mechanics. Non-covalent interactions are typically treated by a combination of point charges, to account for electrostatics, and Lennard-Jones potentials, for dispersive and repulsive interactions [108]-[109]. More than a decade ago, Hobza et al showed that the force field of Cornell et al (often referred to as AMBER) best reproduced *ab initio* data for interaction of DNA base pairs [108]. More recently, Paton and Goodman showed that the OPLS-AA force field performs well for binding energy prediction of both hydrogen bonding and dispersion-bound complexes [109].

The generalized Born model/surface area approach (GB/SA) is another method used to calculate binding free energy, developed by Still et al [110-112], and is widely used in calculating free energy of binding for ligand-receptor complexes [113-115]. In this method, cavitation energy depends on molecular surface area, while relative solvation of separated ligand and receptor compared to their complex is estimated from a generalization of the Born model. When combined with MM methods for calculation of electrostatic and van der Waals interactions, these are referred to as MM-GB/SA methods. The GB/VI (generalized Born/volume integral) model, implemented in recent versions of MOE software [116], is similar to GB/SA in most respects, but calculates the cavitation energy as an integral over molecular volume rather than surface area [112]. MM/GB-VI is a fast and promising method to calculate the interaction energy in solvent. There are many advantages of using this method, the dielectric constant of the solvent is estimated based on the atoms [116] present in the specific complex under study, rather than on idealised values. In addition, this method yields an estimate of binding free energy, unlike all other methods used here that give only interaction energies. The change in entropy on binding is not explicitly included in MM/GB-VI: it has previously been shown that although entropy is essential in calculating absolute binding free energy [114]. it is not essential for estimating the relative binding free energy[117, 118].

1.3.6 Implicit and explicit solvation models:

Solvent can have an important influence on the structure and properties of biomolecules, so in many cases one must consider the effect of solvent in calculations. There are two main methods to predict the solvation effect on biomolecular properties, [119-121] namely implicit (or continuum) and explicit solvation models. Explicit solvent models include the solvent molecule explicitly in the calculation, while the implicit solvent models replace the solvent molecules with a dielectric continuum or similar medium [122, 123]. Each method has its advantages and disadvantages: while explicit solvent models consider the solvent effect in the highest levels of detail, they can be expensive and time consuming. On the other hand, implicit solvation models are able to ‘‘pre-average’’ solvent effect and therefore reduce the need for computationally expensive sampling, making these model widely used in studying biomolecules [124]. Conductor-like Screening Model (COSMO) is a continuum method which is widely used to model solvents, especially water [113, 125, 126]. This method depends on generation of a conducting surface at vdW distance in order to calculate the dielectric screening charges and energies [125].

There are, however, some major disadvantages of implicit models, which cannot treat the effect of hydrogen bond between solvent and solute [127, 128]. According to Suhai et al, continuum solvation models do not provide the correct geometry of some simple biological molecules such as alanine dipeptide, while explicit models do. Because both models (implicit and explicit) have their strengths and weaknesses, a combination of both methods were used by Chalmet et al [129]. In this approach, the first shell of the solvent, which has different effect on the solute from the bulk, was treated explicitly while the remainder was treated as continuum model. However, it has been found that such mixed solvent models do not necessarily give more accurate performance than either pure implicit or explicit models [130].

1.4 Conclusions

Methodological developments, coupled with ever-improving computer hardware, mean that theoretical methods are increasingly used to probe the non-covalent interactions in biological molecules. Benchmark *ab initio* methods, as well as those based around intermolecular perturbation theory, can still only be applied to relatively small model systems, but new theoretical developments mean that applicability is always increasing. Density functional theory requires less computational resources, but only recently have such methods been able to properly balance the importance of electrostatic and dispersion-based interactions. Semi-empirical methods, especially when used together with empirical dispersion and/or hydrogen bonding corrections, and force field methods show much promise for treatment of entire proteins or nucleic acids.

1.5 References:

[1] Mantzourani, E.; Laimou, D.; Matsoukas, M. T.; Tselios, T. *Anti-Inflammatory & Anti-Allergy Agents in Medicinal Chemistry* **2008**, *7*, 294-306

[2] Zhao, Y.; Truhlar, D. *Journal of Chemical Theory and Computation* **2007**, *3*, 289-300

[3] Meyer, E. A.; Castellano, R. K.; Francois Diederich. *Angewandte chemie* **2003**, *42*, 1210-1250

[4] Cerny, J.; Hobza, P. *Physical Chemistry Chemical Physics* **2007**, *9*, 5291.

[5] Hill, J. G.; Platts, J. A. *Journal of chemical theory computation*, **2007**, *3*, 80.

[6] Riley, K. E.; Pitonak, M.; Jurecka, P.; Hobza, P. *Chem. Rev*, **2010**, *110*, 5023.

[7] Alkorta, I.; Rozas, I.; Elguero, J. *Chem. Soc. Rev.*, **1998**, *27*, 163.

[8] Arunan, E.; Desiraju, G. R.; Klein, R. A.; Sadlej, J.; Scheiner, S.; Alkorta, I.; Clary, D. C.; Crabtree, R. H.; Dannenberg, J. J.; Hobza, P.; Kjaergaard, H. G.; Legon, A. C.; Mennucci, B.; Nesbitt, D. J. In *http://media.iupac.org/reports/provisional/abstract11/arunan_prs.pdf*, **2010**.

[9] Waller, M. P.; Robertazzi, A.; Platts, J. A.; Hibbs, D. E.; Williams, P. A. *J. Comput. Chem*, **2006**, *27*, 491.

[10] Grimme, S. *Angewandte Chemie-International Edition*, **2008**, *47*, 3430.

[11] Mantzourani, E. D.; Mavromoustakos, T. M.; Platts, J. A.; Matsoukas, J. M.; Tselios, T. V. *Current Medicinal Chemistry* **2005**, *12*, 1521-1535

[12] Wearsch, P. A.; Cresswell, P. *Current Opinion in Cell Biology* **2008**, *20*, 624-631

[13] Barber, L. D.; Parham, P. *Annual Review of Cell Biology* **1993** *9*, 163-206

[14] Yague, J.; Marina, A.; Vazquez, J.; Castro, J. A. L. d. *Journal of Biological Chemistry* **2001**, *276*, 43699-43707

[15] Werner, H.-J.; Marchetti, O. *J. Phys. Chem. A*, **2009**, *113*, 11580.

[16] Pittner, J.; Hobza, P. *Chem. Phys. Lett.*, **2004**, *390*, 496.

[17] Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Physical Chemistry Chemical Physics*, **2006**, *8*, 1985-1993.

[18] Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M. *J. Chem. Phys.*, **2005**, *122*, 144331.

[19] Sinnokrot, M. O.; Sherrill, C. D. *J. Am. Chem. Soc.*, **2004**, *126*, 7690.

[20] Grimme, S. *Jounral of Comutational Chemistry*, **2004**, *25*, 1463.

[21] Hill, J. G.; Platts, J. A.; Werner, H.-J. *Physical Chemistry Chemical Physics*, **2006**, *8*, 4072-4078.

[22] Halkier, A.; Helgaker, T.; Jorgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.*, **1998**, *286*, 243.

[23] Friesner, R. A. *Proc Natl Acad Sci U S A*, **2005**, *102*, 6648.

[24] Gkionis, K.; Hill, J. G.; Oldfield, S. P.; Platts, J. A. *Journal of Molecular Modeling*, **2009**, *15*, 1051.

[25] Sponer, J.; Riley, K. E.; Hobza, P. *Physical Chemistry Chemical Physics*, **2008**, *10*, 2595.

[26] Hobza, P.; Sponer, J.; Reschel, T. *J. Comput. Chem.*, **1995**, *16*, 1315.

[27] Berka, K.; Laskowski, R.; Riley, K. E.; Hobza, P.; Vondrasek, J. *J. Chem. Theor. Comput.*, **2009**, *5*, 982.

[28] Pitonak, M.; Hesselmann, A. *J. Chem. Theory Comput*, **2009**, *6*, 168.

[29] Pitonak, M.; Neogrady, P.; Cerny, J.; Grimme, S.; Hobza, P. *Chem Phys Chem*, **2009**, *10*, 282.

[30] Werner, H.-J.; Knowles, P. J.; Manby, F. R. *Jounral of Chemical physics*, **2003**, *118*, 8149-8160.

[31] Pulay, P. *Chemical physical letters*, **1983**, *100*, 151-154.

[32] Saebo, S.; Pulay, P. *Annual Review of physcal chemistry*, **1993**, *44*, 213-236.

[33] Tateno, M.; Hagiwara, Y. *Journal of physics: Condens matter*, **2009**, *21*.
[34] Werner, H.-J.; Hartke, B.; Schutz, M. *Chem. Phys.*, **1998**, *239*, 561.
[35] Grimme, S. *Journal of Chemical Physics*, **2003**, *118*, 9095-9102.
[36] Hill, J. G.; Platts, J. A. *Physical Chemistry Chemical Physics*, **2008**, *10*, 2785.
[37] Antony, J.; Grimme, S. *J. Phys. Chem. A*, **2007**, *111*, 4862.
[38] Distasio, R. A.; Head-Gordon, M. *Mol. Phus*, **2007**, *105*, 1073.
[39] Werner, H.-J.; Marchetti, O. *Phys. Chem. Chem. Phys.*, **2008**, *10*, 3400.
[40] Aldler, T. B.; Knizia, G.; Werner, H.-J. *J. Chem. Phys.*, **2007**, *127*, 221106.
[41] Knizia, G.; Aldler, T. B.; Werner, H.-J. *J. Chem. Phys.*, **2009**, *130*, 054104.
[42] McNamara, J. P.; Hillier, I. H. *Physical Chemistry Chemical Physics* **2007**, *9*, 2362-2370
[43] Sharma, R.; McNamara, J. P.; Raju, R. K.; Vincent, M. A.; Hillier, I. H.; Morgado, C. A. *Physical Chemistry Chemical Physics* **2008**, *10*, 2767.
[44] Hohenberg, P.; Kohn, W. *Phys. Rev.*, **1964**, *B136*, 864.
[45] Kristyan, S.; Pulay, P. *Chem. Phys. Lett*, **1994**, *229*, 175.
[46] Perez-Jorda, J. M.; Becka, A. D. *Chem. Phys. Lett*, **1995**, *233*, 134.
[47] Janowski, T.; Pulay, P. *Chem. Phys. Lett.*, **2007**, *447*, 27.
[48] Vydrov, O. A.; Wu, Q.; Voorhis, T. V. *J. Chem. Phys.*, **2008**, *129*, 014106.
[49] Wu, X.; Vargas, M. C.; Nayak, S.; Lotrich, V.; Scoles, G. *Jounral of Chemical physics*, **2001**, *115*, 8748-8757.
[50] Wu, Q.; Yang, W. *Jounral of Chemical physics*, **2002**, *116*, 515-524.
[51] Zimmerli, U.; Parrinello, M.; Koumotsakos, P. *J. Chem. Phys*, **2004**, *120*, 2693.
[52] Grimme, S. *Journal of computaional chemistry*, **2006**, *27*, 1787-1799.
[53] Hepburn, J.; Scoles, G.; Penco, R. *Chem. Phys. Lett.*, **1975**, *36*, 451.
[54] Ahlrichs, R.; Penco, R.; Scoles, G. *Chem. Phys.* , **1977**, *19*, 119.
[55] Hobza, P.; Sandorfy, C. *J. Am. Chem. Soc.*, **1987**, *109*, 1302.
[56] Hobza, P.; Mulder, F.; Sandorfy, C. *J. Am. Chem. Soc.*, **1981**, *103*, 1360.
[57] Meijer, E. J.; Sprik, M. *J. Chem. Phys.*, **1996**, *105*, 8684.
[58] Mooij, W. T. M.; van Duijneveldt, F. B.; van Duijneveldt-van de Rijdt, J. G. C.; van Eijck, B. P. *J. Phys. Chem. A*, **1999**, *103*, 9872.
[59] Mooij, W. T. M.; van Eijck, B. P.; Kroon, J. *J. Phys. Chem. A*, **1999**, *103*, 9883.
[60] Morgado, C. A.; McNamara, J. P.; Hillier, I. H.; Burton, N. A.; Vincent, M. A. *Journal of Chemical Theory and Computation* **2007**, *3*, 1656-1664
[61] Hobza, P.; Cerny, J.; Jurecka, P.; Salahub, D. *J. Comput. Chem.*, **2007**, *28*, 555.
[62] Zhao, Y.; Schultz, N. E.; Truhlar, D. G. *Journal of chemical theory and computation*, **2006**, *2*, 364-382.
[63] Zhao, Y.; Truhlar, D. G. *Journal of chemical theory and computation*, **2008**, *4*, 1849-1868.
[64] Riley, K. E.; Pitonak, M.; Cerny, J.; Hopza, P. *Journal of chemical theory and computation*, **2010**, *6*, 66-80.
[65] Becke, A., D. *Journal of Chemical Physics*, **1993**, *98*, 1372-1377.
[66] Robertazzi, A.; Platts, J. A. *Chemistry-a European Journal* **2006**, *12*, 5747-5756
[67] Kurita, N.; Inoue, H.; Sekino, H. *Chem. Phys. Lett.*, **2003**, *370*, 161.
[68] Lundqvist, B. I.; Hult, E.; Rydberg, H.; Bogicevic, A.; Stromquist, J.; Langreth, D. C. *Progress in Surface Science*, **1998**, *59*, 149.
[69] Lundqvist, B. I.; Dion, M.; Rydberg, H.; Schroder, E.; Langreth, D. C. *Phys. Rev. Lett.*, **2004**, *92*, 246401.
[70] Gulans, A.; Puska, M. J.; Nieminen, R. M. *Phys. Rev. B*, **2009**, *79*, 201105.
[71] Sato, T.; Nakai, H. *J. Chem. Phys.*, **2009**, *131*, 224104.
[72] von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U. *Phys. Rev. Lett*, **2004**, *93*, 153004.
[73] Skylaris, C.-K.; Haynes, P. D.; Mostofi, A. A.; Payne, M. C. *J. Chem. Phys.*, **2005**, *122*, 084119.
[74] Haynes, P. D.; Skylaris, C.-K.; Mostofi, A. A.; Payne, M. C. *Psi-k Newsletter*, **2005**, *72*, 78.
[75] Thiel, W. *Adv. Chem. Phys.*, **1996**, *93*, 703.

[76] Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.*, **1977**, *99*, 4907.
[77] Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.*, **1977**, *99*, 4899.
[78] Stewart, J. J. P. *Journal os Molecular modeling* **2007**, *13*, 1173-1213.
[79] Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.*, **1985**, *107*, 3902.
[80] Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. *Jounral of Comutational Chemistry*, **2006**, *27*, 1101.
[81] Stewart, J. J. P. *J. Comput. Chem.*, **1989**, *10*, 221.
[82] Stewart, J. J. P. *J. Comput. Chem.*, **1989**, *10*, 209.
[83] Feng, F.; Wang, H.; Fang, W.-H.; Yu, J.-g. *Journal of Theortical and Computational Chemistry*, **2009**, *8*, 691.
[84] Tuttle, T.; Thiel, W. *Physical Chemistry Chemical Physics*, **2008**, *10*, 2159-2166.
[85] Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B*, **1998**, *58*, 7260.
[86] Elstner, M.; Frauenheim, T.; Suhai, S. n. *Journal of Molecular Structure (Theochem)*, **2003**, *632*, 29.
[87] Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaziras, E. *J. Chem. Phys.*, **2001**, *114*, 5149.
[88] Zhechkov, L.; Heine, T.; Patchkowskii, S.; DSeifert, G.; Duarte, H. A. *J. Chem. Theory Comput*, **2005**, *1*, 841.
[89] Hobza, P.; Sponer, J.; Otyepka, M.; Zgarbova, M. *Phys. Chem. Chem. Phys.*, **2010**, *12*, 10476.
[90] Puzyn, T.; Suzuki, N.; Heranczyk, M.; Rak, J. *Journal of chemical information and modeling*, **2008**, *48*, 1174-1180.
[91] Aldulaijan, S.; Platts, J. A. *Journal of Molecular Graphics and Modelling*, **2010**, *29*, 240.
[92] Rezac, J.; Fanfrlik, J.; Salahub, D.; Hobza, P. *Jounral of Chemical theory and computation*, **2009**, *5*, 1749-1760.
[93] Korth, M. *J. Chem. Theor. Comput.*, **2010**, *6*, 3808.
[94] Korth, M.; Pitonak, M.; Rezac, J.; Hobza, P. *Journal of chemical theory and computation*, **2010**, *6*, 344.
[95] Stewart, J. J. P. *J. Mol. Model.*, **2009**, *15*, 765.
[96] Boys, S. F.; Bernardi, F. *Mol. Phys*, **1970**, *19*, 553.
[97] Jeziorski, B.; Cwiok, T.; Kolos, W.; Moszynski, R. *J. Mol. Struct.*, **1994**, *307*, 135.
[98] Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.*, **2002**, *362*, 319-325.
[99] Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.*, **2002**, *362*, 319.
[100] Hesselmann, A.; Jansen, G. *Chem. Phys. Lett.*, **2003**, *367*, 778.
[101] Hesselmann, A.; Jansen, G.; Schutz, M. *J. Am. Chem. Soc.*, **2006**, *128*, 11730.
[102] Misquitta, A. J.; Szalewicz, K. *Chem. Phys. Lett.*, **2002**, *357*, 301.
[103] Misquitta, A. J.; Jeziorski, B.; Szalewicz, K. *Phys. Rev. Lett.*, **2003**, *91*, 033201.
[104] Misquitta, A. J.; Podeszwa, R.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.*, **2005**, *123*, 214103.
[105] Williams, H. L.; Chabalowski, C. F. *J. Phys. Chem. A*, **2001**, *105*, 646.
[106] Podeszwa, R.; Bukowski, R.; Szalewicz, K. *J. chem. theory Comput.*, **2006**, *2*, 400.
[107] Hesselmann, A.; Jansen, G.; Schutz, M. *J. Chem. Phys.*, **2005**, *122*, 014103.
[108] Hobza, P.; Kabelac, M.; Sponer, J.; Mejzlik, P.; Vondrasek, J. *Journal of Computational Chemistry* **1997**, *18*, 1136-1150
[109] Paton, R. S.; Goodman, J. M. *Journal of Chemical Information and Modeling* **2009**, *49*, 944-955.
[110] Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *Journal of the American Chemical Society*, **1990**, *112*, 6127-6129.
[111] Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *Journal of physical chemistry A*, **1997**, *101*, 3005-3014.

[112] Labute, P. *Journal of computational chemistry*, **2008**, *29*, 1693-1698.

[113] Anisimov, V. M.; Cavasotto, C. N. *Journal of Computational Chemistry*, **2011**, *32*, 2254-2263.

[114] Zoete, V.; Michielin, O. *Proteins*, **2007**, *67*, 1026-1047.

[115] Zoete, V.; Irving, M.; Michielin, O. *Journal of Molecular Recognition.*, **2010**, *23*, 142-152.

[116] MOE. In <*http://www.chemcomp.com/*>*)* Chemical computing group, **2012**.

[117] Wang, W.; Kollman, P. A. *Journal of Molecular Biology*, **2000**, *303*, 567-582.

[118] Gohlke, H.; Kiel, C.; Case, D. A. *Journal of Molecular Biology*, **2003**, *330*, 891-913.

[119] Davis, M. E.; McCammon, J. A. *Chem . Rev*., **1990**, *90*, 509–521.

[120] Sharp, K. A.; Honig, B. *Annu. Rev. Biophys. Biophys. Chem*, **1990**, *19*, 301-332.

[121] Warshel, A.; Papazyan, A. *Curr. Opin. Struct. Biol.*, **1998**, *8*, 211-217.

[122] Tajkhorshid, E. J., K. J; Suhai, S. *J. Phys. Chem. B.*, **1998**, *102*, 5899-5913.

[123] Baker, N. In *Continuum models for biomolecular solvation: SciTopics*, **2008**.

[124] Dong, F.; Wagoner, J. A.; Baker, N. *Phys. Chem. Chem. Phys.*, **2008**, *10*, 4889-4902.

[125] Klamt, A. *Journal of Physical chemistry*, **1994**, *99*, 2224-2235.

[126] Klamt, A.; Schuurmann, G. *Journal of Chemical Society Perkin Transactions*, **1993**, *2*, 799-805.

[127] Jalkanen, K. J.; Elstner, M.; Suhai, S. *J. Mol. Struc. (Theochem)*, **2004**, *675*, 61-77.

[128] Han, W. J., K.J; Elstner, M; Suhai, S. *J. Phys. Chem. B.*, **1998**, *102*, 2587-2602.

[129] Chalmet, S.; Rinaldi, D.; Ruiz-Lopez. M, F. *Int. J. Quantum Chem*, **2001**, *84*, 559-564.

[130] Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models, 2nd Ed*, Wiley & Sons: Chichester **2004**.

Chapter 2:

Theory:

In this chapter, a brief theoretical background of the methods that were used in this work is presented. Several textbooks [1-10] were used as main references, unless otherwise cited.

2.1 Schrödinger equation:

The time-independent Schrödinger equation is the foundation of Quantum Mechanics theory:

$$\mathrm{H}\Psi = \mathrm{E}\Psi \qquad (2.1)$$

It is a second order partial differential equation. H is the Hamiltonian operator, Ψ is a wavefunction, and E is the energy of the system. This eigenvalue equation can be solved by knowing the function Ψ and the eigenvalue E.

The Hamilton operator includes five contributions to the total energy of a system: the kinetic energies of the electrons and nuclei, the attraction of the electrons to the nuclei, and the interelectronic and internuclear repulsions.

$$H = -\sum_i \frac{\hbar^2}{2m_e}\nabla_i^2 - \sum_k \frac{\hbar^2}{2m_k}\nabla_k^2 - \sum_i \sum_k \frac{e^2 Z_k}{r_{ik}} + \sum_{i<j} \frac{e^2}{r_{ij}} + \sum_{k<l} \frac{e^2 Z_k Z_l}{r_{kl}} \qquad (2.2)$$

i and *j* are electrons, *k* and *l* are nuclei, *ħ* is Planck's constant divided by 2π, m_e is the mass of the electron, m_k is the mass of nucleus *k*, *e* is the charge on the electron, *Z* is an atomic number, r_{kl} is the distance between *k* and *l* nuclei, and ∇^2 is the Laplacian operator:

$$\nabla_i^2 = \frac{\partial^2}{\partial \chi_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2} \qquad (2.3)$$

Where *x, y,* and *z* are Cartesian coordinates.

The Hamilton operator consists of the kinetic energy and potential energy parts. The first two components of (2.2) are the kinetic energy part where the last three components of (2.2) are the potential energy part.

It is impossible to solve the Schrödinger equation for three or more particles because of the correlated motions of particles, so approximations are needed. The Born-Oppenheimer approximation is used to solve Schrödinger equation for many body systems.

2.2 Born-Oppenheimer Approximation:

The idea of this approximation is that the nuclei of molecular systems are moving very slowly compared to electrons: therefore; the nuclei can be considered to be fixed with respect to electrons motion and the electrons depend on any changes on the positions of the nuclei. This difference on the motions is due to the great difference in masses between electrons and nuclei. On this approach, the nuclear kinetic energy is neglected and the nuclear-nuclear repulsion is considered a constant, and the Schrödinger equation is solved for the electrons alone in the electrostatic field of the nuclei. The electronic Hamiltonian includes the first, third and fourth contributions of eq (2.2):

$$H_{electrons} = -\sum_i \frac{\hbar^2}{2m_e} \nabla_i^2 - \sum_i \sum_k \frac{e^2 Z_k}{r_{ik}} + \sum_{i<j} \frac{e^2}{r_{ij}} \qquad (2.4)$$

The total wavefunction for the molecular system $\Psi_{total}(r,R)$ can be divided into an electronic wavefunction $\Psi_{electrons}(r,R)$ and a nuclear wavefunction $\Psi_{nuclei}(r,R)$:

$$\Psi_{total}(r,R) = \Psi_{electrons}(r,R)\Psi_{nuclei}(r,R) \qquad (2.5)$$

Where r is electronic coordinates and R is nuclear coordinates.

The electronic Schrödinger equation is:

$$\mathrm{H}_{elec}\Psi_{elec}(r,R) = \mathrm{E}_{elec}\Psi_{elec}(r,R) \qquad (2.6)$$

Where the electronic coordinates r are independent variables and the nuclear coordinates R are parameters.

2.3 Hartree - Fock Approximation:

Hartree - Fock method is an approximation to solve Schrödinger equation for many-body system. It depends on the one-electron Fock operator $f(i)$ in which the electron-electron repulsion is presented in an average potential $U^{HF}(i)$, in_which the single electron (*i*) has been affected by all the rest of the electrons.

$$f(i) = -\frac{1}{2}\nabla_i^2 - \sum_k \frac{Z_k}{r_{ik}} + U^{HF}(i) \qquad (2.7)$$

Because the solution of one electron affects the other electrons by the average potential term, the Hartree – Fock equation (2.8) need to be solved by Self-consistent Field method (SCF).

$$f(i)\chi(x_i) = \varepsilon\chi(x_i) \quad (2.8)$$

The main disadvantage of this method is that it ignores electron correlation due to the electron-electron repulsion, which is treated as an average, so it does not provide the exact energy E_{exact}:

$$E_{cor} = E_{exact} - E_{HF} \quad (2.9)$$

E_{cor} is the correlation energy, which is the difference between the exact energy E_{exact} and HF energy E_{HF}. It is an important term on calculating chemical properties.

Based on Hartree - Fock method many methods had been developed such as Coupled Cluster (CC), Møller–Plesset Perturbation Theory (MPPT), and Density Functional Theory (DFT).

2.4 Coupled Cluster (CC):

One of the best methods to calculate the electron correlation energy is coupled cluster method (CC). It is a size-consistent method, *i.e.* that the energy of two atoms or molecules at infinite separation equals the sum of the energy of each atom or molecule. On this method the full-correlation interaction wavefunction is included, and writing as:

$$\Psi_{CC} = e^{C}\Psi_{HF} \quad (2.10)$$

The cluster operator C is defined as:

$$C = C_1 + C_2 + C_3 + ...+ C_n = \sum C_i \qquad (2.11)$$

Where n is the total number of electrons, and i is the excitation level.

The coupled cluster is a computationally expensive method; therefore other more economic schemes are used instead such as coupled cluster with double-excitation operator CCD where C is approximated as ($C = C_2$), coupled cluster with single and double-excitation operator CCSD where C is approximated as ($C = C_1 + C_2$). Coupled cluster with single, double and triple-excitation operator CCSDT where C is approximated as ($C = C_1 + C_2 + C_3$), which is highly accurate but very computationally expensive and so only applicable to small model systems.

The CCSD(T) approach, in which triple excitations are included perturbatively, provides almost identical accuracy to CCSDT method with much reduced computational cost. CCSD(T) method is now widely known as the "gold standard" of single-reference calculations [11].

2.5 Møller–Plesset Perturbation Theory (MPPT):

Møller–Plesset presented an alternative theory to solve the problem of electron correlation. MPPT method is a size-consistent method and it is based on the perturbation theory. The Hamiltonian H of a system includes the sum of a zeroth-order Hamiltonian H_0 and a perturbation V :

$$H = H_0 + \lambda V \qquad (2.12)$$

Where λ is a parameter that has a value between 0 and 1. When λ is zero then H is equal to H_0, but if λ is 1 then the H equals to its true value.

The eigenfunctions Ψ_i and the eigenvalues E_i of the Hamiltonian H are expressed in power of λ:

$$\Psi_i = \Psi_i^{(0)} + \lambda\Psi_i^{(1)} + \lambda^2\Psi_i^{(2)} + ... = \sum_{n=0} \lambda^n \Psi_i^{(n)}$$

$$E_i = E_i^{(0)} + \lambda E_i^{(1)} + \lambda^2 E_i^{(2)} + ... = \sum_{n=0} \lambda^n E_i^{(n)}$$

$E_i^{(1)}$ is the first-order correction to the energy, $E_i^{(2)}$ is the second-order correction and so on. From the eigenfunction, the energies can be calculated as:

$$E_i^{(0)} = \int \Psi_i^{(0)} H_0 \Psi_i^{(0)} d\tau \qquad (2.13)$$

$$E_i^{(1)} = \int \Psi_i^{(0)} V \Psi_i^{(0)} d\tau \qquad (2.14)$$

$$E_i^{(2)} = \int \Psi_i^{(0)} V \Psi_i^{(1)} d\tau \qquad (2.15)$$

$$E_i^{(3)} = \int \Psi_i^{(0)} V \Psi_i^{(2)} d\tau \qquad (2.16)$$

The above equations show that in order to obtain an improvement on the Hartree-Fock energy, second-order Møller–Plesset perturbation (MP2) is required. Third-order and fourth-order Møller–Plesset perturbation (MP3) and (MP4) are also available.

MPPT is an *ab initio* method which provides low computational cost, but it is limited to small systems. In addition, while it provides accurate binding energies for hydrogen bonds it is known to overestimate the interaction energy in stacked systems. In order to improve the performance of MP2 and reduce the cost, many methods have been developed such as density-fitted local MP2 (DF-LMP2) [12], which makes use of the local nature of electron correlation. Spin-component scaled MP2 (SCS-MP2) [13], which depends on separation of the correlation energy resulting from anti-parallel and parallel spin pairs of electrons, and assigns two new scaling factors to these contributions. Spin-component scaled for nucleobases (SCSN) was developed by optimising the scaling factors against binding energy of stacked nucleobases, and totally neglects the anti-parallel-spin electron pairs contribution and gives the parallel spin contribution a scaling of 1.76 [14]. SCS-MP2 for molecular interactions (SCS(MI)-MP2) is an improvement of SCS-MP2 approach [15, 16]. In this method the parallel-spin contribution and anti-parallel spin contribution scaled 1.29 and 0.40, respectively [16].

2.6 Density Functional Theory (DFT):

All *ab initio* methods discussed above are computationally expensive and can be used for small molecules or clusters. The main difference between these methods and density functional theory is that DFT calculates the electronic density distribution instead of wavefunction. Density functional theory (DFT) is a very popular method for many reasons such as; it takes in account the electron correlation, it is less expensive than the *ab initio* methods; therefore it can be used to calculate molecules with 100 atoms or above, and in general it provides very accurate results.

DFT approach calculates the energy of a system (E) as a functional of the density. The first model was developed by Thomas-Fermi which contains some basic elements. However the main theorems that underpin modern DFT were set out by

Hohenberg and Kohn. The first theorem is that the ground state energy of an electronic system is written as a functional of the electron density. In other words, in order to calculate the ground state energy and other property of a system, we only need to know the electron density in three-dimensional space and not the full wavefunction. That means that the energy of a system (E) is a function of the density $p(r)$:

$$E[p(r)] = \int V_{ext}(r) p(r) dr + F[p(r)] \quad (2.17)$$

The first term presents the interaction of the electrons with the external potential$V_{ext}(r)$. $F[p(r)]$ includes the kinetic energy of the electrons and the interelectronic interactions.

The second theorem gives a variation principle for the density functionals:

$$\varepsilon_{el}[p(r)] \geq \varepsilon_{el}[p_0(r)] \quad (2.18)$$

p_0 is the true density for the system and p any other density obeying

$$\int p(r)dt = \int p_0(r)dt = N \quad (2.19)$$

The drawback of these theorems is that $F[p(r)]$ is not known. Therefore E is depending on $p(r)$ which is also not known.

The above problem was solved by Kohn and Sham approach. On this approach, $F[p(r)]$ is approximated as the sum of three terms:

$$F[p(r)] = E_{KE}[p(r)] + E_H[p(r)] + E_{XC}[p(r)] \quad (2.20)$$

Where $E_{KE}[p(r)]$ is the kinetic energy, $E_H[p(r)]$ is the electron-electron Coulombic energy, and $E_{XC}[p(r)]$ is the exchange and correlation. $E_{KE}[p(r)]$ is the kinetic energy of a system *non-interacting* electrons with the same density as the real system. The full expression of the Kohn-Sham is:

$$E[p(r)] = \sum_{i=1}^{N} \int \varphi_i(r)dr + \frac{1}{2}\int\int \frac{p(r_1)p(r_2)}{|r_1 - r_2|}dr_1 dr_2 + E_{XC}[p(r)] - \sum_{A=1}^{M} \int \frac{Z_A}{|r - R_A|}p(r)dr \quad (2.21)$$

Kohn and Sham presented the density $p(r)$ of the system as “the sum of the square moduli of a set of one-electron orthonormal orbitals” [8]:

$$p(r) = \sum_{i=1}^{N} |\varphi_i(r)|^2 \quad (2.22)$$

This leads to the one-electron Kohn-Sham equation:

$$\left\{ -\frac{\nabla_1^2}{2} - \left(\sum_{A=1}^{M} \frac{Z_A}{r_{1A}} \right) + \int \frac{p(r_2)}{r_{12}} dr_2 + V_{XC}[r_1] \right\} \varphi_i(r_1) = \varepsilon_i \varphi_i(r_1) \quad (2.23)$$

$V_{XC}[r_1]$ is the exchange-correlation functional and ε_i are orbital energies.

2.7 Semi-empirical methods (SE):

Due to the extremely expensive cost of the *ab initio* methods and the limitation of these methods, semi-empirical methods have been developed. Semi-empirical methods are approximation methods based on the Hartree-Fock theory. In these methods only the valence electrons of a system are considered explicitly, and the Coulomb and exchange integrals that form the most expensive part of *ab initio* methods are replaced by one or more parameters. The first semi-empirical method was CNDO (Complete Neglect of Differential Overlap) and the parameters were developed from *ab initio* calculations. Many approaches have been developed after CNDO approach, such as INDO (Intermediate Neglect of Differential Overlap), NDDO (Neglect of Diatomic Differential Overlap), and MNDO (Modified Neglect of Diatomic Overlap).

2.7.1 Austin model 1 (AM1):

Since MNDO method preformed very poorly in the prediction of hydrogen bond energies and geometries, Dewar and co-workers presented the Austin Model 1 (AM1) semi-empirical method. The main modification was done on the nuclear repulsion term by "adding a stabilization Gaussian function to the MNDO core-core interaction"[17, 18]. On AM1 the nuclear repulsion energy between nuclei A and B is written as:

$$V_N(A,B) = V_{AB}^{MNDO} + \frac{Z_A Z_B}{r_{AB}} \sum_{i=1}^{4} \left\lfloor a_{A,i} e^{-b_{A,i}(r_{AB}-c_{A,i})^2} + a_{B,i} e^{-b_{B,i}(r_{AB}-c_{B,i})^2} \right\rfloor \quad (2.24)$$

Up to 4 parameters for each atom a, b, and c presented Gaussian functions. In the beginning this method described just four elements C, H, O and N and then parameterization for B, F, Mg, Al, Si, P, S, Cl, Zn, Ge, Br, Sn, I, and Hg have been presented.

2.7.2 Parameterized model 3 (PM3):

Parameterized model 3 (PM3), which is also based on MNDO approach, was developed by Stewart by improving the techniques of the parameterization. It has parameters for H, C, N, O, F, Al, Si, P, S, Cl, Br, and I in the beginning but later parameters for Li, Be, Na, Mg, Ca, Zn, Ge, As, Se, Cd, In, Sn, Sb, Te, Hg, Tl, Pb, and Bi have been added. PM3 has two Gaussian functions for each atom instead of four in AM1.

2.7.3 Recife Model 1 (RM1):

Recife Model 1 (RM1) is a new parameterization method of AM1 [17], and can be easily used in any software that contains AM1. The parameterization set of this method contains 1736 important molecules in biochemistry containing ten atoms (C, H, N, O, P, S, F, Cl, Br, and I) using only the s-p basis set with all parameters re-optimized [17, 19].

2.7.4 Parameterized model 6 (PM6):

Parameterized model 6 (PM6) was developed by Stewart [17]. PM6 has parameters for 80 atoms, covering most of the periodic table. PM6 is the most accurate method among the semi-empirical method up to date.

2.7.5 MOZYME:

MOZYME is MOPAC keyword which uses localized molecular orbital (LMO) instead of the standard SCF procedure.

2.8 Force Field (FF) methods: Molecular Mechanics (MM):

The energy of a molecule in the force field (FF) methods is calculated as a function of the nuclear positions only. On these methods, the electronic motion is completely neglected.

The FF energy E_{FF} includes bonded E_{bond} and non-bonded $E_{non-bond}$ energies. The bonded energy contains the bond stretching energy, the bond angle energy and the bond rotation (torsion) energy. The non-bonded energy contains the van der Waals and electrostatic energies.

$$E_{FF} = E_{bond} + E_{non-bond} \qquad (2.25)$$

$$E_{FF} = \sum_{bonds}\frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{angles}\frac{k_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{torsions}\frac{V_n}{2}(1+\cos(nw-\gamma)) + \sum_{i=1}^{N}\sum_{j=i+1}^{N}(4\varepsilon_{ij}[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}})$$
(2.26)

Where k_i and V_n are the force constants. l_i and θ_i are the bond length and the valence angle deviate from the reference values $l_{i,0}$ and $\theta_{i,0}$ respectively. γ is the phase angle. r_{ij} is the distance between atom i and j. ε_{ij} and σ_{ij} are Lennard-Jones parameters. q_i and q_j are the atomic charges on atom i and j.

2.8.1 AMBER:

Assisted Model Building with Energy Refinement (AMBER) is a force field method developed by Kollman for the simulation of peptides and nucleic acids. In this method hydrogen bonding was described explicitly with a 12-10 potential:

$$U_H = \sum_{H-bonds}(\frac{C_{12}}{R^{12}} - \frac{C_{10}}{R^{10}}) \qquad (2.27)$$

AMBER-94 is very popular method and it is all-atom force field.

2.8.2 OPLS:

Optimised Potentials for Liquid Simulations is a force field method developed to model proteins in solution. Many OPLS parameters have been presented to many atoms. OPLS-AA method includes all atoms explicitly, which was developed by William L. Jorgensen [20]. The parameters of the bending and bond stretching are based on AMBER force field [21] except alkane parameters which have been developed by CHARMM. Where most torsional parameters are based on HF/6-31G* calculations [22, 23].

2.9 Continuum Solvation Free Energy:

Solvation free energy (ΔG_{Sol}) is "the free energy change to transfer a molecule from vacuum to solvent." [8]

$$\Delta G_{Sol} = \Delta G_{elec} + \Delta G_{vdw} + \Delta G_{cav} \qquad (2.28)$$

The solvation free energy is divided to three components. First, the electrostatic components (ΔG_{elec}) which is important especially for charged and polar solutes because of the polarisation of the solvent. Second the van der Waals interaction between the solvent and solute (ΔG_{vdw}), which is divided to the dispersion term (ΔG_{disp}) and the repulsive term (ΔG_{rep}). The third component of the solvation free energy is the cavity free energy (ΔG_{cav}), which "is the energy required to form the solute cavity within the solvent."[8]

2.9.1 The electrostatic contribution to the free energy solvation: The Born and Onsager models:

Born and then Onsager contributed to the study of the solvation. Born obtained the electrostatic term of the solvation free energy by placing a charge within a spherical cavity, and this was extended by Onsager to a dipole in a spherical cavity. The Born model can be defined as the work needed to transfer an ion from vacuum to medium, and can be written as:

$$\Delta G_{elec} = -\frac{q^2}{2a}(1-\frac{1}{\varepsilon}) \qquad (2.29)$$

Where q is the charge on the ion, a is the radius of the cavity, and ε is the dielectric constant.

The Conductor-like Screening Model (COSMO) is a continuum approach which depends on generation of a conducting surface at vdW distance in order to calculate the dielectric screening charges and energies by providing a ε-dependent correction factor:

$$f(\varepsilon) = \frac{(\varepsilon - 1)}{(\varepsilon + \frac{1}{2})} \qquad (2.30)$$

Where ε is the dielectric constant of the solvation.

Generalized Born model (GB) is widely used to calculate the electrostatic contribution to the solvation free energy in force field methods. The model includes a system of particles with radii a_i and charges q_i. The total electrostatic free energy for a system can be calculated by taking the sum of the Coulomb energy and the Born free energy of solvation in a medium of relative permittivity ε:

$$G_{elec} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{q_i q_j}{\varepsilon r_{ij}} - \frac{1}{2}(1-\frac{1}{\varepsilon})\sum_{i=1}^{N} \frac{q_i^2}{a_i} \qquad (2.31)$$

Where the first term in (2.31) can be written as:

$$\sum_{i=1}^{N}\sum_{j=i+1}^{N}\frac{q_i q_j}{\varepsilon r_{ij}}=\sum_{i=1}^{N}\sum_{j=i+1}^{N}\frac{q_i q_j}{r_{ij}}-(1-\frac{1}{\varepsilon})\sum_{i=1}^{N}\sum_{j=i+1}^{N}\frac{q_i q_j}{r_{ij}} \quad (2.32)$$

Which is the sum of the Coulomb interaction in *vacuo* and in $(1-\frac{1}{\varepsilon})$.

The total electrostatic energy in the GB model includes three terms and the first term is the Coulomb interaction between charges in *vacuo:*

$$G_{elec}=\sum_{i=1}^{N}\sum_{j=i+1}^{N}\frac{q_i q_j}{r_{ij}}-(1-\frac{1}{\varepsilon})\sum_{i=1}^{N}\sum_{j=i+1}^{N}\frac{q_i q_j}{r_{ij}}-\frac{1}{2}(1-\frac{1}{\varepsilon})\sum_{i=1}^{N}\frac{q_i^2}{a_i} \quad (2.33)$$

The GB equation ΔG_{elec} is the difference between G_{elec} and the Coulomb energy in *vacuo:*

$$\Delta G_{elec}=-(1-\frac{1}{\varepsilon})\sum_{i=1}^{N}\sum_{j=i+1}^{N}\frac{q_i q_j}{r_{ij}}-\frac{1}{2}(1-\frac{1}{\varepsilon})\sum_{i=1}^{N}\frac{q_i^2}{a_i} \quad (2.34)$$

GB is very important method on continuum solvent calculation due to its accuracy (it has good agreement with experiments), simplicity and has atomic forces required for molecular dynamic simulations. This model is widely used on molecular mechanics and semi-empirical quantum mechanics calculations.

2.9.2 The non-electrostatic contribution to the solvation free energy:

The electrostatic component of the solvation free energy has been discussed above. The other two components: the van der Waals and the cavity free energy are usually combined to:

$$\Delta G_{cav}+\Delta G_{vdw}=\gamma A+b \quad (2.35)$$

Where A is the total solvent accessible area and γ and b are constants. The constant b is commonly set to 0, making the cavity and van der Waals terms depended on the solvent accessible area.

2.9.2.1 Generalized Born Surface Area (GB-SA):

Generalized Born model Surface Area (GB-SA) is widely used in calculation the solvation free energy of a solute E_{sol} as a sum of polarization and cavitation terms:

$$\Delta G_{sol} \approx E_{sol} = E_{pol} + E_{cav} \quad (2.36)$$

E_{pol} is a classical electrostatic term which presents the induced charge interactions between the solute and solvent, and E_{cav} presents the non-classical effects, solute-solvent van der Waals interactions and the energy necessary for creating the solute cavity in the solvent. The E_{pol} can be written as:

$$E_{pol} = -\frac{1}{2}\frac{\varepsilon_1^{-1} - \varepsilon^{-1}}{4\pi\varepsilon_0}\sum_{ij} q_i q_j F_{GB}(r_{ij}; R_i^{eff}, R_j^{eff}) \quad (2.37)$$

r_{ij} is the distance between atoms i and j, q_i is the partial charge of atom i, ε_1 is the dielectric constant of solute interior, ε is the dielectric constant of the solvent, ε_0 is the vacuum permittivity, R_i^{eff} is a position only dependent effective Born radius of atom i which presents the degree of which the atom is buried, and F_{GB} is a pairwise interaction function presented by Still.

$$F_{GB}(r_{ij}; G_i Gj) = \frac{1}{r_{ij}} S_{GB}(r_{ij}^2 G_i G_j) \quad (2.38)$$

Where $G_i = \frac{-1}{(2R_i^{eff})}$ and $S_{GB}(t) = (1 + t^{-1}e^{-1}/4)^{\frac{-1}{2}}$. By using these values the (2.37) equation is written as:

$$E_{pol} = \frac{\varepsilon_1^{-1} - \varepsilon^{-1}}{4\pi\varepsilon_0}\{\sum_{ij} q_i^2 G_i - \sum_{i<j} \frac{q_i q_j}{r_{ij}} S_{GB}(r_{ij}^2 G_i G_j)\} \quad (2.39)$$

$\{\}\{G_i\}$ are the Born self energy factors of the individual atoms placed in the solute cavity.

In GB-SA model, the cavity term E_{cav} is a linear combination of exposed atomic surface areas:

$$E_{cav} = \sum_i \sigma_i A_i \quad (2.40)$$

Where A_i is presents the exposed surface area of atom i. And $\{\sigma_i\}$ are a coefficients dependent on the chemical type of each atom.

2.9.2.2 Generalized Born Volume Integral (GB-VI):

In this method [24] the cavitation term E_{cav} is calculated by solvent volume integral $(2R_i G_i)^3$ instead of solvent exposed surface area as on GB-SA:

$$E_{GB-VI} = \tau\{\sum_{i=1}^{n} \frac{q_i^2 G_i}{4\pi\varepsilon_0} - \sum_{i<j}^{n} \frac{q_i q_j S_{GB}(r_{ij}^2 G_i G_j}{4\pi\varepsilon_0 r_{ij}} + \sum_{i=1}^{n} \gamma_i (2R_i G_i)^3\} \quad (2.41)$$

Where $\tau = (\varepsilon_1^{-1} - \varepsilon^{-1})$, $S_{GB}(t) = (1 + t^{-1}e^{-t}/4)^{-1/2}$ and

$$(2R_i G_i)^3 = -\sqrt{R_i^6}\,\frac{3}{4\pi}\int\int\int r^{-6}\delta(x \notin solute)d^3x \quad (2.42)$$

The partial charge q_i must be specified, $\{R_i\}$ and $\{\gamma_i\}$ are parameters that must be estimated from experiment.

2.10 Molecular Dynamic (MD):

Molecular dynamic simulation solves the numerical integration of *Newton's* equations (F= *ma*) of motion for a system for a period of time. The velocities and trajectories for the system on this period of time are saved. Many properties can be evaluated using these trajectories. MD starts by giving each atom on the system some kinetic energy. The force $F_i(t)$ eq (2.43) on atom i can be calculated by the potential energy function based on its position $r_i(t)$:

$$F_i = -\frac{\partial}{\partial r_i} v(....,r_i,....) \quad (2.43)$$

The acceleration $a_i(t)$ of this atom is:

$$a_i(t) = \frac{F_i(t)}{m_i} \quad (2.44)$$

The atomic position can be obtained by:

$$\frac{d^2 r_i}{dr^2}(t) = \frac{F_i(t)}{m_i} \quad (2.45)$$

For each atom there are three components:

$$m_i a_i(t) = \sum_j f_{ij}(t) \quad (2.46)$$

Where $\sum_j f_{ij}(t)$ is the sum over all atoms *j* presenting forces on atom *i* which need to be calculated. In order to do so, it is replaced with difference equations for a small time step which can be solved successfully by Taylor series approximations:

$$x_i(t+\Delta t) = x_i(t) + v_i(t)\Delta t + a_i(t)\frac{\Delta r^2}{2} \quad (2.47)$$

$$x_i(t+\Delta t) = x_i(t) - v_i(t)\Delta t + a_i(t)\frac{\Delta r^2}{2} \quad (2.48)$$

Where x is the coordinate of atom i. By adding the two equations (2.47) and (2.48):

$$x_i(t+\Delta t) = 2x_i(t) - x_i(t-\Delta t) + a_i(t)\Delta t^2 \quad (2.49)$$

And by substitution from (2.45):

$$x_i(t+\Delta t) = 2x_i(t) - x_i(t-\Delta t) + \sum_j \frac{f_{ij}(t)}{m_i} \quad (2.50)$$

This equation above gives the prediction of the position of atom i at time $t+\Delta t$ if we know the position at time t and $t-\Delta t$. This equation is known as the Verlet algorithm and it is used to find the position of each atom at time steps.

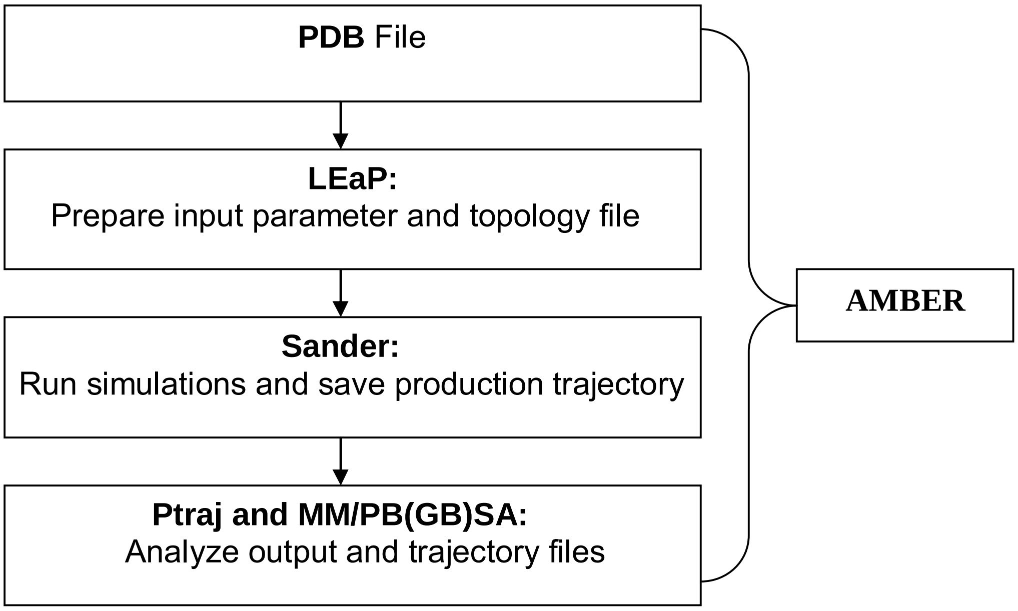
By subtracting equations (2.47) and (2.48):

$$v_i(t) = \frac{x_i(t+\Delta t) - x_i(t-\Delta t)}{2\Delta t} \quad (2.51)$$

Where $v_i(t)$ is the velocity at time t.

2.10.1 Assisted Model Building with Energy Refinement (AMBER):

Assisted Model Building with Energy Refinement (AMBER) is a collection of programs is widely used to study molecular dynamic simulations and to analyze the results for proteins, nucleic acids and carbohydrates. The calculations are divided in to three stages; preparation, simulation and analyzing the output files. Each stage has been done by one or more of AMBER programs. The preparation was done by *LEaP* program, the simulations by *Sander*, and the analyzing by *Ptraj* and *MM/(PB)GBSA* methods.


PDB File
LEaP:
Prepare input parameter and topology file
Sander:
Run simulations and save production trajectory
Ptraj and MM/PB(GB)SA:
Analyze output and trajectory files
AMBER

2.10.2 MM/PB(GB)SA:

Molecular Mechanics Poisson-Boltzmann Surface Area (MM/PBSA) and Molecular Mechanics-Generalized Born Surface Area (MM/GBSA) methods are widely used to calculate protein-ligand binding free energy. In these methods the binding free energy $\Delta G_{\text{Bind, Solv}}$ calculated by the sum of gas-phase (vacuum) contribution $\Delta G_{\text{Bind, Vacuum}}$, desolvation free energy of the system ΔG_{desolv} and entropic contribution -TΔS (see equation 2.52 and Figure 2.1).

$$\Delta G_{bind} = \Delta G_{\text{Bind, Vacuum}} + \Delta G_{desolv} - T\Delta S \qquad (2.52)$$

The vacuum term $\Delta G_{\text{Bind, Vacuum}}$ includes the van der Waals interaction energy ΔG_{vdW} , the electrostatic interaction energy ΔG_{elec} , and the internal energy variation between the two molecules on the complex ΔG_{intra} (bond, angle, and torsional angle energies) between the complex and isolated molecules:

$$\Delta G_{\text{Bind, Vacuum}} = \Delta G_{vdW} + \Delta G_{elec} + \Delta G_{intra} \qquad (2.53)$$

ΔG_{desolv} is the difference between the solvation free energy $\Delta G_{\text{solv, Complex}}$ of the complex and of the isolated part ($\Delta G_{\text{solv, Ligand}}$ and $\Delta G_{\text{solv, Receptor}}$):

$$\Delta G_{desolv} = \Delta G_{\text{solv, Complex}} - (\Delta G_{\text{solv, Ligand}} + \Delta G_{\text{solv, Receptor}}) \qquad (2.54)$$

The solvation free energy includes the electrostatic $\Delta G_{elec.solv}$ and the nonpolar contributions $\Delta G_{np.solv.}$

$$\Delta G_{solv} = \Delta G_{elec.solv} + \Delta G_{np.solv} \qquad (2.55)$$

The vacuum component $\Delta G_{\text{Bind, Vacuum}}$ is calculated by single point energy calculations, while the solvation components $\Delta G_{\text{solv, Complex}}$,$\Delta G_{\text{solv, Ligand}}$ and $\Delta G_{\text{solv, Receptor}}$) are calculated by Poisson-Boltzmann (PB) equation on MM/PBSA, or by Generalized Born (GB) equation on MM/GBSA method which has been found much faster than MM/PBSA .

The entropy component divided to translational S_{trans}, rotational S_{rot} and vibrational S_{vib}. Because of the expensive cost of calculating of the entropy term and the chance of errors, MM-PB (GB)SA method may neglect the entropy contribution especially when “a comparison of states of similar entropy is desired such as two ligands binding to the same protein” [25]

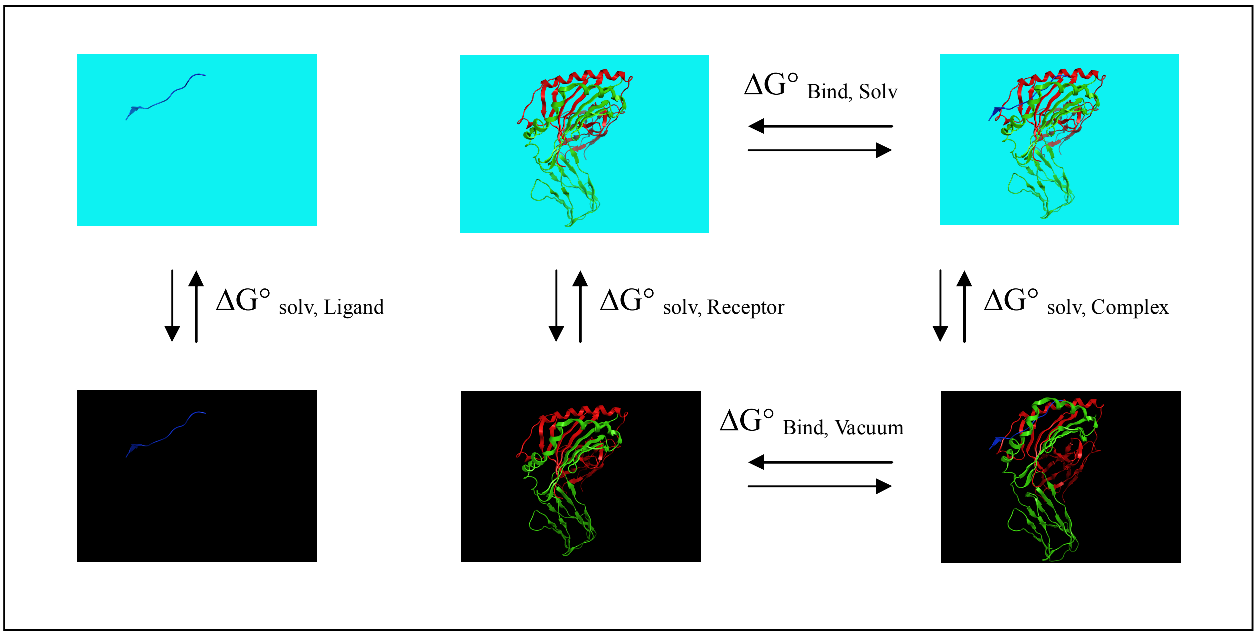


Figure 2.1: Thermodynamic cycle used on MM/PB (GB)SA calculations, black surfaces indicate vacuum and the blue surfaces indicate solvent.

2.11 Statistical methods:

Spearman's rank correlation (r_s) is a statistical method which evaluates the strength link between two sets (X and Y):

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)} \qquad (2.56)$$

It starts by ranking these sets starting from 1 for the smallest value for each set ($rank_x$) and ($rank_y$). d^2 is the difference between the two ranks and n is the number of data pairs. If the value of r_s is 1 or -1, it means that there is a perfect correlation between these two sets. If r_s is equal to 0, there is no correlation.

The Pearson correlation coefficient (r) is similar to Spearman's rank correlation [26]. It evaluates the linear relation between two x and y:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}} \qquad (2.57)$$

n is the number of data pairs. r =1 or -1 means perfect correlation (positive and negative correlation, respectively. r =0 means there is no correlation between these two sets.

Coefficient of Determination (R^2 or r^2) estimates the strength of the linear correlation between x and y. It takes a value between 0 and 1. Another definition of the Coefficient of Determination is that it gives the percent of the variance in data that is explained by the best fitted line. For example, R^2 =0.85 means that 85% of the data are accounted for by the model.

Receiver operating characteristic (ROC) curve is a technique which evaluates a method that predicts positive and negative results [27, 28] for example by using a cutoff. It is important to have information about when the results can be considered positive or negative, for example a cutoff has been used in this work. In order to get ROC curve, the specificity (Sp) and sensitivity (Se) of the data set must be calculated:

Specificity (Sp) = Number of true positives / Number of peptides in positive test set

and

Sensitivity (Se) = Number of true negatives / Number of peptides in negative test set.

A ROC curve can be achieved by plotting specificity against sensitivity for a range of values, calculated by for example the cutoff. The accuracy of the method is calculated by measuring the area under the ROC curve (*a*). *a* may take a value between 1 and 0.5. *a*=1 means a perfect prediction, where a=0.5 means a random prediction.

2.13 References:

[1] Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*, Oxford Science Publications **1987**.

[2] Atkins, P.; Friedman, R. *Molecular Quantum Mechanics*, Oxford University Press **2005**.

[3] Cramer, C. J. *Essentials Computational Chemistry Theories and Models*, Wiley **2004**.

[4] Frenkel, D.; Smit, B. *Understanding Molecular Simulation from Algorithms to Applications*, Academic Press **2002**.

[5] Goodfellow, J. M. *Molecular Dynamics Applications in Molecular Biology*, Macmillan Press Scientific&Medical **1991**.

[6] Goodman, J. M. *Chemical Applications of Molecular Modelling*, Royal Society of Chemistry **1998**.

[7] Hinchliffe, A. *Molecular Modellingfor Beginners*, Wiley **2003**.

[8] Leach, A. R. *Molecular Modelling Principles and Applications*, Prentice Hall **2001**.

[9] Maritz, J. S. *Distribution-Free Statistical Methods*, Chapman & Hall **1995**.

[10] Myers, J. L.; Well, A. *Research Design and Statistical Analysis*, Psychology Press **2003**.

[11] Riley, K. E.; Pitonak, M.; Jurecka, P.; Hobza, P. *Chem. Rev*, **2010**, *110*, 5023-5063.

[12] Werner, H.-J.; Knowles, P. J.; Manby, F. R. *Jounral of Chemical physics*, **2003**, *118*, 8149-8160.

[13] Grimme, S. *Journal of Chemical Physics*, **2003**, *118*, 9095-9102.

[14] Hill, J. G.; Platts, J. A. *Journal of chemical theory computation*, **2007**, *3*, 80-85

[15] Berka, K.; Laskowski, R.; Riley, K. E.; Hobza, P.; Vondrasek, J. *J. Chem. Theor. Comput.*, **2009**, *5*, 982-992.

[16] Distasio, R. A.; Head-Gordon, M. *Mol. Phys*, **2007**, *105*, 1073-1083.

[17] Stewart, J. J. P. *Journal os Molecular modeling* **2007**, *13*, 1173-1213.

[18] Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.*, **1985**, *107*, 3902-3909.

[19] Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. *Jounral of Comutational Chemistry*, **2006**, *27*, 1101-1111.

[20] Jorgensen., W. L.; Maxwell., D. S.; Tirado-Rives., J. *J. Am. Chem. Soc.*, **1996**, *118*, 11225-11236.

[21] Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S. J.; Weiner, P. J. *J. Am. Chem. Soc.*, **1984**, *106*, 765-784.

[22] Jorgensen, W. L.; McDonald, N. A. *Journal of Molecular Structure:THEOCHEM*, **1998**, *424*, 145-155.

[23] Price, M. L. P.; Ostrovsky, D.; Jorgensen, W. L. *Journal of computational chemistry*, **2001**, *22*, 1340-1352.

[24] Labute, P. *Journal of computational chemistry*, **2008**, *29*, 1693-1698.

[25] Case, D. A.; Darden, T. A.; Cheatham, I., T.E. ; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B. P.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Liu, 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, V.; Babin, C.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. In http://ambermd.org/, **2010**.

[26] Roberts, F.; Roberts, D. In http://mathbits.com/,**2012**

[27] Davies, M. N.; Sansom, C. E.; Beazley, C.; Moss, D. S. *Molecular Medicine*, **2003**, *9*, 220-225.

[28] Swets, J. *Science*, **1988**, *3*, 1285-1293.

Chapter 3:

Ab initio, DFT and semi-empirical calculations of Myelin Basic Protein epitope (MBP) to MHC class II:

3.1 Introduction:

In order to create and develop new drugs, we have to understand the way that a drug interacts with its receptor in order to affect the biological system in the body's cells. One important concept is understanding the chemical interactions between the drug and the receptor.[1-3] In most cases, the most significant interactions between drugs and their biological receptors are non-covalent interactions.[4] Although, non-covalent interactions are typically weaker than covalent interactions, collectively they exert an important influence in many properties of biomacromolecules, for example they are well known to affect the structure of proteins, DNA and RNA.[5-8]

Biomacromolecules are very large systems. Because of their size [9], and their flexibility [4, 10], there are many difficulties to use some of the computational methods to study the non-covalent interactions. Empirical potentials, Semi-empirical quantum chemical methods, and *ab initio* HF methods have been used to investigate the non-covalent interactions in biomacromolecules [11]. We have employed many of the methods discussed above to examine in detail the interaction of the immuno-dominant epitope of myelin basic protein (MBP) with Major Histocompatibility Complex (MHC) class II receptor.

3.1.1 Myelin Basic Protein (MBP):

Myelin Basic Protein (MBP) contributes in the myelination (i.e. formation of an insulating sheath) of nerves in the central nervous system (CNS) [12, 13]. It has been found that MBP is one of the proteins which is attacked by the immune system [12, 13], so this protein plays an important role in demyelinating diseases such as Multiple sclerosis (MS) [14]. Immunological studies suggest that a key sequence recognised by the immune system (or epitope) is that found in positions 83 to 99 with primary structure Glu-Asn-Pro-Val-Val-His-

Phe-Phe-Lys-Asn-Ile-Val-Thr-Pro, which is therefore widely used in studies of multiple sclerosis (MS) disease [15].

3.2 Methods:

The X-ray crystallographic coordinates contained in PDB entry 1YMM, corresponding to the study of Hahn et al [15], were obtained from the Protein Data Bank [16]. These were loaded into the Molecular Operating Environment (MOE) software, and protonated according to typical protonation states. All hydrogen positions were optimised using the AMBER94 forcefield, with heavy atoms fixed at their X-ray positions. Particular attention was paid to histidine residues, for which all possible protonation states were checked with both OPLSAA and AMBER94. This analysis revealed a clear preference for neutral histidines in all cases, in agreement with Wucherpfenning who stated that LYS93 is the only charged residue on this peptide [17]. Truncation of the PDB coordinates to individual amino-acids resulted in neutral species, to avoid charge-charge terms dominating binding energies.

Ab initio calculations were performed using the MOLPRO package of programs [18]. DF-LMP2 calculations and SCSN scaling employed the aug-cc-pVTZ orbital and fitting basis sets [19-21]. The Gaussian03 suite of programmes was used to calculate the interaction energies for AM1, PM3, BHandH and MP2/6-31G(0.25d) [22]. The MOPAC programme was used to carry out PM6, RM1 and RM1BH calculations. For larger systems with the RM1 method, we used the MOZYME keyword to accelerate the calculations [23]. MOE was used for OPLS-AA and AMBER94 calculations. For PM3-D and AM1-D methods, we used optimised parameters for H, C, N, and O reported by McNamara et al [9]. Dispersion corrections were calculated following the procedure set out by Grimme [24].

Before we used these modified methods, we tested them by calculating the interaction energies for S22 complexes. The results were exactly the same as the results on the McNamara study. After this, we calculate the dispersion term to S22 complexes and add it to the AM1 and PM3 results. The results were the same as the results in the study Table (3.1). After, we were sure about our scripts, we started to use these new methods to calculate the interaction energies for our peptides.

Table 3.1: PM3-D and AM1-D interaction energies for S22 complexes.

	PM3-D		AM1-D	
complex name	This work	McNamara results	This work	McNamara results
water_dimer	-5.14	-5.14	-7.29	-7.29
phenol_dimer	-7.52	-7.52	-9.76	-9.76
methane_dimer	-1.24	-1.24	-0.94	-0.94
ammonia_dimer	-1.77	-1.77	-3.43	-3.43
benzene_water cs	-3.65	-3.65	-3.43	-3.43
pyrazine_dimer	-4.20	-4.2	-4.57	-4.57
benzene_hcn_cs	-4.43	-4.43	-4.44	-4.44
uracil_dimer_stack	-6.78	-6.78	-10.56	-10.56
uracil_dimer_hb	-20.30	-20.3	-20.15	-20.15
adenine_thymine	-17.33	-17.33	-16.58	-16.58
adenine_thymine_stack	-10.63	-10.63	-12.20	-12.2
ethene_dimer d2d	-3.60	-3.6	-3.31	-3.31
formic_acid_dimer c2h	-18.57	-18.57	-15.45	-15.45
formamide_dimer	-15.37	-15.37	-17.16	-17.16
benzene_ammonia	-2.96	-2.96	-3.00	-3
benzene_methane	-2.42	-2.42	-2.12	-2.12
benzene_dimer c2v	-4.15	-4.15	-3.85	-3.85
benzene_dimer c2h	-4.30	-4.3	-2.90	-2.9
indole_benzene_t-shape_c1	-6.65	-6.65	-7.10	-7.1
indole_benzene_stack_c1	-6.09	-6.09	-4.04	-4.04
2-pyridoxine_2-aminopyridine	-17.52	-17.52	-16.50	-16.5
ethene_ethine c2v	-1.85	-1.85	-1.61	-1.61

3.3 Results and Discussion:

3.3.1 Pairwise calculations:

MOE was used to obtain each individual pairwise interaction between amino acids in the complex, based on distance criteria between peptide and any atom of receptor residue, as defined in the “ligand interactions” procedure used in MOE [25]. A total of 49 interactions were identified by these criteria, and are listed in Table (3.2) and Table (3.3).

Table 3.2: Peptide-protein interactions identified by distance criteria.

Peptide residues	α residues	β residues
Glu85	-	-
Asn86	Ser53, Arg50, Phe51	-
Pro87	Ser53, Ala52, Phe51	-
Val88	Ser53	His81
Val89	Ser53, Phe54	Asn82
His90	Phe24	Tyr78, His81, Asn82
Phe91	Gln9, Phe22, Phe54, Gly58, Asn62	Tyr78
Phe92	Gln9, Asn62	Arg13, Phe26, Asp28, Gln70, Ala71, Tyr78
Lys93	Asn62	-
Asn94	Glu11, Asn62, Val65, Asp66	Arg13
Ile95	Val65, Asn69	Trp61, Ile67
Val96	Asn69	Trp61
Thr97	Asn69, Ile72	Asp57, Trp61
Pro98	Arg76	Pro56, Asp57, Tyr60

Table 3.3: 49 pairwise interaction energies (kcal/mole)

Peptide residues	α residues	β residues	OPLSAA	MP2	PM3	AM1	PM3-D	AM1-D	RM1BH	PM6	RM1	RM1-D	RM1-D(0.5)	BhandH	PM6-D
ASN86	SER53		7.99	13.32	17.81	22.32	15.43	19.95	8.28	11.16	10.35	7.59	9.24	5.24	8.79
ASN86	ARG50		1.65	6.426	5.62	8.626	3.73	6.73	2.64	4.12	6.30	3.42	5.15	1.98	2.23
ASN86	PHE51		-1.41	-1.314	0.61	-0.17	-2.93	-3.71	1.43	0.35	1.44	-2.50	-0.13	-0.77	-3.19
PRO87	SER53		-2.33	-0.635	0.97	-0.05	-1.41	-2.43	-2.01	-4.05	0.22	-2.52	-0.87	-8.35	-6.43
PRO87	PHE51		-1.18	-0.661	0.60	0.007	-0.96	-1.55	0.17	-0.56	0.17	-1.63	-0.54	-1.79	-2.12
PRO87	Ala 52		-2.9	-2.372	-0.52	-1.42	-2.80	-3.70	-0.53	-3.02	-0.5	-3.14	-1.55	-4.21	-5.30
VAL88		HIS81	-3.06	-4.993	-3.08	-0.11	-4.77	-1.79	-5.01	-4.07	-1.16	-3.22	-1.98	-8.06	-5.75
VAL88	Ser53		-1.69	0.221	2.13	-0.47	-3.46	-6.05	6.84	-3.29	2.71	-4.22	-0.06	-5.34	-8.88
VAL89	Ser53		-2.99	-1.456	0.43	0.45	-1.20	-1.18	0.49	0.24	0.49	-1.50	-0.30	-0.15	-1.39
VAL89	Phe54		-1.67	-1.199	-0.30	-0.6	-3.72	-4.02	0.60	-1.53	0.78	-3.02	-0.73	-3.33	-4.95
VAL89		Asn82	-3.87	-1.901	-0.17	-1	-3.44	-4.27	-0.63	-2.77	-0.62	-4.47	-2.16	-4.44	-6.03
His90	phe24		-1.37	-1.019	-0.22	-0.47	-1.07	-1.33	-0.34	-0.77	-0.33	-1.29	-0.71	-1.78	-1.63
His90		tyr78	-2.12	-1.217	0.29	0.63	-2.43	-2.09	0.67	0.28	0.68	-2.25	-0.49	-1.29	-2.43
His90		HIS81	-2.6	-4.063	-1.39	-0.92	-6.18	-5.70	1.21	0.00	1.21	-4.04	-0.88	-1.64	-4.78
His90		Asn82	-3.59	-1.288	1.83	-1.81	-0.74	-4.39	0.13	-1.49	0.15	-3.19	-1.18	-3.49	-4.07
Phe91	Gln9		-4.79	-3.333	-1.04	-1.55	-4.42	-4.92	-1.32	-3.31	-1.32	-5.10	-2.83	-5.55	-6.68
Phe91	Phe22		-0.8	-0.274	-0.57	0.165	-2.96	-2.22	0.49	0.19	0.49	-2.27	-0.61	-0.42	-2.20
Phe91	Phe54		3.44	0.281	-2.10	0.926	-4.93	-1.90	-0.08	-0.59	-0.08	-3.43	-1.42	-0.34	-3.42
Phe91	Gly58		-0.13	0.252	1.02	1.72	-2.89	-2.19	1.65	0.20	1.66	-3.02	-0.21	-0.72	-3.70
Phe91	Asn62		-2.17	-2.194	-0.13	0.092	-2.48	-2.26	-0.58	-1.13	-0.58	-3.07	-1.57	-2.32	-3.49
Phe91		tyr78	-1.15	-1.256	-1.03	-0.22	-3.62	-2.81	0.20	-0.74	0.20	-2.89	-1.03	-1.42	-3.33
PHE92	GLN9		-7.23	-5.017	-1.49	0.138	-5.02	-3.39	-6.63	-7.32	-0.88	-5.21	-2.60	-12.32	-10.85
PHE92		Arg13	-4.15	-5.935	-3.71	-2.65	-10.8	-9.79	-2.77	-5.28	-2.78	-11.20	-6.15	-8.36	-12.42
PHE92		Phe26	-3.13	-2.685	0.14	1.054	-4.30	-3.39	0.29	-0.60	0.29	-4.64	-1.68	-2.64	-5.05
PHE92		Asp28	-0.084	-3.863	-1.72	-2.8	-3.31	-4.39	-2.67	-4.47	-1.93	-3.95	-2.73	-7.52	-6.05
PHE92	Asn62		-2.01	-1.627	-0.36	-0.81	-1.35	-1.81	-0.58	-1.13	-0.58	-1.64	-1.00	-1.54	-2.13
PHE92		Gln70	-1.83	-0.949	0.09	0.478	-1.72	-1.33	0.31	0.24	0.31	-1.55	-0.43	-0.04	-1.57
PHE92		Ala71	-1.07	-0.573	-0.78	-0.01	-2.80	-2.02	0.25	-0.07	0.25	-2.00	-0.65	-0.86	-2.08
PHE92		tyr78	-3.33	-1.278	0.40	0.694	-5.01	-4.72	0.94	-0.54	0.94	-5.40	-1.59	-1.97	-5.95
Lys93	Asn62		-4.79	-6.123	-4.95	-5.09	-6.48	-6.62	-5.34	-6.88	-5.34	-6.95	-5.98	-8.12	-8.40
ASN94	GLU11		-9.33	-14.94	-10.61	-10.9	-12.28	-12.61	-9.70	-12.9	-8.86	-10.50	-9.52	-17.82	-14.58
ASN94	ASN62		22.09	19.76	20.56	26.92	13.03	19.38	4.79	3.80	12.47	11.42	12.04	4.59	-3.729
ASN94		ARG13	-11.86	-11.54	-2.99	1.423	-6.76	-2.35	-10.21	-10.4	-6.83	-11.90	-8.87	-15.34	-14.18
ASN94	Val65		-1.96	-1.12	-0.72	0.15	-4.27	-3.38	0.52	-0.11	0.53	-3.25	-0.98	-1.64	-3.64

ASN94	Asp66		-4.55	-7.86	-6.38	-5.58	-7.93	-7.12	-5.06	-6.94	-5.06	-6.62	-5.68	-8.65	-8.48
Ile95		trp61	-2.84	-2.17	-0.26	0.18	-3.57	-3.13	0.20	-0.41	0.20	-3.33	-1.20	-1.68	-3.71
Ile95	Val65		-1.07	-0.27	0.28	0.52	-1.05	-0.81	0.30	0.26	0.31	-1.11	-0.25	-0.09	-1.07
Ile95		ile67	5.8	3.72	0.51	2.48	-5.78	-3.81	4.86	2.45	4.86	-3.52	1.51	1.49	-3.84
Ile95	asn69		-2.45	-2.99	-0.98	-1.94	-1.99	-2.94	-0.99	-1.95	-1.00	-2.24	-1.49	-2.68	-2.95
VAL96		TRP61	-3.87	-5.25	-1.58	-3.15	-3.19	-4.76	-2.38	-3.91	-0.81	-2.73	-1.57	-7.17	-5.51
VAL96	asn69		-3.53	-1.44	-0.92	-1.33	-2.88	-3.29	-1.24	-2.95	-1.24	-3.49	-2.14	-3.92	-4.91
THR97	ASN69		-4.78	-3.17	0.13	-1.32	-3.71	-5.16	-0.02	-2.52	0.04	-4.31	-1.69	-4.57	-6.36
thr97		asp57	18.58	5.59	3.07	0.81	0.83	-1.41	-2.24	-3.49	0.12	-2.91	-1.08	-2.19	-5.72
thr97		TRP61	-1.46	-0.64	-0.01	0.74	-3.02	-2.27	0.37	0.06	0.37	-2.78	-0.88	-1.18	-2.95
thr97	ile72		-0.72	0.43	-0.96	0.35	-5.20	-3.88	0.76	-0.19	0.76	-4.26	-1.24	-0.81	-4.42
Pro98		Pro56	5.44	2.07	1.58	0.87	-1.58	-2.29	2.61	0.44	2.99	-0.79	1.47	-0.09	-2.73
Pro98		asp57	11.25	-2.91	-3.43	-2.60	-8.75	-7.91	-3.59	-5.88	-0.27	-7.09	-2.99	-9.57	-11.19
Pro98		Tyr60	4.16	2.40	0.78	2.33	-4.50	-2.95	4.31	2.18	4.30	-2.47	1.59	1.56	-3.10
Pro98	Arg76		1.45	0.40	1.12	1.51	-1.19	-0.80	1.73	2.38	1.72	-0.69	0.76	0.71	0.06
MUE [a]			2.08		1.97	2.15	2.33	2.33	2.27	1.68	2.17	2.13	1.36	2.39	3.48
MSE [a]			-0.46		-1.31	-1.74	1.74	1.31	-0.81	0.41	-1.51	1.92	-0.13	2.12	3.46
MAX [a]			5.3		3.21	4.77	9.51	7.54	14.96	15.95	7.29	8.49	7.71	15.17	23.49
MIN [a]			-14.16		-8.54	-13	-4.77	-9.18	-6.62	-4.07	-6.08	-4.4	-5.41	-2.42	-0.36

a. relative to MP2 method.

Based on MP2 results, it has been found that the first residue of our peptide (Glu85) is not interacting with any receptor residues. It interacts with Asn86 residues in the peptide chain Figure (3.1).

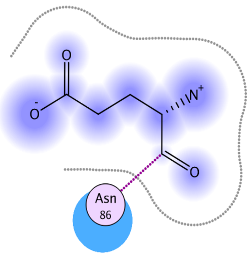


Figure 3.1: Glu85 structure on the peptide chain

The second residue (Asn86) has three destabilization interactions. So, we can see that Asn86 does not interact with the receptor. Pro87 is a stable residue compared with Glu85 and Asn86. Val88 has good interactions with the receptor. Val 89, which is hydrophobic residue [17], has three interactions. The most stable interaction according to our calculations is Val89- Asn82. His90 has four interactions. The more interesting interaction is with His81,which is believed to be a stacking interaction between the two aromatic groups Figure 3.2. His90 is one of the most significant residues for our peptide- receptor interactions [15, 17, 26].

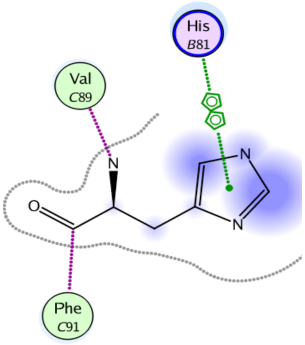


Figure 3.2: His90- His81 stacking interaction.

Phe91 has six interactions. Most of these interactions are stabilizing interactions. The most stable interaction for Phe91 is with Gln9 from the α chain. Phe92, which is hydrophobic residue [17], is the most stabilization residue in the peptide bound to the MHC class II receptor according to MP2 interaction energies, and one of the most significant residues for the peptide- receptor interactions Figure (3.3) [26]. It has eight stabilizing interactions with the receptor residues. According to MOE programme Phe92 has hydrogen bond interaction with Gln9. Phe92 is placed between two aromatic molecules Phe26 and Tyr78. These two residues with Gln70, Ala71, Asp28 and Arg13 create a hydrophobic pocket [17]. Ala71 makes a space which is occupied with Phe92 aromatic group which stabilises the interaction of this residue.

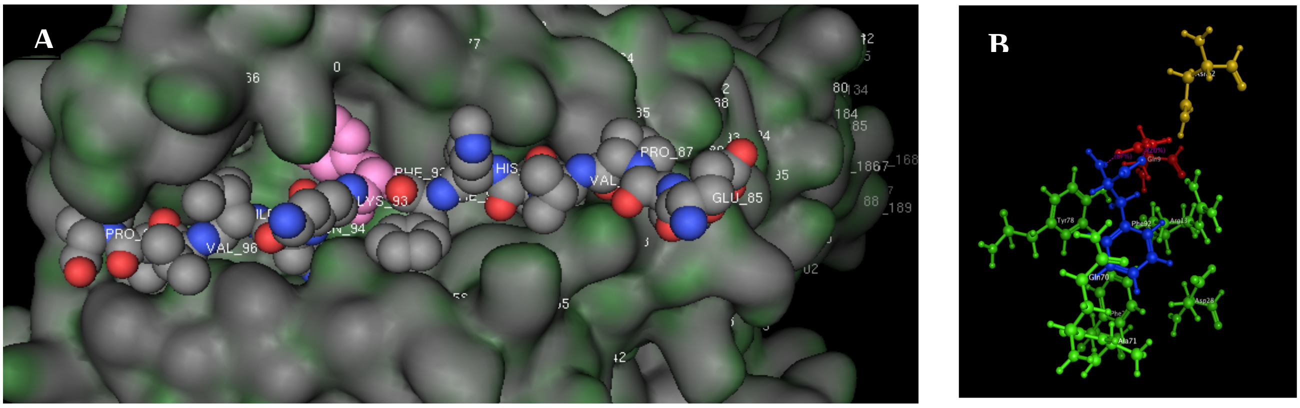


Figure 3.3: A) Phe92 (pink colour) inside the pocket of the MHC residues.
B) Phe92 (blue colour), Gln9 (red colour) and Gln70, Ala71, Asp28, Phe26, Tyr78 and Arg13 (green colour) create a hydrophobic pocket.

Lys93 is the only charged amino acid on this peptide [17]. Lys93 has just one interaction. This interaction is considered one of the significant interactions with the receptor [17, 26]. The position of Lys93 on the receptor prefer positive charge residue [27]. Asn94, which is a polar residue, interacts with five residues. Three of these interactions are Hydrogen-bond interactions. Asn94 is located inside pocket which created by five receptor residues Figure (3.4). From the calculations, Asn94 is the second most stabilizing residue, after Phe92, which bound to the receptor according to MP2 interaction energies. Ile95 has four interactions. Three of these interactions are stabilizing interactions. Val96 has two stabilizing interactions with the receptor. One of these interactions is a hydrogen-bond interaction according to MOE. Thr97 has four interactions. One of them is hydrogen-bond interaction. Pro98, which is the last residue of the peptide, it has three interactions.

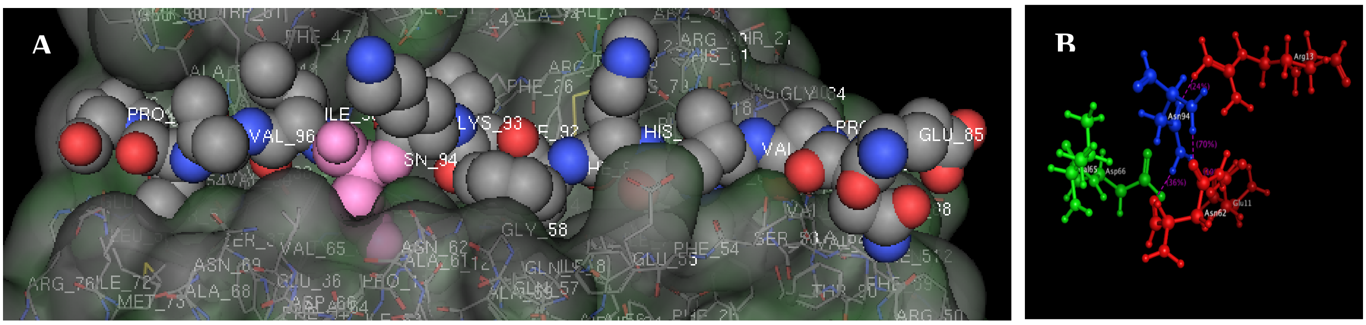


Figure3.4: A)Asn94 (pink colour) in side the pocket. B) Asn94 (blue colour), Asn62, Glu11, and Arg13 (red colour) H-bond interactions, and Val65 and Asp66 (green colour) dispersion interactions.

A subset of nine interactions, Val88-Ser53α, Phe91-Phe54α, Ile95-Ile67β, Thr97-Asp57β, Pro98-Pro56 β, Pro98-Asp57β, Pro98-Tyr60 β, Asn86-Asn62α and Val88-His81β, were selected for study using DF-LMP2 and SCSN methods. Selections were made to cover a range of interaction energies and types. These data were used as a benchmark to test the performance of faster, more approximate methods. As shown in Table (3.4), MP2/6-31G(0.25d) gives reliable results, with MUE of 0.92 kcal/mol when compared with SCSN. Similar performance (MUE = 0.84 kcal/mol) was found when using SCS(MI) as the benchmark method Table (3.5).

Table 3.4: the MUE (mean unsigned error), MSE (mean signed error), MAX and MIN error for several methods compared with SCSN aug-cc-pVTZ for nine pairwise amino acid interactions (kcal/mol).

	MUE	MSE	MAX a	MIN a
PM6-D	7.02	-7.02	-3.22	-13.09
OPLS-AA	6	-4.73	5.32	-17.07
PM3-D	4.22	-3.49	2.13	-8.82
AM1-D	4.17	-1.78	6.64	-6.44
AM1	4.07	1.47	15.96	-4.87
RM1-BH	3.32	-0.12	8.51	-8.1
BHandH	3.19	3.19	8.07	0.41
RM1-D	3.16	-2.48	3.05	-5.59
RM1	2.82	1.38	4.74	-3.36
PM3	2.11	-0.21	2.52	-4.6
PM6	1.81	-1.39	1.83	-6.97
MP2/6-31G(0.25d)	0.92	-0.92	-0.01	-2.11

a) Errors defined as $E_{SCSN} – E_{Method}$, and hence are positive for overbinding relative to SCSN, negative for underbinding.

Table (3.4) also contains data for several more approximate methods. All are considerably worse than MP2/6-31G(0.25d) for these data, although some semi-empirical methods such as PM6, PM3 and RM1 show some promise. Slightly surprisingly, inclusion of dispersion correction (using the default parameters from ref. [23]) actually makes predictions worse in all cases: this aspect will be discussed in more detail below. For these data at least, OLPS-AA does not appear to be a suitable method to predict interaction energy. On the basis of these results, MP2/6-31G(0.25d) was selected as the most appropriate method to use as a benchmark for all pairwise interactions and for larger systems.

Table 3.5: the MUE (mean unsigned error), MSE (mean signed error), MAX and MIN error for several methods compared with SCS(MI) aug-cc-pVTZ for nine pairwise amino acid interactions (kcal/mol).

	MUE	MSE	MAX	MIN
PM6-D	6.66	-6.66	-3.08	-12.62
OPLS-AA	5.99	-4.38	6.52	-16.87
PM3-D	4.19	3.84	8.91	-0.93
AM1-D	4.14	2.13	6.65	-5.44
BHandH	3.55	3.55	9.27	0.51
AM1	3.45	-2.27	4.14	-12.17
RM1-D	3.4	2.84	6.52	-2.53
RM1-BH	3.3	0.23	9.71	-7.77
RM1	2.8	-1.03	4.16	-4.22
PM6	2.03	1.74	7.18	-1.31
PM3	2.01	0.14	8.91	-0.93
MP2/6-31G(0.25d)	0.84	-0.57	1.19	-1.9

Following this test, all 49 pairwise interactions between amino acids were calculated using several methods: results are summarised in Table (3.6), using MP2/6-31G(0.25d) as a benchmark, with full details reported in Table (3.3). This data shows that the OPLS-AA force field method gave good agreement with MP2 for several interactions, but failed for several others such as PRO98-ASP57, for which the interaction energy is -2.91 kcal/mol with MP2/6-31G(0.25d) and +11.25 kcal/mol according to OPLS-AA. The overall MUE error compared with MP2/6-31G(0.25d) is 2.08 kcal/mol, which although rather smaller than that reported in Table (3.4) (for fewer interactions) is rather greater than the 1 kcal/mol generally accepted as “chemical accuracy”.

Table 3.6: MUE, MSE, MAX, and MIN relative to MP2/6-31G(0.25d) for all 49 amino acid interactions identified by MOE (kcal/mol).

	MUE	MSE	MAX	MIN
OPLS-AA [a]	2.08	-0.46	5.3	-14.16
PM3	1.97	-1.31	3.21	-8.54
AM1	2.15	-1.74	4.77	-12.96
PM3-D	2.33	1.74	9.51	-4.77
AM1-D	2.33	1.31	7.54	-9.18
RM1BH	2.27	-0.81	14.96	-6.62
PM6	1.68	0.41	15.95	-4.07
RM1	2.17	-1.51	7.29	-6.08
RM1-D	2.13	1.92	8.49	-4.4
RM1-D(0.7)	1.36	-0.14	7.71	-5.41
BHandH	2.39	2.12	15.17	-2.42
PM6-D	3.48	3.46	23.49	-0.36

a) Cut-off for electrostatic interaction energy of 10 Å employed. Without this cut-off, interaction energies differ by up to 0.17 kcal/mol, MUE identical at the precision shown.

BHandH results in overestimated interaction energies, for instance ∆E for PHE92-GLN9 is -12.3 kcal/mol using BHandH and just -5.01 kcal/mol with MP2/ 6-31G(0.25d). The overall MUE comparing with MP2/6-31G(0.25) calculations is 2.39 kcal/mol, again rather larger than required for our purposes. Semi-empirical methods PM3 and AM1 give similar overall errors, with MUE of 1.97 kcal/mol for PM3 and 2.15 kcal/mol for AM1, and more often than not underestimate interaction energies. Similar performance is found for reparameterisations RM1 and RM1-BH, whereas PM6 shows a slight improvement over other related methods.

Including a dispersion correction, and modifying parameters according to McNamara [9], does not improve performance, with MUE of 2.33 kcal/mol for both AM1-D and PM3-D. To analyse these data in more detail, errors for hydrogen bonded, dispersion bound, and charged interactions were calculated separately, and found to be 1.92, 2.41 and 2.82 kcal/mol for PM3-D, and 2.36, 2.23, and 3.07 kcal/mol for AM1-D, respectively. Notably, Table (3.6) shows that the mean signed error (MSE) changes sign on addition of dispersion, suggesting that the default dispersion correction overcompensates for the shortcomings of the underlying methods in these cases. The form of the dispersion correction given by Grimme [24] is shown in equation (3.1):

$$E_{disp} = -s_6 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{C_6^{ij}}{R_{ij}^6} f_{dmp}(R_{ij}) \qquad (3.1)$$

where s_6 is a global scaling factor that must be optimised for the method to be corrected. Tables (3.4) and (3.6) employed the default scaling factor of 1.4, as in the initial reports of AM1-D and PM3-D method. Varying the s_6 parameter, the optimal combination of method and global scaling was located to be RM1 with s_6 = 0.7, for which an overall MUE of 1.36 kcal/mol was obtained. The response of this overall error to the value of s_6 is shown in Figure (3.5). Repeating this procedure with other semi-empirical methods gave similar curves but slightly higher MUE values, with the exception of PM6 for which the optimal value of s_6 was zero, *i.e.* any attempt to improve on predictions by adding a dispersion term actually gave worse performance.

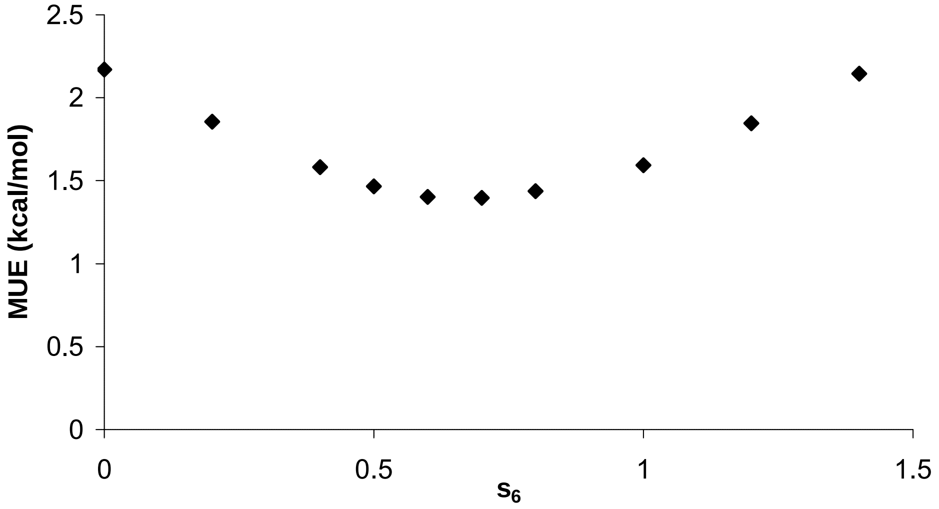


Figure 3.5 Response of RM1-D error as a function of global scaling parameter s_6.

Previous studies [17, 26, 27] indicate that the most important residues for interaction of this MBP epitope with MHC-II are the central residues Val88 to Asn94. In general, our calculations are in agreement with these findings, indicating that most stabilisation of the complex stems from these residues' interactions. Table (3.7) reports the sum of individual pairwise interaction energies from both MP2/6-31G(0.25d) and RM1-D with s_6 = 0.7. Both methods show that Phe92 and Asn94 are particularly strongly stabilising, with significant further contributions from all except the terminal residues. Table (3.7) also reports interaction energies of each peptide residue with *all* receptor residues identified as interacting, calculated with RM1-D. This data shows very similar trends to those from pairwise data, with Phe92 and Asn94 considerably more stabilising than other residues, and terminal Asn86 and Pro98 destabilising the complex. The difference between pairwise and direct interaction energies is the many-body term: no clear trend is apparent in this data, but the size of this term in some cases, *e.g.* more than 4 kcal/mol for Phe92, suggests that conclusions from pairwise calculations should be treated with caution. Summing either pairwise or direct interaction energies gives estimates of the overall stabilisation of the complex of -59.0 and -57.3 kcal/mol, respectively.

Table 3.7: Stabilisation energies due to each residue in peptide (kcal/mol)

	Sum of pairwise		Direct	Many body [a]
	MP2	RM1-D	RM1-D	
Glu85	-	-	-	-
Asn86	+18.44	+11.51	+14.20	+2.69
Pro87	-3.67	-3.30	-3.21	+0.09
Val88	-4.77	-4.05	-2.09	+1.96
Val89	-4.56	-3.41	-3.52	-0.11
His90	-7.59	-4.77	-3.76	+1.01
Phe91	-6.52	-8.05	-8.36	-0.31
Phe92	-21.93	-13.62	-17.85	-4.23
Lys93	-6.12	-5.86	-5.86	0.00
Asn94	-15.71	-21.08	-16.77	+4.33
Ile95	-1.73	-2.86	-1.60	+1.26
Val96	-6.70	-3.93	-3.84	+0.09
Thr97	+2.22	-2.87	-5.36	-2.49
Pro98	+1.97	+3.28	+0.70	-2.58

a. the difference between sum of pairwise and direct results from RM1-D method.

3.3.2 Dipeptides and heptapeptides calculations:

The speed of semi-empirical methods such as RM1 allows us to examine the interaction energies of larger models of the peptide than single amino acids. The peptide was cut into seven dipeptides, as well as two heptapeptides, and the interaction energy with all receptor residues identified in Table (3.3) calculated, as shown in Table (3.8). This data further indicates that stabilisation comes mainly from the central residues, with the dipeptide Lys93-Asn94 particularly strongly bound. Table (3.8) also shows that much greater binding results from the heptapeptide Phe92-Pro98 than from Glu85-Phe91, perhaps unsurprisingly given that this contains both of the most strongly bound single residues.

Table 3.8: Interaction energies of larger models (kcal/mol) by RM1-D

Dipeptides	
Glu85-Asn86	10.62
Pro87-Val88	-6.34
Val89-His90	-7.45
Phe91-Phe92	-17.88
Lys93-Asn94	-58.84
Ile95-Val96	-7.63
Thr97-Pro98	1.72
Sum	-85.79
Heptapeptides	
Glu85-Phe91	-20.16
Phe92-Pro98	-81.83
Sum	-101.99

These data also show that the estimated overall stabilisation of the complex increases as the size of the model peptides increase. Considering only single amino acids, RM1-D estimate is –57.3 kcal/mol, rising to –85.8 kcal/mol from dipeptides and to –102.0 kcal/mol from heptapeptides. Calculating all 14 residues as one molecule with all 29 receptor residues gives an interaction energy of -110.0 kcal/mol, *i.e.* larger still. From these data it seems clear that cutting the peptide into individual amino acids causes significant error in prediction of overall binding energy.

3.3.3 Peptide (MBP)-receptor (MHC II) calculations:

The size of the receptor was then increased by progressively adding more amino acids, using a 4.5Å distance cutoff repeatedly until all receptor residues were included. Table (3.9) shows that there is a large effect of the size of the model on the calculated interaction between peptide and receptor, with the interaction energy of the peptide with the entire receptor calculated at -263.1 kcal/mol. Figure (3.6) shows how interaction energy varies with model size, demonstrating that even with almost 300 receptor residues, or 80% of the entire protein, convergence is not reached. It is notable that the total charge on the peptide-protein complex changes from neutrality to -13 as the size of the receptor model increases, suggesting that long-range electrostatic forces may be playing a significant role here.

Table 3.9: Interaction energy of peptide with progressively larger models of receptor (kcal/mol)

No. amino acids in receptor model	Total charge (e)	ΔE
29	0	-110.02
47	-1	-141.87
135	-6	-211.13
198	-9	-218.63
291	-9	-236.87
325	-12	-256.49
360 (all)	-13	-263.10

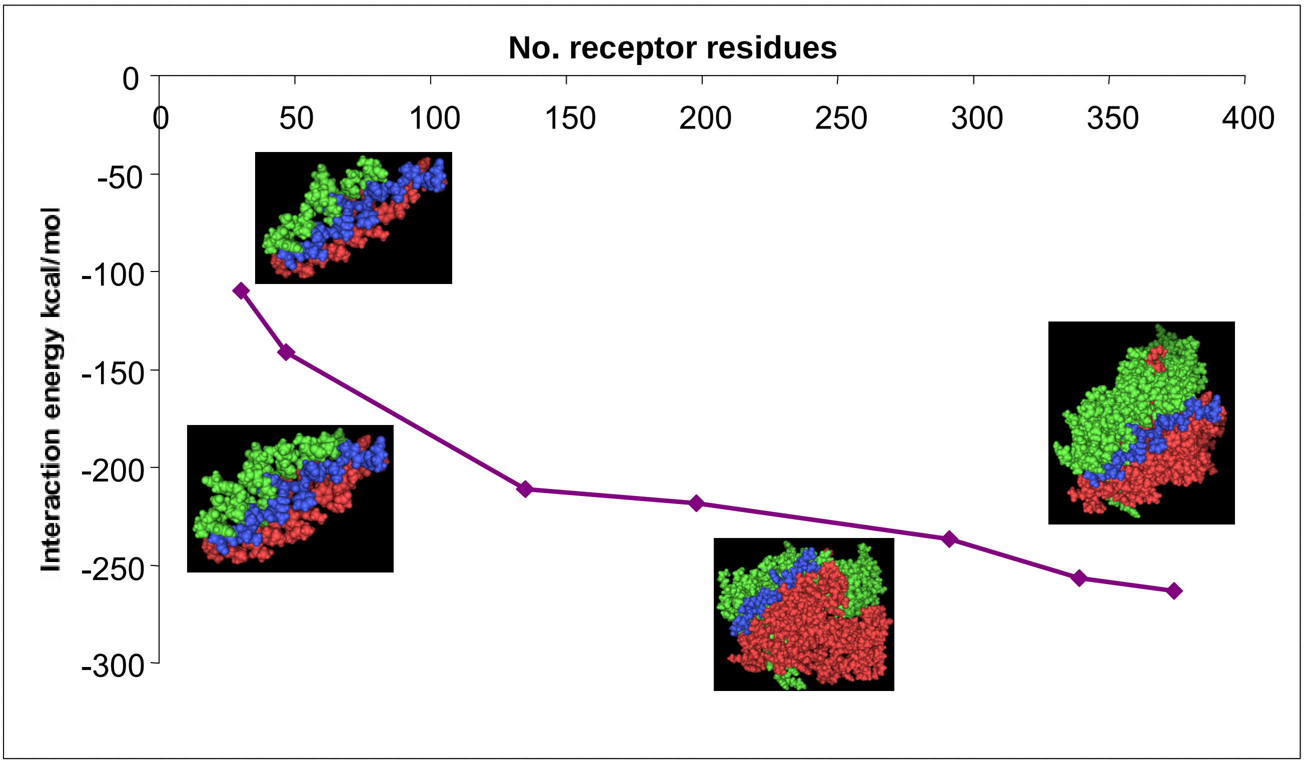


Figure 3.6: The interaction energies for large systems (kcal/mol).

3.3.4 Molecular Dynamics snapshots calculations:

It is well known that both peptide and receptor are flexible systems, such that calculations on the single static X-ray structure may not represent the true nature of this interaction. We have therefore calculated interaction energies for ten low-energy snapshots from molecular dynamics (average RMSD Cα =1.09 Å), taken from reference [28], in order to study the effect of the motion in our peptide-receptor interaction energy (Table (3.10) and Figure (3.7)). Interaction energies were calculated using OPLS-AA and RM1-D using the entire receptor. This data shows that there is a large effect of the motion in our peptide-receptor interaction energy, despite the relatively small RMSD across the snapshots. Detailed analysis of ligand-receptor interactions of each snapshot does not reveal any clear origin for these trends. For instance, snapshot 8 is less strongly bound while snapshot 6 is more strongly bound than average, but snapshot 6 actually has fewer ‘native’ interactions than snapshot 8 Table (3.11). Thus, it seems that the origin of the variation in binding energy is not simple, and cannot be assigned to any single interaction. Slightly surprisingly, in the light of the above results, OPLS-AA shows very similar behaviour to RM1-D results, with a difference in means of 6.9 kcal/mol, and very similar trends in interaction energy across snapshots. More work was done to determine whether longer dynamical simulations are required to properly model peptide-protein binding, and results are reported in chapter 5.

Table 3.10: Interaction energies for ten molecular dynamics snapshots (kcal/mol)

Snapshot	OPLSAA	RM1-D (0.7)
1	-175.19	-159.97
2	-192.53	-193.30
3	-177.48	-167.09
4	-157.17	-195.14
5	-193.89	-208.91
6	-194.36	-211.52
7	-178.60	-174.42
8	-121.08	-137.40
9	-175.37	-186.87
10	-150.88	-150.81
Mean	-171.66	-178.54
Standard deviation	22.98	24.78

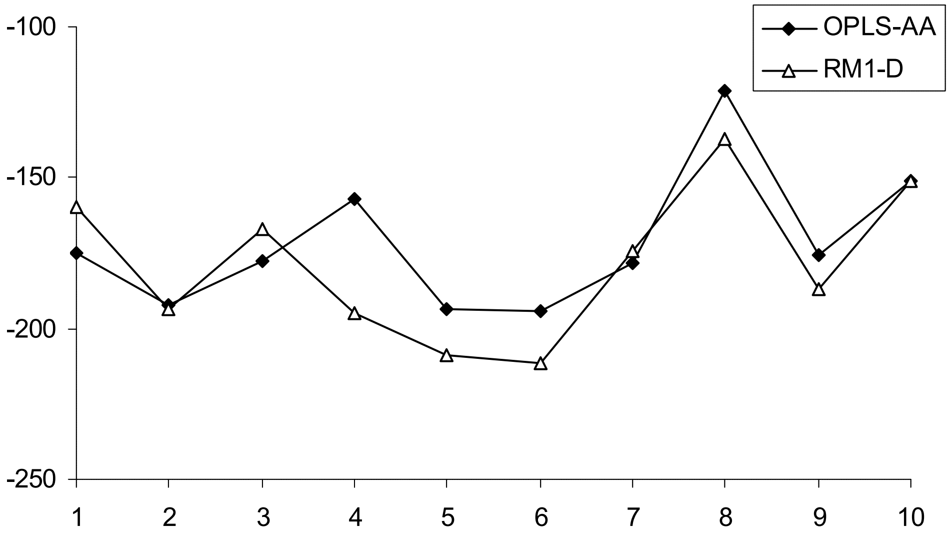


Figure 3.7: Interaction energies for ten molecular dynamics snapshots (kcal/mol).

Table 3.11 Changes in ligand Interactions
for 10 MD snapshots

	1		2	
	new	missing	new	missing
Glu85	-	-	-	-
Asn86		Arg (a)	Gly(a), Ala(a)	
Pro87			Val(b)	
Val88			Val(b)	
Val89			Val(b)	Asn(b)
His90	Thr (b)	Phe, no stacking with His	Thr(b)	His(b)
Phe91	Ala (b)	Gly (a)	Ala(a), Glu(a)	Tyr(b), Phe22(a)
Phe92			Ala(b), Gln(b)	Asp(b)
Lys93	Gln (a)	Asn (a)	Gln(b)	
Asn94		no H-bond with Asn (a)	No H-bond with Asn	
Ile95	Tyr (b)	Ile (b)	Tyr(b), Asp(a), Arg(b)	Ile(b)
Val96	Val (a)		Val(a)	
Thr97	Arg(a), Met(a), Ile(a)	Trp (b)	Met(a), Arg(a)	Ile(a)
Pro98	Ile(a)	Pro(b), Asp(b), Tyr(b)	Ile(a)	Pro(b), Asp(b), Tyr(b)

	3		4	
	new	missing	new	missing
Glu85	-	-	-	-
Asn86	Gly(a), Ala(a)	Arg, Phe	Gly(a), Ala(a)	
Pro87	Val(b)		Val(b)	
Val88	Val(b)			
Val89	Phe31(a), Phe23(a)		Ple(a), Vla(b)	
His90	Asn(a)2H-bonds, Tyr(a), Thr(a)	Phe(a), His(b), Asn(b), Tyr(b)	Thr(b)	Phe(a), No stcaking with His
Phe91	Ala(a), Glu(a)	Phe22(a)		Phe(a)
Phe92	Ala(b)	Gln(b)	Glu(a)	
Lys93	Ile(b), Gln(b)	Asn(a)	Gln(b)	
Asn94				
Ile95	Tyr(b)	Ile(b)	Tyr(b)	
Val96	Val(a)		Val(a)	
Thr97	Arg(a) 2H-bonds, Met(a)	Trp(b)		Ile(a)
Pro98		Arg(a), Pro(b)		Asp(b),Pro(b),Arg(a)

	5		6	
	new	missing	new	missing
Glu85	-	-	-	-
Asn86	Arg(a)		Ala(a), Gly(a)H-bond	
Pro87	Val(b)			Ala(a)
Val88	His(b)	Phe(a), Ala(a)	Phe(a)	
Val89	Val(b)		Phe(a),His(b), Val(b)	
His90	Thr(b)	Phe(a), No stacking with His		
Phe91	Ala(a)	Asn(a)	Ala(a), Glua)	Phe(a), Tyr(b)
Phe92	Ala(b)			Ala(b), Gln(b)
Lys93	Gln(b)		Gln(b)	Asn(a)
Asn94		No H-bond wih Asn		
Ile95		Val(a), Trp(b)	Tyr(b)	Val(a)
Val96	Val(a)		Val(a)	
Thr97	Arg(a)		Met(a), Arg(a)	Trp(b)
Pro98		Pro(b), Asp(b)	Trp(b)	Pro(b)

	7		8	
	new	missing	new	missing
Glu85	-	-	-	-
Asn86	Ala(a), Gly(a)H-bond		Ala(a), Gly(a)	
Pro87		Ala(a)	Val(b)	
Val88	Phe(a)		Val(b)	
Val89	Phe(a),His(b), Val(b)		Val(b)	
His90			Thr(b)	
Phe91	Ala(a), Glua)	Phe(a), Tyr(b)	Ala(a), Glu(a)	
Phe92		Ala(b), Gln(b)	Ala(b)	Gln(a), Asp(b)
Lys93	Gln(b)	Asn(a)	ile(b), Phe(b)	
Asn94				
Ile95	Tyr(b)	Val(a)	Asp(a), Tyr(b)	ile(b)
Val96	Val(a)		Val(a)	Trp(b)
Thr97	Met(a), Arg(a)	Trp(b)	Arg(a), Asp(a), Met(a)	Trp(b)
Pro98	Trp(b)	Pro(b)	ile(a)	Pro(b), Asp(b), Tyr(b)

	9		10	
	new	missing	new	missing
Glu85	-	-	-	-
Asn86	Val(b)	Arg(a)		
Pro87	Val(b)		Val(b)	
Val88			Glu(a)	
Val89	Phe(a)		Val(b)	
His90	Thr(b)	His(b)	Thr(b)	Phe(a)
Phe91		Tyr(b)		
Phe92		Gln(b)		Ala(b)
Lys93			Gln(b)	
Asn94	Ile(a)		Ile(a)	
Ile95		Ile(b), Trp(b)		Trp(b)
Val96	Val(a)	Ile(a), Trp(b), Asp(b)	Val(a)	Trp(b)
Thr97		Arg(a), Met(a)		Ile(a)
Pro98	Ile(a)	Pro(b), Asp(b), Tyr(b)	Ile(b), Trp(b)	Arg(a), Pro(b), Asp(b)

3.4 Conclusions:

We have tested several approximate methods against correlated *ab initio* calculations for their ability to predict the energy of interaction between amino acids, focussing on the interaction of a peptide implicated in multiple sclerosis with its biological MHC receptor. We find that the semi-empirical RM1 approach with additional correction for dispersion effects gives the best reproduction of *ab initio* data, with a mean unsigned error of a little more than 1 kcal/mol over almost 50 interactions after optimisation of the global scaling factor s_6. Performance is similar for several other parameterisations of semi-empirical theory, with RM1 chosen for its slightly better results. The atomistic forcefield OPLS-AA also shows promise, with a mean error of slightly more than 2 kcal/mol.

The computational efficiency of this approach, especially when coupled with the MOZYME method, means that study of larger systems than pairs of amino acids is feasible. Many body effects are significant in some cases, although estimates of complex stabilisation from pairwise interactions and from larger calculations are similar. Increasing the size of model systems used to represent the bound peptide from single amino acids to dipeptides and heptapeptides increases the predicted interaction energy, as does expanding the number of amino acids used to model the receptor. We also show that the interaction energy varies significantly over 10 snapshots taken from a previous molecular dynamics study of this complex, confirming that predictions from a single static structure are unlikely to be representative.

3.5 References:

[1] Mantzourani, E.; Laimou, D.; Matsoukas, M. T.; Tselios, T. *Anti-Inflammatory & Anti-Allergy Agents in Medicinal Chemistry* **2008**, *7*, 294-306

[2] Zhao, Y.; Truhlar, D. *Journal of Chemical Theory and Computation* **2007**, *3*, 289-300

[3] Meyer, E. A.; Castellano, R. K.; Francois Diederich. *Angewandte chemie* **2003**, *42*, 1210-1250

[4] Cerny, J.; Hobza, P. *Physical Chemistry Chemical Physics* **2007**, *9*, 5291-5303

[5] Grimme, S. *Journal of Computational Chemistry* **2004**, *25*, 1463-1473.

[6] Jurecka, P.; Sponer, J.; Cerny, J.; Hobza, P. *Physical Chemistry Chemical Physics*, **2006**, *8*, 1985-1993.

[7] Jurecka, P.; Cerny, J.; Hobza, P.; Salahub, D. *Journal of computaional chemistry*, **2006**, *28*, 555-569.

[8] Eistner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *Jounral of Chemical physics*, **2001**, *114*, 5149-5155.

[9] McNamara, J. P.; Hillier, I. H. *Physical Chemistry Chemical Physics* **2007**, *9*, 2362-2370

[10] Muller-Dethlefs, K.; Hopza, P. *Chemical Reviews*, **2000**, *100*, 143-167.

[11] Hobza, P.; Kabelac, M.; Sponer, J.; Mejzlik, P.; Vondrasek, J. *Journal of Computational Chemistry* **1997**, *18*, 1136-1150

[12] Bielekova, B.; Goodwin, B.; Richert, N.; Cortese, I.; Kondo, T.; Afshar, G.; Gran, B.; Eaton, J.; Antel, J.; Frank, J. A.; McFarland, H. F.; Martin, R. *Nature Medicine*, **2000**, *6*, 1167-1175.

[13] Kappos, L.; Comi, G.; Panitch, H.; Oger, J.; Antel, J.; Conlon, P.; Steinman, L. *Nature Medicine*, **2000**, *6*, 1176-1182.

[14] Berger, T.; Rubner, P.; Schautzer, F.; Egg, R.; Ulmer, H.; Mayringer, I.; Ditiz, E.; Deisenhammer, F.; Reindl, M. *New England Journal of Medicine*, **2003**, *349*, 139-145.

[15] Hahn, M.; Nicholson, M. J.; Pyrdol, J.; Wucherpfenning, K. W. *Nature immunology*, **2005**, *6*, 490-496.

[16] http://www.rcsb.org/pdb. **2010**

[17] Wucherpfenning, K. W.; Sette, A.; Southwood, S.; Oseroff, C.; Matsui, M.; Strominger, J.; Hafler, D. A. *Journal of Experimental Medicine*, **1994**, *179*, 279-290.

[18] Werner, H.-J.; Knowles, P. J. L., R.; Schu¨tz, M.; Celani,; P.; Korona, T. M., F. R.; Rauhut, G.; Amos, R. D.;; Bernhardsson, A. B., A.; Cooper, D. L.; Deegan, M.; J. O.; Dobbyn, A. J. E., F.; Hampel, C.; Hetzer, G.;; Lloyd, A. W. M., S. J.; Meyer, W.; Mura, M. E.;; Nicklass, A. P., P.; Pitzer, R.; Schumann, U.; Stoll,; H.; Stone, A. J. T., R.; Thorsteinsson, T., **2008**, MOLPRO,Version 2008.2; a package of ab initio programs. See: http://www.molpro.net.

[19] Hill, J. G.; Platts, J. A. *Journal of chemical theory computation*, **2007**, *3*, 80-85.

[20] Kendall, R. A.; Dunning, T. H. *Journal of Chemical Physics*, **1992**, *96*, 6796-6806.

[21] Weigend, F.; Kohn, A.; Hattig, C. *Journal of Chemical Physics*, **2002**, *116*, 3175-3183.

[22] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B. S., G. E.; ; Robb, M. A. C., J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J.

J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; and Pople, J. A. *Gaussian, Inc., Wallingford CT*, **2004**, Gaussian 03, Revision C.02, .

[23] Stewart, J. J. P. *Journal of Molecular Modelling*, **2009**, *15*, 765-805.

[24] Grimme, S. *Journal of computational chemistry*, **2006**, *27*, 1787-1799.

[25] http://www.chemcomp.com/journal/cstat.html. **2010**

[26] Mantzourani, E. D.; Mavromoustakos, T. M.; Platts, J. A.; Matsoukas, J. M.; Tselios, T. V. *Current Medicinal Chemistry* **2005**, *12*, 1521-1535

[27] Hausmann, S.; Martin, M.; Gauthier, L.; Wucherpfenning, K. *The Journal of Immunology*, **1999**, 338-344.

[28] Mantzourani, E. D.; Mavromoustakos, T. M.; Platts, J. A.; Brancale, A.; Tselios, T. V. *Journal of molecular graphics and modelling*, **2007**, *26*, 471-481.

Chapter 4:

Calculation of non-covalent interactions in solvent for diverse peptides:

4.1 Introduction:

Major Histocompatibility Complex (MHC) molecules are an important class of receptor in the immune system of all vertebrates: in humans they are termed human leukocyte antigens (HLA). Their role is to bind peptides presented to cell surfaces, hence allowing recognition of self or non-self and stimulating appropriate immune response in the case of non-self.

MHC receptors are generally separated into class I and class II. Both have a single peptide binding site, which in class I is made up of a single amino-acid chain, whereas in class II the active site is located at the junction between two chains [1, 2]. Incorrect recognition of self peptides as being non-self is implicated in a number of auto-immune diseases such as multiple sclerosis and rheumatoid arthritis. The exact mechanism of this is not known but the concept of “molecular mimicry”, in which certain self-peptide sequences are sufficiently similar to non-self sequences to induce immune attack on the body, has been proposed. Prediction of the key binding event between peptide and MHC is therefore desirable, both in understanding the origin of these debilitating diseases and in design of new therapies to treat them. In the previous chapter we examined the binding of a key epitope of myelin basic protein, which is known to bind tightly to MHC II. In this section, we will consider the binding of a more diverse range of peptides to this receptor, to test whether methods can distinguish binding from non-binding peptides

Peptide-receptor interactions always occur in biological solvent [3]: therefore; in order to estimate the interaction energies for these complexes in appropriate ways, solvent must be considered in calculations. Calculating interaction energies for large biological complexes in solvent by computational methods is a challenging task [3]. Many approaches have been tested to estimate the effect of the solvent in these interactions [3] as discussed above.

The speed of semi-empirical and molecular mechanics methods allows examination of the interaction energies of larger models of the peptide than single amino acids, especially when coupled with the MOZYME method. Therefore, we sought biological data to compare against the computational methods, in order to choose the most suitable method for predicting peptide-receptor interactions. IC_{50} data, *i.e.* the concentration required to inhibit 50% of binding of natural peptide in competitive binding, are widely used in such cases [4]. Although it is possible to convert IC_{50} to inhibition constant (Ki), which is directly related to binding free energy, using the Cheng-Prusoff equation [5], we were not able to perform this conversion for the peptide-MHC II complexes under study, due to lack of information about ligand and receptor concentrations in literature data. IC_{50} values are sensitive to conditions such as the temperature and solvent [6, 7], it is therefore preferable to choose sets of IC_{50} data measured in a consistent manner in the same laboratory. We have therefore concentrated on several sets of peptide-MHCII receptor complexes with IC_{50} values measured in the same conditions, and with related X-ray structures published, and used these as tests of possible methods for prediction of peptide-MHCII binding using a variety of statistical techniques. We employed many of the methods discussed above (molecular mechanics methods OPLS-AA, AMBER94, and MM/GB-VI and semi-empirical RM1-D and PM6-DH2 with COSMO [8]) to examine in detail the interaction of three sets of peptides with Major Histocompatibility Complex (MHC) class II receptor, and to compare calculated binding energies to available IC_{50} data.

4.2 Data Sets and Computational Methods:

The first set studied is derived from hen egg lysozyme (HEL), and is based on the complex of 12 amino acids (MKRHGLDNYRGY) with MHC class II (mouse I-Ag7) [4]. The X-ray structure of the peptide-receptor complex was taken from PDB entry 1F3J. IC_{50} data has been reported for analogues of the HEL peptide, in which one or more N-terminal and/or C-terminal residues are truncated to reveal the key residues for binding. IC_{50} values of 1000nM or more are denoted non-binders, and IC_{50} less than 1000nM are binders [4]. This set therefore contains 5 binders and 5 non-binders.

The second set studied is based on a complex of myelin basic protein (MBP)-derived peptide with HLA DRB1*1501 [4, 9]. It contains fourteen amino acids (ENPVVHFFKNIVTP) [4, 9], and the relevant X-ray structure was taken from PDB entry 1BX2 Figure (4.1). In this set, each amino acid is replaced in turn by Ala [9], and values of IC_{50} measured [9]. In this set, all the IC_{50} values show stable interactions according to the 1000 nM cutoff used above, and many interactions have the same value of IC_{50}. However, two peptides have rather higher IC_{50} values, in which Val89 and Phe92 are replaced by Ala. Both amino acids are known to form strong interactions, in pocket 1 and pocket 4 of the binding site, respectively [4, 9, 10] and by replacing these amino acids with Ala, the binding affinity of the peptides decreased [9].

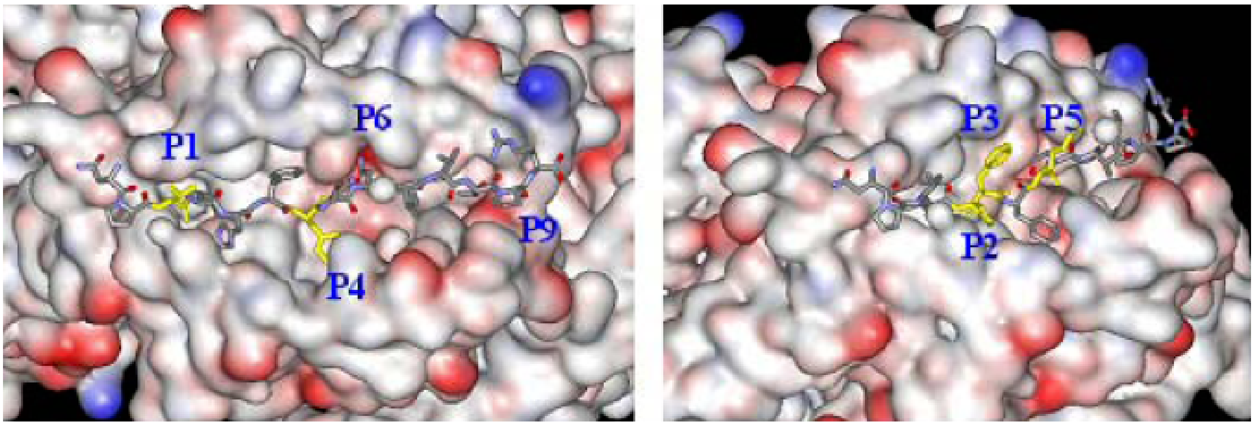


Figure 4.1 Two views of an epitope of myelin basic protein bound to an HLA receptor. P1 and P4 are the main “anchor” residues, while P3 and P5 are contacts to T-cell receptors.

A third was taken from Southwood *et al*’s study [11], and contains 22 peptides with much more diverse sequences than the first two sets interacting with HLA DRB1*0101. In this case, X-ray structures of complexes are not available. Instead, manual docking was performed by mutating amino acids to the relevant sequence in MOE, using the X-ray structure of human class II MHC protein HLa-DR1 in complex with the tight binding peptide A2 (103-117) [12] (PDB code 1AQD) as a template. In order to guide this procedure, possible amino acids that could act as “anchors” within binding pockets were identified by means of sequence-based prediction methods SYFPEITHI and SVMHC, as well as the algorithm set out by Southwood *et al* [11].

SYFPEITHI is a databank and prediction algorithm for peptide-MHC binding, and contains a large range of ligands and peptide motifs, used to predict the peptide binding with MHC receptor [13-15] based on published motifs of amino acids and anchor positions. It calculates a score to identify the amino acid as anchor, auxiliary anchor, preferred residues or if the amino acid has “a negative effect on the binding ability” [14].

SVMHC is a prediction server for MHC class I and II, used to test the ability of peptides to bind with different MHC alleles, and to find the best “binders in a protein sequence” [16]. According to Donnes and Elofsson, the performance of SVMHC and SYFPEITHI for six MHC types common between these methods are compared [16], with SVMHC giving 95% correct predictions and 91% for SYFPEITHI [17]. The final sequence-based prediction method used is the algorithm set out by Southwood *et al*, which is specific for the DRB1*0101 allele MHC class II [11]. Each residue has value based on its position on the receptor, encoded into an in-house awk program to evaluate the most likely binding sites of the peptide based on these values.

The X-ray crystallographic coordinates were obtained from PDB entry 1BX2 for MBP peptide and 1F3J for HEL peptide [4, 18-20]. For the Southwood data set, three prediction methods (SYFPEITHI, SVMHC and Southwood) were used to identify the best peptide anchors that fit in the receptor pockets. For the web-based prediction methods SVMHC [17] and SYFPEITHI [15], the single letter of the residues of a peptide have been typed into the website. Then the correct allele and the nonamer peptide length have been selected. These methods give a score for all possible alignment of a peptide in the receptor. The more positive a value is the more preferable the binding sites. An algorithm method which developed by Southwood et al [11] for the DRB1*0101 allele has also been used. This method was used by an awk script written by Dr Platts. The single letter residue of a peptide under the study was saved in a plain text file. When the program was run, it gives a value for the most likely alignment of a peptide in the receptor. Again the most positive value is the most likely alignment of a peptide in the receptor. We choose the best alignment of peptide in the receptor as a consensus of these methods, and only this alignment was used in further study. Table

(4.1) shows the prediction scores for these three methods. This table also shows that there are seven peptides from 22 peptides (blue colour) where just two methods predict the same anchor residue where the third method predicts another residue as an anchor. We have chosen the alignment which is predicted by two methods or more.

Table 4.1: The scores and the best alignment of peptides for 22 Southwood set according to SYFPEITHI, SVMHC and Southwood algorithm methods.

Peptide No.	Peptide's alignment	SVMHC	SYFPEITHI	AWK
1188.34	HN**W**VNHAVPLAM	0.87	29	35.32
1188.16	KSK**Y**KLATSVLAGL	2.2	26	270.41
1136.47	HHY**F**VDLIGGAMLS	0.38	35	0
1188.32	GLA**Y**KFVVPGAATP	0.55	32	1.34
1136.16	LTS**Q**FFLPALPVFT	-995.9	25	0
	SQF**F**LPALPVFTWL	1.6	16	57.58
27.415	VKY**L**VIVFLIFFDL	0.75	26	0
	YLV**I**VFLIFFDL	1.5	26	0
27.403	**L**VNLLIFHING	1.6	25	31.17
	NLL**I**FHINGKIIK	0.7	31	0
1136.21	PQE**W**KPAITVKVLP	1.2	26	183.62
1136.28	AII**F**LFGPPTALRS	1.49	26	4.8
1136.11	**V**VFPASFFIKL	0.4	24	7.29
	PAS**F**FIKLPIILA	1.1	27	6.8
1136.14	TCF**L**IPLTSQFFLP	0.5	30	8.85
1188.13	AG**L**LGNVSTVLLG	0.23	23	4.92
	AGL**L**GNVSTVLLGG	-0.05	28	7.57
1136.24	SNV**L**ATITTGVLDI	-2.3	28	0.93
1136.12	IK**L**PIILAFATCF	0.6	25	10.81
	KLP**I**ILAFATCFLI	0.79	26	5.29
27.392	SSV**F**NVVNSSIGLI	0.35	23	0.14
	VFN**V**VNSSIGLIM	-0.62	25	0
27.417	**V**KNVIGPFMKA	0.4	31	0
1136.55	PLS**Y**NYIPVNSN	-0.1	17	3.7
1136.71	GST**Y**AASSATSVD	-0.31	27	29.77
1136.38	**S**SIIFGAFPSL	-998.7	25	0
27.388	**M**RKLAILSVSS	0.8	28	285.9
1136.59.01a	**R**VYQEPQVSPP	-996.3	22	0
1136.46	WST**M**YLTHHYFVDL	-1.5	17	0

Coordinates were loaded into MOE, and protonated according to typical protonation states. All hydrogen positions were optimised using the AMBER94 forcefield, with heavy atoms fixed at their X-ray positions. MOE program was used to calculate interaction energies using OPLSAA and AMBER94 force fields with dielectric constant 1 (vacuum), 4, 20 and 78.4 (water) [1, 9]. MOE was used to calculate interaction energies with the Born solvation model, and also binding free energies with the MM/GBVI method. For this method, the dielectric constant is estimated according to the atoms present in the receptor, and a constrained energy minimization performed for ligand atoms [19, 20].

MOPAC was used to carry out semi-empirical calculations, using the RM1-D tested in our previous study (Chapter 3) [10] and also the recent PM6-DH2 method, incorporating corrections for both dispersion and hydrogen bonding [21, 22]. For larger systems we used the MOZYME keyword to accelerate the calculations[2]. COSMO was used to estimate the effect on a solvent [23], with the same values for dielectric constant noted above [11]and NSPA (number of geometrical segments per atom) [19]equal to 122.

Several statistical tests were used to investigate the suitability of different theoretical methods for prediction of peptide-MHCII bonding, using published IC_{50} values as a test. Specifically, we employed the standard Pearson R^2 value against the negative log of IC_{50} values, Spearman’s rank correlation coefficient [24], and area under relative operating characteristic (ROC) curves by using the ROCkit package [25, 26]. In each case, a value of 1.0 indicates the ideal of perfect prediction.

4.3 Results and discussions:

Table (4.2) reports IC_{50} and interaction energies from OPLS-AA, AMBER94, MM/GBVI and RM1-D for the series of peptides based on HEL. According to Tsai (2002), the value 4 of the dielectric constant is suitable to be used in protein interaction [27], and so was employed here. Sequential removal of one to three residues from the C-terminus of the native peptide increases IC_{50}, a trend that is reflected in interaction energies from all methods considered. In contrast, removal of the N-terminal methionine residue actually increases potency: three of the five methods reflect this in increased binding, and the remaining two methods show only very small change in interaction energy. The shortest sequence, KRHGLDNY, is the least potent peptide in this data set, and again all methods predict weak binding for this peptide. From the GB-VI results, we can see that the binding energy is approximately additive: for example, removal of M from the N-terminus of the peptide reduces binding energy by 1 kcal/mol independently of the other residues present. Similarly, simultaneous removal of both M and Y from N- and C-termini reduces binding by 10.5 kcal/mol, a value that is very close to the sum of individual values (1.0 and 9.4 kcal/mol, respectively).

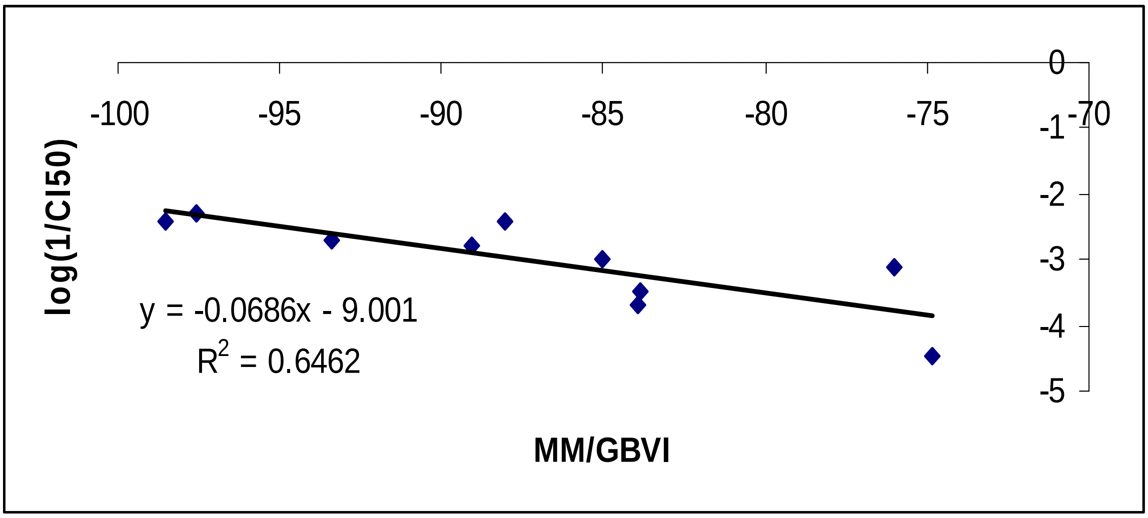
Table 4.2: IC_{50} values (nM) and interaction energies (kcal/mol) for HEL along with R^2, rank correlation, and ROC area for each method.

	IC_{50}	MM/	RM1-D/	PM6-DH2/	OPLS-AA/	AMBER94/
Peptide		GBVI	COSMO	COSMO	Born	Born
MKRHGLDNYRGY	**250**	**-98.55**	**-214.89**	**-347.72**	**-129.53**	**-130.69**
MKRHGLDNYRG	600	-89.11	-205.87	-344.21	-119.61	-120.62
MKRHGLDNYR	1000	-85.05	-188.91	-312.20	-112.30	-111.85
MKRHGLDNY	1250	-75.98	-145.17	-293.10	-99.28	-95.17
KRHGLDNYRGY	200	-97.57	-209.74	-337.18	-138.25	-135.53
KRHGLDNYRG	**250**	**-88.03**	**-200.92**	**-334.00**	**-128.33**	**-125.49**
KRHGLDNYR	5000	-83.94	-183.75	-302.20	-121.03	-116.70
KRHGLDNY	30000	-74.83	-140.01	-282.69	-108.02	-100.01
RHGLDNYRGY	500	-93.37	-150.54	-303.55	-125.39	-120.58
RHGLDNYRG	3000	-83.83	-141.75	-300.10	-115.47	-110.54
R^2		0.65	0.46	0.65	0.43	0.55
Rank correlation		0.88	0.82	0.63	0.80	0.86
*(IC_{50}-average) rank correlation		0.92	0.90	0.78	0.73	0.82
ROC area (cutoff 1000 nM)		1.00	0.93	0.96	0.96	1.00

* Rank correlation for the set after taking the average values for peptides in **bold** (IC_{50} = 250).

Statistical measures across the entire data set clarify the differences in methods. Plotting log(1/IC_{50}) against interaction energy gives some correlation for all methods, but noticeably superior performance for MM/GBVI over others considered Figure (4.2). The pattern is similar, but not as clear cut, when considering rank correlation, whether using raw or averaged data. ROC data shows that MM/GBVI and AMBER94/Born are able to unambiguously separate binders from non-binders with no false positive or negatives, whereas PM6-DH2/COSMO, RM1-D/COSMO and OPLS-AA/Born cannot. However, even those methods give high values, indicating that their predictive ability remains rather good Figure (4.3) (4.4) (4.5) (4.6).

a)



b)

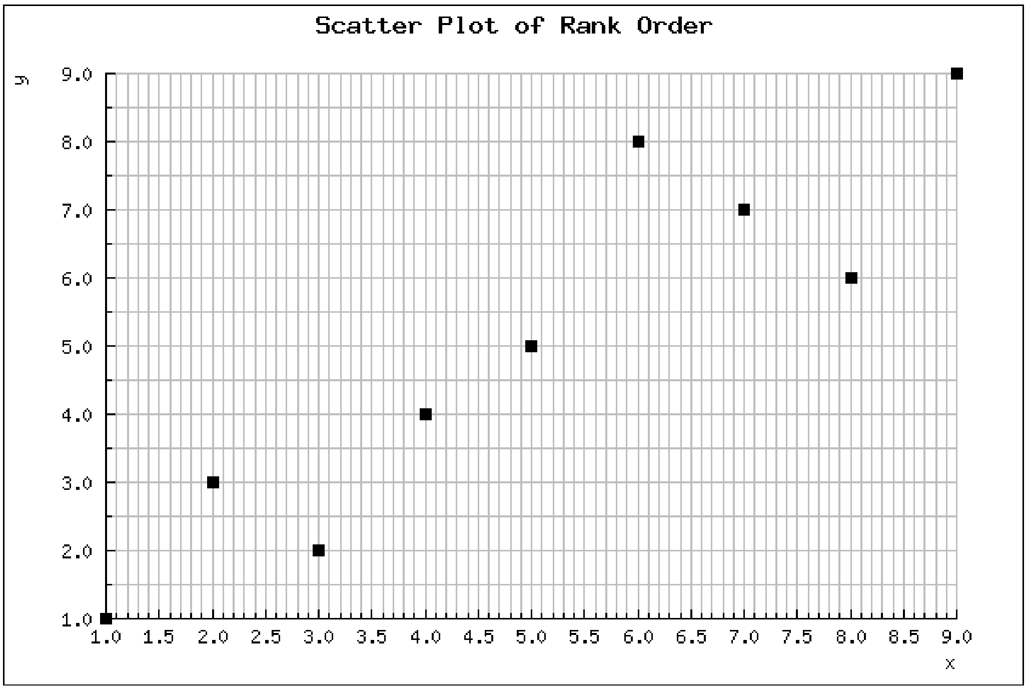


Figure 4.2 a) Linear and b) rank correlations from MM/GB-VI data for HEL data set. ROC curve not shown due to perfect prediction.

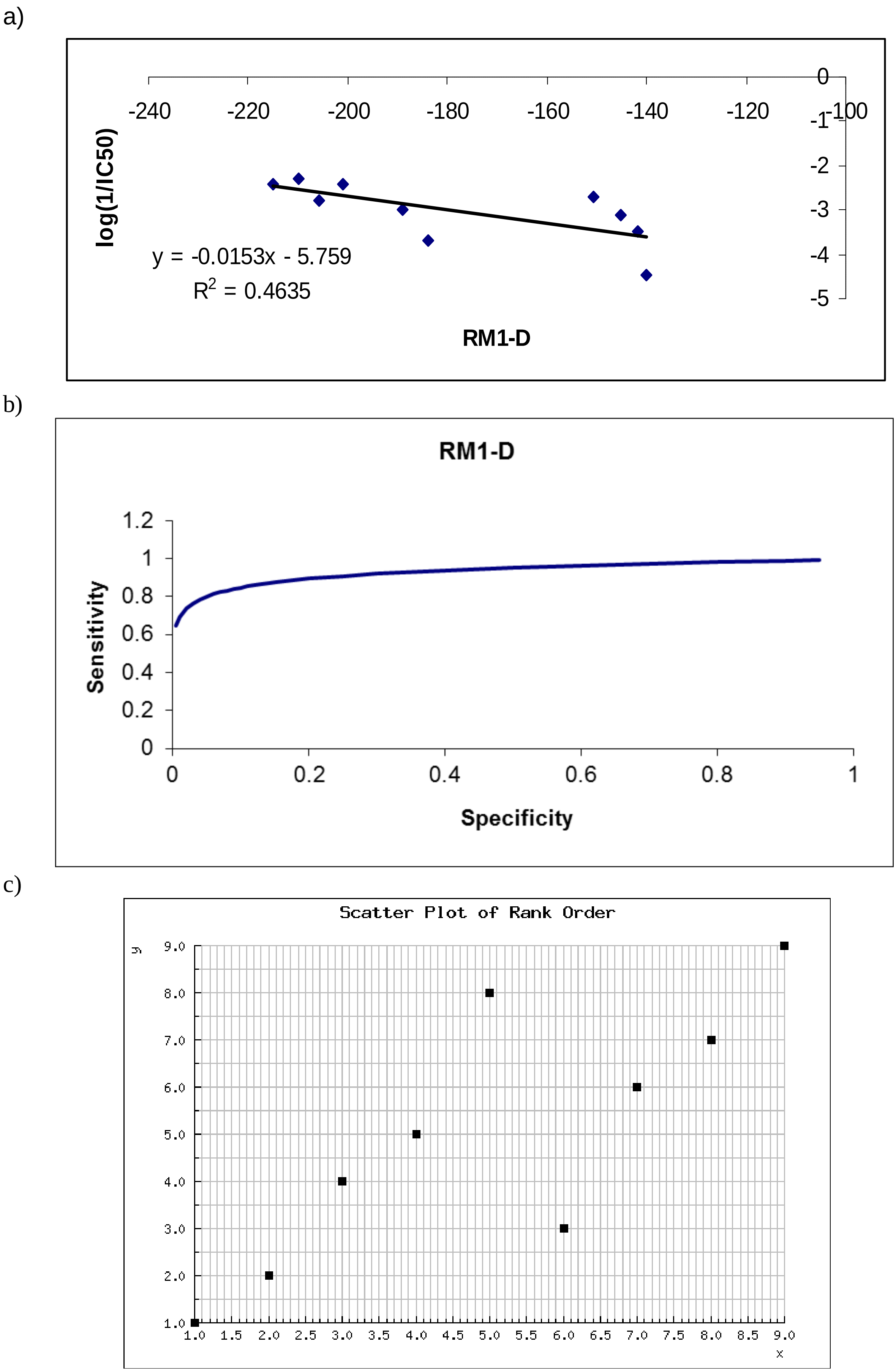


Figure 4.3 a) Linear, b) ROC curve and c) rank correlations from RM1-D data for HEL data set.

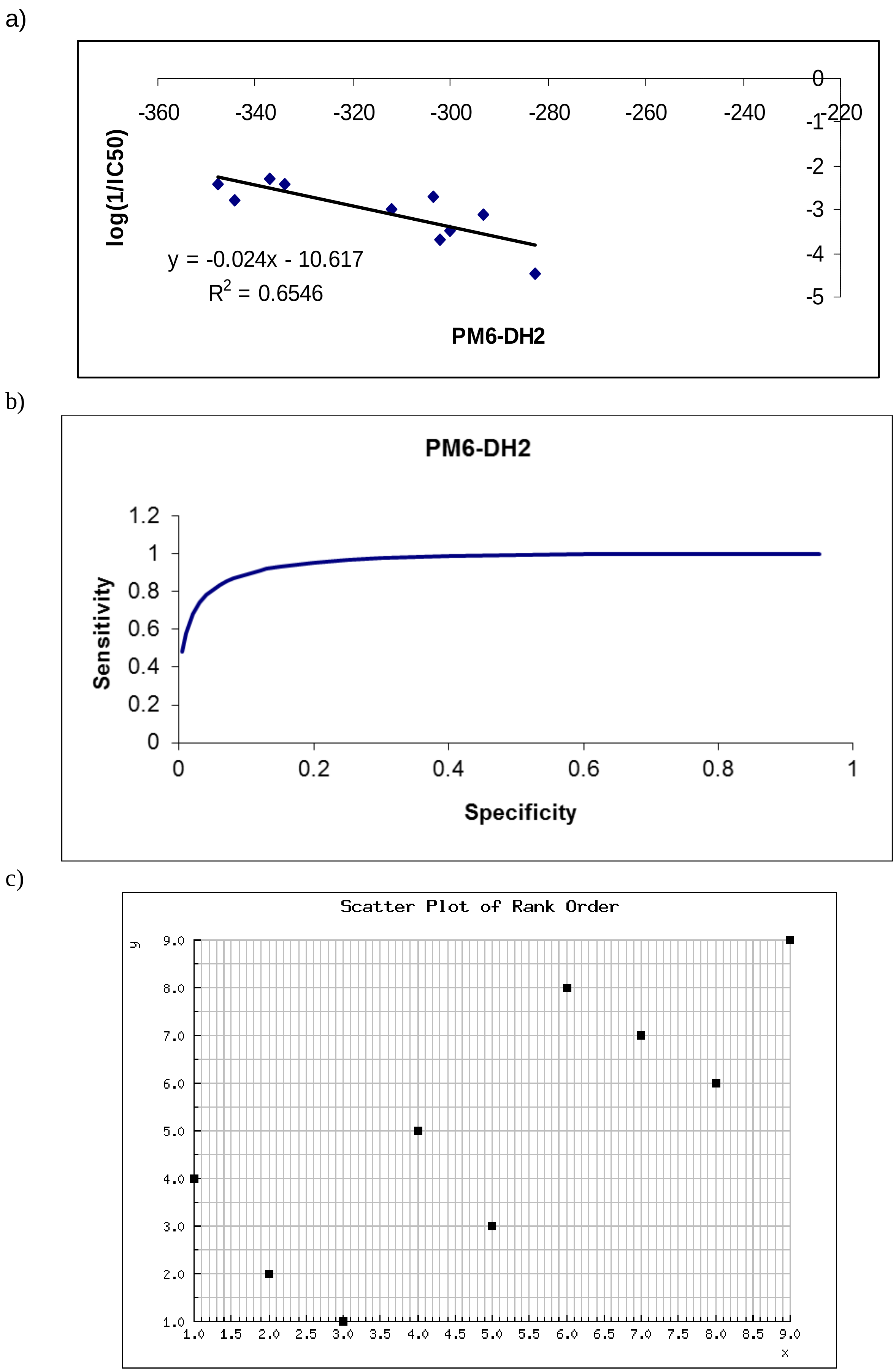


Figure 4.4 a) Linear, b) ROC curve and c) rank correlations from PM6-DH2 data for HEL data set.

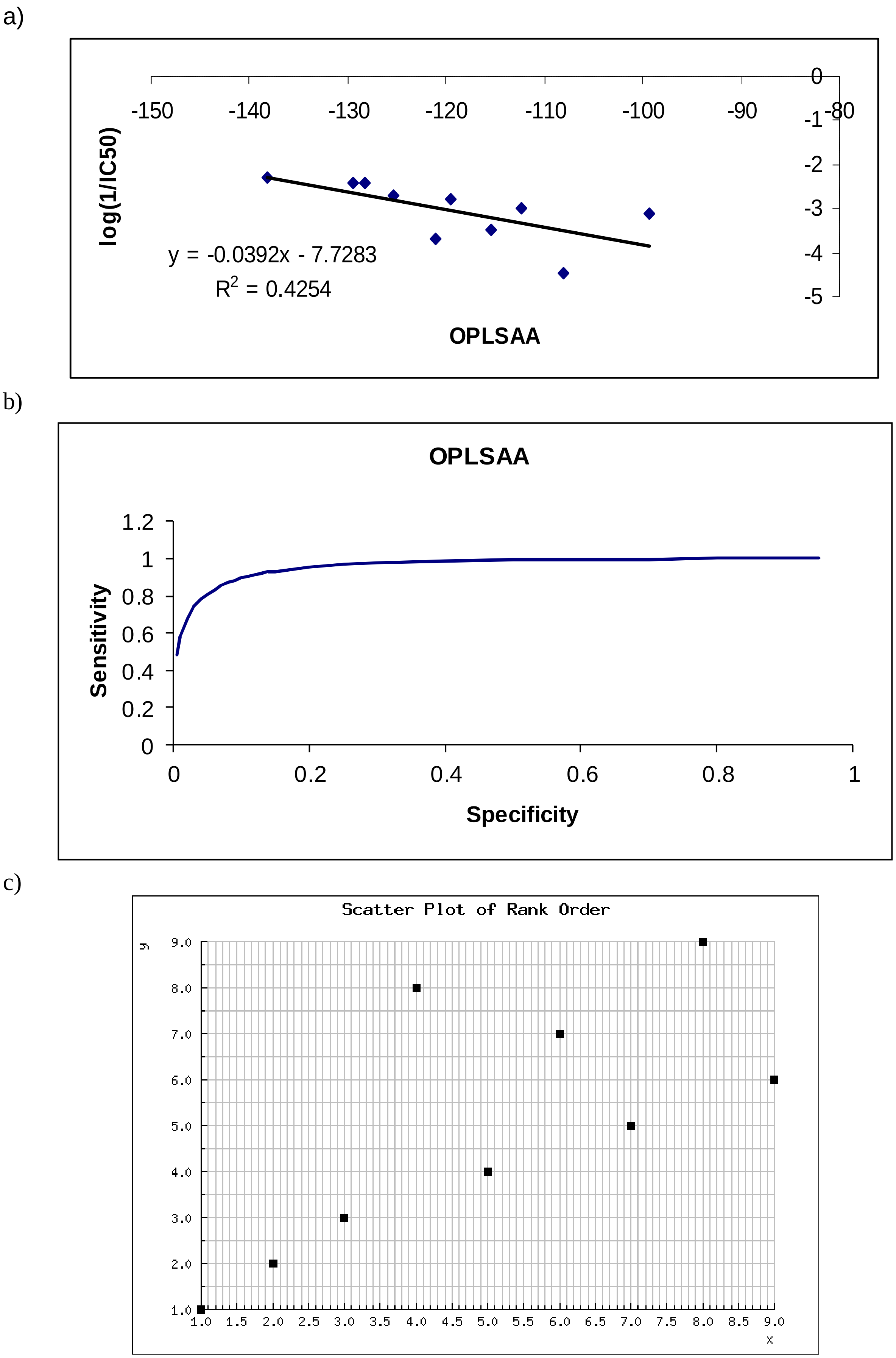
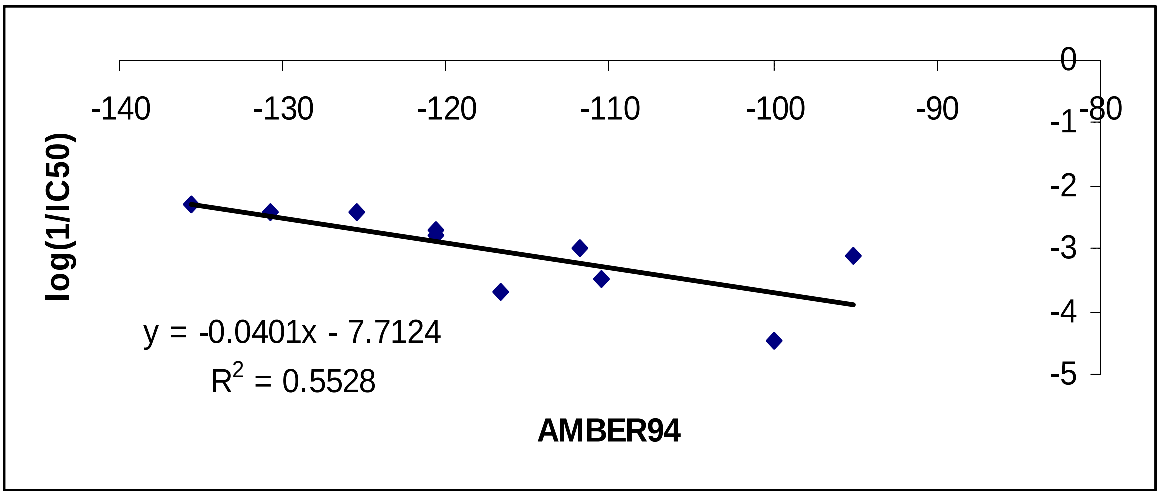


Figure 4.5 a) Linear, b) ROC curve and c) rank correlations from OPLSAA data for HEL data set.

a)



b)

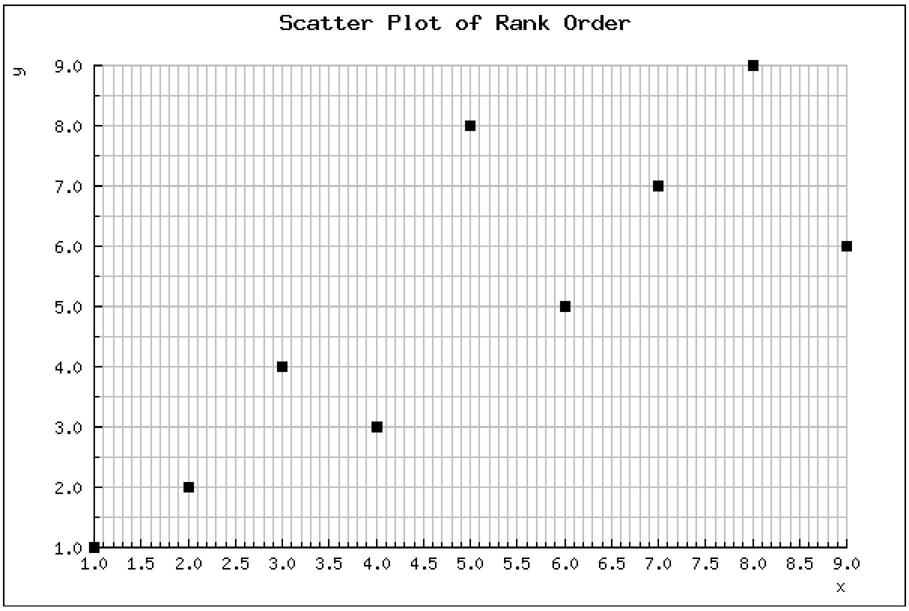


Figure 4.6 a) Linear and b) rank correlations from AMBER94 data for HEL data set. ROC curve not shown due to perfect prediction.

Table (4.3) reports similar data for the data set consisting of peptides based on MBP. In this case, all but two peptides are quite strongly bound to the receptor, and also exhibit very low IC_{50} values. The two exceptions to this are for mutation of Val89 and Phe92, which are well-known to be important as “anchor residues”: mutation of these into alanine significantly increases IC_{50} values. All methods considered predict that the F92A mutation is particularly weakly bound. The instability of the F92A mutant is most marked with the forcefield method: OPLS-AA predicts that this peptide is not bound at all to the receptor. In contrast, the semi-empirical methods succeed in predicting the relatively weak binding of the V89A mutant, whereas with force field methods this mutant does not stand out as being more weakly bound than other peptides.

Table 4.3: IC_{50} values (nM) and interaction energies (kcal/mol) for MBP along with R^2, rank correlation, and ROC area for each method.

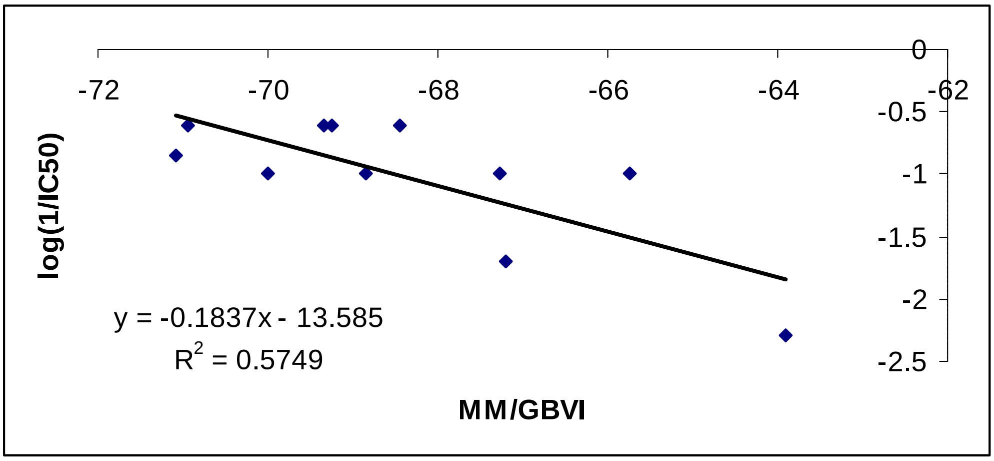
		MM/	RM1-D/	PM6-DH2/	OPLS-AA/	AMBER94/
Peptide	IC_{50}	GBVI	COSMO	COSMO	Born	Born
EAPVVHFFKNIVTP	7	-71.09	-20.38	-124.32	-12.86	-14.50
ENAVVHFFKNIVTP	**10**	**-70.00**	**-18.01**	**-125.35**	**-9.24**	**-15.18**
ENPAVHFFKNIVTP	**10**	**-68.85**	**-19.83**	**-127.46**	**-10.72**	**-18.02**
ENPVAHFFKNIVTP	50	-67.19	-15.66	-118.41	-6.64	-14.23
ENPVVAFFKNIVTP	**10**	**-65.74**	**-17.54**	**-124.17**	**-8.68**	**-13.08**
ENPVVHAFKNIVTP	**10**	**-67.28**	**-18.19**	**-124.50**	**-6.28**	**-13.20**
ENPVVHFAKNIVTP	199	-63.90	-13.41	-117.60	+0.05	-5.55
ENPVVHFFKAIVTP	4	-68.45	-17.22	-113.15	-29.96	-37.80
ENPVVHFFKNAVTP	4	-70.95	-20.91	-128.37	-14.70	-23.28
ENPVVHFFKNIATP	4	-69.25	-18.76	-127.31	-9.29	-16.37
ENPVVHFFKNIVAP	4	-69.35	-21.79	-128.18	-29.98	-34.72
R^2		0.57	0.69	0.22	0.46	0.45
*(IC_{50}-average) rank correlation		1.00	0.90	0.60	1.00	0.90

* Rank correlation for the set after taking the average energies for peptides with IC_{50} = 4 nM (underlined) and for peptides with IC_{50} = 10 nM (**bold**).

The R^2 statistic indicates reasonable correlation between IC_{50} and RM1-D interaction energy, a slightly worse correlation with MM/GBVI data, and poor correlations with OPLS-AA, AMBER94 and PM6-DH2 data. Application of the rank correlation statistic is not straightforward for this data set, since four peptides have IC_{50} = 4 nM and a further four have IC_{50} = 10 nM. Therefore, we took the average energies for the peptides with

IC_{50} = 4 nM and the average energies for the peptides with IC_{50} = 10 nM and used these averages on the calculation of the rank correlation of this set. The standard cutoff of 1000 nM to distinguish binders from non-binders for ROC analysis is inappropriate in this case Figure (4.7) (4.8) (4.9) (4,10) (4.11).

a)



b)

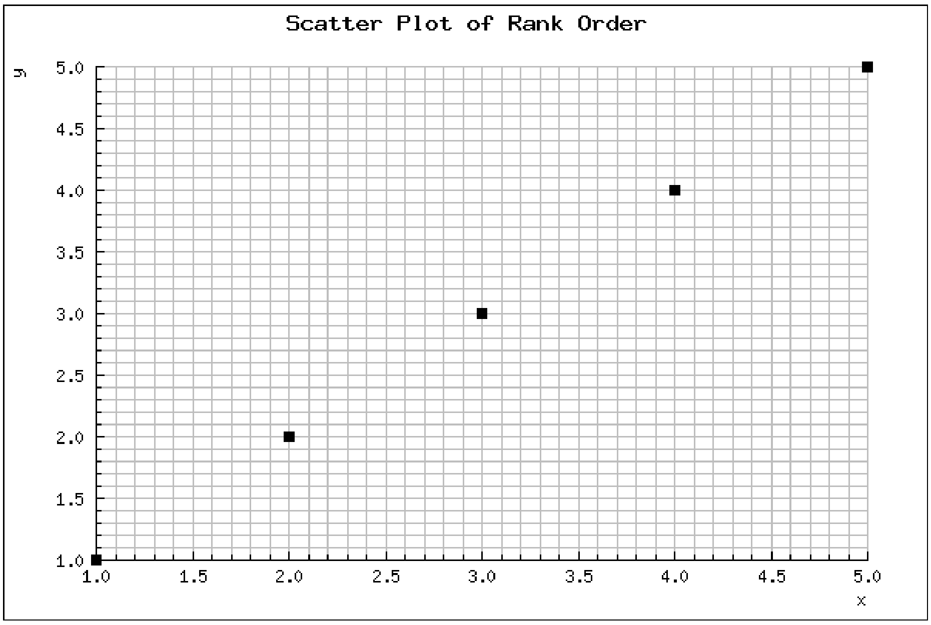
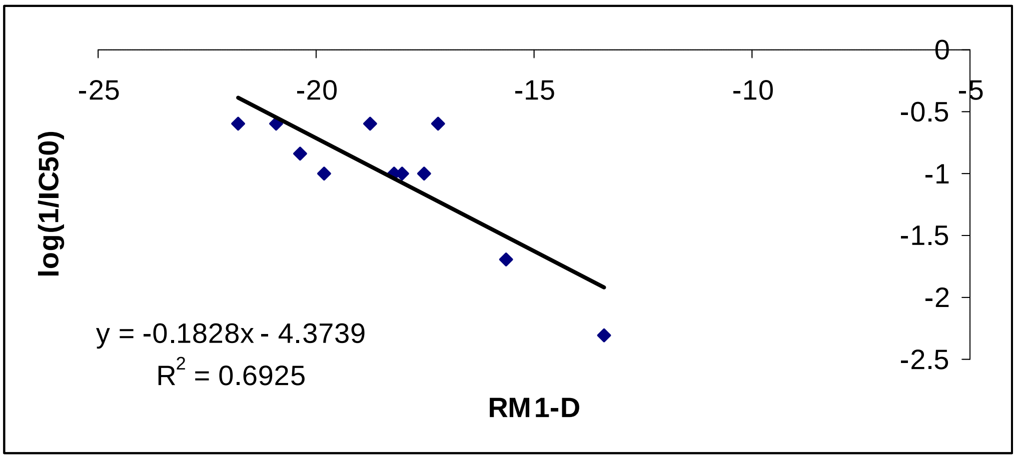


Figure 4.7 a) Linear and b) rank correlations from MM/GBVI data for MBP data set.

a)



b)

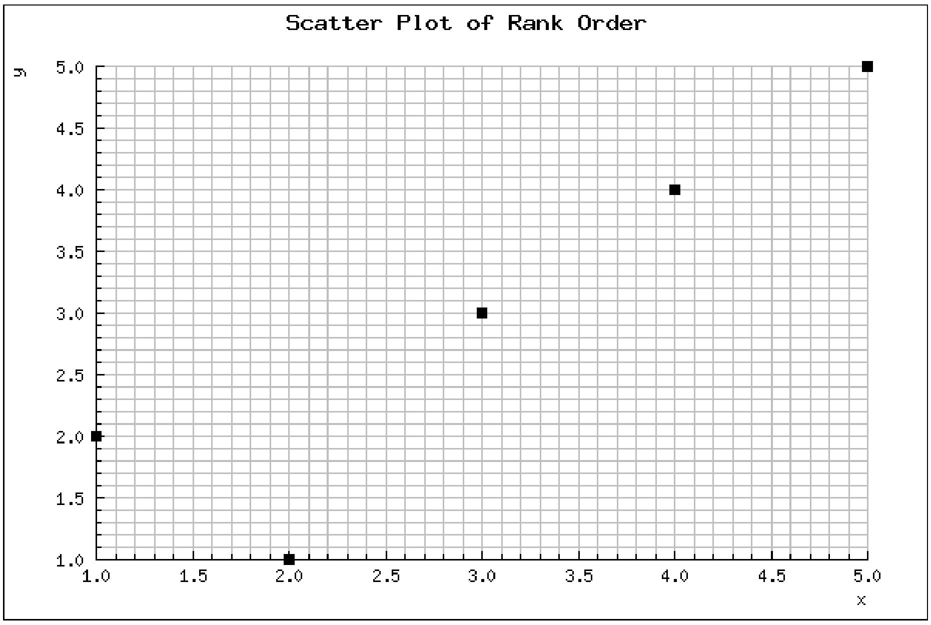
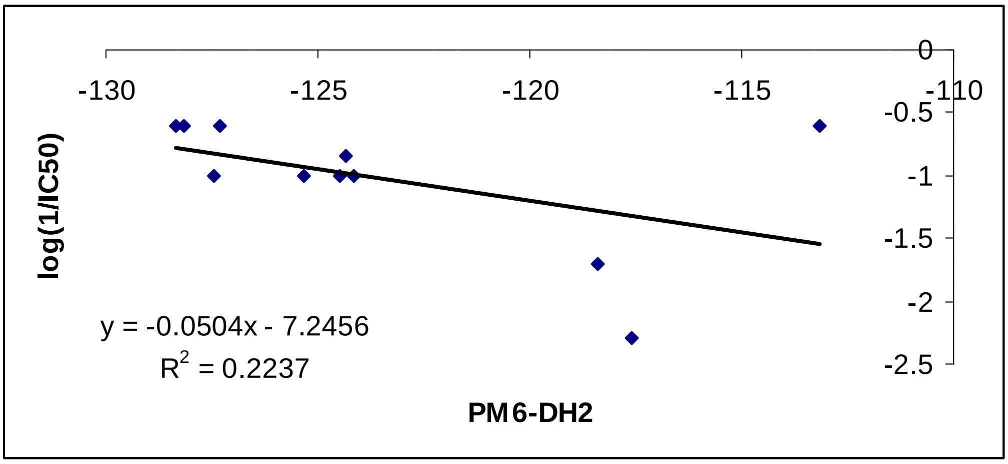


Figure 4.8 a) Linear and b) rank correlations from RM1-D data for MBP data set.

a)



b)

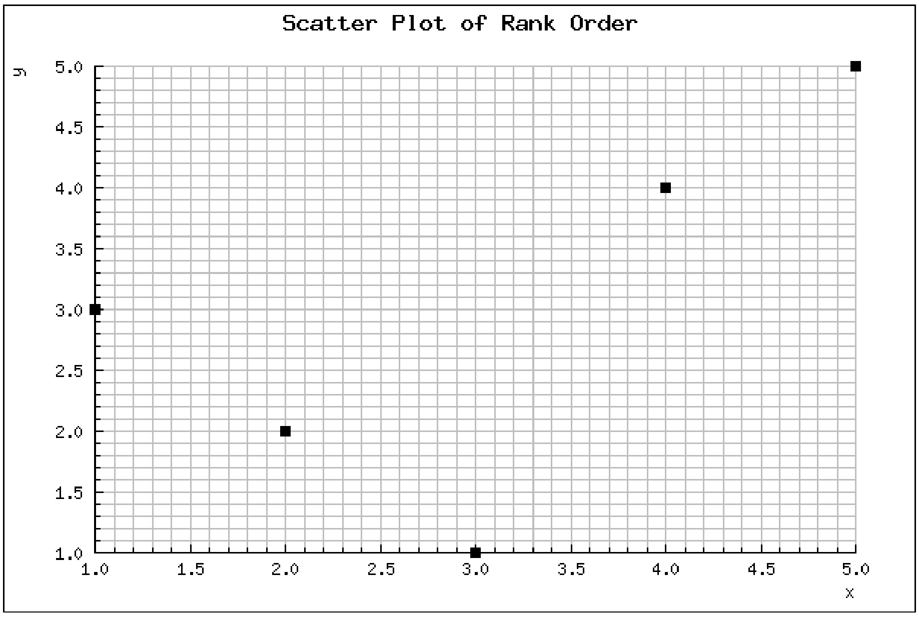


Figure 4.9 a) Linear and b) rank correlations from PM6-DH2 data for MBP data set.

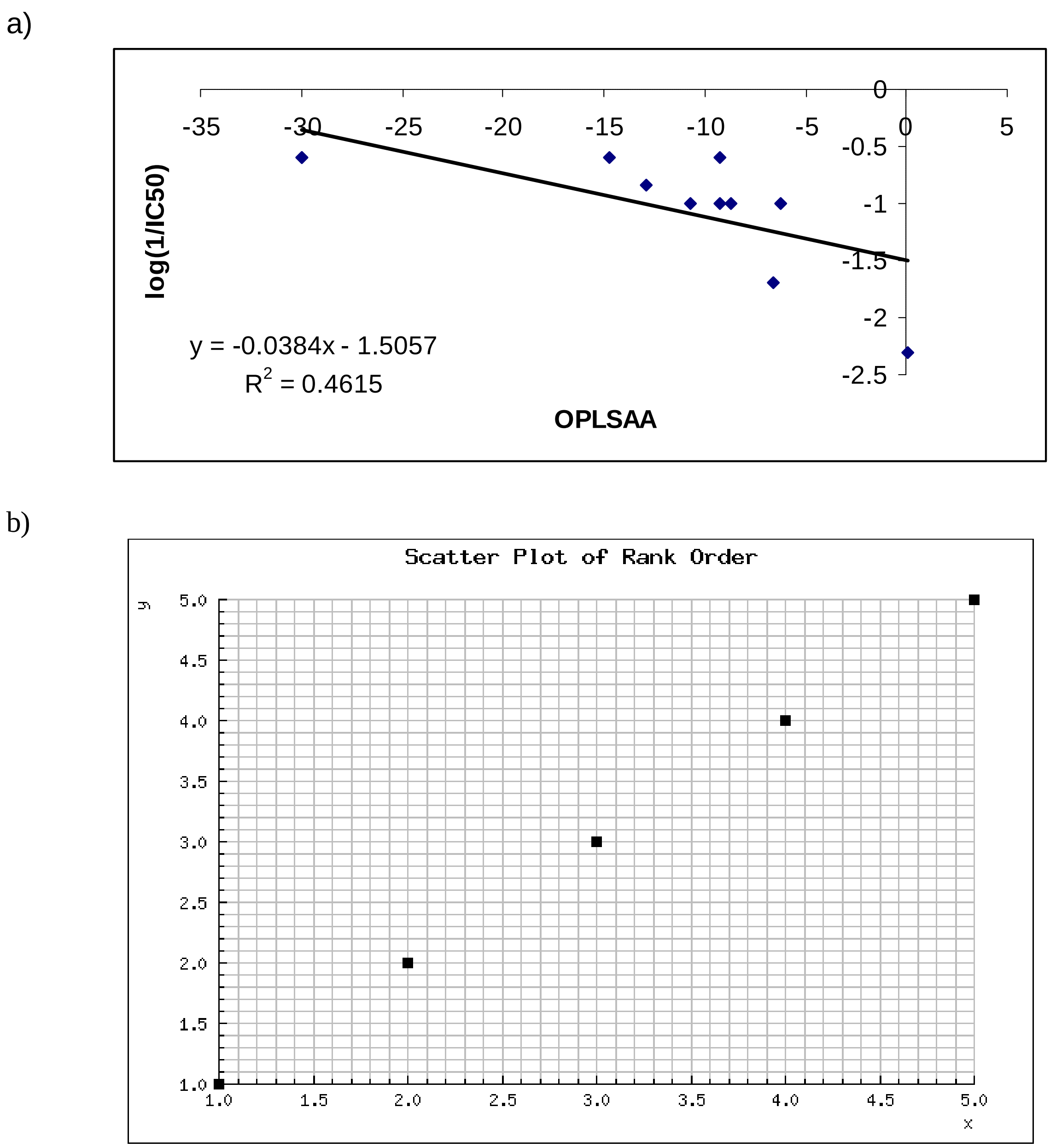
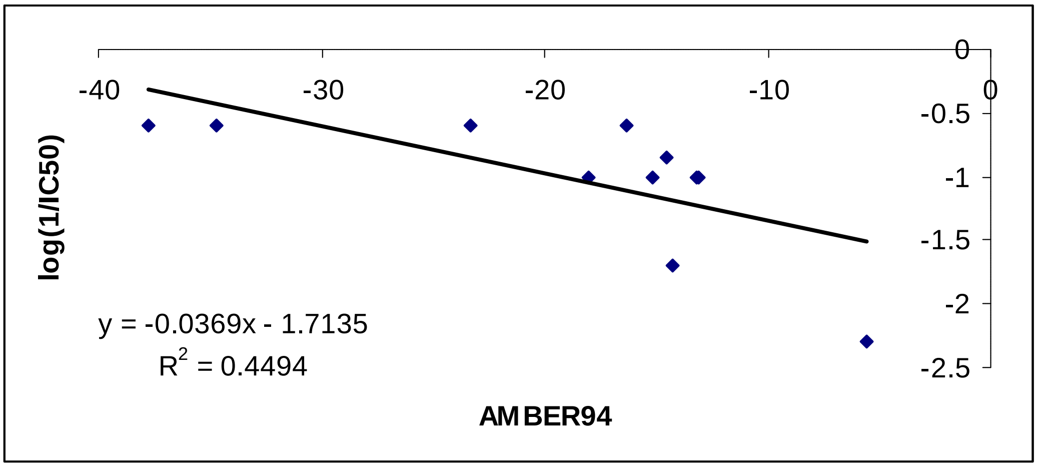


Figure 4.10 a) Linear and b) rank correlations from OPLSAA data for MBP data set.

a)



b)

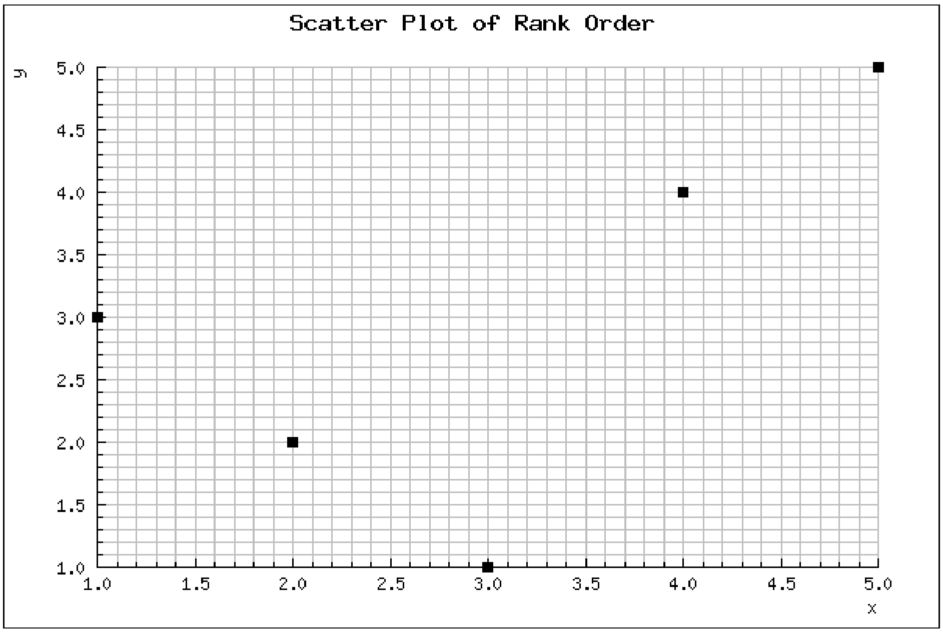


Figure 4.11 a) Linear and b) rank correlations from AMBER94 data for MBP data set.

While the results for HEL and MBP data sets are encouraging, the structural similarities and restricted range of IC_{50} data (particularly for MBP) mean that more stringent tests are required before we can reach any conclusions on the suitability of the methods examined. For this, we employed Southwood *et al*'s set of 22 structurally diverse peptides bound with IC_{50} values ranging from below 2 to over 2000 nM [11]. Initially, SYFPEITHI and SVMHC prediction servers, along with our own implementation of Southwood *et al*'s algorithm, were used to identify the best alignment for each peptide. This alignment was then constructed by manual mutation of PDB structure 1AQD in MOE, and energy minimized. The peptide on 1AQD PDB structure contains 14 residues, with the fourth residue located in pocket 1 of the receptor. So, in order to mutate this peptide to Southwood's peptides, we located the anchor residue in pocket 1 and then mutated the rest of the original peptide from 1AQD to that employed by Southwood *et al*. By using this technique we included the core residues of the peptides (located on pocket 1 to pocket 9 of the receptor, the important binder residues) and some more residues of the peptides to our calculations, but we missed few residues from each peptide on the set. On table (4.4) and (4.5), we underline the residues included in our calculations and identify the residue in pocket 1 in bold red.

Table 4.4: IC_{50} values (nM) and interaction energies (kcal/mol) for Southwood data set along with R^2, rank correlation, and ROC area for each method.

			MM/	RM1-D/	PM6-DH2/	OPLS-AA/	AMBER94/
Peptide No.	Sequence*	IC_{50}	GBVI	COSMO	COSMO	Born	Born
1188.34	HN**W**VNHAVPLAMKLI	14	-40.62	-85.53	-152.33	-143.45	-126.40
1188.16	KSK**Y**KLATSVLAGLL	3.7	-49.04	-182.02	-246.12	-138.83	-139.35
1136.47	THHY**F**VDLIGGAMLSL	2.2	-57.48	-26.04	-101.91	-145.89	-143.04
1188.32	GLA**Y**KFVVPGAATPY	3.1	-42.75	-105.60	-172.91	-129.11	-120.34
1136.16	LTSQF**F**LPALPVFTWL	1.6	-53.28	-71.09	-143.94	-148.22	-138.43
27.415	NVKYLV**I**VFLIFFDL	2011	+17.46	-4.54	-71.56	-93.76	-95.87
27.403	**L**VNLLIFHINGKIIK	78	-13.19	-77.91	-127.79	-168.31	-128.50
1136.21	IPQE**W**KPAITVKVLPA	2.2	-36.32	-130.90	-209.19	-132.11	-117.41
1136.28	LAAII**F**LFGPPTALRS	0.23	-53.05	-90.94	-161.75	-140.97	-135.16
1136.11	VVFPAS**F**FIKLPIILA	0.89	-59.32	-84.21	-146.19	-155.07	-137.81
1136.14	FATCF**L**IPLTSQFFLP	5.3	-24.52	-63.29	-136.09	-133.68	-131.21
1188.13	AGL**L**GNVSTVLLGGV	116	-28.96	-86.62	-149.38	-115.74	-102.50
1136.24	NLSNV**L**ATITTGVLDI	182	-25.61	-27.96	-96.12	-113.74	-96.22
1136.12	IKLP**I**ILAFATCFLIP	105	40.92	-118.30	-150.80	-107.73	-98.42
27.392	SSV**F**NVVNSSIGLIM	41	-38.79	-51.92	-123.90	-133.96	-123.70

27.417	**V**KNVIGPFMKAVCVE	56	-53.73	-100.38	-152.69	-128.45	-126.36
1136.55	QEIDPLS**Y**NYIPVNSN	65	-11.14	-7.50	-78.95	-119.45	-102.80
1136.71	EPQGST**Y**AASSATSVD	5.1	-58.73	-16.20	-96.59	-127.66	-113.00
1136.38	**S**SIIFGAFPSLHSGCC	70	-8.49	-33.79	-85.15	-90.11	-85.43
27.388	**M**RKLAILSVSSFLFV	50	-13.22	-73.82	-125.30	-143.71	-128.80
1136.59.01a	**R**VYQEPQVSPPQRAET	130	+29.36	-28.23	-86.59	-94.42	-110.26
1136.46	LWWST**M**YLTHHYFVDL	68	-9.91	-106.31	-190.35	-135.71	-128.45
R^2			0.54	0.14	0.23	0.36	0.48
Rank correlation			0.78	0.29	0.48	0.66	0.74
ROC area			0.93	0.62	0.75	0.79	0.87

* The underlined residues are the residues which we included in our calculations and the residues on bold red are the residue which located on pocket one of the receptor.

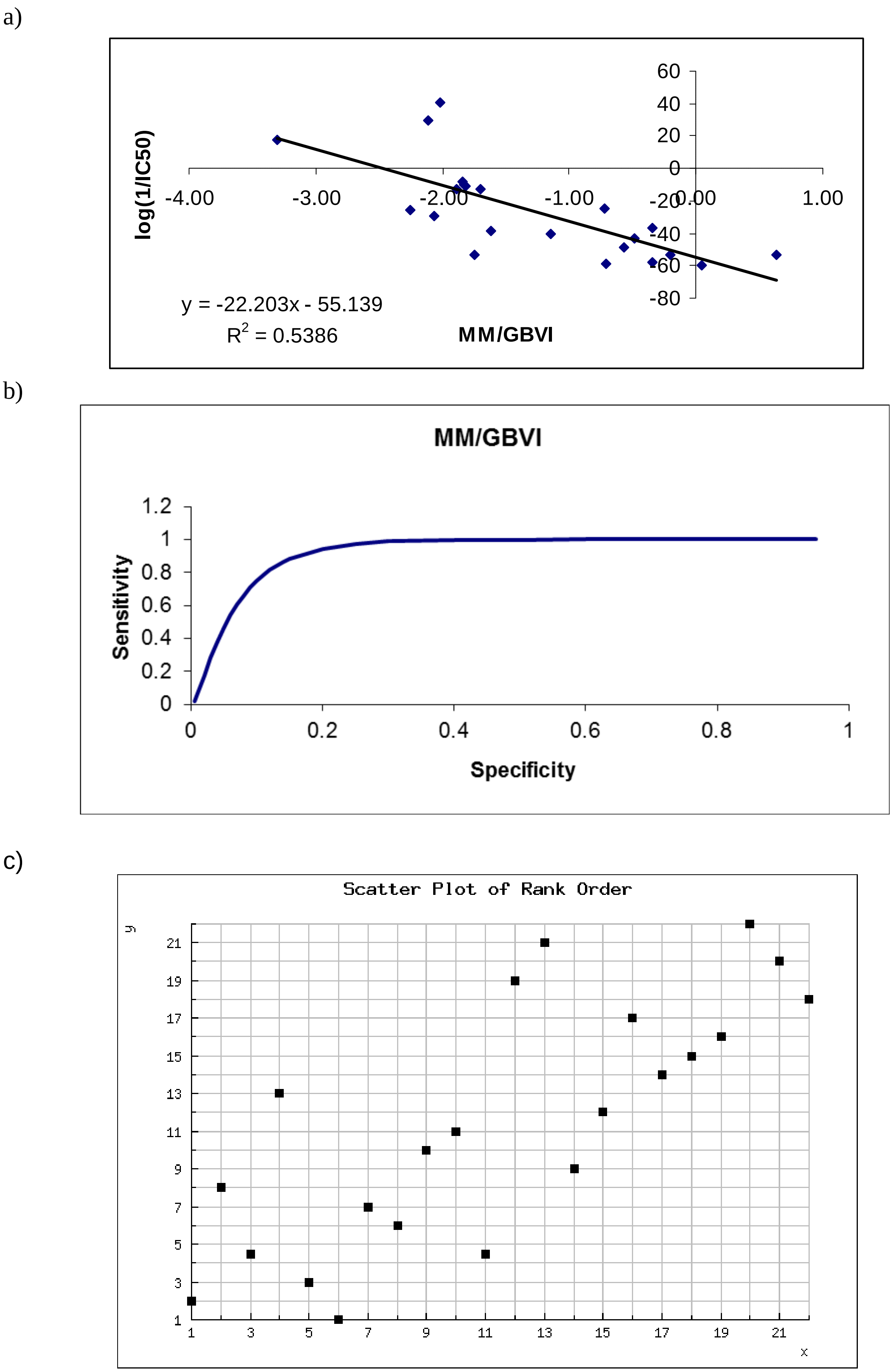


Figure 4.12 a) Linear, b) ROC curve and c) rank correlations from MM/GBVI data for Southwood data set.

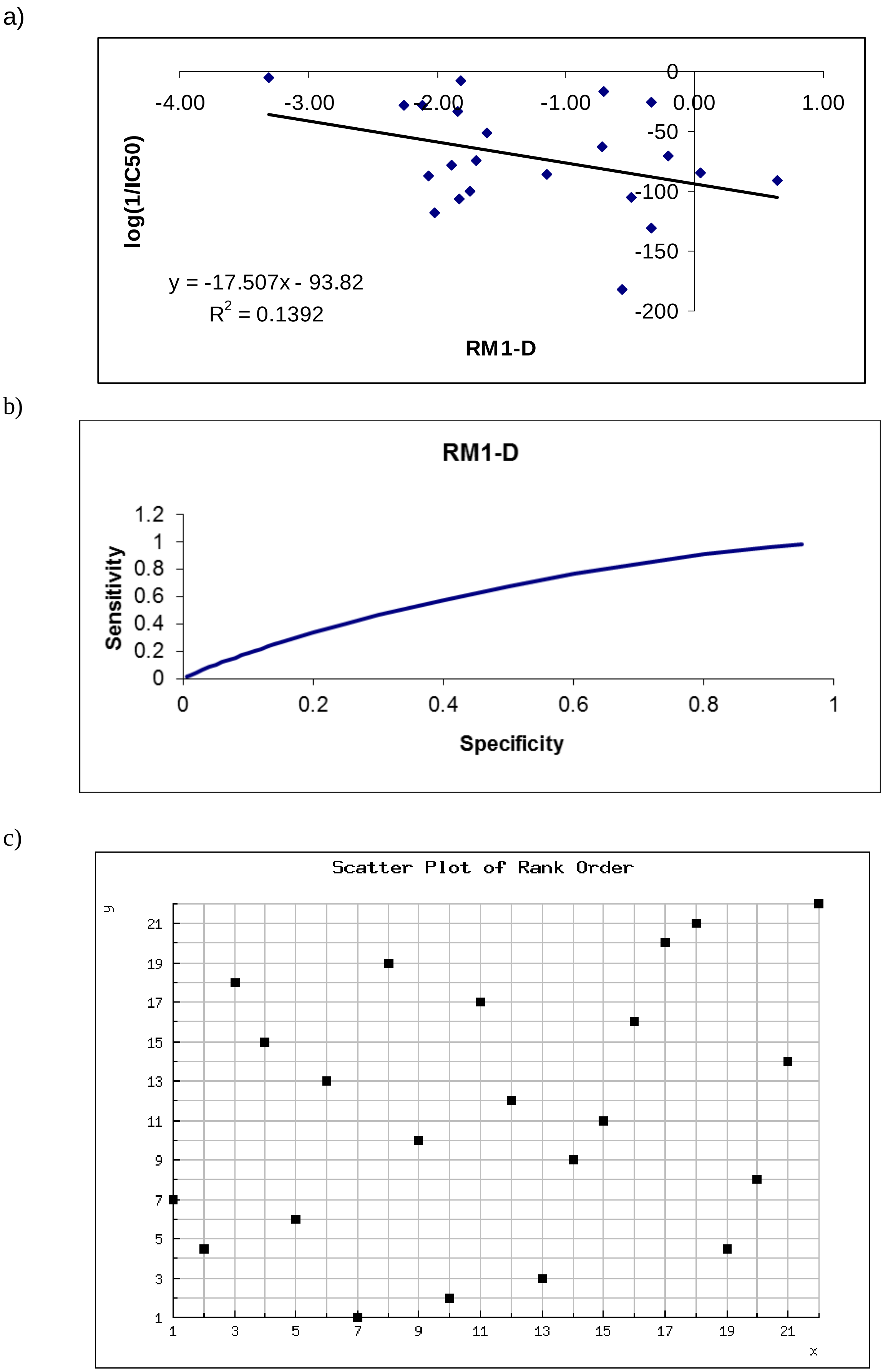


Figure 4.13 a) Linear, b) ROC curve and c) rank correlations from RM1-D data for Southwood data set.

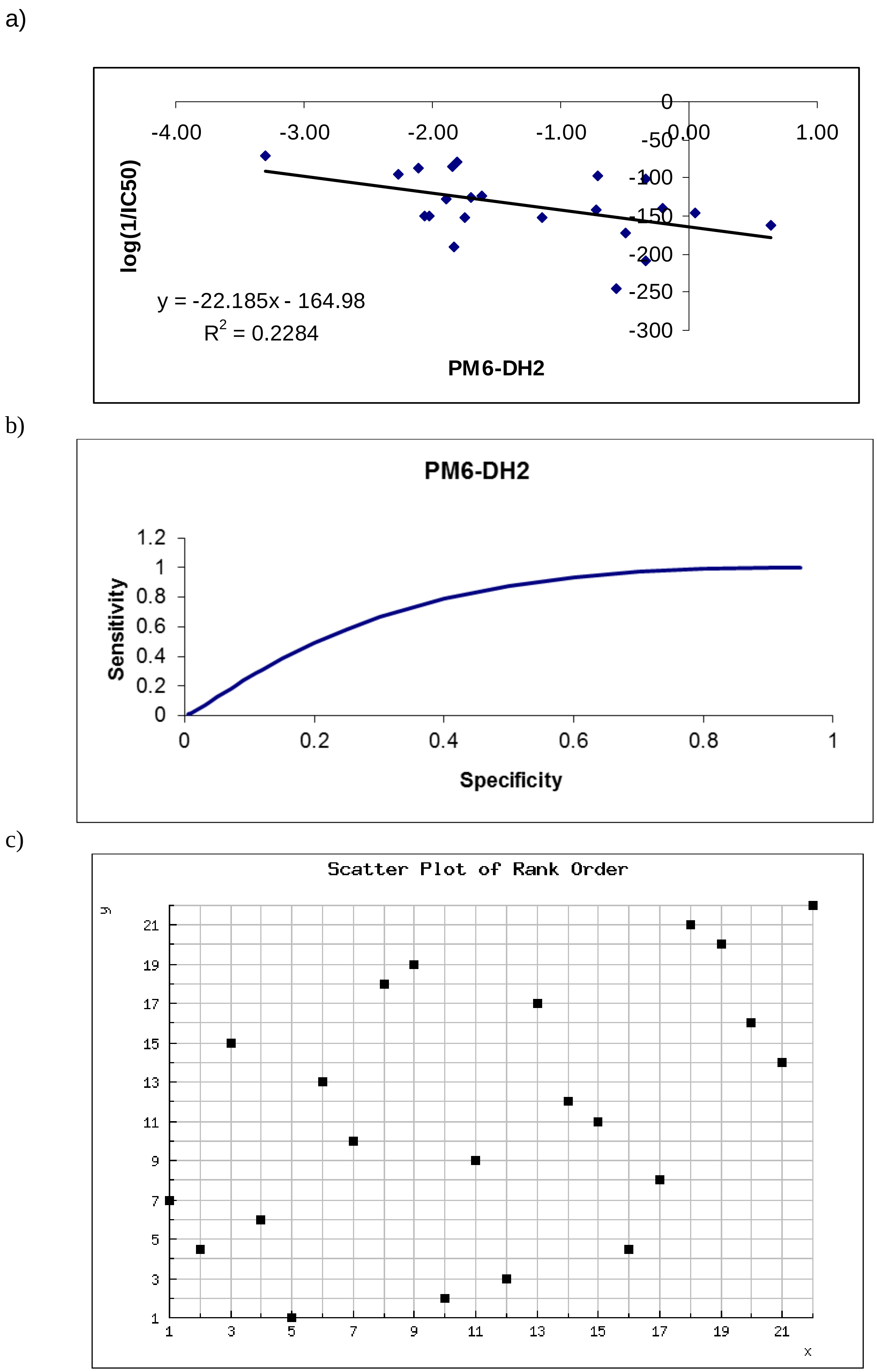


Figure 4.14 a) Linear, b) ROC curve and c) rank correlations from PM6-DH2 data for Southwood data set.

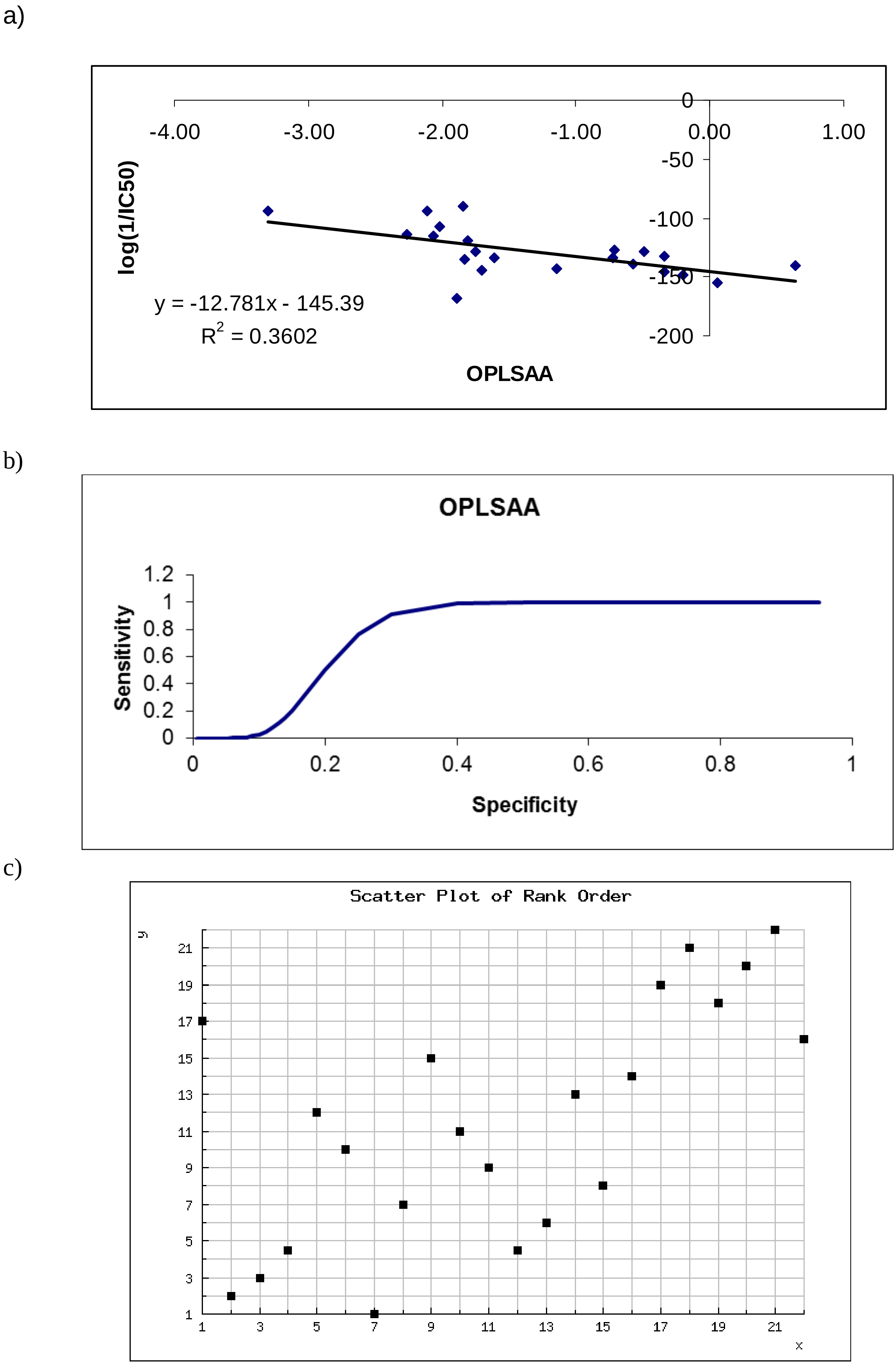


Figure 4.15 a) Linear, b) ROC curve and c) rank correlations from OPLSAA data for Southwood data set.

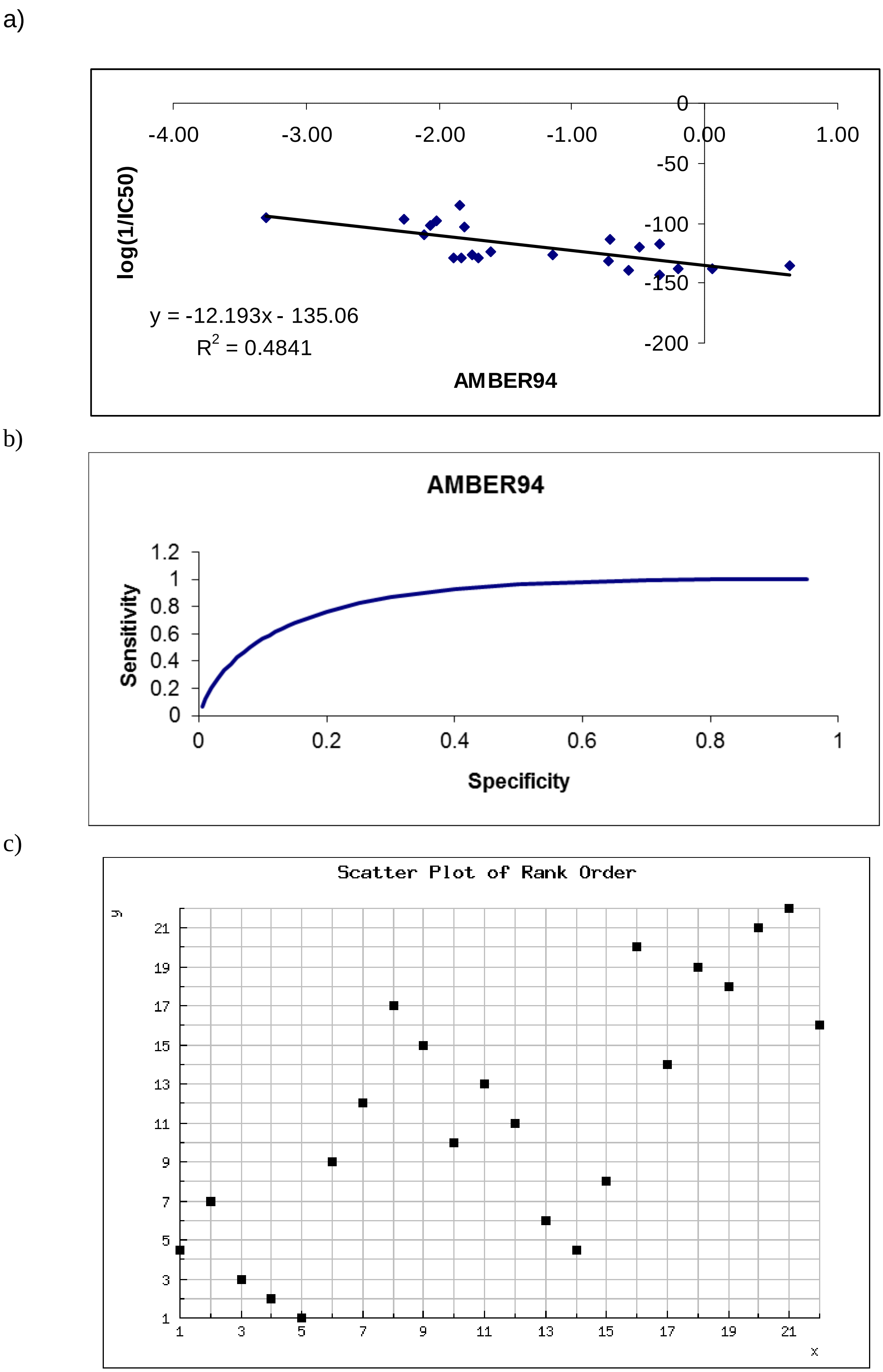


Figure 4.16 a) Linear, b) ROC curve and c) rank correlations from AMBER94 data for Southwood data set.

These structures were then used to examine the performance of the methods discussed above in predicting binding energy for this more challenging set of data (Table 4.3). As in other data sets considered above, all methods clearly identify the peptide with the highest IC_{50} value, namely 27.415, as being particularly weakly bound. Indeed, MM/GBVI predicts this peptide not to be bound at all to the receptor. Across the entire set, statistical measures show promising performance for MM/GBVI and AMBER94/Born methods, with rather worse performance for OPLS-AA/Born and PM6-DH2/COSMO, and poor results from RM1-D/COSMO. The MM/GBVI R^2 value of 0.54 is more than 99.9% significant, and corresponds to a standard error for estimate of log ($1/IC_{50}$) of 0.64 nM. The rank correlation coefficient of 0.78 indicates that this method puts almost 80% of peptides in the correct rank order. For ROC results, we used a cutoff of 50 nM to distinguish binders from non-binders, resulting in 11 peptides in each category, thereby giving a balanced test of predictions. The area under the ROC curve of 0.93 found using MM/GBVI is highly encouraging, indicating that very few false positives/negatives result from this approach. In contrast, the value of 0.62 for RM1-D is only slightly higher than random Figure (4.12) (4.13) (4.14) (4.15) (4.16).

Because of the encouraging performance of MM/GBVI, we then explored whether this method could be used to predict alignment of peptides within the receptor, rather than relying on purely sequence-based methods. To do this, numerous potential binding poses were generated with SYFPEITHI and SVMHC algorithms, and the one with the most negative MM/GBVI interaction energy selected. In 20 of the 22 cases, this agreed with the results from sequence-based prediction methods, but for two peptides (nos. 1136.14 and 1136.16) a lower energy alternative was found from this analysis. For the 1136.14, MM/GBVI predicts Thr as the anchor residue instead of Leu, and for 1136.16, the MM/GBVI predicts Gln as the anchor residue instead of Phe. This is illustrated in

Figure (4.17) for 1136.16: as might be expected, sequence-based predictions place Phe in the hydrophobic environment of pocket 1. However, this leads to placement of Ala into pocket 4, Pro in pocket 6 and Thr in pocket 9, none of which are particularly favourable for binding. With Gln as the residue in pocket 1, a hydrogen bond can form to the side-chain carbonyl (Figure (4.17), bottom left). In addition, this alignment places Leu in pocket 4, Ala in pocket 6 and Val in Pocket 9, all of which contribute to favourable binding. Comparison of Tables (4.4) and (4.5) shows that the second alignment has almost 10 kcal/mol greater binding energy, despite the apparent anomaly of a having relatively polar residue in the hydrophobic pocket 1.

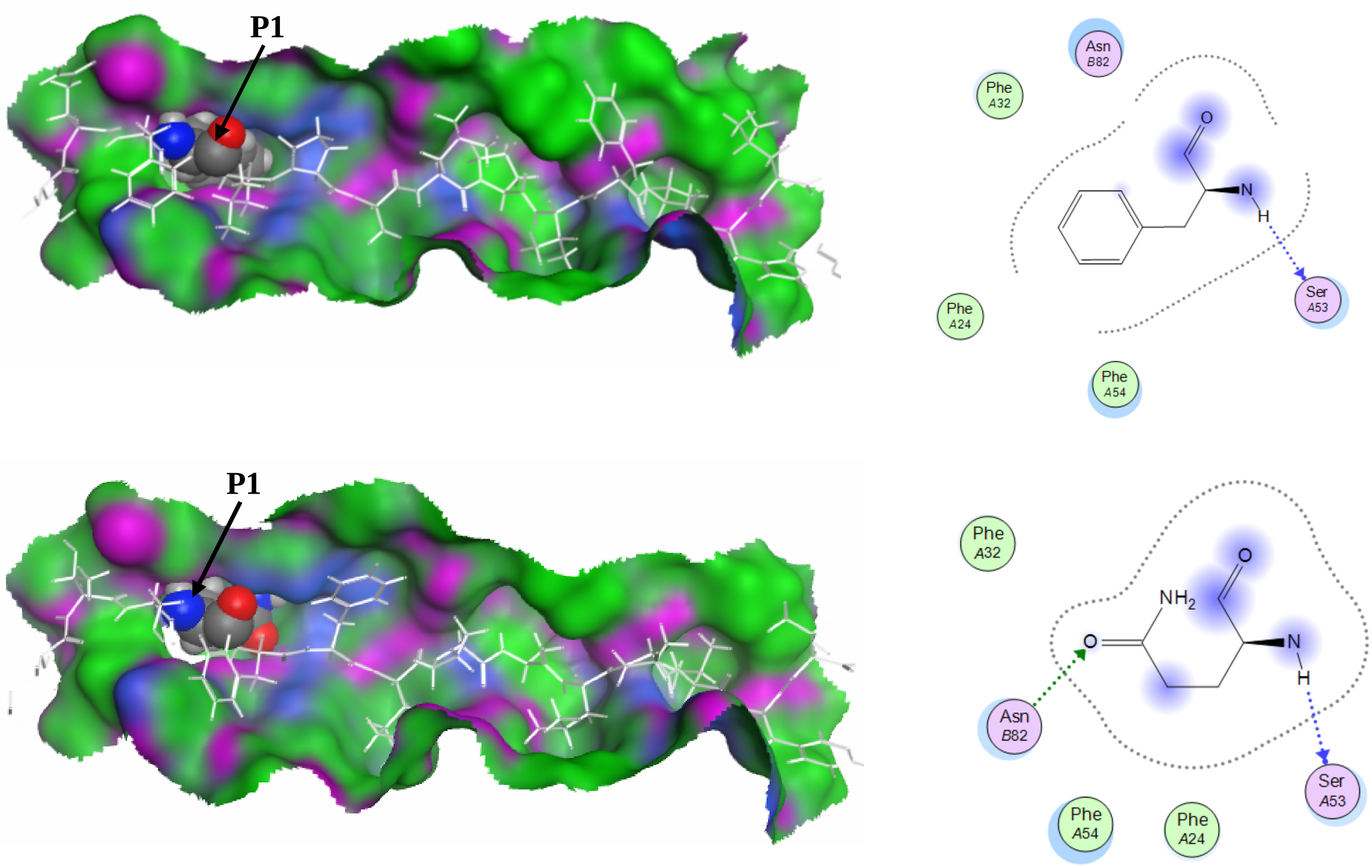


Figure 4.17 3D and 2D ligand interaction views of two possible alignments of peptide 1136.16 in HLA-DR1. Top: Phe in pocket 1; Bottom: Gln in pocket 1. On the left, the MHC receptor is shown as a continuous surface, the residue in pocket 1 as space-filling CPK spheres, and the remainder of the peptide as white wireframe. On the right, blue-shading of the peptide residue indicates exposed atoms.

Using the new values for these two peptides improves all statistical tests slightly, as shown in Table (4.5). MM/GBVI data shows small increases in R^2 and rank correlation coefficient, with plots corresponding to these data shown in Figure (4.18). The area under the ROC curve increases from 0.93 to 0.96, again illustrated in Figure (4.18). The statistics from other methods are barely affected by this change. Thus, we conclude that MM/GBVI interaction energies are a useful addition to sequence-only methods of prediction of peptide-MHC-II binding alignments.

Table 4.5: IC_{50} values (nM) and interaction energies (kcal/mol) for Southwood data set from MM/GBVI alignment along with R^2, rank correlation, and ROC area for each method.[a]

Peptide No.	Sequence*	IC50	MM/ GBVI	RM1-D/ COSMO	OPLS-AA/ Born	AMBER94/ Born	PM6-DH2/ COSMO
1188.34	HN**W**VNHAVPLAMKLI	14	-40.62	-85.53	-143.45	-126.4	-152.33
1188.16	KSK**Y**KLATSVLAGLL	3.7	-49.04	-182.02	-138.83	-139.35	-246.12
1136.47	THHY**F**VDLIGGAMLSL	2.2	-57.48	-26.04	-145.89	-143.04	-101.91
1188.32	GLA**Y**KFVVPGAATPY	3.1	-42.75	-105.6	-129.11	-120.34	-172.91
1136.16	LTS**Q**FFLPALPVFTWL	**1.6**	**-62.43**	**-68.14**	**-146.77**	**-131.54**	**-143.94**
27.415	NVKYLV**I**VFLIFFDL	2011	17.46	-4.54	-93.76	-95.87	-71.56
27.403	**L**VNLLIFHINGKIIK	78	-13.19	-77.91	-168.31	-128.5	-127.79
1136.21	IPQE**W**KPAITVKVLPA	2.2	-36.32	-130.9	-132.11	-117.41	-209.19
1136.28	LAAII**F**LFGPPTALRS	0.23	-53.05	-90.94	-140.97	-135.16	-161.75
1136.11	VVFPAS**F**FIKLPIILA	0.89	-59.32	-84.21	-155.07	-137.81	-146.19
1136.14	FA**T**CFLIPLTSQFFLP	**5.3**	**-64.73**	**-66.8**	**-140.92**	**-130.17**	**-136.09**
1188.13	AGL**L**GNVSTVLLGGV	116	-28.96	-86.62	-115.74	-102.5	-149.38
1136.24	NLSNV**L**ATITTGVLDI	182	-25.61	-27.96	-113.74	-96.22	-96.12
1136.12	IKLP**I**ILAFATCFLIP	105	40.92	-118.3	-107.73	-98.42	-150.80
27.392	SSV**F**NVVNSSIGLIM	41	-38.79	-51.92	-133.96	-123.7	-123.90
27.417	**V**KNVIGPFMKAVCVE	56	-53.73	-100.38	-128.45	-126.36	-152.69
1136.55	QEIDPLS**Y**NYIPVNSN	65	-11.14	-7.5	-119.45	-102.8	-78.95
1136.71	EPQGST**Y**AASSATSVD	5.1	-58.73	-16.2	-127.66	-113	-96.59
1136.38	**S**SIIFGAFPSLHSGCC	70	-8.49	-33.79	-90.11	-85.43	-85.15
27.388	**M**RKLAILSVSSFLFV	50	-13.22	-73.82	-143.71	-128.8	-125.30
1136.59.01a	**R**VYQEPQVSPPQRAET	130	29.36	-28.23	-94.42	-110.26	-86.59
1136.46	LWWST**M**YLTHHYFVDL	68	-9.91	-106.31	-135.71	-128.45	-190.35

R^2	0.56	0.14	0.36	0.47	0.23
Rank correlation	0.79	0.29	0.66	0.74	0.48
ROC area	0.96	0.62	0.8	0.87	0.75

[a] Alignments that differ from Table 3 shown in bold.

* The underlined residues are the residues which we included in our calculations and the residues on bold red are the residue which located on pocket one of the receptor.

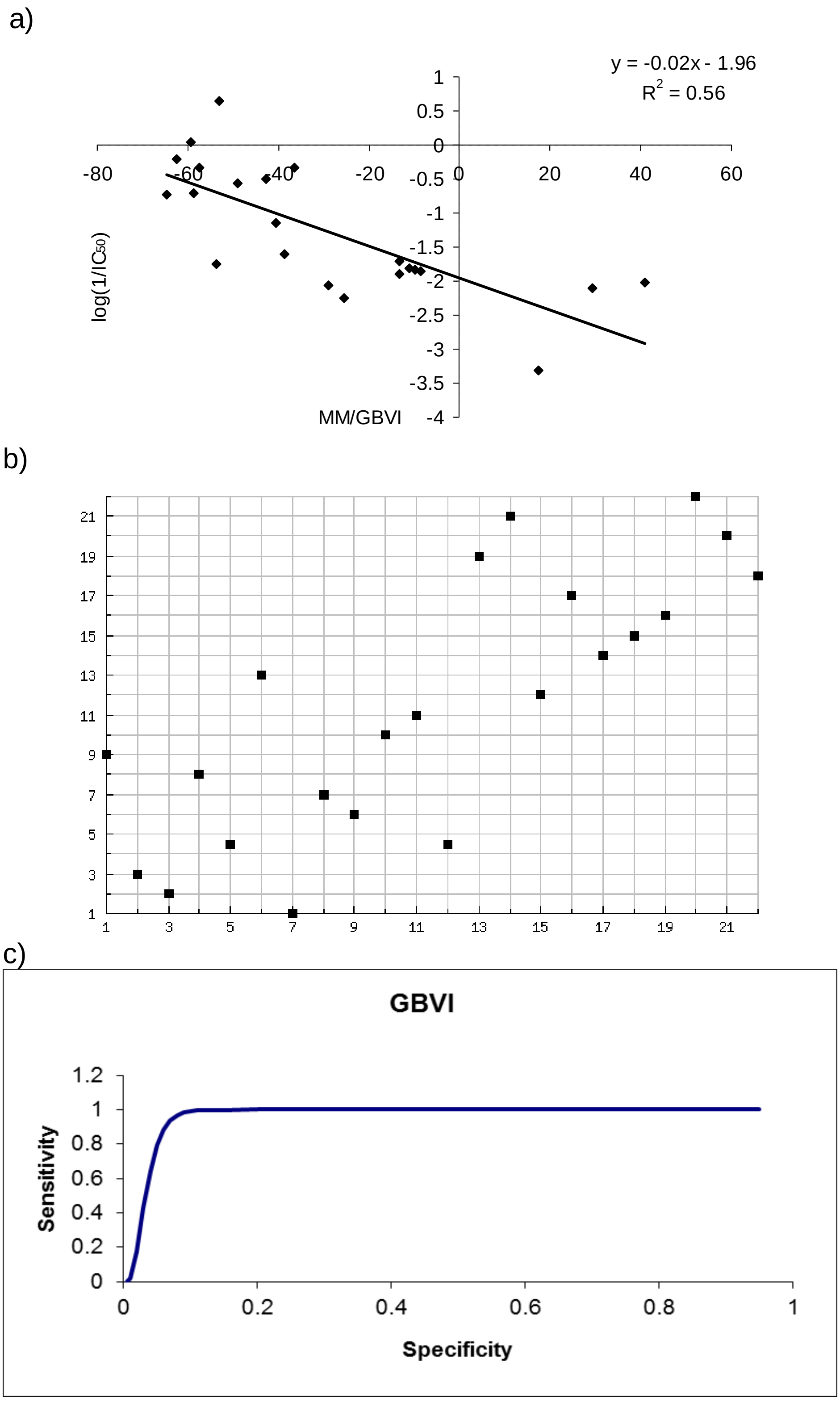


Figure 4.18 a) Linear correlation, b) rank correlation and c) ROC curve from MM/GBVI data for Southwood data set.

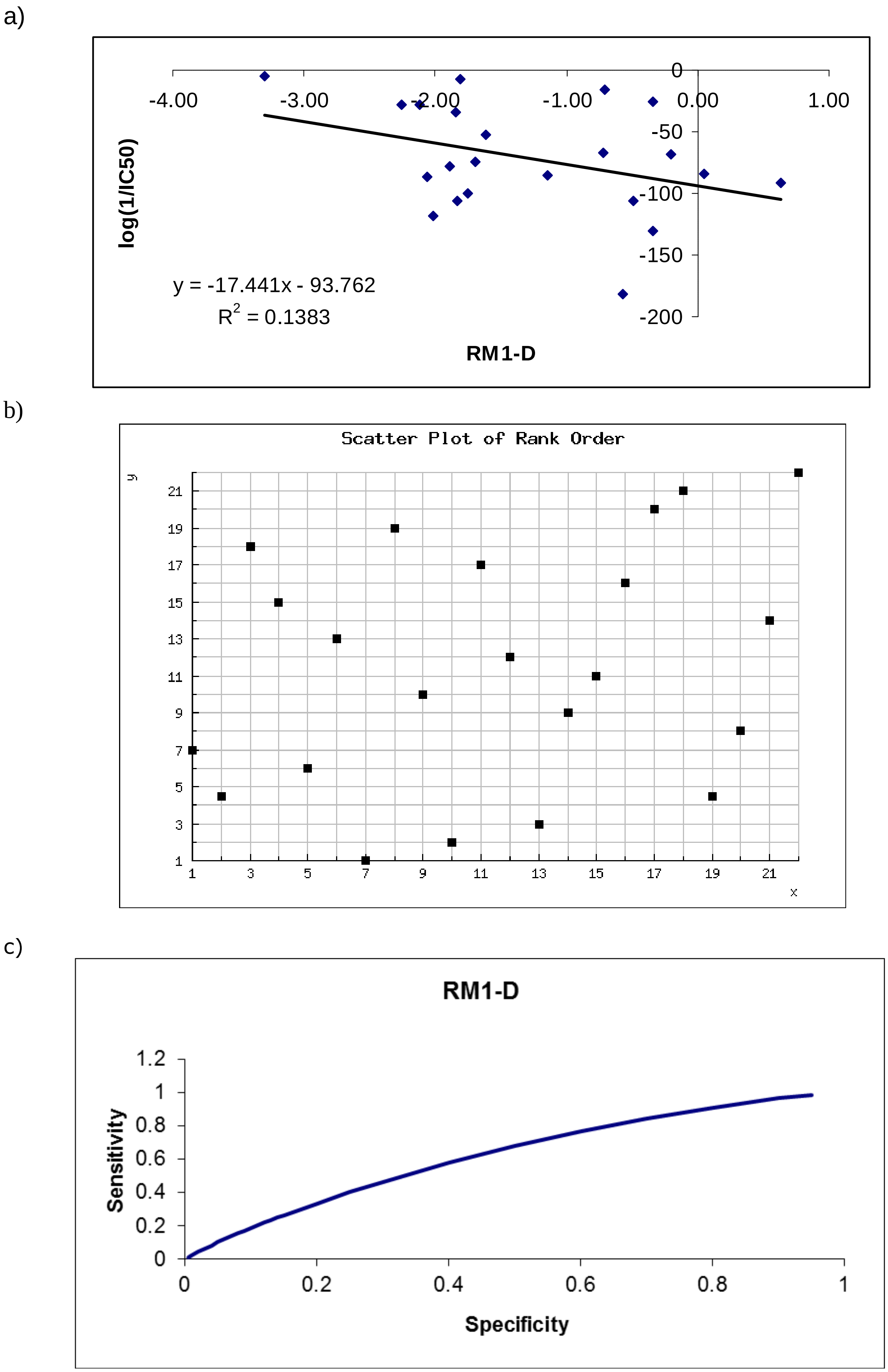


Figure 4.19 a) Linear correlation, b) rank correlation and c) ROC curve from RM1-D data for Southwood data set.

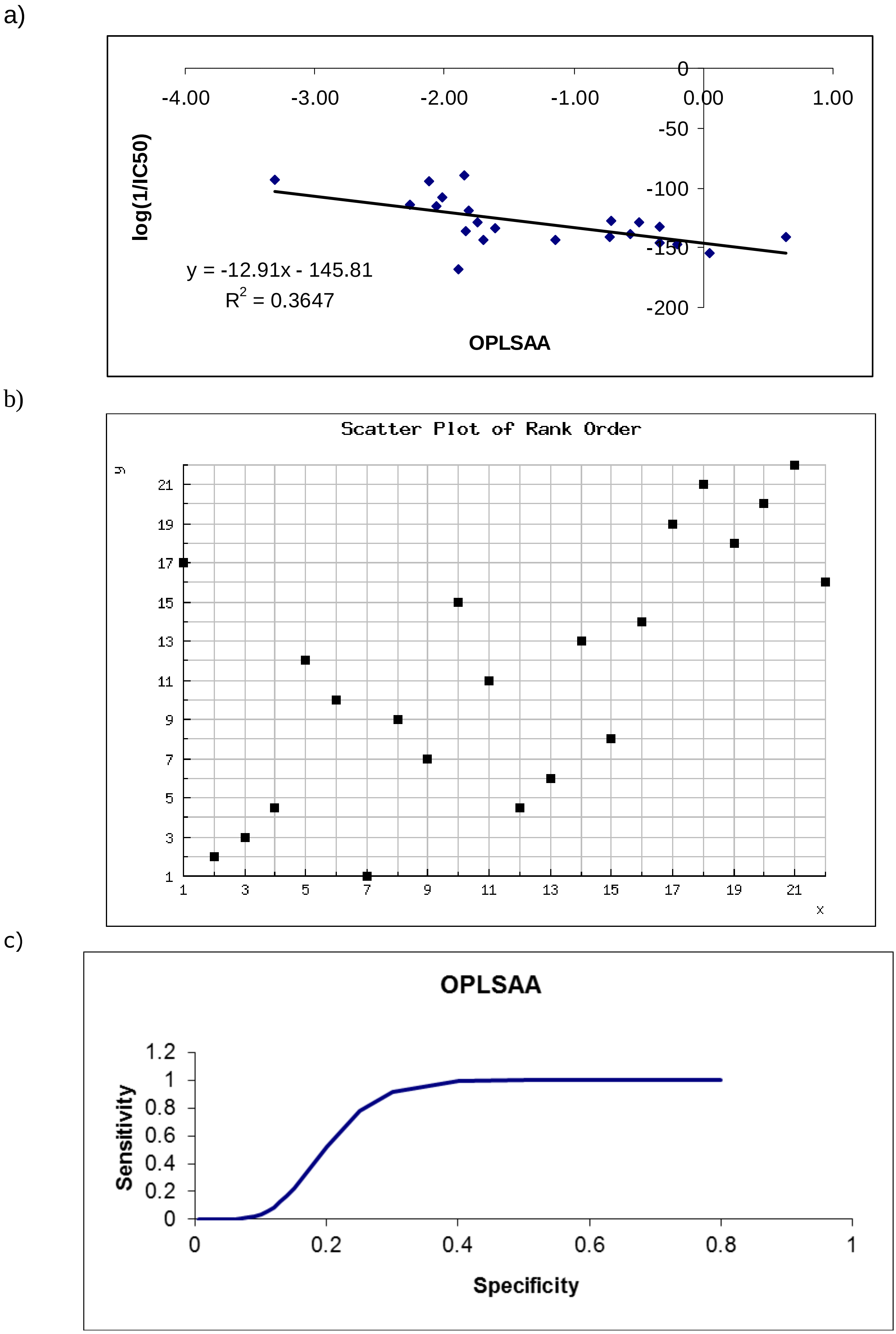


Figure 4.20 a) Linear correlation, b) rank correlation and c) ROC curve from OPLSAA data for Southwood data set.

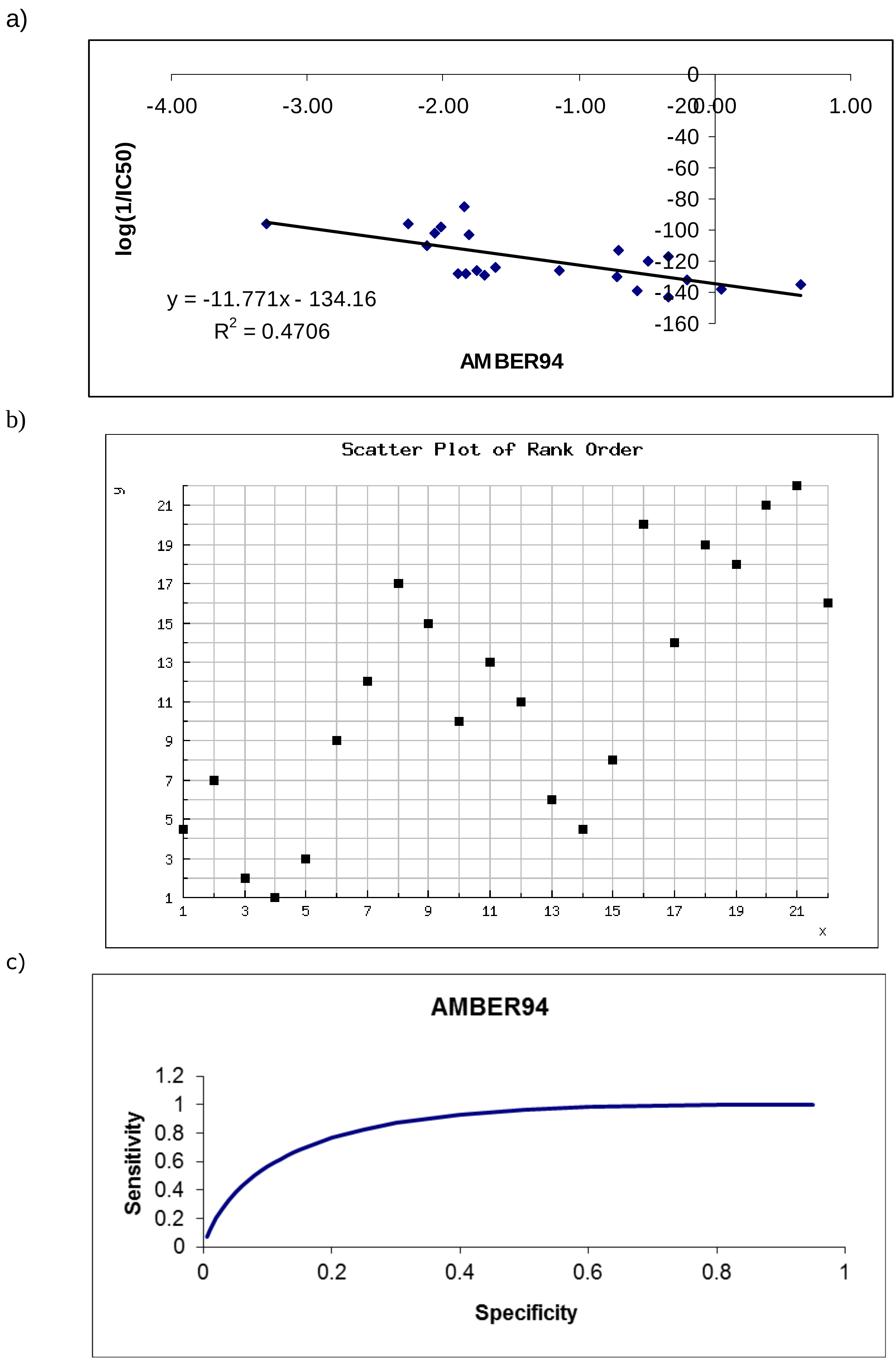


Figure 4.21 a) Linear correlation, b) rank correlation and c) ROC curve from AMBER94 data for Southwood data set.

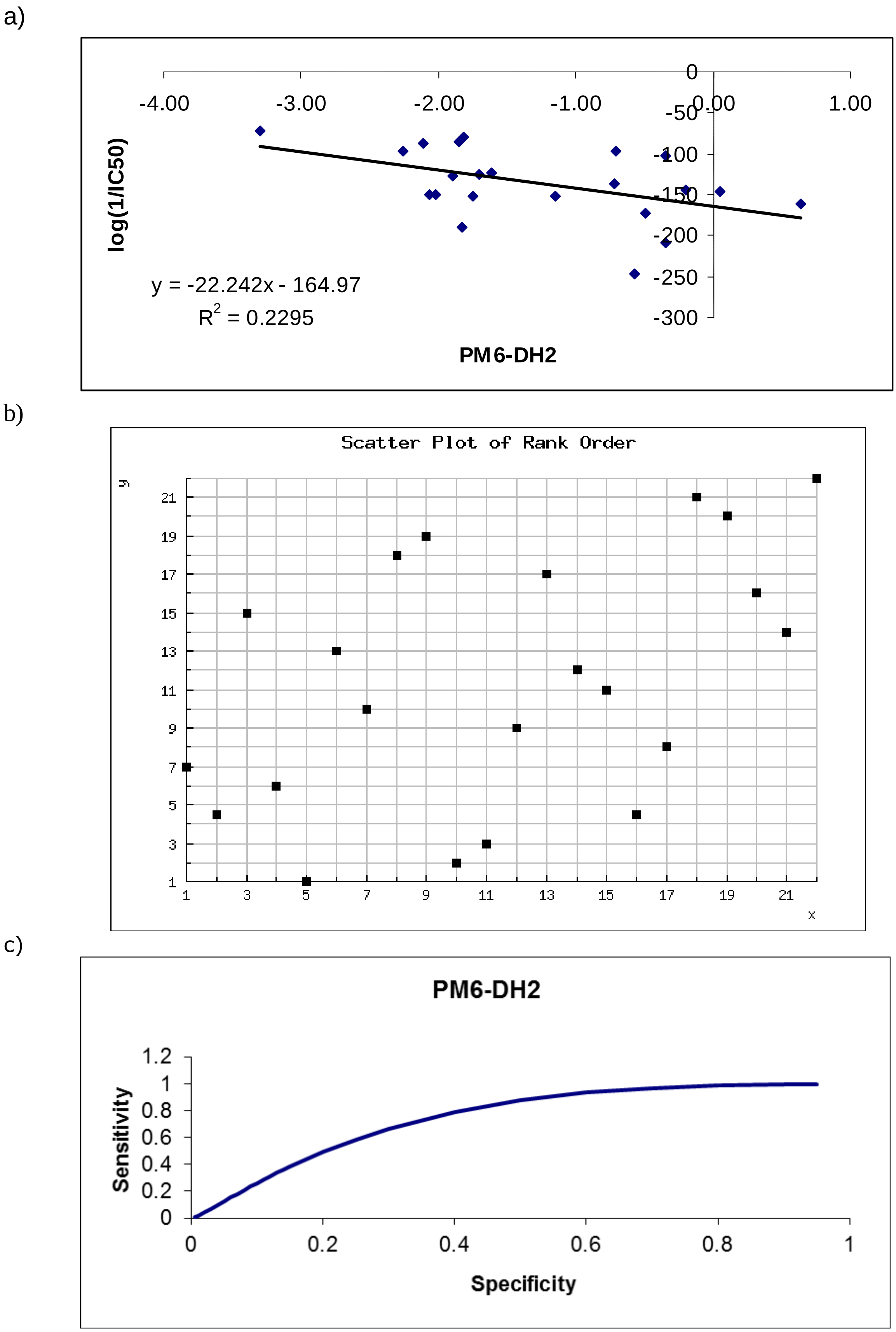


Figure 4.22 a) Linear correlation, b) rank correlation and c) ROC curve from PM6-DH2 data for Southwood data set.

4.5 Conclusion:

We have tested several methods to calculate the interaction energy for peptide-MHC-II complexes for three separate data sets, using IC_{50} data to evaluate the accuracy of each theoretical method. We show that MM/GBVI approach is a promising way to calculate the free energy for peptide-receptor complexes, with reliable performance for all three data sets as measured by three distinct statistical tests. For two data sets where peptides are closely related, HEL and MBP, excellent performance is evident from these statistics, with strongly significant correlation between interaction energy and $\log(1/IC_{50})$ good or perfect ranking of activity, and no false negatives/positives. AMBER94 with a Born model of solvation performs almost as well, while OPLS-AA/Born and RM1-D/COSMO give rather worse performance. MM/GBVI also performs well for the more diverse set of peptides contained in the Southwood data set despite the lack of entropy contributions to these calculations, apparently confirming that such contributions are not required for this set in evaluation of relative binding free energies even for ligands as flexible as peptides.

We also show that this method can be used to predict the anchor residues that reside in receptor binding pockets, and that this approach gives slight improvement in statistics over purely sequence-based prediction methods such as SYFPEITHI or SVMHC. The accuracy of the MM/GBVI approach may stem from the fact that the dielectric constant employed is estimated from the atoms present in the specific complex under study, rather than on an idealised value, or from the use of constrained optimisation that allows ligand and some receptor flexibility while keeping the overall binding mode fixed. Of course, both peptide ligand and protein receptor are flexible objects, such that the single snapshots used here can only be approximations of the entire binding event. We explored the use of molecular dynamics to calculate MM-GB/SA averaged over multiple snapshots, and reported the results in Chapter 5. For now, we have shown that the MM/GBVI approach can deliver reasonable predictions of peptide-MHC binding in a matter of a few seconds on a desktop computer.

4.6 References:

[1] Mantzourani, E. D.; Mavromoustakos, T. M.; Platts, J. A.; Matsoukas, J. M.; Tselios, T. V. *Current Medicinal Chemistry* **2005**, *12*, 1521-1535

[2] Wearsch, P. A.; Cresswell, P. *Current Opinion in Cell Biology* **2008**, *20*, 624-631

[3] Klamt, A. *Journal of Physical chemistry*, **1994**, *99*, 2224-2235.

[4] Harrison, L. C.; Honeyman, M. C.; Termbleau, S.; Gregori, S.; Gallazzi, F.; Augstein, P.; Brusic, V.; Hammer, J.; Adorini, L. *Journal of Experimental Medicine*, **1997**, *185*, 1013-1021.

[5] Cheng, Y.; Prusoff, W. *Biochem Pharmacol* **1973**, *22*, 3099-3108.

[6] In www.bdbiosciences.com.

[7] Tajkhorshid, E. J., K. J; Suhai, S. *Journal of Physical chemistry B*, **1998**, *102*, 5899-5913.

[8] Klamt, A.; Schuurmann, G. *Journal of Chemical Society Perkin Transactions*, **1993**, *2*, 799-805.

[9] Krogsgaard, M.; Wucherpfenning, K. W.; Canella, B.; Hansen, B. E.; Svejgraad, A.; Pyrdol, J.; Ditzel, H.; Raine, C.; Engberg, J.; Fugger, L. *Journal of Experimental Medicine*, **2000**, *191*, 1395-1412.

[10] Aldulaijan, S.; Platts, J. A. *Journal of Molecular Graphics and Modelling* **2010**, *29*, 240-245

[11] Southwood, S.; Sidney, J.; Kondo, A.; Guercio, M.-F. d.; Appella, E.; Hoffman, S.; Kubo, R. T.; Chesnut, R. W.; Gery, H. M.; Sette, A. *The Journal of Immunology*, **1998**, *160*, 3363-3373.

[12] Murthy, V. L.; Stern, L. J. *Structure*, **1997**, *5*, 1385-1397.

[13] In *http:/alkaid001.atspace.com*.

[14] Jalkanen, K. J.; Elstner, M.; Suhai, S. *Journal of Molecular Structure: THEOCHEM*, **2004**, *675*, 61-77.

[15] Rommensee, H.-G.; Bachmann, J.; Emmerich, N. P.; Bachor, O. A.; Stevanovic, S. *Immunogenetics.* , **1999**, *50*, 213-219.

[16] Donnes, P.; Kohlbacher, O. *Nucleic acids research*, **2006**, *34*, web server issue.

[17] Donnes, P.; Elofsson, A. *bioinformatics*, **2002**, *3*.

[18] In *http://www.chemcomp.com/*. **2011**

[19] Labute, P. *Journal of computational chemistry*, **2008**, *29*, 1693-1698.

[20] MOE. *http://www.chemcomp.com/*. **2011**

[21] Rezac, J.; Fanfrlik, J.; Salahub, D.; Hobza, P. *Journal of Chemical theory and computation*, **2009**, *5*, 1749-1760.

[22] Korth, M.; Pitonak, M.; Rezac, J.; Hobza, P. *Journal of chemical theory and computation*, **2010**, *6*, 344-352.

[23] Stewart, J. J. P. *available at http:/OpenMOPAC.net*, **2009**.

[24] In *http://www.wessa.net/rankcorr.wasp*.**2011**

[25] Dorfman, D. D.; Alf, E. *Journal of Mathematical Psychology*, **1969**, *6*, 487-496.

[26] Metz, C. E.; Herman, B. A.; Shen, J.-H. *Stat. Med. Journals*, **1998**, *17*, 1033-1053.

[27] Tsai, C. S. *An introduction to computational biochemistry*. **2002**.

Chapter 5:

Molecular dynamics simulations:

5.1 Introduction:

Proteins are very flexible molecules for which the conformation has an important influence in their functions [1]: therefore using a single x-ray structure (rigid structure) to study and understand their properties such as binding free energy might not be enough. Molecular dynamics simulations are usually useful tools in such cases. Molecular dynamics simulation of biological molecules can be defined as "the science of simulating the motions of a system of particles applied to biological macromolecules", and "gives the fluctuations in the relative positions of the atoms in a protein or in DNA as a function of time"[2]. The first molecular dynamic simulation has been done 35 years ago for bovine pancreatic trypsin inhibitor (BPTI) which proved that proteins are not rigid structures [1, 3]. After that, many investigations were carried out on protein phenomena such as the effect of the solvent [4]. The contribution of theoretical chemistry on biology has been improved after introducing molecular dynamic simulation with supercomputers[2]. By increasing the computer power, the number of molecular dynamic studies on biological macromolecules increased [1, 5].

AMBER suite of programs was used to investigate the influence of the motion on implicit and explicit solvents for the Southwood set. This set, as we mentioned on the previous chapter, contains 22 peptides which interact with MHC class II as a receptor [6]. The binding free energies were calculated by molecular mechanics- generalized -Born surface area (MM/GBSA) method using MM/PB(GB)SA script on AMBER for both implicit and explicit solvent simulations.

5.1.1 Implicit and explicit solvation models:

Solvent can have an important influence on the structure and properties of biomolecules, so in many cases one must consider the effect of solvent in calculations. There are two main methods to predict the solvation effect on biomolecular properties [7-9], namely implicit (or continuum) and explicit solvation models. Explicit solvent models include the solvent molecule explicitly in the calculation, while the implicit solvent models replace the solvent molecules with a dielectric continuum or similar medium [10, 11]. Each method has its advantages and disadvantages: while explicit solvent models consider the solvent effect in the highest levels of detail, they can be expensive and time consuming. On the other hand, implicit solvation models are able to ‘‘pre-average’’ solvent effect and therefore reduce the need for computationally expensive sampling, making these model widely used in studying biomolecules [12].

5.2 Methodology:

The calculations are divided in to three stages; preparation, simulation and analyzing the output files.

5.2.1 Preparation:

The first stage is the preparation which was done by using *LEaP* program which is provided by AMBER. This program is designed to read atoms coordination, topology and force field which are required to run the simulations. In this study ff99SB force field was used, which is recommended by AMBER [13]. *xleap* which is the graphic version of *leap* that allows seeing two-dimensional structure of the molecule under the study was used. We used the PDB files which were created by MOE program for the previous calculations. To load the PDB files, we used “*loadpdb*” command, and “*edit*” command was used to view the structure. The structure was then neutralized using AMBER *xleap* by adding sodium ions (Na+) at positions of high negative electric potential around our molecule using “*addions Na+0*” command [14]. For the explicit solvent calculations, we used truncated octahedral box of water by using TIP3P model [15] with 8 angstroms as buffer dimension. This was done by “*solvateoct TIP3PBOX 8.0*” command. Two types of text files were produced from this stage “*prmtop”* and “*inpcrd”* files. “*Prmtop”* file includes the topology and the parameters of force field. Where “*inpcrd”* is the atoms coordinates file. These two files were used on the next stage of the calculations.

5.2.2 Minimization and molecular dynamic simulation:

The second stage of this experiment was carried out by *sander* which is one of the AMBER programs. Sander “stands for **S**imulated **A**nnealing with **N**MR-**D**erived **E**nergy **R**estraints, but this module is used for a variety of simulations that have nothing to do with NMR refinement”[16]. Table A on the appendix shows the keywords, the descriptions and values which have been used in this study. In order to make the calculations faster [5], bond lengths to hydrogen may be constrained. To do so, SHAKE method [17], which is suitable for macromolecules, [5] is highly recommended for most MD calculations [16]. This method integrates “the Cartesian equations of motion of flexible molecules” depending on atomic coordinates only and not on time [5]. SHAKE method also known as an iterative method, “since it treats the constraints in an iterative way”[5]. The length of the time step Δt in MD simulation depends on the highest frequency motions, which is the bond stretching, happening in the system under the study [5, 16]. SHAKE deletes the bond stretching freedom of bonds to hydrogen, and as a result, a larger timestep can be used [16]. We used NTC= 2, NTC indicates the SHAKE option (see appendix table A), 2 for constrain the bonds involving hydrogen [16]. Also we used a Langevin dynamics approach to control the temperature. Sander stage divides into few steps; minimization, heat, equilibrium, and molecular dynamics simulations. All input files are located on the appendix.

5.2.2.1 Molecular dynamics simulation for implicit solvent (1ns and 6ns):

The minimization step is very important to run stable molecular dynamic simulations. In this step, the structure of the complex was minimized (reached the closest minimum) in order to remove any stress or overlapping of the atoms on the starting structure. The input file is specified as imin = 1 to run the minimization, (maxcyc =1000) is the maximum number of cycles of minimization with (ncyc =500) the first 500 being steepest descent, (ntb =0) not periodic simulation, (cut=12) we used a cutoff of 12 Å, and (igb=1) is to use generalized Born solvation model. See the minimization input file (min.in) on the appendix.

After the minimization, we moved to heating step. In this step imin=0 was used because there is no minimization, igb=1 generalized Born solvation model was used, ntx=1 was used to read the initial coordinates and velocities from *inpcrd* file, ig=-1 for random number seed, ntt=3 the Langevin thermostat was used, and temp0=300 the temperature which the system was kept at. The description of all keywords (table A) and the heat input file are on the appendix.

After this step we run a short molecular dynamic simulation (100ps) to test our complex before running long simulation. imin=0 was used because there is no minimization, igb=1 generalized Born solvation model was used, ntx=1 was used to read the initial coordinates and velocities from *inpcrd* file, ig=-1 for random number seed, ntt=3 the Langevin thermostat was used, and temp0=300 the temperature at which the system was kept.The input file for the short simulations (md1.in) and the description of these keywords are on the appendix. In this step we did not use the SHAKE method because we just run the calculations for short time. After completion of this step, production molecular dynamic simulations for subsequent analysis were run for longer time (1ns); the relevant input file (md2.in) is in the appendix. We used shake option (ntc=2, ntf=2) , turned the minimization off, used 12 Å cutoff, used Born implicit solvent, wrote the coordinates on file every 200 steps, and for temperature Langevin thermostat was used to keep the temperature at 300 K. The molecular dynamic simulation was run for a total of 500000 steps with a 2fs time step.

In order to investigate the stability of the results based on the length of the simulation, some of the implicit solvent simulations were run up to 6 ns. Because of the cost for such calculations, this was done for just for three peptides (1136.11, 27.415, and 27.417) as an additional five molecular dynamic simulations each 1ns, and for longer simulations for two peptides 27.415 and 27.417. The molecular dynamics simulations were continued, starting from the last structure file from the previous simulation. After finishing all the molecular simulations for this step we moved to the analyzing stage.

5.2.2.2 Molecular dynamic for explicit solvent simulation with periodic boundaries (1ns):

It has been mentioned before that for explicit solvent, a truncated octahedral box of water TIP3P model was used. At this stage, the water has not been affected by the solute, which may lead to unstable simulation. Therefore it is important to run a minimization before heating the system slowly to 300 K, to relax the water box during a molecular dynamic equilibrium stage before running the molecular dynamic simulations[14].

This section (explicit solvent) is more complex than the previous section and it takes longer time. The explicit solvent minimization step was divided into two steps; first the protein-ligand complex was held fixed, and allows just the water molecules and ions to minimize (min1.in), and second, the whole system was minimized (min2.in). In order to hold the complex fixed, position restraint (ntr=1) was used based on the “GROUP input” which is classified in the minimization input file (min1.in). A force constant of 500 kcal mol^{-1} angstrom^{-2} and restrain residues 1 through 379 (as an example) are used. Then the second step (min2.in) started using the previous restrt file. After minimizing the system, the simulation was moved to the equilibration stage starting with heating the system from 0K to 300K by using Langevin temperature equilibration method (NTT=3) which is performing very well on “maintaining and equalizing” the system temperature [14]. A weak restraint was used, as in the first step in the minimization, to ensure this heating occurs without any change in the protein-ligand complex. The minimization input files (min1.in) and (min2.in) and the heating input file (heat.in) are in the appendix.

After the heating step, the simulation was moved to the equilibration step using constant pressure in order to relax the water density. After the system reached the 300 K, the restraints on our protein-ligand complex were removed and 100ps simulation time starting from (restrt) file from the heat step. The equilibration input file (equl.in) is in the appendix. After that, the molecular dynamic simulations were run for 1ns time. See the input file (md2.in) in the appendix. This step took a long time to complete, as expected. The molecular dynamic simulation for 1ns in implicit solvent takes on average 24 hours while in explicit solvent it takes on average 72 hours. So, explicit solvent simulation requires three times more than implicit solvent simulation for each 1ns.

5.2.3 Analyzing the results stage:

At this stage, three methods were used to analyze the results. First, (*process_mdout.perl*) script from AMBER [14] was used to extract the energies, and other properties from the md.out files. This script creates a numbers of files that contain the properties of the complexes such as energies and temperatures *vs.* time. Second, MM/GBSA method was used to calculate the binding free energies of these peptide-MHC complexes. By taking the average of the trajectory, many properties such as energies and temperatures of the molecules can be obtained. And because of neglecting entropy term, I focused on calculating the binding free energy using MM/GBSA method. In order to get the backbone atom root mean square deviation (RMSD) *vs*. Time, the third program (*ptraj*) was used, which is a trajectory analysis program on AMBER.

For the implicit solvent simulation for 1ns, several statistical tests were used to investigate the influence of the molecular dynamic simulations on the binding free energies of the Southwood set. The binding free energies were calculated by MM/GBSA method. And also to test the ability of this method on the prediction of peptide-MHCII bonding, using published IC_{50} values as a test. Specifically, I employed the standard Pearson R^2 value against the negative log of IC_{50} values, Spearman's rank correlation coefficient [18], and area under relative operating characteristic (ROC) curves by using the ROCkit package [19, 20]. In each case, a value of 1.0 indicates the ideal of perfect prediction.

5.3 Results and discussions:

The previous chapter concentrated on studying a single structure of each peptide. In this chapter, molecular dynamic simulations were used to investigate the influence of the motion on calculate the binding free energy of peptide-MHC complexes. In this chapter, the calculation was divided to four sections; MD simulations on implicit solvent for 1ns time, MD simulations on implicit solvent for 6ns time, MD simulations on implicit solvent for longer time and MD simulations on explicit solvent for 1ns time.

5.3.1 Implicit solvent simulations (1ns):

First calculations were done on an implicit solvent and were running for 1 ns. Figure (5.1) shows the plots of total energy, potential energy and temperature as a function of time for 1136.11 peptide, and figures A and B on the appendix show these plots for two peptides (27.415 and 27.417) from Southwood's set. The other complexes show the same behavior (figures not shown). These three peptides were chosen to cover a wide activity range: 27.415 has IC_{50}=2011 nM, and is considered as inactive; 27.417 has IC_{50}=56 nM and is considered active; and 1136 has IC_{50} = 0.89 nM is considered a very active peptide.

From the plots in Figure (5.1), it is clear that the simulations were running well and no strange behavior was accrued during the simulations. The temperature for example remained more or less constant during the simulations. Figure (5.2) shows the plots of the backbone rms deviation (RMSD) of these three peptides from initial structures as a function of time. From these plots and Table (5.1), we can see that the peptides structures have been changed during the simulations but they were close to the initial structures. After checking that the simulations were run well, I calculated the binding free energy of Southwood set peptide-MHC complexes using MM/GBSA method provided by AMBER. Table (5.1) shows the free energies of the set by MM/GBSA method.

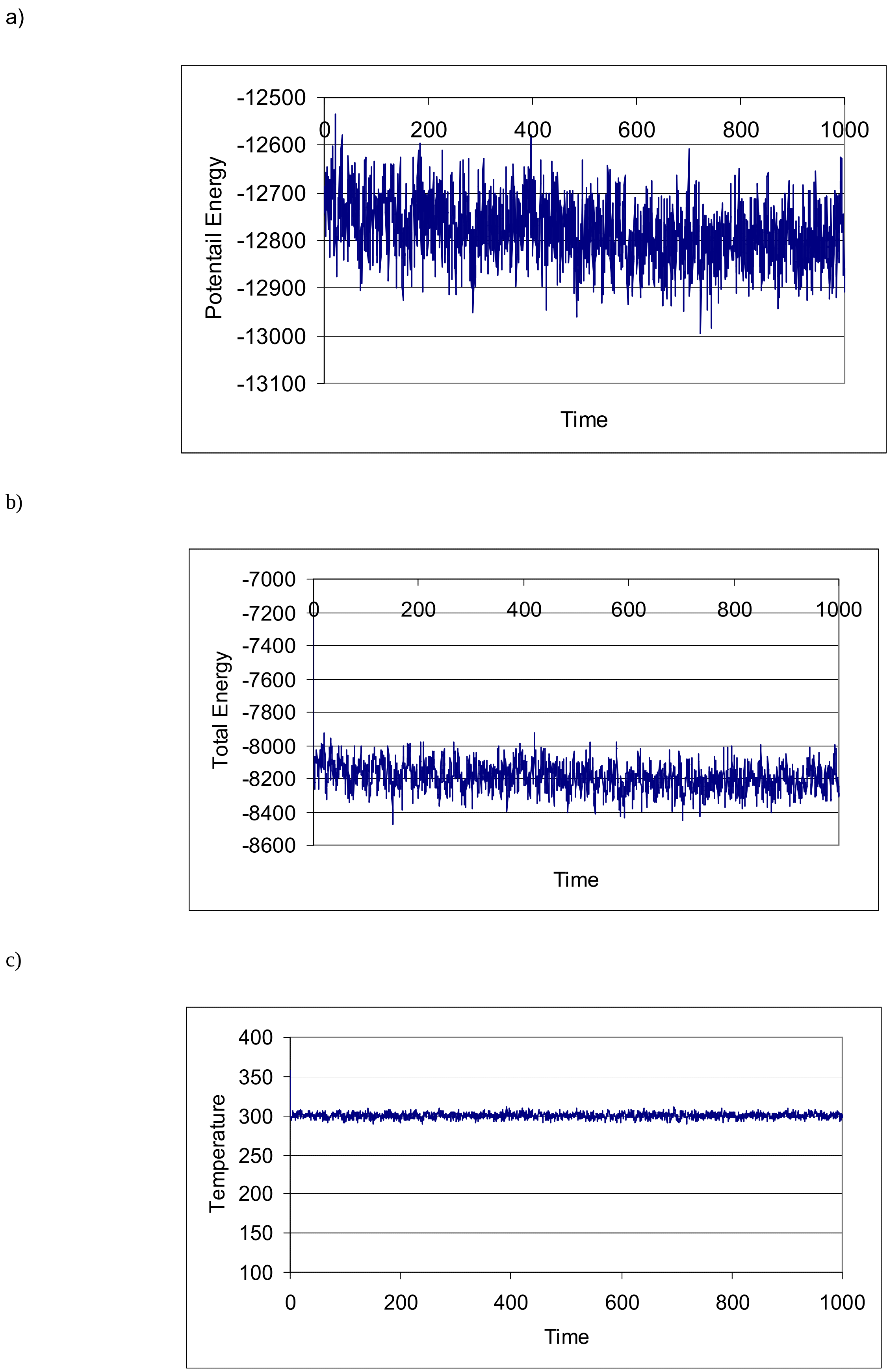
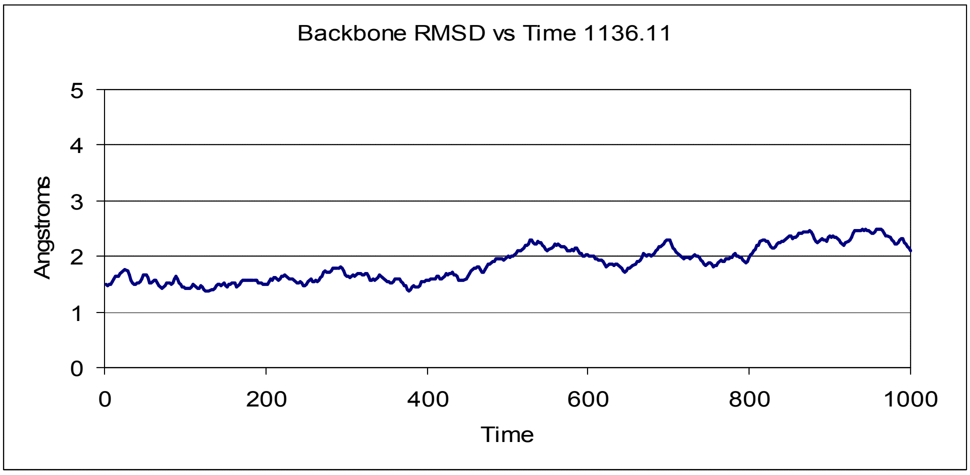
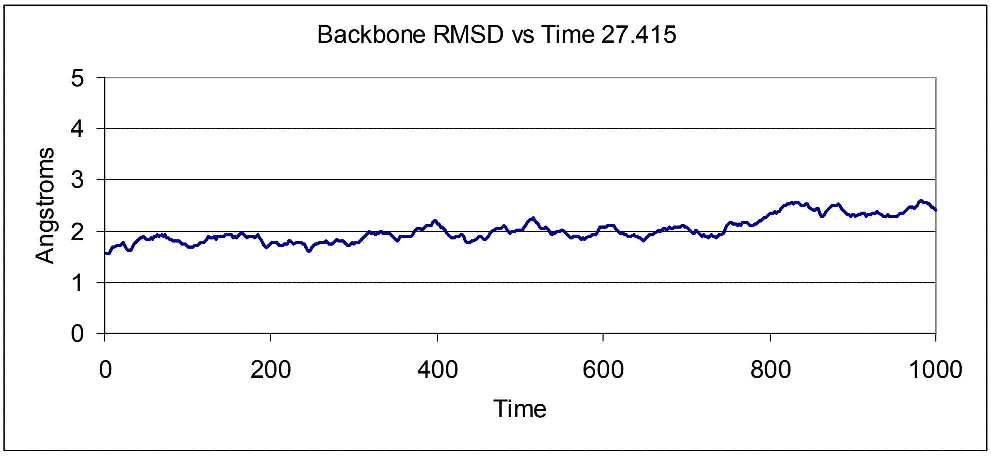


Figure 5.1: The plots of a) potential energy (kcal/mole), b) total energy (kcal/mole) and c) temperature (K) as a function of time of (1136.11) peptide.

a)



b)



c)

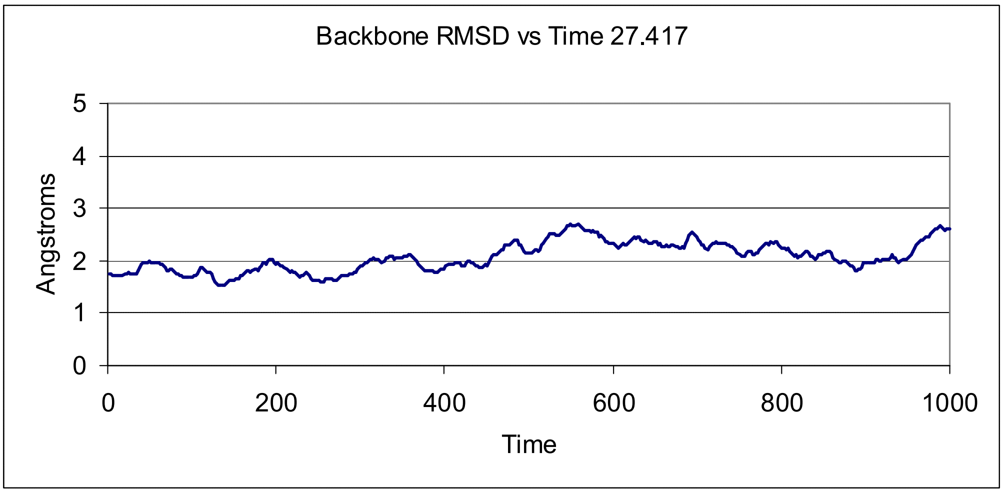


Figure 5.2: The plots of backbone RMSD of a) 1136.11, b) 27.415, and c) 27.417 peptides from initial structures as a function of time.

Table 5.1: MM/GBSA free energies (kcal/mole) and standard deviation values of Southwood set of peptide-MHC complexes after molecular dynamic simulations for 1ns.

Peptide No.	**IC50**	**MD/MM-GBSA**	**Std. Dev.**
1188.34	14	-129.89	4.95
1188.16	3.7	-145.50	7.82
1136.47	2.2	-127.22	5.32
1188.32	3.1	-122.89	6.86
1136.16	1.6	-115.68	7.13
27.415	2011	-88.89	7.40
27.403	78	-119.26	6.61
1136.21	2.2	-130.71	7.50
1136.28	0.23	-121.62	5.36
1136.11	0.89	-126.28	6.00
1136.14	5.3	-113.77	5.69
1188.13	116	-116.23	5.85
1136.24	182	-112.00	6.01
1136.12	105	-106.06	5.09
27.392	41	-120.59	5.49
27.417	56	-107.25	5.48
1136.55	65	-121.05	8.45
1136.71	5.1	-102.48	5.14
1136.38	70	-87.42	5.12
27.388	50	-115.92	6.54
1136.59.01a	130	-107.69	5.76
1136.46	68	-133.30	6.99
Rank correlation		0.54	
R^2		0.28	
ROC area		0.77	

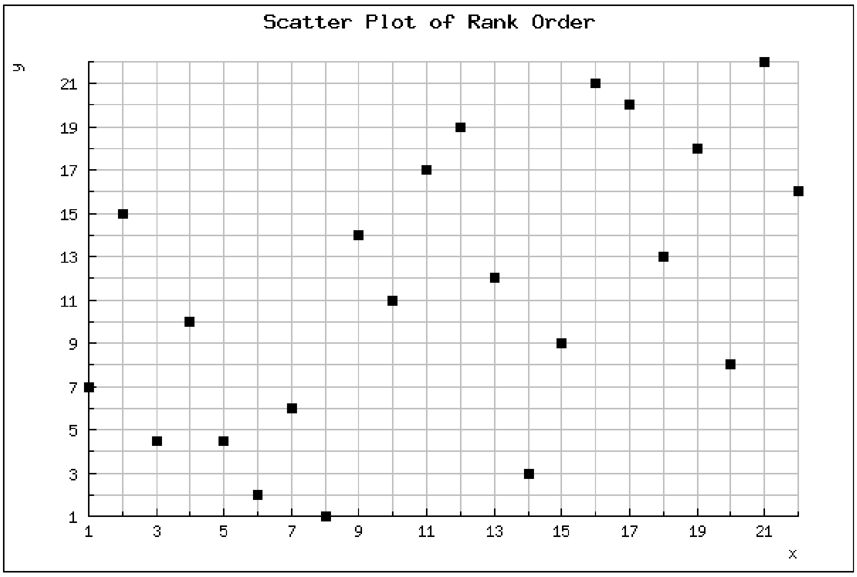


Figure 5.3: The Spearman's rank correlation of the free energy of Southwood set after molecular dynamic simulations on an implicit solvent for 1 ns by MM/GBSA method.

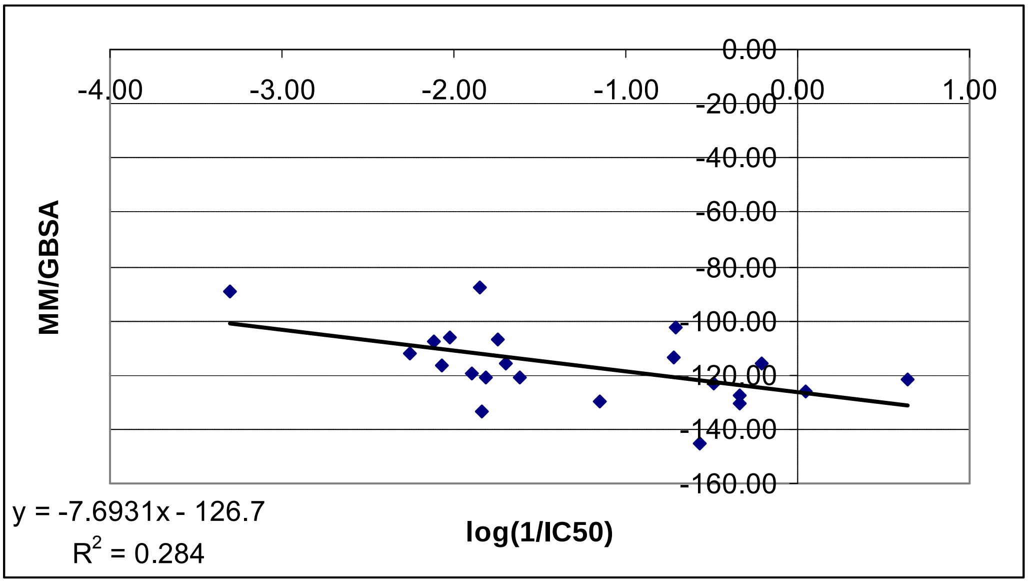


Figure5.4: The standard Pearson R^2 of the free energy of Southwood set after molecular dynamic simulations on an implicit solvent for 1 ns by MM/GBSA method.

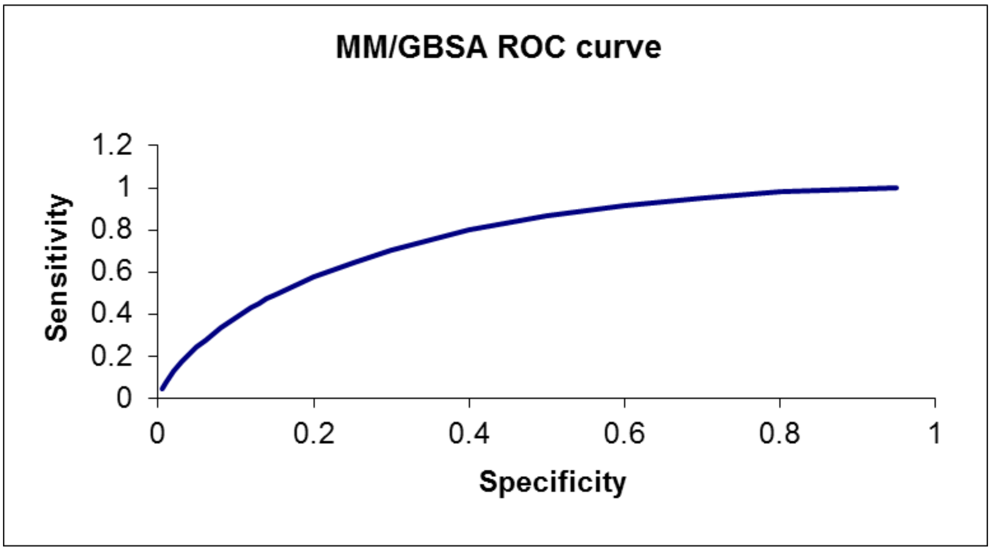


Figure 5.5: The ROC curve of the free energy of Southwood set after molecular dynamic simulations on an implicit solvent for 1 ns by MM/GBSA method.

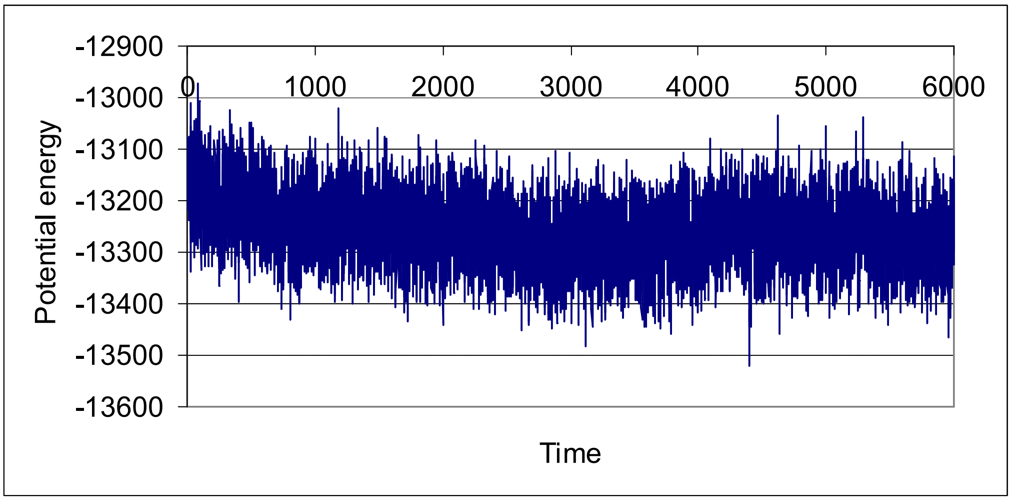
From the statistical tests (Table (5.1), Figure (5.3), (5.4) (5.5)), we can see that the results are less encouraging than the results from our previous calculations of MM/GBVI for a single x-ray structure. The rank correlation coefficient of 0.54 indicates that this method puts slightly more than 50% of peptides in the correct rank order. For ROC results, I used a cutoff of 50 nM to distinguish binders from non-binders was used, resulting in 11 peptides in each category, thereby giving a balanced test of predictions. The area under the ROC curve of 0.77 is less encouraging, comparing with the results from MM/GBVI method in previous chapter. The standard Pearson R^2 value of 0.28 indicates that this approach does not reach the accuracy of MM/GBVI method for a single x-ray structure by MOE program which gives R^2 = 0.56, rank correlation coefficient = 0.79 , and area under the ROC curve = 0.96.

5.3.2 Implicit solvent simulations (6ns):

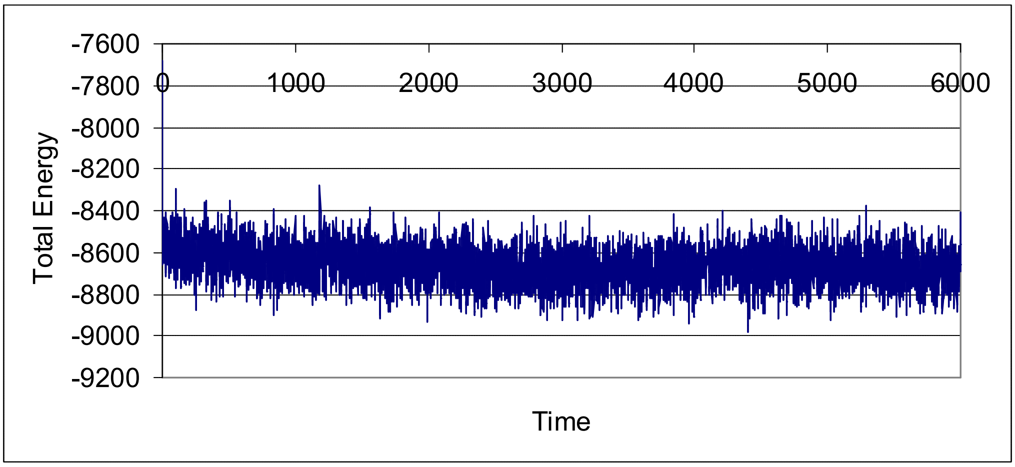
To investigate the stability of the results based on the length of the simulation, the simulations in this section were run up to 6 ns on an implicit solvent. Because of the cost for such calculations, we decided to run the simulations just for three peptides (1136.11, 27.415, and

27.417). It has been found that running several short molecular dynamic simulations provide more sample “conformations in the vicinity of the native structure” [21]. And the results by taking the averages over several molecular dynamic trajectories are different from those obtained from individual trajectories [21]. So, for each peptide, we ran an additional five molecular dynamic simulations each 1ns. Figure (5.6) shows the plots of total energy, potential energy and temperature as a function of time for (27.415) peptide. The other complexes show the same behavior; figures not shown. From these plots, it is clear that the simulations were running well and no strange behavior was accrued during the simulations. The temperature for example remained more or less constant during the simulations. Figure (5.7) shows a plot of backbone RMSD of 27.415 peptide from initial structure as a function of time. The other complexes show the same behavior; figures not shown. From Figures (5.6) and (5.7) and Table (5.2), we can see that the peptides structures have been changed during the simulations but they were close to the initial structures. The most noticeable change happened between 2.5 to 3ns Figure (5.7). After checking that the simulations were run well, we calculated the free energies of the Southwood set peptide-MHC complexes using MM/PBSA method provided by AMBER. Table (5.2) shows the free energies of the set by MM/GBSA method.

a)



b)



c)

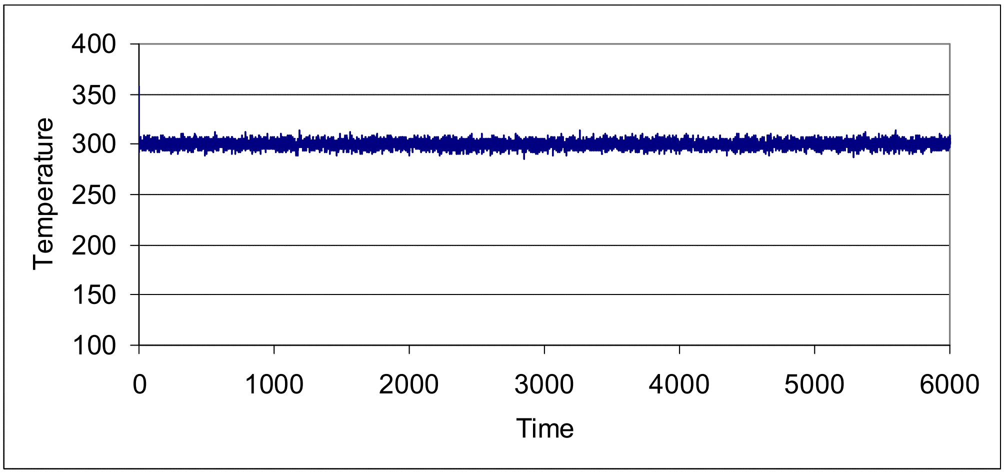


Figure5.6: The plots of a) potential energy (kcal/mole), b) total energy (kcal/mole) and c) temperature (K) as a function of time of (27.415) peptide.

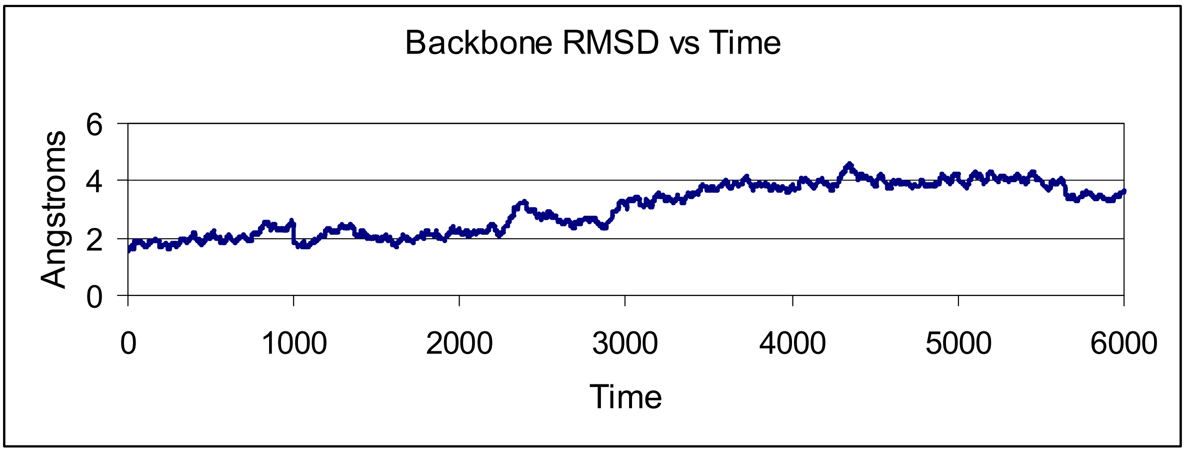


Figure5.7: The plot of backbone RMSD of 27.415 peptide from initial structures as a function of time (6ns).

Table 5.2: MM/GBSA free energies (kcal/mole) and standard deviation values of 1136.11, 27.415, and 27.417 peptides after each 1ns simulations and the averages.

Peptide No:	1136.11		27.417		27.415	
Time ns	MM/GBSA	Std. Dev.	MM/GBSA	Std. Dev.	MM/GBSA	Std. Dev.
1	**-126.28**	**6.00**	**-107.25**	**5.48**	**-88.89**	**7.40**
2	-135.90	6.05	-111.96	5.76	-95.23	5.35
3	-131.17	6.52	-124.15	5.90	-100.24	6.41
4	-142.37	7.30	-105.83	5.31	-100.79	4.51
5	-139.66	6.16	-112.74	6.96	-95.27	6.40
6	-117.45	7.50	-113.05	5.60	-100.25	4.03
Average	**-132.14**	**10.69**	**-112.50**	**8.28**	**-96.78**	**7.14**

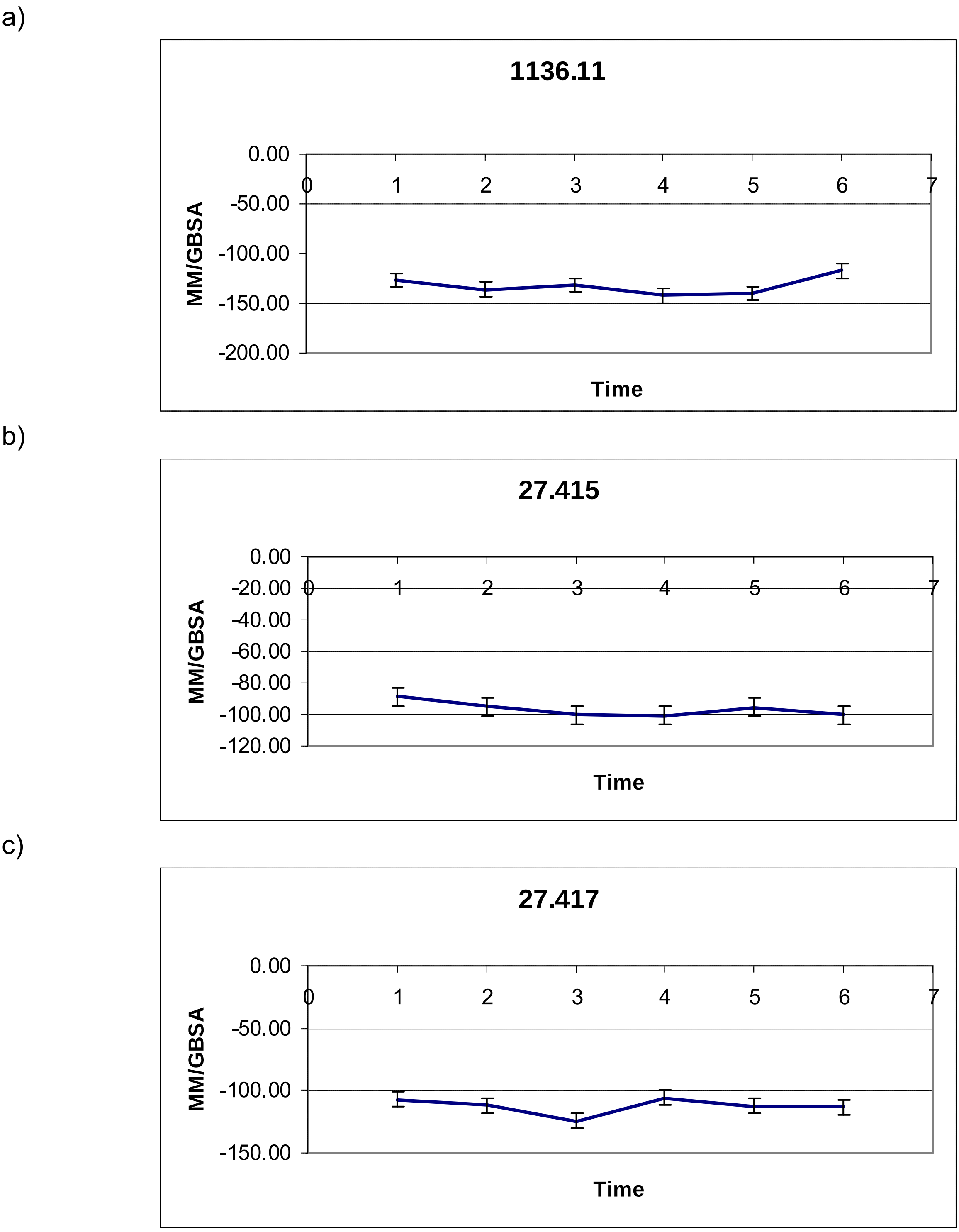


Figure 5.8: The MM/GBSA free energies plots and standard deviation for a) 1136.11, b) 27.415 and c) 27.417 peptides for each 1ns from 1ns to 6ns time.

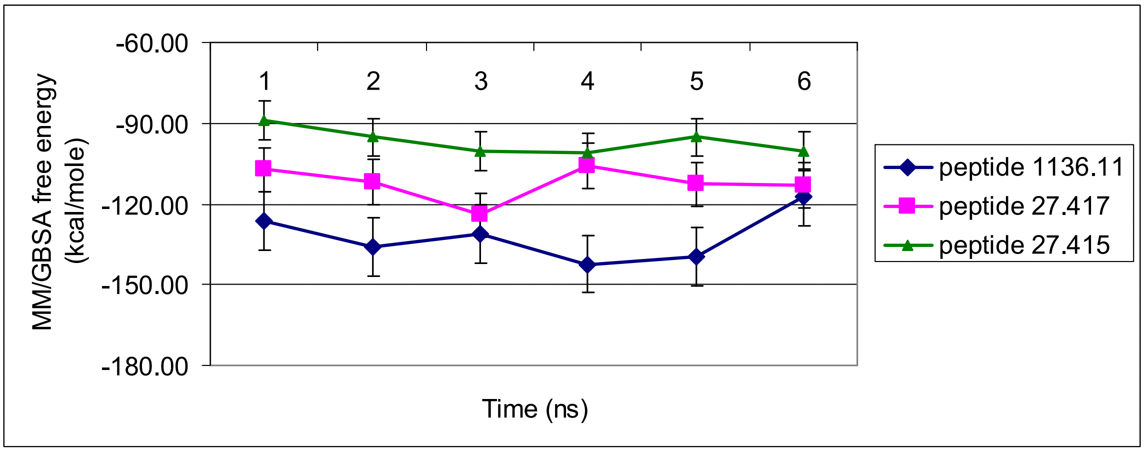


Figure5.9: Compare the MM/GBSA free energies plots and standard deviation for 1136.11, 27.415 and 27.417 peptides during 6ns simulations.

By comparing the results and the energies, average energies and standard deviation values of these three peptides after simulations for 6ns with the results of these peptides after 1ns, it has been found that there is no big influence of running the simulation between 1ns and 6ns time on the results of MM/GBSA free energies of these peptide-MHC complexes. And by comparing the time required for each 1ns with 6X 1ns time (six time longer for 6ns), we found that running the simulations for 6ns does not improve the MM/GBSA calculations sufficiently to warrant repeating such simulations for the remaining 19 peptides. Therefore, we did not run the 6ns simulation for the all peptides on Southwood set. Also, from Table (5.3) and the Figures above (5.8 and 5.9), it is obvious that the non-active peptide (27.415) got lower free energy comparing with the two active peptides. And for these three peptides, MM/GBSA method successfully predicts the correct rank order of these three peptides based on IC_{50} values, and save this order during the whole simulations of 6ns.

5.3.3 Implicit solvent simulations for longer time:

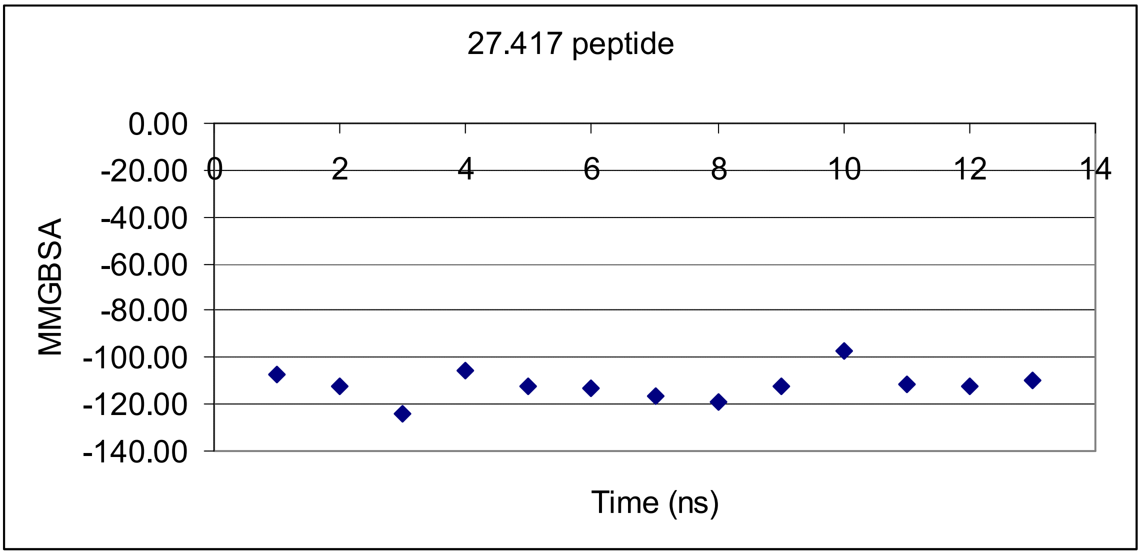
In order to investigate the effect of longer simulations, longer molecular dynamics were run for two peptides 27.415 and 27.417 starting from 6ns .rst files. The results of the binding free energy for these peptides were calculated by MM/GBSA are shown on Table (5.3) and Figure (5.10).

Table 5.3: MM/GBSA free energies (kcal/mole) and standard deviation values of 27.415 (up to 10ns) and 27.417 (up to 13ns) after each 1ns simulations.

	27.417		27.415	
Time ns	MMPBSA	Std. Dev.	MMPBSA	Std. Dev.
1	-107.25	5.48	-88.89	7.40
2	-111.96	5.76	-95.23	5.35
3	-124.15	5.90	-100.24	6.41
4	-105.83	5.31	-100.79	4.51
5	-112.74	6.96	-95.27	6.40
6	-113.05	5.60	-100.25	4.03
7	-116.86	6.35	-89.7	6.09
8	-118.73	4.53	-80.6	6.69
9	-112.68	5.82	-95.91	5.37
10	-97.42	4.99	-87.6	7.62
11	-111.51	6.95	-	-
12	-112.67	6.99	-	-
13	-109.85	6.59	-	-

From these results it is very obvious that the peptide-receptor binding is truly affected by these molecular dynamics simulations. However, the changes on the binding free energies after each 1ns simulations for these two peptides are small especially between the simulations after 1ns and after 13ns simulation time for 27.417 and 10ns for 27.425. The binding free energy for 27.417 after 1 ns simulation is (-107.25) and after 13 ns is (-109.85) kcal/mole. Similar conclusion for 27.415 after 1 ns simulation time the binding free energy is (-88.89) and after 10 ns is (-87.6) kcal/mole. By taking the averages of the binding free energies for these peptides simulations, it has been found that the average binding energy for 27.417 is (-111.9) and for 27.415 is (-93.45) kcal/mole. Table (5.3) also shows that MM/GBSA method successfully predicts the correct rank order of these two peptides based on IC_{50} values, and save this order during the whole simulations times.

a)



b)

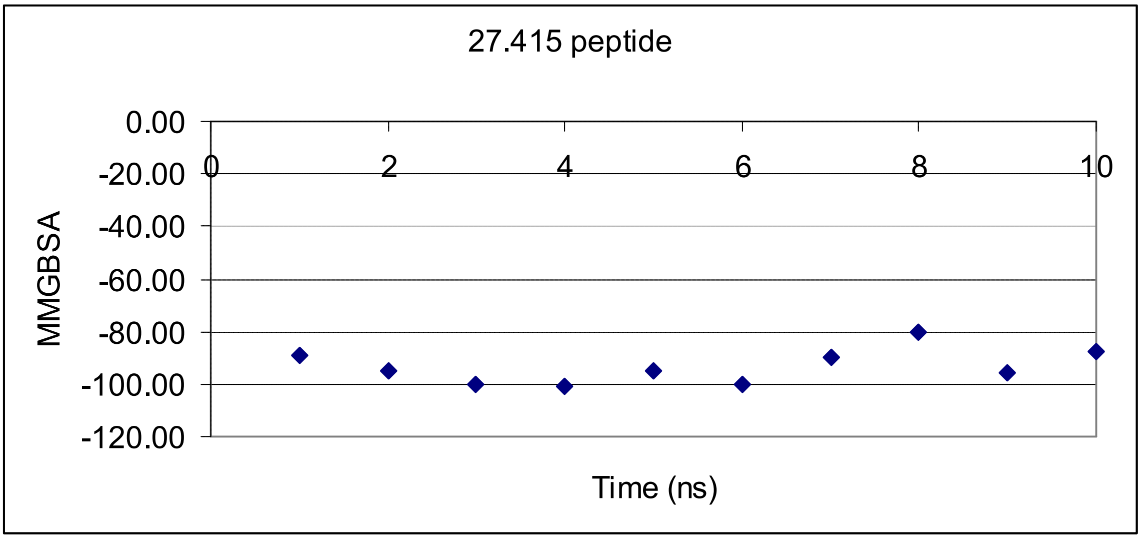


Figure 5.10: The MM/GBSA free energies plots and stander deviation for a) 27.415 peptide from 1ns to 10 ns and b) 27.417 peptide from 1ns to 13ns time.

Figure(5.11) shows the receptor binding site including 27.417 peptide after 1ns, 7ns, and 12ns times. This figure provides more evidence that the peptide-receptor binding is truly affected by these molecular dynamics simulations.

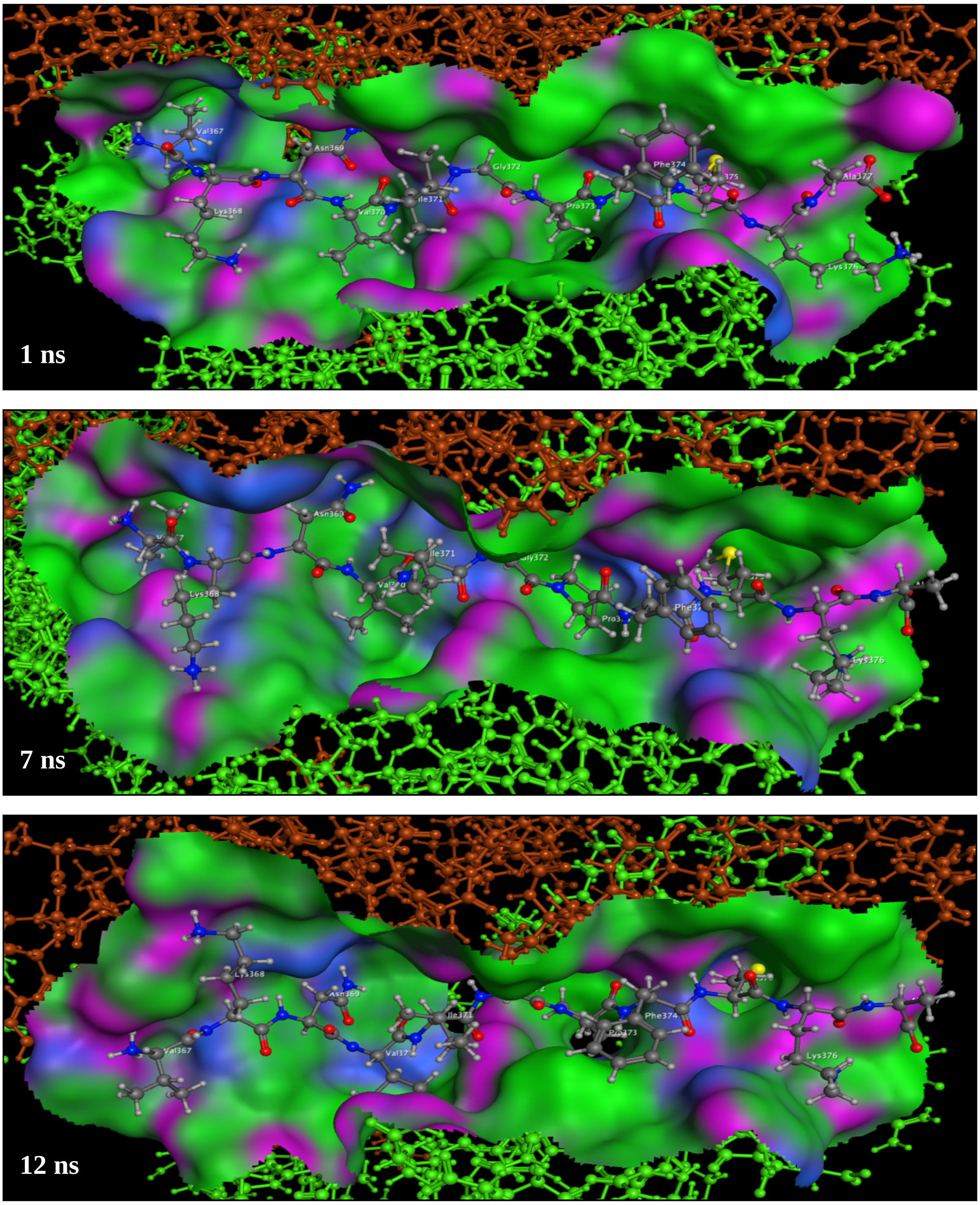


Figure 5.11: the 27.417 peptide in the binding site of the receptor (MHC II) after 1ns, 7ns, and 12ns simulation times.

Based on the results and the lack of time, it has been found that longer simulations take a lot of time and are not worth the extra resource. Therefore, longer simulations were not preferable to be run.

5.3.4 Explicit solvent simulations:

So far, all these simulations were done on implicit solvent. In our final section of the molecular dynamics simulations, we considered the explicit solvent (water) in the calculation. And because of the expensive cost of such simulations, three peptides (1136.11, 27.415 and 27.417) were considered in this section and were running for 1ns time. Figure (5.12) shows the plots of total energy, potential energy and temperature as a function of time for (1136.11) peptide, and figures C and D on the appendix show these plots for two peptides, (27.415 and 27.417) from Southwood's set. From these plots, it is clear that the simulations were running well and no strange behavior was accrued during the simulations. The temperature for example remained more or less constant during the simulations. Figure (5.13) shows the plots of the backbone RMSD of these three peptides from initial structures as a function of time. From these plots and Table (5.4), we can see that the peptides structures have been changed during the simulations but they were close to the initial structures. After checking that the simulations were run well, we calculated the free energies of these three peptides using MM/GBSA method provided by AMBER. Table (5.4) shows the free energies by MM/GBSA method.

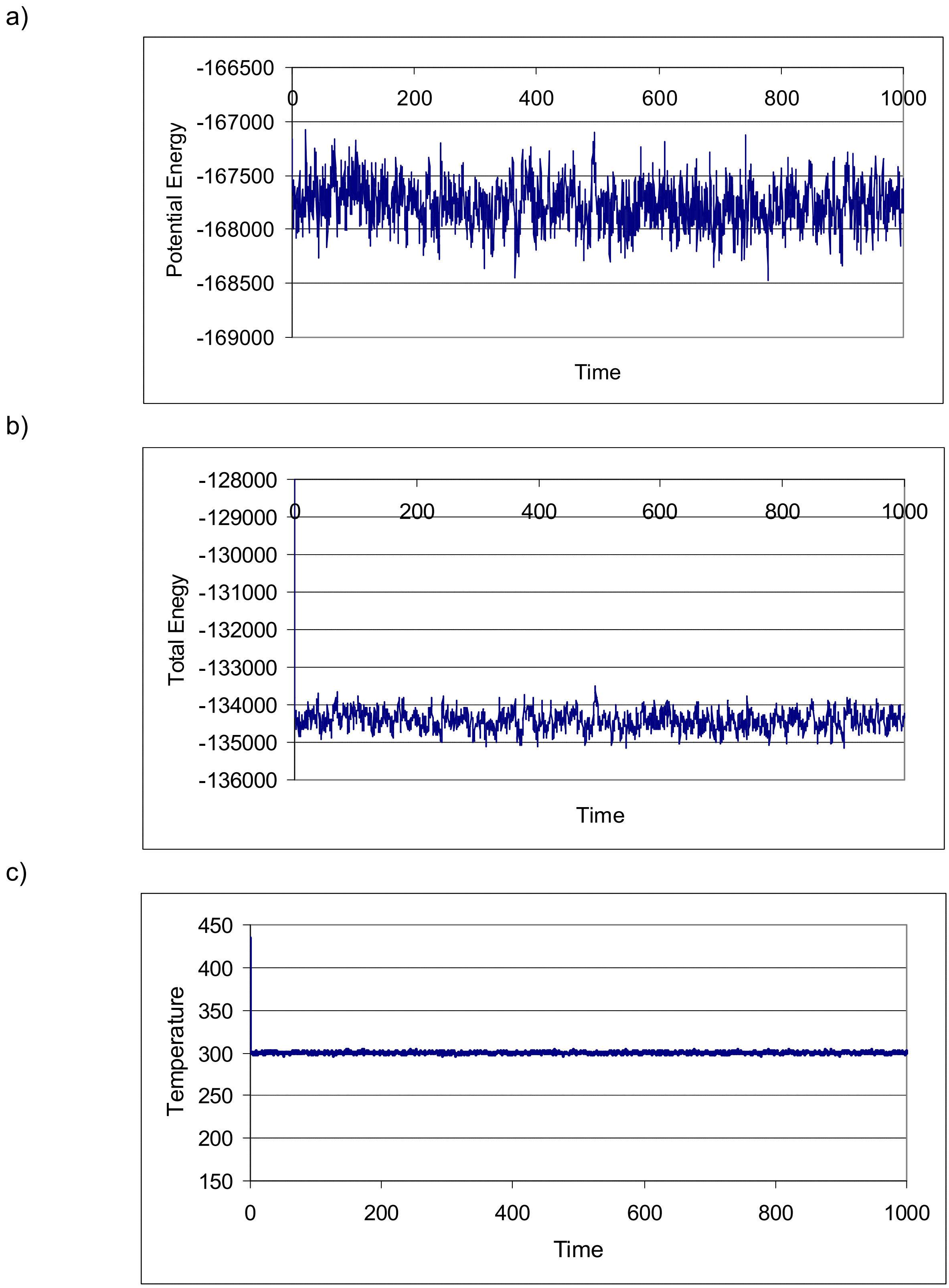


Figure5.12: The plots of a) potential energy (kcal/mole), b) total energy (kcal/mole) and c) temperature (K) as a function of time of (1136.11) peptide on explicit solvent (water).

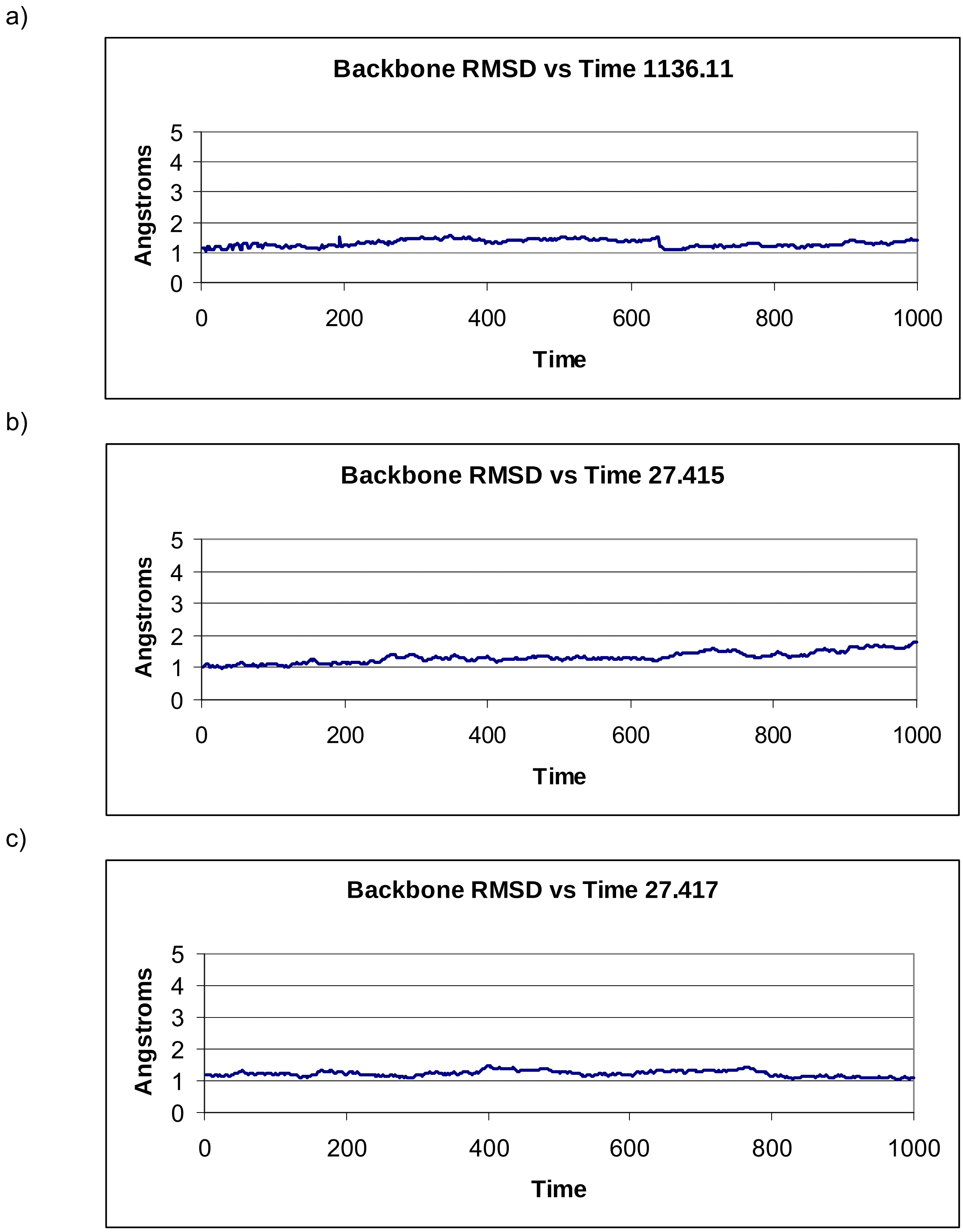


Figure5.13: The plots of backbone RMSD of a) 1136.11,b) 27.415, and c) 27.417 peptides from initial structures as a function of time on explicit solvent (water).

Table 5.4: MM/GBSA free energies (kcal/mole) and standard deviation values of 1136.11, 27.415, and 27.417 peptides after 1ns simulations on implicit and explicit solvent (water).

	Explicit solvent		Implicit solvent	
Peptide No.	MM/GBSA	Std. Dev.	MM/GBSA	Std. Dev.
1136.11	-122.91	7.39	-126.28	6.00
27.415	-111.50	3.70	-88.89	7.40
27.417	-105.38	4.78	-107.25	5.48

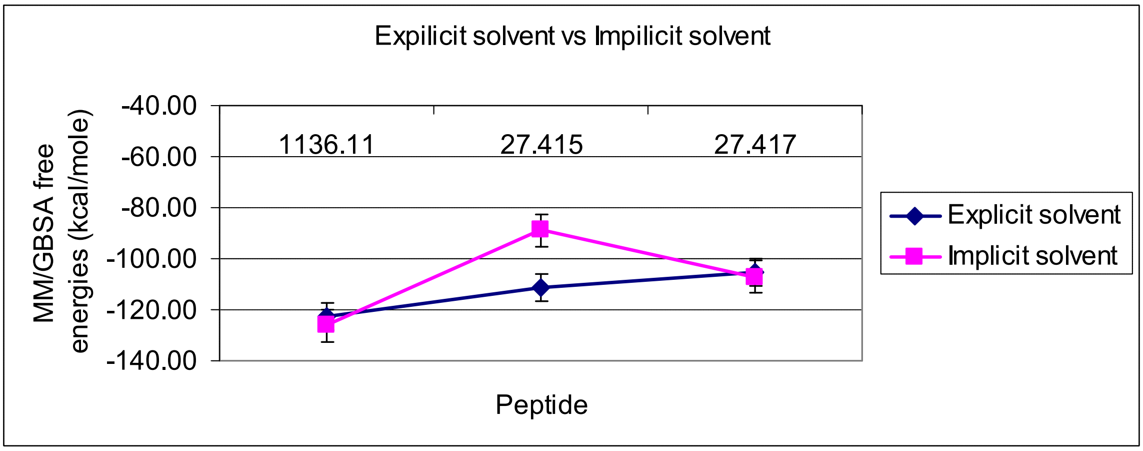


Figure5.14: The MM/GBSA free energies plots and stander deviation for 1136.11, 27.415 and 27.417 peptides for 1ns on explicit solvent and implicit solvent.

We can see from Table (5.4) and Figure (5.14), that the results for the two active peptides (1136.11 and 27.417) are very close. On the other hand, the results for the non-active peptide (27.415) are also not that close. These results are perhaps not too surprising. The interaction between the active peptides and the receptor are strong, therefore the influence of the explicit solvent is not very noticeable comparing with non-active peptide. This may mean that the presence of water molecules between the peptide and the receptor is responsible for the high binding energies for 27.415 peptide Figure (5.15). This peptide has a value of 2011 of IC_{50}, where the cutoff of the active peptides is IC_{50}=50 (peptide with IC_{50} equal to 50 and more is consider non-active peptide).

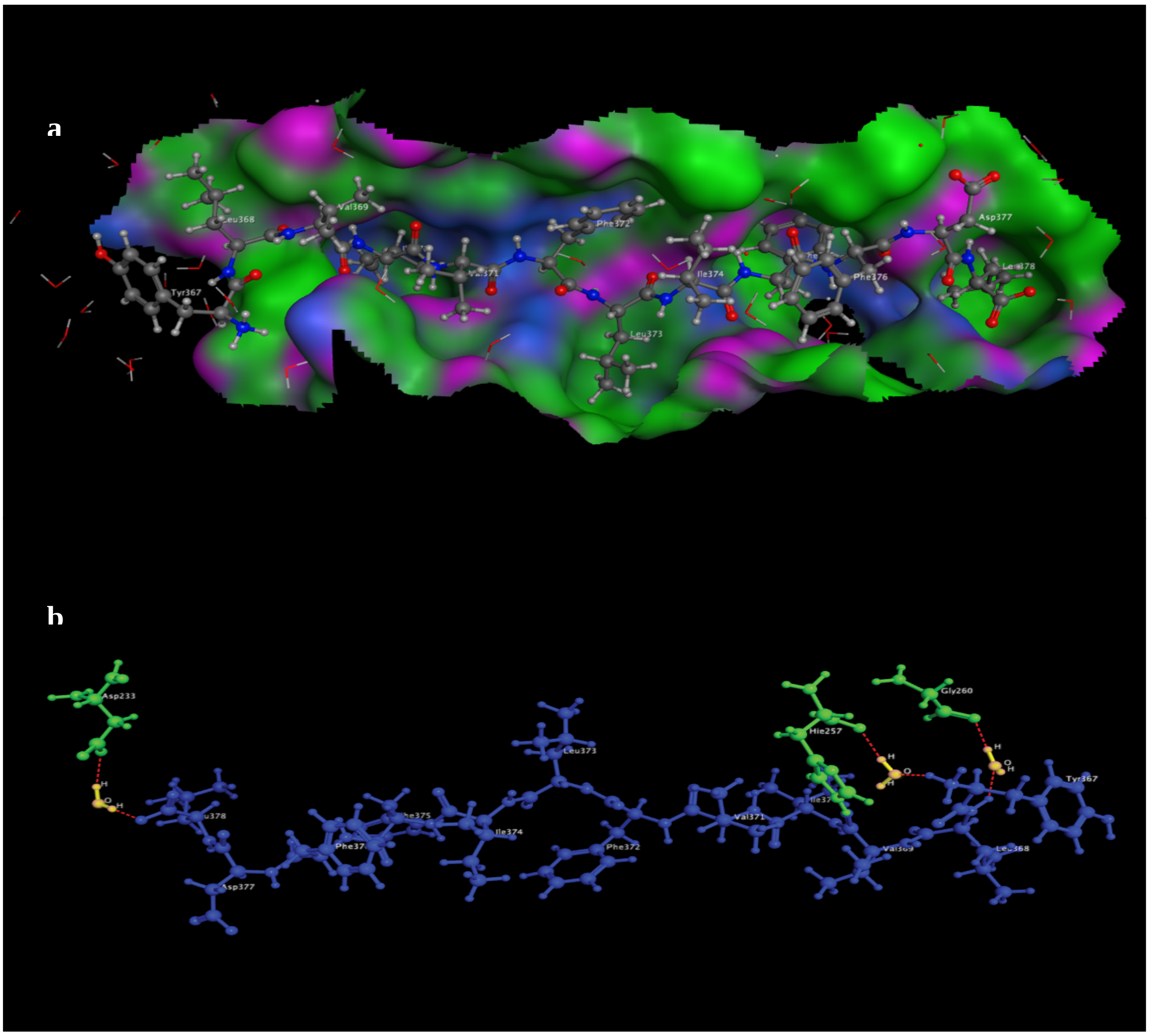


Figure5.15: a) 3D structure of (27.415) peptide and water molecules on the active site of the receptor. b) 3D structure show the hydrogen bonds (red) between the peptide (blue), water molecules (yellow) and three residues from the receptor (green).

5.4 Conclusion:

AMBER suite of programs was used to investigate the influence of the motion in implicit and explicit solvents for Southwood set. The binding free energies were calculated by Molecular Mechanics Generalized -Born Surface Area (MM/GBSA) method. From the statistic methods for implicitly solvated simulation for 1ns, it has been found that the results are less encouraging than the results of using MM/GBVI on MOE, which show that the MM/GBVI approach can deliver reasonable predictions of peptide-MHC binding in a matter of a few seconds on a desktop computer.

Moreover, by increasing the simulation time, and using explicit solvent, the results of MM/GBSA show slightly improving but no big changes were found. Therefore, by considering the cost of these simulations and the results, we found that implicitly solvated simulation for 1ns is suitable for such investigation.

5.5 References:

[1] Karplus, M.; McCammon, J. A. *nature structural biology*, **2002**, *9*, 646-788.

[2] Karplus, M.; Petsko, G. A. *Nature*, **1990**, *347*, 631-639.

[3] McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature*, **1977**, *267*, 585-590.

[4] Brunger, A. T.; Brooks, C. L. I.; Karplus, M. *Proceedings of the National Academy of Sciences in USA.*, **1985**, *82*, 8458-8462.

[5] Gunsteren, W. E. v.; Berendsen, H. J. C. *Angewandte Chemie International Edition*, **1990**, *29*, 992-1023.

[6] Southwood, S.; Sidney, J.; Kondo, A.; Guercio, M.-F. d.; Appella, E.; Hoffman, S.; Kubo, R. T.; Chesnut, R. W.; Gery, H. M.; Sette, A. *The Journal of Immunology*, **1998**, *160*, 3363-3373.

[7] Davis, M. E.; McCammon, J. A. *Chemical Review*, **1990**, *90*, 509–521.

[8] Sharp, K. A.; Honig, B. *Annual Review Biophysics Biophysical Chemistry*, **1990**, *19*, 301-332.

[9] Warshel, A.; Papazyan, A. *Current Opinion Structure Biology*, **1998**, *8*, 211-217.

[10] Tajkhorshid, E. J., K. J; Suhai, S. *Journal of Physical Chemistry B.*, **1998**, *102*, 5899-5913.

[11] Baker, N., **2008**.

[12] Dong, F.; Wagoner, J. A.; Baker, N. *Physical Chemistry Chemical Physcis*, **2008**, *10*, 4889-4902.

[13] Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; III, T. E. C.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. *Computer Physics Communications*, **1995**, *91*, 1-41.

[14] Case, D. A.; Darden, T. A.; Cheatham, I., T.E. ; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B. P.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Liu, 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, V.; Babin, C.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. In *http://ambermd.org/*, **2010**.

[15] Jorgensen , W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *Journal of Chemical Physics*, **1983**, *79*, 926-935.

[16] In *<http://ambermd.org/doc6/html/AMBER-sh-6.7.html>*. **2012**

[17] Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J., C. *Journal of Computational Physics*, **1977**, *23*, 327-341.

[18] In *http://www.wessa.net/rankcorr.wasp*. **2012**

[19] Dorfman, D. D.; Alf, E. *Journal of Mathematical Psychology*, **1969**, *6*, 487.

[20] Metz, C. E.; Herman, B. A.; Shen, J.-H. *Statistical Medicine* , **1998**, *17*, 1033.

[21] Caves, L.; Evanseck, J.; Karplus, M. *Protein science* **1988**, *7*, 7649-7666.

Chapter 6:

Conclusions:

6.1 Conclusions:

On this study several approximate methods have been used to predict the peptide-protein interaction energy in vacuum and in solvent, starting from pairwise interaction, though whole peptide-protein interaction and ending with the influence of the motion by using molecular dynamics simulations.

For the prediction of the pairwise interaction in vacuum, focussing on the interaction of a peptide implicated in multiple sclerosis with its biological MHC receptor, it has been found that the semi-empirical RM1 approach with additional correction for dispersion effects gives the best prediction comparing with *ab initio* methods. This study also shows that increasing the size of model systems used to represent the bound peptide from single amino acids to dipeptides and heptapeptides increases predicting interaction energy, as does expanding the number of amino acids used to model the receptor.

For the prediction of the peptide-protein binding energy including the effect of the solvent, several methods have been tested to calculate the interaction energy for peptide-MHC-II complexes for three separate data sets, using IC_{50} data to evaluate the accuracy of each theoretical method. The results show that MM/GBVI approach is a promising way to calculate the binding energy for peptide-protein systems, with reliable performance for all three data sets as measured by three distinct statistical tests. MM/GBVI can also be used to predict the anchor residues that reside in receptor binding pockets, and this approach gives slight improvement in statistics over purely sequence-based prediction methods such as SYFPEITHI or SVMHC.

Of course, both peptide ligand and protein receptor are flexible objects, such that exploring the use of molecular dynamics simulation to calculate the peptide-protein binding energy averaged over multiple snapshots was required. AMBER suite of programs was used to investigate the influence of the motion in implicit and explicit solvents for Southwood set. The binding energies were calculated by Molecular Mechanics Generalized -Born Surface Area (MM/GBSA) method. From the statistical methods, it has been found that the results are less encouraging than the results of single x-ray or "docked" structure using MM/GBVI on MOE, which shows that the MM/GBVI approach can deliver reasonable predictions of peptide-MHC binding in a matter of a few seconds on a desktop computer.

In conclusion, the prediction of binding energy of peptide-protein system which is mainly controlled by non-covalent interactions is not an easy task considering many factors such as the effect of the solvent, the size of the complex and the running time of the molecular dynamics simulations. These factors play important rolls in such calculations. In this work we tried to clarify the effects of these factors and compare the performance of several approximation methods.

More investigations on this topic are always required: for example, it would be interesting to examine in more detail the role of explicit water in peptide-protein binding. We showed in Chapter 5 that water molecules can come between peptide and receptor, but did not have time to analyse this data in more detail. Analysis of residence times of individual water molecules and their effect on binding energy could be a valuable area for new study. Another possible area of future work would be in design of non-peptide molecules that can bind to MHC-II receptors, as potential leads for new treatments of MS and related auto-immune diseases. The same methods described in this work, especially RM1-D, MM/GBVI and MM/GBSA, can be applied to organic and drug-like molecules. If combined with more traditional drug discovery methods such as automated docking and/or pharmacophore searching, we may be able to suggest new drug leads for further testing.

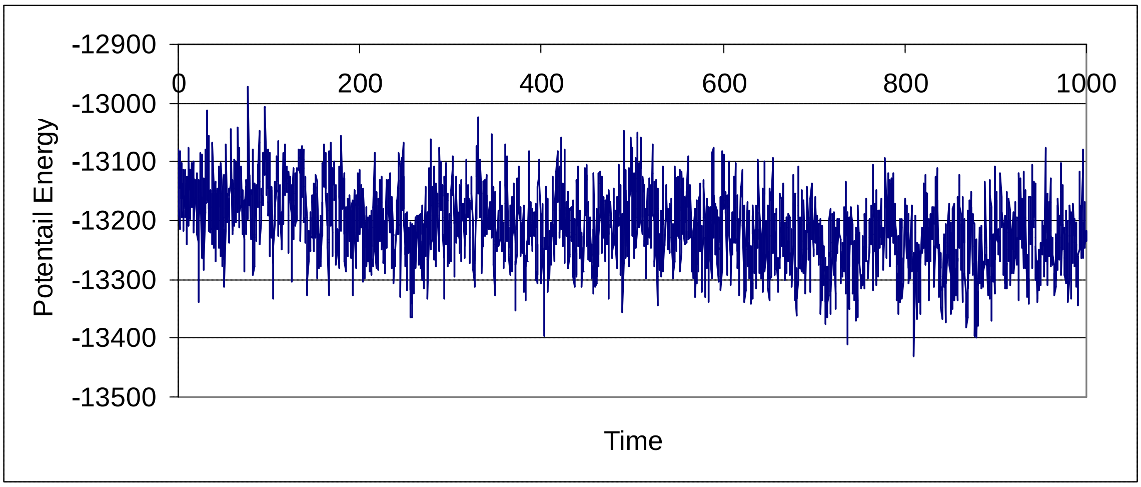
Appendix

keys and descriptions	values used
imin= to run minimization	0= no minimization just MD simulation 1= Perform minimization (no molecular dynamics)
ntx= Option to read the initial coordinates, velocities and box size from the "inpcrd" file	1= X is read formatted with no initial velocity information (default) 7= X, V and BOX(1..3) are read formatted
irest= to restart the run	0= No effect (default) 1= restart calculation
ntpr= every ntpr steps energy and temperature will be printed on mdinfo file	500, 200
ntwr= every NTWR steps during dynamics, the "restrt" file will be written	5000
ntwx= every NTWX steps the coordinates will be written to file "mdcrd".	500, 200
ntf= force evaluation, If SHAKE is used (see NTC), it is not necessary to calculate forces for the constrained bonds	2= bond interactions involving H-atoms omitted (use with NTC=2)
ntc= Flag for SHAKE to perform bond length constraints	2= bonds involving hydrogen are constrained
ntb= Periodic boundary.	0=no periodicity is applied 1= constant volume (default) 2= constant pressure
ntp= Flag for constant pressure dynamics. This option MUST be set to 1 or 2 when Constant Pressure periodic boundary conditions are used (NTB = 2).	0= Used with NTB not = 2 (default) 1= md with isotropic position scaling
igb= Controls the use of the generalized Born model.	1= the pairwise generalized Born model is used
ntr= Flag for restraining specified atoms in Cartesian space using a harmonic potential	0= No position restraints (default) 1= MD with restraint of specified atoms
maxcyc= Maximum number of cycles of minimization	1000
ncyc= After NCYC cycles the method of minimization would be switched from steepest descent to conjugate gradient method.	500
nstlim= Number of MD-steps per NRUN to be performed	25000, 50000, 500000
temp0= Reference temperature at which the system is to be kept, Default 300	300
tempi= Initial temperature	0, 300. the velocities will be calculated from the forces instead
ig= The seed for the random number generator	-1
ntt= For temperature regulation, the Langevin thermostat (NTT=3) is used to maintain the temperature of the system at 300 K This method uses Langevin dynamics with a collision frequency given by gamma_In	3
gamma_In	1
taup= Pressure relaxation time (in ps), when NTP > 0	2
cut	12= a cut off of 12 angstroms.
END= END of this section.	

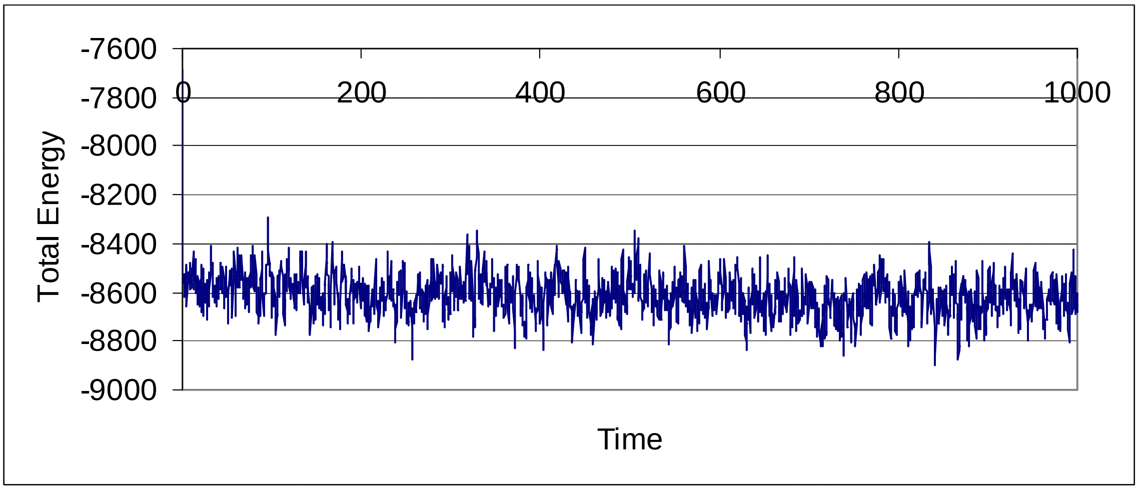
Table A: minimizations and molecular dynamic simulations Keywords, descriptions, and values.

Figure A: The plots of a) potential energy (kcal/mole),b) total energy (kcal/mole) and c) temperature (K) as a function of time of (27.415) peptide.

a)



b)



c)

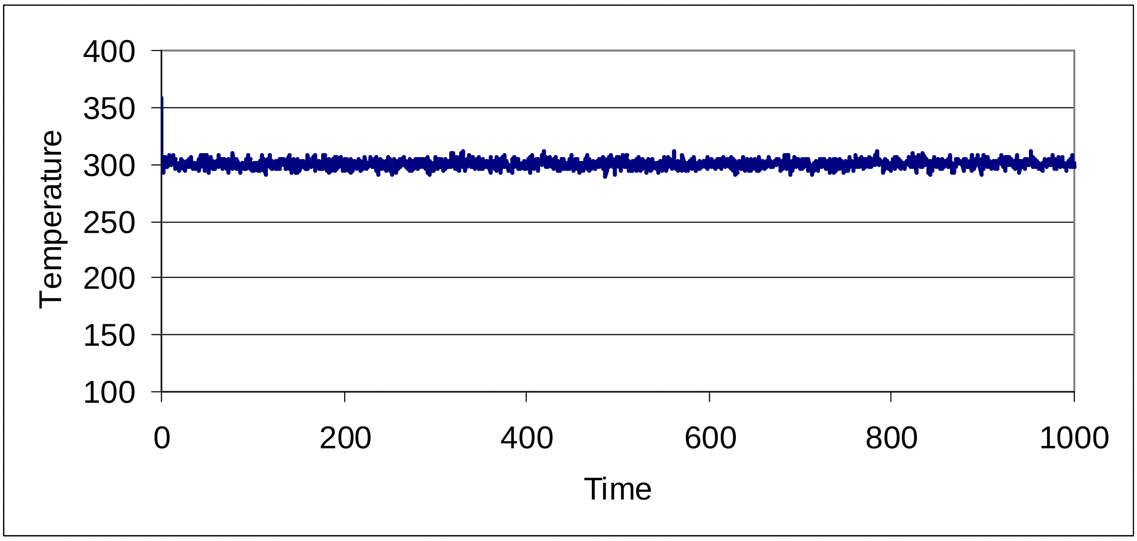
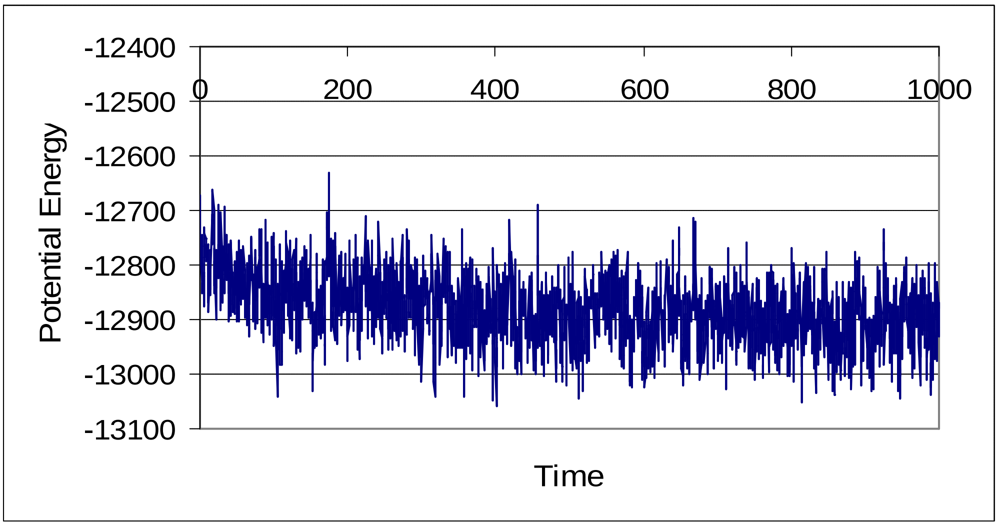
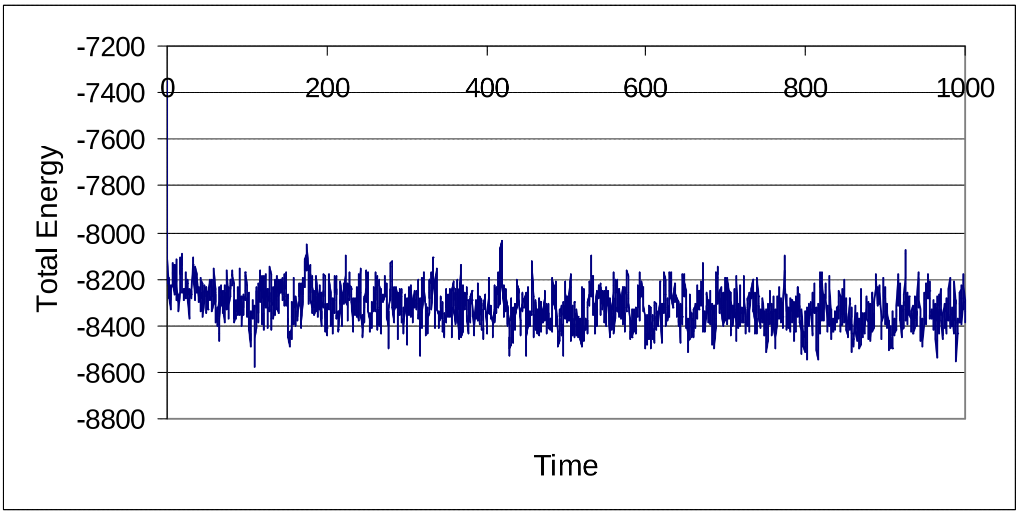


Figure B: The plots of a) potential energy (kacl/mole), b) total energy (kcal/mole) and c) temperature (K) as a function of time of (27.417) peptide.

a)



b)



c)

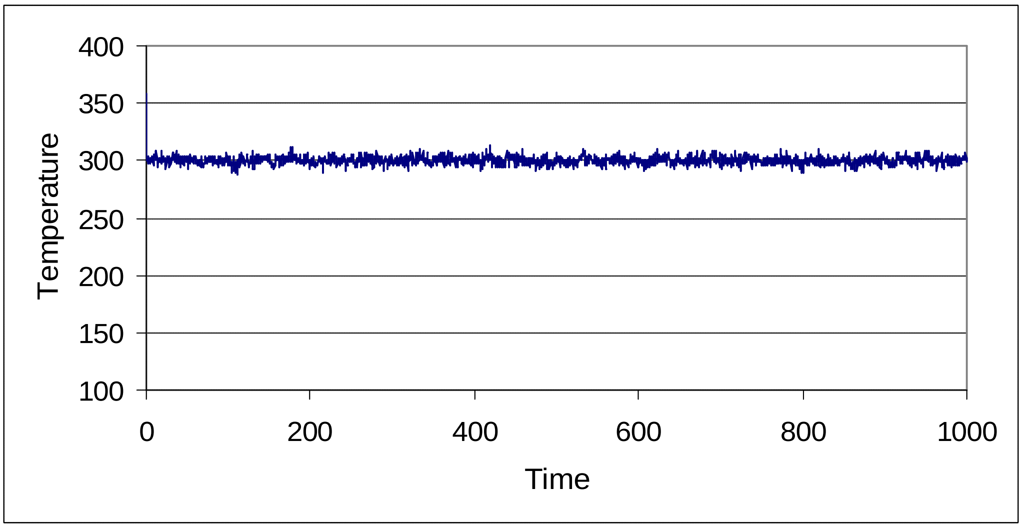
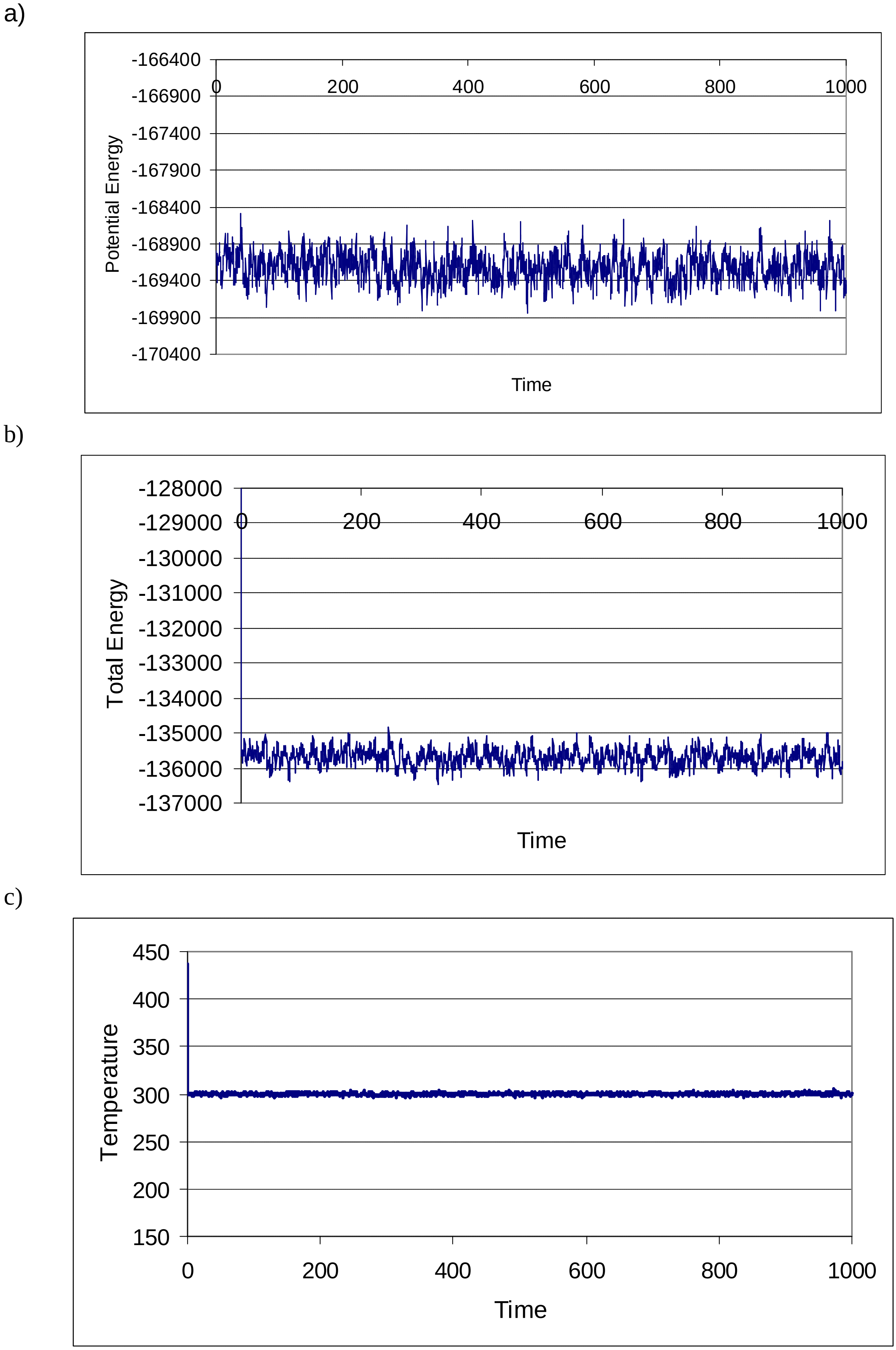


Figure C: The plots of a) potential energy (kcal/mole),b) total energy (kcal/mole) and c) temperature (K) as a function of time of (27.415) peptide on explicit solvent (water).



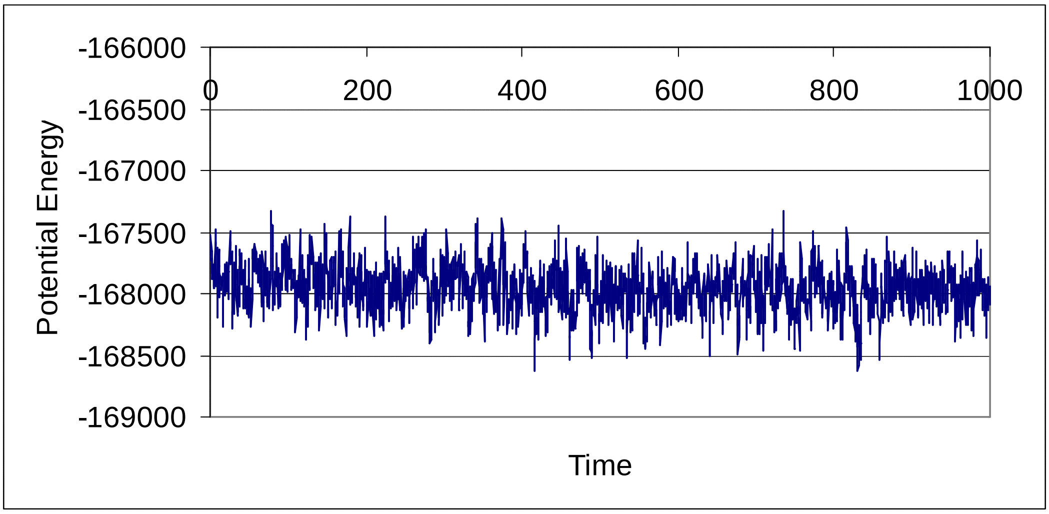
a)

b)

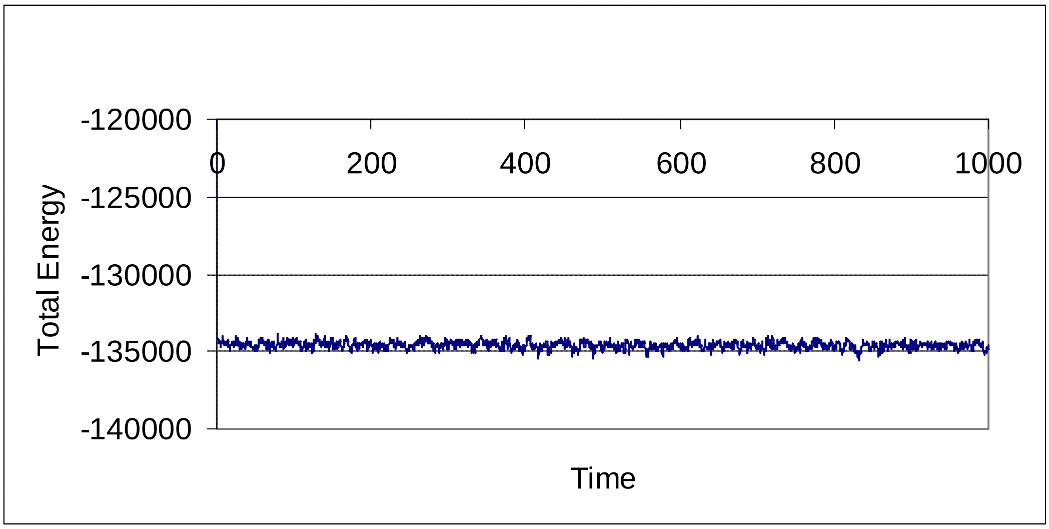
c)

Figure D: The plots of a) potential energy (kcal/mole), b) total energy (kcal/mole) and c) temperature (K) as a function of time of (27.417) peptide on explicit solvent (water).

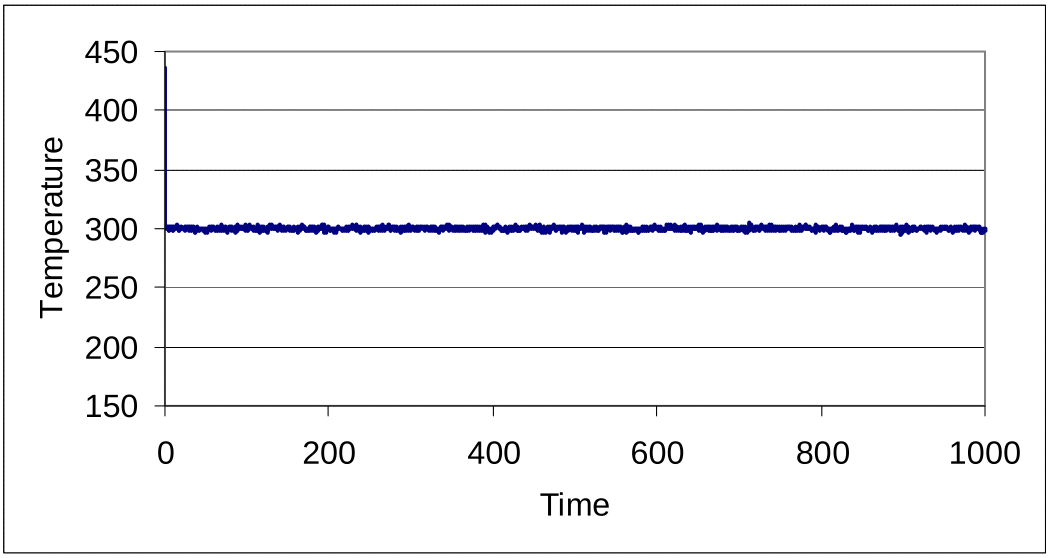
a)



b)



c)



Input files for molecular dynamics simulation in implicit solvent:

The minimization input file (min.in):

```
&cntrl
  imin   = 1,
  maxcyc = 1000,
  ncyc   = 500,
  ntb    = 0,
  igb    = 1,
  cut    = 12
 /
```

The heating input file (heat.in):

```
Heating up from 0 to 300K
 &cntrl
  imin=0,irest=0,ntx=1,
  nstlim=25000,dt=0.002,
  ntc=2,ntf=2,
  cut=12.0, ntb=0,
  ntpr=500, ntwx=500,
  ntt=3, gamma_ln=2.0,
  tempi=0.0, temp0=300.0,
  igb=1, ig=-1
 /
```

Short simulations (md1.in):

```
&cntrl
  imin = 0, ntb = 0,
  igb = 1, ntpr = 100, ntwx = 100,
  ntt = 3, gamma_ln = 1.0,
  tempi = 300.0, temp0 = 300.0
  nstlim = 100000, dt = 0.001,
  cut = 12.0
  ig=-1
 /
```

Simulation input file (md2.in):

```
&cntrl
  imin = 0, ntb = 0,
  igb = 1, ntpr = 200, ntwx = 200,
  ntc = 2, ntf = 2
  ntt = 3, gamma_ln = 1.0,
  tempi = 300.0, temp0 = 300.0
  nstlim = 500000, dt = 0.002,
  cut = 12.0
  ig=-1
 /
```

Input files for molecular dynamic in explicit solvent simulation:

The minimization input file (min1.in):

```
&cntrl
  imin   = 1,
  maxcyc = 1000,
  ncyc   = 500,
  ntb    = 1,
  ntr    = 1,
  cut    = 12
 /
Hold the protein fixed
500.0
RES 1 379                      GROUP input
END
END
```

Minimization the whole system (min2.in):

```
&cntrl
  imin   = 1,
  maxcyc = 2500,
  ncyc   = 1000,
  ntb    = 1,
  ntr    = 0,
  cut    = 12.0
 /
```

The heat input file (heat.in):

```
&cntrl
  imin   = 0,
  irest  = 0,
  ntx    = 1,
  ntb    = 1,
  cut    = 12.0,
  ntr    = 1,
  ntc    = 2,
  ntf    = 2,
  tempi  = 0.0,
  temp0  = 300.0,
  ntt    = 3,
  gamma_ln = 1.0,
  nstlim = 10000, dt = 0.002
  ntpr = 200, ntwx = 200, ntwr = 2000
 /
Keep protein fixed with weak restraints
10.0
RES 1 379
END
END
```

The equilibrium input file (equl.in):

```
&cntrl
  imin = 0, irest = 1, ntx = 7,
  ntb = 2, pres0 = 1.0, ntp = 1,
  taup = 2.0,
  cut = 12, ntr = 0,
  ntc = 2, ntf = 2,
  tempi = 300.0, temp0 = 300.0,
  ntt = 3, gamma_ln = 1.0,
  nstlim = 50000, dt = 0.002,
  ntpr = 500, ntwx = 500, ntwr = 5000
 /
```

Molecular dynamic simulations (md2.in):

```
&cntrl
  imin = 0,
  ntb = 1,
  cut = 8.0, ntr = 0,
  ntc = 2, ntf = 2,
  tempi = 300.0, temp0 = 300.0,
  ntt = 3, gamma_ln = 1.0,
  nstlim = 500000, dt = 0.002,
  ntpr = 200, ntwx = 200, ntwr = 2000
  ig=-1
 /
```