Molecular modeling and simulation for proteins and polymer membranes

Christian, Bope Domilongo

2017

Christian, B. D. (2017). Molecular modeling and simulation for proteins and polymer membranes. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/72871>

<https://doi.org/10.32657/10356/72871>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**MOLECULAR MODELING AND SIMULATION FOR
PROTEINS AND POLYMER MEMBRANES**

BOPE DOMILONGO CHRISTIAN

SCHOOL OF BIOLOGICAL SCIENCES

2017

MOLECULAR MODELING AND SIMULATION FOR PROTEINS AND POLYMER MEMBRANES

BOPE DOMILONGO CHRISTIAN

SCHOOL OF BIOLOGICAL SCIENCES

A thesis submitted to the Nanyang Technological University

in partial fulfillment of the requirement for the degree of

Doctor of Philosophy

2017

Acknowledgements

My heartfelt gratitude goes to my supervisor, A/Prof. Lu Lanyuan as well as my co-supervisors Dr. Zhang Zuoqi (Wuhan University, China) and Dr. Cheng Yu (A*STAR) for this great opportunity to conduct my Ph.D under their supervisions. Without their commitment, immeasurable help, guidance, invaluable suggestions and motivation, this Ph.D journey would not reach a satisfactory end.

I would like also to take this opportunity to extend my gratitude to my thesis advisory committee members A/Prof. Mu Yuguang and A/Prof. Shuzhou Li for their effective, insightful suggestions, substantial input and recommendation.

I am also grateful to my past and present lab mates Dr. Nalaparaju Anjaiah, Dr. Tong Dudu, Dr. Chen Wenwen, Mr. Zhang Haiping, Dr. Zhu Guanhua for their collaborative support and open-minded scientific discussions.

My appreciation to Professor Benoit Morel for the proofreading of my thesis at the Carnegie Mellon University, United State of America.

A special thanks goes to my family, particularly my father Crispin Bope Domilongo may his soul rest in peace and receive the fruits of his huge sacrifices; to my mother Josephine Bokashanga. To my lovely wife Florence Bope and my son Precious Bope Domilongo for their love and prayers.

Last but not the least my gratitude goes to the Singapore International Graduate Award (SINGA), NTU and School of Biological Sciences for the financial support.

Table of Contents

Table of Contents	1
List of Figures.....	3
List of Tables	10
Abbreviations	11
Abstract.....	15
1. General Introduction.....	17
1.1 Multiscale models for biomolecular systems	18
1.2 Wastewater treatment	27
1.2.1 Micropollutants.....	30
1.2.2 Polymer membranes	32
1.3 Outline of the research work	32
2. General Methodologies	36
2.1 Molecular simulations	36
2.1.1 Quantum mechanics.....	36
2.1.2 Molecular mechanics and molecular dynamics	37
2.1.3 Free energy calculation and umbrella sampling	42
2.2 Multiscale parameterization	48
2.2.1 Heterogeneous elastic network models and structure-based models from fluctuation matching targeted at atomistic molecular dynamics simulations.....	48
2.2.2 Fluctuation matching targeted at fine-grained normal modes analysis	51
2.2.3 Fluctuation matching targeted at experimental B-factors	56
2.2.4 Relative-entropy-based fluctuation matching with improved convergence	60
3. An improved fluctuation matching approach based on relative entropy for the elastic network model and a structure-based model of biomacromolecules	61
3.1 Introduction	61
3.2 Theoretical framework and computational details.....	62
3.3 Parameterization with Non-Negativity Constraints	65

3.4 Results and discussions	67
3.5 Comparison between relative entropy with non-negativity force constant and fluctuation matching related methods	81
3.6 Conclusions	83
4. Molecular simulation to investigate the adsorption mechanism of the endocrine disrupting compound 17-α ethinylestradiol at the monomer level.....	84
4.1 Introduction	84
4.2 Modeling and simulations	85
4.3 Results and discussions	90
4.4 Conclusions	96
5. Molecular insight into the interaction between 17-α ethinylestradiol and polymer membranes in wastewater purification	97
5.1 Introduction	97
5.2 Modeling and simulations	99
5.2.1 Atomistic models	99
5.2.2 Membrane solution interface simulation system	100
5.2.3 Solution inside the membrane.....	102
5.3 Results and discussions	106
5.3.1 Adsorption of EE2 onto polymer membranes	106
5.3.2 Potential of mean force.....	109
5.3.3 Mean-squared displacements and diffusion coefficients.....	117
5.3.4 Non-bonded interaction energy	122
5.3.5 Insight into atomistic interaction	124
5.4 Conclusions	130
6. Concluding remarks	131
References	133
Publications	158

List of Figures

Figure 1.1 Four free vibration modes of F-actin (52-protomers, 0.14 μ m length) in planar deformation obtain from FEM method. The angular frequencies are 6.7×10^{-4} , 7.01×10^{-4} , 1.97×10^{-3} and 2.011×10^{-1} cycles/ps.....6

Figure 1.2 Atomistic model of F-actin from Holmes *et al.*, (PDB ID: 1J6Z) [1] constructed from a single G-actin using the multiscale approach. Each domain is represented by one bead labeled as CG1, CG2, CG3 and CG4. (a) 13 units F-actin from the Holmes model (b) single actin monomer, (c) G-actin coarse-grained by domain, binding a nucleotide ADP.....10

Figure 1.3 Water stress map by country. Figure adapted from ref. (<http://www.wri.org/>)...
.....12

Figure 2.1 Free energy surface. (a) Scheme of the multidimensional energy landscape of fluorescent protein. The red arrow represents unfolding pathway. (b) Illustration of the energy landscape along the unfolding pathway onto one reaction coordinates. The experiment start from minimum 1 is able to overcome lower energy barrier B and sample in minimum 2, but it is challenging for the system to overcome the energy barrier C and D. Figure is adapted from [2].....28

Figure 2.2 Principle of umbrella sampling simulation. The red dot represents protein A that is fixed and dot blue represents protein B that is pulled away by its center of mass to generate different initial configurations. (a) Pulling simulation-using SMD to obtain a series of starting configurations for umbrella sampling windows. (b) The starting

independent simulation (blue dot) obtained from SMD, and the black line represents the bias umbrella potential. (c) The result is illustrated as a histogram of configuration, with neighboring windows overlapping in order to have a continuous energy function. Figure is adapted from (www.bevanlab.biochem.vt.edu).....30

Figure 2.3 CG-ENM scheme of lysozyme. Lysozyme is coarse-grained into seven-beads based on the C α atom sequence, as follow: CG1 (1-12), CG2 (13-58), CG3 (59-80), CG4 (81-91), CG5 (92-123), CG6 (124-143), and CG7 (144-164). The seven coarse-grained beads (CG1~CG7) are colored in blue, red, orange, yellow, green, violet and gray, respectively. (a) The side view Illustrates the closure binding motion of domain A (CG1~CG2) and domain B (CG4~CG7). The arrows represent the closure movement. (b) The front view represents a twisting motion between domain A and B. The arrows represent the distortion direction.....39

Figure 2.4 The comparisons of the experimental and calculated B-factors for (a) crambin, (b) lysozyme, (c) trypsin inhibitor, and (d) HIV-1 protease. The red circles represent the experimental values. ANM and HANM using a 15 Å distance cutoff are represented using green and blue curves.....43

Figure 3.1 Comparison of the residual convergence from RE and FM for C α -level ENMs. (a) 1GB1: the RE is denoted by the red curve (with the final residual 0.142) and FM is denoted in green curve (with the residual 0.162). (b) 1GB4: the RE is denoted in red curve (with the final residual 0.169) and FM is denoted in green curve (with the final residual 0.20). 71

Figure 3.2 Comparison of the residual convergence from RE and original FM for eight-bead CG ENMs. (a) 1FD6: the RE is denoted in red curve (with the final residual 0.025) and FM is denoted in green curve (with the final residual 0.030); (b) 1GB1, the RE is denoted in red curve (with the final residual 0.032) and FM is denoted in green curve (with the final residual 0.039)..... 74

Figure 3.3 The relative difference between the spring constants generated by FM and RE methods for the CG-8 scheme. The percentage value on the y-axis is computed as the absolute value of the difference divided by the FM spring constant. The x-axis shows the spring index of the protein. The two proteins (a) and (b) are 1FD6 and 1GB1 respectively. 75

Figure 3.4 Comparison of the residual convergence from RE and FM for C α -level ENMs. (a) 2KDP: the RE is denoted by the red curve (with the final residual 0.21) and FM is denoted in green curve (with the residual 0.23). (b) 1HNN: the RE is denoted in red curve (with the final residual 0.13) and FM is denoted in green curve (with the final residual 0.15).....79

Figure 4.1 Chemical structure and atomistic structure of monomers. (a) PES. (b) PS. (c) EE2. (d) PVDF. (e) PA6-12. (f) testosterone..... 85

Figure 4.2 Root mean squared displacement of the complex structures, PES-EE2, PS-EE2 and PES- testosterone during equilibrated simulations.....69

Figure 4.3 PMF with increasing COM distances between monomers (PES, PVDF and PS) and EE2. (a) PES-EE2 with the binding free energy of -1.5 kcal/mol. (b) PS – EE2 with the binding free energy of -0.75 kcal/mol. (c) PVDF-EE2 with the binding free energy of

0.0 kcal/mol (no interaction). (d) PA6-12-EE2 with the binding free energy of -1.28 kcal/mol. (e) PES-testosterone with the binding free energy of -0.98 kcal/mol.....72

Figure 4.4 Chemical structures and atomic structure of EE2 and its analogs. (a) EE2-AR (EE2 without the p-bond in the benzene ring). (b) Original EE2. (c) EE2-OH (EE2 without the OH group). Dotted circles highlight the difference between EE2, EE2-OH and EE2-AR.....74

Figure 4.5 Potential of mean force (PMF) with increasing COM distances between the PES monomer and EE2 analogues (EE2, EE2-OH and EE2-AR). (a) PES-E2 with the binding free energy of -1.5 kcal/mol. (b) PES-EE2-AR (EE2 without the p-bond in the benzene ring) with the binding free energy of -0.62 kcal/mol. (c) PES-EE2-OH (EE2 without the OH group) with the binding free energy of -1.00 kcal/mol. (d) PES-EE2-CA (EE2 without the triple bond carbon) with the binding free energy of -1.07 kcal/mol.....76

Figure 5.1 Chemical structure of the organic solute molecules and model polymers: (a) EE2. (b) Testosterone. (c) PES. (d) PA612. (e) PVDF. (f) PYS. 99

Figure 5.2 SMD pulling snapshot of EE2 (green) crossing PES membrane (licorice blue) in z-direction. 100

Figure 5.3 Optimized models of polymer membranes in green, organic solute molecules in blue and water molecule in red. (a) PES-EE2. (b) PES-testosterone. (c) PVDF-EE2. (d) PVDF-testosterone. (e) PYS-EE2. (f) PA612-EE2. 102

Figure 5.4 (a) Snapshots of the membrane-solution interface simulation system at 100 ns showing the adsorption mechanism of four organic solute molecules of EE2 (red, pink,

orange and blue) on to the polymeric membrane PES. (b) Corresponding density profile of EE2 (the black line indicates the membrane interface). 107

Figure 5.5 (a) Snapshots of the membrane-solution interface simulation system at 100 ns showing the adsorption mechanism of four organic solute molecules of EE2 (red, pink, orange and blue) on to the polymeric membrane PVDF. (b) Corresponding density profile of EE2 (The black line indicates the membrane interface). 108

Figure 5.6 Potential of mean force (PMF) for EE2 and testosterone crossing PVDF membrane. (a) PVDF-EE2 with binding free energy of -7.79 ± 0.11 kcal/mol. (b) PVDF-testosterone with binding free energy of -6.59 ± 0.5 kcal/mol. 112

Figure 5.7 Potential of mean force (PMF) for EE2 and testosterone crossing PES membrane. (a) PES-EE2 with binding free energy of -15.59 ± 0.61 kcal/mol. (b) PES-testosterone with binding free energy of -13.08 ± 0.67 kcal/mol. 113

Figure 5.8 Potential of mean force (PMF) of EE2 and testosterone crossing the PYS and PA612 membranes. (a) PYS-EE2 with a binding free energy of -10.59 ± 0.11 kcal/mol. (b) PA612-EE2 with a binding free energy of -12.59 ± 0.27 kcal/mol. 115

Figure 5.9 Mean-squared displacements (MSDs) of (a). EE2 in PES and PVDF. (b) Testosterone in PES and PVDF. 117

Figure 5.10 Mean-square displacements (MSDs) of EE2 organic solute in polymeric membranes PES (red), PA612 (green) and PYS (blue). 120

Figure 5.11 A schematic of a hydrogen bond between the PES membrane and EE2 organic solute molecule. (a) PES (green) and EE2 (blue) simulation box with water

molecules. (b) Enlarged view of the formation of HB with a distance of 2.9 Å, water molecules are not shown for the sake of clarity..... 124

Figure 5.12 The hydrogen bond-time autocorrelation function for PES-EE2 (red) and PES-testosterone (green)..... 125

Figure 5.13 The number of hydrogen bonds formed between PES-EE2 and PA612-EE2 over the simulation time. 125

Figure 5.14 A schematic of a hydrogen bond between the PA612 membrane and the EE2 organic solute molecule. (a) PES (green) and EE2 (blue) simulation box with water molecule. (b) Enlarged view of the formation of HBs with a distance of 3.0 Å and 3.0 Å; water molecule are not shown for clarity..... 126

Figure 5.15 A snapshot of π - π interaction between PES and EE2. (a) Simulation box of EE2 in PES membrane. (b) Enlarged view of π - π interaction; the distance was computed from the centroid between the rings and is equal to 4.1 Å. 127

Figure 5.16 Illustration of π - π interaction between PS and EE2. (a) Simulation box of PS-EE2. (b) Enlarged view of π - π interaction; the distance was computed from the centroid and the distance between the rings is equal to 4.0 Å. 128

List of Tables

Table 1.1 Typical unit processes and operations used for wastewater treatment [71].	28
Table 3.1 NMR data set and comparison of the converged residuals between RE and FM) methods. The residual numbers are calculated for C α -level HENMs.	68
Table 3.2 NMR data set and comparison of the converged residuals between RE	69
Table 3.3 Mode Overlap between the NMR and HENM modes calculated by principal component analysis (PCA) for Relative Entropy and Fluctuation Matching of the 1FD6 protein.	77
Table 3.4 NMR data set and comparison of the converged residuals between RE and FM methods of middle size protein. The residual numbers are calculated for C α -level HENMs.	78
Table 3.5 Comparison of fluctuation-matching related methods.	81
Table 5.1 Composition of complex structures systems and simulation conditions.	104
Table 5.2 Binding free energies for PVDF, PA612, PES and PYS polymer with organic solute molecule (EE2 and Testosterone).	93
Table 5.3 Simulated diffusion coefficients (D) of organic small molecule (EE2 and Testosterone) in polymer chains.	102
Table 5.4 Interaction Energy (Electrostatic and van der Waals) for PVDF, PA612, PES and PS polymer with organic solute molecule (EE2 and Testosterone).	105

Abbreviations

AA-MD	all-atom Molecular Dynamics
AA-NMA	all-atom Normal Mode Analysis
AMBER	Assisted model building with energy refinement
ANM	Anisotropic network model
CG	Coarse-grained
CG-ENM	Coarse-grained elastic network models
COM	Center of mass
Cryo-EM	cryo-electron macroscopy
DFT	Density function theory
E1	Estrone
E2	17- β -estradiol
E3	Estriol
EDC	Endocrine disrupting compound
ED-CG	Essential dynamics CG

EE2	17- α ethinylestradiol
EE2-AR	EE2 without double bond (π -bond) in the benzene ring
EE2-CA	EE2 without carbon triple bond
EE2-OH	EE2 without OH
EM	Electron macroscopic
ENM	Elastic network model
ERP	Electron paramagnetic resonance
FDA	US Food and drug administration
FEM	Finite element method
FG	Fine grained
FM	Fluctuation matching
GGA	Gradient-corrected functional
GNM	Gaussian network model
GROMACS	Groningen machine for chemical simulations
HANM	Heterogeneous anisotropic elastic network model

HENM	Heterogeneous ENM
LAMMPS	Large-scale atomic molecular massively parallele simulator
MD	Molecular dynamics
MF	Microfiltration
MM	Molecular mechanics
MS-CG	Multiscale coarse-grained
MUCA	Multicanonical algorithm
NAMD	Nanoscale molecular dynamics
NF	Nano filtration
NMA	Normal mode analysis
NMA-FM	Normal mode analysis fluctuation matching
NMR	Nuclear magnetic resonance spectroscopy
PA6-12	Polyamide 6-12
PCA	Principal Component analysis
PES	Polyether sulfone

PFM	Progressive fluctuation matching
PMF	Potential of mean force
PS	Polysulfone
PVDF	Polyvinylidene fluoride
PYS	Polystyrene
RE	Relative entropy
RO	Reverse osmosis
UF	ultrafiltration
VDW	van der waals
WHAM	Weighted histogram analysis method

Abstract

A computational framework was used to expand the applicable length and time scale of molecular dynamics simulations (MD) by developing a multiscale CG parameterization approach for biomolecules, and an investigation on the adsorption mechanism of an endocrine disrupting compound onto polymer membranes.

For a multiscale model to parameterize CG-ENM force fields and structure-based model, we developed an improved fluctuation matching method, which is based on the relative entropy method, pioneered by Shell and colleagues. This method includes a non-negativity constraint and Newton Raphson's algorithm for iteration. Furthermore, this framework incorporates pairwise force constant correlations, which play an important role in the study of protein dynamics. Similar to the well-established fluctuation matching proposed by Lyman and coworkers, our heterogeneous ENM parameterization using the relative entropy method with non-negativity constraints, including a pairwise force constant correlation which plays an important role in the study of protein dynamics. Furthermore, fluctuation matching based on relative entropy method guarantees global optimum, with fast convergence.

For the adsorption of 17- α ethinylestradiol (EE2), we investigate the adsorption mechanism of EE2 from wastewater using MD simulations of monomers and polymer membranes level to validate experimentally observed results. Our findings from monomers level simulations with the small molecules EE2, and testosterone (polyether sulfone-EE2, polysulfone-EE2, polyvinylidene fluoride-EE2, polyether sulfone-

testosterone) and their analogous structures as well as polymer membranes simulations with the small molecules EE2 and testosterone (polyether sulfone-EE2 polyvinylidene fluoride-EE2 and polyether sulfone-testosterone) shows that the enhancement of binding affinity between PES and EE2 from both simulation set ups (monomer level and polymer membranes level) obtained from binding free energy, mean-squared displacement, diffusion coefficients, and insight atomistic interaction are attributed mainly to the π - π interaction and hydrogen bond. Our simulation findings are in agreement with the experimental results. Furthermore, to verify prediction veracity of our computational framework, additional systems are considered such as polyamide 612 at monomer and polymer membrane level and polystyrene membrane. In total, the binding free energies of all the systems considered in this study are in the order as; PES-EE2 > PES-testosterone > PS-EE2 > PA612-EE2 > PVDF-EE2 > PVDF-testosterone. To the best of our knowledge, these results suggest that the microporous PES hollow fiber membrane system is the most cost-effective approach that can effectively and efficiently remove EE2 at low concentrations.

1. General Introduction

The emergence of the computational approach to study life sciences and chemical systems has shed light on our understanding of various phenomena such as biological and chemical processes. Molecular dynamics (MD) simulations can be defined as computer based method to study the physical movement of biological and chemical systems such as atoms and molecules and it belong to the category of N-body simulation. Traditional MD simulations combined with new methods such as enhanced sampling, elastic network model (ENM) and free energy calculation continue to decipher critical biological and chemical processes.

The overwhelming literature and applications validated by ENM for the last two decades have provided a reliable and effective model to understand the intrinsic dynamics and conformational changes, characterize macromolecular flexibility, as well as analyze structural experimental data of biological and chemical processes.

Real applications of ENM ranges from simple structures, and huge macromolecular machines to complex biomacromolecular structures such as swollen motion of the entire virus capsids [3] or functional motion of the ribosome which was characterized using ENM [4]. Other real applications of ENM include construction of hybrid methods for systems rearrangements and simulations, as well as the combination of information from electron microscopy (EM) and atomic resolution technique to study the dynamics of complex structures [5]. Other interesting applications of ENM includes the study of

skeletal filaments [6-11], lipid membranes [12, 13], proteins [14], viral particles [15] and deoxyribonucleic acid [16].

MD simulations can be considered as a bridge between microscopic length and time scales and the macroscopic environment of experiments. As part of being widely used in life sciences as a powerful technique to study biomolecular structure, dynamics, and the mechanism of simple and complex biological processes, MD simulations have been used successfully to investigate chemical processes in various applications such as biotechnology, the photosynthetic apparatus [17], cellulose and second generation biofuels [18], nanoparticles [19], and polymer composite materials [20] in wastewater treatment for the removal of micropollutants [21, 22].

1.1 Multiscale models for biomolecular systems

Cells and their molecular components occupy a central position in research in life sciences, which is the subject of thousands of publications produced every year. The development of the field of structural Biology to determine the structure of biological macromolecules such as proteins and nucleic acids, has contributed to the understanding of the structure and by extension, the function of cells. Different experimental techniques have been used to determine the structure of biomacromolecular system such as nuclear magnetic resonance spectroscopy (NMR), Small angle scattering, macromolecular crystallography, proteolysis, electron paramagnetic resonance (ERP), cryo-electron microscopy (cryo-EM), multiangle light scattering, ultra-fast laser spectroscopy, mass spectrometry, dual-

polarization interferometry and circular dichroism. A computational approach offers another level of analysis of the mechanism and dynamics of biomacromolecular structure.

For the last three decades, computational studies of large biomacromolecular structures have been improved [23, 24]. The development of the field of scientific computing, increasing computer power, and high parallel computing infrastructures, algorithms, and methods have considerably contributed to increasing the timescale and length of MD simulations over microseconds. In the era of petascale and high throughput computing, molecular dynamics simulations software such as Nano scale molecular dynamics (NAMD) [25], Groningen machine for chemical simulations (GROMACS) [26], Large-scale atomic/molecular massively parallel simulator (LAMMPS) [27], Assisted model building with energy refinement (AMBER) [28] and other have developed cutting edge packages scalable to the development of hardware computing power. However, research on large-scale protein conformational changes require time and length scales beyond the reach of today's most powerful computers. This is especially true for large complex biomacromolecules, protein complexes, and complex biological processes such as protein folding and protein conformational changes. To address the problem of time and length scales, the coarse-grained (CG) model and enhance sampling methods were developed. The CG model is a widely used approach to expand the temporal and spatial scales of molecular simulations by reducing the degrees of freedom of the atomistic model [29] while enhance sampling method enhance phase space sampling [30]. In literature, there are many approaches to develop CG models. The elastic network models (ENM) [31-34] and Gō-like model [35, 36] are the most used types of CG models, with a common feature;

both types of models are designed to stabilize the experimental native structure of the biomolecule under analysis.

Protein conformational changes in the vicinity of their native structures contributes significantly in protein biological function [37, 38], such as enzyme catalysis and ligand binding. Different computational methods have been proposed to analyze biological functions of protein and to capture conformational changes from the experimentally solved native structures. The seminal work of Tirion (1996) [34] has pioneered the development of ENM. Tirion has shown that harmonic potential function and normal mode analysis can be used to study protein functional motion. Many coarse-grained elastic network models (CG-ENM) were proposed [39]. A CG-ENM is a network where the nodes are composed of beads and the edges are represented by the harmonic bond (spring). Each bead or interaction site consists of a group of atoms. The harmonic bonds in a standard ENM are based on Hookean springs with uniform or heterogeneous force constants. Typically, the position of a bead is at the $C\alpha$ position of each amino acids residue of the protein, Phosphate, sugar, and base for nucleotides [40]. Among different methods of CG-ENM, two groups have emerged: the Gaussian network model (GNM) [32-34] and the anisotropic network model (ANM) [31]. GNM is based on the assumption that the pairwise fluctuations in an ENM follow normal distributions in the vicinity of their equilibrium positions. The ANM is an improvement of the GNM method, which incorporates the anisotropic positional fluctuations [31, 38]. One of the advantages of GNM and ANM is that they consider the experimental structures as energy-minimized atomic structures. This energy-minimized structure can be used as an input in normal mode analysis (NMA). The NMA was pioneered by Suezaki and Gō (1975) [41] and

considers the protein as an elastic media. Traditionally, an all-atom molecular dynamics simulation (AA-MD) samples the protein conformational distribution near the native structure using the time integration of Newton's equation of motion. Based on the statistical mechanics principles, the thermodynamic properties of protein conformation changes can be accurately calculated from the sampled conformations, which reveal information into dynamic motions of biological systems. The all-atom normal mode analysis (AA-NMA) computes the mode of the protein based on the assumption that the protein is only undergoing harmonic vibrations. AA-NMA is unable to scale protein motion, which can be considered as a limitation, and AA-NMA is computationally faster than AA-MD. On the other hand, Mark Bathe [42] developed a computational model based on the finite element method (FEM). The FEM is a well-established numerical method applied in the field of engineering to model mechanical properties and stress distribution in complex geometry systems to compute NMA and mechanical properties of protein and their biomacromolecular assemblies (Figure 1.1) [43-45]. Sharing a similar framework with ENM, FEM provides many advantages compared to ATM-NMA such as the elimination of expensive energy minimization prior to NMA. FEM can directly use X-ray data of protein that contain unknown structures [46].

Figure 1.1 Four free vibration modes of F-actin (52-protomers, 0.14 μ m length) in planar deformation obtain from FEM method. The angular frequencies are 6.7×10^{-4} , 7.01×10^{-4} , 1.97×10^{-3} and 2.011×10^{-1} cycles/ps.

The heterogeneous ENM models (HENM) have been proposed to improve the accuracy of CG-ENM, which includes heterogeneous force constants between pairwise beads [47-52]. Compared with the simplified original ENM with a single force constant, HENM can be used to describe the intrinsically non-uniform molecular interactions and is expected to provide a more realistic view of protein dynamics. Nevertheless, the implementation of HENM raises the issue of how to efficiently parameterize the ENM spring constants. It should be noted that the heterogeneous force constants should be well parameterized; otherwise the accuracy and the efficiency of HENM will be compromised and the computational cost will be increased. Xia *et al.*, (2013) [53] developed iterative algorithms known as “Progressive Fluctuation Matching (PFM)” and normal mode fluctuation matching (NM-FM) [51] which are the improvements to the fluctuation matching used to parameterize heterogeneous force constants.

Fluctuation matching is a computational method to parameterize the force constants of the harmonic springs in a heterogeneous CG-ENM. This method is based on the distance fluctuations from an atomistic model using molecular dynamics simulations [7, 47]. The underlying principle of the method is to iteratively adjust the force constants to fit the target fluctuation. The target fluctuation (force constants) can be obtained from AA-MD or AA-NMA. After the original formulation of the method developed by Chu and Voth (2006) [7], Lyman *et al.*, (2008) [50] proposed an improved version of the original fluctuation matching algorithm. Other MD simulations studies using the fluctuation-matching algorithm to parameterized CG-ENM force fields can be found in literature [9, 54, 55]. The target fluctuations can be obtained from NMA instead of the conventional MD simulations [51]. Due to the fact that there exists some similarity between the harmonic potential in an ENM and the Lennard Jones (LJ) potential in a structure-based model, the fluctuation-matching algorithm was also used for developing heterogeneous structure-based models [56, 57]. As a recent extension of the original fluctuation approach, a modified fluctuation method was devised for fitting crystallographic B-factors instead of the force-field-generated fluctuations [53]. Due to the harmonic potential energy function, a single ENM is not suitable to describe the formation and break of protein native contacts.

The structure-based approach is often use to study protein folding mechanisms [35]. In the structure-based approach, the potential energy function is used to stabilize the protein native structure, and the approach is named the “Gō-like model”. The structure-based model is derived from the energy landscape theory [58]. The structure-based models can be constructed using Ca atoms as a scheme of CG potential, while atomistic implementations are also available in literature [59, 60]. The structure-based model adopts

the LJ type of potential to maintain the native contacts between amino acid residues, it also include bonded interaction, bond angles, dihedral. The structure-based model is an extension of the harmonic springs in the ENM. In the early development of structure-based models, the interaction strength of the native LJ interaction is identical, similar to the case of the original ENM. Nevertheless, non-uniform LJ interactions have been developed for CG structure-based models to reflect the heterogeneous physical interactions. Those non-uniform LJ potential parameters are derived from either the amino acid identity [61] or an atomistic force field [56, 57, 62].

Despite its success with the parameterization of CG force constants, HENM and structure-based models; the theoretic foundation of the fluctuation method was not clarified with the original development. Furthermore, the original heuristic iterative algorithm lacks an analysis on the convergence property.

Multiscale modeling is a well-established approach in molecular simulation for developing CG force fields targeting large scale macromolecular systems, on the basis of the information from small scale systems and fine-grained (e.g. atomistic) models [63]. Particularly, a lot of multiscale approaches, such as the force matching method [64, 65] and the relative entropy method [66, 67], aim to reproduce an atomistic configurationally probability distribution. Force matching and relative entropy can be considered as rigorous approaches in terms of statistical mechanics, linking the atomistic and CG force fields. The multiscale coarse-grained (MS-CG) models are used to expand applicable time scale, length scale, and system size of molecular simulation. The MS-CG model can be performed in two ways. The first approach is to define a CG scheme of the monomer units

and obtained CG fluctuation using available methods (fluctuation matching, force matching, relative entropy or Iterative Boltzmann). After this has been done, the parameterization MS-CG simulation can be performed.

The second approach of the MS-CG, which is more interesting in terms of expanding time-scale and system size, is when analyzing a large complex system, for example, skeletal filaments (F-actin, Microtubules) [7, 68]. A prime example is the atomistic model of a 13-monomer F-actin with both nucleotide ATP and ADP studied by Chu *et al.*, (2006) [48] composed of 575,000 atoms, including water molecules. Since F-actin (Figure 1.2) is composed of identical repeat units of G-actin, a good idea would be to first parameterize the interaction within an actin dimer, which includes both intra, and inter-monomer interactions, and then expand the interaction to the whole fiber. After the parameterization of all F-actin, then MS-CG simulation can then be performed.

Figure 1.2 Atomistic model of F-actin from Holmes *et al.*, (PDB ID: 1J6Z) [1] constructed from a single G-actin using the multiscale approach. Each domain is represented by one bead labeled as CG1, CG2, CG3 and CG4. (a) 13 units F-actin from the Holmes model (b) single actin monomer, (c) G-actin coarse-grained by domain, binding a nucleotide ADP.

There are some obvious advantages derived from the interpretation of the fluctuation matching approach using the relative entropy scheme. First of all, the relative entropy approach provides a solid physical foundation for the intuitively developed formulation of fluctuation matching. Additionally, it is revealed that the iterative fluctuation algorithm can be considered as an approximation that neglects the inter-pair correlations, which suggests a possible remedy for the occasional un-converged cases of proteins. Finally, by using the relative entropy formulation, a more robust algorithm that incorporates the inter-

pair correlations can be designed, resulting in better convergence properties for some CG protein models.

1.2 Wastewater treatment

During the last century, the world's population quadrupled, while the use of water has increased faster than the world's population by a factor of seven [69]. The challenge of water scarcity will continue as the world becomes wealthier, and thirstier and as the developing countries expand and diversify their economies [70]. The World Water Council predicted that by 2030; approximately 3.9 billion of the world's population will live in a "water scarce" area (Figure 1.3). In order to meet the world's growing need for clean water, efficient and affordable water treatment technology is required. The varied use of water has contributed to the deterioration of water quality by introducing chemical or biological contaminants. Over the last century, considerable research has been conducted to propose a solution on the challenging problem of water treatment. One of the proposed methods is wastewater reclamation and reuse using membrane technology. The wastewater is known as sewage, which includes liquid waste discharged by households, commercial and industrial facilities. Wastewater recycle methods have been applied to remove solids, organics pollutants (or micropollutants) and nutrient from the sewage [71].

Figure 1.3 Water stress map by country. Figure adapted from ref. (<http://www.wri.org/>)

The role of membrane science and technology in developing a sustainable environment has contributed to the emergence of membrane separation technology as an innovation mechanism for sewage water treatment technology in the global water market [71, 72]. Different separation techniques have been proposed and implemented for the removal of solids, organic pollutants (or micropollutants) and nutrients which can be categorized as, physical, biological and chemical processes (Table 1.1) [71] [73-75].

Table 1.1 Typical unit processes and operations used for wastewater treatment [71].

Types	Physical	Chemical	Biological
Unit Process	Screening	Neutralization	Activated sludge
	Comminution	Chemical coagulation	Trickling filters
	Grinding/maceration	Chemical precipitation	Rotating biological
	Flow equalization	Chemical stabilization	Aerated lagoon
	Mixing	Adsorption	Pond stabilization
	Flocculation	Ion exchange	Sequencing batch
	Sedimentation	Dechlorination	Anaerobic
	Flotation	Disinfection	Up-flow anaerobic
	Aeration	Oxidation/advanced oxidation	Moving-bed biofilm
	Air stripping		Biological nutrient
	Granular-medium filtration		Membrane bioreactors
	Electrodeposition		
	Membrane filtration (Microfiltration (MF), Ultrafiltration (UF), Nano filtration (NF), and remove osmosis (RO))		

1.2.1 Micropollutants

Micropollutants are natural contaminants and anthropogenic that are persistent, as well as bioactive compound such as pesticides, pharmaceutical, plasticizers and other categories such as antibiotics, hormones and endocrine disrupting compound (EDC). Micropollutants are not biodegradable and thus are difficult to remove from wastewater using traditional wastewater treatment technology. The presence of micropollutants in wastewater has raised a global concern that seriously threatens environmental and human health [76]. This can be due to their interference with the physiological function of hormones such as fertility, growth and development. Removal of micropollutants in wastewater is well documented in the literature [73, 77-84].

EDC are exogenous chemicals responsible for adverse health effects in intact organism [85] and have a significant impact on the mechanism and the function of the endocrine system on human health and wildlife, e.g., fishes, mammals, birds, reptiles, amphibians, and invertebrates.[86-89]. EDC may be responsible for (i) reducing the effect of endogenous hormones; (ii) perturbing the synthesis and metabolism of endogenous hormones; (iii) perturbing the synthesis of hormone receptors [90-94]. EDC include a wide range of molecules such as, chlorinated pesticides, herbicides, phthalates, alkylphenols, and steroid hormones, including natural and synthetic estrogens. In a comprehensive review, Frye *et al.*, [95] reported the sources, consequences and actions of EDC. Natural estrogens such as estrone (E1), 17 β -estradiol (E2), and estriol (E3), the synthetic hormone 17 α -ethinylestradiol (EE2), as well as progesterone, and testosterone are abundant in surface water sources [96]. EE2 is used in human oral contraceptives and

hormone therapy replacement such as osteoporosis, menstrual disorders, and prostate cancer in low ng L⁻¹ range [97, 98]. The US Food and Drug administration (FDA) and the National Center for Toxicological Research published a report which classified EE2 as a major prominent estrogenic chemical in EDC [99, 100]. The European Commission's Joint Research Centre in 2015 also listed EE2 in the watch list as a major potentially polluting compound in the aquatic environment [101]. For example Flores-Valverde *et al.*, (2010) [102] shows that EE2 at low concentration (1-29 ng/l) may affect the steroid metabolome in fish (*Rutilus rutilus*) consequently, disrupting steroid biosynthesis. Coe *et al.*, (2008) [103] reported that a small concentration of EE2 (2-10 ng/L) is able to perturb the reproductive hierarchies and success rate of mating in a group of spawning fish. The principal source of EE2 in the aquatic ecosystem is a result of the incomplete removal of EE2 in wastewater treatment [104-108]. It should be noted that on average 70% of pharmaceutical residues in wastewater are from household use, which includes EE2. Also approximately 90% of consumed prescription drugs end up in wastewater.

1.2.2 Polymer membranes

A polymer membrane is a selective permeable barrier used in membrane filtration able to block certain substance while others can pass through. The selectivity depend on the membrane pore sizes, chemical and biological interactions, and thickness. Membranes can be categorized into synthetic membranes and biological membranes [109]. Membrane technology is used in water treatment, water reclamation, and removal of micropollutants. Based on the pore sizes, the membrane technology can be characterized as nano filtration (NF) or reverse osmosis (RO), which represents the prominent membrane, processes technology for the removal of micropollutants. Microfiltration (MF) and ultrafiltration (UF) are other groups of membrane technology and are unable to filter small molecules.

1.3 Outline of the research work

With the increasing timescale reach of computational resources, MD simulations have proven over the last decade to be a reliable technique for understanding the intrinsic dynamics of biomolecules and chemical systems. In this work, we developed methods to study biological and chemical systems using MD simulations and quantum calculations. In the first part of this thesis (chapter 3), a novel multiscale algorithm was developed to improve fluctuation matching method propose by Lyman *et al.*, (2008) [110] to parameterize CG-ENM force fields and structure-based models. The algorithm is based on a relative entropy with a non-negative constraint following the work of Shell *et al.*, (2009) [111]. This guaranteed a global optimum and thus enhanced optimization convergence. Similar to fluctuation matching, the proposed algorithms incorporate spring constant

pairwise correlations. The parameterized CG force fields are obtained by minimizing the relative entropy function of force constants. It should be noted that the force constants should be strictly positive. In order to satisfy aforementioned force constant requirements, a non-negative constraint was incorporated into the algorithm. To validate the improved fluctuation matching method, a comparative analysis was conducted with the established fluctuation matching proposed by Lyman *et al.*, (2008) [110], which is an improvement on the initial algorithm proposed by Chu *et al.*, (2006) [8]. Furthermore, a comparative framework analysis of different variants of well-established fluctuation matching methods and relative entropy were presented.

In the second part (chapter 4), MD simulations were conducted to study the adsorption mechanism of an EDC (EE2 or testosterone) onto polymer membranes at monomer scale. The removal of EE2 from wastewater using polymer membranes as filters is an important step in water reclamation and reuse. In literature, a lot of studies have been conducted to investigate the interaction between estrogens (EE2) and different polymer membranes [21, 74]. In this work we analyzed the atomistic interactions at the monomer level between EE2, testosterone, and five types of polymers namely, polyether sulfone (PES), polysulfone (PS), polyamide 6-12 (PA6-12) and polyvinylidene fluoride (PVDF) at the monomer level. Their interactions were analyzed by calculating the binding free energy between the small organic molecules (EE2, testosterone) and the monomers. The results show that PES has strong interactions with EE2, meaning that PES is able to retain more EE2 compared to PS, PA6-12, and PVDF. The results are consistent with the experimental results found in the literature [22, 112]. In order to determine the major contributors to the interaction, different analogous structures of EE2 were constructed

namely, EE2 without the OH group (EE2-OH), EE2 without the double bond (π -bond) in the benzene ring (EE2-AR), and EE2 without the carbon triple bond (EE2-CA). The binding free energy between the different analogs shows that in the simulation with EE2-AR, the potential of mean force and the binding free energy decrease, compared to the original EE2 structures which indicate high a contribution of π - π bond in the PES-EE2 interaction.

In the third part (chapter 5), a comprehensive molecular dynamics method was developed to study the adsorption of EE2 onto polymer membranes for the removal of micropollutants. To the best of our knowledge, this work outlines for the first time, the adsorption mechanism at the atomistic level between EE2 and polymer membranes. The mechanism presented in this work can be applicable to various micropollutants depending on chemical and atomistic properties. The model was tested on EE2 and testosterone interacting with polymer membranes. It should be noted that different interactions mechanisms should be considered in order to explain qualitative and quantitative atomistic interactions between micropollutants (EE2, testosterone) and polymer membranes. In recent experiments [22, 112] it was observed that microporous PES hollow fiber has high efficiency to retain aromatic hormones especially EE2 from aqueous solution. A computational study is performed to understand and compare the interaction of EE2 with polymer membranes, namely, PES, PVDF, PA6-12, and polystyrene (PYS). The choice of PES and PVDF polymer membranes in this work is mainly to provide molecular insights to our recent experimental observations on the adsorption of EE2. The equilibrium molecular dynamics simulations results based on the membrane-solution interface system show that the interaction of EE2 with PES is specific and strong, whereas

the interaction is non-specific and weak for PVDF. The potential of mean forces (PMFs) and binding free energies are also consistent with the interaction behavior found in experiments. To test the proposed computational framework, the systems of testosterone interacting with PES and PVDF are also investigated as a control. Using similar calculations, the interactions of testosterone with PES and PVDF observed in experiments are also explained. In total, the binding free energies of all the systems considered in this study are in the following order: PES-EE2>PES-testosterone>PA6-12-EE2>PYS-EE2>PVDF-EE2>PVDF-testosterone. The diffusion coefficient results show that the mobility of organic solute in the polymer membranes has a same trend as their binding free energies, indicating that a strong interaction with the membrane means the mobility is small. The non-bonded short-range interaction energies are consistent with the binding free energies and mobility. The formation of hydrogen bonds and π - π interactions are observed between the EE2 and polymer membranes. This molecular dynamics simulations study provides molecular insights into the experimental observations and helps as a computational methodology to screen the polymer membranes materials for EE2 removal from wastewater and can be extended to the other types of micropollutants.

2. General Methodologies

2.1 Molecular simulations

2.1.1 Quantum mechanics

A quantum mechanics approach is used to describe the theory of nature by describing chemical reactions, electronic interactions, low energies of atoms, geometry, charge transfer and electronic excitation of small scale systems. In quantum mechanics, the system is regulated by the wave functions. The properties of the particle are defined as a wave. Mathematically, the equation for finding the wave function of any given system is calculated based on Schrödinger wave equation,

$$H\Psi = E\Psi \quad (2.1)$$

where H is the Hamiltonian operator of the system which represents the potential and kinetic energies, E is the total energy of a state and Ψ is a wave function.

Quantum mechanics calculations were used in chapters 4 and 5 to obtain atomic charge and structure optimizations using the Density function theory (DFT) of chemical structures namely, small molecule organics and polymer structures, prior to MD simulations.

2.1.2 Molecular mechanics and molecular dynamics

2.1.2.1 Molecular mechanics

Molecular mechanics (MM) study the molecule system based on classical mechanics. In MM, the electron degrees of freedom is included, the mass and the atomic charge are located in the nuclei. The accuracy of the MM is based on the mathematical description and parameterization of the potential energy function or force fields. The potential energy function is subject to certain assumptions. For example the Born-Oppenheimer approximation makes it possible to write the energy as a function of nuclear coordinates [113]. The interactions between atoms are categorized into bonded and non-bonded interactions. The bonded interactions are composed of stretching; bending and rotating between covalently bonded atoms. The non-bonded interactions are composed of Van der Waals (VDW) and electrostatic interactions.

The potential energy function in most of the widely used force fields packages is composed of the five terms as described above [114].

$$\begin{aligned}
V(\mathbf{r}^N) = & \sum_{bonds} \frac{k_{bond}}{2} (l_i - l_{i,0})^2 \\
& + \sum_{angles} \frac{k_{angle}}{2} (\theta_i - \theta_{i,0})^2 \\
& + \sum_{torsion} \frac{V_n}{2} (1 + \cos(n\omega - \gamma)) \\
& + \sum_{i < j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \\
& + \sum_{i < j} \sum \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \quad (2.2)
\end{aligned}$$

where, $V(\mathbf{r}^N)$ represents the potential energy, which depends on the location ($\mathbf{r}$) of N atoms, l_0 , θ_0 and γ are the equilibrium constants, while k_{bond} , k_{angle} and V_n represent the force constants. Other parameters such as the constant, ϵ_{ij} and σ in VDW interaction, as well as ϵ_0 and q in electrostatic interaction, are empirical values set in the force fields. These values are obtained from experiments or from theoretical quantum calculations. The role of each term can be summarized as follows: The first term is for the pairwise interactions between atoms. The second term represents the sum of all valence angles in the system. The third term represents the torsional potential. The fourth term represents the non-bonded interactions and the fifth term represents the coulombic interactions.

2.1.2.2 Molecular dynamics simulation

Under physiological conditions, the behavior of biological systems (proteins, nucleic acids) can be described as a dynamics system not static. Therefore, the study of such systems requires the incorporation of a time dependence function. MD simulations consider Newton's equation of motion to determine time dependence, which is suitable to study biological systems. In most of MD approaches, the atoms are treated classically. Let's consider a simple motion of atomic systems. The force $\mathbf{F}_i$ is applied on the atom i of mass m_i to generate an acceleration of $\mathbf{a}_i$, Newton's equation of motion is

$$\mathbf{F}_i = m_i \mathbf{a}_i \quad (2.3)$$

$$\mathbf{a}_i = \frac{d^2 \mathbf{r}_i}{dt^2} \quad (2.4)$$

The force can be expressed as function of the slope of the potential function (U) as well.

$$\mathbf{F}_i = -\Delta_i U = -\frac{d}{dr_i} U \quad (2.5)$$

Combining eq (2.3), (2.4) and (2.5) gives

$$-\frac{dU}{dr_i} = m_i \frac{d^2 \mathbf{r}_i}{dt^2} \quad (2.6)$$

The potential energy function for a system made up of ‘N’ number of particles in three dimensions and dependent on the Cartesian coordinates systems (x, y and z) of all atoms is:

$$U = \sum_{i=1}^N \mathbf{F}_i(x, y, z) \quad (2.7)$$

The potential energy function U is computed as the sum of energy terms, which contribute to the system such as bond vibration, bond angles, torsion angles, VDW and electrostatic interactions. The definition and contribution of each term in the potential energy function is similar has to what has previously been described in Equation (2.2) of the molecular mechanics section [section 2.1.2.1].

$$U = E_{bond} + E_{angle} + E_{torsion} + E_{vdW} + E_{ele} \quad (2.8)$$

The acceleration can be easily determined after the magnitude force on each single atom is obtained. When Newtown’s equation is integrated, the trajectory is generated. It should be noted that the trajectory contains microscopic properties of the atoms during the simulation. The time-average of the system from a series of frames with respect to the simulation can then be extracted from the trajectory.

2.1.2.3 Elastic network and structure-based models

The potential energy function of the ENM model is represented using a quadratic function based on the minimum energy conformation [31, 34, 38, 39, 115]:

$$V = \frac{1}{2} \sum_{ij} \gamma_{ij} (r_{ij} - r_{ij}^0)(r_{ij} - r_{ij}^0) = \frac{1}{2} \sum_{ij} \gamma_{ij} (\Delta r_{ij})^2, \quad (2.9)$$

where γ_{ij} is the spring constant for the pairwise residue (i, j) , r_{ij} is the pairwise distance of the residue pair (i, j) and r_{ij}^0 is their equilibrium distance, generally obtained from the experimental conformation.

The potential energy function of the structure-based models introduced by Clementi *et al.*, (2000) [35] is presented as follows, and is composed of three terms namely, local, contact, and excluded volume interactions [56, 116, 117].

$$V_{protein} = V_{local} + V_{contact} + V_{excluded}. \quad (2.10)$$

V_{local} is the summation of bond length, bonds angle, and dihedral angle potentials, namely,

$$\begin{aligned} V_{local} = & \sum_{bond} k_b (r_{bond} - r_{bond}^0)^2 + \sum_{angle} k_\theta (\theta_{angle} - \theta_{angle}^0)^2 \\ & + \sum_{dihedral} \left\{ k_\phi [1 - \cos(\phi_{dih} - \phi_{dih}^0)] + \frac{1}{2} k_\phi [1 - \cos 3(\phi_{dih} - \phi_{dih}^0)] \right\}, \end{aligned} \quad (2.11)$$

where r_{bond} , θ_{angle} and ϕ_{dih} describes length, angle and dihedral of virtual bond respectively. The superscript “0” represent the values in the native structure (generally taken from the Protein Data Bank <http://www.rcsb.org>).

$V_{contact}$ is the non-bonded potential that describes the short-range interactions of native contact pairs, which are separated by more than three residues in sequence and the

distance between C α atoms is within the define cutoff. Generally a contact is represented by a LJ potential form $V_{contact}$ is described as

$$V_{contact(ENM)} = \sum_{i < j - 3}^{native} \epsilon \left[5 \left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - 6 \left(\frac{r_{ij}^0}{r_{ij}} \right)^{10} \right], \quad (2.12)$$

where r_{ij} is the pairwise beads distance between i and j , while ϵ represents the force constant of the interactions and r_{ij}^0 is a native contact distance. The softer “10-12” modified LJ potential is used here for CG structure-based approach representations [118]. The summation is done for all the native contact pairs.

$V_{excluded}$ is the excluded volume potential associated to non-native pairs, composed of atoms which do not form native pairs, and interact exclusively by repulsion:

$$V_{excluded} = \sum_{i < j - 3}^{non-native} \epsilon_{ex} \left(\frac{d}{r_{ij}} \right)^{12}, \quad (2.13)$$

where d represents the radius of the beads and ϵ_{ex} is the corresponding force constant. The excluded term is incorporated in order to prevent clashes between non-native pairs.

2.1.3 Free energy calculation and umbrella sampling

The time-average of the system from a series of frames with respect to simulation can then be extracted from the trajectory. The length of the trajectory can be in the order of nanoseconds to milliseconds long. The primary objective of MD simulations is to get

significant microstate structures at an identical microscopic or thermodynamic state from simulations, which are as close to observable experimental value as possible. It should be noted that in statistic mechanics, the ensemble averages correspond to experimental observable averages and take place under microscopic condition as described below;

$$\langle A \rangle_{ensemble} = \iint dp^N dr^N A(\mathbf{p}^N, \mathbf{r}^N) p(\mathbf{p}^N, \mathbf{r}^N) \quad (2.14)$$

where $A(\mathbf{p}^N, \mathbf{r}^N)$ represents the observable as a function of the moments (p) and the position (r) of the system, $p(\mathbf{p}^N, \mathbf{r}^N)$ represents the probability density function of the ensemble and is computed using a partition function. In order to compute ensemble averages from MD simulations, all possible states corresponding to the specific thermodynamic constraints have to be sampled. The time average is calculated using the following formula;

$$\langle A \rangle_{time} = \lim_{x \rightarrow \infty} \frac{1}{t} \int_{t=0}^t A(\mathbf{p}^N(t), \mathbf{r}^N(t)) dt \approx \frac{1}{M} A(\mathbf{p}^N, \mathbf{r}^N) \quad (2.15)$$

where t represents simulation time, and M is the time step number of the simulation. The time average obtained from MD simulations and the ensemble average from observed experiments can be considered equal under Ergodic hypothesis. An attempt to sample all possible states using MD simulations will result in an infinite simulation time, which is impossible. Practically, in order to obtain experimentally acceptable information, MD simulations should be able to sample enough representative conformations. However, most biological processes occur on a length and time scale beyond the reach of MD simulations. The capacity to explore a large conformational space with a length and time

scale over nanoseconds for biomolecular systems that contain a lot of degrees of freedom, and huge number of local minimum energy is a limitation of traditional MD simulations. In order to mimic observed experimental value; most MD simulations are simulated at room temperatures (~ 300 K) or at physiological condition (~ 310 K). Simulation at low temperature has a tendency to be trapped in one of the local minimum energy state. This low temperature simulation is unable to sample the conformational space, which is located at a high-energy barrier from the sampled configuration as illustrated in Figure 2.1. Thus, it is difficult to get an accurate canonical distribution at low temperature.

Figure 2.1 Free energy surface. (a) Scheme of the multidimensional energy landscape of fluorescent protein. The red arrow represents unfolding pathway. (b) Illustration of the energy landscape along the unfolding pathway onto one reaction coordinates. The experiment start from minimum 1 is able to overcome lower energy barrier B and sample in minimum 2, but it is challenging for the system to overcome the energy barrier C and D. Figure is adapted from [2].

To overcome the multiple minima problem, and the inaccuracy of thermodynamic averages, a generalized ensemble algorithm was developed. The well-established generalized ensemble methods used in MD simulations are namely the multicanonical algorithm (MUCA), the simulated tempering, and the replicas exchange method (REMD) [119, 120]. The underlying principle of the generalized ensemble algorithm is that each state of the system is weighted using a non-Boltzmann probability weight factor in order to perform a random walk in the potential energy space. The purpose of the random walk is to help the simulation to overcome the energy barrier and be able to sample a large conformational space [121]. The MUCA is considered as one of the best-established generalized ensemble methods and is a robust realization of the umbrella-sampling algorithm. Since the umbrella-sampling algorithm was used in this work, a summary of the algorithm is presented. Further details of the umbrella sampling algorithm can be found in the Kästner (2011) [122] review paper.

The umbrella sampling algorithm was pioneered by Torrie and Valleau (1977) [123]. In this method, the multiple energy minimum problems are overcome by modifying the potential energy function in order to be able to sample high-energy states. This method is suitable for conducting simulations along different defined reaction coordinate ζ 's and can be applied in both MD simulations and Monte Carlo simulations. In this work, the reaction coordinate to perform umbrella sampling was obtained from steered MD simulations [124]. The structure obtained from each reaction coordinate is considered as a starting configuration for a series of umbrella sampling windows, which are run independently as illustrated in Figure 2.2.

Figure 2.2 Principle of umbrella sampling simulation. The red dot represents protein **A** that is fixed and dot blue represents protein **B** that is pulled away by its center of mass to generate different initial configurations. (a) Pulling simulation-using SMD to obtain a series of starting configurations for umbrella sampling windows. (b) The starting independent simulation (blue dot) obtained from SMD, and the black line represents the bias umbrella potential. (c) The result is illustrated as a histogram of configuration, with neighboring windows overlapping in order to have a continuous energy function. Figure is adapted from (www.bevanlab.biochem.vt.edu).

The umbrella sampling algorithm can be summarized as follows;

$$u'(r) = u(r) + W(r) \quad (2.16)$$

where $u'(r)$ is the modified potential function from $u(r)$, $W(r)$ is a quadratic weighted function;

$$W(r) = k(r - r_0)^2 \quad (2.17)$$

where r_0 is the equilibrium state. Configuration state away from the equilibrium state will have large weighted function; therefore the energy will be biased. The extraction of the corresponding Boltzmann average from the non-Boltzmann distribution is obtained using the formula below;

$$\langle A \rangle = \frac{\langle A(\{\vec{R}_i\}) \exp[+ W(r) / k_B T] \rangle_W}{\langle \exp[W(r) / k_B T] \rangle_W} \quad (2.18)$$

The subscript W shows that the average is related to the probability $P_W(\{\vec{R}_i\})$, obtained from the modified energy function $u'(r)$. The weighted histogram analysis method (WHAM) [125] was used to compute the free energy profiles (PMFs) from umbrella sampling trajectories. It is possible to perform many independent umbrella-sampling simulations with different initial velocities in order to increase sampling and obtain accurate PMFs standard error.

2.2 Multiscale parameterization

In the literature, there are many methods to parameterize multiscale models such as the Iterative Boltzmann Inversion [29, 126], force matching [47, 64], fluctuation matching [47, 50, 51], and the relative entropy method [66, 67]. Prior to the parameterization of CG force fields, the definition of the CG scheme is an important step. There is no predefined definition of the CG scheme; the choice depends on the experience or properties under study. There are many methods that can be used to set CG scheme. The default is to consider C- α atom as a CG scheme. MD simulations can be used to determine the partition of the CG beads by using the essential dynamics CG (ED-CG) method [127]. In a review, Saunders *et al.*, (2012) [128] categorized multiscale models into two major groups, the mapping methods to parameterized CG force fields and the bridging methods capable of connecting different scales during simulations.

The parameterization of CG force fields using fluctuation matching MD simulations, fluctuation matching based on normal modes analysis, fluctuation matching to fit experimental B-factors, and the relative entropy method on CG-ENM is presented.

2.2.1 Heterogeneous elastic network models and structure-based models from fluctuation matching targeted at atomistic molecular dynamics simulations

Fluctuation matching (FM) is a well-established method used to obtain a CG force field from a fine-grained model (e.g., an atomistic force field) [47, 57]. In most of the cases, FM has a predefined energy function form and the strength of each pairwise interaction is determined by matching the fluctuations from an atomistic simulation. Having originated

from the idea of quasi-harmonic analysis [37], the FM method approximates CG pair interactions by simple harmonic potentials. The fluctuation matching method was pioneered by Chu et al., (2006) [8] can be summarized as below;

Let 's EIC be the effective internal coordinate, which represent CG bead interactions. $\langle EIC \rangle$ is the equilibrium quantity of EIC and $\langle \delta EIC^2 \rangle$ the fluctuation on the vicinity of mean.

$$\langle \delta EIC_i^2 \rangle = \langle (EIC_i - \langle EIC_i \rangle)^2 \rangle \quad (2.19)$$

where EIC is a bond distance coordinate (bond angle or dihedral angle). EIC fluctuation can be derived from ATM-MD or CG MD simulations. For CG-ENM, the fluctuations of EIC represent the fluctuation of spring distances. Generally, the fluctuating bond distance calculated exclusively represent its own distance bond but also included the force constant of other pairwise bonds in the ENM. However, in the FM method, inter-pairwise correlation is not considered and FM assumes that each force constant considers its own pairwise fluctuation. Substituting the EIC in Equation (2.20) by the pairwise distance r , the force constant is computed using the following recursive equation;

$$k_{ij}^{n+1} = k_{ij}^n + \alpha \left(\langle \Delta r_{ij}^2 \rangle_{CG}^n - \langle \Delta r_{ij}^2 \rangle_{AA} \right) \quad (2.20)$$

Where, k_{ij} represent pairwise (i,j) interaction the force constant in ENM with r_{ij} represent pair distance. The subscript AA mean the target fluctuation is obtained from AA-MD and

n represents iteration number. The value of the scaling factor step size, α , is in range of 0-1.

Chu *et al.*, (2006) [48] pioneered the fluctuation matching method to parameterize a CG force field. The underlying principal is to align the mean value and the fluctuation of the effective internal coordinate used to define a CG model consistent with the values extracted from an atomistic MD simulations. The effective fluctuation from the CG force field is calculated from NMA for computational efficiency. The FM method was tested on 13 units of actin filament (F-actin) illustrated on Figure 1.2. CG models of F-ATP and F-ADP were constructed with parameters obtained from atomistic MD simulations using the FM method. The results revealed that the mechanical properties such as the persistence length of F-actin depend on the assigned inter-monomer cutoff distances. Coupling fluctuation matching with ENM contributes to reasonable temporal and spatial scales for the simulation of the F-actin filament.

Based on the fluctuation matching approach pioneered by Chu *et al.*, (2006) [48], Lyman *et al.*, (2008) [50] developed a variant of the fluctuation matching method to parameterize HENM of proteins. Lyman's algorithm reproduces the fluctuations observed from atomistic simulations. Similarly, Lyman's algorithm fits pair distance fluctuations from an atomistic molecular simulation trajectory. The optimization of force constants used in parameterization is based on the inverse of force constant ($1/k$) to improve convergence efficiency:

$$\frac{1}{k_{ij}^{n+1}} = \frac{1}{k_{ij}^n} - \alpha (\langle \Delta r_{ij}^2 \rangle_{CG}^n - \langle \Delta r_{ij}^2 \rangle_{AA}) \quad (2.21)$$

Moreover, the model was combined with the ED-CG method [127] to define CG force fields under various CG schemes. In the original work of Lyman *et al.*, (2008) this method was successfully tested on different systems such as the ADP-bound actin filament (PDB ID: 1J6Z) [1], myoglobin and the N-terminal Bin/amphiphysin/Rvs (N-BAR) domain of amphiphysin. Other biomolecules studies using the Lyman's algorithm to parameterize CG force fields have also been reported in the literature [9, 54, 129, 130].

Takada and coworkers (2012) [56] used fluctuation matching to parameterize a structure-based model. The results revealed that the CG model is able to reproduce correctly, the fluctuations from an atomistic model of the native structure. Different systems were used to test the method such as RNA and the RNA-ribosome complex. For the RNA related molecular systems, the parameters of the structure-based model were obtained by matching the atomistic fluctuations of sixteen different RNA molecules and ten RNA-protein complexes. It was demonstrated that the CG parameters obtained using fluctuation matching reproduced the native fluctuations from RNA and RNA-protein complexes. Furthermore, a model ribosome system with very considerable amplitudes of fluctuations around the equilibrium structure was tested.

2.2.2 Fluctuation matching targeted at fine-grained normal modes analysis

The NMA method is used to study the vibrational motions of a protein structure around its equilibrium position. All-atom NMA (AA-NMA) is computationally expensive for large biomolecular structures because it is based on all atom force fields. Thus, expensive energy minimization is required prior to AA-NMA and the degree of freedom in the

system is often large. Conversely, for CG-ENM, there is no requirement of energy minimization since the experimental structure is considered to be at the energy minima. Furthermore, the CG model significantly reduces the degrees of freedom, thus reducing the computational time used for NMA. The ENM method is used to study elastic properties, conformational changes and protein folding of biomolecules and large bio-filaments [38, 131-133].

Fine-grained (FG) or CG fluctuations derived from NMA can also be considered as fitting targets in the fluctuation matching approach in the same way fluctuation matching targeting MD-based fluctuations is considered. Xia *et al.*, (2012) [51], pioneered a method known as “normal mode analysis-based fluctuation matching” (NMA-FM) to parameterize CG-ENM of biological systems. The fluctuation targets derived from the FG model (e.g., an atomistic force field) and the CG model are obtained from NMA. The NMA-FM uses NMA to generate FG fluctuation targets, reduce computational time, and converge fast especially for large biomolecules, which can be considered advantageous compared to the fluctuation targets derived from FG-MD simulations. Furthermore, NMA-FM is scalable to the CG scheme, such as using a conventional residue-level ENM as input to parameterize an even more coarse-grained model. These rigorous CG models would be helpful in modeling mechanical properties and conformational fluctuations of large complex protein systems, where each protein domain is CG into one or a few beads. The parameterized force constants in the CG model are obtained from the effective interaction strengths among the CG site pairs. Therefore, NMA-FM can be implemented to analyze the effective pair interactions between two parts of a protein. Below is the

summary of the numerical algorithm for updating the spring constant of the heterogeneous CG-ENM force constants.

- (1) Generate CG-ENM by connecting all pairs of CG sites using initial force constant

$$k_{ij}.$$

- (2) Perform an NMA, and map the fluctuations on the bonds between every pair of CG sites to calculate $\Delta r_{ij,CG}^2$.

- (3) Update all spring constants instantly using the Equation (2.22) until they are converged:

$$\frac{1}{k_{ij}^{n+1}} = \frac{1}{k_{ij}^n} - \varepsilon \left(\Delta r_{ij,FG}^2 - \Delta r_{ij,CG}^2 \right) \quad (2.22)$$

where n represents the interaction number, ε represent a scaling factor that control the size of the steps in the optimization of the springs constants, $\Delta r_{ij,CG}^2$ represent the NMA bond fluctuations and $\Delta r_{ij,FG}^2$ represents a target fluctuation obtained from atomistic NMA. The iterative algorithm converges when the threshold σ is below 10^{-3} (It should be noted that the convergence threshold value depends on accuracy level), σ is defined by

$$\sigma = \left(\left| \Delta r_{ij,FG} \right| - \left| \Delta r_{ij,CG} \right| / \left(\left| \Delta r_{ij,FG} \right| \right) \right).$$

The NMA-FM approach was first tested on two systems namely lysozyme (PDB ID: 2LZM) [134] and adenylate kinase (PDB ID: 1AKE) [135] to generate seven-bead and six-bead ENMs, respectively from C α -based ENMs. Figure 2.3 illustrates rigorous CG-ENM of lysozyme and adenylate kinase, which are able to reproduce the positional

fluctuations from the $C\alpha$ -based ENMs. Furthermore, each CG model reproduces the first two characteristic low-frequency modes from the corresponding FG model. This illustrates the important slow dynamic motions of the proteins, which can be correctly derived from minimalistic CG-ENM. The test set was extended to other six-bead ENM of the S2 subdomain of myosin (PDB ID: 1NKN) [136] from a $C\alpha$ -based ENM, and the corresponding mechanical pulling results were consistent with the experimental information in literature. Apart from $C\alpha$ -based ENMs, standard atomistic force fields can also serve as the FG input in NMA-FM.

Figure 2.3 CG-ENM scheme of lysozyme. Lysozyme is coarse-grained into seven-beads based on the C^{α} atom sequence, as follow: CG1 (1-12), CG2 (13-58), CG3 (59-80), CG4 (81-91), CG5 (92-123), CG6 (124-143), and CG7 (144-164). The seven coarse-grained beads (CG1~CG7) are colored in blue, red, orange, yellow, green, violet and gray, respectively. (a) The side view Illustrates the closure binding motion of domain A (CG1~CG2) and domain B (CG4~CG7). The arrows represent the closure movement. (b) The front view represents a twisting motion between domain A and B. The arrows represent the distortion direction.

2.2.3 Fluctuation matching targeted at experimental B-factors

Thermal B-factors (Debye-Waller factor) of atoms in a biomolecule are the average squared displacements of atoms in the vicinity of their equilibrium positions as a function of the temperature, which measure the stabilities of atoms [133, 137]. Bahar and coworkers compare the coefficient of correlation for 113 proteins systems to the experimental B-factor using GNM and ANM. The correlation coefficients obtained were 0.59 and 0.55 for GNM and ANM respectively [138-140]. However, GNM and ANM models ignore the contribution of the rigid body motions and the effect of crystal packing. This can be considered as form of simulation inaccuracy when compared with the observed experimental value. B-factors are projections of normal modes to atom positions. The ENM methods should be able to perfectly reproduce the experimental B-factors in order to reflect correct protein dynamics. Different methods have been proposed to reproduce experimental B-factors. These methods attempted to reduce the experimental crystallization area like considering the rigid body motions and the crystal packing effect [140-143]. For example Kundu *et al.*, (2002) [140] investigates the dynamics of the protein system in crystal packing by including the effect of nearby molecules which represents the environmental factor of the dynamic in the crystal lattice in GNM which improves the coefficient correlation from 0.59 to 0.66.

Erman (2006) [49] also proposed an algorithm to iteratively update the Kirchhoff matrix of GNM. In the Erman method, the connections of nearby C α atoms in the backbone are fixed, whilst the strength of residues engaged in pairwise interactions is varied until the computed and experimental B-factors are equal. Although the generated effective

B-factors can match the experiment quite well, there are some negative force constants, which pull the structure away from the native equilibrium position. This unphysical observed effect that indicates there may be some over-fitting in the algorithm.

Xia *et al.*, (2013) [53] developed a method named the progressive fluctuation matching (PFM) method to construct heterogeneous anisotropic elastic network models (HANMs). PFM used heterogeneous force constants. The underlying principle of the PFM algorithm is to generate an HANM that fits the experimental B-factor. The PFM algorithm is summarized as follow,

- (1) Let us consider ANM/HANM. First, the harmonic restraint potential is added to each atom so that the computed B-factor is closed to experiment one (outer loop).

The force constant k_i^{res} is computed according to the following equation.

$$k_i^{res} = \alpha k_B T \cdot 8\pi (B_i^{cal} - B_i^{exp}) / (B_i^{cal} B_i^{exp}) \quad (2.23)$$

where α is a numerical scaling factor used to optimize the restraint strength, k_B is the Boltzmann's constant, and T is the temperature.

- (2) The NMA-FM approach discussed in section 2.2.2 is used to parameterize a new HANM without including the restraint potential. The force constant is updated using Equation (2.21) (inner loop).

The PFM method incorporates the functionality of NMA-FM in the first cycle so that the computed B-factor is closed to experiment one. The method iteratively updates the

effective B-factor until it matches the experimental value, using gradually decreasing restraint potentials. It should be noted that PFM and many other methods did not explicitly incorporate the crystal lattice effect, which is convenient for simulation purposes due to technical difficulty [138-140]. The PFM method is able to reproduce accurately the experimental B-factor for proteins.

The PFM algorithm was tested using four proteins systems including crambin (PDB ID: 1CRN) [144], trypsin inhibitor (PDB ID: 5PTI) [145], HIV protease (PDB ID: 1HHP) [146], and lysozyme (PDB ID: 2LZM). The results show a good root mean square displacement (RMSD) between experimentally observed and calculated B-factors, which are 0.060, 0.095, 0.247 and 0.049 Å² respectively, with a very good correlation coefficient of 0.99 for all the protein tested. Figure 2.4 illustrates a comparison between experimentally observed and effective B-factors of the four proteins.

Figure 2.4 The comparisons of the experimental and calculated B-factors for (a) crambin, (b) lysozyme, (c) trypsin inhibitor, and (d) HIV-1 protease. The red circles represent the experimental values. ANM and HANM using a 15 $\text{\AA}$ distance cutoff are represented using green and blue curves.

2.2.4 Relative-entropy-based fluctuation matching with improved convergence

In chapter 3, a theoretical framework of relative entropy combined with the Newton Raphson algorithm to parameterized a heterogeneous elastic network model known as the “Heterogeneous Elastic Network Model Parameterized by Using the Relative Entropy method with Non-negative Constraints” [55] is explained in detail.

3. An improved fluctuation matching approach based on relative entropy for the elastic network model and a structure-based model of biomacromolecules

3.1 Introduction

The relative entropy method is a traditional statistical mechanics method known as the Kullback-Leibler divergence [147]. Mathematically, the relative entropy can be defined as follows:

Let's consider a discrete distribution with a probability function, p_k , and let a second discrete distribution function have a probability q_k . Therefore, the relative entropy of p with respect to q , given by $D(p||q)$ can be described as follows;

$$D(p||q) = \sum_k p_k \log_2 \left(\frac{p_k}{q_k} \right) \quad (3.1)$$

It should be noted that $D(p||q) \neq D(q||p)$, therefore, the relative entropy is not a true metric.

Shell and coworkers [66, 67] applied the relative entropy as a framework for multiscale analysis via a likelihood approach. The relative entropy variant proposed by Shell and coworkers is a robust method for the parameterization of CG-ENM. The underlying principle is to map the target atomistic information (all atoms) into a CG model by minimizing S_{rel} as a function of GC scheme and effective interaction [111, 148].

3.2 Theoretical framework and computational details

Shell and coworkers [66, 67] pioneered a novel CG approach based on minimizing a configuration-space function, known as relative entropy (RE). The advantage of the proposed method is to control the error due to CG in arbitrary microscopic properties and propose a mechanism to mitigate the error. The CG potential energy function is computed by minimizing the RE, S_{rel} . The RE measures the information lost during the CG parameterization. Below is the expression of S_{rel} ,

$$S_{rel} = \int p_{FP}(\mathbf{r}) \ln \left(\frac{p_{FP}(\mathbf{r})}{p_{CG}(M(\mathbf{r}))} \right) d\mathbf{r} + \langle S_{map} \rangle_{FP}, \quad (3.2)$$

where, FP is a “first-principle” model as the reference system described by an atomistic classical potential, the CG model is constructed to reproduce the behavior of the FP model by fewer interaction sites or a simpler form of the interaction potential. The atomistic coordinate $\mathbf{r}$ corresponds to configurationally microstates of the FP model, and $M(\mathbf{r})$ is the CG mapping function on the FP system. In this method, a smaller value of RE corresponds to an acceptable CG model parameter. The method was applied first on the lattice gas. The result revealed that the relative entropy method is robust, applicable to a large variety of systems and reduces CG errors. In order to parameterize a CG model under the relative entropy approach, the optimization corresponds to minimizing the RE value as a function of adjustable features in CG scheme.

The force constant can be optimized in the CG-ENM potential using the HENM to describe the CG system and apply the RE approaches to perform the CG model

parameterization. The potential energy function V_{HENM} is the sum of the harmonic interactions:

$$V_{HENM} = \sum_i^N \sum_{j>i}^N \frac{1}{2} k_{ij} (r_{ij} - r_{ij}^0)^2 \cdot H(r_c - r_{ij}^0), \quad (3.3)$$

where k_{ij} is the heterogeneous force constant of the harmonic springs connecting a pair of particles i and j and N is the total number of particles. r_{ij} is the distance between particles i and j , r_{ij}^0 is the equilibrium distance between particles i and j , r_c is the cut-off distance, and H is the Heaviside unit step function. To minimize the residual RE, the fluctuation bond obtained from the FP configuration should match with the CG output model. Based on Equation (3.2), a canonical expression for the first derivative of the relative entropy is described by

$$\nabla_k S_{rel} = \beta \left\langle \frac{\partial U_{CG}}{\partial k} \right\rangle_{FP} - \beta \left\langle \frac{\partial U_{CG}}{\partial k} \right\rangle_{CG} = \langle \Delta r_{ij}^2 \rangle_{FP} - \langle \Delta r_{ij}^2 \rangle_{CG}, \quad (3.4)$$

where k represents the force constant related to particle i and j , and the two subscripts of k are not shown for simplicity. For both FP and CG models, one can compute the pairwise-distance fluctuations using MD simulations or NMA. The positivity of the second derivative of the Hessian matrix can be considered as proof of the existence of a minimum RE.

For a single parameter case,

$$\frac{\partial^2 S_{rel}}{\partial k^2} = \beta^2 \left\langle \frac{\partial U_{CG}^2}{\partial k} \right\rangle_{CG} - \beta^2 \left\langle \frac{\partial U_{CG}}{\partial k} \right\rangle_{CG}^2. \quad (3.5)$$

The second derivative can be computed from the CG ensemble and is always positive. This is the evidence that the RE possesses a single global minimum for the linear parameters in CG potential. This conclusion can be applied to the multiple parameter case for HENM, where the Hessian matrix replaces the second derivative. The Hessian matrix is constructed during the optimization using Newton's method, on the basis of the gradient in Equation (3.4). By the combination of Equation (3.3) and (3.4),

$$\begin{aligned} H_{mn} &= \beta \left\langle \frac{\partial^2 U_{CG}}{\partial k_m \partial k_n} \right\rangle_{FP} - \beta \left\langle \frac{\partial^2 U_{CG}}{\partial k_m \partial k_n} \right\rangle_{CG} + \beta^2 \left\langle \frac{\partial U_{CG}}{\partial k_m} \frac{\partial U_{CG}}{\partial k_n} \right\rangle_{CG} - \beta^2 \left\langle \frac{\partial U_{CG}}{\partial k_m} \right\rangle_{CG} \left\langle \frac{\partial U_{CG}}{\partial k_n} \right\rangle_{CG} \\ &= \beta^2 \left\langle \Delta r_{ij}^2 \cdot \Delta r_{kl}^2 \right\rangle_{CG} - \beta^2 \left\langle \Delta r_{ij}^2 \right\rangle_{CG} \left\langle \Delta r_{kl}^2 \right\rangle_{CG}, \end{aligned} \quad (3.6)$$

where k_m and k_n are the two force constants for i - j and k - l pairs respectively.

The relative entropy approach provides a rigorous framework for the multiscale optimization of CG-ENM. Equation (3.5) is consistent with the idea of fluctuation matching and the CG spring constants can be obtained by letting the derivative in the equation be zero. The parameterization of CG model is subsequently obtained by the standard Newton's (Newton-Raphson) method using the Hessian matrix in Equation (3.6). Compared with the heuristic fluctuation-matching algorithm in Equation (2.21), The RE method considers the pairwise correlation between two pair-distance fluctuations. The advantage of using relative entropy is that it guarantees the existence and uniqueness of

the HENM solution. Conversely, algorithms based on Equation (2.22) and (3.6) can have convergence problems for molecular systems with strong inter-pairwise correlations.

3.3 Parameterization with Non-Negativity Constraints

A new version of HENM based on relative entropy is proposed to improve convergence in multiscale parameterization of HENM approach. Compared with the traditional HENM parameterization methods such as fluctuation matching, the parameterization of HENM with non-negative constraints using relative entropy all obtained HENM force constant that are strictly positive. Thus, the traditional framework needs to be modified to incorporate the constraint of a positive solution. The Newton-Raphson method with a non-negativity algorithm [149] is used for this goal.

The reduced Hessian matrix H_R is constructed in order to include the non-negativity constraints;

$$[H_R]_{mn} = \begin{cases} \delta_{mn} & \text{if } k_m = 0 \text{ or } k_n = 0, \\ \frac{\partial^2 S_{rel}}{\partial k_m \partial k_n} & \text{otherwise,} \end{cases} \quad (3.7)$$

where δ is the Dirac delta function. In one iteration step, each force constant k is updated by

$$k^{q+1} = P(k^q - \alpha H_R^{-1} \cdot \nabla_k S_{rel}), \quad (3.8)$$

where α represents the step size, q represents the iteration number, and $P(k)$ represents a

$$\text{projection operator: } P(k) = \begin{cases} k & k > 0, \\ 0 & \text{otherwise,} \end{cases}.$$

The iteration stops when k^{q+1} converges to k^q . This algorithm is locally quadratically convergent.

The initial guess of the force constant needs to be computed using the equipartition theorem as describe in Equation (3.9) before the iteration process starts.

$$k_m^0 = \frac{k_B T}{\langle \Delta r_{ij}^2 \rangle_{FP}}, \quad (3.9)$$

where $\langle \Delta r_{ij}^2 \rangle_{FP}$ is the pairwise distance fluctuation in the first principle (FP) ensemble.

In the proposed RE approach to parameterize HENM, in addition to the element of the Hessian matrix, H_{ij} to relate to a single pairwise fluctuations $\langle \Delta r_{ij}^2 \rangle_{CG}$, it also depends on the correlation between two different pairwise fluctuations (e.g., the correlation between $\langle \Delta r_{ij}^2 \rangle_{CG}$ and $\langle \Delta r_{kl}^2 \rangle_{CG}$). The incorporation of this feature in the parameterization of HENM using RE-based parameters improves the convergence compare to the heuristic parameterization algorithm pioneered by *Lyman et al.*, (2008) [50].

The relative entropy method coupled with the projected Newton method using a non-negativity simulation was tested using an experimental NMR protein structure. The data set comprising the structures of 25 proteins were obtained from the Protein Data Bank

(<http://www.rcsb.org/>) illustrated in Table 3.1. In order to select a reference structure from an NMR ensemble, we first aligned the NMR, and the average structures were calculated from the NMR ensemble. Secondly, the lowest RMSD structures were selected as the representative structures for the NMR structure ensembles. The representative structures were considered to compute a target fluctuation for the RE and FM approaches. Two CG schemes were considered, namely the C α -level (Table 3.1) and the eight-bead (CG-8) (Table 3.2). The optimize force constant cut-off was set to 1.5 nm for both methods (RE and FM). The HENM pairwise distance fluctuations were obtained from 10 *ns* of MD simulations.

3.4 Results and discussions

Based on theoretical background as illustrated in section 3.3, the RE method is more robust compare to the FM approach considering the convergence in force constant optimization due to the fact that the inter-spring correlation is an important factor in the algorithm. The following residual to quantify the error in the optimization iterations was defined as follows;

$$\text{Res} = \sum_{\text{springs}} \frac{\left| \langle \Delta r_{ij}^2 \rangle_{FP} - \langle \Delta r_{ij}^2 \rangle_{CG} \right|}{\langle \Delta r_{ij}^2 \rangle_{FP}}. \quad (3.10)$$

Table 3.1 and Table 3.2 illustrate the residual of HENM parameterization using RE and FM for CG-C α and CG-8 beads respectively. It can be observed that the RE produces

better fitting results for the parameterization of HENMs, which is closer to the true global minima compare to FM.

Table 3.1 NMR data set and comparison of the converged residuals between RE and FM methods. The residual numbers are calculated for C α -level HENMs.

Protein PDB ID	Number of Residues	Number of NMR	RE Residual	FM Residual
1B3C	64	40	0.1556	0.1801
2NR2	76	144	0.0920	0.1241
1BB0	56	60	0.1906	0.2022
2KDP	71	47	0.1163	0.1402
1SKT	90	40	0.1890	0.2230
1GB1	56	60	0.1426	0.1621
1GB4	57	47	0.1694	0.2003
1GHC	75	14	0.2649	0.2751
1BRZ	53	43	0.156	0.1328
1Q9P	95	20	0.1952	0.2095
1HSN	79	49	0.1754	0.1980
1MBH	52	50	0.1284	0.1536
1FD6	57	40	0.1476	0.1686
3NLA	73	40	0.1546	0.1808
1BW5	66	50	0.1115	0.1384
1CXW	60	50	0.1706	0.1920
2K39	76	116	0.0967	0.1267
1FCL	56	40	0.1962	0.2230
1OMT	56	50	0.1919	0.2176
1TNP	90	40	0.1310	0.1559
1TNQ	90	40	0.1108	0.1302
1R2U	89	40	0.2431	0.2591
2B88	58	40	0.2274	0.2559
1AFP	51	40	0.1812	0.2045
1NCS	47	46	0.1734	0.1941

Table 3.2 NMR data set and comparison of the converged residuals between RE and FM methods. The residual numbers are calculated for CG-8 level

Protein PDB ID	Number of Residues	Number of NMR models	RE Residual	FM Residual
1FD6	57	40	0.0346	0.0672
1GB1	56	60	0.0379	0.0552
1OMT	56	50	0.0195	0.0459
1FCL	56	40	0.0602	0.0632
1AFP	51	40	0.0248	0.0372
1GB4	57	47	0.0586	0.1005
1B3C	64	40	0.042	0.0205
1GHC	75	14	0.0345	0.1292
1R2U	89	40	0.042	0.0927
1TNP	90	40	0.0618	0.0399
1TNQ	90	40	0.0066	0.0643
3NLA	73	40	0.0752	0.0991

In principle, the projected Newton algorithm guarantees the optimal force constants. In Figure 3.1, the convergence profiles of the $C\alpha$ -CG scheme of the RE and FM methods serve as a function of the iteration step. It has been observed that the RE method converges faster than FM. Even though the RE method converges fast, the FM method can still produce a reasonable final result with a slightly larger residual, which explains the success of the FM algorithm in literature when the $C\alpha$ -CG scheme is considered. Therefore, the force constant the FM method produces are closes to the true global minimum obtain from the RE method.

Figure 3.1 Comparison of the residual convergence from RE and FM for $C\alpha$ -level ENMs. (a) 1GB1: the RE is denoted by the red curve (with the final residual 0.142) and FM is denoted in green curve (with the residual 0.162). (b) 1GB4: the RE is denoted in red curve (with the final residual 0.169) and FM is denoted in green curve (with the final residual 0.20).

Figure 3.2 shows the convergence plot of the CG scheme for eight beads (CG-8) HENM of each NMR structure. Compared with the C- α CG scheme, CG-8 beads show a clear difference between RE and FM in terms of convergence; meaning that although it is challenging to obtain converged results for both approaches due to higher residual fluctuations, the RE converge fast than FM. It can be clearly observed that both residual profiles in Figure 3.2 reach the steady region and the RE residual is observed to be lower than FM. Figure 3.3 compares RE and FEM using CG-8 spring constants from the last iterations in Figure 3.2. Due to the variability in the magnitude of spring constant, the relative differences are calculated as $|k(\text{RE}) - k(\text{FM})|/k(\text{FM})$ and are illustrated in Figure 3.3. The spring constant of RE is in the numerator because it contains a zero value due to the incorporation of a non-negativity constraint. Therefore, the FM spring constant is in the denominator. In Figure 3.3, it appears that the difference between the RE and FM force constant mainly originate from three spring constants, with a 100% relative difference (one spring from 1FD6 and two spring from 1GB1), while other spring constants display a change of less than 10%. Furthermore, a 100% relative differences originates from the zero values of the RE method. The three springs have the largest (or second largest) fluctuations in the corresponding protein model, and the RE algorithm completely removed the three harmonic interactions and produces zero value spring constants. It is observed that the RE approach attempted to generate the three pair fluctuations using other harmonic interactions in the networks. Comparatively, the FM approach can adjust a pair fluctuation by turning the spring constant of the same pair. Due to the use of “one over spring constant” terms in Equation (2.21), the FM algorithm is unable to generate final optimization results close to zero. Practically, it is difficult to get

converged target fluctuations for those pairs with larger fluctuations and weaker interactions. Based on Figure 3.3, we demonstrated that the RE method is able to generate those high fluctuations by using other harmonic interactions in the ENM and therefore produces the optimal fitting results.

Figure 3.2 Comparison of the residual convergence from RE and original FM for eight-bead CG ENMs. (a) 1FD6: the RE is denoted in red curve (with the final residual 0.025) and FM is denoted in green curve (with the final residual 0.030); (b) 1GB1, the RE is denoted in red curve (with the final residual 0.032) and FM is denoted in green curve (with the final residual 0.039).

Figure 3.3 The relative difference between the spring constants generated by FM and RE methods for the CG-8 scheme. The percentage value on the y-axis is computed as the absolute value of the difference divided by the FM spring constant. The x-axis shows the spring index of the protein. The two proteins (a) and (b) are 1FD6 and 1GB1 respectively.

Furthermore, we analyze the difference between RE and FM parameterization methods as a function of their dynamics. The dynamics of protein using ENM can be represented by low frequency modes. Ten first modes from a principal component analysis (PCA) of HENM MD trajectories and NMR ensembles to represent CG and FP proteins dynamics were considered. The mode overlap used to compare mode similarity was described following the work of Yang *et al.*, (2008) [150] using Equation (3.11).

$$O_m(\mathbf{u}, \mathbf{v}) = \frac{1}{m} \sum_{k=1}^m \sum_{p=1}^m |\mathbf{u}^{(k)} \cdot \mathbf{v}^{(p)}|, \quad (3.11)$$

where $\mathbf{u}$ and $\mathbf{v}$ are two sets of modes from different models, $\mathbf{u}^{(k)}$ and $\mathbf{v}^{(p)}$ is the k^{th} and p^{th} slowest modes of $\mathbf{u}$ and $\mathbf{v}$ respectively, and m is the total number of low frequency modes in the calculation ($m=10$). The absolute value of O_m is in the interval of 0 to 1. The similarity between the two modes are proportional, thus the higher this value, the higher the similarity. The result in Table 3.3 illustrates that for both C α and CG-8 cases, the RE method is able to produce low frequency modes that are closer to the NMR-derived ones, compared to the FM approach.

Table 3.3 Mode Overlap between the NMR and HENM modes calculated by principal component analysis (PCA) for Relative Entropy and Fluctuation Matching of the 1FD6 protein.

Mode Overlap between NMR and HENM		
CG Level	Relative Entropy	Fluctuation Matching
C α	0.6652	0.6564
CG-8 beads	0.8539	0.8360

Further validation of the method was conducted using middle size protein such 2KDP and 1HHN show in Table 3.4. In Figure 3.4, the convergence profile of the C α -CG shows that RE method has better convergence compare to FM which is consistent with the results of the small size protein tested in Table 3.1.

Table 3.4 NMR data set and comparison of the converged residuals between RE and FM methods of middle size protein. The residual numbers are calculated for C α -level HENMs.

Protein PDB ID	Number of Residues	Number of NMR models	RE Residual	FM Residual
2KDP	118	40	0.2104	0.2304
1HHN	101	20	0.1315	0.1501

Figure 3.4 Comparison of the residual convergence from RE and FM for Ca-level ENMs. (a) 2KDP: the RE is denoted by the red curve (with the final residual 0.21) and FM is denoted in green curve (with the residual 0.23). (b) 1HNN: the RE is denoted in red curve (with the final residual 0.13) and FM is denoted in green curve (with the final residual 0.15).

3.5 Comparison between relative entropy with non-negativity force constant and fluctuation matching related methods

As illustrated in Table 3.4, we compare the five fluctuation matching-related methods described above. The common underlying principle of all these methods is the concept of multiscale modeling and their aim to parameterize CG-ENM or other structured-based models by fitting a set of fluctuations. The principal difference between different parameterization methods discussed in this thesis is the fitting target and iterative algorithm.

The fluctuation matching approach was pioneered by Chu *et al.*, (2006) [8], and Lyman *et al.*, (2008) [50] implemented the $1/k$ type of formula iteration, which is an improved convergence of the original FM method. In Chu and Lyman's approaches the fitting target is the pair fluctuations obtained from MD simulations, and obtaining fluctuation data can be computationally expensive for large systems. Thus, the NMA-FM is a scalable GC scheme using pairwise fluctuations derived from NMA as fitting target [51]. An improvement of NMA-FM known as "PFM" algorithm addressed the problem of B-factor fitting using ENM [53]. The HENMs obtained using NMA-FM and PFM is considered a multiscale model with different fitting targets (B-factor and pair-fluctuation). This work provides a theoretical interpretation of the FM with a practical algorithm based on relative entropy, and it numerically guarantees the optimal solution for spring constants in FM.

Table 3.5 Comparison of fluctuation-matching related methods

Method	Fitting target	Iterative Formula(s)
Chu et al. [48]	Pair fluctuations from MD simulation	$k_{ij}^{n+1} = k_{ij}^n + \alpha \left(\langle \Delta r_{ij}^2 \rangle_{CG}^n - \langle \Delta r_{ij}^2 \rangle_{AA} \right)$
Lyman et al. [50]	Pair fluctuations from MD simulation	$\frac{1}{k_{ij}^{n+1}} = \frac{1}{k_{ij}^n} - \alpha \left(\langle \Delta r_{ij}^2 \rangle_{CG}^n - \langle \Delta r_{ij}^2 \rangle_{AA} \right)$
NMA-FM [51]	Pair fluctuations from NMA	
PFM [53]	B-factor	$k_i^{res} = \alpha k_B T \bullet 8\pi \left(B_i^{cal} - B_i^{exp} \right) / \left(B_i^{cal} B_i^{exp} \right)$
		(Outer loop) $\frac{1}{k_{ij}^{n+1}} = \frac{1}{k_{ij}^n} - \alpha \left(\langle \Delta r_{ij}^2 \rangle_{CG}^n - \langle \Delta r_{ij}^2 \rangle_{AA} \right)$ (Inner loop)
Relative Entropy [55, 151, 152]	Pair fluctuations from MD simulations	$S_{rel} = \int p_{FP}(\mathbf{r}) \ln \left(\frac{p_{FP}(\mathbf{r})}{p_{CG}(M(\mathbf{r}))} \right) d\mathbf{r} + \langle S_{map} \rangle_{FP}$

3.6 Conclusions

The primary goal of this study was to develop relative entropy based new fluctuation method to parameterize CG ENM and structural-based model that are able to extend the time-scale limitation for proteins and other biomolecules. The main idea of this method is to minimize the relative entropy with a non-negative constraint and thus update the force constants iteratively until it matches the target bond fluctuations. An HENM derived through the constrained relative entropy method, consists of harmonic springs with realistic positive force constants, and yields the calculated bond fluctuations that are basically identical to the experimental ones. The original fluctuation matching [110] has been successfully used in MD simulations to parameterize ENM and structure-based models. The relative entropy with a non-negative constraint algorithm results were compared with those using the original fluctuation matching algorithm. While the original fluctuation matching approach has achieved wide success in developing CG-ENM and structure-based models, incorporating the fluctuation-matching algorithm into the relative entropy scheme has the advantage of guaranteeing a global optimum with fast convergence.

4. Molecular simulation to investigate the adsorption mechanism of the endocrine disrupting compound 17- α ethinylestradiol at the monomer level

4.1 Introduction

MD simulations were conducted to study adsorption mechanism for the removal of micropollutants (EE2 or testosterone) in wastewater treatment. Around 70% of pharmaceutical residues in wastewater, including EE2 are from household used. Following our experiments, results have shown that the microporous polyether sulfone (PES) hollow fiber is able to retain aromatic hormones especially EE2 [22, 112], thus, a computational study was conducted to investigate the adsorption of micropollutants (EE2 or testosterone) onto a single monomer at the atomistic level. In order to have a comprehensive comparison, different types of single monomer structures were considered such as PES, polysulfone (PS), polyamide 612 (PA6-12) and polyvinylidene fluoride (PVDF). It should be noted that PVDF structure was used as a control system. The binding free energy was calculated using umbrella sampling to evaluate the interaction between a small molecule solute (EE2 or testosterone) and a single monomer. The potential of mean force (PMF) results show that PES has a strong binding free energy (strong interaction) with EE2 meaning that PES is able to retain more EE2 compared to PS, PA6-12 and PVDF. The results are consistent with experimental observable [22, 112]. Further analysis was carried out to determine the major contributor to the interaction as well as gain insights into the details information of the interaction. Different analogous structures of EE2 were constructed such as EE2 without the double

bond (π -bond) in the benzene ring (EE2-AR), EE2 without the carbon triple bond (EE2-CA), and EE2 without the OH group (EE2-OH). PMF between different analogs shows that EE2-AR PMF and binding free energy decreased compared to the original EE2 structures, which indicates high contribution of the π - π interaction in to the interaction between PES-EE2.

4.2 Modeling and simulations

The atomistic model of six monomers units of PES, PS, PA6-12, and PVDF as well as the small molecule hormones (testosterone, EE2) as illustrated in Figure 4.1, was constructed using Accelrys Material Studio software (version 2016) [153]. The initial structures were optimized by Density Function Theory (DFT) calculations using Dmol³ package to generate stable conformations. The DFT calculations used the gradient-corrected functional algorithm (GGA) and PW91 with a double zeta plus polarization basis set (DNP) basis set [154]. The atomic charges were assigned using Mulliken population analysis. After the DFT optimization, MD simulations were carried out for each monomer unit and solute molecule solvated in water to study the interaction energy between them in terms of the potential of mean force (PMF, i.e., binding free energy as a function of a reaction coordinate).

Figure 4.1 Chemical structure and atomistic structure of monomers. (a) PES. (b) PS.

(c) EE2. (d) PVDF. (e) PA6-12. (f) testosterone.

All MD simulations were conducted using the GROMACS package, version 4.6.5 [155] with the OPLS force field [156]. The topology file was constructed using *topolbuild* tool (www.gromacs.org/Downloads/User_contributions/Other_software).

The complex structures, PES-EE2, PS-EE2, PA6-12, and PVDF-EE2, and PES-testosterone and the analogous structures of EE2 (PES-EE2-AR, PES-EE2-CA, and PES-EE2-OH) were placed in a cubic box, and solvated with single point charge (SPC) water [157]. For each simulation system, energy minimization was performed to relax the structure using the conjugate gradient algorithm [158]. Prior to the steered molecular dynamics pulling simulations (SMD) to obtain a different initial structure along the reaction coordinate, an MD equilibrium simulation (NPT) of 5 ns in the absence of any restraint were performed. The integration time step of 2 fs was used. The simulation temperature and pressure were maintained at 300 K using the Parrinello-Donadio-Bussi V-rescale thermostat [159] and at 1 atm using the Berendsen barostat [160], respectively. The short-range non-bonded interactions were modeled using Lennard Jones potentials with a cut off at 1.4 nm, and the long-range electrostatic interactions were calculated using the particle mesh Ewald (PME) algorithm [161, 162]. The LINCS algorithm [163] was used to constrain all bond lengths. As shown in Figure 4.2, the root mean square

displacements (RMSD) for the complexes structures reached the equilibrium at 0.4 ns.

Figure 4.2 Root mean squared displacement of the complex structures, PES-EE2, PS-EE2 and PES- testosterone during equilibrated simulations.

The equilibrated trajectory structure with a smallest distance between the monomer and the solute organic molecule of each simulation was aligned in the z direction and used as the initial structure for SMD pulling simulations. The structures were placed in a rectangular box with periodic boundary condition and solvated with water. For each SMD pulling simulation, the center of mass (COM) of EE2, testosterone or analogues EE2 molecules were pulled away from the COM of the monomer along the z -axis over 200 ps. This SMD pulling simulations sample different conformations at various interaction distances with a spring constant of 2000 KJ mol^{-1} and a pulling rate of 0.02 nm . A

position restraint was applied on the monomer during the pulling simulations. The maximum pulling distance considered was set to 1.2 nm. Snapshots of conformations obtained from the SMD pulls within the range of 1.2 nm were considered as the initial conformations for a series of umbrella sampling simulations [164, 165]. The umbrella sampling window spacing was set at 0.1 nm. The total number of windows for each system was defined based on its interaction and free energy landscape profile. The simulation time for each window was set to 10 ns. The weighted histogram analysis method (WHAM) [166] was used to generate the free energy profiles from umbrella sampling trajectories.

The umbrella sampling simulations were performed to compute the free energy change ΔG_{bind} along the reaction coordinate ξ situated the in z -axis. The free energy difference between the lowest free energy in the profile at a close contact and the value of the plateau at large distances stands for the binding free energy of the molecular pair

$$\Delta G_{bind} = (G_{max} - G_{min}) \quad (4.1)$$

4.3 Results and discussions

Potential mean of force (PMF) profiles were computed from a set of umbrella sampling simulations for six complexes of monomers (PES, PS, PA6-12, and PVDF) and small molecules (EE2 or testosterone) including the EE2 analogues with the distance between them as the reaction coordinate. It should be noted that in a calculation of PMF profiles, a plateau region at long separating distances between the molecules and an energy minimum at shorter distances is present. The differences between these two free energies, E_{min} and E_{max} , corresponds to the ΔG_{bind} of binding for monomers and small molecules. The PMF profiles and values of ΔG_{bind} of binding for five complexes are shown in Figure 4.3.

At short distances (< 0.1 pm), highly repulsive free energies can be observed. The low ΔG_{bind} value of PES-EE2 (-1.5 kcal/mol) shows that the two molecules have a high binding free energy, indicating a strong interaction between them. The ΔG_{bind} values of PS-EE2 (-0.75 kcal/mol), PA6-12-EE2 (-1.28 kcal/mol), PES-testosterone (-0.98 kcal/mol) and PVDF-EE2 (0.0 kcal/mol) are notably higher compare to PES - EE2, implying a much weaker interaction between EE2 and PS, PA6-12 or PVDF. The binding free energies of EE2 with the five tested monomers are consistent with the experimental observation that a PES membrane has the highest adsorption capability for EE2.

Figure 4.3 PMF with increasing COM distances between monomers (PES, PVDF and PS) and EE2. (a) PES-EE2 with the binding free energy of -1.5 kcal/mol. (b) PS – EE2 with the binding free energy of -0.75 kcal/mol. (c) PVDF-EE2 with the binding free energy of 0.0 kcal/mol (no interaction). (d) PA6-12-EE2 with the binding free energy of -1.28 kcal/mol. (e) PES-testosterone with the binding free energy of -0.98 kcal/mol.

In experimental conditions, the single monomer exists as a combination of many units to form a polymer membrane and not as a single unit as analyzed in the study. Therefore, molecular simulation may not fully explain adsorption mechanism between polymers and small molecules (EE2, testosterone), which is one limitation of this study. This limitation is addressed in chapter 5 where adsorption of small molecules (EE2 or testosterone) onto polymer membranes was considered. Nevertheless, this study provides important insights into the intermolecular interactions between a single monomer of the various structures and EE2 or testosterone; this sheds light on the possible explanations for the difference in EE2 adsorption between different systems such as PES, PS, PA6-12, and PVDF.

Further simulations were conducted to investigate the interaction mechanism of EE2 adsorption and to understand the major contributing factor, which enhances interaction between EE2 and PES. The binding free energies between PES and EE2 analogous structures were performed using the same approach as the previous calculations. The analogous structures included EE2 without the double bond (π -bond) in the benzene ring (EE2-AR), EE2 without the triple bond carbon (EE2-CA), and EE2 without the OH group (EE2-OH) as shown in Figures 4.4. The EE2-AR analog structure was designed to reduce aromatic π - π interaction by removing its aromatic functional group. The EE2-CA analogue structure was designed to reduce the hydrophobic interaction with the monomer while EE2-OH was designed to reduce the hydrogen bonding ability with the removal of the OH functional group.

Figure 4.4 Chemical structures and atomic structure of EE2 and its analogs. (a) EE2-AR (EE2 without the p-bond in the benzene ring). (b) Original EE2. (c) EE2-OH (EE2 without the OH group). Dotted circles highlight the difference between EE2, EE2-OH and EE2-AR.

Analogous structures PES-EE2-AR, PES-EE2-CA and, PES-EE2-OH have higher ΔG_{bind} values (-0.62 kcal/mol, -1.07 kcal/mol and -1.00 kcal/mol, respectively) compared to PES-EE2 (-1.5 kcal/mol). These ΔG_{bind} mean that the forces of intermolecular interaction between the modified hypothetical EE2 analogues and PES monomer is much lower as compared to the original EE2 structure. In particular, compared with EE2 and EE2-AR has a significantly weaker interaction with PES, indicating a greater contribution of π - π interaction than hydrogen bonding to the high affinity between EE2 and PES.

The PMF profiles for PES - EE2-AR, PES - EE2-CA, PES - EE2-OH and their respective values of ΔG_{bind} are calculated and shown in Figure. 4.5.

Figure 4.5 Potential of mean force (PMF) with increasing COM distances between the PES monomer and EE2 analogues (EE2, EE2-OH and EE2-AR). (a) PES-EE2 with the binding free energy of -1.5 kcal/mol. (b) PES-EE2-AR (EE2 without the p-bond in the benzene ring) with the binding free energy of -0.62 kcal/mol. (c) PES-EE2-OH (EE2 without the OH group) with the binding free energy of -1.00 kcal/mol. (d) PES-EE2-CA (EE2 without the triple bond carbon) with the binding free energy of -1.07 kcal/mol.

4.4 Conclusions

The objective of this study was to investigate adsorption mechanism between small molecules (EE2 or testosterone) and a single monomer (PES, PS, PA6-12, and PVDF) in order to validate the experimental results we obtained from microporous PES hollow fiber membrane that has a high capability of retaining EE2. MD simulations revealed strong binding affinity between PES and EE2, which can be principally associated with the π - π interaction and hydrogen bonding. It should be noted that under experimental conditions, a single monomer structure exists as an aggregation of many units to constitute a polymer membranes. Therefore, molecular simulation using single unit may not fully explain how the various polymers interact with EE2, which is one limitation of this study. The following chapter addresses this limitation by analyzing the interactions between small molecules (EE2 or testosterone) using polymer membranes composed of 30 repeated units and 2 chains for each structure.

5. Molecular insight into the interaction between 17- α ethinylestradiol and polymer membranes in wastewater purification

5.1 Introduction

Different separation techniques have been proposed and implemented for the removal of micropollutants from wastewater treatment including physical processes, biological processes and advanced oxidation processes (AOP). The physical processes can be categorized into two groups, sorption, and membrane technology. Most widely techniques in membrane technology are Microfiltration (MF), Ultrafiltration (UF), Nano filtration (NF), and reverse osmosis (RO) [73-75]. The retention of micropollutants in membrane-based physical processes depends mainly on the type of membrane, the solution chemistry, the characteristic of the substance to be removed, and the interaction between the substance to be removed and the membrane. A lot of research groups have considered various types of polymer membranes, which can be used efficiently in physical processes separation techniques for the removal of micropollutants by taking into consideration the aforementioned features along with the cost effectiveness of the polymer materials. Schäfer *et al.*, (2011) [73] presented in a review paper, a non-exhaustive list of different polymer membranes used in adsorption studies for the removal of micropollutants such as polysulfone, polyester, polyamide, polyether sulfone, polyvinylidene fluoride, polystyrene, polypropylene, cellulose and others. PES membranes have been widely used in environmental and life sciences research because of its low specific protein-binding adsorption properties, and are recommended for filtering biological and pharmaceutical

solutions and have a high adsorption capacity for low-polarity organic solvents. Ng *et al.*, (2005 and 2017) [22, 112] observed that a microporous polyether sulfone hollow fiber has much higher efficiency to retain aromatic hormones especially EE2.

Studies carried out across different countries reveal the presence of EE2 in influents and in effluents of the wastewater treatment process and receiving waters [104, 167-169] with the values from 0.2 to 1.5 ng/L (0.88 – 6.58 pM) using identical surface water. Despite the fact that some studies are able to detect the presence of EE2 in surface water in low concentrations (ng L⁻¹) [170], in many cases EE2 is not detected due to the poor limit of detection (LOD) of 1 ng/L (4.39 pM) and low concentration of EE2 which still constitute a risk for human and wildlife [171-174]. Among the studies which are able to remove EE2, only a few provide atomistic insight information on the interaction mechanism between polymer membranes and EE2 [21]. Bhandari, R.K et al., (2015) [89] presented a non-exhaustive table of estrogenic chemical occurrence in surface water from different countries. The challenge of removing EE2 from water is not only due the small concentration in which it occurs and its physiological activity at such low levels, but also its small size (molecular weight of 296 g/mol) [73]. Membrane technology has been widely used and applied in water treatment and purification due its low cost and reliable performance [175, 176]. Strong atomistic interaction with the membrane can enhance the removal of contaminants from source water.

In this chapter, we present a comprehensive novel and efficient MD simulation framework to enumerate the removal of EE2 from aqueous solution using polymer membranes and provide molecular level information on polymer membrane interactions with EE2.

Molecular systems of EE2 interacting with four types of polymer membrane namely, polyether sulfone (PES), polyvinylidene fluoride (PVDF), polyamide 612 (PA6-12) and polystyrene (PYS) are considered. To test the concept, testosterone interacting with PES and PVDF are also included in the simulation. The choice of polymer membranes tested in this study was based on the results obtained from experiments that were conducted to analyze the sorption between organic solute molecule and polymer membranes [21, 22, 73, 112, 177-179].

5.2 Modeling and simulations

5.2.1 Atomistic models

The structure of polymer membranes is illustrated in Figure 5.1. The solute organic molecules considered are EE2 and testosterone. The atomistic models of the polymer membranes were constructed by the Amorphous Cell module in Materials Studio software [153]. Each membrane is composed of two polymer chains with the degree of polymerization of 30, in a three dimensional periodic cell and the density of the dry polymer for PES, PVDF, PA6-12 and PYS are 1.37 g/cm^3 , 1.78 g/cm^3 , 1.14 g/cm^3 and 1.05 g/cm^3 respectively and pressure was set to 1 bar. The atomistic models of solute organic molecules, EE2 and testosterone were also constructed using Materials Studio. The initial structures were optimized by Density Function Theory (DFT) calculations using the Dmol³ module. The atomic charges were assigned using Mulliken population analysis using the gradient-corrected functional algorithm (GGA) and PW91[180] with the DNP basis set. The annealing procedure was performed to equilibrate the initial structures of the polymer chain to avoid any unstable conformations and prevent the

structure from being trapped in various local minimum [181, 182]. The annealing simulation was performed with initial temperature of 300 K to the maximum of 500 K heating up every 50 K, and heating down every 50 K using NPT ensemble. The annealing procedure was cycled five times.

Figure 5.1 Chemical structure of the organic solute molecules and model polymers:

(a) EE2. (b) Testosterone. (c) PES. (d) PA612. (e) PVDF. (f) PYS.

5.2.2 Membrane solution interface simulation system

To calculate the relative affinity of organic solute molecules with polymer membranes, a membrane solution interface simulation system was setup as shown in Figure 5.2 and used to compute the free energies of organic solute crossing the polymer chain membrane from aqueous solution. The umbrella sampling method [123, 183] was used to calculate the potential mean of force (PMFs), which represents the free energy change ΔG_{bind} of

organic solute crossing polymer chain membranes as a function of the reaction coordinate ξ , situated in the *z*-axis between the organic solute center of mass (COM) and the center of the polymer chains.

Figure 5.2 SMD pulling snapshot of EE2 (green) crossing PES membrane (licorice blue) in *z*-direction.

Two polymers chains and one organic solute molecules (EE2 or testosterone) were placed in a rectangular box size of (3.87 nm 1.93nm 10.64 nm) with periodic boundary condition in three directions and solvated in water [184] and the polymer membranes thickness of ~ 2.0 nm. For each steered MD simulation (SMD) pulling structures. COM of EE2 or testosterone was pulled to cross polymer membranes along the *z*-axis while the position restraint was applied on the polymers chains. The SMD pulling simulations allow the organic solute molecules to sample a range of conformations at different interacting distances. The snapshots obtained from SMD pulls were considered to be the initial conformations for a series of umbrella sampling simulations. The simulation time for each window was set to 10 ns. The weighted histogram analysis method (WHAM) [125] was used to compute the binding free energy profiles from the umbrella sampling trajectories.

In order to have an accurate binding free energy profile and better sampling, 22 independent umbrella-sampling trajectories were considered for each system and used to calculate the mean binding free energies and standard error. The binding free energy difference ΔG_{bind} between the lowest free energy in a profile at a close contact and the value of the plateau at large distances apart represents the binding free energy.

5.2.3 Solution inside the membrane

The simulation system inside the polymer membranes was built up by placing two polymer chains and two solute organic molecules (EE2 or testosterone) in a box, and solvated with single point charge (SPC) water [184]. It should be noted that no water molecules was trapped inside the polymer due to entropically unfavorable condition, and the polymer membranes type considered in this study are amorphous which does not have close pocket to trap water molecule. The Composition of polymer chains and solute molecules in the system, as well as the simulation conditions are listed in Table 5.1. Periodic boundary conditions were applied to the system in all three directions. The energy minimization was performed to relax the structure using a steep descent algorithm. The six optimized polymer membranes cells in a cubic box with organic solute molecules are presented in Figure 5.3. The cubic box size for each systems is; PES-EE2 ((3.26)³ nm³), PES-testosterone ((3.65)³ nm³), PVDF-EE2 ((1.9)³ nm³), PVDF-testosterone ((1.9)³ nm³), PYS-EE2 ((2.98)³ nm³), PA612-EE2 ((3.63)³ nm³).

Figure 5.3 Optimized models of polymer membranes in green, organic solute molecules in blue and water molecule in red. **(a) PES-EE2. (b) PES-testosterone. (c) PVDF-EE2. (d) PVDF-testosterone. (e) PYS-EE2. (f) PA612-EE2.**

The non-equilibrium MD simulations were conducted in order to investigate the solute organic mobility in the presence of different polymer membranes. NPT dynamics was first computed to adjust the density of the system. Furthermore, 100 ns non-equilibrium NVT simulation with an integration time step of 1 fs was performed with an acceleration

of 0.1 nm ps⁻² in y-direction, on the solute organic molecule and water molecule while the position restraint was applied on the polymer backbone atoms in order to mimic the polymer chain's influence on the solute organic molecule's mobility. The diffusion process was computed from the slope of the mean-squared displacement (MSD) of the organic solute molecule as a function of time, using the Einstein relation.

$$D_i = \frac{1}{6} \lim_{t \rightarrow \infty} \frac{d}{dt} \sum_{i=1}^N \left\langle \left| \vec{r}_i(t) - \vec{r}_i(0) \right|^2 \right\rangle \quad (5.1)$$

where D_i is the diffusion coefficient of EE2 or testosterone, $\vec{r}_i(0)$ is its initial position vector, $\vec{r}_i(t)$ is its vector position at time t , and $\left\langle \left| \vec{r}_i(t) - \vec{r}_i(0) \right|^2 \right\rangle$ represents the MSD of EE2 or testosterone.

Table 5.1 Composition of complex structures systems and simulation conditions.

Molecular system	Degree of polymerization of the polymer	Number of polymer chain	Number of organic molecule	Density (g/cm³) at 300 K
PES-EE2	30	2	2	0.997 ± 0.11
PES-testosterone	30	2	2	0.999 ± 0.084
PVDF-EE2	30	2	2	1.062 ± 0.17
PVDF-testosterone	30	2	2	1.057 ± 0.41
Pa612-EE2	30	2	2	0.981 ± 0.089
PS-EE2	30	2	2	0.966 ± 0.016

In addition, 100 ns equilibrium NVT MD simulations were conducted, from which 50 ns trajectories were used to calculate short-range interaction energy ($E_{Interaction}$) between the polymer membranes and the organic solute molecule as follow,

$$E_{Interaction} = E_{Electrostatic} + E_{vdw} \quad (5.2)$$

where $E_{Electrostatic}$ is the electrostatic interaction energy and E_{vdw} is the van der Waals interaction energy.

The hydrogen bond (HB) between polymer membranes and organic solute molecule was computed with a sampling interval of 1ps. The HB autocorrelation and lifetime distribution were calculated via the integration of this function [185, 186].

$$\tau_1 = \int_0^{\infty} C(t) dt \quad (5.3)$$

where τ_1 is the overall HB lifetime and $C(t)$ is the autocorrelation function.

5.3 Results and discussions

5.3.1 Adsorption of EE2 onto polymer membranes

To study the adsorption phenomenon of EE2 molecule into the polymer membrane, the movement of EE2 from solution to the membrane interface in separate equilibrium MD simulations for 100 ns was analyzed. Initially, EE2 molecules were placed in the solution compartment. After the system reaches equilibrium, the EE2 molecules are observed to be near the membrane surface as illustrated in Figure 5.4 (a) for PES membrane and Figure 5.5 (a) for PVDF membrane, respectively. This indicates the existence of an interaction between EE2 and the flexible polymer membranes. Figure 5.4 (b) and 5.5 (b) show the corresponding number density profile of EE2 at 100 ns. The density profile was calculated by dividing the simulation box into small slices in the x-direction with $\Delta x = 0.7$ nm and 0.5 nm for PES and PVDF, respectively and estimating the number densities per nm³ for each slice. From the simulation trajectory of the PES-EE2 system, it was determined that the position of EE2 molecules after the equilibration is same to the position at $t = 100$ ns indicating the strong interaction at the specific position through the simulation time. On the other hand, for PVDF-EE2, we observed the EE2 molecules near the membrane interface after equilibrium, which is similar to the observation at $t = 100$ ns. However, the position of EE2 is not same through the simulation time indicating a weak interaction between PVDF and EE2.

Figure 5.4 (a) Snapshots of the membrane-solution interface simulation system at 100 ns showing the adsorption mechanism of four organic solute molecules of EE2 (red, pink, orange and blue) on to the polymeric membrane PES. (b) Corresponding density profile of EE2 (the black line indicates the membrane interface).

Figure 5.5 (a) Snapshots of the membrane-solution interface simulation system at 100 ns showing the adsorption mechanism of four organic solute molecules of EE2 (red, pink, orange and blue) on to the polymeric membrane PVDF. (b) Corresponding density profile of EE2 (The black line indicates the membrane interface).

5.3.2 Potential of mean force

The binding free energy calculation was conducted to analyze the strength of the interaction between the small molecules, EE2 or testosterone, with polymer membranes. SMD simulation was performed and the free energy costs of the solute molecule moving from the solution to the membrane were analyzed by computing the potential of mean forces (PMFs). The PMFs were computed from a set of umbrella sampling simulations for PES-EE2, PES-testosterone, PVDF-EE2, PVDF-testosterone, PA6-12-EE2 and PYS-EE2 with the reaction coordinates as the distance between them. This is shown in Figure 5.6, 5.7 and 5.8. The binding free energy difference was estimated as,

$$\Delta G_{bind} = (G_{max} - G_{min}) \quad (5.4)$$

where G_{max} is the free energy value when the PMF profile reaches a plateau region at long separating distances between the molecules and G_{min} is the minimum value on the PMF curve. The difference between these two free energies corresponds to ΔG_{bind} , which is the binding free energy for polymer membranes and organic solute molecules. The values of the ΔG of binding free energies for the six complexes are listed in Table 5.2.

Table 5.2 Binding free energies for PVDF, PA612, PES and PYS polymer with organic solute molecule (EE2 and Testosterone).

Complex structures	ΔG (Kcal/mol)
PVDF-EE2	-7.79 ± 0.11
PVDF - Testosterone	-6.59 ± 0.5
PA612-EE2	-12.52 ± 0.27
PES – EE2	-15.9 ± 0.61
PES – Testosterone	-13.08 ± 0.67
PYS – EE2	-10.59 ± 0.11

The highest binding free energy of organic solute molecules to the polymer corresponds to a higher interaction and vice versa. Among the different polymer materials represented on From Table 5.2, PVDF has the weakest interaction with solute molecules. The PMFs of the PVDF polymer membranes shown in Figures 5.6 (a) and 5.6 (b) support to the low interaction nature of these polymer membranes with hydrophobic and hydrophilic compounds and can be used as a control structure for adsorption [73]. The PES-EE2 binding free energy Figure 5.7 (a) is the highest of all the tested polymer membranes, which could be the reason why PES has a strong interaction and is highly efficient to retain more EE2, compared to other structures, as observed in our recent experiments [22, 112]. This highest binding free energy of the EE2 with PES might be due to the fact that the sulfone group R renders the polymer membranes polar and susceptible to forming H-bond as well as π - π interaction [73]. Furthermore the results for PES-testosterone Figure 5.7 (b) explains the experimental observation that PES has higher efficiency to retain EE2 compared to testosterone [22, 73, 112].

Figure 5.6 Potential of mean force (PMF) for EE2 and testosterone crossing PVDF membrane. (a) PVDF-EE2 with binding free energy of -7.79 ± 0.11 kcal/mol. (b) PVDF-testosterone with binding free energy of -6.59 ± 0.5 kcal/mol.

Figure 5.7 Potential of mean force (PMF) for EE2 and testosterone crossing PES membrane. (a) PES-EE2 with binding free energy of -15.59 ± 0.61 kcal/mol. (b) PES-testosterone with binding free energy of -13.08 ± 0.67 kcal/mol.

PA6-12 has a polar nature and is able to donate and accept a hydrogen molecule simultaneously, which can explain the strong interaction between PA6-12-EE2 as shown in Figure 5.8 (b). It should also be noted that the PA6-12 hydrophobicity is due to the large number of methyl sequences (-CH₂-) in PA6-12 repeat units [187]. The PYS-EE2 interaction shown in Figure 5.9 (b) may be due to the nature of the ring in polystyrene, which is electron neutral, and when this polymer interacts with EE2, which has a benzene ring with many electrons; it could contribute to the occurrence of the π - π interaction. The hydrophobic interactions also contribute to the overall interaction between PYS-EE2 [73]. The binding free energies of organic solute molecule crossing the polymer membranes are in the order PES-EE2>PES-testosterone>PA6-12-EE2>PYS-EE2>PVDF-EE2>PVDF-testosterone.

Figure 5.8 Potential of mean force (PMF) of EE2 and testosterone crossing the PYS and PA612 membranes. (a) PYS-EE2 with a binding free energy of -10.59 ± 0.11 kcal/mol. (b) PA612-EE2 with a binding free energy of -12.59 ± 0.27 kcal/mol.

5.3.3 Mean-squared displacements and diffusion coefficients

To further understand the strength of the interactions demonstrated from the binding free energies listed in Table 5.2 and the mobility of organic solute molecules, non-equilibrium MD simulations were carried out. The mean-squared displacements (MSD) were calculated first and the diffusion coefficients of organic solute molecules (EE2 or testosterone), in the presence of the polymer membranes, were computed based on the slope of the MSD curves in the linear range of the log-log scale Equation (5.1). The diffusion coefficients of organic molecules given in Table 5.3 show the mobility of organic solute molecules in the presence of polymer membranes. The position restraint was applied on the polymer membranes to mimic the fixed polymer membrane as occurs in experiments. Figure 5.9 (a) shows that the MSD of EE2 in PVDF is higher than that of EE2 in PES, which means that interaction of EE2 with PES is stronger than the interaction of EE2 with PVDF, which reduces the mobility. A similar trend is observed for testosterone MSD in PES and PVDF as shown in Figure 5.9 (b).

Figure 5.9 Mean-squared displacements (MSDs) of (a). EE2 in PES and PVDF. (b) Testosterone in PES and PVDF.

The diffusion coefficients values given in Table 5.3 shows that diffusion of the EE2 and testosterone organic solute in the PES membrane are less than that in the PVDF membrane. The lower the mobility of organic solute molecules in polymer chains means the higher binding interaction. The diffusion results show that PES has a much higher interaction strength to retain EE2 compared to testosterone which is consistent with experimental results [22, 112]. Similarly, the diffusion coefficient of the EE2 and testosterone organic solute in the PVDF membranes shows that EE2 and testosterone have a high mobility in the PVDF membrane which has weaker interactions [73]. From Table 5.3, it is observed that the diffusion coefficients of EE2 in the PES polymer is smallest compared to other structures, meaning that the mobility EE2 is reduced in the presence of the PES polymer, which explains the high binding free energy observed in Table 5.2.

Table 5.3 Simulated diffusion coefficients (D) of organic small molecule (EE2 and Testosterone) in polymer chains.

Complex structures	D ($\times 10^{-11} m^2/s$)
PVDF-EE2	19.19
PVDF - Testosterone	14.2
PA612-EE2	15.60
PES – EE2	5.78
PES - Testosterone	8.67
PYS – EE2	4.8

A similar pattern is observed between Table 5.2 (binding free energy) and Table 5.3 (diffusion coefficients). In Figure 5.10, a comparison of EE2 MSD in the membranes PES, PYS and PA6-12 are presented which shows that EE2 has less mobility in the presence of PES. The diffusion coefficients of organic solute molecules in the presence of polymer membranes are in the order; PES-EE2>PES-testosterone>PA6-12-EE2>PYS-EE2>PVDF-EE2>PVDF-testosterone. It should be highlighted that long simulation (100 ns) were performed in order to obtain accurate value of diffusion coefficients from the equilibrium trajectories.

Figure 5.10 Mean-square displacements (MSDs) of EE2 organic solute in polymeric membranes PES (red), PA612 (green) and PYS (blue).

5.3.4 Non-bonded interaction energy

Non-bonded interaction energy was computed to provide quantitative details about the interaction energy. The non-bonded (LJ and coulombic) interaction energies of solute organic molecules with polymer membranes based on short-range were calculated and listed in Table 5.4. The interaction energies presented in Table 5.4 are in a similar order to the binding free energy (PMF) and diffusion coefficients, namely, PES-EE2>PES-testosterone>PA6-12-EE2>PYS-EE2>PVDF-EE2>PVDF-testosterone. This shows that PES-EE2 has a strong non-bonded interaction, similar to PMFs binding free energies and the organic solute molecule diffusion coefficient. The interaction energy based on short-range interactions is not the most accurate approach for energetics analysis. However, these values provide an overall view of energy interaction to explain the trend of the binding free energy results shown in Table 5.2 and the diffusion coefficients shown in Table 5.3, which are consistent with the experimental results.

Table 5.4 Interaction Energy (Electrostatic and van der Waals) for PVDF, PA612, PES and PS polymer with organic solute molecule (EE2 and Testosterone).

Complex structures	Kcal/mol
PVDF-EE2	-30.26
PVDF - Testosterone	-21.941
PA612-EE2	-37.29
PES – EE2	-45.10
PES - Testosterone	-43.72
PYS – EE2	-35.941

5.3.5 Insight into atomistic interaction

Insight into the atomistic interaction mechanism was conducted by investigating the hydrogen bonding and π - π interactions. Among the tested polymer membranes PES and PA6-12 are able to form strong hydrogen bonds (HB). The formation of HB for PES-EE2 can be attributed to the sulfone group R which makes the polymer membranes polar and therefore available for H-bond with a distance of 2.9 Å as shown in Figure 5.11. The lifetime of HBs was computed using Equation (5.3). The lifetime of HB for PES-EE2 and PES-testosterone is computed from the average over all autocorrelation functions shown in Figure 5.12 and the values are 164 and 140.85 *ps* for PES-EE2 and PES-testosterone respectively. Figure 5.13 shows the number of HBs formed between PES-EE2 and PA6-12-EE2; it can be observed that PES-EE2 has more HB than PA6-12-EE2, which confirms the fact that PES has a much stronger interaction with the EE2 organic solute molecule.

Figure 5.11 A schematic of a hydrogen bond between the PES membrane and EE2 organic solute molecule. (a) PES (green) and EE2 (blue) simulation box with water molecules. (b) Enlarged view of the formation of HB with a distance of 2.9 Å, water molecules are not shown for the sake of clarity.

Figure 5.12 The hydrogen bond-time autocorrelation function for PES-EE2 (red) and PES-testosterone (green).

Figure 5.13 The number of hydrogen bonds formed between PES-EE2 and PA612-EE2 over the simulation time.

The PA6-12-EE2 HBs, shown in Figure 5.14 have distance of 3.0 Å and 3.2 Å and their occurrences can be explained by the polar nature of PA6-12, as well as its ability to act as both a hydrogen donor and acceptor [21, 73, 177]. The lifetime of HBs was 140 ps. Furthermore; the contribution of π - π interaction in the adsorption mechanism was analyzed as well for PES-EE2 and PYS-EE2 interactions. It has been observed that the formation of π - π interactions between PES-EE2 as shown in Figure 5.15, with a distance of 4.1 Å, was calculated from the two ring centroids. A similar scenario was observed for the interaction between PYS-EE2 as shown in Figure 5.16 with a distance of 4.0 Å, calculated from the two ring centroids.

Figure 5.14 A schematic of a hydrogen bond between the PA612 membrane and the EE2 organic solute molecule. (a) PES (green) and EE2 (blue) simulation box with water molecule. (b) Enlarged view of the formation of HBs with a distance of 3.0 Å and 3.0 Å; water molecule are not shown for clarity.

Figure 5.15 A snapshot of π - π interaction between PES and EE2. (a) Simulation box of EE2 in PES membrane. (b) Enlarged view of π - π interaction; the distance was computed from the centroid between the rings and is equal to 4.1 Å.

Figure 5.16 Illustration of π - π interaction between PYS and EE2. (a) Simulation box of PS-EE2. (b) Enlarged view of π - π interaction; the distance was computed from the centroid and the distance between the rings is equal to 4.0 Å.

5.4 Conclusions

A comprehensive MD simulations approach to analyze the adsorption mechanism of the organic solute molecules EE2 and testosterone using different type of polymer membranes for wastewater purification was developed. Due to the small concentration and size of EE2 (molecular weight of 296 g/mol), the removal of EE2 from wastewater is challenging. Based on the results obtained from binding free energy, diffusion coefficients, non-bonding interaction, and atomistic interaction mechanism shows that EE2 interact strongly with the PES polymer membranes. Thus, this membrane is a suitable for the removal of EE2 from wastewater. These simulation results for PES and PVDF interacting with EE2 and testosterone are consistent with the experimental findings. Our results demonstrate how different types of interaction mechanisms such as hydrogen bonds, and π - π interactions contribute to the adsorption of EE2 onto the polymer membranes. Furthermore, our computational framework has accurately predicted binding free energy between PA612 and PYS polymer membranes with EE2, which explain the use, these polymer membranes in water treatment technology. The interactions between different types of polymer membranes and organic solute molecules using our computational approach and this can be used guide water purification experiments for binding interaction prediction. The proposed comprehensive screening method for the removal of micropollutants in wastewater treatment can be extended to other polymer membranes and organic solute molecules types.

6. Concluding remarks

In this work, MD simulations were used to investigate biological and chemical systems. For the biological systems, we have developed an improved fluctuation matching model to parameterize CG-ENM force fields and structure-based model. The method is based on the relative entropy method proposed by Shell and coworkers [151] in which a non-negativity constraint and Newton-Raphson's iteration were incorporated. Similar to well-established fluctuation matching proposed by Lyman *et al.*, (2008) [110], our heterogeneous ENM parameterization using the relative entropy method with non-negativity constraints and incorporates a pairwise force constant correlation which plays an important role in the study of protein dynamics. Furthermore, fluctuation matching based on relative entropy method guarantees global optimum, with fast convergence.

For the chemical systems, we investigate the adsorption of EE2 from wastewater using molecular dynamics simulations of monomers and polymer membranes to validate experimentally observed results. First, a computational framework at the monomer level was developed to analyze the binding affinity between small organic molecules (EE2 or testosterone) and monomer structures such as PES, PS, PA6-12 and PVDF. Furthermore, analogous structures of EE2 were constructed to deepen our understanding of the major contributing factors to the binding free energy and atomistic interactions. Our findings show that the enhancement of binding affinity between PES and EE2 obtained from the binding free energy is attributed to the formation π - π interactions and hydrogen bonds, which is conformed to experimental findings. The limitation of this approach is the fact that a single monomer structure of each membrane type was used to study this

phenomenon while in real scenarios a polymer exist as a combination of many monomers into for a polymer membranes is used.

To overcome this limitation, comprehensive MD simulations and screening prediction approach framework were designed were to study adsorption of small molecules (EE2 or testosterone) onto polymer membranes such as PES, PYS, PA6-12 and PVDF. Qualitative and quantitative results were obtained from the binding free energy, mean-squared displacement and diffusion coefficients, as well as insights into the atomistic interaction, which helped determine the major contributing factors at atomistic level of the adsorption mechanism. It should be noted that the adsorption mechanisms for both the single monomer structure and the polymer membranes shows that PES and EE2 have a strong binding affinity, which may be enhanced, mainly by π - π interaction and hydrogen bonds. The results are consistent with the observed experimental results. To the best of our knowledge, these results suggest that the microporous PES hollow fiber membrane system is the most cost efficient approach can effectively and efficiently remove EE2 at low concentrations (20 μ g/L). Finally, the proposed framework can be extended to study the adsorption mechanism of other type of micropollutants.

References

1. Holmes, K.C., W. Gebhard, and W. Kabsch, *Atomic model of the actin filament*. Nature, 1990. **347**: p. 44 - 49.
2. Dietz, H. and M. Rief, *Exploring the energy landscape of GFP by single-molecule mechanical experiments*. Proceedings of the National Academy of Sciences of the United States of America, 2004. **101 (46)**: p. 16192-16197.
3. Tama, F. and C.L. Brooks III, *Symmetry, Form, and Shape Guiding Principles for Robustness in Macromolecular Machines*. Annu. Rev. Biophys. Biomol. Struct, 2006. **35**: p. 115-33.
4. Wang, Y., A.J. Rader, I. Bahar, and R.L. Jernigan, *Global ribosome motions revealed with elastic network model*. J Struct Biol, 2004. **147**: p. 302-314.
5. Lopez-Blanco, J.R. and P. Chacón, *iMODFIT: efficient and robust flexible fitting based on vibrational analysis in internal coordinates*. J. Structure Biology, 2013. **184**: p. 216-270.
6. Arkhipov, A., P.L. Freddolino, K. Imada, K. Namba, and S. Klaus, *Coarse-grained molecular dynamics simulations of a rotating bacterial flagellum*. J. Biophys, 2006. **91(12)**: p. 4589-97.
7. Chu, J.W. and G.A. Voth, *Allostery of actin filaments: Molecular dynamics simulations and coarse-grained analysis*. Proceedings of the National Academy of Sciences. U. S. A., 2005. **102(37)**: p. 13111-6.

8. Chu, J.W. and G.A. Voth, *Coarse-grained modeling of the actin filament derived from atomistic-scale simulations*. J. Biophys, 2006. **90**: p. 1572-1582.
9. Fan, J., M.G. Saunders, and G.A. Voth, *Coarse-Graining Provides Insights on the Essential Nature of Heterogeneity in Actin Filaments*. J. Biophys, 2012. **103**: p. 1334-1342.
10. Deriu, M.A., M. Soncini, T.C. Bidone, A. Redaelli, and F.M. Montevercchi, *Coarse Grain Modeling for Microtubule Mechanics*. Materials Science Forum, 2010. **638-642**: p. 629-634.
11. Kim, I.J., J. Kwon, I. Baek, H.S. Park, and S. Na, *Cofilin reduces the mechanical properties of actin filaments: approach with coarse-grained methods*. Physical Chemistry Chemical Physics, 2015. **17**: p. 8148-8158.
12. Ayton, G.S. and V.G. A., *Hybrid coarse-graining approach for lipid bilayers at large length and time scales*. J. Phys. Chem. B, 2009. **113**: p. 4413-4424.
13. Sergei, I. and G.A. Voth, *Multiscale Coarse-Graining of Mixed Phospholipid/Cholesterol Bilayers*. Journal of chemical theory and computation, 2006. **2**: p. 637.
14. Tozzini, V., *Minimalist models for proteins: a comparative analysis*. Quarterly Review of Biophysics, 2010. **43**: p. 333-371.
15. Nguyen, H.D., V.S. Reddy, and C.L. Brooks III, *Invariant polymorphism in virus capsid assembly*. Journal of the American Chemical Society, 2009. **131**: p. 2606-2614.

16. Dans, P.D., J. Walther, H. Gomez, and M. Orozco, *Multiscale simulation of DNA*. Current opinion in structural biology, 2016. **37**: p. 29-45.
17. Chandler, D., J. Stümpfer, M. Sner, S. Sceuring, and K. Schulten, *Light harvesting by lamellar chromatophores in Rhodospirillum*. Biophysical journal, 2014. **106**: p. 2503-2510.
18. Himmel, M.E., A.D. French, and R.J. Woods, *Effect of microfibril twisting on theoretical powder diffraction patterns of cellulose iβ*. Cellulose, 2014. **21**: p. 879-884.
19. Baran, L. and S. Sokolowski, *A comparison of molecular dynamics results fro two models of nanoparticles with fixed and mobile ligands in two-dimension*. Applied surface science, 2017. **396**: p. 1343-1351.
20. Lau, K.T., C. Gu, and D. Hui, *A critical review on nanotube and nanotube/nanoclay related polymer composite* . Composites: Part B, 2006. **37**: p. 425-436.
21. Han, J., W. Qiu, Z. Cao, J. Hu, and W. Gao, *Adsorption of ethinylestradiol (EE2) on polyamide 612: Molecular modeling and effects of water chemistry*. Water Research, 2013. **47**: p. 2273 - 2284.
22. Ng, C.K., C.D. Bope, A. Nalaparaju, Y. Cheng, L. Lu, R. Wang, and B. Cao, *Concentrating synthetic estrogen 17α-ethinylestradiol using microporous polyethersulfone hollow fiber membranes: Experimental exploration and molecular simulation*. Chemical Engineering, 2017. **314**: p. 80-87.

23. Juan, R.P., C.G. Boon, C.C. Keith, L. Bo, C.B. Rafael, R. Till, Y. Hang, W. Zhe, and S. Klaus, *Molecular dynamics simulations of large macromolecular complexes*. Current opinion in structural biology, 2015. **31**: p. 64-74.
24. Schlick, T., R. Collepardo-Guevara, H.L. A., S. Jung, and X. Xiao, *Biomolecular modeling and simulation: a field coming of age*. Quaterly Reviews of Biophysics, 2011. **44**: p. 191-228.
25. James, C.P., B. Rosemary, W. Wei, G. James, T. Emad, V. Elizabeth, C. Christophe, D.S. Robert, K. Laxmikan, and S. Klaus, *Scalable Molecular Dynamics with NAMD*. Journal of Computational Chemistry, 2004. **26**: p. 1781-1802.
26. Sander, P., P. Szilard, S. Roland, L. Per, B. Par, A. Rossen, R.S. Michael, C.S. Jeremy, P.M. Kasson, S. David van der, H. Berk, and L. Erik, *GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit*. J. Bioinformatics, 2013. **29**: p. 845-854.
27. Plimpton, S., *Fast Parallel Algorithm for Short-Range Molecular Dynamics*. Journal of Computational Physics, 1994. **117**: p. 1-19.
28. Case, D.A., T.E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K.M. Merz Jr, A. Onufriev, C. Simmerling, B. Wang, and R.J. Woods, *The Amber biomolecular simulation programs*. J. Computat. Chem., 2005. **26**: p. 1668-1688.
29. Muller-Plathe, F., *Coarse-graining in Polymer Simulation: From The Atomistic To The Mesoscopic Scale and Back*. . Chem. Phys Chem, 2002. **3**: p. 754-769.

30. Bernardi, R.C., M.C.R. Melo, and K. Schulten, *Enhance sampling techniques in molecular dynamics simulations of Biological systems*. Biochim. Biophys Acta, 2015. **1850(5)**: p. 872-877.
31. Atilgan, A.R., S.R. Durell, R.L. Jernigan, M.C. Demirel, O. Keskin, and I. Bahar, *Anisotropic of fluctuation dynamics of proteins with an eleastic network model*. J. Biophys, 2001. **80**: p. 505-515.
32. Bahar, I., A.R. Atilgan, and B. Erman, *Direct evaluation of thermal fluctuation in proteins using a single-parameter harmonic potential*. Fold. Des, 1997. **2**: p. 173-181.
33. Halioglu, T., I. Bahar, and B. Erman, *Gaussian dynamics of folded proteins*. Physical Review Letters, 1997. **79**: p. 3090-3094.
34. Tirion, M.M., *Large amplitude elastic motion in proteins from a single-parameter, atomistic analysis*. J. Phys Rev Lett, 1996. **77**: p. 1905-1908.
35. Clementi, C., H. Nymeyer, and J.N. Onuchic, *Topological and Energetic Factors: What Determines the Structures Details of the Transition State Ensemble and "En-route" Intermediates for Protein Folding? An investigation for Small Globular Proteins*. . Journal of Molecular Biology, 2000. **298**: p. 937-953.
36. Ueda, Y., H. Taketomi, and N. Go, *Studies on Protein folding , unfolding, and fluctuation by Computer-Simulation 2. 3-Dimensional Lattice Model of Lysozyme*. Biopolymers, 1978. **17(6)**: p. 1531-1548.
37. Brooks, B.R., J. Dusanka, and M. Karplus, *Harmonic analysis of large systems. I. Methodology*. Journal of Computational Chemistry, 1995. **16**: p. 1522.

38. Bahar, I., T.R. Lezon, A. Bakan, B. Ahmet, and I.H. Shrivastava, *Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membranes Proteins*. Chem. Rev., 2010. **110**: p. 1463-1497.
39. Lopez-Blanco, J.R. and P. Chacón, *New generation of elastic network models*. Current opinion in structural biology, 2016. **37**: p. 46-53.
40. Knots, T.A., N. Rathore, D.C. Schwartz, and J.J.J. de Pablo, *A coarse grain model for DNA*. J. Chem. Phys, 2007. **126**: p. 084901-12.
41. Suezaki, Y. and N. Gö, *Breathing mode of conformational fluctuations in globular proteins*. Int. J. Peptide Protein Res, 1975. **7**: p. 333-334.
42. Bathe, M., *A finite element framework for computation of protein normal modes and mechanical response*. Proteins: Structure, Function, and Bioinformatics, 2008. **70**: p. 1595-1609.
43. Bathe, K.J., *Finite Element Procedure*. 1996: Prentice-Hall Inc.
44. Kim, D.N., C.T. Nguyen, and M. Bathe, *Conformational dynamics of supramolecular protein assemblies*. Journal of Structural Biology, 2011: p. 261-270.
45. Cook, D.R., S.D. Malkus, E.M. Plesha, and J.R. Witt, *Concepts and applications of finite element analysis.*, ed. F. Edition. 2002: John Wiley & Sons, Inc.
46. Ming, D.K.Y., L.M. A., Z. Huang, and M.J. P., *How to describe protein motion without amino acid sequence and atomic coordinates*. Proceedings of the National Academy of Sciences. U. S. A., 2002. **99**: p. 8620-8625.

47. Chu, J.W., S. Izvekov, and G.A. Voth, *The multiscale challenge for biomolecular systems: coarse-grained modeling*. Molecular Simulation, 2006. **32**(3-4): p. 211-218.
48. Chu, J.W. and G.A. Voth, *Coarse-grained modeling of the actin filament derived from atomistic-scale simulations*. J. Biophys, 2006. **90**(5): p. 1572-1582.
49. Erman, B., *The Gaussian network model: precise predications of residue fluctuations and application to bending problems*. Biophys. J, 2006. **91**: p. 3589-3599.
50. Lyman, E., J. Pfandtner, and G.A. Voth, *Systematic Multiscale Parameterization of Heterogeneous Elastic Network Models of Proteins*. J. Biophys, 2008. **95**: p. 4183-4192.
51. Xia, F. and L. Lu, *Multiscale Coarse-Graining via Normal Mode Analysis*. Journal of chemical theory and computation, 2012. **8**(11): p. 4797-806.
52. Xia, F., D. Tong, Y. Lifeng, W. Dayong, C.H.H. Steven, P. Koehl, and L. Lu, *Identifying Essential Pairwise Interactions in Elastic Network Model Using the Alpha Shape Theory*. J. Comput. Chem, 2014. **35**: p. 1111-1121.
53. Xia, F., D. Tong, and L. Lu, *Robust Heterogeneous Anisotropic Elastic Network Model Precisely Reproduces the Experimental B-factors of Biomolecules*. J. Chem. Theory Comput, 2013. **9**(8): p. 3704-14.
54. Fan, J., M.G. Saunders, E.G. Haddadian, K.F. Freed, E.M. De La Cruz, and G.A. Voth, *Molecular Origins of Cofilin-Linked Changes in Actin Filament Mechanics*. Journal of Molecular Biology, 2013. **425**: p. 1225-1240.

55. Bope, C.D., D. Tong, X. Li, and L. Lu, *Fluctuation matching approach for elastic network model and structure-based model of biomacromolecules*. Progress in Biophysics and Molecular Biology, 2017: p. 1-13.
56. Hori, N. and S. Takada, *Coarse-Grained Structure-Based Model for RNA-Protein Complexes Developed by Fluctuation Matching*. Journal of chemical theory and computation, 2012. **8**(9): p. 3384-94.
57. Li, W., H. Yoshii, N. Hori, T. Kameda, and S. Takada, *Multiscale methods for protein folding simulations*. Methods, 2010. **52**(1): p. 106-14.
58. Onuchic, J.N. and P.G. Wolynes, *Theory of protein folding*. J. Current opinion in structural biology, 2004. **14**(1): p. 70-5.
59. Sutto, L., I. Mereu, and F.L. Gervasio, *A Hybrid All-Atom Structure-Based Model for Protein Folding and Large Scale Conformational Transitions*. Journal of Chemical Theory and Computation, 2011. **7**(12): p. 4208-4217.
60. Whitford, P.C., J.K. Noel, S. Gosavi, A. Schung, K.Y. Sanbonmatsu, and J.N. Onuchic, *An All-atom Structure-Based Potential for Proteins: Bridging Minimal Models with All-atom Empirical Forcefields*. Proteins: Structure, Function, and Bioinformatics, 2009. **75**: p. 430-441.
61. Karanicolas, J. and C.L. Brooks, *The origins of asymmetry in the folding transition states of protein L and protein G*. Protein Science, 2002. **11**(10): p. 2351-2361.
62. Li, W.F., P.G. Wolynes, and S. Takada, *Frustration, specific sequence dependence, and nonlinearity in large-amplitude fluctuations of allosteric proteins*.

- Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(9): p. 3504-3509.
63. Voth (Ed.), G.A., *Coarse-Graining of Condensed Phase and Biomolecular Systems*. 2008, Boca Raton: CRC Press.
64. Izvekov, S. and G.A. Voth, *A Multiscale Coarse-Graining Method for Biomolecular Systems*. The Journal of Physical Chemistry B 2005. **109**: p. 2469-2473.
65. Noid, W.G., J.W. Chu, G.S. Ayton, V. Krishna, S. Izvekov, G.A. Voth, A. Das, and H.C. Andersen, *The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models*. Journal of Chemical Physics, 2008. **128**(24).
66. Chaimovich, A. and M.S. Shell, *Coarse-graining errors and numerical optimization using a relative entropy framework*. Journal of Chemical Physics, 2011. **134**(9): p. 094112.
67. Shell, M.S., *Systematic coarse-graining of potential energy landscapes and dynamics in liquids*. Journal of Chemical Physics, 2012. **137**(8): p. 084503.
68. Deriu, M.A., S. Enemark, M. Soncini, F.M. Montevercchi, and A. Redaelli, *Tubulin: from atomistic structure to supramolecular mechanical properties*. J. Mater Sci, 2007. **42**(8864-8872).
69. Report, H.D., *United Nations Development Programme (UNDP), 2006, ch 4*. 2006.
70. Coping with water scarcity, C.o.t.T.-f.C., *United Nations (UN) Water. Food and Agricultural Association (FAO)*. 2007.

71. Zhang, T.C., R.Y. Surampalli, S. Vigneswaran, R.D. Tyagi, S.L. Ong, and C.M. Kao, *Membrane Technology and Environmental Applications*. 2012: American Society of Civil Engineers.
72. Caetano, A., M.N. De Pinho, E. Drioli, and H. Muntau, *Membrane Technology: Applications to Industrial Wastewater Treatment*. 1995: Kluwer Academic.
73. Schäfer, A.I., I. Akanyeti, and J.C. Semião, *Micropollutant sorption to membrane polymers: A review of mechanisms for estrogens*. *Advances in Colloid and Interface Science*, 2011. **164**: p. 100-117.
74. Silva, C.P., M. Otero, and V. Esteves, *Processes for elimination of estrogenic steroid hormones from water: A review*. *Environmental Pollution*, 2012. **165**: p. 38-58.
75. Rajindar, S., *Membrane Technology and Engineering for Water Purification: Application, Systems Design and Operation*. Second ed. 2015: Elsevier.
76. Bergman, A., J.J. Heindel, S. Jobling, K.A. Kidd, and T. Zoeller, *State of the science of endocrine disrupting chemical 2012*. United Nations Environmental Programme and the World Health Organization, 2013.
77. Dickenson, E.R.V., S.A. Snyder, D.L. Sedlak, and J.E. Drewes, *Indicator compounds for assessment of wastewater effluent contributions to flow and water quality*. *Water Research*, 2011. **45(3)**: p. 1199-1212.
78. Loos, R., R. Carvalho, D.C. Antonio, S. Comero, G. Locoro, S. Tavazzi, B. Paracchini, M. Ghiani, T. Lettieri, L. Blaha, B. Jarosova, S. Voorspoels, K. Servaes, P. Haglund, J. Fick, R.H. Lindberg, D. Schwesig, and B.M. Gawlik, *EU-*

- wide monitoring survey on emerging polar organic contaminants in wastewater treatment plant effluents*. Water Research, 2013. **47(17)**: p. 6475-6487.
79. Jobling, S., M. Nolan, C.R. Tyler, G. Brighty, and J.P. Sumpter, *Widespread sexual disruption in wide fish*. Environ. Sci. Technol, 1998. **32(17)**: p. 2498-2506.
 80. Brodin, T., J. Fick, M. Jonsson, and J. Klaminder, *Dilute concentrations of a psychiatric drug alter behavior of fish from natural* . Science, 2013. **339(6121)**: p. 814-815.
 81. Miège, C., J.M. Choubert, L. Ribeiro, M. Eusèbe, and M. Coquery, *Fate of pharmaceuticals and personal care products in wastewater treatment plants - Conceptions of a database and first results*. Environ. Pollut., 2009. **157(5)**: p. 1721-1726.
 82. Luo, Y., W. Guo, N.H. H., L.D. Nghiem, F.I. Hai, J. Zhang, S. Liang, and X.C. Wang, *A review on the occurrence of micropollutants in the aquatic environment and their fate and removal during wastewater treatment*. Science Total Environment, 2014. **473-474**: p. 619-641.
 83. Carballa, M., F. Omil, J.M. Lema, M. Llompart, C. Carcía-Jares, I. Rodriguez, M. Gómez, and T. Ternes, *Behavior of pharmaceuticals, cosmetics and hormones in a sewage treatment plant*. Water Research, 2004. **38(12)**: p. 2918-2926.
 84. Zorita, S., L. Mårtensson, and L. Mathiasson, *Occurrence and removal of pharmaceuticals in municipal sewage treatment system in the south of Sweden*. Sci. Total Environ., 2009. **407(8)**: p. 2760-2770.

85. Clouzot, L., B. Marrot, P. Doumenq, and N. Roche, *17 α - Ethinylestradiol: An Endocrine Disrupter of Great Concern. Analytical Methods and Removal Process Anpllyed to Water Purifiaction. A Review*. Environmental Prograss, 2008. **27**: p. 383 - 396.
86. Harrison, P.T.C., P. Holmes, and C.D.N. Humfrey, *Reproductive health in humans and wildlife: are adverse trends associated with environmental chemical exposures?* The Science of the Total Environment, 1997. **205**: p. 97-106.
87. Iguchi, T., H. Watanabe, and Y. Katsu, *Developmental Effects of Estrogenic Agents on Mice, Fish, and Frogs: A Mini-Review*. Hormones and Behavior, 2001. **40**: p. 248 - 251.
88. Taylor, M.R. and P.T.C. Harrison, *Ecological Effects of Endocrine Disruption: Current Evidence and Research Priorities*. Chemosphere, 1999. **39**: p. 1237 - 1248.
89. Bhandari, R.K., S.L. Deen, D.K. Holliday, C.M. Jandegian, C.D. Kassotis, S.C. Nagel, D.E. Tillitt, F.S. vom Saal, and C.S. Rosenfeld, *Effect of the environment estrogenic contaminants bisphenol A and 17 α -ethinyl estradiol on sexual development and adult behaviors in aquatic wildlife species*. General and Comparative Endocrinology, 2015. **214**: p. 195 - 219.
90. Caliman, F.A. and M. Gavrilescu, *Pharmaceuticals, Personal Care Products and Endocrine Disrupting Agent in the Environment- A Review*. Clean, 2009. **37**: p. 277 - 303.

91. Koch, C.A. and E. Diamanti-Kandarakis, *Introduction to Endocrine Disrupting Chemicals -is it time to act?* Rev Endocr Metab Disord, 2015. **16**: p. 269 - 270.
92. Sweeney, M.F., N. Hasan, A.M. Soto, and C. Sonnenschein, *Environmental endocrine disruptors: Effects on the human male reproductive system.* Rev Endocr Metab Disord, 2015. **16**: p. 341 - 357.
93. Roy, J.R., S. Chakraborty, and T.R. Chakraborty, *Estrogen-like endocrine disrupting chemicals affecting puberty in humans- a review.* Medical Science Monitor, 2009. **15**: p. RA137 - 145.
94. Mendes, J.J.A., *The endocrine disruptors: a major medical challenge.* Food and Chemical Toxicology, 2002. **40**: p. 781 - 788.
95. Frye, C., E. Bo, G. Calamandrei, L. Calzã, F. Dessì-Fulgheri, M. Fernández, L. Fusani, O. Kah, M. Kajta, Y. Le Page, H.B. Patisaul, A. Venerosi, A.K. Wojtowicz, and G.C. Panzica, *Endocrine Disrupters: A review of some sources, effects, and mechanisms of actions on behavior and neuroendocrine systems.* . Neuroendocrinol, 2012. **24(1)**: p. 144-159.
96. Racz, L.A. and R.K. Goel, *Fate and removal of estrogens in municipal wastewater.* Journal of Environmental Monitoring, 2010. **12**: p. 58-70.
97. Datapharm. *Ethinylestradiol.* EMC Medecine Guides (UK) 2008; Available from: <http://www.medecine.org.uk/guides/ethinylestradiol>.
98. USFDA. *U.S Food and Drug Administration, Drug Databases - Drugs@FDA Approved Drug Products.* 2012; Available from: <http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm>.

99. Ding, D., L. Xu, H. Fang, H. Hong, R. Perkins, S. Harris, E.D. Bearden, L. Shi, and W. Tong, *The EDKB: an established knowledge base for endocrine disrupting chemicals*. BMC Bioinformatics, 2010. **11**.
100. USFDA. *U.S. Food and Drug Administration and National Center for Toxicological Research, Endocrine Disruptor Knowledge Base (EDKB)*. 2012; Available from: <http://www.fda.gov/ScienceResearch/BioinformaticsTools/EndocrineDisruptorKnowledgebase/>.
101. Carvalho, R.L., L.A. Ceriani, A. Ippolito, and T. Lettieri, *Development of the First Watch List Under the Environmental Quality Standards Directive*. European Commission Joint Research Center, Ispra, Italy, 2015.
102. Flores-Valverde, A.M., J. Horwood, and E.M. Hill, *Disruption of the Steroid Metabolome in Fish Caused by Exposure to the Environmental Estrogen 17 α -Ethinylestradiol*. Environmental Science & Technology, 2010. **44(9)**: p. 3552-3558.
103. Coe, T.S., P.B. Hamilton, D. Hodgson, G.C. Paull, J.R. Stevens, K. Sumner, and C.R. Tyler, *An Environmental Estrogen Alters Reproductive Hierarchies, Disrupting Sexual Selection in Group-Spawning Fish*. . Environmental Science & Technology, 2008. **42(13)**: p. 5020-5025.
104. Desbrow, C., E.J. Routledge, G.C. Brighty, J.P. Sumpter, and M. Waldock, *Identification of Estrogenic Chemical in STW Effluent. 1. Chemical Fraction and*

- in Vitro Biological Screening*. Environmental Science & Technology, 1998. **32**: p. 1549 - 1558.
105. Braga, O., G.A. Smythe, A.I. Schäfer, and A.J. Feitz, *Steroid estrogens in primary and tertiary wastewater treatment plants*. Environmental Science & Technology, 2005. **39 (9)**(3351 - 3358).
 106. Johnson, A.C. and J.P. Sumpter, *Removal of Endocrine-Disrupting Chemicals in Activated Sludge Treatment Works*. Environ. Sci. Technol, 2001. **35 (24)**: p. 4697 - 4703.
 107. Gomes, R.L., M.D. Scrimshaw, and J.N. Lester, *Determination of endocrine disrupters in sewage treatment and receiving waters*. Trends in Analytical Chemistry, 2003. **22 (10)**: p. 697 - 707.
 108. Leech, D.M., M.T. Snyder, and R.G. Wetzel, *Natural organic matter and sunlight accelerate the degradation of β -17 estradiol in water*. Science of the Total Environment, 2009. **407**(2087 - 2092).
 109. Mulder, M., *Basic principles of membrane technology.*, ed. 2. 1996: Kluwer Academic.
 110. Lyman, E., J. Pfaendtner, and G.A. Voth, *Systematic Multiscale Parameterization of Heterogeneous Elastic Network Models of Proteins*. Biophysical journal, 2008. **95**: p. 4183-4192.
 111. Chaimovich, A. and M.S. Shell, *Anomalous waterlike behavior in spherically-symmetric water models optimized with the relative entropy*. Phys. Chem. Chem. Phys, 2009. **11**: p. 1901-1915.

112. Ng, C.K. and B. Cao, *What Exactly Are You Filtering Out?* Environ. Sci. Technol, 2015. **49**: p. 5259-5260.
113. Leach, A.R., *Molecular Modelling: principles and applications*. 2001: Prentice Hall.
114. Ponder, J.W. and D.A. Case, *Force fields for protein simulations*. Advances in Protein Chemistry, 2003. **66**: p. 27-85.
115. Hinsen, K., *Analysis of Domain Motions by Approximate Normal Mode Calculation*. Proteins: Structure, Function, and Genetics, 1998. **33**: p. 417-429.
116. P., J. and H.U.H. E, *Modeling Structural Flexibility of Proteins with Go-Models*. J. Chem. Theory Comput, 2012. **8**(6): p. 2127-2133.
117. Lammert, H., A. Schug, and J. Onuchic, N., *Robustness and generalization of structures-based models fro protein folding and function*. Proteins: Structure, Function and Bioinformatics, 2009. **77**: p. 881-891.
118. Noel, J.K. and J.N. Onuchic, *The Many Faces of Structures-Based Potentials: From Protein Folding Landscapes to Structural Characterization of Complex Biomolecules*, ed. D. N. 2012.
119. Sugita, Y., A. Mitsutake, and Y. Okamoto, *Generalized-ensemble algorithms for folding simulations.*, ed. e.b.W. Janke. 2008, Berlin: Springer-Verlag.
120. Frenkel, D. and B. Smit, *Understanding molecular simulation: From algorithm to applications.*, ed. ed. 2002, San Diego, California: Academic Press.

121. Monticelli, L. and E. Salonen, *Biomolecular simulations: Methods and Protocols, Methods in Molecular Biology*. Springer Science + Business Media New York, 2013. **924**.
122. Kästner, J., *Umbrella sampling*. WIREs Computational Molecular Science, 2011. **1**: p. 932-942.
123. Torrie, G.M. and J.P. Valleau, *Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling*. J. Comput. Phys., 1977. **23**: p. 187-199.
124. Izrailev, S., B. Stepaniants, D. Isralewitz, H. Lu, F. Molnar, W. Wriggers, and K. Schulten, *Steered Molecular Dynamics*. Springer-Verlag: Berlin, 1998. **Vol. 4**.
125. Hub, J.S., B.L. de Groot, and D. Van Der Spoel, *g_wham-A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates*. J. Chem. Theory Comput, 2010. **2010**: p. 3713-3720.
126. Reith, D., M. Putz, and F. Muller-Plathe, *Deriving Effective Mesoscale Potentials from Atomistic Simulations*. Journal of Computational Chemistry, 2003. **24**: p. 1624-1636.
127. Zhang, Z., L. Lanyuan, W. Noid, G., G. Krishna, J. Pfandtner, and G.A. Voth, *A systematic methodology for defining coarse-grained sites in large biomolecules*. J. Biophys, 2008. **95**: p. 5073-5083.
128. S., M.G. and G.A. Voth, *Coarse-graining of multiprotein assemblies*. Current opinion in structural biology, 2012. **22**(2): p. 144-50.

129. Grafmuller, A., E.G. Noya, and G.A. Voth, *Nucleotide-Dependent Lateral and Longitudinal Interactions in Microtubules*. Journal of Molecular Biology, 2013. **425**: p. 2232-2246.
130. Zhang, Z., Y.K. Sanbonmatsu, and G.A. Voth, *Key Intermolecular Interaction in the E. coli 70S Ribosome Revealed by Coarse-grained Analysis*. Journal of the American Chemical Society, 2011. **133**: p. 16828-16838.
131. Deriu, M.A., M. Soncini, M. Orsi, M. Patel, J.W. Essex, F.M. Montecvecchi, and A. Redaelli, *Anisotropic elastic network modeling of entire microtubules*. Biophysical journal, 2010. **99**(7): p. 2190-9.
132. Bahar, I. and A. Rader, *Coarse-grained normal mode analysis in structural biology*. Current opinion in structural biology, 2005. **15**: p. 586-592.
133. Dykeman, E.C. and O.F. S., *Normal mode analysis and application in biological physics*. J. Phys.:Condens. Matter, 2010. **22** **423202**.
134. Weaver, L.H. and B.W. Matthews, *Structure of Bacteriophage T4 Lysozyme Refined at 1.7 Å Resolution*. J. Mol. Biol, 1987. **193**: p. 189 - 199.
135. Müller, C.W. and G.E. Schulz, *Structure of Complex Between Adenylate Kinase from Escherichia coli and the Inhibitor Ap5A Refined at 1.9 Å Resolution A model for Catalytic Transition State*. J. Mol. Biol, 1992. **224**: p. 159 - 177.
136. Li, L., J.H. Brown, L. Reshtnikova, A. Blazsek, L. Farkas, N. Nyltray., and C. Cohen, *Visualization of an unstable coiled coil from the scallop myosin rod*. Nature, 2003. **424**: p. 341 - 345.

137. Kuriyan, J. and W.I. Weis, *Rigid protein motion as a model for crystallographic temperature factors*. Proceedings of the National Academy of Sciences. U. S. A., 1991. **88**: p. 2773-2777.
138. Yang, L.W., X. Lui, C.J. Jursa, M. Holliman, A.J. Rader, H.A. Karimi, and I. Bahar, *iGNM: a database of protein functional motions based on Gaussian Network Model*. Bioinformatics, 2015. **21**: p. 2978-2987.
139. Eyal, E., L.W. Yang, and I. Bahar, *Anisotropic network model: systematic evaluation and a new web interface*. Bioinformatics, 2006. **22**: p. 2619-2627.
140. Kundu, S., J.S. Melton, D.C. Sorensen, and J.G.N. Phillips, *Dynamics of Proteins in Crystals: Comparison of Experiment with Simple Models*. J. Biophys, 2002. **83**: p. 723-732.
141. Schomaker, V. and K.N. Trueblood, *On the Rigid-Body Motion of Molecules in crystals*. Acta Crystallogr. B, 1968. **24**: p. 63-76.
142. Harata, K., Y. Abe, and M. Muraki, *Crystallographic Evaluation of Internal Motion of Human α -Lactalbumin Refined by Full-matrix Least-squares Methods*. Journal of Molecular Biology, 1999. **287**: p. 347-358.
143. Song, G. and R.L. Jernigan, *vGNM: A Better Model for Understanding the Dynamics of Proteins in Crystals*. J. Mol. Biol, 2007. **369**: p. 880-893.
144. Teeter, M.M., *Water structure of a hydrophobic protein at atomic resolution: Pentagon rings of water molecules in crystals of crambin*. Proceedings of the National Academy of Sciences. U. S. A., 1984. **81**: p. 6014 - 6018.

145. Wlodawer, A., J. Walter, and R. Huber, *Structure of bovine pancreatic trypsin inhibitor: Results of joint neutron and X-ray refinement of crystal form II*. J. Mol. Biol, 1984. **180**: p. 301 - 329.
146. Spinelli, S., Q.Z. Lui, P.M. Alzari, P.H. Hirel, and R.J. Poljak, *The Three-dimensional structure of the aspartyl protease from the HIV-1 isolate BRU*. Biochimie, 1991. **73**: p. 1391 - 1396.
147. Kullback, S. and R.A. Leibler, *On information and sufficiency*. Ann. Math. Stat., 1951. **22(79)**.
148. Carmichael, S.P. and M.S. Shell, *A new multiscale algorithm and its application to coarse-grained Peptide models for self-assembly*. J. Phys. Chem. B, 2012. **116(29)**: p. 8383-8393.
149. Press, W.H., S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, ed. S. Edition. 1992: Cambridge University Press.
150. Yang, L., G. Song, A. Carriquiry, and R.L. Jernigan, *Close Correspondance between the Essential Protein Motions from Principal Component Analysis of Multiple HIV-1 Protease structures and Elastic Network Modes*. Structure, 2008.
151. Shell, M.S., *Systematic coarse-graining of potential energy landscapes and dynamics in liquids*. J. chem. phys, 2012. **137(8)**: p. 084503.
152. Chaimovich, A. and M.S. Shell, *Coarse-graining errors and numerical optimization using a relative entropy framework*. J. chem. phys, 2011. **134(9)**: p. 094112.

153. Accelrys, *Discovery Studio Modeling Environment, Release 8.0*. San Diego: Accelrys Software Inc, 2016.
154. Perdrew, J.P. and Y. Wang, *Accurate and simple analytic representation of the electron-gas correlation energy*. J. Phys. Chem. B, 1992. **45**: p. 13244 - 13249.
155. Sander, P., P. Szilard, S. Roland, L. Per, B. Par, A. Rossen, R.S. Michael, C.S. Jeremy, P.M. Kasson, S. David van der, H. Berk, and L. Erik, *GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit*. J. Bioinformatics, 2013. **29**: p. 845-854.
156. Jorgensen, W.L. and J. Tirado Rives, *The OPLS [optimized potential for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin*. J. Am. Chem. Soc, 1988. **110**: p. 1657-1666.
157. Berendsen, H.J.C., J.P.M. Postma, W.F. van Gunsteren, and H. J., *Interaction models for water in relation to protein hydration*. . In Intermolecular Forces, B. Pullman (ed), Dordrecht, The Netherlands, 1981: p. 331-332.
158. Zimmermann, K., *ORAL: All Purpose Molecular Mechanics Simulator and Energy Minimizer*. J. Comput. Chem, 1991. **12**: p. 310-319.
159. Bussi, G., D. Donadio, and M. Parrinello, *Canonical sampling through velocity rescaling*. J. Chem. Phys, 2007. **126**: p. 014101-7.
160. Berendsen, H.J.C., J.P.M. Postma, W.F. Van Gunsteren, A. DiNola, and J.R. Haak, *Molecular dynamics with coupling to an external bath*. J. Chem. Phys, 1984. **81**: p. 3684-3690.

161. Darden, T., D. York, and L. Pedersen, *Particle mesh Ewald: An $N \log(N)$ method for Ewald sums*. J. Chem. Phys, 1993. **98**: p. 10089-10092.
162. Essmann, U., L. Perera, M.L. Berkowitz, T. Darden, H. Lee, and L. Pedersen, *A smooth particle mesh Ewald method*. J. Chem. Phys, 1995. **103**: p. 8577-8593.
163. Hess, B., H. Bekker, H.J.C. Berendsen, and J.G.E.M. Fraaije, *LINCS: a Linear Constraint Solver for Molecular Simulation*. J. Comput. Chem, 1997. **18**: p. 1463-1472.
164. Patey, G.N. and J.P. Valleau, *The free energy of spheres with dipoles: Monte carlo with multistage sampling*. J. Chem. Phys. Lett, 1973. **21**: p. 297-300.
165. Torrie, G.M. and J.P. Valleau, *Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling*. J. Comput. Phys., 1977. **23**: p. 187-199.
166. Hub, S.J., L.B. De Groot, and D. Van der Spee, *g_wham- A Free Weighted Histogram Analysis Implemented Including Robust Error and Autocorrelation Estimation*. J. Chem. Theory Comput, 2010. **6**: p. 3713-3720.
167. Lui, R., J.L. Zhou, and W. A., *Simultaneous determination of endocrine disrupting phenolic compounds and steriods in water by solid-phase extraction-gas chromatography-mass spectrometry*. Journal of Chromatography A, 2004. **1022**: p. 179 - 189.
168. Zhou, J.L., R. Lui, A. Wilding, and A. Hibberd, *Sorption of Selected Endocrine Disrupting Chemicals to Different Aquatic Colloids*. Environ. Sci. Technol, 2007. **41**: p. 206 - 213.

169. Baronti, C., R. Curini, G. D'Ascenzo, A. Corcia, A. Gentili, and R. Samperi, *Monitoring Natural and Synthetic Estrogens at Activated Sludge Sewage Treatment Plants and in a Receiving River Water*. Environmental Science & Technology, 2000. **34 (24)**: p. 5059 - 5066.
170. Conroy, O., A.E. Sáez, D. Quanrud, W. Ela, and R.G. Arnold, *Changes in estrogen/anti-estrogen activities in ponded secondary effluent*. Science of the Total Environment, 2007. **382**: p. 311 - 323.
171. Loos, R., G. Locoro, S. Comero, S. Contini, D. Schwesig, F. Werres, P. Balsa, O. Gans, S. Weiss, L. Blaha, M. Bolchi, and B.M. Gawlik, *Pan-European survey on the occurrence of selected polar organic persistent pollutants in ground water*. Water Research, 2010. **44**: p. 4115 - 4126.
172. Loos, R., B.M. Gawlik, G. Locoro, E. Rimaviciute, S. Contini, and G. Bidoglio, *EU-wide survey of polar organic persistent pollutants in European river waters*. Environmental Pollution, 2009. **157**: p. 561 - 568.
173. Williams, R.J., A.C. Johnson, S.J.J. L., and R. Kanda, *Steroid Estrogens Profiles along River Stretches Arising from Sewage Treatment Works Discharges*. Environ. Sci. Technol, 2003. **37**: p. 1744 - 1750.
174. Fine, D.D., G.P. Breidenbach, P.T. L., and S.R. Hutchins, *Quantitation of estrogens in ground water and swine lagoon samples using solid-phase extraction, pentafluorobenzyl/trimethylsilyl derivatizations and gas chromatography-negative ion chemical ionization tandem mass spectrometry*. Journal of Chromatography A, 2003. **1017**: p. 167 - 185.

175. Degremont, G., *The Water Treatment Handbook*. tenth ed. 2007, Paris: Lavoisier.
176. Worch, E., *Adsorption Technology in Water Treatment: Fundamentals, Processes, and Modeling*. 2012, Germany: De Gruyter.
177. Han, J., W. Qiu, S. Meng, and W. Gao, *Removal of ethinylestradiol (EE2) from water via adsorption on aliphatic polyamides*. *Water Research*, 2012. **46**: p. 5715-5724.
178. Schöntag, J.M., B.S. Pizzolatti, V.H. Jangada, and F. de Souza, *Water quantity produced by polystyrene granules as a media filter on rapid filters*. *Journal of Water Process Engineering*, 2015. **5**: p. 118-126.
179. Han, J., W. Qiu, J. Hu, and W. Gao, *Chemisorption of estrone in nylon microfiltration membranes: Adsorption mechanism and potential use for estrone removal from water*. *Water Research*, 2012. **46**: p. 873-881.
180. Perdrew, J.P. and Y. Wang, *Accurate and simple analytic representation of the electron-gas correlation energy* *J. Phys. Chem. B*, 1992. **45**: p. 13244-13249.
181. Brunello, G., S.G. Lee, S.S. Jang, and Y. Qi, *A molecular dynamics simulation study of hydrated sulfonated poly(ether ether ketone) for application to polymer electrolyte membrane fuel cells: Effect of water*. *J. Renewable and Sustainable Energy*, 2009. **1**: p. 033101.
182. Jang, S.S., V. Molinero, T. Çağm, and W.A. Goddard III, *Nanophase-Segregation and Transport in Nafion 117 from Molecular Dynamics Simulations: Effect of Monomeric Sequence*. *J. Phys. Chem. B*, 2004. **108**: p. 3149-3157.

183. Patey, G.N. and J.P. Valleau, *The free energy of spheres with dipoles: Monte carlo with multistage sampling*. J. Chem. Phys. Lett, 1973. **21**: p. 297-300.
184. Berendsen, H.J.C., J.P.M. Postma, W.F. Van Gunsteren, and J. Hermans, *Interaction models for water in relation to protein hydration*. Intermolecular Forces, B. Pullman (ed), Dordrecht, The Netherlands, 1981: p. 331-342.
185. Van Der Spoel, D., P.J. Van Maaren, P. Larson, and N. Tîmneanu, *Thermodynamics of Hydrogen Bonding in Hydrophilic and Hydrophobic Media*. J. Phys. Chem. B, 2006. **110**: p. 4393-4398.
186. Luzar, A., *Resolving the hydrogen bond dynamics conundrum*. J. Chem. Phys, 2000. **113**: p. 10663-10675.
187. Kohan, M.I., *Nylon Plastic Handbook*. 1995, New York, USA: Hanser Gardner.

Publications

1. **Christian Domilongo Bope**, Dudu Tong, Xiuting Li, Lanyuan Lu. Fluctuation matching approach for elastic network model and structure-based model of biomacromolecules, *Progress in Biophysics and Molecular Biology* **128** (2017), 100-112.
2. Chun Kiat Ng, **Christian Domilongo Bope**, Anjaiah Nalaparaju, Yuan Cheng, Lanyuan Lu, Rong Wang, Bin Cao. Concentrating synthetic estrogen 17 α -ethinyl estradiol using microporous polyether sulfone hollow polymer membranes: Experimental exploration and molecular simulation. *Chemical Engineering Journal* **314** (2017), 8087.
3. **Christian Domilongo Bope**, Anjaiah Nalaparaju, Chun Kiat Ng, Yuan Cheng, Lanyuan Lu. Molecular Simulation Study for Understanding and Comparing the Interaction of 17 α -ethinylestradiol (EE2) with Polymer Membranes in Wastewater Purification. *Journal of Molecular Simulation* (submitted).