The Pennsylvania State University
The Graduate School
Department of Chemical Engineering

MOLECULAR DESIGN IN CHEMICAL
AND BIOLOGICAL SYSTEMS

A Thesis in
Chemical Engineering
by
Andreas Lehmann

© 2004 Andreas Lehmann

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2004
We approve the thesis of Andreas Lehmann.

Costas D. Maranas  
Professor of Chemical Engineering  
Thesis Adviser  
Chair of Committee

J. Larry Duda  
Professor of Chemical Engineering

Kristen Fichthorn  
Merrell R. Fenske Chair Professor of  
Chemical Engineering and Professor of Physics

Barbara J. Garrison  
Shapiro Professor of Chemistry

Andrew L. Zydney  
Interim Department Head, Chemical Engineering  
Endowed Bio-Chair and Professor of Chemical Engineering
Abstract

The focus of this thesis is the development of methods for computational molecular design with the goal to aid the experimentalist and reduce laboratory cost. Two topical complexes are treated: small molecules that are of interest in the chemical industry and much larger protein molecules in biological systems. These two types of molecules require the use of different structure-property relationships and methodical approaches of how the search is conducted.

In the first part, the combination of quantum chemical methods with optimization techniques for molecular design of small molecules is examined. A hydrofluorocarbon refrigerant design example and a solvent design example illustrate the proposed framework. The hydrofluorocarbon compounds are optimized for their heats of formation and the potential solvents are searched for capacity, selectivity and environmental safety. In both examples, a genetic algorithm is applied to generate and screen candidate molecules. The molecular properties are evaluated using a combination of quantum chemical calculations and group contribution methods. The feasibility of the proposed approach for small molecules is assessed and it is found that establishing a proper trade-off between accuracy of the quantum chemical method and computational expense is vital.

In the second part, optimization techniques are combined with force field methods for protein design. The focus here lies on making a clear distinction in how to account for different protein properties such as protein stability and function. A framework is proposed for predicting promising mutants of a protein with improved binding functionality
for a known ligand of this protein. The method is applied to two design examples. The first example validates the method by comparing computationally designed mutations with experimental data. In the second example, predictions are given for promising multiple mutations of the plant protein concanavalin A, which will lead to enhanced binding of glucose.
# Table of Contents

List of Tables .................................................. viii

List of Figures .................................................. x

Acknowledgments .................................................. xii

Chapter 1. Introduction .......................................... 1

Chapter 2. Small-Molecule Design Background ................. 7

2.1 Quantum Chemistry .......................................... 7

2.2 Basis Functions and Basis Sets ............................. 8

2.3 Levels of Theory ............................................ 10

2.4 Initial Geometry and Molecular Representation ............ 11

2.5 Optimization Procedure: A Genetic Algorithm .......... 12

Chapter 3. Molecular Design using Quantum Chemical Calculations for Property Estimation ........................................ 18

3.1 Optimization Procedure ..................................... 18

3.2 Case Study 1: Design of Hydrofluorocarbons .............. 22

3.2.1 Problem Definition ...................................... 22

3.2.2 Results of Case Study 1 ................................. 33

3.3 Case Study 2: Solvent Design ............................... 41

3.3.1 Problem Definition ...................................... 41
3.3.2 Results of Case Study 2 ........................................ 47
3.4 Summary and Conclusions ........................................... 51

Chapter 4. Protein Design Background .................................. 54

4.1 Protein Structure and Rotamer Libraries .......................... 55
4.2 Energy Function for Protein Stability ............................... 58
4.3 Calculating Protein-Ligand Binding Affinities ................. 60
4.4 Computational Protein Design for Improved Ligand Binding ... 63
4.5 Scoring Functions for Protein-Ligand Binding .................. 65
4.6 A Mixed-Integer Linear Programming Formulation to Find Optimal Rotamer Combinations ................................. 67

Chapter 5. A Test System: Mutations of the Acyl-Coenzyme A Binding Protein-Dodecanoyl-Coenzyme A System ................... 70

Chapter 6. Concanavalin A - Glucose Binding Site Optimization ........ 80

6.1 Concanavalin A .................................................. 81
6.1.1 cDNA for pre-pro-ConA ..................................... 83
6.1.2 Crystal Structures of Concanavalin A - Sugar Complexes ... 84
6.2 Competitive Binding and the Binding Affinity .................. 85
6.3 Prediction of Mutants Through Computational Protein Design ... 91
6.4 Future Research .................................................. 101

Chapter 7. Conclusions .................................................. 109
Appendix. 100 Best Concanavalin A Mutants for Improved Glucose Binding as Predicted by Maximizing the Absolute PLP Binding Score 111

References 117
List of Tables

3.1 Experimental and Calculated Data for $\Delta H_f^0$ of Various Hydrofluorocarbons 25

3.2 Ranking of 10 Molecules Closest to a Target Value of $\Delta H_f^0_{\text{target}} = -150$ kcal/mol, Calculated from the GCM based MILP Model ................................. 34

3.3 GA Performance for Different Parameter Sets, Averaged Over a Number of 1000 Runs for each GA .................................................................................. 37

3.4 Ranking of 10 Molecules Closest to a Target Value of $\Delta H_f^0_{\text{target}} = -150$ kcal/mol, Calculated from the Quantum Chemical GA ............................. 39

3.5 Comparison of Quantum Chemical $K_{\text{OW}}$ Results from Solvation Calculations for Case Study 2 ................................................................. 49

3.6 Case Study 2: The 10 Globally Best Molecules based on GCM and the 10 Best Molecules based on Quantum Chemical Solvation Calculations 50

4.1 Amino Acid Names, Three-letter Codes, and One-letter Codes ........ 56

4.2 Energy Function of Looger and Hellinga .................................................. 61

5.1 Experimental Binding Affinities of Mutated ACBP-Dodecanoyl-CoA and PLP-designed Mutants of ACBP- Palmitoyl-CoA ........................................... 72

5.2 Experimental Binding Affinities of Mutated ACBP-Dodecanoyl-CoA and FlexX-designed Mutants of ACBP-Palmitoyl-CoA ......................................... 72

5.3 Experimental Binding Affinities of Mutated ACBP-Dodecanoyl-CoA and CHARMM-designed Mutants of ACBP-Palmitoyl-CoA ....................................... 73
6.1 Comparison of Binding Energies of Two ConA-Sugar Complexes . . . . 86
6.2 Average Binding Scores of the Best 100 Designs based on Maximizing Binding Scores Compared with Binding Scores of 100 Random Designs . 95
A.1 List of 100 Best Designs based on Maximizing Absolute PLP Scores . 116
# List of Figures

1.1 Molecular Design Flowsheet ............................................. 2

3.1 Flowchart of the Molecular Design Approach ........................... 20

3.2 Development of the Normalized Total Fitness of a Population of 20
Molecules for different GA Parameters ...................................... 35

3.3 Development of the Normalized Total Fitness of a Population of 50
Molecules for different GA Parameters ...................................... 36

4.1 Protein Backbone ................................................................ 57

4.2 Rotamer Example: Glutamic Acid ......................................... 59

5.1 Experimental Binding Affinities of Mutated ACBP-Dodecanoyl-CoA and
PLP-designed Mutants of ACBP-Palmitoyl-CoA .............................. 74

5.2 Experimental Binding Affinities of Mutated ACBP-Dodecanoyl-CoA and
FLexX-designed Mutants of ACBP-Palmitoyl-CoA ............................ 75

5.3 Experimental Binding Affinities of Mutated ACBP-Dodecanoyl-CoA and
CHARMM-designed Mutants of ACBP-Palmitoyl-CoA ........................ 76

6.1 Influence of Dextran and Glucose Binding Constants on the Fluorescent
Response of a PEG Hydrogel Biosensor ....................................... 89

6.2 ConA Binding Site in the Wild Type and the Best PLP- Based Design .. 92

6.3 Frequency of Occurrence of Amino Acids in the PLP-based Protein Designs
and in the Lectin Protein Family for Residue Leu99 .......................... 97
6.4 Frequency of Occurrence of Amino Acids in the PLP-based Protein Designs
and in the Lectin Protein Family for Residue Ser168 .......... 98

6.5 Frequency of Occurrence of Amino Acids in the PLP-based Protein Designs
and in the Lectin Protein Family for Residue Thr226 .......... 99

6.6 Frequency of Occurrence of Amino Acids in the PLP-based Protein Designs
and in the Lectin Protein Family for Residue Arg228 .......... 100

6.7 Histogram of Hellinga Energies for 100 RANDOM Residue Positions,
ConA:Glucose .......................................................... 102

6.8 Histogram of Hellinga Energies for 100 OPTIMIZED Residue Positions,
ConA:Glucose .......................................................... 103

6.9 Hellinga Energy vs. CHARMM Scores for ConA:Glucose ........ 104

6.10 Hellinga Energy vs. PLP Scores for ConA:Glucose ............ 105

6.11 Hellinga Energy vs. FlexX Scores for ConA:Glucose ............ 106
Acknowledgments

I thank my thesis advisor, Costas D. Maranas, for the guidance and patience he has shown me during my time at Penn State. I am also grateful to all members of the Maranas group for inspiration and enlightening discussions on a wide variety of topics. I am most thankful and indebted to my wife Mila, who tolerated a long distance relationship during these years.
Chapter 1

Introduction

Molecular design is an iterative process of finding molecules that possess one or more desired properties. To save time and resources, it is of great interest to carry out part of this search for potentially new molecules on a computer rather than experimentally, generate a list of candidate substances and then conduct experiments based on this shorter list. Comparing a number of candidate substances (structures) computationally is based on a quantitative model of the structure-property relationship. The quality of models for this relationship is often constrained either by computational cost or by its state of development. In this thesis, methods for computational design are introduced, which address these shortcomings of the models. Figure 1.1 illustrates the basic idea of all molecular design. After defining property targets and constraints, a new molecule is generated by the algorithm. This molecule is now evaluated by a property estimation method, which models the structure-property relationship. Next, it is tested if the current molecule fulfills the defined targets and constraints, in which case the search is over. If not, the algorithm generates more molecules using an optimization method until the targets are met and the constraints are satisfied. The molecules that are investigated here fall into two classes: small molecules that are typically found in applications in the chemical industry and are, therefore, referred to as chemical systems, and, on the other hand, significantly larger protein molecules found in biological systems. For these two
Fig. 1.1. Molecular Design Flowsheet
classes of molecules, different property estimation methods and different optimization methods are used.

In the small-molecule design part, quantum chemistry is introduced for the first time as a property estimation method. Previous molecular design studies [17, 69, 115, 33] typically employed group contribution methods (GCM) [113, 47, 29] for this purpose. In these methods, the parameters for the contribution of molecular groups to the property of the entire molecule (or polymer repeat unit) are obtained by fitting the group contribution model to experimental data for a set of chemical compounds. However, if a certain atom type, molecular group type, or a certain type of chemical bonding is not present in the experimental set, the GCM will not account for this chemical information, thus limiting the predictive capabilities of these methods. Quantum chemical ab initio methods (calculations from first principles) obtain molecular properties from the most fundamental level of molecular information: the location of the nuclei and the number of electrons. From this input, in principle, molecular level information about any system can be predicted (e.g., molecular energies, electronic charge distributions, dipole and higher moments, vibrational frequencies, or molecular structure). Quantum chemical methods can provide molecular level properties with an accuracy that lies within the limits of experimental error. Although at the price of high computational cost, quantum chemical methods offer intriguing advantages for molecular design. They do not depend on a particular class of compounds and as more methods for accessing properties (the combination of quantum calculations and subsequent evaluations) become available, the potential to predict the properties of unknown molecules is growing. Frequently, the accuracy of the results is limited by computational time rather than by the chosen method.
Here, the use of *ab initio* methods for molecular design is explored through two example problems.

In the protein design part, an optimization-based framework is introduced to computationally re-design the binding site of existing proteins with the goal to directly improve protein-ligand binding. Protein re-design refers to changes in the amino acid sequence of the protein. Improving the function of proteins presents a significant challenge because the relationship between the sequence, the resulting protein structure (fold), and, hence, the function of the protein is not yet well understood. Consequently, there is no general method available for quantifying protein function from sequence information alone. Protein design is either based on random methods such as DNA shuffling, which does not require specific structural knowledge of the protein in order to modify its function, or proteins are re-designed based on knowledge of the protein structure. If the structure and the sites that are related to the protein function are known, specific mutations can be introduced to modify functionality. Introducing a single-site mutation is experimentally costly, and this is even more so for multiple-site mutations. It is, therefore, desirable to predict promising mutations computationally to reduce experimental cost. This presents several challenges. No universally applicable method is known to quantify even the simplest protein function, namely protein-ligand binding, and how it relates to the protein structure. Other functions such as enzymatic activity are even harder to quantify. Changes in protein structure, which occur far away from the active site, may affect the protein’s function or stability.

Computational protein design based on structural information relies on rotamer libraries, which represent lists of all side-chain conformations of all amino acids at each
residue position that is to be re-designed. A rotamer is a conformation of dihedral angles of the amino acid side-chain, which depends on characteristic backbone dihedral angles observed in the given protein structure. Rotamer libraries are obtained from statistical analysis of experimental Protein Data Bank (PDB)[7] structures. Depending on which rotamer library is chosen, the number of rotamers varies between 153 [67] to 320 [89] rotamers per design position, which leads to a remarkable combinatorial explosion if a large number of design positions is included in the design. In the design step (compare Figure 1.1), the optimization method that is typically used is dead-end elimination (DEE) [24] and the property that is typically estimated is the protein energy in the global minimum energy conformation (GMEC). DEE is a pruning method in which rotamers and rotamer pairs that cannot be part of the GMEC are eliminated over a number of computational cycles. There are several problems associated with this approach. The argument for seeking the GMEC is that a protein, irrespective of its function, must first be stable. Because the protein function, such as enzymatic activity level, is very difficult to assess computationally, stability is employed as a surrogate for protein fitness in most studies. Furthermore, if the protein function is not enzymatic activity but ligand binding, typically the binding affinity is not modeled directly because DEE requires energy terms which are unified for the entire protein-ligand complex. That is, the same energy terms are used for intra-protein interactions and for protein-ligand interactions (e.g. [99, 100, 65]). Lastly, DEE produces one GMEC. However, the most functional protein does not necessarily have to be the most stable one [75] and, due to uncertainties in modeling, the mathematical optimum does not have to be the biological optimum. For these reasons, it is of considerable interest to be able to generate a list of solution designs
rather than just one optimal design. In the work presented here, the simplest protein function, protein-ligand binding, is targeted for computational optimization. Emphasis is placed on distinguishing between different objectives for the protein: preserving protein stability and improving protein-ligand binding. The presented framework accounts for both objectives.
Chapter 2

Small-Molecule Design Background

In this chapter, the basic principles of quantum chemistry are highlighted, the molecular representation for small-molecule design is introduced, and genetic algorithms for exploring molecular alternatives are discussed.

2.1 Quantum Chemistry

The objective of quantum chemistry is to find the best possible approximation of the wave function \( \psi(r, R) \) in the electronic Schrödinger equation of a system containing one or more molecules.

\[
\hat{H}\psi(r, R) = E(R)\psi(r, R)
\]  \hspace{1cm} (2.1)

Here, \( \hat{H} \) is the Hamiltonian operator, \( E(R) \) is the energy of the system, \( R \) is the vector of the nuclear coordinates and \( r \) refers to the coordinates of the electrons. The wave function \( \psi(r, R) \) is postulated to contain all physical information of the system; once it is known, in principle, all other physical information can be obtained from it. Because of the complexity of Eq. 2.1, no analytical solution is known for systems of practical interest and \( \psi \) is approximated computationally. Most methods are based on the variation principle, which states that any trial wave function \( \phi \) estimating the true \( \psi \) can only yield an energy \( E \geq E_0 \), where \( E_0 \) is the ground state energy of the system. Variational methods
such as the Hartree-Fock (HF) method or density-functional theory (DFT) approximate \( \psi \) by varying \( \phi \) until \( E \) converges to a local minimum, using the self-consistent field (SCF) method. Following the Born-Oppenheimer approximation, nuclei are treated as fixed because they are about 1000 time heavier than electrons and, thus, move much slower than electrons. Therefore, \( E \) and \( \psi \) depend only parametrically on \( R \) - for each \( R \) , \( \psi(r, R) \) is different. In a geometry optimization step, \( R \) is modified using a non-linear gradient method until \( E \) converges to a local minimum. For each \( R \) , the SCF procedure is solved again. Hence, there are two nested energy minimization problems, resulting in significant computational expense. Quantum chemical results from non-optimized geometries are physically meaningless. For a thorough introduction to the large field of quantum chemistry, the reader is referred to the literature [61, 104, 83, 60].

2.2 Basis Functions and Basis Sets

A one-electron wave function is called an orbital. The mathematical form of orbitals is derived from analytical solutions of the hydrogen atom. The electronic wave function of a molecule \( \psi \) is a function of the orbitals of the \( n \) (even) electrons, doubly occupying \( n/2 \) molecular orbitals (Pauli principle):

\[
\psi = f(\psi_1, \ldots, \psi_{n/2})
\]  

(2.2)
where \( f(\psi_1, \ldots, \psi_{n/2}) \) is the Slater determinant [61]. Molecular orbitals are formed as linear combinations of atomic orbitals [60]:

\[
\psi_i = \sum_{\mu=1}^{m} c_{\mu i} \phi_{\mu}
\]  

(2.3)

where \( \psi_i \) is the \( i \)-th molecular orbital, \( c_{\mu i} \) are the coefficients of the linear combination, \( \phi_{\mu} \) is the \( \mu \)-th atomic orbital, and \( m \) is the number of atomic orbitals. In quantum chemical codes, the atomic orbitals \( \phi_{\mu} \) are represented by basis functions, which are themselves linear combinations of Gaussian Type Orbitals (GTO) \( g_t \) with

\[
\phi_{\mu} = \sum_{t} d_{\mu t} \cdot g_t \quad \text{and} \quad g_t = N \cdot e^{-\alpha r^2} l^m m^n
\]  

(2.4)

where \( N \) is a normalization constant, \( \alpha \) is an exponent, \( x, y \) and \( z \) are Cartesian coordinates, and \( l, m \) and \( n \) are exponents whose integer combinations determine the atomic orbital type \( t \) (\( s, p, d \) or \( f \)) that a particular \( g_t \) represents. The functions \( g_t \) are primitive gaussians and their combination to \( \phi_{\mu} \) is a contracted gaussian, wherein the parameters \( d_{\mu t}, N \) and \( \alpha \) are fixed for each orbital type \( t \) in each atom type that a software accounts for. A collection of basis functions consisting of subsets of basis functions for each atom type is a basis set. The size of a basis set depends on how many basis functions are used to represent its atoms. A basis set that is used in this study, 6-31G(d), assigns 15 basis functions to approximate the 5 orbitals (\( 1s, 2s, 2p_x, 2p_y, \) and \( 2p_z \)) of each second-row atom (e.g., C, F) and assigns 2 basis functions to each hydrogen atom. This is considered a relatively small basis set. The large basis set 6-311+(3df,2p) assigns 34 basis functions
to second-row atoms and 9 basis functions to hydrogen atoms [28]. In quantum chemical codes that are based on the variation principle, the coefficients $c_{\mu i}$ in Eq. 2.3 are evaluated by minimizing the energy of a given molecular geometry, which reduces the problem of finding an arbitrary wave function $\phi$ to the simpler one of optimizing the $c_{\mu i}$. Clearly, the larger the basis set, the more accurate is the approximation of the true wave function. However, the price for greater accuracy is much greater computational cost.

2.3 Levels of Theory

Many computational methods have been developed over the last five decades for the calculation of the approximate wave function or electron density. Distinguished in their levels of mathematical and numerical complexity, examples include Hartree-Fock theory (HF), density-functional theories (DFT), Møller-Plesset perturbation theory (MP$^1$), or configuration interaction (CI). A model chemistry is the combination of a level of theory with a basis set [28]. For example, HF/STO-3G is a comparatively simple model chemistry, whereas QCISD(T)/6-311+G(3df,2p) involves a more complex level of theory (configuration interaction) and a considerable higher number of basis functions. The latter one may require hours of computing time for determining the energy even for molecules containing only two or three atoms. The choice of a model chemistry depends on what is to be investigated. The most accurate results can be obtained by using a very high level of theory and a large basis set. However, the high computational cost of very accurate model chemistries usually prohibit their use in most practical cases. This is the reason for the growing success of DFT calculations. For many applications,
DFT provide good results in a reasonable time. The choice of a model chemistry often
has an empirical component. Many theoretical studies devote considerable attention to
the comparison of experimental data with predictions of a number of different model
chemistries and discuss their performances for the system of interest.

2.4 Initial Geometry and Molecular Representation

For successful geometry optimization, non-linear gradient methods require a good
 initial guess for $\mathbf{R}$ in Eq. 2.1. Most quantum chemical codes have a graphical user in-
terface for building molecules to accompany the chemical intuition of the user. In an
optimization problem, however, new molecules are generated by the solver, and the initial
guess for $\mathbf{R}$ must also be generated computationally. The quantum chemistry package
Gaussian 98 [30] provides the subroutine Model Builder [34, 88], which generates an ini-
tial geometry solely from molecular connectivity information by assigning standard bond
lengths and bond angles to adjacent atoms according to their types. For representing
molecules, a connectivity matrix $\mathbf{M}$ was chosen. In $\mathbf{M}$, each row contains information
about one atom of the molecule. The first-column element $m_{i1}$ represents the atom
type through its atomic number. For example, for carbon $m_{i1} = 6$. All other columns
contain the connectivity information as binary variables: $m_{ij+1} = 1$ if the atom in row
$i$ is connected to the atom in row $j$, $m_{ij+1} = 0$ otherwise. For example, the molecular
matrix for $F_2C\equiv CFH$ is given by:

$$M = \begin{bmatrix}
6 & 0 & 1 & 1 & 1 & 0 & 0 \\
6 & 1 & 0 & 0 & 0 & 1 & 1 \\
9 & 1 & 0 & 0 & 0 & 0 & 0 \\
9 & 1 & 0 & 0 & 0 & 0 & 0 \\
9 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 \\
\end{bmatrix}$$

(2.5)

Note that this representation only intrinsically encodes multiple bonds. This way of modeling is similar to the quantum chemical notion of the molecule as a collection of nuclei surrounded by an electron cloud. It is important to note that bonds are not an input but rather a result of the electronic structure calculation.

### 2.5 Optimization Procedure: A Genetic Algorithm

The molecular design problem is a discrete optimization problem: atom A is either connected to atom B or not and it is of a given type (i.e., carbon) or not. Commonly used discrete optimization methods are mathematical programming methods or guided random search methods such as simulated annealing or genetic algorithms. Mathematical programming techniques [74] such as MILP require a complete analytical model of the optimization problem. Since the approximate solution of the Schrödinger equation is not available in analytical form, these methods cannot be used directly to solve the quantum chemical model of the molecular system. A genetic algorithm (GA), first introduced by
Holland [42] and discussed by Michalewicz[72], is used in this study. By maintaining a population of chromosomes and applying crossovers and mutations to them, ranked by a certain fitness measure, GA’s constantly sample the search space and are intrinsically parallel in nature[72].

**Chromosome Representation:** The molecular candidates are represented as bit strings. These are one-dimensional fields of binary elements, i.e., chromosomes, on which the GA operates. For evaluating the fitness, chromosomes are translated into the molecular matrix M (Eq. 2.5). The combination of bit string representation and translation procedure encodes the constraints of the optimization problem. After randomly creating a number of chromosomes, one obtains an initial population and can apply the GA, which consists of the following steps:

1. Calculation of the *fitness* of each chromosome
2. *Selection* of chromosomes for the new generation
3. Random application of the *crossover operator* to members of the population
4. Random application of the *mutation operator* to members of the population

**Fitness Evaluation:** For each chromosome, the fitness is a measure of how far this chromosome deviates from the optimization goal. In this study, the fitness expresses how far the property \( P_i \) of a particular chromosome \( i \) deviates from the given target \( P_{\text{target}} \). For a vanishing deviation, the fitness \( F_i \) should approach one and it should decrease as the deviation increases. This behavior is achieved by defining \( F_i \) in form of a Gaussian
function [115]:

\[ F_i = \exp \left( -\alpha \frac{(P_{\text{target}} - P_i)^2}{(P_{\text{target}})^2} \right) \]  

(2.6)

Here, \( \alpha \) is a parameter of the algorithm, which determines how steeply the Gaussian curve decreases with larger deviations. If the design involves properties that have either a lower bound or an upper bound, a sigmoidal fitness function is used

\[ F_i = \left\{ 1 + \exp \left( -\beta \left[ \frac{P_i - P_{F=0.5}}{P_{\text{Range}}} \right] \right) \right\}^{-1} \]  

(2.7)

where \( P_{F=0.5} \) is the property value for which the evaluated fitness is 0.5. The lower or upper limit of the property constraints are placed at this point [115]. Note that the limit is not a bound, i.e., values below the lower bound or above the upper bound can still yield acceptable fitness values, depending on the slope \( \beta \) of the resulting S-curve. For multiple objectives, Eqns. 2.6 or 2.7 are formulated for each objective and the total fitness of each chromosome is obtained by averaging over all property objectives. From the fitness of each chromosome, the total fitness of the population is determined by

\[ F_{tot} = \sum_{i=1}^{\text{pop\_size}} F_i \]  

(2.8)

where \( \text{pop\_size} \) is the size of the population.

**Selection** The selection process ensures that the fittest chromosomes have better chances for survival while the unfit chromosomes do not die out immediately, but only
after a number of generations. A probability \( p_i \) for selection into the next generation is

defined to be proportional \([38]\) to the fitness of chromosome \( i \):

\[
p_i = \frac{F_i}{F_{tot}} \quad (2.9)
\]

After calculating a cumulative probability \( q_i \),

\[
q_i = \sum_{j=1}^{i} p_j \quad (2.10)
\]

and generating a random number \( r \) with \( 0 \leq r < 1 \), chromosome \( i \) is selected into the

new population if \( q_{i-1} < r \leq q_i \). This is repeated \( \text{pop.size} \) times in order to obtain the population of the next generation.

**Crossover:** Chromosomes of the new population are randomly selected for crossover with a probability \( p_c \). During crossover, two chromosomes exchange their sequences of bits following a randomly selected position \( m \) of the bit string. For example, for \( m = 6 \), and \( S_1 \) and \( S_2 \) before crossover:

\[
S_1 = (1 \ 0 \ 0 \ 1 \ 1 \ 1| \ 1 \ 0 \ 0 \ 1) \quad (2.11)
\]

\[
S_2 = (0 \ 1 \ 0 \ 1 \ 0 \ 1| \ 1 \ 1 \ 1 \ 1)
\]

leads to \( S'_1 \) and \( S'_2 \) after crossover:

\[
S'_1 = (1 \ 0 \ 0 \ 1 \ 1 \ 1| \ 1 \ 1 \ 1 \ 1) \quad (2.12)
\]

\[
S'_2 = (0 \ 1 \ 0 \ 1 \ 0 \ 1| \ 1 \ 0 \ 0 \ 1)
\]
**Mutation:** Chromosomes of the new population are randomly selected for mutation with a probability $p_m$. Mutation is the change of one bit ($0\rightarrow1$ or $1\rightarrow0$) of a chromosome at a randomly selected position of the bit string.

**Tuning the GA** Several parameters influence the performance of the GA: the population size $\text{pop\_size}$, the probability of crossover $p_c$, the probability of mutations $p_m$ and the number of generations $N_{\text{generations}}$. Before the GA is used to operate on a population of molecules that are evaluated by costly quantum chemical calculations, it is important to tune these parameters for optimal performance of the algorithm, which is a compromise between rapid convergence (evaluation of few molecules) and extensive sampling of the search space (evaluation of many molecules). If the algorithm converges too fast, the confidence in the solution is smaller. After finding the $n$ best molecules using a rapid property evaluation method such as group contribution methods (GCM), information about the performance of the GA can be obtained by evaluating how many of the best solutions the GA finds and at what computational expense. For this purpose, the GA is run for a number of generations while the same GCM is used for the property evaluation. Using the fitness $F_i$ as a measure, a record of $n$ all-time-best (GA) individuals is maintained. The GA is a random search technique and it is unlikely that it finds all globally best molecules. Let $N$ be the set of the $n$ globally best molecules. The percentage $P_{n,\text{best}}$ of the $n$ globally best molecules that the GA finds is given by

$$
P_{n,\text{best}} = 100\epsilon/n
$$

(2.13)
where $\nu$ is the number of molecules found by the GA that are elements of $N$. Running the GCM-based GA $k$ times, $P_{n,\text{best}}$ is averaged by

$$P_{n,\text{best},\text{avg}} = \frac{1}{k} \sum_{i=1}^{k} P_{n,\text{best}}(i) \quad (2.14)$$

In this study, $n = 10$. For $k \geq 1000$, $P_{10,\text{best},\text{avg}}$ fluctuates only in the first decimal place. Another important performance measure of the GA is how many function evaluations the algorithm needs to achieve a certain $P_{n,\text{best}}$. The parameter $N_{\text{functions}}$ monitors how often the evaluation subroutine is called during one completion of the GA, i.e., how many candidate molecules pass the filter discussed below. This is of interest because, ultimately, costly quantum chemical calculations are to be used for these function evaluations. The objective of the empirical tuning procedure is to maximize $P_{10,\text{best},\text{avg}}$.

**Filter** For avoiding unnecessary evaluations and by applying a GCM, molecules are screened for proximity to the optimization goal if the property is to be evaluated by quantum chemistry. A filter only accepts molecules with

$$P_{\text{target}} - \Delta P \leq P_{i,\text{GCM}} \leq P_{\text{target}} + \Delta P \quad (2.15)$$

for evaluation by the quantum chemical method. In Eq. 2.15, $P_{i,\text{GCM}}$ is the property $P$ of chromosome $i$, based on a GCM calculation and $\Delta P$ is a tolerance that ensures that candidate molecules whose $P_{i,\text{GCM}}$ lies outside of this tolerance are highly unlikely to have a $P_i$ based on quantum chemistry within the tolerance. The choice of $\Delta P$ depends on both, the accuracies of the GCM and the quantum chemical method.
Chapter 3

Molecular Design using Quantum Chemical Calculations for Property Estimation

3.1 Optimization Procedure

The solution of the quantum chemical problem as part of the property evaluation of a molecule is found by numerical methods and not by analytical structure-property relationships. Because of the ability to handle non-analytical functions, a genetic algorithm (GA) was chosen as the optimization procedure. The GA generates molecules and treats the quantum chemical evaluation subroutine of the molecules as a black box. It is important to note that since GA’s are directed random search methods, one cannot claim global optimality. For this reason and because ab initio calculations are very time consuming, it is important to tune the performance of the GA before engaging in costly quantum chemical property evaluations. Performance is governed by a number of adjustable parameters such as probabilities of crossover and mutation, population size and number of generations.

For GA tuning, it is favorable to have a list of the globally best, second best, etc. molecules for evaluating how many of these globally best candidates the GA is able to find with a particular set of parameters. Generating such a list using quantum chemistry, however, is exactly the problem to be solved. It is much less time consuming to generate the list when using group contribution methods (GCM) for rapid property evaluation. In
this study, two different approaches for GCM-based GA tuning are used. If the GCM is
given in linear form or its constraints can be transformed into a linear form [69], a list of
molecules can be generated by solving a GCM-based mixed-integer linear programming
(MILP) form of the problem to global optimality and comparing the results to those
of a GCM-based GA for the same molecular search space as in the original problem.
This approach is applied to the first example, and the MILP formulation is given in
Section 3.2.1. If the GCM cannot be given in linear form, as in the second example,
one can sample either the entire search space or a reduced search space and find the
GCM-based globally optimal molecules by exhaustive enumeration or other GCM-based
search techniques.

Figure 3.1 illustrates the proposed approach. First, a list with \( n \) GCM-based
globally best molecules is found either by solving an MILP model or by exhaustive
enumeration (Step 1). Next, it is evaluated how many of these best solutions are found
by the GA with a particular set of parameters (Step 2). The GA is tuned such that it
finds the maximal average number of globally best solutions when using the GCM for
property evaluation. Step 3 describes the actual optimization procedure. All molecules
are first screened using the GCM. Molecules with GCM values within a targeted range,
depending on the GCM’s accuracy and on the optimization goals, are selected for further
evaluation by the \textit{ab initio} calculation. This reduces the number of candidate molecules
that need to be evaluated in the computationally expensive \textit{ab initio} step. The number of
candidate molecules is further reduced by comparing them with a database and checking
if a particular molecule was evaluated in a previous generation or GA run. In this
case, the result is retrieved from the database. After generating an initial geometry,
1. Generate List of n Globally Best Molecules
   - based on group contribution methods (GCM) using globally optimal search methods
     (MILP, exhaustive enumeration etc.)

2. GCM-based Tuning of the Genetic Algorithm (GA)
   - maximize percentage of best candidates found by the GCM-based GA
     that is also part of the n globally best candidates identified in 1.

3. Application of Tuned GA using Quantum Chemistry
   - Randomly generate pop_size chromosomes
   - Build molecular matrices from chromosomes
   - Evaluation of all molecules of a generation by GCM,
     filtering molecules with properties close to target
   - Check DATABASE for molecules that were previously
     evaluated by quantum chemical calculations
   - Parallel submission of filtered molecules

   Feed NEW molecules in DATABASE

   Repeat for a number of generations,
   RECORD BEST candidates of each generation

   Performed by quantum chemical code

   Optimization of molecular geometry
   until minimum SCF energy is found

   Property evaluation of molecule i

   GA evaluation, selection, crossover, mutation
   and generation of new population

   END

Fig. 3.1. Flowchart of the Molecular Design Approach
the candidate molecules are submitted to a quantum chemical software. Note that the molecules are submitted individually, which results in a parallel execution of this most time consuming step. Subsequent to geometry optimization, the candidate molecules can be subjected to further quantum chemical evaluations as in the second example, where each candidate molecule is solvated in water and octanol, respectively. After finding the final properties, the fitness of the candidate molecules is evaluated by the GA. By repeating this GA cycle for a number of generations, the best candidates are recorded. All program parts are linked. While the evaluation routines run, the GA pauses and waits for result files. If a candidate molecule causes problems, it can be treated individually and the result files are then submitted to the GA for continuation.

In the following sections, two design examples are presented. In the first problem, the molecular design of hydrofluorocarbons is addressed by minimizing the deviation between a target property and the property of the designed “solution” molecules. The ideal gas heat of formation ($\Delta H_f^0$) at standard conditions is chosen as the target property, motivated by an interest in chemically stable hydrofluorocarbon refrigerants. The second example focusses on the design of solvents for liquid-liquid extraction. The optimized properties are the solvent limiting capacity $C_{\infty,A}$, the limiting selectivity $S_{\infty,A,B}$ and the environmental safety of the solvent, represented by the octanol-water partition coefficient $K_{\text{OW}}$. In this example, $K_{\text{OW}}$ is evaluated by a quantum chemical method [63] while $C_{\infty,A}$ and $S_{\infty,A,B}$ are evaluated using the UNIFAC group contribution method [29, 87]. It is emphasized that these examples were selected primarily as benchmark problems to explore the applicability of the proposed approach.
3.2 Case Study 1: Design of Hydrofluorocarbons

3.2.1 Problem Definition

The first example is motivated by the search for alternative, chlorine free refrigerants. One of the properties of interest here is the heat of formation $\Delta H_f^0$ of the candidate substance, which is used as an indicator for the stability of the compound. To this end, molecules with $\Delta H_f^0$ closest to a target value $\Delta H_{f,\text{Target}}^0$ are identified using the GA-based optimization procedure.

**Heat of Formation** The heat of formation is obtained using the method of Curtiss et. al. [20], by first evaluating the dissociation (atomization) energy $\sum D_0$, i.e., the energy difference between the molecule and its dissociated atoms. For example, for the molecule $A_x B_y H_z$,

$$\sum D_0 = x \cdot \varepsilon_0(A) + y \cdot \varepsilon_0(B) + z \cdot \varepsilon_0(H) - \varepsilon_0(A_x B_y H_z) - \varepsilon_{\text{ZPE}}(A_x B_y H_z) \quad (3.1)$$

where $\varepsilon_0(X)$ is the energy of the particle (atom or molecule) $X$ and $\varepsilon_{\text{ZPE}}(A_x B_y H_z)$ is the zero-point energy of the molecule. Variables $x$, $y$ and $z$ represent the number of atoms $A$, $B$ or $H$ in the example molecule. The zero-point energy is a part of the ground-state energy that accounts for molecular vibrations persisting even at 0 K. All values in Eq. 3.1 are calculated using Gaussian 98 [30]. Note that the model chemistry must be the same for all particles $X$. The enthalpy of formation of the molecule at 0 K
is given by

$$\Delta H^0_f(A, B, H, 0K) = x \cdot \Delta H^0_f(A, 0K) + y \cdot \Delta H^0_f(B, 0K) + z \cdot \Delta H^0_f(H, 0K) - \sum D_0 \quad (3.2)$$

Here, the dissociation energy is subtracted from the sum of the widely accepted values for the 0 K heats of formation of gaseous atoms $\Delta H^0_f(X, 0K)$, which are tabulated in Chase et. al. [14]. $\Delta H^0_f(A, B, H, 0K)$ is also corrected for the standard state (298 K):

$$\Delta H^0_f(A, B, H, 298K) = \Delta H^0_f(A, B, H, 0K)$$

$$+ \left[ H^0(A, B, H, 298K) - H^0(A, B, H, 0K) \right]$$

$$- x \cdot \left[ H^0(A, 298K) - H^0(A, 0K) \right]_{st} \quad (3.3)$$

$$- y \cdot \left[ H^0(B, 298K) - H^0(B, 0K) \right]_{st}$$

$$- z \cdot \left[ H^0(H, 298K) - H^0(H, 0K) \right]_{st}$$

In this equation, $[H^0(A, B, H, 298K) - H^0(A, B, H, 0K)]$ is evaluated by the quantum chemistry software and $[H^0(X, 298K) - H^0(X, 0K)]_{st}$ are obtained from tabulated values[14].

**Quantum Chemical Model** A number of preliminary runs were first carried out in order to find a model chemistry which balances acceptable accuracy and computational expense. The $\Delta H^0_f$ were predicted for several hydrofluorocarbons using B3LYP/6-31G(d) for both the geometry optimization and the final energy calculation. Bauschlicher [5] reports an average error of 5.18 kcal/mol for $\sum D_0$ using this model chemistry on the G2 test set of 55 molecules [21]. When comparing results for $\sum D_0$ and $\Delta H^0_f$, the
influence of the average error of \( H^0(A, B, H, 298K) - H^0(A, B, H, 0K) \) in Eq. 3.3 on the average error in \( \Delta H^0_f \) is neglected. For the set of 21 molecules that were calculated for this work, an average error of 4.53 kcal/mol was found. The results are summarized in Table 3.1. Curtiss et. al. [20] reported average errors of 2.43 kcal/mol for the 55 molecules of the G2 test set. However, these results were obtained using the B3LYP/6-311+G(3df,2p) model chemistry. This basis set was deemed too computationally expensive for the purpose of using it as a subroutine within an optimization loop. For comparison, the heat of formation values as calculated from the GCM are provided in Table 3.1. The average error for this method was found to be 12.23 kcal/mole.

**Bit String Representation and Objective Function for Case Study 1** For the problem at hand, the search space is constrained to hydrofluorocarbons with a maximum of three carbon atoms and possible multiple bonds between them. The bit strings contain 10 digits, each of which can assume the value 0 or 1 (hence, the name bits). For example, a randomly generated string \( S \) may look as:

\[
S = (1 \ 0 \ 0 \ 1 \ 1 \ 1 \ 0 \ 0 \ 1)
\]  

(3.4)

The elements of \( S \) are defined as follows. The first two positions define the number of carbon atoms in the molecule by the equation:

\[
\text{Number of Carbons} = 1 + S(1) \cdot 2^{S(2)}
\]  

(3.5)
<table>
<thead>
<tr>
<th>Formula</th>
<th>Name</th>
<th>$\Delta H^0_{f, \text{EXP}}$ [kcal/mol]</th>
<th>$\Delta H^0_{f, \text{BALYP/6-31G(d)}}$ (Abs. Dev.) [kcal/mol]</th>
<th>$\Delta H^0_{f, \text{GCM}}$ (Abs. Dev.) [kcal/mol]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH₃F</td>
<td>fluoromethane</td>
<td>-56.00 [14]</td>
<td>-56.02 (0.08)</td>
<td>-62.16 (6.16)</td>
</tr>
<tr>
<td>CH₂F₂</td>
<td>difluoromethane</td>
<td>-59.00 [62]</td>
<td>(3.99)</td>
<td>(3.16)</td>
</tr>
<tr>
<td>CHF₃</td>
<td>trifluoromethane</td>
<td>-107.7 [14]</td>
<td>-107.20 (0.50)</td>
<td>-109.03 (1.33)</td>
</tr>
<tr>
<td>CF≡CH</td>
<td>fluoroethylene</td>
<td>30.0 [15]</td>
<td>28.45 (1.55)</td>
<td>2.67 (27.33)</td>
</tr>
<tr>
<td>CF₂=CH₂</td>
<td>1,1-difluoroethene</td>
<td>-32.4 [55]</td>
<td>(0.13)</td>
<td>(4.72)</td>
</tr>
<tr>
<td>CF₂=CH</td>
<td>trifluoroethene</td>
<td>-113.3 ± 2.0 [53]</td>
<td>-121.73 (8.43)</td>
<td>-135.16 (21.86)</td>
</tr>
<tr>
<td>CF₂=CF₂</td>
<td>tetrafluoroethene</td>
<td>-157.4 [109]</td>
<td>-166.38 (8.98)</td>
<td>-184.37 (26.97)</td>
</tr>
<tr>
<td>CF₃=CH₂</td>
<td>1,1,1-trifluoroethane</td>
<td>-157.4 [15]</td>
<td>(8.98)</td>
<td>(26.97)</td>
</tr>
<tr>
<td>CF₃=CH₃</td>
<td>1,1,1,1-tetrafluoroethane</td>
<td>-178.94 ± 0.76 [122]</td>
<td>(3.07)</td>
<td>(16.01)</td>
</tr>
<tr>
<td>CF₄=CF₄</td>
<td>hexafluoroethane</td>
<td>-178.9 ± 0.4 [54]</td>
<td>(3.11)</td>
<td>(15.97)</td>
</tr>
<tr>
<td>CF₅=CF₅</td>
<td>heptfluoroethene</td>
<td>-312.1 [15]</td>
<td>-329.06 (7.86)</td>
<td>-350.63 (15.57)</td>
</tr>
<tr>
<td>CFH=CH₂CH₃</td>
<td>trans-1-fluoro-1-propene</td>
<td>-14.3 [3]</td>
<td>-37.64 (2.66)</td>
<td>-44.01 (2.71)</td>
</tr>
<tr>
<td>CFH=CH₃CH₃</td>
<td>cis-1-fluoro-1-propene</td>
<td>-14.1 [3]</td>
<td>-37.68 (2.62)</td>
<td>-44.01 (2.71)</td>
</tr>
<tr>
<td>CH₂=CH=CF₃</td>
<td>3,3,3-trifluoropropene</td>
<td>-146.0 ± 2.0 [106]</td>
<td>(12.37)</td>
<td>(15.59)</td>
</tr>
<tr>
<td>CF₅=CF₅=CF₅</td>
<td>hexafluoropropene</td>
<td>-146.79 ± 1.6 [52]</td>
<td>(3.25)</td>
<td>(8.91)</td>
</tr>
<tr>
<td>CH₃=CF₂=CH₂</td>
<td>2,2-difluoropropene</td>
<td>-144.1 ± 1.6 [19]</td>
<td>(5.54)</td>
<td>(6.62)</td>
</tr>
<tr>
<td>CF₃=CF=CF₃</td>
<td>octafluoropropane</td>
<td>-275.26 [82]</td>
<td>(7.55)</td>
<td>(9.88)</td>
</tr>
<tr>
<td>CF₃=CF₂=CF₂ = 2.2-difluoropropane</td>
<td>-129.8 ± 3.0 [121]</td>
<td>-131.00 (1.20)</td>
<td>-120.99 (8.81)</td>
<td></td>
</tr>
<tr>
<td>CF₃=CF₂=CF₃ = octafluoropropane</td>
<td>-142.6 ± 2.0 [84]</td>
<td>-140.88 (4.68)</td>
<td>-106.40 (19.80)</td>
<td></td>
</tr>
<tr>
<td>CF₃=CF₂=CF₂ = octafluoropropane</td>
<td>-126.55 ± 2.10 [57]</td>
<td>(4.32)</td>
<td>(20.15)</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1. Experimental and Calculated Data for $\Delta H^0_f$ of Various Hydrofluorocarbons;
Values in Parentheses Show Absolute Deviations from Experimental Values
The third and fourth positions define the number of bonds between the first and the second carbon atom based on the formula:

\[
\text{Number of Bonds between } C_1 \text{ and } C_2 = S(3) \cdot 2^0 + S(4) \cdot 2^1
\] (3.6)

The case \(S(3) = S(4) = 0\) encodes a double bond on each side of the middle atom in case of a molecule with three carbon atoms. If the molecule has only two carbon atoms, they are connected by a double bond between them. The last six positions of \(S\) define the number of fluorine atoms on each carbon:

\[
\text{Number of Fluorines on } C_1 = S(5) \cdot 2^0 + S(6) \cdot 2^1
\] (3.7)

\[
\text{Number of Fluorines on } C_2 = S(7) \cdot 2^0 + S(8) \cdot 2^1
\] (3.8)

\[
\text{Number of Fluorines on } C_3 = S(9) \cdot 2^0 + S(10) \cdot 2^1
\] (3.9)

The translation of the bit strings into connectivity matrices is described next. First, all carbon atoms are connected in a molecular backbone. In the next step, the bond multiplicities between carbon atoms are assigned, which permits the calculation of the number of remaining bonds at each carbon atom (three at most). These bonds are then filled with the number of fluorine atoms assigned to each carbon atom through the bit string. When all fluorine atoms are connected, the remaining bonds of each carbon atom are filled with hydrogen atoms. Based on these definitions, one recognizes \(S\) as a bit string representation of \(\text{F}_2\text{C}=\text{CFH}\). Note that one cannot attach three fluorine
atoms to the first carbon atom as required by $S(5) = S(6) = 1$ because the carbon atoms are connected by a double bond, which has higher priority. Thus, in this problem representation, two bit strings can result in the same molecular connectivity matrix, which produces some redundancy. However, this formulation ensures that no bit string can result in an infeasible molecule. The objective function in this example is written in the form of Eq. 2.6 with a target value of $\Delta H_0^f, \text{target} = -150 \text{ kcal/mol}$.

**GA Tuning: An MILP Model based on a Group Contribution Method for Rapid Property Calculation** For tuning the GA, a significantly faster evaluation method is needed. In this case study, a group contribution method proposed by Joback and Reid [47] was used as a surrogate for the heat of formation predictions based on quantum chemical calculations. The advantage of Joback’s and Reid’s method is its simplicity, which stems from the assumption that each group $i$ contributes additively by a value $\Delta h_0^{f,i}$ to $\Delta H_0^f$ of the molecule. The approximating expression is

$$\Delta H_0^f = 68.29 + \sum n_i \cdot \Delta h_0^{f,i}$$

(3.10)

where $n_i$ is the number of occurrences of group $i$ in the molecule. This simple additivity form results in a relatively straightforward MILP formulation [12, 90]. The key advantage of this MILP formulation is that efficient solvers such as CPLEX or OSL (accessed via GAMS [31, 32]) can identify the globally optimal molecule that is closest to the given target value. In addition, the MILP framework can be used to generate a list of $n$-best solutions [69], knowledge of which provides the basis for tuning. As an objective, one seeks to minimize the deviation of the molecule’s heat of formation from a given target
value:
\[
\min \left| 1 - \frac{\Delta H^0_f}{\Delta H^0_{f,\text{target}}} \right|
\]  
(3.11)

The nonlinear objective is recast in a linear form [69] as follows:

\[
\min s
\]  
(3.12)

subject to
\[
s \geq \left( 1 - \frac{\Delta H^0_f(L_f, FC_{\text{tot}})}{\Delta H^0_{f,\text{target}}} \right), \quad s \geq \left( \frac{\Delta H^0_f(L_f, FC_{\text{tot}})}{\Delta H^0_{f,\text{target}}} - 1 \right)
\]

The dependencies on \(L_f\) and \(FC_{\text{tot}}\) are given in Eqs. 3.16 and 3.17. Molecules are modeled as a molecular graph with up to three possible vertexes. Each vertex can assume one of the types listed in Joback and Reid [47], excluding the fluorine atoms. Information on how the vertexes are connected is encoded by the adjacency matrix \(A = (a_{ijk})\) [112], whose elements are binary variables with

\[
a_{ijk} = \begin{cases} 
1 & \text{if vertex } i \text{ is connected to vertex } j, \\
\text{forming a bond of multiplicity } k & \text{and } i, j, k = 1, 2, 3 \\
0 & \text{otherwise}
\end{cases}
\]  
(3.13)

Note that \(i\) and \(j\) range from 1 to 3 because the molecular graph is limited to three vertexes and \(k\) ranges from 1 to 3 because carbon atoms can be connected by single, double or triple bonds. Similarly, a vertex type binary \(y_{il}\) is defined, which determines
the type of the group that occupies a vertex:

\[
y_{il} = \begin{cases} 
1 & \text{if vertex } i \text{ is of type } l \\
0 & \text{otherwise}
\end{cases} \quad i = 1, 2, 3 \\
\text{and} \\
\quad l = (-CH3), \ldots, (\equiv C-) 
\] (3.14)

The indices \( l \) span over all the groups [47] that contain carbon atoms. The number of fluorine atoms is calculated by evaluating the number of unoccupied sites on a vertex that is occupied by a carbon containing group. For example, if \( y_{1,(=CH-)} = y_{2,(=C<)} = a_{122} = 1 \) and all other binary variables are zero, there is one site unoccupied on the first vertex and two on the second one. The third vertex has no carbon group, therefore, no fluorine atoms can be attached to it. This combination of binary variables yields FHC=CF\(_2\), 1,2,2-trifluoroethylene. This can be accomplished by writing the following constraints:

\[
FC_i = \sum_l 4y_{il} - \sum_{k=1}^{3} \left( \sum_{j=1}^{j<i} k \cdot a_{ijk} + \sum_{j>i}^{3} k \cdot a_{ijk} \right) - \sum_l y_{il} \cdot H_l \quad \text{with} \quad i = 1, 2, 3 \quad (3.15)
\]

\[
FC_{\text{tot}} = \sum_{i=1}^{3} FC_i 
\] (3.16)

Here, \( FC_i \) is the number of fluorine atoms attached to vertex \( i \) and parameter \( H_l \) contains the number of hydrogen atoms pertaining to group \( l \). \( FC_{\text{tot}} \) represents the total number of fluorine atoms in the molecule. The first term in Eq. 3.15 sets up four available sites on vertex \( i \), if the vertex is occupied. The second term subtracts the number of bonds that are formed with other carbon atoms and the third term subtracts the bonds that
are formed with hydrogen atoms. If any sites remain unoccupied, fluorine atoms are
attached to vertex $i$. Eq. 3.16 adds the fluorine atoms of the entire molecule. Because
fluorine atoms are not treated as a group, Eq. 3.10 is rewritten as

$$
\Delta H^0_f = 68.29 + \sum_l L_l \cdot \Delta h^0_{f,l} + F C_{\text{tot}} \cdot \Delta h^0_{f,(\text{F})}
$$

where $L_l$ is the number of group $l$ in the molecule. At least one vertex should be occupied:

$$
\sum_{i=1}^{3} \sum_l y_{il} \geq 1
$$

At most one group $l$ can be present at a vertex $i$ or the vertex remains empty:

$$
\sum_l y_{il} \leq 1 \quad i = 1, 2, 3
$$

A molecular tree graph has exactly one vertex more than it has edges, i.e., the molecule
has exactly one carbon group more than it has connections between these groups:

$$
\sum_{i=1}^{3} \sum_{j=i+1}^{3} \sum_{k=1}^{3} a_{ijk} = \sum_{i=1}^{3} \sum_l y_{il} - 1
$$

The molecule should be connected, i.e., when vertices $i$ and $j$ are occupied, there should
be a connection of bonds leading from $i$ to $j$. This can be achieved by enforcing that if
vertex $j$ is occupied, there is at least one vertex $i$ with $i < j$ connected to $j$ via an edge
(bond) $a_{ijk}$:

$$
\sum_{i=1}^{j-1} \sum_{k=1}^{3} a_{ijk} = \sum_l y_{jl} \quad j = 2, 3
$$
Vertexes can only be connected by one type of bond at a time:

\[
\sum_{k=1}^{3} a_{ijk} \leq 1 \quad \begin{cases} 
    i=1,2,3 \\
    j=i+1, \ldots, 3
\end{cases}
\]  

(3.22)

The number of bonds of type \( k \) that a group \( l \) can form with other carbon groups is limited and varies between the groups. In order to account for this, the parameter \( CB_{lk} \) is introduced. For example, \( CB_{(>C<),1} = 2 \), since this model is restricted to unbranched hydrofluorocarbons. For double and triple bonds, \( CB_{lk} \) must balance the number of edges with bond multiplicity \( k \):

\[
\sum_{j=1}^{i-1} a_{ijk} + \sum_{j=i+1}^{3} a_{ijk} = \sum_{l} y_{il} \cdot CB_{lk} \quad \begin{cases} 
    i=1,2,3 \\
    k=2,3
\end{cases}
\]  

(3.23)

For single bonds, the number of edges \( a_{ij1} \) can be smaller than \( CB_{11} \), because the vertex \( i \) can be an ending vertex of the molecule. In this case, not all the possible connections \( CB_{11} \) to other carbon groups would be used. Instead, these connections could be used to attach fluorine atoms or hydrogen atoms:

\[
\sum_{j=1}^{i-1} a_{ij1} + \sum_{j=i+1}^{3} a_{ij1} \leq \sum_{l} y_{il} \cdot CB_{11} \quad i = 1,2,3
\]  

(3.24)

The following constraint uses the continuous variable \( L_l \) to count the number of each group \( l \) in the molecule:

\[
L_l = \sum_{i=1}^{3} y_{il} \quad l = (\text{CH3}), \ldots, (\equiv \text{C}-)
\]  

(3.25)
Reducing the number of degenerate solutions, some constraints are added to tighten the formulation. The following two constraints enforce that vertexes 1 and 3 are ending vertexes by restraining their maximal number of adjacent groups to one:

\[
\sum_{j=2}^{3} \sum_{k=1}^{3} a_{1jk} \leq 1 \quad (3.26)
\]

\[
\sum_{i=1}^{2} \sum_{k=1}^{3} a_{i3k} \leq 1 \quad (3.27)
\]

Finally, two constraints are added to force vertex 2 to be occupied if vertex 3 is occupied,

\[
\sum_{l} y_{2l} \geq \sum_{l} y_{3l} \quad (3.28)
\]

and to force vertex 1 to be occupied if vertex 2 is occupied.

\[
\sum_{l} y_{1l} \geq \sum_{l} y_{2l} \quad (3.29)
\]

This MILP formulation enables one to solve the problem to global optimality through a branch-and-bound algorithm [27]. The basic principle of branch-and-bound algorithms is a tree search, in which first all the binary variables are relaxed to be continuous with lower bounds of 0 and upper bounds of 1. The search tree is generated by consecutively fixing the binary variables to 0 and 1, thus creating two branches for each binary variable. In each branch, there is a continuous subproblem, which is solved as a linear program (LP) to optimality. The key advantage is that, in a minimization problem, branches with
higher optimal solutions to the continuous linear subproblem can be pruned as globally suboptimal if there are other branches with lower optimal solutions.

For generating more than one globally optimal molecule, efficient integer cuts are incorporated. If $y_{il}^{\text{sol}}$ is the optimal solution, then the constraint

$$
\sum_{(i,j):y_{il}^{\text{sol}}=1} y_{il} + \sum_{(i,j):y_{il}^{\text{sol}}=0} (1 - y_{il}) \leq \left( \sum_{i,l} 1 \right) - 1
$$

(3.30)

makes $y_{il}^{\text{sol}}$ infeasible when the problem is solved again. Thus, looping the solution procedure $n$ times and accumulating $(n - 1)$ integer cuts in the process, generates a list of the $n$ best solutions to the problem.

### 3.2.2 Results of Case Study 1

**Solutions of the MILP Model** Solving the MILP model based on the GCM for a target of $\Delta H_{f,\text{target}}^0 = -150$ kcal/mol (Fig. 3.1, Step 1) produced a list of ten molecules (Table 3.2) that are closest to the target value. Table 3.2 provided the basis for evaluating the GA when using the GCM for property evaluation. Based on the additivity assumption, the GCM cannot predict if a particular conformation of atoms in a molecule will be energetically stable. Therefore, there is no certainty if all the molecules in Table 3.2 can exist under standard conditions. The molecules ranked 3rd, 9th and 10th, however, do exist as indicated in Table 3.1.

**GA Tuning Based on the GCM** Figures 3.2 and 3.3 show the evolution of the normalized total fitness of the population ($F_{\text{tot}}$; Eq. 2.8, divided by pop size). In these
<table>
<thead>
<tr>
<th>Rank</th>
<th>Molecule</th>
<th>$\Delta H_{f,GCM}^0$ [kcal/mol]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>CF≡C–CF$_3$</td>
<td>-149.651</td>
</tr>
<tr>
<td>2.</td>
<td>CF$_2$≡C–CF$_2$</td>
<td>-150.399</td>
</tr>
<tr>
<td>3.</td>
<td>CF$_3$H</td>
<td>-157.165</td>
</tr>
<tr>
<td>4.</td>
<td>CF$_2$≡CF–CH$_3$</td>
<td>-142.433</td>
</tr>
<tr>
<td>5.</td>
<td>CF$_2$≡CH–CFH$_2$</td>
<td>-140.093</td>
</tr>
<tr>
<td>6.</td>
<td>CH$_2$≡CF–CF$_2$H</td>
<td>-139.393</td>
</tr>
<tr>
<td>7.</td>
<td>CFH≡CH–CF$_2$H</td>
<td>-139.015</td>
</tr>
<tr>
<td>8.</td>
<td>CFH$_2$–CF$_2$H</td>
<td>-162.098</td>
</tr>
<tr>
<td>9.</td>
<td>CH$_2$≡CH–CF$_3$</td>
<td>-137.882</td>
</tr>
<tr>
<td>10.</td>
<td>CH$_3$≡CF$_3$</td>
<td>-162.928</td>
</tr>
</tbody>
</table>

Table 3.2. Ranking of 10 Molecules Closest to a Target Value of $\Delta H_{f,\text{target}}^0 = -150$ kcal/mol, Calculated from the GCM based MILP Model

runs, the parameters of the GA were varied. All runs exhibit the expected behavior: the normalized total fitness of the population is increasing. A GA is converged if the total fitness assumes a value that cannot be improved over a reasonable number of generations. The examples typically show convergence after 10 to 20 generations. The GA tries to maximize the total fitness of a population; it does not strive for a great diversity of specific individuals. Therefore, a converged population frequently has a high number of equal individuals with fitness values close to one and the algorithm may converge into a suboptimal population with little probability of improving. Figures 3.2 and 3.3 show that the runs with parameters $p_c = 0.25$ and $p_m = 0.01$ (solid curves, Fig. 3.2a and Fig. 3.3b) quickly converged to populations with a high normalized total fitness and exhibited small fluctuation around this value. This is attributed to few changes in these populations after 20 generations. This set of parameters is unfavorable because it does
Fig. 3.2. Development of the Normalized Total Fitness of a Population of 20 Molecules for different GA Parameters: If $p_m$ is too small (case b) the fitness of the population may deteriorate. The population of case a) is converged with many identical molecules.
Fig. 3.3. Development of the Normalized Total Fitness of a Population of 50 Molecules for different GA Parameters: The total fitness of populations a) and c) fluctuates stronger than in case b), indicating little diversity of the b)-population. Populations a) and c) are better candidates for a set of GA parameters that ensures sufficient sampling of the search space.
not generate enough diversity in its populations to sample a variety of molecules. The
dashed curves in Figs. 3.2 and 3.3 show larger fluctuations. This can be attributed
to a more random behavior of the algorithm, which frequently produces below-average
individuals, but also bears higher possibility of finding above-average individuals.

Because the best molecules are sought as solutions of the optimization problem,
it is not as important to find a good final population, but to find good individuals in
the course of running the algorithm. As indicated above, the parameter combination
$ p_c = 0.25 $ and $ p_m = 0.01 $ did not promise to maintain a successful (high normal fitness)
and diverse population, whereas other parameter combinations did (e.g., Fig. 3.3a and
c). For deciding on the set of GA parameters, a large number of parameter combinations
were evaluated by trial and error while recording $ P_{10, \text{best, avg}} $ and $ N_\text{functions} $. Table 3.3
shows the performance of the GA for an average set of parameters (upper part) and
the final set of parameters (lower part). The GA was run a 1000 times for all sets
of parameters that were tested to average over the different outcomes arising from the
random search method. The frequency of occurrence of crossovers and mutations agrees
well with their probabilities, for example in the upper part: population size × number of generations × crossover probability = number of crossovers = 150. This value is close to the computational average of 149.794 over 1000 runs. The set of parameters $p_c = 0.25$ and $p_m = 0.01$ showed only average performance ($P_{10,\text{best,avg}} = 55.49\%$) as expected from the discussion of Figs. 3.2 and 3.3. The lower part of Table 3.3 shows the set of parameters that was chosen as the best performing set. As evidenced by the high mutation probability ($p_m = 0.03$), the total fitness of this GA fluctuated, but with almost 9 out of 10 best molecules this GA exhibited very good performance. As one would expect, better sampling of the search space requires more function evaluations, which is also observed in Table 3.3.

**GA Based on DFT Calculations** Table 3.4 lists the best 10 molecules that were found by the GA, using Gaussian 98 with the B3LYP/6-31G(d) model chemistry for the calculation of $\Delta H_f^0$. Five of these molecules were also found by the GCM-based MILP.
<table>
<thead>
<tr>
<th>Rank</th>
<th>Molecule</th>
<th>$\Delta H^0_{f,\text{B3LYP/6-31G(d)}}$ [kcal/mol]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>CH$_2$=CH-CF$_3$</td>
<td>-150.004</td>
</tr>
<tr>
<td>2.</td>
<td>CF$_2$=C=CF$_2$</td>
<td>-138.093</td>
</tr>
<tr>
<td>3.</td>
<td>CFH$_2$-CH$_2$-CF$_2$H</td>
<td>-166.214</td>
</tr>
<tr>
<td>4.</td>
<td>CF$_2$=CF$_2$</td>
<td>-166.382</td>
</tr>
<tr>
<td>5.</td>
<td>CF$_2$=CF$_2$CH$_3$</td>
<td>-132.754</td>
</tr>
<tr>
<td>6.</td>
<td>CF$_2$H-CH$_2$-CFH$_2$ (conformer with 3.)</td>
<td>-167.397</td>
</tr>
<tr>
<td>7.</td>
<td>CF=CF-CF$_3$</td>
<td>-132.403</td>
</tr>
<tr>
<td>8.</td>
<td>CF$_3$H</td>
<td>-168.408</td>
</tr>
<tr>
<td>9.</td>
<td>CH$_3$-CF$_2$-CH$_3$</td>
<td>-130.983</td>
</tr>
<tr>
<td>10.</td>
<td>CFH$_2$-CF$_2$-CH$_3$</td>
<td>-169.939</td>
</tr>
</tbody>
</table>

Table 3.4. Ranking of 10 Molecules Closest to a Target Value of $\Delta H^0_{f,\text{target}} = -150$ kcal/mol, Calculated from the Quantum Chemical GA
Their ranks in Table 3.2 are listed in parentheses in Table 3.4. Note that the molecule from Table 3.1 that seems closest to the target value, 3,3,3-trifluoropropene (CH$_2$=CH–CF$_3$), was found by the DFT-based GA. The molecules ranked 4th, 8th and 9th in Table 3.4 are also listed in Table 3.1, indicating they can be synthesized. The list was compiled from the results (10 best molecules) of three runs of the GA. Each run took about 2.5 to 3 hours of wallclock time with evaluation times between 248.6 and 3,617.2 CPU seconds per molecule. The Gaussian 98 evaluations run parallel in each generation.

A number of molecules presented problems to the Model Builder of Gaussian 98 as the subsequent geometry optimization did not converge. The cause lies in the nature of Model Builder [88], which was designed to account for most cases involving organic molecules. It is not guaranteed to give reasonable initial geometries in all cases. For molecules that created this problem (temporarily no result files available for the GA), the GA paused and the initial geometry was improved manually. After submitting the result to the GA, the algorithm continued. Note that molecules ranked third and sixth in Table 3.4 have the same configuration but not the same conformation. Their molecular matrices (Eq. 2.5) are different, which caused the Model Builder to generate different initial geometries, converging into different local optima in the geometry optimization.
3.3 Case Study 2: Solvent Design

3.3.1 Problem Definition

The objective of this case study is to identify a solvent for the liquid-liquid extraction of benzene from cyclohexane. Solvent selection for liquid-liquid extraction or extractive distillation is based on various solvent criteria such as solvent selectivity, capacity, cost, safety requirements, wastewater load, environmental requirements and several more. Out of these, capacity, selectivity and environmental fate are chosen exemplarily as guiding properties for the solvent design case study. The solvent capacity $C_A$ is a measure of how well the solvent $S$ can dissolve a solute $A$ that is to be recovered from the mixture $A-B$, where $B$ is the carrier. Solvent selectivity $S_{A,B}$ indicates the preference of $S$ for $A$ compared to $B$. For a “first order” characterization of solvents, it is common practice to apply the limiting values of $C_A$ and $S_{A,B}$ at infinite dilution of $A$ and $B$ (e.g. Hradetzky et. al. [44]):

$$C_{\infty,A,S} = \frac{1}{\gamma_{\infty,A,S}}$$

(3.31)

$$S_{\infty,A,B} = \frac{\gamma_{\infty,B,S}}{\gamma_{\infty,A,S}}$$

(3.32)

For most solvents, capacity and selectivity are competing properties (solvents with high selectivity show only low capacity, and vice versa), which led to the formulation of their product $\omega$ as a more realistic basis for evaluation:

$$\omega = C_{\infty,A,S} \cdot S_{\infty,A,B}$$

(3.33)
The environmental impact of a solvent has been correlated to the octanol-water partition coefficient $K_{\text{OW}}$ \[^{114}\]. It is a measure of hydrophobicity because it describes the equilibrium partition between water and a nearly water-immiscible liquid phase. Also, 1-octanol is a good surrogate for the lipids in aquatic and animal biota, and the organic matter in soils and sediments. Furthermore, increasing values of $K_{\text{OW}}$ have been found to correlate with increasing bioaccumulation in the food chain (Lin and Sandler \[^{64}\]). The same authors have correlated $K_{\text{OW}}$ to the ratio of infinite dilution activity coefficients in pure water and pure 1-octanol:

$$\log_{10} K_{\text{OW},d} = b + a \log_{10} \frac{\gamma_{W,\infty}}{\gamma_{i,\infty}}$$  \hspace{1cm} (3.34)

with $a = -0.68$ and $b = 0.91$. In the same paper, the authors evaluated $\log K_{\text{OW}}$ for 40 compounds using quantum chemical solvation calculations and a previously developed group-contribution solvation (GCS) model\[^{63}\], which will be discussed in the next paragraph. Based on ideas of this GCS model, they also devised a group contribution method for the rapid calculation of $K_{\text{OW}}$ (GCSKOW). In this example problem, $C_{\infty,A,S}$ and $S_{\infty,A,B}$ are evaluated using the UNIFAC group contribution model and $K_{\text{OW}}$ using quantum chemistry (GCS) or, for filtering molecules, with the GCSKOW model.

**Infinite Dilution Partition Coefficients** Lin and Sandler \[^{63}\] developed a method to obtain infinite dilution partition coefficients $\frac{\gamma_{\infty} S}{\gamma_{\infty} S}$ based on complex quantum chemical solvation calculations. These coefficients are a measure of how a solute $S$ at infinite dilution partitions between solvents 1 and 2. Lin’s and Sandler’s approach is based
on the idea of combining the UNIQUAC\cite{1} activity coefficient (\(\gamma\)) model with the free energy of solvation (\(\Delta G^{\text{sol}}\)), which is available from quantum chemistry. \(\Delta G^{\text{sol}}\) is found by assuming that a single solute molecule is placed into a solvent, which is modeled as an electric continuum represented by four physical constants: dielectric constant, ionization potential, refractive index, and density. Models based on this assumption are called continuum solvation models\cite{111}. The solvation free energy change of a molecule placed in a fixed position into the continuum solvent \(\Delta G^{*\text{sol}}\) (the asterisk indicates the fixed position) consists of four components, assuming molecular rotation and vibration effects are neglected\cite{111, 63}:

\[
\Delta G^{*\text{sol}} = \Delta G^{\text{cav}} + \frac{\Delta G^{\text{el}} + \Delta G^{\text{dis}} + \Delta G^{\text{rep}}}{\Delta G^{\text{chg}}} \tag{3.35}
\]

The cavitation contribution \(\Delta G^{\text{cav}}\) is the work needed to form a sufficient cavity in the solvent for transferring a molecule from the gas phase into the solvated state. The electrostatic component \(\Delta G^{\text{el}}\) represents the contribution from the electrostatic charge distribution that arises on the molecular surface and its electrostatic interaction with the solvent. The dispersion contribution \(\Delta G^{\text{dis}}\) results from London dispersion attractions between solute and solvent. The repulsion contribution \(\Delta G^{\text{rep}}\) results from quantum-mechanical repulsions between solute and solvent. For a more in-depth explanation of these terms, the reader is referred to the literature\cite{61, 111}. Note that the three latter components arise because of the charges of a molecule while the first is due to the molecular size and shape only. Lin and Sandler summarize the latter three terms as the charging free energy \(\Delta G^{\text{chg}}\). The UNIQUAC activity coefficient model also distinguishes
between a *combinatorial* term that is based only on size and shape of the interacting molecules and a *residual* term that accounts for the molecular interaction:

\[
\ln \gamma_i = \ln \gamma_{i i}^{\text{comb}} + \ln \gamma_{i j}^{\text{res}}(u_{ii}, u_{ij}, u_{ji})
\]  

(3.36)

In Eq. 3.36, the interaction parameters \( u_{ii}, u_{ij}, \) and \( u_{ji} \) need to be determined from experimental data. Lin and Sandler[63] combined \( \Delta G_{\text{chg}} \) and the UNIQUAC model, eliminating the parameters \( u_{ii}, u_{ij}, \) and \( u_{ji} \) in order to avoid the need for experimental data. They arrived at the following expressions for the infinite dilution activity coefficients of a solute \( S \) in a solvent \( I \)

\[
RT \ln \gamma_{S/1}^\infty = RT \ln \gamma_{S/1}^{\infty, \text{comb}} + RT q_S (\tau_S - \tau_I) + (\Delta G_{\text{chg}}^{S/S} - \Delta G_{\text{chg}}^{S/I})
\]

(3.37)

and for the infinite dilution partition coefficient of a solute \( S \) in a solvent \( I \) and a solvent \( 2 \)

\[
RT \ln \frac{\gamma_{S/2}^\infty}{\gamma_{S/2}^{\infty, \text{comb}}} = RT \ln \frac{\gamma_{S/1}^{\infty, \text{comb}}}{\gamma_{S/2}^{\infty, \text{comb}}} + RT q_S (\tau_2 - \tau_1) + (\Delta G_{\text{chg}}^{S/S} - \Delta G_{\text{chg}}^{S/2})
\]

(3.38)

with

\[
\frac{\Delta G_{\text{chg}}^{i/i}}{RT} = q_i (\tau_i - 1 + \ln \tau_i)
\]

(3.39)
Eq. 3.39 can be used to access $\tau_i$ from $\Delta G_{i}^{\text{chg}}$ of a solvent molecule solvated in a dielectric continuum of itself. Here, $q_i$ is the relative Van-der-Waals surface of species $i$. The charging free energies $\Delta G_{i}^{\text{chg}}$ can be obtained from quantum chemical continuum solvation calculations of a single molecule $S$ solvated in a solvent $i$. It should be noted that the molecular structure parameters $r_i$ and $q_i$ of the UNIQUAC-based combinatorial term in Eqns. 3.37 and 3.38 are obtained from optimized molecular geometries of solutes and solvents, that is, from quantum chemistry. Unfortunately, activity coefficients calculated from this model do not agree well with experimental activity coefficients [63]. Therefore, Lin and Sandler adjusted a scale factor $\alpha$ that accounts for the size of each atom in the solvation calculation according to the functional group that the atom belongs to. These scale factors were adjusted for different solvents (water, acetonitrile, n-octanol and n-hexane) until sufficient agreement with experimental data was achieved. For example, a carbon in a -CH$_X$ group will be assigned a different $\alpha$ value than a carbon in a -CN group. Also, the $\alpha$ value of a carbon in a -CN group in water is different from a carbon in a -CN group in n-octanol. Lin and Sandler [63] found that infinite dilution partition coefficients $\gamma_{S/1}^\infty$ were in good agreement with experimental data. Combining Eqns. 3.38 and 3.34 yields a quantum chemistry-based model for $K_{OW}$ with only one adjustable parameter $\alpha$.

Quantum Chemical Continuum Solvation Model  For generating starting geometries, the molecular matrix $M$ (Eq. 2.5) was submitted to the Model Builder of Gaussian 98, followed by an inexpensive Hartree-Fock(HF)/STO-3G geometry optimization. For obtaining the charging free energies, the GCS method of Lin and Sandler [63] was used,
including the same quantum chemistry software: General Atomic and Molecular Electronic Structure System [98] (GAMESS). The equilibrium geometry of the solute in vacuum was obtained by a geometry optimization using HF with a Dunning-Hay double zeta valence (GAMESS: DZV) basis [25] with added polarization functions (DZP). Furthermore, one additional set of diffuse and polarization functions was added with exponents of one third of the exponent of each most diffuse DZP function, respectively, giving rise to a DZPsp(df) basis. The same vacuum equilibrium geometry was used in subsequent solvation calculations without further optimization. The solvation calculations were carried out at the same HF/DZPsp(df) level.

**Bit String Representation and Objective Function for Case Study 2** The search space for this example includes saturated aliphatic molecules with up to five carbon atoms in the backbone and the following functional groups: N(-C≡N), O(-C=O), -OH, -NO₂, -Clᵣ (n = 1, 2, 3). Each molecule can contain one of these functional groups or none, multiple functional groups are excluded. The bit string contains eight digits. Similar to the bit strings in the first example, the first three positions define the number of carbon atoms, positions four, five and six code for the functional group type and positions seven and eight code for the carbon atom to which the functional group is attached. The following constraints are implemented in the routine that translates the bit strings into the molecular matrix M: For avoiding redundancy, functional groups can only be attached to a carbon atom number \( C_i \) with

\[
C_i \leq \left\lfloor \frac{C_{\text{total}}}{2} \right\rfloor + 1 \quad (3.40)
\]
where $C_{\text{total}}$ is the total number of carbon atoms in the molecule and $[x]$ stands for the integer part of a real number $x$. For example, if the candidate molecule has five carbon atoms, the functional groups can only be attached to carbon atoms number one, two or three. Groups $\text{N}(-\text{C}≡\text{N})$ and $\text{-Cl}_3$ can only be attached to the first carbon atom and $\text{O}(-\text{C}=\text{O})$ cannot be attached to the first carbon atom. This limitation of group $\text{O}(-\text{C}=\text{O})$ arises from Lin and Sandler’s method, in which $\alpha$ values for ketones are parametrized but not for aldehydes. The objective function in this example is written in the form of Eq. 2.7 with a lower limit of $\omega = -1.5$, a slope of $\beta_\omega = 1.0$, an upper limit of $\log K_{\text{OW}} = 2.0$ and a slope of $\beta_{\log K_{\text{OW}}} = 7.0$. Thus, the algorithm favors individuals with high $\omega$ and low $\log K_{\text{OW}}$ values.

### 3.3.2 Results of Case Study 2

**Accuracy of the Quantum Chemical Model** When attempting to reproduce data of Lin and Sandler [64], it was found that the accuracy of their model cannot fully be reproduced because the results of the continuum solvation calculation are sensitive to the Cartesian coordinates of the input geometry. For different initial guesses, the geometry optimization using the DZPsp(df) basis will find geometries with almost identical internal coordinates but with different Cartesian coordinates. In the continuum solvation calculation, the grid for the molecular cavity is set up based on Cartesian coordinates of the nuclei. Differences in the grid can lead to differences in the $\Delta G^{\text{el}}, \Delta G^{\text{dis}}$ and $\Delta G^{\text{rep}}$ terms of Eq. 3.35 in the order of 0.1-0.2 kcal/mole each [71]. Since $K_{\text{OW}}$ depends on the difference between two charging free energies (Eqns. 3.34 and 3.38), this uncertainty can be magnified or cancelled when repeating the calculation with different initial Cartesian coordinates.
coordinates. Table 3.5 shows a comparison between Lin’s and Sandler’s results [64] and data that were reproduced for this work. The data can be reproduced only with a root mean square (RMS) of 0.199 log $K_{\text{OW}}$ units as opposed to the authors RMS of 0.114 (for the molecules shown). This difference in accuracy is attributed to the optimization of the group scale factor $\alpha$ (Section 3.3.1), which was performed based on the authors Cartesian coordinates for each optimized molecular geometry. These coordinates are different from the Cartesian coordinates obtained for this work because in the optimization procedure initial coordinates are generated automatically. For comparison: Among other properties, $K_{\text{OW}}$ data were used for the parametrization of the widely used COSMO-RS[50] statistical mechanics method, which is based on screening charge distributions obtained from quantum chemical COSMO[51] calculations. Eckert and Klant [26] report an RMS deviation of 0.471 log $K_{\text{OW}}$ units for a parametrization set of 301 compounds.

**Best Molecules and GA Tuning Based on UNIFAC and GCSKOW Group Contribution Methods**

Solving the fitness for the entire search space, which encompasses 52 molecules for this small example, yielded the 10 globally best compounds given in Table 3.6. Based on an analysis similar to the one in Section 3.2.2, the following GA parameters were chosen for Case Study 2: $\text{pop.\ size} = 30$, $N_{\text{generations}} = 20$, $p_c = 0.25$ and $p_m = 0.03$. 
<table>
<thead>
<tr>
<th>Compound</th>
<th>$\Delta \Delta G_{\text{Lin}}^{\text{chg}}$ [64]</th>
<th>log $K_{\text{OW, Lin}}$</th>
<th>$\Delta \Delta G_{\text{this work}}^{\text{chg}}$ [kcal/mol]</th>
<th>log $K_{\text{OW, this work}}$</th>
<th>log $K_{\text{OW, EXP}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>methane</td>
<td>1.884</td>
<td>1.07</td>
<td>1.889</td>
<td>0.956</td>
<td>1.09 [97]</td>
</tr>
<tr>
<td>ethane</td>
<td>3.004</td>
<td>1.84</td>
<td>2.994</td>
<td>1.674</td>
<td>1.81 [103]</td>
</tr>
<tr>
<td>propane</td>
<td>3.77</td>
<td>2.39</td>
<td>3.981</td>
<td>2.322</td>
<td>2.36 [97]</td>
</tr>
<tr>
<td>butane</td>
<td>4.55</td>
<td>2.96</td>
<td>4.832</td>
<td>2.902</td>
<td>2.89 [97]</td>
</tr>
<tr>
<td>methanol</td>
<td>-0.927</td>
<td>-0.81</td>
<td>-1.163</td>
<td>-1.106</td>
<td>-0.77 [103]</td>
</tr>
<tr>
<td>ethanol</td>
<td>-0.013</td>
<td>-0.17</td>
<td>-0.031</td>
<td>-0.363</td>
<td>-0.31 [96]</td>
</tr>
<tr>
<td>propanol</td>
<td>0.762</td>
<td>0.39</td>
<td>1.034</td>
<td>0.340</td>
<td>0.25 [96]</td>
</tr>
<tr>
<td>nitromethane</td>
<td>-0.022</td>
<td>-0.13</td>
<td>-0.001</td>
<td>-0.326</td>
<td>-0.33 [96]</td>
</tr>
<tr>
<td>nitroethane</td>
<td>0.640</td>
<td>-0.33</td>
<td>0.734</td>
<td>0.150</td>
<td>0.18 [96]</td>
</tr>
<tr>
<td>nitropropane</td>
<td>1.199</td>
<td>0.74</td>
<td>1.799</td>
<td>0.897</td>
<td>0.87 [96]</td>
</tr>
<tr>
<td>acetonitrile</td>
<td>-0.650</td>
<td>-0.55</td>
<td>-0.663</td>
<td>-0.701</td>
<td>-0.34 [96]</td>
</tr>
<tr>
<td>propionitrile</td>
<td>0.251</td>
<td>0.09</td>
<td>0.341</td>
<td>-0.031</td>
<td>0.16 [96]</td>
</tr>
<tr>
<td>acetone</td>
<td>-0.185</td>
<td>-0.23</td>
<td>-0.102</td>
<td>-0.235</td>
<td>-0.24 [96]</td>
</tr>
<tr>
<td>chloropropane</td>
<td>3.145</td>
<td>2.07</td>
<td>3.613</td>
<td>2.184</td>
<td>2.04 [96]</td>
</tr>
<tr>
<td>dichloromethane</td>
<td>2.097</td>
<td>1.38</td>
<td>2.252</td>
<td>1.343</td>
<td>1.25 [96]</td>
</tr>
<tr>
<td>1,1-dichloroethane</td>
<td>2.623</td>
<td>1.79</td>
<td>2.891</td>
<td>1.802</td>
<td>1.79 [103]</td>
</tr>
<tr>
<td>chloroform</td>
<td>2.814</td>
<td>1.98</td>
<td>2.584</td>
<td>1.699</td>
<td>1.9 [96]</td>
</tr>
<tr>
<td>1,1,1-trichloroethane</td>
<td>3.185</td>
<td>2.31</td>
<td>3.173</td>
<td>2.156</td>
<td>2.49 [103]</td>
</tr>
<tr>
<td>tetrachloromethane</td>
<td>3.869</td>
<td>2.82</td>
<td>3.085</td>
<td>2.181</td>
<td>2.64 [46]</td>
</tr>
<tr>
<td>2-propanol</td>
<td>0.297</td>
<td>0.1</td>
<td>1.102</td>
<td>0.403</td>
<td>0.05 [96]</td>
</tr>
<tr>
<td>2-chloropropane</td>
<td>3.041</td>
<td>2.02</td>
<td>3.420</td>
<td>2.073</td>
<td>1.9 [96]</td>
</tr>
<tr>
<td>RMS</td>
<td>0.114</td>
<td></td>
<td>0.199</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.5. Comparison of Quantum Chemical $K_{\text{OW}}$ Results from Solvation Calculations for Case Study 2
<table>
<thead>
<tr>
<th>Rank</th>
<th>Compound</th>
<th>$\omega_{\text{UNIFAC}}$</th>
<th>$\log K_{\text{OW, GCS} KOW}$</th>
<th>$\log K_{\text{OW, GCS}}$</th>
<th>$\log K_{\text{OW, EXP}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>acetonitrile</td>
<td>1.800</td>
<td>-0.380</td>
<td>-0.701</td>
<td>-0.34 [96]</td>
</tr>
<tr>
<td>2.</td>
<td>acetone</td>
<td>1.668</td>
<td>-0.114</td>
<td>-0.236</td>
<td>-0.24 [96]</td>
</tr>
<tr>
<td>3.</td>
<td>propionitrile</td>
<td>1.895</td>
<td>0.150</td>
<td>-0.031</td>
<td>0.16 [96]</td>
</tr>
<tr>
<td>4.</td>
<td>2-butanone</td>
<td>2.106</td>
<td>0.416</td>
<td>0.256</td>
<td>0.29 [96]</td>
</tr>
<tr>
<td>5.</td>
<td>nitromethane</td>
<td>0.860</td>
<td>-0.154</td>
<td>-0.326</td>
<td>-0.33 [96]</td>
</tr>
<tr>
<td>6.</td>
<td>nitroethane</td>
<td>1.264</td>
<td>0.375</td>
<td>0.150</td>
<td>0.18 [96]</td>
</tr>
<tr>
<td>7.</td>
<td>methanol</td>
<td>0.409</td>
<td>-0.697</td>
<td>-1.106</td>
<td>-0.77 [103]</td>
</tr>
<tr>
<td>8.</td>
<td>butyronitrile</td>
<td>1.962</td>
<td>0.679</td>
<td>0.605</td>
<td>0.60 [96]</td>
</tr>
<tr>
<td>9.</td>
<td>2-nitropropane</td>
<td>1.573</td>
<td>0.755</td>
<td>0.93 [43]</td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>2-pentanone</td>
<td>2.430</td>
<td>0.945</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.6. Case Study 2: The 10 Globally Best Molecules based on GCM and the 10 Best Molecules based on Quantum Chemical Solvation Calculations. The rankings are the same from 1. to 8. and the two last compounds switch ranks when changing the $\log K_{\text{OW}}$ evaluation method.
**GA Based on Continuum Solvation Calculations** The tuned GA generated almost the same list of 10 best candidate solvents as the one in the previous step. The resulting compounds are also listed in Table 3.6. The only difference is that 2-pentanone and 2-nitropropane switched ranks. The $\omega$ values in both cases are the same because capacity and selectivity were evaluated using UNIFAC in both cases. The log $K_{OW}$ values differ because they were obtained from different methods. Although the quantum chemical method (RMS=0.199) is less accurate than the GCSKOW model (RMS=0.14), the trends are clearly the same. For comparison, experimental log $K_{OW}$ values are provided. It is noteworthy that nitromethane, a classic solvent for the benzene-cyclohexane system [86], is among the suggested candidate solvents. The difference in computational expense is striking: the GCM-based GA runs in less than a minute and the quantum chemistry-based GA runs in 18 to 36 hours.

### 3.4 Summary and Conclusions

In this work, the use of *ab initio* calculations for the property evaluation in molecular design was explored. Two case studies were presented to assess the feasibility of the proposed approach. The first example involves the design of hydrofluorocarbons, best matching a particular $\Delta H_f^0$ target value and the second example is a solvent design study, for which three criteria were chosen for solvent selection: capacity, selectivity and environmental fate, represented by the octanol-water partition coefficient $K_{OW}$. In both examples, a genetic algorithm (GA) was deployed as the optimization procedure, which calls a quantum chemical code as a subroutine to evaluate the properties of selected candidate molecules. Furthermore, group contribution methods were applied for various
reasons: to tune the GA, to filter molecules whose deviation from the target value or upper bounds is considered too large and as an additional property evaluation method (case study 2). By running the algorithm several times, lists of 10 molecules with the highest fitness values were compiled for each example.

In case study 1, all molecules in Table 3.4 are known to exist in the ground state because they were obtained by DFT \textit{ab initio} calculations. All their geometry optimizations converged into local minima. Since the thermal energy contribution from 0 K to 298 K is small compared to the dissociation energy \( \sum D_j \), DFT results support the existence of these compounds. However, no information is obtained whether a compound can be synthesized.

In case study 2, all molecules in the search space are known to exist. Nitromethane, a known extraction solvent of the benzene-cyclohexane system, was identified as a candidate solvent. The fitness function (Eqns. 2.6 and 2.7) can readily be extended to more than one property, thus, accounting for more than one design objective.

Searching for candidates via genetic algorithms allows for incorporating multiple evaluation methods within one design application. Any property evaluation method can be used in this approach because the GA treats the property evaluation as a black box. The applicability of the proposed approach depends on properties that are accessible for calculation from quantum chemical models. Due to computational expense, quantum models are typically used to calculate very few molecules, frequently just one, and obtain a bulk property from other relationships such as statistical mechanics. This limits their range of applicability. For example, a diffusion coefficient results from the interaction of many particles and cannot readily be obtained by quantum chemical models.
Presently, a meaningful application for engineering applications is also limited by the accuracy of the existing quantum chemical methods, which in many instances still lies within the range of group contribution methods and does not necessarily warrant the computational expense of quantum chemistry. However, GCM depend on experimental data for group parametrization. This disadvantage will likely be overcome by quantum chemical methods in the future [95]. For example, Eckert and Klamt [26] have used COSMO-RS to predict VLE, LLE, SLE data, partition coefficients, and vapor pressures. In COSMO-RS, only atoms are parametrized, not functional groups. With the advent of more accurate and more efficient \textit{ab initio} methods, opportunities for a fruitful combination of optimal molecular design and quantum chemical property prediction are likely to emerge. The proposed approach can be a valuable tool for automatically building databases of molecules that combine a number of desired properties.
Chapter 4

Protein Design Background

Protein engineering is the process of identifying proteins with a desired function not found in nature by either modifying known proteins or by de-novo protein design. Known proteins can be modified by changing their DNA sequence, which codes for the protein. In a subsequent step, the DNA is transmitted to a host cell such as *E.coli* and the protein is expressed by the host. There are two major methods of altering the DNA sequence of proteins based on prior knowledge, i.e., aside from randomly changing the sequence and testing: DNA recombination and site-directed mutagenesis. In DNA recombination, small pieces of DNA from two or more parent sequences are cleaved and recombined, thus, obtaining DNA for proteins that combines favorable properties of their parent proteins. Typically, the parent sequences stem from proteins of similar function in different organisms, and they display large percentages of sequence identity. Site-directed mutagenesis refers to changing the DNA sequence of one given protein at specific residue positions. The changes are usually based on some knowledge about the protein’s structure and function, for example the location of the binding site. It is of considerable interest to support this search for new proteins computationally for reducing experimental cost and for furthering our understanding of protein function. The computational method introduced here aims at supporting protein design by mutagenesis through predicting multiple promising mutations of a given protein. The special focus
of this method lies in treating protein stability and protein function as two different objectives. The framework presented here allows to incorporate different functions to model the protein’s energy and the protein’s binding affinity with a known ligand. This is of interest because of the limitations of current models for either objective [4, 37].

4.1 Protein Structure and Rotamer Libraries

Modifying protein function requires changing the protein’s structure through changes in the sequence. Computational protein design is based on experimentally determined protein structures because the relationship between protein sequence, protein folding, and, therefore, the protein function is not yet fully understood. It is in most cases impossible to predict a correct (within experimental accuracy of RMSD=2.0Å) structure only from information about the protein sequence. Many structures are deposited in the Brookhaven Protein Data Bank (PDB)[7], a public resource.

Proteins are long macromolecular chains with an amino acid repeat unit of the form \( -((\text{NH})-(\text{H-C}_\alpha -\text{R})-(\text{C}=\text{O}))- \), where the side-chain residues R, which are attached to the \( \alpha \)-carbon atoms, can take 20 different forms. Amino acid nomenclature is provided in Table 4.1. The long central chain is the protein backbone (see also Figure 4.1). There is little rotation around the peptide bond \( \text{C}^\alpha-\text{N} \), which keeps the dihedral angle \( \omega \) (H-N-\( \text{C}_{n-1} \)-O) close to 180°. The flexibility of the backbone arises mainly from the dihedral angles \( \phi \) and \( \psi \). Starting from a PDB structure of a protein-ligand complex, the protein is modified by keeping the backbone fixed and exploring how placing different side chains in a residue position affects the system. The side chains are also flexible. Depending on its length, a side chain can have up to four rotational bonds with dihedral angles
<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Three-letter code</th>
<th>One-letter code</th>
<th>Charge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala</td>
<td>A</td>
<td>Nonpolar (hydrophobic)</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>Methionine</td>
<td>Met</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>Proline</td>
<td>Pro</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Trp</td>
<td>W</td>
<td></td>
</tr>
<tr>
<td>Valine</td>
<td>Val</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn</td>
<td>N</td>
<td>Polar, uncharged</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gln</td>
<td>Q</td>
<td></td>
</tr>
<tr>
<td>Glycine</td>
<td>Gly</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>Serine</td>
<td>Ser</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>Asp</td>
<td>D</td>
<td>Acidic</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>Glu</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg</td>
<td>R</td>
<td>Basic</td>
</tr>
<tr>
<td>Histidine</td>
<td>His</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys</td>
<td>K</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1. Amino Acid Names, Three-letter Codes, and One-letter Codes
Fig. 4.1. Protein Backbone
χₙ, χ₂, etc. (see Figure 4.2). Side-chain rotamer libraries consist of a list of discrete side-chain conformations and their associated probabilities determined from their frequency of occurrence in the PDB [89]. In most cases, these conformations correspond to “rotamers” or local minima on potential energy maps with frequencies predictable from conformational analysis of organic molecules. The discreetness of rotamers is enforced by barriers of 4-10 kcal/mol. As more high-resolution structures have become available in recent years, it has become possible to determine rotamer preferences as a function of backbone conformation. Through a rotamer library, combinations of backbone dihedral angles can be translated into a number of discrete rotamer conformation for each side chain. In this work, the rotamer library of Dunbrack and Cohen [89] was used. Once the rotamer library is built for a given protein, interactions of the side-chain rotamers with the fixed backbone, with each other, and with a ligand can be explored.

4.2 Energy Function for Protein Stability

Protein stability is enforced during the design process by maintaining energy changes within a limit from a minimum energy of the protein. The energy contributions are modeled using the energy function of Looger and Hellinga [66, 65]. The combined energy of one combination of rotamers (one design) is given by

\[ E_H = \sum_i E_{rb}(i) + \sum_i \sum_j E_{rt}(i,j) \]  (4.1)

where \( E_H \) designates the energy based on Looger and Hellinga, \( E_{rb}(i) \) describes a rotamer-backbone interaction between the amino acid side chain at residue position \( i \)
Fig. 4.2. Rotamer Example: Glutamic Acid
and the backbone and $E_{rr}(i,j)$ describes a rotamer-rotamer interaction between the side chains of residues $i$ and $j$. In Loogers and Hellingas model, $E_{rb}(i)$ and $E_{rr}(i,j)$ are comprised of terms for van-der-Waals forces, hydrogen bonding, entropic contributions in the form of immobilization penalties for side chains and a solvation contribution, which accounts for how much surface of the side chain rotamer is exposed to the solvent or buried into the protein. Atomic solvation parameters (ASP, e.g., [120, 81]) are used to evaluate if exposure of an atom to the surface is favorable or unfavorable for protein stability. These parameters model the Gibbs free energy of solvation of a solute molecule:

$$\Delta G_{solv} = \sum_i \sigma_i A_i \quad (4.2)$$

Here, $A_i$ is the conformation-dependent accessible surface area of each atom $i$. The parameter $\sigma_i$ represents the contribution to $\Delta G_{solv}$ of each atom $i$ per unit accessible area. All energy terms are given in more detail in Table 4.2.

4.3 Calculating Protein-Ligand Binding Affinities

The binding affinity is the standard Gibbs free energy $\Delta G^0$ of the reversible association reaction of a receptor $R$ and a ligand $L$ when forming a receptor-ligand complex $R-L$. As the calculation of the binding affinity is at the center of any computational design of receptors or ligands, accurate modeling of this property is of utmost importance. However, the problem of obtaining accurate binding affinities is far from being resolved and no single method has been identified that is superior for all receptor-ligand binding problems [37, 2]. Three general approaches are used: free-energy perturbation
<table>
<thead>
<tr>
<th>Physical Effect</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Van-der-Waals energy</td>
<td>( \varepsilon_{ij} \cdot \left( \frac{r_{ij}}{d_{ij}} \right)^{12} - 2 \left( \frac{r_{ij}}{d_{ij}} \right)^{6} )</td>
</tr>
<tr>
<td>(Lennard-Jones-type Potential)</td>
<td></td>
</tr>
<tr>
<td>Electrostatic potential</td>
<td>( \frac{q_i q_j}{\epsilon d_{ij}} )</td>
</tr>
<tr>
<td>Hydrogen bonds</td>
<td>( f(d_{ij}, \theta) ) with: ( d_{ij} = \text{donor}(i)\text{-acceptor}(j) ) distance ( \theta = \text{donor}(i)\text{-H-acceptor}(j) ) angle</td>
</tr>
<tr>
<td>Frozen-core penalty</td>
<td>Penalty term if residue is “trapped” in the core of the protein, amino acid dependent</td>
</tr>
<tr>
<td>Solvation</td>
<td>Atomic solvation procedure (see also section 4.2) and parameters (ASP) from CHARMM [11]</td>
</tr>
</tbody>
</table>

Table 4.2. Energy Function of Looger and Hellinga [66, 65], \( i \) and \( j \) are indeces for atoms
calculations (FEP) and thermodynamic integration (TI), approaches based on additive free-energy contributions, and regression-based empirical scoring functions. FEP and TI methods are the only methods with a sound theoretical basis in statistical thermodynamics [36]. However, the computational expense for these simulations makes them prohibitive from a design point of view. Additive free-energy approaches assume that different contributions to $\Delta G^\circ$ can be calculated separately. The contributions can be defined differently in different models and are typically added in a “master equation”, for example [2]:

$$\Delta G^\circ = \Delta G_{\text{solvent}} + \Delta G_R + \Delta G_L + \Delta G_{R \leftrightarrow L} + \Delta G_{\text{motion}}$$  \hspace{1cm} (4.3)

$\Delta G_{\text{solvent}}$ accounts for the difference in free energy due to the changes in water solvation for $R$ and $L$ upon complex formation. The following two terms represent the conformational free energy changes of $R$ and $L$, $\Delta G_{R \leftrightarrow L}$ stands for the intermolecular interactions between $R$ and $L$ in the complex (van der Waals and electrostatics) and $\Delta G_{\text{motion}}$ includes the changes in rotational, translational and vibrational motion. Each of these contribution terms has enthalpic and entropic contributions, except $\Delta G_{R \leftrightarrow L}$, which is purely enthalpic, and, thus ($P\Delta V$ is negligible in solution), purely energetic. Care must be taken when modeling contributions for a particular system in order to account for effects like enthalpy-entropy compensation (see section 6.1.2). The individual contributions are obtained from force fields such as CHARMM and other atomic models, e.g., for the solvation contributions, from the solvent accessible area of individual atoms. Empirical scoring functions are based on the additivity assumption as well. They differ
from the additive $\Delta G^\circ$ contribution methods in that the parameters for the contribution terms are obtained from experimental data of receptor-ligand complexes. Also, the physical basis of the regression equations can be more obscure than in the approaches discussed above [37].

4.4 Computational Protein Design for Improved Ligand Binding

Some accounts are published of improved protein-ligand binding sites using computer-aided design strategies. Hellinga and Richards [41] developed the DEZYMER algorithm, which builds new ligand binding sites into a protein of known three-dimensional structure by altering sequence and rotamers but leaving the protein backbone intact. This approach works well for simple-ligand binding sites such as metal ions, which display well-defined coordination geometries and tight *intrinsic* binding [23]. Also, more recent work of Hellinga's group [70, 6] indicates that this method has not been used to design binding sites for larger ligands such as glucose or in a less defined binding situation. Another approach for improving protein-ligand binding was pursued by Mayo and co-workers. The authors found mutants of the $\alpha M/\beta 2$ integrin I domain with improved binding affinity for the ligand iC3b [99] and mutants of calmodulin (CaM) with improved binding specificity for smooth muscle myosin light chain kinase (smMLCK) over six other target peptides [100]. In both cases, the optimization goal was the stabilization of either the protein alone [99] or the protein-ligand complex [100]. That is, side chain rotamers were evaluated seeking proteins with lower total energies than that of the the wild type, and eliminating large numbers of rotamers based on the Dead-End Elimination theorem (DEE) [24]. The mutants were screened experimentally for improved binding affinity or
specificity. The residues were apparently varied depending on the availability of complex structures. No structure was available for the ΩMβ2 integrin I - iC3b complex. Consequently, the design procedure was limited to residues in the protein’s hydrophobic core and residues close to the binding site were kept as wild type. For the CaM - smMLCK complex, a structure is available (1cdl.pdb), which suggests only one binding mode. The design procedure included 24 fully buried CaM residues that are within 4Å of the target peptide.

A great driving force for development of binding functions (Section 4.3) is the field of protein-ligand docking, which designates the problem of predicting the “correctly” bound complex when given a protein, a ligand and their atomic coordinates [40]. Generally, binding occurs at the intended binding site of the protein, however, there is no guarantee and alternative binding sites may be found. Docking methods aim at the search of ligands as prospective drugs (lead generation) and are highly dependent on good search algorithms and good predictions of the binding affinity. In a recent publication, Taylor and Burnett [105] propose a docking strategy called DARWIN, which is based on a genetic algorithm and the CHARMM energy function. One of their examples is ConA - α-D-mannopyranoside docking. Distinguishing between an energy function used during energy minimization of the ligand and an evaluation function for scoring, the authors observe that local minimizations of candidate dockings are very important because a small displacement of the ligand can cause van der Waals collisions with a large rise in energy, therefore overlooking good candidates with seemingly unfavorable energy values (false negatives).
4.5 Scoring Functions for Protein-Ligand Binding

The protein design example is a reversed docking problem without the search for the binding site. Three functions used in the docking literature for rapid scoring of design alternatives were chosen: the piecewise-linear potential (PLP) [35], the FlexX score [92], which is based on Böhm’s score [8], and the CHARMM energy function [11]. The PLP score is

\[
S_{\text{PLP, hb}, ij} = \begin{cases} 
0 & d_{ij} > 3.4\text{ Å} \\
(-22.67 + 6.67 \cdot d_{ij}) & 3.1 < d_{ij} \leq 3.4 \\
-0.5 & 2.6 < d_{ij} \leq 3.1 \\
(15.34 - 6.67 \cdot d_{ij}) & 2.3 < d_{ij} \leq 2.6 \\
(20 - 8.69 \cdot d_{ij}) & d_{ij} \leq 2.3 
\end{cases}
\]  

(4.4)

for nondirectional hydrogen bonds between donor \(i\), acceptor \(j\). Donor and acceptor are heavy atoms, no hydrogen atoms are used, \(d_{ij}\) is the distance between \(i\) and \(j\). PLP has a similar function for steric interactions between heavy atoms:

\[
S_{\text{PLP, st}, ij} = \begin{cases} 
0 & d_{ij} > 5.5\text{ Å} \\
(-2.2 + 0.4 \cdot d_{ij}) & 4.5 < d_{ij} \leq 5.5 \\
-0.4 & 3.6 < d_{ij} \leq 4.5 \\
(6.8 - 2.0 \cdot d_{ij}) & 3.4 < d_{ij} \leq 3.6 \\
(20 - 5.88 \cdot d_{ij}) & d_{ij} \leq 3.4 
\end{cases}
\]  

(4.5)
The FlexX score is given by

$$\Delta G_{\text{binding}} = \Delta G_{\text{hb}} \sum_{\text{hb}} f(\Delta R, \Delta \alpha) + \Delta G_{\text{lipo}} \sum_{\text{lipocont.}} f^*(\Delta R) \quad (4.6)$$

Here, $f(\Delta R, \Delta \alpha)$ and $f^*(\Delta R)$ are scaling functions, penalizing deviations from ideal geometries with

$$f(\Delta R, \Delta \alpha) = f_1(\Delta R)f_2(\Delta \alpha) \quad (4.7)$$

$$f_1(\Delta R) = \begin{cases} 
1 & \Delta R \leq 0.2\text{Å} \\
1 - (\Delta R - 0.2)/0.4 & \Delta R \leq 0.6 \\
0 & \Delta R > 0.6
\end{cases} \quad (4.8)$$

$$f_2(\Delta \alpha) = \begin{cases} 
1 & \Delta \alpha \leq 30^\circ \\
1 - (\Delta \alpha - 30)/50 & \Delta \alpha \leq 80^\circ \\
0 & \Delta \alpha > 80^\circ
\end{cases} \quad (4.9)$$

$$f^*(\Delta R) = \begin{cases} 
0 & \Delta R > 0.6\text{Å} \\
1 - (\Delta R - 0.2)/0.4 & 0.2 < \Delta R \leq 0.6 \\
1 & -0.2 < \Delta R \leq 0.2 \\
1 - (\Delta R - 0.2)/0.4 & 0.2 < \Delta R \leq 0.6 \\
(\Delta R + 0.6)/0.2 & \Delta R > 0.6
\end{cases} \quad (4.10)$$
In $f(\Delta R, \Delta \alpha)$, $\Delta R$ is the deviation from an ideal hydrogen bond distance of 1.9 Å and $\Delta \alpha$ is the deviation from 180°. In $f^*(\Delta R)$, the ideal difference is formed with atoms of lipophilic groups and the ideal distance is the sum of the van-der-Waals radii, each increased by 0.6 Å. FlexX is an all-atom score. The CHARMM function that was used as a score includes a van-der-Waals term, the electrostatic potential, and ASP’s to account for de-solvation upon binding (see also Table 4.2).

4.6 A Mixed-Integer Linear Programming Formulation to Find Optimal Rotamer Combinations

After generating rotamer libraries depending on the protein’s backbone structure, and calculating energy values and rotamer-ligand binding scores, an optimal combination of rotamers must be found. For each residue-design position, there are 19 possible amino acids (proline is excluded), which could be placed into the position, each with all its rotamers, resulting in up to 320 potential rotamers per design position for the rotamer library that was used in this work [89]. This is achieved by using a mixed-integer linear programming (MILP) formulation of the problem, in which the choice of a particular rotamer $r$ at position $i$ is represented by the binary variable $y(i,r)$, which can only assume the integer values 0 or 1. The problem is then written as

$$\min \left( S = \sum_i \sum_r y(i,r) S_{rb}(i,r) \right)$$  \hspace{1cm} (4.11)
subject to:

$$E = \sum_{i} \sum_{r} y(i, r) E_{rb}(i, r) + \sum_{i} \sum_{r} \sum_{j>i} \sum_{s} y(i, r) \cdot y(j, s) \cdot E_{rl}(i, r, j, s)$$  \hspace{1cm} (4.12)$$

$$E \leq E_{H,\text{min}} + \Delta E$$  \hspace{1cm} (4.13)$$

Only one rotamer can be picked at each design position:

$$\sum_{r} y(i, r) = 1 \quad \text{for all } i, y(i, r) \in \{0, 1\}$$  \hspace{1cm} (4.14)$$

$E_{H,\text{min}}$ in Eq. 4.13 is obtained in a preceding energy minimization step, in which the score $S$ is not considered:

$$E_{H,\text{min}} = \min(E)$$  \hspace{1cm} (4.15)$$

Eq. 4.12 is nonlinear due to the product of binary variables $y(i, r)$ and $y(j, s)$.

$$w(i, r, j, s) = y(i, r) \cdot y(j, s) \quad \text{with} \quad 0 \leq w(i, r, j, s) \leq 1$$  \hspace{1cm} (4.16)$$

Linearity in the continuous subproblems in the branch-and-bound tree search (see also Section 3.2.1) greatly improves the model because the optimal solution of a linear problem is proven to be globally optimal, and because linear subproblems are much easier to solve.

Recognizing that the continuous variable $w(i, r, j, s)$ in Eq. 4.16 can only assume binary
values, a number of linear constraints is formulated, which enforce the same outcome.

\[ w(i, r, j, s) \leq y(i, r) \quad \text{for all } i, j > i, r, s \]  
(4.17)

\[ w(i, r, j, s) \leq y(j, s) \quad \text{for all } i, j > i, r, s \]  
(4.18)

\[ w(i, r, j, s) \geq y(i, r) + y(j, s) - 1 \quad \text{for all } i, j > i, r, s \]  
(4.19)

\[ \sum_r \sum_s w(i, r, j, s) = 1 \quad \text{for all } i, j > i \]  
(4.20)

The index \( j \) is always enforced to be larger than index \( i \). This avoids redundant entries in the array \( w(i, r, j, s) \) and in the number of constraints that have to be generated. The same purpose is served by the following constraint:

\[ w(i, r, j, s) = 0 \quad \text{for all } i, j \leq i \]  
(4.21)

Using \( w(i, r, j, s) \), Eq. 4.12 can now be recast in linear form:

\[ E = \sum_i \sum_r y(i, r) E_{rb}(i, r) + \sum_i \sum_r \sum_{j>i} \sum_s w(i, r, j, s) \cdot E_{rr}(i, r, j, s) \]  
(4.22)
Chapter 5

A Test System: Mutations of the Acyl-Coenzyme A Binding Protein- Dodecanoyl-Coenzyme A System

To test the capabilities of the design procedure, a comparison with experimental data is provided. Unfortunately, only a few data exist of mutated proteins with experimental binding energies. Typically, when introducing site-directed protein mutations, researchers are interested in other protein functions, such as enzymatic activity. Results from the design procedure are compared with data from Kragelund et. al. [59], who measured binding affinities of the system acyl-coenzyme A binding protein (ACBP)-dodecanoyl-coenzyme A (dodecanoyl-CoA) after introducing one specific point mutation in various positions close to the binding site. At positions 21Asp, 28Tyr, 32Lys, and 54Lys, the authors introduced more than one point mutation, thus, enabling one to rank the mutants with respect to their $\Delta G_{\text{bind}}$ values. Using the optimization framework introduced in Chapter 4, it was attempted to computationally reproduce the experimental ranking.

The experimental system ACBP-dodecanoyl-CoA was modeled using the PDB structure of the system ACBP-palmitoyl-CoA (1aca.pdb), in which the acyl part of the ligand contains a $\text{C}_{15}$ alkane residual instead of a $\text{C}_{12}$ residual, as in the experimental system. In ACBP, ligand binding is very tight ($\Delta G_{\text{bind,exp}} = -11.25$ kcal/mol) and the preference for long-chain acyl-CoA is very high. There are 18 residues within 4 Å of any atom of the ligand, which contains 127 atoms itself. These residue positions are $\text{9Ala}$,
10Glu, 12Val, 13Lys, 15Leu, 18Lys, 21Asp, 22Glu, 24Met, 25Leu, 27Ile, 28Tyr, 31Tyr, 32Lys, 50Lys, 53Ala, 54Lys, and 73Tyr. Of these, 21Asp, 28Tyr, 32Lys, and 54Lys were taken to be residue-design positions with a complete rotamer set. The other positions are treated as preserved design positions, thus, making them part of the rotamer library with the reduced rotamer set of their respective wild type (WT) residue. The mutant binding experiments of Kragelund et al. were modeled using the following protocol: All amino acids in the design positions were fixed, except for one at a time, which had its full rotamer range of all amino acids. This one position was, thus, free to “mutate”. Note that all other positions, which are close to the binding site, still have the rotamers of their respective amino acids.

The optimal rotamer combination was identified by minimizing the binding scores and keeping the Hellinga energy as a constraint to enforce protein stability. This minimization was carried out using the MILP formulation given in Section 4.6. Subsequently, the protein mutants were rebuilt and the ligand and binding site (within 4 Å of any atom of the ligand) were geometry-optimized using the CHARMM energy function in order to relax the discrete rotamer geometries in conjunction with the ligand. The MILP is formulated in such a way that each amino acid can be selected once in each MILP solution. Through integer cuts, lists were generated of the best designed mutants, the second best, etc. Each design was then scored by all scoring functions. The results are provided in Tables 5.1, 5.2, and 5.3 and in Figures 5.1, 5.2, and 5.3.

The computationally designed mutants in Table 5.1 were obtained by minimizing the PLP score, and Tables 5.2 and 5.3 were produced by minimizing the FlexX score and
<table>
<thead>
<tr>
<th>Mutant</th>
<th>$\Delta G_{\text{bind,exp}}$ [kcal/mol]</th>
<th>PLP Score</th>
<th>FlexX Score</th>
<th>CHARMM Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>D21A</td>
<td>-11.9 ± 0.4</td>
<td>-143.01</td>
<td>-12.20</td>
<td>-38.60</td>
</tr>
<tr>
<td>WT21=D</td>
<td>-11.25 ± 0.27</td>
<td>-151.65</td>
<td>-14.07</td>
<td>-43.89</td>
</tr>
<tr>
<td>D21H</td>
<td>-10.4 ± 0.3</td>
<td>-153.14</td>
<td>-12.40</td>
<td>-44.82</td>
</tr>
<tr>
<td>WT28=Y</td>
<td>-11.25 ± 0.27</td>
<td>-149.91</td>
<td>-11.37</td>
<td>-43.92</td>
</tr>
<tr>
<td>Y28A</td>
<td>-8.91 ± 0.16</td>
<td>-129.72</td>
<td>-11.83</td>
<td>-37.58</td>
</tr>
<tr>
<td>Y28F</td>
<td>-8.7 ± 0.4</td>
<td>-148.54</td>
<td>-10.87</td>
<td>-45.04</td>
</tr>
<tr>
<td>Y28N</td>
<td>-8.47 ± 0.03</td>
<td>-144.05</td>
<td>-9.55</td>
<td>-40.79</td>
</tr>
<tr>
<td>WT32=R</td>
<td>-11.25 ± 0.27</td>
<td>-151.65</td>
<td>-14.07</td>
<td>-43.89</td>
</tr>
<tr>
<td>K32R</td>
<td>-9.51 ± 0.04</td>
<td>-150.04</td>
<td>-13.63</td>
<td>-43.71</td>
</tr>
<tr>
<td>K32A</td>
<td>-7.98 ± 0.02</td>
<td>-143.48</td>
<td>-11.82</td>
<td>-51.03</td>
</tr>
<tr>
<td>K32E</td>
<td>-7.83 ± 0.07</td>
<td>-138.61</td>
<td>-9.91</td>
<td>-50.95</td>
</tr>
<tr>
<td>WT54=R</td>
<td>-11.25 ± 0.27</td>
<td>-151.65</td>
<td>-14.07</td>
<td>-43.89</td>
</tr>
<tr>
<td>K54M</td>
<td>-9.85 ± 0.08</td>
<td>-140.05</td>
<td>-9.65</td>
<td>-45.16</td>
</tr>
<tr>
<td>K54A</td>
<td>-9.46 ± 0.11</td>
<td>-126.94</td>
<td>-10.15</td>
<td>-41.85</td>
</tr>
</tbody>
</table>

Table 5.1. Binding affinities of experimental mutants of ACBP with dodecanoyl-CoA [59] and mutants designed by maximizing the PLP score of the ACBP-palmitoyl-CoA system. The designs were found by maximizing the PLP score, and the resulting mutant-ligand complexes were then evaluated three times, using each score individually.

<table>
<thead>
<tr>
<th>Mutant</th>
<th>$\Delta G_{\text{bind,exp}}$ [kcal/mol]</th>
<th>PLP Score</th>
<th>FlexX Score</th>
<th>CHARMM Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>D21A</td>
<td>-11.9 ± 0.4</td>
<td>-146.42</td>
<td>-13.12</td>
<td>-43.74</td>
</tr>
<tr>
<td>WT21=D</td>
<td>-11.25 ± 0.27</td>
<td>-153.20</td>
<td>-15.33</td>
<td>-41.90</td>
</tr>
<tr>
<td>D21H</td>
<td>-10.4 ± 0.3</td>
<td>-153.14</td>
<td>-12.40</td>
<td>-44.82</td>
</tr>
<tr>
<td>WT28=Y</td>
<td>-11.25 ± 0.27</td>
<td>-142.10</td>
<td>-14.87</td>
<td>-32.56</td>
</tr>
<tr>
<td>Y28A</td>
<td>-8.91 ± 0.16</td>
<td>-127.57</td>
<td>-10.99</td>
<td>-42.34</td>
</tr>
<tr>
<td>Y28F</td>
<td>-8.7 ± 0.4</td>
<td>-151.68</td>
<td>-13.71</td>
<td>-41.40</td>
</tr>
<tr>
<td>Y28N</td>
<td>-8.47 ± 0.03</td>
<td>-122.42</td>
<td>-9.49</td>
<td>-43.38</td>
</tr>
<tr>
<td>WT32=R</td>
<td>-11.25 ± 0.27</td>
<td>-153.88</td>
<td>-15.36</td>
<td>-41.34</td>
</tr>
<tr>
<td>K32R</td>
<td>-9.51 ± 0.04</td>
<td>-148.58</td>
<td>-13.76</td>
<td>-43.87</td>
</tr>
<tr>
<td>K32A</td>
<td>-7.98 ± 0.02</td>
<td>-141.08</td>
<td>-11.48</td>
<td>-51.51</td>
</tr>
<tr>
<td>K32E</td>
<td>-7.83 ± 0.07</td>
<td>-140.84</td>
<td>-12.36</td>
<td>-42.79</td>
</tr>
<tr>
<td>WT54=R</td>
<td>-11.25 ± 0.27</td>
<td>-153.20</td>
<td>-15.33</td>
<td>-41.90</td>
</tr>
<tr>
<td>K54M</td>
<td>-9.85 ± 0.08</td>
<td>-142.66</td>
<td>-11.74</td>
<td>-44.34</td>
</tr>
<tr>
<td>K54A</td>
<td>-9.46 ± 0.11</td>
<td>-140.87</td>
<td>-12.84</td>
<td>-48.27</td>
</tr>
</tbody>
</table>

Table 5.2. Binding affinities of experimental mutants of ACBP with dodecanoyl-CoA [59] and mutants designed by maximizing the FlexX score of the ACBP-palmitoyl-CoA system. The designs were found by maximizing the FlexX score, and the resulting mutant-ligand complexes were then evaluated three times, using each score individually.
<table>
<thead>
<tr>
<th>Mutant</th>
<th>$\Delta G_{\text{bind,exp}}$ [kcal/mol]</th>
<th>PLP Score</th>
<th>PlexX Score</th>
<th>CHARMM Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>D21A</td>
<td>-11.9 ± 0.4</td>
<td>-130.06</td>
<td>-10.72</td>
<td>-42.57</td>
</tr>
<tr>
<td>WT21=D</td>
<td>-11.25 ± 0.27</td>
<td>-142.50</td>
<td>-13.64</td>
<td>-40.89</td>
</tr>
<tr>
<td>D21H</td>
<td>-10.4 ± 0.3</td>
<td>-134.23</td>
<td>-10.33</td>
<td>-41.57</td>
</tr>
<tr>
<td>WT28=Y</td>
<td>-11.25 ± 0.27</td>
<td>-142.50</td>
<td>-13.64</td>
<td>-40.89</td>
</tr>
<tr>
<td>Y28A</td>
<td>-8.91 ± 0.16</td>
<td>-127.51</td>
<td>-9.69</td>
<td>-42.15</td>
</tr>
<tr>
<td>Y28F</td>
<td>-8.7 ± 0.4</td>
<td>-141.49</td>
<td>-12.23</td>
<td>-40.78</td>
</tr>
<tr>
<td>Y28N</td>
<td>-8.47 ± 0.03</td>
<td>-141.88</td>
<td>-9.74</td>
<td>-48.32</td>
</tr>
<tr>
<td>WT32=R</td>
<td>-11.25 ± 0.27</td>
<td>-133.47</td>
<td>-13.47</td>
<td>-35.97</td>
</tr>
<tr>
<td>K32R</td>
<td>-9.51 ± 0.04</td>
<td>-137.32</td>
<td>-9.71</td>
<td>-51.80</td>
</tr>
<tr>
<td>K32A</td>
<td>-7.98 ± 0.02</td>
<td>-132.76</td>
<td>-11.25</td>
<td>-45.42</td>
</tr>
<tr>
<td>K32E</td>
<td>-7.83 ± 0.07</td>
<td>-137.44</td>
<td>-11.88</td>
<td>-46.03</td>
</tr>
<tr>
<td>WT54=R</td>
<td>-11.25 ± 0.27</td>
<td>-142.50</td>
<td>-13.64</td>
<td>-40.89</td>
</tr>
<tr>
<td>K54M</td>
<td>-9.85 ± 0.08</td>
<td>-134.96</td>
<td>-10.99</td>
<td>-46.49</td>
</tr>
<tr>
<td>K54A</td>
<td>-9.46 ± 0.11</td>
<td>-119.86</td>
<td>-10.88</td>
<td>-45.64</td>
</tr>
</tbody>
</table>

Table 5.3. Binding affinities of experimental mutants of ACBP with dodecanoyl-CoA [59] and mutants designed by maximizing the CHARMM score of the ACBP-palmitoyl-CoA system. The designs were found by maximizing the CHARMM score, and the resulting mutant-ligand complexes were then evaluated three times, using each score individually.
Fig. 5.1. Binding affinities of experimental mutants of ACBP with dodecanoyl-CoA [59] and mutants designed by maximizing the PLP score of the ACBP-palmitoyl-CoA system. The designs were found by maximizing the PLP score, and the resulting mutant-ligand complexes were then evaluated three times, using each score individually. The bars show absolute $\Delta G_{\text{bind}}$, data and score values. The PLP scores are scaled by 0.1 and the CHARMM scores are scaled by 0.3.
Fig. 5.2. Binding affinities of experimental mutants of ACBP with dodecanoyl-CoA [59] and mutants designed by maximizing the FlexX score of the ACBP-palmitoyl-CoA system. The designs were found by maximizing the FlexX score, and the resulting mutant-ligand complexes were then evaluated three times, using each score individually. The bars show absolute $\Delta G_{\text{bind}}$ data and score values. The PLP scores are scaled by 0.1 and the CHARMM scores are scaled by 0.3.
Fig. 5.3. Binding affinities of experimental mutants of ACBP with deconoyl-CoA [59] and mutants designed by maximizing the CHARMM score of the ACBP-palmitoyl-CoA system. The designs were found by maximizing the CHARMM score, and the resulting mutant-ligand complexes were then evaluated three times, using each score individually. The bars show absolute $\Delta G_{\text{bind}}$ data and score values. The PLP scores are scaled by 0.1 and the CHARMM scores are scaled by 0.3.
the CHARMM score, respectively. The ranking is considered successful if the optimization solutions can reproduce the experimental ranking or parts of it. For example, for the PLP scoring of the mutations in position 28 in Table 5.1, alanine (A) was found on the bottom of the design list, and not in between tyrosine (Y) and asparagine (N) as in the experimental ranking. This is considered an unsuccessful ranking for alanine. However, if the unsuccessful alanine ranking was removed from the list, all other mutations of position 28 would be ranked correctly through the design procedure. Therefore, the wild type residue tyrosine (Y) and the site mutations phenylalanine (F) and asparagine (N) are considered as successfully ranked. Successful rankings are printed in bold font and unsuccessful rankings are printed in italic font throughout Tables 5.1, 5.2, and 5.3. Figures 5.1, 5.2, and 5.3 further illustrate these rankings.

The tables show that the CHARMM score appears not to be well suited to either combine promising rotamers from the rotamer library to maximize binding functionality (Table 5.3) and it is also not a good score for evaluating designs obtained by other scores (Tables 5.1 and 5.2) for the ACBP-palmitoyl-CoA system. Note in this context that the CHARMM score is different from the CHARMM energy function, which is used for the relaxation geometry optimization. In both functions, the solvation contribution $\Delta G_{\text{solv}}$ to $\Delta G_{\text{bind}}$ through the loss of solvent accessibility upon binding is quantified by using ASP's. However, in the CHARMM score, this contribution was modeled as

$$\Delta G_{\text{ASP,prot-lig-comp}} = -\frac{1}{2} \left( \Delta G_{\text{ASP,prot-lig}} - \Delta G_{\text{ASP,prot}} - \Delta G_{\text{ASP,lig}} \right)$$  (5.1)
for the protein-ligand complex and

\[ \Delta G_{\text{ASP,rot-lig-comp}} = \Delta G_{\text{ASP,rot}} - \Delta G_{\text{ASP,rot-lig}} \] (5.2)

for the rotamer-ligand complex. Each term in Eqns. 5.1 and 5.2 is modeled by obtaining the accessible surface area for the molecule or the complex given in the index of the term and then calculating \( \Delta G_{\text{solv}} \) from Eqn. 4.2. In contrast, in the CHARMM energy function \( \Delta G_{\text{solv}} \) (Eqn. 4.2) is modeled by obtaining the accessible surface area only once for the protein-ligand complex. The CHARMM score results show that this very simple model of \( \Delta G_{\text{solv}} \) for individual rotamers (Eqns. 5.1 and 5.2) does not improve the approximation of \( \Delta G_{\text{bind}} \) for ACBP-palmitoyl-CoA.

Tables 5.1 and 5.2 show that both, the PLP and the FlexX score perform much better in predicting the ranking of mutations at various positions of ACBP. For positions 28Tyr, 32Lys, and 54Lys, both scores rank the wild type residue highest, except for the PLP ranking of a FlexX design in position 28. Note especially, that the PLP design leads to many correct rankings using both scores in positions 28, 32, and 54. The ranking of alanine (A) residues is not very good with either score. One possible reason for this is that alanine residues replace larger residues in all positions, thus, reducing the potential number of contacts with the ligand. Since the backbone is fixed in the optimization framework, subtle changes in the backbone, which could lead to an increased number of protein-ligand contacts with other residues, cannot be identified using this model. Figures 5.1, 5.2, and 5.3 show that the successful rankings are reproduced qualitatively, but not quantitatively. For example, in Figure 5.1 at Position 28, the wild type residue
lysine leads to significantly stronger binding compared with the mutants than predicted by
the PLP score. This confirms that binding scores are not a good quantitative description
of the binding affinity, but that binding scores allow for successful ranking of design
alternatives.

In conclusion, both PLP and FlexX are suitable functions to score protein-ligand
binding of ACBP-palmitoyl-CoA and to identify promising mutations from the rotamer
library. The CHARMM score is not well-suited for either task in this system.
Chapter 6

Concanavalin A - Glucose Binding Site Optimization

The research presented in this chapter is motivated by the search for mutants of concanavalin A (ConA) with improved binding for glucose in order to further develop a biosensor application that is suitable for noninvasive glucose monitoring of diabetes patients. This search for improved mutants is aided by computational modeling of the binding process to predict candidate mutants and to better understand the binding process of the ConA - glucose system. One area of great prospect and need for biomolecules lies in biosensor applications because proteins can detect ligands or substrates with great specificity. Russel et. al. [94] have developed a fluorescence biosensor that is based on a photopolymerized poly(ethylene glycol) (PEG) hydrogel. Incorporated in the PEG is physically immobilized fluorescein isothiocyanate dextran (FITC-dextran) and tetramethylrhodamine isothiocyanate ConA (TRITC-ConA) is chemically conjugated into the hydrogel network using an α-acryloyl, ω-N-hydroxysuccinimidyl ester of PEG-propionic acid. In the absence of glucose, TRITC-ConA binds with FITC-dextran, and when the system is excited with monochromatic light, the FITC fluorescent response is quenched through fluorescence resonance energy transfer. Competitive glucose binding to TRITC-ConA liberates FITC-dextran, resulting in increased FITC fluorescence, which is proportional to the glucose concentration. Fluorescence-based glucose assay approaches have appeared in the literature as a means to continuously monitor glucose levels of diabetes
patients. Diabetes mellitus is a disease that inflicts an estimated 16 million Americans. Although the administration of insulin and control of diet have allowed patients to lead nearly normal lives, rigorous maintenance of glucose levels is necessary to slow the progression of long-term complications associated with diabetes. Current commercial methods of blood glucose measurement involving calorimetric strips or electrochemical sensors require a blood sample, which is both uncomfortable and aesthetically unpleasant and opens the body to infection. As a result, patient compliance with rigorous glucose testing is poor. A noninvasive monitoring approach could lead to better compliance and reduce the risk of infection. However, the ConA - glucose interaction needs improvement for making this system suitable for a biosensor application. Specifically, glucose sensitivity and specificity need improvement to arrive at suitable detection times and to better quantify glucose concentration.

6.1 Concanavalin A

The source of ConA is the jack bean (Canavalia ensiformis). ConA is poor in sulfur-containing amino acids and devoid of disulfide bonds. Its natural ligands are presumably glycans of higher order N-linked to asparagine. In its native state, ConA is a tetramer above pH 7 and a homodimer below pH 6. Each subunit consists of a 26 kDa monomer of 237 residues. Each monomer binds independently to the ligands leading to an enhanced avidity of the oligomeric ConA. Two divalent metal ions, a calcium atom and a transition metal (usually manganese), are tightly bound near the binding pocket. Each metal ion is coordinated by two water molecules. Although these metal ions do not participate directly in the binding of carbohydrates, their presence is a prerequisite
for sugar binding. The metal ions help position amino acid residues (aspartic acid and asparagine), which form three key hydrogen bonds with the carbohydrate. A structural water near the binding site bridges Asn14 and Arg228 with an oxygen of the ligand [80]. Plant ConA is produced through a series of highly unusual posttranslational processing steps, which renders it challenging to produce recombinant mutants in a prokaryotic host such as \textit{E. coli}. The protein is derived from a 290-residue glycosylated precursor polypeptide (pre-pro-ConA), which undergoes a complex series of post-translational reactions over a time period of several weeks, eventually leading to the formation of the mature 237- residue carbohydrate-binding protein. Processing steps include removal of the N-terminal signal sequence (residues 1-29) to produce pro-ConA, peptide-bond cleavage on the carboxyl side of Asn residues 148, 163 and 281 and the religation of the mature N-terminal fragment (residues 164-281) with the mature C-terminal fragment (residues 30-148) [10, 13]. Also, whereas mature ConA is not a glycoprotein, plant pro-ConA carries one N-linked oligosaccharide of the oligomannose type and this glycosylated pro-ConA is unable to bind to a cross-linked dextran. Deglycosylation is another necessary step in planta to activate pro-ConA. However, deglycosylation on one side and cleavage and religation on the other side seem to be independent of each other [91]. It has been shown that non-glycosylated recombinant pro-ConA expressed in \textit{E. coli} is active without polypeptide cleavage [73]. Furthermore, the posttranslational processing apparently occurs without a large change in the three-dimensional structure of pro-ConA [13]. Based on these facts, there are four major steps in the course of this research:
1. Preparation and expression of recombinant mature ConA: Based on the above facts it seems possible to rearrange the cDNA of pre-pro-ConA into a cDNA of mature ConA and express it in *E. coli*. Successful expression of active ConA is a prerequisite for the construction of ConA mutants.

2. Prediction of promising ConA mutants based on computational modeling: For reducing experimental effort and for gaining insight into the ConA - sugar binding process, it is desirable to support the search for promising ConA mutants through predictions of promising candidates based on computational protein design.

3. Preparation of ConA mutants: If mature ConA can be expressed from the rearranged cDNA, construction and expression of mutants is possible.

4. Evaluation of ConA mutants for applicability for a glucose biosensor: Candidate mutants will be immobilized in a PEG hydrogel, tagged with a fluorophore and tested for their capability to detect glucose at physiological concentrations.

6.1.1 cDNA for pre-pro-ConA

cDNA for pre-pro-ConA from 1985(!) was obtained from Prof. D.M. Carrington. The cDNA was amplified through PCR and the sequence was confirmed by Dr. Alex Horswill, currently in the group of Prof. Stephen J. Benkovic, Department of Chemistry, here at the Pennsylvania State University. Thanks are extended to these researchers for their help and expertise in the early stages of this project.
6.1.2 Crystal Structures of Concanavalin A - Sugar Complexes

X-ray crystal structures of the ConA - methyl α-D-mannopyranoside complex [76] (2.0 Å resolution, 5cna.pdb) and of the ConA - methyl α-D-glucopyranoside complex [9] (2.0 Å resolution, 1gic.pdb) are available. ConA - mannoside exists as a tetramer and ConA - glucoside exists as a dimer. All subunits contain one bound sugar, a Ca$^{2+}$ ion and a Mn$^{2+}$ ion, which are essential for binding the sugars. Bradbrook et. al. [9] summarized a number of experimental calorimetric literature data comparing mannoside and glucoside binding around pH=7 and T=300K: $\Delta \Delta G = 0.8 \pm 0.1$, $\Delta \Delta H = 1.6 \pm 0.3$ and $\Delta (T \Delta S) = 0.8 \pm 0.2$ (all in kcal/mol). Mannoside binds with a stronger affinity than glucoside (more favorable $\Delta G$ and also $\Delta H$), although glucoside binds with a more favorable entropy. Enthalpy and entropy are correlated for systems with weak intermolecular interactions. It can be interpreted that a gain of intermolecular binding is accompanied by a loss in degrees of freedom. This effect is called enthalpy-entropy compensation [37].

Bradbrook et. al. carried out “static” and dynamic (MD) simulations in an attempt to quantify enthalpic contributions to Eq. 4.3. In the “static” calculations, the authors energy-minimized (1) the positions of all added hydrogen atoms and (2) all atoms within a 12 Å sphere around the binding site and subsequently calculated $\Delta H_{R \leftrightarrow L}$, averaged for all subunits. For both minimizations, they found no difference between $\Delta H_{R \leftrightarrow L,\text{gluc}}$ and $\Delta H_{R \leftrightarrow L,\text{mann}}$ greater than standard deviation, that is, $\Delta H_{R \leftrightarrow L} = 0$. Similar results were obtained in preliminary CHARMM calculations for this work, as shown in Table 6.1. Stronger mannose binding could not be confirmed in these “static”
calculations based on the original PDB geometry and a geometry optimization. In the MD simulations, the authors did not succeed in quantifying a discernable difference \( \Delta \Delta H \) for all of the enthalpy terms in Eq. 4.3 due to the fact that the spread from averaging was in the same order of magnitude as the observed \( \Delta \Delta H \) terms. They were able to show that the interaction term favors mannoside binding \( \Delta H_{R \leftrightarrow L} = 4.9 \pm 3.6 \) kcal/mol, which was offset by the configuration term \( \Delta H_L = -5.5 \pm 3.1 \), favoring glucose binding. However, a noteworthy effect was observed through the MD simulations in some subunits, but not others: Specifically, hydrogen bonds between the the C2 hydroxyl and Leu99 and water mediated hydrogen bonds between the C2 hydroxyl and Thr226 were observed in the mannoside complex but not the glucoside complex. The C2 hydroxyl group in mannoside finds more potential binding partners for hydrogen bonding as the sugar moves in the binding side. All these observations suggest that: (1) the better binding affinity for mannoside compared to glucoside is a dynamic phenomenon, (2) structural waters may play a role, (3) enthalpy calculations alone can be misleading, and (4) the crystal structure, although the only one available, is not an ideal starting point.

### 6.2 Competitive Binding and the Binding Affinity

In this section, it is discussed why increasing the binding affinity is expected to improve the existing biosensor application of Russel et. al. [94]. Competitive binding has been modeled by various authors [93, 118, 45]. The glucose sensor under consideration is described by five types of components, namely the labeled ConA receptor, the labeled dextran ligand (which competes with glucose for the ConA binding sites), glucose and
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>no optimization, E(ConA+metals+H₂O+sugar)</td>
<td>-20.0</td>
<td>-20.1</td>
<td>-21.7</td>
</tr>
<tr>
<td>- E(H₂O+sugar)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- E(H₂O+ConA+metals)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ E(H₂O)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no optimization, E(ConA+metals+sugar)</td>
<td>-20.1</td>
<td>-20.1</td>
<td>-21.7</td>
</tr>
<tr>
<td>- E(sugar)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- E(ConA+metals)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sugar optimized, E(ConA+metals+sugar)</td>
<td>-56.6</td>
<td>-52.9</td>
<td>-63.8</td>
</tr>
<tr>
<td>- E(sugar)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- E(ConA+metals)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental Value[22]</td>
<td>-5.1</td>
<td>-5.1</td>
<td>-2.7</td>
</tr>
</tbody>
</table>

Table 6.1. Comparison of binding energies of two ConA-sugar complexes: For all complexes, hydrogens atoms were added to the PDB data using CHARMM. The locations of all heavy atoms were then fixed and the hydrogen atom locations were energy minimized. The term “optimization” in this table refers to the optimization of the sugar geometry. Different columns are shown for structures using Zn²⁺ and Mg²⁺ ions as substitutes for the original Mn²⁺ ions, since Mn is currently not parametrized in CHARMM. Zn²⁺ is another possible ion in the Mn²⁺ binding site [48].
two complexes: ConA:Dextran and ConA:Glucose. The equilibrium concentration of each component \([\text{ConA}], [\text{Dex}], [\text{Glc}], [\text{ConA:Dex}]\) and \([\text{ConA:Glc}]\) is determined by the binding constants \(K_{\text{Dex}}\) and \(K_{\text{Glc}}\):

\[
K_{\text{Dex}} = \frac{[\text{ConA:Dex}]}{[\text{Dex}] \cdot [\text{ConA}]} \tag{6.1}
\]

\[
K_{\text{Glc}} = \frac{[\text{ConA:Glc}]}{[\text{Glc}] \cdot [\text{ConA}]} \tag{6.2}
\]

Writing the bulk concentrations of ConA, dextran and glucose as \([\text{ConA}]_0\), \([\text{Dex}]_0\), and \([\text{Glc}]_0\) with

\[
[\text{ConA}]_0 = [\text{ConA}] + [\text{ConA:Glc}] + [\text{ConA:Dex}] \tag{6.3}
\]

\[
[\text{Dex}]_0 = [\text{Dex}] + [\text{ConA:Dex}] \tag{6.4}
\]

\[
[\text{Glc}]_0 = [\text{Glc}] + [\text{ConA:Glc}] \tag{6.5}
\]

yields a system of five equations with five unknown equilibrium concentrations, which is determined through the parameters \(K_{\text{Dex}}, K_{\text{Glc}}, [\text{ConA}]_0, [\text{Dex}]_0, \) and \([\text{Glc}]_0\). \([\text{ConA}]\) is a solution of the cubic equation

\[
a_3[\text{ConA}]^3 + a_2[\text{ConA}]^2 + a_1[\text{ConA}] + a_0 = 0 \tag{6.6}
\]
with

\[ a_3 = -K_{Dex} K_{Glc} \]  

(6.7)

\[ a_2 = ([\text{ConA}]_0 - [\text{Dex}]_0 - [\text{Glc}]_0) K_{Dex} K_{Glc} - K_{Dex} - K_{Glc} \]  

(6.8)

\[ a_1 = ([\text{ConA}]_0 - [\text{Glc}]_0) K_{Glc} + ([\text{ConA}]_0 - [\text{Dex}]_0) K_{Dex} - 1 \]  

(6.9)

\[ a_0 = [\text{ConA}]_0 \]  

(6.10)

The equilibrium concentration of [Dex], which drives the fluorescent response, can then be obtained through

\[ [\text{Dex}] = \frac{[\text{Dex}]_0}{K_{Dex} [\text{ConA}] + 1} \]  

(6.11)

Using data from [94] ([ConA]_0 = 20 \cdot 10^{-6} \text{ M}, [Dex]_0 = 8.695 \cdot 10^{-6} \text{ M}, changing [Glc]_0 from 0 to 33.34 \cdot 10^{-3} \text{ M}) and [93] (K_{Dex} = 1.5 \cdot 10^4 \text{ M}^{-1}, K_{Glc} = 320 \text{ M}^{-1}), and solving the model presented here, the influence of K_{Glc} and K_{Dex} on the fluorescent response can be estimated. The fluorescent response is measured as a relative increase compared to an initial value at [Glc]_0 = 0. With increasing glucose concentration, more FTTC- dextran is released and the fluorescence of these ligands is not quenched through the vicinity of TRITC-ConA anymore, thus, causing an increase in relative fluorescence.

Figure 6.1 shows several curves of [Dex] for various parameter combinations of K_{Glc} and K_{Dex}. The original curve is based on the data of [94]. The other curves were obtained by doubling the values for K_{Glc}, K_{Dex}, respectively and for K_{Glc} and K_{Dex} combined.
Fig. 6.1. Influence of dextran and glucose binding constants on the fluorescent response of a PEG hydrogel biosensor
The first thing to note is that the original [Dex] curve is not linear in the presented range of [Glc]₀, whereas the relative fluorescent response is [94]. This shows that [Dex] and the relative fluorescence are not directly proportional, indicating that other factors in the system also influence the response. When increasing $K_{\text{Dex}}$ only, the relative change in [Dex] increases as well. However, this was not observed by [94] who found that increasing $K_{\text{Dex}}$ through the use of mannosylated dextran decreases the relative fluorescence and attributed this to the reduction in the absolute value of [Dex] compared to the original case. Also, increasing $K_{\text{Dex}}$ resulted in a substantial increase in equilibration time of the sensor (40min vs. 10-12min for each 200mg/dL (11.11mM) increase in [Glc]₀). As indicated in Figure 6.1, increasing $K_{\text{Glc}}$ would yield a more desirable outcome. The relative increase in [Dex] would increase as well as the absolute value of [Dex]. Since the driving force for glucose diffusion increases, the response time is likely to decrease. Increasing the binding constant $K_{\text{Glc}}$ translates to increasing the Gibbs free energy of binding $\Delta G^\circ$, also called the binding affinity:

$$\Delta G^\circ = -RT \ln K_{\text{Glc}} \quad (6.12)$$

The difference in Gibbs free energy $\Delta G^\circ$ between the dissociated protein and ligand and the protein-ligand complex has a negative value. It is desirable to increase $|\Delta G^\circ|$ of ConA - glucose binding by finding one or more mutants of ConA with this improved functionality.
6.3 Prediction of Mutants Through Computational Protein Design

The design positions for this Problem were chosen through capturing all residues, which have atoms within a distance of 5 Å from the glucose ligand. These positions are Tyr12, Asn14, Thr97, Gly98, Leu99, Tyr100, Ser168, Ala207, Asp208, Thr226, Gly227, Arg228. Of these, positions Leu99, Ser168, Thr226, Arg228 were chosen as design residue positions, which means that all rotamers of all amino acids (except proline) are considered at these positions. At the other positions, all rotamers of the wild type amino acid are explored. ConA binds monosaccharides and polysaccharides. Monosaccharides fit completely into the binding site (see Figure 6.2). However, polysaccharides do not fit, but bind into the pocket with one sugar while the other sugars align towards the Tyr12 position. For this reason in particular, residue positions Ser168 and Thr226 were chosen as design-residue positions. Rotamers are evaluated with respect to their contribution to the total energy of the protein and with respect to their contribution to the binding affinity with the ligand. The energy contributions are modeled using the energy function of Looger and Hellinga [66]. The combined energy is given by:

$E_\text{H} = \sum_i E_{\text{rb}}^{(i)}(i) + \sum_i \sum_j E_{\text{rr}}^{(i,j)}(i,j)$ \hspace{1cm} (6.13)

Note that the above equation gives the energy of one design combination of rotamers. The rotamer library allows for calculation of energy terms for all rotamers, that is $E_{\text{rb}}^{(i,r)}$ for all rotamers $r$ at residue position $i$ and $E_{\text{rr}}^{(i,r,j,s)}$ for the interaction of rotamer $r$ at residue position $i$ with rotamer $s$ at residue position $j$. Based on the assumption that lower energy combinations of rotamers are more stable, one needs to find the rotamer
Fig. 6.2. ConA binding site in the wild type (left) and the best PLP-based design. Grey/red/white: ligand methyl alpha-D-glucopyranoside; residues [format: wild type - residue # - mutation]: Tyr12Tyr (turquoise), Asn14Asn (yellow), Gly98Gly (violet), Leu99Lys (green), Tyr100Tyr (white), Ser168Trp (red), Thr226Val (blue), Arg228Arg (magenta)
combination with the lowest energy through optimization methods such as dead-end elimination (DEE), Monte Carlo methods or mixed-integer linear programming (MILP). The latter method is used in the framework presented here. In order to describe the interaction between the protein and the ligand, a score \( S(i, r) \) is calculated, which describes the contribution to the total \( \Delta G^\circ \) of each rotamer \( r \) at residue position \( i \). Scoring functions are not necessarily used to accurately predict \( \Delta G^\circ \) data but rather to rank binding between several ligands: A stronger binder will score higher. Scoring functions do not perform equally well for different systems. Consensus scoring is, therefore, a recommended method where information from various scores is used to assess the protein-ligand interaction [102]. In the presented analysis, three different scores were used to predict favorable mutations: the partial linear potential (PLP) [35] score, the FlexX [92] score and the CHARMM [11, 68] energy function. These scoring functions contain terms to account for hydrogen bonds, steric interactions, van-der-Waals forces, electrostatic (Coulomb) interactions, and solvation based on Atomic Solvation Parameters (ASP). Each potential contains some of these terms but not all. For example, the PLP score contains only a term for non-directional hydrogen bonds and a term for steric interactions, thus, many physical phenomena of protein-ligand binding are not accounted for. However, in this analysis it is not attempted to decide on what is the best scoring function for all proteins. In light of the present uncertainty surrounding \( \Delta G^\circ \) predictions, it is decided on what appears to be the best scoring function for the ConA - glucose system. In a first step, the MILP form of the protein design problem was solved, which was introduced in Section 4.6. Solving this MILP, the rotamer combination that minimizes the score was identified. The energy \( E_H \) was constrained to remain within 10 kcal/mol
of the global minimum in order to ensure that the protein remains stable when residues are changed. The value of $\Delta E = 10$ kcal/mol was obtained in a trial-and-error search; it was identified as the value for which all scores show improvement for the designed mutants (consensus scoring). The MILP problem can be solved again by excluding the previous solution. By repeating this process a number of times, a list of the $n = 100$ best solutions was compiled. This protein design problem was solved using each score and, thus, three different lists with 100 best designs were produced. Presently, there are no experimental data available for $\Delta G^\circ$ of ConA mutants. The lists obtained in the protein design step were evaluated by testing whether the designed mutants display better binding then randomly generated mutants. The random case was generated by assigning a randomly chosen amino acid to each residue design position $i$ (the asterisk indicates that the choice of amino acids in the design positions is random but fixed) and then solving for the minimum energy combination of rotamers of these randomly chosen amino acids:

$$\min \left( \sum_i \sum_r y(i^*, r) E_{i^*, r} + \sum_i \sum_{j>i} \sum_s y(i^*, r) \cdot y(j^*, s) \cdot E_{i^*, r, j^*, s} \right)$$

(6.14)

After the minimum combination of rotamers has been found, the binding scores $S(i^*)$ were evaluated. This was also repeated 100 times, thus, obtaining a list of 100 random designs. Note the major difference between the designed mutants and the randomly designed mutants: in the protein design case, the particular amino acid that is placed at each design position $i$ is determined by maximizing the binding score. In the random
Table 6.2. Average binding scores of the best 100 designs based on maximizing binding scores compared with binding scores of 100 random designs (data in parentheses are standard deviations). The scores on which the designs are based are shown in each row. The columns show the scores by which each design was evaluated. Example: The designs in the fourth row were found by maximizing the PLP score. The evaluation of these designs using the other scores are shown in columns three (CHARMM) and five (FlexX). Instances that scored higher than the random designs are shown in boldface.

design case, the particular amino acid in a design position $i^*$ is randomly chosen but fixed. Only the rotamers of this amino acid are energy minimized and the best rotamer combination is identified.

For modeling the ConA-glucose system with this method, the crystal structure of ConA complexed with methyl $\alpha$-D-glucopyranoside (1gic.pdb), provided in the Protein Data Bank [7], was prepared. Water molecules were removed. Hydrogen atoms were added and the geometry was optimized using CHARMM. Table 6.2 compares average scores of random designs and of protein designs based on maximizing (absolute) protein-ligand binding scores. As expected, proteins that are designed using binding scores show better binding scores than the random designs. Note that designs based on maximizing the PLP score also have improved CHARMM and FlexX scores. This is not the case for the CHARMM and the FlexX based designs. Therefore, the PLP score appears to be
a better basis for finding ConA mutants with improved glucose binding and the FlexX score appears not to be well suited for this task. The complete list of 100 best PLP designs is given in Table A.1 in the Appendix. It was tested how the protein design results based on maximizing the PLP score compare to the variation in the amino acid sequences of the protein family. Based on a sequence alignment obtained from ClustalW [110], it was compared how often a particular amino acid is selected for a design position by the design procedure with how often this amino acid is selected by nature for the same design position. This can provide insight at which residue positions the model prefers other selections than nature. Knowledge of these differences can potentially identify a better design or shortcomings of the model. The results are shown in Figures 6.3, 6.4, 6.5, and 6.6.

Figure 6.6 shows that the PLP-based protein designs have a strong preference for arginine in position 228 as compared to the protein family. This is not surprising considering that the wild type protein has an arginine at that position. Thus, the design confirms arginine as a very good side chain solution for position 228 in ConA. Figure 6.4 shows a similar comparison for Ser168. It is notable that nearly 30\% of all lectins have serine in this position with an otherwise broad distribution of amino acids. The protein designs do not share the preference for serine and display an equally broad distribution. Note that three large residues have the biggest shares: arginine, lysine and tryptophan. This may be attributed to an opening in the binding site geometry formed by the closest residues surrounding the ligand as shown in Figure 6.2. This opening appears more closed when larger residues such as tryptophan are placed in the residue positions close to the opening (168 or 226) and, thus, enable more interactions with the ligand.
Fig. 6.3. Frequency of occurrence of amino acids in the PLP-based protein designs and in the lectin protein family for residue Leu99 (wild type residue)
Fig. 6.4. Frequency of occurrence of amino acids in the PLP-based protein designs and in the lectin protein family for residue Ser168 (wild type residue)
Fig. 6.5. Frequency of occurrence of amino acids in the PLP-based protein designs and in the lectin protein family for residue Thr226 (wild type residue)
Fig. 6.6. Frequency of occurrence of amino acids in the PLP-based protein designs and in the lectin protein family for residue Arg228 (wild type residue)
Figures 6.7 and 6.8 show histograms of 100 runs optimized by maximizing the PLP binding score, but with the energy constraint given in Eq. 4.22. This shows that the constraint works for this system as it forces the energy to a smaller average. Figures 6.9, 6.10, and 6.11 show plots of Hellings-energies versus the respective scores. From the irregular pattern, it is concluded that these types of energies are not correlated for this system. This is not surprising as ConA-glucose is a loosely binding system.

6.4 Future Research

This research aims at providing mutants of ConA with improved binding for glucose to further develop an existing biosensor application and at developing a computational framework to predict improved candidate mutants for reducing experimental cost. Given the time required to prepare a mutant with one point mutation, or, much more so, multiple mutations, and given the limitations of rational protein design [85], it is of interest to further develop a reliable computational tool to predict multiple point mutations. The project should consist of two phases: first, the preparation of recombinant mature ConA and, in the second and longer phase, the prediction, preparation, and testing of candidate mutants.

1. Preparation of recombinant mature ConA: The natural expression of ConA is very time consuming and not completely understood. Therefore, the expression of ConA mutants in planta appears to be a less favorable route for obtaining mutants. In the first phase of this project, it is necessary to determine if it is possible to prepare cDNA of mature ConA, for example through rearranging the cDNA of pre-pro-ConA, and express it in prokaryotic cells such as *E. coli*. Based on the information
Fig. 6.7. Histogram of Hellinga Energies for 100 RANDOM Residue Positions, ConA:Glucose, \( E_{\text{avg}} = -12.92 \text{ kcal/mol} \), \( E_{\text{min}} = -38.56 \text{ kcal/mol} \)
Fig. 6.8. Histogram of Hellinga Energies for 100 OPTIMIZED Residue Positions, ConA:Glucose, $E_{\text{avg}} = -34.13$ kcal/mol, $E_{\text{min}} = -38.56$ kcal/mol
Fig. 6.9. Hellinga Energy vs. CHARMM Scores for ConA:Glucose
Fig. 6.10. Hellinga Energy vs. PLP Scores for ConA:Glucose
Fig. 6.11. Hellinga Energy vs. FlexX Scores for ConA:Glucose
about ConA’s posttranslational processing provided in Section 6.1, correct folding of recombinant mature ConA seems possible. Successful and fast preparation of active ConA is a prerequisite for the preparation of ConA mutants.

2. Prediction, preparation, and testing of candidate mutants:

(a) Prediction of promising ConA mutants based on computational modeling:

Given the uncertainty of methods for rapid and accurate binding affinity calculations as well as the multitude of methods for calculating the structural energy of proteins, it seems advantageous to modularize these calculations and determine which combination of methods for calculating protein energies and protein-ligand binding affinities are appropriate for a given system. The framework presented here is a first step in this direction and can be applied to other systems. One of the presented scoring functions, the PLP score, appears to be a promising method to identify candidate mutants for the ConA-glucose system. Validation of these results through preparation and testing of predicted mutants and further model development in close collaboration with experimental project partners would provide further insights in the ConA-glucose binding process. Fields for improvement of the current design process are: extending the range of models for protein-ligand binding and protein structural energy, incorporating backbone flexibility into the model, and extending the search space of design residue positions to regions outside of the binding site. A great advantage of the ConA-glucose system is the fact that protein function almost directly translates into binding affinity. This
is not the case with most protein design applications, where the function is enzymatic activity and where the relationship between protein function and properties that are available to the modeler, e.g., binding affinity or overall protein - substrate complex stability, is much more obscure.

(b) Preparation of ConA mutants: If mature ConA can be expressed from the rearranged cDNA, candidate ConA mutants with one or more point mutations should be prepared.

(c) Evaluation of ConA mutants for applicability for a glucose biosensor: Candidate mutants can be immobilized in a PEG hydrogel, tagged with a fluorophore and tested for their capability to detect glucose at physiological concentrations.
Chapter 7

Conclusions

In this work, optimization-based frameworks for molecular design were introduced, targeting the design of two different classes of molecules: small "chemical" molecules and large proteins.

In the small-molecule design part, for the first time, quantum chemical methods are used for property estimation to design hydrofluorocarbon compounds with targeted heats of formation, and aliphatic hydrocarbon compounds were designed as optimal extraction solvents. The introduced framework uses a genetic algorithm to explore the search space, and this algorithm calls all property evaluation routines. As a proof of concept, known molecules were confirmed in two example calculations. Advantages of this concept are that molecules can be designed with regard to multiple objectives and that any form of property estimation method can be incorporated because the optimization algorithm treats the evaluation routines as black boxes. Therefore, a combination of property estimation methods can be utilized to search for better designs. For example, in cases where no fast and reliable estimation method and no experimental data are available, quantum chemical methods can be invoked. Limitations of using quantum chemistry are clearly the computational expense and the reduced ability to obtain bulk properties that are of interest to engineers. However, as more quantum-based property
estimation methods are being developed, the presented approach could be a valuable front end to automatically extend molecular databases.

In the protein design part, a framework is introduced to computationally predict mutated proteins with improved binding functionality. This framework extends previous approaches in that it targets directly the protein function as a design objective and maintains protein stability as a constraint. As an example, predictions of mutated concanavalin A designs were obtained and indicators are discussed as to why these predictions are promising. There are two advantages to this approach. Through utilizing a mixed-integer linear programming step, in which rotamer energies and binding scores are introduced as parameters, the method becomes independent from how the parameters are obtained. Consequently, the methods of how to obtain energies and binding scores are exchangeable. Especially in the case of binding scores, this is crucial because of their limited reliability at this point. Secondly, the method distinguishes between protein stability and protein function as different objectives. Particularly in systems with weak binding such as concanavalin A - glucose, this is valuable because a minimal total energy of the complex does not necessarily mean that the system is optimized for binding. This approach can easily be extended to other functional objectives for proteins, such as properties related to substrate specificity or enzymatic activity. Limitations of this approach arise from the limited reliability of the energy and scoring functions. Also, there is presently no systematic way of telling as to what energy difference is appropriate for the energy constraint.
Appendix

100 Best Concanavalin A Mutants for
Improved Glucose Binding as Predicted by
Maximizing the Absolute PLP Binding Score

This appendix provides the solution list of 100 predicted ConA mutants with improved glucose binding.

<table>
<thead>
<tr>
<th>Soln. #</th>
<th>99</th>
<th>168</th>
<th>226</th>
<th>228</th>
<th>PLP Score</th>
<th>FlexX Score</th>
<th>CHARMM Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>LYS</td>
<td>TRP</td>
<td>CYS</td>
<td>ARG</td>
<td>-36.62</td>
<td>-4.63</td>
<td>-31.00</td>
</tr>
<tr>
<td>2</td>
<td>LYS</td>
<td>TRP</td>
<td>VAL</td>
<td>ARG</td>
<td>-39.76</td>
<td>-5.82</td>
<td>-37.24</td>
</tr>
<tr>
<td>3</td>
<td>GLN</td>
<td>TRP</td>
<td>VAL</td>
<td>ARG</td>
<td>-35.91</td>
<td>-4.62</td>
<td>-37.52</td>
</tr>
<tr>
<td>4</td>
<td>THR</td>
<td>TRP</td>
<td>VAL</td>
<td>ARG</td>
<td>-42.60</td>
<td>-5.31</td>
<td>-35.43</td>
</tr>
<tr>
<td>5</td>
<td>LEU</td>
<td>TRP</td>
<td>VAL</td>
<td>ARG</td>
<td>-38.37</td>
<td>-5.51</td>
<td>-37.74</td>
</tr>
<tr>
<td>6</td>
<td>GLN</td>
<td>LYS</td>
<td>CYS</td>
<td>ARG</td>
<td>-35.84</td>
<td>-4.19</td>
<td>-38.84</td>
</tr>
<tr>
<td>7</td>
<td>GLN</td>
<td>ARG</td>
<td>LYS</td>
<td>ARG</td>
<td>-33.45</td>
<td>-4.55</td>
<td>-36.33</td>
</tr>
<tr>
<td>8</td>
<td>GLN</td>
<td>ARG</td>
<td>VAL</td>
<td>ARG</td>
<td>-33.36</td>
<td>-4.46</td>
<td>-35.43</td>
</tr>
<tr>
<td>9</td>
<td>GLN</td>
<td>ARG</td>
<td>CYS</td>
<td>ARG</td>
<td>-33.40</td>
<td>-4.57</td>
<td>-36.48</td>
</tr>
<tr>
<td>10</td>
<td>GLN</td>
<td>GLU</td>
<td>LYS</td>
<td>ARG</td>
<td>-40.00</td>
<td>-4.65</td>
<td>-53.84</td>
</tr>
<tr>
<td>11</td>
<td>LYS</td>
<td>LYS</td>
<td>GLU</td>
<td>ARG</td>
<td>-36.77</td>
<td>-5.52</td>
<td>-58.05</td>
</tr>
<tr>
<td>12</td>
<td>GLN</td>
<td>THR</td>
<td>VAL</td>
<td>ARG</td>
<td>-34.43</td>
<td>-4.71</td>
<td>-35.77</td>
</tr>
<tr>
<td>13</td>
<td>GLN</td>
<td>ASN</td>
<td>CYS</td>
<td>ARG</td>
<td>-33.94</td>
<td>-4.87</td>
<td>-36.42</td>
</tr>
</tbody>
</table>

Table A.1: List of 100 Best Designs based on Maximizing Absolute PLP Scores
<table>
<thead>
<tr>
<th>Soln. #</th>
<th>99</th>
<th>168</th>
<th>226</th>
<th>228</th>
<th>PLP Score</th>
<th>FlexX Score</th>
<th>CHARMM Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>LYS</td>
<td>TRP</td>
<td>CYS</td>
<td>LYS</td>
<td>-35.34</td>
<td>-4.64</td>
<td>-34.27</td>
</tr>
<tr>
<td>15</td>
<td>LYS</td>
<td>TRP</td>
<td>LYS</td>
<td>ARG</td>
<td>-35.13</td>
<td>-4.96</td>
<td>-35.70</td>
</tr>
<tr>
<td>16</td>
<td>LYS</td>
<td>TRP</td>
<td>ILE</td>
<td>ARG</td>
<td>-36.59</td>
<td>-5.38</td>
<td>-35.52</td>
</tr>
<tr>
<td>17</td>
<td>LYS</td>
<td>LYS</td>
<td>CYS</td>
<td>ARG</td>
<td>-34.84</td>
<td>-4.54</td>
<td>-36.53</td>
</tr>
<tr>
<td>18</td>
<td>LYS</td>
<td>LYS</td>
<td>LYS</td>
<td>ARG</td>
<td>-39.72</td>
<td>-4.72</td>
<td>-36.59</td>
</tr>
<tr>
<td>19</td>
<td>LYS</td>
<td>GLN</td>
<td>CYS</td>
<td>ARG</td>
<td>-36.35</td>
<td>-4.94</td>
<td>-33.20</td>
</tr>
<tr>
<td>20</td>
<td>LYS</td>
<td>GLU</td>
<td>CYS</td>
<td>ARG</td>
<td>-40.33</td>
<td>-4.49</td>
<td>-40.79</td>
</tr>
<tr>
<td>21</td>
<td>LYS</td>
<td>GLU</td>
<td>LYS</td>
<td>ARG</td>
<td>-38.26</td>
<td>-4.75</td>
<td>-38.80</td>
</tr>
<tr>
<td>22</td>
<td>LYS</td>
<td>GLU</td>
<td>ARG</td>
<td>ARG</td>
<td>-38.50</td>
<td>-4.81</td>
<td>-38.20</td>
</tr>
<tr>
<td>23</td>
<td>LYS</td>
<td>ASN</td>
<td>CYS</td>
<td>ARG</td>
<td>-36.14</td>
<td>-4.71</td>
<td>-31.90</td>
</tr>
<tr>
<td>24</td>
<td>LYS</td>
<td>CYS</td>
<td>LYS</td>
<td>ARG</td>
<td>-34.95</td>
<td>-4.63</td>
<td>-32.81</td>
</tr>
<tr>
<td>25</td>
<td>LYS</td>
<td>ASP</td>
<td>CYS</td>
<td>ARG</td>
<td>-37.67</td>
<td>-4.85</td>
<td>-35.91</td>
</tr>
<tr>
<td>26</td>
<td>LYS</td>
<td>LYS</td>
<td>ASP</td>
<td>ARG</td>
<td>-37.58</td>
<td>-5.74</td>
<td>-48.40</td>
</tr>
<tr>
<td>27</td>
<td>LYS</td>
<td>ASP</td>
<td>LYS</td>
<td>ARG</td>
<td>-37.38</td>
<td>-4.77</td>
<td>-34.70</td>
</tr>
<tr>
<td>28</td>
<td>LYS</td>
<td>GLN</td>
<td>LYS</td>
<td>ARG</td>
<td>-36.01</td>
<td>-4.77</td>
<td>-32.75</td>
</tr>
<tr>
<td>29</td>
<td>LYS</td>
<td>VAL</td>
<td>CYS</td>
<td>ARG</td>
<td>-35.73</td>
<td>-5.26</td>
<td>-34.47</td>
</tr>
<tr>
<td>30</td>
<td>LYS</td>
<td>CYS</td>
<td>CYS</td>
<td>ARG</td>
<td>-35.93</td>
<td>-5.27</td>
<td>-34.95</td>
</tr>
<tr>
<td>31</td>
<td>LYS</td>
<td>LYS</td>
<td>VAL</td>
<td>ARG</td>
<td>-36.58</td>
<td>-5.21</td>
<td>-37.54</td>
</tr>
<tr>
<td>32</td>
<td>LYS</td>
<td>ARG</td>
<td>CYS</td>
<td>ARG</td>
<td>-35.44</td>
<td>-4.71</td>
<td>-36.10</td>
</tr>
<tr>
<td>33</td>
<td>LYS</td>
<td>VAL</td>
<td>LYS</td>
<td>ARG</td>
<td>-35.82</td>
<td>-5.04</td>
<td>-33.62</td>
</tr>
<tr>
<td>34</td>
<td>LYS</td>
<td>GLN</td>
<td>ILE</td>
<td>ARG</td>
<td>-36.35</td>
<td>-5.07</td>
<td>-33.74</td>
</tr>
</tbody>
</table>

Table A.1: List of 100 Best Designs based on Maximizing Absolute PLP Scores
<table>
<thead>
<tr>
<th>Soln. #</th>
<th>99</th>
<th>168</th>
<th>226</th>
<th>228</th>
<th>PLP Score</th>
<th>FlexX Score</th>
<th>CHARMM Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>LYS</td>
<td>LEU</td>
<td>LYS</td>
<td>ARG</td>
<td>-35.84</td>
<td>-5.04</td>
<td>-34.19</td>
</tr>
<tr>
<td>36</td>
<td>LYS</td>
<td>THR</td>
<td>CYS</td>
<td>ARG</td>
<td>-34.67</td>
<td>-4.35</td>
<td>-30.95</td>
</tr>
<tr>
<td>37</td>
<td>LYS</td>
<td>LYS</td>
<td>THR</td>
<td>ARG</td>
<td>-36.68</td>
<td>-5.24</td>
<td>-37.06</td>
</tr>
<tr>
<td>38</td>
<td>LYS</td>
<td>ILE</td>
<td>CYS</td>
<td>ARG</td>
<td>-36.03</td>
<td>-5.27</td>
<td>-34.24</td>
</tr>
<tr>
<td>39</td>
<td>LYS</td>
<td>ASN</td>
<td>VAL</td>
<td>ARG</td>
<td>-36.56</td>
<td>-4.90</td>
<td>-32.66</td>
</tr>
<tr>
<td>40</td>
<td>LYS</td>
<td>ASN</td>
<td>LYS</td>
<td>ARG</td>
<td>-35.78</td>
<td>-5.23</td>
<td>-35.47</td>
</tr>
<tr>
<td>41</td>
<td>LYS</td>
<td>ARG</td>
<td>LYS</td>
<td>ARG</td>
<td>-38.31</td>
<td>-4.65</td>
<td>-31.50</td>
</tr>
<tr>
<td>42</td>
<td>LYS</td>
<td>GLY</td>
<td>CYS</td>
<td>ARG</td>
<td>-34.73</td>
<td>-4.59</td>
<td>-31.87</td>
</tr>
<tr>
<td>43</td>
<td>LYS</td>
<td>SER</td>
<td>CYS</td>
<td>ARG</td>
<td>-36.29</td>
<td>-5.04</td>
<td>-33.62</td>
</tr>
<tr>
<td>44</td>
<td>LYS</td>
<td>LEU</td>
<td>CYS</td>
<td>ARG</td>
<td>-36.12</td>
<td>-4.95</td>
<td>-33.48</td>
</tr>
<tr>
<td>45</td>
<td>LYS</td>
<td>THR</td>
<td>ILE</td>
<td>ARG</td>
<td>-36.39</td>
<td>-4.85</td>
<td>-33.08</td>
</tr>
<tr>
<td>46</td>
<td>LYS</td>
<td>ASP</td>
<td>ARG</td>
<td>ARG</td>
<td>-37.33</td>
<td>-4.82</td>
<td>-34.93</td>
</tr>
<tr>
<td>47</td>
<td>LYS</td>
<td>THR</td>
<td>ASP</td>
<td>ARG</td>
<td>-36.14</td>
<td>-4.68</td>
<td>-34.86</td>
</tr>
<tr>
<td>48</td>
<td>LYS</td>
<td>THR</td>
<td>THR</td>
<td>ARG</td>
<td>-36.34</td>
<td>-4.97</td>
<td>-33.49</td>
</tr>
<tr>
<td>49</td>
<td>LYS</td>
<td>CYS</td>
<td>THR</td>
<td>ARG</td>
<td>-36.39</td>
<td>-5.00</td>
<td>-33.87</td>
</tr>
<tr>
<td>50</td>
<td>LYS</td>
<td>CYS</td>
<td>ILE</td>
<td>ARG</td>
<td>-36.39</td>
<td>-5.15</td>
<td>-34.31</td>
</tr>
<tr>
<td>51</td>
<td>LYS</td>
<td>GLN</td>
<td>VAL</td>
<td>ARG</td>
<td>-36.23</td>
<td>-5.24</td>
<td>-34.59</td>
</tr>
<tr>
<td>52</td>
<td>LYS</td>
<td>ARG</td>
<td>VAL</td>
<td>ARG</td>
<td>-37.64</td>
<td>-5.38</td>
<td>-39.44</td>
</tr>
<tr>
<td>53</td>
<td>LYS</td>
<td>CYS</td>
<td>ARG</td>
<td>ARG</td>
<td>-37.20</td>
<td>-5.08</td>
<td>-36.13</td>
</tr>
<tr>
<td>54</td>
<td>LYS</td>
<td>GLU</td>
<td>VAL</td>
<td>ARG</td>
<td>-37.85</td>
<td>-4.29</td>
<td>-34.10</td>
</tr>
<tr>
<td>55</td>
<td>LYS</td>
<td>THR</td>
<td>VAL</td>
<td>ARG</td>
<td>-36.64</td>
<td>-4.96</td>
<td>-33.87</td>
</tr>
</tbody>
</table>

Table A.1: List of 100 Best Designs based on Maximizing Absolute PLP Scores
<table>
<thead>
<tr>
<th>Soln. #</th>
<th>99</th>
<th>168</th>
<th>226</th>
<th>228</th>
<th>PLP Score</th>
<th>FlexX Score</th>
<th>CHARMM Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
<td>LYS</td>
<td>VAL</td>
<td>VAL</td>
<td>ARG</td>
<td>-36.51</td>
<td>-4.94</td>
<td>-33.21</td>
</tr>
<tr>
<td>57</td>
<td>LYS</td>
<td>GLN</td>
<td>THR</td>
<td>ARG</td>
<td>-35.91</td>
<td>-5.34</td>
<td>-33.92</td>
</tr>
<tr>
<td>58</td>
<td>LYS</td>
<td>LYS</td>
<td>ILE</td>
<td>ARG</td>
<td>-36.51</td>
<td>-5.17</td>
<td>-37.36</td>
</tr>
<tr>
<td>59</td>
<td>LYS</td>
<td>ALA</td>
<td>VAL</td>
<td>ARG</td>
<td>-36.28</td>
<td>-5.25</td>
<td>-34.61</td>
</tr>
<tr>
<td>60</td>
<td>LYS</td>
<td>ASP</td>
<td>VAL</td>
<td>ARG</td>
<td>-37.71</td>
<td>-4.86</td>
<td>-35.63</td>
</tr>
<tr>
<td>61</td>
<td>LYS</td>
<td>CYS</td>
<td>VAL</td>
<td>ARG</td>
<td>-36.45</td>
<td>-5.10</td>
<td>-33.72</td>
</tr>
<tr>
<td>62</td>
<td>LYS</td>
<td>LEU</td>
<td>VAL</td>
<td>ARG</td>
<td>-34.91</td>
<td>-4.50</td>
<td>-31.88</td>
</tr>
<tr>
<td>63</td>
<td>LYS</td>
<td>THR</td>
<td>LYS</td>
<td>ARG</td>
<td>-35.99</td>
<td>-5.09</td>
<td>-34.11</td>
</tr>
<tr>
<td>64</td>
<td>LYS</td>
<td>VAL</td>
<td>ILE</td>
<td>ARG</td>
<td>-36.45</td>
<td>-5.15</td>
<td>-34.55</td>
</tr>
<tr>
<td>65</td>
<td>LYS</td>
<td>CYS</td>
<td>ASP</td>
<td>ARG</td>
<td>-37.13</td>
<td>-3.85</td>
<td>-28.89</td>
</tr>
<tr>
<td>66</td>
<td>LYS</td>
<td>ILE</td>
<td>VAL</td>
<td>ARG</td>
<td>-36.17</td>
<td>-5.17</td>
<td>-33.52</td>
</tr>
<tr>
<td>67</td>
<td>LYS</td>
<td>GLY</td>
<td>VAL</td>
<td>ARG</td>
<td>-34.87</td>
<td>-4.56</td>
<td>-32.26</td>
</tr>
<tr>
<td>68</td>
<td>LYS</td>
<td>SER</td>
<td>VAL</td>
<td>ARG</td>
<td>-36.52</td>
<td>-4.96</td>
<td>-33.65</td>
</tr>
<tr>
<td>69</td>
<td>LYS</td>
<td>GLU</td>
<td>ILE</td>
<td>ARG</td>
<td>-38.91</td>
<td>-5.26</td>
<td>-38.41</td>
</tr>
<tr>
<td>70</td>
<td>CYS</td>
<td>TRP</td>
<td>VAL</td>
<td>ARG</td>
<td>-36.29</td>
<td>-4.93</td>
<td>-36.86</td>
</tr>
<tr>
<td>71</td>
<td>LEU</td>
<td>ARG</td>
<td>VAL</td>
<td>ARG</td>
<td>-34.42</td>
<td>-5.01</td>
<td>-35.24</td>
</tr>
<tr>
<td>72</td>
<td>LEU</td>
<td>ARG</td>
<td>CYS</td>
<td>ARG</td>
<td>-34.61</td>
<td>-5.26</td>
<td>-36.38</td>
</tr>
<tr>
<td>73</td>
<td>THR</td>
<td>LYS</td>
<td>GLU</td>
<td>ARG</td>
<td>-40.37</td>
<td>-4.39</td>
<td>-52.42</td>
</tr>
<tr>
<td>74</td>
<td>VAL</td>
<td>TRP</td>
<td>VAL</td>
<td>ARG</td>
<td>-36.12</td>
<td>-4.17</td>
<td>-35.51</td>
</tr>
<tr>
<td>75</td>
<td>ARG</td>
<td>LYS</td>
<td>CYS</td>
<td>ARG</td>
<td>-33.50</td>
<td>-4.60</td>
<td>-36.33</td>
</tr>
<tr>
<td>76</td>
<td>THR</td>
<td>GLU</td>
<td>LYS</td>
<td>ARG</td>
<td>-36.64</td>
<td>-4.34</td>
<td>-40.57</td>
</tr>
</tbody>
</table>

Table A.1: List of 100 Best Designs based on Maximizing Absolute PLP Scores
<table>
<thead>
<tr>
<th>Soln. #</th>
<th>99</th>
<th>168</th>
<th>226</th>
<th>228</th>
<th>PLP Score</th>
<th>FlexX Score</th>
<th>CHARMM Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>77</td>
<td>ARG</td>
<td>GLU</td>
<td>CYS</td>
<td>ARG</td>
<td>-36.13</td>
<td>-3.95</td>
<td>-39.74</td>
</tr>
<tr>
<td>78</td>
<td>ARG</td>
<td>GLU</td>
<td>LYS</td>
<td>ARG</td>
<td>-38.63</td>
<td>-5.10</td>
<td>-45.70</td>
</tr>
<tr>
<td>79</td>
<td>ILE</td>
<td>TRP</td>
<td>VAL</td>
<td>ARG</td>
<td>-33.41</td>
<td>-3.08</td>
<td>-33.76</td>
</tr>
<tr>
<td>80</td>
<td>ARG</td>
<td>GLU</td>
<td>ARG</td>
<td>ARG</td>
<td>-36.09</td>
<td>-4.49</td>
<td>-38.99</td>
</tr>
<tr>
<td>81</td>
<td>THR</td>
<td>TRP</td>
<td>CYS</td>
<td>ARG</td>
<td>-37.59</td>
<td>-4.88</td>
<td>-35.98</td>
</tr>
<tr>
<td>82</td>
<td>THR</td>
<td>TRP</td>
<td>LYS</td>
<td>ARG</td>
<td>-36.49</td>
<td>-4.93</td>
<td>-37.65</td>
</tr>
<tr>
<td>83</td>
<td>ARG</td>
<td>THR</td>
<td>VAL</td>
<td>ARG</td>
<td>-33.82</td>
<td>-4.25</td>
<td>-32.20</td>
</tr>
<tr>
<td>84</td>
<td>ARG</td>
<td>VAL</td>
<td>VAL</td>
<td>ARG</td>
<td>-33.82</td>
<td>-4.34</td>
<td>-31.72</td>
</tr>
<tr>
<td>85</td>
<td>ARG</td>
<td>GLN</td>
<td>VAL</td>
<td>ARG</td>
<td>-33.47</td>
<td>-4.33</td>
<td>-32.54</td>
</tr>
<tr>
<td>86</td>
<td>ARG</td>
<td>THR</td>
<td>CYS</td>
<td>ARG</td>
<td>-33.48</td>
<td>-4.25</td>
<td>-31.88</td>
</tr>
<tr>
<td>87</td>
<td>ARG</td>
<td>GLU</td>
<td>VAL</td>
<td>ARG</td>
<td>-35.79</td>
<td>-4.21</td>
<td>-33.45</td>
</tr>
<tr>
<td>88</td>
<td>ARG</td>
<td>CYS</td>
<td>VAL</td>
<td>ARG</td>
<td>-33.51</td>
<td>-4.46</td>
<td>-32.48</td>
</tr>
<tr>
<td>89</td>
<td>ARG</td>
<td>GLN</td>
<td>CYS</td>
<td>ARG</td>
<td>-34.08</td>
<td>-4.41</td>
<td>-33.74</td>
</tr>
<tr>
<td>90</td>
<td>ARG</td>
<td>ARG</td>
<td>CYS</td>
<td>ARG</td>
<td>-35.20</td>
<td>-4.60</td>
<td>-34.37</td>
</tr>
<tr>
<td>91</td>
<td>ARG</td>
<td>ASN</td>
<td>CYS</td>
<td>ARG</td>
<td>-33.13</td>
<td>-4.43</td>
<td>-32.54</td>
</tr>
<tr>
<td>92</td>
<td>GLU</td>
<td>ARG</td>
<td>CYS</td>
<td>ARG</td>
<td>-35.05</td>
<td>-5.31</td>
<td>-36.50</td>
</tr>
<tr>
<td>93</td>
<td>GLU</td>
<td>ARG</td>
<td>VAL</td>
<td>ARG</td>
<td>-34.93</td>
<td>-5.39</td>
<td>-35.70</td>
</tr>
<tr>
<td>94</td>
<td>THR</td>
<td>TRP</td>
<td>VAL</td>
<td>LYS</td>
<td>-36.71</td>
<td>-4.26</td>
<td>-37.34</td>
</tr>
<tr>
<td>95</td>
<td>THR</td>
<td>LYS</td>
<td>LYS</td>
<td>ARG</td>
<td>-37.57</td>
<td>-5.61</td>
<td>-37.21</td>
</tr>
<tr>
<td>96</td>
<td>THR</td>
<td>LYS</td>
<td>THR</td>
<td>ARG</td>
<td>-35.87</td>
<td>-3.81</td>
<td>-34.46</td>
</tr>
<tr>
<td>97</td>
<td>THR</td>
<td>LYS</td>
<td>CYS</td>
<td>ARG</td>
<td>-36.98</td>
<td>-3.88</td>
<td>-35.95</td>
</tr>
</tbody>
</table>

Table A.1: List of 100 Best Designs based on Maximizing Absolute PLP Scores
<table>
<thead>
<tr>
<th>Soln. #</th>
<th>99</th>
<th>168</th>
<th>226</th>
<th>228</th>
<th>PLP Score</th>
<th>FlexX Score</th>
<th>CHARMM Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>98</td>
<td>THR</td>
<td>ARG</td>
<td>LYS</td>
<td>ARG</td>
<td>-35.81</td>
<td>-4.57</td>
<td>-33.58</td>
</tr>
<tr>
<td>99</td>
<td>THR</td>
<td>ARG</td>
<td>THR</td>
<td>ARG</td>
<td>-36.38</td>
<td>-4.60</td>
<td>-33.45</td>
</tr>
<tr>
<td>100</td>
<td>THR</td>
<td>ARG</td>
<td>CYS</td>
<td>ARG</td>
<td>-35.33</td>
<td>-4.44</td>
<td>-33.62</td>
</tr>
</tbody>
</table>

Table A.1: List of 100 Best Designs based on Maximizing Absolute PLP Scores
References


[60] J.K. Labanowski. Simplified introduction to ab initio basis sets. Technical report, Ohio Supercomputer Center, 1224 Kinnear Rd., Columbus, OH 43212-1163, 1996. E-mail: jkl@osc.edu, jkl@ohstpy.bitnet, URL(several)
http://www.ccl.net/cca/documents/basis-sets/basis.html.


Vita