DEEPSAM
(Diffusion Equation Evolutionary Programming Simulated Annealing Method)

A NEW HYBRID EVOLUTIONARY ALGORITHM
FOR FINDING THE LOWEST MINIMA OF POTENTIAL SURFACES:
APPROACH AND APPLICATIONS

Thesis submitted for the degree of
"Doctor of Philosophy"

By

Moshe Goldstein

Submitted to the Senate of the Hebrew University
March 2009 – Nissan 5769
This work was carried out under the supervision of

Professor R. Benny Gerber
Acknowledgements

First and most, I thank my advisor and mentor, Prof. R. Benny Gerber, for all his support, encouragement and insightful discussions during all these intellectually challenging years of work together, in times of failure as well as in times of success: in scientific research like in Life, we do not progress without a balanced dose of both. Benny always had the wisdom and the human sensibility to give the right advice at the right moment. Only with his wise and knowledgeable direction this research could safely arrive to its successful end.

Thank you, Benny! Thank you for all!

I thank my close friend and research partner, Dr. Erick Fredj, for suggesting me and Benny the first idea from which the algorithm presented here started to be devised and developed. Erick was my unofficial assistant-advisor, with whom I openly discussed many of the problems I had to deal with during this research. He also read and commented most of this Thesis.

I thank Prof. Amiram Goldblum for his suggestion of using a set of cyclic peptides from one of his papers as the Proof of Concept of the algorithm presented here.

I thank all the members of Benny's research group for all the time we enjoyed together; I especially thank Dr. Elad Segev, a fellow PhD student, for building with me the "kit" of the Ubiquitin+13 test-case, and for reading and commenting the section where it is described and discussed. I also thank Dr. Eric Brown, one of Benny’s Post-Docs, for introducing me to Python, with which I have written most of DEEPSAM’s code.

I thank my close friend Dr. Tova Feldmann, for her comments on an initial draft of this Thesis.

I thank Dr. Ora Schueler-Furman, for her comments about a preliminary version of part of this thesis. I also thank her and Dr. Debbie Shalev, for helping me transform "by-hand" the CIF files of the cyclic peptides used by us, into PDB format. I acknowledge help in the CIF-to-PDB issue, from Prof. H Bernstein and Georgi Darakev, both from the RASMOL group.
I thank Prof. Jay W. Ponder, head of the TINKER project, for his help concerning the use of the package. I thank Alex Herbert, developer of the MaxCluster program, for his help concerning that program.

I thank the System Administrators (Gregory, Michael, Max), at the Fritz Haber Research Center (FH), for their help any time I needed them. I thank Mrs. Geula Levy, the Administrative Manager of FH, for her prompt and kind help in every issue.

I thank the staff of the Jerusalem College of Technology’s Computer Science Dept. who always were ready to make things easier in order to let me proceed with this research: the Chair of the Dept, Dr. Haim Dayan; the Secretary, Mrs. Chanah Twitto; my colleagues – Ezra Dashtt, Dr. Michael Dreyfuss, Dr. Rafi Yehezkael, Dr. Yaakov HaCohen-Kerner, Dr. Ariel Stulman.

Last but not least, my family: I want to thank my wife Etka and my children Tzipi, Shlomo and Naomi, for their infinite patience and tolerance to a husband and father who loves them a lot, but too much time dedicated to the research work describe in this thesis instead of enjoying the quality time that they deserve. I want to thank my oldest daughter Tzipi for her help when I did not know how to do something with Word or PowerPoint – she is a real MS-Office expert! I also want to thank my second mother Perla, and my sister Sara (Buluk), who in spite of the geographic distance always have been close to me.

Last, I would like thank the Almighty for letting me arriving to this milestone of my life:

ב穰ך אנת יִלְוַי מַלְךַת הַעוֹלָמִים שְׁחֵיתִי וּכְיָמִי וְיָגוֹנֵנִי לֻםְךָ יָהּ.
Dedication

It is an honor for me to dedicate this work
to the memory of my beloved Parents

Professor Dr. Nelson Goldstein z"l
Dr. Tziporah-Pola Katzcovicz-Goldstein z"l
Abstract

A key factor in the properties of biological molecules is their structure. This research has focused on the problem of protein structure prediction due to the critical role of peptides and proteins in the normal biological functionality of cells and organisms. At sufficiently low temperatures, though not always, the structure of the minimal free energy corresponds to the global minimum of its force field (FF) – its Potential Energy Surface (PES), which is a function expressing the potential energy interactions inside the molecule. If the FF is the true (exact) one, its global minimum certainly is the correct structure at $T \to 0$ K. The problem of predicting the native (or folded) structure of a polypeptide given its sequence or some unfolded structure is the problem of finding that global minimum. This problem is computationally hard because of its exponential size: we know that a polypeptide's PES has an estimated number of at least $10^n$ minima, where $n$ is the polypeptide's sequence length. Protein structure prediction methods which strictly rely on the assumed correctness of Force Fields (potential energy functions) are ab-initio methods. The approach taken in this research aims at building a purely ab-initio global minimum search method. The method presented here hybridizes several well-established algorithms which have complementary advantages, resulting in a more powerful tool. Three well-known algorithms are combined in my hybrid method, called DEEPSAM (Diffusion Equation Evolutionary Programming Simulated Annealing Method): Evolutionary Programming (EP), the Diffusion Equation Method (DEM), and Simulated Annealing (SA). By smoothing the potential function, DEM contributes to search space reduction. By using simultaneously an ensemble of candidate solutions (a population of conformations), EP contributes to a wide exploration of the search space. By using DEM hybridized with SA and/or a Newton-like local minimization method, four special mutation operators are dynamically chosen to be used. This contributes to long step size and small step size exploitation of the neighborhood of conformations in search space. It is worth noting that because of the population-oriented approach, this algorithm provides us not only with the deepest minimum found, but also with an ensemble of deep lying minima structures that
includes the deepest one. All those minima structures are close to each other energetically and perhaps geometrically as well.

DEEPSAM was tested in three stages:

1. DEEPSAM was run upon a set of seven cyclic and non-cyclic peptides, having between six and nine amino acid residues. For all but one of the examples, the hybrid algorithm finds minima deeper than those obtained by a very extensive "brute force" scan. In all cases, the computed global minimum is in very good accord with the experimental structure. The results show that starting from any unfolded structure; when a reliable potential function is available, DEEPSAM is a powerful structure predictor. If the starting structure is a relatively good local minimum found by some other method (for example, the low-resolution step of Rosetta), we can expect a better performance.

2. DEEPSAM was run upon the protein Ubiquitin, charged +13, in the gas phase. Using the amber98 force field, DEEPSAM was able to find almost the same structure found by a long series of Molecular Dynamics (MD) simulations. DEEPSAM was able to find the same structure in a very short time relative to the time it took those MD simulations to get the same structure. The final structure found by the MD simulation was experimentally confirmed by Dr. Thomas Wyttenbach from UCSB. This indirectly confirms DEEPSAM's result too.

3. DEEPSAM was run upon the protein Crambin, a relatively small thionin protein (46 amino acids long), which contains all the secondary structures present in more complex proteins. The purpose of these runs was to try DEEPSAM as a tool for testing the reliability (and validity) of three well-known forces fields (CHARMM27, AMBER99 and OPLSAA). DEEPSAM was run upon Crambin three times, with one different force field each time. The results suggest that DEEPSAM may be used as a force field tester.

Populations of small size (in most cases, only five conformations) were enough to get these very good results; the parallel design of the algorithm contributed to its time efficiency. These two aspects, together, contributed to relatively modest requirements of computer resources.

Following the success of the algorithm, a series of challenging future work is underway.
In conclusion, a very promising algorithm which uses relatively small computer resources, is now available for predicting the structure of bio-molecules.
# Table of Contents

1 – Introduction ............................................................... 1

2 – Research Goals ........................................................... 8

3 – Method ...................................................................... 10

   I. Reducing search space by PES smoothing ......................... 11

   II. PES exploitation by small-step and long-step exploration .... 14

   III. PES sampling and exploration ....................................... 17

   IV. DEEPSAM selection operators and self-adaptability ........... 25

   V. Stopping Criterion ....................................................... 26

4 – Results ....................................................................... 28

   I. Proof of Concept – Cyclic and Non-Cyclic Peptides .......... 28

      I.a. Background .............................................................. 28

      I.b. Analysis and Results ................................................ 29

      I.c. Conclusions ............................................................. 57

   II. Predicting Structure in the Gas Phase – Ubiquitin+13 ......... 59

      II.a. Background .............................................................. 59

      II.b. Analysis and Results ................................................ 61

      II.c. Conclusions ............................................................. 63

   III. About Force Fields Reliability – Crambin ..................... 64

      III.a. Background .............................................................. 64

      III.b. Analysis and Results ................................................ 65

      III.c. Conclusions ............................................................. 69

5 – Concluding Remarks ................................................... 70

6 – References .................................................................... 72
<table>
<thead>
<tr>
<th>Abbr.</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMBER</td>
<td>Assisted Model Building with Energy Refinement force fields</td>
</tr>
<tr>
<td>CASP</td>
<td>Critical Assessment of protein Structure Prediction</td>
</tr>
<tr>
<td>CHARMM</td>
<td>Chemistry at HARvard using Molecular Mechanics</td>
</tr>
<tr>
<td>CPU</td>
<td>Central Processing Unit</td>
</tr>
<tr>
<td>DEEPSAM</td>
<td>Diffusion Equation Evolutionary Programming Simulated Annealing Method</td>
</tr>
<tr>
<td>DEM</td>
<td>Diffusion Equation Method</td>
</tr>
<tr>
<td>DEMSA</td>
<td>Diffusion Equation Method Simulated Annealing mutation operator</td>
</tr>
<tr>
<td>EA</td>
<td>Evolutionary Algorithm</td>
</tr>
<tr>
<td>EP</td>
<td>Evolutionary Programming</td>
</tr>
<tr>
<td>ESI</td>
<td>Electro-Spray-Ionization</td>
</tr>
<tr>
<td>FF</td>
<td>Force Field</td>
</tr>
<tr>
<td>GA</td>
<td>Genetic Algorithm</td>
</tr>
<tr>
<td>gRMSD</td>
<td>RMSD of two conformations according to the MAXSUB found</td>
</tr>
<tr>
<td>GSA</td>
<td>Genetic Simulated Annealing</td>
</tr>
<tr>
<td>HP</td>
<td>Hydrophobic-Polar protein folding model</td>
</tr>
<tr>
<td>IMS</td>
<td>Ion-Mobility Spectrometry</td>
</tr>
<tr>
<td>IPG</td>
<td>Initial Population Generation algorithm</td>
</tr>
<tr>
<td>L-BFGS</td>
<td>Limited-memory Broyden-Fletcher-Goldfarb-Shanno local minimization method</td>
</tr>
<tr>
<td>LEP</td>
<td>Levy Evolutionary Programming</td>
</tr>
<tr>
<td>LS</td>
<td>Local Search (local minimization) algorithm</td>
</tr>
<tr>
<td>MAXSUB</td>
<td>Maximal Superposable Sub-structure</td>
</tr>
<tr>
<td>MC</td>
<td>Monte Carlo optimization method</td>
</tr>
<tr>
<td>MD</td>
<td>Molecular Dynamics</td>
</tr>
<tr>
<td>MDSA</td>
<td>Molecular Dynamics Simulated Annealing</td>
</tr>
<tr>
<td>MMFF</td>
<td>Molecular Mechanics Force Field</td>
</tr>
<tr>
<td>MS</td>
<td>Mass Spectrometry</td>
</tr>
<tr>
<td>NMA</td>
<td>Normal Modes Analysis</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear Magnetic Resonance</td>
</tr>
<tr>
<td>OPLSAA</td>
<td>Optimized Potential for Liquid Simulations force field</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>PES</td>
<td>Potential Energy Surface</td>
</tr>
<tr>
<td>RMSD</td>
<td>Root Mean Square Deviation</td>
</tr>
<tr>
<td>RVN</td>
<td>Randomized Vector Normal mode technique</td>
</tr>
<tr>
<td>SA</td>
<td>Simulated Annealing optimization method</td>
</tr>
<tr>
<td>SSI</td>
<td>Simple Sequential Inhibition process algorithm</td>
</tr>
</tbody>
</table>
1 – Introduction

It is well known that structure determines the physical, biological and chemical properties of bio-
molecules\(^1\). While the structures of bio-molecules are of great importance, theoretical *structure prediction* is a daunting task. Despite extensive research activities, inspiring ideas, and major progress, this problem remains open to a large extent. Because of the critical role of peptides and proteins in the normal biological functionality of cells and organisms, we have considered the problem of *protein structure prediction*\(^2\) as the focus of the research presented in this Thesis.

Proteins can be described by a hierarchy of three structural levels. At the *first structural level*, a protein is a polymer consisting of a *sequence* of naturally occurring amino-acids; this amino-acid sequence, which is determined by the gene that encodes it,\(^1\) is what is called its *primary* structure. Actually, such linear structure may be viewed as a word in a 20-letter alphabet language – each letter being the 1-letter name of an amino-acid.- Upon this structural level, sequence alignment algorithms were developed and used for protein sequence alignment/similarity studies.\(^6,7\) At the *second structural level*, residues in the protein sequence are assembled in well-defined repetitive 3-dimensional structural patterns: \(\alpha\)-*helices* - formed by the hydrogen bonding of the carbonyl oxygen of a residue with the amide nitrogen of another residue four positions further in the sequence, \(\beta\)-*sheets* - formed by the hydrogen bonding of backbone groups of distant residues in the sequence, \(\beta\)-*turns* - formed by the hydrogen bonding of the carbonyl oxygen of a residue with the amide nitrogen of another residue three positions further in the sequence, provoking direction reversal. All those together constitute what is called the protein’s *secondary* structure. At the *third structural level*, protein structure constitutes what is called the *tertiary* structure - a sequence of secondary structure components, that part of them may be connected by unstructured sub-sequences called *coils* or *loops*, which gets packed into a compact 3-dimensional conformation. One of an enormous multitude of such conformations, the *folded* (or *native*)
structure, is that which most typically governs the biological function of any protein molecule. In fact, the term tertiary structure is commonly used to refer to the folded structure – in what follows, those two terms will be used interchangeably.

The transformation process from primary to tertiary structure is called protein folding, where secondary structure patterns are intermediate states between them. Several diseases (Alzheimer and others) are thought to be provoked by proteins’ malfunction caused by misfolding. A better understanding of the protein folding process, and our capability of predicting the native structure given the sequence (or any unfolded structure) of a peptide or protein, will allow us to understand also misfolding and the diseases it provokes, providing the necessary knowledge to more appropriate drug design.

Two complementary thermodynamics-based approaches to protein folding can be identified: Anfinsen’s approach, which emphasizes thermodynamic stability, and (b) Levinthal’s approach which emphasizes kinetics. Anfinsen experimentally demonstrated that a protein, unfolded using chemical denaturants, or by thermal denaturation, spontaneously refolds to its native conformation. This experiment resulted in the principle that the protein’s tertiary structure is already encoded in its primary structure. According to Anfinsen’s thermodynamic hypothesis, a protein’s native structure is the unique conformation in thermodynamic equilibrium at the lowest possible free energy - the folded state of a protein is the conformation with the most thermodynamic stability. This means that the protein native structure is the conformation for which the free energy is in its lowest value - its global minimum. Flexible macromolecules, like proteins and peptides, have many “meta-stable” local minimum conformations separated by low-energy barriers. At sufficiently low temperatures, though not always, the structure of the minimal free energy corresponds to the global minimum of its force field (FF) – its Potential Energy Surface (PES), which is a function expressing the potential energy interactions inside the molecule. If the FF is the true (exact) one, its global minimum certainly is the correct structure at $T \to 0 \text{ K}$. In any case, at finite $T$, both the global minimum and the (local) minimum structures are generally populated. The FFs actually available to us are Molecular Mechanics FFs (MMFFs)
whose PES is the Born-Oppenheimer ground-state energy\textsuperscript{20}. The Born-Oppenheimer approximation allows representing the potential energy of the molecule as a classical mechanics expression, meaning that an MMFF is a potential energy function which is dependent on the atoms’ nuclei positions.

\[ E_{FF} = E_b + E_{nb} \]

\[ E_b = \sum_{bonds} k_b (b - b_0)^2 + \sum_{angles} k_\theta (\theta - \theta_0)^2 + \sum_{torsions} k_\phi [\cos(n\phi + \delta) + 1] \]

\[ E_{nb} = E_{vdw} + E_{el} \]

\[ E_{vdw} = \sum_{nonbond pairs} \left[ \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}} \right] \]

\[ E_{el} = \sum_{nonbond pairs} \left[ \frac{q_i q_j}{r_{ij}} \right] \]

(1)

where \( E_b \) is the bonding energy term and \( E_{nb} \) is the non-bonding energy term.

\textbf{Figure 1.} Force Field interactions

See that each term in eq 1\textsuperscript{20,117} depends on constant values got from experimental data or from Quantum Mechanics calculations. Those values, \((k_b, b_0, k_\theta, \theta_0, \ldots)\), which are related to each atom (or interaction between pairs of atoms) in the molecule, are the \textit{parameters} of the MMFF. The MMFF's \textit{parameterization} determines how good it is as an approximation of the \textit{correct} FF. If the FF is correct, the structure is correct – since one does not know how good FFs are, this is the important and also severe limitation of the FF approach to molecular structure. In spite of this

\* picture taken from Ponder, J.W.; Case, D.A\textsuperscript{20}
important limitation and inaccuracy, we argue that FFs are probably the best modeling tools available to us – thus, from this point we will assume that the FFs at hand are correct (see section 4.III where the reliability of MMFFs is reconsidered).

The PES function depends on the 3D positions of the atomic nuclei in all-atoms force fields (or Cα nuclei in united-atoms force fields). Given a protein composed by \( n \) (meaning hundreds or thousands of) atoms, its PES is a function \( f \) whose domain is the \( n \)-dimensional space of 3D atomic coordinates, and its range is the set of corresponding \( f \)'s values. Note that every point, in this Cartesian \( n \)-dimensional space, represents a conformation of the molecular system in 3D space - actually, the potential has \( 3n-6 \) relevant degrees of freedom and not \( 3n \) (since there are 6 translational-rotational degrees of freedom for the entire molecule, seen as a whole).

The number of thermodynamically meta-stable states of proteins is estimated to be exponential on the number \( n \) of amino-acids in the protein sequence. Assuming that the average possible “preferred” meta-stable states of each one of the \( n \) amino-acids of the protein sequence is \( q \), the estimated number of possible meta-stable states will be at least \( q^n \). Several estimates of that average can be found in the literature\(^5,11,12,15,16,17\), from \( q=2 \) through \( q=10 \). Levinthal argued that it is not possible to reach the native structure by random sampling of the conformational space; considering that proteins fold quickly and spontaneously, it would take too much time to sample the estimated huge number of thermodynamically meta-stable states. This apparent contradiction, known as the Levinthal’s paradox, suggests that protein folding happens independently of the number of meta-stable states - instead of sorting out all the possible conformations, Levinthal argued that proteins naturally fold through the “fastest-folding pathways” towards the kinetically accessible most stable conformation. The Levinthal’s paradox became the base of protein folding methods, which try to identify the alternative pathways from an unfolded structure of a protein towards its native structure.

Following Anfinsen’s thermodynamic hypothesis, the *protein structure prediction* problem is a search problem\(^22\) whose search space is the set of conformations for which the values of \( f \) are potential energy local minima (most of them unfolded) - because of the existence of a multitude of
local minima, this problem is also called the “multiple-minima” problem. The best (or optimum) from all 3D conformations in the whole search space is the global minimum of $f$. Thus, the protein structure prediction problem is that of finding the unique PES’s global minimum, which corresponds to the thermodynamically most stable conformation, which is the protein's native structure. Provided a reliable PES is available, the Anfinsen’s hypothesis became the base of \textit{ab-initio} protein structure identification methods, which aim at getting the structure from the FF (assuming it is correct). We note at the outset that the approach taken by us is a purely \textit{ab-initio} one: a FF is assumed to be available, and to be correct, thus the structure (at $T \rightarrow 0$ K) is found to be the global minimum. This is, indeed, a limitation of our approach, as it is of other methods that rely strictly on the PES. A range of approaches and methods have been developed for the protein structure identification problem. Actually, those are search/optimization algorithms: (a) Monte Carlo(MC),\textsuperscript{4,24} Simulated Annealing(SA),\textsuperscript{22,24,25} Tabu Search\textsuperscript{26}, Evolutionary Algorithms(EAs),\textsuperscript{22,26-32} and Potential Smoothing Methods\textsuperscript{33} which are global search algorithms, and (b) Grid Search, Steepest Descent, Conjugate Gradients, Newton-Raphson, etc., which are local search algorithms.\textsuperscript{34} The scientific literature about this kind of algorithms is immense – only those relevant to our discussion will be referred later.

Another approach to the protein structure identification problem is based on the assumption that proteins are not structured at random - structures shared by families of proteins have been identified from studies made on the structural information already available in protein databases. Those structures, called \textit{folds}, are assumed to “reflect the physical-chemical properties of strings of amino acids in sequences as well as limitations on the types of interactions allowed within a folded polypeptide chain”,\textsuperscript{1} and are used to structurally classify proteins: “Proteins of the same fold have the same major secondary structures in the same arrangement with the same topological connections, with some small variations typically in the loop region. In some sense, folds are distinct templates of protein structures”.\textsuperscript{13} It is argued that given a known protein, if the sequence (or sub-sequences) of another protein exactly aligns to it in more than 40% of its amino-acids, without large insertions or deletions, both proteins can usually have the same structure.\textsuperscript{1} Making
use of databases containing structural information of known proteins or folds, two methods have been developed to deal with the protein structure identification problem: “Comparative Modeling” (also called “Homology Modeling”)\(^3\text{-}\text{2,}\text{36}\) and “Threading”.\(^1\text{-}\text{3}\) In Comparative Modeling, the \((\text{target})\) unknown protein is aligned with \((\text{template})\) known proteins, from which the best aligned one is selected; based on this alignment, the target backbone structure is generated; then, the side chains are accordingly placed. In Threading, folds are used as template structures without sequence-aligning them with the target. Actually, the target is forced to structurally align to the structure of every one of the known folds; the most suitable fold is selected as the target structure. Y. Zhang and J. Skolnick\(^36\) argue that “because an accurate theory for the prediction of protein structure on the basis of physical principles does not yet exist, comparative modeling/threading approaches are the only reliable strategy for high-resolution tertiary structure prediction”. In fact, those approaches alone are not the panacea either: “all methods of model building based on a preexisting structure, whether found by sequence homology or by threading, suffer from massive feedback and bias”;\(^1\) also, if a threading method “fails to find any fold with a significantly high score, nothing has been learned about the sequence”. A successful method which is neither an homology method nor a threading one, but borrows from both and from the FF approach, is Rosetta.\(^1\text{-}\text{35-}\text{42}\) It is especially useful when it is not possible to find structures homologous to the target, or even when the structures found in protein databases are not more than 30% similar to the target. A Rosetta run is a two-step process – a low-resolution step followed by a high-resolution refinement step. In the low-resolution step, instead of using an all-atom representation of the target protein, a reduced representation is used: an all-atom representation is used for amino-acid backbones, side-chains are represented by a "centroid" which is the average mass center of the amino-acids, and torsion angles are the only allowed changing. This representation reduces the size of the problem as a function of the number of represented atoms in the molecular system. In the high-resolution step, an all-atom representation of the target protein is used. Starting from an initially unstructured chain, the low-resolution conformational space to be searched is that in which each new conformation is built by assembling
together fragments (short sub-sequences of length 9 and 3 residues) taken from corresponding sub-sequences of randomly selected known protein structures – this ensures that the local energy level of each fragment is close to optimal. Rosetta’s fragments assembly approach reminds H. A. Scheraga’s “build-up” procedure, which incrementally builds hopefully optimal protein structures, starting by assembling together optimal amino-acid conformations. At both, the low-resolution step and the high-resolution step, specially designed potential functions are used in thousands of Monte Carlo Simulated Annealing trajectories which are computed in parallel - the underlying assumption is that the native structure would be found if enough sampling of the conformation space is performed.
2 – Research Goals

We wanted to have a good tool for structure prediction – a tool with which we will be able (a) to find a good approximation of the native structure of a protein or peptide, given a force field and a starting (unfolded) structure, and (b) to find an ensemble of structures close enough structurally and energetically to the found approximation of the native structure.

According to the "No-Free-Lunch" theorems, there is no universal problem solver; in complex problems like the one we are dealing with, it is desirable to hybridize different approaches: hybridization is “a key factor for achieving superior performance”. According to D. E. Goldberg and S. Voessner, the motivation for algorithm hybridization is to achieve increased efficiency: an “adequate solution quality in minimum time” or “maximum quality in specified time”. Our knowledge of the problem, and of the different algorithms at hand, should allow us to usefully combine together their advantages in order to obtain better solutions than those produced using each one separately - "a hybrid combines the global searcher (the GA) with other methods which exploit problem specific knowledge to generate better solutions than either could have come up with on its own." Theories and taxonomies of hybridization can be found in the literature. In the hybridization of EAs with local searchers, the population-oriented characteristic of the EA part of the hybrid contributes its global search space exploration capabilities, while the individual-oriented characteristic of the local search part contributes its local neighborhood exploitation capabilities – ideally, hybridization would provide us with both, good search space sampling together with good local improvement of the sampled points (a discussion of exploration and exploitation in EAs can be found elsewhere).

Thus, the goal of this research was to develop a new hybrid PES-based global optimization algorithm, called DEEPSAM (Diffusion Equation Evolutionary Programming Simulated Annealing Method), which combines the complementary advantages of three well-established
methods: Evolutionary Programming (EP), Diffusion Equation Method (DEM) and Simulated Annealing (SA). The idea underlying such a combination is that with each component-algorithm overcoming others' typical difficulties in the search for the minimum will result in a more powerful tool.

The rest of this Thesis is as follows: in the Method chapter, we describe DEEPSAM and the role of each one of its hybridized component-algorithms; in the Results chapter, we describe three application systems upon which DEEPSAM was tried, including the analysis of the computed results and their comparison with experimental data. In the Concluding Remarks chapter, we express our conclusions and a preview of further developments.
3 – Method

DEEPSAM is an Evolutionary Programming (EP) algorithm hybridized with Diffusion Equation Method (DEM) and Simulated Annealing (SA). DEEPSAM’s flowchart can be seen in Figure 2.

In DEEPSAM’s design we had in mind three main goals: (a) search space, which scales exponential with the number of residues, should be reduced; (b) knowledge acquired about the topography of the PES should be exploited, locally and globally; (c) widely distributed different sub-areas of the PES should be simultaneously explored.

Figure 2. DEEPSAM flowchart
I. Reducing search space by PES smoothing

Two different approaches can be identified in order to reduce the search space: (a) at the structure representation level - the number of atoms in the molecular system’s representation is reduced by using united-atom force fields, the lattice model, the HP model, etc.\textsuperscript{53} or the more recent coarse-grained models;\textsuperscript{54} (b) at the potential energy function level - the number of minima to sample is reduced by using function smoothing.\textsuperscript{33,43,55,56} The latter is the approach adopted by us as DEEPSAM’s mechanism for search space reduction. A smoothing method is a deterministic global optimization algorithm which, given a PES function $f(x)$ ($x \in \nabla^0$), and a symbolic function deformation operator $T(t)$ where $t$ is the smoothing level (or rate of function deformation) and $\Delta$ is the Laplacian, it is

$$T(t) = \exp(t\Delta)$$  \hspace{1cm} (2)

composed by two essential steps:

(a) Function smoothing: $T(t)$ is applied upon $f(x)$ such that a smoothed PES function $F(x,t)$ is

$$F(x,t) = T(t)f(x)$$  \hspace{1cm} (3)

while the initial un-smoothed PES function $F(x,0)$ is

$$F(x,0) = f(x)$$  \hspace{1cm} (4)

The operator $T(t)$ successively applied, such that

$$F_i(x,t_i) = T(t_i)F_{i-1}(x,t_{i-1})$$  \hspace{1cm} (5)

where every specific smoothing level $t_i$ is determined by a smoothing protocol

$$t_i = t_{\text{max}} \left( \frac{i}{M} \right)^2 \hspace{1cm} (i = 0,1,\ldots,M)$$  \hspace{1cm} (6)

where $t_{\text{max}}$ is some arbitrarily chosen high smoothing level, and $M$ is the number of generated smoothed PES functions $F_i$ ($i$ in $[1,\ldots,M]$). As can be seen in Figure 3, we may expect that at any smoothed PES $F_i$, there will be less minima to sample than at $F_{i-1}$, meaning that the size of the optimization problem at $F_i$ will be much smaller, significantly reducing the search effort.

(b) A reversal procedure is executed, starting at $F_M$ which is the smoothest PES (determined by
$t_{\text{max}}$, the highest smoothing level). As can be seen in Figure 3, a local minimization algorithm \(LS\) is applied upon some molecular structure at \(F_M\), and for each \(F_i\) (\(i\) in \([M .. 1]\)), the calculated local minimum is projected into \(F_{i-1}\) and \(LS\) is applied on it at \(F_{i-1}\).

![Figure 3. Function smoothing](image)

As can be seen in Figure 3, the smoothest possible PES function has only one minimum; that smoothed PES can be generated at an extremely high \(t_{\text{max}}\):

\[
t_{\text{max}} > \frac{1}{6} d^2 m(m+1)(2m+1)
\]

where \(d\) is the longest bond length, and \((m + 2)\) is the number of bonds. Note that independent of which initial conformation is chosen, starting the reversal procedure at the smoothest possible PES and after computing its unique minimum, the final minimum calculated at the original un-smoothed PES will be always the same one. But, as can be seen in Figure 3, the calculated final minimum will not necessarily be the global minimum. Even if Nakamura et al.’s "Two-Stage Method"\textsuperscript{57} algorithm tries to deal with this pitfall, using the highest possible \(t_{\text{max}}\) is a disadvantage because the physical meaning of smoothed PESs is gradually lost, relative to that of the un-smoothed PES. Because of that, the \(t_{\text{max}}\) actually used should be chosen with cautiousness.

\textsuperscript{†} from Pappu, R. V., Hart, R. K. and Ponder, J.W.\textsuperscript{58} (reprinted - “Adapted” or “in part” - with permission © 1998 American Chemical Society)
Trying to determine the conditions in which the final (local) minimum is the global minimum, J. W. Ponder and coworkers\textsuperscript{58-60} identified three possible events in every smoothing transformation step (see Figure 4):

1. *merging* - minima close to each other at \( F_i \) merge into one minimum, reducing the overall number of minima at \( F_{i+1} \); energy barriers cluster into one *basin* and the *neighborhood* affected by the *LS* become wider than in a less smoothed PES.

2. *shifting* - given a minimum A at \( F_i \), its corresponding minimum at \( F_{i+1} \) is a slightly structural *shift* relative to A.

3. *crossing* - given two unrelated minima B and C at \( F_i \), where \( F_i(B, t_i) < F_i(C, t_i) \), their corresponding minima \( B' \) and \( C' \) at \( F_{i+1} \) *cross* each other, being \( F_{i+1}(B', t_{i+1}) > F_{i+1}(C', t_{i+1}) \).

![Figure 4. Function smoothing: merging, shifting, crossing\textsuperscript{‡}](image)

J. W. Ponder and coworkers found that if the smoothing process is a sequence of *merges*, the global minimum is kept, and the local minimum computed at \( F_M \) is related to the global minimum, which is the local minimum computed when back in the un-smoothed PES. This finding suggests that in such cases, a smoothing method converges to the global minimum. DEEPSAM, which has been built upon TINKER,\textsuperscript{61} a molecular modeling package developed by J. W. Ponder's group,
uses TINKER pss program which implements the smoothing method, as a variation of Nakamura et al.'s "Two-Stage Method",\textsuperscript{57} which itself is a variation of H. A. Scheraga and coworkers' Diffusion Equation Method (DEM)'s.\textsuperscript{25,33,43,55,56}

II. PES exploitation by small-step and long-step exploration

In order to 	extit{exploit} the knowledge acquired about the topography of the PES, DEEPSAM uses a combination of function smoothing, a local minimization method, which allows small step size exploration of a PES, and Simulated Annealing (SA), which is a global minimization method which allows long step size exploration of a PES. In SA, given a conformation $C_i$ in search space, which corresponds to the current step $i$ of the search, a new candidate conformation $C_{i+1}$ is generated in a neighborhood of $C_i$. In order to prevent to get stuck in a local minimum, and to allow to “jump” over high energy barriers, the acceptance of a candidate conformation $C_{i+1}$ is controlled by the Boltzmann probability-based Metropolis criterion: the probability of acceptance $P(C_{i+1})$ is as specified in eq 8.

$$P(C_{i+1}) = \begin{cases} 1, & \text{if } F(C_{i+1}) \leq F(C_i) \\ e^{\frac{F(C_i) - F(C_{i+1})}{T}}, & \text{otherwise} \end{cases}$$

That is: if $F(C_{i+1}) \leq F(C_i)$, the new conformation is accepted; otherwise, it is accepted according to the Boltzmann-like probability.

At the beginning of an SA run, when the “temperature” parameter value of $T$ is high, the acceptance of relatively "bad" candidate conformations is almost purely random. During the run, a series of decreasing $T$ values is generated by a cooling protocol, as specified in eq 9.

$$T_{i+1} = \xi T_i \cdot 0 < \xi < 1$$

The rate of cooling, $\xi$, will determine how quickly (or slowly) the value of $T$ will decrease during an SA run. Like in metal annealing, a relatively slow cooling will give better results than a relatively quick one. The actual effect of a cooling protocol is to focus the search, reducing the
probability of acceptance of relatively bad conformations, until SA stops at the "frozen" state.

J. W Ponder and coworkers\textsuperscript{59} found a correspondence between the smoothing protocol in smoothing methods and the cooling protocol in SA: a high temperature $T$ in SA’s cooling protocol determines a wide covering of the search landscape, from which a Boltzmann-like stochastic choice is done; a high smoothing level $t$ determines a wide covering of the search landscape from which a deterministic choice is done. In both cases, a slow protocol will provoke a more detailed scan of the PES.

Taking into account that, in general, a smoothing method does not get to the global minimum because of shifting and crossing, and SA may get it only using a very slow cooling protocol, we realized that instead of applying a full DEM’s reversal procedure, we may get an approximation of the global minimum computing only two local minima (see Figure 5) instead of $2n+1$ minima: (a) a minimum $M_i$ will be computed at a given smoothed PES $F_i$ taking some given structure as the initial structure for some given minimization procedure $LS_i$; (b) the other minimum $M_0$ will be computed projecting $M_i$ on the un-smoothed PES $F_0$, and taking it as the initial structure for some other minimization procedure $LS_0$ ($LS_0$ and $LS_i$ are not necessarily different algorithms). $LS_0$ will do a relatively long "jump" over energy barriers of the original un-smoothed PES $F_0$ while $LS_i$ will do a relatively short "jump" within the neighborhood of the conformation computed at $F_i$ (we already said that search neighborhoods are wider in smoother PESs). This two-step procedure was designed as a modification of TINKER’s pss program, in combination with TINKER’s implementation of Molecular Dynamics Simulated Annealing (MDSA) – thus, the name given to it is DEMSA (Diffusion Equation Method with Simulated Annealing). We implemented four alternative DEMSA procedures (in what follows, the acronym \textit{nw} refers to the L-BFGS quasi-Newton procedure, and the acronym \textit{sa} refers to MDSA): \textit{nwnwpss - nw} is used both, as $LS_0$ and $LS_i$, allowing to do a controlled "jump" in search space; \textit{sanwpss - nw} is used as $LS_i$, and \textit{sa} is used as $LS_0$, starting \textit{sa} at $F_0$ from a hopefully good position determined by \textit{nw} at $F_i$; \textit{nwsapss - sa} is used as $LS_i$, and \textit{nw} is used as $LS_0$, allowing to locally refine at $F_0$ the structure found by \textit{sa} at $F_i$; \textit{sasapss - sa} is used as both, $LS_0$ and $LS_i$, starting \textit{sa} at $F_0$ from a structure.
whose place in the search space would be relatively far from the starting structure upon which sa was run at $F_i$. Figure 5 shows the application of two DEMSA procedures over the same conformation (one drawn in red and the other in green).

![Figure 5. DEMSA](image)

DEEPSAM will choose which one of those DEMSA procedures to use according to the length of the “jump” (the step size) it decides is needed at any given moment of the computation process.
III. PES sampling and exploration

In order to explore widely distributed different sub-areas of the PES, DEEPSAM uses the Evolutionary Programming (EP)\(^{27-29}\) type of EAs. Like other EAs, EP's search is the process of evolutionary change of an ensemble of \(n > 1\) candidate solutions – a population. In most cases found in the EA literature, the initial ensemble of candidate solutions (the initial population) is randomly generated in a problem-independent manner. Relatively little research has been published in this aspect of EAs.\(^{62-67}\) In our case, the initial population cannot be randomly generated – it must be an ensemble of physically feasible conformations, all of them local minima of the PES. To be able to start the search from a relatively well distributed and physically feasible initial population, a special algorithm has been developed; it is described in section III.a. What characterizes EP, and distinguishes it from other kinds of EAs, is that the only variation operator used is mutation. In order to be able to cover the PES with a well distributed ensemble of conformations, a PES sampling mechanism has been developed by us. This sampling mechanism uses EP's population-oriented approach together with a combination of DEMSA-based mutation operators inspired by Ch. Y. Lee and X. Yao's\(^ {68}\) Levy-based mutation operators; they are described in sections III.b and III.c.

III.a. Initial Population Generation

Given a population \(P\) of \(n\) individuals \(\{a_1, \ldots, a_n\}\), generated by some search space exploration procedure, population diversity is a term which refers to how widely distributed are those \(n\) individuals over the search space. We talk about good population diversity when the distance \(D(a_i, a_j)\) between any pair \((a_i, a_j)\) of individuals in the population \(P\) is greater than some minimum distance \(\Delta > 0\). Good population diversity will promote the exploitation of un-clustered local sub-areas of the search space, by local search procedure(s). In our problem context, the population is an ensemble of \(n > 1\) molecular conformations and the distance measure is the RMSD among them, meaning how much structurally different is each conformation relative to any other in the ensemble.
A well distributed ensemble of conformations is needed during the first iterations of an EA run. For that we developed an Initial Population Generation algorithm (IPG) whose goal was to generate an initial population of physically feasible conformations, with good diversity. IPG, which is similar, in its approach, to the "Simple Sequential Inhibition Process" (SSI) algorithm recently proposed by H. Maaranen, K. Miettinen and A. Penttinen, creates an ensemble of local minima conformations whose RMSD between them and a given initial (unfolded) molecular conformation and among themselves are at least 2.5 Å in the case of peptides of length less than 10 amino acids, or at least 4 Å in the case of peptides of length at least 10 amino acids. In order to ensure physical feasibility, IPG uses Molecular Dynamics (MD) and Normal Modes Analysis (NMA) as its conformation generation means.

MD simulations are strictly classical. The positions for all atoms are updated by simultaneous integration of the Newtonian equations of motion, using the Verlet integrator. The forces applied on the moving atoms are calculated from the gradient of the pairwise additive potential, using all atoms. Given an initial conformation, the entire system is thermalized at the desired temperature ($T=800K$). This is accomplished initializing the dynamics from that initial conformation by imparting random velocities selected from a Maxwell Boltzmann distribution corresponding to a mean kinetic energy $kT$. The total energy of the system is followed periodically to check conservation. The thermal equilibrium of the system is propagated in time for a period of 10-20 ps in its ground state potential. The thermally equilibrated system is then used as the initial state of the conformational space sampling promoted by the MD simulation. In IPG, candidate conformations for the initial population of a DEEPSAM run are chosen equally spaced in simulated time – a conformation snapshot every 1 ps. DEEPSAM uses TINKER's dynamic program for its MD simulations.

NMA has been known for years to be applicable for the description of the movements of biological molecules. The basic idea behind NMA is to model a molecule as a collection of harmonic oscillators (atoms) coupled together by springs (inter-atomic model). A molecule made up of $N$ atoms has $3N-6$ normal modes. The low frequency (large displacement) modes of a
protein are thought to be the most biologically significant and in fact may play a crucial role in
binding pocket activity.

NMA has several limitations: (a) it requires a huge amount of computer memory for even medium
sized proteins; (b) the motion of a protein is assumed purely harmonic, in which case its motion
can be exactly expressed as a superposition of normal modes. However, protein motions are
known to be fluid-like and an-harmonic, so the energy potential near equilibrium must be
approximated by a harmonic function. Still, NMA has provided a great deal of insight into the
nature of collective motions in proteins.

The starting point for NMA is one particular stable conformation of the system, representing a
local minimum of the PES. One then constructs a harmonic approximation of the potential well
around this conformation. A harmonic potential well has the form described by eq 10.

$$U(r) = \frac{1}{2} (r - R) \cdot K(R) \cdot (r - R)$$  \hspace{1cm} (10)

where $R$ is an $3N$-dimensional vector and $N$ is the number of atoms describing the stable
conformation at the center of the well and $r$ is an equally $3N$-dimensional vector representing the
current conformation. The symmetric and positive semi definite matrix $K$ (the Hessian matrix)
describes the shape of the potential well. Thus, a harmonic model for a potential well consists of $R$
and $K$.

NMA is a suitable method for studying large amplitude deformational motions of bio-molecules.
In many cases it has been found that functionally important transition pathways of bio-molecules
often follow the trajectories of one of the low frequency normal modes. In IPG we have
implemented and applied a new randomized vector normal mode (RVN) technique to generate
conformational changes described by normal modes. Given a conformation $A$, we generate a new
conformation $B_{Map}$ which is a perturbation from $A$; a first order approximation of $B_{Map}$ can be
obtained using eq 11.

$$B_{Map} = A + \sum_{i=1}^{M} C_i v_i$$  \hspace{1cm} (11)
Here $A$ is the $3N$-dimensional vector giving us the coordinates of the conformation $A$ and $M$ is a small set of dominant normal modes (those chosen randomly from a Gaussian distribution over the low frequency normal modes (between 500 and 1300 cm$^{-1}$)) which are used to generate $B_{Map}$. $v_j$ corresponds to the normalized eigenvector of the Hessian matrix obtained from conformation $A$. $C_i$ are the random normal mode coordinates obtained from a Gaussian distribution. The mapped structure, $B_{Map}$ will be accepted by IPG if the RMSD between it and the source structure $A$ is greater that some given $\Delta > 0$ (see below a detailed description of IPG).

This NMA-based method has been implemented as part of our population program, whose code is based on TINKER’s vibrate program which is written in Fortran77.

IPG was designed to solve the problem of finding an initial population of $\text{popsize}$ conformations with RMSD relative to the initial conformation, and among themselves, greater than a minimum structural distance $\Delta > 0$. The C-like pseudo-code of IPG is as follows:

```
1. Initialize parameters;
2. looptimes = MDlooptimes;
    genproc = k ps MD simulation at a 800 K temperature, taking a conformation snapshot every ps;
    selectconfs = [ ];
3. for (i = looptimes; i > 0; i--) {
    genconfs = generate k confs running genproc;
    selectconfs = selectconfs + genconfs;
    selectconfs = get ride of confs whose RMSD to others in selectconfs, and the initial one, is less than $\Delta$;
    if (nrofselectedconfs >= popsize)
        return selectconfs; /* we are done */
}
4. if (genproc == NMAprocedure) {
    missingconfs = randomly select missing confs from the set of non-selected confs;
    selectconfs = selectconfs + missingconfs;
    return selectconfs; /* we are done */
}
else {
    looptimes = NMAlooptimes;
    genproc = NMAprocedure;
    goto 3;
}
```
As can be seen in IPG's pseudo-code, it works as follows: (1) starting from a given initial conformation, an MD-based iterative process is launched – it generates a set of conformations taken from the sequence generated by the dynamics simulation; (2) all the RMSD unacceptable conformations are deleted from the set; (3) if the size of the remaining set is smaller than the required *popsize*, another iterative process is launched in which our RVN method randomly generates conformations that are added to the set; (4) all the RMSD unacceptable conformations are deleted from the set; (5) if the size of the remaining set continues to be smaller than the required *popsize*, our IPG will complete the *popsize* population by randomly choosing conformations from those that were deleted in (2) and (4). In order to delete the unacceptable conformations from the set, a directed *complete* graph, represented by a directed tree, represents the set of conformations. The root of the tree represents the given initial conformation, the nodes \(a_1, \ldots, a_n\) represent the conformations in the set, and the weights assigned to the directed edges \((a_i, a_j)\) are the RMSD between \(a_i\) and \(a_j\).

The tree is scanned in preorder. A node \(a_n\), its in-going edge and all its out-going edges are deleted if the RMSD between the node and its father-node is unacceptable. Because of efficiency reasons, each node and all its out-going edges are built on-demand, only after the previous sibling-node in the tree has been visited and inspected.
Thus, most, and hopefully all, the conformations of the set result of running our IPG, will be a population of $n$ conformations with RMSD $\geq \Delta$, among themselves and also relative to the initial conformation. That is, we can expect to start the DEEPSAM’s loop with a good initial population diversity.

III.b. Stochastic value transformation: LEP mutation operators

The Levy probability distribution is a range of stable probability distribution functions. Its full form has two parameters: its scaling factor $\gamma > 0$, and its shape $\alpha$ ($0.0 < \alpha < 2.0$). Assigning 1 to $\gamma$, its form is as can be seen in eq 12.

$$L_{\alpha}(y) = \frac{1}{\pi} \int_0^\infty e^{-q^{\alpha}} \cos(qy) dq \quad y \in R \quad (12)$$

The Levy distribution’s instance corresponding to $\alpha = 1.0$ is equivalent to the Cauchy distribution, and when $\alpha \rightarrow 2.0$ it corresponds to the Gaussian distribution; for other $\alpha$ values, Levy distribution’s instances are between Cauchy and the Gaussian. In real-valued EAs, Gaussian mutation is generally used.27,28 Ch. Y. Lee and X. Yao68 proposed Levy-based mutation operators. Each one of those mutation operators is determined by values of Levy distribution’s $\alpha$ parameter. This kind of mutation operator characterizes a special kind of EP, called Levy Evolutionary Programming (LEP). According to Ch. Y. Lee and X. Yao,68 a population of $n$ candidate solutions is represented as a set of $n$ vector pairs $(\vec{x}_i, \vec{\sigma}_i), i = 1,...,n,$ such that

$$\vec{x}_i = \{x_{i,1},...,x_{i,n}\} \quad \vec{\sigma}_i = \{\sigma_{i,1},...,\sigma_{i,n}\} \quad (13)$$

For each parent candidate solution $(\vec{x}_i, \vec{\sigma}_i)$, an offspring candidate solution $(\vec{x}_i', \vec{\sigma}_i')$ is generated by randomly changing the parent as presented in Eqs 14, 15 and 16.

$$\vec{\sigma}_{i,j}' = \vec{\sigma}_{i,j} \exp\left\{r' \cdot N(0,1) + \pi N(0,1)\right\} \quad (14)$$
At each iteration, LEP applies a set of four different mutation operators over every candidate solution in the current population: Cauchy mutation ($\alpha = 1.0$), Gaussian mutation ($\alpha = 2.0$), $\alpha = 1.3$ and $\alpha = 1.7$, each one having its own step size. It is well known that Gaussian mutation is better for generating random moves with relatively short step sizes than Cauchy mutation, which is better for generating random moves with relatively long step sizes. The search is expected to benefit from the simultaneous use of random moves with different step sizes because at any given moment we cannot know how far each candidate solution is from the global minimum. Applying those four Levy mutations upon each one of the $n$ candidate solutions in the current population, the generated set of $4n$ offspring candidate solutions is expected to be well distributed over the search space.

**III.c. Stochastic function transformation: DEEPSAM mutation operators**

We adopted LEP Levy mutation operators’ idea with an important modification: instead of using stochastic value transformations, we use stochastic function transformations. Given a range $[0.0, t_{\text{max}}]$ of PES smoothing levels, let's partition it into $n$ overlapping sub-ranges and let an $M$-step smoothing protocol be over each one of them. Let’s assign each of the $n$ smoothing sub-ranges to each of the $n$ conformations of the current population in ascending order of conformation energy. For each one of the $n$ overlapping sub-ranges, and for some specific value of $\alpha$ (which determines a specific Levy distribution instance), an actual PES smoothing level will be randomly chosen from the $M$-step smoothing protocol. Let’s represent the set of $n$ actual PES smoothing levels as a pair $(\bar{t}, \bar{\sigma})$, such that

$$\bar{t} = \{t_1, \ldots, t_n\} \quad \bar{\sigma} = \{\sigma_1, \ldots, \sigma_n\}$$  \hspace{1cm} (17)
For the current "parent" pair $\left( i, \sigma \right)$, a new "offspring" pair $\left( i', \sigma' \right)$ is randomly generated as presented in Eqs. 16, 18 and 19:

$$\sigma'_i = \sigma_i \exp \left\{ \tau' N(0,1) + \alpha N_i(0,1) \right\}$$  \hspace{1cm} (18)$$

$$t'_i = \text{mod}_n \left( t_i + \sigma'_i L_i(\alpha) \right)$$  \hspace{1cm} (19)$$

Using the procedure described above, $\alpha$-determined Levy distribution instances are used to randomly choose actual PES smoothing levels for each one of the $n$ overlapping sub-ranges. Each one of those actual smoothing levels determines the actual smoothed PES $F_i$ upon which the LS$_i$ of the DEMSA procedure will be applied. For each conformation in the current population (call it a parent-conformation), instead of randomly changing its coordinates like in LEP mutation operators, DEEPSAM randomly changes actual PES smoothing levels upon which specific DEMSA procedures are generated; those dynamically generated DEMSA procedures are DEEPSAM’s actual mutation operators. In Figure 7 we see two parent-conformations (one in red and the other in green) upon which three DEMSA procedures are applied, generating three offspring conformations each.

As can be seen in Figure 7, the step size of each DEEPSAM mutation operator actually used depends on the PES-smoothing level applied and the kind of DEMSA procedure used.
IV. DEEPSAM selection operators and self-adaptability

Two general kinds of selection operators are used in EAs: parent selection and survivor selection. The former refers to techniques for selecting individuals of the current population to be parents of the next generation population and the latter refers to techniques for selecting offspring individuals to survive and be part of the next generation population.

DEEPSAM, like most implementations of EP, uses all the conformations of the current population as parent-conformations. DEEPSAM's survivor selection operator is inspired by a family-oriented selection operator proposed by D. Thierens and D. Goldberg's. For each one of the \( n \) parent-conformations, (a) five DEEPSAM mutation operators are stochastically generated by using Levy distribution instances determined by five corresponding \( \alpha \)-values (1.0 (Cauchy), 2.0 (Gaussian), and three additional values between them); (b) a family of six conformations is created by applying those five mutation operators upon the parent-conformation, generating one offspring-conformation each – all together \( n \) families. For each one of those \( n \) families, the offspring conformations and their parent compete among themselves, selecting the survivor-conformation as follows: (a) if the best offspring is better than the parent, it is selected as the survivor of its family; (b) otherwise, the most structurally similar to the parent is selected according to the Metropolis acceptance criterion. The set of \( n \) selected survivor-conformations will be the new population to be used in the next iteration.

In addition to this survivor selection operator, DEEPSAM uses a second selection operator inspired by GSA, which is a GA hybridized with SA. GSA keeps the best local minimum found since the beginning of the computation – it is called the 'best-so-far'; when the algorithm stops, the 'best-so-far' is the computed global minimum. Instead of keeping only one 'best-so-far' minimum, DEEPSAM keeps a 'best-so-far' population of local minima found since the beginning of the computation. At the end of each iteration, an elitist selection operator is applied upon a set of \( 2n \) conformations: the \( n \) "survivor" conformations result of the current iteration and the \( n \) conformations of the current 'best-so-far' population. The 'best-so-far' population will always contain the best (not necessarily different) \( n \) conformations found since the beginning of the
computation. The best conformations are prevented to be lost, what makes sure that the sequence of 'best-so-far' populations since the beginning of the computation is monotonically descendent in terms of energy. This means that at the end of DEEPSAM's computation the 'best-so-far' population will contain an ensemble of the energetically deepest conformations found so far – the ensemble towards which the computation converged (including the computed global minimum). Because of that we may expect DEEPSAM to achieve the goal of getting an ensemble of conformations energetically (and hopefully structurally) close to each other, including the computed global minimum.

For each family, according to the result of the selection operator, the algorithm adapts itself in response to what happens during the computation, deciding which type of DEMSA procedure to use in the next iteration: taking into account the properties described above for each DEMSA procedure, each one of the DEMSA procedures to be actually applied upon the conformations of the new population, are selected depending on the energy comparison between the parent-conformation and the best offspring-conformation of the previous iteration. Detail treatment of Self-Adaptation in Evolutionary Algorithms can be found elsewhere.\textsuperscript{28,29,72-77}

Depending on the size of the molecular system, and because of the all-atom molecular representation in use, DEMSA’s resource (memory and computation time) requirements may be relatively expensive – a reason to try to restrict the population size. It is known that small populations provoke a lack of population diversity and premature convergence. To deal with this, the micro-GAs\textsuperscript{78-81} approach has been adopted in DEEPSAM: if at some iteration, all the best-so-far population has not changed, the algorithm re-initialize itself – a completely new population is generated, taking the best-so-far conformation as its initial conformation; the algorithm is restarted from this re-initialized population.

V. Stopping Criterion

Few recent papers\textsuperscript{82-85} try to determine stopping criteria for EAs - DEEPSAM implementation uses a traditional stopping criterion: a priori arbitrary determination of CPU time/iterations limit.
4 – Results

I. Proof of Concept – Cyclic and Non-Cyclic Peptides

I.a. Background

In order to test our algorithm, we wanted to choose a set of peptides with the following characteristics: (a) their sequence should be long enough to have a non-trivial number of local minima; (b) their sequence should be short enough to allow relatively short run times, which will enable a thorough analysis (and "post-mortem" debugging) of each run; (c) we were interested in trying our algorithm upon cyclic versions whose native structures are already known – cyclic peptides also are less flexible than non-cyclic ones; (d) we were interested in trying our algorithm also upon non-cyclic versions. Thus, as our *proof of concept*, we chose six short peptides from the set of cyclic peptides used by A. Rayan, H. Senderowitz and A. Goldblum. The X-Ray structures of those peptides can be seen in Figure 8.

\[
\begin{align*}
\text{(GLY-HIS-GLY-ALA-TYR-GLY)} & \quad \text{(GLY-LUE-GLY-GLY-LEU-GLY)} \\
\end{align*}
\]

**Figure 8a.** X-Ray structures of short cyclic peptides used to test our algorithm
A. Rayan, H. Senderowitz and A. Goldblum\textsuperscript{86} stochastically construct cyclic peptides conformations using standard bond lengths, standard bond angles, and backbone angles taken from a database of proteins. In our case, we used TINKER's protein program to create structures "by-hand" corresponding to the given sequences; part of the structures were created in linear non-cyclic form, and part in head-to-tail cyclic form.

I.b. Analysis and Results

We consider now the results we got, starting from an unfolded structure. Note that in non-cyclic

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig8b.png}
\caption{X-Ray structures of short cyclic peptides used to test our algorithm}
\end{figure}
cases, the initial structures are \textit{linear}. For each of those peptides:

1. The structure created by-hand was locally minimized, becoming the initial conformation for all subsequent calculations.

2. In the Introduction to this Thesis we said that a protein’s PES has an estimated number of minima, exponential on the length of the sequence. According to Rayan, A., Senderowitz, H., and Goldblum, A.\textsuperscript{86} that estimation should not be different for peptides, even in the case of cyclic peptides where cyclization is hoped to promote higher stability and conformational restriction. Inspired by what was done by A. Rayan, H. Senderowitz and A. Goldblum,\textsuperscript{86} given the initial conformation, we used TINKER's \textit{scan} program in order to do a relatively exhaustive sampling of the PES. Considering that \textit{scan} samples the PES by random "jumps", we arbitrarily decided that it will be enough to let \textit{scan} do at least 40000 random "jumps". The energetically deepest conformation found after those 40000 "jumps" over the PES, was considered by us the \textit{scan}'s global minimum.

3. According to what has been described in section 2 above, given the \textit{same} initial conformation, we run DEEPSAM on it and the energetically deepest local minimum found was considered by us DEEPSAM's global minimum.

4. We run DEEPSAM again, taking \textit{scan}'s global minimum as the new initial conformation.

5. We analyze the results we got from (1)-(4).

6. We compare our results with X-Ray experimental data (visualized in Figure 8) – we will show good structural agreement between them.

Taking into account the importance of solvation on bio-molecules’ function, we also run DEEPSAM upon the same set of peptides using an implicit solvent model. Results in this context will be also presented.

The FF used by us in all the calculations done upon those peptides is TINKER's implementation of OPLSAA which was especially developed to deal with the modeling of proteins and peptides, and it is one of the mostly used for this purpose.

As widely recognized, the low energy structures of bio-molecules are of great importance for the
determination of their biological function. Because of that, assuming PES correctness, energy is the fundamental criterion used in protein structure identification algorithms. It is encouraging that DEEPSAM's energy minimization effectively leads to a very good approximation of the native structure, even with population size as small as five conformations only. When we come to analyze the results that we got from our calculations, structural aspects become of most great importance. We chose two cases which will be analyzed in detail; for the other four cases, the analysis provided will be in a more summarized form; see Figure 9 for a visualization of those two chosen structures.

Our analysis will be as follows: first, we present DEEPSAM's calculated global minima, and we compare them with the corresponding scan's calculated global minima and with the corresponding X-RAY crystallographic structures; second, we do a similar analysis for structures we got when the initial structure upon which DEEPSAM was run again, was scan's global minimum; third, we analyze run-times; finally, we analyze the evolution of the computation process.

![Initial structure vs. Scan's Global minimum](image)

<table>
<thead>
<tr>
<th>Initial structure</th>
<th>Scan's Global minimum</th>
</tr>
</thead>
</table>

**Figure 9a.** cyclic-(GLY-HIS/GLY-ALA-TYR-GLY)
<table>
<thead>
<tr>
<th>DEEPSAM's global minimum</th>
<th>DEEPSAM's run from <em>scan</em>'s global minimum</th>
</tr>
</thead>
</table>

**Figure 9b.** cyclic-(GLY-HIS-GLY-ALA-TYR-GLY)

<table>
<thead>
<tr>
<th>Initial structure</th>
<th><em>scan</em>'s Global minimum</th>
</tr>
</thead>
</table>

**Figure 9c.** non cyclic-(GLY-HIS-GLY-ALA-TYR-GLY)
Figure 9d. non cyclic-(GLY-HIS-GLY-ALA-TYR-GLY)

Figure 9e. cyclic-(ILE-PRO-ILE-PHE-PRO-TYR-PRO)
Starting from unfolded structures. In Table 1, energy data may be found for DEEPSAM and \textit{scan} calculations, starting from the same initial local minimum. We can see that in five from the six cases, DEEPSAM succeeds to find conformations with the same energy, or energetically deeper, than those found by \textit{scan}; see also that the only case in which DEEPSAM did not succeed to find a conformation energetically deeper than that found by \textit{scan}, the difference is only 0.107 Kcal/Mol.

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Initial Local Minimum</th>
<th>Scan Global Minimum</th>
<th>DEEPSAM Global Minimum</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyclic-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>-125.389</td>
<td>-162.191</td>
<td>-164.948</td>
<td>-2.757</td>
</tr>
<tr>
<td>non-Cyclic-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>-151.625</td>
<td>-303.799</td>
<td>-303.799</td>
<td>0.000</td>
</tr>
<tr>
<td>Cyclic-GLY-LUE-GLY-GLY-LEU-GLY</td>
<td>-126.648</td>
<td>-155.694</td>
<td>-156.443</td>
<td>-0.749</td>
</tr>
<tr>
<td>non-Cyclic-PRO-PRO-PHE-PHE-LEU-ILE-ILE-LEU-VAL</td>
<td>-150.476</td>
<td>-272.993</td>
<td>-273.297</td>
<td>-0.304</td>
</tr>
<tr>
<td>Cyclic-ILE-PRO-ILE-PHE-PRO-TYR-PRO</td>
<td>-49.037</td>
<td>-76.244</td>
<td>-76.137</td>
<td>0.107</td>
</tr>
<tr>
<td>Cyclic-ALA-ILE-PRO-PHE-ASN-SER-LEU</td>
<td>-146.630</td>
<td>-175.172</td>
<td>-175.713</td>
<td>-0.541</td>
</tr>
<tr>
<td>Cyclic-SER-PHE-LEU-PRO-VAL-ASN-LEU</td>
<td>-141.718</td>
<td>-166.308</td>
<td>-167.868</td>
<td>-1.560</td>
</tr>
</tbody>
</table>

Table 2 presents a measure of structural similarity (RMSD) between global minima computed by DEEPSAM and by \textit{scan}. Note that in spite of being energetically close, there are cases that their RMSD is at least 4 Å, which is a significant structural dissimilarity. Also note that taking into account only the backbone heavy atoms, the structural dissimilarities are smaller, meaning that the higher overall RMSD is influenced by side chains.
Table 2: RMSD comparison (in Å) among DEEPSAM and scan Global Minima structures

<table>
<thead>
<tr>
<th>Peptide</th>
<th>DEEPSAM-Scan Global Minima all-atoms RMSD</th>
<th>DEEPSAM-Scan Global Minima backbone RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyclic-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>0.932</td>
<td>0.932</td>
</tr>
<tr>
<td>non-Cyclic-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>0.271</td>
<td>0.001</td>
</tr>
<tr>
<td>Cyclic-GLY-LUE-GLY-GLY-LEU-GLY</td>
<td>1.774</td>
<td>0.033</td>
</tr>
<tr>
<td>non-Cyclic-PRO-PRO-PHE-PHE-LEU-ILE-ILE-ILE-LEU-VAL</td>
<td>4.612</td>
<td>3.111</td>
</tr>
<tr>
<td>Cyclic-ILE-PRO-ILE-PHE-PRO-TYR-PRO</td>
<td>4.901</td>
<td>2.213</td>
</tr>
<tr>
<td>Cyclic-ALA-ILE-PRO-PHE-ASN-SER-LEU</td>
<td>4.062</td>
<td>1.910</td>
</tr>
<tr>
<td>Cyclic-SER-PHE-LEU-PRO-VAL-ASN-LEU</td>
<td>3.796</td>
<td>2.133</td>
</tr>
</tbody>
</table>

See Figure 10 for a visualization of the superposition of those global minima.

![Figure 10a. Scan and DEEPSAM global minima superposition (blue and red, respectively)](image-url)
It is well known that RMSD alone is not a good measure of structure similarity; because of that, we wanted to identify a set of superposable amino acids in better alignment than the overall structure. For this purpose, we used the MaxCluster program,\textsuperscript{94} which implements well known structure comparison algorithms,\textsuperscript{95-98} especially MaxSub,\textsuperscript{95} which has been adopted by CASP as one of its structure comparison tools. MaxSub measures structural similarity by identifying maximal superposable sub-structures (MAXSUBs) within an arbitrarily determined distance cutoff (only the backbone's heavy atoms are considered for the structural superposition). Determined the MAXSUB superposition, we get three similarity measures: its rank \( r \) (0.0 \leq r \leq 1.0), its MAXSUB RMSD, and its gRMSD. The gRMSD is the RMSD of the two conformations when they are superposed according to the MAXSUB found within the specified distance cutoff. The difference between gRMSD and the optimal RMSD, tells us how much the MAXSUB superposition moved the two conformations relative to the optimal overall superposition. According to a rule-of-thumb suggested by A. Rayan, H. Senderowitz and A. Goldblum,\textsuperscript{86} a good superposition vis-à-vis the size of the peptides we are dealing with, is that whose RMSD is 1.0 Å or less. Following this rule-of-thumb, we would like to determine distance cutoffs for which the RMSDs of the MAXSUBs found, are at most 1.0 Å. In order to determine those distance cutoffs,
we run *MaxCluster* several times, each time assigning the distance cutoff a value from the decreasing sequence [4.5, 4.0, 3.5, 3.0, 2.5, 2.0, 1.5, 1.0, 0.5]. For all the distance cutoffs for which it is possible to find MAXSUBs, we choose all those for which their MAXSUBs are proper subsets of the peptide sequence. From all of them, we choose all those for which their RMSDs are at most 1.0 Å and they are the longest alignments. If there is more than one, we will choose that whose distance cutoff is the tightest of all of them. If no one is a proper subset of the peptide sequence, we will choose that with the tightest distance cutoff. If the RMSDs of all the MAXSUBs found are greater than 1.0 Å, we choose all those with the tightest RMSD. There may be cases where the same MAXSUB, with the same RMSD, is found for different distance cutoffs – that with the tightest distance cutoff will be chosen.

In Table 3 we can see that even in cases where the overall superposition is not in good agreement, it is possible to find a sub-structure whose superposition is in very good agreement, giving a better perspective of the similarity between structures, than only the rough overall RMSD.

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Distance cutoff</th>
<th>Alignment</th>
<th>Rank</th>
<th>RMSD</th>
<th>gRMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyclic-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>1.0</td>
<td>GHG_YG</td>
<td>0.667</td>
<td>0.510</td>
<td>0.840</td>
</tr>
<tr>
<td>Cyclic-GLY-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>0.5</td>
<td>GHGAYG</td>
<td>1.000</td>
<td>0.001</td>
<td>0.009</td>
</tr>
<tr>
<td>Cyclic – GLY-LUE-GLY-GLY-LEU-GLY</td>
<td>0.5</td>
<td>GLGGLG</td>
<td>0.998</td>
<td>0.021</td>
<td>0.063</td>
</tr>
<tr>
<td>non-Cyclic-PRO-PRO-PHE-PHE-LEU-Ile-LEU-Val</td>
<td>1.5</td>
<td>P___LIILV</td>
<td>0.490</td>
<td>0.968</td>
<td>4.742</td>
</tr>
<tr>
<td>Cyclic ILE-PRO-ILE-PHE-PRO-TYR-PRO</td>
<td>1.0</td>
<td><strong>IFPY</strong></td>
<td>0.446</td>
<td>0.545</td>
<td>4.454</td>
</tr>
<tr>
<td>Cyclic ALA-ILE-PRO-PHE-ASN-SER-LEU</td>
<td>1.5</td>
<td>AIPFNSL</td>
<td>0.702</td>
<td>0.741</td>
<td>1.701</td>
</tr>
<tr>
<td>Cyclic SER-PHE-LEU-PRO-VAL-ASN-LEU</td>
<td>1.0</td>
<td><strong>LPVN</strong></td>
<td>0.464</td>
<td>0.509</td>
<td>3.214</td>
</tr>
</tbody>
</table>

In Table 3: MaxSub comparison (in Å) among DEEPSAM and scan Global Minima structures

In Figure 11, we can see visualizations of MAXSUB superpositions. Note that the thicker "sticks" in the displayed picture, represent the MAXSUB.
DEEPSAM is a population-oriented algorithm. All the conformations in the population evolve together, in parallel, towards the hopefully global minimum. The algorithm maintains two kinds of populations at each iteration, (a) an ensemble of conformations (call it the current population) is generated based on the ensemble of the previous iteration, and (b) the ensemble of best-so-far conformations. At the end of each iteration, the best-so-far ensemble is updated with all those conformations of the current population which are energetically deeper than the worst one in it. It
is clear that the ensemble of best-so-far conformations is ensured to be energetically decreasing; this way, we can expect the ensemble to converge to some neighborhood of the lowest energy conformation found. We also may expect that those conformations will be structurally close to each other, but we cannot warrant it in general (it is clear that two dissimilar structures can have the same energy value!).

For all our peptides we compared among the conformations of the best-so-far ensemble using *MaxCluster*’s multi-conformation structure comparison feature – see Table 4.

**Table 4:** Comparison (in Å) between the global minimum and the other conformations of the best-so-far population

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Conf1</th>
<th>Conf2</th>
<th>Conf3</th>
<th>Conf4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyclic-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>0.056</td>
<td>0.051</td>
<td>0.057</td>
<td>0.073</td>
</tr>
<tr>
<td>non-Cyclic-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>0.700</td>
<td>1.232</td>
<td>0.664</td>
<td>0.057</td>
</tr>
<tr>
<td>Cyclic-GLY-LUE-GLY-GLY-LEU-GLY</td>
<td>0.388</td>
<td>0.067</td>
<td>0.049</td>
<td>0.033</td>
</tr>
<tr>
<td>non-Cyclic-PRO-PRO-PHE-PHE-LEU-ILE-ILE-LEU-VAL</td>
<td>1.194</td>
<td>0.106</td>
<td>0.148</td>
<td>0.039</td>
</tr>
<tr>
<td>Cyclic-ILE-PRO-ILE-PHE-PRO-TYR-PRO</td>
<td>0.121</td>
<td>0.044</td>
<td>0.210</td>
<td>0.018</td>
</tr>
<tr>
<td>Cyclic-ALA-ILE-PRO-PHE-ASN-SER-LEU</td>
<td>0.039</td>
<td>0.029</td>
<td>0.027</td>
<td>0.019</td>
</tr>
<tr>
<td>Cyclic-SER-PHE-LEU-PRO-VAL-ASN-LEU</td>
<td>0.937</td>
<td>0.713</td>
<td>0.601</td>
<td>0.589</td>
</tr>
</tbody>
</table>

See that we got not only the global minimum, but also an ensemble of structurally close conformations that includes the global minimum. A visualization of the superposition of all the conformations of the best-so-far ensemble can be seen in Figure 12, for the two chosen peptides.
In Table 5, we show the agreement we found between the structures predicted by DEEPSAM and their corresponding X-RAY crystallographic structures (available from the Cambridge Structural Database).

For the chosen peptides, a visualization of the superposition of DEEPSAM's calculated global minima and the X-RAY crystallographic structures can be seen in Figure 13.
<table>
<thead>
<tr>
<th>Peptide</th>
<th>Backbone RMSD</th>
<th>Distance cutoff</th>
<th>MAXSUB Alignment</th>
<th>MAXSUB Rank</th>
<th>MAXSUB RMSD</th>
<th>gRMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>cyc-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>1.854</td>
<td>1.5</td>
<td>_HGAYG</td>
<td>0.548</td>
<td>1.108</td>
<td>1.976</td>
</tr>
<tr>
<td>non-cyc-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>1.677</td>
<td>1.5</td>
<td>_HGAYG</td>
<td>0.653</td>
<td>0.821</td>
<td>1.951</td>
</tr>
<tr>
<td>cyc-GLY-LUE-GLY-GLY-LEU-GLY</td>
<td>1.722</td>
<td>2.0</td>
<td>_LGGLG</td>
<td>0.679</td>
<td>0.970</td>
<td>1.686</td>
</tr>
<tr>
<td>non-cyc-PRO-PRO-PHE-PHE-LEU-ILE-ILE-LEU-VAL</td>
<td>3.869</td>
<td>2.0</td>
<td><strong>FFLI</strong></td>
<td>0.339</td>
<td>1.149</td>
<td>7.812</td>
</tr>
<tr>
<td>cyc-ILE-PRO-ILE-PHE-PRO-TYR-PRO</td>
<td>2.554</td>
<td>1.5</td>
<td>IPI__YP</td>
<td>0.627</td>
<td>0.561</td>
<td>3.164</td>
</tr>
<tr>
<td>cyc-ALA-ILE-PRO-PHE-ASN-ROTLEU</td>
<td>2.547</td>
<td>2.0</td>
<td>A_P_NS_</td>
<td>0.506</td>
<td>0.756</td>
<td>2.393</td>
</tr>
<tr>
<td>cyc-SER-PHE-LEU-PRO-VAL-ASN-ROTLEU</td>
<td>2.133</td>
<td>1.0</td>
<td><strong>LPVN</strong></td>
<td>0.443</td>
<td>0.565</td>
<td>3.287</td>
</tr>
</tbody>
</table>

**Figure 13a.** Superposition of DEEPSAM global minimum and X-RAY structures (the X-RAY structure is in blue, and the DEEPSAM global minimum structure is in red)
**Figure 13b.** Superposition of DEEPSAM global minimum and X-RAY structures (the X-RAY structure is in blue, and the DEEPSAM global minimum structure is in red)

A visualization of their MAXSUB superposition can be seen in Figure 14.

**Figure 14a.** MAXSUB Superposition of DEEPSAM global minimum and X-RAY structures (the X-RAY structure is in blue, and the DEEPSAM global minimum structure is in red)
From Table 5 and Figures 13 and 14, it is possible to see that the structures predicted by DEEPSAM for GLY-HIS-GLY-ALA-TYR-GLY in both, the case of cyclic initial structure and the case of non-cyclic initial structure, are similar one to the other, and to the X-Ray structure. Following this observation, we compared both final structures. The results of the superposition of both are shown in Table 6 and in Figure 15.

**Table 6: Comparing (in Å) Global Minima of cyclic and non-cyclic GLY-HIS-GLY-ALA-TYR-GLY**

<table>
<thead>
<tr>
<th>Backbone RMSD</th>
<th>Distance cutoff</th>
<th>MAXSUB Alignment</th>
<th>MAXSUB Rank</th>
<th>MAXSUB RMSD</th>
<th>MAXSUB gRMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.011</td>
<td>2.0</td>
<td>GHGAYG</td>
<td>0.891</td>
<td>0.715</td>
<td>0.794</td>
</tr>
</tbody>
</table>

**Figure 14b.** MAXSUB Superposition of DEEPSAM global minimum and X-RAY structures (the X-RAY structure is in blue, and the DEEPSAM global minimum structure is in red)

**Figure 15.** Superposition of cyclic and non-cyclic GLY-HIS-GLY-ALA-TYR-GLY DEEPSAM global minima (the cyclic is in blue, and the non-cyclic is in red)
It seems that two H-bonds were created during the computation process (see Figures 9c and 9d): (GLY1.N – GLY1.2H ⋅⋅⋅ GLY6.OXT) and (HIS2.ND1 – HIS2.HD1 ⋅⋅⋅ GLY6.O). This means that DEEPSAM was able to "cyclize" a structure that in the beginning was linear.

**Starting from scan's global minimum structures.** In this section, our presentation will be similar to that in the previous section. In Table 7, energy data may be found for all DEEPSAM calculations, when the starting conformations are the *scan* global minima.

**Table 7: Energy comparison (in Kcals/Mol) among structures computed by DEEPSAM (from the *scan* global minimum)**

<table>
<thead>
<tr>
<th>Peptide</th>
<th>DEEPSAM Global Minimum</th>
<th>Scan Global Minimum</th>
<th>DEEPSAM Global Minimum starting from scan</th>
</tr>
</thead>
<tbody>
<tr>
<td>cyc-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>-164.948</td>
<td>-162.191</td>
<td>-164.995</td>
</tr>
<tr>
<td>cyc-GLY-LUE-GLY-GLY-LEU-GLY</td>
<td>-156.443</td>
<td>-155.694</td>
<td>-156.443</td>
</tr>
<tr>
<td>cyc-ILE-PRO-ILE-PHE-PRO-TYR-PRO</td>
<td>-76.137</td>
<td>-76.244</td>
<td>-77.468</td>
</tr>
<tr>
<td>cyc-ALA-ILE-PRO-PHE-ASN-SER-LEU</td>
<td>-175.713</td>
<td>-175.172</td>
<td>-176.808</td>
</tr>
<tr>
<td>cyc-SER-PHE-LEU-PRO-VAL-ASN-LEU</td>
<td>-167.868</td>
<td>-166.308</td>
<td>-169.360</td>
</tr>
</tbody>
</table>

Table 8 presents RMSD structural similarity measures between the global minima computed by DEEPSAM, starting from the *scan* global minimum, and the global minima computed starting from the initial structure.
Table 8: RMSD comparison (in Å) among global minima computed by DEEPSAM from the initial local minimum and from the scan global minima.

<table>
<thead>
<tr>
<th>Peptide</th>
<th>DEEPSAM from Scan all-atoms RMSD</th>
<th>DEEPSAM from Scan Backbone RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>cyc-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>0.539</td>
<td>0.051</td>
</tr>
<tr>
<td>non-cyc-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>1.193</td>
<td>0.001</td>
</tr>
<tr>
<td>cyc-GLY-LUE-GLY-GLY-LEU-GLY</td>
<td>1.185</td>
<td>0.001</td>
</tr>
<tr>
<td>non-cyc-PRO-PRO-PHE-PHE-LEU-ILE-ILE-LEU-VAL</td>
<td>4.935</td>
<td>3.368</td>
</tr>
<tr>
<td>cyc-ILE-PRO-ILE-PHE-PRO-TYR-PRO</td>
<td>4.639</td>
<td>2.239</td>
</tr>
<tr>
<td>cyc-ALA-ILE-PRO-PHE-ASN-SER-LEU</td>
<td>0.956</td>
<td>0.327</td>
</tr>
<tr>
<td>cyc-ER-PHE-MEU-MRO-MAL-AS-MEU</td>
<td>3.927</td>
<td>2.160</td>
</tr>
</tbody>
</table>

Table 9 presents MAXSUB comparison among the same global minima whose RMSD comparison was presented in Table 8.

Table 9: MAXSUB comparison (in Å) among global minima computed by DEEPSAM from the initial local minimum and from the scan global minima.

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Distance cutoff</th>
<th>Alignment</th>
<th>Rank</th>
<th>RMSD</th>
<th>gRMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>cyc-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>0.5</td>
<td>GHGAYG</td>
<td>0.997</td>
<td>0.027</td>
<td>0.073</td>
</tr>
<tr>
<td>non-cyc-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>1.0</td>
<td>GHGAYG</td>
<td>1.000</td>
<td>0.001</td>
<td>0.010</td>
</tr>
<tr>
<td>cyc-GLY-LUE-GLY-GLY-LEU-GLY</td>
<td>0.5</td>
<td>GLGLGL</td>
<td>1.000</td>
<td>0.001</td>
<td>0.010</td>
</tr>
<tr>
<td>non-cyc-PRO-PRO-PHE-PHE-LEU-ILE-ILE-LEU-VAL</td>
<td>1.0</td>
<td>___<em>III</em></td>
<td>0.324</td>
<td>0.167</td>
<td>4.364</td>
</tr>
<tr>
<td>cyc-ILE-PRO-ILE-PHE-PRO-TYR-PRO</td>
<td>1.0</td>
<td>___<em>IFPY</em></td>
<td>0.440</td>
<td>0.560</td>
<td>4.515</td>
</tr>
<tr>
<td>cyc-ALA-ILE-PRO-PHE-ASN-SER-LEU</td>
<td>0.5</td>
<td>AIPFNSL</td>
<td>0.983</td>
<td>0.066</td>
<td>0.117</td>
</tr>
<tr>
<td>cyc-SER-PHE-LEU-PRO-VAL-ASN-LEU</td>
<td>1.0</td>
<td>___<em>LPVN</em></td>
<td>0.471</td>
<td>0.487</td>
<td>3.296</td>
</tr>
</tbody>
</table>

From the data presented in Tables 7, 8, and 9, we see that in all the cases of our test set, when DEEPSAM starts its calculations from the minimum found by scan, it found minima deeper than both, the minimum found by scan and the minimum found by DEEPSAM itself, when the
starting state was at the initial conformation. From the structural comparisons, we see that in cases 1, 2, 3, and 6, the newly found minima were very close to those found when the conformation from which DEEPSAM started was the initial conformation. We also see that in the other three cases (4, 5, and 7), the structures found starting from that found by \textit{scan}, are very different from those found by DEEPSAM when it started at the initial conformation.

Having these newly found minima, we compare them with the X-RAY crystallographic structures. Table 10 presents the results of these comparisons.

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|c|c|c|c|}
\hline
Peptide & Backbone RMSD & Distance cutoff & MAXSUB Alignment & MAXSUB Rank & MAXSUB RMSD & MAXSUB gRMSD \\
\hline
\textit{cyc-GLY-HIS-GLY-ALA-TYR-GLY} & 1.840 & 1.5 & GHG___ & 0.499 & 0.075 & 4.495 \\
\textit{non-cyc-GLY-HIS-GLY-ALA-TYR-GLY} & 1.677 & 1.5 & __GAY_ & 0.498 & 0.095 & 1.906 \\
\textit{cyc-GLY-LUE-GLY-GLY-LEU-GLY} & 1.722 & 2.5 & _LGGLG & 0.726 & 0.970 & 1.686 \\
\textit{non-cyc-PRO-PRO-PHE-PHE-LEU-ILE-ILE-LEU-VAL} & 2.588 & 1.0 & PPFF_____ & 0.329 & 0.596 & 5.562 \\
\textit{cyc-ILE-PRO-ILE-PHE-PRO-TYR-PRO} & 1.732 & 1.5 & I_IFPYP & 0.702 & 0.783 & 0.942 \\
\textit{cyc-ALA-ILE-PHE-ASN-SER-LEU} & 2.575 & 2.0 & A_P_NS_ & 0.513 & 0.706 & 2.383 \\
\textit{cyc-SER-PHE-LEU-PRO-VAL-ASN-LEU} & 1.080 & 0.5 & __FLPVNL & 0.742 & 0.210 & 0.799 \\
\hline
\end{tabular}
\caption{MaxSub Comparison (in Å) among experimental data and DEEPSAM’s global minima structures starting from the \textit{scan} global minima.}
\end{table}

For the chosen peptides, a visualization of the superposition of calculated global minima (starting from \textit{scan} results) and the X-RAY crystallographic structures can be seen in Figure 16.
A visualization of their MAXSUB superposition can be seen in Figure 17.
cyclic-(GLY-HIS-GLY-ALA-TYR-GLY)
non cyclic-(GLY-HIS-GLY-ALA-TYR-GLY)

cyclic-(ILE-PRO-ILE-PHE-PRO-TYR-PRO)

Figure 17. MAXSUB Superposition of global minimum starting from scan and X-RAY structures (X-RAY structure is blue, and DEEPSAM global minimum is red)

From all what we saw in this and in the previous section (see tables 5 and 10, above), we conclude that DEEPSAM may be used as a structure predictor starting its computations from any unfolded structure, but if we can start DEEPSAM's calculations from a relatively good local minimum found by some other method (for example, the low-resolution step of Rosetta), we may expect it to get better predicted structures and in less time (for run time analysis, see below).
Run times. The data presented above are the result of running DEEPSAM and scan over our test set of peptides. DEEPSAM is population-oriented and essentially parallel. In the flow chart diagram, in Figure 2, it can be seen that DEEPSAM's design implements two levels of parallelism: at each iteration, \( n \) sets of five independent DEMSA mutation operators are independently applied on the \( n \) conformations of the current population. As we explained above, we use five different values of the parameter \( \alpha \) of the Levy distribution in order to probabilistically generate five actual smoothing levels that determine five different smoothed PESs upon which five instances of DEMSA are applied. DEEPSAM's parallel design, which was implemented using the Python\(^{99}\) programming language, may be deployed in a grid system like mosix\(^{100}\) – a possibility that we may try in the future. Assuming the initial population has already been created, each DEEPSAM iteration generates five new offspring conformations for every one parent conformation. If \( n \) is the size of the population, each iteration generates \( 5n \) new conformations. Assuming \( m \) is the number of iterations of a given run, the number of conformations generated in that run is \( 5n \times m \). Taking into account the initial population, and assuming there were no population re-initializations, all the conformations generated in a DEEPSAM run is \( 5n \times m + n \). In most of the runs presented here, the population size is \( n = 5 \); that is, the total number of conformations generated is \( 25m + 5 \). The algorithm dynamically changes itself - in the same iteration, different DEMSAs run in parallel, each one with its own computational cost. At each iteration, the main program waits until all the DEMSAs finish; this means that the actual run time of a given iteration is determined by the DEMSA with the longest run time. In fact, this is what is recorded in DEEPSAM's log file as the actual run time of each iteration. Then, in order to have an estimation of DEEPSAM's run time, (a) we take the average run time of the iterations of a given run, and (b) we calculate the overall time based on that average. Scan is individual-oriented and essentially sequential. Then, in order to have an estimation of scan's run time, we also calculate the average time that took scan to generate each conformation, and calculate the overall time based on that average. Then, we compare them.
Table 11 presents DEEPSAM's estimated average and overall run time for every one of the peptides in our test set, until the deepest conformation was found.

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Nr of iterations to find deepest conformation</th>
<th>Nr of generated conformations</th>
<th>Average iteration run-time</th>
<th>Overall run-time</th>
</tr>
</thead>
<tbody>
<tr>
<td>cyc-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>11</td>
<td>280</td>
<td>294.557</td>
<td>3240.127</td>
</tr>
<tr>
<td>non-cyc-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>6</td>
<td>155</td>
<td>677.166</td>
<td>4062.996</td>
</tr>
<tr>
<td>cyc-GLY-LEU-GLY-GLY-LEU-GLY</td>
<td>145</td>
<td>3630</td>
<td>243.618</td>
<td>35324.610</td>
</tr>
<tr>
<td>non-cyc-PRO-PRO-PHE-PHE-LEU-ILE-ILE-LEU-VAL</td>
<td>20</td>
<td>505</td>
<td>1894.553</td>
<td>37891.060</td>
</tr>
<tr>
<td>cyc-ILE-PRO-ILE-PHE-PRO-TYR-PRO</td>
<td>32</td>
<td>805</td>
<td>828.470</td>
<td>26511.040</td>
</tr>
<tr>
<td>cyc-ALA-ILE-PRO-PHE-ASN-SER-LEU</td>
<td>48</td>
<td>1205</td>
<td>316.451</td>
<td>15189.648</td>
</tr>
<tr>
<td>cyc-SER-PHE-LEU-PRO-VAL-ASN-LEU</td>
<td>55</td>
<td>1380</td>
<td>593.274</td>
<td>32630.070</td>
</tr>
</tbody>
</table>

Table 12 presents scan's estimated average and overall run time for every one of the peptides in our test set, until the deepest conformation was found.

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Nr of generated conformations until deepest conformation is found</th>
<th>Average Conformation generation time (in seconds)</th>
<th>Overall run-time (in seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cyc-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>37483</td>
<td>2.084</td>
<td>78099.578</td>
</tr>
<tr>
<td>non-cyc-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>3931</td>
<td>6.388</td>
<td>25112.676</td>
</tr>
<tr>
<td>cyc-GLY-LEU-GLY-GLY-LEU-GLY</td>
<td>13375</td>
<td>1.864</td>
<td>24931.000</td>
</tr>
<tr>
<td>non-cyc-PRO-PRO-PHE-PHE-LEU-ILE-ILE-LEU-VAL</td>
<td>51231</td>
<td>18.828</td>
<td>964577.268</td>
</tr>
<tr>
<td>cyc-ILE-PRO-ILE-PHE-PRO-TYR-PRO</td>
<td>30767</td>
<td>8.895</td>
<td>273672.465</td>
</tr>
<tr>
<td>cyc-ALA-ILE-PRO-PHE-ASN-SER-LEU</td>
<td>20303</td>
<td>5.283</td>
<td>107260.749</td>
</tr>
<tr>
<td>cyc-SER-PHE-LEU-PRO-VAL-ASN-LEU</td>
<td>9360</td>
<td>5.590</td>
<td>52322.400</td>
</tr>
</tbody>
</table>

It is important to note that even if the run times data presented in tables 11 and 12 should be
taken only as estimated times, the difference between DEEPSAM's and scan's run times is so significant, that we may conclude that DEEPSAM is an effective computational procedure.

**History of the Search.** Given one specific case from our test set, the cyc-GLY-HIS-GLY-ALA-TYR-GLY, how does DEEPSAM's computation evolve, energetically and structurally? As we already emphasized, our algorithm is population-oriented, meaning that all the conformations in the population evolve in parallel, and it will be useful for us to look on the evolution of the whole population together.

At each iteration, our algorithm maintains two populations: (a) a special population, called the "best-so-far" population, which keeps the best conformations found from the beginning of the run, being an energetically decreasing ensemble of conformations, and (b) the "current" population, upon which the DEMSA mutation operators are applied, followed by the application of r-select operators on the offspring conformations result of the DEMSA operators, which generates the new "current" population. Plotting the evolution of all the conformations together, shows us how the population as a whole evolves during the computation process. In what follows we show three graphs of two types: RMSD vs. iteration and energy vs. iteration.

![History of Search: RMSD vs Iteration](image)

**Figure 18. History of Search**

In Figure 18 a graph of the first type is presented – it shows how the RMSD changed relative to
the initial conformation structure, giving us an insight on how structures evolved.

**Figure 19. History of Search**

In Figure 19 a graph of the second type is presented – it shows how energy levels evolved, giving us an insight on the structure of the sampled PES areas during the computation process. The conformation called "conf5" is the energetically best conformation in the population.

**Figure 20. History of Search**
In Figure 20, another graph of the first type is presented, showing the evolution of the "best-so-far" population; actually, this graph shows us how the algorithm converges to the final predicted global minimum conformation. This graph also allows us to identify the set of lowest lying conformations found, which includes the predicted global minimum.

In Figures 19 and 20, we can see that at the 11\textsuperscript{th} iteration, the algorithm already finds the global minimum.

**Starting from unfolded structures, using an implicit solvent model.** As we already said above, solvation is of great importance in the function of bio-molecules; because of that, we also run DEEPSAM upon the same set of peptides, using an implicit solvent model. Following the same above methodology, we used the MAXSUB concept in order to compare the predicted against the experimental structures. As can be seen in Table 13, using the GBSA implicit solvent model, the results obtained show a good agreement between the structures predicted by DEEPSAM and their corresponding X-RAY crystallographic structures.

<table>
<thead>
<tr>
<th>Table 13: Comparison (in Å) among experimental data and DEEPSAM's Global Minima, using the GBSA implicit solvent model.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peptide</td>
</tr>
<tr>
<td>cyc-GLY-HIS-GLY-ALA-TYR-GLY</td>
</tr>
<tr>
<td>non-cyc-GLY-HIS-GLY-ALA-TYR-GLY</td>
</tr>
<tr>
<td>cyc-GLY-LUE-GLY-GLY-LEU-GLY</td>
</tr>
<tr>
<td>non-cyc-PRO-PRO-PHE-PHE-LEU-ILE-ILE-LEU-VAL</td>
</tr>
<tr>
<td>cyc-ILE-PRO-ILE-PHE-PRO-TYR-PRO</td>
</tr>
<tr>
<td>cyc-ALA-ILE-PRO-PHE-ASN-SER-LEU</td>
</tr>
<tr>
<td>cyc-SER-PHE-LEU-PRO-VAL-ASN-LEU</td>
</tr>
</tbody>
</table>

For the chosen peptides, a visualization of the superposition of DEEPSAM's calculated global minima and the X-RAY crystallographic structures can be seen in Figure 21. A visualization of their MAXSUB superposition can be seen in Figure 22.
Like in the results without solvent, in Table 13 and Figures 21 and 22, it is possible to see that the structures predicted by DEEPSAM for GLY-HIS-GLY-ALA-TYR-GLY in both, the case of cyclic initial structure and the case of non-cyclic initial structure, are similar one to the other, and to the X-Ray structure. Also here, like in the case without solvent, we compared both final structures. The results of the superposition of both are shown in Table 14 and in Figure 23. It seems that one H-bond was created during the computation process (see Figure 22): (HIS2.NE2 – HIS2.HE2 ⋅⋅⋅ GLY6.O). This means that also in the case with the GBSA implicit solvent model, DEEPSAM was apparently able to "cyclize" a structure that in the beginning was linear.

![Superposition of DEEPSAM's global minimum and X-RAY structures, with GBSA implicit solvent (the X-RAY structure is in blue, and the DEEPSAM's global minimum structure is in red)](image)

**Figure 21.** Superposition of DEEPSAM's global minimum and X-RAY structures, with GBSA implicit solvent (the X-RAY structure is in blue, and the DEEPSAM's global minimum structure is in red)
Figure 22. MAXSUB Superposition of DEEPSAM global minimum and X-RAY structures, with GBSA implicit solvent (the X-RAY structure is in blue, and the DEEPSAM's global minimum structure is in red).

Table 14: Comparing (in Å) DEEPSAM's Global Minima of cyclic and non-cyclic GLY-HIS-GLY-ALA-TYR-GLY, with GBSA implicit solvent.

<table>
<thead>
<tr>
<th>Backbone RMSD</th>
<th>Distance cutoff</th>
<th>MAXUSB Alignment</th>
<th>MAXSUB Rank</th>
<th>MAXSUB RMSD</th>
<th>MAXSUB gRMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.414</td>
<td>1.0</td>
<td>_HGAYG</td>
<td>0.651</td>
<td>0.554</td>
<td>1.888</td>
</tr>
</tbody>
</table>
Figure 23. Superposition of cyclic and non-cyclic GLY-HIS-GLY-ALA-TYR-GLY DEEPSAM's global minima, with GBSA implicit solvent (the cyclic is in blue, and the non-cyclic is in red)

All what we saw in this section strengthen even more what we already said for the cases without solvent: DEEPSAM may be used as a structure predictor starting its computations from any unfolded structure.

**Comparing DEEPSAM results with those of its component methods.** In most of this chapter, we compared DEEPSAM results with a "brute force" PES sampling algorithm, implemented by TINKER's *scan* program. In this section we will present results we got by running DEEPSAM component methods over the cyclic and non-cyclic versions of the GLY-HIS-GLY-ALA-TYR-GLY peptide; we will compare them with the already presented DEEPSAM results for the same peptide in both versions.

As was already said in chapter 3 above, at each DEEPSAM's iteration, one of a set of \( n \) overlapping sub-ranges of the range \([0.0, t_{\text{max}}]\) of PES smoothing levels is assigned to each one of the \( n \) conformations of the current population. DEEPSAM uses a Levy-Distribution-based method to stochastically generate five different PES smoothing levels from within the assigned sub-range, generating a corresponding set of five DEMSA mutation operators which may be of four possible kinds: *nwnwpss*, *nwsapss*, *sanwpss*, or *sasapss*.

DEEPSAM's results were compared with (a) the results got from running an *nwsapss* DEMSA operator (implemented combining TINKER's *anneal* program with TINKER’s *pss* program), (b)
the results got from running TINKER’s *anneal* program (an MDSA implementation), and (c) the results got from running TINKER’s *pss* program (a DEM implementation).

In order to try to do a meaningful comparison, the initial structures were the same as those used in our DEEPSAM’s runs, and parameters were assigned values in the hope that those component methods will run longer and will get deep minima. Comparison is presented in tables 15 and 16.

**Table 15: Comparing Energy (in Kcal/Mol) among DEEPSAM and its Component Methods.**

<table>
<thead>
<tr>
<th>Peptide</th>
<th>DEEPSAM minimum</th>
<th><em>anneal</em> minimum</th>
<th><em>Pss</em> minimum</th>
<th><em>nwsapss</em> minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>cyc-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>-164.948</td>
<td>-153.229</td>
<td>-145.818</td>
<td>-113.365</td>
</tr>
</tbody>
</table>

**Table 16: Comparing Run-time (in seconds) among DEEPSAM and its Component Methods.**

<table>
<thead>
<tr>
<th>Peptide</th>
<th>DEEPSAM minimum</th>
<th><em>anneal</em> minimum</th>
<th><em>Pss</em> minimum</th>
<th><em>nwsapss</em> minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>cyc-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>3240</td>
<td>7200</td>
<td>1500</td>
<td>2220</td>
</tr>
<tr>
<td>non-cyc-GLY-HIS-GLY-ALA-TYR-GLY</td>
<td>4062</td>
<td>4500</td>
<td>2100</td>
<td>4080</td>
</tr>
</tbody>
</table>

It is important to note that even if the data presented in tables 15 and 16 show that DEEPSAM’s results are better than those of the component methods checked against it, and that we have reasons to conclude that DEEPSAM is indeed an effective structure prediction computational procedure, these data should be considered preliminary comparison results that need to be part of an additional wider study to be pursued in the future.

**I.c. Conclusions**

We tested DEEPSAM on seven small peptide structures, five of them in cyclic form and the other two in non-cyclic form, which we took as our Proof of Concept. For each one of them, using TINKER’s implementation of the OPLSAA force field parameterization, we got a conformation with energy at least as deep as the deepest minimum found by a relatively extensive *scan* conformational search of the corresponding PES (in only one case, DEEPSAM did not find a
minimum as deep as that found by scan). In all cases, DEEPSAM found an energetically deeper conformation if the scan minimum was taken as the initial conformation of the run. This suggests that it will be beneficial that the initial conformation of a DEEPSAM run will be a good local minimum, or an ensemble of good local minima (like those generated by the low-resolution step of Rosetta), for which we may expect to get better results. We also found that the structures got by DEEPSAM are in good agreement with experimental data, either in calculations without solvent or with the GBSA implicit solvent model. Obviously, agreement between the algorithm results and experiment depends on the validity of the force field, and also on the environment effects that affect X-RAY and NMR experiments. This good agreement supports not only the results got by the algorithm, but also the acceptable validity of the force field used for those specific peptides. It is very important to note that our method ends not only with the deepest minimum found, but with an ensemble of deep lying minima structures.

We also found that our algorithm seems to be better than its component algorithms, what tells us that combining DEM with MDSA together with EP, is a successful and effective idea.

From the data presented above, we can see that our algorithm seems to be an effective structure predictor in spite of the fact that the population size used in our runs was small (five conformations per population, in most of the cases), requiring relatively modest computer resources.
II. Predicting Structure in the Gas Phase – Ubiquitin +13

II.a. Background

Biomolecules function in aqueous environments. For this reason, it is appropriate to study them in solution. It also seems appropriate to mimic the structure of such biomolecule-solvent systems and their dynamics using explicit solvent representations or implicit solvent models\textsuperscript{101}. Even if these computational approaches are reasonable representations of the environments of such systems, the following challenges at hand: (a) it is difficult to identify how intra-molecular interactions and surface-solvent interactions influence on the folding process\textsuperscript{102-104}; (b) this kind of calculations may be prohibitively expensive in their requirements of run time and computing resources. Less complex scenarios have been made possible in order to allow a deeper understanding of intra-molecular interactions inside a protein or peptide – proteins in the gas phase\textsuperscript{102-105} (in-vacuum). This kind of scenarios also provides the possibility of studying better how bio-molecules and their solvent interact by gradually re-inserting solvent molecules into the system. Mass spectrometry (MS) has allowed the study of properties of proteins in the absence of water. Ion-Mobility Spectrometry\textsuperscript{105} (IMS) techniques are used to study the structure of protein ions (protonated protein molecules) in vacuum – in the gas phase. The ion’s drift time ($t_D$) – the time the ion spends traveling through the buffer gas (helium, in most cases) across the drift tube under a weak electric field – expresses the ion’s mobility, which depends on its orientation averaged collision cross section with the buffer gas.

The measured drift time $t_D$ determines the ion's cross section as given in eq \textsuperscript{20}\textsuperscript{103}

$$Q_{avg}^{1,1} = \frac{(18\pi)^{\frac{1}{2}}}{16} \left[ \frac{1}{m_p} + \frac{1}{m} \right]^{\frac{1}{2}} \frac{ze}{(k_B T)^{\frac{1}{2}}} \frac{t_D V}{L^2 \rho}$$  \hspace{1cm} (20)$$

where $m$ and $m_p$ are the masses of the ion and buffer gas, $ze$ is the charge, $\rho$ is the buffer gas number density, $L$ is the drift tube length and $V$ is the voltage across the drift tube. Based on the measured ion's collision cross section values, the ion's structure is determined comparing them to those collision cross section values predicted from model structures obtained by Molecular
Dynamics (MD) simulations, global energy minimization methods, X-ray crystallography or by NMR techniques.

The "Ubiquitin system", as it is called by A. Varshavsky, is a protein with very important physiological functions in all eukaryotes. Vijay-Kunar et al provide a very rich description of structural (and also some physiological) details of this important protein; here we will relate to only a few of them. Ubiquitin (see Fig. 24) is a relatively short protein (76 residues) with a very strong structural stability over a wide range of pH and temperature values. Reasons for its stability are: pronounced hydrophobic core formed by residues from the α-helices and the β-sheets, hydrogen-bonding interactions between the helices and the turns, hydrogen bonding that involves a symmetric arrangement of the two helices and two reverse turns.

Figure 24. Ubiquitin PDB X-ray Structure (PDB 1UBQ)

Elad Segev et al studied a protonated state of the Ubiquitin protein, in the gas phase - the Ubiquitin +13 protein ion. In the next section we describe our use of DEEPSAM as structure predictor in the gas phase.
II.b. Analysis and Results

Elad Segev et al\textsuperscript{108} describe a series of solvent-free MD simulations for Ubiquitin +13 protein ion unfolding pathways as they occur in Electro-Spray-Ionization (ESI) MS experiments, reporting a "striking accord" between the 2137 Å\textsuperscript{2} cross section computed for the final elongated structure obtained by those MD simulations and the ESI-IMS experimental 2115 Å\textsuperscript{2} cross section (see Fig. 25). Those MD simulations were started from a partially unfolded +13 charge state conformation called "A state", which is stable at low pH methanol-water solution as experimentally described\textsuperscript{108}.

\textbf{Table 17:} Ubiquitin +13 charge state protonation

<table>
<thead>
<tr>
<th>AA</th>
<th>Atom Name</th>
<th>Original Partial Charge</th>
<th>New Partial Charge</th>
<th>Atom Charge Change</th>
<th>Original AA Partial Charge</th>
<th>New AA Partial Charge</th>
<th>AA Charge Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLU16</td>
<td>OE1</td>
<td>-0.8188</td>
<td>-0.3188</td>
<td>+0.5000</td>
<td>-1.0000</td>
<td>0.0000</td>
<td>+1.0000</td>
</tr>
<tr>
<td></td>
<td>OE2</td>
<td>-0.8188</td>
<td>-0.3188</td>
<td>+0.5000</td>
<td>-1.0000</td>
<td>0.0000</td>
<td>+1.0000</td>
</tr>
<tr>
<td>GLU18</td>
<td>OE1</td>
<td>-0.8188</td>
<td>-0.3188</td>
<td>+0.5000</td>
<td>-1.0000</td>
<td>0.0000</td>
<td>+1.0000</td>
</tr>
<tr>
<td></td>
<td>OE2</td>
<td>-0.8188</td>
<td>-0.3188</td>
<td>+0.5000</td>
<td>-1.0000</td>
<td>0.0000</td>
<td>+1.0000</td>
</tr>
<tr>
<td>PRO19</td>
<td>1HB</td>
<td>0.0253</td>
<td>0.1953</td>
<td>0.1700</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2HB</td>
<td>0.0253</td>
<td>0.1953</td>
<td>0.1700</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1HG</td>
<td>0.0213</td>
<td>0.1913</td>
<td>0.1700</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2HG</td>
<td>0.0213</td>
<td>0.1913</td>
<td>0.1700</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1HD</td>
<td>0.0391</td>
<td>0.2091</td>
<td>0.1700</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2HD</td>
<td>0.0391</td>
<td>0.1891</td>
<td>0.1500</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASP21</td>
<td>OD1</td>
<td>-0.8014</td>
<td>-0.3014</td>
<td>0.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OD1</td>
<td>-0.8014</td>
<td>-0.3014</td>
<td>0.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLU24</td>
<td>OE1</td>
<td>-0.8188</td>
<td>-0.3188</td>
<td>+0.5000</td>
<td>-1.0000</td>
<td>0.0000</td>
<td>+1.0000</td>
</tr>
<tr>
<td></td>
<td>OE2</td>
<td>-0.8188</td>
<td>-0.3188</td>
<td>+0.5000</td>
<td>-1.0000</td>
<td>0.0000</td>
<td>+1.0000</td>
</tr>
<tr>
<td>LYS29</td>
<td>1HZ</td>
<td>0.3400</td>
<td>0.0100</td>
<td>-0.3300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2HZ</td>
<td>0.3400</td>
<td>0.0050</td>
<td>-0.3350</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3HZ</td>
<td>0.3400</td>
<td>0.0050</td>
<td>-0.3350</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASP32</td>
<td>OD1</td>
<td>-0.8014</td>
<td>-0.3014</td>
<td>0.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OD1</td>
<td>-0.8014</td>
<td>-0.3014</td>
<td>0.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLU34</td>
<td>OE1</td>
<td>-0.8188</td>
<td>-0.3188</td>
<td>+0.5000</td>
<td>-1.0000</td>
<td>0.0000</td>
<td>+1.0000</td>
</tr>
<tr>
<td></td>
<td>OE2</td>
<td>-0.8188</td>
<td>-0.3188</td>
<td>+0.5000</td>
<td>-1.0000</td>
<td>0.0000</td>
<td>+1.0000</td>
</tr>
<tr>
<td>ASP39</td>
<td>OD1</td>
<td>-0.8014</td>
<td>-0.3014</td>
<td>0.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OD1</td>
<td>-0.8014</td>
<td>-0.3014</td>
<td>0.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLU51</td>
<td>OE1</td>
<td>-0.8188</td>
<td>-0.3188</td>
<td>+0.5000</td>
<td>-1.0000</td>
<td>0.0000</td>
<td>+1.0000</td>
</tr>
<tr>
<td></td>
<td>OE2</td>
<td>-0.8188</td>
<td>-0.3188</td>
<td>+0.5000</td>
<td>-1.0000</td>
<td>0.0000</td>
<td>+1.0000</td>
</tr>
<tr>
<td>ASP52</td>
<td>OD1</td>
<td>-0.8014</td>
<td>-0.3014</td>
<td>0.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OD1</td>
<td>-0.8014</td>
<td>-0.3014</td>
<td>0.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASP58</td>
<td>OD1</td>
<td>-0.8014</td>
<td>-0.3014</td>
<td>0.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OD1</td>
<td>-0.8014</td>
<td>-0.3014</td>
<td>0.5000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLU64</td>
<td>OE1</td>
<td>-0.8188</td>
<td>-0.3188</td>
<td>+0.5000</td>
<td>-1.0000</td>
<td>0.0000</td>
<td>+1.0000</td>
</tr>
<tr>
<td></td>
<td>OE2</td>
<td>-0.8188</td>
<td>-0.3188</td>
<td>+0.5000</td>
<td>-1.0000</td>
<td>0.0000</td>
<td>+1.0000</td>
</tr>
<tr>
<td>GLY76</td>
<td>O</td>
<td>-0.7855</td>
<td>-0.2855</td>
<td>+0.5000</td>
<td>-1.0000</td>
<td>0.0000</td>
<td>+1.0000</td>
</tr>
<tr>
<td></td>
<td>OXT</td>
<td>-0.7855</td>
<td>-0.2855</td>
<td>+0.5000</td>
<td>-1.0000</td>
<td>0.0000</td>
<td>+1.0000</td>
</tr>
</tbody>
</table>

Original Total Partial Charge \(+1.0000\)
New Total Partial Charge \(+13.0000\)
We decided to try DEEPSAM as predictor of Ubiquitin +13 protein ion's unfolded structure in the gas phase, starting its run from Ubiquitin's PDB X-ray structure (see Figure 24) with the same protonation used in Elad Segev's MD simulations (see Table 17). The TINKER's implementation of the amber98 Force Field parameterization was used in this run. We run DEEPSAM with a population of size 5; the best-so-far final population had three different conformations, the best of which can be seen in Fig. 26.

Dr. Thomas Wyttenbach, from Prof. M. T. Bower's group at UCSB did the cross sections' calculations for us; the results were as follows: (a) 2140 Å² - DEEPSAM's global minimum, (b) 2179 Å² - the second best, and (c) 2168 Å² - the third. Thomas concluded that taking into account that the smallest cross section value differs from the largest only by less than 2%, which is below the average instruments’ error, those structures are experimentally indistinguishable (see Fig 27).
II.c. Conclusions

Taking into account Thomas' conclusions, and that the cross section value of Elad Segev's final structure (2137 Å²) and that of our global minimum structure (2140 Å²) are almost identical, we may conclude that both structures are experimentally indistinguishable.

Thus, this allows us to establish that in the absence of X-ray crystallography and homological data for bio-molecules in the gas phase, a purely \textit{ab-initio} structure predictor like DEEPSAM has a very important role to play in the context of bio-molecules in vacuum and as generator of \textit{model structures} to be used for comparing between experimentally obtained cross sections and the cross sections computed from DEEPSAM-predicted structures.
III. About Force Fields Reliability – Crambin

III.a. Background

Convinced that DEEPSAM is a good structure predictor, we decided to go a step further. Our purpose at this stage was to try it as a tool for testing the reliability of force fields relative to a specific molecular system. Taking into account that DEEPSAM's IPG algorithm generates an initial population of conformations far enough from each other and from the initial conformation, the plan to achieve our goal was as follows: (a) choose a protein for which its structure is already known; (b) choose three well known force fields; (c) for each one of those three force fields, run DEEPSAM upon the chosen protein, and check if DEEPSAM is able to "return" to the original structure. Comparing the final structures found using those three different force fields; we will be able to compare the appropriateness (or reliability) of those force fields to the test protein.

The three force fields we chose were AMBER99, CHARMM27, and OPLSAA. The protein chosen by us was *Crambin*, one of the most studied Thionins. Thionins are a family of relatively small strongly homologous polypeptides that share several structural characteristics: (a) they are tightly compact because of three to four disulfide bridges between pairs of Cysteines; (b) hydrophobic and hydrophilic amino acids can be found in similar places in their sequences; (c) their secondary structures are found in similar places in their sequences; and (d) their tertiary structures have essentially the same three-dimensional shape. Thionins can be found on leaves and seeds of plants, and most of them act as effective plant defensos. Thionins have been of great interest because of their relatively small length, their interesting biological function, and their shared structural characteristics.

*Crambin* is a thionin whose biological function is not known yet. It is 46-residues long and has the following characteristics: (a) very low percentage of charged residues; (b) strong hydrophobicity; (c) 46% of its residues are part of its two α-helices; (d) 17% of its residues are part of its β-sheet structure composed of two short anti-parallel β-strands; (e) the other residues are extended and random coil turns connecting among the α-helix and β-sheet structures; (f) three
neutral disulfide bridges connect between \( \beta \)-strands and the two \( \alpha \)-helices. See Figure 28.

Figure 28. Crambin PDB X-ray Structure (1CRN)

As we already said in the Introduction, the Force Field \(^{20,116,117}\) (FF) of a molecular structure – its Potential Energy Surface (PES) – is a function expressing the potential energy interactions inside the molecule. Also, it was emphasized that the algorithm presented here is a purely \textit{ab-initio} protein structure prediction method whose fundamental assumption is that a FF for the specific molecule is \textit{available} and it is \textit{correct}. Based on this assumption, we said that in most cases the FF's global minimum \textit{is} the correct structure of the molecule at sufficiently low temperatures (\( T \rightarrow 0 \text{ K} \)). We also said in the Introduction that the \textit{parameterization} of a Molecular Mechanics Force Field (MMFF) determines how good it is as an approximation of the \textit{correct} FF; that is, how \textit{reliable} it is to allow a structure prediction to be closer to the \textit{correct} one. Here we will show that DEEPSAM may be used as a tool to check an MMFF reliability.

\textbf{III.b. Analysis and Results}

We used Crambin's PDB X-ray structure as the initial conformation of our three DEEPSAM runs using TINKER's implementation of the parameterizations of our three chosen MMFFs (AMBER99, CHARMM27 and OPLSAA). In what follows, we present and compare the results of our three DEEPSAM runs.
<table>
<thead>
<tr>
<th>Minimized Initial Structure</th>
<th>Global Minimum Structure</th>
<th>RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1164.2968 Kcal/Mol</td>
<td>-1221.3176 Kcal/Mol</td>
<td>4.8244 Å</td>
</tr>
</tbody>
</table>

**Figure 29a:** results with AMBER99

---

MaxSub Alignment: **TCCPSI**VAR**S**FN**V**CRL**P**GT**P**EAIC**A**TY**T**GC**I**I**P**GATC**P**GD**Y**AN**§**
MaxSub Pairs = 29   MaxSub = 0.510   MaxSub RMSD = 1.087

**Figure 29b.** Superposition initial and final structures (AMBER99)
(the initial structure is in red; the computed final structure is in blue)

§ the aligned residues are in blue.
Minimized Initial Structure
-608.9103 Kcal/Mol

Global Minimum Structure
-689.4586 Kcal/Mol

RMSD
4.3804 Å

**Figure 30a:** results with CHARMM27

MaxSub Alignment: TTCCPSIVARSNFVNVCRLPGTPEAICATYTGCIIPGATC PGDYAN
MaxSub Pairs = 33  MaxSub Rank = 0.575  MaxSub RMSD = 1.132

**Figure 30b.** Superposition of initial and final structures (CHARMM27)
(the initial structure is in red; the computed final structure is in blue)
The results show that (a) the initial structures of CHARMM27 and OPLSAA were very similar; (b) all final structures, calculated in every one of the force fields, are energetically deeper than the corresponding initial structure; (c) the final structures are significantly different than the corresponding initial structures, especially in the case of OPLSAA; (c.1) with Amber99 and
Charmm27, our algorithm succeeded to restore both \( \alpha \)-helices; (c.2) with OPLSAA, our algorithm succeeded to restore only one \( \alpha \)-helix and no \( \beta \)-sheet; (c.3) with Amber99, our algorithm succeeded to restore only a small portion of the anti-parallel \( \beta \)-sheet; (c.4) with Charmm27, our algorithm succeeded to restore all the anti-parallel \( \beta \)-sheet, but a \( \beta \)-strand which did not exist in the initial structure was generated; (d) for each force field, we can see that the final structure is significantly different than the other final structures.

**III.c. Conclusions**

Thus, from the test done we conclude that DEEPSAM may be used as *force field* tester: (a) all the three force fields were found **significantly inaccurate**; (b) the CHARMM27 was shown to be the most appropriate for the test molecule - see that in this case the lowest RMSD, the biggest MaxSub and the highest MaxSub rank were obtained.
5 – Concluding Remarks

In this Thesis I presented a new hybrid evolutionary algorithm named DEEPSAM (Diffusion Equation Evolutionary Programming Simulated Annealing Method), which makes effective integrative use of the advantages of three well-known optimization methods: Evolutionary Programming (EP), Simulated Annealing (SA) and the Diffusion Equation Method (DEM).

DEEPSAM was tested in three different scenarios:

1. Seven small peptide structures (part of them cyclic and part of them non-cyclic): Very good agreement with experimental data was shown. It is worth noting that because of the fact that DEEPSAM is a purely *ab-initio* method, its structure predictions necessarily depend on the validity of the force field in use.

2. A protonated state of the Ubiquitin protein +13 in the Gas Phase: the structure found by DEEPSAM was in agreement with experimental data – according to cross section values calculations, the structures found by DEEPSAM and those found experimentally were indistinguishable.

3. Three well-known force field parameterizations (FF) were tested for reliability against a test-protein, Crambin. Running DEEPSAM upon Crambin, we were able to decide which of the three FFs was the better model to describe the PES of our test protein.

It is worth noting that DEEPSAM seems to be time and space efficient because of two reasons: populations of small size (in most cases, only five conformations) were enough to get these very good results, and the parallel design of the algorithm contributed to its time efficiency.

Following the success of the DEEPSAM's algorithm, future work is already underway: continuation of code development and cooperation with other groups for DEEPSAM's integration into GAMESS (Marc Gordon), TINKER (Jay W. Ponder), Rosetta. (Ora Furman – David Baker's group) All this is very encouraging for future applications: peptides and proteins in solution and
in the Gas Phase, Sacharides, Nucleic Acids, as well as the use of low resolution structures from Rosetta as initial structures for DEEPSAM.

In conclusion, a very promising algorithm is now available for finding the structure of biomolecules – (a) it can be expected to be a good structure predictor, (b) due to its structure prediction capabilities, it has a very important role to play in the Gas Phase context, where there is neither X-ray crystallography data nor homologous data for bio-molecules, and (c) it can be used as a force field tester against structural data.
6 – References


Mackerell, A.D.Jr.: Empirical Force Fields for Biological Macromolecules: Overview and
## תוכן עניינים

<table>
<thead>
<tr>
<th>מספר</th>
<th>פרק</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>מבוא</td>
</tr>
<tr>
<td>2</td>
<td>מטרות המחקר</td>
</tr>
<tr>
<td>3</td>
<td>השיטה</td>
</tr>
<tr>
<td>4</td>
<td>תוצאת</td>
</tr>
<tr>
<td>5</td>
<td>הערות סיכום</td>
</tr>
<tr>
<td>6</td>
<td>ביבליוגרפיה</td>
</tr>
</tbody>
</table>

### פרק 1-

<table>
<thead>
<tr>
<th>מספר</th>
<th>ענין</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>מבוא</td>
</tr>
</tbody>
</table>

### פרק 2-

<table>
<thead>
<tr>
<th>מספר</th>
<th>ענין</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>מטרות המחקר</td>
</tr>
</tbody>
</table>

### פרק 3-

<table>
<thead>
<tr>
<th>מספר</th>
<th>ענין</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>השיטה</td>
</tr>
</tbody>
</table>

### פרק 4-

<table>
<thead>
<tr>
<th>מספר</th>
<th>ענין</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>תוצאת</td>
</tr>
</tbody>
</table>

### פרק 5-

<table>
<thead>
<tr>
<th>מספר</th>
<th>ענין</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>הערות סיכום</td>
</tr>
</tbody>
</table>

### ביבליוגרפיה

<table>
<thead>
<tr>
<th>מספר</th>
<th>ענין</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>ביבליוגרפיה</td>
</tr>
</tbody>
</table>
The DEEPSAM algorithm was tested in three modes: (a) DEEPSAM analyzed each sequence with a unique scheme, allowing the identification of six non-cyclical peptide runs on a set of seven peptides. The integrated algorithm found only one exception among the examples. Amino acids in lengths of six to nine were more intensively searched when compared to a brute force search method.

When presented with a non-repeated structure, the potential for reliability was determined. If the initial structure was not a structure at all, we could expect better performance of DEEPSAM. For example, when DEEPSAM is applied to the Ubiquitin protein, it is expected to perform better due to the use of amber98. DEEPSAM was able to find the same structure as the structure found using dynamics. The result of this experiment confirmed the outcome of the DEEPSAM algorithm.

For proteins of a size of 13 or larger, Ubiquitin, the DEEPSAM algorithm performed well. For smaller proteins, such as Crambin, typically found in the thionin family, DEEPSAM was able to find the potential structure. It is notable that with small populations, the algorithm contributed to the efficiency of running parallel computing. The results showed the potential of DEEPSAM as a tool for testing force fields.

In conclusion, the DEEPSAM algorithm guarantees a relatively short computing time, suitable for biological molecules.
تلكיר

הוגשו המשיפיעים יוצרי בתקנון המולקולות הביולוגיות והם המבוך שללה, המחקר הזה התמקד בהע-Pacific וולקזד בפלנקטונים בהנומרו של תניי אירונגרמיים. bulundטטרוזוט מוסקף נמוך, אשר לא תומי. המבוך באגרניי החופשית המינימואל מתאימים למינימום הגלובאלי של השדה החום (FF, לחלך, השטח הפוטנציאלי, של פנס שיווי), אם השטח הפוטנציאלי של המינימום המינימלי מתאים למינימום הגלובאלי של השדה החום. אם הFF-ה שלח, T → 0 K.

בעיית ניבוי המבוך האטומי (ואコミ) של פולי-פיופיס כשטחת הרץ של מינימום הגלובאלי של המבוך האטומי של פולי-

הלוע לבעיית מציאת המינימום הגלובאלי, אך היאเบינה uname הקטן ב였다 האקﺳפטואוניאלי: יועד לשמש המומר של מינימום של הפוטנציאל של פוליפונטיזיאל. פונקציית הפוטנציאל של פולי המוערך של מינימום של מספר-

לעות ממבר המבוך שלח, פוליפונטיזיאל. פונקציית הפוטנציאל של פולי המוערך של מינימום של מספר-

שיטות ab-initio על הנחת הנכונות של שדות כח הן שיטות שנתמכות אך ורק, על ניבוי מבוך של חלבון.

הגישה במחקר זה חותרת לבניית שיטת (אובלי פונקציית פוטנציאל) DEM (Diffusion Equation Method) שיטה ליה綜合יו פוליפונטיזיאל. הפוליפונטיזיאל שואג במחזור של כל החומרים בצורת תופונקצייה דיפוזיה של תעודה מגשימית, מ-DEM ו-SA, שתי שיטותールית במישון באמצעות DEM וה-SA שיאווסט מינימיזציה, לשני יהLOUR dem hybridizes, EP (Evolutionary Programming) ו-SA (Simulated Annealing Method).

באמצעות פישות שטוחות פוליפונטיזיאל DEM שיטה ה-DEM של תורת לממדות מרחב חפוזות. שיטות פוליפונטיזיאל DEM שיטה ה-DEM של תורת לממדות מרחב חפוזות. שיטות פוליפונטיזיאל DEM של תורת לממדות מרחב חפוזות. שיטות פוליפונטיזיאל DEM של תורת לממדות מרחב חפוזות. שיטות פוליפונטיזיאל DEM של תורת לממדות מרחב חפוזות. שיטות פוליפונטיזיאל DEM של תורת לממדות מרחב חפוזות. שיטות פוליפונטיזיאל DEM של תורת לממדות מרחב חפוזות. שיטות פוליפונטיזיאל DEM של תורת לממדות מרחב חפוזות. שיטות פוליפונטיזיאל DEM של תורת לממדות מרחב חפוזות. שיטות פוליפונטיזיאל DEM של תורת לממדות מרחב חפוזות. שיטות פוליפונטיזיאל DEM של תורת לממדות מרחב חפוזות. שיטות פוליפונטיזיאל DEM של תורת לממדות מרחב חפוזות. שיטות פוליפונטיזיאל DEM של תורת לממדות מרחב חפוזות. שיטות פוליפונטיזיאל DEM של תורת לממדות מרחב חפוזות. שיטות פוליפונטיזיאל DEM של תורת לממדות מרחב חפוזות. שיטות פוליפונטיזיאל DEM של תורת לממדות מרחב חפוזות. שיטות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של תicerca לשטוחות פוליפונטיזיאל DEM של ת التعا 넘어 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת שהור 이상 הנעוצמה בחרת嗝ה
הקדשה

כבוד ראשון להקדיש עבודה זו
לזכר של חורי האהובים

פרופ' ד"ר גלוסטיני ז"ל
דר ציפורה (מולח) בכוכבי-גולדהטיני ז"ל
אני מודה לחברי הסגל של המחלקה למדעי המחשב של בית ספר גבוה לטכנולוגיה-מכן, ובנוסף, על מח好きなי המחלקה לתנור טויטו, החיים דייןר, רפי יחזקאל, ד"ר מיכאל דרייפוסר, עזרא דשטמר, הקולגות שלי: ד"ר רפי יוחיאלי, ד"ר עקיב הנחה-קרנר, ו.tele רחאל שטולמן.


נתיחה והי לי ליצד.

לבסוף, אני רוצה להודות לבארא עולם שאפשר לי להגיעה لأنן דרכ' התachat ביתי:

ברוך אתה ו酩שו בולעל והיהי כל יום ויום לתוך זה.
שלמי תודה

ארוית, כל העונג של הודות למפתיה, פורפ', ברי גרבר, על כל התמיכה וה美媒ד, על אן ספור שיתוף
ודיכים שקימונים בשמוך כל הסגנונות האול, שוインターハוק פלאטשמאף אנטי-קוקואלי, בברבע
ענשו וברועו הצלחת: במחצית מ裁定, בצבעי ביכר, על התוודעות בלו מונח פלאטשמאף של שיתוף.
לברנוני זהי ההכלה וה الرحمن את העקת הצעה, עז מסייס מענה, ברבע
הנכם. כלisiert העיתון והוסיכלים, המחקר הזה יכל לגדוש לסופו המוצלח.

תודה, תודה רבה רבה, תודהصبحันו, אנחנו מקריב, לב!! תודה מכל!!

ברצוני לוחות להחלף ושיקוף לעvoie המחבר, די"ארק פרדואר,'על ישיתאים ואותו, או יבי-ברועו
התחזית למון האלגוריתמים שטמעו פמא התתעה, אריק זהי המחבר המשנה על-רשבים, שלי,
שיאית כימית ידוע על הרבח מתבעות שאיתן התמדות, לארץ עם קרא וויר
على חלום נרחבות של התויה האירוח.

אני מודע לפורפ', עמיור גולדברג על הענה להכ乸 מחסנית碎片ים צייכליים מאדם מספריים,
כאמצויות לאימוץ התוכן האלגוריתם המובא שאך.

אני מודע לכל תיבר, קובץ המחבר של בינ על כל התמכיות ויביח. אני מודע מבויה של "ר"אלעד
שגב, את האלגוריתמים בוביצי המחבר של, על ישית הקשעה חנקן המחבר-זואופ, של
על קריינות ויזואליים על הפסקת תיבי זה, ונינו מקריה זה. אני מודע="ארק
בראון, את הפוסט-דוקטוריים של בינ, שחייק על אט שפת ההכנת חוסון, שורק חוק ש
DEEPSAM כותב ביה.

אני מודע לליידיתBush, "ר" סובב פלדמן, על הערתיה על גרסה ראשונית של התויה האירוח.
איני מודע ל"ד"ארק פרדואר, על הعراضת על גרסה ראשונית של תולח מציחי התויה האירוח. אני ובモaza ל
ולו"ד"ר שליש על סערת יב으며 אסימיות הצייכליים מפורמט CIO וסערת PDB.
PDB-ל CIO-ל, RASMOL, רביעי המרה מ
הל ברשימים, ומר אוזנים דריבב, מבית ג'ון ש
לו מודע "לא" פרדאור, את הפרוייקט של שולחן TINKER, על עטיה, ובכלא שעישור החל מ
ה aş ואף למודע LA, פרדואר, ברורה, מפתית לתוכני, לע עטה ובכלא המ
שקור לתוכני התויה.

אני מודע ל"ג" גלורה, מיכאל שמחא, מנלי המצרך של תורני יב, לע עטיה ובכלא של טקלה
ומודע לאלאס, שמלאי המצרך של תורני יב, על עטיה ובכלא של טקלה
בחפירות המוחטיבים על המצרך. אני ובモaza ל, אלמנו האדומית של המצרך, על
על עטיה המיתית בכלי עניין.
עבודה זו נעשתה בהדריכה של:

פרופ' יוסי בני גרבר
DEEPSAM

(Diffusion Equation Evolutionary Programming Simulated Annealing Method)

אלגוריתם אבולוציוני משולב

למציאת מינימות תומכות ביותר של משטחים פוטנציאלים:

גישא ואפליקציות

תyor לשומ קבלת תואר דוקטור לפילוסופיה

מאת

משה גולדשטיין

הוגש לסינט האוניברסיטה העברית בירושלים

אפריל 2009 – ניס תשל"ט