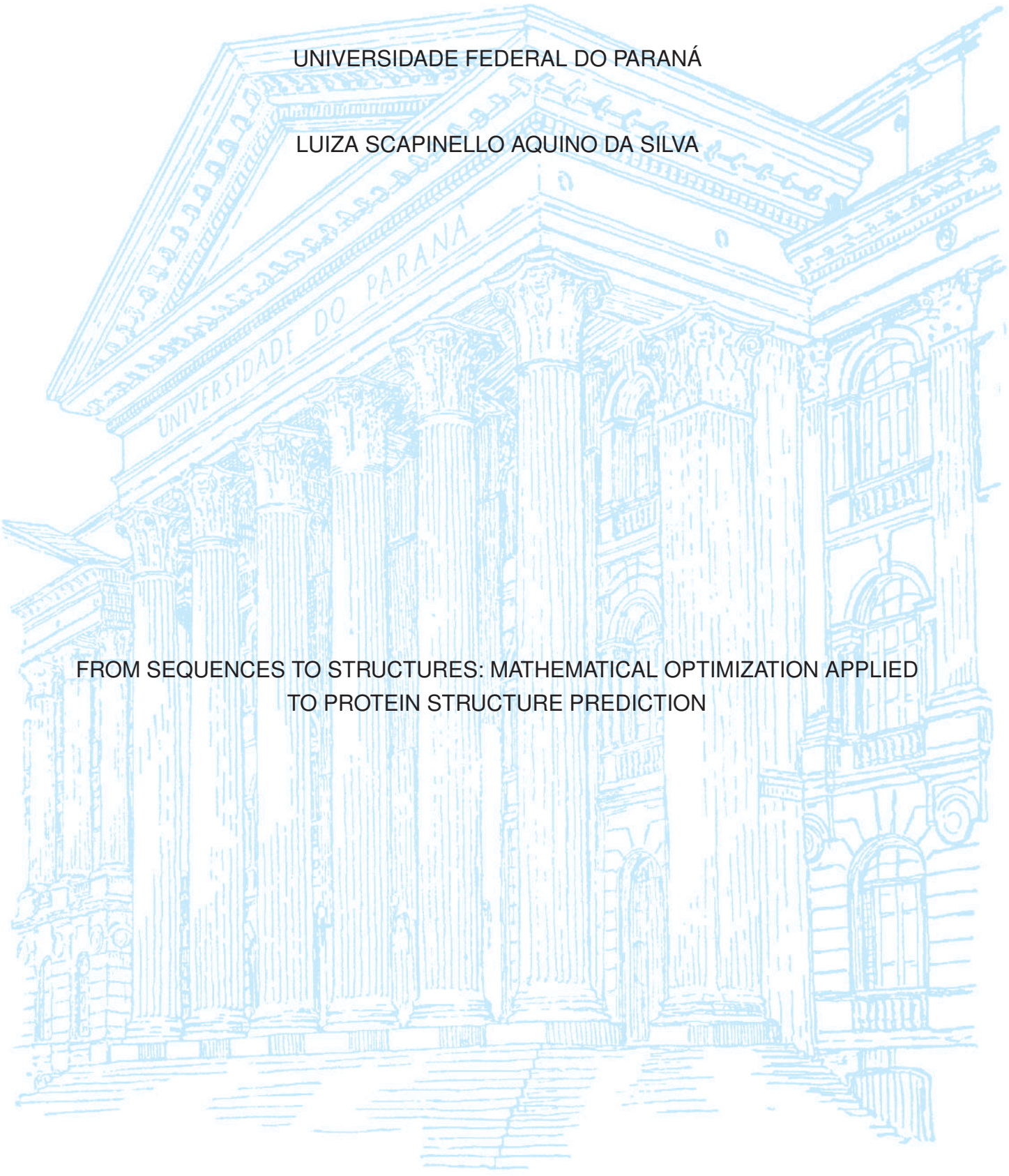




UNIVERSIDADE FEDERAL DO PARANÁ

LUIZA SCAPINELLO AQUINO DA SILVA

FROM SEQUENCES TO STRUCTURES: MATHEMATICAL OPTIMIZATION APPLIED
TO PROTEIN STRUCTURE PREDICTION

CURITIBA

2026

LUIZA SCAPINELLO AQUINO DA SILVA

FROM SEQUENCES TO STRUCTURES: MATHEMATICAL OPTIMIZATION APPLIED
TO PROTEIN STRUCTURE PREDICTION

Tese apresentada ao curso de Pós-Graduação
em Engenharia Elétrica, Setor de Tecnologia,
Universidade Federal do Paraná, como requisito
parcial à obtenção do título de Doutor em En-
genharia Elétrica.

Orientador: Dr. Leandro dos Santos Coelho

Coorientador: Dr. Laio Oriel Seman

CURITIBA

2026

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIA E TECNOLOGIA

Silva, Luiza Scapinello Aquino da

From sequences to structures: mathematical optimization applied to protein structure prediction / Luiza Scapinello Aquino da Silva. – Curitiba, 2026.

1 recurso on-line : PDF.

Tese (Doutorado) - Universidade Federal do Paraná, Setor de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica.

Orientador: Leandro dos Santos Coelho

Coorientador: Laio Oriel Seman

1. Proteínas. 2. Otimização matemática. 3. Algoritmos. I. Universidade Federal do Paraná. II. Programa de Pós-Graduação em Engenharia Elétrica. III. Coelho, Leandro dos Santos. IV. Seman, Laio Oriel. V. Título.

Bibliotecário: Douglas Lenon da Silva CRB-9/1892

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação ENGENHARIA ELÉTRICA da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **LUIZA SCAPINELLO AQUINO DA SILVA**, intitulada: **From sequences to structures: Mathematical optimization applied to protein structure prediction**, sob orientação do Prof. Dr. LEANDRO DOS SANTOS COELHO, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutora está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 24 de Fevereiro de 2026.

Assinatura Eletrônica

25/02/2026 21:49:05.0

LEANDRO DOS SANTOS COELHO

Presidente da Banca Examinadora

Assinatura Eletrônica

24/02/2026 18:37:04.0

GIDEON VILLAR LEANDRO

Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica

25/02/2026 21:51:46.0

HAMED YAZDANPANA

Avaliador Externo (UNIVERSIDADE DE SÃO PAULO - USP)

Assinatura Eletrônica

24/02/2026 19:00:38.0

ROBERTO ZANETTI FREIRE

Avaliador Externo (UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ)

For Artemis, the cat, the best companion I could have asked for

ACKNOWLEDGMENTS

This thesis is a reflection of the constant support and appreciation that I received from my family and loved ones during this ambitious academic pursuit. I am very grateful to my parents, who encouraged my curiosity and supported my education efforts from the beginning. I have learned from experience the importance of asking questions and how different points of view can enhance our understanding of complex problems.

My most profound appreciation goes to Professor Dr. Leandro dos Santos Coelho and Dr. Laio Oriel Seman, my Ph.D. advisors and mentors, for their time, effort, and understanding in helping me succeed in my studies. Their vast wisdom and wealth of experience have inspired me throughout my studies. In addition, I would like to thank Dr. Viviana Cocco Mariani for her technical assistance during my research.

"Hope and curiosity about the future seemed better than guarantees. The unknown was always so attractive to me...and still is."

Hedy Lamarr

"The more I study, the more insatiable do I feel my genius for it to be."

Ada Lovelace

RESUMO

As proteínas são componentes fundamentais da vida, atuando como blocos de construção em diversos processos biológicos. Consequentemente, compreender a sua formação é um campo de amplo alcance e relevância. O problema de previsão da estrutura de proteínas (PSP) se baseia no uso da sequência linear dos aminoácidos e em funções matemáticas baseadas em física e biologia para determinar sua estrutura nativa, possuindo esta o mínimo de energia livre para garantir a estabilidade da conformação. À vista disso, a principal contribuição dessa tese de doutorado é a execução de uma abordagem pioneira para PSP baseada em otimização matemática que utiliza o modelo de representação de proteínas AB off-lattice, o qual faz uso das características hidrofílicas dos aminoácidos para classificá-los e calcular suas posições. A formulação quebra o cenário energético de dobramento de proteínas, transformando a questão do PSP convencionalmente não linear em um problema de otimização bilinear (BL), uma vez que a natureza intrinsecamente não convexa do PSP continua sendo uma limitação que desafia a eficácia dos métodos de otimização global. Essa formulação foi obtida através da introdução de variáveis e restrições auxiliares que descrevem a relação complexa entre o espaço conformacional da proteína e seu cenário energético, introduzindo uma decomposição matematicamente estruturada que ainda não havia sido explorada neste contexto. Essa reformulação oferece um novo caminho para codificar padrões de interação, dependências geométricas e contribuições energéticas dentro de uma preposição algébrica tratável, permitindo a aplicação de ferramentas avançadas de programação matemática a um problema biofísico classicamente intratável. Múltiplas abordagens usando heurísticas aparecem na literatura com sucesso variado em relação ao PSP; assim, o problema de otimização BL é comparado às metaheurísticas mais utilizadas para validar o método sugerido. O modelo BL proposto demonstrou notável precisão na identificação das conformações de energia mínima global em um conjunto de dados de referência fornecido pelo Protein Data Bank (PDB). Comparada aos métodos heurísticos tradicionais, a abordagem BL viabilizou soluções exatas, reduzindo a probabilidade de aprisionamento em mínimos locais, com melhor acurácia, especialmente quando testada em proteínas grandes e com custo computacional até cinco vezes menor do que o das metaheurísticas tradicionais. Essa tese destaca o potencial de reformular o problema tradicionalmente não linear de PSP; a transformação bilinear oferece um caminho para metodologias que podem determinar a solução global, desafiando os paradigmas atuais de PSP.

Palavras-chave: Predição de estruturas de proteínas; modelo AB off-lattice; otimização bilinear; algoritmo exato.

ABSTRACT

Proteins are fundamental components of life and act as building blocks in various biological processes. Consequently, understanding their formation is a field of broad scope and relevance. The protein structure prediction (PSP) problem is based on the use of the linear amino acid sequence that constitutes a protein and on mathematical functions based on physics and biology to determine its native structure, which holds the minimum free energy to guarantee the stability of the conformation. This process is essential to clarify the functions and actions of a protein. In light of this, the main contribution of this Doctoral thesis is the implementation of a pioneering approach to PSP based on mathematical optimization that uses the AB off-lattice protein representation model, which exploits the hydrophilic characteristics of amino acids to classify them and calculate their positions. The formulation breaks the intricate energetic scenario of protein folding, reformulating the traditionally nonlinear PSP problem as a bilinear optimization problem, although the inherently highly non-convex nature of PSP continues to pose challenges for global optimization. The BL design was obtained through the establishment of auxiliary variables and constraints that describe the complex relationship between the protein's conformational space and its energy landscape, introducing a mathematically structured decomposition that had not yet been explored in this context. This reformulation provides a novel pathway to encode interaction patterns, geometric dependencies, and energetic contributions within a tractable algebraic preposition, enabling the application of advanced mathematical programming tools to a classically intractable biophysical problem. Multiple approaches using heuristics appear in the literature with varying success in relation to the PSP; thus, the bilinear (BL) optimization problem is compared to the most widely used metaheuristics with the best results in the studied application to validate the suggested method. The proposed BL model demonstrated remarkable accuracy in identifying global minimum energy conformations in a reference dataset provided by the Protein Data Bank (PDB). Compared to traditional heuristic methods, the BL approach provided exact solutions, reducing the likelihood of local minima blockage, with satisfactory performance metrics, i.e., better accuracy, especially when tested on larger proteins (composed of more than 100 amino acids) and with a computational runtime cost up to five times lower than traditional metaheuristics. This thesis highlights the potential to reformulate the traditionally non-linear PSP problem, as bilinear transformation offers a path to methodologies that can determine the global solution, challenging current PSP paradigms.

Key-words: Protein structure prediction; AB off-lattice model; bilinear optimization; exact algorithm.

LIST OF FIGURES

FIGURE 1 – General chemical structure of an amino acid	16
FIGURE 2 – Protein's biological synthesis process	17
FIGURE 3 – Structure of the thesis	24
FIGURE 4 – Types of angles in the AB off-lattice model	30
FIGURE 5 – Example of convex and concave functions	34
FIGURE 6 – Comparison of the original model versus the BL approach	59
FIGURE 7 – Comparison graph of the running time for the smallest proteins . .	67
FIGURE 8 – Comparison graph of the running time for the biggest proteins . .	67
FIGURE 9 – Comparison graph of the RMSD for each protein	76

LIST OF TABLES

TABLE 1 – Essencial amino acids	28
TABLE 2 – Most notorious papers on DI for PSP	49
TABLE 3 – Recent papers on metaheuristics approaches to PSP	54
TABLE 4 – Hiperparameters setup for the compared metaheuristics	65
TABLE 5 – Selected proteins	65
TABLE 6 – Time (seconds) taken to complete the run	66
TABLE 7 – Native energy optimization results	69
TABLE 8 – RMSD optimization results	73
TABLE 9 – Paired t-test results (Best RMSD)	77

LIST OF ABBREVIATIONS AND ACRONYMS

ABC Artificial Bee Colony

AI Artificial Intelligence

ANN Artificial Neural Network

BL Bilinear

CNN Convolutional Neural Network

DE Differential Evolution

DL Deep Learning

EA Evolutionary Algorithm

GA Genetic Algorithm

IMP Interior Point Method

KKT Karush–Kuhn–Tucker

LP Linear Programming

ML Machine Learning

MRFO Manta Ray Foraging Optimization

NLP Nonlinear Programming

NP Non-deterministic Polynomial Time

PDB Protein Data Bank

PSO Particle Swarm Optimization

PSP Protein Structure Prediction

RAM Random-Access Memory

RF Random Forest

RMSD Root-Mean-Square Deviation

SI Swarm Intelligence

SQP Sequential Quadratic Programming

SVM Support Vector Machine

std Standard Deviation

LIST OF SYMBOLS

Angstrom = 100 picometers

A Hydrophobic residue

B Hydrophilic (polar) residue

n Sequence length

θ_i Curvature angle of the i^{th} residue

β_i Torsion angle of the i^{th} residue

ξ_i Bipolar variable of the i^{th} residue

Φ Intramolecular potential energy function

$V_1(\theta_i)$ Potential curvature energy

$V_2(r_{ij}, \xi_i, \xi_j)$ Non-bonded pairwise interaction between the i^{th} and j^{th} residues

r_{ij} Distance between the i^{th} and the j^{th} residues

$C(\xi_i, \xi_j)$ Level of attraction between residues ξ_i and ξ_j

(x_i, y_i, z_i) Cartesian coordinates for a residue i

x Vector that contains the optimization variables for the problem

$f(x)$ Function to be optimized

$h(x)$ Equality constrain

$g(x)$ Inequality constrain

C Convex set

$H(f)(x)$ Hessian matrix

c_{IMP} Column vector of coefficients for the problem

A_{IMP} Matrix of the coefficients of the constraints

b_{IMP} Vector that defines the right-hand side of the constraints

ϕ Barrier Function

$F(x)$ Modified objective function

M Barrier parameter

y Dual variables

sv Slack variables

L Lagrangian function

d Search direction

H_k Hessian approximation of the Lagrangian function

lb Lower bound

ub Upper bound

ro_i Positions of residues i in the original structure

rp_i Positions of residues i in the predicted structure

s Representation of $\sin(\cdot)$ as a transformed BL variable

c Representation of $\cos(\cdot)$ as a transformed BL variable

$\rho_{i,j} = r_{i,j}^2$ Square of the distance r_{ij}

w Bilinear term for $x \cdot y$

CONTENTS

1	INTRODUCTION	15
1.1	MOTIVATION	20
1.2	OBJECTIVES	22
1.3	CONTRIBUTIONS	22
1.4	DOCUMENT STRUCTURE	23
2	PROTEIN STRUCTURE PREDICTION	25
2.1	AB OFF-LATTICE MODEL	27
3	COMPUTATIONAL APPROACHES	33
3.1	NON-LINEAR PROGRAMMING	35
3.1.1	Karush–Kuhn–Tucker conditions	37
3.1.2	Interior Point Method	38
3.1.3	Sequential Quadratic Programming	40
3.2	LINEAR PROGRAMMING RELAXATION	41
3.3	ROOT MEAN SQUARE DEVIATION	44
4	RELATED WORKS	46
4.1	MACHINE AND DEEP LEARNING IN THE PSP PROBLEM	46
4.2	METAHEURISTICS IN THE PSP PROBLEM	50
5	BILINEAR OPTIMIZATION IN PSP	55
5.1	MAIN CODE	60
6	RESULTS AND DISCUSSION	64
7	CONCLUSION	78
7.1	FUTURE WORK	79
7.2	CONTRIBUTIONS	80
	REFERENCES	82
APPENDIX A	MAIN PYTHON CODE	98

1 INTRODUCTION

Proteins are intricate macromolecules that perform numerous essential functions within all known living organisms (Whitford, 2013). They are composed of amino acid chains linked by peptide bonds and exhibit exceptional structural and functional diversity, ranging from fibrous, which are elongated and insoluble for usually structural support, to globular conformations, that are compact and spherical, commonly seen performing dynamic functions (Kurochkina, 2012). This permits them to adapt to a plethora of cellular roles in different environments. As they are the primary constituents of the dry mass of cells, they have an important activity in virtually every biological process, being a central character in the construction, maintenance, protection, and repair of biological tissues (Motmaen *et al.*, 2023).

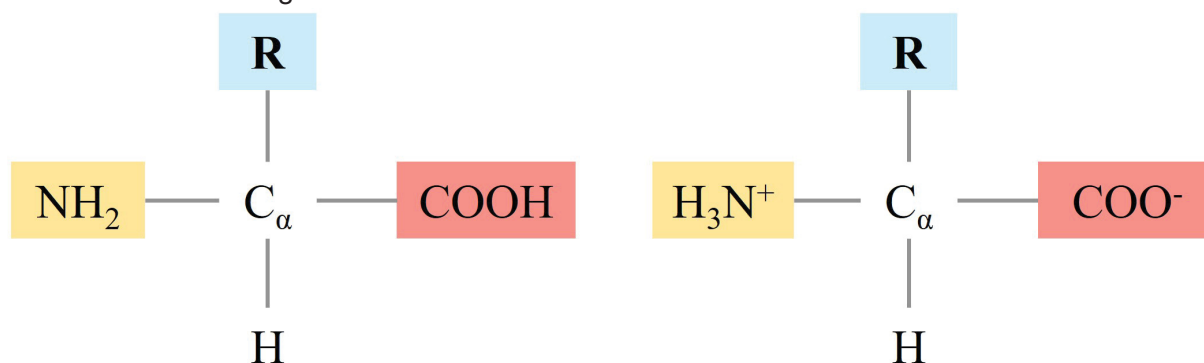
As their conformation varies, so does their function. On one hand, fibrous proteins have a structural role, i.e., they are part of something, for instance, the collagen in tendons and the keratin in the hair. On the other hand, other types of proteins have a dynamic function, acting as molecular machines responsible for catalyzing biochemical reactions, transmitting and interpreting cellular signals, transporting molecules across membranes, and regulating gene expression (Xu; Zhang, 2012). Enzymes enhance metabolic processes with remarkable selectivity and efficiency, whereas transport proteins enable the movement of ions, metabolites, and other molecules essential for cellular homeostasis. Signaling and receptor proteins facilitate communication between and across cells, orchestrating intricate physiological responses (Ranganathan *et al.*, 2023).

During biosynthesis, a protein undergoes a complex folding process in which its one-dimensional amino acid sequence is molded into a specific three-dimensional conformation (Waudby *et al.*, 2019). The information encoded in its linear sequence dictates the spatial arrangement of atoms within the molecule, ultimately generating a well-defined and unique native structure that determines the protein's function. This folding process is governed by a delicate balance of intramolecular forces, including hydrogen bonding, hydrophobic interactions, van der Waals forces, and electrostatic attractions, all of which collectively stabilize the final conformation (Fasman, 2012).

Each amino acid consists of a central carbon atom, known as the α -carbon, to which four distinct groups are attached: an amino group ($-\text{NH}_2$), a carboxyl group ($-\text{COOH}$), a hydrogen atom, and a variable side chain, denoted as the R group (Li *et al.*, 2025). Figure 1 illustrates two general structures of an α -amino acid, one of which represents an amino acid in an aqueous environment. In this representation, the amino acid is in its zwitterionic form (dipolar ion), in which the amino group has an additional

proton ($-\text{NH}_3^+$) and the carboxyl group is negatively ionized ($-\text{COO}^-$). This dual charge distribution gives the molecule both acidic and basic properties, contributing to the overall stability of a protein.

Figure 1 – General chemical structure of an amino acid



SOURCE: The author (2025)

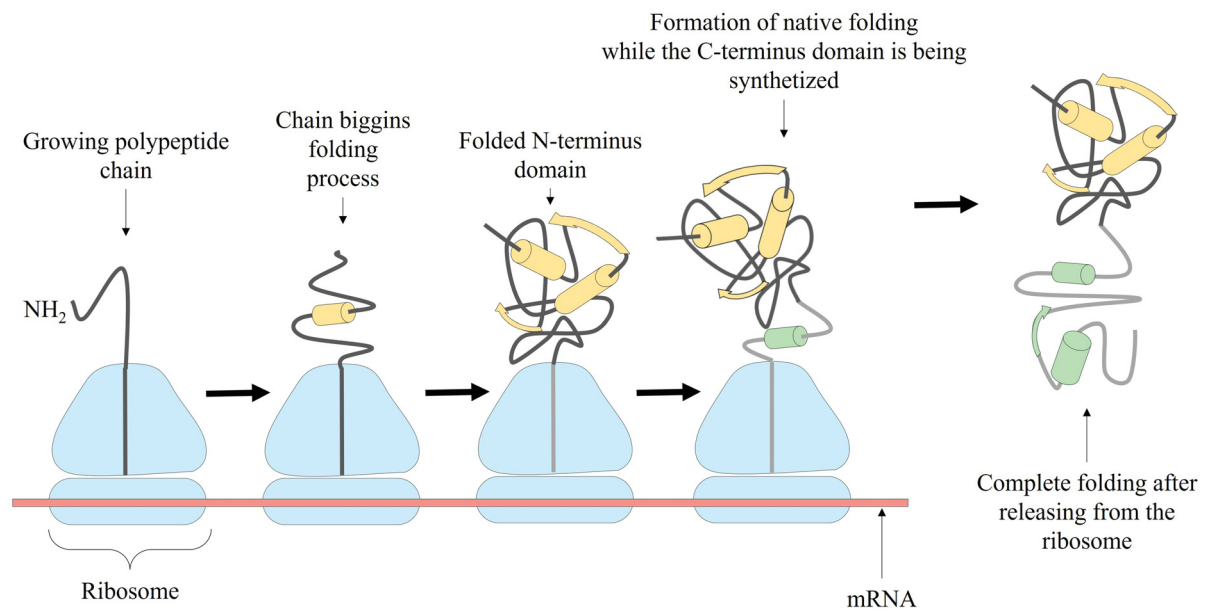
There are twenty known amino acids that form proteins, each characterized by distinct physicochemical properties derived from the composition of its side chain, or R group (Ferina; Daggett, 2019). The amino acids vary in size, shape, charge distribution, polarity, hydrophobicity, and chemical reactivity, characteristics that collectively influence their interactions with each other and the environment. Through the ordered polymerization of these twenty amino acids in varying sequences and combinations, cells synthesize an immense diversity of proteins, individually endowed with unique structural and functional attributes.

The beginning of the chain, known as the N-terminus, retains a free amino group ($-\text{NH}_3^+$ under physiological conditions), while the opposite end, known as the C-terminus, bears a free carboxyl group ($-\text{COO}^-$). Therefore, a protein is always represented in the form of the 'N' to 'C' direction (Branden; Tooze, 2012). This orientation is significant to a protein's synthesis due to the fact that it determines the sequence in which amino acids are read and incorporated by the ribosomes during translation.

Hence, while the sequence of the R group defines a protein's unique chemical identity and functionality, the peptide backbone serves as the structural framework of a protein. The rigidity of the peptide bond constrains rotational freedom and contributes to the defined geometry of a protein's secondary structures. Therefore, peptide bond formation not only links amino acids into polymers but also establishes the physicochemical basis for higher levels of structural organization that eventually regulate protein function. Figure 2 demonstrates how a protein is made in the ribosome.

The principle that sequence determines structure, and consequently, structure dictates function, forms the theoretical foundation upon which modern protein research is built. For a protein to perform its function effectively, its constituent atoms must

Figure 2 – Protein's biological synthesis process



SOURCE: The author (2025)

adopt a highly specific three-dimensional arrangement corresponding to the most thermodynamically stable, or energetically favorable, state (Anfinsen, 1973). This precise structural configuration is known as the native conformation, and its attainment is the result of a complex folding process guided by intramolecular interactions that collectively minimize the free energy of the system.

Protein misfolding occurs when a polypeptide fails to attain or maintain its native conformation. These misfolded proteins often expose hydrophobic residues that are normally buried within the protein core, promoting self-association and the formation of insoluble aggregates, which can disrupt cellular homeostasis, interfere with normal biochemical pathways, and trigger cytotoxic effects that contribute to the development of several disorders (Hipp; Hartl, 2024). Some neurodegenerative diseases exemplify the pathological consequences of protein misfolding and aggregation, such as Alzheimer's disease, amyotrophic lateral sclerosis, and Parkinson's disease (Selvaraj *et al.*, 2025).

Bioinformatics constitutes a keystone of modern biological research, offering computational methodologies for systematic analysis of protein sequences, structures, and interactions (Li *et al.*, 2024). Through its integrative and interdisciplinary nature, bioinformatics complements traditional experimental biology by providing an efficient, cost-effective, and scalable means of exploring complex biological systems. Although conventional laboratory techniques such as X-ray crystallography (Ilari; Savino, 2008) and nuclear magnetic resonance (Wuethrich, 1989) remain indispensable for empirical validation, they are often resource-intensive, time-consuming, and limited in throughput. As a solution, bioinformatics uses computational power to predict, model, and

simulate molecular processes, enabling the generation of testable hypotheses prior to experimental verification (Terdalkar, 2025).

Protein Structure Prediction (PSP) represents a central and long-standing challenge in bioinformatics, primarily due to the immense conformational space that a polypeptide chain can theoretically adopt (Kuhlman; Bradley, 2019). Moreover, the dynamic nature of proteins, which can take on many conformations depending on the physiological situation, adds another level of intricacy (Gough; Kalodimos, 2024). Therefore, determining a protein's three-dimensional structure from its amino acid sequence has been a fundamental goal in molecular biology, as structure dictates function and underlies nearly all biochemical processes. PSP has seen transformative progress through Machine Learning (ML) and integrated bioinformatics approaches, advances that have made accurate, large-scale predictions feasible (Malhotra *et al.*, 2025).

In recent years, computational approaches, ML more specifically, have been gaining appreciation in biological areas, assessing the activities that occur in a cell, organ, or environment with the goal of having a comprehensive view of the recorded proteins (Goshisht, 2024). ML algorithms are particularly well-suited to the challenges posed by modern biological data, which are often high-dimensional, heterogeneous, and noisy. This suitability can be explained by how ML easily finds subtle patterns that might be missed by traditional statistical methods, and consequently, reveals intricate, nonlinear relationships within biology-related datasets.

Among the various strategies developed to address this challenge, *ab initio* models represent a fundamental and conceptually elegant approach (Calvet *et al.*, 2023). These models seek to predict protein structures relying on the physicochemical properties of amino acids and the thermodynamic principle that a protein's native state corresponds to its global free-energy minimum. Unlike template-based or homology modeling methods, which depend on known structures as references, *ab initio* prediction aims to simulate the folding process itself—exploring conformational space through energy minimization, molecular dynamics, or stochastic optimization techniques (Filgueiras *et al.*, 2022).

Within the *ab initio* framework, models of protein folding can be broadly categorized as lattice or off-lattice, each differing in how conformational space is represented and explored. The first kind discretizes protein structures onto a predefined grid, simplifying the folding problem by restricting possible residue positions to fixed lattice points (Wang; Zhou, 2024). While computationally efficient and useful for conceptual studies of folding kinetics, these models impose geometric constraints that limit their capacity to capture the continuous and flexible nature of real protein conformations.

In contrast, off-lattice models operate in continuous three-dimensional space,

allowing residues (amino acids) to adopt any spatial position permitted by physical and chemical constraints. They aim to minimize the total potential energy of the system by considering a comprehensive set of force-field parameters and energy terms, including bond angles and torsional rotations (Misir, 2024). These models can generate structures that more closely approximate experimental observations. However, their increased accuracy comes at a high computational cost, as the energy surface becomes vast and complex, containing numerous local minima. To overcome this challenge, metaheuristic optimization algorithms are frequently employed to efficiently search for low-energy conformations that correspond to the protein's native state (Oladejo, 2024).

The metaheuristic algorithms evaluate protein structures using energy or fitness functions that approximate the protein's potential energy. Then, these approaches can identify near-native conformations with remarkable efficiency compared to exhaustive search techniques by iteratively minimizing the energy functions (Jana *et al.*, 2018b). Ultimately, solving the PSP problem through such models enhances the understanding of the fundamental principles governing protein folding and also contributes to decoding misfolding mechanisms in proteins.

Other methodologies that have reshaped the landscape of protein structure analysis are deep learning strategies such as AlphaFold (Jumper *et al.*, 2021b) and RoseTTAFold (Bord, 2021). They have demonstrated unprecedented accuracy by integrating sequence information, evolutionary patterns, and structural constraints to infer the intricate relationships between the amino acid sequence and the protein's three-dimensional conformation, along with its biological function. Despite these remarkable achievements, the PSP problem remains fundamentally unsolved (Moore *et al.*, 2022).

For proteins known as orphan proteins, de novo sequences, or black proteome elements that do not exhibit identifiable sequence homology or evolutionary information in current databases, the prediction performance of artificial intelligence (AI)-based techniques is still restricted. In these situations, the lack of co-evolutionary indicators or numerous sequence alignments drastically lowers model accuracy and confidence. Furthermore, these methods might have trouble with dynamic structures, sections that are inherently disordered, conformational flexibility, and proteins that experience substantial structural changes brought on by partners or ligands.

In the context presented, the PSP problem becomes a promising avenue for exploration. At its core, PSP seeks to determine the native three-dimensional conformation of a protein solely from its linear amino acid sequence, a task made difficult by the astronomical size of the conformational search space and the complex energy landscape that governs protein folding. The motivation behind this thesis arises from the need to overcome these persistent challenges by introducing an innovative bilinear (BL) optimization-based framework. This method aims to improve the accuracy and

robustness of structure prediction while offering a principled mathematical formulation capable of enabling exact or near-exact solution strategies. In doing so, it contributes not only to computational efficiency but also to the theoretical foundations of PSP, opening new pathways for rigorous, optimization-driven approaches in structural bioinformatics.

The potential impact of this research is accentuated by the potential to recast the PSP problem as an exact optimization task, thereby enabling the systematic application of advanced mathematical programming and global optimization techniques. By embedding this novel BL optimization methodology within the AB off-lattice model, the framework achieves a balance between representational simplicity and biophysical relevance: it reduces the complexity of full atomic detail while preserving the essential geometric and energetic principles governing protein folding. This dual advantage enhances both the tractability and the interpretability of the PSP formulation, as the Bilinear reformulation allows for exact global solutions where heuristics fail.

Understanding and elucidating the three-dimensional structure of proteins is therefore of outstanding importance, as structural information provides the foundation for deciphering molecular mechanisms, biological functions, and intermolecular interactions. Crucially, the relevance of structural knowledge extends beyond theoretical inquiry to numerous scientific and technological applications. In drug discovery and rational drug design, detailed structural models enable the identification of functional pockets and guide the engineering of molecules capable of modulating protein activity with high specificity.

Moreover, in medical and disease-related research, abnormalities in protein conformation, whether through misfolding, aggregation, or mutation-induced instability, play central roles in disorders such as Alzheimer's (Tecalco-Cruz *et al.*, 2022), Parkinson's (Padilla-Godínez *et al.*, 2021), and cystic fibrosis (McDonald *et al.*, 2023). Likewise, in synthetic biology and bioengineering, structural insights empower the design of novel proteins with tailored catalytic properties, enhanced stability, or entirely new functionalities. Consequently, understanding protein conformation holds profound implications not only for fundamental biology but also for translational research and biomedical innovation.

1.1 MOTIVATION

As previously said, PSP has a great impact in biomedicine, especially in drug discovery and testing, and in disease-related studies. This is due to the fact that accurately predicting the three-dimensional structures of proteins signifies a crucial advancement in computational biology (Teixeira *et al.*, 2026). These predictions offer insights into the underlying principles governing protein folding and the relationships between structure and function, with implications throughout both scientific and medical

domains. Comprehending protein structures enables scientists to design innovative biomolecules with designated roles, direct therapeutic strategies, and elucidate the molecular causes of diseases (Yin *et al.*, 2026).

Beyond its biological and medical relevance, the PSP problem also exerts significant influence on the fields of computer science and engineering. Its inherent complexity positions PSP as a benchmark challenge for developing advanced algorithms, optimization techniques, and high-performance computational architectures. Moreover, the problem inspires the creation of new computational models capable of handling large-scale, non-convex problems, pushing the boundaries of algorithm design and computational efficiency (Ashrafuzzaman, 2026).

Given the centrality of protein structure to biological function, the ability to accurately predict a protein's native conformation from its amino acid sequence has long been one of the most compelling and challenging problems in bioinformatics. Despite the extraordinary progress achieved by deep learning models such as AlphaFold and RoseTTAFold, several fundamental obstacles persist. Thus, although transformative, current AI-based predictors do not fully resolve the PSP problem, particularly in contexts where data-driven inference is constrained or insufficient.

In light of these challenges, *ab initio* approaches, that is, those relying solely on first principles of physics and chemistry without the need for homologous sequences, remain an essential avenue of exploration. These models attempt to infer protein structure from the intrinsic energetic and geometric properties of amino acids and their interactions. In this context, simplified models like the AB off-lattice model are essential as they diminish the complexity of the folding process while maintaining its fundamental physical factors.

Motivated by the need for new methods that overcome the limitations of state-of-the-art predictors while retaining computational tractability, this thesis proposes an innovative bilinear optimization-based approach for the PSP problem, which reformulates protein folding as an exact optimization problem. Integrating this formulation with the AB off-lattice model provides a structured, analytically grounded framework capable of approximating the energy landscape while offering theoretical guarantees unavailable in conventional heuristic or deep-learning (DL)-based approaches.

In summary, the motivation for this thesis arises from the convergence of three major factors: the biological importance of protein structure; the limitations of existing predictive models in scenarios lacking evolutionary information; and the opportunity to advance protein structure prediction through rigorous optimization-based methodologies. By developing an exact bilinear optimization framework within an *ab initio* off-lattice model, this research seeks to contribute to the foundational understanding of protein folding while offering new computational tools for tackling one of the most enduring

challenges in computational biology.

1.2 OBJECTIVES

The primary aim of this thesis is the proposal of a cutting-edge computational framework for protein structure prediction, employing mathematical optimization techniques, particularly bilinear optimization, to properly predict protein structures using an AB off-lattice model.

The specific objectives of this research are:

1. To analyze the literature on the PSP problem, including the accuracy and computational cost of model architectures, deep learning algorithms, and metaheuristics;
2. To design, implement, and test a systematic scheme for transforming PSP's nonlinear optimization problem into a BL optimization format, which reduces computational time by up to five times in proteins smaller than 100 residues;
3. To perform numerical experiments and computational analysis of the BL optimization methodology for different protein sizes up to 100 amino acids;
4. To evaluate and compare the performance of the BL optimization technique against prevalent metaheuristic methods regarding accuracy, reliability, and computational efficiency;
5. To evaluate and compare the performance of the BL optimization technique against metaheuristic methods using protein-relevant metrics such as root mean square deviation (RMSD) and statistical metrics;
6. For proteins smaller than 50 amino acids, the achieved RMSD must be lower than 5Å;
7. For proteins bigger than 50 amino acids, the achieved RMSD must be lower than 10Å.

1.3 CONTRIBUTIONS

For the sake of predicting protein structures more accurately, the thesis presents a novel method based on bilinear optimization that provides a more realistic description of protein folding energy landscapes. This study illustrates that PSP can be analyzed via an exact bilinear structure, offering novel perspectives on the development of scalable optimization algorithms, promoting the investigation of hybrid mathematical-computational methodologies, and enhancing the field of computational engineering

by showcasing the transformation of intricate real-world challenges into organized optimization problems.

By decomposing the non-linear PSP formulation into a BL model, it enables a more systematic exploration of conformational space while preserving essential biophysical principles. This reformulation enhances the accuracy of protein structure prediction along with a lower computational runtime cost when compared to classical approaches. The BL optimization-based approach demonstrates compelling performance, surpassing the accuracy of several established metaheuristic methods traditionally employed in PSP research (Silva *et al.*, 2024). Through a detailed comparative evaluation using diverse datasets from the Protein Data Bank (PDB) (PDB, 2023), the methodology exhibits robustness, consistency, and improved predictive capability relative to existing computational frameworks.

Importantly, the contributions of this thesis extend beyond computational biology and hold significant implications for engineering and computer science. The formulation and solution of the BL model advance the development of algorithms capable of efficiently addressing high-dimensional, non-convex optimization problems, which is a class of challenges widely encountered in engineering design, control systems, signal processing, and machine learning. The transformation of a complex nonlinear PSP problem into a bilinear optimization one constitutes a significant contribution that transcends Biology, marking a progression in optimization methods and the resolution of large-scale nonconvex issues.

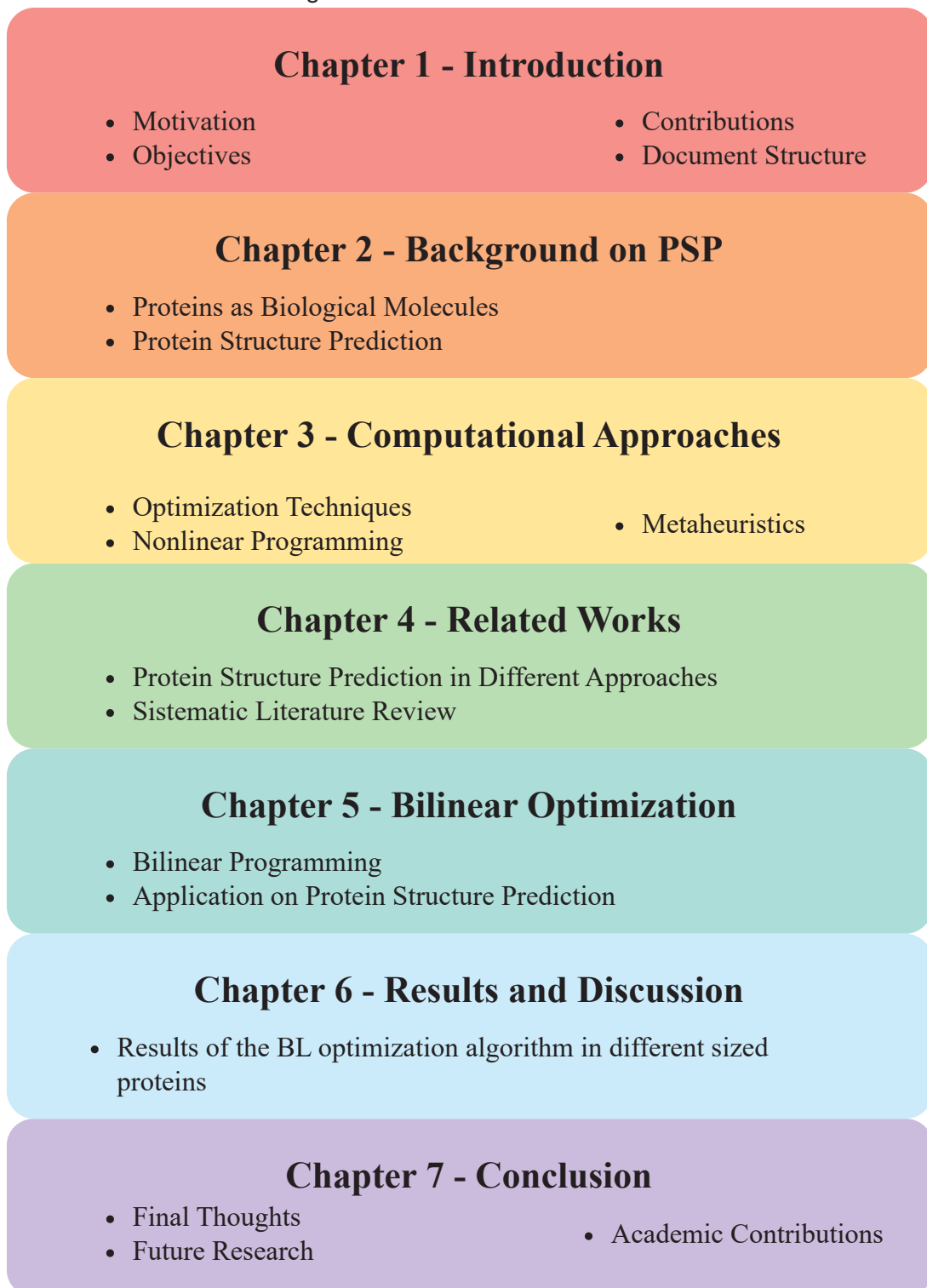
1.4 DOCUMENT STRUCTURE

The remainder of this doctoral thesis is structured into six chapters, which are detailed in the following. Chapter 2 details the main theoretical aspects involving the protein structure prediction problem. Ergo, this chapter elucidates the background regarding the physical and biological aspects of the PSP problem, including a presentation of the mathematical computation of the AB off-lattice model. In Chapter 3, a thorough explanation of the fundamental concepts in relation to optimization techniques, with an emphasis on non-linear programming is expected. Next, Chapter 4 is dedicated to a comprehensive review of related works.

The subsequent chapter, Chapter 5, marks a significant advancement in the field by introducing BL programming and optimization techniques within the context of PSP. This innovation aims to enhance the accuracy, reliability, and efficiency of strategies for deciphering the complexities of protein folding. The consequences of the proposed model are explored in depth in Chapter 6, where the results are shown alongside a discussion of their meaning for the PSP application. Finally, Chapter 7 concludes this thesis and delineates some research prospects. In Figure 3, the thesis's structure is

presented.

Figure 3 – Structure of the thesis



SOURCE: The author (2025)

2 PROTEIN STRUCTURE PREDICTION

Protein folding is the physical process through which a linear polypeptide chain acquires its native three-dimensional structure, which is the specific conformation required for its biological function (Song *et al.*, 2018). The computational challenge of modeling or predicting this process is known as the protein folding problem, or PSP problem. Due to the astronomically large number of conformations that a protein can theoretically adopt, PSP is regarded as one of the most difficult and long-standing problems in computational biology, involving both combinatorial complexity and intricate biophysical interactions.

As already expressed, a protein's primary structure, defined by its linear residue sequence, encodes all the information necessary to determine its unique, functional three-dimensional form. Moreover, the rotational degrees of freedom around peptide and side-chain bonds allow the protein chain to explore a vast conformational landscape during folding. Despite this flexibility, proteins reliably adopt a specific native structure, corresponding to the state of lowest free energy (Finkelstein *et al.*, 2022).

Because the fundamental principle "sequence dictates structure, and structure determines function" lies at the core of molecular biology, the PSP problem has long been recognized as a critical challenge in biochemistry and computational biology (Pearce; Zhang, 2021b). For much of the 2000s, progress in accurate structure prediction appeared to plateau, hindered by limited experimental data and the difficulty of capturing long-range residue interactions.

This trajectory shifted in the early 2010s with the emergence of coevolutionary approaches, which used patterns in multiple sequence alignments to infer residue-residue interactions and produced the first significant increases in prediction accuracy in years (Dos Santos *et al.*, 2018). A decade of gradual progress ensued, ultimately leading to a significant breakthrough with the introduction of DeepMind's AlphaFold2 at the onset of the 2020s (Marcu *et al.*, 2022). AlphaFold2 introduced a transformative deep-learning architecture that dramatically improved prediction quality, achieving performance widely described as having virtually solved the problem for many single-domain proteins.

Computational approaches to the PSP problem are traditionally classified into three major categories: template-based modeling, template-free modeling, and hybrid methods. The first one is also known as homology or comparative modeling; it relies on the existence of experimentally determined protein structures with sequences similar to the target protein. By aligning the unknown sequence with these established templates, the technique deduces the probable three-dimensional conformation (Binbay

et al., 2023). Consequently, template-based modeling builds models by replicating and optimizing the structural grids of other similar proteins, referred to as templates, that have been found in the Protein Data Bank (PDB).

In contrast, template-free methods, frequently referred to as *ab initio* or *de novo* modeling approaches (Dhingra *et al.*, 2020), do not depend on structural databases and instead attempt to predict protein structures from first principles, using physico-chemical models, optimization strategies, or deep-learning architectures to explore the conformational landscape. Template-free modeling seeks to predict protein structures without the use of global arrangement structures; they have historically relied on physics, knowledge-based energy functions, and intensive sampling operations to generate protein structure models (Mufassirin *et al.*, 2023). Finally, hybrid methods integrate both strategies.

Coarse-grained protein structures are often generated through lattice-based models, while such formulations simplify the search space and make computation more tractable, they provide only an approximate representation of real protein geometry and therefore offer limited biological fidelity. When combined with modern protein-specific scoring functions, including those derived from empirical biophysics and machine-learning frameworks, off-lattice models achieve a significantly higher descriptive accuracy compared to their lattice-based counterparts. As a result, off-lattice modeling holds greater academic and practical relevance for PSP, as it offers the capacity to capture structural subtleties essential to depict leading chains more accurately.

Anfinsen, Haber, and White's formulation of the Thermodynamic Hypothesis emerged from pioneering in vitro refolding experiments, which revealed that a fully denatured protein could spontaneously regain its native conformation. This observation demonstrated a fundamental principle: the biological activity of a protein is inextricably linked to its precise three-dimensional arrangement of atoms, highlighting that structure is the physical basis of function (Anfinsen *et al.*, 1961). A decade later, Anfinsen strengthened this conclusion using both in vitro and in vivo approaches, showing that proteins reliably return to a single, energetically favorable fold when environmental conditions permit. These findings established the now-classical view that all necessary information for correct folding is encoded in the amino acid sequence itself, and that the native state corresponds to the global minimum of the protein's Gibbs free energy landscape (Anfinsen, 1973).

This thermodynamic control perspective profoundly shaped modern biophysics, leading to the notion that folding is not random, but directed by intramolecular forces such as hydrogen bonding, hydrophobic interactions, van der Waals contacts, and electrostatics. Because the correct fold represents the state of lowest free energy, protein folding can be interpreted as a self-organization process driven by thermodynamics,

not requiring external templating or cellular machinery for many small globular proteins. With this theoretical foundation well established, the Anfinsen hypothesis continues to influence fields such as computational protein design, molecular simulation, materials engineering, and structure-based drug discovery, where predicting or engineering stable conformations relies fundamentally on free-energy optimization principles (Hirata *et al.*, 2018).

Therefore, according to Anfinsen’s thermodynamic hypothesis, the native conformation of a protein corresponds to the state of minimum Gibbs free energy (Ahmad; Khan, 2005), many computational approaches are built around the principle of energy minimization. These methods evaluate candidate structures using an energy function designed to approximate the biophysical forces that drive folding (Chivian *et al.*, 2003). Under this framework, Anfinsen’s theory implies that the primary amino acid sequence contains all the structural information needed to converge toward the native fold, without external templating, guiding the search for the lowest-energy arrangement of residues.

The computational interpretation of this idea led to the development of search algorithms capable of sampling and scoring millions of potential conformations. However, the sheer size of conformational space presents a major challenge, as even moderately sized proteins have astronomically many possible states. To navigate this landscape efficiently, modern techniques often adopt hierarchical or coarse-grained representations, progressively refining models from simplified abstractions toward atomic detail. This multiscale strategy reflects how proteins fold in nature, forming local secondary structures while also establishing long-range contacts that shape the global topology, with both processes occurring cooperatively and often in parallel (Torrìsi *et al.*, 2020).

In other words, computer simulations were created to rate protein conformations and subsequently search through possible states for the one with the lowest free energy. Using simpler, usually coarser abstractions to progressively direct the search is one way to solve the grand issue of conformational search space size explosion, since proteins seem to fold both locally and non-locally simultaneously while progressively producing more complicated forms.

2.1 AB OFF-LATTICE MODEL

The *ab initio* PSP problem can be approached using either lattice-based or off-lattice modeling frameworks (Boiani; Parpinelli, 2020). Off-lattice models relax spatial constraints and allow continuous variation in bond angles and distances, enabling a more realistic and detailed representation of protein conformations. For a model to be effective, however, it must balance abstraction with expressiveness, providing a tractable search space while still reflecting the biophysical forces and energy landscape that guide folding. The development of the AB off-lattice model emerged precisely from this need,

offering a reduced yet physically grounded model capable of describing conformational behavior and defining an energy function suitable for evaluating candidate structures.

At the core of the AB off-lattice formulation lies the hydrophobic–polar abstraction, one of the earliest and most influential simplifications in protein modeling. Rather than representing the full biochemical diversity of the twenty amino acids, this kind of model categorizes residues into two fundamental groups according to their affinity for the aqueous environment: hydrophobic residues, which tend to cluster internally to minimize solvent exposure, and polar residues, which are more stable when exposed to water, also known as hydrophilic (Agrahari; Chatterjee, 2020). This dichotomy is rooted in the thermodynamic driving force of protein folding, where the burial of hydrophobic residues is a primary determinant of native structure formation.

In this representation, the emergent structure arises from the interplay between residue interactions and chain geometry. Hydrophobic residues increasingly attract one another, forming compact cores characteristic of folded proteins, whereas polar residues tend to remain solvent-facing, stabilizing the outer surfaces (Hameduh *et al.*, 2025). The resulting behavior is governed by an appropriate energy function that evaluates pairwise interactions, chain bending angles, and spatial arrangement, ensuring that conformations compatible with hydrophobic collapse are energetically favored. Despite its minimalism, the AB abstraction reproduces essential folding phenomena, such as secondary-structure-like motifs, core formation, and transition between intermediate states (Badaczewska-Dawid; Kolinski, 2024).

The AB off-lattice model was first introduced in 1993 by Stillinger *et al.* (Stillinger *et al.*, 1993) as an insightful representation of protein folding dynamics. In this formulation, the twenty naturally occurring amino acids are reduced into two fundamental categories: A (hydrophobic) and B (polar or hydrophilic), as summarized in Table 1.

Table 1 – Essential amino acids

Name	Hydro affinity	Name	Hydro affinity
Alanine	Hydrophobic	Leucine	Hydrophobic
Arginine	Polar	Lysine	Polar
Asparagine	Polar	Methionine	Hydrophobic
Aspartic acid	Polar	Phenylalanine	Hydrophobic
Cysteine	Hydrophobic	Proline	Hydrophobic
Glutamine	Polar	Serine	Polar
Glutamic acid	Polar	Threonine	Polar
Glycine	Hydrophobic	Tryptophan	Hydrophobic
Histidine	Polar	Tyrosine	Polar
Isoleucine	Hydrophobic	Valine	Hydrophobic

Therefore, in the AB off-lattice model, a protein is represented by a binary sequence vector $\{A, B\}$, where each entry denotes whether a residue is hydrophobic

(A) or polar/hydrophilic (B). This abstraction enables the amino acid chain to be encoded compactly while preserving the hydrophobic effect. Additionally, the geometry of the chain is defined through its bond (or curvature) angles, represented by

$$\theta = \{\theta_1, \theta_2, \dots, \theta_{n-2}\} \quad (2.1)$$

where n is the total number of residues, and each angle θ_i specifies the directional change between consecutive bonds along the chain. In the off-lattice formulation, these angles vary continuously within the $[-180^\circ, 0) \cup (0, 180^\circ]$ range, negative values represent clockwise rotations and positive values correspond to counterclockwise rotations. An angle of zero is excluded, as it would yield a fully collinear arrangement of three successive residues, offering no geometric bend.

In addition to curvature angles, the three-dimensional representation of the protein also requires consideration of torsion angles, denoted by

$$\beta = \{\beta_1, \beta_2, \dots, \beta_{n-3}\} \quad (2.2)$$

Each torsion angle β_i specifies the rotation around the bond connecting residues $(i + 1)$ and $(i + 2)$, thereby determining how the chain twists out of the plane defined by the corresponding curvature angles. Together, the curvature and torsion angles define the full spatial embedding of the chain in $\mathbb{R}^3$.

For a protein chain of length n , the complete conformation is thus described by a total of $(2n - 5)$ continuous angular parameters: $(\theta_1, \theta_2, \dots, \theta_{n-2}, \beta_1, \beta_2, \dots, \beta_{n-3})$. Figure 4 illustrates the geometric role of the curvature angles θ and torsion angles β for two hypothetical protein fragments, displaying how these parameters collectively determine the three-dimensional folding pattern of the chain.

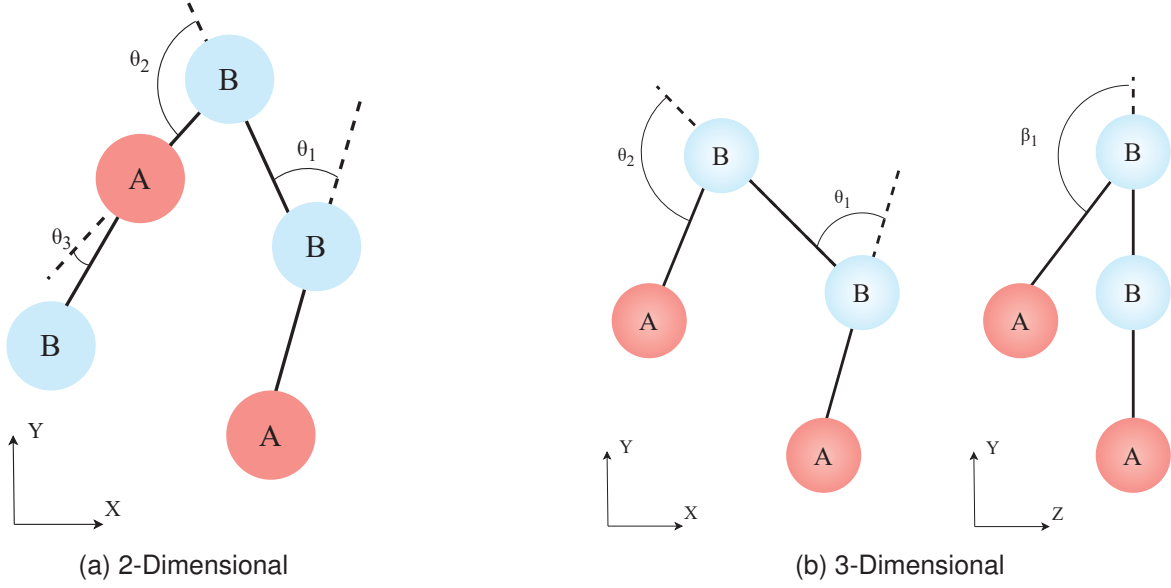
The sequence of amino acids along the structure is represented through a set of binary variables $\{\xi_i\}_{i=1}^n$. If the i^{th} amino acid is of type 'A', then ξ_i is equal to one; if it is of type 'B', then ξ_i is negative one. Hence, this binary encoding allows the interaction terms in the energy function to explicitly depend on the residue types.

Using this representation, the intramolecular potential energy Φ of a protein of length n can be expressed as

$$\Phi = \underbrace{\sum_{i=1}^{n-2} V_1(\theta_i)}_{\text{curvature (bond angle) energy}} + \underbrace{\sum_{i=2}^{n-1} \sum_{j=i+2}^n V_2(r_{ij}, \xi_i, \xi_j)}_{\text{nonbonded pairwise interactions}}, \quad (2.3)$$

where $V_1(\theta_i)$ is the curvature or bond-angle potential and $V_2(r_{ij}, \xi_i, \xi_j)$ is the pairwise Lennard Jones (Lenhard *et al.*, 2024) type interaction potential between residues i and j , with r_{ij} denoting the Euclidean distance between them.

Figure 4 – Types of angles in the AB off-lattice model



SOURCE: The author (2025)

The curvature energy term penalizes deviations from linearity between successive bonds and encourages locally smooth chain conformations by assigning lower energy to configurations in which the curvature angle θ_i is small, i.e., closer to a straight bond alignment. Accordingly, the curvature energy is defined as

$$V_1(\theta_i) = \frac{1 - \cos(\theta_i)}{4}, \quad i = 1, 2, \dots, n - 2. \quad (2.4)$$

The non-bonded interaction term ($V_2(\cdot)$) is modeled using a species-dependent Lennard-Jones (12,6) potential, which interprets that two interacting particles exhibit repulsion at minimal distances, attraction at moderate distances, and cease interaction at infinite distances. In more detail, this formulation incorporates a steep repulsive force at short inter-residue distances and a weaker attractive component at longer distances, representing van der Waals interactions. The potential is defined as

$$V_2(r_{ij}, \xi_i, \xi_j) = 4 \left(r_{ij}^{-12} - C(\xi_i, \xi_j) r_{ij}^{-6} \right), \quad (2.5)$$

where r_{ij} denotes the Euclidean distance between residues i and j . $C(\xi_i, \xi_j)$ is a residue-type interaction coefficient that modulates the attractive term; it encodes the hydrophobic or hydrophilic character of the residues based on the AB representation and is given by

$$C(\xi_i, \xi_j) = \frac{1}{8} (1 + \xi_i + \xi_j + 5 \xi_i \xi_j). \quad (2.6)$$

This expression yields three distinct interaction strengths:

$$\begin{aligned} C(\xi_i, \xi_j) &= 1.0 && \text{for an AA pair (strong attraction),} \\ C(\xi_i, \xi_j) &= -0.5 && \text{for an AB or BA pair (weak repulsion),} \\ C(\xi_i, \xi_j) &= 0.5 && \text{for a BB pair (weak attraction).} \end{aligned}$$

The inter-residue distance r_{ij} is computed using the Euclidean norm applied to the Cartesian coordinates (x_i, y_i, z_i) and (x_j, y_j, z_j) of residues i and j :

$$r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}. \quad (2.7)$$

The Cartesian coordinates for residue i in the AB off-lattice model are determined directly from the curvature angles θ and torsion angles β . For a chain embedded in three-dimensional space, the coordinates (x_i, y_i, z_i) are recursively computed as

$$(x_i, y_i, z_i) = \begin{cases} (0, 0, 0) & \text{for } i = 1 \\ (0, 1, 0) & \text{for } i = 2 \\ (\cos(\theta_1), \sin(\theta_1) + 1, 0) & \text{for } i = 3 \\ (x_{i-1} + \cos(\theta_{i-2}) \cos(\beta_{i-3}), y_{i-1} + \sin(\theta_{i-1}) \cos(\beta_{i-3}), \\ \quad z_{i-1} + \sin(\beta_{i-3})) & \text{for } 4 \leq i \leq n \end{cases} \quad (2.8)$$

The first three residues are initialized in the plane $z = 0$, establishing a fixed reference frame for the remainder of the chain. For all subsequent residues ($i \geq 4$), the position of residue i is established from residue $i - 1$ by applying a rotation determined by the curvature angle θ_{i-2} and torsion angle β_{i-3} . The curvature angle dictates how sharply the chain bends within the local plane, while the torsion angle controls the out-of-plane twisting, so facilitating a complete three-dimensional representation of the polypeptide backbone.

Given this geometric representation, a complete conformation of a protein of length n is uniquely determined by the set of angular variables

$$(\theta_1, \theta_2, \dots, \theta_{n-2}, \beta_1, \beta_2, \dots, \beta_{n-3}),$$

which map directly to Cartesian coordinates via the recurrence above. The PSP task in the AB off-lattice model consequently consists of identifying the configuration of these angles that minimizes the total potential energy Φ . This entails identifying the global minimizer of the AB-model energy landscape, resulting in the conformation that most closely resembles the native, lowest-free-energy structure of the protein.

The computational analysis of the AB off-lattice model becomes more tractable due to the reduction in complexity (Lee *et al.*, 2008) and is more like the natural physical environment in which proteins fold. Nevertheless, this simplicity also introduces important limitations. The binary hydrophobic–hydrophilic categorization of residues and the heavily abstracted interaction potentials fail to capture the rich chemical diversity and context-dependent behavior of the 20 natural amino acids. As a result, the model may overlook critical physicochemical effects that drive folding, potentially leading to

oversimplified solutions or loss of structural specificity (Zhang *et al.*, 2020). Moreover, the predictive performance of the AB model is strongly dependent on the chosen energy function; inaccuracies, coarse approximations, or incomplete parameterizations can propagate through the optimization process and yield systematically biased or erroneous folding predictions (Bošković; Brest, 2016).

For a more comprehensive explanation, the complete objective function for the PSP problem in the AB off-lattice model is mathematically expressed as:

$$\min \quad \Phi = \sum_{i=1}^{n-2} V_1(\theta_i) + \sum_{i=2}^{n-1} \sum_{j=i+2}^n V_2(r_{ij}, \xi_i, \xi_j) \quad (2.9a)$$

$$\text{where} \quad V_1(\theta_i) = \sum_{i=2}^{n-1} \frac{1 - \cos(\theta_i)}{4} \quad (2.9b)$$

$$V_2(r_{ij}, \xi_i, \xi_j) = 4(r_{ij}^{-12} - C(\xi_i, \xi_j), r_{ij}^{-6}) \quad (2.9c)$$

$$\text{subjected to:} \quad (2.9d)$$

$$(x_i, y_i, z_i) = \begin{cases} (0, 0, 0) & \text{for } i = 1 \\ (0, 1, 0) & \text{for } i = 2 \\ (\cos(\theta_1), \sin(\theta_1) + 1, 0) & \text{for } i = 3 \\ (x_{i-1} + \cos(\theta_{i-2}) \cos(\beta_{i-3}), y_{i-1} + \sin(\theta_{i-1}) \cos(\beta_{i-3}), z_{i-1} + \sin(\beta_{i-3})) & \text{for } 4 \leq i \leq n \end{cases} \quad (2.9e)$$

$$\theta_i \in [-180^\circ, 0) \cup (0, 180^\circ] \quad (2.9f)$$

$$\beta_i \in [-180^\circ, 0) \cup (0, 180^\circ] \quad (2.9g)$$

$$r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (2.9h)$$

$$C(\xi_i, \xi_j) \in -0.5, 0.5, 1 \quad (2.9i)$$

3 COMPUTATIONAL APPROACHES

The prediction of protein structure ultimately requires searching a vast conformational space to identify structures with minimal free energy, a process that lies at the heart of computational optimization. Broadly speaking, an optimization problem seeks the best solution among a set of feasible alternatives, guided by an objective function and subject to defined constraints (Agrawal *et al.*, 2018). In this context, the “optimal” solution corresponds to the configuration that minimizes (or maximizes) the chosen objective, such as a protein’s estimated energy.

Formally, both general and domain-specific optimization problems can be expressed in the following form:

$$\min f(x) \tag{3.1a}$$

$$\text{subjected to: } h_i(x) = 0, \quad i = 1, \dots, m_h, \tag{3.1b}$$

$$g_j(x) \leq 0, \quad j = 1, \dots, m_g, \tag{3.1c}$$

where, the vector $x = (x_1, \dots, x_n)$ gathers the problem’s optimization variables, $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is the objective function to be minimized, and the functions $h_i(x) = 0$, and $g_j(x) \leq 0$ are respectively the equality and inequality constraint. In practice, each variable in x often has lower and upper bounds, which are either explicitly defined or implicitly induced by the constraints (Bazaraa *et al.*, 2013).

Suppose a vector x^* has the lowest objective value of all the vectors inside the feasible region. In that case, the set of points that satisfy every constraint is referred to as the optimal solution for the problem, or the region within the constraint borders. One challenge with this kind of problem is that there might be several local minima (local maxima) if the feasible region and the goal function are not both convex in a minimization problem (or, analogically, concave in a maximization problem). Whereas a global minimum is as good as all conceivable points, a local minimum has a solution that is at least as excellent as any point nearby (Ryu; Yin, 2022).

The formulation of mathematical relationships among choice variables, constraints, and the objective function may provide an optimization problem that is either convex or concave. The two classes are intimately connected, as the negative of a convex function is a concave function, and conversely. Another important property is closure under addition, which states that the sum of convex functions remains convex, and the same holds for concave functions (Wilhelm *et al.*, 2023).

The features of convex optimization problems are especially appealing. Any local minimum is also a global minimum if the feasible set is convex and the objective

function is convex, guaranteeing the existence of a single globally optimal solution. Non-convex optimization problems, on the other hand, are far more difficult since they may have several local minima, saddle points, and flat areas, which makes it harder for optimization algorithms to find the global optimum consistently.

A central class within mathematical optimization is convex optimization, which encompasses problems such as linear programming (LP) and least-squares minimization. LP problems, in particular, are characterized by linear objective functions and linear equality and inequality constraints, allowing them to be solved efficiently even at large scales (Dantzig, 2002). In optimization theory, the notion of convexity applies to both sets and functions, and they are defined as convex under specific criteria.

A set $C \subseteq \mathbb{R}^n$ is a convex set if given any two points $x_1, x_2 \in C$, and $x_1 \neq x_2$, then

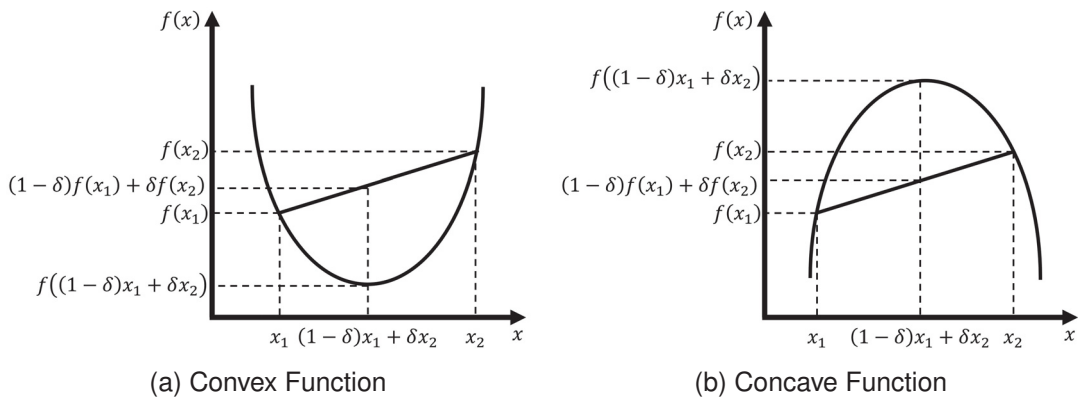
$$(1 - \delta)x_1 + \delta x_2 \in C, \quad \forall \delta \in [0, 1] \quad (3.2)$$

Similarly, a function $f \subseteq \mathbb{R}$ is convex if its domain C is a convex set and if for two points $x_1, x_2 \in C$:

$$f((1 - \delta)x_1 + \delta x_2) \leq (1 - \delta)f(x_1) + \delta f(x_2), \quad \forall \delta \in [0, 1] \quad (3.3)$$

This inequality implies that the chord connecting $(x_1, f(x_1))$ and $(x_2, f(x_2))$ lies above the graph of the function, a property that geometrically distinguishes convex from concave functions. Convex optimization problems benefit from this structure, since every local minimum is also a global minimum, and powerful polynomial-time algorithms exist for large classes of convex problems. This characteristic is depicted in Figure 5.

Figure 5 – Example of convex and concave functions



SOURCE: The author (2024)

A standard analytical tool for assessing whether a function is convex or concave is the Hessian matrix (Byrd *et al.*, 2011). For a twice continuously differentiable scalar function $f(x)$, the Hessian matrix $\nabla^2 f(x)$ is defined as the matrix of second-order partial derivatives. The Hessian matrix can be written as follows:

$$H(f)(x) = \nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \quad (3.4)$$

A function f is said to be convex over a convex set $\mathcal{D} \subseteq \mathbb{R}^n$ if its Hessian matrix is positive semidefinite for all $\mathbf{x} \in \mathcal{D}$, that is, $\mathbf{z}^\top \nabla^2 f(\mathbf{x}) \mathbf{z} \geq 0$, $\forall \mathbf{z} \in \mathbb{R}^n$, $\forall \mathbf{x} \in \mathcal{D}$. If the Hessian is positive definite for all $\mathbf{x} \in \mathcal{D}$, then f is strictly convex, guaranteeing the existence of a unique global minimizer.

In a similar vein, if the function's Hessian is negative semidefinite over the domain, it is concave. The function is considered strictly convex if the Hessian is positive definite, which suggests that the global minimizer is unique. On the other hand, an indefinite Hessian suggests a non-convex function with several local extrema and saddle points. Finding globally optimal solutions becomes much more challenging when the Hessian is indefinite, the function is non-convex, and the optimization landscape may contain several local optima and saddle points (Cox, 2020).

Similarly, the function is considered concave if its Hessian is negative semidefinite over the domain. If the Hessian is positive definite, the function is said to be strictly convex, which suggests that the global minimizer is unique. Conversely, an indefinite Hessian indicates a non-convex function, implying the presence of saddle points and multiple local extrema. When the Hessian is indefinite, the function is non-convex, and the optimization landscape may contain multiple local optima and saddle points, significantly increasing the difficulty of finding globally optimal solutions.

However, many real-world scientific problems, which include PSP, do not exhibit convexity. Energy landscapes are typically rugged, multimodal, and highly nonlinear. These characteristics place such problems outside the domain of convex optimization and motivate the use of more general frameworks. This leads naturally to nonlinear programming (NLP), which, unlike LPs, allows nonlinear objective functions and nonlinear constraints, often resulting in non-convex search spaces with multiple local optima.

3.1 NON-LINEAR PROGRAMMING

At its core, nonlinear programming seeks an optimal solution, maximizing or minimizing an objective function subject to equality and inequality constraints. In the context of the PSP problem, the objective is typically formulated as a minimization task, where the goal is to find the conformation associated with the lowest free energy, which aligns directly with Anfinsen's thermodynamic hypothesis.

NLP is a branch of mathematical optimization dedicated to problems in which the objective function, at least one constraint, or both contain nonlinear expressions (Avriel, 2020). Nonlinear functions may include polynomial terms, exponentials, trigonometric expressions, or any combination of these. Such nonlinearities introduce significant computational challenges, including non-convexity and multiple local minima, which complicate the search for globally optimal solutions (Molina-Pérez *et al.*, 2024a). Due to its landscape complexity, NLP solutions are highly sensitive to the initial starting point and the specific algorithm employed. Consequently, two optimization runs initialized differently may converge to entirely distinct solutions (Li *et al.*, 2020).

As a result, solving NLP problems typically requires specialized optimization algorithms capable of handling nonlinear relationships, non-convexity, and potentially high-dimensional search spaces. Depending on the mathematical structure of the problem, different classes of methods may be employed. Gradient-based approaches (Molina-Pérez *et al.*, 2024b) are effective when the objective function is smooth and differentiable, while evolutionary algorithms (Kirdar *et al.*, 2023) provide robust derivative-free alternatives suited for irregular or noisy landscapes. Interior-point methods (Espaas; Vassiliadis, 2024), however, are widely used for constrained nonlinear problems due to their strong theoretical guarantees and efficiency in large-scale settings.

A common strategy for handling nonlinearities in optimization problems is to reformulate them into simpler structures, one of which is the bilinear form. Nevertheless, transforming a nonlinear problem into a bilinear one is not always straightforward, as the feasibility of this conversion depends heavily on the nature of the nonlinear terms involved (Fischetti; Monaci, 2020). One widely used technique is local linearization, where the nonlinear function is approximated around a specific operating point.

In some cases, nonlinear expressions can be simplified through the introduction of auxiliary variables, enabling a reformulation better suited for optimization. This transformation may convert inherently nonlinear relationships into bilinear constraints or at least into forms that interact more favorably with solvers. A typical approach is to replace nonlinear terms with the product of two newly defined variables, effectively capturing the original relationship in a bilinear structure (Zhen *et al.*, 2022).

More broadly, nonlinear functions can be approximated using a variety of analytical techniques, such as polynomial expansions, Taylor series, or piecewise-linear approximations, to achieve a representation that is simpler and more solver-friendly. These approximations do not eliminate nonlinearity but can significantly improve the tractability of the optimization problem by reshaping it into a form where bilinear or near-linear methods can be applied effectively.

Introducing methods that are the foundation of robust global optimization techniques, as is the case with bilinear optimization, is crucial because they represent

the state-of-the-art for efficiently solving convex or locally-convex subproblems that arise during the global search process. Sequential Quadratic Programming (SQP), for instance, provides a forceful structure by iteratively solving quadratic programming subproblems, while the Interior Point Method (IPM) is widely used as a highly scalable and robust engine for solving the resulting convex LP problems within the LP relaxations of bilinear programs.

3.1.1 Karush–Kuhn–Tucker conditions

The Karush–Kuhn–Tucker (KKT) conditions form a fundamental set of optimality criteria used in constrained nonlinear programming. They generalize the method of Lagrange multipliers to handle inequality constraints and are widely applied in optimization problems across engineering and machine learning. The KKT provides necessary conditions for a solution to be optimal under mild regularity assumptions, and in certain cases, such as convex problems, these conditions also become sufficient for optimality (Clemptner, 2016).

In general, a nonlinear optimization problem may include both equality and inequality constraints. The KKT conditions unify these constraints through the introduction of Lagrange multipliers associated with each one. These multipliers quantify how sensitive the objective function is to changes in the constraints, allowing the optimizer to balance the goal of minimizing the objective against the requirement to satisfy the feasible region (Andreani *et al.*, 2024).

Accordingly, taking the minimization problem in Equation 3.1, the gradient of the Lagrangian function, which consists of the objective function and the constraints multiplied by the Lagrange multipliers, in relation to x must be equal to zero.

$$\nabla f(x) + \sum_{i=1}^{m_h} \lambda_i h_i(x) + \sum_{j=1}^{m_g} \mu_j g_j(x) = 0 \quad (3.5)$$

where λ_i and μ_j are the Lagrange multipliers representing the equality and inequality constraints, respectively.

Another key component of the KKT framework is complementary slackness, which links the inequality constraints to their associated multipliers. Intuitively, it states that either a constraint is active at the optimum (meaning it holds with equality) or its multiplier must be zero. This principle formalizes the idea that only constraints limiting the optimal solution exert influence on the objective value. Inactive constraints, by contrast, impose no pressure on the solution and therefore play no role in determining local optimality. Mathematically speaking,

$$\mu_j g_j(x) = 0, \quad j = 1, \dots, m_g \quad (3.6)$$

and

$$\mu_j \geq 0, \quad j = 1, \dots, m_g \quad (3.7)$$

Taken together, the KKT conditions form a useful tool for characterizing and verifying optimal solutions in constrained optimization problems. However, several challenges remain. In many nonlinear settings, the existence of multiple local minima can make it difficult to identify the true global optimum, even when the KKT conditions are satisfied. Moreover, solving the KKT system can be computationally demanding, often requiring sophisticated numerical methods and careful initialization to ensure convergence to a meaningful solution (Dempe; Zemkoho, 2014).

3.1.2 Interior Point Method

The Interior-Point Method is an optimization technique widely used to solve both LP and NLP problems. In contrast to classical approaches such as the simplex method (Aldino; Ulfa, 2021), which progresses along the boundary of the feasible region, IPM navigates through the interior of that region, which gives the method its name (Nemirovski; Todd, 2008). The fundamental principle behind IPM consists of beginning from a strictly interior point, which is a point that satisfies all constraints with strict inequalities, and following a continuous path that remains within the feasible region while iteratively approaching the optimal solution. Given a primal LP problem:

$$\text{minimize } c_{IMP}^T x \quad \text{subject to } A_{IMP}x = b_{IMP}, x \geq 0 \quad (3.8)$$

where the notation c_{IMP}^T indicates the transpose of the vector c_{IMP} , which is a column vector of coefficients, and every one of its elements corresponds to each variable in the objective function. A_{IMP} is a matrix representing the coefficients of the constraints, while b_{IMP} is a vector that defines the right-hand side of the bounds.

The IPM often relies on a "barrier function" to ensure that iterates remain strictly within the feasible region throughout the optimization process. The barrier function imposes a penalty that grows rapidly as the solution approaches the boundary, discouraging any step that would violate the constraints. Among the various barrier functions, the logarithmic barrier is one of the most widely used (Aquib, 2021). It is defined as follows:

$$\phi = - \sum_{i=1}^m \log(x_i). \quad (3.9)$$

Then, the modified objective function can be rewritten as

$$F(x) = c_{IMP}^T x - M \sum_{i=1}^k \log(x_i) \quad (3.10)$$

where M is the barrier parameter, controlling the proximity to the boundary of the feasible region. As M approaches zero, the barrier function becomes an exact penalty function for the infeasibility of x .

The steps for using the IPM begin with initialization, in which the method requires a start with an initial point strictly within the feasible region and not lying on the boundary of the constraints. Then, using the gradient and the Hessian of $F(x)$, it is possible to determine the direction in which the function decreases most rapidly. At each step, a modified version of the original problem that includes the barrier function has to be solved; this is called the iteration step. Consequently, the iteration and updated steps are repeated until the solution converges to an optimal point. Convergence is typically determined by checking if the changes in the objective function or the point are below a certain threshold (Jiang *et al.*, 2020).

The IPM is particularly useful for large-scale problems and for those where the feasible region is complex. It is flexible and can be adapted to a range of optimization problems. A natural extension of the IPM is the primal-dual interior point method (Mehrotra, 1992), combining the fundamentals of IPM with the concept of duality. This method is highly effective for solving LP problems and more general convex optimization problems. It seeks to solve both the primal and dual problems simultaneously, which often results in better computational efficiency (Bao *et al.*, 2024). Given a primal LP problem by Equation 3.8, its dual is

$$\text{maximize } b_{IMP}^T y \quad \text{subject to } A_{IMP}^T y \leq c_{IMP} \quad (3.11)$$

where y is the dual variable of x . The central idea of IPM is to solve both primal and dual problems simultaneously by following the central path, which is a trajectory of feasible solutions lying strictly inside the feasible region (Papp; Varga, 2025). As with the original IPM, a logarithmic barrier function is used to ensure that solutions remain within the feasible region. However, in this case, the objective is to minimize the primal objective function plus the barrier term, while also maximizing the dual objective function. Therefore, the associated Lagrangian function is given by:

$$L(x, y, sv) = c_{IMP}^T x - b_{IMP}^T y + M \sum_{i=1}^k \log(x_i) \quad (3.12)$$

where sv represents the slack variables which are associated with the constraints of the dual problem. The complementary slackness conditions are a set of optimality conditions that characterize the solution to an LP problem. These conditions establish relationships between the primal and dual variables and play a crucial role in the duality theory of LP (Shintaku *et al.*, 2024). The primal problem seeks to minimize a linear objective function subject to linear equality and inequality constraints. In contrast, the dual problem seeks to maximize another linear function subject to its constraints. The

complementary slackness conditions relate the primal and dual variables and indicate when each constraint is "active" or "inactive" at the optimal solution (Tracy; Manchester, 2024).

One of the most popular algorithms for solving LP and NLP problems using IPM is the Primal-Dual Predictor-Corrector Algorithm. It iteratively updates primal and dual variables along the central path, adjusting the barrier parameter μ to ensure convergence. The complexity of IPM is typically polynomial, with the number of iterations required depending on the problem's size and properties. For LP problems, the complexity is often bounded by a polynomial in the number of variables and constraints. However, for specific NLP problems, the complexity can be higher due to the additional computational burden of solving nonlinear equations in each iteration (Wright, 1997).

Complementary slackness establishes an algebraic relationship between the optimal primal and dual solutions. These conditions arise naturally from the KKT method. Intuitively, complementary slackness links the activity of constraints in the primal problem to the values of the corresponding dual variables (Karim; Solomonik, 2022). Formally, the conditions state that for each inequality constraint, either the constraint is active (i.e., holds with equality at the optimum) and the associated dual variable may be strictly positive, or the constraint is inactive (i.e., has slack) and its dual variable must be zero. Conversely, if a dual variable is positive, the corresponding primal constraint must be binding; if the dual variable is zero, the constraint cannot be active.

Therefore, complementary slackness guarantees that at optimality, a primal variable and the slack of its associated dual constraint cannot both be strictly positive simultaneously, ensuring consistency between primal and dual feasibility, and helps reveal which constraints determine the optimal solution's structure.

3.1.3 Sequential Quadratic Programming

Another optimization technique for solving NLP is the Sequential Quadratic Programming, which iteratively solves a sequence of quadratic subproblems approximating the original nonlinear optimization problem, often achieving fast and reliable convergence to a local optimum (Boggs; Tolle, 1995). At each iteration, SQP constructs a quadratic approximation of the objective function, typically using a second-order Taylor expansion, and the constraints around the current point. The resulting quadratic subproblem is then solved to produce a search direction that satisfies the linearized constraints. This direction is subsequently used to update the current iterate, guiding the solution progressively closer to optimality (Na *et al.*, 2023).

At each iteration k , SQP solves the following quadratic programming subprob-

lem:

$$\min \quad \frac{1}{2}d^T H_k d + \nabla f(x_k)^T d \quad (3.13a)$$

$$\text{subjected to: } h_p(x_k) + \nabla h_p(x_k)^T d = 0, \quad p = 1, \dots, m_{SQP} \quad (3.13b)$$

$$lb \leq x_k^q + d^q \leq ub, \quad q = 1, \dots, n_{SQP} \quad (3.13c)$$

where x_k is the current variable being iterated in the search direction d for the objective function $f(x_k)$. Moreover, H_k represents the Hessian approximation of the Lagrangian function, while $h_p(x_k)$ are the equality constraints. Finally, lb and ub are the lower and upper bounds on the variables.

Sequential Quadratic Programming achieves convergence when the norm of the gradient of the Lagrangian becomes sufficiently small, indicating that the necessary conditions for optimality are nearly satisfied. Convergence is also assessed through feasibility and optimality criteria, most notably the satisfaction of the KKT conditions. Moreover, SQP methods are widely regarded for their high accuracy and strong theoretical foundations. They are particularly effective for nonlinear optimization problems with both equality and inequality constraints, thanks to their ability to incorporate constraint information directly into the quadratic subproblems (Liu; Reynolds, 2021).

3.2 LINEAR PROGRAMMING RELAXATION

While IPMs and SQP are highly efficient and robust tools for solving convex optimization problems or for rapidly locating a local optimum in nonlinear, constrained landscapes, their power diminishes significantly when confronting problems with highly non-convex objective functions, such as the energy minimization of a protein structure. The central limitation is their reliance on local search principles, as both methods are inherently driven by gradient and Hessian information from a starting point, which causes them to converge to the nearest minimum, ergo, a local optimum, and they possess no intrinsic mechanism to explore the entire feasible space for the global optimum.

This deficiency is particularly critical in computational biology, where the protein folding landscape is characterized by a vast number of energy minima, making it impossible to guarantee that the identified structure represents the true lowest energy state. Consequently, an option is to shift toward methods of global optimization, applying techniques such as bilinear optimization and LP relaxation to transform the non-convex problem into a solvable, or closely approximated, convex form that provides a certifiable lower bound or the exact global solution.

Linear Programming relaxation is a fundamental technique in mathematical optimization, which involves taking a difficult optimization problem, often an integer programming or a combinatorial optimization problem, and relaxing its constraints so

that variables are allowed to take continuous values instead of only integer values (Takabe; Hukushima, 2016). Therefore, it is used to approximate and analyze problems that are inherently combinatorial or involve discrete decision variables, transforming the original problem into an LP, which can be solved efficiently in polynomial time, providing an approximate solution or a bound for the original problem.

Many optimization problems impose integrality constraints such as $x_i \in \{0, 1\}$ or $x_i \in \mathbb{Z}$. These constraints typically render the problem as non-deterministic polynomial time-hard (NP-hard). In computer theory, a computational problem L is classified as NP-hard if, for every problem K that can be solved in NP, there exists a polynomial-time reduction from K to L, meaning that, assuming a solution for L requires one unit of time, L's solution can be used to solve K in polynomial time (Hochba, 1997).

As the feasible region becomes a discrete, non-convex set cannot be explored efficiently using standard convex optimization methods. LP relaxation mitigates this issue by loosening the integrality requirements, permitting the variables to take on continuous values within the same limits. This could be done, for instance, by replacing $x_i \in \{0, 1\}$ with $x_i \in [0, 1]$. The relaxation converts the original problem into a linear program, whose feasible region is a convex polyhedron encompassing all possible integer solutions as extreme or interior points (Roux *et al.*, 2025).

Well-known algorithms as the case of the simplex method and interior-point approaches, can effectively solve LPs in polynomial time. Also, a theoretical bound on the optimal value of the original integer program is provided by the optimal value of the relaxed problem, as the LP relaxation produces an upper bound for maximization problems and a lower bound for minimization ones (Sundar *et al.*, 2022). These bounds are essential for pruning the search space and speeding up convergence to the global optimum in branch-and-bound, branch-and-cut, and related exact optimization frameworks (Du *et al.*, 2025).

Beyond its role in exact solution methods, LP relaxation also offers structural insight into the underlying optimization problem. Which constraints are binding, if the formulation represents the intended geometric connections, and whether more legitimate inequalities or cutting planes might be needed to tighten the relaxation, can all be found in the relaxed solution.

When the ideal relaxed solution is already integral, the LP relaxation is said to be tight, when the integer program can be solved directly by its relaxation. In other situations, the relaxation could be loose, producing fractional answers that are very distinct from any possible integer solution (Dokka *et al.*, 2021). This gap serves as a marker for the formulation's quality and encourages the addition of reformulations, restrictions, or specialized relaxations such as McCormick envelopes for bilinear terms.

The convex and concave surface representation of a bilinear function over a rectangular domain is known as the McCormick envelope (Ye; Scott, 2023). The McCormick relaxation creates the tightest convex underestimator and concave overestimator that can be written using linear inequalities given a bilinear term xy with upper and lower bounds. These linear inequalities, known as the McCormick envelopes, enable the embedding of bilinear optimization problems into an LP or mixed-integer linear programming paradigm by forming a polyhedral outer approximation of the bilinear term (Leyffer; Manns, 2025).

However, the size of the variable domains has a significant impact on the quality of the McCormick relaxation. Large domains result in weaker envelopes and, hence, less accurate approximations of the underlying nonconvex function. Domain reduction techniques have been integrated into hybrid global–local optimization techniques and sophisticated metaheuristics to overcome this constraint.

Specifically, the integration of piecewise McCormick relaxations and their bilinear generalizations has led to significant improvements in domain reduction-based metaheuristics that have demonstrated success in scheduling and sequencing issues (Assis *et al.*, 2017). These piecewise relaxations considerably tighten the approximation by splitting the variable domains into smaller subregions and using McCormick envelopes inside each partition, albeit at the expense of adding more binary variables and constraints.

To improve the tractability of bilinear terms, various relaxation strategies have been devised in addition to McCormick-based approaches. One particularly useful example is multiparametric disaggregation. This technique uses parametric partitioning to break down a bilinear term into a series of auxiliary variables that encode detailed information about the variable interactions.

Multiparametric disaggregation, when used methodically, can produce relaxations that are provably tighter than traditional McCormick enclosures while preserving a structure that can be solved by mixed-integer programming solvers (Rocha *et al.*, 2021). This method has proven more effective than McCormick relaxations in several situations, particularly those involving highly nonlinear or irregular bilinear interactions, both in terms of computational performance and solution quality.

Overall, the techniques covered in this chapter emphasize the crucial difficulty of striking a balance between accuracy and tractability when working with bilinear and combinatorial structures. While McCormick envelopes and their generalizations introduce increasingly tight approximations capable of capturing the complex interdependence of the underlying variables, LP relaxation provides the initial layer of convexification. When combined, these methods provide the mathematical foundation needed to handle the intricacy of the PSP formulation created in this thesis.

For clarification, in the domain of global optimization, the core role of LP relaxation is to convert the difficult non-convex problem into a simpler convex problem that can be solved efficiently, often using approaches such as IPMs. This is achieved through a process where all non-linear, non-convex terms are replaced by a new variable and a set of linear inequalities, a technique commonly implemented through McCormick Envelopes. By relaxing the non-convex constraints, the feasible region is enlarged to form a convex polytope.

Critically, because the original feasible set is contained within this relaxed, enlarged set, solving the LP relaxation turns out to be an objective value that is a mathematically guaranteed lower bound on the true global optimum (for a minimization problem). This lower bound is essential, as it serves as the foundation for modern global optimization algorithms, such as Branch-and-Bound (Scavuzzo *et al.*, 2024), providing the necessary benchmark to prune vast portions of the search space and efficiently converge to the global solution.

3.3 ROOT MEAN SQUARE DEVIATION

The PSP problem is frequently evaluated using the Root-Mean-Square Deviation (RMSD), a widely adopted metric for quantifying structural similarity. RMSD measures the average distance between corresponding residues (or amino acids in this case) after optimally superimposing two structures, allowing a direct comparison between a predicted conformation and its reference (native) structure. The similarity is typically expressed in Angström (Å) or, more commonly in recent computational studies, in nanometers (nm) (Maruyama *et al.*, 2023).

The RMSD is computed as:

$$RMSD = \sqrt{\frac{\sum_{i=1}^n |ro_i - rp_i|^2}{n}} \quad (3.14)$$

where ro_i and rp_i represent the positions of amino acid i in the original and the predicted structures to be compared. Also, n is the number of residues involved, the same number as the sequence's length. Consequently, identical structures have an RMSD value of zero. The higher the RMSD value, the more divergent the structures are (Filipe; Loura, 2022).

Despite its intuitive interpretation and widespread use, RMSD has limitations that affect its reliability as a standalone evaluation metric. It is highly sensitive to localized structural deviations, such as flexible loops or terminal regions, which can significantly inflate the RMSD value even when the overall fold is preserved. RMSD also grows naturally with protein size, making comparisons across proteins of different lengths difficult (Sargsyan *et al.*, 2017).

It is also important to mention that RMSD is purely geometric and does not incorporate biochemical or functional considerations (Kasprzak *et al.*, 2025). Two structures may differ slightly in space yet preserve all relevant interactions, leading to a deceptively high RMSD. Conversely, a structure with a low RMSD might incorrectly reproduce key interactions such as hydrogen bonds, salt bridges, or hydrophobic packing.

4 RELATED WORKS

As previously discussed, PSP plays a fundamental role in understanding biological function and has far-reaching implications for drug discovery and disease treatment. Consequently, this area of interest has attracted sustained attention across structural biology, bioinformatics, and computational biology over the past few decades (Schauperl; Denny, 2022). Nevertheless, the immense conformational search space and the highly intricate energy landscapes associated with protein folding pose significant computational challenges, motivating the development and continuous refinement of diverse modeling strategies and algorithmic techniques (Custódio *et al.*, 2014).

Among these, the AB off-lattice model, which strikes a compromise between simplicity and the capacity to capture crucial interactions, has become a compelling tool for researching protein folding mechanisms and predicting protein structures (Silva *et al.*, 2024). The AB off-lattice model first emerged in the work of Stillinger *et al.* (Stillinger *et al.*, 1993), which introduced a continuum model to examine the generic properties of protein folding. By providing insights that are hard to gain with more complex representations, their work opened the door for simplified models to understand the fundamental concepts of protein folding.

Several artificial intelligence approaches have been developed to tackle the difficulties posed by the PSP problem. The three main computing frameworks PSP uses are metaheuristics, ML, and DL. Additionally, methods for protein structure prediction, such as homology-based, threading, and AB initio, can be divided into groups according to how biologically approximate they are (Márquez-Chamorro *et al.*, 2015). This chapter compresses some of the most essential work on the PSP problem using the AB off-lattice model in the last decade.

4.1 MACHINE AND DEEP LEARNING IN THE PSP PROBLEM

Numerous studies have been conducted to date that have used conventional ML techniques to address a variety of computational biology-related issues, including protein classification, structure prediction, and interaction prediction (Cheng *et al.*, 2008). The most popular ML techniques for classifying protein interactions are support vector machines (SVMs), random forests (RFs), and artificial neural networks (ANNs). These algorithms use hand-engineered features as inputs, primarily generated from underlying protein sequences using distribution patterns of amino acids, physicochemical qualities, and evolutionary data (Jha *et al.*, 2022).

Recently, significant breakthroughs due to the growing usage of ANNs for

structure prediction pipelines have been witnessed, which replace calculations formerly based on energy models and sampling techniques. These models have extracted physical contacts from the evolutionary record, distilled sequence-structure patterns from known structures, and incorporated templates from homologs in the PDB (AlQuraishi, 2021). Notably, using contact information in protein quality assessment is beneficial for PSP, as demonstrated by a single-model quality assessment method that applies residue-residue contact information predicted by two protein ANN methods to generate a new feature score for quality assessment (Cao *et al.*, 2017).

SVM classifier based on physicochemical features and position-specific scoring matrix has been applied to extended early-stage PSP (Fabian; Stapor, 2017). Similar techniques have also been used to predict a protein's secondary structure by recognizing patterns of amino acid sequences (Kurniawan *et al.*, 2017). Therefore, it is safe to state that SVMs have been widely used in the PSP problem, demonstrating a high degree of generalization (Akhter; Shehu, 2018).

An improved SVM with fuzzy has also been applied to secondary protein structure prediction, where it iterates the class centers in the feature space to create an optimal separation hyperplane. Then, the K-nearest neighbor algorithm assigns large membership values to sample points near this hyperplane and small membership values to outliers (Xie *et al.*, 2018). Furthermore, after extracting the key characteristics from the initial high-dimensional input vectors using a robust principal component analysis approach, SVM was experimented with a radial basis function kernel (Zakaria *et al.*, 2020).

Other interesting approaches include a method that generates pairwise sequence alignments for more accurate template-based modeling was proposed, in which an ML model is trained using the structural alignment of known homologs (Makigaki; Ishida, 2020). Also, an ANN model for secondary protein structure prediction based on swarm optimization has been evaluated; it offers the level of detail and insight required to focus on the search while accounting for both the local and global views (Akbar *et al.*, 2021a).

Hybrid recurrent neural networks have been applied to the PSP problem (Ema *et al.*, 2022). The sequence embedding from a supervised transformer protein language model is incorporated into a multi-scale network with knowledge distillation for single-sequence PSP, predicting inter-residue two-dimensional geometry, which is then used to reconstruct three-dimensional structures with energy minimization (Wang *et al.*, 2022). A simplified attention-based NN for protein angle prediction and visualization from amino acid sequences can be used to identify the important interactions between various protein sequence components that contribute to its folding, making it more effective on a small dataset while handling the complexity and variety of the structural data

(Franceschina, 2023).

The accuracy of PSP has greatly increased with the current use of deep neural networks for end-to-end model training and spatial constraint prediction (Pearce; Zhang, 2021a). DL techniques enable the extraction of complex features from protein sequence data. In 2020, noticeable progress was made in solving the PSP problem. The AlphaFold (Varadi *et al.*, 2022) program is a deep learning system that looks for alterations at various residues that appear to be correlated—the residues may not be sequential in the main chain—by scouring vast databases of related DNA sequences from various animals. The calculation of a contact map is made possible by the observed correlations, which suggest that the residues may be physically close to one another.

Then, AlphaFold 2, the program's second iteration, won the 2020 Critical Assessment of Protein Structure Prediction (CASP14) (Jumper *et al.*, 2021a). In essence, the first AlphaFold blends a guiding potential obtained from pattern recognition with local physics. Researchers discovered that the algorithm tended to overestimate interactions between adjacent residues. To further enhance its performance, they modified it in the second edition.

Deep convolutional neural networks (CNNs) have, likewise, been vastly applied to the PSP problem in the last decade. By gradually incorporating convergent protein property representations from lower levels, deep CNNs may learn high-level relational characteristics from paired protein sequences. Additionally, several CNNs can obtain robust protein structure similarities from various angles (Bai; Yang, 2017). A variety of CNN designs, including dropouts, dilated convolutions, and wide residual networks, using publicly available datasets was created, examining several deep learning architectures to increase the accuracy of medium- and long-range interaction (Adhikari, 2020). Convolutional residual neural networks have also been successfully applied to the PSP problem (Xu *et al.*, 2021).

Recently, an end-to-end differentiable recurrent geometric network that uses a protein language model to learn hidden structural information from unaligned proteins was established, in which a linked geometric module compactly depicts the C_α backbone geometry invariantly (Chowdhury *et al.*, 2022). To forecast proteins that undergo conformational changes or for which few homologous sequences are known, a modified version of the AlphaFold2 algorithm was developed, integrating experimental distance restraint information into its network design (Stahl *et al.*, 2023).

A pipeline for uniform protein single- and multichain multiple sequence alignment construction was reported using iterative alignment search through genomic and metagenome sequence databases. Furthermore, data analyses also reveal that this model's advantage is in its balanced alignment search, efficient model selection, and ability to integrate large metagenomics databases (Zheng *et al.*, 2024). Other research

employs a two-dimensional fusion deep CNN and a supervised transformer protein language model for single-sequence PSP, where the protein transformer binding site matrix is predicted by combining the two training features, and the three-dimensional structure is rebuilt by applying energy minimization (Zhou *et al.*, 2024).

More recently, an open-source Jupyter Notebook inside Google Colaboratory was developed based on AlphaFold. This program is a command-line tool that facilitates the usage of AlphaFold 2 while exposing its advanced options, including the shortening of turnaround times of experiments by its optimized management of AlphaFold 2 models (Kim *et al.*, 2025).

Even though the creation of AlphaFold was a breakthrough towards solving the PSP problem, several researchers address challenges that still need to be solved. Some of these challenges include problems regarding proteins with limited evolutionary data or complex molecular interactions (Park *et al.*, 2025). It is important to remember that projected structural models are not validated and cannot be regarded as the ground truth, even though they have the potential to greatly expedite and influence fundamental research (Varadi *et al.*, 2025).

Several reviews have also been published exploring the application of DL methodologies and tools in the PSP problem (Meng *et al.*, 2025). Comparative overviews also delve into works derived from the AlphaFold (Malhotra *et al.*, 2025). More extensive reviews are also worth noting, as they analyse not only ANN-based methods, but also those which incorporate evolutionary computation, because in light of the rapid and substantial advances in AI-based methods for PSP, it is important to systematically survey this progress in order to synthesize existing research insights and provide informed guidance for future developments (Zhan *et al.*, 2025).

This section lists recent PSP works based on the 3D AB off-lattice, given in Table 2.

Table 2 – Most notorious papers on DI for PSP

Model	Architecture	Article
AlphaFold	Spacial CNN	(Senior <i>et al.</i> , 2019)
AlphaFold 2	Evolutionary Transformer	(Jumper <i>et al.</i> , 2020)
RoseTTAFold	Three-track Neural Network ResNet	(Board, 2021)
AlphaFold 3	Pair Transformer + Diffusion Model	(Abramson <i>et al.</i> , 2024)

AlphaFold-based methods represent the current state of the art in data-driven protein structure prediction, particularly for backbone modeling. AlphaFold 2 and AlphaFold 3 achieve exceptionally high accuracy for single-chain proteins and many complexes, generally outperforming earlier versions and alternative deep learning approaches, with a typical accuracy ranking of AlphaFold 3 $\geq$ AlphaFold 2 > RoseTTAFold

> AlphaFold 1. However, despite their success, these methods exhibit notable limitations: they rely heavily on evolutionary information and therefore struggle with orphan proteins that lack homologous sequences, and they provide limited insight into protein dynamics and alternative conformational states.

From a practical perspective, usability also differs substantially among these tools. While AlphaFold 2 is open-source and widely adopted, AlphaFold 3 is primarily accessible through restricted server-based platforms, which can limit reproducibility and large-scale experimentation. In contrast, RoseTTAFold offers a comparatively faster and more accessible open-source alternative, albeit with a modest trade-off in predictive accuracy. These considerations highlight that, despite their impressive performance, DL-based predictors are not universally applicable and motivate the continued exploration of complementary approaches.

4.2 METAHEURISTICS IN THE PSP PROBLEM

Classical gradient-based mathematical techniques are generally unsuitable for addressing the PSP problem, as its energy function is typically non-differentiable and highly irregular (Song *et al.*, 2018). Consequently, population-based metaheuristics, particularly evolutionary algorithms (EAs) and swarm intelligence (SI) methods, have been extensively explored in the literature as effective alternatives for tackling PSP (Dubey *et al.*, 2016). These approaches are well-suited to navigating the complex, high-dimensional, and non-convex energy landscapes characteristic of protein folding (Jana *et al.*, 2018a).

Within the SI paradigm, particle swarm optimization (PSO) models have also been widely used, as they are proposed to collectively learn and share information among particles to efficiently traverse high-dimensional search spaces (Falae, 2025). While artificial bee colony (ABC) algorithms emulate foraging behaviors to balance exploration and exploitation when searching for low-energy protein conformations (Akram *et al.*, 2024). More recently, bio-inspired swarm methods such as manta ray foraging optimization (MRFO) have been proposed, simulating chain, cyclone, and somersault foraging strategies to enhance exploration and convergence in complex, multimodal landscapes (Zhao *et al.*, 2020).

Among EA, genetic algorithms (GAs) are among the earliest and most widely adopted approaches for PSP; they apply crossover and mutation operators to explore the conformational space and refine candidate structures (Hoque *et al.*, 2009). Differential Evolution (DE) has also been investigated due to its strong global search capability and simple parameterization, showing competitive performance in continuous and mixed representations of protein conformations (Narloch; Dorn, 2021). Collectively, these metaheuristic strategies have demonstrated robustness in coping with the multimodality

and ruggedness of PSP energy landscapes, albeit often at the cost of high computational effort and limited guarantees of global optimality (Siqueira; Venske, 2021).

In 2014, an approach combining simulated annealing and tabu search algorithms was introduced (Lin *et al.*, 2014), where a variety of tactics were effectively used to find the ideal conformation quickly. In the same year gravitational search algorithm, bat algorithm, PSO, and ABC were the four algorithms compared (Parpinelli *et al.*, 2014), in which two metrics were used to assess the algorithms: processing time and solution quality. The PSO algorithm offered the overall optimum balance, according to the results. Furthermore, PSO and the gravitational search algorithm were promised to develop even better solutions with additional iterations.

A combination of the PSO, GA, and tabu search algorithms was also presented (Zhou *et al.*, 2014); tests reveal that the suggested approach performs more accurately in determining the protein sequence energy value than individual algorithms, particularly for lengthy protein sequences. Nevertheless, the proposed technique requires additional processing costs and function evaluations. Other metaheuristics, such as the firefly-based model, were also applied (Maher *et al.*, 2014).

The PSP problem was also recast as a numerical optimization challenge using the AB off-lattice model (Li *et al.*, 2015a). Recent advancements have introduced a computational approach to enhance the accuracy of predicting native-like three-dimensional protein structures, using conformational preferences of amino acid residues and secondary structure information from templates stored in the PDB (Borguesan *et al.*, 2015), represented as an angle probability list. This study presents two knowledge-based prediction methods: GA and PSO, to test diverse protein case studies. Stereochemical and structural analyses of the predicted structures indicate that employing the Angle Probability List significantly improves the metaheuristics' efficiency by narrowing the conformational search space.

Several metaheuristic algorithms have been applied to address the PSP problem, yet the optimal choice of algorithm is often unclear; therefore, fitness landscape analysis was proposed to offer a solution by characterizing the problem's structural features to recommend the most suitable algorithm (Jana *et al.*, 2017). DE was considered one of the best search techniques (Narloch; Parpinelli, 2017a). To improve accuracy for the prediction of a protein's tertiary structure with particular emphasis on accurately modeling coils and turns, a structure pattern library was built from experimentally determined PDB structures to constrain the conformational search space and integrated with an adapting differential evolution metaheuristic, whose experimental results show improved structural accuracy measured by lower RMSD (Oliveira *et al.*, 2017).

An approach that balances the conflict of exploration and exploitation in the evolutionary process by implementing it in an interleaved manner was also suggested,

which used qualitative and quantitative comparisons with existing hybridizations of PSO and DE, as well as comprehensive learning PSO (Jana *et al.*, 2018d). A hybrid approach integrating DE and local refinement of protein structures through fragment replacements was executed for PSP, utilizing the coarse-grained protein conformation representation within the Rosetta environment (Varela; Santos, 2019).

A DE algorithm that integrates cooperative global exploration and local exploitation through multiple mutation strategies and an underestimation-assisted selection mechanism was employed to predict proteins' 3D structure (Zhou *et al.*, 2019). The method alternates between global and local phases within each generation, using adaptive underestimation models and self-adjusted parameters to efficiently filter candidate solutions and accelerate convergence. Another DE-based proposal was a surrogate-assisted differential evolution algorithm for PSP that aims to reduce the high computational cost by integrating gene expression programming to generate diverse candidate conformations, radial basis function guides the DE search efficiently, and covariance matrix adaptation evolution strategy to enhance exploration of the conformational space (Rakhshani *et al.*, 2019).

Grasshopper optimization algorithm and some chaotic variations were proposed for the PSP problem within the scope of AB-OFF lattice model, where the idea was to bridge the mechanisms of the metaheuristic between exploratory and exploitative stages, while using the variants in distinct chaotic maps (Saxena; Kumar, 2020). GA was also efficiently employed in order to predict proteins' tertiary structures from amino acid sequences, with experimental results indicating that the proposed algorithm can achieve comparable or superior structural accuracy, albeit with higher computational cost, highlighting a trade-off between accuracy and runtime and motivating the inclusion of a post-processing enhancement step to refine and stabilize predicted structures (Yousef *et al.*, 2020).

A computational framework for protein secondary structure prediction that combines PSO, fuzzy sets, and a neural network model to address limitations of existing prediction tools was proposed, trained, and evaluated on six standard benchmark datasets, demonstrating robust performance across multiple statistical measures and cross-validation schemes (Akbar *et al.*, 2021b). A many-objective evolutionary algorithm was also applied to solve the PSP problem, addressing the limitations of single energy-based formulations, as the algorithm formulates PSP as a four-objective optimization problem (Lei *et al.*, 2021).

Other metaheuristics were also used in the PSP problem and were proven to be efficient, as was the case of the Jaya algorithm (Silva, 2021) and hybridized whale optimization (Saxena *et al.*, 2022). PSO and tabu search were once more combined in building neighborhood functions for a streamlined PSP approach (Shuchun *et al.*,

2022). More interesting mentions in recent algorithm applied to the PSP problem include the butterfly optimization algorithm (Gonzalez-Sanchez *et al.*, 2023), the elitist animal migration optimization (Ezgi, 2023), and the ingenious crow search algorithm (Alshamrani *et al.*, 2023) have been utilized for the PSP problem.

An adaptive archive-based mutation strategy was proposed to balance exploration and exploitation, modeling the PSP problem as a multi-objective with three knowledge-based energy functions and EA. This strategy uses different mutation operators in various evolutionary stages adaptively. To speed up convergence, a mixed individual transfer technique is suggested for sharing search data across several populations (Hong *et al.*, 2024).

The application of PSO for PSP and refinement as an alternative to costly experimental methods was investigated, presenting a SI-based computational framework that searches the protein conformational space by optimizing energy functions to identify low-energy, stable structures (Falae, 2025). Another approach to address the NP-hard protein folding problem was proposed using a hybrid bio-inspired optimization algorithm that combines the jumping spider optimization algorithm with golden ratio-based simulated annealing for energy minimization in peptide structure prediction (Soto-Monterrubbio *et al.*, 2025). It further enhances this framework by integrating CNN for fragment prediction.

This section lists recent PSP works based on the 3D AB off-lattice, given in Table 3.

Table 3 – Recent papers on metaheuristics approaches to PSP

Metaheuristic	Articles
DE	(Narloch; Parpinelli, 2017b) (Oliveira <i>et al.</i> , 2017) (Jana <i>et al.</i> , 2018c) (Zhou <i>et al.</i> , 2019) (Rakhshani <i>et al.</i> , 2019) (Varela; Santos, 2022) (Stubs Parpinelli <i>et al.</i> , 2023)
GA	(Sar; Acharyya, 2014) (Rashid <i>et al.</i> , 2015) (Borkar; Mahajan, 2018) (Yousef <i>et al.</i> , 2020) (Marchi; Parpinelli, 2021)
PSO	(Băutu; Luchian, 2010) (Chen <i>et al.</i> , 2011) (Jana; Sil, 2013) (Song <i>et al.</i> , 2018) (Akbar <i>et al.</i> , 2021b) (Lei <i>et al.</i> , 2021) (Falae, 2025)
ABC	(Li <i>et al.</i> , 2015a) (Li <i>et al.</i> , 2015b) (Saxena; Kumar, 2020) (Suma; Vinod Chandra, 2021) (Sekhar <i>et al.</i> , 2023)

5 BILINEAR OPTIMIZATION IN PSP

This chapter presents a novel methodology designed to improve the computational treatment of the PSP problem. The proposed approach reformulates the classical AB off-lattice model into a bilinear optimization framework. To this end, the chapter provides a detailed, step-by-step description of the procedures used to perform this transformation and to integrate the resulting formulation into an optimization-driven PSP pipeline. The motivation for this reformulation stems from the inherent complexity of the AB off-lattice model, which is known to be NP-hard and challenging to optimize directly (Dorn *et al.*, 2014).

By expressing the PSP problem as a bilinear optimization problem, the goal is to create a structure that is more amenable to advanced mathematical programming techniques while preserving the essential physical characteristics of protein folding. This transformation opens new opportunities for exploring rugged energy landscapes, facilitating more accurate identification of near-native conformations. Accordingly, this method uses extensive datasets available in PDB to refine predictions and improve accuracy.

As previously seen, the bond or curvature angles between amino acids are given by vector $\theta = (\theta_1, \dots, \theta_{n-2})$ with n being the sequence length and $\theta_i \in [-\pi/2, \pi/2]$ for all $i \in \mathcal{I} = \{1, \dots, n-2\}$. Additionally, the sequencing of amino acids depends on the torsion angles, which are collected by vector $\beta = (\beta_1, \dots, \beta_{n-3})$, where $\beta_j \in [-\pi/2, \pi/2]$ for all $j \in \mathcal{J} = \{1, \dots, n-3\}$. Therefore, the n length protein conformation determines the $(2n-5)$ optimization variables in the problem, which are the angle parameters given by θ and β .

Firstly, it is possible to represent positions using Cartesian coordinates rather than angles and trigonometric functions to address sequencing problems. In this representation, the variables capture the projections, specifically the $\sin(\cdot)$ and $\cos(\cdot)$ (Camponogara *et al.*, 2019), and the idea is to represent them as their own variable. The formal expression for this variable transformation is given by:

$$\begin{cases} s_{\theta,i} = \sin(\theta_i) \\ c_{\theta,i} = \cos(\theta_i) \end{cases}, \quad \forall i \in \mathcal{I} \quad (5.1)$$

and

$$\begin{cases} s_{\beta,j} = \sin(\beta_j) \\ c_{\beta,j} = \cos(\beta_j) \end{cases}, \quad \forall j \in \mathcal{J} \quad (5.2)$$

For the change of variable to be consistent, it is necessary to assume the

following constraints:

$$\begin{cases} s_{\theta,i}^2 + c_{\theta,i}^2 = 1 \\ s_{\theta,i} \in [-1, 1] \\ c_{\theta,i} \in [0, 1] \end{cases}, \quad \forall i \in \mathcal{I} \quad (5.3)$$

and

$$\begin{cases} s_{\beta,j}^2 + c_{\beta,j}^2 = 1 \\ s_{\beta,j} \in [-1, 1] \\ c_{\beta,j} \in [0, 1] \end{cases}, \quad \forall j \in \mathcal{J} \quad (5.4)$$

Next, given that $\theta_i, \beta_j \in [-\pi/2, \pi/2]$, it is logical to restrict $c_{\theta,i}, c_{\beta,j} \in [0, 1]$. In three dimensions, the Cartesian coordinates for a residue of amino acid i in an AB off-lattice model are modeled in terms of bilinear terms involving the decision variables; that being said, they are defined by the following equations:

$$(x_i, y_i, z_i) = \begin{cases} (0, 0, 0) & \text{for } i = 1 \\ (0, 1, 0) & \text{for } i = 2 \\ (c_{\theta,1}, s_{\theta,1} + 1, 0) & \text{for } i = 3 \\ (x_{i-1} + c_{\theta,i-2}c_{\beta,i-3}, y_{i-1} + s_{\theta,i-1}c_{\beta,i-3}, z_{i-1} + s_{\beta,i-3}) & \text{for } 4 \leq i \leq n \end{cases} \quad (5.5)$$

By following the same encoding ξ_i as discussed in section 2.1, the intramolecular potential energy function Φ for any protein sequence of length n is expressed as

$$\Phi = V_1 + \sum_{i=2}^{n-1} \sum_{j=i+2}^n V_2(\tilde{\rho}_{ij}, \xi_i, \xi_j) \quad (5.6)$$

where V_1 is the potential curvature energy, defined by:

$$V_1 = \sum_{i=2}^{n-1} \frac{1 - c_{\theta,i}}{4} \quad (5.7)$$

The non-connected interaction V_2 adopts a species-dependent Lennard-Jones form. This includes a weakly attractive component at long distances (representing the van der Waals energy) and a strongly repulsive component at short distances, ensuring that:

$$V_2(\tilde{\rho}_{ij}, \xi_i, \xi_j) = 4(\tilde{\rho}_{i,j}^6 - C(\xi_i, \xi_j)\tilde{\rho}_{i,j}^3) \quad (5.8)$$

and

$$C(\xi_i, \xi_j) = \frac{1}{8}(1 + \xi_i + \xi_j + 5\xi_i\xi_j) \quad (5.9)$$

$$\rho_{i,j}\tilde{\rho}_{i,j} = 1 \quad (5.10)$$

where $\rho_{i,j} = r_{i,j}^2$ is the variable representing in this case the distance between residues and its counterpart $\tilde{\rho}_{i,j} = r_{i,j}^{-2}$, which implies that $\tilde{\rho}_{i,j} = 1/\rho_{i,j}$. Notice that $\rho_{i,j}$ is the

squared Euclidian distance $r_{i,j}$ between the i^{th} and j^{th} residues of the polypeptide chain, mathematically:

$$\rho_{ij} = (x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2 \quad (5.11)$$

By combining the stated equations and relations, the following BL optimization problem is formulated:

$$\min \quad \Phi = V_1 + \sum_{i=2}^{n-1} \sum_{j=i+2}^n V_{2,i,j} \quad (5.12a)$$

$$\text{subjected to: } s_{\theta,i}^2 + c_{\theta,i}^2 = 1, \forall i \in \mathcal{I} \quad (5.12b)$$

$$s_{\beta,j}^2 + c_{\beta,j}^2 = 1, \forall j \in \mathcal{J} \quad (5.12c)$$

$$(x_i, y_i, z_i) = \begin{cases} (0, 0, 0) & \text{for } i = 1 \\ (0, 1, 0) & \text{for } i = 2 \\ (c_{\theta,1}, s_{\theta,1} + 1, 0) & \text{for } i = 3 \\ (x_{i-1} + c_{\theta,i-2}c_{\beta,i-3}, y_{i-1} + s_{\theta,i-1}c_{\beta,i-3}, \\ \quad z_{i-1} + s_{\beta,i-3}) & \text{for } 4 \leq i \leq n \end{cases} \quad (5.12d)$$

$$\begin{cases} \rho_{i,j} = (x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2 \\ \rho_{i,j}\tilde{\rho}_{i,j} = 1 \\ \tilde{\rho}_{2,i,j} = \tilde{\rho}_{i,j}^2 \\ \tilde{\rho}_{3,i,j} = \tilde{\rho}_{2,i,j}\tilde{\rho}_{i,j} \\ V_{2,i,j} = 4\tilde{\rho}_{3,i,j}(\tilde{\rho}_{3,i,j} - C(\xi_i, \xi_j)) \end{cases}, \quad \begin{matrix} i = 1, \dots, n-2, \\ j = i+2, \dots, n \end{matrix} \quad (5.12e)$$

$$\begin{cases} s_{\theta,i} \in [-1, 1] \\ c_{\theta,i} \in [0, 1] \end{cases}, \quad \forall i \in \mathcal{I} \quad (5.12f)$$

$$\begin{cases} s_{\beta,j} \in [-1, 1] \\ c_{\beta,j} \in [0, 1] \end{cases}, \quad \forall j \in \mathcal{J} \quad (5.12g)$$

$$\begin{cases} x_i \in [x_i^{\min}, y_i^{\max}] \\ y_i \in [y_i^{\min}, y_i^{\max}] \\ z_i \in [z_i^{\min}, z_i^{\max}] \end{cases}, \quad \forall i \in \mathcal{I} \quad (5.12h)$$

$$\begin{cases} \rho_{i,j} \in [\rho_{i,j}^{\min}, \rho_{i,j}^{\max}] \\ \tilde{\rho}_{i,j} \in [\tilde{\rho}_{i,j}^{\min}, \tilde{\rho}_{i,j}^{\max}] \\ \tilde{\rho}_{2,i,j} \in [(\tilde{\rho}_{i,j}^{\min})^2, (\tilde{\rho}_{i,j}^{\max})^2] \\ \tilde{\rho}_{3,i,j} \in [(\tilde{\rho}_{i,j}^{\min})^3, (\tilde{\rho}_{i,j}^{\max})^3] \end{cases}, \quad \begin{matrix} i = 1, \dots, n-2, \\ j = i+1, \dots, n \end{matrix} \quad (5.12i)$$

$$\text{with } V_1 = \sum_{i=2}^{n-1} \frac{1 - c_{\theta,i}}{4} \quad (5.12j)$$

The bounds on the x_i , y_i , and z_i variables can be computed as follows:

$$\text{Lower bounds: } \begin{cases} (x_1^{\min}, y_1^{\min}, z_1^{\min}) = (0, 0, 0) \\ (x_2^{\min}, y_2^{\min}, z_2^{\min}) = (0, 1, 0) \\ (x_3^{\min}, y_3^{\min}, z_3^{\min}) = (-1, 0, 0) \\ (x_i^{\min}, y_i^{\min}, z_i^{\min}) = (-i + 2, -i + 3, -i + 3) \quad \text{for } 4 \leq i \leq n \end{cases} \quad (5.13)$$

and

$$\text{Upper bounds: } \begin{cases} (x_1^{\max}, y_1^{\max}, z_1^{\max}) = (0, 0, 0) \\ (x_2^{\max}, y_2^{\max}, z_2^{\max}) = (0, 1, 0) \\ (x_3^{\max}, y_3^{\max}, z_3^{\max}) = (1, 2, 0) \\ (x_i^{\max}, y_i^{\max}, z_i^{\max}) = (i - 2, i - 1, i - 3) \quad \text{for } 4 \leq i \leq n \end{cases} \quad (5.14)$$

Consequently, it is possible to infer bounds on the $\rho_{i,j}$ variables, assuming that $\rho_{i,j,l}^{\min} > 0$, $\tilde{\rho}_{i,j,l}^{\max} = 1/\rho_{i,j,l}^{\min}$ and $\tilde{\rho}_{i,j,l}^{\min} = 1/\rho_{i,j,l}^{\max}$.

For $i = 1, \dots, n - 2$, and $j = i + 2, \dots, n$, let:

$$\rho_{i,j}^{\min} = \rho_{i,j,x}^{\min} + \rho_{i,j,y}^{\min} + \rho_{i,j,z}^{\min}, \quad (5.15)$$

where

$$\begin{cases} \rho_{i,j,l}^{\min} = 0, & \text{if } [l_i^{\min}, l_i^{\max}] \cap [l_j^{\min}, l_j^{\max}] = \emptyset \\ \rho_{i,j,l}^{\min} = \min\{|l_j^{\min} - l_i^{\max}|, |l_i^{\min} - l_j^{\max}|\}^2, & \text{otherwise} \end{cases}, \quad \text{for } l \in \{x, y, z\}. \quad (5.16)$$

and

$$\rho_{i,j}^{\max} = \rho_{i,j,x}^{\max} + \rho_{i,j,y}^{\max} + \rho_{i,j,z}^{\max}, \quad (5.17)$$

where

$$\rho_{i,j,l}^{\max} = \max\{l_j^{\max} - l_i^{\min}, l_i^{\max} - l_j^{\min}\}^2, \quad l \in \{x, y, z\}. \quad (5.18)$$

An alternative way of computing bounds for the variables consists of minimizing (maximizing) the value of a given variable to obtain a lower (upper) bound while being subject to the constraints of the baseline problem (5.12). The resulting problem must be solved optimally to yield a valid bound. Thus, the computation of such optimization-based bounds (Castro, 2015) rests on effectively solving a relaxation of the constraint set, for instance, a convex set.

To this end, an LP relaxation can be obtained by replacing the BL terms with McCormick envelopes (McCormick, 1976). Consider a general BL terms $x \cdot y$ for which lower (upper) bounds x^L (x^U) and y^L (y^U) are known. Then, the McCormick envelope uses a new variable $w = x \cdot y$ to replace the BL term, with w being constrained or enveloped by the following linear underestimators and overestimators:

$$\begin{cases} w \geq x^L y + x y^L - x^L y^L \\ w \geq x^U y + x y^U - x^U y^U \\ w \leq x^U y + x y^L - x^U y^L \\ w \leq x y^U + x^L y - x^L y^U \end{cases} \quad (5.19)$$

The end product of solving an LP issue is a lower (upper) bound for every variable. In addition to producing optimization-based bounds for the decision variables, McCormick envelopes can generate LP relaxations when incorporated into a spatial branch-and-bound algorithm. A piecewise McCormick envelope technique, which produces a Mixed-Integer Linear Programming relaxation, is used to provide tighter relaxations by splitting the domain of a single variable into partitions and then estimating the BL term inside each partition (Castro; Grossmann, 2014).

Figure 6 explores a more detailed comparison of how the bilinear approach differs from the original AB off-lattice model.

Figure 6 – Comparison of the original model versus the BL approach

Original	Bilinear Approach
Variables $\theta_i \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \forall i \in \mathcal{I}$ $\beta_j \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \forall j \in \mathcal{J}$	Variables $s_{\theta,i} = \sin(\theta_i) \in [-1, 1] \forall i \in \mathcal{I}$ $c_{\theta,i} = \cos(\theta_i) \in [0, 1] \forall i \in \mathcal{I}$ $s_{\beta,j} = \sin(\beta_j) \in [-1, 1] \forall j \in \mathcal{J}$ $c_{\beta,j} = \cos(\beta_j) \in [0, 1] \forall j \in \mathcal{J}$
Objective $\min \quad \Phi = \sum_{i=1}^{n-2} V_1(\theta_i) + \sum_{i=2}^{n-1} \sum_{j=i+2}^n V_2(r_{ij}, \xi_i, \xi_j)$ $V_1(\theta_i) = \sum_{i=2}^{n-1} \frac{1 - \cos(\theta_i)}{4}$ $V_2(r_{ij}, \xi_i, \xi_j) = 4(r_{ij}^{-12} - C(\xi_i, \xi_j), r_{ij}^{-6})$	Objective $\min \quad \Phi = V_1 + \sum_{i=2}^{n-1} \sum_{j=i+2}^n V_{2,i,j}$ $V_1 = \sum_{i=2}^{n-1} \frac{1 - c_{\theta,i}}{4}$ $V_2 = 4(\tilde{\rho}_{i,j}^6 - C(\xi_i, \xi_j)\tilde{\rho}_{i,j}^3)$
Coordinates $(x_i, y_i, z_i) = \begin{cases} (0, 0, 0) & \text{for } i = 1 \\ (0, 1, 0) & \text{for } i = 2 \\ (\cos(\theta_1), \sin(\theta_1) + 1, 0) & \text{for } i = 3 \\ (x_{i-1} + \cos(\theta_{i-2})\cos(\beta_{i-3}), y_{i-1} + \sin(\theta_{i-1})\cos(\beta_{i-3}), \\ \quad z_{i-1} + \sin(\beta_{i-3})) & \text{for } 4 \leq i \leq n \end{cases}$	Coordinates $(x_i, y_i, z_i) = \begin{cases} (0, 0, 0) & \text{for } i = 1 \\ (0, 1, 0) & \text{for } i = 2 \\ (c_{\theta,1}, s_{\theta,1} + 1, 0) & \text{for } i = 3 \\ (x_{i-1} + c_{\theta,i-2}c_{\beta,i-3}, y_{i-1} + s_{\theta,i-1}c_{\beta,i-3}, \\ \quad z_{i-1} + s_{\beta,i-3}) & \text{for } 4 \leq i \leq n \end{cases}$

SOURCE: The author (2026)

5.1 MAIN CODE

In Appendix A, the complete main code is provided. It implements a nonlinear continuous optimization model built in Pyomo to generate a three-dimensional conformation of a protein-like chain under the off-lattice AB framework. The main idea is to represent the backbone geometry through angular variables and then determine the Cartesian coordinates of each residue by propagating the chain sequentially in space. The algorithm searches for a set of angles that yields a physically plausible fold while minimizing an energy-inspired objective function.

To ensure a consistent geometric interpretation of the angular variables, the model uses sine and cosine decision variables and enforces unit-circle constraints, guaranteeing that each pair $(\cos(\cdot), \sin(\cdot))$ corresponds to a valid angle, based on Equations 5.1 to 5.4. In addition, the first residues and initial angular values are fixed, which removes rigid-body degrees of freedom such as translation and rotation. This anchoring step is essential because it provides a unique reference frame and prevents the solver from exploring equivalent solutions that differ only by global transformations, which is according to Equation 5.5.

Once the reference frame is defined, the positions of the remaining residues are constructed recursively. Each new coordinate depends on the previous residue and on the current angular parameters, which effectively encodes a kinematic chain in three dimensions. This procedure produces a continuous backbone trajectory and implicitly enforces the connectivity of the polymer, meaning that the full spatial configuration is determined by the optimized angle set. Bounds are also imposed on the coordinate variables to restrict the feasible region and improve numerical stability during nonlinear optimization.

To model long-range interactions between non-adjacent residues, the code defines variables representing squared pairwise distances between selected residue pairs. Additional auxiliary variables are introduced to represent inverse-distance terms through algebraic constraints. The residue pairs considered exclude immediate neighbors, reflecting the fact that nonlocal contacts are typically responsible for driving folding behavior rather than consecutive backbone connections.

The objective function combines two main components. The first term acts as a regularization or smoothness penalty over the angular variables, discouraging degenerate or unrealistic backbone configurations. The second term captures sequence-dependent interactions through a coefficient that depends on whether a residue pair is of type A–A, B–B, or A–B, consistent with the AB model's hydrophobic/polar interpretation. Together, these terms define a trade-off between maintaining a coherent geometry and promoting favorable contacts depending on the amino-acid class pattern in the

sequence.

After defining the model, the code applies several Pyomo transformation passes aiming to simplify the algebraic system and strengthen the continuous relaxation through bound tightening and constraint propagation. These include converting constraints into variable bounds when possible, inducing linearity in simplified expressions, propagating fixed variables, and deactivating trivial constraints. Such preprocessing steps can significantly improve solver robustness, especially for nonlinear and nonconvex models with many auxiliary variables and bilinear relations. The resulting nonlinear program is then solved using an interior-point optimizer (KNITRO), producing an optimized set of sine/cosine variables. These are converted back into explicit angles, which yields the final angle representation of the chain.

Finally, the optimized angles are used to reconstruct 3D coordinates and generate a new structure file based on a reference PDB entry. The pipeline retrieves the experimental structure, computes the corresponding backbone distances, rebuilds the optimized coordinates, and writes an updated PDB-like output. The quality of the predicted conformation is assessed by computing the RMSD relative to the reference structure, and the final report includes both the optimized objective value and the structural deviation metric. The pseudocode for the proposed model is shown below.

Algorithm 1 Bilinear optimization model for 3D AB off-lattice folding

Require: Chain length N , AB sequence $seq = (seq_1, \dots, seq_N)$ with $seq_i \in \{A, B\}$

Ensure: Optimized angles (θ, β) , coordinates $\{\mathbf{r}_i\}_{i=1}^N$, objective value, RMSD

Initialize optimization model

Define residue indices $i \in \{1, \dots, N\}$ and coordinate dimensions $d \in \{x, y, z\}$

Decision variables (angle representation via sine/cosine)

For each i : define $(c_{\theta_i}, s_{\theta_i}) \in [-1, 1]^2$ and $(c_{\beta_i}, s_{\beta_i}) \in [-1, 1]^2$

For each residue i : define Cartesian coordinates $\mathbf{r}_i = (x_i, y_i, z_i) \in \mathbb{R}^3$

Trigonometric feasibility constraints

for $i = 1$ to $N - 1$ **do**

 Enforce $c_{\theta_i}^2 + s_{\theta_i}^2 = 1$

 Enforce $c_{\beta_i}^2 + s_{\beta_i}^2 = 1$

end for

Fix reference frame (remove rigid-body symmetries)

Set $\mathbf{r}_1 = (0, 0, 0)$

Set $\mathbf{r}_2 = (0, 1, 0)$

Fix initial angles (e.g., $\theta_1 = \frac{\pi}{2}$ and $\beta_1 = \frac{\pi}{2}$)

Algorithm 1 Bilinear optimization model for 3D AB off-lattice folding (continued)

Backbone propagation constraints

Compute $\mathbf{r}_3$ from $\mathbf{r}_2$ using $(c_{\theta_1}, s_{\theta_1})$

for $i = 4$ to N **do**

Enforce recursive construction:

$$x_i = x_{i-1} + c_{\theta_{i-2}} \cdot c_{\beta_{i-3}}$$

$$y_i = y_{i-1} + s_{\theta_{i-2}} \cdot c_{\beta_{i-3}}$$

$$z_i = z_{i-1} + s_{\beta_{i-3}}$$

end for

Impose bounds on (x_i, y_i, z_i) to restrict the feasible region

Define interacting residue pairs

$$\mathcal{P} \leftarrow \{(i, j) \mid 1 \leq i \leq N-2, i+2 \leq j \leq N\}$$

Distance and inverse-distance auxiliary variables

for all $(i, j) \in \mathcal{P}$ **do**

Define squared distance:

$$r_{ij} = \|\mathbf{r}_j - \mathbf{r}_i\|^2$$

Introduce reciprocal variable $\hat{r}_{ij}$ such that:

$$r_{ij} \hat{r}_{ij} = 1$$

Define higher-order terms:

$$r_{ij}^{(2)} = \hat{r}_{ij}^2, \quad r_{ij}^{(3)} = r_{ij}^{(2)} \hat{r}_{ij}$$

end for

Sequence-dependent interaction coefficient

if $seq_i = A$ and $seq_j = A$ **then**

return $C(s_i, s_j) = 1$

else if $seq_i = B$ and $seq_j = B$ **then**

return $C(s_i, s_j) = 0.5$

else

return $C(s_i, s_j) = -0.5$

end if

Objective function

Define bending/regularization term:

$$v_1 \geq \frac{1}{4} \sum_{i=2}^{N-2} (1 - c_{\theta_i})$$

Define nonlocal interaction term:

$$v_2 \geq 4 \sum_{(i,j) \in \mathcal{P}} r_{ij}^{(3)} (r_{ij}^{(3)} - C(s_i, s_j))$$

Minimize total energy surrogate:

$$\min v_1 + v_2$$

Solve the nonlinear program

Apply preprocessing (bound tightening and constraint propagation)

Solve using KNITRO until convergence

Algorithm 1 Bilinear optimization model for 3D AB off-lattice folding (continued)

Recover angles and validate structure

for $i = 1$ to $N - 2$ **do**

$\theta_i \leftarrow \arctan 2(s_{\theta_i}, c_{\theta_i})$

end for

for $i = 1$ to $N - 3$ **do**

$\beta_i \leftarrow \arctan 2(s_{\beta_i}, c_{\beta_i})$

end for

Reconstruct coordinates $\{\mathbf{r}_i\}$ from (θ, β)

Compare with reference structure and compute RMSD

return (θ, β) , $\{\mathbf{r}_i\}$, objective value, RMSD

6 RESULTS AND DISCUSSION

This chapter provides a clear analysis of the data and makes links between the study's findings and its goals by thoroughly presenting and analyzing the results. A server with an AMD Ryzen 9 5950X 16-Core (16 cores, 32 threads), 128 GB of random-access memory (RAM), and Ubuntu 24.04.1 LTS 64-bit was used for all trials. The tables give a concise overview of each algorithm's performance, and the outcomes of using different metaheuristic algorithms in conjunction with the suggested BL model are shown in detail in those lists.

Key performance metrics are included in the tables, along with the best (minimum), mean, and standard deviation (std) values for two crucial components: the RMSD, a measure of structural alignment, and the objective function, which represents the protein's inherent energy. These metrics are essential markers of the stability and efficacy of the optimization process, providing insight into the solutions' competence.

Exact optimization techniques aim to find the best answer to a problem within a workable solution space. These methods are frequently employed when deterministic evaluation of the objective function and constraints is feasible (Bollapragada *et al.*, 2023). The mean and standard deviation are statistical measurements used in mathematical optimization that characterize a group of values' central tendency and variability. However, since exact optimization looks for a deterministic solution rather than a probabilistic one, these statistical metrics have no bearing on the subject.

Metaheuristics attempt to reproduce fundamental physical, biological, or natural laws by employing a range of operators. In order to provide a practical solution in an acceptable amount of processing time, metaheuristics rely on problem knowledge. Natural evolutionary processes, including genetic mutation, reproduction, and survival of the fittest, inspire evolutionary architects. Differential evolution and genetic algorithm are two of the most well-known EAs. SI algorithms, like ABC and PSO, simulate living creatures' collective behavior and interactions.

The MealPy library was used to program the metaheuristic algorithms used in this research (Van Thieu; Mirjalili, 2023). These algorithms are among the most well-known and respected methods for solving the PSP problem (Oladejo, 2024). This decision is based on their well-established reputation and proven ability to solve comparable optimization problems. Every metaheuristic was executed 100 times, with 1000 epochs in each run. The parametric settings of the contrasted metaheuristics are compiled in Table 4. A methodical trial-and-error approach was used to find the ideal hyperparameter values. A variety of hyperparameter combinations were carefully selected

employing iterative experimentation, informed by empirical evaluation on validation sets.

Table 4 – Hiperparameters setup for the compared metaheuristics

DE		GA	
Population	50	Population	50
Crossover Probability	0.5	Crossover Probability	0.85
Differential Weight	0.5	Mutation Probability	0.1
PSO		ABC	
Population	50	Population	50
Local Coefficient	2.05	Limit before abandoning a food source	0.6
Global Coefficient	3.5		
Positive Constant (α)	0.4		

The bilinear programming model was implemented using the Pyomo mathematical optimization framework (Hart *et al.*, 2011) and solved with the Knitro nonlinear interior-point trust-region solver (Byrd *et al.*, 2006). The optimization process was executed for up to 1,000 iterations, after which the solution was considered converged.

A small set of proteins with varying lengths was randomly selected from the PDB to evaluate and compare the optimization performance under different experimental conditions. Table 5 reports the length n of each protein along with the corresponding problem dimension, which is equal to $2n - 5$. Each PDB entry is uniquely identified by a four-character PDB identifier, which is used to reference the selected protein structures.

Table 5 – Selected proteins

PDB ID	Length	Dimension
1CB3 (Demarest <i>et al.</i> , 1999)	13	21
1EDP (Andersen <i>et al.</i> , 1992)	17	29
1EDN (Janes <i>et al.</i> , 1994)	21	37
1AGT (Krezel <i>et al.</i> , 1995)	38	71
1CRN (Teeter, 1984)	46	87
2N35 (Jacobs <i>et al.</i> , 2015)	52	107
1CB9 (Dementieva <i>et al.</i> , 1999)	60	115
1HVV (Misura <i>et al.</i> , 2001)	75	145
1PTF (Jia <i>et al.</i> , 1994)	88	171
2G13 (Tsai <i>et al.</i> , 2007)	98	191

Table 6 reports the execution time required by each metaheuristic (DE, GA, PSO, and ABC), as well as by the proposed bilinear approach to reach their respective final solutions for each protein instance. Overall, the BL approach consistently outperforms the population-based metaheuristics in terms of computational efficiency, achieving substantial reductions in runtime across all tested proteins.

Table 6 shows that the BL approach consistently achieves the shortest execution time across all proteins, being approximately five times faster than the population-based

Table 6 – Time (seconds) taken to complete the run

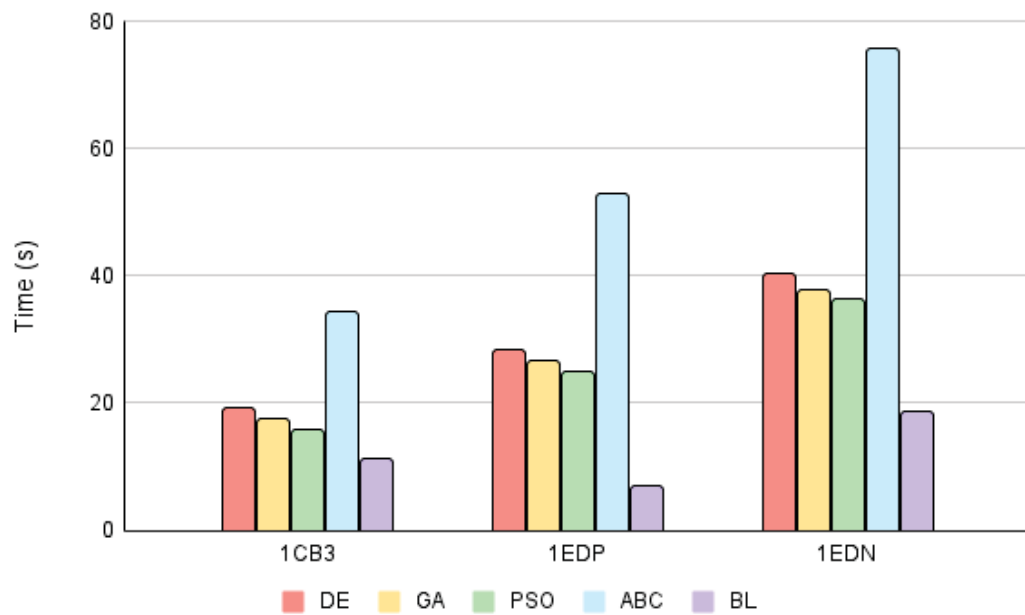
Protein	DE	GA	PSO	ABC	BL
1CB3	19.20	17.47	15.82	34.21	11.27
1EDP	28.17	26.45	24.85	52.69	6.97
1EDN	40.14	37.53	36.27	75.62	18.42
1AGT	113.39	137.35	176.84	223.05	26.31
1CRN	159.48	160.95	303.73	332.59	54.58
2N35	201.18	196.80	195.55	402.19	54.67
1CB9	262.94	276.48	257.38	529.29	117.22
1HVV	399.75	399.15	395.12	810.04	171.56
1PTF	547.48	545.89	543.70	1112.59	183.12
2G13	665.29	662.71	660.54	1355.56	277.67

metaheuristics in most cases. This speedup is mainly due to the algorithmic structure, which, instead of maintaining and updating a population of candidate solutions, BL solves a deterministic nonlinear programming problem using an interior-point trust-region method, which typically converges in a relatively small number of iterations. In contrast, the evolutionary and swarm methods require many function evaluations per iteration (proportional to the population size) and additional overhead for selection, crossover/mutation, velocity updates, or scout phases, which accumulate significantly as the protein length increases. Moreover, the BL formulation can exploit gradient and curvature information, enabling more directed progress toward stationary points, whereas metaheuristics spend substantial time exploring the search space stochastically. This effect becomes more pronounced for larger proteins (e.g., 1PTF and 2G13), where the combinatorial complexity increases and population-based exploration becomes increasingly expensive, while BL remains comparatively efficient.

For smaller proteins, the BL formulation reduces execution time by approximately 30–60% compared to DE, GA, and PSO. The difference is more pronounced when BL is compared to ABC, where the improvement amounts to a time reduction of approximately 75%. Figure 7 offers a visual representation of the difference in running time in a graph expressing the values referenced to the three smallest proteins.

For larger and more complex proteins, the gain becomes even more pronounced, with runtime reductions often exceeding 65–75% and reaching up to 80% when compared to ABC. Notably, while the execution time of metaheuristic methods grows rapidly with problem size, the BL approach exhibits a more moderate scaling behavior, indicating improved robustness with respect to dimensionality. These numbers can be better seen in Figure 8, where it is possible to note that DE, GA, and PSO have similar outputs, while ABC represents the worst time expectation, and BL is notably the fastest. These notions highlight the advantage of the proposed mathematical reformulation in delivering quality solutions at a significantly lower computational cost, achieving the

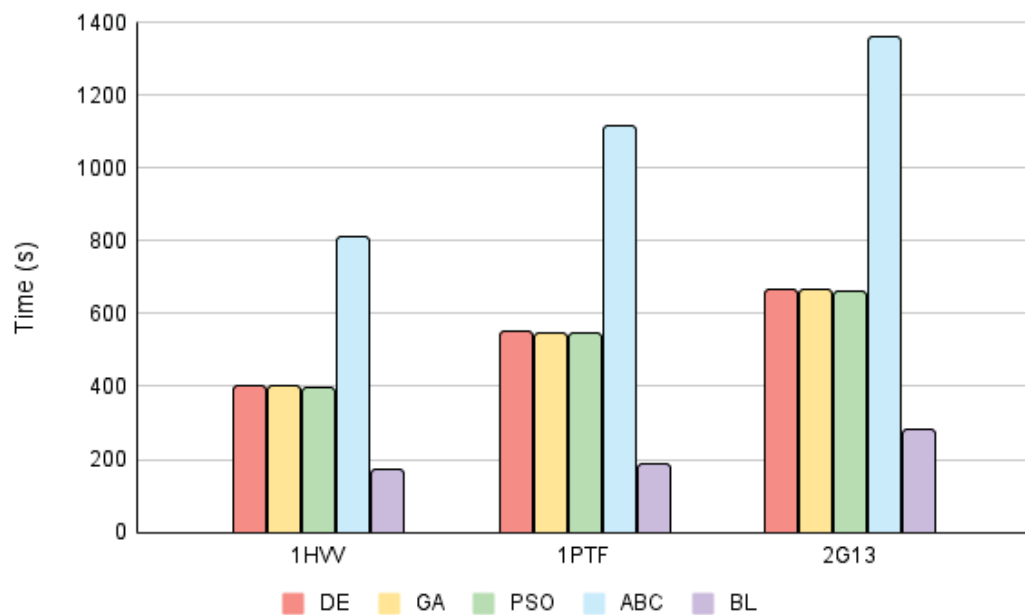
Figure 7 – Comparison graph of the running time for the smallest proteins



SOURCE: The author (2026)

specific objective number three.

Figure 8 – Comparison graph of the running time for the biggest proteins



SOURCE: The author (2026)

Table 7 presents the native energy values of the selected protein instances. The proteins considered span a wide range of lengths, from short peptides of 13 amino acids, which are generally regarded as relatively simple optimization cases, to larger proteins approaching 100 amino acids, posing substantially greater computational and

optimization challenges. This diversity in protein size allows for a more comprehensive assessment of the methods' performance across varying levels of structural complexity and provides important context for the subsequent RMSD analysis. Best results are in bold. The underlined results represent the best mean and standard deviation for the compared metaheuristics, as those metrics are not applied to the BL approach since it is an exact model that does not have this kind of variation.

Table 7 – Native energy optimization results

Protein	Metric	DE	GA	PSO	ABC	BL
1CB3	Best	-5.759	-0.848	-0.022	-1.473	-6.789
	mean $\pm$ std	<u>-4.235 $\pm$ 0.598</u>	-1.968 $\pm$ 1.039	0.801 $\pm$ <u>0.473</u>	0.400 $\pm$ 0.627	-
1EDP	Best	-7.657	-1.265	1.257	0.210	-10.354
	mean $\pm$ std	<u>-4.459 $\pm$ 1.068</u>	-1.761 $\pm$ <u>0.439</u>	2.173 $\pm$ 0.511	1.991 $\pm$ 0.479	-
1EDN	Best	-9.349	-0.058	2.327	1.392	-16.127
	mean $\pm$ std	<u>-5.201 $\pm$ 1.334</u>	-1.044 $\pm$ <u>0.288</u>	4.022 $\pm$ 0.814	3.389 $\pm$ 0.592	-
1AGT	Best	-15.080	1.753	9.009	6.771	-42.791
	mean $\pm$ std	<u>-9.308 $\pm$ 3.230</u>	4.4396 $\pm$ <u>1.027</u>	112.704 $\pm$ 78.394	6.771 $\pm$ 18.543	-
1CRN	Best	-23.277	10.014	95.117	19.051	-79.503
	mean $\pm$ std	<u>-9.861 $\pm$ 9.755</u>	17.136 $\pm$ 10.883	851.338 $\pm$ 644.444	170.206 $\pm$ 133.631	-
2N35	Best	-4.238	100.238	262.599	58.755	-40.477
	mean $\pm$ std	<u>9.278 $\pm$ 17.414</u>	107.447 $\pm$ 57.446	2850.773 $\pm$ 2172.797	640.789 $\pm$ 442.603	-
1CB9	Best	-9.759	272.322	511.046	74.798	-68.082
	mean $\pm$ std	<u>57.579 $\pm$ 93.743</u>	717.090 $\pm$ 536.605	9151.722 $\pm$ 7182.942	2804.050 $\pm$ 1819.864	-
1HVV	Best	0.697	991.692	8467.447	2928.539	-49.744
	mean $\pm$ std	<u>1304.949 $\pm$ 1809.622</u>	8400.311 $\pm$ 5556.535	5.419 $\cdot 10^4 \pm 3.556 \cdot 10^4$	2.099 $\cdot 10^4 \pm 1.224 \cdot 10^4$	-
1PTF	Best	-4.331	1.084 $\cdot 10^4$	4.303 $\cdot 10^4$	1.634 $\cdot 10^4$	-84.056
	mean $\pm$ std	<u>6736.274 $\pm$ 6756.273</u>	3.247 $\cdot 10^4 \pm 2.070 \cdot 10^4$	1.581 $\cdot 10^5 \pm 7.855 \cdot 10^4$	7.281 $\cdot 10^4 \pm 3.974 \cdot 10^4$	-
2G13	Best	-9.862	1.310 $\cdot 10^4$	1.164 $\cdot 10^5$	2.710 $\cdot 10^4$	-74.457
	mean $\pm$ std	<u>1.955 $\cdot 10^4 \pm 1.718 \cdot 10^4$</u>	7.712 $\cdot 10^4 \pm 4.138 \cdot 10^4$	3.241 $\cdot 10^5 \pm 1.537 \cdot 10^5$	1.582 $\cdot 10^5 \pm 8.447 \cdot 10^4$	-

Among many proteins, the BL optimization approach consistently reaches the lowest energy levels (often known as the "Best" findings). The negative energy values are a reflection of the forces that are involved in the conformation process. For instance, the hydrophobic effect, in which the apolar side chains are protected in the protein's interior to minimize unfavorable interactions with water, is represented as negative values in the energy formulation. The van der Waals interactions also add to the negative expectations for the energy, when the close packing of atoms in the native state enhances stabilizing short-range forces (Galano-Frutos; Sancho, 2024).

The minimization result demonstrates the proposed approach's resilience and versatility in synthesizing various protein architectures. Meanwhile, GA performed more consistently in the smaller proteins, suggesting it can be used in certain optimization circumstances when variation during repeatability may be more significant than minimization. The outcomes also draw attention to the difficulties unique to each algorithm, particularly in the extreme situations exemplified by proteins 1PTF and 2G13, where the BL performs noticeably better than the others.

The "mean $\pm$ std" metrics, in addition to the "Best" results, offer a more detailed knowledge of the performance of the compared metaheuristics. Significant variation has been seen among the proteins, especially in proteins 1CRN, 2N35, 1CB9, 1HVV, 1PTF, and 2G13, where the standard values are high. Higher standard deviations in the results for these proteins suggest more significant variability in the optimization results. The complex structural properties of the proteins may cause this more excellent dispersion in certain situations, making it more difficult for some algorithms to converge steadily.

The DE algorithm reliably displays competitive mean values for various proteins. Its consistent achievement of energy optimizations is commendable, as evidenced by the several instances where its standard deviation is smaller. DE functions consistently across various protein architectures based on its capacity to successfully explore solution spaces. On the other hand, ABC shows inconsistent performance across proteins, with mean values ranging from positive to negative. The corresponding standard deviation values draw attention to variations in the optimization results, indicating that although ABC can perform exceptionally well in some circumstances, its performance can be less consistent.

For this reason, a rigorous assessment of ABC's applicability for various protein architectures is necessary. GA demonstrates remarkable homogeneity in mean values for proteins 1EDP and 1EDN, suggesting that it helps attain steady-state energy levels. The more significant standard deviation values in other instances, especially for larger proteins, imply that GA might be more susceptible to structural intricacies and need parameter fine-tuning for best results. With a competitive mean performance for protein 1CB3, PSO shines out, yet its higher standard deviation values suggest some degree of

uncertainty in the data. PSO's efficacy varies according to the situation; it performs best in particular situations and shows more remarkable dispersion in others.

The results in Table 7 reveal a noticeable advantage of the BL approach in terms of energy minimization across all protein instances. For every protein considered, the BL formulation achieves the lowest (most negative) best energy, often by a wide margin relative to metaheuristic competitors. For smaller proteins such as 1CB3 and 1EDP, BL improves upon the best-performing metaheuristic (DE) by approximately 18% and 35%, respectively. These gains become substantially larger as protein size increases: for 1EDN, the BL approach attains a best energy of -16.127 compared to DE second best performance with -9.349, corresponding to an improvement of roughly 73%. For medium-sized proteins, the BL formulation achieves the best energy reductions of approximately 184% (1AGT) and 241% (1CRN) relative to DE, indicating a markedly deeper descent into the energy landscape.

The contrast is even more striking for larger proteins (2N35, 1CB9, 1HVV, 1PTF, and 2G13), where swarm- and evolution-based methods frequently converge to positive or enormous energy values, suggesting entrapment in poor local minima or numerical instability. In these cases, the BL approach not only maintains stable negative energies but also improves the best solutions by one to three orders of magnitude compared to GA, PSO, and ABC. For example, in 1PTF, BL reaches -84.056, whereas PSO and ABC report best energies on the order of 10^4 , and in 2G13, BL improves upon GA by more than two orders of magnitude.

The best free energy values indicate the protein's most likely shape under the specified environmental conditions, which also capture the subtleties of the protein's folded state. In order to reduce their exposure to the watery environment, the hydrophobic amino acids group together inside proteins. Often referred to as the apolar nucleus, this hydrophobic core is a crucial structural component of many proteins in their folded condition. The greater surface area in larger proteins frequently results in more solvent-exposed areas, which may affect how the protein interacts with water, affecting how soluble and stable the protein is. Additionally, larger proteins could have a more complex interior region packing, particularly in the hydrophobic core. As the protein's size rises, achieving a well-packed core becomes more challenging, and stability depends on a delicate balance of pressures.

The quality of the energy function and how well it represents the genuine energetics of protein structures determine the association between the native energy and RMSD. In theory, a good energy function should give higher energy to structures that deviate more from the native structure and lower energy to structures that are closer to it. However, the association is not always clear-cut because it depends on several variables, such as the intricacy of the protein folding problem, the precision of

the energy function, and the techniques employed for structure prediction. A poorer association with RMSD can occasionally result from the energy function needing to be more precise to rank native-like structures lower in energy consistently. Table 8 shows the best, mean, and standard deviation for each protein's RMSD.

Table 8 – RMSD optimization results

Protein	Metric	DE	GA	PSO	ABC	BL
1CB3	Best	3.448	2.966	3.228	2.682	2.619
	mean $\pm$ std	4.652 ± 0.575	4.441 ± 0.671	4.441 ± 0.718	4.374 ± 0.683	-
1EDP	Best	4.031	6.342	4.630	3.951	3.840
	mean $\pm$ std	5.848 ± 1.209	10.519 ± 2.277	7.181 ± 1.975	7.396 ± 1.742	-
1EDN	Best	4.576	9.459	4.673	4.651	4.015
	mean $\pm$ std	6.997 ± 1.513	13.125 ± 1.122	7.868 ± 1.7075	8.120 ± 1.846	-
1AGT	Best	10.318	17.480	9.541	9.688	4.987
	mean $\pm$ std	13.525 ± 2.168	25.686 ± 3.100	14.511 ± 2.859	14.995 ± 2.633	-
1CRN	Best	7.672	11.957	5.072	9.060	4.165
	mean $\pm$ std	14.533 ± 4.014	22.584 ± 4.501	12.280 ± 2.731	13.542 ± 2.946	-
2N35	Best	8.451	10.133	6.824	7.430	5.908
	mean $\pm$ std	16.690 ± 4.298	19.980 ± 5.035	12.667 ± 2.918	12.993 ± 2.645	-
1CB9	Best	12.635	12.979	11.783	11.158	9.756
	mean $\pm$ std	20.427 ± 4.572	19.976 ± 4.852	15.923 ± 2.5266	15.088 ± 2.507	-
1HVV	Best	9.432	10.763	8.583	9.590	6.412
	mean $\pm$ std	16.368 ± 4.275	16.961 ± 5.1371	18.549 ± 3.819	17.446 ± 3.783	-
1PTF	Best	13.936	12.779	13.152	13.055	10.102
	mean $\pm$ std	25.576 ± 8.177	23.420 ± 5.799	18.256 ± 3.328	19.719 ± 3.394	-
2G13	Best	16.728	5.906	13.133	14.508	3.663
	mean $\pm$ std	25.049 ± 5.443	12.906 ± 5.068	19.248 ± 3.703	19.923 ± 3.230	-

A complex correlation was found in this study on protein energy optimization between the mean values produced by metaheuristic algorithms and the RMSD values from the Best findings. Interestingly, the BL Approach regularly shows RMSD values smaller than the mean values of these algorithms, even though it may not consistently produce the least RMSD compared to other metaheuristics in the Best results. Compared to the metaheuristics, the BL approach's reduced RMSD suggests a higher proximity between individual optimization runs, indicating a more consistent and dependable performance.

The BL Approach is a more dependable method for reaching optimal protein energy conformations since it places more weight on proximity to the Best outcomes than mean performance. This careful analysis emphasizes how crucial it is to assess optimization algorithms in terms of their general performance consistency and the very best results, as this is a crucial component for the dependability and repeatability of experimental results.

An analysis of Table 8 shows valuable information regarding the effectiveness of optimization strategies for the native RMSD. The results show that the proposed BL approach consistently achieves the lowest best RMSD for all protein instances, with particularly pronounced gains as protein length increases. For the five smallest proteins (1CB3, 1EDP, 1EDN, 1AGT, and 1CRN), the BL formulation yields RMSD values that are either below or very close to 5 Å, a threshold commonly associated with structurally meaningful and biologically plausible predictions. In these cases, the BL approach improves the best RMSD by approximately 12–50% relative to the strongest competing metaheuristic, with the most notable reductions observed for 1AGT (from 9.541 to 4.987, improvement of about 48%) and 1CRN (from 5.072 to 4.165, improvement of about 18%). These results comply with the specific objective number 6.

The difficulty of the algorithms to optimize the protein structure increases with the size of the protein since larger proteins have greater difficulty establishing thermodynamic stability. Nevertheless, as protein size increases, the BL approach continues to demonstrate strong structural accuracy. For larger proteins (2N35, 1CB9, 1HVV, 1PTF, and 2G13), the proposed method consistently attains RMSD values below or close to 10 Å, while competing metaheuristics frequently exceed this range. For instance, in 1HVV, the BL approach reduces the best RMSD from 8.583 (PSO) to 6.412, corresponding to an improvement of approximately 25%, and in 2G13, it achieves a best RMSD of 3.663, improving upon GA by nearly 38%. These results indicate that the BL formulation effectively preserves predictive accuracy even as the dimensionality of the problem increases, and are in compliance with the specific objective number 7.

Notably, for larger proteins such as 1AGT, 1CB9, 1PTF, and 2G13, the BL optimization achieves best RMSD values that are comparable to, or significantly better

than, those obtained by evolutionary and swarm-based methods, despite requiring substantially lower computational time, as shown in Table 6. These results suggest that the proposed bilinear reformulation is able to preserve structural accuracy while offering a favorable trade-off between solution quality and computational cost, particularly as problem dimensionality increases. Overall, the BL approach attains competitive or superior best-case RMSD values for several protein instances, particularly for medium and large-sized proteins, while exhibiting deterministic behavior due to its single-run formulation. In contrast, population-based metaheuristics display larger variance, as reflected by their standard deviations, indicating sensitivity to initialization and stochastic search dynamics.

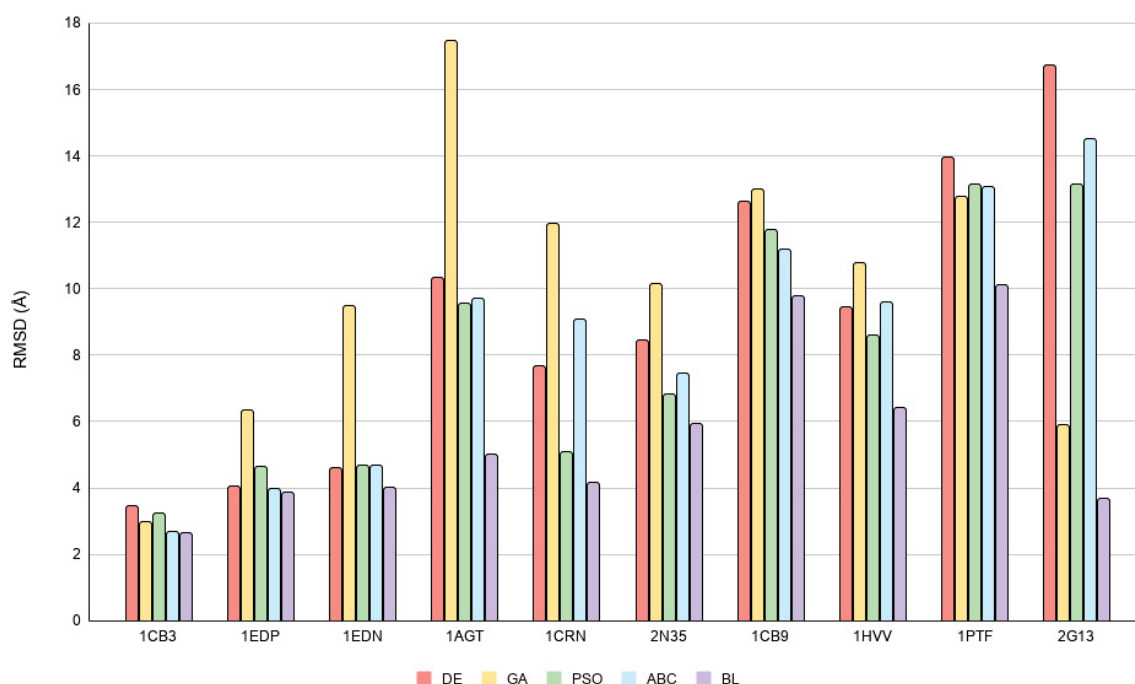
Additionally, several protein instances exhibit clear quantitative advantages for the proposed BL approach in terms of best-case structural accuracy. Most notably, for protein 2G13, the BL approach achieves a best RMSD of 3.663, improving upon the next best result (GA, 5.906) by approximately 38%, representing a substantial gain in structural fidelity. For 1EDN, the BL formulation reduces the best RMSD from 4.576 (DE) to 4.015, corresponding to an improvement of roughly 12%. More moderate but consistent gains are observed for 1PTF and 1CB9, where the BL approach improves the best RMSD by approximately 5% and 3%, respectively, compared to the strongest competing metaheuristics. For 1AGT, the improvement is more incremental (around 2%), yet still demonstrates competitive accuracy.

In contrast, for smaller proteins such as 1CB3 and 1EDP, certain population-based metaheuristics achieve slightly lower best RMSD values, likely benefiting from stochastic exploration and multiple-run variability; nevertheless, the BL approach remains competitive while offering a deterministic solution and significantly lower computational cost. These results in RMSD can be better seen in Figure 9.

The experimental results demonstrate that the proposed bilinear reformulation constitutes an effective strategy for mitigating the challenges posed by local minima in PSP. By explicitly modeling the energetic interactions through auxiliary variables and structured constraints, the BL approach reshapes the search landscape, reducing the reliance on stochastic exploration and limiting premature convergence to suboptimal local solutions.

This is evidenced by its consistent ability to attain competitive or superior RMSD values, particularly for medium- and large-scale protein instances, while maintaining significantly lower computational costs compared to population-based metaheuristics. Although stochastic methods may occasionally achieve marginally better solutions for very small proteins due to repeated random sampling, the deterministic nature, scalability, and robustness of the BL formulation make it particularly well suited to addressing the local-minimum challenge inherent in PSP, offering a reliable and computationally

Figure 9 – Comparison graph of the RMSD for each protein



SOURCE: The author (2026)

efficient alternative for accurate protein structure optimization.

The scalability of the proposed BL approach is a critical consideration when extending protein structure prediction to larger systems exceeding 200 amino acids. As protein length increases, the dimensionality of the BL formulation grows rapidly due to the introduction of auxiliary variables and constraints required to model pairwise interactions and geometric dependencies. This leads to a polynomial but steep increase in problem size, which can place substantial demands on memory usage and solver time when applied to very large proteins. While the results presented in this work demonstrate favorable scaling behavior up to approximately 100 amino acids, where the BL approach maintains both computational efficiency and structural accuracy, further scaling beyond 200 residues is expected to require additional strategies, such as interaction sparsification, domain decomposition, or hierarchical optimization, to remain tractable.

Despite these challenges, an important advantage of the BL formulation lies in its reduced sensitivity to initialization when compared to stochastic metaheuristic methods. Since the optimization is driven by a deterministic solver, the solution process is less dependent on random initial conformations and does not rely on repeated runs to escape poor local minima. However, initialization still plays a non-negligible role in large-scale instances, as ill-conditioned starting points may slow convergence or bias the solver toward suboptimal regions of the feasible space, particularly in highly

non-convex landscapes. In practice, this suggests that informed initialization strategies, as is the case with coarse-grained pre-folding or secondary-structure-guided constraints, could significantly enhance convergence speed and solution quality for larger proteins.

To assess whether the observed differences in structural accuracy are statistically meaningful, the best RMSD values obtained for each protein were compared across optimization methods (Table 8). Since all algorithms were evaluated on the same set of proteins, the results form paired observations, allowing a direct per-protein comparison between the BL approach and each metaheuristic. Therefore, paired two-sided Student’s t-tests were performed between BL and DE, GA, PSO, and ABC using the best RMSD values across the 10 proteins, testing the null hypothesis that the mean RMSD difference is zero. This analysis provides quantitative evidence of whether BL yields a consistent improvement in RMSD performance beyond what may be expected from random variability across instances.

Table 9 – Paired t-test results (Best RMSD)

Comparison	mean difference (other-BL)	t-stat	p-value
DE vs BL	3.549	3.011	0.0147
GA vs BL	4.511	4.121	0.0026
PSO vs BL	2.496	2.849	0.0191
ABC vs BL	3.000	2.887	0.0180

The t-test results in Table 9 indicate that the BL approach achieves statistically significant improvements in RMSD when compared against the metaheuristics in most pairwise comparisons. In general, the mean differences show that BL tends to produce better structural accuracy across the tested proteins, rather than occasional isolated wins. The smallest p-values, $p < 0.05$, support rejecting the null hypothesis that BL and the competing method have equivalent average performance, meaning the observed RMSD reductions are unlikely to be due to random variation alone. The strongest evidence appears in comparisons where the RMSD gaps are larger (e.g., BL vs GA and BL vs DE), suggesting that BL provides not only improved best-case solutions but also a consistent shift toward lower-error predictions. Therefore, it is safe to assume that the statistical test reinforces the conclusion that BL offers a reliable and measurable improvement in solution quality, rather than an improvement that is purely anecdotal or dependent on a single protein instance.

7 CONCLUSION

This thesis aimed at the proposition of a novel computational framework for protein structure prediction grounded in an exact bilinear optimization approach applied to the AB off-lattice model. It introduced a bilinear reformulation of the original nonlinear problem, which provides a mathematically tractable structure that enables the use of exact optimization methodologies. This constitutes a pioneering step toward reliably identifying the global minimum of the AB model's energy function, a long-standing challenge given the model's intrinsic nonlinearity and NP-hard nature.

Approaches such as AlphaFold 2 have poor performance on antibody–antigen and TCR–antigen complexes, which are useful for vaccines and immunotherapy (Pratt *et al.*, 2025). The novel BL model opens a new pathway for discussion and application on proteins whose sequences or evolutionary information are hardly trackable in current databases, becoming a helpful tool for designing proteins for vaccines, expanding the number of starting points for the nanoparticles needed for neutralizing responses, in addition to modeling vaccine constructs and assessing stability. These kinds of applications are important in the Brazilian biotechnology scenario, as the country is a leading researcher in tropical diseases, such as dengue and malaria, and also other worldwide viral diseases, as is the case of SARS-CoV-2.

As previously discussed, recent breakthroughs in AI-based protein structure prediction have transformed the field. However, these methods rely heavily on evolutionary information derived from multiple sequence alignments. Consequently, their predictive accuracy drops substantially when applied to orphan proteins, *de novo* sequences, or highly divergent peptides, for which no homologous sequences exist. In these cases, data-driven inference becomes fundamentally limited, as there is no statistical signal from which to learn structural patterns. By grounding prediction in physicochemical principles rather than evolutionary priors, the BL optimization approach remains applicable even in the complete absence of sequence homology. Thus, for proteins beyond the reach of current AI systems, the method presented in this thesis stands as a principled and interpretable alternative, reinforcing its potential strategic value to future bioinformatics research and applications.

Beyond improvements in predictive accuracy, the BL method also demonstrated substantial gains in computational efficiency. While metaheuristic algorithms often require a large number of evaluations and extended runtime to approach high-quality solutions, the bilinear approach achieved convergence significantly faster. In the conducted experiments, the bilinear method reached feasible high-quality solutions in about a third of the time that metaheuristic baselines typically required by the metaheuristic

baselines when applied to medium or large aminoacids sequence. When used in the smallest protein tested, the time difference was not drastic, although an improvement was still shown. This decrease in computational time highlights the practical feasibility of the proposed technique, especially in large-scale or time-critical PSP applications.

The comparative performance of the assessed optimization techniques is further illustrated by experimental results, which further show the bilinear strategy's consistency and robustness. The bilinear technique consistently produced RMSD scores of about -46 ± 27 , exceeding the mean RMSD of top metaheuristics, which might range in the 10^5 size, despite variability in RMSD values seen between test sequences. This computing time reduction highlights the suggested method's practicality, especially in large-scale or time-sensitive PSP applications.

The advantages of the suggested bilinear framework, such as faster convergence, lower RMSD, and more consistent performance compared to conventional metaheuristic approaches, are demonstrated by this emphasis on both accuracy and computing efficiency. These features demonstrate the method's promise as a promising avenue for future study in precise and hybrid optimization approaches for PSP.

7.1 FUTURE WORK

There are several promising avenues for extending the contributions of this thesis. One natural direction is the development of hybrid models that integrate the proposed bilinear optimization framework with deep learning-based structural predictors. By using DL to enhance the realism and initialization quality of the model, the bilinear optimization methodology can transition from solving an exact problem on a simplified potential to solving a near-exact problem on a potential landscape that is orders of magnitude closer to the true protein folding energy function, thereby uniting the mathematical rigor of global optimization with biological realism.

Ultimately, the combination of the bilinear approach with DL can be explored as the bilinear optimization method guarantees the global optimum for a defined mathematical model; the critical challenge remains the fidelity of that model to real-world physics. Specifically, DL models can be trained on extensive structural and physicochemical data to perform several key functions, e. g., generating physically-informed energy parameters that more accurately reflect complex side-chain interactions and serving as a sophisticated search guidance heuristic that provides highly relevant initial structures for the bilinear solver, dramatically cutting down the Branch-and-Bound time.

Furthermore, the integration of the proposed method into hybrid metaheuristic-exact pipelines also offers an exciting research direction. Metaheuristics could be used to explore broad regions of the search space, while the bilinear model performs

local refinements to high-precision optima—effectively combining global exploration with local optimality guarantees. As some researchers suggest that future advancements will rely on creating models based on basic physicochemical principles that provide more precise and thorough predictions for a greater range of biological systems (Park *et al.*, 2025).

7.2 CONTRIBUTIONS

The author published works connected to the study proposal while working on her PhD.

DA SILVA, LUIZA SCAPINELLO AQUINO; SEMAN, LAIO ORIEL; CAMPONOGARA, EDUARDO; MARIANI, VIVIANA COCCO; COELHO, LEANDRO DOS SANTOS. Bilinear optimization of protein structure prediction: An exact approach via AB off-lattice model. *COMPUTERS IN BIOLOGY AND MEDICINE*, v. 176, p. 108558, 2024. DOI: 10.1016/j.compbimed.2024.108558

Papers on different use of deep learning approaches and metaheuristics in different applications were published, both in conference proceeding and in peer-reviewed journals.

AQUINO, LUIZA SCAPINELLO; SEMAN, LAIO ORIEL; MARIANI, VIVIANA COCCO; COELHO, LEANDRO DOS SANTOS; STEFENON, STEFANO FRIZZO; GONZALES, GABRIEL VILLARRUBIA. Spatiotemporal wind energy forecasting: A comprehensive survey and a deep equilibrium-based case study with StemGNN. *IEEE ACCESS*, v. 1, 2025. DOI: 10.1109/ACCESS.2025.3586997.

SEMAN, LAIO ORIEL; **AQUINO, LUIZA SCAPINELLO**; STEFENON, STEFANO FRIZZO; YON, KIN-CHOONG; MARIANI, VIVIANA COCCO; COELHO, LEANDRO DOS SANTOS. Simultaneously anomaly detection and forecasting for predictive maintenance using a zero-cost differentiable architecture search-based network. *COMPUTERS & INDUSTRIAL ENGINEERING*, v. 1, p. 111412, 2025. DOI: doi.org/10.1016/j.cie.2025.111412.

AQUINO, LUIZA SCAPINELLO; AGOTTANI, LUIS FERNANDO RODRIGUES; SEMAN, LAIO ORIEL; MARIANI, VIVIANA COCCO; COELHO, LEANDRO DOS SANTOS; GONZALES, GABRIEL VILLARRUBIA. Fault Detection in Power Distribution Systems Using Sensor Data and Hybrid YOLO with Adaptive Context Refinement. *APPLIED SCIENCES*, v. 15, n. 16, p. 9186, 2025. DOI: doi.org/10.3390/app15169186.

STEFENON, STEFANO FRIZZO; SEMAN, LAIO ORIEL; **DA SILVA, LUIZA SCAPINELLO AQUINO**; MARIANI, VIVIANA COCCO; COELHO, LEANDRO DOS SANTOS. Hypertuned temporal fusion transformer for multi-horizon time series forecasting of dam level in hydroelectric power plants. *INTERNATIONAL JOURNAL OF ELECTRI-*

CAL POWER & ENERGY SYSTEMS, v. 157, p. 109876, 2024. DOI:doi.org/10.1016/j.ijepes.2024.109876.

STEFENON, STEFANO FRIZZO; SEMAN, LAIO ORIEL; **AQUINO, LUIZA SCAPINELLO**; COELHO, LEANDRO DOS SANTOS. Wavelet-Seq2Seq-LSTM with attention for time series forecasting of level of dams in hydroelectric power plants. ENERGY, v. 274, p. 127350, 2023. DOI: doi.org/10.1016/j.energy.2023.127350.

DA SILVA, LUIZA SCAPINELLO AQUINO; LUCIO, YAN LIEVEN SOUZA; COELHO, LEANDRO DOS SANTOS; MARIANI, VIVIANA COCCO; RAO, RAVIPUDI VENKATA. A comprehensive review on Jaya optimization algorithm. ARTIFICIAL INTELLIGENCE REVIEW, v. 56, n. 5, p. 4329-4361, 2022. DOI: doi.org/10.1007/s10462-022-10234-0.

COELHO, LEANDRO; **SCAPINELLO AQUINO DA SILVA, LUIZA**; DE GEUS, LEONARDO ALEXANDRE; COCCO MARIANI, VIVIANA . Comparison of convolutional neural networks approaches applied to the diagnosis of Alzheimer's disease. In: ESANN 2025 proceedings, 2025, Bruges, Belgium.

AQUINO, LUIZA SCAPINELLO AQUINO; SEMAN, L. O.; MARIANI, V. C.; COELHO, LEANDRO DOS SANTOS. A Novel Spatiotemporal Wind Power Forecasting Model Based on Hybrid Graph Neural Network Approach (COBEM 2025). In: 28th International Congress of Mechanical Engineering (COBEM 2025), 2025, Curitiba-PR, Brazil.

LUCIO, YAN LIEVEN SOUZA; **DA SILVA, LUIZA SCAPINELLO AQUINO**; MARIANI, VIVIANA COCCO; COELHO, LEANDRO DOS SANTOS. Manta Ray Foraging Optimization Approaches on Multivariable PID Controller Tuning. In: 2023 IEEE Congress on Evolutionary Computation (CEC), 2023, Chicago-IL, USA. DOI: 10.1109/CEC53210.2023.10254147.

DA SILVA, LUIZA SCAPINELLO AQUINO; LOPES, ALLAN; SEMAN, LAIO ORIEL; MARIANI, VIVIANA COCCO COELHO, LEANDRO DOS SANTOS. Graph Neural Network Applied to Bearing Fault Diagnosis. In: 27th ABCM International Congress of Mechanical Engineering (COBEM 2023), 2023, Florianopolis-SC, Brazil.

REFERENCES

- ABRAMSON, Josh; ADLER, Jonas; DUNGER, Jack; EVANS, Richard; GREEN, Tim; PRITZEL, Alexander; RONNEBERGER, Olaf; WILLMORE, Lindsay; BALLARD, Andrew J; BAMBRICK, Joshua, *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. **Nature**, Nature Publishing Group UK London, v. 630, n. 8016, p. 493–500, 2024.
- ADHIKARI, Badri. DEEPCON: protein contact prediction using dilated convolutional neural networks with dropout. **Bioinformatics**, Oxford University Press, v. 36, n. 2, p. 470–477, 2020.
- AGRAHARI, Gargi; CHATTERJEE, Abhijit. Speed-up of Monte Carlo simulations by preparing starting off-lattice structures that are close to equilibrium. **The Journal of Chemical Physics**, AIP Publishing, v. 152, n. 4, 2020.
- AGRAWAL, Akshay; VERSCHUEREN, Robin; DIAMOND, Steven; BOYD, Stephen. A rewriting system for convex optimization problems. **Journal of Control and Decision**, Taylor & Francis, v. 5, n. 1, p. 42–60, 2018.
- AHMAD, Basir; KHAN, Rizwan Hasan. Protein Folding: From Hypothesis Driven to Data Mining. **Pakistan Journal of Biological Sciences**, v. 8, n. 3, p. 487–492, 2005.
- AKBAR, Sana; PARDASANI, Kamal Raj; KHAN, Farhan. Swarm optimization-based neural network model for secondary structure prediction of proteins. **Network Modeling Analysis in Health Informatics and Bioinformatics**, Springer, v. 10, n. 1, p. 33, 2021.
- AKBAR, Sana; PARDASANI, Kamal Raj; PANDA, Nihar Ranjan. PSO based neuro-fuzzy model for secondary structure prediction of protein. **Neural Processing Letters**, Springer, v. 53, n. 6, p. 4593–4612, 2021.
- AKHTER, Nasrin; SHEHU, Amarda. From extraction of local structures of protein energy landscapes to improved decoy selection in template-free protein structure prediction. **Molecules**, MDPI, v. 23, n. 1, p. 216, 2018.
- AKRAM, Faiz; AZIZ, Shafaque; KHAN, Nayyar Ahmed; FAIZI, Syed Akramah; RAZA, Khalid. Integrating Artificial Bee Colony Algorithms for Deep Learning Model Optimization: A Comprehensive Review. **Solving with Bees: Transformative Applications of Artificial Bee Colony Algorithm**, Springer, p. 73–102, 2024.
- ALDINO, Ahmad Ari; ULFA, Marchamah. Optimization of Lampung Batik Production Using the Simplex Method. **BAREKENG: Jurnal Ilmu Matematika Dan Terapan**, v. 15, n. 2, p. 297–304, 2021.
- ALQURAISHI, Mohammed. Machine learning in protein structure prediction. **Current opinion in chemical biology**, Elsevier, v. 65, p. 1–8, 2021.
- ALSHAMRANI, Ahmad M; SAXENA, Akash; SHEKHAWAT, Shalini; ZAWBAA, Hossam M; MOHAMED, Ali Wagdy. Performance evaluation of ingenious crow search optimization algorithm for protein structure prediction. **Processes**, MDPI, v. 11, n. 6, p. 1655, 2023.

- ANDERSEN, Niels H; CHEN, Chinpan; MARSCHNER, Thomas M; KRYSTEK JR, Stanley R; BASSOLINO, Donna A. Conformational isomerism of endothelin in acidic aqueous media: a quantitative NOESY analysis. **Biochemistry**, ACS Publications, v. 31, n. 5, p. 1280–1295, 1992.
- ANDREANI, Roberto; SCHUVERDT, Maria L; SECCHIN, Leonardo D. On enhanced KKT optimality conditions for smooth nonlinear optimization. **SIAM Journal on Optimization**, SIAM, v. 34, n. 2, p. 1515–1539, 2024.
- ANFINSSEN, Christian B. Principles that govern the folding of protein chains. **Science**, American Association for the Advancement of Science, v. 181, n. 4096, p. 223–230, 1973.
- ANFINSSEN, Christian B; HABER, Edgar; SELA, Michael; WHITE JR, Frederick H. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 47, n. 9, p. 1309–1314, 1961.
- AQUIB, Mohd. A comparative analysis of interior-point methods in convex optimization. **Int. J. Adv. Res. Ideas Innov. Technol.**, v. 7, n. 3, p. 353–358, 2021.
- ASHRAFUZZAMAN, Mohammad. Artificial intelligence in protein science and engineering. In: ARTIFICIAL Intelligence in Biomaterials Design and Development. [S.l.]: Elsevier, 2026. P. 251–275.
- ASSIS, Leonardo Salsano de; CAMPONOGARA, Eduardo; ZIMBERG, Bernardo; FERREIRA, Enrique; GROSSMANN, Ignacio E. A piecewise McCormick relaxation-based strategy for scheduling operations in a crude oil terminal. **Computers & Chemical Engineering**, v. 106, p. 309–321, 2017.
- AVRIEL, Mordecai. Nonlinear programming. In: MATHEMATICAL Programming for Operations Researchers and Computer Scientists. Boca Raton, USA: CRC Press, 2020. P. 271–367.
- BADACZEWSKA-DAWID, Aleksandra E; KOLINSKI, Andrzej. Importance of Secondary Structure Data in Large Scale Protein Modeling Using Low-Resolution SURPASS Method. In: PREDICTION of Protein Secondary Structure. New York, USA: Springer, 2024. P. 55–78.
- BAI, Lin; YANG, Lina. A unified deep learning model for protein structure prediction. In: 3RD IEEE International Conference on Cybernetics (CYBCONF). Lausanne, Switzerland: IEEE, 2017. P. 1–6.
- BAO, Zhiyuan; WAN, Yujian; HU, Zechun; MUJEEB, Asad. DC-based security constraints formulation: a perspective of primal-dual interior point method. **Electric Power Systems Research**, Elsevier, v. 228, p. 110058, 2024.
- BĂUTU, Andrei; LUCHIAN, Henri. Protein structure prediction in lattice models with particle swarm optimization. In: INTERNATIONAL Conference on Swarm Intelligence. Brussels, Belgium: Springer, 2010. P. 512–519.
- BAZARAA, Mokhtar S; SHERALI, Hanif D; SHETTY, Chitharanjan M. **Nonlinear programming: theory and algorithms**. Hoboken, New Jersey, United States of America: John Wiley & sons, 2013.

BINBAY, Furkan Ayberk; RATHOD, Dhruv Chetanbhai; GEORGE, Ajay Abisheck Paul; IMHOF, Diana. Quality assessment of selected protein structures derived from homology modeling and AlphaFold. **Pharmaceuticals**, MDPI, v. 16, n. 12, p. 1662, 2023.

BOARD, Advisory. RoseTTAFold: Accurate Protein Structure Prediction Accessible to All. **Institute for Protein Design**, 2021.

BOGGS, Paul T; TOLLE, Jon W. Sequential Quadratic Programming. **Acta Numerica**, v. 4, p. 1–51, 1995.

BOIANI, Mateus; PARPINELLI, Rafael Stubs. A GPU-based hybrid jDE algorithm applied to the 3D-AB protein structure prediction. **Swarm and Evolutionary Computation**, Elsevier, v. 58, p. 100711, 2020.

BOLLAPRAGADA, Raghu; KARAMANLI, Cem; KEITH, Brendan; LAZAROV, Boyan; PETRIDES, Socratis; WANG, Jingyi. An adaptive sampling augmented Lagrangian method for stochastic optimization with deterministic constraints. **Computers & Mathematics with Applications**, Elsevier, v. 149, p. 239–258, 2023.

BORGUESAN, Bruno; SILVA, Mariel Barbachan e; GRISCI, Bruno; INOSTROZA-PONTA, Mario; DORN, Márcio. APL: An angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. **Computational biology and chemistry**, Elsevier, v. 59, p. 142–157, 2015.

BORKAR, Pradnya S; MAHAJAN, AR. Genetic algorithm-based approach for RNA secondary structure prediction. In: **PROGRESS in Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2016, Volume 1**. Singapore: Springer, 2018. v. 1, p. 397–408.

BOŠKOVIĆ, Borko; BREST, Janez. Differential evolution for protein folding optimization based on a three-dimensional AB off-lattice model. **Journal of Molecular Modeling**, Springer, v. 22, p. 252, 2016.

BRANDEN, Carl Ivar; TOOZE, John. **Introduction to protein structure**. [S.l.]: Garland Science, 2012.

BYRD, Richard H; CHIN, Gillian M; NEVEITT, Will; NOCEDAL, Jorge. On the use of stochastic hessian information in optimization methods for machine learning. **SIAM Journal on Optimization**, SIAM, v. 21, n. 3, p. 977–995, 2011.

BYRD, Richard H; NOCEDAL, Jorge; WALTZ, Richard A. Knitro: An integrated package for nonlinear optimization. In: **LARGE-SCALE Nonlinear Optimization. Nonconvex Optimization and Its Applications**. Berlin, Germany: Springer, 2006. v. 83. P. 35–59.

CALVET, Laura; BENITO, Sergio; JUAN, Angel A; PRADOS, Ferran. On the role of metaheuristic optimization in bioinformatics. **International Transactions in Operational Research**, Wiley Online Library, v. 30, n. 6, p. 2909–2944, 2023.

CAMPONOGARA, Eduardo; SEMAN, Laio Oriel; GILI, Luiz Carlos. Mixed-integer bilinear and piecewise-linear models for designing switching strategies of multilevel power converters. **Computers & Electrical Engineering**, v. 77, p. 88–108, 2019.

CAO, Renzhi; ADHIKARI, Badri; BHATTACHARYA, Debswapna; SUN, Miao; HOU, Jie; CHENG, Jianlin. QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. **Bioinformatics**, Oxford University Press, v. 33, n. 4, p. 586–588, 2017.

- CASTRO, Pedro. Tightening piecewise McCormick relaxations for bilinear problems. **Computers & Chemical Engineering**, v. 72, p. 300–311, 2015.
- CASTRO, Pedro M.; GROSSMANN, Ignacio E. Optimality-based bound contraction with multiparametric disaggregation for the global optimization of mixed-integer bilinear problems. **Journal of Global Optimization**, v. 59, p. 277–306, 2014.
- CHEN, Xin; LV, Mingwei; ZHAO, Lihui; ZHANG, Xudong. An improved particle swarm optimization for protein folding prediction. **International Journal of Information Engineering and Electronic Business**, Modern Education and Computer Science Press, v. 3, n. 1, p. 1, 2011.
- CHENG, Jianlin; TEGGE, Allison N; BALDI, Pierre. Machine learning methods for protein structure prediction. **IEEE reviews in biomedical engineering**, IEEE, v. 1, p. 41–49, 2008.
- CHIVIAN, Dylan; ROBERTSON, Timothy; BONNEAU, Richard; BAKER, David. Ab initio methods. **Methods of biochemical analysis**, v. 44, p. 547–57, 2003.
- CHOWDHURY, Ratul; BOUATTA, Nazim; BISWAS, Surojit; FLORISTEAN, Christina; KHARKAR, Anant; ROY, Koushik; ROCHEREAU, Charlotte; AHDRIITZ, Gustaf; ZHANG, Joanna; CHURCH, George M, *et al.* Single-sequence protein structure prediction using a language model and deep learning. **Nature Biotechnology**, Nature Publishing Group US New York, v. 40, n. 11, p. 1617–1623, 2022.
- CLEMPNER, Julio B. Necessary and sufficient Karush–Kuhn–Tucker conditions for multiobjective Markov chains optimality. **Automatica**, Elsevier, v. 71, p. 135–142, 2016.
- COX, Gregory. Almost sure uniqueness of a global minimum without convexity. **The Annals of Statistics**, JSTOR, v. 48, n. 1, p. 584–606, 2020.
- CUSTÓDIO, Fábio Lima; BARBOSA, Helio JC; DARDENNE, Laurent Emmanuel. A multiple minima genetic algorithm for protein structure prediction. **Applied Soft Computing**, Elsevier, v. 15, p. 88–99, 2014.
- DANTZIG, George B. Linear programming. **Operations research**, INFORMS, v. 50, n. 1, p. 42–47, 2002.
- DEMAREST, Stephen J; HUA, Yuxin; RALEIGH, Daniel P. Local interactions drive the formation of nonnative structure in the denatured state of human α -lactalbumin: a high resolution structural characterization of a peptide model in aqueous solution. **Biochemistry**, ACS Publications, v. 38, n. 22, p. 7380–7387, 1999.
- DEMENTIEVA, Daria V; BOCHAROV, Eduard V; ARSENIEV, Alexander S. Two forms of cytotoxin II (cardiotoxin) from *Naja naja oxiana* in aqueous solution: spatial structures with tightly bound water molecules. **European Journal of Biochemistry**, Wiley Online Library, v. 263, n. 1, p. 152–162, 1999.
- DEMPE, Stephan; ZEMKOHO, Alain B. KKT reformulation and necessary conditions for optimality in nonsmooth bilevel optimization. **SIAM Journal on Optimization**, SIAM, v. 24, n. 4, p. 1639–1669, 2014.
- DHINGRA, Surbhi; SOWDHAMINI, Ramanathan; CADET, Frédéric; OFFMANN, Bernard. A glance into the evolution of template-free protein structure prediction methodologies. **Biochimie**, Elsevier, v. 175, p. 85–92, 2020.
- DOKKA, Trivikram; LETCHFORD, Adam N; MANSOOR, M Hasan. On the complexity of surrogate and group relaxation for integer linear programs. **Operations Research Letters**, Elsevier, v. 49, n. 4, p. 530–534, 2021.

- DORN, Márcio; SILVA, Mariel Barbachan e; BURIOL, Luciana; LAMB, Luis. Three-dimensional protein structure prediction: Methods and computational strategies. **Computational biology and chemistry**, Elsevier, v. 53, p. 251–276, 2014.
- DOS SANTOS, Ricardo Nascimento; JIANG, Xianli; MARTÍNEZ, Leandro; MORCOS, Faruck. Coevolutionary signals and structure-based models for the prediction of protein native conformations. In: **COMPUTATIONAL Methods in Protein Evolution**. New York, USA: Springer, 2018. P. 83–103.
- DU, Shuhan; TONG, Junbo; FAN, Wenhui. Learning efficient branch-and-bound for solving Mixed Integer Linear Programs. **Applied Soft Computing**, Elsevier, v. 172, p. 112863, 2025.
- DUBEY, Sandhya PN; BALAJI, S; GOPALAKRISHNA KINI, N; SATHISH KUMAR, M. A comparative study of various meta-heuristic algorithms for Ab initio protein structure prediction on 2D hydrophobic-polar model. In: 1. **PROCEEDINGS of Fifth International Conference on Soft Computing for Problem Solving: SocProS 2015**. Singapore: Springer, 2016. v. 2, p. 387–399.
- EMA, Romana Rahman; KHATUN, A; HOSSAIN, MA; AKHOND, Mostafijur Rahman; HOSSAIN, Nazmul; ARAFAT, Md Yasir. Protein secondary structure prediction using hybrid recurrent neural networks. **J. Comput. Sci**, v. 18, n. 7, p. 599–611, 2022.
- ESPAAS, Thomas A; VASSILIADIS, Vassilios S. Higher-order interior point methods for convex nonlinear programming. **Computers & Chemical Engineering**, Elsevier, v. 180, p. 108475, 2024.
- EZGI, Deniz Ülker. Elitist animal migration optimization for protein structure prediction based on 3D off-lattice model. **International Journal of Advanced Computer Science and Applications**, Science and Information (SAI) Organization Limited, v. 14, n. 3, 2023.
- FABIAN, Piotr; STĄPOR, Katarzyna. Developing a new SVM classifier for the extended ES protein structure prediction. In: **FEDERATED Conference on Computer Science and Information Systems**. Prague, Czech Republic: IEEE Press, 2017. P. 169–172.
- FALAE, Esther Adekemi. Particle swarm optimization for protein structure prediction and refinement. In: **NATURE-INSPIRED Metaheuristic Algorithms**. Boca Raton, Florida, EUA: CRC Press, 2025. P. 187–202.
- FASMAN, Gerald D. **Prediction of protein structure and the principles of protein conformation**. Berlim, Germany: Springer Science & Business Media, 2012.
- FERINA, Jennifer; DAGGETT, Valerie. Visualizing protein folding and unfolding. **Journal of molecular biology**, Elsevier, v. 431, n. 8, p. 1540–1564, 2019.
- FILGUEIRAS, Juan Luis; VARELA, Daniel; SANTOS, José. Energy minimization vs. deep learning approaches for protein structure prediction. In: **SPRINGER, 1. INTERNATIONAL Work-Conference on the Interplay Between Natural and Artificial Computation**. Puerto de la Cruz, Spain: [s.n.], 2022. v. 1, p. 109–118.
- FILIFE, Hugo AL; LOURA, Luís MS. Molecular dynamics simulations: advances and applications. **Molecules**, MDPI, v. 27, n. 7, p. 2105, 2022.
- FINKELSTEIN, Alexei V; BOGATYREVA, Natalya S; IVANKOV, Dmitry N; GARBUZYNSKIY, Sergiy O. Protein folding problem: enigma, paradox, solution. **Biophysical reviews**, Springer, v. 14, n. 6, p. 1255–1272, 2022.

- FISCHETTI, Matteo; MONACI, Michele. A branch-and-cut algorithm for mixed-integer bilinear programming. **European Journal of Operational Research**, Elsevier, v. 282, n. 2, p. 506–514, 2020.
- FRANCESCHINA, Mateus Antonio. Attention based neural networks for protein structure prediction. In: ANNUAL Research and Innovation Conference. Edmonton, Alberta, Canada: Concordia University of Edmonton, 2023. P. 1–6.
- GALANO-FRUTOS, Juan José; SANCHO, Javier. Energy, water, and protein folding: A molecular dynamics-based quantitative inventory of molecular interactions and forces that make proteins stable. **Protein Science**, Wiley Online Library, v. 33, n. 2, e4905, 2024.
- GONZALEZ-SANCHEZ, Belen; VEGA-RODRÍGUEZ, Miguel A; SANTANDER-JIMÉNEZ, Sergio. A multi-objective butterfly optimization algorithm for protein encoding. **Applied Soft Computing**, Elsevier, v. 139, p. 110269, 2023.
- GOSHISHT, Manoj Kumar. Machine learning and deep learning in synthetic biology: Key architectures, applications, and challenges. **ACS omega**, ACS Publications, v. 9, n. 9, p. 9921–9945, 2024.
- GOUGH, Nancy R; KALODIMOS, Charalampos G. Exploring the conformational landscape of protein kinases. **Current Opinion in Structural Biology**, Elsevier, v. 88, p. 102890, 2024.
- HAMEDUH, Tareq; MILLER, Andrew D; HEGER, Zbynek; HADDAD, Yazan. The proteomic code: Novel amino acid residue pairing models “encode” protein folding and protein-protein interactions. **Computers in Biology and Medicine**, Elsevier, v. 190, p. 110033, 2025.
- HART, William E; WATSON, Jean-Paul; WOODRUFF, David L. Pyomo: modeling and solving mathematical programs in Python. **Mathematical Programming Computation**, Springer, v. 3, p. 219–260, 2011.
- HIPP, Mark S; HARTL, F Ulrich. Interplay of proteostasis capacity and protein aggregation: implications for cellular function and disease. **Journal of Molecular Biology**, Elsevier, v. 436, n. 14, p. 168615, 2024.
- HIRATA, Fumio; SUGITA, Masatake; YOSHIDA, Masasuke; AKASAKA, Kazuyuki. Perspective: Structural fluctuation of protein and Anfinsen’s thermodynamic hypothesis. **The Journal of Chemical Physics**, AIP Publishing, v. 148, n. 2, 2018.
- HOCHBA, Dorit S. Approximation algorithms for NP-hard problems. **ACM Sigact News**, ACM, v. 28, n. 2, p. 40–52, 1997.
- HONG, Jun; ZHAN, Zhi-Hui; HE, Langchong; XU, Zongben; ZHANG, Jun. Protein Structure Prediction Using A New Optimization-Based Evolutionary and Explainable Artificial Intelligence Approach. **IEEE Transactions on Evolutionary Computation**, IEEE, 2024.
- HOQUE, Md Tamjidul; CHETTY, Madhu; SATTAR, Abdul. Genetic algorithm in ab initio protein structure prediction using low resolution model: a review. **Biomedical data and applications**, Springer, p. 317–342, 2009.
- ILARI, Andrea; SAVINO, Carmelinda. Protein structure determination by x-ray crystallography. **Bioinformatics: Data, Sequence Analysis and Evolution**, Springer, p. 63–87, 2008.

JACOBS, Steven A; GIBBS, Alan C; CONK, Michelle; YI, Fang; MAGUIRE, Diane; KANE, Colleen; O'NEIL, Karyn T. Fusion to a highly stable consensus albumin binding domain allows for tunable pharmacokinetics. **Protein Engineering, Design and Selection**, Oxford University Press, v. 28, n. 10, p. 385–393, 2015.

JANA, Nanda Dulal; DAS, Swagatam; SIL, Jaya. **A Metaheuristic Approach to Protein Structure Prediction: Algorithms and Insights from Fitness Landscape Analysis**. New York, USA: Springer, 2018. v. 31.

_____. Backgrounds on protein structure prediction and metaheuristics. In: **A Metaheuristic Approach to Protein Structure Prediction: Algorithms and Insights from Fitness Landscape Analysis**. New York, USA: Springer, 2018. P. 1–28.

_____. The Lévy distributed parameter adaptive differential evolution for protein structure prediction. **A Metaheuristic Approach to Protein Structure Prediction: Algorithms and Insights from Fitness Landscape Analysis**, Springer, p. 151–167, 2018.

JANA, Nanda Dulal; SIL, Jaya. Hybrid particle swarm optimization technique for protein structure prediction using 2D off-lattice model. In: **SWARM, Evolutionary, and Memetic Computing: 4th International Conference (SEMCCO 2013)**. Chennai, India: Springer, 2013. P. 193–204.

JANA, Nanda Dulal; SIL, Jaya; DAS, Swagatam. Hybrid metaheuristic approach for protein structure prediction. **A Metaheuristic Approach to Protein Structure Prediction: Algorithms and Insights from Fitness Landscape Analysis**, Springer, Berlin, Germany, p. 197–206, 2018.

_____. Selection of appropriate metaheuristic algorithms for protein structure prediction in AB off-lattice model: a perspective from fitness landscape analysis. **Information Sciences**, Elsevier, v. 391, p. 28–64, 2017.

JANES, Robert W; PEAPUS, Diane H; WALLACE, BA. The crystal structure of human endothelin. **Nature Structural Biology**, Nature, v. 1, n. 5, p. 311–319, 1994.

JHA, Kanchan; SAHA, Sriparna; SINGH, Hiteshi. Prediction of protein–protein interaction using graph neural networks. **Scientific Reports**, Nature Publishing Group UK London, v. 12, n. 1, p. 8360, 2022.

JIA, Zongchao; VANDONSELAAR, Margaret; HENGSTENBERG, Wolfgang; QUAIL, J Wilson; DELBAERE, Louis TJ. The 1·6 structure of histidine-containing phosphotransfer protein HPr from *Streptococcus faecalis*. **Journal of Molecular Biology**, Elsevier, v. 236, n. 5, p. 1341–1355, 1994.

JIANG, Haotian; KATHURIA, Tarun; LEE, Yin Tat; PADMANABHAN, Swati; SONG, Zhao. A faster interior point method for semidefinite programming. In: **61ST annual symposium on foundations of computer science (FOCS)**. Durham, North Carolina, United States of America: IEEE Press, 2020. P. 910–918.

JUMPER, John; EVANS, Richard; PRITZEL, Alexander, *et al.* Highly accurate protein structure prediction with AlphaFold. **Nature**, Nature Publishing Group, v. 596, n. 7873, p. 583–589, 2021.

JUMPER, John; EVANS, Richard; PRITZEL, Alexander; GREEN, Tim; FIGURNOV, Michael; RONNEBERGER, Olaf; TUNYASUVUNAKOOL, Kathryn; BATES, Russ; ŽÍDEK, Augustin; POTAPENKO, Anna, *et al.* Applying and improving AlphaFold at CASP14. **Proteins: Structure, Function, and Bioinformatics**, v. 89, n. 12, p. 1711–1721, 2021.

JUMPER, John; EVANS, Richard; PRITZEL, Alexander; GREEN, Tim; FIGURNOV, Michael; TUNYASUVUNAKOOL, Kathryn; RONNEBERGER, Olaf; BATES, Russ; ŽÍDEK, Augustin; BRIDGLAND, Alex, *et al.* AlphaFold 2. **Fourteenth Critical Assessment of Techniques for Protein Structure Prediction**, DeepMind London, UK, p. 13, 2020.

KARIM, Samah; SOLOMONIK, Edgar. Efficient preconditioners for interior point methods via a new schur complement-based strategy. **SIAM Journal on Matrix Analysis and Applications**, SIAM, v. 43, n. 4, p. 1680–1711, 2022.

KASPRZAK, Y; RÜCKERT, J; LUDOLPH, N; HÜBNER, CG; PAULSEN, H. Hydrogen bonds vs RMSD: Geometric reaction coordinates for protein folding. **The Journal of Chemical Physics**, AIP Publishing, v. 162, n. 7, 2025.

KIM, Gyuri; LEE, Sewon; LEVY KARIN, Eli; KIM, Hyunbin; MORIWAKI, Yoshitaka; OVCHINNIKOV, Sergey; STEINEGGER, Martin; MIRDITA, Milot. Easy and accurate protein structure prediction using ColabFold. **Nature Protocols**, Nature Publishing Group UK London, v. 20, n. 3, p. 620–642, 2025.

KIRDAR, Gözdenur; ÖZTÜRK, Diclehan Tezcaner; TESTIK, Murat Caner. Regression-based multiobjective optimization approaches for the autoclave composite parts placement problem: a mixed integer nonlinear programming model and an evolutionary algorithm. **Soft Computing**, Springer, v. 27, n. 19, p. 14045–14062, 2023.

KREZEL, Andrzej M; KASIBHATLA, Chandrasekhar; HIDALGO, Patricia; MACKINNON, Roderick; WAGNER, Gerhard. Solution structure of the potassium channel inhibitor agitoxin 2: caliper for probing channel geometry. **Protein Science**, Wiley Online Library, v. 4, n. 8, p. 1478–1489, 1995.

KUHLMAN, Brian; BRADLEY, Philip. Advances in protein structure prediction and design. **Nature Reviews Molecular Cell Biology**, Nature Publishing Group UK London, v. 20, n. 11, p. 681–697, 2019.

KURNIAWAN, I; HARYANTO, T; HASIBUAN, LS; AGMALARO, MA. Combining PSSM and physicochemical feature for protein structure prediction with support vector machine. **Journal of Physics: Conference Series**, IOP Publishing, v. 835, n. 1, p. 012006, 2017.

KUROCHKINA, N. Common structural characteristics of fibrous and globular proteins. **Proteomics Research Journal**, Nova Science Publishers, Inc., v. 3, n. 4, p. 373, 2012.

LEE, Jinwoo; JOO, Keehyoung; KIM, Seung-Yeon; LEE, Jooyoung. Re-examination of structure optimization of off-lattice protein AB models by conformational space annealing. **Journal of Computational Chemistry**, v. 29, n. 14, p. 2479–2484, 2008.

LEI, Zhenyu; GAO, Shangce; ZHANG, Zhiming; ZHOU, MengChu; CHENG, Jiujun. MO4: A many-objective evolutionary algorithm for protein structure prediction. **IEEE Transactions on Evolutionary Computation**, IEEE, v. 26, n. 3, p. 417–430, 2021.

- LENHARD, Johannes; STEPHAN, Simon; HASSE, Hans. On the History of the Lennard-Jones Potential. **Annalen der Physik**, Wiley Online Library, v. 536, n. 6, p. 2400115, 2024.
- LEYFFER, Sven; MANNS, Paul. McCormick envelopes in mixed-integer PDE-constrained optimization. **Mathematical Programming**, Springer, p. 1–46, 2025.
- LI, Bai; CHIONG, Raymond; LIN, Mu. A balance-evolution artificial bee colony algorithm for protein structure optimization based on a three-dimensional AB off-lattice model. **Computational biology and chemistry**, Elsevier, v. 54, p. 1–12, 2015.
- LI, Bai; LIN, Mu; LIU, Qiao; LI, Ya; ZHOU, Changjun. Protein folding optimization based on 3D off-lattice model via an improved artificial bee colony algorithm. **Journal of molecular modeling**, Springer, v. 21, p. 1–15, 2015.
- LI, Huili; NAEEM, Anisa; YOUSAF, Shamaila; ASLAM, Adnan; TCHIER, Fairouz; TOLA, Keneni Abera. Topological analysis and predictive modeling of amino acid structures with implications for bioinformatics and structural biology. **Scientific Reports**, Nature, v. 15, n. 1, p. 638, 2025.
- LI, Qian; LIU, San-Yang; YANG, Xin-She. Influence of initialization on the performance of metaheuristic optimizers. **Applied Soft Computing**, Elsevier, v. 91, p. 106193, 2020.
- LI, Qing; HU, Zhihang; WANG, Yixuan; LI, Lei; FAN, Yimin; KING, Irwin; JIA, Gengjie; WANG, Sheng; SONG, Le; LI, Yu. Progress and opportunities of foundation models in bioinformatics. **Briefings in Bioinformatics**, Oxford University Press, v. 25, n. 6, p. bbae548, 2024.
- LIN, Xiaoli; ZHANG, Xiaolong; ZHOU, Fengli. Protein structure prediction with local adjust tabu search algorithm. In: BMC bioinformatics. Berlin, Germany: Springer, 2014. v. 15, p. 1–11.
- LIU, Zhe; REYNOLDS, Albert. Robust multiobjective nonlinear constrained optimization with ensemble stochastic gradient sequential quadratic programming-filter algorithm. **SPE Journal**, OnePetro, v. 26, n. 04, p. 1964–1979, 2021.
- MAHER, Brian; ALBRECHT, Andreas A; LOOMES, Martin; YANG, Xin-She; STEINHÖFEL, Kathleen. A firefly-inspired method for protein structure prediction in lattice models. **Biomolecules**, MDPI, v. 4, n. 1, p. 56–75, 2014.
- MAKIGAKI, Shuichiro; ISHIDA, Takashi. Sequence alignment using machine learning for accurate template-based protein structure prediction. **Bioinformatics**, Oxford University Press, v. 36, n. 1, p. 104–111, 2020.
- MALHOTRA, Yuktika; JOHN, Jerry; YADAV, Deepika; SHARMA, Deepshikha; RAWAL, Kamal; MISHRA, Vaibhav; CHATURVEDI, Navaneet, *et al.* Advancements in protein structure prediction: A comparative overview of AlphaFold and its derivatives. **Computers in Biology and Medicine**, Elsevier, v. 188, p. 109842, 2025.
- MARCHI, Felipe; PARPINELLI, Rafael Stubs. A multi-objective approach to the protein structure prediction problem using the biased random-key genetic algorithm. In: IEEE PRESS. IEEE Congress on Evolutionary Computation (CEC). Krakow, Poland: [s.n.], 2021. P. 1070–1077.
- MARCU, Ștefan-Bogdan; TĂBÎRCĂ, Sabin; TANGNEY, Mark. An overview of AlphaFold's breakthrough. **Frontiers in artificial intelligence**, Frontiers Media SA, v. 5, p. 875587, 2022.

- MÁRQUEZ-CHAMORRO, Alfonso E; ASECIO-CORTÉS, Gualberto; SANTIESTEBAN-TOCA, Cosme E; AGUILAR-RUIZ, Jesús S. Soft computing methods for the prediction of protein tertiary structures: A survey. **Applied Soft Computing**, Elsevier, v. 35, p. 398–410, 2015.
- MARUYAMA, Yutaka; IGARASHI, Ryo; USHIKU, Yoshitaka; MITSUTAKE, Ayori. Analysis of protein folding simulation with moving root mean square deviation. **Journal of Chemical Information and Modeling**, ACS Publications, v. 63, n. 5, p. 1529–1541, 2023.
- MCCORMICK, Garth P. Computability of Global Solutions To Factorable Nonconvex Solutions: Part I: Convex Underestimating Problems. **Mathematical Programming**, v. 10, p. 147–175, 1976.
- MCDONALD, Eli Fritz; MEILER, Jens; PLATE, Lars. CFTR folding: from structure and proteostasis to cystic fibrosis personalized medicine. **ACS Chemical Biology**, ACS Publications, v. 18, n. 10, p. 2128–2143, 2023.
- MEHROTRA, Sanjay. On the implementation of a primal-dual interior point method. **SIAM Journal on optimization**, SIAM, v. 2, n. 4, p. 575–601, 1992.
- MENG, Yajie; ZHANG, Zhuang; ZHOU, Chang; TANG, Xianfang; HU, Xinrong; TIAN, Geng; YANG, Jialiang; YAO, Yuhua. Protein structure prediction via deep learning: an in-depth review. **Frontiers in Pharmacology**, Frontiers Media SA, v. 16, p. 1498662, 2025.
- MISIR, Mustafa. Algorithm selection for protein structure prediction on 2d ab off-lattice model. In: 1. PROCEEDINGS of the 8th International Conference on Computational Biology and Bioinformatics. Kyoto, Japan: [s.n.], 2024. P. 75–83.
- MISURA, Kira MS; SCHELLER, Richard H; WEIS, William I. Self-association of the H3 region of syntaxin 1A. Implications for intermediates in SNARE complex assembly. **Journal of Biological Chemistry**, ASBMB, v. 276, n. 16, p. 13273–13282, 2001.
- MOLINA-PÉREZ, Daniel; MEZURA-MONTES, Efrén; PORTILLA-FLORES, Edgar Alfredo; VEGA-ALVARADO, Eduardo; CALVA-YAÑEZ, Bárbara. A differential evolution algorithm for solving mixed-integer nonlinear programming problems. **Swarm and Evolutionary Computation**, Elsevier, v. 84, p. 101427, 2024.
- MOLINA-PÉREZ, Daniel; PORTILLA-FLORES, Edgar Alfredo; MEZURA-MONTES, Efrén; VEGA-ALVARADO, Eduardo; CALVA-YAÑEZ, María Bárbara. Efficiently handling constraints in mixed-integer nonlinear programming problems using gradient-based repair differential evolution. **PeerJ Computer Science**, PeerJ Inc., v. 10, e2095, 2024.
- MOORE, Peter B; HENDRICKSON, Wayne A; HENDERSON, Richard; BRUNGER, Axel T. The protein-folding problem: Not yet solved. **Science**, American Association for the Advancement of Science, v. 375, n. 6580, p. 507–507, 2022.
- MOTMAEN, Amir; DAUPARAS, Justas; BAEK, Minkyung; ABEDI, Mohamad H; BAKER, David; BRADLEY, Philip. Peptide-binding specificity prediction using fine-tuned protein structure prediction networks. **Proceedings of the National Academy of Sciences**, National Academy of Sciences, v. 120, n. 9, e2216697120, 2023.

MUFASSIRIN, MM Mohamed; NEWTON, MA Hakim; SATTAR, Abdul. Artificial intelligence for template-free protein structure prediction: a comprehensive review. **Artificial Intelligence Review**, Springer, v. 56, n. 8, p. 7665–7732, 2023.

NA, Sen; ANITESCU, Mihai; KOLAR, Mladen. Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming. **Mathematical Programming**, Springer, v. 202, n. 1, p. 279–353, 2023.

NARLOCH, Pedro Henrique; DORN, Márcio. Evaluating the success-history based adaptive differential evolution in the protein structure prediction problem. In: SPRINGER. INTERNATIONAL Conference on the Applications of Evolutionary Computation. Online event: [s.n.], 2021. P. 194–209.

NARLOCH, Pedro Henrique; PARPINELLI, Rafael Stubs. Diversification strategies in differential evolution algorithm to solve the protein structure prediction problem. In: INTELLIGENT Systems Design and Applications: 16th International Conference on Intelligent Systems Design and Applications. Porto, Portugal: Springer, 2017. P. 125–134.

_____. The protein structure prediction problem approached by a cascade differential evolution algorithm using ROSETTA. In: 2017 Brazilian Conference on Intelligent Systems. Uberlandia, Minas Gerais, Brazil: IEEE Press, 2017. P. 294–299.

NEMIROVSKI, Arkadi S; TODD, Michael J. Interior-point methods for optimization. **Acta Numerica**, Cambridge university press, v. 17, p. 191–234, 2008.

OLADEJO, Sunday O. Metaheuristics for protein structure prediction: review and empirical analysis. In: IEEE, 1. IEEE International Conference on Artificial Intelligence and Mechatronics Systems. New York, USA: [s.n.], 2024. P. 1–6.

OLIVEIRA, Mariana; BORGUESAN, Bruno; DORN, Marcio. SADE-SPL: A Self-Adapting Differential Evolution algorithm with a loop Structure Pattern Library for the PSP problem. In: 2017 IEEE Congress on Evolutionary Computation (CEC). Donostia, Spain: IEEE Press, 2017. P. 1095–1102.

PADILLA-GODÍNEZ, Francisco J; RAMOS-ACEVEDO, Rodrigo; MARTÍNEZ-BECERRIL, Hilda Angélica; BERNAL-CONDE, Luis D; GARRIDO-FIGUEROA, Jerónimo F; HIRIART, Marcia; HERNÁNDEZ-LÓPEZ, Adriana; ARGÜERO-SÁNCHEZ, Rubén; CALLEA, Francesco; GUERRA-CRESPO, Magdalena. Protein misfolding and aggregation: the relatedness between Parkinson's disease and hepatic endoplasmic reticulum storage disorders. **International journal of molecular sciences**, MDPI, v. 22, n. 22, p. 12467, 2021.

PAPP, Dávid; VARGA, Anita. Interior-point algorithms with full Newton steps for nonsymmetric convex conic optimization. **SIAM Journal on Optimization**, SIAM, v. 35, n. 4, p. 2490–2517, 2025.

PARK, Sanggeun; MYUNG, Sojung; BAEK, Minkyung. Advancing protein structure prediction beyond AlphaFold2. **Current Opinion in Structural Biology**, Elsevier, v. 90, p. 102985, 2025.

PARPINELLI, Rafael Stubs; BENITIEZ, Cesar MV; CORDEIRO, Jelson; LOPES, Heitor Silverio. Performance Analysis of Swarm Intelligence Algorithms for the 3D-AB off-lattice Protein Folding Problem. **J. Multiple Valued Log. Soft Comput.**, v. 22, n. 3, p. 267–286, 2014.

PDB. **Protein Data Bank**. United States of America: RCBS, 2023. Accessed: August 2023. Available from: <<https://www.rcsb.org/>>.

PEARCE, Robin; ZHANG, Yang. Deep learning techniques have significantly impacted protein structure prediction and protein design. **Current opinion in structural biology**, Elsevier, v. 68, p. 194–207, 2021.

_____. Toward the solution of the protein structure prediction problem. **Journal of Biological Chemistry**, ASBMB, v. 297, n. 1, 2021.

PRATT, Olivia S; ELLIOTT, Luc G; HAON, Margaux; MESDAGHI, Shahram; PRICE, Rebecca M; SIMPKIN, Adam J; RIGDEN, Daniel J. AlphaFold 2, but not AlphaFold 3, predicts confident but unrealistic β -solenoid structures for repeat proteins. **Computational and structural biotechnology journal**, Elsevier, v. 27, p. 467–477, 2025.

RAKSHANI, Hojjat; IDOUMGHAR, Lhassane; LEPAGNOT, Julien; BRÉVILLIERS, Mathieu. Speed up differential evolution for computationally expensive protein structure prediction problems. **Swarm and Evolutionary Computation**, Elsevier, v. 50, p. 100493, 2019.

RANGANATHAN, Srivastav; LIU, Junlang; SHAKHNOVICH, Eugene. Enzymatic metabolons dramatically enhance metabolic fluxes of low-efficiency biochemical reactions. **Biophysical journal**, Elsevier, v. 122, n. 23, p. 4555–4566, 2023.

RASHID, Mahmood A; KHATIB, Firas; HOQUE, Md Tamjidul; SATTAR, Abdul. An enhanced genetic algorithm for ab initio protein structure prediction. **IEEE Transactions on Evolutionary Computation**, IEEE, v. 20, n. 4, p. 627–644, 2015.

ROCHA, Leandro Pohlmann; CAMPONOGARA, Eduardo; SEMAN, Laio Oriel. Multiparametric disaggregation relaxation of bilinear terms for the operational management of crude oil supply. In: IEEE International Conference on Automation Science and Engineering. Lyon, France: IEEE Press, 2021.

ROUX, Daniel de; CARR, Robert; RAVI, R. Instance-specific linear relaxations of semidefinite optimization problems. **Mathematical Programming Computation**, Springer, p. 1–51, 2025.

RYU, Ernest K; YIN, Wotao. **Large-scale convex optimization: algorithms & analyses via monotone operators**. Cambridge, United Kingdom: Cambridge University Press, 2022.

SAR, Enakshi; ACHARYYA, Sriyankar. Genetic algorithm variants in predicting protein structure. In: 2014 International Conference on Communication and Signal Processing. Melmaruvathur, India: IEEE Press, 2014. P. 321–325.

SARGSYAN, Karen; GRAUFFEL, Cédric; LIM, Carmay. How molecular size impacts RMSD applications in molecular dynamics simulations. **Journal of chemical theory and computation**, ACS Publications, v. 13, n. 4, p. 1518–1524, 2017.

SAXENA, Akash; KUMAR, Rajesh. Chaotic variants of grasshopper optimization algorithm and their application to protein structure prediction. **Applied nature-inspired computing: Algorithms and case studies**, Springer, p. 151–175, 2020.

SAXENA, Akash; SHEKHAWAT, Shalini; SHARMA, Ajay; SHARMA, Harish; KUMAR, Rajesh. Application of hybridized whale optimization for protein structure prediction. In: 2021 International Conference on Innovative Computing and Communications. Delhi, India: Springer, 2022. v. 2, p. 141–150.

SCAVUZZO, Lara; AARDAL, Karen; LODI, Andrea; YORKE-SMITH, Neil. Machine learning augmented branch and bound for mixed integer linear programming.

Mathematical Programming, Springer, p. 1–44, 2024.

SCHAUPERL, Michael; DENNY, Rajiah Aldrin. AI-based protein structure prediction in drug discovery: impacts and challenges. **Journal of Chemical Information and Modeling**, ACS Publications, v. 62, n. 13, p. 3142–3156, 2022.

SEKHAR, Sivagnanam Rajamanickam Mani; MATT, Siddesh Gaddadevara; KUMARA MAHADEVACHAR, Vasantha. Protein tertiary structure prediction by integrating ant colony optimization with path relinking and structure knowledge. **International Journal of Information Technology**, Springer, v. 15, n. 3, p. 1399–1405, 2023.

SELVARAJ, Chandrabose; VIJAYALAKSHMI, Periyasamy; DESAI, Dipali; MANOHARAN, Jayakumar. Proteostasis imbalance: Unraveling protein aggregation in neurodegenerative diseases and emerging therapeutic strategies. **Advances in Protein Chemistry and Structural Biology**, Elsevier, v. 146, p. 1–34, 2025.

SENIOR, Andrew W; EVANS, Richard; JUMPER, John; KIRKPATRICK, James; SIFRE, Laurent; GREEN, Tim; QIN, Chongli; ŽÍDEK, Augustin; NELSON, Alexander WR; BRIDGLAND, Alex, *et al.* Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). **Proteins: structure, function, and bioinformatics**, Wiley Online Library, v. 87, n. 12, p. 1141–1148, 2019.

SHINTAKU, Yuichi; INAOKA, Tatsuhiko; TERADA, Kenjiro. A primal–dual interior point method to implicitly update Gurson–Tvergaard–Needleman model. **Computational Mechanics**, Springer, p. 1–18, 2024.

SHUCHUN, Yu; XIANXIANG, Li; XUE, Tian; MING, Pang. Protein structure prediction based on particle swarm optimization and tabu search strategy. **BMC bioinformatics**, BioMed Central, v. 23, n. 10, p. 1–10, 2022.

SILVA, Luiza Scapinello Aquino da. **Otimização Jaya e variantes aplicada ao problema de predição de estruturas de proteínas**. 2021. PhD thesis – Universidade Federal do Paraná, Curitiba, PR, Brazil.

SILVA, Luiza Scapinello Aquino da; SEMAN, Laio Oriel; CAMPONOGARA, Eduardo; MARIANI, Viviana Cocco; SANTOS COELHO, Leandro dos. Bilinear optimization of protein structure prediction: An exact approach via AB off-lattice model. **Computers in biology and medicine**, Elsevier, v. 176, p. 108558, 2024.

SIQUEIRA, Lucas; VENSKE, Sandra. Ab initio protein structure prediction using evolutionary approach: a survey. **Revista de Informática Teórica e Aplicada**, v. 28, n. 2, p. 11–24, 2021.

SONG, Shuangbao; JI, Junkai; CHEN, Xingqian; GAO, Shangce; TANG, Zheng; TODO, Yuki. Adoption of an improved PSO to explore a compound multi-objective energy function in protein structure prediction. **Appl. Soft Comput.**, v. 72, p. 539–551, 2018.

SOTO-MONTEERRUBIO, Diego A; PERAZA-VÁZQUEZ, Hernán; PEÑA-DELGADO, Adrián F; GONZÁLEZ-HERNÁNDEZ, José G. Enhanced Methodology for Peptide Tertiary Structure Prediction Using GRSA and Bio-Inspired Algorithm. **International Journal of Molecular Sciences**, MDPI, v. 26, n. 15, p. 7484, 2025.

STAHL, Kolja; GRAZIADEI, Andrea; DAU, Therese; BROCK, Oliver; RAPPSILBER, Juri. Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning. **Nature Biotechnology**, Nature Publishing Group US New York, v. 41, n. 12, p. 1810–1819, 2023.

STILLINGER, Frank H; HEAD-GORDON, Teresa; HIRSHFELD, Catherine L. Toy model for protein folding. **Physical Review E**, APS, v. 48, n. 2, p. 1469, 1993.

STUBS PARPINELLI, Rafael; BOIANI, Mateus; DIAS, André EP. A massively parallel speciation-based differential evolution algorithm applied to the 3D-AB protein structure prediction. **Concurrency and Computation: Practice and Experience**, Wiley Online Library, v. 35, n. 17, e6745, 2023.

SUMA, LS; VINOD CHANDRA, SS. Mining of structural motifs in proteins using artificial bee colony optimization framework for druggability. **Journal of Bioinformatics and Computational Biology**, World Scientific, v. 19, n. 05, p. 2150025, 2021.

SUNDAR, Kaarthik; SANJEEVI, Sujeevraja; NAGARAJAN, Harsha. Sequence of polyhedral relaxations for nonlinear univariate functions. **Optimization and Engineering**, Springer, v. 23, n. 2, p. 877–894, 2022.

TAKABE, Satoshi; HUKUSHIMA, Koji. Statistical mechanical analysis of linear programming relaxation for combinatorial optimization problems. **Physical Review E**, APS, v. 93, n. 5, p. 053308, 2016.

TECALCO–CRUZ, Angeles C; PEDRAZA-CHAVERRI, José; BRIONES-HERRERA, Alfredo; CRUZ-RAMOS, Eduardo; LÓPEZ–CANOVAS, Lilia; ZEPEDA–CERVANTES, Jesús. Protein degradation-associated mechanisms that are affected in Alzheimer s disease. **Molecular and cellular biochemistry**, Springer, v. 477, n. 3, p. 915–925, 2022.

TEETER, Martha M. Water structure of a hydrophobic protein at atomic resolution: Pentagon rings of water molecules in crystals of crambin. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 81, n. 19, p. 6014–6018, 1984.

TEIXEIRA, Samantha K; LIMA, Angélica N; RESENDE-LARA, Pedro Túlio; OLIVEIRA, Luciana R de; TOLEDO, Marcelo AF de. Protein Structure Prediction Methods. **Unlocking Protein Structure-Function Relationships: Computational and Experimental Approaches and Applications**, Springer, p. 3–50, 2026.

TERDALKAR, Samir. Bioinformatics: An Effective Teaching Resource in Biological Sciences. **The Voice of Creative Research**, v. 7, n. 2, p. 173–178, 2025.

TORRISI, Mirko; POLLASTRI, Gianluca; LE, Quan. Deep learning methods in protein structure prediction. **Computational and Structural Biotechnology Journal**, Elsevier, v. 18, p. 1301–1310, 2020.

TRACY, Kevin; MANCHESTER, Zachary. On the Differentiability of the Primal-Dual Interior-Point Method. **arXiv preprint arXiv:2406.11749**, 2024.

TSAI, Yingssu; SAWAYA, Michael R; CANNON, Gordon C; CAI, Fei; WILLIAMS, Eric B; HEINHORST, Sabine; KERFELD, Cheryl A; YEATES, Todd O. Structural analysis of CsoS1A and the protein shell of the Halothiobacillus neapolitanus Carboxysome. **PLoS Biology**, Public Library of Science San Francisco, USA, v. 5, n. 6, p. 144, 2007.

VAN THIEU, Nguyen; MIRJALILI, Seyedali. MEALPY: An open-source library for latest meta-heuristic algorithms in Python. **Journal of Systems Architecture**, Elsevier, v. 139, p. 102871, 2023.

- VARADI, Mihaly; ANYANGO, Stephen; DESHPANDE, Mandar; NAIR, Sreenath; NATASSIA, Cindy; YORDANOVA, Galabina; YUAN, David; STROE, Oana; WOOD, Gemma; LAYDON, Agata, *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. **Nucleic acids research**, Oxford University Press, v. 50, n. D1, p. d439–d444, 2022.
- VARADI, Mihaly; TSENKOV, Maxim; VELANKAR, Sameer. Challenges in bridging the gap between protein structure prediction and functional interpretation. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 93, n. 1, p. 400–410, 2025.
- VARELA, Daniel; SANTOS, José. Crowding differential evolution for protein structure prediction. In: SPRINGER, 1. INTERNATIONAL Work-Conference on the Interplay Between Natural and Artificial Computation. Almeria, Spain: [s.n.], 2019. P. 193–203.
- _____. Niching methods integrated with a differential evolution memetic algorithm for protein structure prediction. **Swarm and Evolutionary Computation**, Elsevier, v. 71, p. 101062, 2022.
- WANG, Wenkai; PENG, Zhenling; YANG, Jianyi. Single-sequence protein structure prediction using supervised transformer protein language models. **Nature Computational Science**, Nature Publishing Group US New York, v. 2, n. 12, p. 804–814, 2022.
- WANG, Youle; ZHOU, Xiangzhen. Efficient quantum algorithm for lattice protein folding. **Quantum Science and Technology**, IOP Publishing, v. 10, n. 1, p. 015056, 2024.
- WAUDBY, Christopher A; DOBSON, Christopher M; CHRISTODOULOU, John. Nature and regulation of protein folding on the ribosome. **Trends in biochemical sciences**, Elsevier, v. 44, n. 11, p. 914–926, 2019.
- WHITFORD, David. **Proteins: structure and function**. Hoboken, New Jersey, United States of America: John Wiley & Sons, 2013.
- WILHELM, Matthew E; WANG, Chenyu; STUBER, Matthew D. Convex and concave envelopes of artificial neural network activation functions for deterministic global optimization. **Journal of Global Optimization**, Springer, v. 85, n. 3, p. 569–594, 2023.
- WRIGHT, Stephen J. **Primal-dual interior-point methods**. Philadelphia, Pennsylvania, United States of America: SIAM, 1997.
- WUETHRICH, Kurt. The development of nuclear magnetic resonance spectroscopy as a technique for protein structure determination. **Accounts of chemical research**, ACS Publications, v. 22, n. 1, p. 36–44, 1989.
- XIE, Shangxin; LI, Zhong; HU, Hailong. Protein secondary structure prediction based on the fuzzy support vector machine with the hyperplane optimization. **Gene**, Elsevier, v. 642, p. 74–83, 2018.
- XU, Dong; ZHANG, Yang. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 80, n. 7, p. 1715–1735, 2012.
- XU, Jinbo; MCPARTLON, Matthew; LI, Jin. Improved protein structure prediction by deep learning irrespective of co-evolution information. **Nature Machine Intelligence**, Nature Publishing Group UK London, v. 3, n. 7, p. 601–609, 2021.

YE, Jason; SCOTT, Joseph K. Extended McCormick relaxation rules for handling empty arguments representing infeasibility. **Journal of Global Optimization**, Springer, v. 87, n. 1, p. 57–95, 2023.

YIN, Tianxiang; CHEN, Yunxuan; SU, Hongyu; WANG, Yuhang; ZHAO, Yixin; DUAN, Chengxu; LIU, Jingrui. Protein Structure Prediction Powered by Artificial Intelligence: From Biochemical Foundations to Practical Applications. **Frontiers in Molecular Biosciences**, Frontiers, v. 13, p. 1767821, 2026.

YOUSEF, Mohamad; ABDELKADER, Tamer; EL-BAHNASY, Khaled. An efficient protein structure prediction using genetic algorithm. In: *INNOVATIVE Smart Healthcare and Bio-Medical Systems*. Boca Raton, USA: CRC Press, 2020. P. 87–106.

ZAKARIA, Nur Aini; SHAH, Zuraini Ali; KASIM, Shahreen. Protein structure prediction using robust principal component analysis and support vector machine. **International Journal of Data Science**, v. 1, n. 1, p. 14–17, 2020.

ZHAN, Zhi-Hui; HONG, Jun; LI, Jian-Yu; WANG, Cheng; HE, Langchong; XU, Zongben; ZHANG, Jun. Artificial intelligence-based methods for protein structure prediction: a survey. **Artificial Intelligence Review**, Springer, v. 58, n. 10, p. 328, 2025.

ZHANG, Lizhong; MA, He; QIAN, Wei; LI, Haiyan. Protein structure optimization using improved simulated annealing algorithm on a three-dimensional AB off-lattice model. **Computational Biology and Chemistry**, Elsevier, v. 85, p. 107237, 2020.

ZHAO, Weiguo; ZHANG, Zhenxing; WANG, Liying. Manta ray foraging optimization: An effective bio-inspired optimizer for engineering applications. **Engineering Applications of Artificial Intelligence**, Elsevier, v. 87, p. 103300, 2020.

ZHEN, Jianzhe; MARANDI, Ahmadreza; MOOR, Danique de; HERTOOG, Dick den; VANDENBERGHE, Lieven. Disjoint bilinear optimization: A two-stage robust optimization perspective. **INFORMS Journal on Computing**, INFORMS, v. 34, n. 5, p. 2410–2427, 2022.

ZHENG, Wei; WUYUN, Qiqige; LI, Yang; ZHANG, Chengxin; FREDDOLINO, P Lydia; ZHANG, Yang. Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data. **Nature Methods**, Nature, v. 21, n. 2, p. 279–289, 2024.

ZHOU, Changjun; HOU, Caixia; WEI, Xiaopeng; ZHANG, Qiang. Improved hybrid optimization algorithm for 3D protein structure prediction. **Journal of molecular modeling**, Springer, v. 20, p. 1–12, 2014.

ZHOU, Xiao-Gen; PENG, Chun-Xiang; LIU, Jun; ZHANG, Yang; ZHANG, Gui-Jun. Underestimation-assisted global-local cooperative differential evolution and the application to protein structure prediction. **IEEE transactions on evolutionary computation**, IEEE, v. 24, n. 3, p. 536–550, 2019.

ZHOU, Yanlin; TAN, Kai; SHEN, Xinyu; HE, Zheng; ZHENG, Haotian. A Protein Structure Prediction Approach Leveraging Transformer and CNN Integration. In: *7TH International Conference on Advanced Algorithms and Control Engineering (ICAACE)*. Shanghai, China: IEEE Press, 2024. P. 749–753.

APPENDIX A – MAIN PYTHON CODE

```

from Bio.PDB import *
from pyomo.environ import *
import requests
import math
import numpy as np
from RMSDfrom3DAB.getRMSD import *

def fetch_sequence(pdb_id):
    url = f"https://www.rcsb.org/fasta/entry/{pdb_id}"
    r = requests.get(url)
    return ''.join(r.text.split('\n')[1:])

def map_to_AB(sequence):
    hydrophobic = "IVLPCMAG"
    mapped_sequence = ''
    for aa in sequence:
        if aa in hydrophobic:
            mapped_sequence += 'A'
        else:
            mapped_sequence += 'B'
    return mapped_sequence

# Fetch 1
pdb_id = "1CB3"

'''
pbd_id used:
1CB3
1EDP
1EDN
1AGT
1CRN
2N35
1CB9
1HWV
1PTF
2G13
1A07
133L
10GS
'''

```

```
# Map to AB
ab_sequence = map_to_AB(sequence)
proteina = [*ab_sequence]
sequencia = [0] + proteina
N = len(sequencia)
n = len(sequencia)

from pyomo.environ import *
#import pyomo as pyo

model = ConcreteModel()

model.N = RangeSet(1, N)
model.D = RangeSet(0, 2) # for 3D coordinates x, y, z

model.cos_x = Var(model.N, within=Reals, bounds=(-1, 1))
model.sin_x = Var(model.N, within=Reals, bounds=(-1, 1))
model.cos_b = Var(model.N, within=Reals, bounds=(-1, 1))
model.sin_b = Var(model.N, within=Reals, bounds=(-1, 1))
model.position = Var(model.N, model.D, within=Reals,
                     bounds=(-100, 100))

def trig_x_rule(model, k):
    return model.cos_x[k] * model.cos_x[k] + model.sin_x[k] * model.sin_x[k]
    == 1
model.trig_x_constr = Constraint(range(1,N), rule=trig_x_rule)

def trig_b_rule(model, k):
    return model.cos_b[k] * model.cos_b[k] + model.sin_b[k] * model.sin_b[k]
    == 1
model.trig_b_constr = Constraint(range(1,N), rule=trig_b_rule)

model.add_constraint0a = Constraint(expr=model.cos_x[1] == 0)
model.add_constraint0b = Constraint(expr=model.sin_x[1] == 1)
model.add_constraint0c = Constraint(expr=model.cos_b[1] == 0)
model.add_constraint0d = Constraint(expr=model.sin_b[1] == 1)

model.add_constraint1a = Constraint(expr=model.position[1, 0] == 0)
model.add_constraint2a = Constraint(expr=model.position[1, 1] == 0)
model.add_constraint3a = Constraint(expr=model.position[1, 2] == 0)

model.add_constraint1b = Constraint(expr=model.position[2, 0] == 0)
model.add_constraint2b = Constraint(expr=model.position[2, 1] == 1)
model.add_constraint3b = Constraint(expr=model.position[2, 2] == 0)

model.add_constraint1c = Constraint(expr=model.position[3, 0]
                                   == model.cos_x[1])
```

```

model.add_constraint2c = Constraint(expr=model.position[3, 1]
                                   == 1 + model.sin_x[1])
model.add_constraint3c = Constraint(expr=model.position[3, 2] == 0)

def pos_x_rule(model, j):
    return model.position[j, 0] == model.position[j-1, 0]
                                   + model.cos_x[j-2] * model.cos_b[j-3]
model.pos_x_constr = Constraint(range(4,N), rule=pos_x_rule)

def pos_y_rule(model, j):
    return model.position[j, 1] == model.position[j-1, 1]
                                   + model.sin_x[j-1] * model.cos_b[j-3]
model.pos_y_constr = Constraint(range(4,N), rule=pos_y_rule)

def pos_z_rule(model, j):
    return model.position[j, 2] == model.position[j-1, 2] + model.sin_b[j-3]
model.pos_z_constr = Constraint(range(4,N), rule=pos_z_rule)

# Define r and r_hat
model.r = Var(range(1, N), range(1, N), within=NonNegativeReals)
model.r_hat = Var(range(1, N), range(1, N), within=NonNegativeReals)
model.r2 = Var(range(1, N), range(1, N), within=NonNegativeReals)
model.r3 = Var(range(1, N), range(1, N), within=NonNegativeReals)

# Additional variables
model.v1 = Var(within=Reals)
model.v2 = Var(within=Reals)

for i in range(4, N):
    model.position[i, 0].setlb(-(i-2))
    model.position[i, 0].setub(i-2)

    model.position[i, 1].setlb(-(i-3))
    model.position[i, 1].setub(i-1)

    model.position[i, 2].setlb(-(i-3))
    model.position[i, 2].setub(i-3)

# Define your expression
# Constraints for r, r_hat, r2, r3 variables
trupla = []
for i in range(1,n-2):
    for j in range(i+2, n):
        trupla.append((i, j))

def r_constraints_rule(model, i, j):
    return model.r[i, j] == sum((model.position[j, d]

```

```

        - model.position[i, d]) ** 2 for d in [0,1,2])
model.r_constraints = Constraint(trupla, rule=r_constraints_rule)
def r_hat_constraint(model, i, j):
    return model.r[i, j] * model.r_hat[i, j] == 1
model.r_hat_con = Constraint(trupla, rule=r_hat_constraint)
def r2_constraint(model, i, j):
    return model.r_hat[i, j] * model.r_hat[i, j] == model.r2[i, j]
model.r2_con = Constraint(trupla, rule=r2_constraint)
def r3_constraint(model, i, j):
    return model.r2[i, j] * model.r_hat[i, j] == model.r3[i, j]
model.r3_con = Constraint(trupla, rule=r3_constraint)

def ccc(i, j):
    aai, aaj = sequencia[i], sequencia[j]
    if aai == "A" and aaj == "A":
        return 1
    elif aai == "B" and aaj == "B":
        return 0.5
    else:
        return -0.5

model.v1_con = Constraint(expr=model.v1 >= (1/4) * sum((1 - model.cos_x[i])
                                                    for i in range(2, n-1)))
model.v2_con = Constraint(expr=model.v2 >= 4*sum(model.r3[i, j] * (model.r3[i, j]
                                                                - ccc(i, j)) for (i, j) in trupla))

model.objective = Objective(expr=model.v1 + model.v2, sense=minimize)

TransformationFactory('contrib.constraints_to_var_bounds').apply_to(model)
TransformationFactory('contrib.induced_linearity').apply_to(model)
TransformationFactory('contrib.propagate_fixed_vars').apply_to(model)
TransformationFactory('contrib.propagate_zero_sum').apply_to(model)
TransformationFactory('contrib.propagate_eq_var_bounds').apply_to(model)
TransformationFactory('contrib.deactivate_trivial_constraints').apply_to(model)

solver = SolverFactory('ipopt')#knitro',
        executable = '/home/seman/Downloads/ampl.linux-intel64/knitro')
# set max iterations
solver.options['maxit'] = 5000

results = solver.solve(model, tee=True)# will display solver output

model = ConcreteModel()

model.N = RangeSet(1, N)
model.D = RangeSet(0, 2) # for 3D coordinates x, y, z

```

```

model.cos_x = Var(model.N, within=Reals, bounds=(-1, 1))
model.sin_x = Var(model.N, within=Reals, bounds=(-1, 1))
model.cos_b = Var(model.N, within=Reals, bounds=(-1, 1))
model.sin_b = Var(model.N, within=Reals, bounds=(-1, 1))
model.position = Var(model.N, model.D, within=Reals, bounds=(-100, 100))

def trig_x_rule(model, k):
    return model.cos_x[k] * model.cos_x[k] + model.sin_x[k] * model.sin_x[k]
    == 1
model.trig_x_constr = Constraint(range(1,N), rule=trig_x_rule)

def trig_b_rule(model, k):
    return model.cos_b[k] * model.cos_b[k] + model.sin_b[k] * model.sin_b[k]
    == 1
model.trig_b_constr = Constraint(range(1,N), rule=trig_b_rule)

model.add_constraint0a = Constraint(expr=model.cos_x[1] == 0)
model.add_constraint0b = Constraint(expr=model.sin_x[1] == 1)
model.add_constraint0c = Constraint(expr=model.cos_b[1] == 0)
model.add_constraint0d = Constraint(expr=model.sin_b[1] == 1)

model.add_constraint1a = Constraint(expr=model.position[1, 0] == 0)
model.add_constraint2a = Constraint(expr=model.position[1, 1] == 0)
model.add_constraint3a = Constraint(expr=model.position[1, 2] == 0)

model.add_constraint1b = Constraint(expr=model.position[2, 0] == 0)
model.add_constraint2b = Constraint(expr=model.position[2, 1] == 1)
model.add_constraint3b = Constraint(expr=model.position[2, 2] == 0)

# Add constraints
model.add_constraint1c = Constraint(expr=model.position[3, 0]
    == model.cos_x[1])
model.add_constraint2c = Constraint(expr=model.position[3, 1]
    == 1 + model.sin_x[1])
model.add_constraint3c = Constraint(expr=model.position[3, 2] == 0)

def pos_x_rule(model, j):
    return model.position[j, 0] == model.position[j-1, 0]
    + model.cos_x[j-2] * model.cos_b[j-3]
model.pos_x_constr = Constraint(range(4,N), rule=pos_x_rule)

def pos_y_rule(model, j):
    return model.position[j, 1] == model.position[j-1, 1]
    + model.sin_x[j-1] * model.cos_b[j-3]
model.pos_y_constr = Constraint(range(4,N), rule=pos_y_rule)

```

```

def pos_z_rule(model, j):
    return model.position[j, 2] == model.position[j-1, 2] + model.sin_b[j-3]
model.pos_z_constr = Constraint(range(4,N), rule=pos_z_rule)

# Define r and r_hat
model.r = Var(range(1, N), range(1, N), within=NonNegativeReals)
model.r_hat = Var(range(1, N), range(1, N), within=NonNegativeReals)
model.r2 = Var(range(1, N), range(1, N), within=NonNegativeReals)
model.r3 = Var(range(1, N), range(1, N), within=NonNegativeReals)

# Additional variables
model.v1 = Var(within=Reals)
model.v2 = Var(within=Reals)

for i in range(4, N):
    model.position[i, 0].setlb(-(i-2))
    model.position[i, 0].setub(i-2)

    model.position[i, 1].setlb(-(i-3))
    model.position[i, 1].setub(i-1)

    model.position[i, 2].setlb(-(i-3))
    model.position[i, 2].setub(i-3)

# Define your expression
# Constraints for r, r_hat, r2, r3 variables
trupla = []
for i in range(1,n-2):
    for j in range(i+2, n):
        trupla.append((i,j))

def r_constraints_rule(model, i, j):
    return model.r[i, j] == sum((model.position[j, d]
                                - model.position[i, d]) ** 2 for d in [0,1,2])
model.r_constraints = Constraint(trupla, rule=r_constraints_rule)
def r_hat_constraint(model, i, j):
    return model.r[i, j] * model.r_hat[i, j] == 1
model.r_hat_con = Constraint(trupla, rule=r_hat_constraint)
def r2_constraint(model, i, j):
    return model.r_hat[i, j] * model.r_hat[i, j] == model.r2[i, j]
model.r2_con = Constraint(trupla, rule=r2_constraint)
def r3_constraint(model, i, j):
    return model.r2[i, j] * model.r_hat[i, j] == model.r3[i, j]
model.r3_con = Constraint(trupla, rule=r3_constraint)

def ccc(i,j):

```

```

    aai, aaj = sequencia[i], sequencia[j]
    if aai == "A" and aaj == "A":
        return 1
    elif aai == "B" and aaj == "B":
        return 0.5
    else:
        return -0.5

model.v1_con = Constraint(expr=model.v1 >= (1/4) * sum((1 - model.cos_x[i])
    for i in range(2, n-1)))
model.v2_con = Constraint(expr=model.v2 >= 4*sum(model.r3[i, j] *
    (model.r3[i, j]
    - ccc(i, j)) for (i, j)
    in trupla))

model.objective = Objective(expr=model.v1 + model.v2, sense=minimize)
solver = SolverFactory('knitro',
    executable='/home/knitro/knitro-13.2.0-Linux-64/knitroampl/knitroampl')
results = solver.solve(model, tee=True)# will display solver output

print("Objective value:", model.objective())
# tamanho 1 a N
xizes = [(np.arctan2(model.sin_x[i](), model.cos_x[i]())) for i in range(1, N-2)]
# tamanho 1 a N
bezoes = [(np.arctan2(model.sin_b[i](), model.cos_b[i]())) for i in range(1, N-3)]

resultado_angles = xizes + bezoes

print(resultado_angles)

proteinName = pdb_id.lower()
length = str(len(proteina))
chainLetter = 'A'

pdbl.retrieve_pdb_file(proteinName, pdir='.', file_format='mmCIF')

middlePDBFile = []
[coordinatesPDB, headerPDBFile, tailPDBFile] = readCIFFile(proteinName,
    chainLetter,
    middlePDBFile)

angles3DAB = [(float(x)) for x in resultado_angles]

# join sequence
chain = ''.join(proteina)
#print(chain)
obj3DAB = calculate3DAB(angles3DAB, chain)

```

```
#print(obj3DAB)

lengthsBetweenCAPDB = lengthsFromCoordinates(coordinatesPDB)
newCoordinates = calculateNewCoordinates(angles3DAB, lengthsBetweenCAPDB)

print(middlePDBFile)
buildNewPDBFile(newCoordinates, headerPDBFile, tailPDBFile, middlePDBFile,
                proteinName)

rmsd = getRMSD(proteinName)

objMCA = str(model.objective())
obj3DAB = '{:.3 f}'.format(obj3DAB)
rmsd = '{:.3 f}'.format(rmsd)

print('Name:      ' + proteinName)
print('Length:    ' + length)
print('Energy MCA: ' + objMCA)
print('Energy TCC: ' + obj3DAB)
print('RMSD:      ' + rmsd)
```