Centro de Investigación y de Estudios Avanzados
del Instituto Politécnico Nacional

Laboratorio de Tecnologías de Información

Utilización de esquemas de evaluación alternativos en metaheurísticas para lidiar con los principales retos asociados con la predicción de la estructura de proteínas basada en el modelo HP

Tesis que presenta:

Mario Garza Fabre

Para obtener el grado de:

**Doctor en Ciencias
en Computación**

Dr. Eduardo A. Rodríguez Tello, Co-Director
Dr. Gregorio Toscano Pulido, Co-Director

Cd. Victoria, Tamaulipas, México. Mayo, 2014

Center for Research and Advanced Studies
from the National Polytechnic Institute

Information Technology Laboratory

On the use of alternative evaluation schemes in metaheuristics to handle the main search difficulties involved with the prediction of protein structures under the HP model

Thesis by:

Mario Garza Fabre

as the fulfillment of the
requirement for the degree of:

**Ph.D.
in Computer Science**

Dr. Eduardo A. Rodríguez Tello, Co-Director
Dr. Gregorio Toscano Pulido, Co-Director

Cd. Victoria, Tamaulipas, México. May, 2014

The thesis of Mario Garza Fabre is approved by:

Dr. Abel García Nájera

Dr. Iván López Arévalo

Dr. Ricardo Landa Becerra

Dr. José Gabriel Ramírez Torres

Dr. Eduardo A. Rodríguez Tello, Committte Co-chair

Dr. Gregorio Toscano Pulido, Committte Co-chair

Cd. Victoria, Tamaulipas, México., May 13 2014

To my friend, who calls himself the pingüinonito, Mario Alberto.

Acknowledgements

First of all, I want to thank God for having given me the opportunity to live this dream. *"In all thy ways acknowledge him, and he shall direct thy paths" (Proverbs 3:6)*.

I am particularly grateful to my wife Gisela and my son Mario Alberto, for their patience, their understanding, and because they have represented a constant source of motivation for me to give my best throughout the course of my graduate studies. I thank my parents, Mario and Malú, who taught me that goals can be achieved through effort, and to whom I owe largely the person I am today.

I would like to thank Dr. Gregorio Toscano Pulido and Dr. Eduardo A. Rodriguez Tello, my thesis supervisors, for their wise guidance, their patience, and especially for the confidence they have reposed in me and the freedom they gave me during the development of this research project.

I would like to thank Dr. Abel García Nájera, Dr. José Gabriel Ramírez Torres, Dr. Ricardo Landa Becerra and Dr. Iván López Arévalo, for participating as reviewers of this thesis. Their comments have certainly contributed to improve the quality of this research. I would also like to acknowledge Dr. José Santos Reyes for serving as a member of the evaluation committee during my pre-doctoral examination. His valuable observations were very beneficial to the completion of this project.

I especially thank my colleagues and friends at CINVESTAV-Tamaulipas, who lived this great and unforgettable adventure with me. I thank the professors at this research center, for their valuable advice and for sharing their knowledge and experience with me. I acknowledge all the staff at CINVESTAV-Tamaulipas, for all the support and facilities provided during my stay in this institution.

I acknowledge support from CONACyT through a scholarship to pursue graduate studies at CINVESTAV-Tamaulipas. I thank also financial support from CONACyT projects 105060 and 99276.

Contents

List of Figures

List of Tables

List of Algorithms

Publications

Garza-Fabre, M., Rodriguez-Tello, E., and Toscano-Pulido, G. Comparing Alternative Energy Functions for the HP Model of Protein Structure Prediction. In *IEEE Congress on Evolutionary Computation*. New Orleans, LA, USA, June 2011, pp. 2307–2314.

Garza-Fabre, M., Toscano-Pulido, G., and Rodriguez-Tello, E. Comparative Study of Alternative Energy Functions for the HP Model of Protein Structure Prediction. In *International Conference on Bioinformatics & Computational Biology*, vol. 2. CSREA Press, Las Vegas, NV, USA, July 2011, pp. 618–624.

Garza-Fabre, M., Rodriguez-Tello, E., and Toscano-Pulido, G. Multiobjectivizing the HP Model for Protein Structure Prediction. In *Evolutionary Computation in Combinatorial Optimization*, vol. 7245 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Málaga, Spain, April 2012, pp. 182–193.

Garza-Fabre, M., Toscano-Pulido, G., and Rodriguez-Tello, E. Locality-based Multiobjectivization for the HP Model of Protein Structure Prediction. In *Genetic and Evolutionary Computation Conference*. ACM, Philadelphia, PA, USA, July 2012, pp. 473–480.

Garza-Fabre, M., Rodriguez-Tello, E., and Toscano-Pulido, G. An Improved Multiobjectivization Strategy for HP Model-Based Protein Structure Prediction. In *Parallel Problem Solving from Nature*, vol. 7492 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Taormina, Italy, September 2012, pp. 82–92.

Garza-Fabre, M., Rodriguez-Tello, E., and Toscano-Pulido, G. Comparative Analysis of Different Evaluation Functions for Protein Structure Prediction under the HP Model. *Journal of Computer Science and Technology 28*, 5 (2013), 868–889.

Garza-Fabre, M., and Toscano-Pulidoand E. Rodriguez-Tello, G. Handling Constraints in the HP Model for Protein Structure Prediction by Multiobjective Optimization. In *IEEE Congress on Evolutionary Computation*. Cancún, México, June 2013, pp. 2728–2735.

Garza-Fabre, M., Toscano-Pulido, G., and Rodriguez-Tello, E. Multi-objectivization, Fitness Landscape Transformation and Search Performance: A Case of Study on the HP model for Protein Structure Prediction. *European Journal of Operational Research* **(under review)**.

Garza-Fabre, M., Rodriguez-Tello, E., and Toscano-Pulido, G. Constraint-Handling Through Multi-Objective Optimization: the Hydrophobic-Polar Model for Protein Structure Prediction. *Computers & Operations Research* **(under review)**.

Resumen

Utilización de esquemas de evaluación alternativos en metaheurísticas para lidiar con los principales retos asociados con la predicción de la estructura de proteínas basada en el modelo HP

por

Mario Garza Fabre
Doctor en Ciencias del Laboratorio de Tecnologías de Información
Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, 2014
Dr. Gregorio Toscano Pulido, Co-Director
Dr. Eduardo A. Rodríguez Tello, Co-Director

La bioinformática es el campo de la ciencia que comprende el uso de técnicas computacionales para abordar problemas de índole biológica. Una de las áreas de investigación más significativas en bioinformática involucra la determinación de la estructura tridimensional de las proteínas. Conocer la estructura proteica es un paso fundamental para el entendimiento de estos elementos esenciales de la vida. Sin embargo, dadas las limitaciones de lo métodos experimentales existentes, los enfoques computacionales se han convertido en una pieza clave para el análisis de tales estructuras. Este proyecto de tesis se enfoca en el modelo hidrofóbico-polar (HP), una versión abstracta pero aún desafiante del problema de predicción de la estructura de proteínas (PSP). Este modelo se basa en el hecho de que las interacciones hidrofóbicas entre aminoácidos juegan un papel determinante durante el proceso de plegamiento de las proteínas. Desde el punto de vista computacional, PSP utilizando el modelo HP representa un problema difícil de optimización combinatoria, un problema que se ha demostrado pertenece a la clase NP-completo. La neutralidad, multimodalidad e infactibilidad son tres propiedades que caracterizan los paisajes de aptitud (*fitness landscapes*) del modelo HP, siendo éstas las principales fuentes de dificultad que deben ser consideradas durante el diseño de metaheurísticas para resolver este problema. El trabajo de investigación reportado en este documento involucra el análisis de esquemas de evaluación alternativos para lidiar con estas dificultades inherentes

del planteamiento convencional del modelo HP. La premisa básica de este trabajo es que, mediante el cambio del esquema de evaluación, será posible influir en el orden de preferencia que existe entre las soluciones candidatas, impactando así en las características del paisaje de aptitud.

La primera etapa de este proyecto se centró en el tema de la neutralidad. La neutralidad se origina por la baja capacidad de discriminación asociada con el esquema de evaluación convencional (función objetivo) del modelo HP. Esto genera grandes mesetas (plateaus) en el paisaje de aptitud, donde un algoritmo puede fallar al identificar una dirección apropiada de búsqueda. Se realizó un estudio comparativo detallado para evaluar el potencial de discriminación, la compatibilidad con el problema, y la efectividad para guiar la búsqueda, de un conjunto de esquemas de evaluación alternativos que se han propuesto en la literatura. El uso de alternativas de discriminación de grano fino, compatibles con el problema, mejoró sustancialmente el desempeño promedio de los algoritmos considerados. Se encontró también que es posible tomar ventaja y explotar la neutralidad inherente del modelo HP.

La segunda parte de esta tesis reporta la primera aplicación de técnicas de optimización multiobjetivo para resolver PSP específicamente utilizando el modelo HP. Planteamientos multiobjetivo se proponen para lidiar con la multimodalidad del problema, es decir, con la existencia de múltiples óptimos locales en el paisaje de aptitud. Los efectos de la transformación del problema fueron investigados a fondo, así como su impacto en el comportamiento de los algoritmos de búsqueda. Como resultado, se encontró que esta transformación introduce incomparabilidad entre soluciones, lo que conlleva un incremento en la neutralidad del paisaje de aptitud. La neutralidad añadida permite que un algoritmo pueda desplazarse hacia clases de aptitud inferiores, evitando de este modo quedar atrapado en óptimos locales. Consecuentemente, la implementación de los enfoques multiobjetivo propuestos ha provocado una mejora significativa en el desempeño de los algoritmos estudiados.

Con base en los hallazgos mencionados, la etapa final de este proyecto exploró el uso de una estrategia basada en optimización multiobjetivo para hacer frente a las grandes áreas de infactibilidad que presentan los paisajes de aptitud del modelo HP. Este problema con restricciones fue replanteado como un problema sin restricciones, al ser éstas tratadas como un criterio de optimización adicional. Un análisis detallado de la transformación del problema reveló que una fracción considerable de la infactibilidad es convertida en neutralidad. Esta neutralidad define nuevas rutas en el paisaje de

aptitud, rutas que no sólo pueden ser más cortas, sino que pueden también ser aprovechadas con la finalidad de escapar de óptimos locales. Mediante la evaluación de diferentes mecanismos, el estudio realizado resalta la importancia de introducir un sesgo adecuado cuando se implementa el enfoque multiobjetivo para manejo de restricciones. Finalmente, la efectividad del enfoque multiobjetivo propuesto se demostró en términos del desempeño de los algoritmos de búsqueda considerados.

Abstract

On the use of alternative evaluation schemes in metaheuristics to handle the main search difficulties involved with the prediction of protein structures under the HP model

by

Mario Garza Fabre
Ph.D. from the Information Technology Laboratory
Center for Research and Advanced Studies from the National Polytechnic Institute, 2014
Dr. Gregorio Toscano Pulido, Co-advisor
Dr. Eduardo A. Rodríguez Tello, Co-advisor

Bioinformatics is the field of science that encompasses the use of computational techniques to address questions of biological significance. One of the most active research areas in bioinformatics is that of protein structure determination. To gain knowledge about protein structure is a fundamental step towards the understanding of such important building blocks of life. Given the limitations of existing experimental methods, however, computational approaches have become the cornerstone of protein structure analysis. This thesis project focuses on the hydrophobic-polar (HP) model, a simplified yet challenging representation of the protein structure prediction (PSP) problem. This model captures the fact that hydrophobic interactions among amino acids constitute a major determinant of the functional conformation of proteins. From the computational point of view, the HP model for the prediction of protein structure gives rise to a challenging combinatorial optimization problem that has been proved to be NP-complete. Neutrality, multimodality and infeasibility are three characterizing properties of the fitness landscapes of PSP under the HP model, representing the main sources of difficulty to be addressed when designing metaheuristic algorithms for solving this problem. The research reported in this thesis is concerned with the analysis of alternative evaluation schemes to cope with these inherent difficulties of the conventional problem formulation. The basic premise is that, by changing the evaluation scheme, it will be possible to influence the comparability relation among candidate solutions in order to impact on the characteristics of the fitness landscape.

The first part of this project dealt with the neutrality issue. Neutrality relates to the existence of large plateaus in the fitness landscape due to the poor discrimination provided by the conventional evaluation scheme (original optimization objective) of the HP model. In such plateaus, metaheuristics could fail to identify a promising direction, leading the search process to be driven almost at random. A detailed comparative analysis was conducted to evaluate the discrimination ability, the compatibility with the problem, and the effectiveness to guide the search process for a set of alternative evaluation schemes which have been proposed in the literature. The use of more fine-grained and problem-compatible evaluation schemes enhanced the average performance of search algorithms. It was also found that it is possible to take advantage of the inherent neutrality of the HP model.

The second part of this thesis reports the first application of multi-objective optimization methods to the particular HP model of the PSP problem. Alternative multi-objective formulations of the problem were proposed to cope with multimodality; *i.e.*, the presence of multiple local optima in the fitness landscape. The effects of the problem transformation have been investigated, as well as the impact of the proposed approaches on the search behavior of metaheuristic algorithms. It was found that the problem transformation introduces incomparability among solutions, which translates into landscape neutrality. This added neutrality allows algorithms to move through inferior fitness classes as a means to avoid stagnating at local optima. Consequently, the use of the proposed multi-objective formulations has been reflected as a significant increase in search performance.

In the light of the above findings, the last part of this research explored the use of multi-objective optimization to deal with the large areas of infeasibility that the landscapes of the HP model involve. This constrained problem is transformed into an unconstrained one by defining an additional optimization criterion to account for the problem constraints. An analysis of the problem transformation revealed that an important fraction of infeasibility is converted into neutrality, defining potentially shorter paths to move through the landscape which can also be exploited to escape from local optima. By evaluating different mechanisms, this study highlights the relevance of introducing a proper search bias when handling constraints by multi-objective optimization. The implementation of the proposed strategy significantly improved the performance of the considered search algorithms.

1
Introduction

1.1 Problem statement and motivation

Proteins are fundamental elements of living organisms, performing an astonishing range of biological functions. They are involved, for example, in transport, structural, enzymatic, hormonal, regulatory and defensive processes. A protein is chain-like molecule, represented by a linear sequence defined over a set of 20 different building blocks called amino acids. The three-dimensional structure of a protein is known to be one of the major determinants of its distinctive functional properties. Therefore, studying the structure of proteins becomes critical for the understanding of such important biological macromolecules. Advances in molecular biology and genome projects during the last decades have led to an exponential growth in the number of newly discovered protein sequences. Given the limitations of experimental methods to determine the structure of proteins, however, there is currently a huge gap between the number of identified protein sequences and the number of known protein structures.[1] Thus, computational techniques have been developed in order to bridge such an ever-increasing gap.

[1] There are currently a total of $51,616,950$ protein sequence entries in the UniProtKB/TrEMBL database (http://www.uniprot.org/statistics/TrEMBL, January 22, 2014). In contrast, there exist only $97,591$ experimentally determined structures in the Protein Data Bank (http://www.rcsb.org/pdb, February 8, 2014).

It is generally accepted that the amino-acid sequence encodes all the information related to the three-dimensional structure of a protein. In other words, it is the specific configuration of amino acids in a protein which determines how it folds into a unique and compact three-dimensional conformation, often referred to as the *native state*. Among all the possible conformations that a protein can adopt, it is believed that its native state corresponds to the one minimizing the overall free-energy [1, 2]. Hence, the process of inferring the functional, energy-minimizing conformation for a protein molecule from its linear sequence of amino acids can be posed as an optimization problem. This problem is referred to as the *protein structure prediction* (PSP) problem in the specialized literature, and represents one of the most active and challenging research areas in the field of bioinformatics.

The difficulty of the PSP problem stems not only from the fact that proteins are very flexible, which causes that the space of potential conformations is extremely large; but also, the lack of a complete understanding of protein folding has prevented the development of models that can accurately capture all characteristics and the forces that this complex and elusive natural "optimization" process involves. Furthermore, the analysis of a protein conformation at atomic resolution, using the more detailed and realistic models, is a computationally-intensive task which can be prohibitive even for relatively small proteins. Therefore, simplified protein models have been proposed as an attempt to alleviate, to a certain extent, the computational intractability of the PSP problem, while still providing valuable insight to advance the understanding of the most general and essential principles governing the folding process [28, 38, 86, 120, 180]. One of such reduced representations of the PSP problem is the so-called *hydrophobic-polar* (HP) *model*, the focus of this research project [61, 126].

In the HP model, protein chains consist of only two types of amino acids, *hydrophobic* (H) and *polar* (P). This model abstracts the so-called *hydrophobic effect*: whereas H amino acids tend to clump together at the core of the protein to hide from the aqueous environment, P amino acids are usually found at the outer surface of the molecule. Therefore, the hydrophobicity of amino acids constitutes a major stabilizing force determining the native conformation of proteins, a fact that is well captured by the HP model. Also, this model discretizes the conformational space, so that a valid protein structure is modeled as a self-avoiding embedding of the protein chain on a given lattice. Under the HP model, thus, PSP is defined as the problem of finding a self-avoiding conformation of the protein where the interaction among H amino acids on the lattice is maximized (assuming that maximizing H-H interactions will result in the formation of a compact hydrophobic core). In spite of the significant simplifications and assumptions made in the HP model, this abstract formulation inherits, to a large extent, the challenging nature of the original problem. From the computational point of view, the prediction of protein structures under the HP model represents a hard combinatorial optimization problem that has been shown to be NP-complete [10, 43]. Such a complexity has motivated the use of a diversity of metaheuristic approaches to address this problem (refer to Section 2.3.3.4 for a review on the application of metaheuristic algorithms to this problem).

Three main sources of difficulty can be identified with regard to the design of metaheuristic algorithms for solving the HP model of the PSP problem: the *neutrality*, *multimodality* and *infeasibility* of its fitness landscapes. Neutrality originates from the poor discrimination associated with the conventional evaluation scheme (original optimization objective) of the HP model. A low discrimination translates into large plateaus of neutral (*i.e.*, incomparable) solutions, on which metaheuristic algorithms may drift due to the lack of a search direction to guide the optimization process. Multimodality, in turn, refers to the existence of multiple local optima in the fitness landscape. Hence, metaheuristic algorithms, mainly local search-based methods, can easily stagnate at a suboptimal solution. Finally, the infeasibility difficulty relates to the fact that, using the existing problem representations, a significant portion of the solution space encodes non-self-avoiding protein structures. A search algorithm could invest a considerable amount of computational effort in exploring the large areas of infeasibility of the fitness landscape. Otherwise, handling infeasible areas as inaccessible may result in a complicated landscape from the perspective of the search algorithm. Thus, metaheuristic algorithms need to be equipped with explicit mechanisms to cope effectively with these issues.

1.2 The proposal

Metaheuristic algorithms can be broadly classified as single-solution-based and population-based metaheuristics [218]. In either case, and regardless of the search mechanisms implemented, the success of these algorithms relies largely on a proper solution representation, as well as on an effective

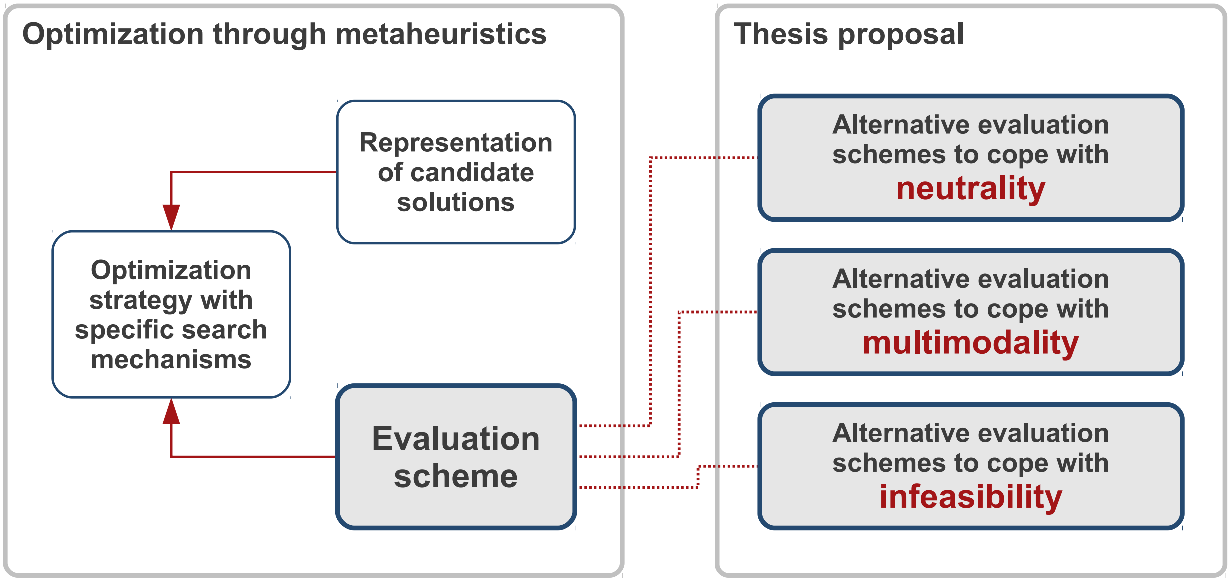

Optimization through metaheuristics
Representation of candidate solutions
Optimization strategy with specific search mechanisms
Evaluation scheme
Thesis proposal
Alternative evaluation schemes to cope with neutrality
Alternative evaluation schemes to cope with multimodality
Alternative evaluation schemes to cope with infeasibility

Figure 1.1: Illustrating the relationship between the fundamental components on which metaheuristic-based optimization relies (left), and the way this thesis deals with the main difficulties of the studied optimization problem by changing one of such components, namely, the evaluation scheme (right).

evaluation scheme (see left part of Figure 1.1). The evaluation scheme plays a critical role in defining the characteristics of the fitness landscape, being the responsible for assessing the quality (fitness) of the candidate solutions and determining how they compare with respect to each other. In this way, the evaluation scheme is a problem-specific design component that provides metaheuristic algorithms with this valuable information in order to guide the search process. By changing the evaluation scheme, therefore, it will be possible to transform the fitness landscape of the problem and impact on the behavior of search algorithms. As illustrated in Figure 1.1, this research project is concerned with the analysis of how the use of alternative evaluation schemes can contribute to dealing with the neutrality, multimodality and infeasibility of the HP model's fitness landscapes and, thus, to the development of more efficient metaheuristics for solving this problem.

Previous efforts have been reported on the use of alternative evaluation schemes to deal with the neutrality of the HP model [9,26,45,98,122,142]. Nevertheless, the literature lacks solid experimental evidence supporting the effectiveness of most of these alternative problem formulations, and there is not a complete understanding of their influence on the performance of metaheuristics. Part of this thesis project engages in a detailed comparative analysis as an attempt to answer these questions.

Regarding multimodality, this research reports for the first time (to the author's knowledge) the use of alternative evaluation schemes to address such a difficulty of the HP model's fitness landscapes. The multi-objectivization of the HP model is proposed, where alternative multi-objective formulations of the problem are used to cope with this issue. This research work produces also the first efforts on the application of multi-objective optimization techniques to this particular problem.

Finally, the infeasibility issue has been previously tackled from the perspective of the evaluation scheme by implementing a penalty strategy. Despite its simplicity, an inherent drawback of such an approach lies in the need for defining the severity of the penalties to be applied, which has been regarded to be a difficult optimization problem itself [161, 186]. Alternatively, this research explores for the first time the use of multi-objective optimization to handle the constraints of the HP model.

1.3 Research hypothesis

The underlying hypothesis on which this research work is based is as follows:

> By implementing alternative evaluation schemes, it will be possible to alter essential characteristics of the fitness landscape in order to address the neutrality, multimodality and infeasibility challenges which arise when designing metaheuristic algorithms for solving the protein structure prediction problem under the HP model.

1.4 Aim of the thesis

The main objective of this research project can be stated as follows:

> To contribute to the fields of computer science and bioinformatics, particularly in the areas of optimization, metaheuristic algorithms and protein structure prediction, through the study of alternative evaluation schemes, tailored to the HP model of the protein structure prediction problem, and the analysis of how they contribute to overcoming the neutrality, multimodality and infeasibility difficulties associated with this problem.

1.5 Outline of the thesis

Besides the present introductory chapter, this thesis has been organized into five other chapters. A general overview of the remaining chapters is provided below:

- Chapter 2 provides background concepts and sets the notation used in this document. Also, previous work on the application of metaheuristic algorithms to the HP model is summarized. In addition, this chapter describes the adopted performance assessment methodology, detailing the considered test instances, the implemented performance measures, the methodology followed during the statistical significance analyses, and the utilized experimental platform.

- Chapter 3 addresses the neutrality of the HP model's fitness landscapes. This chapter studies different alternative energy (evaluation) functions which have been reported in the specialized literature to cope with this issue. The considered approaches are described in detail and an in-depth comparative analysis explores their discrimination capabilities, their consistency with respect to the original problem's definition, and their effectiveness to guide the search process.

- In Chapter 4, the multi-objectivization of the HP model is proposed as a means of dealing with multimodality. Three different multi-objective evaluation schemes for the HP model are investigated, all of them based on the decomposition of the original optimization criterion of the problem. A thorough analysis of the potential effects of the (single-objective to multi-objective) problem transformation is conducted. Also, it is evaluated the extent to which multi-objectivization impacts on the search performance of metaheuristic algorithms.

- Chapter 5 proposes the use of multi-objective optimization as a constraint-handling strategy for the HP model. This originally constrained single-objective problem is restated as an unconstrained multi-objective problem by treating constraints as a supplementary optimization criterion. The effects of this problem transformation are carefully analyzed. Different mechanisms to provide this strategy with a proper search bias are investigated. Finally, the suitability of the proposed approach is evaluated in terms of the performance of search algorithms.

- Finally, Chapter 6 provides concluding remarks, summarizes the main achieved findings and contributions, and highlights some possible directions for future research which can be derived from the findings of the present work.

It is worthy to mention at this point that the provided discussion of the related work has been split and it is covered throughout the different chapters of this document in order to make each performed study more self-contained (see Sections 2.3.3.4, 3.2, 4.2 and 5.2).

2

Background concepts and performance assessment methodology

2.1 Introduction

The purpose of this chapter is to provide a series of basic concepts and definitions in order to familiarize the reader with the main topics related to this research, as well as with the notation and performance assessment methodology adopted in the remaining chapters of this document. The particular case of study of this project, the HP model for the prediction of protein structures, represents a challenging optimization problem. Therefore, this chapter begins by covering in Section 2.2 fundamental definitions from the field of optimization. Then, Section 2.3 offers a general overview on proteins and the protein structure prediction problem, with special emphasis on the HP model of this problem addressed in this thesis. Finally, Section 2.4 concludes this chapter by describing in detail the HP model test instances, the performance measures, the statistical significance testing methodology, and the experimental platform utilized during the experiments of this research work.

2.2 Optimization

This section introduces some background concepts related to optimization which are essential for the sake of self-containedness of this thesis document. The remainder of this section is organized as follows. Sections 2.2.1 and 2.2.2 provide a formal definition of the single-objective and multi-objective optimization problems, respectively. Multi-objectivization, the transformation from a single-objective to a multi-objective optimization problem, is discussed in Section 2.2.3. Finally, basic concepts with regard to fitness landscapes and fitness landscape analysis are presented in Section 2.2.4.

2.2.1 Single-objective optimization

Without loss of generality, a *single-objective optimization problem* can be formally stated as follows:

$$\begin{aligned} \text{Minimize} \quad & f(\mathbf{x}), \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{X}_\mathcal{F}, \end{aligned} \tag{2.1}$$

where $\mathbf{x}$ is a *solution vector*; $\mathcal{X}_\mathcal{F}$ denotes the *feasible set*, *i.e.*, the set of all *feasible solution vectors* in the search space $\mathcal{X}$, $\mathcal{X}_\mathcal{F} \subsetneq \mathcal{X}$; and $f : \mathcal{X} \rightarrow \mathbb{R}$ is the objective function to be optimized. The aim is thus to find the feasible solution(s) yielding the optimum value for the objective function. That is, the problem is to find $\mathbf{x}^* \in \mathcal{X}_\mathcal{F}$ such that $f(\mathbf{x}^*) = \min\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}_\mathcal{F}\}$.

2.2.2 Multi-objective optimization

Without loss of generality, a *multi-objective optimization problem* can be formally defined as follows:

$$\begin{aligned} \text{Minimize} \quad & \mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_k(\mathbf{x})]^T, \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{X}_\mathcal{F}, \end{aligned} \tag{2.2}$$

where $\mathbf{f}(\mathbf{x})$ is the *objective vector* and $f_i : \mathcal{X} \rightarrow \mathbb{R}$ is the i-th objective function, $i \in \{1, 2, \ldots, k\}$. Rather than searching for a single optimal solution, the task in multi-objective optimization is to identify a set of trade-offs among the, usually conflicting, problem objectives. More formally, the goal is to find a set of *Pareto-optimal solutions* $\mathcal{P}^*$, such that

$$\mathcal{P}^* \quad = \quad \{\mathbf{x}^* \in \mathcal{X}_{\mathcal{F}} \mid \nexists \mathbf{x} \in \mathcal{X}_{\mathcal{F}} : \mathbf{x} \prec \mathbf{x}^*\}. \tag{2.3}$$

The symbol "$\prec$" denotes the *Pareto-dominance* relation [176]:

$$\begin{aligned} \mathbf{x} \prec \mathbf{x}' \quad \Leftrightarrow \quad & \forall i \in \{1, 2, \ldots, k\} : f_i(\mathbf{x}) \leq f_i(\mathbf{x}') \quad \wedge \\ & \exists j \in \{1, 2, \ldots, k\} : f_j(\mathbf{x}) < f_j(\mathbf{x}'). \end{aligned} \tag{2.4}$$

If $\mathbf{x} \prec \mathbf{x}'$, then $\mathbf{x}$ is said to *dominate* $\mathbf{x}'$. Otherwise, $\mathbf{x}'$ is said to be *nondominated* with respect to $\mathbf{x}$, denoted by $\mathbf{x} \not\prec \mathbf{x}'$. The image of $\mathcal{P}^*$ in the objective space is the so-called *Pareto-optimal front*, usually referred to as the *trade-off surface*. Some of these concepts are illustrated in Figure 2.1.

2.2.3 Multi-objectivization

Multi-objectivization[1] refers to the process of reformulating an originally single-objective optimization problem in terms of two or more objective functions [119].[2] Two main directions exist to perform such a single-objective to multi-objective transformation: (i) by incorporating additional information in the form of *supplementary objectives*, also referred to as artificial or helper objectives [18, 105]; or (ii) by means of the *decomposition* of the original objective function of the problem [85, 119]. In either case, the goal remains to solve the original problem, so that the original optima are to be also Pareto-optimal with regard to the new alternative multi-objectivized formulation.

When multi-objectivization is based on the addition of supplementary objectives, the single-objective problem is restated as a multi-objective problem of the form $\mathbf{f}(\mathbf{x}) = [f(\mathbf{x}), g_1(\mathbf{x}), \ldots, g_h(\mathbf{x})]^T$;

[1] The term *multi-objectivization* was originally coined by Knowles *et al.* [119], but the first studies on this kind of problem transformation date back to the work of Louis and Rawlins [145].

[2] It should be noted that originally multi-objective problems have been also multi-objectivized in the literature [96].

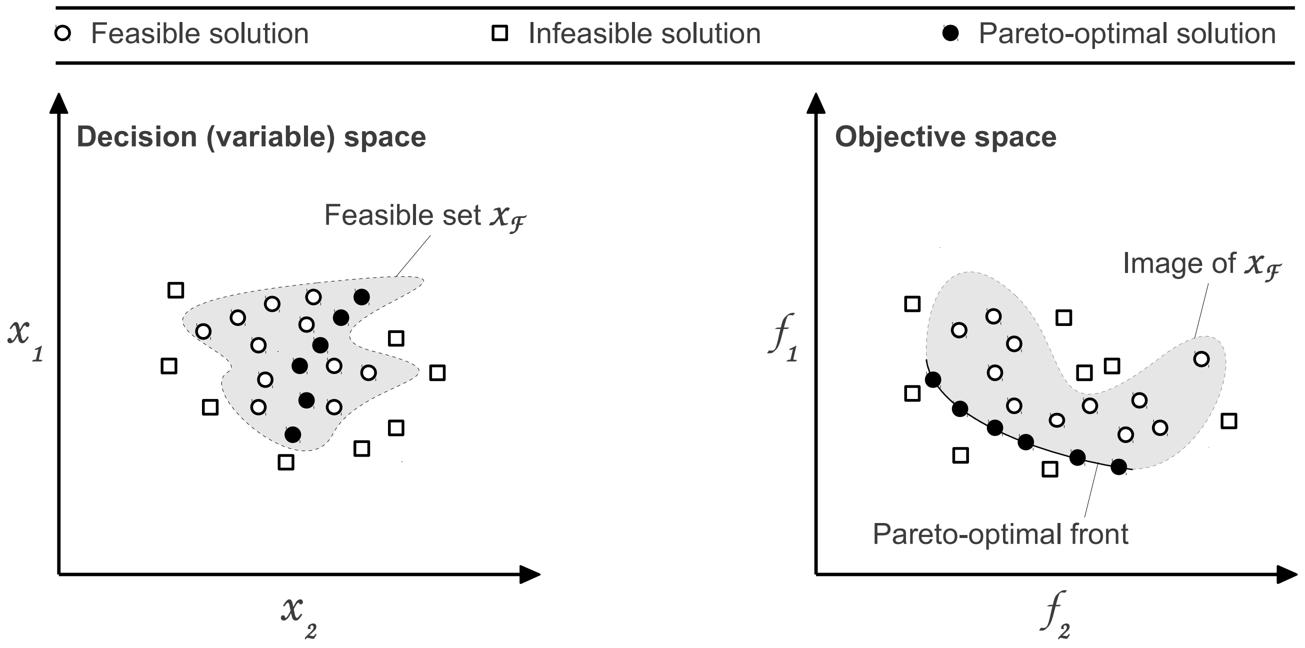

○ Feasible solution
□ Infeasible solution
● Pareto-optimal solution
Decision (variable) space
Feasible set $\mathcal{X}_{\mathcal{F}}$
x_1
x_2
Objective space
Image of $\mathcal{X}_{\mathcal{F}}$
f_1
f_2
Pareto-optimal front

Figure 2.1: Illustration of basic concepts of multi-objective optimization.

where f is the original objective function of the problem and g_i denotes the i-th supplementary objective, $1 \leq i \leq h$. In the literature, this has been the most extensively studied approach to multi-objectivization. In a recent review [202], Segura *et al.* make a distinction between multi-objectivization proposals where supplementary objectives are problem-dependent and are based solely on information from the evaluated solution [18,82,103,105,137,138,222], and those where they act as diversity measures [22,23,166,200,203–205,208,221,237]. The authors give also separate treatment to studies where additional objectives are implemented to handle constraints [161,195,234].

Finally, in multi-objectivization by decomposition, the original objective is fragmented into several different components, each to be treated as an objective function under the new alternative formulation. Formally, the problem is restated in terms of $d \geq 2$ objectives, $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), \ldots, f_d(\mathbf{x})]^T$, such that the sum of all the new objectives equals the original objective function; *i.e.*, $f(\mathbf{x}) = \sum_{i=1}^{d} f_i(\mathbf{x})$, for all $\mathbf{x} \in \mathcal{X}_{\mathcal{F}}$.[3] Some works reported in this direction include [8,50–52,55,84,85,119,174,213,233].

[3] Though other different decompositions are possible, this definition ensures that the original optimum coincides with one of the Pareto-optima in the multi-objective version of the problem [119].

2.2.4 Fitness landscapes

The notion of a fitness landscape, first introduced by Wright [239], has been found to be useful in understanding the most essential characteristics of certain optimization problems, or problem classes. By analyzing the fitness landscape, it is possible to gain further insight into problem difficulty as a means of explaining, or even predicting, the performance of search algorithms. Fitness landscape analysis is expected to provide important clues for guiding the development of more competitive search mechanisms, which are able to deal with (or to take advantage of) the particular characteristics of the landscape associated with the given optimization task. Some fundamental definitions on this topic, which are relevant according to the scope of this research project, are presented below. For a more comprehensive treatment of the topic the reader can be referred to [107,150,182,214,227,232].

A *fitness landscape* can be generically defined in terms of a triplet $(\mathcal{X}, \mathcal{N}, \xi)$. The first element, $\mathcal{X}$, represents the set of all potential solutions to the problem, *i.e.*, the *search space*. The notion of connectedness among solutions in $\mathcal{X}$ is introduced by the so-called *neighborhood structure*, $\mathcal{N} : \mathcal{X} \to 2^{\mathcal{X}}$, a function mapping each possible solution $\mathbf{x} \in \mathcal{X}$ to a set of solutions $\mathcal{N}(\mathbf{x}) \subseteq \mathcal{X}$.[4] Hence, $\mathcal{N}(\mathbf{x})$ is referred to as the neighborhood of $\mathbf{x}$ and each solution $\mathbf{x}' \in \mathcal{N}(\mathbf{x})$ is called a *neighbor* of $\mathbf{x}$. Finally, ξ denotes the evaluation scheme, which consists of i) a measure (or set of measures) to serve as an indicator of the quality or "height" of the different candidate solutions; and ii) a mechanism to impose an ordering relation given the adopted quality measure(s). As the evaluation scheme, in single-objective optimization a *fitness function* (usually directly related to the problem's objective function) is considered, and a simple ordering based on such fitness function sets the preference relation among solutions. A search algorithm can thus be thought of as navigating in order to find the highest peak of the fitness landscape [159], *i.e.*, the solution with the overall best fitness value.[5] In the multi-objective context, however, a number of (conflicting) criteria determine the quality of solutions, so that defining an ordering relation is not as straightforward as in the single-objective case. The partial order induced by the Pareto-dominance relation is assumed in this research.

[4] $2^{\mathcal{X}}$ refers to the power set of $\mathcal{X}$, usually also denoted by $\mathcal{P}(\mathcal{X})$.

[5] While an objective function can be either minimized or maximized, in this research project a fitness function is assumed to be always maximized (the goal is to search for the fittest candidate solution).

2.2.4.1 Neutrality

The fitness landscape of a problem can be studied in terms of different properties, being *neutrality* of particular importance given the purposes of the present study. The standard definition of neutrality, in the single-objective case, refers to the degree to which a landscape contains connected areas of equal fitness [182]. Considering a broader notion to cover also the multi-objective case, neutrality can be understood as the result of the *incomparability* that the adopted evaluation scheme ξ induces. The term incomparability is used in this study to indicate the situation where no preferences can be imposed between a pair of solutions, so that these solutions are considered equivalent when evaluated under ξ. In this way, two different solutions $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ are said to be *neutral* (*i.e.*, incomparable), denoted by $neutral(\mathbf{x}_1, \mathbf{x}_2)$, if either they share the same fitness value (single-objective case), or they are nondominated, in the Pareto sense, with respect to each other (multi-objective case).

Having defined neutrality, a series of related basic concepts can be introduced as follows. The *neutral neighborhood* of a solution $\mathbf{x} \in \mathcal{X}$, $\mathcal{N}_n(\mathbf{x})$, is given by the subset of all its *neutral neighbors*, *i.e.*, $\mathcal{N}_n(\mathbf{x}) = \{\mathbf{x}' \in \mathcal{N}(\mathbf{x}) \mid neutral(\mathbf{x}, \mathbf{x}')\}$. The total number of neutral neighbors of $\mathbf{x}$, *i.e.*, the cardinality of $\mathcal{N}_n(\mathbf{x})$, is known as the *neutrality degree* of $\mathbf{x}$, and the ratio of the neutrality degree to the size of the neighborhood is referred to as the *neutrality ratio*. A *neutral fitness landscape* is characterized by a large number of solutions presenting a high degree of neutrality. This leads to (potentially large) connected areas of incomparable solutions called *plateaus*, more formally referred to as *neutral networks*. This is a common scenario in problems for which, despite involving a huge search space, only a reduced number of different fitness values can be assigned, as is the case of the HP model of the protein structure prediction problem studied herein. Consider the *neutrality graph* $G = (\mathcal{X}, \mathcal{E}_n)$ where $\mathcal{E}_n = \{(\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{X}^2 \mid \mathbf{x}_2 \in \mathcal{N}_n(\mathbf{x}_1)\}$. Each connected component of the graph G corresponds to a different neutral network. In other words, a neutral network is a connected subgraph $G' = (\mathcal{X}', \mathcal{E}'_n)$ of G, $\mathcal{X}' \subseteq \mathcal{X}$ and $\mathcal{E}'_n \subseteq \mathcal{E}_n$, where i) there exists a path connecting any pair of solutions $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}'$, and ii) there exists no edge $(\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{E}_n \setminus \mathcal{E}'_n$ such that $\mathbf{x}_1 \in \mathcal{X}'$ and $\mathbf{x}_2 \in \mathcal{X} \setminus \mathcal{X}'$. The neutral network of a solution $\mathbf{x}$, *i.e.*, the neutral network to which $\mathbf{x}$ belongs, will be denoted in this study as $NN(\mathbf{x})$. Finally, another important concept is that of a *neutral* *walk*. A neutral walk from $\mathbf{x}_1$ to $\mathbf{x}_k$ refers to a sequence of solutions $\langle \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k \rangle$ such that $\mathbf{x}_{i+1} \in \mathcal{N}_n(\mathbf{x}_i)$, $1 \leq i < k$. That is, a neutral walk represents a path on a neutral network.

2.3 Proteins and protein structure prediction

This section introduces fundamental concepts with regard to the biological aspects of this research work, as well as provides a formal definition of the particular problem this thesis focuses on. This section proceeds as follows. Section 2.3.1 discusses the essentials of proteins, their structure and the functional diversity of these important biological macromolecules. Section 2.3.2 covers basic notions about protein structure determination and describes the computational approaches to perform this task. A formal definition of the HP model of the protein structure prediction problem is presented in Section 2.3.3. Section 2.3.3 includes also a review on the diversity of metaheuristic approaches which have been reported in the specialized literature to address this problem.

2.3.1 Proteins

Proteins are at the heart of cellular function, making possible most of the key processes associated with life (the diversity of protein functions are discussed later in Section 2.3.1.2). Amino acids, the building blocks of proteins, are all of them consistent with the general structure presented in Figure 2.2 (left part of the figure). Each amino acid has a central carbon atom (C$_\alpha$) which is covalently

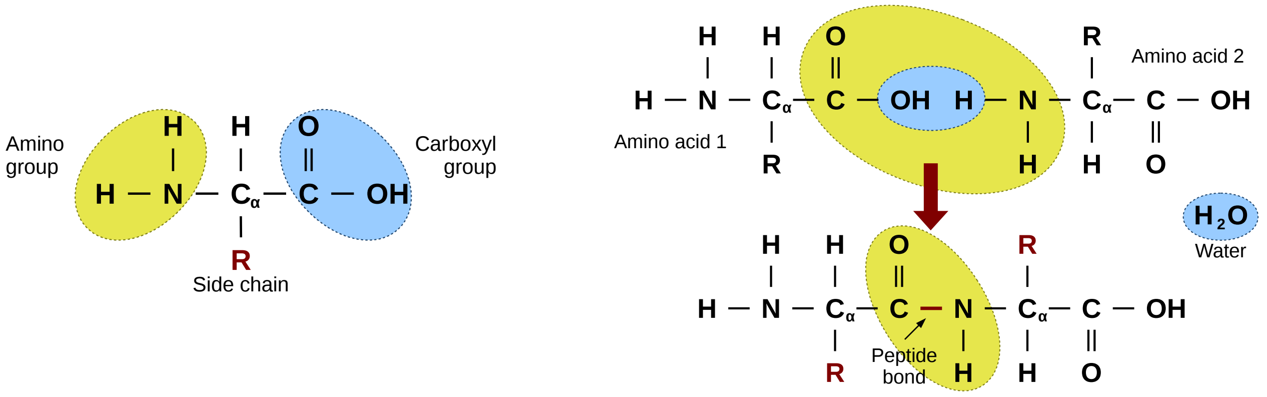


Figure 2.2: General structure of all amino acids (left). The peptide bond formation process (right).

Table 2.1: The 20 amino acids and their standardized representations in one and three letters. It is indicated whether each amino acid can be classified as hydrophobic or polar [4, 115, 129, 209, 212, 216, 238]; some amino acids (C, G, S, T, and Y) are classified differently depending on the authors.

Amino acid	**Three-letter representation**	**One-letter representation**	**Hydrophobic ● / Polar ○**
Alanine	Ala	A	●
Arginine	Arg	R	○
Asparagine	Asn	N	○
Aspartic acid	Asp	D	○
Cysteine	Cys	C	● [4, 216] ○ [115, 129, 209, 212]
Glutamic acid	Glu	E	○
Glutamine	Gln	Q	○
Glycine	Gly	G	● [4, 129, 209, 216] ○ [115, 212]
Histidine	His	H	○
Isoleucine	Ile	I	●
Leucine	Leu	L	●
Lysine	Lys	K	○
Methionine	Met	M	●
Phenylalanine	Phe	F	●
Proline	Pro	P	●
Serine	Ser	S	● [4, 216] ○ [115, 129, 209, 212]
Threonine	Thr	T	● [4, 216] ○ [115, 129, 209, 212]
Tryptophan	Trp	W	●
Tyrosine	Tyr	Y	● [4, 209, 216] ○ [115, 129, 212]
Valine	Val	V	●

bonded to a carboxyl group (COOH), to an amino group (NH_2), to a hydrogen atom (H) and to a radical (R) group or side chain. There are 20 amino acids commonly found in proteins, each of which has a distinctive R group that is responsible for its particular chemical properties. Table 2.1 lists the 20 amino acids and their corresponding standardized one-letter and three-letter representations.

In proteins, amino acids are held together by *peptide bonds*. Hence, protein chains are also referred to as *polypeptides*. The peptide bond is formed when the carboxyl group of an amino acid reacts with the amino group of another, releasing a water molecule. The elements of a polypeptide chain are, therefore, amino acid *residues*. This process is illustrated in the right part of Figure 2.2.

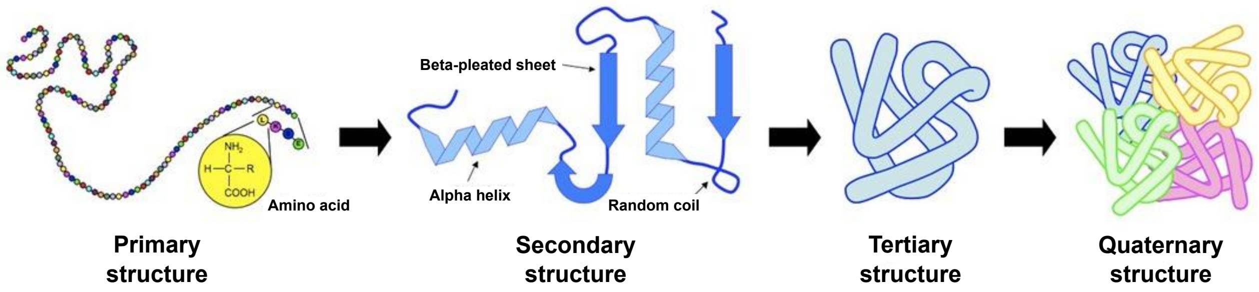


Figure 2.3: Four distinct levels of protein structure complexity (figure taken from http://www.vce.bioninja.com.au/aos-1-molecules-of-life/biomolecules/proteins.html).

2.3.1.1 *The structure of proteins*

Proteins present complex structures commonly described in terms of four hierarchical levels of organization. The linear sequence of amino acid residues in the polypeptide chain constitutes the *primary structure* of a protein. The *secondary structure* describes the arrangement of amino acids within certain areas of a polypeptide chain into motifs such as α helices, β sheets, and coils (also called loops). The *tertiary structure* defines the overall folding of the protein chain in three-dimensional space, where secondary structure elements are packed into globular domains. It is the tertiary, three-dimensional conformation which is essential to the function of the protein molecule. Finally, some proteins are composed of multiple (two ore more) polypeptide chains called subunits.[6] The quaternary structure describes the spatial arrangement and position of each of the subunits in a multiunit protein. The stabilizing forces that hold the polypeptide subunits together are the same forces that are responsible for tertiary structure stabilization. A major force stabilizing the quaternary structure is the hydrophobic interaction among nonpolar side chains at the contact regions of the subunits.

2.3.1.2 *Diversity of proteins and their function*

According to the diversity of their biological functions, proteins can be mainly classified as:

- *Hormonal*. Usually transported through the blood, hormones are messenger proteins that transmit signals from one cell to another to coordinate certain activities. An example of a

[6] A protein molecule that consists of a single polypeptide chain is said to be *monomeric*. Proteins made up of more than one polypeptide chain (as many of the large ones) are called *oligomeric*.

hormonal protein is *insulin*, which regulates glucose metabolism by controlling the blood-sugar concentration. Other examples of hormonal proteins include *oxytocin* and *somatotropin*.

- *Enzymatic*. Enzymes are responsible for catalyzing the thousands of chemical reactions of the living cell. Examples include *lactase*, which breaks down the sugar lactose found in milk, and *pepsin*, a digestive enzyme that works in the stomach to break down proteins in food.

- *Structural*. Also known as fibrous or support proteins, structural proteins are necessary components of the body. Examples include *keratin*, *collagen*, and *elastin*. Keratin is the main structural component in hair, nails, teeth and skin. Collagen and elastin provide support for connective tissues such as tendons and ligaments.

- *Defensive*. Antibodies, or *immunoglobulins*, are specialized proteins produced by the immune system. They defend the body from antigens, such as bacteria, viruses and other harmful microorganisms, rendering them inactive.

- *Storage*. Storage proteins mainly store mineral ions such as iron and potassium. *Ovalbumin* and *casein* are storage proteins found in breast milk and egg whites, respectively, that play an important role in embryonic development.

- *Transport*. Transport proteins carry vital materials (molecules) to the cells. *Hemoglobin*, for example, transports oxygen through the blood. *Myoglobin*, in turn, absorbs oxygen from hemoglobin and then releases it to the muscles. Other examples of transport proteins are *calbindin* and *cytochromes*.

- *Receptor*. Receptor proteins are located at outer part of the cells. They regulate substances, nutrients and signals that enter and leave the cells. Some receptors, for example, activate enzymes, while others stimulate endocrine glands to secrete epinephrine and insulin to regulate blood sugar levels. An example is the *acetylcholine* receptor.

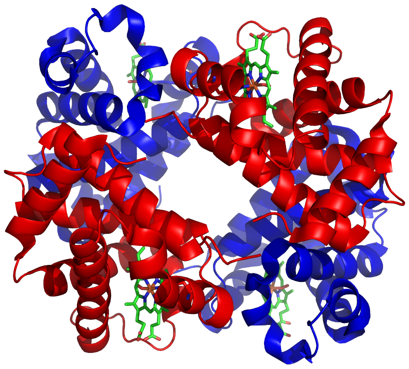


Figure 2.4: The tertiary, three-dimensional structure of hemoglobin, a globular protein (this figure was taken from http://en.wikipedia.org/wiki/File:1GZX_Haemoglobin.png).

- *Contractile*. Also known as motor proteins, contractile proteins perform mechanical work and are responsible for movement. They regulate the strength and speed of heart and muscle contractions. Contractile proteins include *actin* and *myosin*.

In addition, proteins are commonly classified into three main groups based on their molecular shape and solubility [215]:

- *Globular (spheroproteins).* A globular protein consists of peptide chains that fold generally into spherical ("globe-like") shapes. The folding of globular proteins is such that most amino acids with hydrophobic side chains are on the inside (at the core) of the protein, whereas most hydrophilic side chains are on the surface of the molecule. This particular characteristic makes globular proteins soluble in aqueous solutions, which allows them to travel through blood or other body fluids to sites where their function is needed. Globular proteins comprise the most varied type of proteins, being involved in functions such as catalysis, transport and regulation (these functions and protein examples have been described above). Figure 2.4 shows the three-dimensional structure of hemoglobin, a protein which falls into the globular category.

- *Structural (scleroproteins)*. Structural or fibrous proteins, as described previously in this section, perform structural functions that provide support and external protection (e.g., keratin, collagen, and elastin). They form muscle fiber, tendons, connective tissue and bone . These proteins present simple, regular, and linear structures (with an elongated and narrow shape) that tend to aggregate together to form macromolecular structures. Structural proteins tend not to denature as easily as globular proteins and, generally, are inert and water-insoluble.

- *Membrane*. A membrane protein is a protein that is found associated with a membrane system of a cell. Membrane protein structure is somewhat opposite to that of globular proteins, with most of the hydrophobic amino acid side chains oriented outwards. Thus, such proteins tend to be water-insoluble and they usually have fewer hydrophobic amino acids than the globular ones. These proteins are much less well-understood, because they are difficult to purify and study. Membrane proteins play several roles including relaying signals within cells, allowing cells to interact, and transporting molecules. Rhodopsin is an example of a membrane protein.

2.3.2 Protein structure determination

In the 1950s, Christian Anfinsen studied the properties of *ribonuclease A* (or RNase A, a 24-residue protein) and observed that its polypeptide chain could fold spontaneously into a unique three-dimensional structure, its *native conformation* [2]. Likewise, Anfinsen's experiments showed that after denaturation, *i.e.*, the structural change of the molecule caused by extreme conditions (*e.g.*, temperature or pH changes), the protein chain was able to refold back to its native conformation upon return to normal conditions. From these findings, Anfinsen proposed his theory of protein folding, stating that the native conformation of a protein is determined by the totality of interactions that occur at the atomic level and, hence, by the chemical properties of its specific amino acid sequence, in a given environment. Among all possible conformations that a protein can adopt, it is believed that its native state corresponds to the one minimizing the overall free-energy; *i.e.*, the thermodynamically most stable state of the molecule. This is the so-called *thermodynamic hypothesis* [1].

In the 1960s, Cyrus Levinthal introduced the argument that there are far too many possible conformations for a linear sequence of amino acids that finding the most stable thermodynamic structure, through pure random sampling of the energy landscape, would require a period of time far greater than the age of the universe [179]. In nature, however, proteins adopt their native conformation quickly (on the order of seconds or less), which suggests that very specific pathways must be followed during the folding process. This is known as the *Levinthal's paradox* [130, 131]. As it was aptly pointed out by Lesk [129], the observation that each protein folds spontaneously into a unique three-dimensional native conformation implies that nature has a well-defined algorithm for predicting the structure of proteins from their amino acid sequences.

In structural biology, there exist experimental methods to determine the three-dimensional structure of proteins. The most representative approaches are X-ray Crystallography (CRX) and Nuclear Magnetic Resonance Spectroscopy (NMR). In CRX, the protein is purified and crystallized, then subjected to an intense beam of X-rays. This produces distinctive characteristic patterns of spots, which are then analyzed to determine the distribution of electrons in the protein. The resulting map of the electron density is then interpreted to determine the location of each atom. In NMR, the purified protein is placed in a strong magnetic field and then probed with radio waves. This produces information that is used to build a model of the protein defining the location of each atom. Out of the $97,591$ experimentally determined structures in the Protein Data Bank,[7] $86,321$ were obtained using CRX, $10,231$ using NMR, and the remaining structures were determined using some other alternative techniques (*e.g.*, electron microscopy). Experimental methods to determine protein structures can be expensive, time consuming and their applicability is usually restricted to small proteins or to those with specific properties. For example, the process of crystallization in CRX is difficult and can impose limitations on the types of proteins that may be studied by this method. Similarly, the NMR technique is currently limited to small protein molecules because of the computational costs it involves. Therefore, computational approaches have become valuable tools for studying the structure of proteins and, hence, play a major role in advancing the understanding of such essential building blocks of life. The computational approaches for protein structure determination can be broadly divided into two main categories: (i) template-based methods, and (ii) template-free methods. These approaches are respectively discussed in Sections 2.3.2.1 and 2.3.2.2.

[7] From http://www.rcsb.org/pdb. February 8, 2014.

2.3.2.1 Computational template-based methods

Template-based methods aim to determine the structure of new proteins based on existing information of other experimentally determined protein structures. Known structures serve as templates to find the structure of new proteins with common characteristics. Template-based methods are often referred to as comparative modeling or knowledge-based methods. This category includes the homology modeling and fold recognition approaches, which are briefly described below [129]:

- Homology modeling. These methods construct an atomic model based on known structures of proteins related at the polypeptide chain level. The basic premise of these methods is that proteins with similar sequences usually have similar structures. Thus, a known structure can be assigned to a new polypeptide chain with a high degree of confidence if the two proteins present a high degree of correspondence (*i.e.*, *identity*) in their amino acids sequences.

- Fold recognition. Given a library of known (experimentally determined) structures, fold recognition methods try to detect which of them shares a folding pattern that could represent the most reasonable model for the new protein chain. The main basis of these methods is that, in some cases, proteins can share similar structures even in the absence of sequence matching.

While the use of previous knowledge is the major advantage of template-based methods, it is also their major drawback [112]. The complete reliance on a database of known structures means that it is possible to accurately model structures similar to those already known, but it is impossible to discover entirely new protein folds using these strategies. Only those proteins for which there exist appropriate templates can be modeled. Moreover, protein models obtained using these methods usually need to be subjected to further refinement by applying template-free approaches [125].

2.3.2.2 Computational template-free methods

Template-free methods are also commonly referred to as *ab initio* or *de novo* strategies. In contrast to template-based methods, template-free approaches attempt to determine the native structure of a protein using only the information from its amino acid sequence (knowledge from other already known protein structures is not taken into consideration). Given a particular configuration of the amino acid sequence, these approaches rely on the understanding of the basic principles that control protein folding in order to infer the native, energy-minimizing conformation of the protein molecule.

On the one hand, template-free methods constitute a more general approach to protein structure determination. Given that these methods do not depend on the knowledge and availability of related protein structures, the application range of these methods is wider (in essence, any protein could equally be modeled). Unlike the template-based modeling, ab initio modeling could help answer the basic questions on how and why a protein adopts the specific structure out of many possibilities [127]. Also, these methods are commonly used in a final refinement stage for models obtained by template-based techniques, or those experimentally determined at low resolution [125]. On the other hand, template-free methods are known to represent the less accurate approach to structure prediction. This is not only due to the high complexity of the prediction problem and the lack of efficient algorithms for sampling the huge conformational space, but also because of the currently incomplete understanding of the essential principles that govern the folding process. Therefore, template-free methods impose great challenges from both the biological and computational perspectives.

As expressed by Kelm *et al.* [112], a template-free method aims to simulate the protein folding process in silico. This is generally implemented by encoding the rules of chemistry and physics in an energy function, and then exploring a protein chain's conformational space while minimizing this designed energy function. In the literature, there have been reported detailed physics-based energy functions which attempt to capture the interactions between pairs of atoms to compute the three-dimensional structure of proteins [196]. These models are thus to be referred to as physics-based all-atom energy models. Among the most representative of such models, it is possible to mention ECEPP (*Empirical Conformational Energy Program for Peptides*) [3, 162, 170], OPLS (*Optimized Potentials for Liquid Simulations*) [108], CHARMM (*Chemistry at HARvard Macromolecular Mechanics*) [19, 20], AMBER (*Assisted Model Building with Energy Refinement*) [41, 189], and GROMOS (*GROningen MOlecular Simulation*) [198, 199]. These models contain terms associated with bond lengths, angles, torsion angles, van der Waals, and electrostatics interactions, and the major difference between them lies in the selection of atom types and the interaction parameters [127].

Physics-based all-atom models, as those mentioned above, have largely been used within the framework of molecular dynamics (MD) simulations [197]. MD samples the space of potential conformations by simulating the natural process of protein folding. MD starts from an arbitrarily chosen conformation that reacts and changes progressively as a result of the existing forces at the atomic level. Assuming that the description of the forces at the atomic level is accurate (which is not the case), following the trajectory of the system should lead to the native protein conformation. MD is a very computationally demanding technique. It involves a full atomic representation of proteins and considers the interaction between all pairs of atoms in the amino acids (whose number grows rapidly as the length of the protein chain increases), as well as the interaction between these atoms and those of the surrounding environment. The complexity of MD simulations has motivated the use of alternative approaches based on energy minimization through pathways that do not necessarily follow the natural protein folding process, such is the case of metaheuristic algorithms [46, 54, 64, 148, 153].

Besides pure physics-based energy models, there exist also knowledge-based approaches. Although knowledge-based models do not directly use previously determined protein structures as in template-based methods, they involve statistical models that capture the properties of native conformations, as observed in protein structure databases. These properties include, for example, the tendencies of certain amino acids to interact with one another and with the solvent, or secondary structure propensities [127, 172]. According to Ngan *et al.* [172], *in essence, the physics-based functions aim at predicting the native structure of a given sequence by mimicking the energetics of protein folding, whereas the knowledge-based functions bypass this intermediate step by directly making statistical inferences on what are observed in the database*. One of the consistently best-performing prediction techniques in CASP[8] competitions is ROSETTA [185]. The ROSETTA method relies on the use of a library of fragments representing observed (from a database of known protein structures) local structures for all short segments of the protein sequence. In a first stage, these fragments are assembled by means of the simulated annealing search procedure [114], the candidate solutions being evaluated on the basis of a scoring function derived from conformational statistics of known protein

[8] CASP (Critical Assessment of Structure Prediction) experiments aim at monitoring and establishing the current state-of-the-art in protein structure prediction [165].

structures. In the second stage, however, low-resolution conformations obtained through fragment assembly are subject to all-atom refinement using the ROSETTA's physics-based energy model.

Due to the flexibility and the number of atoms in protein molecules, the space of potential conformations is excessively large. This makes studies at atomic resolution to some extent prohibitive even for relatively short protein sequences. In addition to the use of all-atom protein models, therefore, it is possible to find in the literature structure prediction approaches implementing a range of other models characterized by the varying levels of detail (or resolution) used to represent amino acids and their interactions. Simplified or coarse-grained models have been developed in order to capture the most important characteristics of proteins. These models discretize the conformational space, and some of the atoms are ignored or groups of them are treated as united pseudo-atoms [14]. By sacrificing atomistic details and imposing constraints on the possible conformations, coarse-grained models lower the computational complexity of protein simulations, enabling the study of the most general and essential principles governing the folding process [28, 38, 86, 120, 180]. To a certain degree, however, these models compromise the accuracy and biological plausibility of the predicted conformations. Nevertheless, as pointed out by Homouz [90], this is the price that has to be paid to capture the main features of protein folding over reasonable times.

One of the most extensively explored simplified models for protein structure prediction is the *hydrophobic-polar* (HP) *model* proposed by Dill [61, 126]. This model abstracts the fact that the hydrophobicity of amino acids is a major determinant of the folded state of proteins. Only two types of amino acids are considered, hydrophobic (H) and polar (P), and the problem is reduced to that of finding a self-avoiding embedding of the protein chain on a given lattice such that the interaction among H amino acids is maximized. Notwithstanding, even under such a simplified model protein structure prediction represents a challenging optimization problem [10, 43]. The HP model is the particular case of study of this research project and is further described in Section 2.3.3.

Extensions to the HP model have been reported in the literature. For example, rather than considering only attractive forces between H amino acids, *i.e.*, H-H interactions, repulsive forces in H-P and P-P interactions are also involved in the so-called "shifted" HP model [27]. The shifted HP model is also commonly referred to as the functional protein model [13, 53, 88, 121].

The HPNX model [17] is a more fine-grained energy model which extends the HP model by further classifying polar amino acids as positively charged (P), negatively charged (N) and neutral (X). The HPNX model aims to maximize not only the interaction among H amino acids, but captures also electrostatic interactions between positively and negatively charged residues. An improvement of the HPNX model was recently proposed, called the hHPNX model [93]. In the hHPNX model, the H group of amino acids splits into two distinct categories, h and H, in order to give separate treatment to some hydrophobic amino acids such as alanine and valine. This separation is based on the observation that these amino acids present different interaction behaviors than other hydrophobic residues [44]. There are also studies on lattice models that use larger alphabets (amino acid groups). For example, it is possible to consider interactions between the whole set of 20 amino acid types [132]. Finally, there exists a variety of off-lattice coarse-grained protein models, models that do not constrain the space of potential conformations to the possible embeddings of a protein chain on a given lattice. For comprehensive reviews on the diversity of such models the reader is referred to [14, 90, 120].

As commented before, the use of coarse-grained protein models sacrifices, to a certain extent, the practical usefulness of the achieved predictions. As expressed by Blaszczyk *et al.* [14], none of the existing models is able to simplify an atomistic system without losing its important features, such as structural details, characteristic interactions and dynamics. Therefore, rather than functioning as stand-alone protein structure prediction approaches, coarse-grained models are commonly used within the framework of a hierarchical or multi-scale methodology. In [173], for instance, Ołdziej *et al.* used a coarse-grained model, the physics-based united-residue (UNRES) model [134–136], during the first stage of the prediction process. Then, the lowest-energy coarse-grained structures were converted to an all-atom representation and optimized based on the ECEPP force field [3, 162, 170]. Dayem Ullah *et al.* [56] proposed a hierarchical approach where the first stage relies on the HP lattice model to find compact structures maximizing the interaction among H-type amino acids. In a later stage, the obtained compact structures served as the starting points to further optimize the protein structure for the input sequences by employing simulated annealing and a 20 amino acid pairwise interactions energy function [11]. The initialization of the simulated annealing procedure with HP model-based compact conformations significantly improved the overall performance of the proposed prediction scheme [56]. Other interesting studies on the use of this kind of hierarchical methodologies for protein structure prediction include [87, 116, 190, 241]. Finally, it is important to conclude this section by highlighting that the success achieved through hierarchical methodologies for protein structure prediction, justifies the work of the continuously growing community (as seen from the number of reported studies) that concentrates on improving coarse-grained protein models, as well as on the development of more effective strategies to search their corresponding landscapes.

2.3.3 Protein structure prediction under the HP model

The 20 different amino acids found in proteins can be classified primarily as *hydrophobic* or *polar* on the basis of their affinity for water. This is determined from the hydrophobicity index, a measure of the relative hydrophobicity of amino acids (or how soluble the amino acids are in water). Table 2.1, presented previously in Section 2.3.1, details whether each of the 20 amino acids falls into these categories. Polar (P) or *hydrophilic* amino acids are usually found at the outer surface of proteins. By interacting with the aqueous environment, P amino acids contribute to the solubility of the molecule. In contrast, hydrophobic (H) or *nonpolar* amino acids tend to pack on the inside of proteins, where they interact with one another to form a water-insoluble core. Such a phenomenon, usually referred to as *hydrophobic collapse*, is one of the major driving forces during the folding process of globular proteins (as discussed in Section 2.3.1.2). The hydrophobic collapse represents the reasoning and motivation behind the hydrophobic-polar (HP) model for protein structure prediction [61, 126].

In the HP model, proteins are abstracted as chains of H- and P-type beads. Protein sequences, which are originally defined over a 20-letter alphabet, are now of the form $S = \langle a_1, a_2, \ldots, a_\ell \rangle$, where $a_i \in \{H, P\}$ denotes the i-th amino acid and ℓ is the length of the sequence. The number of H and P amino acids in S are here referred to as ℓ_H and ℓ_P, respectively. A protein conformation is represented in this model as an embedding of the protein chain on a given lattice (both the two-dimensional square and three-dimensional cubic lattices are considered in this research project). With the aim of emulating the so-called hydrophobic collapse, the goal in the HP model is to maximize the interaction among H amino acids in the lattice. Such interactions are to be referred to as H-H

topological contacts. Two H amino acids a_i and a_j are said to form a topological contact if they are nonconsecutive in S (*i.e.*, $|j - i| \geq 2$) but adjacent in the lattice. The objective is thus to find a lattice embedding of the protein chain where the number of H-H topological contacts, $HHtc$, is maximized. Adhering to the notation of the field, an energy function, to be minimized, is defined as the negative of $HHtc$; *i.e.*, maximizing $HHtc$ is equivalent to minimizing such an energy function.

2.3.3.1 The HP model as an optimization problem

Let $\mathcal{X}$ be the set of all potential protein conformations, *i.e.*, the search space, and let $\mathcal{X}_\mathcal{F}$ denote the subset of all the feasible states ($\mathcal{X}_\mathcal{F} \subsetneq \mathcal{X}$). PSP under the HP model can be more formally stated as the problem of finding $\mathbf{x}^* \in \mathcal{X}_\mathcal{F}$ such that $E(\mathbf{x}^*) = \min\{E(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}_\mathcal{F}\}$. The energy function $E : \mathcal{X} \to \mathbb{R}$ maps each possible conformation $\mathbf{x} \in \mathcal{X}$ to an energy value:

$$E(\mathbf{x}) \quad = \quad \sum_{a_i, a_j} e(a_i, a_j), \tag{2.5}$$

where

$$e(a_i, a_j) \quad = \quad \begin{cases} -1, & \text{if } a_i \text{ and } a_j \text{ are both } H \text{ amino acids} \\ & \text{and they form a topological contact;} \\ & \\ 0, & \text{otherwise.} \end{cases}$$

As an example, the optimal structure for a protein sequence of length $\ell = 20$ on the two-dimensional square lattice is presented in Figure 2.5. This example corresponds to protein sequence 2d4, one of the test instances for the HP model considered in this study (refer to Section 2.4.1 for details about the adopted test cases). The prediction of protein structures based on the HP model is a hard combinatorial optimization problem which has been proved to be NP-complete [10, 43].

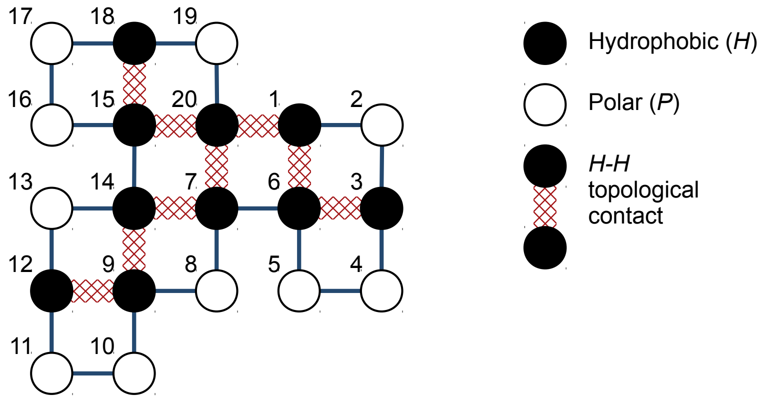


Figure 2.5: Optimal conformation for sequence 2d4 of length $\ell = 20$ on the two-dimensional square lattice. Black and white beads denote H and P amino acids, respectively. Amino acids have been numbered from 1 to ℓ according to their positions in the protein sequence S. The energy of this conformation is $E = -9$, since there are 9 H-H topological contacts, $HHtc = 9$.

2.3.3.2 Problem constraints

For a protein conformation to be considered feasible (*i.e.*, valid), its corresponding embedding on the lattice is required to satisfy two different properties: *connectivity* and *self-avoidance*. On the one hand, connectivity requires consecutive amino acids in the protein sequence to be placed at adjacent positions of the lattice. On the other hand, the self-avoidance property implies that the conformation has to be free of collisions; *i.e.*, two different amino acids can not be assigned to the same lattice position. While connectivity is implicitly satisfied by using an internal coordinates representation, as described in Section 2.3.3.3, such a representation scheme can not ensure the self-avoidance of the encoded conformations; refer to the example provided in Figure 2.6. Therefore, an explicit mechanism is required to be implemented in order to address the self-avoidance constraint.

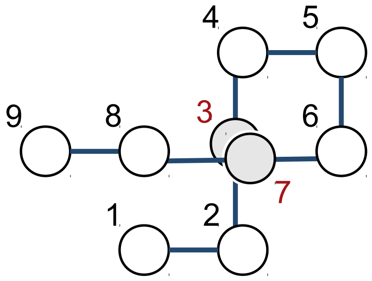


Figure 2.6: An infeasible protein conformation. A collision was produced when amino acids 3 and 7 were mapped to the same lattice position.

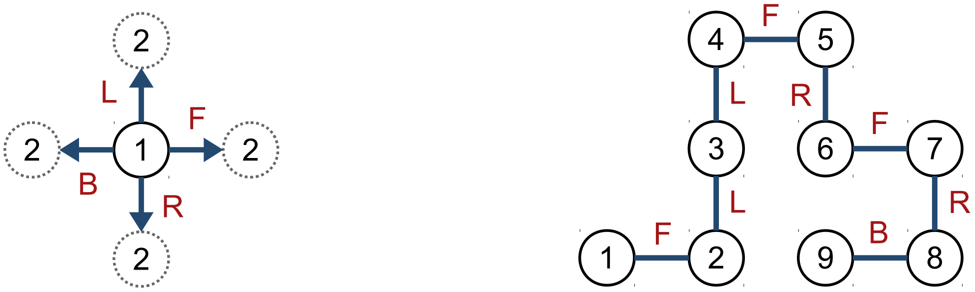


Figure 2.7: Internal coordinates representation based on absolute moves. Illustration of the encoding scheme (left). An example conformation encoded as $FLLFRFRB$ (right).

2.3.3.3 Conventional problem representations

In the literature, the vast majority of the metaheuristic approaches which have been proposed for the HP model of the PSP problem are based on the use of an *internal coordinates representation* of the protein conformations. Using internal coordinates, a protein conformation is encoded as a sequence of moves specifying the lattice position for each amino acid with regard to the preceding one; the position of the first amino acid is fixed. Two alternative encoding schemes can be adopted, namely, the *absolute moves encoding* [224] and the *relative moves encoding* [177].

Based on a global reference system defined by the lattice, the absolute moves encoding represents three-dimensional conformations (in the cubic lattice) as sequences in $\{F, B, L, R, U, D\}^{\ell-1}$. These symbols ($F$,$B$,$L$,$R$,$U$ and D) are used to denote the forward, backward, left, right, up and down moves from one amino acid to the next. Only moves $\{F, B, L, R\}$ are allowed in the two-dimensional case (square lattice). An example of the absolute moves encoding is presented in Figure 2.7.

In the relative moves encoding, conformations are represented as sequences in $\{F, L, R, U, D\}^{\ell-2}$ for the three-dimensional lattice, and $\{F, L, R\}^{\ell-2}$ for the two-dimensional one. In contrast to the absolute encoding, the relative alternative implements a local reference system which rotates at each encoding decision (other than F). No backward (B) moves are allowed, ensuring that the encoded conformation will always be *one-step self-avoiding*. Note that only $\ell - 2$ encoding decisions need to be taken by assuming the first move to be forward (F). An example is provided in Figure 2.8.

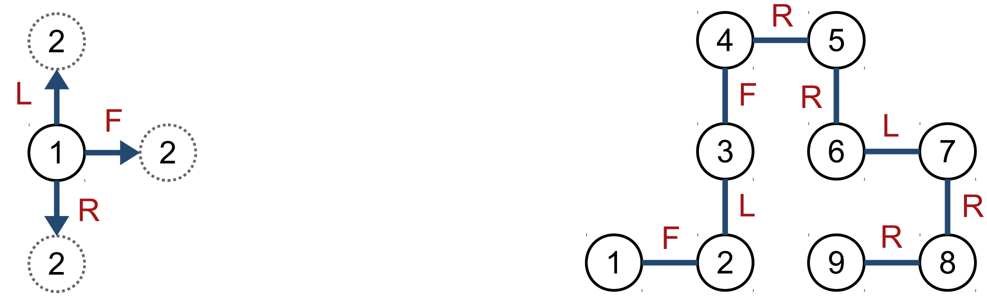


Figure 2.8: Internal coordinates representation based on relative moves. Illustration of the encoding scheme (left). An example conformation encoded as $FLFRRLRR$ (right).

2.3.3.4 *Related work: metaheuristic approaches for the HP model*

As mentioned in Section 2.3.3.1, the HP model for protein structure prediction has been proved to be an NP-complete problem [10, 43]. Though exact methods have been proposed in the literature to cope with this particular problem [30, 94], the exponential growth in the number of possible protein conformations has limited the application of these methods to relatively small problem instances. The plethora of the proposed approaches for the HP model rely on the use of a diversity of metaheuristic algorithms [140, 248]. The *genetic algorithm* (GA) has been, by far, the most widely used metaheuristic approach to deal with the HP model [81, 89]. The seminal work of Unger and Moult investigated the application of GAs to solve the two-dimensional HP model based on the square lattice [225, 226], later extending their research to the three-dimensional model based on the cubic lattice [224]. Unger and Moult showed that GAs can be more effective than *Monte Carlo* (MC) methods, since GAs are less likely to be trapped in local optima (mainly due to the nature of the crossover operator [223]). Note, however, that some MC-based approaches have achieved very competitive results in the literature; see, for example, the *replica exchange Monte Carlo* algorithm proposed by Thachuk *et al.* [219]. In fact, Unger and Moult's algorithm can be considered to be a hybridization of a GA with MC. Metropolis filters are used at the output of variation operators, so that improving solutions are always accepted but the acceptance of non-improving offspring is based on a nondeterministic decision. These filters restrict also the acceptance of infeasible solutions; if an infeasible solution is generated, the variation operators iterate until a feasible solution is obtained.

Patton *et al.* [177] implemented a GA based on the work of Unger and Moult. The algorithm of Patton *et al.* does not filter infeasible solutions, but penalizes infeasible solutions according to the number of conflicts they present. Using different test protein sequences, the GA of Patton *et al.* was found to achieve equal or better results than those of Unger and Moult's algorithm, but with only about 10% of their total number of objective function evaluations. A fundamental difference between the GA of Patton *et al.* and that of Unger and Moult is the encoding scheme. Both proposals use an internal coordinates representation. However, whereas Unger and Moult adopted the absolute moves encoding, Patton *et al.* utilized the relative moves encoding (the absolute and relative moves encoding schemes were described in Section 2.3.3.3). The studies of Unger and Moult [224–226] and Patton *et al.* [177] are among the most important works in the specialized literature. This is not only due to the pioneering nature of these works, but also because the proposed absolute and relative encoding schemes represent the basis on which the vast majority of the metaheuristic approaches for the HP model proposed so far have been designed. Despite the fact that the relative moves encoding reduces the size of the conformational space and all the encoded conformations are *one-step self-avoiding* (see Section 2.3.3.3), there is not clear evidence in the literature which suggests the superiority of one of these encoding schemes with respect to the other in practice; even contradictory arguments and results have been reported on this regard [42, 122].

There are other early works that use GAs in order to solve the HP model of the PSP problem; refer, for instance, to [113, 122, 123]. Broadly, the reported GAs differ mainly in the used representation, the fitness (evaluation) function, the treatment of infeasible solutions and the implemented search operators. The work of Krasnogor *et al.* [122] has been particularly relevant because it analyzed the impact that the choice of certain design components could have on the performance of GAs when solving this problem, namely, the encoding of solutions, the evaluation function and the constraint-handling mechanism. In the literature, different implementations for each of these components have been proposed. However, as the authors aptly pointed out [122]:

> "*... prior researchers selected an encoding without explicit numerical comparisons. Since other algorithmic parameters were also chosen differently, it is difficult to asses the impact that the choice of encoding has on a genetic algorithm's performance.*"

Despite the significant growth of the field since then, such a statement remains true at the present time and it generalizes to the different design components defining the characteristics of the fitness landscape, which exert thus a direct influence on the behavior of metaheuristic algorithms. For example, several authors have proposed alternative formulations of the evaluation function for the HP model in oder to guide more effectively the search process [9, 26, 32, 45, 98, 122, 142]. Nevertheless, in most of the cases the merits of these alternative evaluation functions have only been partially investigated. Therefore, Chapter 3 of this thesis is concerned with the analysis and comparison of such different strategies. Similarly, there is not a clear consensus with respect to the most appropriate treatment for the large number of infeasible solutions that the landscapes in the HP model involve [42, 57, 66, 122, 192]. By proposing and analyzing the advantages of a new constraint-handling mechanism for the HP model, and by comparing it with respect to representative approaches from the literature, Chapter 5 is intended to contribute in providing further insight into this matter.

More recent studies on the application of GAs to the HP model of the PSP problem include [24, 42, 45, 47, 92, 142]. Also, it has become popular in recent years to address this problem by hybridizing GAs with a local improvement strategy [7, 9, 31–34, 98–101, 143, 169]. These are the so-called *memetic algorithms* [163, 164]. It has also been reported in the literature the use of multimeme algorithms [124], where the evolutionary method is hybridized with a set of local search heuristics that are self-adaptively selected and applied according to each particular problem instance, search stage or individual in the population [121, 178]. Finally, the HP model of the PSP problem has also been tackled by coupling GAs with other metaheuristics such as *tabu search* (TS) [80], see for example [98, 183, 246], and the *artificial bee colony* algorithm [110, 111], refer for instance to [229].

Alongside GAs, a variety of metaheuristic approaches have been proposed for PSP under the HP model, including TS (without hybridizing to GAs) [15, 175], other TS-based hybrids [184, 249], *ant colony optimization* [35, 37, 63, 83, 95, 168, 211], *immune-based algorithms* [48, 49, 53, 57], *particle swarm optimization* [21, 109, 152], *differential evolution* [12, 104, 141, 192] and *estimation of distribution algorithms* [29, 191]. The HP model has also been dealt through recently developed metaheuristic paradigms such as the firefly-inspired algorithm [149, 247], artificial plant optimization [25] and the *energy-landscape paving* technique [133]. Cellular automata have also been used to achieve not only the final folded conformation of HP model protein sequences, but also to model the temporal and dynamic folding process which emerges as a consequence of amino acid interactions [193, 194].

Finally, it is important to remark that, according to the author's revision of the state-of-the-art, all previous studies have addressed the HP model of the PSP problem from the single-objective optimization perspective. It was not until the research work developed in Chapters 4 and 5 of this thesis that this particular problem is approached by using multi-objective optimization techniques.

2.4 Performance assessment

This section presents the performance assessment methodology followed during the development of this research. First, Section 2.4.1 details all the considered test instances for the two- and three-dimensional variants of the HP model (square and cubic lattices, respectively). Section 2.4.2 defines the adopted performance measures. The methodology on which the conducted statistical significance analyses were based is described in Section 2.4.3. Finally, Section 2.4.4 summarizes the main characteristics of the experimental platform utilized during the course of this project.

2.4.1 Test instances

A total of 30 well-known benchmark sequences for the HP model have been considered in this research project: 15 out of them are for the two-dimensional square lattice, and the remaining 15 are for three-dimensional cubic lattice. Tables 2.2 and 2.3 present the full HP protein sequences, their length (ℓ) and the optimal or best-known energy value (E^*) reported in the literature, to the best of the author's knowledge [53, 100, 121, 219, 240, 245].

2.4.2 Performance measures

Although alternative (either single-objective or multi-objective) formulations of the HP model are investigated in Chapters 3, 4 and 5, it is important to remark that the focus of this research project remains always to solve the original problem, as it was introduced in Section 2.3.3. Therefore, all

Table 2.2: Considered test sequences for the HP model. Two-dimensional square lattice.

Seq.	**Full HP protein sequence**	ℓ	E^*
2d1	$H_2P_5H_2P_3HP_3HP$	18	-4
2d2	$HPHPH_3P_3H_4P_2H_2$	18	-8
2d3	$PHP_2HPH_3PH_2PH_5$	18	-9
2d4	$HPHP_2H_2PHP_2HPH_2P_2HPH$	20	-9
2d5	$H_3P_2HPHPHP_2HPHPHP_2H$	20	-10
2d6	$H_2P_2HP_2HP_2HP_2HP_2HP_2HP_2H_2$	24	-9
2d7	$P_2HP_2H_2P_4H_2P_4H_2P_4H_2$	25	-8
2d8	$P_3H_2P_2H_2P_5H_7P_2H_2P_4H_2P_2HP_2$	36	-14
2d9	$P_2HP_2H_2P_2H_2P_5H_{10}P_6H_2P_2H_2P_2HP_2H_5$	48	-23
2d10	$H_2(PH)_4H_3P(HP_3)_3(P_3H)_3PH_4(PH)_4H$	50	-21
2d11	$P_2H_3PH_8P_3H_{10}PHP_3H_{12}P_4H_6PH_2PHP$	60	-36
2d12	$H_{12}PHPH(P_2H_2P_2H_2P_2H)_3PHPH_{12}$	64	-42
2d13	$H_4P_4H_{12}P_6(H_{12}P_3)_3HP_2H_2P_2H_2P_2HPH$	85	-53
2d14	$P_6HPH_2P_5H_3PH_5PH_2P_4H_2P_2H_2PH_5PH_{10}PH_2PH_7P_{11}H_7P_2HPH_3P_6HPH_2$	100	-48
2d15	$P_3H_2P_2H_4P_2H_3PH_2PH_2PH_4P_8H_6P_2H_6P_9HPH_2PH_{11}P_2H_3PH_2PHP_2HPH_3P_6H_3$	100	-50

Table 2.3: Considered test sequences for the HP model. Three-dimensional cubic lattice.

Seq.	**Full HP protein sequence**	ℓ	E^*
3d1	$HPHP_2H_2PHP_2HPH_2P_2HPH$	20	-11
3d2	$H_2P_2HP_2HP_2HP_2HP_2HP_2HP_2H_2$	24	-13
3d3	$P_2HP_2H_2P_4H_2P_4H_2P_4H_2$	25	-9
3d4	$P_3H_2P_2H_2P_5H_7P_2H_2P_4H_2P_2HP_2$	36	-18
3d5	$P_2H_3PH_3P_3HPH_2PH_2P_2HPH_4PHP_2H_5PHPH_2P_2H_2P$	46	-35
3d6	$P_2HP_2H_2P_2H_2P_5H_{10}P_6H_2P_2H_2P_2HP_2H_5$	48	-31
3d7	$H_2(PH)_4H_3P(HP_3)_3(P_3H)_3PH_4(PH)_4H$	50	-34
3d8	$PH(PH_3)_2P(PH_2PH)_2H(HP)_3(H_2P_2H)_2\ PHP_4(H(P_2H)_2)_2$	58	-44
3d9	$P_2H_3PH_8P_3H_{10}PHP_3H_{12}P_4H_6PH_2PHP$	60	-55
3d10	$H_{12}PHPH(P_2H_2P_2H_2P_2H)_3PHPH_{12}$	64	-59
3d11	$P(HPH_2PH_2PHP_2H_3P_3)_3(HPH)_3P_2H_3P$	67	-56
3d12	$P(HPH)_3P_2H_2(P_2H)_6H(P_2H_3)_4P_2(HPH)_3\ P_2HP(PHP_2H_2P_2HP)_2$	88	-72
3d13	$P_2H_2P_5H_2P_2H_2PHP_2HP_7HP_3H_2PH_2P_6HP_2HPHP_2HP_5H_3P_4H_2PH_2P_5H_2P_4H_4PHP_8H_5P_2HP_2$	103	-58
3d14	$P_3H_3PHP_4HP_5H_2P_4H_2P_2H_2(P_4H)_2P_2HP_2H_2P_3H_2PHPH_3P_4H_3P_6H_2P_2$ $HP_2HPHP_2HP_7HP_2H_3P_4HP_3H_5P_4H_2(PH)_4$	124	-75
3d15	$HP_5HP_4HPH_2PH_2P_4HPH_3P_4HPHPH_4P_{11}HP_2HP_3HPH_2P_3H_2P_2HP_2HPHPHP_8HP_3$ $H_6P_3H_2P_2H_3P_3H_2PH_5P_9HP_4HPHP_4$	136	-83

the obtained experimental results are to be evaluated in terms of the conventional energy function of the HP model (E) or, which is equivalent, according to the total number of H-H topological contacts ($HHtc$) in the obtained protein conformations.

Two additional performance measures have been considered, both computed over multiple independent executions of the implemented search algorithms. First, the *relative root mean square error* (RMSE) is determined individually for each given test instance t as follows:

$$\text{RMSE}(t) = 100\%\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\frac{E_r(t) - E^*(t)}{E^*(t)}\right)^2}, \tag{2.6}$$

where $E_r(t)$ denotes the energy of the best solution found during a single execution r, R is the total number of executions carried out, and $E^*(t)$ is the optimal (or best known) energy value for instance t (as indicated in Section 2.4.1). RMSE indicates the performance scored for the particular instance t in a 0% to 100% scale, being RMSE$(t) = 0\%$ the preferred value for this measure.

Finally, the *overall relative root mean square error* (O-RMSE) measure extends RMSE in order to assess the overall performance of the studied approaches, considering all the test instances. Having defined RMSE, O-RMSE can be formally defined as follows:

$$\text{O-RMSE} = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\text{RMSE}(t), \tag{2.7}$$

where $\mathcal{T}$ is the set of all the adopted test instances. Thus, O-RMSE $= 0\%$ suggests the ideal situation where the optimal solution for each instance was reached during all the performed executions.

2.4.3 Statistical significance testing

In the experiments presented in this document, the statistical significance analyses were conducted as follows. First, *D'Agostino-Pearson's omnibus* K^2 test was used to evaluate the normality of data distributions. For normally distributed data, either *ANOVA* or the *Welch's t* parametric tests were used depending on whether the variances across the samples were homogeneous (*homoskedasticity*) or not. This was investigated using the *Bartlett's* test. For non-normal data, the nonparametric *Kruskal-Wallis* test was adopted. Finally, a significance level of $\alpha = 0.05$ has been considered.

2.4.4 Experimental platform

The algorithms implemented in this research project were coded in ANSI C and compiled with *gcc* using the optimization flag -$O3$. All experiments performed were run sequentially on the Neptuno cluster at the Information Technology Laboratory, CINVESTAV-Tamaulipas. This cluster is equipped with 10 InfiniBand interconnected nodes, each of which features 8 cores running at 2.66 GHz, has a total of 16 GB of RAM, and uses the CentOS distribution of the Linux operating system.

3

Using alternative energy functions to cope with the neutrality of the HP model's fitness landscapes

3.1 Introduction

Metaheuristic algorithms rely on an effective evaluation scheme, which can distinguish each candidate solution from the others, to make the most appropriate choice at each iteration. This is of utmost importance in order to guide the search process towards promising regions of the solution space. Nevertheless, the conventional evaluation scheme of the HP model, consisting of the energy function defined in Section 2.3.3.1, induces a very poor discrimination among potential conformations. Therefore, there could be many different conformations for a given protein sequence presenting the same energy value (this scenario is illustrated through Figure 3.1). More precisely, given a protein sequence S, with length ℓ and optimal energy value E^*, there can be at most $|E^*| + 1$ available

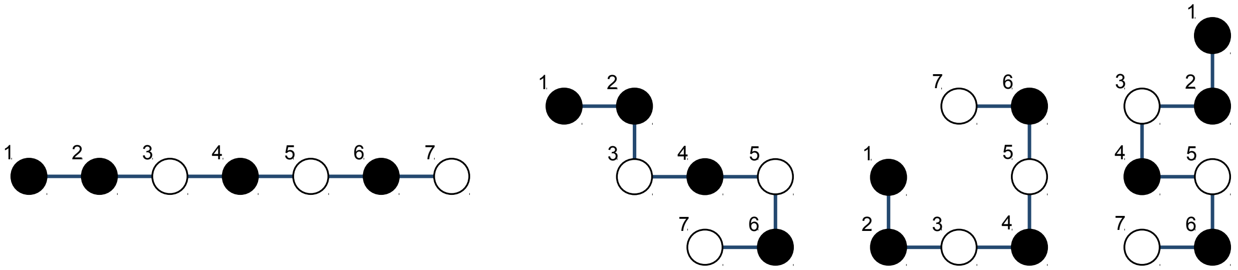


Figure 3.1: Four different structures for protein sequence HHPHPHP on the two-dimensional square lattice. All the four protein conformations present the same energy value, namely, $E = 0$.

energy levels to classify a search space of size[1] $|\mathcal{X}| = 4^{\ell-1}$. As an example, consider sequence 2d1, the smallest of the test cases adopted for this research project (refer to Section 2.4.1 for details). In this case, $\ell = 18$ and $E^* = -4$, so that there are only five different energy levels which can be used to discriminate among a total of $4^{17} = 17,179,869,184$ potential solutions. Note, however, that some equally ranked conformations could present better chances than others to be further improved.

From the fitness landscape perspective, the low discrimination provided by the conventional energy function of the HP model translates into the existence of a considerable number of solutions with a high neutrality degree. As discussed in greater detail in Section 2.2.4, this produces a neutral fitness landscape characterized by large plateaus of incomparable solutions. In such plateaus, metaheuristic algorithms, especially trajectory (local search-based) methods, could fail to detect and exploit promising search directions, leading the optimization process to be oriented almost at random.

For this reason, alternative energy functions for the HP model have been proposed in the literature [9, 26, 45, 98, 122, 142]. The aim of these alternative formulations of the energy function is to provide a more fine-grained discrimination among candidate conformations, as a means of guiding the search process of metaheuristics in a more effective manner. In most of the cases, however, the proposal of these alternative evaluation approaches was not supported, or it was only partially supported, by solid experimental evidence. Moreover, to the best of the author's knowledge, a comparative study devoted to exploring the impact of using different of such approaches has never been reported.

[1] The given size of the search space assumes the use of the two-dimensional square lattice and the absolute moves representation of the protein conformations, see Section 2.3.3.3.

This chapter is intended to formally analyze and compare different formulations of the HP model's energy function. Seven energy formulations are considered: the conventional energy function of the HP model, and six alternative proposals from the literature. Given that the purpose of using an alternative energy function is mainly to address the weak discrimination of the conventional formulation, an in-depth investigation of the discrimination potential for each of the studied approaches is first conducted. Then, an essential property, HP-compatibility, is introduced and explored for each considered function. This property reflects whether or not an alternative energy function is consistent with the original definition of the problem. Finally, an assessment of the practical usefulness of the studied evaluation approaches within two different metaheuristic algorithms is carried out.

This chapter is organized as follows. In Section 3.2, the alternative energy functions for the HP model considered in this study are described. The experimental results are presented in Section 3.3. Finally, Section 3.4 provides the conclusions as well as some possible directions for future research.

3.2 Alternative energy functions for the HP model

The purpose of this section is to describe in detail the studied alternative formulations of the HP model's energy function. Six different approaches that have been proposed in the specialized literature are covered by this study, each of which is discussed separately below.

3.2.1 Krasnogor et al., 1999

Given two conformations with the same number of H-H topological contacts, it is possible that one of them has better characteristics than the other. Based on this observation, Krasnogor *et al.* proposed the following distance-dependent energy function [122]:[2]

$$E_{K99}(\mathbf{x}) \quad = \quad \sum_{a_i, a_j} e(a_i, a_j), \tag{3.1}$$

[2] This energy function is denoted as E_{K99} in order to distinguish this alternative formulation from the conventional formulation of the HP model's energy function, defined as E in Section 2.3.3.1. Acronyms similar to K99, which is based on first author's initial and publication year, have also been used within the definition of all other alternative energy formulations described in subsequent sections.

where

$$e(a_i, a_j) = \begin{cases} -1, & \text{if } a_i \text{ and } a_j \text{ are both } H \text{ and } d(a_i, a_j) = 1; \\ -1/(d(a_i, a_j)^k \ell_H), & \text{if } a_i \text{ and } a_j \text{ are both } H \text{ and } d(a_i, a_j) > 1; \\ 0, & \text{otherwise.} \end{cases}$$

In the above definition, $d(a_i, a_j)$ denotes the distance on the lattice between amino acids a_i and a_j, and a value of $d(a_i, a_j) = 1$ indicates that a_i and a_j form a topological contact. The authors suggest to use a value of $k = 4$ for the square lattice and $k = 5$ for the cubic and triangular lattices [122].

According to Krasnogor *et al.* [122], this formulation of the energy function preserves the conventional rank ordering of the conformations, at the same time it enables a finer level of distinction among conformations with the same number of H-H topological contacts. In [122], this function was analyzed using a genetic algorithm. Only five relatively short protein sequences (with less than 50 amino acids) were considered. Experiments were performed for the two-dimensional square and triangular lattices, as well as for the three-dimensional cubic lattice. Although no detailed results are provided, it was pointed out that no significant improvements in performance were obtained by using this modified energy function. However, the authors suggest that the advantages of using this function can become more evident for larger protein sequences and by implementing this approach within a local search strategy. The relevance of using this proposal needs to be further investigated.

3.2.2 Custódio et al., 2004

Given that the aim in the HP model is only to maximize the interactions between H amino acids, the positioning of P amino acids is not directly optimized. This may result in unnatural structures for sequences with long P segments and, particularly, when such P segments are located at the ends of the protein chain [45]. An example of this scenario is presented in Figure 3.2.

Custódio *et al.* proposed a modified energy function based on the assumption that it may be preferable for an H amino acid to have a P neighbor rather than to be in contact with the aqueous solvent [45]. In the proposed function, the energy of a conformation is computed as the weighted sum

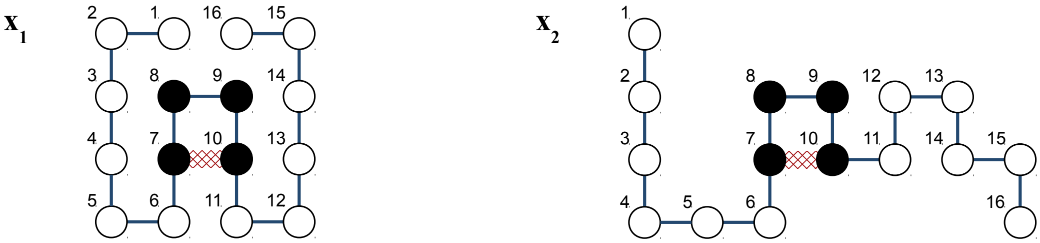


Figure 3.2: Two conformations with the same number of H-H topological contacts ($HHtc = 1$). However, the structure on the left is more natural-like (globular) than the one on the right.

of the number of H-H contacts (HHc), H-P contacts (HPc) and H-solvent contacts (HSc). A free lattice location (not assigned to any amino acid) is said to be occupied by the solvent. Formally, the energy of a conformation $\mathbf{x}$ is given by:

$$E_{C04}(\mathbf{x}) \quad = \quad \omega_1 HHc + \omega_2 HPc + \omega_3 HSc, \tag{3.2}$$

where ω_1, ω_2 and ω_3 denote the relative importance of HHc, HPc and HSc, respectively. Although not specified in [45], these weighting coefficients were set to $\omega_1 = 0$, $\omega_2 = 10$ and $\omega_3 = 40$ for the reported experiments.[3] Thus, given these weights, the minimization of (3.2) penalizes H-P and H-solvent contacts, H-P contacts being favored over H-solvent contacts, while H-H interactions are not penalized (H-H contacts have no contribution to the energy value using these weights).

Custódio *et al.* [45] evaluated the suitability of this proposal by using a genetic algorithm. A total of 10 instances for the three-dimensional cubic lattice were considered (7 sequences of length $\ell = 27$ and the remaining 3 sequences are of length $\ell = 64$). The proposed function allowed to improve the performance of the implemented algorithm for some of the adopted instances. The reported results also suggest that this function presents a greater tendency to form more natural-like conformations.

3.2.3 Lopes and Scapin, 2006

Lopes and Scapin [142, 143] proposed an alternative energy function for the HP model which is based on the concept of *radius of gyration*. The radius of gyration is a measure of the compactness of

[3] This information was obtained through personal communication with the authors.

conformations; the more compact the conformation is, the lower the value is for this measure. The proposed function is defined in (3.3):

$$E_{L06}(\mathbf{x}) \quad = \quad HnLB \cdot RadiusH \cdot RadiusP. \tag{3.3}$$

The $HnLB$ term comprises the number of H-H topological contacts in the conformation ($HHtc$) and a penalty factor which takes into account the violation of the self-avoiding constraint. Formally:

$$HnLB \quad = \quad HHtc - (NC \cdot PW), \tag{3.4}$$

where NC is the number of collisions (defined in [142, 143] as the number of lattice nodes assigned to more than one amino acid) in the conformation and PW is the penalty weight. The value of PW depends on the chain length, ℓ, and it can be computed as $PW = (0.033 \times \ell) + 1.33$, as derived based on empirical observations of the authors during preliminary experimentation [143].[4]

Before defining the $RadiusH$ and $RadiusP$ terms, let us first define RgH as the radius of gyration for H amino acids:

$$RgH \quad = \quad \sqrt{\frac{\sum\limits_{a|a=H} \left[(x_a - X_H)^2 + (y_a - Y_H)^2 + (z_a - Z_H)^2\right]}{\ell_H}}, \tag{3.5}$$

where x_a, y_a and z_a are the lattice coordinates of amino acid a. The X_H, Y_H and Z_H terms denote the arithmetic mean of the coordinates for all H amino acids. Analogously, we can compute RgP, the radius of gyration for P amino acids, by considering P rather than H amino acids in (3.5).

[4] During the experimentation reported in this chapter, only solutions encoding feasible protein conformations have been considered. Therefore, the penalty factor in (3.4) was simply omitted.

Once RgH has been defined, the $RadiusH$ term measures how compact the hydrophobic core of the conformation is. The $RadiusH$ term is given by:

$$RadiusH \;=\; MaxRgH - RgH, \tag{3.6}$$

where $MaxRgH$ denotes the radius of gyration for H amino acids in a totally unfolded conformation; *i.e.*, $MaxRgH$ represents the maximum possible RgH value for the given protein sequence.

Finally, the $RadiusP$ term aims to push P amino acids away from the hydrophobic core. Given the previously defined RgH and RgP measures, the $RadiusP$ term is computed as:

$$RadiusP \;=\; \begin{cases} 1, & \text{if } (RgP - RgH) \geq 0; \\ \frac{1}{1-(RgP-RgH)}, & \text{otherwise.} \end{cases} \tag{3.7}$$

The $RadiusP$ term will always lie in the range $[0,1]$. A value of $(RgP - RgH) > 0$ means that P amino acids are more exposed than H amino acids. This is a convenient scenario, so that the $RadiusP$ term has no contribution to the final energy value ($RadiusP = 1$). Otherwise, $(RgP - RgH) < 0$ suggests that H amino acids are more spread than the P ones, so that $RadiusP$ is used to penalize the energy value of the conformation. Note that (3.3) is to be maximized.[5]

Lopes and Scapin argue that the above described function provides an adequate discrimination among conformations with the same number of H-H topological contacts [142, 143]. This function was implemented within a genetic algorithm in order to solve several instances on the two-dimensional square lattice. However, no results are provided on the advantages of using this proposed variant of the energy function with regard to the conventional energy formulation of the HP model.

3.2.4 Berenboym and Avigal, 2008

Berenboym and Avigal proposed an alternative energy function called the *global energy* [9]. In this function, each pair of nonconsecutive H amino acids contributes to the energy value even if they are

[5] The negative of (3.3) can be used as an energy-minimization formulation of the problem which adheres to the notation commonly used in this field.

not topological neighbors. Formally, the global energy for a given conformation $\mathbf{x}$ is defined as:

$$E_{B08}(\mathbf{x}) = \sum_{a_i,a_j} e(a_i,a_j), \tag{3.8}$$

where

$$e(a_i,a_j) = \begin{cases} \frac{-1}{(x_{a_i}-x_{a_j})^2+(y_{a_i}-y_{a_j})^2+(z_{a_i}-z_{a_j})^2}, & \text{if } a_i \text{ and } a_j \text{ are both } H \text{ and } |j-i| \geq 2; \\ 0, & \text{otherwise.} \end{cases}$$

Here, x_{a_i}, y_{a_i} and z_{a_i} denote the lattice coordinates of amino acid a_i. In [9], the effects of using a local search operator within a genetic algorithm were analyzed for both, the conventional and the proposed global energy functions. However, an explicit comparison to demonstrate the advantages of using a particular energy function was not reported. This issue needs to be further explored.

3.2.5 Cebrián et al., 2008

Cébrian *et al.* [26] proposed a variant of the HP model's energy function, which was later further explored in [65]. The proposed energy formulation measures the deviation that each pair of H amino acids presents with respect to the unit distance (*i.e.*, topological contact distance). Let $d(a_i,a_j)^2 = (x_{a_i}-x_{a_j})^2+(y_{a_i}-y_{a_j})^2+(z_{a_i}-z_{a_j})^2$ be the lattice distance between amino acids a_i and a_j, and let $dv(a_i,a_j) = d(a_i,a_j)^2 - 1$ denote its deviation from the unit distance. The energy value of a conformation $\mathbf{x}$ is given by:

$$E_{C08}(\mathbf{x}) = \sum_{a_i,a_j|a_i=H,a_j=H} dv(a_i,a_j)^k, \tag{3.9}$$

where $k \geq 1$ is a parameter of the function, whose larger values give more weight to unit distances. A value of $k = 2$ was adopted for this study, since this value provided the best behavior according to the results reported in [26]. $E_{C08}(\mathbf{x}^*) = 0$ would refer to the ideal (potentially unrealistic) scenario where all pairs of H amino acids are at a unit distance in conformation $\mathbf{x}^*$. In [26, 65], no experimental results to support the advantages of using the proposed energy function were reported.

3.2.6 Islam and Chetty, 2009

In [98, 99, 101], the authors reported a memetic algorithm with a modified energy function which incorporates two additional measures: H-*compliance* and P-*compliance*.

H-*compliance* measures the proximity of H amino acids to the center of a hypothetical cuboid (or rectangle in a two-dimensional space) enclosing all H amino acids, which is to be denoted by the reference point (x_r, y_r, z_r). Formally, this measure is given by:

$$H\text{-}\mathit{compliance}(\mathbf{x}) = \frac{\sum_{a|a=H} (x_r - x_a)^2 + (y_r - y_a)^2 + (z_r - z_a)^2}{\ell_H}, \tag{3.10}$$

where x_a, y_a and z_a denote the lattice coordinates of amino acid a.

P-*compliance* is a measure of how close P amino acids are to the boundaries of a hypothetical cuboid enclosing all P amino acids. Such a cuboid is defined by x_{min}, x_{max}, y_{min}, y_{max}, z_{min} and z_{max}. The P-*compliance* measure is formally given by:

$$P\text{-}\mathit{compliance}(\mathbf{x}) = \frac{\sum_{a|a=P} \min \left\{ \begin{array}{l} |x_{min} - x_a|, |x_{max} - x_a|, |y_{min} - y_a|, \\ |y_{max} - y_a|, |z_{min} - z_a|, |z_{max} - z_a| \end{array} \right\}}{\ell_P}. \tag{3.11}$$

Finally, the energy of a given conformation $\mathbf{x}$ is defined as:

$$E_{I09}(\mathbf{x}) = \alpha E(\mathbf{x}) + H\text{-}\mathit{compliance}(\mathbf{x}) + P\text{-}\mathit{compliance}(\mathbf{x}), \tag{3.12}$$

where E is the conventional energy function of the HP model defined in (2.5), see Section 2.3.3.1, and α is to be large enough in order to ensure that this will be the dominant term in (3.12).

In [98], the authors demonstrated the advantages of using the proposed energy function using an 85-length HP protein sequence on the two-dimensional square lattice. However, the influence of using this function should be further explored for a larger set of test cases.

3.3 Comparative analysis

In this section, seven different formulations of the energy function for the HP model are evaluated and compared:[6] D85, the conventional energy function of the HP model [61,126]; and the alternative energy formulations K99, reported by Krasnogor *et al.* [122]; C04, reported by Custódio *et al.* [45]; L06, reported by Lopes and Scapin [142, 143]; B08, reported by Berenboym and Avigal [9]; C08, reported by Cébrian *et al.* [26]; and I09, reported by Islam and Chetty [98, 99, 101].

It is important to remark that even when an alternative energy function is implemented, the goal of the optimization process remains to maximize the number of H-H topological contacts ($HHtc$), which is the singular objective in the HP model (see Section 2.3.3). In this study, the purpose of using alternative formulations of the energy function is to guide the search process in a more effective manner while solving the original problem. For all the experiments reported in this chapter, protein conformations are encoded using an internal coordinates representation based on absolute moves. Details on this problem representation are provided in Section 2.3.3.3. Moreover, only solutions encoding feasible protein conformations have been considered during all the analyses presented in this chapter. All the feasible conformations (either analyzed directly in Sections 3.3.1 and 3.3.2, or used as the initial solutions for the search algorithms implemented in Sections 3.3.3 and 3.3.4) have been randomly generated using the backtracking procedure proposed in [42].

The remaining of this section is organized as follows. First, important properties of the studied energy functions are examined in Sections 3.3.1 and 3.3.2. Then, the effectiveness of these approaches to guide the optimization process is evaluated in Sections 3.3.3 and 3.3.4.

3.3.1 Degree of discrimination

The discrimination potential is an important property of the evaluation scheme which impacts directly on the behavior of metaheuristic algorithms. That is, if it is not possible to set preferences among candidate solutions, then the progress in the search process could become practically dominated by

[6] For convenience, a three-letter acronym has been assigned to refer to each of the studied energy functions. The adopted acronyms are based on first author's initial and publication year.

random decisions. In this section, the degree of discrimination that the studied energy functions provide is investigated. This is done by analyzing the distribution of ranks that these approaches induce on a set of protein conformations. A ranking expresses the relationship among a set of elements according to a given property. In the context of this study, protein conformations are to be ranked and the property to set such a relationship corresponds to the energy value. Given a set of protein conformations, the first ranking position is assigned to the conformation with the best energy value, the next ranking position to the one with the second best energy value, and so on. If two or more conformations present the same energy, then they will share the same rank.

The *relative entropy* (RE) measure proposed by Corne and Knowles was adopted [40]. Given a set of n ranked conformations (there are at most n ranks, and at least 1), the relative entropy of the distribution of ranks D is defined as:

$$\text{RE}(D) = \frac{\sum_r \frac{D(r)}{n} \log(\frac{D(r)}{n})}{\log(1/n)}, \tag{3.13}$$

where $D(r)$ denotes the number of conformations with rank r. RE(D) tends to 1 as approaching to the ideal situation where each conformation has a different rank (*i.e.*, the maximum possible discrimination). On the other hand, when all the conformations share the same ranking position (*i.e.*, the poorest discrimination), RE(D) takes a value of zero.

In this experiment, $1,000$ different feasible conformations were generated at random. For each of the studied functions, these solutions were evaluated and ranked to finally compute the RE measure. A total of 100 repetitions of this experiment were performed for all the test instances. The overall statistics of this experiment are presented in Figure 3.3. Instance-specific results are provided in Figure 3.4. From Figure 3.3, it can be seen that some of the functions discriminate stronger than others. The obtained results are quite similar for both the two- and the three-dimensional lattices. In all test cases, the conventional energy function of the HP model, D85, achieved the lowest RE values. This confirms the poor discrimination capabilities of this function, which has been the main factor motivating the exploration of alternative approaches. Among the alternative functions, C04

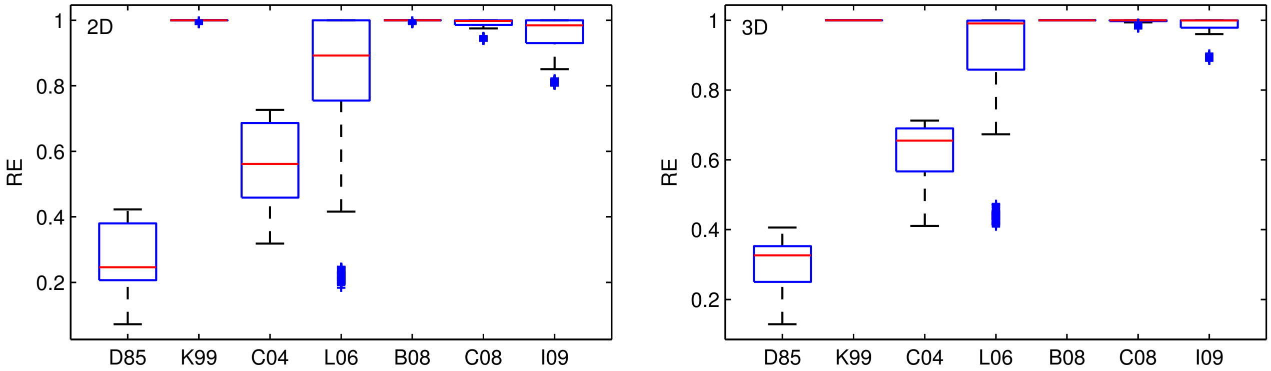


Figure 3.3: Relative entropy (RE) of the distribution of ranks obtained by using the different energy functions analyzed. Overall statistics for all two- (left) and three-dimensional (right) test cases.

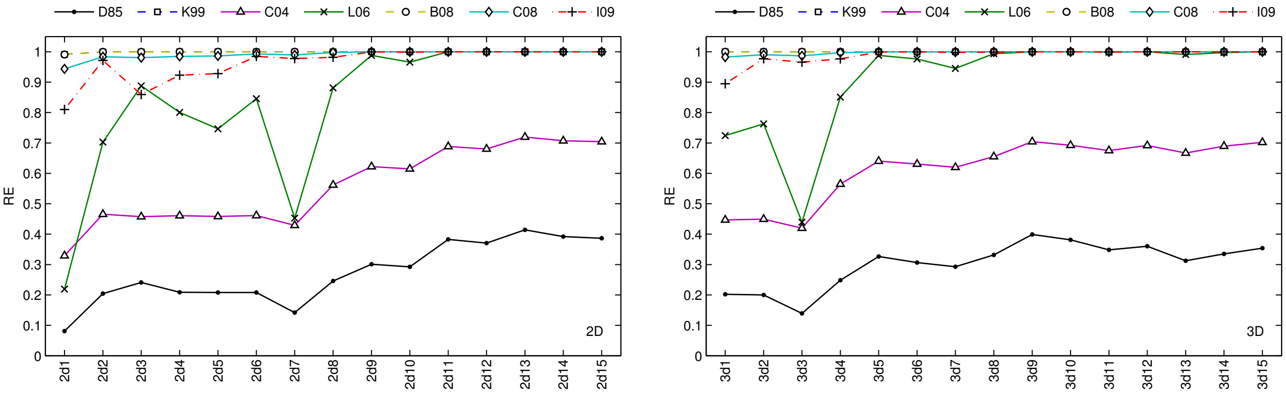


Figure 3.4: Relative entropy (RE) obtained by the different energy functions analyzed. Average of 100 independent repetitions for all two- (left) and three-dimensional (right) test instances.

presented the worst performance in terms of discrimination. Function L06 reached high RE values most of the time. However, this function presented a moderate discrimination for the shortest test sequences (see Figure 3.4). Regarding function I09, it is possible to note that the obtained RE values were almost always above 0.9, which indicates a strong discrimination. Finally, it is important to remark the high degree of discrimination provided by functions K99, B08 and C08. Functions K99 and B08 are the most discriminating functions according to the obtained results, followed by C08 which suffered slight decreases for some of the test instances.

The above results can be better understood by analyzing the histograms with the distribution of ranks achieved by each of the studied energy functions. Figure 3.5 presents such histograms for a single repetition of this experiment regarding sequence 2d4 on the two-dimensional square

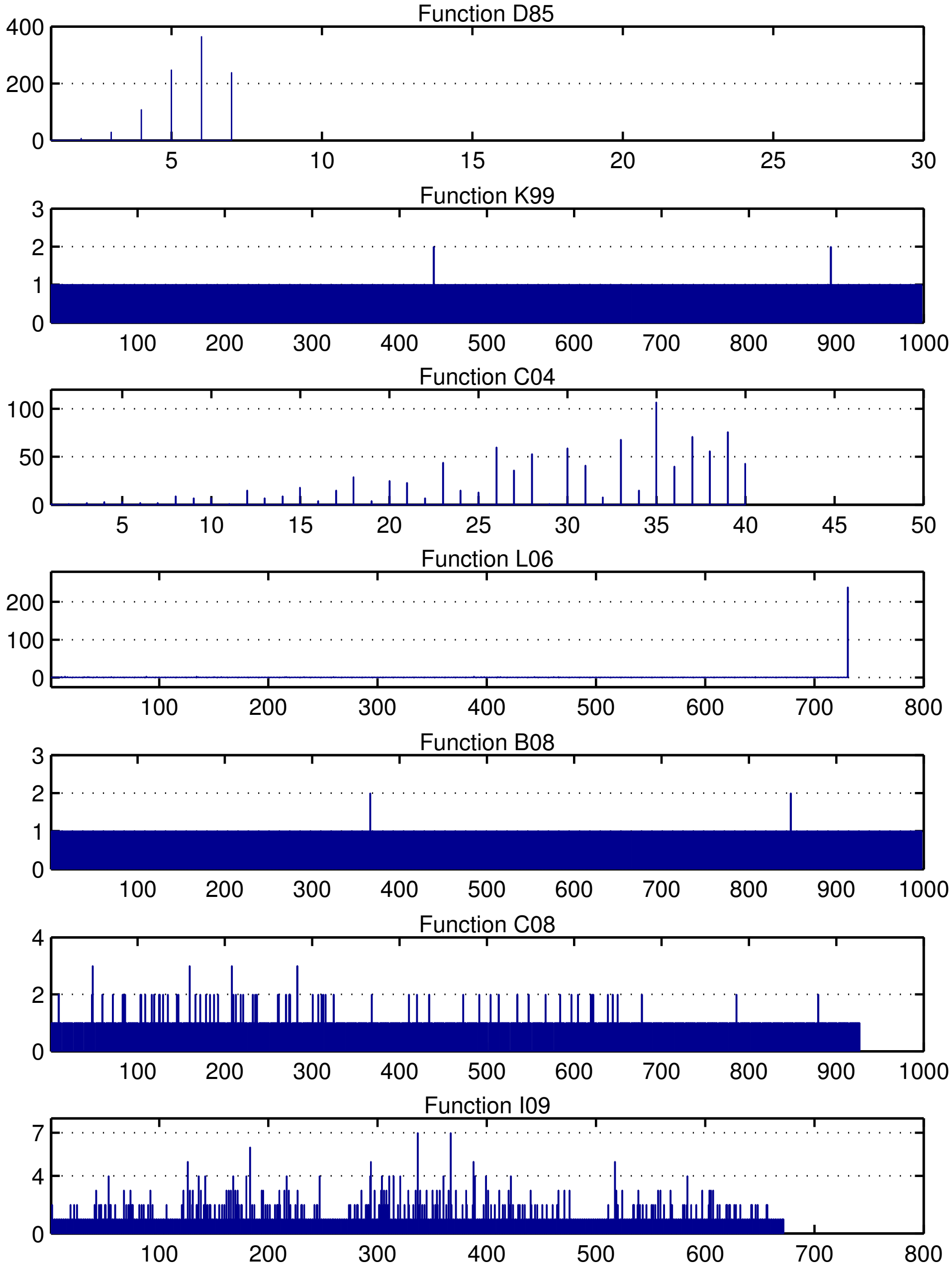


Figure 3.5: Density of the distribution of ranks achieved by the studied energy functions. Results for a single repetition, sequence 2d4, two-dimensional square lattice.

lattice (similar results were observed for other test instances). From Figure 3.5, it is possible to note how poor the distribution of ranks achieved by function D85 is. Only seven different ranking positions were induced to classify the $1,000$ generated conformations. It can be seen that there are almost 400 conformations sharing the sixth rank. As stated in the preamble of this chapter, using function D85 there can be only $|E^*| + 1$ different energy levels. Therefore, no matter the amount of generated conformations, the maximum number of ranks which can be assigned through function D85 is 10, since $E^* = -9$ for this benchmark sequence (2d4). The second worst scenario is presented by function C04, where only 40 different ranking positions were produced, out of which one was assigned to more than 100 conformations. Functions L06 and I09 enabled a more fine-grained discrimination, since about 730 and 680 different ranking positions were occupied to classify the totality of conformations, respectively. In the case of function I09, a maximum of seven conformations were assigned to the same rank. On the other hand, the histogram for function L06 presents a high peak indicating that there are about 250 equally ranked conformations. Function L06 is defined as the product of three terms, out of which one corresponds to the number of H-H topological contacts, $HHtc$ (see Section 3.2.3). Thus, all conformations for which $HHtc = 0$ will share the same energy value, $E_{L06} = 0$. This can be seen as a drawback; function L06 will not be able to discriminate among these conformations even if some of them present better characteristics than the others. Finally, the histograms for K99, B08 and C08 confirm the high degree of discrimination that these functions provide. Function C08 allowed roughly 930 different ranking positions to be assigned. K99 and B08 exhibited the strongest discrimination among all the studied energy functions. The corresponding histograms for these functions reveal that almost every conformation was mapped to a different ranking position. Only a few ranks were assigned to at most two conformations.

3.3.2 HP-compatibility

Alternative energy functions for the HP model are used in order to perform a more effective exploration through the space of potential protein conformations. Nevertheless, as stated at the beginning of Section 3.3, these functions need to remain consistent with the original optimization objective of the HP model of the PSP problem. This original objective consists in minimizing the conventional energy function, here referred to as function D85, by maximizing the total number of H-H topological contacts, $HHtc$ (see Section 2.3.3). Therefore, an important issue to be investigated is whether or not these alternative energy formulations are consistent with such an original objective.

The alternative energy functions should not contradict the conventional function D85 at the time of discriminating among potential conformations. Otherwise, the search process could be oriented towards solutions which differ from the original optima in the HP model (false optima could potentially be introduced). In this study, functions that meet this requirement, *i.e.*, not contradicting function D85, are said to feature the *HP-compatibility* property or, in other words, they are *HP-compatible*. Thus, HP-compatibility can be defined as the capability of an alternative energy function to preserve the conventional rank ordering among potential protein conformations. More formally:[7]

> An alternative energy function $\hat{E} : \mathcal{X}_{\mathcal{F}} \to \mathbb{R}$ is said to be HP-compatible if and only if $\hat{E}(\mathbf{x}_1) < \hat{E}(\mathbf{x}_2) \Rightarrow E(\mathbf{x}_1) \leq E(\mathbf{x}_2)$ for every pair of conformations $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}_{\mathcal{F}}$. Otherwise, if there exists at least a pair of conformations $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}_{\mathcal{F}}$ such that $E(\mathbf{x}_1) < E(\mathbf{x}_2)$ but $\hat{E}(\mathbf{x}_1) > \hat{E}(\mathbf{x}_2)$, then function $\hat{E}$ is not HP-compatible.

Note that the case where $E(\mathbf{x}_1) = E(\mathbf{x}_2)$ but $\hat{E}(\mathbf{x}_1) \neq \hat{E}(\mathbf{x}_2)$ is not considered a contradiction. This is a convenient scenario, since the aim of using the alternative function $\hat{E}$ is precisely to enable a more fine-grained discrimination than that achieved through the conventional function E.

In this section, the HP-compatibility property is explored for all the alternative energy functions considered. An experiment was conducted where $1,000$ different feasible structures were generated at random, and all pairwise comparisons among them were performed. The percentage of such comparisons where the alternative function agrees with (does not contradict) the conventional one was computed. The resulting measure is to be referred to as the *relative compatibility* (RC). Although a value of RC $= 100\%$ does not guarantee the HP-compatibility property for a given function, RC $< 100\%$ is enough to disprove it. To some extent, by focusing on the RC values it is possible to inquire into the severity of the cases where the HP-compatibility property is not satisfied. For all the selected instances, 100 repetitions of this experiment were performed. The average RC obtained for each instance is depicted in Figure 3.6. Figure 3.7 provides the overall statistics of this experiment.

From Figures 3.6 and 3.7, it is possible to note that functions K99 and I09 showed 100% of agreement with the conventional HP energy function for all the instances of this experiment. These

[7] By convention, this definition assumes that lower energy values correspond to higher quality conformations. In this definition, E denotes the conventional energy function (D85) of the HP model described in Section 2.3.3.1.

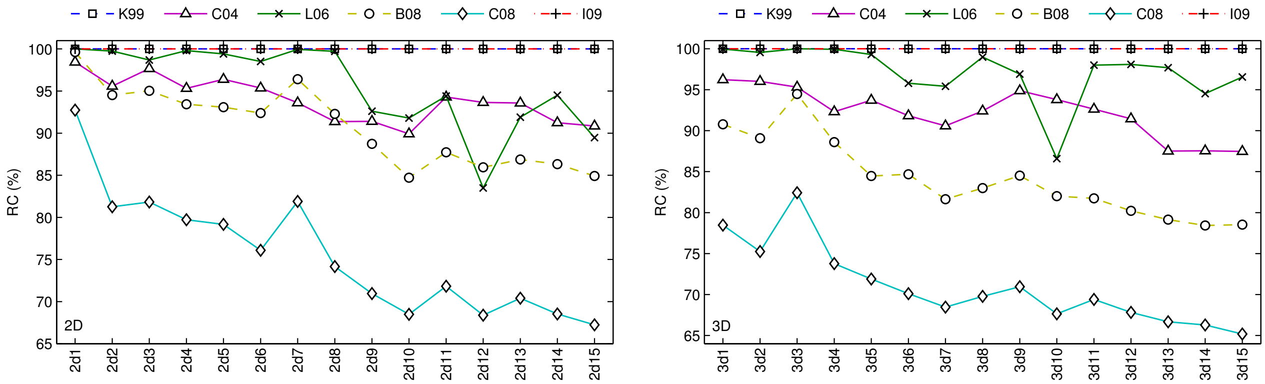


Figure 3.6: Relative compatibility (RC) obtained by each of the alternative energy functions analyzed. Average results for all two-dimensional (left) and three-dimensional (right) test instances.

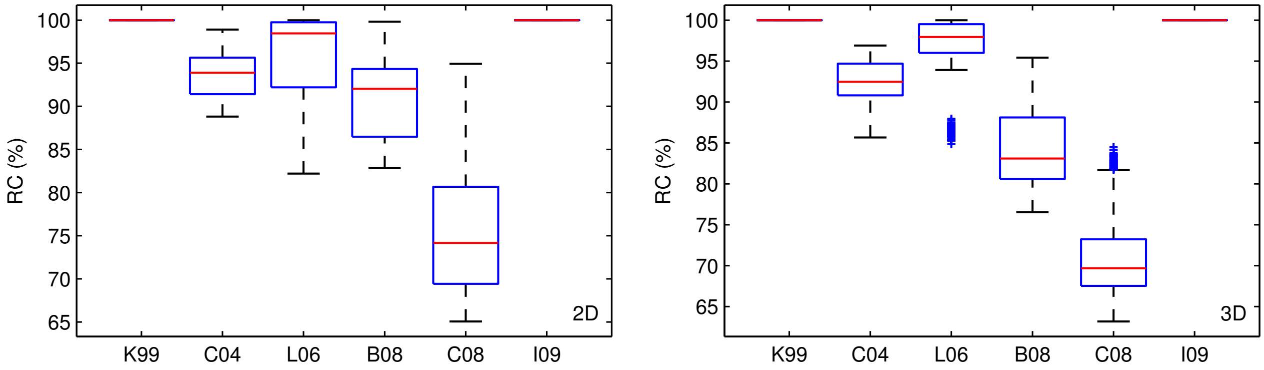


Figure 3.7: Relative compatibility (RC) obtained by each of the alternative energy functions analyzed. Overall statistics for all two-dimensional (left) and three-dimensional (right) test cases.

results suggest, but do not ensure, that functions K99 and I09 are HP-compatible. On the other hand, the obtained results reveal that functions C04, L06, B08 and C08 do not present the HP-compatibility property, which becomes more evident with the increasing problem size. Function L06 scored very competitive results for the shortest two- and three-dimensional test sequences. However, its performance declined for the largest test cases, especially when facing sequences 2d12 and 3d10. The average RC values obtained by L06 were almost always above 95%. The performance of function C04 gradually decreased as the problem size increased. The RC values achieved by this approach ranged from 90% to 95% most of the time. Function B08 presented the second worst overall behavior in this experiment. In the two-dimensional instances, the performance of B08 was above RC $= 90\%$ for the shortest sequences but at around 85% for the largest ones. Regarding the three-dimensional instances, function B08 obtained RC values below 85% in most of the cases.

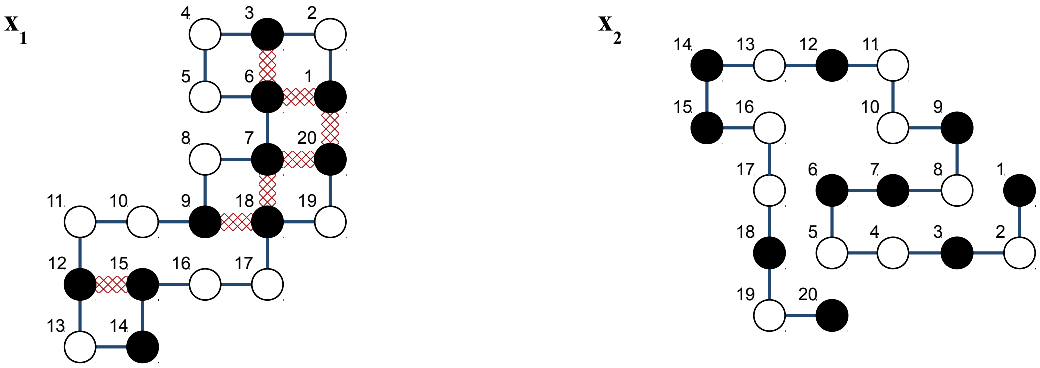


Figure 3.8: Two conformations $\mathbf{x}_1$ and $\mathbf{x}_2$ for sequence 2d4 on the two-dimensional square lattice. This figure illustrates the situation where function C08 contradicts function D85; that is, $[E(\mathbf{x}_1) = -7] < [E(\mathbf{x}_2) = 0]$ but $[E_{C08}(\mathbf{x}_1) = 5548] > [E_{C08}(\mathbf{x}_2) = 5308]$.

Finally, it can be highlighted the poor performance exhibited by function C08. This approach achieved the lowest RC values for all the adopted test cases. The average RC obtained by function C08 was roughly 75% for two-dimensional benchmarks, while it was at about 70% for the three-dimensional cases. Figure 3.8 presents an example scenario where function C08 contradicts the conventional function D85. In this example, a couple of two-dimensional conformations $\mathbf{x}_1$ and $\mathbf{x}_2$ for sequence 2d4 are compared with respect to each other by using functions D85 and C08. As a result, the conventional energy function D85 prefers conformation $\mathbf{x}_1$ (with $HHtc = 7$) to $\mathbf{x}_2$ (with $HHtc = 0$), while function C08 induces the opposite order of preference between these solutions.

The low RC values obtained by some functions, particularly C08, suggest serious implications. The lower the RC value, the more likely that the global optimum induced by the alternative function differs from the global optimum of the original problem. Therefore, alternative functions which are not HP-compatible cannot be expected to steer the search process in an effective manner.

3.3.3 Search performance using a basic local search algorithm

A *best improvement local search* (BILS) algorithm was implemented in order to evaluate the effectiveness of the studied energy functions at guiding the search process, see Algorithm 1. BILS starts with a randomly generated conformation, denoted by $\mathbf{x}$. In a greedy manner, $\mathbf{x}$ is iteratively replaced by the best among all the improving conformations defined in the neighborhood of $\mathbf{x}$, $\mathcal{N}(\mathbf{x})$.

The search process stops when, given the current conformation $\mathbf{x}$ and the adopted neighborhood structure, it is not possible to achieve an improvement, *i.e.*, $\mathbf{x}$ is locally optimal.

Algorithm 1 Best improvement local search (BILS) algorithm.

```
$choose\ \mathbf{x} \in \mathcal{X}_{\mathcal{F}}\ uniformly\ at\ random$
repeat
  $\mathbf{x} \leftarrow Best\_Improvement(\mathcal{N}(\mathbf{x}))$
until $no\ improvement\ is\ possible$
```

As stated at the beginning of Section 3.3, only solutions encoding feasible conformations are considered in this study. Hence, the initial solutions for the BILS algorithm were generated using the backtracking procedure proposed in [42]. The implemented neighborhood structure $\mathcal{N}(\mathbf{x})$ is defined by all feasible conformations which can be reached through single 1-variable perturbations of $\mathbf{x}$; *i.e.*, $\mathcal{N}(\mathbf{x}) = \{\mathbf{x}' \in \mathcal{X}_{\mathcal{F}} \mid H_d(\mathbf{x}, \mathbf{x}') = 1\}$, where $H_d(\mathbf{x}, \mathbf{x}')$ denotes the Hamming distance between $\mathbf{x}$ and $\mathbf{x}'$. Given a protein sequence of length ℓ, the size of such a neighborhood is $|\mathcal{N}(\mathbf{x})| = 3(\ell - 1)$ in the two-dimensional square lattice and $|\mathcal{N}(\mathbf{x})| = 5(\ell - 1)$ for the three-dimensional case.

The motivation for using such a simple BILS algorithm is as follows. On the one hand, BILS seems to be a suitable algorithm for analyzing the impact of varying the evaluation scheme. Once the neighborhood structure has been defined, the behavior and performance of the algorithm will be mainly determined by the discrimination capabilities of the different energy functions. *"A local search is effective if it is able to find good local minima"* [16]. BILS stops at a local optimum, and the characteristics of such a local optimum will depend on the used discrimination method. Moreover, due to the low degree of discrimination provided by some of the functions, the search process can be expected to stop early (after a reduced number of iterations). On the other hand, no additional parameters of the algorithm have to be adjusted, which avoids affecting (neither negatively nor positively) the behavior induced by the studied energy functions through parameter settings.

The behavior of the BILS algorithm was evaluated when using each of the different studied energy functions. Figure 3.9 presents the results obtained for all the two-dimensional instances, while the results for the three-dimensional case are provided in Figure 3.10. Plots in these figures show the average number of H-H topological contacts ($HHtc$) achieved by the BILS algorithm as the search

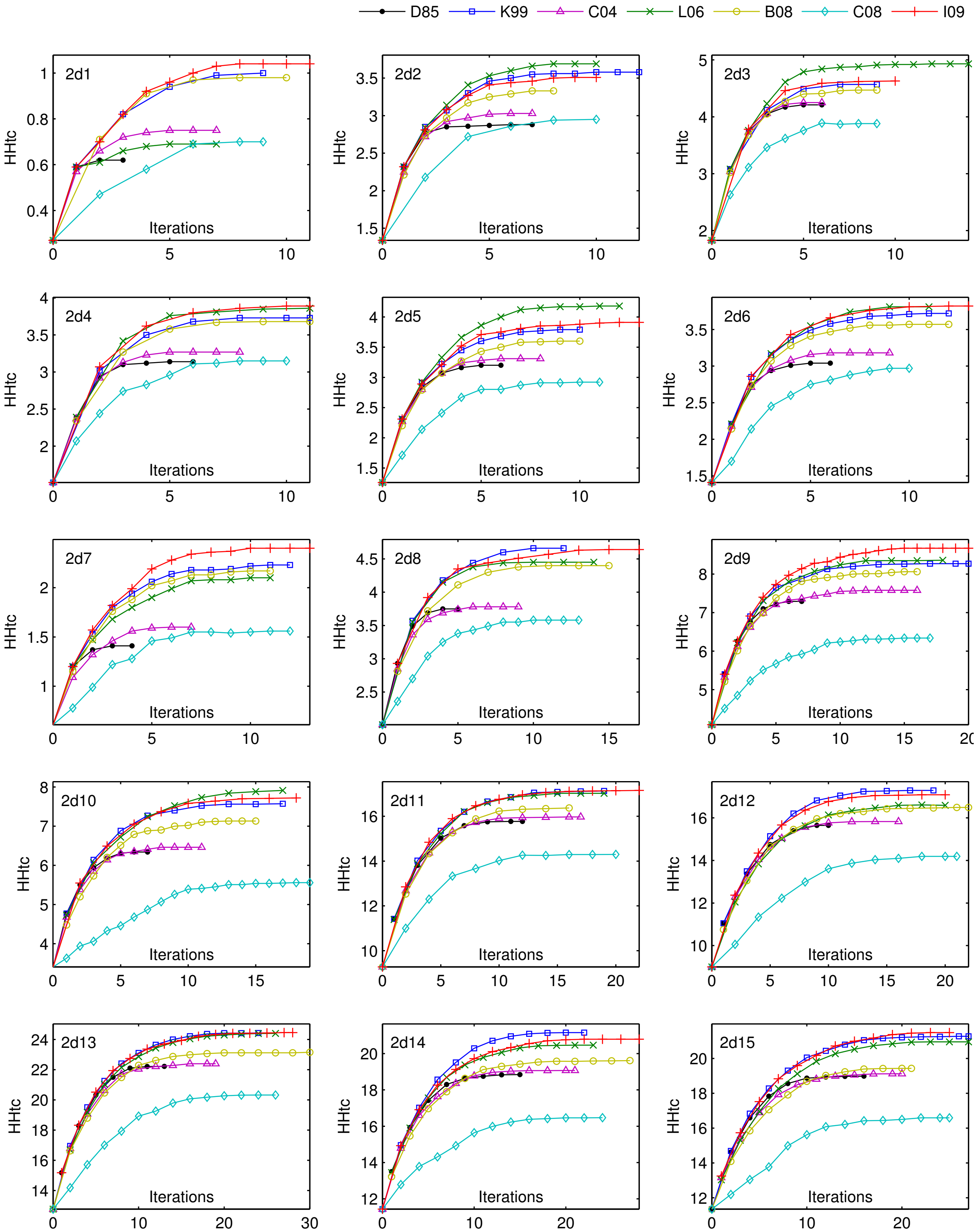


Figure 3.9: Results of BILS when using the studied energy functions. Number of H-H topological contacts ($HHtc$) obtained at each iteration. Average of 100 independent runs. Each plot presents the results for a particular two-dimensional instance. Legend is provided at the top of the figure.

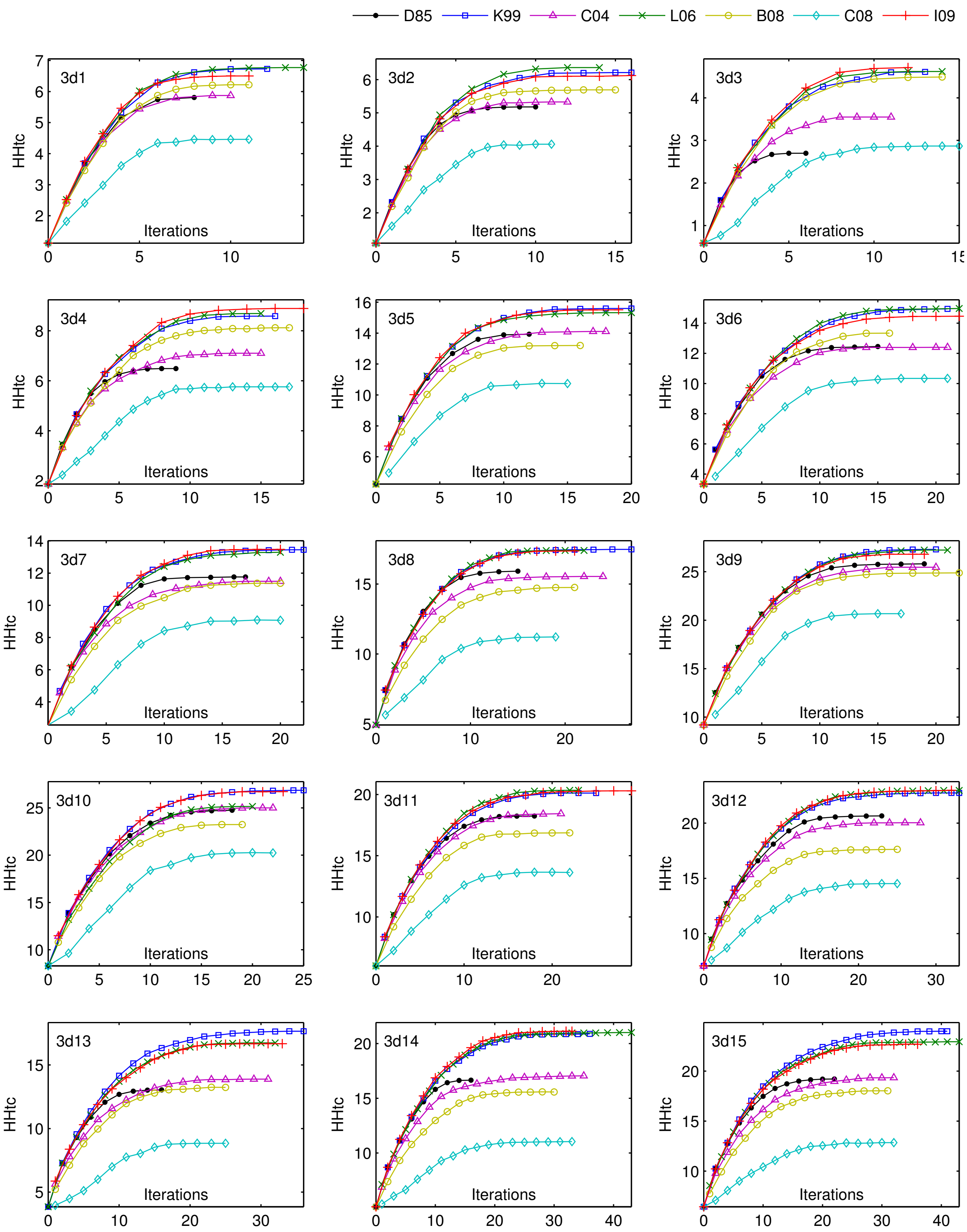


Figure 3.10: Results of BILS when using the studied energy functions. Number of H-H topological contacts ($HHtc$) obtained at each iteration. Average of 100 independent runs. Each plot presents the results for a particular three-dimensional instance. Legend is provided at the top of the figure.

progressed (at each iteration), for each considered test case. These results were computed from a total of 100 independent executions performed for each configuration of the experiment. From Figures 3.9 and 3.10 it is possible to derive some general conclusions. As expected, the conventional energy function D85 presented a limited performance for this experiment. For all the test instances (except for sequence 3d9), the algorithm reached the lowest number of iterations due to the poor discrimination that function D85 provides. In most cases, however, the poorest performance of the algorithm was obtained when using function C08. Although functions B08 and C04 behaved better than function D85 in most of the two-dimensional instances, these functions reported a poorer search performance than D85 for some of the three-dimensional test cases. Function L06 obtained very competitive results most of the time. L06 allowed the algorithm to score the highest $HHtc$ values for some of the test cases (*e.g.*, 2d3, 2d5, 2d10, 3d2), while showing a slightly inferior performance for some other instances (*e.g.*, 2d1, 2d7, 3d10). Finally, it is possible to highlight the promising behavior that functions I09 and K99 consistently exhibited for all the considered test cases.

More detailed information and the results of the statistical significance analysis are provided in Tables 3.1 and 3.2. For the different analyzed energy functions and all the adopted test cases, these tables detail the best obtained energy value (E_b), the number of times that this solution was found (ν), and the arithmetic mean ($\bar{E}$). Also, the obtained values for the overall relative root mean square error (O-RMSE) measure are presented at the bottom of the tables. In these tables, values **marked** $+$ highlight a statistically significant increase in performance achieved by the alternative energy function with regard to the conventional function D85. Conversely, values **marked** $-$ indicate that a statistically significant performance decrease was obtained as a consequence of using the alternative formulation. In addition, the best average performance (lowest average energy) for each of the instances and the best (lowest) O-RMSE value have been **shaded** in these tables. Tables 3.1 and 3.2 confirm the superiority that functions K99, I09 and L06 have shown in this experiment. In the vast majority of the instances, it can be seen from the tables that functions K99, I09 and L06 significantly improved the performance of the BILS algorithm with respect to the conventional function D85. There were no significant differences between functions D85 and C04 except for sequences 3d3 and 3d4, in both cases favoring C04. Function B08 scored significantly better results

Table 3.1: Detailing the performance results obtained by the BILS algorithm when using the seven studied energy formulations for the HP model. Two-dimensional test cases.

	D85		**K99**		**C04**		**L06**		**B08**		**C08**		**I09**	
Seq.	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$
2d1	-3 (1)	-0.6	-3 (2)	**-1.0+**	-3 (1)	-0.8	-3 (2)	-0.7	-3 (2)	**-1.0+**	-2 (12)	-0.7	-3 (3)	**-1.0+**
2d2	-7 (1)	-2.9	-7 (3)	**-3.6+**	-7 (1)	-3.0	-7 (3)	**-3.7+**	-7 (3)	**-3.3+**	-7 (1)	-3.0	-7 (2)	**-3.5+**
2d3	-8 (1)	-4.2	-8 (1)	-4.6	-8 (1)	-4.2	-8 (1)	**-4.9+**	-7 (4)	-4.5	-7 (1)	-3.9	-8 (1)	**-4.6+**
2d4	-7 (1)	-3.1	-7 (3)	**-3.7+**	-7 (1)	-3.3	-7 (2)	**-3.9+**	-7 (3)	**-3.7+**	-7 (1)	-3.1	-7 (3)	**-3.9+**
2d5	-6 (7)	-3.2	-7 (1)	**-3.8+**	-6 (7)	-3.3	-7 (4)	**-4.2+**	-7 (1)	-3.6	-6 (6)	-2.9	-7 (4)	**-3.9+**
2d6	-7 (2)	-3.0	-7 (1)	**-3.7+**	-7 (1)	-3.2	-7 (1)	**-3.8+**	-7 (1)	**-3.6+**	-6 (2)	-3.0	-7 (2)	**-3.8+**
2d7	-5 (1)	-1.4	-7 (1)	**-2.2+**	-6 (1)	-1.6	-7 (1)	**-2.1+**	-7 (1)	**-2.2+**	-7 (1)	-1.6	-7 (1)	**-2.4+**
2d8	-7 (2)	-3.8	-8 (1)	**-4.7+**	-7 (2)	-3.8	-7 (6)	**-4.5+**	-7 (7)	**-4.4+**	-7 (4)	-3.6	-9 (1)	**-4.6+**
2d9	-12 (3)	-7.3	-12 (4)	**-8.3+**	-12 (2)	-7.6	-13 (1)	**-8.4+**	-12 (4)	**-8.1+**	-11 (7)	**-6.3−**	-15 (1)	**-8.7+**
2d10	-10 (6)	-6.3	-13 (1)	**-7.6+**	-11 (1)	-6.5	-12 (3)	**-7.9+**	-13 (1)	**-7.1+**	-11 (1)	**-5.6−**	-13 (2)	**-7.7+**
2d11	-22 (2)	-15.8	-24 (3)	**-17.1+**	-22 (1)	-16.0	-24 (3)	**-17.0+**	-25 (1)	-16.4	-24 (1)	**-14.3−**	-25 (1)	**-17.1+**
2d12	-22 (1)	-15.7	-24 (1)	**-17.3+**	-22 (1)	-15.8	-22 (2)	**-16.6+**	-21 (4)	**-16.5+**	-22 (1)	**-14.2−**	-23 (1)	**-17.1+**
2d13	-30 (2)	-22.2	-35 (1)	**-24.4+**	-30 (3)	-22.4	-35 (1)	**-24.4+**	-31 (2)	-23.1	-33 (1)	**-20.3−**	-35 (1)	**-24.4+**
2d14	-28 (1)	-18.8	-30 (1)	**-21.1+**	-26 (4)	-19.1	-29 (1)	**-20.5+**	-28 (1)	-19.6	-25 (1)	**-16.5−**	-29 (1)	**-20.8+**
2d15	-26 (2)	-19.0	-28 (3)	**-21.3+**	-29 (1)	-19.1	-28 (1)	**-20.9+**	-26 (1)	-19.4	-22 (5)	**-16.6−**	-27 (3)	**-21.5+**
O-RMSE	67.33%		61.47%		66.26%		61.52%		63.30%		69.58%		**60.92%**	

Table 3.2: Detailing the performance results obtained by the BILS algorithm when using the seven studied energy formulations for the HP model. Three-dimensional test cases.

	D85		**K99**		**C04**		**L06**		**B08**		**C08**		**I09**	
Seq.	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$
3d1	-10 (2)	-5.8	-11 (2)	**-6.7+**	-10 (1)	-5.9	-11 (1)	**-6.8+**	-11 (1)	-6.2	-9 (2)	**-4.5−**	-10 (4)	**-6.5+**
3d2	-9 (4)	-5.2	-11 (1)	**-6.2+**	-10 (1)	-5.3	-10 (3)	**-6.4+**	-9 (6)	**-5.7+**	-7 (7)	**-4.1−**	-10 (1)	**-6.1+**
3d3	-7 (2)	-2.7	-9 (1)	**-4.6+**	-7 (3)	**-3.5+**	-8 (2)	**-4.6+**	-9 (1)	**-4.5+**	-7 (3)	-2.9	-8 (2)	**-4.7+**
3d4	-12 (2)	-6.5	-13 (2)	**-8.6+**	-13 (1)	**-7.1+**	-14 (1)	**-8.7+**	-14 (2)	**-8.1+**	-13 (1)	**-5.8−**	-15 (1)	**-8.9+**
3d5	-22 (1)	-13.9	-23 (1)	**-15.6+**	-21 (1)	-14.1	-22 (2)	**-15.3+**	-22 (1)	-13.2	-17 (2)	**-10.7−**	-22 (1)	**-15.5+**
3d6	-19 (4)	-12.4	-22 (1)	**-14.9+**	-19 (1)	-12.4	-21 (3)	**-15.0+**	-19 (4)	**-13.3+**	-18 (1)	**-10.3−**	-21 (3)	**-14.5+**
3d7	-18 (2)	-11.8	-20 (2)	**-13.4+**	-18 (1)	-11.5	-22 (1)	**-13.3+**	-17 (2)	-11.4	-17 (1)	**-9.1−**	-18 (4)	**-13.5+**
3d8	-23 (2)	-15.9	-25 (1)	**-17.5+**	-22 (2)	-15.6	-24 (1)	**-17.4+**	-23 (1)	**-14.8−**	-19 (1)	**-11.2−**	-24 (2)	**-17.4+**
3d9	-36 (2)	-25.8	-38 (4)	**-27.3+**	-36 (1)	-25.5	-36 (3)	**-27.2+**	-36 (1)	-24.9	-32 (1)	**-20.7−**	-38 (2)	-26.8
3d10	-34 (1)	-24.8	-38 (1)	**-26.9+**	-36 (1)	-25.0	-35 (2)	-25.1	-33 (1)	**-23.2−**	-33 (1)	**-20.2−**	-37 (1)	**-26.7+**
3d11	-28 (2)	-18.2	-31 (1)	**-20.1+**	-26 (2)	-18.4	-29 (1)	**-20.3+**	-27 (1)	**-16.9−**	-25 (1)	**-13.6−**	-31 (2)	**-20.3+**
3d12	-29 (5)	-20.6	-31 (6)	**-22.7+**	-30 (1)	-20.0	-33 (1)	**-22.9+**	-27 (2)	**-17.6−**	-22 (1)	**-14.5−**	-34 (2)	**-23.0+**
3d13	-22 (1)	-13.0	-28 (1)	**-17.6+**	-22 (1)	-13.9	-24 (4)	**-16.7+**	-22 (2)	-13.2	-17 (1)	**-8.8−**	-24 (3)	**-16.7+**
3d14	-24 (2)	-16.6	-30 (1)	**-20.9+**	-29 (1)	-17.1	-32 (1)	**-21.0+**	-26 (1)	**-15.6−**	-20 (1)	**-11.0−**	-34 (1)	**-21.1+**
3d15	-30 (1)	-19.2	-36 (1)	**-24.0+**	-30 (1)	-19.3	-35 (2)	**-22.9+**	-28 (2)	-18.0	-21 (2)	**-12.8−**	-32 (3)	**-22.7+**
O-RMSE	65.61%		**58.87%**		64.54%		59.17%		64.34%		73.09%		59.22%	

than function D85 in 9 out of the 15 two-dimensional instances, and 4 of the three-dimensional instances. Note, however, that this function was significantly outperformed by function D85 in 5 of the largest three-dimensional test cases. Finally, it can also be confirmed the poor performance presented by function C08. Function C08 performed significantly worse than function D85 for the largest two-dimensional instances and for all but one of the three-dimensional cases.

3.3.4 Search performance using the iterated local search algorithm

In Section 3.3.3, a basic local search algorithm was employed as a first step in analyzing the effectiveness of the studied energy functions at guiding the search process. Through local search it is possible to converge towards local optima. However, the performance of these algorithms is usually unsatisfactory in terms of finding global optimum solutions [16, 218]. Therefore, it is required to implement additional strategies to foster exploration and to allow the search process to escape from local optima. A possible strategy consists in iteratively applying local search each time starting from a different initial solution, such as it is done in the iterated local search (ILS) algorithm [146,147,157].

In this section, a basic ILS algorithm is used for inquiring into the suitability of the studied energy functions (outlined in Algorithm 2). The ILS algorithm starts with a feasible conformation generated at random[8], denoted as $\mathbf{x}$. Then, a local search strategy (embedded heuristic) is applied to $\mathbf{x}$ until a local optimum $\mathbf{x}^*$ is found. At each iteration, a perturbation $\mathbf{x}'$ of the current local optimum $\mathbf{x}^*$ is obtained and used as a starting point of another round of local search. The new local optimum solution found $\mathbf{x}'^*$ may be accepted as the new incumbent solution $\mathbf{x}^*$, based on a given acceptance criterion. This iterative procedure is repeated until a given stop condition is met.

In order to implement the ILS algorithm, three basic components have to be defined, namely, the embedded local search heuristic, the perturbation strength and the acceptance criterion. In this study, these components are defined as follows. The best improvement local search (BILS) algorithm, as described and implemented in Section 3.3.3, is adopted as the embedded heuristic. Six different values for the perturbation strength are considered: $\{2, 3, 4, 6, 8, 10\}$. Here, the perturbation

[8] Initial feasible solutions were generated using the backtracking algorithm proposed in [42].

Algorithm 2 Iterated local search (ILS) algorithm.

```
$choose\ \mathbf{x} \in \mathcal{X}_{\mathcal{F}}\ uniformly\ at\ random$
$\mathbf{x}^* \leftarrow LocalSearch(\mathbf{x})$
repeat
    $\mathbf{x}' \leftarrow Perturbation(\mathbf{x}^*)$
    $\mathbf{x}'^* \leftarrow LocalSearch(\mathbf{x}')$
    $\mathbf{x}^* \leftarrow AcceptanceCriterion(\mathbf{x}^*, \mathbf{x}'^*)$
until $< stop\ condition >$
```

strength refers to the number of encoding positions in the conformation which are to be affected by the perturbation. Three different acceptance criteria are explored: (i) IMP, the new local optimum $\mathbf{x}'^*$ is accepted if it has a better energy value than the incumbent solution $\mathbf{x}^*$; (ii) IEQ, the new local optimum $\mathbf{x}'^*$ is accepted if it is at least as good as the incumbent solution $\mathbf{x}^*$; and (iii) ALL, the new local optimum $\mathbf{x}'^*$ is always accepted. The three different acceptance criteria, together with the six considered values for the perturbation strength, lead to a total of 18 parameter configurations of the ILS. All these parameter configurations were evaluated in order to identify the most appropriate conditions for the compared approaches. In all the cases, the algorithm was allowed to run until a maximum number of 5×10^5 solution evaluations was reached, and a total of 50 independent executions were performed. Figure 3.11 presents the overall relative root mean square error (O-RMSE) obtained by the studied energy functions for the different parameter settings of the ILS.

Among the alternative energy functions, Figure 3.11 shows that K99, L06 and I09 consistently competed at the top of the ranking for the different parameter configurations of the ILS. In the two-dimensional case, the performance of function B08 was competitive for most of the ILS configurations. In contrast, this function exhibited a low performance in all cases when facing the three-dimensional instances. Function C08 obtained the worst (higher) O-RMSE values in most of the cases, followed by function C04. Functions C08 and C04 are thus the worst performers of this experiment. Regarding the conventional energy function D85, an interesting behavior can be observed when comparing the results obtained using the different acceptance criteria. While the ranking among the alternative energy functions remains consistent in most of the cases from one acceptance criterion to another, there was a significant increase in the performance of function D85 when using the IEQ acceptance

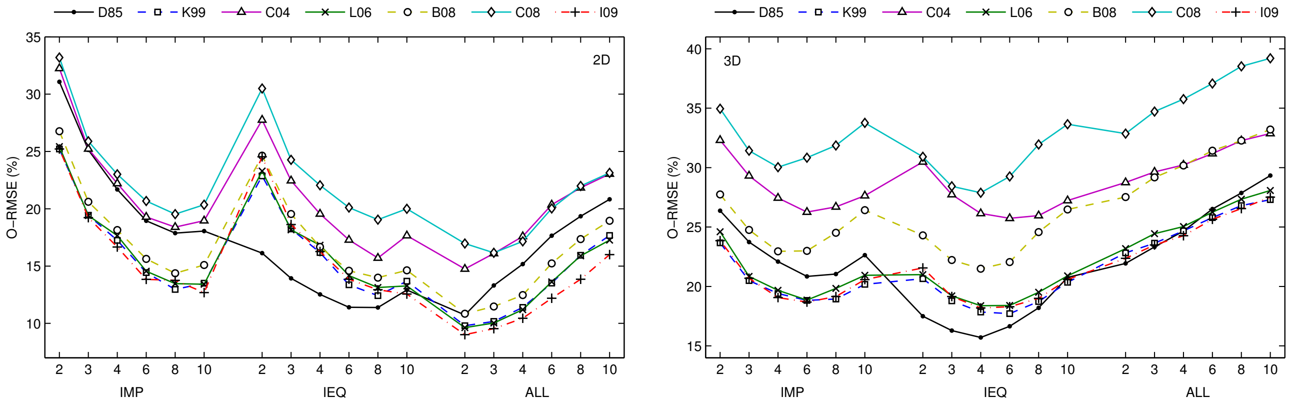


Figure 3.11: Overall relative root mean square error (O-RMSE) obtained for all parameter configurations of the ILS algorithm. Two-dimensional (left) and three-dimensional test cases.

criterion. The IEQ acceptance criterion allowed the algorithm to exploit the low discrimination associated with function D85 as a means of enhancing exploration and escaping from local optima.

In order to provide a more detailed analysis, the parameter adjustment which allowed each of the studied energy functions to reach the lowest O-RMSE value has been selected.[9] Tables 3.3 and 3.4 detail the obtained results for all two-dimensional and three-dimensional test cases. The information in these tables is organized as described in Section 3.3.3 with regard to Tables 3.1 and 3.2. From Table 3.3, it is possible to observe that function I09 reached the lowest average energy for 11 (73.33%) out of the 15 two-dimensional instances, obtaining the best O-RMSE value. In 5 of the instances, the improvements obtained by function I09 were statistically significant with respect to the conventional energy function D85. The second best performer was function K99, which showed the best average performance for 7 of the instances and significantly improved the results of function D85 in 3 other cases. Function L06 achieved significantly better results than function D85 for 5 of the instances, but there was a significant difference against function L06 in 4 of the largest test cases. Slightly similar results were obtained by function B08. Although the conventional function D85 does not present a remarkable performance, the results of this function are still considered competitive. Finally, the poorest performance was obtained by functions C04 and

[9] For the two-dimensional instances, the ALL acceptance criterion and a perturbation strength of 2 were chosen for all the studied approaches. In the three-dimensional case, the IEQ acceptance criterion was selected for all the energy functions. A perturbation strength of 4 was used for all functions except C04, for which this parameter was set to 6.

Table 3.3: Detailing the results obtained by the ILS algorithm when using the seven studied energy formulations for the HP model. Two-dimensional test cases.

	D85		**K99**		**C04**		**L06**		**B08**		**C08**		**I09**	
Seq.	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$
2d1	-4 (50)	**-4.0**	-4 (50)	**-4.0**	-4 (49)	**-4.0**	-4 (50)	**-4.0**	-4 (50)	**-4.0**	-4 (50)	**-4.0**	-4 (50)	**-4.0**
2d2	-8 (50)	**-8.0**	-8 (50)	**-8.0**	-8 (49)	**-8.0**	-8 (50)	**-8.0**	-8 (50)	**-8.0**	-8 (49)	**-8.0**	-8 (50)	**-8.0**
2d3	-9 (44)	-8.9	-9 (47)	-8.9	-9 (46)	-8.9	-9 (50)	**-9.0+**	-9 (49)	**-9.0**	-9 (48)	**-9.0**	-9 (47)	-8.9
2d4	-9 (50)	**-9.0**	-9 (50)	**-9.0**	-9 (50)	**-9.0**	-9 (50)	**-9.0**	-9 (50)	**-9.0**	-9 (50)	**-9.0**	-9 (50)	**-9.0**
2d5	-10 (50)	**-10.0**	-10 (50)	**-10.0**	-10 (49)	**-10.0**	-10 (50)	**-10.0**	-10 (50)	**-10.0**	-10 (35)	**-9.7—**	-10 (50)	**-10.0**
2d6	-9 (47)	-8.9	-9 (50)	**-9.0**	-9 (42)	-8.8	-9 (50)	**-9.0**	-9 (50)	**-9.0**	-9 (44)	-8.9	-9 (50)	**-9.0**
2d7	-8 (36)	-7.7	-8 (47)	**-7.9+**	-8 (25)	**-7.5—**	-8 (50)	**-8.0+**	-8 (50)	**-8.0+**	-8 (50)	**-8.0+**	-8 (50)	**-8.0+**
2d8	-14 (2)	-12.3	-14 (2)	-12.4	-13 (1)	**-11.2—**	-14 (7)	**-12.9+**	-14 (9)	**-12.8+**	-14 (4)	-12.2	-14 (15)	**-13.0+**
2d9	-21 (3)	-18.8	-22 (1)	**-19.6+**	-20 (1)	**-17.4—**	-21 (4)	**-19.5+**	-21 (5)	-19.1	-21 (1)	**-17.5—**	-22 (1)	**-20.1+**
2d10	-20 (1)	-18.2	-21 (1)	-18.3	-19 (1)	**-16.8—**	-21 (1)	-18.4	-19 (1)	**-17.1—**	-17 (3)	**-15.3—**	-21 (1)	**-18.7+**
2d11	-33 (7)	-31.3	-34 (1)	**-31.5**	-33 (1)	**-29.3—**	-33 (2)	**-30.7—**	-33 (3)	-30.9	-32 (1)	**-27.5—**	-34 (2)	-31.0
2d12	-34 (1)	-30.5	-35 (2)	**-31.2+**	-34 (1)	**-29.3—**	-35 (1)	**-32.1+**	-33 (1)	**-29.3—**	-28 (8)	**-26.5—**	-35 (2)	**-32.2+**
2d13	-46 (3)	-42.6	-47 (1)	**-43.0**	-45 (1)	**-40.2—**	-46 (1)	**-41.9—**	-46 (1)	-42.2	-44 (1)	**-37.1—**	-46 (1)	-42.9
2d14	-42 (3)	**-38.6**	-41 (2)	-38.2	-38 (1)	**-34.4—**	-40 (3)	**-37.2—**	-40 (3)	**-37.2—**	-39 (1)	**-32.6—**	-41 (4)	-38.4
2d15	-42 (1)	-39.1	-42 (3)	-39.0	-39 (1)	**-35.1—**	-41 (1)	**-38.2—**	-41 (2)	**-37.0—**	-35 (1)	**-30.6—**	-42 (5)	**-39.3**
O-RMSE	10.73%		9.80%		14.75%		9.63%		10.84%		16.16%		**9.02%**	

Table 3.4: Detailing the results obtained by the ILS algorithm when using the seven studied energy formulations for the HP model. Three-dimensional test cases.

	D85		**K99**		**C04**		**L06**		**B08**		**C08**		**I09**	
Seq.	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$
3d1	-11 (50)	**-11.0**	-11 (50)	**-11.0**	-11 (50)	**-11.0**	-11 (50)	**-11.0**	-11 (50)	**-11.0**	-11 (50)	**-11.0**	-11 (50)	**-11.0**
3d2	-13 (50)	**-13.0**	-13 (50)	**-13.0**	-13 (37)	**-12.7—**	-13 (50)	**-13.0**	-13 (49)	**-13.0**	-13 (43)	**-12.8—**	-13 (50)	**-13.0**
3d3	-9 (50)	**-9.0**	-9 (50)	**-9.0**	-9 (41)	**-8.8—**	-9 (50)	**-9.0**	-9 (50)	**-9.0**	-9 (49)	**-9.0**	-9 (50)	**-9.0**
3d4	-18 (49)	**-18.0**	-18 (43)	**-17.8—**	-18 (8)	**-16.2—**	-18 (38)	**-17.6—**	-18 (27)	**-17.3—**	-18 (16)	**-16.8—**	-18 (37)	**-17.6—**
3d5	-33 (4)	**-31.1**	-32 (7)	**-30.1—**	-31 (3)	**-27.9—**	-34 (1)	**-30.3—**	-33 (1)	**-30.0—**	-30 (2)	**-26.8—**	-33 (1)	**-30.1—**
3d6	-31 (13)	**-29.6**	-31 (3)	**-28.8—**	-29 (5)	**-26.4—**	-31 (5)	**-29.0—**	-31 (2)	**-28.4—**	-31 (2)	**-27.7—**	-31 (7)	**-28.9—**
3d7	-32 (1)	**-29.2**	-32 (1)	**-28.5—**	-30 (1)	**-25.5—**	-32 (3)	**-28.2—**	-32 (1)	**-27.3—**	-31 (1)	**-23.8—**	-30 (8)	**-28.1—**
3d8	-40 (2)	**-36.2**	-39 (2)	-35.5	-36 (1)	**-31.2—**	-39 (1)	**-34.5—**	-40 (1)	**-33.5—**	-35 (1)	**-29.9—**	-40 (1)	**-35.1—**
3d9	-52 (1)	**-48.3**	-51 (3)	-48.0	-49 (1)	**-44.5—**	-50 (5)	**-47.2—**	-50 (3)	**-46.3—**	-49 (4)	**-44.8—**	-51 (4)	-47.6
3d10	-56 (1)	**-50.2**	-52 (6)	**-48.0—**	-49 (1)	**-43.7—**	-55 (1)	**-48.6—**	-52 (1)	**-45.6—**	-50 (2)	**-41.6—**	-54 (1)	**-49.2—**
3d11	-45 (1)	**-41.4**	-45 (1)	**-39.6—**	-40 (1)	**-35.4—**	-44 (2)	**-39.5—**	-43 (1)	**-37.1—**	-40 (1)	**-32.8—**	-45 (1)	**-39.5—**
3d12	-57 (2)	**-50.5**	-54 (2)	**-48.1—**	-50 (1)	**-41.4—**	-54 (1)	**-48.2—**	-50 (3)	**-43.0—**	-42 (1)	**-35.7—**	-54 (2)	**-48.1—**
3d13	-47 (1)	**-40.6**	-45 (1)	**-39.2—**	-40 (1)	**-31.8—**	-45 (2)	**-38.6—**	-42 (1)	**-36.4—**	-38 (1)	**-29.0—**	-44 (4)	**-39.0—**
3d14	-55 (1)	**-49.4**	-53 (2)	**-47.0—**	-49 (1)	**-38.2—**	-55 (1)	**-45.8—**	-47 (1)	**-39.5—**	-47 (1)	**-33.1—**	-57 (1)	**-46.4—**
3d15	-61 (1)	**-53.8**	-59 (1)	-52.0	-50 (1)	**-40.1—**	-58 (3)	**-50.0—**	-55 (1)	**-44.7—**	-48 (1)	**-37.0—**	-57 (1)	**-49.0—**
O-RMSE	**15.71%**		17.72%		25.74%		18.38%		21.49%		27.88%		18.19%	

C08, whose results were significantly worse than those of the conventional function D85 in most of the cases. A quite different scenario can be observed regarding the three-dimensional test cases. It can be seen from Table 3.4 that the conventional function D85 scored the best average performance for all the considered test cases. The statistical analysis indicates that function D85 significantly outperformed all the alternative energy formulations in the vast majority of the cases. Among the alternative functions, the best results were obtained by K99, followed by functions I09 and L06. Finally, the worst overall behavior was presented by functions B08, C04 and particularly C08.

The obtained results confirm that an effective evaluation scheme is essential in order to guide the search process towards high quality conformations. For most parameter configurations of the ILS algorithm, the best results were obtained using alternative energy functions which provide a fine-grained discrimination. Nevertheless, a particular acceptance criterion (IEQ, in this case) increased the performance of the ILS algorithm when using the conventional energy function, D85. Using such an acceptance criterion, the results of function D85 were statistically superior compared to those obtained by the different alternative functions. This suggests that it is possible to take advantage of the low degree of discrimination provided by the conventional energy formulation of the HP model.

3.4 Discussion and conclusions

The conventional energy function of the HP model is known to provide a very poor discrimination among potential conformations. Nevertheless, an effective evaluation scheme is an essential component of metaheuristics, being the responsible for steering the search process towards promising regions of the solution space. Therefore, alternative formulations of the energy function have been proposed in the literature to cope with this issue. This chapter presented the results of a comparative study where seven different formulations of the HP model's energy function were considered.

The first step in this study was concerned with the analysis of the degree of discrimination that each of the considered energy functions provides. Through such an analysis, it was possible to confirm the poor discrimination capabilities of the conventional energy function of the HP model, D85, which has been the main motivation for exploring alternative energy formulations. All the alternative functions were found to provide a more fine-grained discrimination. From the obtained results, the most discriminative functions are K99 and B08, followed by C08 and I09, in this order.

The HP-compatibility property was defined and investigated for each of the alternative energy functions. This important property refers to the capability of an alternative energy function to preserve a rank ordering among potential conformations which is consistent with the original objective function of the HP model. The obtained results suggest (but does not prove) that functions K99 and I09 feature this property. Very competitive results in this regard were also obtained by function L06. However, this was not the case for functions C04, B08 and particularly C08, which obtained the worst results in the conducted experiment. Alternative energy functions which are not HP-compatible may not be able to guide the search process properly since they can potentially introduce false optima.

The effectiveness of the studied energy functions to guide the search process was examined using a best improvement local search (BILS) algorithm. The conventional energy function D85 exhibited a low performance for this experiment. In most of the adopted test cases, however, the worst performance of the algorithm was obtained when using the alternative function C08. Also, functions B08 and C04 showed a poor search performance for most of the instances. In contrast, the alternative functions I09, L06 and K99 consistently presented a very promising behavior.

In order to further explore the suitability of the studied energy functions, a more sophisticated metaheuristic algorithm, called the iterated local search (ILS), was implemented. In most of the cases, the findings obtained using the ILS were similar to those obtained in the previous experiment using the BILS algorithm. Among the alternative energy functions, K99, I09 and L06 consistently exposed a promising behavior, while functions B08, C04 and particularly C08 presented the worst overall performance in this test. On the other side, the results obtained for the conventional function D85 suggest that, using a proper acceptance criterion, it is possible to take advantage of the neutrality that the low discrimination of this function injects into the fitness landscape.

From this study, it is possible to derive some general conclusions. First, intensity of discrimination does not necessarily imply effectiveness at guiding the search process. Even when functions K99, B08, C08 and I09 were all identified to provide a strong discrimination, only K99 and I09 presented a promising search behavior. In contrast, functions B08 and C08 showed a poor search performance in most of the cases. Such a poor performance can be explained by the fact that functions B08 and C08 are not HP-compatible. Function C04 is also not HP-compatible; the low discrimination capabilities of C04 gives further explanation to the reduced search performance obtained when using this function. Finally, function L06 obtained very competitive results in terms of both, degree of discrimination and HP-compatibility. As a consequence, function L06 consistently competed at the top of the ranking regarding search performance together with functions K99 and I09. Therefore, the degree of discrimination and the HP-compatibility property were found to be useful as a means of explaining the success or failure of the studied energy functions at guiding the search process.

The conventional energy function D85 presented a limited search performance for the BILS algorithm and for most parameter configurations of the ILS. This confirms the relevance of exploring alternative, more fine-grained evaluation schemes for the HP model. There exists evidence in the literature, however, which suggests that the neutrality of the fitness landscape can be exploited in order to design more competitive search algorithms [39, 154–156, 228, 231, 242]. The performance that function D85 achieved when using some parameter configurations of the ILS provides additional support to this idea. Furthermore, Chapter 4 demonstrates that, by introducing even more neutrality into the fitness landscape, it is possible to deal effectively with multimodality. Therefore, future work will focus on investigating how to benefit from a fine-grained discrimination, at the same time that the inherent neutrality of the HP model can be exploited. Finally, an interesting research direction involves the evaluation of how some characteristics of the fitness landscape (*e.g.*, neutrality, ruggedness [182, 228, 236]) change when using the different energy functions (as it is evaluated in Chapters 4 and 5 in the context of other different problem transformations). Such an analysis would certainly be helpful to further support the findings of the study presented in this chapter.

4

Addressing the multimodality of the HP model's fitness landscapes through multi-objectivization

4.1 Introduction

The term *multi-objectivization* was originally coined by Knowles *et al.* to refer to the process of reformulating a single-objective optimization problem in terms of two or more objective functions, *i.e.*, as a multi-objective problem [119]. It is commonly assumed that the higher the number of objective functions, the more difficult a problem is; and this is usually the case [72–75, 97, 118]. A single-objective to multi-objective transformation, however, has served as the basis for the development of more competitive search algorithms. A number of successful applications of multi-objectivization have been reported in the literature [202]. This transformation can be either based on the addition of new *supplementary objectives* [18, 105], or it can be based on the *decomposition* of the original objective function of the problem [85, 119], see Section 2.2.3. In either case, multi-objectivization may result in fundamental changes to the problem's fitness landscape. Since the performance of search algorithms is dictated by their interaction with the underlying fitness landscape [236], multi-objectivization can thus significantly impact on the ability of these algorithms to solve a given optimization task.

This chapter explores for the first time the multi-objectivization of the HP model of the protein structure prediction problem. The originally single-objective HP model is restated in an alternative multi-objective form by decomposing the conventional energy (objective) function of the problem into two separate objectives. Three different strategies to perform such a decomposition are investigated: the *parity decomposition* (PD), the *locality decomposition* (LD) and the *H-subsets decomposition* (HD). As discussed further in Section 4.2, decomposition introduces plateaus of incomparable solutions, an effect that can be exploited in order to overcome search difficulties such as that of becoming trapped in local optima [85, 119]. In this way, this study inquires into the suitability of multi-objectivization for dealing with the multimodality of the HP model's fitness landscapes.

The remainder of this chapter proceeds as follows. Related work is reviewed in Section 4.2. In Section 4.3, the PD, LD and HD formulations of the HP model are introduced. Section 4.4 presents a thorough analysis of the potential effects of the problem transformation. It is investigated how multi-objectivization influences the comparability relation among solutions, and how such an alteration in the comparability of solutions impacts on an essential property of the fitness landscape: *neutrality*. Then, a detailed comparative study is presented in Section 4.5, where the three multi-objectivization proposals for the HP model are evaluated with respect to each other and with respect to the conventional single-objective problem formulation. Such a comparative study concentrates on search performance and two different metaheuristic algorithms are considered, namely, a single-solution-based algorithm and a population-based algorithm. Finally, Section 4.6 discusses the main findings and conclusions of this study, as well as highlights some possible directions for future research.

4.2 Related work

Recently, a considerable number of successful applications of multi-objectivization have been reported in the literature. For a recent review on applications of multi-objectivization, the reader can be referred to [202]. Multi-objectivization has been largely studied in the context of well-known combinatorial problems such as the traveling salesman problem [103, 105, 119, 139]; the job-shop scheduling problem [105, 137]; the bin packing problem [201, 205]; the vehicle routing problem [235]; and the shortest path and minimum spanning tree problems [171]. Also, multi-objectivization has found interesting applications in the fields of mobile communications [200, 204, 207]; computational mechanics [82]; computer vision [233]; power system operation planning [222]; underwater sensor networks [243]; structural topology optimization [208]; computer aided manufacturing [36]; classifier parameter tuning [181]; robotics [166]; and data mining [102]. Multi-objectivization has been found to be useful also for multimodal [60], large scale [206] and constrained optimization [195]. Finally, multi-objectivization has also been proposed to deal with problems from the field of bioinformatics, such as those related to gene regulatory networks [220] and, as in the present research, to protein structure prediction [8, 50–52, 55, 84, 174, 213, 217]. Note, however, that previous multi-objectivization studies to deal with the protein structure prediction problem focus on detailed energy models. It was not until the present study that this concept is applied to the particular HP model of this problem.

The study reported in this chapter concentrates on the decomposition approach to multi-objectivization. As formally described in Section 2.2.3, this approach involves restating the single-objective problem in terms of two or more objective functions, such that the sum of all the new objectives equals the original optimization criterion of the problem. It has been demonstrated that the only possible effect of decomposition is to introduce plateaus in the search landscape [85]. That is, originally comparable solutions may become incomparable (mutually nondominated in terms of the Pareto-dominance relation) with regard to the new multi-objectivized problem formulation. Such an effect can be potentially exploited as a means of escaping from local optima [85, 119].

4.3 Multi-objectivization of the HP model

Three different multi-objectivization schemes for the HP model are proposed and analyzed in this chapter: the parity decomposition, the locality decomposition and the H-subsets decomposition. As the name of these approaches suggests, the three multi-objectivization proposals are based on the decomposition of the original energy (objective) function of the HP model. Decomposition, as discussed in Section 4.2, has the potential effect of introducing incomparability among candidate solutions. In this way, these alternative multi-objective formulations of the problem can be useful as a means of accepting degrading moves (*i.e.*, the replacement of a solution with an inferior one) and, thus, can be implemented as a mechanism to prevent search algorithms from becoming trapped in local optima. The parity, locality and H-subsets decompositions are described in detail in Sections 4.3.1, 4.3.2 and 4.3.3, respectively.

4.3.1 Parity decomposition

In the two-dimensional square and three-dimensional cubic lattices, which are the two variants of the HP model covered by this research project, adjacencies (topological contacts) are only possible between amino acids whose positions in the protein sequence are of opposite parity (this is illustrated in Figure 4.1). Based on this fact, a two-objective formulation of the HP model, $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x})]^T$, is defined over the set of all potential feasible protein conformations $\mathbf{x} \in \mathcal{X}_{\mathcal{F}}$:

$$f_1(\mathbf{x}) = \sum_{a_i, a_j} e(a_i, a_j), \qquad \text{for } \; i \equiv 0 \pmod 2,\ i < j; \tag{4.1}$$

$$f_2(\mathbf{x}) = \sum_{a_i, a_j} e(a_i, a_j), \qquad \text{for } \; i \equiv 1 \pmod 2,\ i < j; \tag{4.2}$$

where both $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are to be minimized and $e(a_i, a_j)$ represents the conventional energy contributions of the HP model as defined in Section 2.3.3.1. That is, the objective function f_1 accounts only for H-H topological contacts between pairs of amino acids a_i, a_j, where i, the sequence position of amino acid a_i, is even $(i < j)$. On the contrary, f_2 is defined for those cases where such the i-th sequence position is odd. Note that the sum of the two new alternative objectives equals the conventional energy function of the HP model presented in Section 2.3.3.1 (*i.e.*, $E(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}_{\mathcal{F}}$), which is in accordance with the decomposition approach for multi-objectivization, see Section 2.2.3. Figure 4.1 presents the optimal conformation for the test

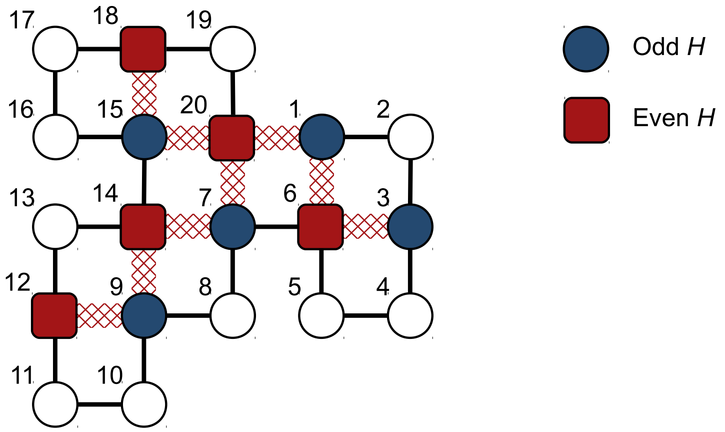


Figure 4.1: Parity decomposition. Interactions on the lattice are only possible between amino acids whose sequence positions are of opposite parity. In this example conformation, $f_1 = 0$ and $f_2 = -9$.

protein sequence 2d4 on the two-dimensional square lattice. In the particular case of this conformation, the values for the objective functions are $f_1 = 0$ and $f_2 = -9$.

4.3.2 Locality decomposition

In this multi-objectivization scheme, the conventional energy function of the HP model is decomposed based on the locality notion of amino acid interactions. An H-H topological contact between amino acids a_i and a_j can be considered to represent either a *local* or a *nonlocal* interaction. This classification depends upon whether the sequence distance between a_i and a_j (*i.e.*, $|j - i|$) is within a given maximum δ, see Figure 4.2. From this, a two-objective problem formulation, $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x})]^T$, is defined for every potential feasible protein conformation $\mathbf{x} \in \mathcal{X}_\mathcal{F}$:

$$f_1(\mathbf{x}) = \sum_{a_i,a_j} e(a_i, a_j), \qquad \text{for } \; j - i \leq \delta, i < j; \tag{4.3}$$

$$f_2(\mathbf{x}) = \sum_{a_i,a_j} e(a_i, a_j), \qquad \text{for } \; j - i > \delta, i < j; \tag{4.4}$$

where functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are both to be minimized and $e(a_i, a_j)$ denotes the conventional energy contributions defined in Section 2.3.3.1. Thus, the objective function f_1 is defined for all

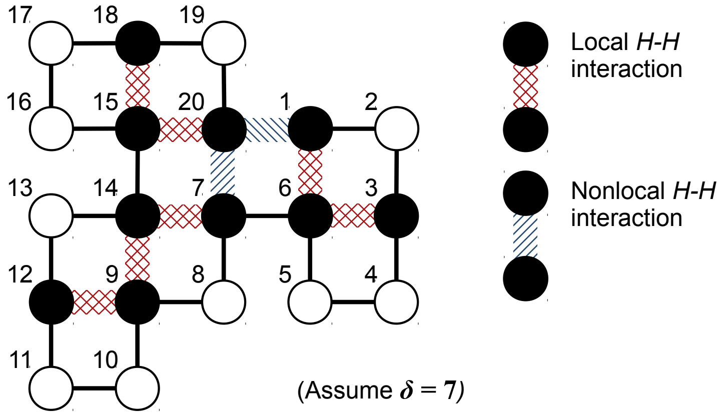


Figure 4.2: Locality decomposition ($\delta = 7$). This structure presents seven local and two nonlocal H-H interactions. Therefore, in the particular case of this example, $f_1 = -7$ and $f_2 = -2$.

the local interactions, whereas function f_2 accounts for the nonlocal ones. The evaluation of the protein conformation provided as an example in Figure 4.2, under this alternative formulation, leads to objective values $f_1 = -7$ and $f_2 = -2$. Note that $E(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}_{\mathcal{F}}$, which is consistent with the decomposition approach for multi-objectivization, as defined in Section 2.2.3.

It is worthy to mention that parameter δ plays a decisive role for the behavior of this proposal. Therefore, the influence of varying this parameter needs to be investigated. All odd values for δ in the range $[3, 21]$ are evaluated in this study (refer to Sections 4.5.1.1 and 4.5.2.1 for details).[1]

4.3.3 H-subsets decomposition

In the H-subsets decomposition, all H amino acids in the protein sequence are first assigned to one of two possible groups, $\mathcal{H}_1$ or $\mathcal{H}_2$. Groups $\mathcal{H}_1$ and $\mathcal{H}_2$ are to be referred to as the H-subsets, and such an assignment of H amino acids to these groups is to be called the H-subsets formation process. Figure 4.3 illustrates one of different H-subsets formation strategies which are explored in this study (described later at the end of this section). Once the H-subsets have been formed, an alternative

[1] In the two-dimensional square and the three-dimensional cubic lattices, a topological contact can occur if and only if the sequence distance between the amino acids is odd and at least equal to 3.

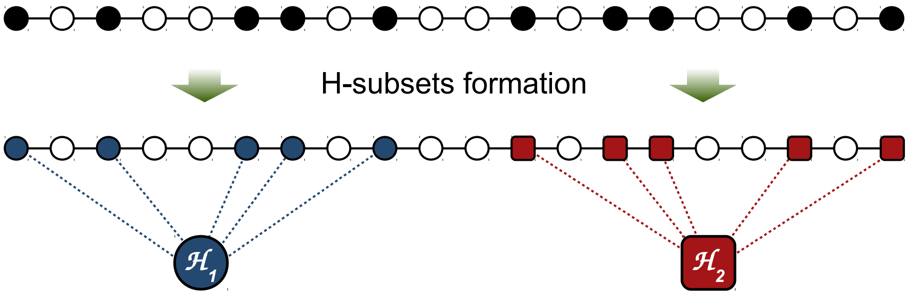


Figure 4.3: H-subsets decomposition. The H-subsets formation process based on the FIX strategy.

two-objective formulation $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x})]^T$ of the problem can be defined as follows ($\mathbf{x} \in \mathcal{X}_{\mathcal{F}}$):

$$f_1(\mathbf{x}) = \sum_{a_i, a_j \in \mathcal{H}_1} e(a_i, a_j) + \sum_{a_i, a_j \in \mathcal{H}_2} e(a_i, a_j), \tag{4.5}$$

$$f_2(\mathbf{x}) = \sum_{a_i \in \mathcal{H}_1, a_j \in \mathcal{H}_2} e(a_i, a_j), \tag{4.6}$$

where $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are both minimization functions and $e(a_i, a_j)$ denotes the conventional energy contributions of the HP model (Section 2.3.3.1). In this way, function f_1 accounts for H-H topological contacts where the two amino acids belong to the same H-subset, either $\mathcal{H}_1$ or $\mathcal{H}_2$. On the contrary, f_2 is defined for H-H topological contacts between amino acid pairs where each amino acid belongs to a different H-subset. Notice that $E(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}_{\mathcal{F}}$, which adheres to the definition of the decomposition approach for multi-objectivization, as provided in Section 2.2.3.

Given the assignment of H amino acids to the H-subsets depicted in Figure 4.3, the structure exemplified in Figure 4.4 presents four H-H topological contacts defined between amino acids belonging to the same H-subset, while the remaining five H-H interactions occur between amino acids from different H-subsets. In this particular example, the objective values are $f_1 = -4$ and $f_2 = -5$.

The H-subsets formation process plays a major role for this decomposition proposal. Different strategies can be adopted in order to accomplish this task; three of them are considered in this study (these strategies are analyzed in Sections 4.5.1.2 and 4.5.2.2):

- FIX: the first half of H amino acids in S are assigned to $\mathcal{H}_1$, all others to $\mathcal{H}_2$, as shown in Figure 4.3 (for an odd number of H amino acids, the one in the middle is assigned randomly);

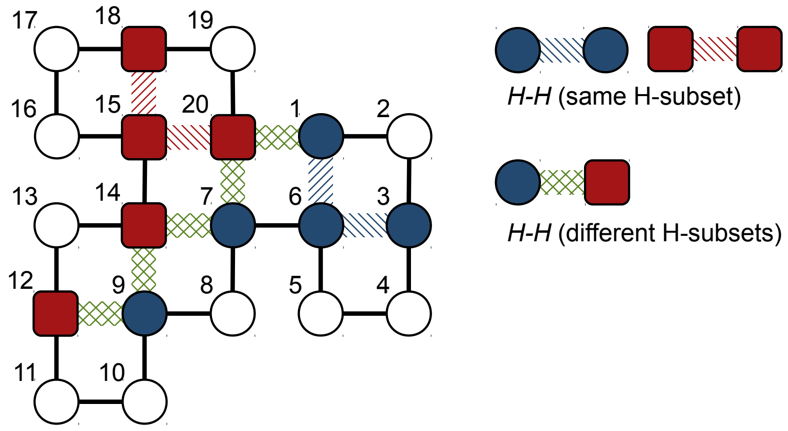


Figure 4.4: H-subsets decomposition. In this example, four H-H topological contacts are defined between amino acids from the same H-subset, while the other five are given between amino acids from different H-subsets. Therefore, $f_1 = -4$ and $f_2 = -5$.

- RND: each H amino acid can be assigned to $\mathcal{H}_1$ or to $\mathcal{H}_2$ with equal probability;

- DYN$_{\boldsymbol{k}}$: it is based on the above described RND strategy. However the H-subsets are dynamically and independently recomputed after k iterations of the search algorithm without achieving an improvement. Different values for k are explored, $k \in \{0, 10, 20, 30\}$, where $k = 0$ refers to the recomputation of the H-subsets at each iteration of the algorithm.

4.4 Effects of multi-objectivization

This section is devoted to investigating the effects that can be achieved by multi-objectivization. Although three different multi-objectivization schemes for the HP model are proposed in this research work, only the locality decomposition (defined in Section 4.3.2), using a value of $\delta = 7$, is considered in this section due to the high computational demands of the performed analyses. The locality decomposition has been selected to illustrate the effects of multi-objectivization because, as it will be shown in subsequent sections of this chapter, this approach provides a quite promising behavior. For convenience, hereafter the locality decomposition will be simply referred to as the multi-objective (MO) formulation of the problem. Similarly, two (relatively) small test instances, sequences 2d4 and

3d1, are investigated in this section (refer to Section 2.4.1 for details).[2] It is expected, however, that other different multi-objectivization proposals and test cases can be explored with similar results.

As stated in Section 2.3.3.1, the quality of a candidate solution in the HP model is evaluated in terms of an energy function, E, defined as the negative of the total number of H-H topological contacts that the encoded protein structure presents, $HHtc$. Nevertheless, the use of positive rather than negative values, as well as the adoption of the term fitness (to be maximized) rather than that of energy (to be minimized), is considered more appropriate for the analysis here reported. Therefore, in the remainder of this section the fitness of a solution $\mathbf{x}$, $Fitness(\mathbf{x})$, will assume the value of

$$Fitness(\mathbf{x}) \;=\; HHtc(\mathbf{x}) \;=\; -E(\mathbf{x}). \tag{4.7}$$

It is worthy to mention at this point that the term fitness is used in this study to refer to the quality of solutions under the conventional single-objective (SO) evaluation scheme of the HP model.[3] In addition, it is important to briefly introduce the concept of a *fitness class*; a solution $\mathbf{x} \in \mathcal{X}_{\mathcal{F}}$ will be said to belong to the fitness class c if it presents a fitness value of $Fitness(\mathbf{x}) = c$.

The analyses conducted in this section are based on an initial set of sampled solutions. Hence, the implemented sampling methodology is first introduced in Section 4.4.1. In Section 4.4.2, it is investigated how multi-objectivization influences the comparability relation among solutions. Finally, Section 4.4.3 evaluates the extent to which such an alteration in the comparability of solutions can be translated into fundamental changes to the fitness landscape structure of the problem.

4.4.1 Sampling of initial solutions

The implemented sampling strategy was conceived by taking into account the following considerations: (i) a sample $\mathcal{S}$ of M different feasible solutions for the given problem instance are to be generated; (ii) the M generated solutions are to be, if possible, evenly distributed over the different

[2] Notwithstanding, given the absolute moves encoding described in Section 2.3.3.3, the size of the search space for these (relatively) small test instances is enormous, namely 4^{19} for sequence 2d4, and 6^{19} for sequence 3d1.

[3] Although alternative multi-objective formulations of the HP model are investigated in this study, the goal remains always to solve the original single-objective problem.

available fitness classes (all fitness classes should be well represented in the collected sample); and finally, (iii) the diversity among solutions belonging to the same fitness class should be maximized.

Algorithm 3 outlines the adopted sampling strategy. The procedure starts by initializing the sample set $\mathcal{S}$ and by identifying the set of all possible fitness classes for the given problem instance, $\mathcal{FC}$ (lines 1 and 2 in Algorithm 3). Iteratively, a search algorithm is executed and all solutions that this algorithm reaches during the search process are kept in $\mathcal{U}$ (line 4). Then, the subset $\mathcal{U}_c$ of solutions in $\mathcal{U}$ belonging to each possible fitness class $c \in \mathcal{FC}$ is identified (line 6). Finally, the solution $\hat{\mathbf{x}} \in \mathcal{U}_c$ that best contributes to increasing the diversity in $\mathcal{S}$, if any, is included in the sample (lines 7 to 9). This process continues until the required sample has been completed.

Algorithm 3 Sampling of the initial solution sets.

```
Input: $M$
Output: $\mathcal{S}$
 1: $\mathcal{S} \leftarrow \emptyset$
 2: $\mathcal{FC} \leftarrow \{Fitness(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}_{\mathcal{F}}\}$
 3: while $|\mathcal{S}| < M$ do
 4:   $\mathcal{U} \leftarrow search\_algorithm()$
 5:   for all $c \in \mathcal{FC}$ do
 6:     $\mathcal{U}_c \leftarrow \{\mathbf{x} \in \mathcal{U} \mid Fitness(\mathbf{x}) = c\}$
 7:     $\hat{\mathbf{x}} \leftarrow \arg\max_{\mathbf{x} \in \mathcal{U}_c} diversity(\mathbf{x}, \mathcal{S})$
 8:     if $diversity(\hat{\mathbf{x}}, \mathcal{S}) > 0$ then
 9:       $\mathcal{S} \leftarrow \mathcal{S} \cup \{\hat{\mathbf{x}}\}$
10:     end if
11:   end for
12: end while
```

Any metaheuristic algorithm could be adopted as the embedded search method. An Iterated Local Search (ILS) algorithm [146, 147, 157], based on the SO problem formulation, was used in this study. The implemented ILS was described in detail in Section 3.3.4, where the best performing parameter settings for this algorithm were also identified. Such best ILS settings identified in Section 3.3.4 were considered in the present study. Due to its distinctive exploration behavior, the ILS method can potentially reach a different local optimum at each iteration. Each time the ILS was invoked during the sampling procedure, this algorithm was allowed to run for a total of 5×10^5 solution evaluations.

The diversity contribution estimates have been partially based on the diversification mechanism proposed by Chira [31]. Instead of measuring diversity in genotype (encoding) space, in [31] diversity was computed from the *contact fingerprint* of candidate solutions. The contact fingerprint for a solution is given by the binary vector $\mathbf{cf}$, where each component $cf_i \in \{0, 1\}$ indicates whether a particular pair of amino acids in the encoded structure defines a topological contact or not. Vector $\mathbf{cf}$ considers as many components as the total number of amino acid pairs which can potentially form a topological contact.[4] The use of the contact fingerprint rather than the encoding of solutions certainly fosters the development of more effective diversity promotion mechanisms. This can be explained by the fact that very different encodings may represent the same protein structure (after rotation or reflection, *i.e.*, the so-called isomorphic conformations [91, 210]). It is important to note, however, that significantly different structures may also present the same contact fingerprint vector if they share the same set of topological contacts. This has motivated the use of a more fine-grained version of this approach, which is referred to in this study as the *distance fingerprint*. The distance fingerprint for a given solution is defined by vector $\mathbf{df}$, each of whose components df_i measures the distance between the lattice coordinates of a particular pair of amino acids. The Manhattan distance was employed for this sake. A total of $\binom{\ell}{2} - 2\ell + 3$ components describe the distance fingerprint vector $\mathbf{df}$ (*i.e.*, only amino acid pairs (a_i, a_j) such that $|j - i| \geq 3$ require to be considered). Finally, the diversity contribution for a new candidate $\mathbf{x}$ with respect to the already collected sample $\mathcal{S}$, $diversity(\mathbf{x}, \mathcal{S})$, has been computed as the minimum Hamming distance (H_d) between the distance fingerprint vector of $\mathbf{x}$ and that of any $\mathbf{x}' \in \mathcal{S}$ with the same fitness value as $\mathbf{x}$. Formally,

$$diversity(\mathbf{x}, \mathcal{S}) = \min\{H_d(\mathbf{df}(\mathbf{x}), \mathbf{df}(\mathbf{x}')) \mid \mathbf{x}' \in \mathcal{S} \wedge Fitness(\mathbf{x}) = Fitness(\mathbf{x}')\}. \tag{4.8}$$

The size of the sample was set to $M = 1,000$ for both the 2d4 and 3d1 instances. By following the above described sampling methodology, it is (ideally) expected to generate sample sets such that about $M/|\mathcal{FC}|$ different solutions represent each possible fitness class $c \in \mathcal{FC}$. As shown in Table 4.1, this was the case of the sample set constructed for the three-dimensional instance 3d1, where a

[4] For an amino acid pair (a_i, a_j) to form a topological contact, i and j need to be of opposite parity and $|j - i| \geq 3$.

Table 4.1: Details of the sample sets generated for instances 2d4 and 3d1. Instance 2d4 involves 10 different fitness classes, while instance 3d1 involves 12.

	Fitness class												
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Total**
Sequence 2d4	119	118	119	119	119	119	119	119	47	2	-	-	$1,000$
Sequence 3d1	84	83	83	83	83	83	83	84	84	84	83	83	$1,000$

total of 83 or 84 different solutions were produced for each of the $|\mathcal{FC}| = 12$ available fitness classes. Note, however, that not all fitness classes for some of the instances can be equally sampled because of the funnel-like energy landscape which characterizes the HP model [62]. As detailed in Table 4.1, only a reduced number of solutions with a high fitness value (fitness classes 8 and 9) were obtained when sampling the search space of instance 2d4. Therefore, a greater number of representatives for the remaining fitness classes were accepted in order to complete the M required solutions.

4.4.2 Incomparability of solutions

The three multi-objectivization proposals described in Section 4.3 are all of them based on the decomposition of the original objective function of the HP model. As stated in Section 4.2, the only possible effect that can be achieved through decomposition is that originally comparable solutions may become incomparable (nondominated in terms of the Pareto-dominance relation) under the new multi-objective formulation of the problem. Consider the example provided in Figure 4.5. In this figure, conformation $\mathbf{x}_1$ (to the left) presents 9 H-H topological contacts, while conformation $\mathbf{x}_2$ (to the right) involves only 3. These originally comparable solutions (*i.e.*, $\mathbf{x}_1$ is clearly superior to $\mathbf{x}_2$) have become mutually nondominated when comparing them under the multi-objective formulation defined by the locality decomposition ($\delta = 7$). The goal of this section is not only to illustrate such a potential effect of decomposition, but also to explore the extent to which solutions belonging to different fitness classes can become incomparable as a consequence of this problem transformation.

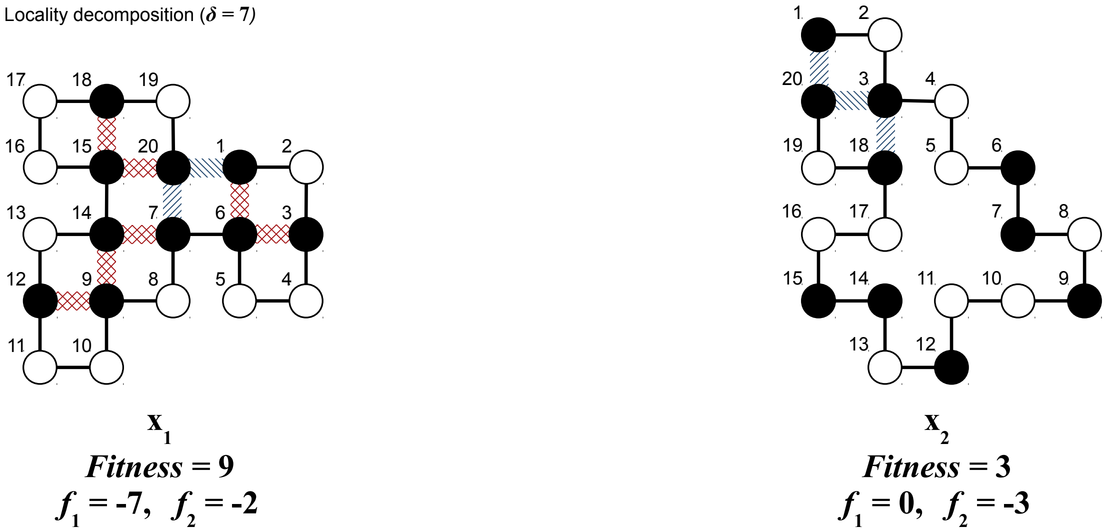


Figure 4.5: This figure illustrates how a pair of originally comparable solutions $\mathbf{x}_1$ (left) and $\mathbf{x}_2$ (right), with $Fitness(\mathbf{x}_1) = 9$ and $Fitness(\mathbf{x}_2) = 3$, can become incomparable by multi-objectivization. The objective values obtained by the locality decomposition ($\delta = 7$) are shown for both $\mathbf{x}_1$ and $\mathbf{x}_2$.

A sample set of $M = 1,000$ different candidate solutions was generated by implementing the methodology detailed in Section 4.4.1. Then, all possible pairwise comparisons among the sampled solutions were performed. From this, it was computed the ratio of the number of incomparable solution pairs found to the total number of pairwise comparisons carried out; this measure is to be referred to as the *incomparability ratio* (IR). IR is thus defined in the range $[0, 1]$, and IR $= 1$ indicates that all the evaluated solution pairs were found to be incomparable. This experiment was replicated for both the conventional single-objective HP model formulation, SO, and the alternative multi-objective formulation, MO.[5] The comparison of solutions under the MO formulation relies on the Pareto-dominance relation and the locality decomposition. Both the 2d4 and 3d1 instances were considered. The obtained results are summarized in Table 4.2. As it can be seen from this table, $70,616$ out of the $\binom{1,000}{2} = 499,500$ total pairs of sampled solutions for instance 2d4 became incomparable when evaluated under the MO formulation. This represents an IR increase of 0.14 with regard to the conventional SO formulation. Similarly, multi-objectivization increased the IR measure by 0.13 when focusing on the 3d1 instance. Such an increase of 0.13 results from the $63,760$ comparable solution pairs for which the original preference relation has been suppressed.

[5] By definition, only solutions having the same fitness value are incomparable under the SO problem formulation.

Table 4.2: Incomparability ratio (IR) presented by the studied SO and MO formulations. For both the 2d4 and 3d1 test instances, a total of $\binom{1,000}{2} = 499,500$ solution pairs have been evaluated.

	SO	**MO**	**Increase (MO−SO)**
Sequence 2d4	$0.11\ (57,132)$	$0.26\ (127,748)$	$0.14\ (70,616)$
Sequence 3d1	$0.08\ (41,168)$	$0.21\ (104,928)$	$0.13\ (63,760)$

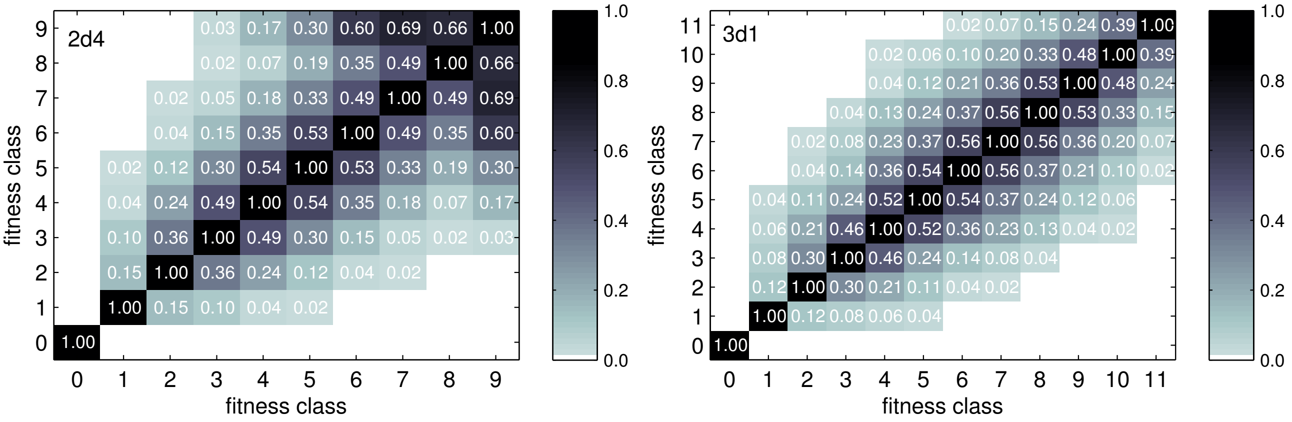


Figure 4.6: Incomparability ratio (IR) computed separately for each possible pair of fitness classes. Two-dimensional 2d4 instance (left) and three-dimensional 3d1 instance (right). In the plots, IR values are highlighted with colors; the darker the color, the higher the IR value.

Finally, the results obtained using the MO formulation are broken down in Figure 4.6 in order to gain further insights into the likelihood of incomparability taking place among the different fitness classes of the considered test instances. Figure 4.6 shows the IR measure computed separately for each possible pair of fitness classes (*i.e.*, fitness classes with respect to the conventional SO formulation). Heat maps in this figure are symmetric along the diagonal. From this figure, it is possible to see that multi-objectivization makes incomparability possible even for pairs of solutions which are distant with respect to their fitness values. For example, incomparability has been introduced between fitness classes 3 and 9 of instance 2d4. A similar scenario can be found with regard to fitness classes 4 and 10 from the collected sample for sequence 3d1. Note, however, that the closer the fitness classes for the selected solution pairs, the higher the indicated IR values (in the plots, the highest IR values appear close to the diagonal). This can be understood by the fact that the increase in the fitness distance between a pair of solutions increases also the probability for one of these solutions to be dominated (in the Pareto sense) by the other under the alternative multi-objectivized formulation. Finally, it should be observed that in the case of both the 2d4 and 3d1 instances, no solution at fitness class 0 became incomparable with respect to any other solution from a higher fitness class. This is due to the fact that any solution $\mathbf{x}_1$ with $Fitness(\mathbf{x}_1) = 0$, $f_1(\mathbf{x}_1) = 0$ and $f_2(\mathbf{x}_1) = 0$ after applying decomposition, will always be dominated by any other solution $\mathbf{x}_2$ with $Fitness(\mathbf{x}_2) > 0$, no matter how $Fitness(\mathbf{x}_2)$ is decomposed into the new set of objectives.

4.4.3 Fitness landscape analysis

As discussed in Section 4.4.2, multi-objectivization by decomposition exerts an influence on the comparability relation over the search space, in such a way that solutions from different fitness classes may become incomparable when evaluated under the new multi-objective formulation of the problem. This notion of incomparability is equivalent to that of neutrality used in the context of fitness landscapes. Two solutions are said to be neutral (*i.e.*, incomparable) if either they share the same fitness value (single-objective case), or they are Pareto-nondominated with respect to each other (multi-objective case), see Section 2.2.4. Hence, the potential effect of multi-objectivization, previously described in terms of introducing incomparability among solutions, will be referred in this section to as that of increasing the neutrality in the fitness landscape. This section is intended to contribute in understanding and, to some extent, quantifying such an effect of multi-objectivization on the fitness landscapes of the HP model.

As detailed in Section 2.2.4, three important components define a fitness landscape: $(\mathcal{X}, \mathcal{N}, \xi)$. While the search space $\mathcal{X}$ and the neighborhood structure $\mathcal{N}$ were kept constant in this study,[6] ξ has been varied from the conventional single-objective (SO) evaluation scheme of the HP model to the alternative multi-objective (MO) evaluation scheme based on the locality decomposition ($\delta = 7$) and the Pareto-dominance relation. By analyzing and comparing the landscapes induced by the SO and MO evaluation schemes, it will then be possible to evaluate the extent to which multi-

[6] $\mathcal{X}$ is given by the implemented absolute moves encoding (described in Section 2.3.3.3). Likewise, $\mathcal{N}(\mathbf{x})$ is defined by all solutions which can be reached through a single change in the encoding of $\mathbf{x}$. Thus, $|\mathcal{N}(\mathbf{x})| = 3(\ell - 1)$ and $|\mathcal{N}(\mathbf{x})| = 5(\ell - 1)$ in the two- and three-dimensional cases, ℓ denoting the length of the protein sequence.

objectivization has impacted on essential characteristics of the problem, those related to neutrality. Neutrality is here investigated by focusing on different properties of neutral networks (NNs); since neutral fitness landscapes, as those in the HP model, are known to be mainly described by their NNs [155]. Nevertheless, NNs in a neutral fitness landscape can be of a considerable size, so that their exhaustive exploration becomes computationally prohibitive even for relatively small problem instances. In the literature, NNs are usually sampled through neutral walks, *i.e.*, series of (neutral) neighboring solutions. In this study, however, an alternative approach was taken, as described below.

Given a sample set $\mathcal{S}$ of M different candidate solutions, collected following the methodology previously detailed in Section 4.4.1, the neutral network $NN(\mathbf{x})$ for each solution $\mathbf{x} \in \mathcal{S}$ has been partially computed based on the $pNN()$ procedure outlined in Algorithm 4. As shown in this algorithm, $NN(\mathbf{x})$ is constructed recursively in a depth-first manner by allowing this procedure to reach a maximum defined depth level ($maxDepth$). The initially given solution $\mathbf{x}$ is assumed to be at depth level 0, so that $depthLevel = 0$ is used in the first call to $pNN()$. At each call to the $pNN()$ method, $NN(\mathbf{x})$ is first initialized to the graph containing no edges and including the provided solution $\mathbf{x}$ as the only node (line 1 in Algorithm 4). If the maximum allowed depth level has not been reached (line 2), the sub-network $NN(\mathbf{x}')$ for every neutral neighbor $\mathbf{x}'$ of $\mathbf{x}$ is obtained from a subsequent execution of the $pNN()$ method (by giving $\mathbf{x}'$ as the new starting point and by increasing the value of $depthLevel$, see line 4). The resulting sub-network $NN(\mathbf{x}')$ is then merged with the parent network $NN(\mathbf{x})$ by means of a graph union operation, here denoted as $\bigcup$ (line 5).[7] Finally, edge $(\mathbf{x}, \mathbf{x}')$ is included in the edge set of $NN(\mathbf{x})$ in order to establish the linkage between $NN(\mathbf{x})$ and the $NN(\mathbf{x}')$ sub-network. This strategy of partially computing the NN for a given solution $\mathbf{x}$, is equivalent to traversing all possible neutral walks departing from $\mathbf{x}$, by restricting the length of the walks to the maximum defined depth level ($maxDepth$).

The 2d4 and 3d1 test instances have been considered for this analysis, and the size of the initial sample sets was fixed to $M = 1,000$ in both cases. Thus, a total of $1,000$ (potentially different) NNs for each of the instances have been explored by using both, the SO and MO evaluation schemes,

[7] Given two graphs $G_1 = (\mathcal{V}_1, \mathcal{E}_1)$ and $G_2 = (\mathcal{V}_2, \mathcal{E}_2)$, the graph union operation $G_1 \bigcup G_2$ produces a third graph $G_3 = (\mathcal{V}_3, \mathcal{E}_3)$ such that $\mathcal{V}_3 = \mathcal{V}_1 \cup \mathcal{V}_2$ and $\mathcal{E}_3 = \mathcal{E}_1 \cup \mathcal{E}_2$.

Algorithm 4 $pNN()$ - Partial NN computation

```
Input: $\mathbf{x}$, $depthLevel$, $maxDepth$
Output: $NN(\mathbf{x})$
$NN(\mathbf{x}) \leftarrow (\mathcal{V}, \mathcal{E}) : \mathcal{V} = \{\mathbf{x}\},\ \mathcal{E} = \emptyset$
if $depthLevel < maxDepth$ then
    for all $\mathbf{x}' \in \mathcal{N}_n(\mathbf{x})$ do
        $NN(\mathbf{x}') \leftarrow pNN(\mathbf{x}', depthLevel + 1, maxDepth)$
        $NN(\mathbf{x}) \leftarrow NN(\mathbf{x}) \bigcup NN(\mathbf{x}')$
        $\mathcal{E} \leftarrow \mathcal{E} \cup \{(\mathbf{x}, \mathbf{x}')\}$
    end for
end if
```

as the bases for neutrality verification. In this way, changing the problem formulation from SO to MO will be reflected as an alteration in the properties of the sampled NNs. In the remainder of this section, the neutral network for a given solution $\mathbf{x}$ will be either referred to as $NN_{SO}(\mathbf{x})$ or $NN_{MO}(\mathbf{x})$, depending on whether the neutrality relation among solutions was determined based on the SO or MO evaluation schemes during the network computation. Finally, in order to overcome the high computational cost of the conducted analysis, the maximum allowed depth level was set to $maxDepth = 10$ and $maxDepth = 7$ for the 2d4 and 3d1 test sequences, respectively.[8]

4.4.3.1 Average neutrality ratio

As a means of evaluating the increase on neutrality caused by multi-objectivization, the *average neutrality ratio* (ANR) of the sampled NNs is investigated. The ANR is defined as the mean of the neutrality ratios (as defined in Section 2.2.4) considering all solutions in a NN [227, 228]. This measure assumes values in the range $[0, 1]$, where 1 corresponds to the highest neutrality. Figure 4.7 contrasts the ANR values obtained when using the SO and MO formulations (*i.e.*, the ANR values computed from $NN_{SO}(\mathbf{x})$ and $NN_{MO}(\mathbf{x})$, for all $\mathbf{x} \in \mathcal{S}$). In this figure, the ANR values appear organized according to the fitness class of the solution given as the starting point for the NN sampling. In addition, the mean of the ANR values in each fitness class is indicated for both the SO and MO formulations. From Figure 4.7, a general tendency can be perceived with regard to

[8] Despite the use of such low $maxDepth$ values, the resulting NNs were considerably large, see Section 4.4.3.2.

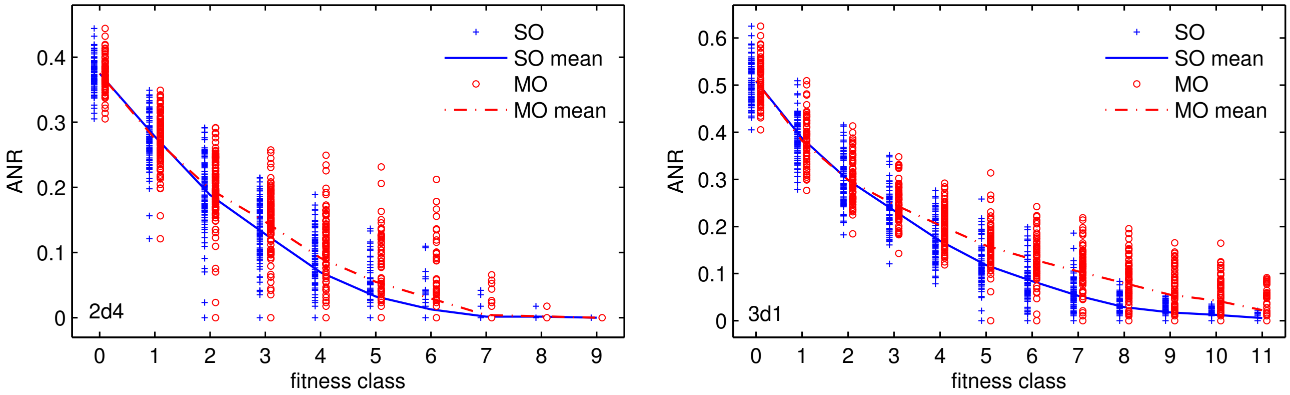


Figure 4.7: Average neutrality ratio (ANR) of the sampled neutral networks. Two-dimensional 2d4 instance (left) and three-dimensional 3d1 instance (right).

the neutrality of the HP model's fitness landscapes. The ANR rapidly decreases with the increase in fitness. That is, while poor quality solutions (with low fitness values) are usually surrounded by a considerable number of neutral neighbors, leading to large NNs (see Section 4.4.3.2), solutions at the highest fitness classes tend to be more isolated and enclosed by infeasible states. The reader is referred to Sections 5.4.2.1 and 5.4.2.2 for further discussion on this issue. In most fitness classes, it is evident from the plots that there was a slight increase in the ANR measure as a consequence of using the MO formulation (fitness classes 2 to 7 of instance 2d4, and fitness classes 3 to 11 of instance 3d1). It is important to note that no increase in the ANR is possible for NNs at fitness class 0. This is due to the fact that, as discussed in Section 4.4.2, a solution at this fitness class can not become neutral with respect to any other solution at a higher fitness class. After multi-objectivization, any solution $\mathbf{x}$ with $Fitness(\mathbf{x}) = 0$ will still be considered inferior (dominated in the Pareto sense) with regard to any solution $\mathbf{x}'$ with $Fitness(\mathbf{x}') > 0$. Thus, NNs for solutions at fitness class 0 will be exactly the same regardless of whether they are computed based on the SO or MO formulation (this applies also for subsequent analyses presented in Sections 4.4.3.2 and 4.4.3.3).

4.4.3.2 Size of the neutral networks

Despite the minor increases in the average neutrality ratio (ANR) obtained by the use of a multi-objective problem formulation, a small variation in the neutrality degree of solutions can still con-

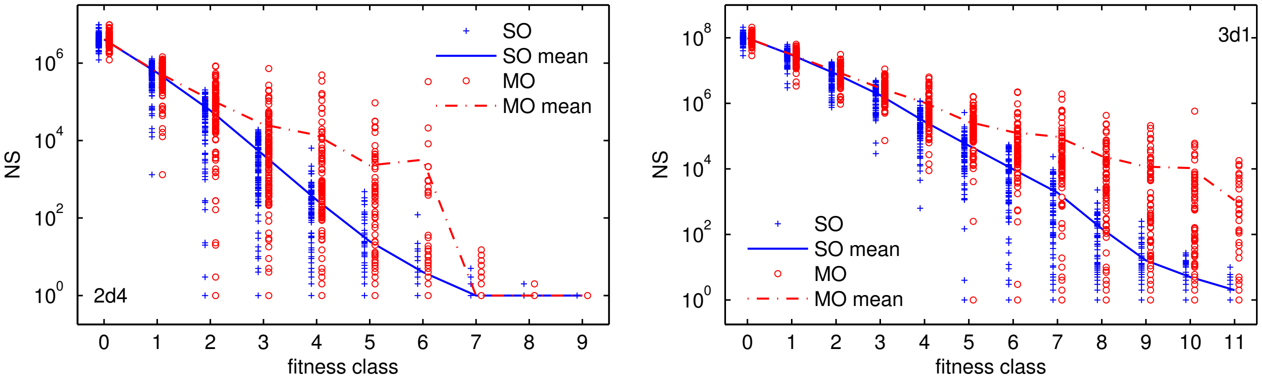


Figure 4.8: Size of the sampled neutral networks (NS). Two-dimensional 2d4 instance (left) and three-dimensional 3d1 instance (right).

tribute significantly to increasing the size (number of solutions) of a NN. For each sampled solution $\mathbf{x} \in \mathcal{S}$ of instances 2d4 and 3d1, the size of the computed $NN_{SO}(\mathbf{x})$ and $NN_{MO}(\mathbf{x})$ networks is analyzed in this section through Figure 4.8. The results are presented separately for each fitness class, and the arithmetic mean in each of the cases is also indicated. The plots, given in a logarithmic (base 10) scale, expose the high neutrality that characterizes the fitness landscapes in the HP model. Even when the sampling of NNs was restricted in this study by setting a maximum allowed depth level, as stated in Section 4.4.3, it is possible to see from the plots that NNs at fitness class 0 involve above 10^6 and around 10^8 solutions for the 2d4 and 3d1 instances, respectively. From Figure 4.8, it is also possible to confirm that, as suggested in Section 4.4.3.1, the size of the NNs is usually larger for lower fitness classes, but neutrality tends to decrease as higher quality solutions are considered. An important increase in the size of the NNs can be observed in most of the cases due to the use of the MO formulation. As the plots indicate, NNs computed based on the MO formulation can be several orders of magnitude larger than those computed based on the SO formulation.

To go further in this analysis, the *neutral network size ratio* (NSR) is defined in this study as the ratio of the size of $NN_{SO}(\mathbf{x})$ to that of $NN_{MO}(\mathbf{x})$. In fact, $NN_{MO}(\mathbf{x})$ will always be a supergraph containing all nodes and edges of $NN_{SO}(\mathbf{x})$, but including also those nodes and edges which result from the neutrality introduced by the multi-objectivization. Thus, $NN_{MO}(\mathbf{x})$ will have at least the same size as $NN_{SO}(\mathbf{x})$, so that NSR is defined in the range $[0, 1]$ and NSR $= 1$ indicates that no change in the NN size was achieved when varying the problem formulation. Figure 4.9 shows the

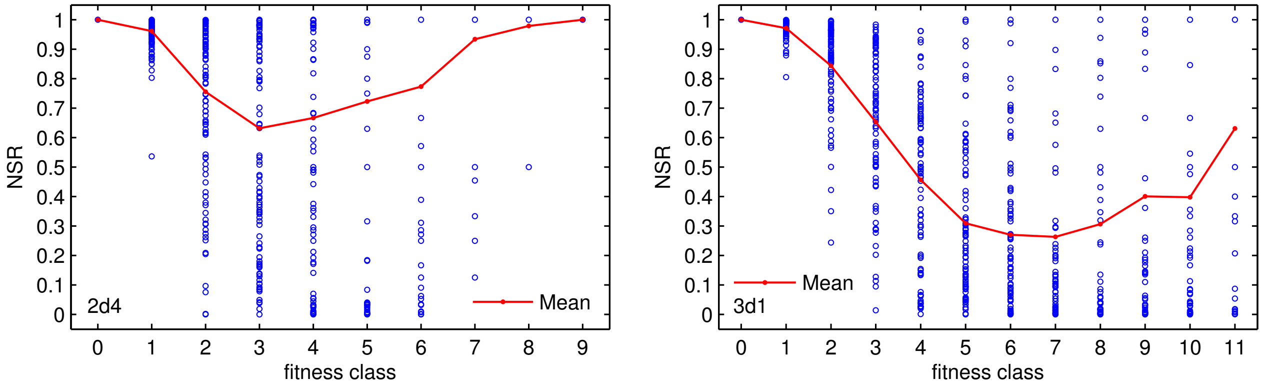


Figure 4.9: Neutral network size ratio (NSR). Two-dimensional 2d4 instance (left) and three-dimensional 3d1 instance (right).

NSR for all the sampled NNs along with the mean values calculated for each of the fitness classes. Multi-objectivization led to an important rise in the size of the explored NNs for most fitness classes of the 2d4 instance. The sharpest increase in the NN size can be observed at fitness class 3, for which the lowest average NSR value has been scored. In average, the size of $NN_{SO}(\mathbf{x})$ is only about 65% of the size of $NN_{MO}(\mathbf{x})$ when $Fitness(\mathbf{x}) = 3$. The impact of using the MO formulation becomes more evident when focusing on the 3d1 instance. The average NSR values for most fitness classes are below 0.5. This evinces that $NN_{MO}(\mathbf{x})$ at least doubled the size of $NN_{SO}(\mathbf{x})$ in the vast majority of the cases. Finally, it is possible to note from the results of both the 2d4 and 3d1 instances that a considerable number of very low NSR values (close to 0) have been accounted for, indicating a highly significant increase in the size of the corresponding NNs.

4.4.3.3 Connectivity between neutral networks

The observed increments with regard to the size of the NNs, as analyzed in Section 4.4.3.2, can be understood as the result of allowing *neutral connections* to be established between NNs. While in the single-objective case all solutions in a NN share, by definition, the same fitness value, in a multi-objectivization scenario such a strict definition of a NN can no longer be supported. That is, given the adopted notion of neutrality for the multi-objective case, which is based on the Pareto-dominance relation (see Section 2.2.4), a NN constructed using as the basis the MO formulation

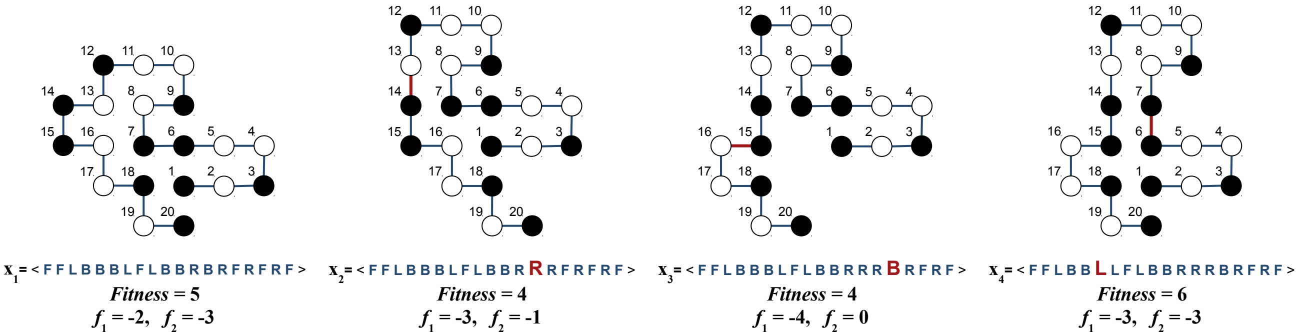

x1 = < F F L B B B L F L B B R B R F R F R F >
Fitness = 5
f1 = -2, f2 = -3
x2 = < F F L B B B L F L B B R R R F R F R F >
Fitness = 4
f1 = -3, f2 = -1
x3 = < F F L B B B L F L B B R R R B R F R F >
Fitness = 4
f1 = -4, f2 = 0
x4 = < F F L B B L L F L B B R R R B R F R F >
Fitness = 6
f1 = -3, f2 = -3

Figure 4.10: Neutral walk, based on the MO formulation, from a solution $\mathbf{x}_1$ with $Fitness(\mathbf{x}_1) = 5$ to a solution $\mathbf{x}_4$ with $Fitness(\mathbf{x}_4) = 6$. The connection between $\mathbf{x}_1$ and $\mathbf{x}_4$ was possible by traversing inferior solutions $\mathbf{x}_2$ and $\mathbf{x}_3$ with $Fitness(\mathbf{x}_2) = Fitness(\mathbf{x}_3) = 4$. Neutral connections were formed between NNs at fitness classes 5, 4 and 6, leading to a single NN. The f_1 and f_2 objective values, obtained by means of the LD multi-objectivization ($\delta = 7$), are presented for each solution.

may involve solutions from varying fitness classes, thus connecting the corresponding NNs. Consider two NNs, NN_1 and NN_2, such that the fitness of NN_1 is $Fitness(NN_1) = A$ and the fitness of NN_2 is $Fitness(NN_2) = B$, $A \neq B$. If, as a result of using the MO formulation, at least a solution $\mathbf{x}_1$ from NN_1 comes to be neutral with respect to a neighboring solution $\mathbf{x}_2$ which belongs to NN_2, then this neutral connection between NN_1 and NN_2 will merge the two NNs together into a single NN where both the fitness classes A and B are represented. To further illustrate these ideas, refer to the example provided in Figure 4.10. This figure presents a series of neutral moves between neighboring solutions, *i.e.*, a neutral walk, based on the MO formulation. The neutrality that the MO formulation introduced between these solutions led to the formation of neutral connections between three different NNs, namely $NN(\mathbf{x}_1)$ at fitness class 5, $NN(\mathbf{x}_2, \mathbf{x}_3)$ at fitness class 4 and $NN(\mathbf{x}_4)$ at fitness class 6. In this way, the four solutions, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ and $\mathbf{x}_4$, and all solutions belonging to their respective NNs, became part of a single larger NN as a consequence of multi-objectivization.

From the above introduced notion of neutral connections, it becomes relevant for this study to investigate the extent to which neutral connections took place, between the different fitness classes, during the performed sampling of NNs. Figure 4.11 summarizes the results obtained for the 2d4 and 3d1 instances. The NNs constructed for each fitness class c were analyzed (*i.e.*, c is the fitness of the initially given solution for the NN computation), and the plots indicate whether neutral connections

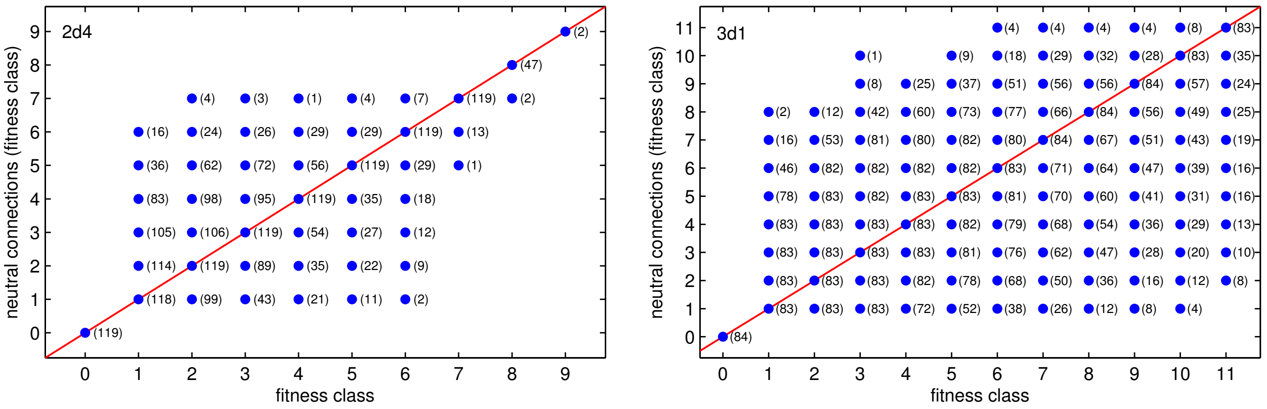


Figure 4.11: Neutral connections established between the different fitness classes. Two-dimensional 2d4 instance (left) and three-dimensional 3d1 instance (right).

were identified between these NNs and NNs at each other possible fitness class c'. The total number of neutral connections found of each type, if any, is shown in parentheses. Diagonals in these plots are used only as a reference (*i.e.*, all NNs connect to themselves at their corresponding fitness classes) to illustrate the single-objective case, so that all other connections not appearing along the diagonal are due to the landscape transformation. As an example, Figure 4.11 indicates, regarding sequence 2d4, that only 16 out of the 118 sampled NNs for fitness class 1 formed neutral connections to NNs at fitness class 6. Through this analysis it is then possible to gain an insight into the diversity of fitness classes that a NN, computed based on the MO formulation, may involve. Figure 4.11 highlights that a significant number of neutral connections were originated by multi-objectivization. On the one hand, NNs from fitness classes 2 to 6 of instance 2d4 presented neutral connections to all fitness classes between 1 and 7 (at least one connection in each of the cases can be observed from the plot). Note also that only a few neutral connections, all of them to inferior fitness classes, were produced from NNs at fitness classes 7 and 8. No NNs at fitness class 9 connected to others. On the other hand, a higher number of neutral connections were generated with regard to the 3d1 test instance. Neutral connections were established, in one direction or the other, between almost all pairs of fitness classes. That is, even though the NNs computed for fitness classes 2 to 5 did not form connections to fitness class 11, multiple connections from class 11 to all such lower fitness classes were created. This points out the fact that inferior fitness classes are easier to reach than the superior ones (because of the funnel-like search landscape that characterizes the studied problem [62]).

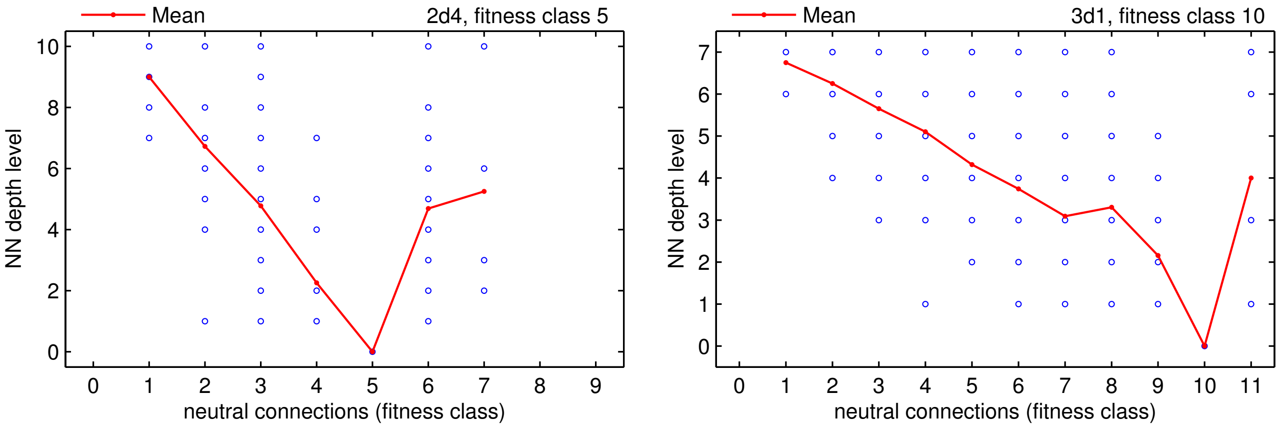


Figure 4.12: Neutral connections formed in relation to the depth level reached during the NNs sampling. Fitness class 5 of instance 2d4 (left) and fitness class 10 of instance 3d1 (right). Connections to the respective fitness classes (5 and 10) occurred at depth level 0 (starting point of the sampling).

A neutral connection from a fitness class c to a fitness class c' indicates that, given an arbitrary solution $\mathbf{x}$ with $Fitness(\mathbf{x}) = c$, a neutral walk departing from $\mathbf{x}$ could potentially lead to a solution $\mathbf{x}'$ with $Fitness(\mathbf{x}') = c'$. Nevertheless, as Figure 4.11 suggests, the more distant the fitness classes c and c', the lower the likelihood that these classes can connect to each other through a neutral walk (in the plots, higher number of neutral connections are shown closer to the diagonal). Such a behavior is certainly accentuated if the length of the walks is bounded (as it was done in this study by allowing a maximum depth level to be reached during the NNs computation). In addition, although (relatively) distant fitness classes can directly connect to each other, *i.e.*, in a single step of the neutral walk, the increase in the fitness distance between a pair of solutions decreases the probability for these solutions to become incomparable after multi-objectivization, as analyzed at the end of Section 4.4.2. Thus, the connection between distant fitness classes is more likely to occur through a series of intermediate states. This point can be better explained by considering Figure 4.12. As an example, this figure considers fitness class 5 for instance 2d4, and fitness class 10 for instance 3d1, in order to illustrate how the neutral connections to the different fitness classes arose as each allowed depth level was reached during the NNs computation.[9] The mean depth level at which connections to the different fitness classes were produced is also provided. It is possible to see from the plots that neutral connections to different fitness classes were given directly at depth

[9] Similar results were obtained for the different fitness classes of the considered test instances.

level 1. Classes $\{2, 3, 4, 6\}$ were directly connected from NNs at fitness class 5 of instance 2d4. Similarly, classes $\{4, 6, 7, 8, 9, 11\}$ were connected in a single neutral step from NNs at fitness class 10 of instance 3d1. Note, however, that the mean depth level values in these plots confirm the above suggested tendency, that the distance between fitness classes is closely related to the number of neutral steps that will be usually required to connect different NNs.

Finally, it is important to address the question of how the formation of neutral connections may impact on the behavior of a search algorithm. As it has been seen, multi-objectivization can affect the neutrality relation for a given solution $\mathbf{x}$ in two possible directions: (i) $\mathbf{x}$ becoming neutral with respect to an inferior solution, *i.e.*, connection to a lower fitness class; or (ii) $\mathbf{x}$ becoming neutral with respect to a superior solution, *i.e.*, connection to a higher fitness class. In the former scenario, multi-objectivization can be thought of as enhancing the mobility of the search algorithm. By connecting to lower fitness classes, the algorithm is allowed to traverse landscape areas which were originally inaccessible under the conventional SO evaluation scheme. In the neutral walk illustrated in Figure 4.10, for example, to move from fitness class 5 ($\mathbf{x}_1$) to fitness class 6 ($\mathbf{x}_4$) it was required to accept a degrading move to fitness class 4 ($\mathbf{x}_2$ and $\mathbf{x}_3$). In this way, the movement through inferior solutions constitutes the basis for a potential strategy to escape from local optima. It should be noted, therefore, that the design of the search algorithm will play a critical role for achieving success through the problem transformation. The new defined neutral paths can only be exploited if the algorithm is designed so as to accept moves between neutral solutions (or, in the words of Barnett, if the algorithm is able to crawl the NNs [6]). The formation of neutral connections to superior fitness classes, the later scenario, can be analyzed from two different perspectives. On the one hand, these connections (indeed all neutral connections in general) reduce what is called *selective pressure* in the context of evolutionary optimization [5], which can boost the exploration behavior of an algorithm. On the other hand, these connections can be interpreted as an important loss in gradient information. Rather than benefiting from multi-objectivization, for instance, a search algorithm based on a strictly-better acceptance criterion could easily stagnate due to its inability to perform a proper discrimination. By relaxing the comparability relation among solutions, thus, multi-objectivization may also hinder the ability of an algorithm to identify a promising search direction.

4.5 Search performance

The aim of this section is to investigate the influence that multi-objectivization can exert on the search behavior of metaheuristic algorithms. To this end, the three proposed multi-objectivization schemes, based on the parity (PD), locality (LD) and H-subsets (HD) decompositions, are evaluated and compared with respect to the conventional single-objective (SO) formulation of the HP model. Two different metaheuristics are considered, namely, a single-solution-based algorithm and a population-based algorithm. The corresponding analyses are presented in Sections 4.5.1 and 4.5.2.

4.5.1 Analysis for a single-solution-based algorithm

In this section, a basic single-solution-based evolutionary algorithm (EA), the so-called (1+1) EA, is used for inquiring into the impact that multi-objectivization can have on search performance. The general structure of the implemented (1+1) EA is sketched in Algorithm 5. In this algorithm, an initial parent individual $\mathbf{x}$ is first generated at random. At each generation, an offspring $\mathbf{x}'$ is created by randomly and independently mutating $\mathbf{x}$ at each encoding position with probability p_m. The new individual $\mathbf{x}'$ is rejected only if it is strictly worse than the parent individual $\mathbf{x}$, otherwise $\mathbf{x}'$ is accepted as the starting point for the next generation. Such an acceptance criterion will be either based on a single-objective discrimination between $\mathbf{x}$ and $\mathbf{x}'$, based on the conventional SO formulation of the HP model, or it will be based on the Pareto-dominance relation when using the alternative PD, LD and HD multi-objective formulations of the problem. In this way, the performance variations to be observed in the (1+1) EA will be attributed to the change in the problem's formulation.

Algorithm 5 Basic (1+1) evolutionary algorithm.

```
choose $\mathbf{x} \in \mathcal{X}_{\mathcal{F}}$ uniformly at random
repeat
    $\mathbf{x}' \leftarrow mutate(\mathbf{x})$
    if $\mathbf{x}'$ not worse than $\mathbf{x}$ then
        $\mathbf{x} \leftarrow \mathbf{x}'$
    end if
until $< stop\ condition >$
```

In the multi-objective optimization context, the use of a nondominated solution archive is usually assumed essential for driving the search process effectively [117, 144]. Therefore, a variant of the (1+1) EA is also considered with the purpose of evaluating the impact of archiving on the behavior of the proposed multi-objective formulations of the HP model. The archiving (1+1) EA, as detailed in Algorithm 6, uses an external archive to store the nondominated solutions (in the Pareto sense) found along the evolutionary process. It is important to realize that the implemented archiving strategy does not lead to a population-based search method. The nondominated solutions archive influences only the acceptance criterion of the algorithm, but no genetic material from this archive is exploited during optimization. The archiving-based acceptance criterion affects the behavior of the algorithm in such a way that the offspring $\mathbf{x}'$ is only accepted if it is not dominated by any individual in the archive. If accepted, $\mathbf{x}'$ is included in the archive and all individuals dominated by $\mathbf{x}'$, and those mapping to the same objective vector $\mathbf{f}(\mathbf{x}')$, are removed.[10] Note also that archiving makes sense only in multi-objective scenarios, so that no results are to be reported on the application of the archiving (1+1) EA to the conventional SO problem formulation.

Algorithm 6 Archiving (1+1) evolutionary algorithm.

```
$choose\ \mathbf{x} \in \mathcal{X}_{\mathcal{F}}\ uniformly\ at\ random$
$\mathcal{A} \leftarrow \{\mathbf{x}\}$
**repeat**
    $\mathbf{x}' \leftarrow mutate(\mathbf{x})$
    **if** $\nexists \hat{\mathbf{x}} \in \mathcal{A} : \hat{\mathbf{x}} \prec \mathbf{x}'$ **then**
        $\mathcal{A} \leftarrow \{\hat{\mathbf{x}} \in \mathcal{A} : \mathbf{x}' \nprec \hat{\mathbf{x}} \wedge \mathbf{f}(\hat{\mathbf{x}}) \neq \mathbf{f}(\mathbf{x}')\} \cup \{\mathbf{x}'\}$
        $\mathbf{x} \leftarrow \mathbf{x}'$
    **end if**
**until** $<stop\ condition>$
```

In both the basic and the archiving variants of the (1+1) EA, individuals encode protein conformations using an internal coordinates representation based on absolute moves (as described in Section 2.3.3.3). Moreover, only individuals encoding feasible protein conformations are considered during the search process (infeasible individuals are always discarded, see Section 5.6.1.1). The initial feasible individuals are generated using a backtracking procedure [42]. In all the cases, the mutation

[10] It is important to remark that the size of the external archive has not been bounded in this algorithm.

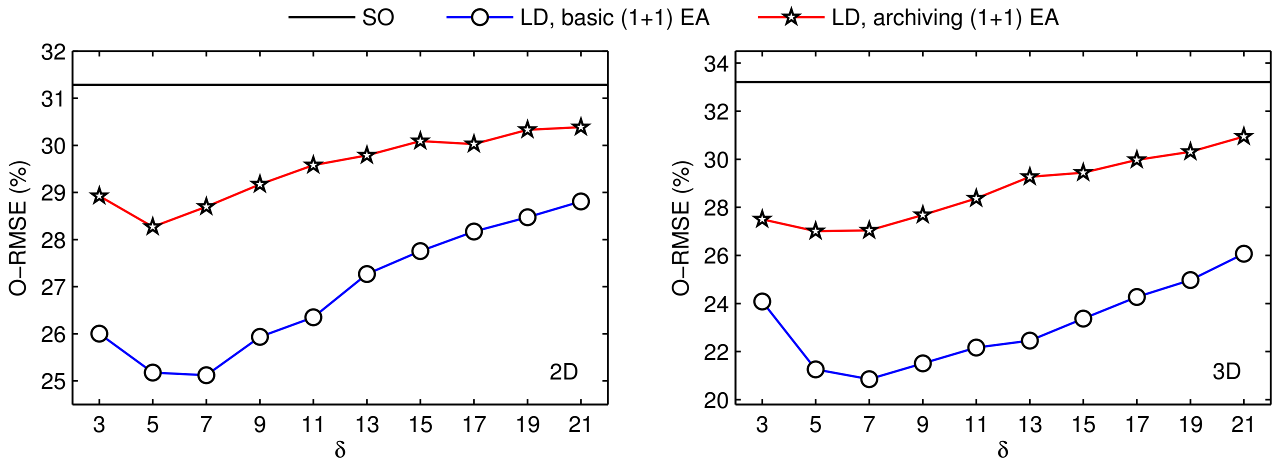


Figure 4.13: Locality decomposition. Evaluating the impact of varying parameter δ on the performance of the (1+1) EA. Two-dimensional (left) and three-dimensional (right) test cases.

probability was fixed to $p_m = \frac{1}{\ell-1}$, where $\ell - 1$ denotes the length of the individuals encoding. Finally, a maximum number of 5×10^5 solution evaluations was adopted as the stopping condition, and a total of 100 independent executions were performed for all two- and three-dimensional instances.

The remainder of this section is organized as follows. The proposed LD and HD formulations are sensitive to the adjustment of some parameters. Therefore, the influence of varying such parameters is first evaluated in Sections 4.5.1.1 and 4.5.1.2. Then, the effects of using the archiving strategy within the (1+1) EA are explored in Section 4.5.1.3. Finally, a detailed comparative analysis among the four studied formulations of the HP model (SO, PD, LD and HD) is presented in Section 4.5.1.4.

4.5.1.1 Settings for the locality decomposition

Given the importance of parameter δ for the behavior of the proposed LD formulation, the proper adjustment of this parameter needs to be investigated. Figure 4.13 presents the overall root mean square error, O-RMSE (Section 2.4.2), scored by LD for 10 different values of δ. Results are provided for both the basic and the archiving variants of the (1+1) EA. In addition, the results of the conventional SO formulation are shown as a baseline. It is evident from Figure 4.13 that an important increase in performance has been obtained by using LD. For the different values of δ, LD reached the best results when using the basic, non-archiving variant of the algorithm. However, even

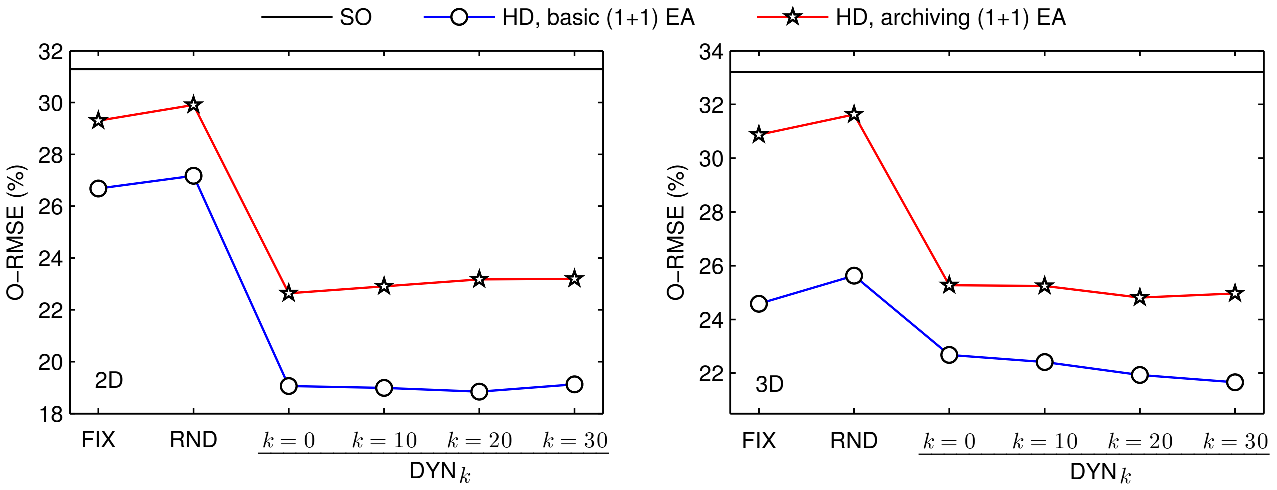


Figure 4.14: H-subsets decomposition. Evaluating different H-subsets formation strategies. Results for the (1+1) EA. Two-dimensional (left) and three-dimensional (right) test cases.

using the archiving (1+1) EA, LD performed better in all cases compared to the SO formulation. It can be seen from the plots that the lowest O-RMSE values were scored at around $\delta = 7$. Note also that the performance of the algorithms gradually declined with the increasing value of δ. Therefore, the distance parameter δ was set to 7 for further analyses presented in Sections 4.5.1.3 and 4.5.1.4.

4.5.1.2 Settings for the H-subsets decomposition

An important issue for the proposed HD formulation is how the H-subsets formation process is carried out. Different strategies have been described in Section 4.3.3: FIX, RND and DYN$_k$, for $k \in \{0, 10, 20, 30\}$. Figure 4.14 indicates the overall root mean square error (O-RMSE) achieved by the HD formulation when using these strategies. Both the basic and the archiving variants of the (1+1) EA are considered. The performance of the SO formulation is also shown as a reference. From Figure 4.14 it is possible to note that, regardless of the H-subsets formation strategy and the variant of the (1+1) EA used, the proposed HD performed better in all the cases when compared with respect to the conventional SO formulation. The lowest O-RMSE values were obtained when using the DYN$_k$ strategy. This suggests that the effect of decomposition for allowing algorithms to escape from local optima can be further enhanced by changing the search landscape dynamically throughout the evolutionary process. For the two-dimensional instances, no important differences in performance

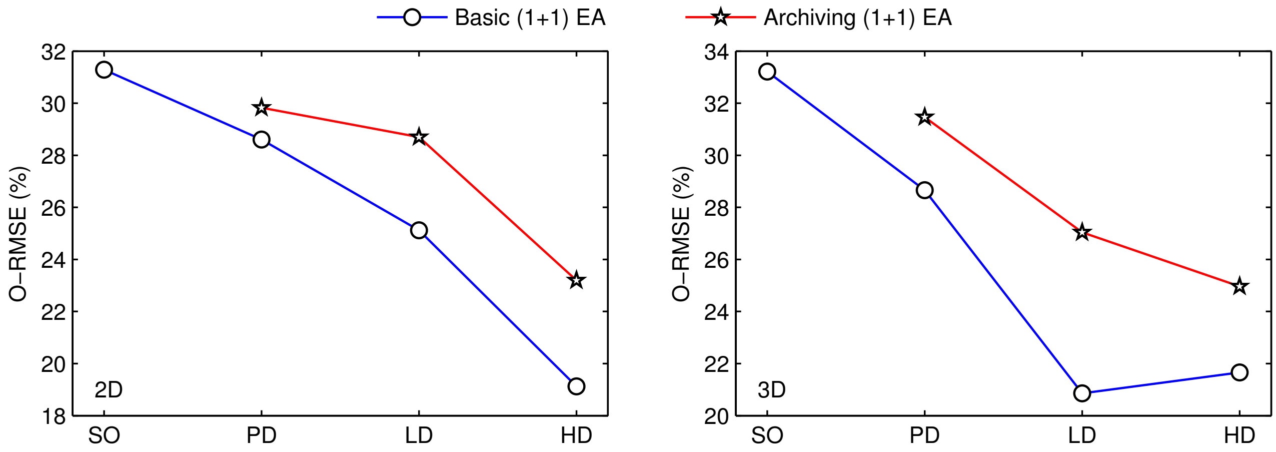


Figure 4.15: Evaluating the effects of using the archiving strategy within the (1+1) EA on the behavior of the proposed PD, LD and HD methods. Two-dimensional (left) and three-dimensional (right) test cases. The performance of the SO formulation is shown as a baseline.

could be observed when varying k. Regarding the three-dimensional case, the algorithms showed a slight positive response to the increase in the value of k. Based on these observations, the DYN$_k$ strategy, with $k = 30$, was adopted for the experiments presented in Sections 4.5.1.3 and 4.5.1.4.

4.5.1.3 The impact of archiving

This section analyzes whether or not implementing the archiving strategy can be beneficial to the performance of the proposed multi-objectivization schemes. Figure 4.15 contrasts the performance of the basic and the archiving variants of the (1+1) EA (in terms of the O-RMSE measure) when using the three multi-objective proposals (PD, LD and HD). Also, the results of the basic (1+1) EA when applying the conventional SO formulation of the HP model are presented as a reference.

From Figure 4.15, it can be seen first that an important increase in performance was obtained through multi-objectivization. The PD, LD and HD multi-objective proposals, either using the basic or the archiving (1+1) EA, scored better results when compared with respect to the SO formulation. HD reached the lowest O-RMSE values at solving the two-dimensional instances. In contrast, the LD formulation performed the best in the three-dimensional case when using the basic (1+1) EA.

Although competitive, the performance of PD, LD and HD was negatively affected by the use of the archiving strategy within the (1+1) EA. Better results were obtained in all the cases by using the basic, non-archiving variant of the algorithm. This is contrary to what can be expected in multi-objective optimization, where archiving is assumed to be essential for converging towards a set of trade-offs among the conflicting problem objectives [117, 144]. Nevertheless, in spite of being alternatively modeled and treated as a multi-objective problem, the HP model is actually a single-objective optimization problem. Maintaining an approximation set of nondominated solutions becomes not as important in this scenario since, by definition, the goal remains to solve the original single-objective problem. The performance decrease originated from the use of the archiving strategy can also be explained by the fact that archiving induces the opposite effect than that of decomposition. As pointed out by Handl *et al.*, the effect of decomposition may be partially reversed through archiving [85]. Whereas decomposition introduces plateaus of incomparable solutions, archiving can make some parts of these plateaus (again) inaccessible (this depends upon what solutions are in the archive). Therefore, the benefits of multi-objectivization for allowing algorithms to escape from local optima could potentially be mitigated by means of archiving. In this way, archiving was found to be obstructive rather than beneficial for the three proposed multi-objectivization schemes.

4.5.1.4 Comparative analysis

This section compares in detail the four studied evaluation schemes for the HP model (SO, PD, LD and HD). The use of the proposed LD and HD formulations requires the adjustment of some parameters. The influence that varying these parameters has on the search performance of the (1+1) EA was first analyzed in Sections 4.5.1.1 and 4.5.1.2. In addition, it was found in Section 4.5.1.3 that better results were obtained in all the cases when using the basic (1+1) EA, rather than the archiving variant of this algorithm. Therefore, the best identified parameter settings for LD and HD, as well as the basic (1+1) EA, were adopted for the analysis conducted in this section.

Tables 4.3 and 4.4 detail the obtained results for all two- and three-dimensional test instances. For each of the instances, these tables show the best obtained energy value (E_b), the number of performed executions where this solution quality was reached (ν) and the arithmetic mean of the scored energy values ($\bar{E}$). Also, the overall relative root mean square error (O-RMSE) is presented at the bottom of the tables in order to evaluate the general performance of the different formulations analyzed.

Table 4.3: Results scored by the basic (1+1) EA when using the four studied formulations of the HP model: SO, PD, LD and HD. Two-dimensional test cases.

			SO		**PD**		**LD**		**HD**	
Seq.	ℓ	E^*	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$
2d1	**18**	**-4**	-4 (4)	-2.72	-4 (6)	-2.74	-4 (3)	-2.73	-4 (4)	**-2.79**
2d2	**18**	**-8**	-8 (19)	-6.91	-8 (32)	-7.15	-8 (35)	-7.27	-8 (94)	**-7.93**
2d3	**18**	**-9**	-8 (16)	-7.06	-8 (62)	-7.59	-9 (3)	-7.52	-9 (62)	**-8.62**
2d4	**20**	**-9**	-9 (9)	-7.01	-9 (9)	-7.34	-9 (16)	-7.59	-9 (66)	**-8.55**
2d5	**20**	**-10**	-9 (3)	-7.02	-10 (2)	-7.23	-9 (1)	-7.13	-10 (4)	**-7.92**
2d6	**24**	**-9**	-8 (16)	-6.90	-9 (1)	-6.94	-9 (3)	-7.42	-9 (8)	**-7.51**
2d7	**25**	**-8**	-7 (32)	-5.90	-8 (6)	-5.95	-8 (8)	-6.27	-8 (26)	**-6.78**
2d8	**36**	**-14**	-13 (1)	-10.04	-13 (1)	-10.36	-13 (5)	-10.79	-13 (1)	**-11.29**
2d9	**48**	**-23**	-18 (6)	-14.44	-19 (3)	-15.70	-21 (1)	-16.71	-21 (3)	**-18.57**
2d10	**50**	**-21**	-18 (2)	-13.88	-18 (1)	-14.22	-19 (1)	-15.52	-20 (1)	**-17.06**
2d11	**60**	**-36**	-30 (2)	-24.58	-32 (1)	-25.97	-33 (1)	-28.32	-33 (3)	**-30.33**
2d12	**64**	**-42**	-29 (1)	-24.21	-30 (1)	-25.54	-32 (1)	-27.12	-32 (7)	**-29.21**
2d13	**85**	**-53**	-41 (1)	-34.13	-42 (1)	-35.08	-44 (1)	-38.59	-47 (1)	**-41.56**
2d14	**100**	**-48**	-41 (1)	-31.28	-39 (4)	-32.97	-39 (4)	-35.38	-43 (2)	**-37.65**
2d15	**100**	**-50**	-40 (1)	-31.95	-40 (3)	-33.37	-42 (1)	-35.83	-40 (19)	**-38.31**
O-RMSE			31.28%		28.60%		25.12%		**19.12%**	

Table 4.4: Results scored by the basic (1+1) EA when using the four studied formulations of the HP model: SO, PD, LD and HD. Three-dimensional test cases.

			SO		**PD**		**LD**		**HD**	
Seq.	ℓ	E^*	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$
3d1	**20**	**-11**	-11 (65)	-10.59	-11 (87)	-10.84	-11 (99)	-10.99	-11 (100)	**-11.00**
3d2	**24**	**-13**	-13 (28)	-11.59	-13 (50)	-12.09	-13 (89)	-12.89	-13 (100)	**-13.00**
3d3	**25**	**-9**	-9 (71)	-8.66	-9 (77)	-8.77	-9 (96)	-8.96	-9 (100)	**-9.00**
3d4	**36**	**-18**	-18 (12)	-15.43	-18 (23)	-16.22	-18 (68)	-17.42	-18 (88)	**-17.84**
3d5	**46**	**-35**	-30 (2)	-24.36	-30 (2)	-25.94	-33 (1)	**-28.60**	-31 (2)	-28.48
3d6	**48**	**-31**	-29 (2)	-23.18	-29 (3)	-24.84	-31 (1)	**-27.64**	-30 (1)	-27.27
3d7	**50**	**-34**	-25 (6)	-21.07	-27 (1)	-22.92	-29 (1)	**-25.27**	-29 (1)	-24.78
3d8	**58**	**-44**	-35 (1)	-27.71	-36 (1)	-30.09	-38 (1)	**-33.64**	-35 (6)	-32.60
3d9	**60**	**-55**	-48 (1)	-38.14	-48 (1)	-41.11	-47 (8)	-44.72	-49 (1)	**-45.02**
3d10	**64**	**-59**	-45 (1)	-36.20	-46 (2)	-38.84	-50 (2)	**-45.46**	-48 (3)	-43.28
3d11	**67**	**-56**	-40 (1)	-30.98	-42 (1)	-33.70	-41 (3)	**-37.94**	-40 (2)	-36.90
3d12	**88**	**-72**	-48 (2)	-37.29	-54 (1)	-41.39	-53 (3)	**-48.46**	-50 (1)	-45.33
3d13	**103**	**-58**	-41 (1)	-30.68	-42 (1)	-32.52	-43 (1)	**-37.09**	-40 (2)	-36.26
3d14	**124**	**-75**	-48 (1)	-35.45	-48 (4)	-38.28	-52 (1)	**-46.44**	-48 (2)	-43.42
3d15	**136**	**-83**	-51 (2)	-38.58	-53 (1)	-43.98	-59 (1)	**-50.27**	-56 (1)	-48.66
O-RMSE			33.21%		28.66%		**20.85%**		21.66%	

Table 4.5: Statistical analysis for comparing the performance of the (1+1) EA when using the four studied HP model's formulations.

	Two-dimensional instances															Three-dimensional instances															
	2d1	2d2	2d3	2d4	2d5	2d6	2d7	2d8	2d9	2d10	2d11	2d12	2d13	2d14	2d15	3d1	3d2	3d3	3d4	3d5	3d6	3d7	3d8	3d9	3d10	3d11	3d12	3d13	3d14	3d15	Overall
PD/SO			+	+				+	+		+	+	+	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+	+	**23**+ **0**−
LD/SO		+	+	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	**28**+ **0**−
HD/SO		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	**29**+ **0**−
LD/PD						+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	**25**+ **0**−
HD/PD		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	**29**+ **0**−
HD/LD		+	+	+	+		+	+	+	+	+	+	+	+	+		+	+	+		−	−	−		−	−	−	−	−	−	**16**+ **9**−

Additionally, the best (lowest) average energy for each of the instances and the best (lowest) O-RMSE values have been **shaded** in these tables. As shown in Table 4.3, the proposed HD multi-objectivization reached the best average performance for all the 15 two-dimensional instances. This is reflected as an O-RMSE decrease of $(31.28 - 19.12) = 12.16\%$ with respect to the conventional SO formulation of the HP model. It can also be noted that LD and PD presented a lower average energy than the SO formulation in all the cases, improving the O-RMSE measure by 6.16% and by 2.68%, respectively. The proposed LD formulation achieved the lowest average energy for 10 out of the 15 three-dimensional instances, see Table 4.4. The best results for the remaining test instances were obtained by using HD. In general, the results of the multi-objective PD, LD and HD formulations are found to be quite competitive. These approaches improved the O-RMSE measure by 4.55%, 12.36% and 11.55% with respect to the conventional SO formulation, respectively.

Finally, Table 4.5 outlines how the four different formulations compare statistically with respect to each other in all the test cases. Each row in this table compares two formulations, say A and B, which is denoted as "A/B". If a significant performance difference exists between A and B for a particular instance, the corresponding cell is marked either + or − depending on whether such a difference favors A or not. Empty cells indicate that there was not a statistically important difference between the compared approaches. The rightmost column presents the overall results of this analysis. Out of a total of 30 test instances, it can be seen from Table 4.5 that the multi-objective PD, LD and HD approaches significantly outperformed the conventional SO formulation in 23, 28 and 29 of the cases, respectively. By comparing among the proposed multi-objectivizations, the results of LD for 25 of the instances were statistically superior to those obtained by PD. Compared with respect to PD, HD significantly increased the performance of the algorithm for all but one of the test sequences (2d1). Finally, HD was found to perform significantly better than LD in 16 of the instances, while there was a significant difference in favor of LD for 9 of the three-dimensional test cases.

4.5.2 Results for a population-based algorithm

This section explores the extent to which multi-objectivization can impact on the behavior of a basic genetic algorithm (GA). The basic structure of the implemented GA is presented in Algorithm 7. First, an initial parent population $\mathcal{P}$ of size N is randomly generated. At each generation, the fittest individuals in $\mathcal{P}$ are selected for mating ($selection\text{-}for\text{-}variation$). Then, a children population $\mathcal{P}'$ is created by applying the genetic operators to the selected parents $\hat{\mathcal{P}}$. Finally, the parent and children populations are combined and the best individuals are selected to survive in order to form the new parent population ($selection\text{-}for\text{-}survival$). The selection process of the GA, which is responsible for guiding the search, will depend upon the problem formulation to be used. On the one hand, when using the conventional SO formulation, selection is to be driven by the energy value of the candidate individuals. On the other hand, when applying the GA to the proposed multi-objective formulations (PD, LD and HD), the discrimination among individuals is to be based on *nondominated sorting* and *crowding distance*, as in the *Non-dominated Sorting Genetic Algorithm II*, NSGA-II [59]. Hence, the change in the problem formulation will be determinant for the behavior of this algorithm.

Algorithm 7 Genetic algorithm.

```
choose $\mathcal{P} \subset \mathcal{X}_\mathcal{F} : |\mathcal{P}| = N$ uniformly at random
while $< stop\ condition >$ do
    $\hat{\mathcal{P}} \leftarrow selection\text{-}for\text{-}variation(\mathcal{P})$
    $\mathcal{P}' \leftarrow variation(\hat{\mathcal{P}})$
    $\mathcal{P} \leftarrow selection\text{-}for\text{-}survival(\mathcal{P} \cup \mathcal{P}')$
end while
```

Roughly, the functioning of the nondominated sorting procedure is as follows (Figure 4.16). The nondominated individuals are initially identified and isolated into the first nondominated layer, $\mathcal{L}_1$.

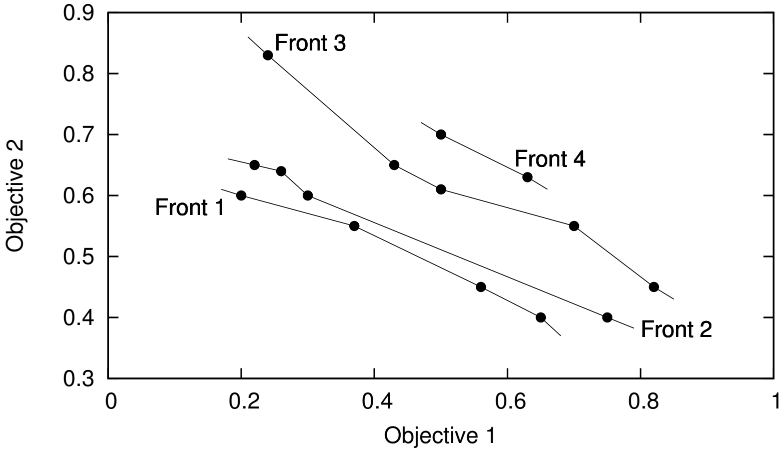


Figure 4.16: Functioning of the nondominated sorting procedure [59]. Individuals are organized into nondominated fronts, or layers, based on the Pareto-dominance relation (minimization assumed).

From the remainder of the population, the now nondominated solutions are identified and assigned to the second nondominated layer, $\mathcal{L}_2$. This process is repeated until each individual in the population is classified. At the $selection\text{-}for\text{-}survival$ stage, individuals are selected layer by layer, starting from $\mathcal{L}_1$, until completing the required number of individuals. Whenever the number of individuals in the layer under consideration exceeds the available capacity of the population, the crowding distance measure is used as a secondary discrimination criterion. This allows to promote population diversity.[11]

In the GA, protein conformations are encoded using an internal coordinates representation based on absolute moves (Section 2.3.3.3). Binary tournament selection was employed as mating strategy. The implemented genetic operators are as follows. One-point crossover is applied according to a given probability p_c. In mutation, each encoding position is randomly and independently perturbed with probability p_m. Only individuals encoding feasible protein conformations are accepted during the search process. Refer to Section 5.6.1.1 for further details on the treatment given to infeasible individuals. The initial feasible populations are generated through the backtracking strategy reported in [42]. In all cases, a maximum number of 5×10^5 solution evaluations was defined as the stopping condition and the reported results were computed over a total of 100 independent GA executions.

[11] The crowding distance is a measure of the density of individuals. This measure is computed locally as the proximity between neighboring points in the objective space [59].

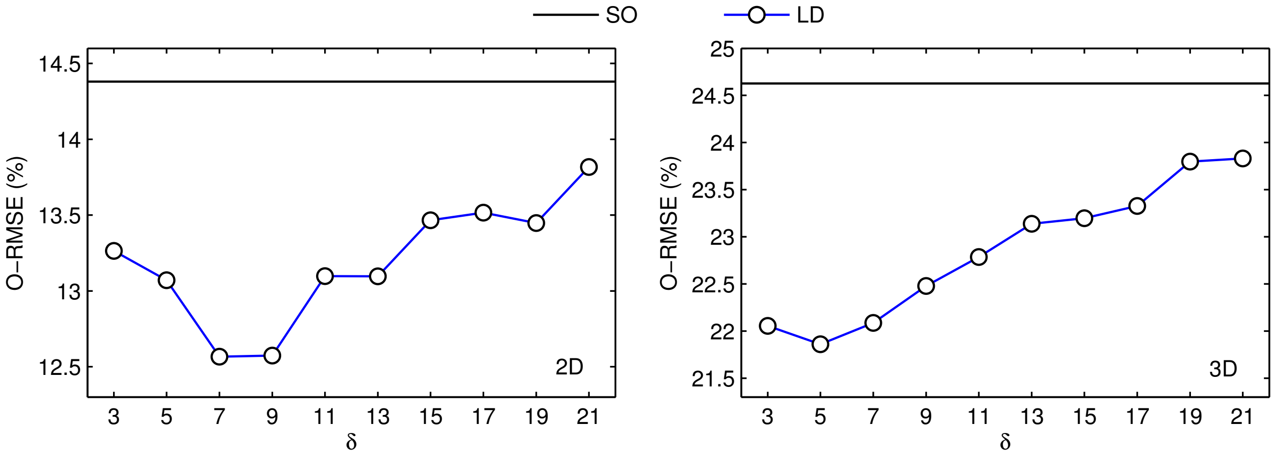


Figure 4.17: Locality decomposition. Evaluating the impact of varying the distance parameter δ on the performance of the GA. Two-dimensional (left) and three-dimensional (right) test cases.

The remainder of this section is organized as follows. Sections 4.5.2.1 and 4.5.2.2 are concerned with the proper adjustment of parameters for the proposed LD and HD formulations. In Section 4.5.2.3, the four studied formulations of the HP model are evaluated under different settings for the implemented GA. Finally, a detailed comparative analysis is presented in Section 4.5.2.4.

4.5.2.1 Settings for the locality decomposition

This section inspects how changing the value of the LD's distance parameter δ can affect the performance of the implemented GA. As indicated in Section 4.3.2, a total of 10 values for δ are explored. Figure 4.17 presents the overall root mean square error (O-RMSE) obtained by the GA when using the different considered values for δ.[12] Results of the SO formulation are shown as a baseline. The best performance of the GA on the two-dimensional instances was obtained when using $\delta = 7$, while $\delta = 5$ provided the best behavior for the three-dimensional case. These settings have been adopted for the analyses presented later in Sections 4.5.2.3 and 4.5.2.4. The performance of the GA tended to decrease as δ was increased. Note, however, that the proposed LD improved the results with respect to the conventional SO formulation regardless of the value chosen for δ.

[12] For each considered δ value, the O-RMSE presented in Figure 4.17 corresponds to the best performance obtained when evaluating a set of different parameter configurations of the GA, see Section 4.5.2.3.

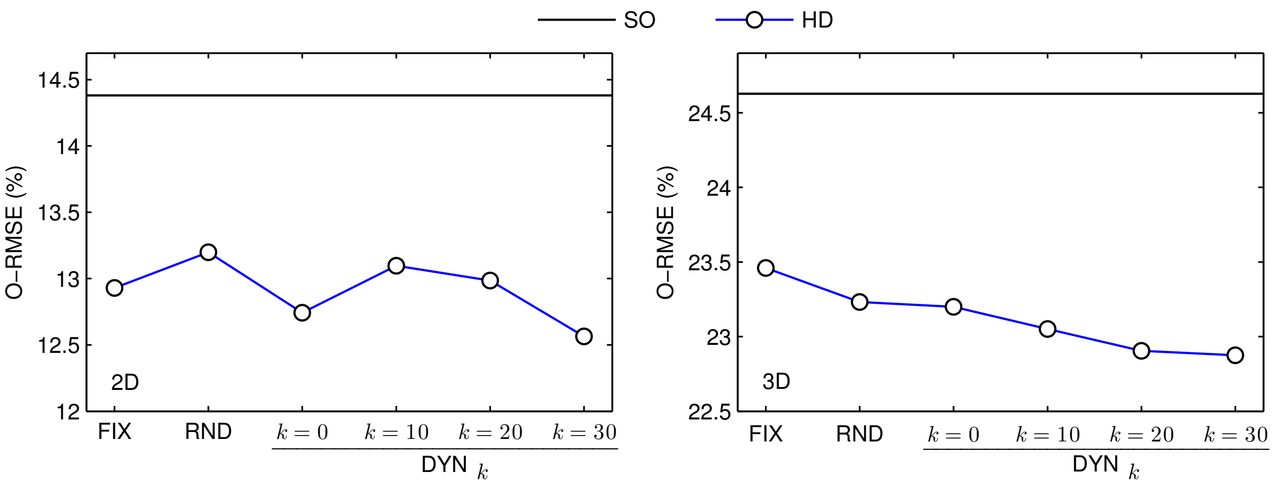


Figure 4.18: H-subsets decomposition. Evaluating different H-subsets formation strategies. Results for the GA. Two-dimensional (left) and three-dimensional (right) test cases.

4.5.2.2 Settings for the H-subsets decomposition

In the HD formulation, H amino acids are first organized into the $\mathcal{H}_1$ and $\mathcal{H}_2$ groups, the H-subsets. Different strategies to perform this task are investigated in this study: FIX, RND and DYN$_k$, for $k \in \{0, 10, 20, 30\}$; refer to Section 4.3.3 for details. Figure 4.18 displays how the performance of the GA (measured in terms of the O-RMSE) was affected by the use of these H-subsets formation strategies.[13] The performance of the conventional SO formulation is presented as a reference. Only minor variations on the GA's performance can be observed as a consequence of using the different H-subsets formation strategies. It can be seen, however, that the performance of the algorithm tends to improve when applying the DYN$_k$ strategy, particularly when focusing on the three-dimensional test cases. This suggests that dynamically changing the problem formulation enhances the ability of multi-objectivization for allowing the GA to escape from local optima, so that a positive effect is achieved in terms of the efficient exploration of the search landscape. In all the cases, the proposed HD formulation decreased the O-RMSE with regard to the SO formulation. The DYN$_k$ strategy with $k = 30$ was adopted for later experiments reported in Sections 4.5.2.3 and 4.5.2.4.

[13] For each H-subsets formation strategy, the O-RMSE value presented in Figure 4.18 corresponds to the best performance obtained when evaluating different parameter configurations of the GA, see Section 4.5.2.3.

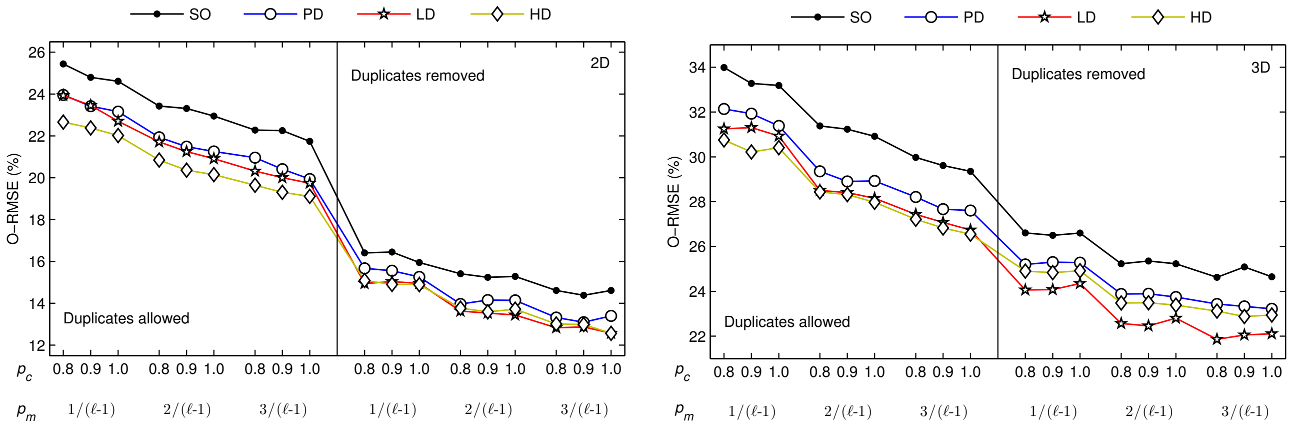


Figure 4.19: Evaluating the SO, PD, LD and HD formulations under different parameter settings for the implemented GA. Two-dimensional (left) and three-dimensional (right) test cases.

4.5.2.3 Settings for the genetic algorithm

Different parameter settings for the GA are evaluated to identify the most appropriate conditions for the compared approaches. Three values for the recombination and mutation probabilities were considered: $p_c \in \{0.8, 0.9, 1.0\}$ and $p_m \in \{\frac{1}{\ell-1}, \frac{2}{\ell-1}, \frac{3}{\ell-1}\}$. Also, the effects of preventing duplicate individuals (clones) from the population are analyzed. Thus, a total of 18 parameter configurations for the GA are investigated. The population size was fixed to $N = 100$ in all the cases.

Figure 4.19 presents the overall root mean square error (O-RMSE) obtained by the four studied formulations of the HP model when using the different GA settings. From the plots, it can be noted that the proposed PD, LD and HD multi-objectivizations performed better than the conventional SO formulation for all the different parameter configurations of the GA. It is also possible to see that LD and HD scored better results than PD in most of the cases. The lowest O-RMSE value for both the two- and the three-dimensional instances was obtained by using the LD formulation, although the superiority of LD over HD becomes more evident when focusing on the three-dimensional case. Note, however, that HD tends to perform better than LD if duplicate individuals are allowed to remain in the population. Some general observations can be made regarding the behavior of the GA. On the one hand, the algorithm seemed not to be seriously affected when varying the recombination probability. On the other hand, the GA responded positively to the increased mutation rate, being

$p_m = \frac{3}{\ell-1}$ the value which provided the best performance in all the cases. Finally, the results were significantly improved in all the cases when duplicate individuals were removed from the population.

For the detailed analysis presented in Section 4.5.2.4, the settings which allowed each of the compared formulations to reach the lowest O-RMSE value were selected. Specifically, duplicates removal was enabled and the mutation probability was set to $p_m = \frac{3}{\ell-1}$ in all cases. The recombination probability was set to (i) two-dimensional case: $p_c = 0.9$ for SO and PD, $p_c = 1.0$ for LD and HD; and (ii) three-dimensional case: $p_c = 0.8$ for SO and LD, $p_c = 1.0$ for PD, $p_c = 0.9$ for HD.

4.5.2.4 Comparative analysis

This section presents a detailed comparative analysis among the four studied evaluation schemes for the HP model. The results reported in this section consider the best parameter adjustment for the proposed LD and HD formulations, as investigated in Sections 4.5.2.1 and 4.5.2.2, and the best performing GA settings for each of the compared approaches, as derived in Section 4.5.2.3.

The results for all two- and three-dimensional test cases are provided in Tables 4.6 and 4.7. The information in these tables is organized in the same manner as in Tables 4.3 and 4.4 described in Section 4.5.1.4. Table 4.6 indicates that the use of the proposed PD, LD and HD multi-objective formulations improved the average performance of the GA for most of the two-dimensional instances. In most cases, the lowest average energy values were obtained by using the LD and HD formulations. LD and HD allowed the GA to reach also the lowest O-RMSE values, decreasing this measure by about 1.8% with respect to the SO formulation. The benefits of multi-objectivization are more perceptible from Table 4.7. The three multi-objective proposals exceeded, or at least met (all the four compared methods scored an unbeatable performance when facing instance 3d2), the results scored by the conventional SO formulation at solving the three-dimensional test cases. The most remarkable performance was presented by the proposed LD formulation, which reported the lowest average energy value for all but one of the three-dimensional instances. PD, LD and HD improved the O-RMSE measure by 1.41%, 2.77% and 1.75% with regard to the SO formulation, respectively.

Table 4.8 illustrates how the four studied HP model's formulations are statistically compared with respect to each other in all the adopted test instances. The interpretation of this table is the

Table 4.6: Results scored by the GA when using the studied SO, PD, LD and HD formulations. Two-dimensional test cases.

			SO		**PD**		**LD**		**HD**	
Seq.	ℓ	E^*	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$
2d1	**18**	**-4**	-4 (89)	-3.89	-4 (94)	-3.94	-4 (96)	**-3.96**	-4 (96)	**-3.96**
2d2	**18**	**-8**	-8 (100)	**-8.00**	-8 (100)	**-8.00**	-8 (100)	**-8.00**	-8 (100)	**-8.00**
2d3	**18**	**-9**	-9 (97)	-8.97	-9 (97)	-8.97	-9 (98)	-8.98	-9 (100)	**-9.00**
2d4	**20**	**-9**	-9 (100)	**-9.00**	-9 (100)	**-9.00**	-9 (100)	**-9.00**	-9 (100)	**-9.00**
2d5	**20**	**-10**	-10 (90)	-9.80	-10 (99)	-9.98	-10 (99)	-9.98	-10 (100)	**-10.00**
2d6	**24**	**-9**	-9 (86)	-8.86	-9 (91)	-8.91	-9 (92)	**-8.92**	-9 (89)	-8.89
2d7	**25**	**-8**	-8 (68)	-7.65	-8 (56)	-7.52	-8 (85)	**-7.84**	-8 (80)	-7.80
2d8	**36**	**-14**	-14 (1)	-11.22	-13 (9)	-11.30	-13 (15)	-11.53	-13 (17)	**-11.54**
2d9	**48**	**-23**	-21 (7)	-17.92	-22 (2)	-18.29	-21 (10)	-18.22	-21 (11)	**-18.30**
2d10	**50**	**-21**	-21 (1)	-17.99	-21 (3)	-18.15	-21 (4)	-18.24	-21 (4)	**-18.42**
2d11	**60**	**-36**	-32 (11)	-29.14	-33 (7)	**-30.11**	-34 (1)	-30.05	-33 (3)	-29.91
2d12	**64**	**-42**	-35 (4)	-30.31	-36 (1)	-31.15	-37 (3)	**-31.77**	-36 (1)	-30.82
2d13	**85**	**-53**	-47 (1)	-40.96	-49 (1)	-41.88	-48 (2)	**-41.99**	-47 (2)	-41.69
2d14	**100**	**-48**	-40 (3)	-35.04	-41 (1)	-35.74	-42 (2)	**-35.78**	-42 (1)	-35.34
2d15	**100**	**-50**	-41 (2)	-35.91	-42 (1)	**-37.07**	-42 (1)	-36.80	-42 (2)	-36.78
O-RMSE			14.38%		13.10%		12.57%		**12.56%**	

Table 4.7: Results scored by the GA when using the studied SO, PD, LD and HD formulations. Three-dimensional test cases.

			SO		**PD**		**LD**		**HD**	
Seq.	ℓ	E^*	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$
3d1	**20**	**-11**	-11 (99)	-10.99	-11 (100)	**-11.00**	-11 (100)	**-11.00**	-11 (100)	**-11.00**
3d2	**24**	**-13**	-13 (100)	**-13.00**	-13 (100)	**-13.00**	-13 (100)	**-13.00**	-13 (100)	**-13.00**
3d3	**25**	**-9**	-9 (95)	-8.95	-9 (99)	-8.99	-9 (99)	-8.99	-9 (100)	**-9.00**
3d4	**36**	**-18**	-18 (28)	-16.44	-18 (33)	-16.58	-18 (41)	**-16.78**	-18 (37)	-16.68
3d5	**46**	**-35**	-31 (2)	-27.24	-31 (1)	-27.59	-32 (2)	**-27.84**	-33 (1)	-27.44
3d6	**48**	**-31**	-30 (1)	-26.12	-30 (2)	-26.74	-31 (1)	**-27.00**	-31 (1)	-26.58
3d7	**50**	**-34**	-31 (1)	-26.60	-30 (8)	-26.93	-31 (3)	**-27.46**	-30 (9)	-26.77
3d8	**58**	**-44**	-37 (2)	-32.01	-39 (1)	-32.42	-38 (2)	**-33.33**	-38 (3)	-32.77
3d9	**60**	**-55**	-48 (1)	-41.85	-50 (1)	-42.42	-48 (3)	**-43.54**	-51 (1)	-43.02
3d10	**64**	**-59**	-53 (1)	-45.31	-52 (3)	-46.59	-54 (3)	**-48.28**	-53 (1)	-46.74
3d11	**67**	**-56**	-40 (3)	-35.79	-42 (2)	-36.56	-44 (1)	**-37.42**	-43 (1)	-36.75
3d12	**88**	**-72**	-50 (1)	-42.88	-53 (1)	-44.69	-54 (1)	**-46.24**	-54 (1)	-44.24
3d13	**103**	**-58**	-41 (1)	-33.32	-41 (1)	-33.88	-39 (6)	**-34.80**	-40 (2)	-34.79
3d14	**124**	**-75**	-51 (1)	-39.12	-51 (1)	-41.33	-50 (2)	**-42.73**	-52 (1)	-41.95
3d15	**136**	**-83**	-51 (3)	-42.94	-57 (1)	-44.45	-53 (3)	**-45.61**	-56 (1)	-45.35
O-RMSE			24.63%		23.22%		**21.86%**		22.88%	

Table 4.8: Statistical analysis for comparing the performance of the GA when using the four studied HP model's formulations.

	Two-dimensional instances															Three-dimensional instances															
	2d1	2d2	2d3	2d4	2d5	2d6	2d7	2d8	2d9	2d10	2d11	2d12	2d13	2d14	2d15	3d1	3d2	3d3	3d4	3d5	3d6	3d7	3d8	3d9	3d10	3d11	3d12	3d13	3d14	3d15	Overall
PD/SO					+						+	+	+	+	+						+				+		+		+	+	11+ 0−
LD/SO					+		+	+			+	+	+	+	+					+	+	+	+	+	+	+	+	+	+	+	19+ 0−
HD/SO					+		+	+		+	+		+		+			+					+	+	+	+	+	+	+	+	16+ 0−
LD/PD							+					+										+	+	+	+	+	+	+	+		10+ 0−
HD/PD							+																					+			2+ 0−
HD/LD												−										−			−		−				0+ 4−

same as for Table 4.5 described at the end of Section 4.5.1.4. The statistical analysis reveals that the improvements achieved by PD, LD and HD were significant with respect to the conventional SO formulation in 11, 19 and 16 of the adopted test cases, respectively. It can also be noted from Table 4.8 that the advantages of multi-objectivization tend to become more evident as the size of the problem (sequence length) increases. HD performed statistically better than PD in only 2 out of the 30 instances (2d7 and 3d13). Finally, LD presented the best overall behavior, scoring significantly better results in 10 and 4 of the instances when compared with respect to PD and HD, respectively.

4.6 Discussion and conclusions

Multi-objectivization concerns the restatement of a single-objective problem in an alternative multi-objective form, which can facilitate the process of finding a solution to the original problem. This concept represents a current and promising research direction which has led to the development of more competitive search mechanisms. In this chapter, multi-objectivization was applied to the particular case of study of this research project, the HP model for protein structure prediction. Three different multi-objectivization schemes for the HP model were proposed, all of them based on the decomposition of the conventional objective function of the problem. The first approach, called the parity decomposition (PD), relies on the fact that topological interactions on the lattice are only possible between amino acids whose sequence positions are of opposite parity. In the second proposal, the locality decomposition (LD), the decomposition of the HP model's objective is carried out by segregating local from nonlocal amino acid interactions. Such a locality notion relates to the sequence distance between the interacting amino acids. Finally, in the H-subsets decomposition (HD) hydrophobic amino acids are first organized into different groups, the H-subsets. Then, the HP model's energy function is decomposed based on the correspondence of amino acids to the H-subsets.

The study presented in this chapter has been divided into two main parts. The first part was devoted to the investigation of the potential effects of multi-objectivization, as a means of gaining insight into how this transformation can influence the behavior of search algorithms. When multi-objectivization is achieved through the decomposition of the original objective function of the problem, as considered in this study, incomparability among candidate solutions can be introduced [85]. That is, originally comparable solutions may become incomparable when evaluated under the new multi-objective formulation of the problem. As a first step in understanding and quantifying such an effect, it was explored the extent to which incomparability may arise between different fitness classes. To this end, a large set of sampled solution pairs were evaluated, out of which a considerable number became incomparable as a consequence of the problem transformation. It was also found that the more distant the fitness classes for a given pair of solutions, the lower the likelihood that the comparability relation between these solutions can be affected by multi-objectivization.

Introducing incomparability among solutions can be alternatively understood as increasing the neutrality of the fitness landscape. Therefore, a detailed fitness landscape analysis was conducted to investigate how multi-objectivization impacts on such an important problem characteristic. A large number of neutral networks (NNs) were sampled, and the main findings achieved during their analysis can be summarized as follows. By rising the neutrality degree of solutions, multi-objectivization led to the formation of neutral connections between NNs from different fitness classes. That is, when originally comparable neighboring solutions become incomparable by multi-objectivization, their corresponding NNs are merged together into larger connected neutral areas of the landscape.

The aforementioned effects of multi-objectivization may lead to different implications from the perspective of a search algorithm. On the one hand, the introduction of neutrality into the fitness landscape can be reflected as an enhancement in the exploration behavior of algorithms. That is, by reducing the so-called *selective pressure* [5], an algorithm can be allowed to move through inferior fitness classes as a means of escaping from local optima. On the other hand, the increase in neutrality can also be understood as an important loss in gradient information, so that multi-objectivization may prevent search algorithms from identifying promising directions in some of the cases.

Understanding the potential effects and consequences of multi-objectivization could lead to the design of more effective search algorithms. To the best of the author's knowledge, in most of the reported applications of multi-objectivization the original evaluation scheme of the problem is completely replaced by the alternative multi-objectivized one. This corresponds also to the approach assumed in the second part of this study, as it will be discussed later in this section. While the complete replacement of the evaluation scheme has reported very promising results in the literature, the above analysis suggests that a better strategy may involve applying multi-objectivization only under certain conditions during the search process; *e.g.*, when stagnation at a local optimum has been detected. Such a strategy could benefit from the potential effects of multi-objectivization as a means of escaping from local optima, at the same time that the gradient information is preserved in order to drive the search process in an effective manner. In the context of the HP model for protein structure prediction, a high degree of neutrality is inherently induced by the conventional evaluation scheme, as it was observed along the conducted analysis. Thus, another interesting approach may consist of alternating the use of multi-objectivization with the use of some other problem formulation specifically designed to cope with neutrality; such as the alternative energy functions evaluated in Chapter 3. Exploring these kind of strategies related to the partial (rather than total) use of multi-objectivization can be seen as a promising direction for future research.

The second part of this chapter analyzed the advantages of multi-objectivization in terms of the performance of search algorithms. The three proposed multi-objectivization schemes for the HP model were compared and evaluated with respect to the conventional single-objective formulation of the problem. Two different evolutionary algorithms (EAs) were considered, namely, a basic (1+1) EA and a genetic algorithm (GA). In this way, both single-solution-based and population-based search methods have been covered in this analysis. As a result, the use of alternative multi-objective formulations of the problem significantly increased the average performance of the implemented EAs in most of the conducted experiments. The obtained results not only demonstrate the effectiveness of the proposed multi-objectivization schemes for the HP model, but also give further support to the suitability of multi-objectivization to address the multimodality challenge of fitness landscapes. In this study, only basic EAs were considered. From the obtained results, however, it is expected that the proposed multi-objectivization schemes for the HP model can be used to improve also the performance of established state-of-the-art algorithms for solving this problem. This issue needs to be thoroughly investigated in order to derive more general conclusions.

To the best of the author's knowledge, the proposed PD, LD and HD formulations represent the first efforts on the use of multi-objective optimization methods to address the particular HP model of the protein structure prediction problem. In addition, no previous work has been reported where the potential effects of multi-objectivization are investigated through the explicit sampling and evaluation of the characteristics of the fitness landscape (in this case by focusing on neutrality). Although such an analysis focused on a particular case of study, most of the findings regarding the fitness landscape transformation can be generalized to other problem domains. In this way, by using the HP model of the PSP problem as an example, this research work is expected to contribute to the general understanding of multi-objectivization.

5

Handling infeasible protein conformations by multi-objective optimization

5.1 Introduction

Evolutionary computation methods and other metaheuristic algorithms have been successfully used to solve complex optimization problems which arise in a diversity of scientific and engineering applications. Often, however, optimization involves not only to reach the best value for a given objective function (or set of objective functions), but also to satisfy a certain set of predefined requirements called constraints. Therefore, additional mechanisms need to be implemented within these algorithms in order to search effectively through this kind of constrained solution spaces.

In the HP model of the protein structure prediction problem, as it is discussed in detail in Section 2.3.3.2, a feasible protein conformation is defined as an embedding of the protein chain on a given lattice, such that this embedding presents *connectivity* and *self-avoidance*. The connectivity property is implicitly satisfied by using an internal coordinates representation of the protein conformations, either based on absolute or relative moves (as described in Section 2.3.3.3). One of the main sources of difficulty in this problem, however, lies in the fact that using the existing problem representations a significant portion of the solution space encodes infeasible (non-self-avoiding) protein structures. Hence, it is important to devise effective mechanisms for handling the self-avoidance constraint. Two main research directions have been adopted to cope with this issue. On the one hand, the search can be confined to the space of only feasible, self-avoiding protein conformations. On the other hand, infeasible protein conformations can also be taken into consideration throughout the optimization process. From the literature, however, it is not possible to identify a clear consensus on which of the two directions, *i.e.*, to avoid or to consider infeasible conformations, could lead to the development of more efficient metaheuristic algorithms for solving this problem [42, 57, 66, 122, 192].

Premised upon the belief that infeasible conformations can provide valuable information for guiding the search process, this research work inquires into the use of multi-objective optimization as an alternative constraint-handling strategy for the HP model. Particularly, the self-avoidance constraint of the HP model is treated as a supplementary optimization criterion. Therefore, this originally constrained single-objective problem is transformed into an unconstrained multi-objective problem.[1] Using such an alternative multi-objective formulation, infeasible solutions can become incomparable with respect to feasible ones, having thus better opportunities for participating during the search process. In the literature, this effect of considering infeasible protein conformations during optimization has been achieved by implementing a penalty strategy [57, 113, 122, 142, 177]. In contrast to the penalty approach, which represents also one of the most widely used techniques in the constraint-handling literature, the multi-objective method, in essence, does not require the fine-tuning of parameters such as the penalty factors;[2] in the penalty strategy, finding the right balance between objective function and penalty values has been regarded to be a difficult optimization problem itself [161, 186].

In the first part of this chapter, a detailed analysis is conducted in order to investigate the potential effects of the problem transformation from the perspective of the fitness landscape. More specifically, it is evaluated how the use of the multi-objective problem formulation impacts on *neutrality*, an

[1] The process of restating a single-objective problem as a multi-objective problem is usually referred to as *multi-objectivization* in the specialized literature [119], as discussed previously in Section 2.2.3 and Chapter 4.

[2] As it will be seen later in this chapter, however, the multi-objective constraint-handling strategy may also require additional parameters or the combination with other mechanisms for biasing purposes.

important property of the fitness landscape. Such an analysis is expected to contribute to the general understanding of the functioning of the multi-objective constraint-handling technique. It has been argued that the multi-objective approach to constraint-handling could be rather ineffective if a search bias towards the feasible region is not introduced [188]. Therefore, the second part of this chapter is concerned with the study of different mechanisms which can be employed for providing the multi-objective strategy with such a search bias. Finally, the third and last part of this research work explores the suitability of the multi-objective method by focusing on the performance of search algorithms. A comparative analysis is presented where the multi-objective approach is evaluated with respect to two commonly adopted techniques from the specialized literature; namely, the use of a rejecting strategy where only feasible solutions are considered, and the use of a penalty strategy. Both single-solution-based and population-based algorithms have been utilized along this analysis.

The remainder of this chapter is structured as follows. Related work is reviewed in Section 5.2. The proposed multi-objective constraint-handling technique is described in Section 5.3. Section 5.4 presents the analysis with regard to the fitness landscape transformation. The search bias issue is addressed in Section 5.5. The comparative study which focuses on performance is covered in Section 5.6. Finally, Section 5.7 discusses the main findings and presents the conclusions of this study.

5.2 Related work

In the literature, two basic directions have been taken to address the self-avoidance constraint which relates to the feasibility of protein conformations in the HP model. On the one hand, the search can concentrate on the feasible space; that is, considering only solutions encoding self-avoiding conformations. This is usually accomplished either (i) by adapting the variation operators to iterate until new feasible conformations are generated, *i.e.*, infeasible conformations are always rejected [31, 33, 49, 53, 66, 224]; (ii) by using specialized operators which are closed on the feasible space, *i.e.*, always transforming feasible conformations into other feasible conformations [42, 128, 219]; or (iii) by implementing repairing procedures in order to convert from infeasible to feasible conformations [29, 42, 106, 192]. These three constraint-handling strategies can be referred to as the *rejecting*, *preserving* and *repairing strategies*, respectively, according to the classification presented by Talbi [218]. A combination of different strategies can also be adopted. In [31, 33], for example, the crossover operator was adapted to reject infeasible conformations, while mutation was based on the specialized pull move transformation proposed in [128]. On the other hand, infeasible conformations can also be allowed to participate during the search process. This is commonly achieved by implementing a *penalty strategy*, where the energy value of a candidate conformation suffers a decrease according to the number of collisions (overlaps) in encoded protein structure [57, 113, 122, 142, 177].

It has been argued that the path from one compact feasible conformation to another, can be significantly shorter if the search is allowed to proceed through the space of infeasible conformations [122]. This has been, perhaps, the main motivation for applying penalty strategies when solving the HP model of the PSP problem. An example of this scenario was given by Krasnogor *et al.* [122]. Also, the authors provided some guidelines on how to design an appropriate penalty function for the HP model. Nevertheless, no experimental results were reported to support these recommendations. In [66], Duarte-Flores and Smith compared between the performance of two variants of a genetic algorithm (GA). The first GA implemented a rejecting strategy, iterating crossover and mutation until feasible offspring were generated. The second GA used a penalty function adhering to the guidelines provided by Krasnogor *et al.* [122]. As a result, a better performance was observed from the use of the first variant of the GA, where infeasible conformations were always discarded.

Using a GA, Cotta compared the use of a penalty function with respect to two alternative constraint-handling approaches [42]. In the first, referred to as the feasible-space approach, the crossover and mutation operators were adapted to produce only feasible offspring. In the second alternative, infeasible offspring were accepted as the output of variation operators, but they were subsequently processed using a repairing procedure. Both the two alternative approaches were based on a backtracking algorithm. As reported in [42], better results were obtained in most of the cases using the penalty method when compared to the feasible-space approach. In contrast, the best overall performance of the implemented GA was obtained by using the repairing procedure.

Almeida *et al.* [57] explored the influence of allowing infeasible conformations on the performance of an immune-based algorithm (IA). In a first variant of the IA, only feasible conformations were permitted. In a second variant of the IA, infeasible individuals were accepted during the initialization process and when new random individuals were required as a consequence of applying the *aging operator* (infeasible individuals were penalized). However, in both variants of the IA the *hypermutation* and *hypermacromutation* operators were adapted to produce only feasible conformations; infeasible mutations were always rejected. The inclusion of infeasible conformations slightly increased the performance of the IA in most of the cases, while significantly reducing the computational effort.

Santos and Diéguez [192] evaluated the advantages of incorporating a repairing strategy into their differential evolution (DE) algorithm. In an initial DE implementation, all infeasible conformations were simply penalized and assigned a fitness value of 0. In a second DE variant, two different repairing operators were implemented; the first working on the amino acid coordinates (phenotype space), and the second acting on the conformation encoding (genotype space). These repairing operators, however, were not based on a backtracking strategy as those explored by Cotta [42]; whenever the position of the colliding amino acids could not be repaired, the infeasible conformations were allowed to remain in the population. The reported results indicate that the use of the repairing operators improved the search performance of the proposed DE algorithm.

Summarizing, there is not strong evidence in the literature (from the author's point of view) regarding whether it can be better to allow or to prevent infeasible protein conformations from being considered during the search process. Rather, from the above described works, it is possible to note that very different and, to some extent, contradictory results have been reported in this respect. One of the aims of the present study is to contribute in providing further insight into this matter.

Finally, it is important to remark that the use of multi-objective optimization concepts for handling constraints is not a novel idea. For recent reviews on this topic, the reader can be referred to [161, 202]. Nevertheless, it was not until the research work reported in this chapter, to the best of the author's knowledge, that the multi-objective approach to constraint-handling is applied to the particular HP model of the protein structure prediction problem.

5.3 Constraint-handling by multi-objective optimization

It is the author's belief that considering infeasible protein conformations during optimization can boost the performance of metaheuristics for solving PSP under the HP model (arguments on this respect have also been given in the literature [122]). Therefore, it is important to devise new constraint-handling mechanisms, which allow these algorithms to exploit the vast amount of infeasibility that the HP model involves, as a means of steering the search process in a more effective manner.

The use of multi-objective optimization is here explored as an alternative constraint-handling strategy for the HP model. The HP model is restated in multi-objective form by incorporating an additional objective function which accounts for the problem constraints. In this way, this originally constrained single-objective optimization problem is transformed into a unconstrained multi-objective one. More formally, a two-objective formulation of the problem, $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x})]^T$, is defined as follows ($\mathbf{x} \in \mathcal{X}$):

$$f_1(\mathbf{x}) = E(\mathbf{x}), \tag{5.1}$$

$$f_2(\mathbf{x}) = Collisions(\mathbf{x}), \tag{5.2}$$

where the objective functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are both to be minimized; $E(\mathbf{x})$ represents the conventional energy (objective) function of the HP model, as defined in Section 2.3.3.1; and $Collisions(\mathbf{x})$ denotes the total number of colliding amino acid pairs (a_i, a_j) such that both a_i and a_j were mapped to the same lattice coordinates in the protein structure encoded by $\mathbf{x}$.

Using the above described multi-objective formulation of the HP model, all feasible solutions $\mathbf{x} \in \mathcal{X}_{\mathcal{F}}$ will feature a value of $f_2(\mathbf{x}) = 0$. Thus, the original characteristics of the feasible areas of the fitness landscape are preserved. That is, the Pareto-dominance relation induces the conventional rank ordering among feasible solutions based on the original optimization objective (f_1). Moreover, since feasible solutions present the best possible value in f_2, an infeasible solution (with $f_2 > 0$) will never be preferred over a feasible solution under the multi-objective formulation. In general, however, this alternative formulation of the problem will lead to the explicit consideration of trade-offs between the two defined criteria, f_1 and f_2. An infeasible conformation $\mathbf{x}_1$ may become incomparable, *i.e.*, nondominated in the Pareto sense, with respect to a feasible conformation $\mathbf{x}_2$. This depends upon how $\mathbf{x}_1$ and $\mathbf{x}_2$ compare to each other with regard to the primary objective function f_1. Therefore, the multi-objective approach for handling constraints allows infeasible protein conformations to compete against feasible ones, being potentially accepted and exploited during the search process.

5.4 Fitness landscape transformation

Whereas infeasible solutions are usually regarded and treated as inferior, or even as inadmissible solutions during the search process, such a distinction between feasible and infeasible solutions is not captured when handling constraints by multi-objective optimization. As discussed in Section 5.3, the multi-objective strategy for handling constraints allows infeasible solutions to become incomparable, under certain conditions, with respect to feasible ones. Such an effect originated from the problem transformation leads to an increase in the neutrality of the fitness landscape. That is, given a feasible solution $\mathbf{x} \in \mathcal{X}_{\mathcal{F}}$, some of the surrounding infeasible solutions may become incomparable (*i.e.*, neutral) with regard to $\mathbf{x}$, thus becoming members of its neutral neighborhood, $\mathcal{N}_n(x)$.

In this section, an analysis is conducted with the aim of investigating the extent to which the use of the multi-objective constraint-handling strategy impacts on the neutrality property of fitness landscapes in the HP model. A fitness landscape is defined by a triplet $(\mathcal{X}, \mathcal{N}, \xi)$, as described in detail in Section 2.2.4. Two variants of the evaluation scheme ξ have been considered: the conventional single-objective (SO) formulation of the problem, and the alternative multi-objective (MO) formulation that handles constraints. In this way, by analyzing and comparing the landscapes induced by these two different evaluation schemes, it will be possible to assess and to gain further understanding of the effects that the studied problem transformation involves.[3]

[3] Both $\mathcal{X}$ and $\mathcal{N}$ were fixed during the analysis here presented. $\mathcal{X}$ is defined by the relative moves encoding described in Section 2.3.3.3. $\mathcal{N}(\mathbf{x})$ is given by all possible single-variable perturbations of $\mathbf{x}$. Thus, $|\mathcal{N}(\mathbf{x})| = 2(\ell - 2)$ and $|\mathcal{N}(\mathbf{x})| = 4(\ell - 2)$ in the two- and three-dimensional lattices, respectively.

Table 5.1: Details of the sample sets generated for instances 2d4 and 3d1. Instance 2d4 involves 10 different fitness classes, while instance 3d1 involves 12.

	Fitness class												
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Total**
Sequence 2d4	119	119	119	119	119	118	118	119	48	2	-	-	$1,000$
Sequence 3d1	84	83	83	84	83	84	83	83	83	84	83	83	$1,000$

The analysis presented in this section draws inspiration from the previous study reported in Section 4.4, where the effects of another alternative single-objective to multi-objective transformation of the problem were examined. Such as in Section 4.4, the term fitness is used in this section for referring to the quality of (feasible) candidate solutions under the conventional SO evaluation scheme of the HP model.[4] The fitness of a solution $\mathbf{x} \in \mathcal{X}_{\mathcal{F}}$ is given by the total number of H-H topological contacts that its encoded protein structure presents, $Fitness(\mathbf{x}) = HHtc(\mathbf{x})$. Similarly, $\mathbf{x}$ is said to belong to the fitness class c, if it holds that $Fitness(\mathbf{x}) = c$. The remaining of this section proceeds as follows. The implemented sampling methodology and the adopted settings are first described in Section 5.4.1. Then, the results of the performed analysis are discussed in Section 5.4.2.

5.4.1 Sampling methodology and settings

Due to the high computational costs involved, and given also the space requirements for reporting results, this analysis has focused on two (relatively) small problem instances: the 2d4 and 3d1 test sequences for the two- and three-dimensional lattices, respectively (see Section 2.4.1).[5] It is expected, however, that similar results can be obtained by replicating this analysis to other different problem instances. The performed analysis relies on an initial sample of solutions for each of the considered test cases. These samples have been collected following the methodology described previously in Chapter 4 (refer to Section 4.4.1). The size of the sample was fixed to $M = 1,000$ for both the 2d4 and 3d1 instances. Details of the obtained sample sets are provided in Table 5.1.

[4] Although an alternative multi-objective formulation of the HP model is implemented as a constraint handling strategy, the goal remains always to solve the original single-objective problem.

[5] The search space (based on the used relative moves encoding) is, however, vast: 3^{18} for 2d4, and 5^{18} for 3d1.

Fitness landscapes have been explored in this study by partially computing the neutral network (NN) of all the sampled solutions. The implemented methodology for the partial computation of the NNs has been described in detail in Section 4.4.3. The maximum allowed depth level in this procedure was set to $maxDepth = 5$ (for both the 2d4 and 3d1 instances) in order to alleviate the computational burden. The NN of the sampled solutions has been computed by using both, the SO and MO evaluation schemes, as the bases for neutrality verification. In this way, changing the problem formulation from SO to MO reflects as an alteration in the properties of the obtained NNs. In the remainder of this section, the NN for a given solution $\mathbf{x}$ will be either referred to as $NN_{SO}(\mathbf{x})$ or $NN_{MO}(\mathbf{x})$, depending on whether it has been computed based on the SO or MO evaluation schemes. It is important to note that under the SO evaluation scheme all infeasible solutions are assumed to be inferior to any feasible one, so that $NN_{SO}(\mathbf{x})$ comprises only feasible states. In contrast, $NN_{MO}(\mathbf{x})$ may also involve infeasible nodes; this is because using the MO formulation of the problem an infeasible solution may become part of the neutral neighborhood, and therefore of the NN, for a feasible solution, as discussed in the preamble of Section 5.4.

5.4.2 Results

This section presents the results of the fitness landscape analysis performed in order to investigate the effects of implementing the MO constraint-handling strategy. In Section 5.4.2.1, this is investigated in terms of how the problem transformation impacted on the neutrality degree of solutions. This is captured by means of the average neutrality ratio. Then, Section 5.4.2.2 evaluates the extent to which such an alteration on the neutrality degree of solutions lead to the increase in the size of the NNs. Finally, the increase in the size of the NNs are explained in Section 5.4.2.3 as a result of the connectivity that the problem transformation establishes between different NNs.

5.4.2.1 Average neutrality and infeasibility ratios

In this section, the *average neutrality ratio* (ANR) is explored as a first step in analyzing the degree to which neutrality is affected by the studied problem transformation. The ANR measure was previously

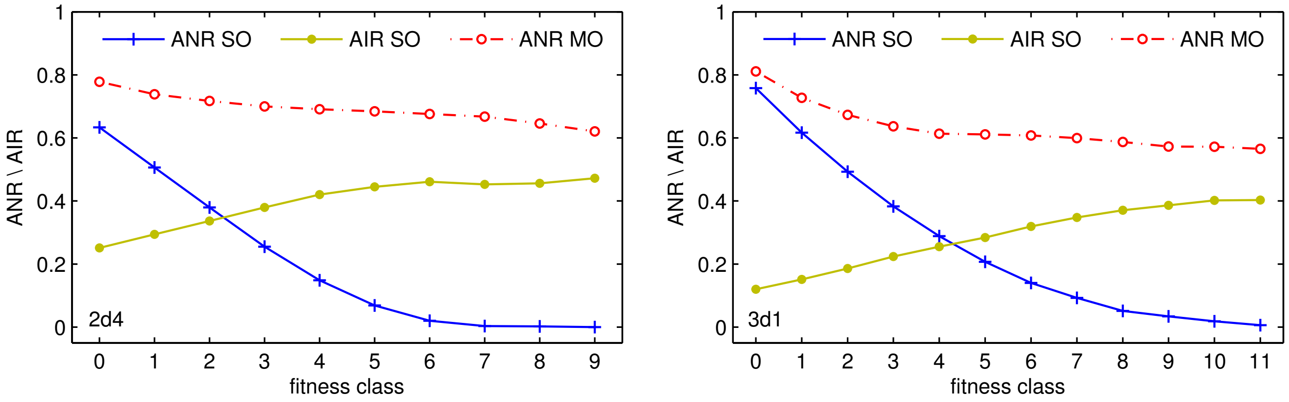


Figure 5.1: Average neutrality (ANR) and infeasibility (AIR) ratio of the sampled neutral networks. Two-dimensional 2d4 instance (left) and three-dimensional 3d1 instance (right).

defined in Section 4.4.3.1. The ANR measure was computed for all the sampled NN_{SO} and NN_{MO} networks, so that it will be possible to contrast the neutrality that the two different evaluation schemes (SO and MO) produce. In addition, the *average infeasibility ratio* (AIR) is investigated for the SO problem formulation. AIR is defined analogously to ANR, but calculated from the number of infeasible (rather than neutral) neighbors of solutions in the NN.[6] The obtained results are presented in Figure 5.1, where the mean ANR and AIR values appear organized (in the x-axis of the plots) based on the fitness class of the sampled solution used as the starting point for the NN computation.

From the ANR and AIR values obtained through the use of the SO formulation ("ANR SO" and "AIR SO" curves in the plots, respectively), it is possible to highlight some general tendencies with regard to the neutrality and infeasibility of the HP model's fitness landscapes. On the one hand, the poorer the quality of a solution, the greater tends to be the number of neutral mutations that the solution can produce. This is suggested by the high ANR values scored for the lowest fitness classes, which rapidly decreased with the increase in fitness. These results confirm previous findings discussed in Section 4.4.3.1. On the other hand, infeasibility becomes more abundant as superior fitness classes are considered. Solutions at the best fitness classes are usually surrounded by infeasible neighbors; as the obtained AIR values indicate, between 40% and 50% of the neighborhood for solutions at the best fitness classes is composed of infeasible states. The above observations can be explained by the

[6] It is worthwhile to remember that the MO evaluation scheme does not distinguish between feasible and infeasible solutions. Hence, the AIR measure applies only to the SO formulation of the problem.

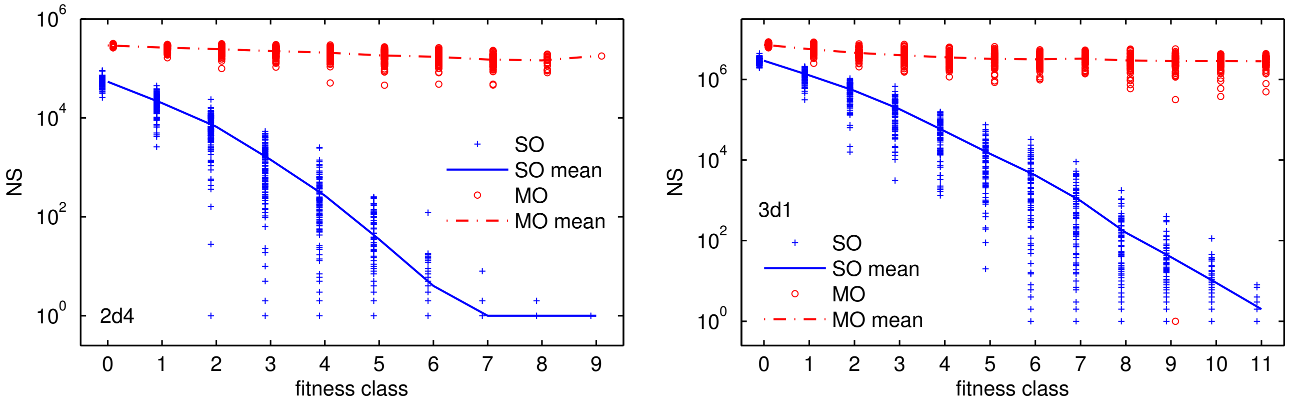


Figure 5.2: Size of the sampled neutral networks (NS). Two-dimensional 2d4 instance (left) and three-dimensional 3d1 instance (right).

fact that fitness, given in terms of the number of H-H topological contacts in this case, is directly related to the compactness of the encoded protein conformations. The higher the fitness value, the more compact the encoded protein structure tends to be. Hence, it is reasonable to conjecture that most perturbations to the encoding of a compact conformation could lead either to an infeasible solution, or to a less folded state of the protein chain which worsens the fitness. Finally, the use of the MO formulation of the problem reports an important increase in the ANR measure, with respect to the SO formulation, for all fitness classes of the two considered test instances. By introducing incomparability between feasible and infeasible solutions, a substantial fraction of the infeasibility has been translated into landscape neutrality as a consequence of the problem transformation.

5.4.2.2 *Size of the neutral networks*

This section investigates how the change in the problem formulation has impacted on the size (number of solutions) of the NNs. Figure 5.2 reports the size of each computed NN_{SO} and NN_{MO} network, as well as the arithmetic mean for the different fitness classes. Notice that results in the plots are presented in a logarithmic (base 10) scale. Figure 5.2 reveals that fitness landscapes in the HP model are characterized by a vast amount of the neutrality, as it has also been pointed out in Section 4.4.3.2. Even though a low maximum allowed depth level ($maxDepth = 5$) was defined during the NN sampling, using the SO formulation the NNs at fitness class 0 are composed of about 10^5 and 10^6 solutions for the 2d4 and 3d1 instances, respectively. The high neutrality which can be found at

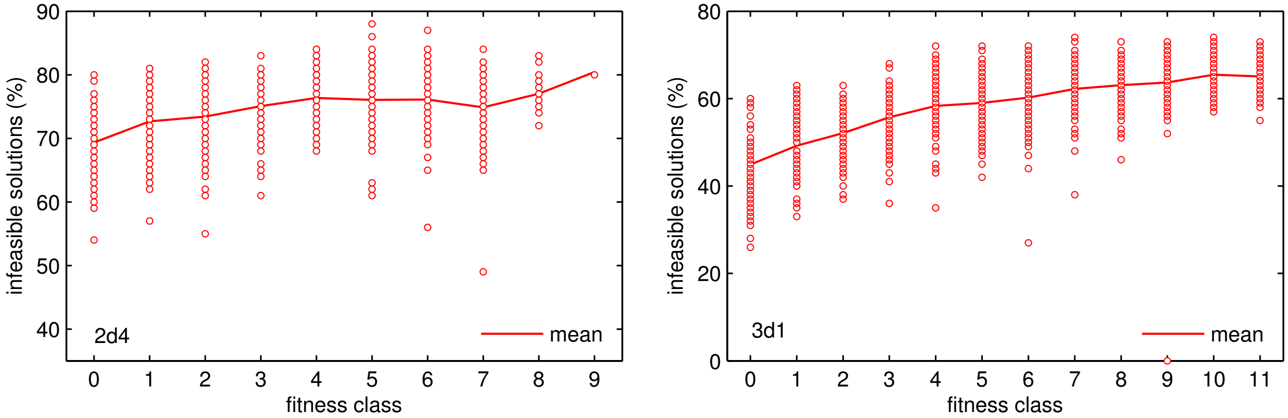


Figure 5.3: Percentage of infeasible solutions in the NNs sampled using the MO formulation. Instance 2d4 (left) and instance 3d1 (right).

low fitness classes leads to the formation of large NNs. On the contrary, solutions at the best fitness classes tend to be more isolated and enclosed by infeasible states. The increase in the neutrality ratios originated from the use of the MO formulation, as analyzed in Section 5.4.2.1, has led to a significant rise in the size of the sampled NNs. NNs computed based on the MO formulation can be several orders of magnitude larger than those computed based on the SO formulation. This behavior becomes more evident as higher fitness classes are considered. It is important to realize that, given a solution $\mathbf{x}$, $NN_{MO}(\mathbf{x})$ will always be a supergraph (presenting at least the size) of $NN_{SO}(\mathbf{x})$. It is worth mentioning that only slight variations in the size of NN_{MO} networks can be perceived from the plots across the different fitness classes. Finally, it becomes relevant for this study the question of to what extent the sampled NN_{MO} networks are composed of infeasible solutions. Figure 5.3 addresses this question. Despite that a feasible solution was given in all the cases as the starting point for the NNs exploration, the bulk of NN_{MO} networks consists of infeasible states. According to Figure 5.3, between 70% and 80% of the nodes, in average, were found to be infeasible when focusing on the two-dimensional instance (respectively, between 50% and 70% for the three-dimensional case).

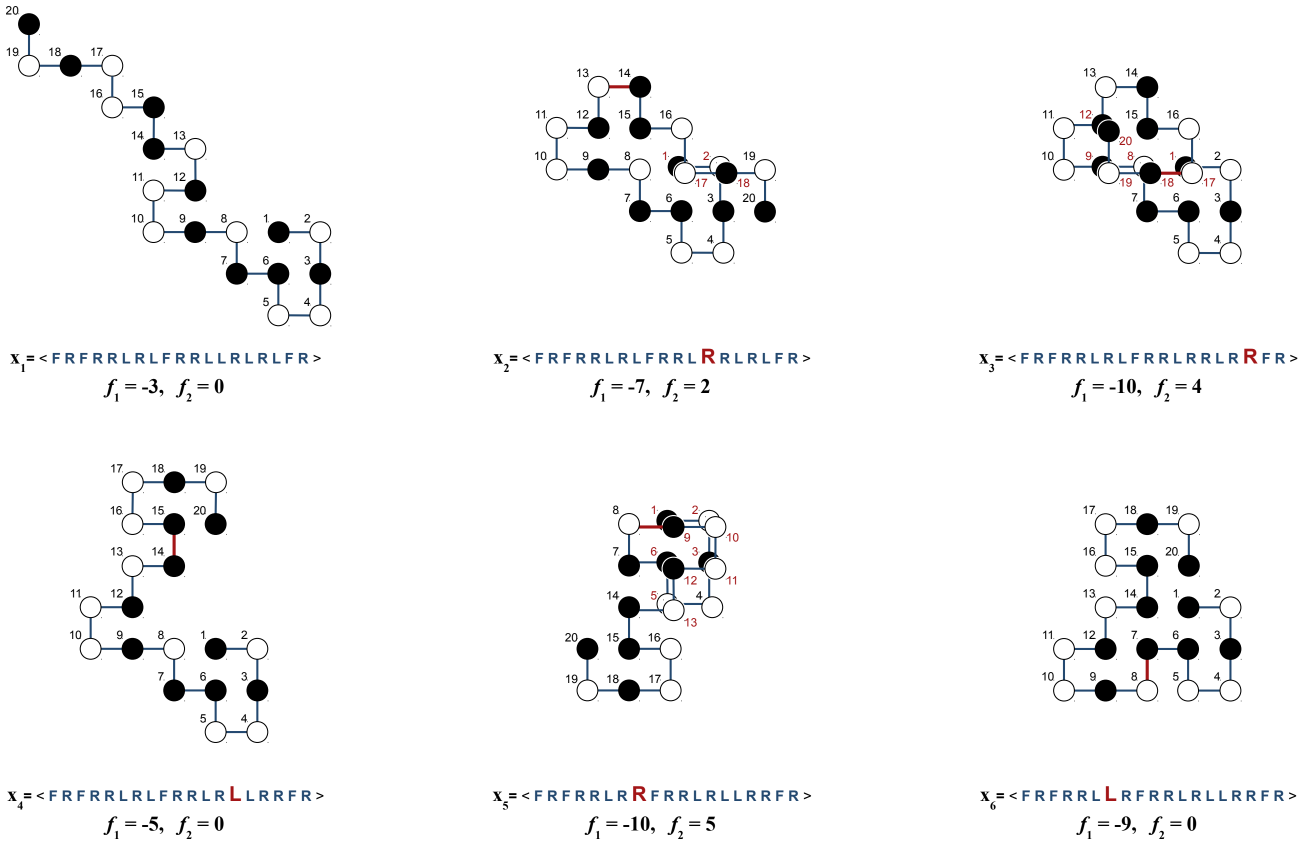
5.4.2.3 *Connectivity between neutral networks*

The introduction of neutrality into the fitness landscape leads to the formation of *neutral connections* between NNs. As defined in Section 4.4.3.3, a neutral connection between NN_1 and NN_2 implies that, as a result of the problem transformation, (at least) a solution $\mathbf{x}_1$ from NN_1 became part of the neutral neighborhood of another solution $\mathbf{x}_2$ which belongs to NN_2. Through such a neutral connection, NN_1 and NN_2 are merged together into a larger NN involving all nodes and edges of the original networks (plus the new edge(s) giving rise to the neutral connection).

In the particular context of the landscape transformation induced by the studied MO constraint-handling strategy, a neutral connection between two NNs can occur, if and only if, one of the two networks is feasible and the other infeasible (incomparability can only be introduced between a feasible and an infeasible state, see Section 5.3). Note, however, that two feasible NNs can be merged through a succession of neutral connections. More precisely, the linkage between two feasible networks NN_1 and NN_k can be given in the form of a sequence $\langle NN_1, NN_2, \ldots, NN_{k-1}, NN_k \rangle$, such that each NN_i is neutrally connected to NN_{i+1}, $1 \leq i < k$, $k \geq 3$, and at least NN_2 and NN_{k-1} are infeasible. In general, a minimum number of $m - 1$ infeasible NNs need to be traversed in order to connect m feasible NNs. To support these ideas, an example is provided in Figure 5.4. This figure illustrates a neutral walk, based on the MO formulation, from a feasible solution $\mathbf{x}_1$ with $Fitness(\mathbf{x}_1) = 3$, to another feasible solution $\mathbf{x}_6$ with $Fitness(\mathbf{x}_6) = 9$ (the global optimum for instance 2d4).[7] In this example, the feasible $NN(\mathbf{x}_1)$ and $NN(\mathbf{x}_6)$ networks have been connected by establishing intermediate neutral connections to (and between) four other different NNs; namely, the infeasible networks $NN(\mathbf{x}_2)$, $NN(\mathbf{x}_3)$ and $NN(\mathbf{x}_5)$, and the feasible network $NN(\mathbf{x}_4)$. The six solutions ($\mathbf{x}_1$ to $\mathbf{x}_6$), and all solutions in their respective NNs, become members of the same NN under the MO evaluation scheme. Therefore, the observed increase in the size of the sampled NNs is not exclusively due to the addition of a significant number of infeasible nodes, as analyzed in Section 5.4.2.2, but is also a result of the combination with other feasible NNs. It should thus be noted that, as in the example, NNs resulting from the use of the MO formulation may involve solutions at different fitness classes (in the original problem formulation) and varying degrees of infeasibility.

To elaborate further on this matter, this section analyzes how the use of the MO formulation during the performed sampling produced neutral connections between NNs from distinct fitness classes. For each fitness class c, the plots in Figure 5.5 indicate whether and how many of the NNs

[7] It is important to remember that both f_1 and f_2 are to be minimized, and that $f_1(\mathbf{x}) = E(\mathbf{x}) = -Fitness(\mathbf{x})$.



x_1= < FRFRRLRLFRRLLRLRLFR >
$f_1 = -3, \; f_2 = 0$

x_2= < FRFRRLRLFRRLRRLRLFR >
$f_1 = -7, \; f_2 = 2$

x_3= < FRFRRLRLFRRLRRLRRFR >
$f_1 = -10, \; f_2 = 4$

x_4= < FRFRRLRLFRRLRLLRRFR >
$f_1 = -5, \; f_2 = 0$

x_5= < FRFRRLRRFRRLRLLRRFR >
$f_1 = -10, \; f_2 = 5$

x_6= < FRFRRLLRFRRLRLLRRFR >
$f_1 = -9, \; f_2 = 0$

Figure 5.4: Neutral walk, based on the MO formulation, connecting six different solutions for instance 2d4. The encoding (based on relative moves) and objective values are provided for each solution.

computed for solutions at this fitness class formed neutral connections to NNs at each other possible fitness class c'. More detailed information about the interpretation of these plots can be found in Section 4.4.3.3. It is important to clarify that only connections to feasible NNs have been accounted for in this analysis. It can be seen from the figure that neutral connections were created between almost each possible pair of fitness classes of the adopted instances. As the only exceptions, no connections to fitness class 9 were identified when sampling the NNs for fitness classes 1 and 2 of instance 2d4. More connections appear indicated below (rather than above) the diagonals in the plots. Indeed, the vast majority of the sampled NNs for the different fitness classes formed neutral connections to NNs at all other lower fitness classes. As suggested in Section 4.4.3.3, inferior fitness classes are easier to reach than the superior ones because of the funnel-like search landscape of the

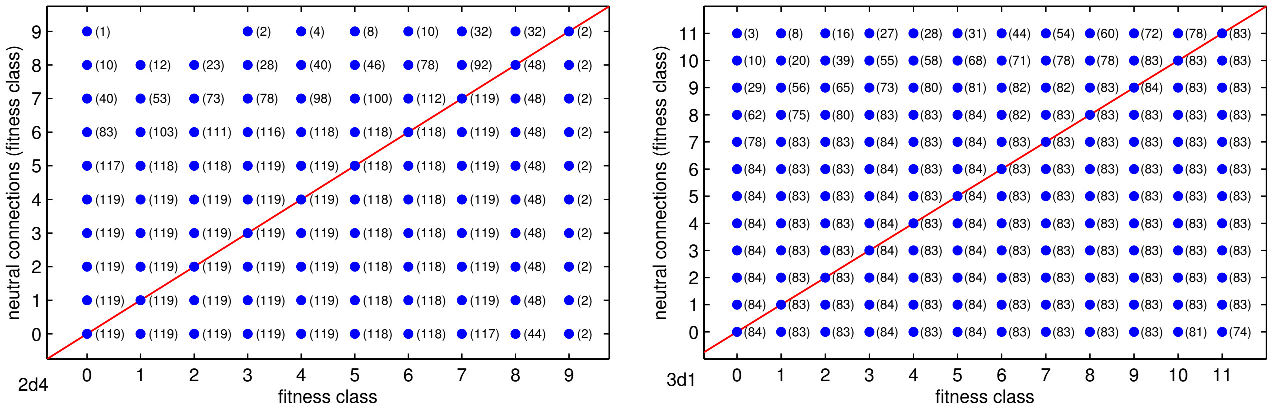


Figure 5.5: Neutral connections formed between fitness classes. Two-dimensional 2d4 instance (left) and three-dimensional 3d1 instance (right).

problem [62]. Despite the use of a considerably low value for parameter $maxDepth$ during the NN sampling procedure, namely $maxDepth = 5$, such a reduced number of allowed neutral steps was still enough to establish connections between even the most distant fitness classes. That is, Figure 5.5 reveals that 1 out of the 119 explored NNs at fitness class 0 of instance 2d4, the worst fitness class, reached the optimum solution at fitness class 9. Note also that the 2 computed NNs from fitness class 9 connected to fitness class 0. A similar behavior can be observed with regard to instance 3d1. Figure 5.5 shows that 3 out of the 84 NNs from fitness class 0 merged with NNs at the best fitness class, *i.e.*, 11, and that a total of 74 connections between these fitness classes occurred in the opposite direction. All such connections were achieved after a maximum number of $maxDepth$ successive neutral moves from the solution given as the starting point for the NNs computation.

Although neutral connections were established between NNs at distant fitness classes, as discussed above, it is possible to see from Figure 5.5 that the number of neutral connections tends to decrease as the distance between fitness classes increases (higher numbers of neutral connections are shown close to the diagonal). Therefore, the distance in fitness relates to the likelihood that two NNs can connect. This is particularly true when the number of allowed intermediate neutral connections is bounded (as it was done in this study with the use of parameter $maxDepth$). Similar observations have been previously made within the context of the analysis conducted in Section 4.4.3.3.

The formation of neutral connections between two feasible networks NN_1 and NN_2 can be understood as the definition of (previously nonexistent) neutral paths bridging the corresponding regions of the feasible space. By allowing movement across infeasible areas, any solution $\mathbf{x}_1$ from NN_1 could potentially be reached through a neutral walk departing at an arbitrary solution $\mathbf{x}_2$ belonging to NN_2. On the one hand, this can be particularly relevant when dealing with problems which present multiple disconnected feasible regions. On the other hand, even in connected feasible spaces, the length of the shortest path between two feasible solutions can be significantly greater if this path considers only feasible intermediate states [122]. In the example provided in Figure 5.4, the feasible solutions $\mathbf{x}_1$ and $\mathbf{x}_6$ differ exactly in $d = 5$ encoding positions. By exhaustive enumeration, it was found that all the $d! = 120$ possible shortest paths (of length d) connecting $\mathbf{x}_1$ and $\mathbf{x}_6$ involve infeasible solutions.[8] Therefore, the shortest feasible path between this pair of solutions is necessarily longer (of length greater than d). It is worth mentioning that 11 out of the 120 shortest paths between $\mathbf{x}_1$ and $\mathbf{x}_6$ represent neutral paths under the studied MO problem formulation (one of them illustrated in Figure 5.4). In a related analysis reported in [66], Duarte-Flores and Smith computed all possible shortest paths from a set of near-optimal solutions to the global optimum of a particular HP model's instance on the triangular lattice. As a result, only about 12% (on average) of the explored paths were found to be strictly feasible. The fact that most of the shortest paths between feasible regions traverse infeasible areas emphasizes the advantages of using the MO constraint-handling strategy. Furthermore, a path within the boundaries of the feasible space may require the explicit movement towards inferior fitness classes, especially when this path connects different *basins of attraction*.[9] The alternative fitness landscape induced by the MO strategy may potentially define neutral paths between basins of attraction, which can be exploited as a means of escaping from local optima.

[8] More specifically, 24 out of these 120 shortest paths involve 2 infeasible solutions, other 48 paths include 3 infeasible solutions, and all the 4 intermediate points are infeasible for the remaining 48 paths.

[9] The basin of attraction of a local optimum $\mathbf{x}$, involves all the areas of the fitness landscape which lead (or tend to lead) directly to $\mathbf{x}$ when optimizing based on gradient information.

5.5 Introducing a search bias

By defining trade-offs between the quality and feasibility of candidate solutions, the multi-objective (MO) approach to handle constraints allows for the exploitation of useful information from infeasible areas of the fitness landscape. Despite the potential advantages of the MO strategy in terms of the landscape transformation, as analyzed at the end of Section 5.4.2.3, its lack of a proper search bias may also lead to detrimental effects on the ability of search algorithms for locating promising regions of the feasible space. More specifically, if a bias towards the feasible region is not introduced, a significant fraction of the computational effort can be invested in evaluating infeasible solutions. Depending on the particular characteristics of the problem under consideration, an unbiased search based on the MO method could even fail to reach any feasible solution at all [188].

In this section, the importance of coupling the MO constraint-handling strategy to an effective biasing mechanism is investigated. Three different biasing methods for the MO strategy are to be evaluated in terms of how their implementation impacts on the performance of search algorithms. A basic single-solution-based evolutionary algorithm (EA), called the (1+1) EA, and a basic genetic algorithm (GA), a population-based technique, have been considered.[10] Details of these algorithms and the adopted settings are provided in Sections 5.6.2 and 5.6.3. The analyses with regard to the three considered biasing approaches are separately presented in Sections 5.5.1, 5.5.2 and 5.5.3.

5.5.1 Archiving

In evolutionary multi-objective optimization, maintaining a repository with the current approximation of the Pareto-optimal set, and thus of the Pareto front, is usually assumed to be a crucial issue [117, 144]. Hereafter, this kind of nondominated solutions repository is to be called *archive*, and the way this archive is constructed, updated and utilized during the search process will be referred to as the *archiving strategy*. This section analyzes the extent to which an archiving strategy can influence the behavior of the implemented (1+1) EA when using the MO constraint-handling technique.

[10] It is important to note that some of the studied biasing methods are only suitable, and thus are only analyzed here, either for (1+1) EA or for the GA.

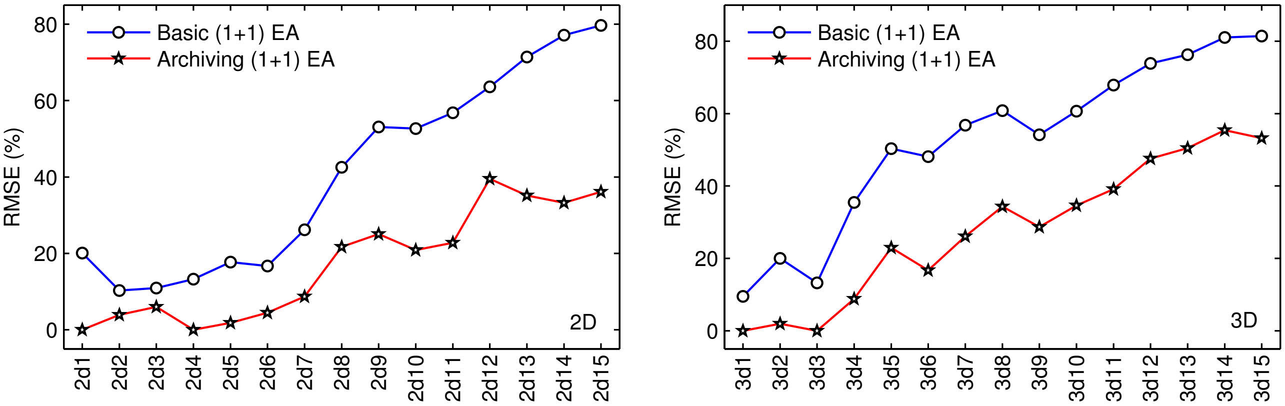


Figure 5.6: RMSE obtained by the basic and archiving variants of the (1+1) EA when using the MO strategy. Two-dimensional (left) and three-dimensional test instances (right).

Rather than functioning as a source of genetic material, *i.e.*, as a population, in the archiving (1+1) EA the archive is used only with the aim of introducing a bias in the selection process. In order to be accepted, a new candidate individual must represent a competitive trade-off between the two defined optimization objectives. This is determined by comparing the new individual with respect to the whole Pareto front approximation stored in the archive. In this way, although (strictly speaking) an explicit bias towards the feasible region is not being applied, archiving restricts the movement of the algorithm, allowing it to concentrate on promising regions, either feasible or infeasible, of the fitness landscape. This archiving variant of the (1+1) EA is described in detail in Section 4.5.1.

Figure 5.6 contrasts the performance of the basic and archiving variants of the (1+1) EA, using the MO strategy, for all two- and three-dimensional test instances. Results are reported in terms of the relative root mean square error, RMSE, computed over a total of 31 independent executions of each experiment. In all the cases, the two algorithms were run for a maximum number of 10^6 solution evaluations. As it can be seen from the plots, the use of the archiving strategy within the (1+1) EA has led to a significant improvement in the RMSE measure for all the 30 adopted test cases (lower RMSE is preferred). It is also possible to observe that the benefits of archiving tend to become more evident as the size of the problem (length of the input protein sequence) increases.

From the above discussed results, archiving has been found to be essential for guiding the search process effectively when the MO constraint-handling strategy is implemented. In the words of Handl *et al.* [85]: "*this way of using an archive yields a negative efficiency preserving strategy, i.e., it prevents degradation of solutions. We say there is degradation if the current solution is replaced at some later iteration by one that it dominates. Such degradation prevents convergence and can lead to endless cycling between solutions that are not mutually incomparable*". Without archiving, therefore, the (1+1) EA based on the MO strategy may drift through the search landscape, moving away from or moving towards the feasible region in a bias-free manner.

Finally, it should be noted that archiving strategies can also be used in the context of population-based methods (where archives are usually referred to as secondary populations). The considered GA, however, relies on an elitist selection scheme which inherently preserves in the population the current Pareto front approximation (or at least part of it due to the fixed population size). Thus, the biasing effects obtained through archiving are implicitly incorporated in such an algorithm.

5.5.2 Feasibility rules

One of the simplest, yet effective and widely used constraint-handling methods, consists in defining a set of rules on which the discrimination among individuals is to be based. This approach is commonly referred to as the use of *feasibility rules* in the specialized literature [151, 160, 161, 167, 244]. The popularity of this method stems not only from its parameter-free nature, but also from its ability to be combined with other constraint-handling mechanisms, as reviewed in [161]. One of the most representative works on this topic was reported by Deb [58], where it was proposed a GA implementing a binary tournament selection operator which relies on the following three criteria:

1. If comparing between two feasible solutions, the one with the best objective function value is to be preferred.

2. If comparing between two infeasible solutions, the one with the lowest infeasibility degree is to be preferred.

3. If comparing between a feasible and an infeasible solution, the feasible one is to be preferred.

The first criterion can be generalized to the case where solutions presenting the same degree of constraint violation are considered. This more general case involves the comparison between feasible individuals, as in the original rule, but covers also the case where two infeasible individuals with the same infeasibility degree are being compared. Such a later scenario has not been accounted for in the originally proposed set of rules [158]. This extended version of the first criterion is implicitly satisfied when handling constraints by multi-objective (MO) optimization. The MO approach, however, lacks the bias towards the feasible region that the second and third discrimination criteria represent.

This section explores how the use of simple feasibility rules based on the MO method can help in guiding the search process effectively in the implemented GA.[11] More specifically, the preference relation between two solutions $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ will depend on the following three criteria:

1. If $\mathbf{x}_1 \prec \mathbf{x}_2$, then solution $\mathbf{x}_1$ is to be preferred.
2. If $\mathbf{x}_2 \prec \mathbf{x}_1$, then solution $\mathbf{x}_2$ is to be preferred.
3. Otherwise, *i.e.*, $\mathbf{x}_1 \not\prec \mathbf{x}_2$ and $\mathbf{x}_2 \not\prec \mathbf{x}_1$, the solution with the lowest degree of constraint violation (lowest f_2 value) is to be preferred.

That is, individuals will be first compared based on the Pareto-dominance relation. Whenever no preferences can be imposed by using the Pareto-dominance relation, the degree of infeasibility of the solutions will be adopted as a secondary discrimination criterion. The implementation of these rules required the adaptation of both the $selection\text{-}for\text{-}variation$ and $selection\text{-}for\text{-}survival$ processes of the GA. On the one hand, in $selection\text{-}for\text{-}variation$ the binary tournament selection operator was simply equipped with the new defined set of rules. On the other hand, the $selection\text{-}for\text{-}survival$ process is performed by means of *nondominated sorting* [59]. As treated more extensively in the preamble of Section 4.5.2, the nondominated sorting procedure works by defining layers of nondominated individuals, which are then iteratively included (from the best down to the worst layer) until completing the new GA population. If this iterative selection procedure faces a nondominated layer containing more individuals than the number of free slots in the population, the secondary

[11] Implementing this approach within the (1+1) EA results in over-penalization; once a feasible solution is reached, infeasible solutions would not be considered anymore. Thus, the conducted analysis focuses only on the GA.

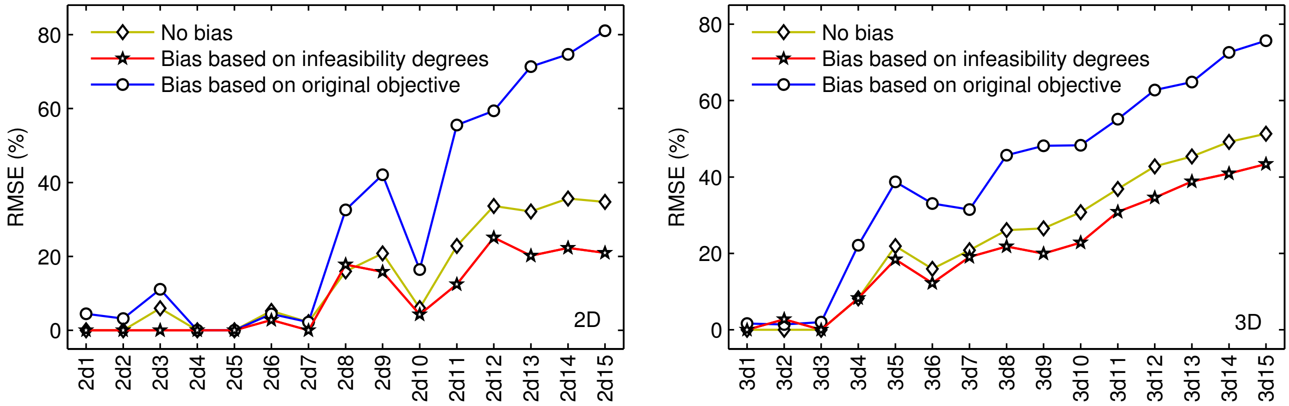


Figure 5.7: Introducing a search bias in the GA by using feasibility rules. RMSE obtained for all the two-dimensional (left) and three-dimensional test instances (right).

criterion based on infeasibility degrees is applied in order to choose the remaining survivors. Since only the last considered nondominated layer is discriminated based on such an infeasibility-based criterion, a significant portion of the infeasible individuals could potentially be selected.[12] This reduces the selection pressure and, thus, can contribute in overcoming *premature convergence*, a problem usually related to the use of the feasibility rules approach for handling constraints [161].

The performance of the GA using the MO constraint-handling strategy was evaluated with and without incorporating the above described infeasibility-based secondary criterion. In addition, a third variant of the GA was considered where such a secondary discrimination criterion is based on the original objective function, and not on the infeasibility degrees. In this way, it will be possible to analyze not only the importance of having a search bias, but also the effects that can be achieved if this bias favors either one or the other of the two optimization objectives defined by the MO strategy. Figure 5.7 presents the obtained results, in terms of the RMSE, for all two- and three-dimensional test cases.[13] In all the experiments, a maximum number of 10^6 evaluations was used as the stopping condition and 31 repetitions were performed. The introduction of a search bias towards the feasible region allowed the GA to score the best RMSE values in most of the cases. Although no important differences can be appreciated for the smallest problem instances, the advantages of introducing this

[12] The number of selected feasible individuals will always match the number of considered nondominated layers. This is because there can be at most one feasible solution per nondominated set computed based on the MO strategy.

[13] For each of the instances, the results presented in Figure 5.7 correspond to the lowest RMSE values obtained by evaluating a set of different parameter configurations of the GA, refer to Section 5.6.3.1 for details.

bias are more clear when focusing on the hardest ones (rightmost part of the plots). It is interesting to observe from the figure that, rather than benefiting optimization, biasing the search according to the original objective has impacted negatively on the performance of the GA. Realize that in a set of nondominated solutions, computed based on the MO strategy, there can be at most one feasible solution; all other solutions within the set are infeasible and, by definition, strictly better than the feasible member with regard to the original objective. Therefore, the discrimination of nondominated individuals based solely on the original objective will favor those individuals which, despite showing a prominent behavior for this criterion, represent the poorest trade-offs in terms of infeasibility. Consequently, the search process can be guided away from the feasible space.

5.5.3 Proportional bias

In multi-objective optimization, introducing a bias can be understood as the articulation of preferences to capture the relative importance of the different optimization criteria. Consider a two-objective problem, denoted by the objective vector $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x})]^T$. If f_2 was determined to be a more important objective function than f_1, it can be hypothesized that the addition of *noise* to f_1, a noise which is proportional and directly dependent on f_2, would produce a biasing effect in order to favor f_2 during the search process. That is, the incorporation of noise into f_1 relaxes the selection pressure with regard to this objective, since the real contribution of f_1 to guiding the search is reduced. Moreover, by relating the injected noise with function f_2, the selection pressure with respect to f_2 is strengthened, yielding a bias. From this reasoning, the above multi-objective problem, where f_2 is assumed to be the highest priority objective, can alternatively be stated in terms of the objective vector $\mathbf{f}'(\mathbf{x}) = [f'_1(\mathbf{x}), f'_2(\mathbf{x})]^T$, such that

$$f'_1(\mathbf{x}) = f_1(\mathbf{x}) + \omega \left(\frac{f_2(\mathbf{x})}{f_2^{max}} \right) (f_1^{max} - f_1^{min}), \tag{5.3}$$

$$f'_2(\mathbf{x}) = f_2(\mathbf{x}), \tag{5.4}$$

where f_1^{max} and f_1^{min} are the maximum and minimum known values for function f_1 (since the beginning of the search), and f_2^{max} is the maximum known value for function f_2.[14] In (5.3), factor $(f_1^{max} - f_1^{min})$ represents the maximum known difference between f_1 values, so that by using this value it would be possible to alter the preference relation between any pair of solutions with regard to f_1. Factor $\left(\frac{f_2(\mathbf{x})}{f_2^{max}}\right)$ allows the incorporated noise to be proportional to the scored f_2 performance; *i.e.*, this factor tends to 1 as worse f_2 values are considered (minimization assumed). Therefore, the better the solution in objective f_2, the lower the perturbation to its f_1 value. Finally, the *bias strength* ω is a user defined parameter introduced with the aim of evaluating the impact of further controlling the magnitude of the applied noise. Using this strategy, two different solutions that are incomparable (nondominated) with respect to their original $\mathbf{f}$ objective vectors, could be discriminated (in favor of the best f_2 performance) if compared based on their alternative objective vectors $\mathbf{f}'$. This strategy can thus be implemented within an optimization algorithm in order to set a search bias.

This section tests the ability of the above described strategy to provide the multi-objective (MO) constraint-handling approach with an effective search bias. The objective function f_2 in the MO problem formulation, which accounts for the degree of constraint violation, is to be defined as the most important criterion in order to bias the search towards the feasible region. It should be emphasized that, under the studied MO formulation, the use of this proportional biasing mechanism will only affect infeasible individuals; *i.e.*, it holds that $f_2(\mathbf{x}) = 0$ for all feasible individuals $\mathbf{x} \in \mathcal{X}_{\mathcal{F}}$, so that no noise can be added to their f_1 values. In contrast to Sections 5.5.1 and 5.5.2, the analysis here presented focuses on the two implemented search algorithms, namely the (1+1) EA and the GA. In these algorithms, the alternative objective vectors $\mathbf{f}'$ of all individuals are to be computed at each iteration to serve as the basis for driving selection. A large set of values in the range $[0, 2]$ have been explored for the bias strength parameter ω, where $\omega = 0$ indicates that no bias is to be applied. In all the cases, the algorithms were run for a total of 10^6 solution evaluations, 31 independent executions were performed, and results are to be evaluated in terms of the O-RMSE measure.

[14] Alternatively, f_1^{max}, f_1^{min} and f_2^{max} could be computed from the current population or Pareto front approximation. These values could even be fixed if this problem-dependent information is known a priori.

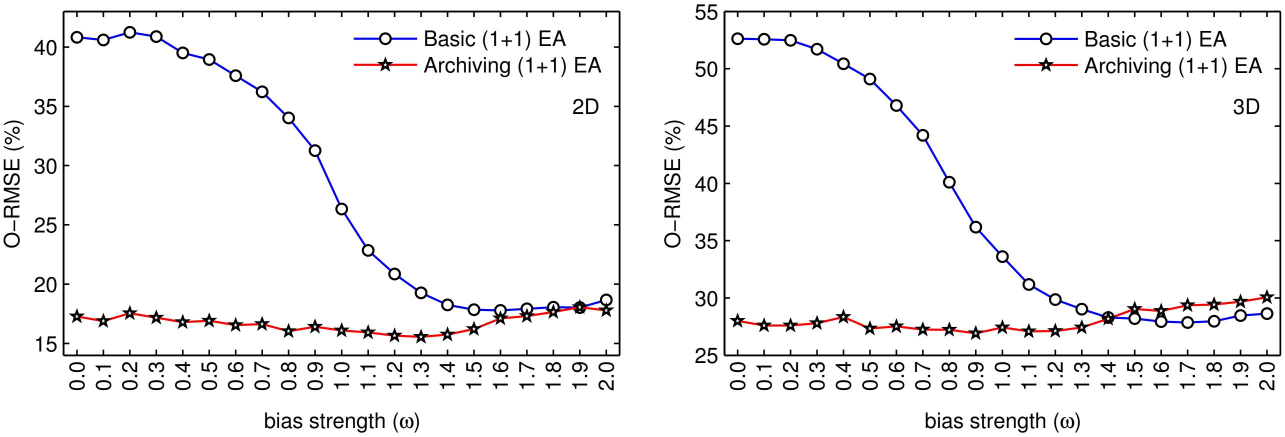


Figure 5.8: Introducing a proportional bias in the MO strategy. O-RMSE obtained by the basic and archiving (1+1) EA for the two-dimensional (left) and three-dimensional test instances (right).

The results for the (1+1) EA are shown in Figure 5.8. In addition to the basic (1+1) EA, the evaluation of the proportional biasing mechanism covers also the archiving variant of the (1+1) EA, as described and analyzed in Section 5.5.1. Hence, the archiving (1+1) EA studied in this section integrates two different biasing methods (*i.e.*, archiving and the proportional bias). Figure 5.8 confirms the need for an effective biasing strategy when the handling of constraints is approached by multi-objective optimization. Without a bias ($\omega = 0$), the basic (1+1) EA scored considerably high O-RMSE values for both the two- and the three-dimensional test instances. Note, however, that the performance of this algorithm was gradually improved with the increasing value of ω. The best performance for the basic (1+1) EA was reached at $\omega = 1.6$ for the two-dimensional instances, and $\omega = 1.7$ for three-dimensional case. The O-RMSE measure was decreased by more than 23% in both cases with respect to the corresponding results at $\omega = 0$. Due to the implicit bias that the archiving (1+1) EA involves, the rewards of implementing the proportional biasing strategy were not as remarkable as those for the basic version of this algorithm. Nevertheless, most of the explored ω values allowed the archiving (1+1) EA to achieve slight but still appreciable decreases in the O-RMSE. While the basic (1+1) EA performed the best for high ω values ($\omega > 1.5$), a less strength of the proportional bias was required when using the self-biasing archiving (1+1) EA (whose performance deteriorated for the highest ω values). The archiving (1+1) EA showed its best performance when using a value of $\omega = 1.3$ (two-dimensional case), and $\omega = 0.9$ (three-dimensional case).

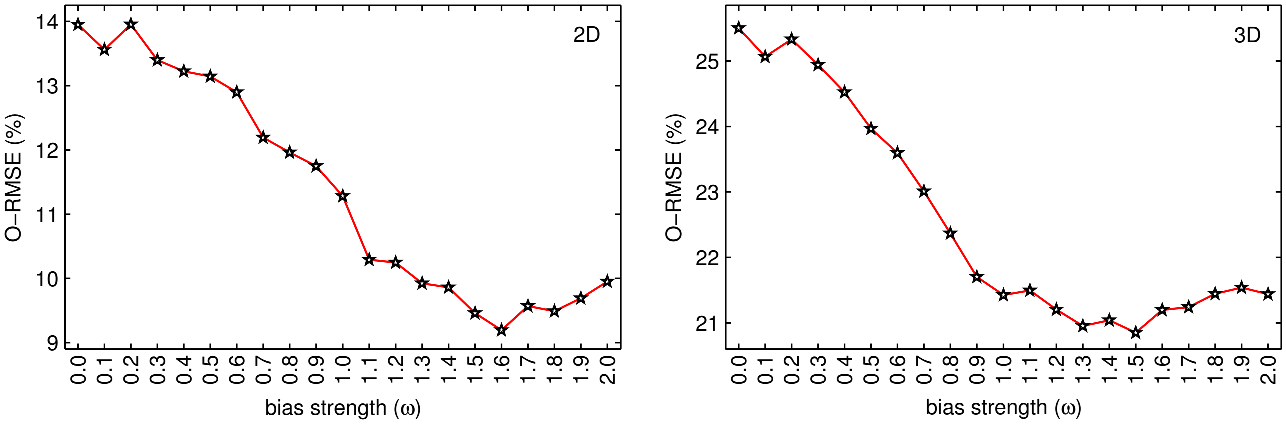


Figure 5.9: Introducing a proportional bias in the MO strategy. O-RMSE obtained by the GA for the two-dimensional (left) and three-dimensional test instances (right).

Finally, Figure 5.9 presents the obtained results with regard to the GA.[15] The incorporation of the proportional biasing mechanism has led to a significant enhancement in the performance of the GA. The behavior of the GA exhibits a clear tendency to improve with the increase in the bias strength parameter ω. It is interesting to note, however, that once the best O-RMSE values were reached at $\omega = 1.6$ and $\omega = 1.5$ (for the two- and three-dimensional cases, respectively), this tendency changes and the GA's performance begins to decline for higher ω values. From this, and given that the above analyzed (1+1) EA suffered a performance decrease when using the highest ω values as well, it is possible to say that the excessive bias could also be detrimental to the search efficiency. In this particular context, the increase in ω tends to lead to over-penalization. Therefore, defining the proper amount of search bias could be a non-trivial, problem- and algorithm-dependent task.

5.6 Impact on search performance

This section investigates the suitability of the multi-objective optimization (MO) strategy for handling constraints in the HP model. To this end, the MO strategy is evaluated and compared with respect to two different constraint-handling approaches usually adopted in the specialized literature, namely, the rejection of infeasible protein conformations and the application of penalties. These approaches

[15] For each considered ω value, the O-RMSE in Figure 5.9 corresponds to the best performance obtained by evaluating a set of different parameter configurations for the GA; details provided in Section 5.6.3.1.

are to be referred in this section to as the reject (RJ) and penalty function (PF) strategies and are described in detail in 5.6.1. As discussed in Section 5.5, introducing a proper search bias is crucial for the success of the MO strategy. This issue is further addressed in this section by evaluating the different biasing mechanisms studied in Section 5.5 with respect to each other. The comparative analysis presented in this section focuses on the impact that the various studied constraint-handling methods have on the performance of search algorithms. Two different evolutionary algorithms (EAs) have been considered, namely, a basic single-solution-based EA and a population-based EA. The corresponding analyses are covered in Sections 5.6.2 and 5.6.3.

5.6.1 Baseline constraint-handling methods

The purpose of this section is to describe the two constraint-handling strategies for the HP model which have been taken as a baseline in this study. The first strategy, described in Section 5.6.1.1, is based on the rejection of solutions encoding infeasible protein conformations. The second considered approach is based on the application of penalties and is detailed in Section 5.6.1.2.

5.6.1.1 Reject strategy

A basic reject strategy (RJ) is considered where only feasible protein conformations are accepted during the search process. A single-solution-based evolutionary algorithm (EA), the (1+1) EA, and a genetic algorithm (GA) are used in this study; refer to Sections 5.6.2 and 5.6.3 for details. In order to implement the RJ strategy, the variation operators of these algorithms were adapted as follows. In the (1+1) EA, once mutation is to be applied to a particular encoding position (determined based on a given probability), all possible perturbations to this position are evaluated in random order until a feasible conformation is obtained. If no change in this position leads to a feasible conformation, the original value is restored. The GA uses a one-point crossover operator. In this operator, all possible crossover points are explored in random order until feasible children are produced; otherwise, either one or both of the parents are copied unchanged. The mutation operator of the GA was adapted in the same manner as described above for the (1+1) EA. Note that such a persistent application of the variation operators involves an additional computational effort (*i.e.*, a significant number of infeasible individuals could potentially be verified and discarded without consuming objective function evaluations, on which the stopping criterion of the implemented algorithms relies). Furthermore, the RJ strategy requires the algorithms to be provided with initial feasible individuals. The backtracking procedure proposed in [42] was used for generating such initial feasible individuals.

The above described RJ strategy is equivalent to the one analyzed by Duarte-Flores and Smith within a GA [66]. Similar strategies have also been adopted in the context of different search metaheuristics. For example, the *hypermutation* and *hypermacromutation* mechanisms, as implemented in some immune system-based algorithms for the HP model reported in the literature, operate in a similar feasibility-preserving fashion. These operators iteratively apply a series of mutations to the input solution and infeasible solutions encountered during this process are always discarded [49, 53, 57].

5.6.1.2 Penalty function

A constraint-handling strategy based on the use of a penalty function (PF) has been considered in this study. In the PF strategy, the energy (objective) value of a candidate solution is penalized according to the number of collisions that the encoded protein conformation presents. More formally, PSP under the HP model is restated as the problem of minimizing an alternative objective function $f(\mathbf{x})$ defined as follows ($\mathbf{x} \in \mathcal{X}$):

$$f(\mathbf{x}) \quad = \quad E(\mathbf{x}) + \rho \times \zeta \times Collisions(\mathbf{x}), \tag{5.5}$$

where $E(\mathbf{x})$ denotes the conventional energy function of the HP model introduced in Section 2.3.3.1. $Collisions(\mathbf{x})$ refers to the total number of amino acid pairs (a_i, a_j) in $\mathbf{x}$ such that a_i and a_j collide at the same lattice position (as used also in the proposed MO strategy, see Section 5.3). Finally, the value of ζ is to be large enough that, assuming a *penalty weight* of $\rho = 1$, it holds that $f(\mathbf{x}_i) \leq 0, \forall \mathbf{x}_i \in \mathcal{X}_\mathcal{F}$ while $f(\mathbf{x}_j) > 0, \forall \mathbf{x}_j \in \mathcal{X} \setminus \mathcal{X}_\mathcal{F}$. By defining the penalty weight ρ within the range $[0, 1]$, it will then be possible to move from an *under-penalization* scenario ($\rho = 0$), where comparisons are only based on the original objective function of the problem, to an *over-penalization*

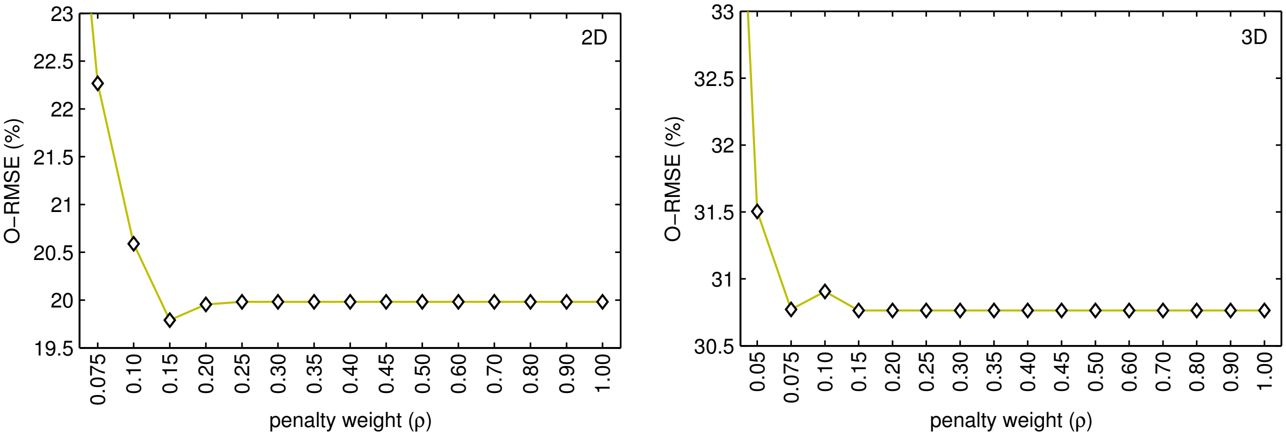


Figure 5.10: Impact of varying the penalty weight (ρ) of the PF method on the (1+1) EA's performance. Two- (left) and three-dimensional (right) instances.

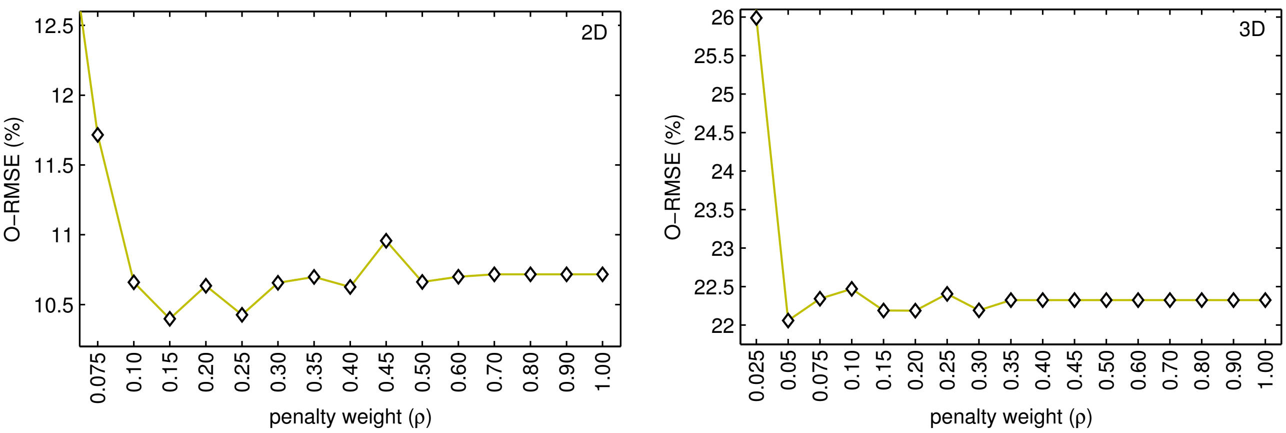


Figure 5.11: Impact of varying the penalty weight (ρ) of the PF method on the performance of the GA. Two- (left) and three-dimensional (right) instances.

scenario ($\rho = 1$), where the penalty term dominates discrimination [187]. In this study, ζ was set to $\zeta = 2\ell_H + 2$ for the two-dimensional square lattice and $\zeta = 4\ell_H + 2$ for the three-dimensional cubic lattice. These values represent upper bounds on the number of H-H topological contacts that can be formed in the corresponding lattices and have also been considered in [122]. It should be noted that ζ depends on the total number of hydrophobic amino acids in the protein sequence, ℓ_H.

With the aim of investigating the importance of the penalty weight ρ, and also to enable a more reliable comparative analysis in Sections 5.6.2 and 5.6.3, different settings for this parameter are here explored. Figures 5.10 and 5.11 show the performance scored by the (1+1) evolutionary algorithm

(EA) and the genetic algorithm (GA) when using the PF method with a series of different ρ values in the range $[0, 1]$. Performance is expressed in terms of the O-RMSE measure, computed over a total of 31 independent repetitions for each experiment. In general, the worst behavior of both the (1+1) EA and the GA was exhibited when no penalties were applied ($\rho = 0$).[16] Figure 5.10 indicates that $\rho = 0.15$ allowed the (1+1) EA to achieve its best performance at solving the two-dimensional instances, and all considered ρ values in the range $[0.15, 1]$ produced the best results for the three-dimensional case. Regarding the GA, it is possible to observe from Figure 5.11 that the lowest O-RMSE values were reached by using $\rho = 0.15$ and $\rho = 0.05$ for the two- and three-dimensional test instances, respectively. The best performing settings for the PF method, as described above, have been considered during the comparative analysis conducted in Sections 5.6.2 and 5.6.3.

5.6.2 Analysis for a single-solution-based algorithm

A basic single-solution-based evolutionary algorithm (EA), the so-called (1+1) EA, has been implemented in order to assess the impact of using the studied constraint-handling methods. Five different constraint-handling approaches are considered, the reject (RJ) and penalty function (PF) strategies taken as the baseline, and three variants of the proposed multi-objective (MO) technique originated from the use of the biasing mechanisms analyzed in Section 5.5: (i) MO+AR, where archiving is used to bias the search process; (ii) MO+PB, where a proportional bias is introduced; and (iii) MO+AR+PB, which combines both the archiving and the proportional biasing mechanisms.

The functioning of the (1+1) EA has been previously described through Algorithm 5 in Section 4.5.1. In this algorithm, the acceptance criterion, and thus the discrimination among candidate individuals, depends upon the constrain-handling strategy to be applied. On the one hand, it can be based on the one-dimensional objective (energy) value of the candidate conformations, either including penalties or not (PF and RJ approaches, respectively). On the other hand, acceptance will be based on the Pareto-dominance relation when applying the MO strategy. In this way, the search behavior and performance of this algorithm will be determined by each of the different studied

[16] In Figures 5.10 and 5.11, the results obtained when using the lowest considered ρ values (leftmost data) have not been displayed in order to highlight details in the most relevant part of the plots.

techniques. Also, the archiving variant of the (1+1) EA, on which the MO+AR and MO+AR+PB approaches rely (see Sections 5.5.1 and 5.5.3), is further described in Section 4.5.1.

In all the cases, an internal coordinates representation based on the relative moves encoding has been implemented.[17] Details on this representation are provided in Section 2.3.3.3. The mutation probability was fixed to $p_m = \frac{1}{\ell-2}$, where $\ell - 2$ denotes the length of the individuals encoding. Finally, a maximum number of 10^6 solution evaluations was adopted as the stopping condition. The proportional biasing mechanism, which leads to the MO+PB and MO+AR+PB strategies, requires the adjustment of the bias strength parameter, ω. Similarly, the PF method requires the fine-tuning of the penalty weight, ρ. The analysis conducted in this section considers the best performing settings for these parameters, as they were respectively derived in Section 5.5.3 and 5.6.1.2.

5.6.2.1 Comparative analysis

Figure 5.12 shows the online (throughout the search) performance achieved by the (1+1) EA when using the studied constraint-handling approaches. Performance is expressed in terms of the overall relative root mean square error, O-RMSE, computed from 100 independent executions of each experiment.[18] Results are reported in steps of $50,000$ solution evaluations until completing the maximum number of 10^6 evaluations defined as the stopping condition. From these figures, it is possible to see that the lowest O-RMSE values, in both the two- and the three-dimensional test cases, were reached by using MO+AR+PB. Therefore, the use of archiving together with the introduction of a proportional bias constitutes a more effective biasing strategy when compared to the separate use of these mechanisms. MO+PB presented a more accelerated convergence than MO+AR at the first stages of the search. This can be explained by the fact that, given that MO+AR does not explicitly bias the search towards the feasible region (as discussed in Section 5.5.1), this method invests more effort in exploring infeasible states. It is worth noting, however, that such an investment has paid

[17] The use of relative rather than absolute moves to represent protein conformations provides an advantage in terms of constraint-handling. As detailed in Section 2.3.3.3, using relative moves all possible solution encodings represent one-step self-avoiding conformations. This is the motivation for using such an encoding scheme in this chapter.

[18] Notice that, while previous analyses considering different parameter settings for the compared approaches were based on 31 repetitions of the experiments, detailed analyses using the best performing settings are based on 100 repetitions in order to compare more representative performance samples.

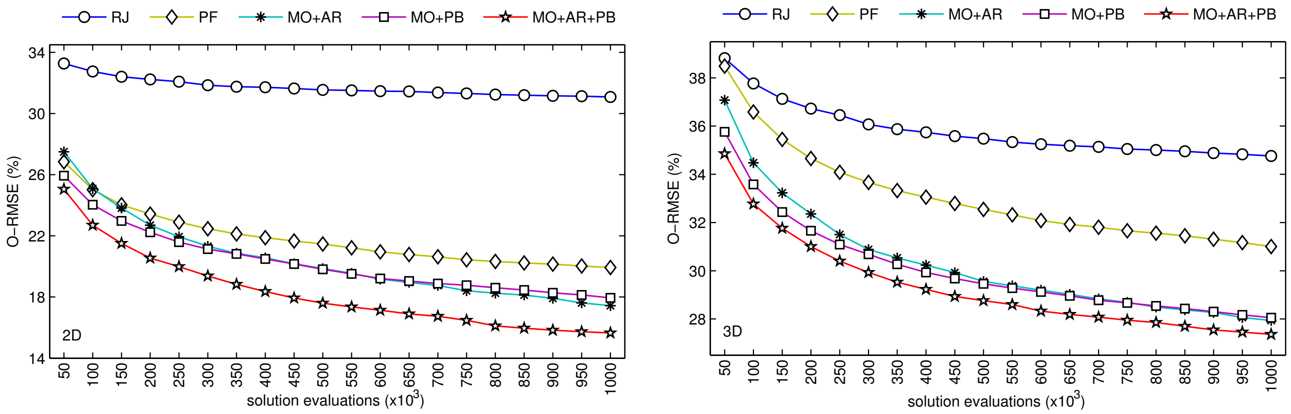


Figure 5.12: O-RMSE scored by the (1+1) EA as the search process progressed. Two-dimensional (left) and three-dimensional (right) instances.

off; the slope in the corresponding curve is more pronounced, indicating that MO+AR exhibits a greater tendency to improve. This allowed MO+AR to score the second best O-RMSE values at the end of the search process. Finally, PF provided a more competitive behavior for the (1+1) EA when compared to the use of RJ, which obtained the poorest overall performance.

To further compare the studied constraint-handling approaches, Tables 5.2 and 5.3 detail the results for all two- and three-dimensional test instances at the end of the search process (after 10^6 solution evaluations). The results for each of the instances are given in terms of the best obtained energy value (E_b), the number of performed executions where this solution was found (ν), and the arithmetic mean ($\bar{E}$). In addition, the O-RMSE measure is provided at the bottom of the tables. The lowest average energy obtained for each of the instances, as well as the best O-RMSE values, appears **shaded** in these tables. As it can be seen from the tables, the use of the three multi-objective strategies improved the average performance of the algorithm in the vast majority of the cases with respect to the RJ and PF methods. An interesting behavior can be observed with regard to the MO+AR and MO+PB approaches. While MO+AR tends to perform better than MO+PB for the shortest test sequences, MO+PB scored more competitive results for the largest ones. This suggests that, by not explicitly introducing a search bias, MO+AR yields a broader exploration. Nevertheless, an explicit and more effective bias seems to be required if the hardness of the problem instances increases. Note, however, that MO+AR was found when analyzing Figure 5.12 to present a greater

Table 5.2: Results obtained by the (1+1) EA when using the studied constraint-handling strategies. Two-dimensional test cases.

			RJ		PF		MO+AR		MO+PB		MO+AR+PB	
Seq.	$\boldsymbol{\ell}$	$\boldsymbol{E^*}$	$\boldsymbol{E_b}$ **($\boldsymbol{\nu}$)**	$\boldsymbol{\bar{E}}$	$\boldsymbol{E_b}$ **($\boldsymbol{\nu}$)**	$\boldsymbol{\bar{E}}$	$\boldsymbol{E_b}$ **($\boldsymbol{\nu}$)**	$\boldsymbol{\bar{E}}$	$\boldsymbol{E_b}$ **($\boldsymbol{\nu}$)**	$\boldsymbol{\bar{E}}$	$\boldsymbol{E_b}$ **($\boldsymbol{\nu}$)**	$\boldsymbol{\bar{E}}$
2d1	**18**	**-4**	-4 (6)	-2.62	-4 (54)	-3.54	-4 (100)	**-4.00**	-4 (86)	-3.86	-4 (100)	**-4.00**
2d2	**18**	**-8**	-7 (65)	-6.62	-8 (54)	-7.54	-8 (84)	-7.84	-8 (83)	-7.83	-8 (86)	**-7.86**
2d3	**18**	**-9**	-8 (50)	-7.46	-9 (43)	-8.42	-9 (70)	-8.70	-9 (18)	-8.18	-9 (86)	**-8.86**
2d4	**20**	**-9**	-8 (16)	-6.67	-9 (64)	-8.60	-9 (98)	**-8.97**	-9 (33)	-8.29	-9 (95)	-8.95
2d5	**20**	**-10**	-8 (32)	-7.24	-10 (39)	-8.99	-10 (86)	**-9.86**	-10 (44)	-9.10	-10 (58)	-9.55
2d6	**24**	**-9**	-9 (1)	-7.03	-9 (46)	-8.45	-9 (83)	**-8.83**	-9 (33)	-8.30	-9 (69)	-8.69
2d7	**25**	**-8**	-8 (1)	-5.68	-8 (15)	-7.01	-8 (43)	-7.41	-8 (24)	-7.19	-8 (50)	**-7.49**
2d8	**36**	**-14**	-12 (7)	-9.82	-13 (10)	-11.28	-14 (1)	-11.36	-13 (8)	-11.34	-14 (2)	**-11.58**
2d9	**48**	**-23**	-19 (2)	-14.88	-20 (2)	-16.79	-23 (1)	-17.69	-20 (2)	-17.41	-21 (2)	**-17.83**
2d10	**50**	**-21**	-19 (1)	-14.78	-21 (1)	-16.49	-20 (6)	-17.04	-21 (1)	-17.46	-21 (2)	**-17.77**
2d11	**60**	**-36**	-32 (1)	-26.12	-32 (3)	-28.20	-34 (1)	-27.81	-32 (2)	-28.75	-33 (2)	**-28.81**
2d12	**64**	**-42**	-31 (1)	-24.00	-31 (7)	-26.04	-33 (1)	-26.25	-32 (3)	**-28.13**	-33 (2)	-26.68
2d13	**85**	**-53**	-41 (1)	-34.50	-44 (1)	-37.75	-45 (1)	-35.54	-45 (1)	**-39.03**	-46 (1)	-38.78
2d14	**100**	**-48**	-38 (1)	-29.69	-41 (2)	-34.55	-41 (1)	-32.90	-40 (1)	-34.74	-42 (1)	**-34.78**
2d15	**100**	**-50**	-40 (1)	-31.30	-40 (1)	-34.70	-40 (1)	-32.60	-41 (2)	-35.73	-42 (3)	**-36.23**
O-RMSE			31.08%		19.93%		17.43%		17.95%		**15.65%**	

Table 5.3: Results obtained by the (1+1) EA when using the studied constraint-handling strategies. Three-dimensional test cases.

			RJ		PF		MO+AR		MO+PB		MO+AR+PB	
Seq.	$\boldsymbol{\ell}$	$\boldsymbol{E^*}$	$\boldsymbol{E_b}$ **($\boldsymbol{\nu}$)**	$\boldsymbol{\bar{E}}$	$\boldsymbol{E_b}$ **($\boldsymbol{\nu}$)**	$\boldsymbol{\bar{E}}$	$\boldsymbol{E_b}$ **($\boldsymbol{\nu}$)**	$\boldsymbol{\bar{E}}$	$\boldsymbol{E_b}$ **($\boldsymbol{\nu}$)**	$\boldsymbol{\bar{E}}$	$\boldsymbol{E_b}$ **($\boldsymbol{\nu}$)**	$\boldsymbol{\bar{E}}$
3d1	**20**	**-11**	-11 (58)	-10.43	-11 (93)	-10.92	-11 (100)	**-11.00**	-11 (95)	-10.95	-11 (99)	-10.99
3d2	**24**	**-13**	-13 (13)	-11.07	-13 (54)	-12.37	-13 (94)	**-12.94**	-13 (43)	-12.16	-13 (77)	-12.74
3d3	**25**	**-9**	-9 (58)	-8.42	-9 (97)	-8.97	-9 (100)	**-9.00**	-9 (98)	-8.98	-9 (100)	**-9.00**
3d4	**36**	**-18**	-18 (14)	-15.24	-18 (25)	-16.08	-18 (46)	-16.97	-18 (23)	-16.37	-18 (57)	**-17.27**
3d5	**46**	**-35**	-28 (2)	-24.06	-30 (3)	-25.33	-32 (1)	**-27.30**	-30 (1)	-25.80	-31 (2)	-26.89
3d6	**48**	**-31**	-28 (2)	-22.70	-29 (1)	-24.00	-30 (1)	**-26.04**	-30 (1)	-24.71	-29 (7)	-25.49
3d7	**50**	**-34**	-27 (1)	-21.15	-27 (3)	-22.19	-30 (1)	**-24.68**	-28 (2)	-23.17	-28 (2)	-23.87
3d8	**58**	**-44**	-33 (3)	-26.74	-34 (3)	-28.75	-37 (1)	-29.76	-34 (7)	-29.97	-37 (1)	**-30.50**
3d9	**60**	**-55**	-46 (2)	-38.30	-47 (3)	-40.67	-48 (2)	-40.00	-48 (1)	-41.13	-49 (2)	**-42.01**
3d10	**64**	**-59**	-45 (2)	-34.88	-47 (1)	-36.44	-48 (1)	-38.86	-50 (1)	**-39.01**	-50 (1)	-38.78
3d11	**67**	**-56**	-38 (3)	-30.62	-41 (1)	-32.22	-40 (2)	-33.30	-41 (1)	-33.60	-42 (1)	**-33.97**
3d12	**88**	**-72**	-46 (1)	-36.44	-49 (1)	-37.15	-47 (5)	-38.96	-48 (1)	**-40.15**	-51 (1)	-40.14
3d13	**103**	**-58**	-39 (1)	-29.25	-38 (2)	-29.67	-39 (1)	-29.70	-39 (1)	**-33.10**	-40 (1)	-31.01
3d14	**124**	**-75**	-45 (2)	-33.32	-46 (2)	-34.11	-46 (1)	-34.29	-52 (1)	**-39.28**	-50 (1)	-36.26
3d15	**136**	**-83**	-51 (1)	-37.66	-51 (1)	-37.94	-50 (1)	-38.11	-51 (1)	**-43.41**	-51 (2)	-40.42
O-RMSE			34.76%		31.00%		27.94%		28.05%		**27.36%**	

Table 5.4: Statistical significance analysis for comparing the performance of the (1+1) EA when implementing the different analyzed constraint-handling approaches.

	Two-dimensional instances															Three-dimensional instances															
	2d1	2d2	2d3	2d4	2d5	2d6	2d7	2d8	2d9	2d10	2d11	2d12	2d13	2d14	2d15	3d1	3d2	3d3	3d4	3d5	3d6	3d7	3d8	3d9	3d10	3d11	3d12	3d13	3d14	3d15	Overall
PF / RJ	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+					**26**+ **0**−
MO+AR / RJ	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+				**27**+ **0**−
MO+PB / RJ	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	**30**+ **0**−
MO+AR+PB / RJ	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	**30**+ **0**−
MO+AR / PF	+	+	+	+	+	+	+		+	+			−	−	−	+	+		+	+	+	+	+		+	+	+				**19**+ **3**−
MO+PB / PF	+	+	−	−		−	+		+	+		+	+		+						+	+	+		+	+	+	+	+	+	**17**+ **3**−
MO+AR+PB / PF	+	+	+	+	+	+	+	+	+	+			+		+	+	+		+	+	+	+	+	+	+	+	+	+	+	+	**26**+ **0**−
MO+PB / MO+AR	−		−	−	−	−	−				+	+	+	+	+	−	−		−	−	−	−		+				+	+	+	**9**+ **12**−
MO+AR+PB / MO+AR			+		−	−				+	+		+	+	+		−		+		−	−		+				+	+	+	**11**+ **5**−
MO+AR+PB / MO+PB	+		+	+	+	+	+		+			−					+		+	+	+	+						−	−	−	**12**+ **4**−

tendency to improve as the search process progresses. The MO+AR strategy could thus be expected to meet or even exceed the results of MO+PB for the largest test cases if the algorithm is allowed to run for a higher number of solution evaluations. The best overall performance was exposed by the MO+AR+PB method. By combining the advantages of the two different biasing mechanisms, MO+AR+PB decreased the O-RMSE measure by 15.43% and 4.28% with respect to RJ and PF in the two-dimensional instances, respectively, and by 7.4% and 3.64% in the three-dimensional case.

Finally, Table 5.4 complements the information provided in Tables 5.2 and 5.3 in order to highlight whether the performance differences between the studied approaches were statistically significant or not. Each row in this table compares two strategies, say A and B, which is denoted as "A / B". If a significant performance difference exists between A and B for a particular instance, the corresponding cell is either marked + or marked − depending on whether such a difference was in favor of, or against A. Unmarked cells indicate that there was not a statistically important difference between A and B. The rightmost column of the table summarizes the results of this analysis. As shown in Table 5.4, PF and MO+AR significantly outperformed RJ in 26 and 27 of the instances. Both MO+PB and MO+AR+PB achieved a statistically significant performance increase with regard to RJ for all the 30 adopted test sequences. The MO+AR, MO+PB and MO+AR+PB strategies scored significantly better results than PF in 19, 17 and 26 instances, respectively. Nevertheless, MO+AR and MO+PB were each significantly surpassed by PF in 3 of the two-dimensional test cases. By comparing among the multi-objective strategies, it is first possible to confirm that MO+PB was statistically superior to MO+AR in 9 of the largest (hardest) test cases, while significantly inferior to MO+AR for 12 of the smallest (easiest) ones. Finally, the table indicates that MO+AR+PB significantly improved the performance for 11 and 12 of the instances with respect to MO+AR and MO+PB, but there were still important differences favoring MO+AR and MO+PB respectively in 5 and 4 of the cases.

5.6.3 Analysis for a population-based algorithm

As the population-based method, the genetic algorithm (GA) described in detail in Section 4.5.2 has been considered. The implementation of the different constraint-handling strategies influences the selection process (at both the $selection\text{-}for\text{-}variation$ and the $selection\text{-}for\text{-}survival$ stages), which is a major determinant of the GA's behavior. Hence, by evaluating the performance of the GA, it will be possible to inquire into the advantages of using the four studied constraint-handling approaches: the reject (RJ) and penalty function (PF) methods adopted as a reference, and the multi-objective approaches introducing a search bias by means of feasibility rules (MO+FR) and proportional biasing (MO+PB). When using the RJ and PF methods, selection is to be based on a single-objective discrimination among candidate individuals. In contrast, when applying the MO+FR and MO+PB strategies, the Pareto-dominance relation is used to impose a partial order among individuals. The *nondominated sorting* procedure is used at the $selection\text{-}for\text{-}survival$ stage, as it is implemented within the *Non-dominated Sorting Genetic Algorithm II*, NSGA-II [59]. This procedure is outlined in Section 4.5.2. Originally, NSGA-II uses also the so-called *crowding distance* measure as a secondary discrimination criterion in order to promote population diversity [59]. In this study, however, this measure has not been incorporated to avoid attributing the performance that the GA achieves through the use of the multi-objective strategies to such a diversification mechanism.

An internal coordinates representation based on relative moves has been adopted (Section 2.3.3.3). The implemented mating strategy and variation operators are the same as used in Section 4.5.2. It is worthy to mention that preliminary testing was conducted in order to explore the effects of preventing duplicate individuals from the population. As a result, the performance of the different analyzed constraint-handling methods was significantly improved in all the cases when duplicate individuals were removed from the population; this mechanism was enabled for all the reported experiments. Finally, a maximum number of 10^6 evaluations was used as the termination criterion.

The four studied constraint-handling strategies are first evaluated in Section 5.6.3.1 under different parameter settings for the GA. The purpose of such an initial evaluation is to identify the most appropriate GA conditions for each of the approaches, to be adopted during the more detailed comparative analysis presented later in Section 5.6.3.2.

5.6.3.1 Settings for the genetic algorithm

The RJ, PF, MO+FR and MO+PB strategies are evaluated under different GA conditions. Three recombination and mutation probabilities were considered: $p_c \in \{0.8, 0.9, 1.0\}$, $p_m \in \{\frac{1}{\ell-2}, \frac{2}{\ell-2}, \frac{3}{\ell-2}\}$. Thus, a total of 9 configurations of the implemented GA are investigated. The population size was fixed to $N = 100$ in all the cases and a total of 31 repetitions for each experiment were performed. Figure 5.13 plots the O-RMSE scored by the four studied constraint-handling approaches when using the different GA settings. The PF and MO+PB strategies require the tuning of the penalty weight (ρ) and the bias strength (ω) parameters, respectively. For each evaluated configuration of the GA, the results of PF and MO+PB reported in Figure 5.13 correspond to the best O-RMSE obtained by considering a diverse set of values for the respective parameters (see Section 5.5.3 and 5.6.1.2).

It is evident from Figure 5.13 that both MO+FR and MO+PB achieved lower O-RMSE values in all cases when compared with respect to RJ and PF. The MO+PB strategy tends to perform better than MO+FR for most GA settings, particularly when focusing on the two-dimensional instances. Finally, the plots indicate that the use of PF yields better results in comparison to the use of RJ in most cases. In general, no clear tendency in the GA's performance can be distinguished with respect to the variation in the recombination probability. It is possible to observe, however, that regardless of the constraint-handling strategy used the GA responded positively to the increased mutation rate.

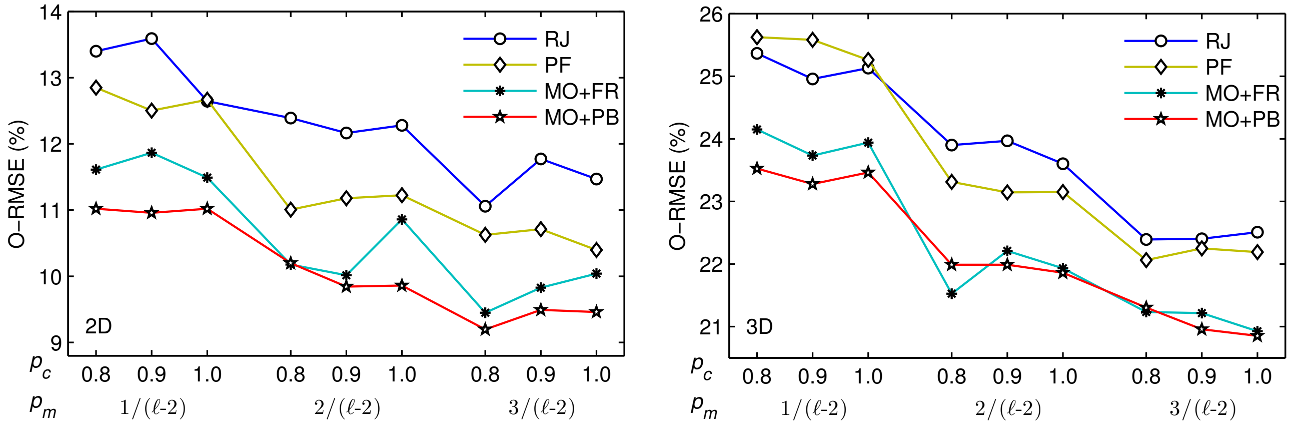


Figure 5.13: Evaluating the studied constraint-handling mechanisms under different parameter settings for the implemented GA. Two-dimensional (left) and three-dimensional (right) test instances.

For further analyses presented in Section 5.6.3.2, the settings for the GA which allowed each of the compared approaches to reach the lowest O-RMSE value have been selected. The selected recombination probabilities are as follows: (i) two-dimensional instances, $p_c = 0.8$ for RJ, MO+FR and MO+PB, and $p_c = 1.0$ for PF; (ii) three-dimensional instances, $p_c = 0.8$ for RJ and PF, and $p_c = 1.0$ for MO+FR and MO+PB. The mutation probability was set to $p_m = \frac{3}{\ell-2}$ in all the cases.

5.6.3.2 Comparative analysis

A detailed comparative analysis among the RJ, PF, MO+FR and MO+PB strategies is presented in this section. In the experimentation here reported, the best performing parameter settings for PF and MO+PB are considered; refer to Section 5.5.3 and 5.6.1.2 for details. Likewise, the best performing GA conditions for each of the compared approaches were adopted (Section 5.6.3.1).

Figure 5.14 shows the online convergence (measured in terms of the O-RMSE) presented by the GA when using the different constraint-handling approaches. The progress in the search is reported in slots of $50,000$ solution evaluations until reaching the maximum allowed number of 10^6 evaluations. This figure is quite revealing in several respects. First, it is possible to note from the plots that the best results at the end of the search process were obtained by using the multi-objective strategies (MO+FR and MO+PB), in both the two- and the three-dimensional test cases. The RJ method, which exhibited the worst performance at the end, scored the best O-RMSE values at the beginning

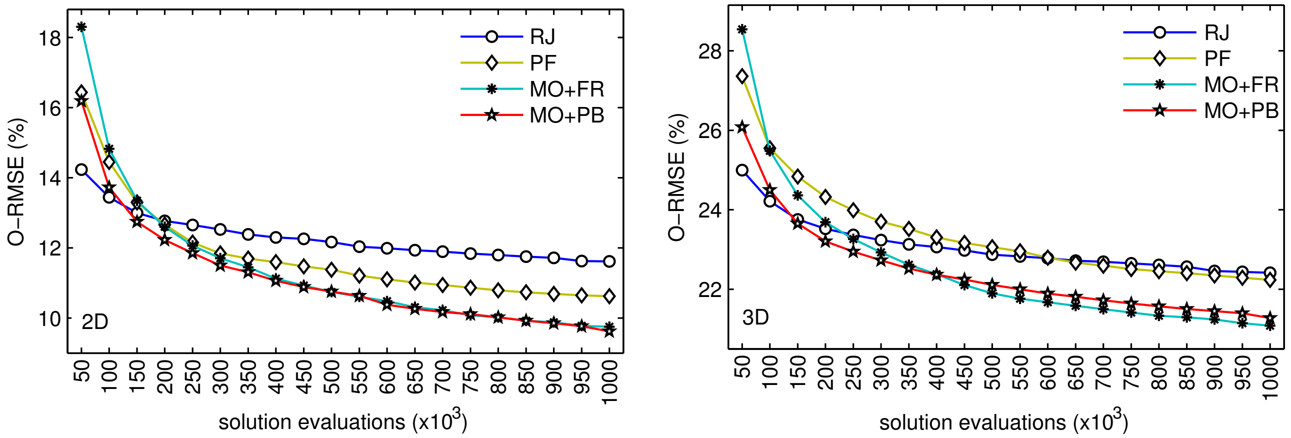


Figure 5.14: O-RMSE scored by the implemented GA as the search process progressed. Two-dimensional (left) and three-dimensional (right) test instances.

of the search. Thus, the use of RJ enabled a faster convergence towards moderate-quality individuals. Given that PF, MO+FR and MO+PB invest an additional amount of effort in evaluating infeasible protein conformations, these strategies require more time (*i.e.*, they consume more objective function evaluations) to locate promising regions of the solution space. By allowing the algorithm to move through infeasible states, however, these methods are more likely to reach better results at the end of the optimization process; as it can be perceived from the slope in the corresponding convergence curves. Finally, although MO+FR and MO+PB competed with the best O-RMSE values at the end, it is important to observe that MO+FR showed a significantly inferior performance at the first stages of the search (indeed the poorest performance among all the four compared techniques). This is because the bias introduced in MO+FR is not as restrictive as that involved in MO+PB and PF, so that MO+FR dedicates more resources to the exploration of infeasible regions of the landscape.

Tables 5.5 and 5.6 detail the above presented results of the GA after 10^6 solution evaluations. The interpretation of these tables is the same as for Tables 5.2 and 5.3, refer to Section 5.6.2. Both the MO+FR and MO+PB strategies reached a better average energy for most of the instances, thereby lowering the O-RMSE, in comparison with RJ and PF. While MO+PB scored the best O-RSME value for the two-dimensional instances, MO+FR obtained the lowest value for this measure in the three-dimensional case. Even though no definite conclusions can be drawn regarding the superiority of the multi-objective methods with respect to each other, it is possible to see from the tables that

Table 5.5: Details of the results obtained by the implemented GA when using the studied constraint-handling strategies. Two-dimensional test cases.

			RJ		PF		MO+FR		MO+PB	
Seq.	ℓ	E^*	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$
2d1	**18**	**-4**	-4 (98)	-3.98	-4 (100)	**-4.00**	-4 (100)	**-4.00**	-4 (100)	**-4.00**
2d2	**18**	**-8**	-8 (100)	**-8.00**	-8 (100)	**-8.00**	-8 (100)	**-8.00**	-8 (100)	**-8.00**
2d3	**18**	**-9**	-9 (100)	**-9.00**	-9 (100)	**-9.00**	-9 (99)	-8.99	-9 (100)	**-9.00**
2d4	**20**	**-9**	-9 (99)	-8.99	-9 (100)	**-9.00**	-9 (100)	**-9.00**	-9 (100)	**-9.00**
2d5	**20**	**-10**	-10 (97)	-9.94	-10 (100)	**-10.00**	-10 (100)	**-10.00**	-10 (100)	**-10.00**
2d6	**24**	**-9**	-9 (93)	-8.93	-9 (86)	-8.86	-9 (94)	-8.94	-9 (96)	**-8.96**
2d7	**25**	**-8**	-8 (57)	-7.57	-8 (82)	-7.82	-8 (95)	**-7.95**	-8 (90)	-7.90
2d8	**36**	**-14**	-14 (2)	-11.76	-14 (2)	-12.11	-14 (4)	-11.95	-14 (3)	**-12.27**
2d9	**48**	**-23**	-22 (2)	-18.90	-22 (2)	-18.96	-22 (7)	**-19.67**	-22 (6)	-19.58
2d10	**50**	**-21**	-21 (15)	-19.19	-21 (15)	-19.17	-21 (33)	**-20.14**	-21 (31)	-19.77
2d11	**60**	**-36**	-33 (3)	-30.10	-34 (2)	-30.78	-35 (1)	**-31.37**	-34 (2)	-30.90
2d12	**64**	**-42**	-39 (1)	**-33.34**	-38 (1)	-32.47	-37 (1)	-31.49	-38 (2)	-32.82
2d13	**85**	**-53**	-48 (2)	-42.96	-47 (3)	-43.04	-49 (2)	-43.20	-48 (1)	**-43.59**
2d14	**100**	**-48**	-40 (2)	-35.70	-41 (2)	-36.46	-43 (1)	**-37.12**	-43 (1)	-37.02
2d15	**100**	**-50**	-43 (2)	-37.61	-44 (1)	-38.09	-44 (2)	**-39.47**	-43 (1)	-38.91
O-RMSE			11.61%		10.62%		9.75%		**9.62%**	

Table 5.6: Details of the results obtained by the implemented GA when using the studied constraint-handling strategies. Three-dimensional test cases.

			RJ		PF		MO+FR		MO+PB	
Seq.	ℓ	E^*	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$	E_b (ν)	$\bar{E}$
3d1	**20**	**-11**	-11 (100)	**-11.00**	-11 (100)	**-11.00**	-11 (100)	**-11.00**	-11 (100)	**-11.00**
3d2	**24**	**-13**	-13 (100)	**-13.00**	-13 (98)	-12.98	-13 (98)	-12.96	-13 (97)	-12.96
3d3	**25**	**-9**	-9 (98)	-8.98	-9 (99)	-8.99	-9 (100)	**-9.00**	-9 (100)	**-9.00**
3d4	**36**	**-18**	-18 (32)	-16.54	-18 (31)	-16.57	-18 (35)	**-16.84**	-18 (40)	-16.79
3d5	**46**	**-35**	-32 (1)	-27.78	-34 (1)	-28.45	-32 (4)	**-28.92**	-32 (1)	-28.34
3d6	**48**	**-31**	-30 (1)	-26.94	-31 (1)	-27.18	-31 (1)	-27.39	-30 (4)	**-27.45**
3d7	**50**	**-34**	-31 (2)	**-27.91**	-31 (1)	-27.75	-32 (1)	-27.40	-31 (1)	-27.78
3d8	**58**	**-44**	-38 (2)	-33.30	-38 (1)	-33.07	-41 (1)	**-34.52**	-38 (2)	-33.72
3d9	**60**	**-55**	-49 (2)	-43.61	-49 (1)	-44.02	-50 (1)	-44.45	-49 (1)	**-44.60**
3d10	**64**	**-59**	-53 (1)	-47.02	-53 (1)	-47.15	-51 (2)	-45.63	-52 (3)	**-47.54**
3d11	**67**	**-56**	-42 (2)	-37.71	-44 (1)	-37.78	-44 (3)	**-38.68**	-44 (1)	-38.28
3d12	**88**	**-72**	-52 (4)	-44.75	-50 (5)	-44.97	-54 (1)	**-47.30**	-53 (1)	-46.36
3d13	**103**	**-58**	-40 (1)	-34.36	-43 (1)	-33.76	-40 (5)	**-35.35**	-41 (1)	-35.16
3d14	**124**	**-75**	-53 (1)	-41.34	-50 (2)	-41.38	-51 (1)	**-43.56**	-53 (1)	-43.08
3d15	**136**	**-83**	-53 (1)	-44.44	-54 (2)	-44.69	-55 (1)	**-46.94**	-55 (2)	-46.28
O-RMSE			22.42%		22.24%		**21.08%**		21.27%	

Table 5.7: Statistical analysis for comparing the performance of the GA when using the different constraint-handling approaches analyzed.

	Two-dimensional instances															**Three-dimensional instances**															
	2d1	**2d2**	**2d3**	**2d4**	**2d5**	**2d6**	**2d7**	**2d8**	**2d9**	**2d10**	**2d11**	**2d12**	**2d13**	**2d14**	**2d15**	**3d1**	**3d2**	**3d3**	**3d4**	**3d5**	**3d6**	**3d7**	**3d8**	**3d9**	**3d10**	**3d11**	**3d12**	**3d13**	**3d14**	**3d15**	**Overall**
PF / RJ							+	+			+	−		+						+											**5**+ **1**−
MO+FR / RJ							+		+	+	+	−		+	+					+	+	−	+	+	−	+	+	+	+	+	**15**+ **3**−
MO+PB / RJ							+	+	+	+	+		+	+	+					+	+			+			+	+	+	+	**15**+ **0**−
MO+FR / PF							+		+	+	+	−		+	+					+			+		−	+	+	+	+	+	**13**+ **2**−
MO+PB / PF						+			+	+					+								+	+			+	+	+	+	**10**+ **0**−
MO+PB / MO+FR								+		−	−	+								−			−		+		−				**3**+ **5**−

MO+FR achieved a better $\bar{E}$ value in most cases. Finally, despite the poorer overall performance of RJ, this strategy outperformed the other three approaches at solving one of the two-dimensional instances (2d12) and a couple of the three-dimensional test cases (3d2 and 3d7).

Table 5.7 outlines the results of the statistical significance analysis and is interpreted as described in Section 5.6.2 with regard to Table 5.4. As it can be observed from Table 5.7, no significant performance differences between the four compared approaches were found when dealing with the smallest test instances. The four studied methods scored similarly competitive results. PF was significantly superior to RJ in 5 of the instances, but significantly inferior at solving the 2d12 instance. Both MO+FR and MO+PB significantly increased the performance of the GA in 15 of the test cases with respect to RJ. In addition, these multi-objective strategies statistically outperformed PF in 13 and 10 instances. Nonetheless, MO+FR presented a significantly lower performance in 3 and 2 of the test cases in comparison to RJ and PF, respectively. Finally, there was a statistically significant difference between the multi-objective strategies for 8 of the instances; in 3 out of these cases such a significant difference favors MO+PB, while it favors MO+FR in the 5 remaining cases.

5.7 Discussion and conclusions

The multi-objective (MO) approach to constraint-handling has been investigated in the context of the HP model for protein structure prediction (PSP), a highly constrained optimization problem. The HP model was reformulated as an unconstrained multi-objective problem by treating constraints as an additional objective function. Rather than discriminating feasible from infeasible solutions, the MO strategy defines trade-offs between quality (original objective function) and feasibility. This gives infeasible solutions the opportunity to be considered and exploited during optimization.

In the first part of this study, a thorough fitness landscape analysis was conducted in order to evaluate the effects that the (single-objective to multi-objective) problem transformation involves. As a result, it was found that a significant portion of the infeasibility translates into landscape neutrality. Under the MO problem formulation, it is possible for an infeasible solution to become part of the neutral neighborhood of a feasible solution. This has prompted an important increase in the neutrality degree of solutions and, consequently, in the size of the neutral networks (NNs). Such a landscape transformation has led to the establishment of neutral connections between feasible and infeasible NNs. Through a series of neutral connections, however, it is possible to bridge different regions of the feasible space, potentially belonging to diverse fitness classes (this can be especially useful when dealing with disjoint feasible spaces). By being allowed to traverse (originally inaccessible) infeasible areas, a search algorithm can thus exploit these neutral connections in the form of new neutral paths to navigate the landscape. The new defined neutral paths can not only be shorter than the existing feasible paths, but can also play a central role in helping the algorithm to escape from local optima.

Despite the aforementioned advantages that the alternative multi-objective landscape entails, an excessive increase in neutrality may also prevent a search algorithm from moving in the correct direction. The conducted landscape analysis not only reported a considerable growth in the size of the NNs due to the use of the MO problem formulation. It was also found that these NNs are mainly composed of infeasible solutions. Without a proper search bias, therefore, the computational resources can be exhausted by exploring uninteresting areas of the solution space (as it was also pointed out by Runarsson and Yao [188]). From the fitness landscape perspective, providing the

MO strategy with a search bias can be understood as the removal of part of the neutrality that this strategy originally introduces. The goal becomes, thus, to benefit from having access to the infeasible areas of the landscape, at the same time that most of the effort is invested in exploiting promising search directions. The second part of this chapter studied the effectiveness of different mechanisms for biasing the search towards the feasible region, which can be coupled to the MO constraint-handling strategy. Three different biasing mechanisms were evaluated; namely, the use of an archiving strategy, the incorporation of a secondary discrimination criterion (use of feasibility rules), and the application of a proportional bias dependent on the degree of constraint violation. On the one hand, the results of such an evaluation confirmed the need for performing a well-biased search when using the MO strategy. The behavior of the considered search algorithms was significantly improved with the implementation of all the three studied biasing approaches. On the other hand, it was also possible to observe that a very strong bias could lead to override the positive effects of the landscape transformation. Thus, the task of identifying the most appropriate amount of bias for a particular problem and search algorithm, could be not as straightforward as might be thought.

In the last part of this study, the MO constraint-handling strategy was further explored by carrying out a comparative analysis where two different approaches from the literature were considered; namely, a rejecting strategy (RJ) where the search is confined to the space of only feasible conformations, and a penalty function (PF) where infeasible solutions are penalized according to the number of conflicts they present. The different strategies were evaluated in terms of the performance of basic evolutionary algorithms. As a result, the use of the MO strategy significantly improved the performance of the implemented algorithms when compared with respect to both the RJ and PF methods. This highlights the suitability of the proposed MO approach. It was also found that PF scored better results in most cases with regard to the RJ strategy. The fact that both MO and PF performed better than the RJ strategy, and that RJ requires a considerable amount of additional computational resources, gives further support to the belief that considering infeasible protein conformations may contribute to the design of more competitive algorithms for solving the HP model of the PSP problem; this has been a subject of concern in the specialized literature.

To the best of the author's knowledge, the results of this research project represent the first efforts on the use of multi-objective optimization methods to face the constraint-handling requirement which arises when dealing with the HP model of the PSP problem. Basic evolutionary algorithms have been used in this study for evaluating the suitability of this approach. From the obtained results, it is expected that the MO strategy can be incorporated as a means of improving the performance of established state-of-the-art algorithms for solving this problem. This issue needs to be investigated in order to derive more general conclusions. Furthermore, the present study explored for the first time, as far as the author is aware, the potential effects of implementing the MO constraint-handling strategy through a fitness landscape analysis. Although such an analysis focused on a particular case of study, the HP model of the PSP problem, similar effects to those observed with regard to the landscape transformation can be expected from the use of the MO strategy in other problem domains. Therefore, the performed analysis contributes to the general understanding of the MO approach for handling constraints.

6

Conclusions and future work

This thesis project was concerned with the analysis and design of alternative evaluation schemes to face the main optimization challenges involved with the prediction of protein structures under the HP model. These challenges relate to the neutrality, multimodality and infeasibility that characterizes the fitness landscapes of this hard combinatorial problem. The present document has reported all the efforts made and the results obtained during the development of this research work. The purpose of this concluding chapter is to summarize the main achieved findings and contributions, as well as to discuss some possible directions that can be taken in order to extend this research.

6.1 Main findings and contributions

Neutrality

- Chapter 3 presented a comparative study of several alternative energy functions proposed in the literature with the primary aim of addressing the neutrality of the HP model. The discrimination capabilities, the consistency with the original problem's definition, and the effectiveness of the different functions to guide the search process, were the focus of the performed study.

- The results of the conducted analysis indicate that both, the discrimination potential and the compatibility with the problem, are two major factors determining the ability of an evaluation function to drive effectively the search process. Functions excelling on the two mentioned criteria consistently presented a promising behavior, improving the performance of search algorithms with respect to the conventional problem formulation. Finally, it was observed that, through a proper algorithm design, it is also possible to take advantage of the inherent neutrality of the HP model's fitness landscape. This provides further support to previous findings on this respect that have been recently reported in the literature [39, 154–156, 228, 231, 242].

- To the best of the author's knowledge, this is the first reported comparative analysis of alternative evaluation schemes for the HP model.

- Preliminary results of this research work have been published in the following specialized international conferences:

  - IEEE Congress on Evolutionary Computation, CEC. New Orleans, LA, USA. 2011 [68].
  - International Conference on Bioinformatics & Computational Biology, BIOCOMP.
    Las Vegas, NV, USA. 2011 [77].

  Also, the full study, as presented in Chapter 3, has been recently published in the Journal of Computer Science and Technology (JCST) [71].

Multimodality

- Chapter 4 explored the suitability of multi-objectivization for dealing with multimodality. Multi-objectivization was accomplished by means of the decomposition of the original energy function of the HP model. Three different multi-objective formulations of the problem were proposed. The effects of the problem transformation were analyzed in detail, and the advantages of the multi-objective formulations in terms of the performance of search algorithms were investigated by comparing with respect to the conventional single-objective formulation of the HP model.

- Multi-objectivization introduces incomparability among solutions, increasing the neutrality of the fitness landscape. Neutral connections are established, which merges neutral networks into larger connected areas of neutrality. On the one hand, this enhances the exploration behavior of search algorithms, allowing them to move through inferior fitness classes as a means of escaping from local optima. On the other hand, this effect translates into the loss of gradient information, which may prevent algorithms from identifying promising directions. Therefore, this understanding of multi-objectivization could guide the design of more effective algorithms.

- The three proposed multi-objectivization schemes were found to significantly increase the average performance of search algorithms with regard to the conventional single-objective problem formulation. In this way, the effectiveness of the proposed approaches was demonstrated and further support was given to the suitability of multi-objectivization to address multimodality.

- According to the performed revision of the literature, this research reports for the first time the application of multi-objective optimization methods to the particular HP model of the protein structure prediction problem. In addition, no previous work has been found where the effects of multi-objectivization are investigated through the explicit sampling and evaluation of the fitness landscape. The findings of such an analysis are generalizable to other problem domains, contributing, thus, to advance the general understanding of multi-objectivization.

- The research work reported in Chapter 4 has led to three publications in the following international conferences:

  - European Conference on Evolutionary Computation in Combinatorial Optimization, EvoCOP. Málaga, Spain. 2012 [70].

  - Genetic and Evolutionary Computation Conference, GECCO. Philadelphia, PA, USA. 2012 [78].

  - International Conference on Parallel Problem Solving From Nature, PPSN. Taormina, Italy. 2012 [69].

Also, the full study has been recently submitted to the special issue on Evolutionary Multi-objective Optimization of The European Journal of Operational Research (EJOR) [76].

Infeasibility

- In Chapter 5, the use of multi-objective optimization as a constraint-handling strategy for the HP model was proposed. By treating constraints as an additional objective function, the HP model was restated as an unconstrained multi-objective problem. The impact of the problem transformation on the fitness landscape was analyzed. Also, the relevance of introducing a proper search bias when using this strategy was explored. Finally, the suitability of the multi-objective approach was evaluated in terms of its ability to effectively guide the search process.

- The multi-objective approach to handle constraints allows for the consideration of trade-offs between quality and feasibility, so that it is possible to traverse (and to exploit useful information from) infeasible areas of the fitness landscape. An important fraction of infeasibility translates into neutrality. The introduced neutrality defines new, potentially shorter paths to move through the landscape, which can also be exploited as a means of escaping from local optima.

- The lack of search bias can lead to the investment of a considerable amount of computational effort in evaluating infeasible solutions. Through the use of different biasing mechanisms, it was possible to impact favorably on the ability of the multi-objective strategy to guide the search process. This highlights the relevance of introducing a bias; but defining the proper amount of bias to be applied could be a non-trivial, problem- and algorithm-dependent task.

- The effectiveness of the multi-objective strategy was demonstrated through a comparative analysis with respect to commonly adopted techniques from the literature. The use of the multi-objective strategy significantly improved the performance of the implemented algorithms. The obtained results support also the belief that considering infeasible solutions during optimization may contribute to the design of more competitive metaheuristics for solving the HP model.

- This research work produced the first efforts with regard to the use of multi-objective optimization for handling the constraints of the HP model, to the author's knowledge. Also, this is the first time that a fitness landscape analysis has been conducted with the aim of investigating the effects of the multi-objective constraint-handling strategy. Although focused on the particular case of study of this research, such an analysis is expected to contribute to the general understanding of the functioning of the multi-objective constraint-handling technique.

- Preliminary results of this work have been presented in the international conference IEEE Congress on Evolutionary Computation (CEC), held in Cancún, México, in 2013 [79]. Also, the full study presented in Chapter 5 has been recently submitted and is currently under review as a candidate for publication in the Computers & Operations Research (COR) Journal [67].

General comments

Chapters 3, 4 and 5 presented encapsulated efforts to face a very specific difficulty (neutrality, multimodality and infeasibility, respectively) with regard to the design of metaheuristic algorithms for predicting protein structures under the HP model. The isolated nature and lack of integration of these efforts can be seen as the main weakness of this thesis. Hence, to address such a weakness should be considered as an imperative future research direction, as discussed further in Section 6.2.

Nevertheless, it has been possible during the development of this project to identify some well-defined connections between the studies reported in the different chapters of this document. These connections can be summarized as follows. Chapter 3 explored the use of fine-grained evaluation schemes to break the neutrality of the fitness landscape. A fine-grained discrimination was found to effectively improve the performance of search algorithms by facilitating convergence in the direction of local optima. However, breaking neutrality also results in a more rugged landscape, which increases multimodality. Therefore, a fine-grained evaluation scheme will usually need to be accompanied by an effective mechanism to escape from local optima if success is to be achieved in terms of global convergence. Whereas Chapter 3 sought to reduce the neutrality of the landscape, it was precisely through the addition of even more neutrality that the multi-objective formulations proposed in Chapters 4 and 5 succeeded in dealing with multimodality and infeasibility. The multi-objective formulations of the HP model, both, the ones investigated in Chapter 4 and the one analyzed in Chapter 5, transform the fitness landscape of the problem in such a way that new neutral paths are introduced. In both cases, the introduced neutral paths can be exploited as a means of escaping from local optima, contributing thus to cope with multimodality. While the new neutral paths produced by the transformations studied in Chapter 4 were all defined within the feasible region, the neutral paths analyzed in Chapter 5 traverse infeasible areas of the landscape. Neutral paths traversing infeasible landscape areas can potentially be shorter than paths confined to the feasible space, which can further contribute to increase the search efficiency of metaheuristics for solving the studied problem.

6.2 Future work

Potential extensions to this research work can be broadly described as follows:

- It seems important to complement the understanding of the alternative energy functions studied in Chapter 3, by evaluating how they impact on the characteristics of the fitness landscape; such as it was done in Chapters 4 and 5 to investigate other problem transformations.

- In Chapters 4 and 5, detailed fitness landscape analyses were carried out with the aim of inquiring into the effects of using alternative multi-objective formulations of the HP model (either to deal with multimodality or to handle infeasibility). The conducted analyses focused on neutrality. Extending these analyses to evaluate the problem transformation from the perspective of other different landscape properties, *e.g.*, ruggedness [230], can certainly be seen as a relevant research direction that will contribute to build a more comprehensive picture.

- The fitness landscape analyses presented in Chapters 4 and 5 contribute to the general understanding of the studied single-objective to multi-objective transformations. That is, the achieved understanding is generalizable in the sense that similar effects to those observed can be expected from the application of these particular transformations to other different optimization problems. It seems important, however, to replicate these analyses to problems from other different application domains in order to further support the acquired understanding.

- As discussed at the end of Section 6.1, isolated efforts have been made in this thesis in order to cope with the neutrality, multimodality and infeasibility of the HP model's fitness landscapes. Despite the fruitfulness of such isolated efforts, it remains open the question of how all these efforts can be integrated and exploited to handle simultaneously the multiple challenges that this problem poses. This represents a worthy subject for future research. Potential directions to be followed in this regard include:

  - The design of self-adaptive mechanisms which can alternate between the different evaluation schemes in response to the particular conditions at each stage of the optimization.
  - The implementation of the proposed multi-objective evaluation schemes (either those studied in Chapters 4 or the one reported in Chapter 5) within a supplementary local search operator that can be incorporated into any other algorithm operating with a different evaluation scheme (*e.g.*, the fine-grained evaluation functions considered in Chapter 3).
  - The design of new alternative evaluation schemes for the HP model which can deal effectively with all the three issues (*i.e.*, neutrality, multimodality and infeasibility).

- Finally, it would be interesting to explore whether the multi-objective evaluation schemes proposed in Chapters 4 and 5 can be incorporated as a means of improving the performance of established state-of-the-art algorithms for solving the HP model of the protein structure prediction problem; either by replacing the original evaluation scheme that these algorithms implement, or coupled within a supplementary local search mechanism (as described above).

Bibliography

[1] Anfinsen, C. Principles that Govern the Folding of Protein Chains. *Science 181*, 4096 (1973), 223–230. (cited on pages 2 and 20)

[2] Anfinsen, C., Redfield, R., Choate, W., Page, J., and Carroll, W. Studies on the Gross Structure, Cross-Linkages, and Terminal Sequences in Ribonuclease. *Journal of Biological Chemistry 207*, 1 (1954), 201–210. (cited on pages 2 and 20)

[3] Arnautova, Y., Jagielska, A., and Scheraga, H. A New Force Field (ECEPP-05) for Peptides, Proteins, and Organic Molecules. *The Journal of Physical Chemistry B 110*, 10 (2006), 5025–5044. (cited on pages 23 and 26)

[4] Bacardit, J., Stout, M., Hirst, J., and Krasnogor, N. Data Mining in Proteomics with Learning Classifier Systems. In *Learning Classifier Systems in Data Mining*, vol. 125 of *Studies in Computational Intelligence*. Springer Berlin Heidelberg, 2008, pp. 17–46. (cited on pages ix and 16)

[5] Back, T. Selective Pressure in Evolutionary Algorithms: A Characterization of Selection Mechanisms. In *IEEE Congress on Evolutionary Computation*. 1994, pp. 57–62. (cited on pages 92 and 110)

[6] Barnett, L. Netcrawling - Optimal Evolutionary Search with Neutral Networks. In *IEEE Congress on Evolutionary Computation*, vol. 1. 2001, pp. 30–37. (cited on page 92)

[7] Bazzoli, A., and Tettamanzi, A. A Memetic Algorithm for Protein Structure Prediction in a 3D-Lattice HP Model. In *Applications of Evolutionary Computing*, vol. 3005 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Coimbra, Portugal, 2004, pp. 1–10. (cited on page 33)

[8] Becerra, D., Sandoval, A., Restrepo-Montoya, D., and Nino, L. A Parallel Multi-Objective Ab Initio Approach for Protein Structure Prediction. In *IEEE International Conference on Bioinformatics and Biomedicine*. Hong Kong, China, 2010, pp. 137–141. (cited on pages 12 and 71)

[9] Berenboym, I., and Avigal, M. Genetic Algorithms with Local Search Optimization for Protein Structure Prediction Problem. In *Genetic and Evolutionary Computation Conference*. ACM, Atlanta, GA, USA, 2008, pp. 1097–1098. (cited on pages 4, 33, 40, 45, 46, and 48)

[10] Berger, B., and Leighton, T. Protein Folding in the Hydrophobic-Hydrophilic (HP) Model is NP-complete. In *International Conference on Research in Computational Molecular Biology*. ACM, New York, NY, USA, 1998, pp. 30–39. (cited on pages 3, 25, 28, and 31)

[11] Berrera, M., Molinari, H., and Fogolari, F. Amino Acid Empirical Contact Energy Definitions for Fold Recognition in the Space of Contact Maps. *BMC Bioinformatics 4* (2003), 8. (cited on page 26)

[12] Bitello, R., and Lopes, H. A Differential Evolution Approach for Protein Folding. In *IEEE Symposium on Computational Intelligence, Bioinformatics and Computational Biology*. Toronto, Canada, 2006, pp. 1–5. (cited on page 33)

[13] Blackburne, B., and Hirst, J. Evolution of Functional Model Proteins. *The Journal of Chemical Physics 115*, 4 (2001), 1935–1942. (cited on page 25)

[14] Blaszczyk, M., Gront, D., Kmiecik, S., Ziolkowska, K., Panek, M., and Kolinski, A. Coarse-Grained Protein Models in Structure Prediction. In *Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes*, Springer Series in Bio-/Neuroinformatics. Springer Berlin Heidelberg, 2014, pp. 25–53. (cited on pages 25 and 26)

[15] Blazewicz, J., Lukasiak, P., and Milostan, M. Application of Tabu Search Strategy for Finding Low Energy Structure of Protein. *Artificial Intelligence in Medicine 35*, 1-2 (2005), 135–145. (cited on page 33)

[16] Blum, C., and Roli, A. Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison. *ACM Computing Surveys 35*, 3 (2003), 268–308. (cited on pages 56 and 61)

[17] Bornberg-Bauer, E. Chain Growth Algorithms for HP-type Lattice Proteins. In *International Conference on Computational Molecular Biology*, RECOMB '97. ACM, Santa Fe, New Mexico, USA, 1997, pp. 47–55. (cited on page 26)

[18] Brockhoff, D., Friedrich, T., Hebbinghaus, N., Klein, C., Neumann, F., and Zitzler, E. Do Additional Objectives Make a Problem Harder? In *Genetic and Evolutionary Computation Conference*. ACM, London, England, 2007, pp. 765–772. (cited on pages 11, 12, and 69)

[19] Brooks, B., Brooks, C., Mackerell, A., Nilsson, L., Petrella, R., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R., Post, C., Pu, J., Schaefer, M., Tidor, B., Venable, R., Woodcock, H., Wu, X., Yang, W., York, D., and Karplus, M. CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry 30*, 10 (2009), 1545–1614. (cited on page 23)

[20] Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S., and Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry 4*, 2 (1983), 187–217. (cited on page 23)

[21] Băutu, A., and Luchian, H. Protein Structure Prediction in Lattice Models with Particle Swarm Optimization. In *Swarm Intelligence*, vol. 6234 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2010, pp. 512–519. (cited on page 33)

[22] Bui, L., Abbass, H., and Branke, J. Multiobjective Optimization for Dynamic Environments. In *IEEE Congress on Evolutionary Computation*, vol. 3. Edinburgh, UK, 2005, pp. 2349–2356. (cited on page 12)

[23] Bui, L., Nguyen, M., Branke, J., and Abbass, H. Tackling Dynamic Problems with Multiobjective Evolutionary Algorithms. In *Multiobjective Problem Solving from Nature*, Natural Computing Series. Springer Berlin Heidelberg, 2008, pp. 77–91. (cited on page 12)

[24] Bui, T., and Sundarraj, G. An Efficient Genetic Algorithm for Predicting Protein Tertiary Structures in the 2D HP Model. In *Genetic and Evolutionary Computation Conference*. ACM, Washington DC, USA, 2005, pp. 385–392. (cited on page 33)

[25] Cai, X., Wu, X., Wang, L., Kang, Q., and Wu, Q. Hydrophobic-Polar Model Structure Prediction with Binary-Coded Artificial Plant Optimization Algorithm. *Journal of Computational and Theoretical Nanoscience 10*, 6 (2013), 1550–1554. (cited on page 33)

[26] Cebrián, M., Dotú, I., Van Hentenryck, P., and Clote, P. Protein Structure Prediction on the Face Centered Cubic Lattice by Local Search. In *AAAI Conference on Artificial Intelligence - Volume 1*. AAAI Press, Chicago, IL, USA, 2008, pp. 241–246. (cited on pages 4, 33, 40, 46, and 48)

[27] Chan, H., and Dill, K. Comparing Folding Codes for Proteins and Polymers. *Proteins: Structure, Function, and Bioinformatics 24*, 3 (1996), 335–344. (cited on page 25)

[28] Chandru, V., DattaSharma, A., and Kumar, V. The Algorithmics of Folding Proteins on Lattices. *Discrete Applied Mathematics 127*, 1 (2003), 145–161. (cited on pages 2 and 25)

[29] Chen, B., Li, L., and Hu, J. A Novel EDAs Based Method for HP Model Protein Folding. In *IEEE Congress on Evolutionary Computation*. Trondheim, Norway, 2009, pp. 309–315. (cited on pages 33 and 115)

[30] Chen, M., and Huang, W. A Branch and Bound Algorithm for the Protein Folding Problem in the HP Lattice Model. *Genomics, Proteomics, Bioinformatics 3*, 4 (2005), 225–230. (cited on page 31)

[31] Chira, C. A Hybrid Evolutionary Approach to Protein Structure Prediction with Lattice Models. In *IEEE Congress on Evolutionary Computation*. New Orleans, LA, USA, 2011, pp. 2300–2306. (cited on pages 33, 79, 115, and 116)

[32] Chira, C., and Hatami, N. Hybrid Evolutionary Algorithm with a Composite Fitness Function for Protein Structure Prediction. In *Intelligent Data Engineering and Automated Learning*, vol. 7435 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2012, pp. 184–191. (cited on page 33)

[33] Chira, C., Horvath, D., and Dumitrescu, D. An Evolutionary Model Based on Hill-Climbing Search Operators for Protein Structure Prediction. In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, vol. 6023 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2010, pp. 38–49. (cited on pages 33, 115, and 116)

[34] Chira, C., Horvath, D., and Dumitrescu, D. Hill-Climbing Search and Diversification Within an Evolutionary Approach to Protein Structure Prediction. *BioData Mining 4*, 1 (2011), 1–17. (cited on page 33)

[35] Chu, D., Till, M., and Zomaya, A. Parallel Ant Colony Optimization for 3D Protein Structure Prediction using the HP Lattice Model. In *IEEE International Parallel and Distributed Processing Symposium*. Denver, CO, USA, 2005, p. 193b. (cited on page 33)

[36] Churchill, A., Husbands, P., and Philippides, A. Multi-Objectivization of the Tool Selection Problem on a Budget of Evaluations. In *Evolutionary Multi-Criterion Optimization*,

vol. 7811 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Sheffield, UK, 2013, pp. 600–614. (cited on page 71)

[37] Citrolo, A., and Mauri, G. A Hybrid Monte Carlo Ant Colony Optimization Approach for Protein Structure Prediction in the HP Model. In *Italian Workshop on Artificial Life and Evolutionary Computation*, vol. 130 of *EPTCS*. Milan, Italy, July 2013, pp. 61–69. (cited on page 33)

[38] Clementi, C. Coarse-grained Models of Protein Folding: Toy Models or Predictive Tools? *Current Opinion in Structural Biology 18*, 1 (2008), 10–15. (cited on pages 2 and 25)

[39] Collard, P., Clergue, M., and Defoin-Platel, M. Synthetic Neutrality for Artificial Evolution. In *Artificial Evolution*, vol. 1829 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Dunkerque, France, 2000, pp. 254–265. (cited on pages 67 and 156)

[40] Corne, D., and Knowles, J. Techniques for Highly Multiobjective Optimisation: Some Nondominated Points are Better than Others. In *Genetic and Evolutionary Computation Conference*, vol. 1. ACM, London, UK, 2007, pp. 773–780. (cited on page 49)

[41] Cornell, W., Cieplak, P., Bayly, C., Gould, I., Merz, K., Ferguson, D., Spellmeyer, D., Fox, T., Caldwell, J., and Kollman, P. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society 117*, 19 (1995), 5179–5197. (cited on page 23)

[42] Cotta, C. Protein Structure Prediction Using Evolutionary Algorithms Hybridized with Backtracking. In *Artificial Neural Nets Problem Solving Methods*, vol. 2687 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2003, pp. 321–328. (cited on pages 32, 33, 48, 56, 61, 94, 102, 114, 115, 116, 117, and 139)

[43] Crescenzi, P., Goldman, D., Papadimitriou, C., Piccolboni, A., and Yannakakis, M. On the Complexity of Protein Folding. In *ACM Symposium on Theory of Computing*. ACM, Dallas, TX, USA, 1998, pp. 597–603. (cited on pages 3, 25, 28, and 31)

[44] Crippen, G. Prediction of Protein Folding From Amino Acid Sequence Over Discrete Conformation Spaces. *Biochemistry 30*, 17 (1991), 4232–4237. (cited on page 26)

[45] Custódio, F., Barbosa, H., and Dardenne, L. Investigation of the Three-dimensional Lattice HP Protein Folding Model Using a Genetic Algorithm. *Genetics and Molecular Biology 27*, 4 (2004), 611–615. (cited on pages 4, 33, 40, 42, 43, and 48)

[46] Custódio, F., Barbosa, H., and Dardenne, L. Full-atom Ab Initio Protein Structure Prediction with a Genetic Algorithm Using a Similarity-based Surrogate Model. In *IEEE Congress on Evolutionary Computation*. Barcelona, Spain, July 2010, pp. 1–8. (cited on page 24)

[47] Custódio, F., Barbosa, H., and Dardenne, L. A Multiple Minima Genetic Algorithm for Protein Structure Prediction. *Applied Soft Computing 15*, Feb. (2014), 88–99. (cited on page 33)

[48] Cutello, V., Morelli, G., Nicosia, G., and Pavone, M. Immune Algorithms with Aging Operators for the String Folding Problem and the Protein Folding Problem. In *Evolutionary Computation in Combinatorial Optimization*, vol. 3448 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Lausanne, Switzerland, 2005, pp. 80–90. (cited on page 33)

[49] Cutello, V., Morelli, G., Nicosia, G., Pavone, M., and Scollo, G. On Discrete Models and Immunological Algorithms for Protein Structure Prediction. *Natural Computing 10*, 1 (2011), 91–102. (cited on pages 33, 115, and 139)

[50] Cutello, V., Narzisi, G., and Nicosia, G. A Class of Pareto Archived Evolution Strategy Algorithms Using Immune Inspired Operators for Ab-Initio Protein Structure Prediction. In *Applications of Evolutionary Computing*, vol. 3449 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Lausanne, Switzerland, 2005, pp. 54–63. (cited on pages 12 and 71)

[51] Cutello, V., Narzisi, G., and Nicosia, G. A Multi-Objective Evolutionary Approach to the Protein Structure Prediction Problem. *Journal of The Royal Society Interface 3*, 6 (2006), 139–151. (cited on pages 12 and 71)

[52] Cutello, V., Narzisi, G., and Nicosia, G. Computational Studies of Peptide and Protein Structure Prediction Problems via Multiobjective Evolutionary Algorithms. In *Multiobjective Problem Solving from Nature*, Natural Computing Series. Springer Berlin Heidelberg, 2008, pp. 93–114. (cited on pages 12 and 71)

[53] Cutello, V., Nicosia, G., Pavone, M., and Timmis, J. An Immune Algorithm for Protein Structure Prediction on Lattice Models. *IEEE Transactions on Evolutionary Computation 11*, 1 (2007), 101–117. (cited on pages 25, 33, 34, 115, and 139)

[54] Datta, A., Talukdar, V., Konar, A., and Jain, L. Neuro-swarm Hybridization for Protein Tertiary Structure Prediction. *International Journal of Hybrid Intelligent Systems 5*, 3 (aug 2008), 153–159. (cited on page 24)

[55] Day, R., Zydallis, J., and Lamont, G. Solving the Protein Structure Prediction Problem Through a Multi-Objective Genetic Algorithm. In *IEEE/DARPA International Conference on Computational Nanoscience*. San Juan, PR, USA, 2002, pp. 32–35. (cited on pages 12 and 71)

[56] Dayem Ullah, A., Kapsokalivas, L., Mann, M., and Steinhöfel, K. Protein Folding Simulation by Two-Stage Optimization. In *Computational Intelligence and Intelligent Systems*, vol. 51 of *Communications in Computer and Information Science*. Springer Berlin Heidelberg, 2009, pp. 138–145. (cited on pages 26 and 27)

[57] de Almeida, C., Gonçalves, R., and Delgado, M. A Hybrid Immune-Based System for the Protein Folding Problem. In *Evolutionary Computation in Combinatorial Optimization*, vol. 4446 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Valencia, Spain, 2007, pp. 13–24. (cited on pages 33, 114, 116, and 139)

[58] Deb, K. An Efficient Constraint Handling Method for Genetic Algorithms. *Computer Methods in Applied Mechanics and Engineering 186*, 2-4 (2000), 311–338. (cited on page 131)

[59] Deb, K., Agrawal, S., Pratab, A., and Meyarivan, T. A Fast Elitist Non-dominated Sorting Genetic Algorithm for Multi-objective Optimization: NSGA-II. In *Parallel Problem Solving from Nature*, vol. 1917 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Paris, France, 2000, pp. 849–858. (cited on pages vii, 101, 102, 132, and 146)

[60] Deb, K., and Saha, A. Finding Multiple Solutions for Multimodal Optimization Problems Using a Multi-objective Evolutionary Approach. In *Genetic and Evolutionary Computation Conference*. ACM, Portland, OR, USA, 2010, pp. 447–454. (cited on page 71)

[61] Dill, K. Theory for the Folding and Stability of Globular Proteins. *Biochemistry 24*, 6 (1985), 1501–9. (cited on pages 2, 25, 27, and 48)

[62] Dill, K., and H., C. From Levinthal to Pathways to Funnels. *Nature Structural Biology 4*, 1 (January 1997), 10–19. (cited on pages 80, 90, and 127)

[63] Dorigo, M., and Di Caro, G. In *New Ideas in Optimization*. McGraw-Hill, London, UK, 1999, ch. The Ant Colony Optimization Meta-heuristic, pp. 11–32. (cited on page 33)

[64] Dorn, M., Buriol, L., and Lamb, L. A Hybrid Genetic Algorithm for the 3-D Protein Structure Prediction Problem Using a Path-Relinking Strategy. In *IEEE Congress on Evolutionary Computation*. New Orleans, LA, USA, June 2011, pp. 2709–2716. (cited on page 24)

[65] Dotú, I., Cebrián, M., Van Hentenryck, P., and Clote, P. On Lattice Protein Structure Prediction Revisited. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 8*, 6 (2011), 1620–1632. (cited on page 46)

[66] Duarte-Flores, S., and Smith, J. Study of Fitness Landscapes for the HP model of Protein Structure Srediction. In *IEEE Congress on Evolutionary Computation*, vol. 4. Canberra, Australia, 2003, pp. 2338–2345. (cited on pages 33, 114, 115, 116, 128, and 139)

[67] Garza-Fabre, M., Rodriguez-Tello, E., and Toscano-Pulido, G. Constraint-Handling Through Multi-Objective Optimization: the Hydrophobic-Polar Model for Protein Structure Prediction. *Computers & Operations Research (under review)*. (cited on page 159)

[68] Garza-Fabre, M., Rodriguez-Tello, E., and Toscano-Pulido, G. Comparing Alternative Energy Functions for the HP Model of Protein Structure Prediction. In *IEEE Congress on Evolutionary Computation*. New Orleans, LA, USA, June 2011, pp. 2307–2314. (cited on page 156)

[69] Garza-Fabre, M., Rodriguez-Tello, E., and Toscano-Pulido, G. An Improved Multiobjectivization Strategy for HP Model-Based Protein Structure Prediction. In *Parallel Problem Solving from Nature*, vol. 7492 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Taormina, Italy, September 2012, pp. 82–92. (cited on page 157)

[70] Garza-Fabre, M., Rodriguez-Tello, E., and Toscano-Pulido, G. Multiobjectivizing the HP Model for Protein Structure Prediction. In *Evolutionary Computation in Combinatorial Optimization*, vol. 7245 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Málaga, Spain, April 2012, pp. 182–193. (cited on page 157)

[71] Garza-Fabre, M., Rodriguez-Tello, E., and Toscano-Pulido, G. Comparative Analysis of Different Evaluation Functions for Protein Structure Prediction under the HP Model. *Journal of Computer Science and Technology 28*, 5 (2013), 868–889. (cited on page 156)

[72] Garza-Fabre, M., Toscano-Pulido, G., and Coello Coello, C. A. Alternative Fitness Assignment Methods for Many-Objective Optimization Problems. In *Artificial Evolution*, vol. 5975 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Strasbourg, France, October 2009, pp. 146–157. (cited on page 69)

[73] Garza-Fabre, M., Toscano-Pulido, G., and Coello Coello, C. A. Ranking Methods for Many-Objective Optimization. In *MICAI 2009: Advances in Artificial Intelligence*,

vol. 5845 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Guanajuato, México, November 2009, pp. 633–645. (cited on page 69)

[74] Garza-Fabre, M., Toscano-Pulido, G., and Coello Coello, C. A. Two Novel Approaches for Many-Objective Optimization. In *IEEE Congress on Evolutionary Computation*. Barcelona, Spain, July 2010, pp. 1–8. (cited on page 69)

[75] Garza-Fabre, M., Toscano-Pulido, G., Coello Coello, C. A., and Rodriguez-Tello, E. Effective Ranking + Speciation = Many-Objective Optimization. In *IEEE Congress on Evolutionary Computation*. New Orleans, LA, USA, June 2011, pp. 2115–2122. (cited on page 69)

[76] Garza-Fabre, M., Toscano-Pulido, G., and Rodriguez-Tello, E. Multiobjectivization, Fitness Landscape Transformation and Search Performance: A Case of Study on the HP model for Protein Structure Prediction. *European Journal of Operational Research (under review)*. (cited on page 158)

[77] Garza-Fabre, M., Toscano-Pulido, G., and Rodriguez-Tello, E. Comparative Study of Alternative Energy Functions for the HP Model of Protein Structure Prediction. In *International Conference on Bioinformatics & Computational Biology*, vol. 2. CSREA Press, Las Vegas, NV, USA, July 2011, pp. 618–624. (cited on page 156)

[78] Garza-Fabre, M., Toscano-Pulido, G., and Rodriguez-Tello, E. Locality-based Multiobjectivization for the HP Model of Protein Structure Prediction. In *Genetic and Evolutionary Computation Conference*. ACM, Philadelphia, PA, USA, July 2012, pp. 473–480. (cited on page 157)

[79] Garza-Fabre, M., and Toscano-Pulidoand E. Rodriguez-Tello, G. Handling Constraints in the HP Model for Protein Structure Prediction by Multiobjective Optimization. In *IEEE Congress on Evolutionary Computation*. Cancún, México, Jun. 2013, pp. 2728–2735. (cited on page 159)

[80] Glover, F. Future Paths for Integer Programming and Links to Artificial Intelligence. *Computers & Operations Research 13*, 5 (1986), 533–549. (cited on page 33)

[81] Goldberg, D. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Boston, MA, USA, 1989. (cited on page 31)

[82] Greiner, D., Emperador, J., Winter, G., and Galván, B. Improving Computational Mechanics Optimum Design Using Helper Objectives: An Application in Frame Bar Structures. In *Evolutionary Multi-Criterion Optimization*, vol. 4403 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Matshushima, Japan, 2007, pp. 575–589. (cited on pages 12 and 71)

[83] Guo, H., Lu, Q., Wu, J., Huang, X., and Qian, P. Solving 2D HP Protein Folding Problem by Parallel Ant Colonies. In *Biomedical Engineering and Informatics*. Tianjin, China, 2009, pp. 1–5. (cited on page 33)

[84] Handl, J., Lovell, S., and Knowles, J. Investigations into the Effect of Multiobjectivization in Protein Structure Prediction. In *Parallel Problem Solving from Nature*, vol. 5199 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Dortmund, Germany, 2008, pp. 702–711. (cited on pages 12 and 71)

[85] Handl, J., Lovell, S., and Knowles, J. Multiobjectivization by Decomposition of Scalar Cost Functions. In *Parallel Problem Solving from Nature*, vol. 5199 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Dortmund, Germany, 2008, pp. 31–40. (cited on pages 11, 12, 69, 70, 71, 98, 109, and 131)

[86] Hart, W., and Newman, A. *Protein Structure Prediction with Lattice Models*. Chapman and Hall/CRC Computer and Information Science Series, 2005. (cited on pages 2 and 25)

[87] Heath, A., Kavraki, L., and Clementi, C. From Coarse-Grain to All-Atom: Toward Multiscale Analysis of Protein Landscapes. *Proteins: Structure, Function, and Bioinformatics 68*, 3 (2007), 646–661. (cited on page 27)

[88] Hirst, J. The Evolutionary Landscape of Functional Model Proteins. *Protein Engineering 12*, 9 (1999), 721–726. (cited on page 25)

[89] Holland, J. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI, USA, 1975. (cited on page 31)

[90] Homouz, D. Simulating Protein Folding in Different Environmental Conditions. In *Protein Conformational Dynamics*, vol. 805 of *Advances in Experimental Medicine and Biology*. Springer International Publishing, 2014, pp. 171–197. (cited on pages 25 and 26)

[91] Hoque, M., Chetty, M., and Dooley, L. Non-Isomorphic Coding in Lattice Model and its Impact for Protein Folding Prediction Using Genetic Algorithm. In *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. Sun Valley, ID, USA, 2006, pp. 1–8. (cited on page 79)

[92] Hoque, M., Chetty, M., Lewis, A., and Sattar, A. Twin Removal in Genetic Algorithms for Protein Structure Prediction Using Low-Resolution Model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 8*, 1 (2011), 234–245. (cited on page 33)

[93] Hoque, M., Chetty, M., and Sattar, A. Extended HP Model for Protein Structure Prediction. *Journal of Computational Biology 16*, 1 (2009), 85–103. (cited on page 26)

[94] Hsieh, S., and Lai, D. A New Branch and Bound Method for the Protein Folding Problem Under the 2D-HP Model. *IEEE Transactions on NanoBioscience 10*, 2 (June 2011), 69–75. (cited on page 31)

[95] Hu, X., Zhang, J., and Li, Y. Flexible Protein Folding by Ant Colony Optimization. In *Computational Intelligence in Biomedicine and Bioinformatics*, vol. 151 of *Studies in Computational Intelligence*. Springer Berlin Heidelberg, 2008, pp. 317–336. (cited on page 33)

[96] Ishibuchi, H., Hitotsuyanagi, Y., Nakashima, Y., and Nojima, Y. Multiobjectivization from Two Objectives to Four Objectives in Evolutionary Multi-Objective Op-

timization Algorithms. In *World Congress on Nature and Biologically Inspired Computing*. Kitakyushu, Japan, 2010, pp. 502–507. (cited on page 11)

[97] Ishibuchi, H., Tsukamoto, N., and Nojima, Y. Evolutionary Many-Objective Optimization: A Short Review. In *IEEE Congress on Evolutionary Computation*. Hong Kong, June 2008, pp. 2424–2431. (cited on page 69)

[98] Islam, M., and Chetty, M. Novel Memetic Algorithm for Protein Structure Prediction. In *Advances in Artificial Intelligence*, vol. 5866 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2009, pp. 412–421. (cited on pages 4, 33, 40, 47, and 48)

[99] Islam, M., and Chetty, M. Clustered Memetic Algorithm for Protein Structure Prediction. In *IEEE Congress on Evolutionary Computation*. Barcelona, Spain, 2010, pp. 1–8. (cited on pages 33, 47, and 48)

[100] Islam, M., and Chetty, M. Clustered Memetic Algorithm With Local Heuristics for Ab Initio Protein Structure Prediction. *IEEE Transactions on Evolutionary Computation 17*, 4 (2013), 558–576. (cited on pages 33 and 34)

[101] Islam, M., Chetty, M., and Murshed, M. Novel Local Improvement Techniques in Clustered Memetic Algorithm for Protein Structure Prediction. In *IEEE Congress on Evolutionary Computation*. New Orleans, LA, USA, 2011, pp. 1003–1011. (cited on pages 33, 47, and 48)

[102] Jacques, J., Taillard, J., Delerue, D., Jourdan, L., and Dhaenens, C. The Benefits of Using Multi-Objectivization for Mining Pittsburgh Partial Classification Rules in Imbalanced and Discrete Data. In *Genetic and Evolutionary Computation Conference*. ACM, Amsterdam, The Netherlands, 2013, pp. 543–550. (cited on page 71)

[103] Jähne, M., Li, X., and Branke, J. Evolutionary Algorithms and Multi-Objectivization for the Travelling Salesman Problem. In *Genetic and Evolutionary Computation Conference*. ACM, Montreal, Canada, 2009, pp. 595–602. (cited on pages 12 and 71)

[104] Jana, N., and Sil, J. Protein Structure Prediction in 2D HP Lattice Model Using Differential Evolutionary Algorithm. In *Information Systems Design and Intelligent Applications*, vol. 132 of *Advances in Intelligent and Soft Computing*. Springer Berlin Heidelberg, Visakhapatnam, India, 2012, pp. 281–290. (cited on page 33)

[105] Jensen, M. Helper-Objectives: Using Multi-Objective Evolutionary Algorithms for Single-Objective Optimisation. *Journal of Mathematical Modelling and Algorithms 3*, 4 (2004), 323–347. (cited on pages 11, 12, 69, and 71)

[106] Johnson, C. M., and Katikireddy, A. A Genetic Algorithm with Backtracking for Protein Structure Prediction. In *Genetic and Evolutionary Computation Conference* (Seattle, WA, USA, 2006), ACM, pp. 299–300. (cited on page 115)

[107] Jones, T. *Evolutionary Algorithms, Fitness Landscapes and Search*. PhD thesis, University of New Mexico, Albuquerque, USA, May 1995. (cited on page 13)

[108] Jorgensen, W., and Tirado-Rives, J. The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *Journal of the American Chemical Society 110*, 6 (1988), 1657–1666. (cited on page 23)

[109] Kanj, F., Mansour, N., Khachfe, H., and Abu-Khzam, F. Protein Structure Prediction in the 3D HP Model. In *IEEE/ACS International Conference on Computer Systems and Applications*. 2009, pp. 732–736. (cited on page 33)

[110] Karaboga, D. An idea based on Honey Bee Swarm for Numerical Optimization. Tech. Rep. TR06, Erciyes University, October 2005. (cited on page 33)

[111] Karaboga, D., Gorkemli, B., Ozturk, C., and Karaboga, N. A comprehensive survey: Artificial bee colony (abc) algorithm and applications. *Artificial Intelligence Review* (2012), 1–37. (cited on page 33)

[112] Kelm, S., Choi, Y., and Deane, C. Protein Modeling and Structural Prediction. In *Springer Handbook of Bio-/Neuroinformatics*. Springer Berlin Heidelberg, 2014, pp. 171–182. (cited on pages 22 and 23)

[113] Khimasia, M., and Coveney, P. Protein Structure Prediction as a Hard Optimization Problem: The Genetic Algorithm Approach. *Molecular Simulation 19*, 4 (1997), 205–226. (cited on pages 32, 114, and 116)

[114] Kirkpatrick, S., Gelatt, C., and Vecchi, M. Optimization by Simulated Annealing. *Science 220*, 4598 (1983), 671–680. (cited on page 24)

[115] Klug, W., and Cummings, M. *Concepts of Genetics*. Pearson Education. Prentice Hall, 2003. (cited on pages ix and 16)

[116] Kmiecik, S., Gront, D., and Kolinski, A. Towards the high-resolution protein structure prediction. fast refinement of reduced models with all-atom force field. *BMC Structural Biology 7*, 1 (2007), 43. (cited on page 27)

[117] Knowles, J., and Corne, D. Properties of an Adaptive Archiving Algorithm for Storing Nondominated Vectors. *IEEE Transactions on Evolutionary Computation 7*, 2 (2003), 100–116. (cited on pages 94, 98, and 129)

[118] Knowles, J., and Corne, D. Quantifying the Effects of Objective Space Dimension in Evolutionary Multiobjective Optimization. In *Evolutionary Multi-Criterion Optimization*, vol. 4403 of *Lecture Notes in Computer Science*. Matshushima, Japan, March 2007, pp. 757–771. (cited on page 69)

[119] Knowles, J., Watson, R., and Corne, D. Reducing Local Optima in Single-Objective Problems by Multi-objectivization. In *Evolutionary Multi-Criterion Optimization*. Springer-Verlag, Zurich, Switzerland, 2001, pp. 269–283. (cited on pages 11, 12, 69, 70, 71, and 114)

[120] Kolinski, A., and Skolnick, J. Reduced Models of Proteins and their Applications. *Polymer 45*, 2 (2004), 511–524. (cited on pages 2, 25, and 26)

[121] Krasnogor, N., Blackburne, B., Burke, E., and Hirst, J. Multimeme algorithms for protein structure prediction. In *Parallel Problem Solving from Nature*, vol. 2439 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Granada, Spain, 2002, pp. 769–778. (cited on pages 25, 33, and 34)

[122] Krasnogor, N., Hart, W., Smith, J., and Pelta, D. Protein Structure Prediction With Evolutionary Algorithms. In *Genetic and Evolutionary Computation Conference*. Morgan Kaufman, Orlando, FL, USA, 1999. (cited on pages 4, 32, 33, 40, 41, 42, 48, 114, 116, 118, 128, and 140)

[123] Krasnogor, N., Pelta, D., Lopez, P., and de la Canal, E. Genetic Algorithms for the Protein Folding Problem: A Critical View. In *Engineering of Intelligent Systems*. ICSC Academic Press, 1998, pp. 353–360. (cited on page 32)

[124] Krasnogor, N., and Smith, J. Emergence of Profitable Search Strategies Based on a Simple Inheritance Mechanism. In *Genetic and Evolutionary Computation Conference*. Morgan Kaufmann Publishers, San Francisco, CA, USA, 2001, pp. 432–439. (cited on page 33)

[125] Larsen, A., Wagner, J., Jain, A., and Vaidehi, N. Protein Structure Refinement of CASP Target Proteins Using GNEIMO Torsional Dynamics Method. *Journal of Chemical Information and Modeling 54*, 2 (2014), 508–517. (cited on pages 22 and 23)

[126] Lau, K., and Dill, K. A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins. *Macromolecules 22*, 10 (1989), 3986–3997. (cited on pages 2, 25, 27, and 48)

[127] Lee, J., Wu, S., and Zhang, Y. Ab Initio Protein Structure Prediction. In *From Protein Structure to Function with Bioinformatics*. Springer Netherlands, 2009, pp. 3–25. (cited on pages 23 and 24)

[128] Lesh, N., Mitzenmacher, M., and Whitesides, S. A Complete and Effective Move Set for Simplified Protein Folding. In *International Conference on Research in Computational Molecular Biology*. ACM, Berlin, Germany, 2003, pp. 188–195. (cited on pages 115 and 116)

[129] Lesk, A. *Introduction to Protein Science: Architecture, Function, and Genomics*. Oxford University Press, 2010. (cited on pages ix, 16, 21, and 22)

[130] Levinthal, C. Are There Pathways for Protein Folding? *Journal de Chimie Physique*, 65 (1968), 44–45. (cited on page 21)

[131] Levinthal, C. How to Fold Graciously. In *Mossbauer Spectroscopy in Biological Systems: Proceedings of a Meeting Held at Allerton House, Monticello, Illinois*. University of Illinois Press, 1969, pp. 22–24. (cited on page 21)

[132] Li, H., Tang, C., and Wingreen, N. Designability of Protein Structures: A Lattice-Model Study Using the Miyazawa-Jernigan Matrix. *Proteins: Structure, Function, and Bioinformatics 49*, 3 (2002), 403–412. (cited on page 26)

[133] Liu, J., Song, B., Liu, Z., Huang, W., Sun, Y., and Liu, W. Energy-Landscape Paving for Prediction of Face-Centered-Cubic Hydrophobic-Hydrophilic Lattice Model Proteins. *Physical Review E 88*, 5 (Nov 2013), 052704. (cited on page 33)

[134] Liwo, A., Kazmierkiewicz, R., Czaplewski, C., Groth, M., Ołdziej, S., Wawak, R., Rackovsky, S., Pincus, M., and Scheraga, H. United-Residue Force Field for Off-Lattice Protein-Structure Simulations: III. Origin of Backbone Hydrogen-Bonding Cooperativity in United-Residue Potentials. *Journal of Computational Chemistry 19*, 3 (1998), 259–276. (cited on page 26)

[135] Liwo, A., Ołdziej, S., Pincus, M., Wawak, R., Rackovsky, S., and Scheraga, H. A United-Residue Force Field for Off-Lattice Protein-Structure Simulations. I. Functional Forms and Parameters of Long-Range Side-Chain Interaction Potentials from Protein Crystal Data. *Journal of Computational Chemistry 18*, 7 (1997), 849–873. (cited on page 26)

[136] Liwo, A., Pincus, M., Wawak, R., Rackovsky, S., Ołdziej, S., and Scheraga, H. A United-Residue Force Field for Off-Lattice Protein-Structure Simulations. II. Parameterization of Short-Range Interactions and Determination of Weights of Energy Terms by Z-Score Optimization. *Journal of Computational Chemistry 18*, 7 (1997), 874–887. (cited on page 26)

[137] Lochtefeld, D., and Ciarallo, F. Helper-Objective Optimization Strategies for the Job-Shop Scheduling Problem. *Applied Soft Computing 11*, 6 (2011), 4161–4174. (cited on pages 12 and 71)

[138] Lochtefeld, D., and Ciarallo, F. Multiobjectivization via Helper-Objectives With the Tunable Objectives Problem. *IEEE Transactions on Evolutionary Computation 16*, 3 (2012), 373–390. (cited on page 12)

[139] Lochtefeld, D., and Ciarallo, F. An Analysis of Decomposition Approaches in Multiobjectivization via Segmentation. *Applied Soft Computing* (2014). In Press. (cited on page 71)

[140] Lopes, H. Evolutionary Algorithms for the Protein Folding Problem: A Review and Current Trends. In *Computational Intelligence in Biomedicine and Bioinformatics*, vol. 151 of *Studies in Computational Intelligence*. Springer Berlin / Heidelberg, 2008, pp. 297–315. (cited on page 31)

[141] Lopes, H., and Bitello, R. A Differential Evolution Approach for Protein Folding Using a Lattice Model. *Journal of Computer Science and Technology 22*, 6 (2007), 904–908. (cited on page 33)

[142] Lopes, H., and Scapin, M. An Enhanced Genetic Algorithm for Protein Structure Prediction Using the 2D Hydrophobic-Polar Model. In *Artificial Evolution*, vol. 3871 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Lille, France, 2006, pp. 238–246. (cited on pages 4, 33, 40, 43, 44, 45, 48, 114, and 116)

[143] Lopes, H., and Scapin, M. A Hybrid Genetic Algorithm for the Protein Folding Problem Using the 2D-HP Lattice Model. In *Success in Evolutionary Computation*, vol. 92

of *Studies in Computational Intelligence*. Springer Berlin / Heidelberg, 2008, pp. 121–140. (cited on pages 33, 43, 44, 45, and 48)

[144] López-Ibáñez, M., Knowles, J., and Laumanns, M. On Sequential Online Archiving of Objective Vectors. In *Evolutionary Multi-Criterion Optimization*, vol. 6576 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Ouro Preto, Brazil, 2011, pp. 46–60. (cited on pages 94, 98, and 129)

[145] Louis, S., and Rawlins, G. Pareto Optimality, GA-easiness and Deception. In *International Conference on Genetic Algorithms*. Morgan Kaufmann, 1993, pp. 118–123. (cited on page 11)

[146] Lourenço, H., Martin, O., and Stützle, T. Iterated Local Search. In *Handbook of Metaheuristics*, vol. 57 of *International Series in Operations Research & Management Science*. Kluwer Academic Publishers, 2002, pp. 321–353. (cited on pages 61 and 78)

[147] Lourenço, H., Martin, O., and Stützle, T. Iterated Local Search: Framework and Applications. In *Handbook of Metaheuristics*, vol. 146 of *International Series in Operations Research & Management Science*. Springer US, 2010, pp. 363–397. (cited on pages 61 and 78)

[148] M., M., M., B., and D., D. On Potential Energy Models for EA-based Ab Initio Protein Structure Prediction. *Evolutionary Computation 18*, 2 (2010), 255–275. (cited on page 24)

[149] Maher, B., Albrecht, A., Loomes, M., Yang, X., and Steinhöfel, K. A Firefly-Inspired Method for Protein Structure Prediction in Lattice Models. *Biomolecules 4*, 1 (2014), 56–75. (cited on page 33)

[150] Malan, K., and Engelbrecht, A. A Survey of Techniques for Characterising Fitness Landscapes and Some Possible Ways Forward. *Information Sciences 241* (2013), 148–163. (cited on page 13)

[151] Mani, A., and Patvardhan, C. A Novel Hybrid Constraint Handling Technique for Evolutionary Optimization. In *IEEE Congress on Evolutionary Computation*. Trondheim, Norway, 2009, pp. 2577–2583. (cited on page 131)

[152] Mansour, N., Kanj, F., and Khachfe, H. Particle Swarm Optimization Approach for Protein Structure Prediction in the 3D HP Model. *Interdisciplinary Sciences: Computational Life Sciences 4*, 3 (Sep. 2012), 190–200. (cited on page 33)

[153] Mansour, N., Kehyayan, C., and Khachfe, H. Scatter Search Algorithm for Protein Structure Prediction. *International Journal of Bioinformatics Research and Applications 5*, 5 (September 2009), 501–515. (cited on page 24)

[154] Marmion, M., Dhaenens, C., Jourdan, L., Liefooghe, A., and Vérel, S. NILS: A Neutrality-Based Iterated Local Search and Its Application to Flowshop Scheduling. In *Evolutionary Computation in Combinatorial Optimization*, vol. 6622 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Torino, Italy, 2011, pp. 191–202. (cited on pages 67 and 156)

[155] Marmion, M., Dhaenens, C., Jourdan, L., Liefooghe, A., and Vérel, S. On the Neutrality of Flowshop Scheduling Fitness Landscapes. In *Learning and Intelligent Optimization*, vol. 6683 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2011, pp. 238–252. (cited on pages 67, 84, and 156)

[156] Marmion, M., Dhaenens, C., Jourdan, L., Liefooghe, A., and Vérel, S. The Road to VEGAS: Guiding the Search Over Neutral Networks. In *Genetic and Evolutionary Computation Conference*. ACM, Dublin, Ireland, 2011, pp. 1979–1986. (cited on pages 67 and 156)

[157] Martin, O., Otto, S., and Felten, E. Large-Step Markov Chains for the Traveling Salesman Problem. *Complex Systems 5*, 3 (1991), 299–326. (cited on pages 61 and 78)

[158] Menchaca-Mendez, A., and Coello Coello, C. A. A New Proposal to Hybridize the Nelder-Mead Method to a Differential Evolution Algorithm for Constrained Optimization. In *IEEE Congress on Evolutionary Computation*. Trondheim, Norway, 2009, pp. 2598–2605. (cited on page 132)

[159] Merz, P. Memetic Algorithms and Fitness Landscapes in Combinatorial Optimization. In *Handbook of Memetic Algorithms*, vol. 379 of *Studies in Computational Intelligence*. Springer Berlin Heidelberg, 2012, pp. 95–119. (cited on page 13)

[160] Mezura-Montes, E., Coello Coello, C., and Tun-Morales, E. Simple Feasibility Rules and Differential Evolution for Constrained Optimization. In *MICAI 2004: Advances in Artificial Intelligence*, vol. 2972 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, México City, México, April 2004, pp. 707–716. (cited on page 131)

[161] Mezura-Montes, E., and Coello Coello, C. A. Constraint-handling in Nature-inspired Numerical Optimization: Past, Present and Future. *Swarm and Evolutionary Computation 1*, 4 (2011), 173–194. (cited on pages 5, 12, 114, 117, 131, and 133)

[162] Momany, F., McGuire, R., Burgess, A., and Scheraga, H. Energy Parameters in Polypeptides. VII. Geometric Parameters, Partial Atomic Charges, Nonbonded Interactions, Hydrogen Bond Interactions, and Intrinsic Torsional Potentials for the Naturally Occurring Amino Acids. *The Journal of Physical Chemistry 79*, 22 (1975), 2361–2381. (cited on pages 23 and 26)

[163] Moscato, P. On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts: Towards Memetic Algorithms. Tech. Rep. C3P Report 826, Caltech Concurrent Computation Program, Pasadena, CA, USA, 1989. (cited on page 33)

[164] Moscato, P., and Cotta, C. A Gentle Introduction to Memetic Algorithms. In *Handbook of Metaheuristics*, vol. 57 of *International Series in Operations Research &amp; Management Science*. Springer New York, 2003, pp. 105–144. (cited on page 33)

[165] Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. Critical Assessment of Methods of Protein Structure Prediction (CASP) — Round X. *Proteins: Structure, Function, and Bioinformatics 82* (2014), 1–6. (cited on page 24)

[166] Mouret, J. Novelty-Based Multiobjectivization. In *New Horizons in Evolutionary Robotics*, vol. 341 of *Studies in Computational Intelligence*. Springer Berlin / Heidelberg, 2011, pp. 139–154. (cited on pages 12 and 71)

[167] Muñoz Zavala, A., Hernández Aguirre, A., and Villa Diharce, E. Constrained Optimization via Particle Evolutionary Swarm Optimization Algorithm (PESO). In *Genetic and Evolutionary Computation Conference* (Washington DC, USA, 2005), pp. 209–216. (cited on page 131)

[168] Nardelli, M., Tedesco, L., and Bechini, A. Cross-lattice Behavior of General ACO Folding for Proteins in the HP Model. In *ACM Symposium on Applied Computing*, SAC '13. ACM, Coimbra, Portugal, 2013, pp. 1320–1327. (cited on page 33)

[169] Nazmul, R., and Chetty, M. A Knowledge-Based Initial Population Generation in Memetic Algorithm for Protein Structure Prediction. In *Neural Information Processing*, vol. 8227 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013, pp. 546–553. (cited on page 33)

[170] Nemethy, G., Gibson, K., Palmer, K., Yoon, C., Paterlini, G., Zagari, A., Rumsey, S., and Scheraga, H. Energy Parameters in Polypeptides. 10. Improved Geometrical Parameters and Nonbonded Interactions for use in the ECEPP/3 Algorithm, with Application to Proline-Containing Peptides. *The Journal of Physical Chemistry 96*, 15 (1992), 6472–6484. (cited on pages 23 and 26)

[171] Neumann, F., and Wegener, I. Can Single-Objective Optimization Profit from Multiobjective Optimization? In *Multiobjective Problem Solving from Nature*, Natural Computing Series. Springer Berlin Heidelberg, 2008, pp. 115–130. (cited on page 71)

[172] Ngan, S., Hung, L., Liu, T., and Samudrala, R. Scoring Functions for De Novo Protein Structure Prediction Revisited. In *Protein Structure Prediction*, vol. 413 of *Methods in Molecular Biology*. Humana Press, 2008, pp. 243–281. (cited on page 24)

[173] Ołdziej, S., Czaplewski, C., Liwo, A., Chinchio, M., Nanias, M., Vila, J., Khalili, M., Arnautova, Y., Jagielska, A., Makowski, M., Schafroth, H., Kaźmierkiewicz, R., Ripoll, D., Pillardy, J., Saunders, J., Kang, Y., Gibson, K., and Scheraga, H. Physics-based Protein-structure Prediction Using a Hierarchical Protocol based on the UNRES Force Field: Assessment in Two Blind Tests. *Proceedings of the National Academy of Sciences of the United States of America 102*, 21 (2005), 7547–7552. (cited on page 26)

[174] Olson, B., and Shehu, A. Multi-Objective Stochastic Search for Sampling Local Minima in the Protein Energy Surface. In *International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*. ACM, Washington DC, USA, September 2013, pp. 430–439. (cited on pages 12 and 71)

[175] Pardalos, P., Liu, X., and Xue, G. Protein Conformation of a Lattice Model Using Tabu Search. *Journal of Global Optimization 11*, 1 (1997), 55–68. (cited on page 33)

[176] Pareto, V. *Cours d'Economie Politique*. Droz, Genève, 1896. (cited on page 11)

[177] Patton, A., Punch III, W., and Goodman, E. A Standard GA Approach to Native Protein Conformation Prediction. In *International Conference on Genetic Algorithms*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995, pp. 574–581. (cited on pages 30, 32, 114, and 116)

[178] Pelta, D., and Krasnogor, N. Multimeme Algorithms Using Fuzzy Logic Based Memes For Protein Structure Prediction. In *Recent Advances in Memetic Algorithms*, vol. 166 of *Studies in Fuzziness and Soft Computing*. Springer Berlin Heidelberg, 2005, pp. 49–64. (cited on page 33)

[179] Pevsner, J. *Bioinformatics and Functional Genomics*. Wiley, 2009. (cited on page 21)

[180] Pierri, C., De Grassi, A., and Turi, A. Lattices for Ab Initio Protein Structure Prediction. *Proteins: Structure, Function, and Bioinformatics 73*, 2 (2008), 351–361. (cited on pages 2 and 25)

[181] Pilat, M., and Neruda, R. Multiobjectivization for Classifier Parameter Tuning. In *Genetic and Evolutionary Computation Conference*. ACM, Amsterdam, The Netherlands, 2013, pp. 97–98. (cited on page 71)

[182] Pitzer, E., and Affenzeller, M. A Comprehensive Survey on Fitness Landscape Analysis. In *Recent Advances in Intelligent Engineering Systems*, vol. 378 of *Studies in Computational Intelligence*. Springer Berlin Heidelberg, 2012, pp. 161–191. (cited on pages 13, 14, and 67)

[183] Rashid, M., Newton, M., Hoque, M., and Sattar, A. A Local Search Embedded Genetic Algorithm for Simplified Protein Structure Prediction. In *IEEE Congress on Evolutionary Computation* (June 2013), pp. 1091–1098. (cited on page 33)

[184] Rashid, M., Newton, M., Hoque, M., Shatabda, S., Pham, D., and Sattar, A. Spiral Search: A Hydrophobic-Core Directed Local Search for Simplified PSP on 3D FCC Lattice. *BMC Bioinformatics 14*, Suppl 2 (2013), S16. (cited on page 33)

[185] Rohl, C., Strauss, C., Misura, K., and Baker, D. Protein structure prediction using rosetta. In *Numerical Computer Methods, Part D*, vol. 383 of *Methods in Enzymology*. Academic Press, 2004, pp. 66–93. (cited on page 24)

[186] Runarsson, T., and Yao, X. Stochastic Ranking for Constrained Evolutionary Optimization. *IEEE Transactions on Evolutionary Computation 4*, 3 (2000), 284–294. (cited on pages 5 and 114)

[187] Runarsson, T., and Yao, X. Constrained Evolutionary Optimization: The Penalty Function Approach. In *Evolutionary Optimization*, vol. 48 of *International Series in Operations Research & Management Science*. Springer US, 2003, pp. 87–113. (cited on page 140)

[188] Runarsson, T., and Yao, X. Search Biases in Constrained Evolutionary Optimization. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 35*, 2 (2005), 233–243. (cited on pages 115, 129, and 152)

[189] Salomon-Ferrer, R., Case, D., and Walker, R. An Overview of the Amber Biomolecular Simulation Package. *Wiley Interdisciplinary Reviews: Computational Molecular Science 3*, 2 (2013), 198–210. (cited on page 23)

[190] Samudrala, R., Xia, Y., Huang, E., and Levitt, M. Ab Initio Protein Structure Prediction Using a Combined Hierarchical Approach. *Proteins: Structure, Function, and Bioinformatics Suppl 3* (1999), 194–198. (cited on page 27)

[191] Santana, R., Larranaga, P., and Lozano, J. Protein Folding in Simplified Models With Estimation of Distribution Algorithms. *IEEE Transactions on Evolutionary Computation 12*, 4 (2008), 418–438. (cited on page 33)

[192] Santos, J., and Diéguez, M. Differential Evolution for Protein Structure Prediction Using the HP Model. In *Foundations on Natural and Artificial Computation*, vol. 6686 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2011, pp. 323–333. (cited on pages 33, 114, 115, and 117)

[193] Santos, J., Villot, P., and Dieguez, M. Cellular Automata for Modeling Protein Folding Using the HP Model. In *IEEE Congress on Evolutionary Computation*. Cancún, México, June 2013, pp. 1586–1593. (cited on page 34)

[194] Santos, J., Villot, P., and Diéguez, M. Protein folding with cellular automata in the 3d hp model. In *Genetic and Evolutionary Computation Conference*. ACM, Amsterdam, The Netherlands, 2013, pp. 1595–1602. (cited on page 34)

[195] Saxena, D., and Deb, K. Trading on Infeasibility by Exploiting Constraints Criticality Through Multi-objectivization: A System Design Perspective. In *IEEE Congress on Evolutionary Computation*. Singapore, 2007, pp. 919–926. (cited on pages 12 and 71)

[196] Scheraga, H. Simulations of the Folding of Proteins: A Historical Perspective. In *Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes*, Springer Series in Bio-/Neuroinformatics. Springer Berlin Heidelberg, 2014, pp. 1–23. (cited on page 23)

[197] Scheraga, H., Khalili, M., and Liwo, A. Protein-Folding Dynamics: Overview of Molecular Simulation Techniques. *Annual Review of Physical Chemistry 58*, 1 (2007), 57–83. (cited on page 24)

[198] Schmid, N., Eichenberger, A., Choutko, A., Riniker, S., Winger, M., Mark, A., and Gunsteren, W. Definition and Testing of the GROMOS Force-Field Versions 54A7 and 54B7. *European Biophysics Journal 40*, 7 (2011), 843–856. (cited on page 23)

[199] Scott, W., Hünenberger, P., Tironi, I., Mark, A., Billeter, S., Fennen, J., Torda, A., Huber, T., Krüger, P., and van Gunsteren, W. The GROMOS Biomolecular Simulation Program Package. *The Journal of Physical Chemistry A 103*, 19 (1999), 3596–3607. (cited on page 23)

[200] Segredo, E., Segura, C., and Leon, C. A Multiobjectivised Memetic Algorithm for the Frequency Assignment Problem. In *IEEE Congress on Evolutionary Computation*. New Orleans, LA, USA, 2011, pp. 1132–1139. (cited on pages 12 and 71)

[201] Segredo, E., Segura, C., and León, C. Memetic Algorithms and Hyperheuristics Applied to a Multiobjectivised Two-dimensional Packing Problem. *Journal of Global Optimization* (2013), 1–26. (cited on page 71)

[202] Segura, C., Coello Coello, C., Miranda, G., and León, C. Using Multi-Objective Evolutionary Algorithms for Single-Objective Optimization. *4OR 11*, 3 (2013), 201–228. (cited on pages 12, 69, 71, and 117)

[203] Segura, C., Coello Coello, C., Segredo, E., Miranda, G., and Leon, C. Improving the Diversity Preservation of Multi-Objective Approaches Used for Single-Objective Op-

timization. In *IEEE Congress on Evolutionary Computation*. Cancún, México, 2013, pp. 3198–3205. (cited on page 12)

[204] Segura, C., Segredo, E., González, Y., and León, C. Multiobjectivisation of the Antenna Positioning Problem. In *International Symposium on Distributed Computing and Artificial Intelligence*, vol. 91 of *Advances in Intelligent and Soft Computing*. Springer Berlin / Heidelberg, Salamanca, Spain, 2011, pp. 319–327. (cited on pages 12 and 71)

[205] Segura, C., Segredo, E., and León, C. Parallel Island-Based Multiobjectivised Memetic Algorithms for a 2D Packing Problem. In *Genetic and Evolutionary Computation Conference*. ACM, Dublin, Ireland, 2011, pp. 1611–1618. (cited on pages 12 and 71)

[206] Segura, C., Segredo, E., and León, C. Analysing the Robustness of Multiobjectivisation Approaches Applied to Large Scale Optimisation Problems. In *EVOLVE - A Bridge between Probability, Set Oriented Numerics and Evolutionary Computation*, vol. 447 of *Studies in Computational Intelligence*. Springer Berlin Heidelberg, 2013, pp. 365–391. (cited on page 71)

[207] Segura, C., Segredo, E., and León, C. Scalability and Robustness of Parallel Hyperheuristics Applied to a Multiobjectivised Frequency Assignment Problem. *Soft Computing 17*, 6 (2013), 1077–1093. (cited on page 71)

[208] Sharma, D., Deb, K., and Kishore, N. Customized Evolutionary Optimization Procedure for Generating Minimum Weight Compliant Mechanisms. *Engineering Optimization 46*, 1 (2014), 39–60. (cited on pages 12 and 71)

[209] Shatabda, S., Hakim Newton, M., Rashid, M., Pham, D., and Sattar, A. The Road Not Taken: Retreat and Diverge in Local Search for Simplified Protein Structure Prediction. *BMC Bioinformatics 14*, Suppl 2 (2013), S19. (cited on pages ix and 16)

[210] Shatabda, S., Newton, M., Rashid, M., and Sattar, A. An Efficient Encoding for Simplified Protein Structure Prediction Using Genetic Algorithms. In *IEEE Congress on Evolutionary Computation*. Cancún, México, June 2013, pp. 1217–1224. (cited on page 79)

[211] Shmygelska, A., and Hoos, H. An Ant Colony Optimization Algorithm for the 2D and 3D Hydrophobic Polar Protein Folding Problem. *BMC Bioinformatics 6*, 1 (2005), 30. (cited on page 33)

[212] Snustad, D., and Simmons, M. *Principles of Genetics*. Wiley, 2003. (cited on pages ix and 16)

[213] Soares Brasil, C., Botazzo Delbem, A., and Ferraz Bonetti, D. Investigating Relevant Aspects of MOEAs for Protein Structures Prediction. In *Genetic and Evolutionary Computation Conference*. ACM, Dublin, Ireland, 2011, pp. 705–712. (cited on pages 12 and 71)

[214] Stadler, P. Fitness Landscapes. In *Biological Evolution and Statistical Physics*, vol. 585 of *Lecture Notes in Physics*. Springer Berlin / Heidelberg, 2002, pp. 183–204. (cited on page 13)

[215] Stoker, H. *Organic and Biological Chemistry*. Cengage Learning, 2009. (cited on page 19)

[216] Stout, M., Bacardit, J., Hirst, J., Krasnogor, N., and Blazewicz, J. From HP Lattice Models to Real Proteins: Coordination Number Prediction Using Learning Classifier Systems. In *Applications of Evolutionary Computing*, vol. 3907 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2006, pp. 208–220. (cited on pages ix and 16)

[217] Sudha, S., Baskar, S., and Krishnaswamy, S. Multi-Objective Approach for Protein Structure Prediction. In *Swarm, Evolutionary, and Memetic Computing*, vol. 8298 of *Lecture Notes in Computer Science*. Springer International Publishing, 2013, pp. 511–522. (cited on page 71)

[218] Talbi, E. *Metaheuristics: From Design to Implementation*. Wiley Publishing, 2009. (cited on pages 3, 61, and 116)

[219] Thachuk, C., Shmygelska, A., and Hoos, H. A Replica Exchange Monte Carlo Algorithm for Protein Folding in the HP Model. *BMC Bioinformatics 8*, 1 (2007), 342. (cited on pages 31, 34, and 115)

[220] Thomas, S., and Jin, Y. Single and Multi-objective in Silico Evolution of Tunable Genetic Oscillators. In *Evolutionary Multi-Criterion Optimization*, vol. 7811 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013, pp. 696–709. (cited on page 71)

[221] Tran, T., Brockhoff, D., and Derbel, B. Multiobjectivization with NSGA-II on the Noiseless BBOB Testbed. In *Genetic and Evolutionary Computation Conference*. ACM, Amsterdam, The Netherlands, 2013, pp. 1217–1224. (cited on page 12)

[222] Trivedi, A., Sharma, D., and Srinivasan, D. Multi-objectivization of Short-term Unit Commitment Under Uncertainty Using Evolutionary Algorithm. In *IEEE Congress on Evolutionary Computation*. Brisbane, Australia, 2012, pp. 1–8. (cited on pages 12 and 71)

[223] Unger, R. The Genetic Algorithm Approach to Protein Structure Prediction. In *Applications of Evolutionary Computation in Chemistry*, vol. 110 of *Structure & Bonding*. Springer Berlin / Heidelberg, 2004, pp. 2697–2699. (cited on page 31)

[224] Unger, R., and Moult, J. Genetic algorithm for 3d protein folding simulations. In *International Conference on Genetic Algorithms*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993, pp. 581–588. (cited on pages 30, 31, 32, and 115)

[225] Unger, R., and Moult, J. Genetic Algorithms for Protein Folding Simulations. *Journal of Molecular Biology 231*, 1 (1993), 75–81. (cited on pages 31 and 32)

[226] Unger, R., and Moult, J. On the Applicability of Genetic Algorithms to Protein Folding. In *Hawaii International Conference on System Sciences*. Hawaii, USA, Jan 1993, pp. 715–725. (cited on pages 31 and 32)

[227] Vanneschi, L., Pirola, Y., Mauri, G., Tomassini, M., Collard, P., and Verel, S. A Study of the Neutrality of Boolean Function Landscapes in Genetic Programming. *Theoretical Computer Science 425*, 0 (2012), 34–57. (cited on pages 13 and 85)

[228] Vanneschi, L., Tomassini, M., Collard, P., Vérel, S., Pirola, Y., and Mauri, G. A Comprehensive View of Fitness Landscapes with Neutrality and Fitness Clouds. In *Genetic*

Programming, vol. 4445 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Valencia, Spain, 2007, pp. 241–250. (cited on pages 67, 85, and 156)

[229] Vargas Benitez, C., Stubs Parpinelli, R., and Lopes, H. Parallelism, Hybridism and Coevolution in a Multi-Level ABC-GA Approach for the Protein Structure Prediction Problem. *Concurrency and Computation: Practice and Experience 24*, 6 (2012), 635–646. (cited on page 33)

[230] Vassilev, V., Fogarty, T., and Miller, J. Smoothness, Ruggedness and Neutrality of Fitness Landscapes: from Theory to Application. In *Advances in Evolutionary Computing*, Natural Computing Series. Springer Berlin Heidelberg, 2003, pp. 3–44. (cited on page 160)

[231] Vérel, S., Collard, P., and Clergue, M. Scuba Search: When Selection Meets Innovation. In *IEEE Congress on Evolutionary Computation*. Portland, OR, USA, 2004, pp. 924–931. (cited on pages 67 and 156)

[232] Verel, S., Collard, P., Tomassini, M., and Vanneschi, L. Fitness Landscape of the Cellular Automata Majority Problem: View from the Olympus. *Theoretical Computer Science 378*, 1 (2007), 54–77. (cited on page 13)

[233] Vite-Silva, I., Cruz-Cortés, N., Toscano-Pulido, G., and de la Fraga, L. Optimal Triangulation in 3D Computer Vision Using a Multi-objective Evolutionary Algorithm. In *Applications of Evolutionary Computing*, vol. 4448 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Valencia, Spain, 2007, pp. 330–339. (cited on pages 12 and 71)

[234] Watanabe, S., and Sakakibara, K. Multi-objective Approaches in a Single-objective Optimization Environment. In *IEEE Congress on Evolutionary Computation*, vol. 2. Edinburgh, UK, 2005, pp. 1714–1721. (cited on page 12)

[235] Watanabe, S., and Sakakibara, K. A Multiobjectivization Approach for Vehicle Routing Problems. In *Evolutionary Multi-Criterion Optimization*, vol. 4403 of *Lecture Notes in*

Computer Science. Springer Berlin / Heidelberg, Matshushima, Japan, 2007, pp. 660–672. (cited on page 71)

[236] WATSON, J. An Introduction to Fitness Landscape Analysis and Cost Models for Local Search. In *Handbook of Metaheuristics*, vol. 146 of *International Series in Operations Research & Management Science*. Springer US, 2010, pp. 599–623. (cited on pages 67 and 70)

[237] WESSING, S., PREUSS, M., AND RUDOLPH, G. Niching by Multiobjectivization with Neighbor Information: Trade-offs and Benefits. In *IEEE Congress on Evolutionary Computation* (Cancún, México, 2013), pp. 103–110. (cited on page 12)

[238] WEST, M., AND HECHT, M. Binary Patterning of Polar and Nonpolar Amino Acids in the Sequences and Structures of Native Proteins. *Protein Science 4*, 10 (1995), 2032–2039. (cited on pages ix and 16)

[239] WRIGHT, S. The Roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution. *Proceedings of the 6th International Congress of Genetics 1* (1932), 356–366. (cited on page 13)

[240] WÜST, T., LI, Y., AND LANDAU, D. Unraveling the Beautiful Complexity of Simple Lattice Model Polymers and Proteins Using Wang-Landau Sampling. *Journal of Statistical Physics 144* (2011), 638–651. (cited on page 34)

[241] XIA, Y., HUANG, E., LEVITT, M., AND SAMUDRALA, R. Ab initio construction of protein tertiary structures using a hierarchical approach. *Journal of Molecular Biology 300*, 1 (2000), 171–185. (cited on page 27)

[242] YU, T., AND MILLER, J. Finding Needles in Haystacks Is Not Hard with Neutrality. In *Genetic Programming*, vol. 2278 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, Kinsale, Ireland, 2002, pp. 46–54. (cited on pages 67 and 156)

[243] YU, Z., XIAO, C., AND ZHOU, G. Multi-objectivization-based Localization of Underwater Sensors Using Magnetometers. *IEEE Sensors Journal* (2013). In Press. (cited on page 71)

[244] Yuan, X., Su, A., Yuan, Y., Nie, H., and Wang, L. An Improved PSO for Dynamic Load Dispatch of Generators with Valve-Point Effects. *Energy 34*, 1 (2009), 67–74. (cited on page 131)

[245] Zhang, J., Kou, S. C., and Liu, J. S. Biopolymer Structure Simulation and Optimization via Fragment Regrowth Monte Carlo. *The Journal of Chemical Physics 126*, 22 (2007), 225101. (cited on page 34)

[246] Zhang, X., Wang, T., Luo, H., Yang, J., Deng, Y., Tang, J., and Yang, M. 3D Protein Structure Prediction with Genetic Tabu Search Algorithm. *BMC Systems Biology 4*, Suppl 1 (2010), S6. (cited on page 33)

[247] Zhang, Y., Wu, L., and Wang, S. Solving Two-Dimensional HP Model by Firefly Algorithm and Simplified Energy Function. *Mathematical Problems in Engineering 2013*, Article ID 398141 (2013), 052704. (cited on page 33)

[248] Zhao, X. Advances on Protein Folding Simulations Based on the Lattice HP models with Natural Computing. *Applied Soft Computing 8*, 2 (2008), 1029–1040. (cited on page 31)

[249] Zhou, C., Hou, C., Zhang, Q., and Wei, X. Enhanced Hybrid Search Algorithm for Protein Structure Prediction Using the 3D-HP Lattice Model. *Journal of Molecular Modeling 19*, 9 (2013), 3883–3891. (cited on page 33)