COPPE
Instituto Alberto Luiz Coimbra de
Pós-Graduação e Pesquisa de Engenharia
UFRJ

NEW METHODS FOR THE DETERMINATION OF THE THREE-DIMENSIONAL STRUCTURE OF PROTEINS AND NANOPARTICLES

Gabriel Pineschi Braun

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientadores: Nelson Maculan Filho
Carlile Campos Lavor

Rio de Janeiro
Março de 2024

NEW METHODS FOR THE DETERMINATION OF THE THREE-DIMENSIONAL STRUCTURE OF PROTEINS AND NANOPARTICLES

Gabriel Pineschi Braun

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Orientadores: Nelson Maculan Filho
Carlile Campos Lavor

Aprovada por: Prof. Nelson Maculan Filho
Prof. Carlile Campos Lavor
Prof. Felipe Maia Galvão França
Prof. Pedro Henrique González Silva

RIO DE JANEIRO, RJ – BRASIL
MARÇO DE 2024

Braun, Gabriel Pineschi

New methods for the determination of the three-dimensional structure of proteins and nanoparticles/Gabriel Pineschi Braun. – Rio de Janeiro: UFRJ/COPPE, 2024.

X, 40 p.: il.; $29,7$cm.

Orientadores: Nelson Maculan Filho
Carlile Campos Lavor

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2024.

Referências Bibliográficas: p. 36 – 39.

1. Distance geometry. 2. Mathematical programming. 3. Nuclear magnetic resonance. I. Maculan Filho, Nelson *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

NOVOS MÉTODOS PARA A DETERMINAÇÃO DA ESTRUTURA TRIDIMENSIONAL DE PROTEÍNAS E NANOPARTÍCULAS

Gabriel Pineschi Braun

Março/2024

Orientadores: Nelson Maculan Filho
Carlile Campos Lavor

Programa: Engenharia de Sistemas e Computação

Através da utilização de dados experimentais obtidos por técnicas como a ressonância magnética nuclear e o método da função de distribuição de pares, torna-se possível determinar estruturas tridimensionais de proteínas e nanoestruturas. Estes são exemplos de instâncias do problema de Geometria de Distâncias Não Associadas (uDGP), onde o objetivo principal é determinar as posições de pontos específicos com base em um conjunto de valores de distância que não foram previamente atribuídos a pares de pontos específicos. Apresentam-se, nesta dissertação, novas formulações de programação matemática e uma abordagem heurística para resolver o uDGP. Nossos resultados demonstram o desempenho superior dos modelos propostos em comparação com os métodos existentes documentados na literatura. Esses modelos têm um imenso potencial para uso prático, oferecendo soluções eficazes para a determinação de estruturas em aplicações envolvendo moléculas, nanopartículas e proteínas.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

NEW METHODS FOR THE DETERMINATION OF THE THREE-DIMENSIONAL STRUCTURE OF PROTEINS AND NANOPARTICLES

Gabriel Pineschi Braun

March/2024

Advisors: Nelson Maculan Filho
Carlile Campos Lavor

Department: Systems Engineering and Computer Science

Utilizing experimental data from techniques like nuclear magnetic resonance and the pair distribution function method, it becomes possible to derive three-dimensional protein structures and nanostructures. These are examples of instances of the Unassigned Distance Geometry problem (uDGP), where the primary objective is to determine the positions of certain points based on a set of distance values that have not been previously assigned to specific point pairs. This dissertation presents new mathematical programming formulations and a heuristic approach to solve the uDGP. Our findings demonstrate the superior performance of the proposed models in comparison to existing methods documented in the literature. These models hold immense potential for practical use, offering effective solutions for the determination of structures in applications involving molecules, nanoparticles, and proteins.

Contents

List of Figures

List of Tables

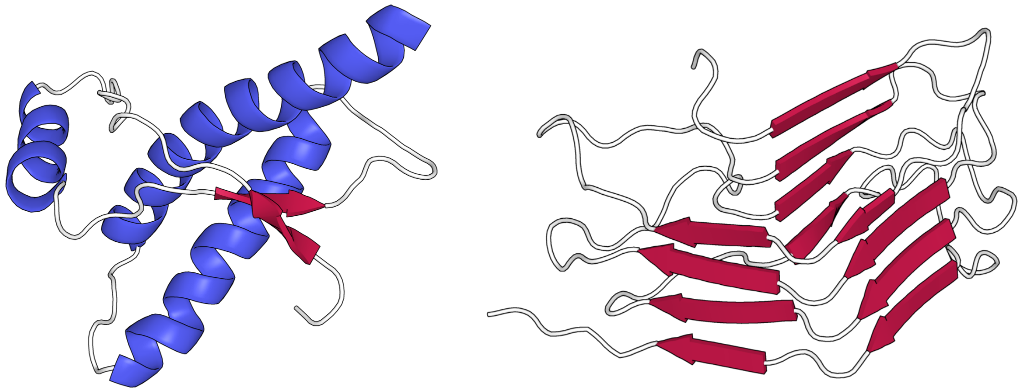
Chapter 1

Introduction

1.1 Motivation

Determining the three-dimensional structures of molecules, proteins, and nanoparticles is crucial for understanding their chemical and physical properties. Moreover, they provide insight into molecular interactions, catalyst mechanisms, and binding sites, aiding in drug development, materials design, and catalysis [8].

Consider the prion protein (PrPc) as an illustration. It interacts with the cells of certain animals, but when its spatial structure is altered, it gives rise to a modified protein known as prion scrapie (PrPSc), as depicted in Figure 1.1. While this altered protein maintains the same composition as its original form, the structural change leads to the emergence of a disease known as Bovine Spongiform Encephalopathy, commonly referred to as Mad Cow Disease [9].



(a) Normal prion (PrPc)

(b) Diseased prion (PrPSc)

Figure 1.1: Normal and abnormal conformations of a prion protein. **(a)** Experimentally determined structure of a normal prion protein [1]. In this conformation, most amino acids participate in alpha-helix and less than 5% in beta sheets. **(b)** Calculated structure of an abnormal prion protein [2]. The abnormal protein is misfolded, where the majority of alpha helices are converted into beta sheets.

X-ray crystallography was the first method developed for the determination of molecular structures [10]. The underlying principle of this technique is that the crystalline atoms cause a beam of X-rays to diffract into many specific directions. By measuring the angles and intensities of these diffracted beams, a crystallographer can produce a three-dimensional picture of the density of electrons within the crystal. From this electron density map, the positions of the atoms in the crystal are determined [11].

The scope of X-ray crystallography, as implied by its name, is confined to the examination of molecules capable of crystallization in solid form. Given that numerous proteins defy isolation as solids and are exclusively present in solution, novel techniques emerged to determine three-dimensional structures exploiting distance information between atom pairs provided by experiments such as Nuclear Magnetic Resonance (NMR) [8]. The data generated from these experiments is a list of distances between pairs of atoms within the structure, allowing proteins to be studied in solution.

Accurate determination of atomic positions within nanoparticles also holds considerable significance, primarily because their size and three-dimensional structure are the main factors dictating their properties [12]. Figure 1.2 illustrates examples depicting the structure of representative nanoparticles.

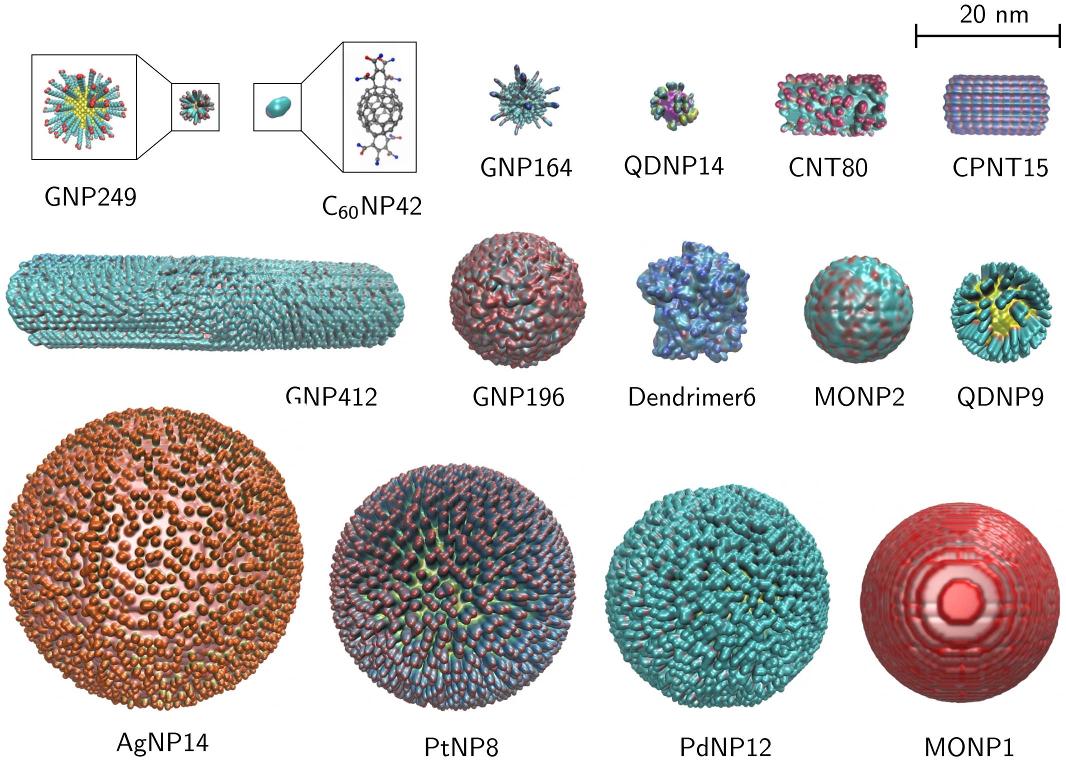


Figure 1.2: Visualization of 15 representative nanoparticle structures in the PubVINAS database [3].

For the case of nanoparticles, higher accuracy is required when determining their structure. The level of precision needed can vary depending on the specific context, but in the case of solid-state systems, it often consists of determining the positions of all atoms to better than 2% for each interatomic distance within that nanostructure [13]. In certain cases, even finer resolution may be necessary. This high level of precision is crucial because the properties of nanostructured materials and intricate molecules are extremely sensitive to minor alterations in the distances between atoms. Therefore, achieving a precise understanding of the nanostructure is essential for the design and comprehension of these materials [14].

X-ray crystallography is also the prevailing method for determining the structure of nanoparticles, however, many structures are also challenging or impractical to obtain through the growth of a single crystal or even a polycrystalline sample [15]. For nanomaterial applications, distance data from NMR analysis used for protein structure determination has too low resolution, with uncertainties of the order of one angstrom. In those cases, the structures can be determined from atomic pair distribution function (PDF) analysis data [5]. This data also consists of a list of distances between pairs of atoms within the structure.

For both nanoparticles and proteins, when their structures cannot be isolated as crystals, determining their three-dimensional conformation necessitates the reconstruction of the structure from a set of distances between pairs of atoms. In consequence, a new problem emerges: the task of deducing the three-dimensional structure from this distance data, commonly referred to as the Distance Geometry Problem (DGP).

1.2 Goals and Contribution

This work aims to present novel mathematical programming formulations and a heuristic approach to address the unassigned variant of the Distance Geometry Problem (uDGP). In this particular variation, only distance values are available, without any accompanying information linking them to specific pairs of atoms. Therefore, this approach enables the direct formulation of data from NMR and PDF as an uDGP, eliminating the need for experimental efforts to associate each calculated distance with its corresponding atom pair.

The primary contribution of this research lies in the development of a novel technique for solving the Unassigned Distance Geometry Problem, enabling precise determination of molecular and nanoparticle structures based solely on distance data. Consequently, this study demonstrates the practicality and effectiveness of applying the uDGP method to accurately infer the structures of both molecules and nanoparticles from NMR and PDF data.

1.3 Dissertation Organization

The organization of this work can be outlined as follows. Initially, Chapter 2 presents the formal definition of the Distance Geometry problem, and the assigned and unassigned variants. Additionally, Chapter 3 provides a more in-depth exploration of its principal applications within the realm of molecular structure determination.

Moving forward, Chapter 4 provides an overview of the current methods employed in solving the Unassigned Distance Geometry Problem (uDGP). Then, Chapter 5 presents our novel mathematical programming model and heuristic.

Subsequently, Chapter 6 elaborates on the instances utilized for testing and validating our approach and Chapter 7 outlines the experiments conducted to validate our method and discusses the resultant findings.

Finally, Chapter 8 concludes this work by summarizing the achieved results and proposing future avenues of exploration within the domain of Distance Geometry Problems.

Chapter 2

Distance Geometry Problem

The field of Distance Geometry (DG) delves into challenges associated with the notion of distance from a geometric standpoint. Presently, work in this area focuses around the task of determining a set of coordinates within a geometric space based on a set of provided distances.

The inception of this field occurred when MENGER [16] utilized the concept of distance to define several geometric principles like congruence and set convexity. Distance Geometry was formally recognized as an emerging research field following the contributions of BLUMENTHAL [17] in 1953.

The interest in the DGP resides in its wealth of applications: molecular structure and conformation, wireless sensor networks, statics, dimensionality reduction, and robotics, as well as in the study of the related mathematical theory [8].

This chapter provides a formal definition of the Distance Geometry Problem (DGP). In Section 2.1, we introduce the concept of assigned and unassigned classes within DGP.

2.1 The assigned and unassigned classes

The Distance Geometry Problem (DGP) can be categorized into two distinct classes: Assigned (aDGP) and Unassigned (uDGP) [8, 18]. The differentiating factor between these categories lies in whether the association between distances and the corresponding pairs of vertices is provided or not.

2.1.1 The Assigned Distance Geometry Problem (aDGP)

In the Assigned Distance Geometry Problem (aDGP), the input data consists of a list of distances, all of which are assigned to the specific pairs of vertices they correspond to.

Definition 1 (Assigned Distance Geometry Problem, aDGP [19])**.** Given an integer $K > 0$ and a simple, non-directed graph $G = (V, E, d)$ whose edges weights are given by a non-negative function $d : E \to [0, \infty)$, find a function $x : V \to \mathbb{R}^K$ such that

$$\forall (i, j) \in E,\ \|x_i - x_j\| = d_{ij} \tag{2.1}$$

where $x_i = x(v_i), x_j = x(v_j)$ and $\|x_i - x_j\|$ is the *Euclidean distance* between the coordinates x_i and x_j.

Definitions 2, 3 and 4, presented below, apply to all classes of the Distance Geometry Problem.

Definition 2 (realization [10])**.** The x function, which maps the vertices of the graph in a DGP to coordinates in the $\mathbb{R}^K$ space, is denominated a **realization** of G in $\mathbb{R}^K$.

Definition 3 (valid realization)**.** If x satisfies the system of Equations 2.1, then x is a **valid realization**.

Definition 4 (framework)**.** The pair (G, x), where x is a valid realization, is called a **framework**.

An example of an aDGP instance and solution is presented in Figure 2.2.

$$\{d_{ij}\} = \begin{bmatrix} 0 & 1.50 & 2.45 & 2.87 \\ 1.50 & 0 & 1.50 & 2.45 \\ 2.45 & 1.50 & 0 & 1.50 \\ 2.87 & 2.45 & 1.50 & 0 \end{bmatrix} \xRightarrow{\text{aDGP}}$$

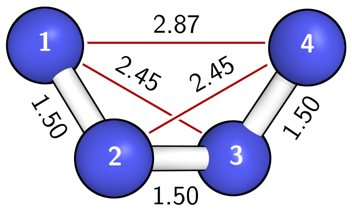


Figure 2.1: Example of an aDGP instance and solution with 4 atoms.

As demonstrated in the work by Saxe in 1980, the complexity of this problem within a Euclidean space with of dimension K is NP-Hard [20]. This is result is derived through a reduction that associates this problem with the 3-Satisfiability Problem.

2.1.2 The Unassigned Distance Geometry Problem (uDGP)

In the unassigned distance geometry problem (uDGP), only the distances are provided, lacking the information about the specific pairs of vertices to which these distances correspond. Therefore, in addition to finding the realization for the vertices, it is also necessary to associate the input distances to vertex pairs they correspond to.

Definition 5 (Unassigned Distance Geometry Problem, uDGP [19])**.** Given an integer $K > 0$, a set of vertices V and a list of distance values $d_1, d_2, \ldots, d_m$, find an injective function $g : \{1, \ldots, m\} \to \{v_i v_j : i = 1, \ldots, n-1;\ j = i+1, \ldots, n\}$ and a function $x : V \to \mathbb{R}^K$ such that $\forall \{i, j\} \in g(\{1, \ldots m\})$,

$$\|x_i - x_j\| = \delta_{ij} \tag{2.2}$$

and

$$\delta_{ij} = d_{g^{-1}(i,j)} \tag{2.3}$$

where $x_i = x(v_i), x_j = x(v_j)$ and $\|x_i - x_j\|$ is the *Euclidean distance* between the coordinates x_i and x_j.

Here, the x function is also the realization of the graph G associated to the uDGP, and g is an assignment function that defines a set $E \subset v \times V$, the edges of G. An example of uDGP is presented in Figure 2.2.

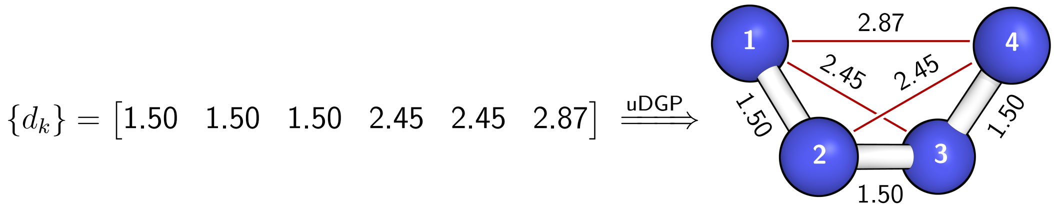

$\{d_k\} = [1.50 \quad 1.50 \quad 1.50 \quad 2.45 \quad 2.45 \quad 2.87]$
uDGP
1
2
3
4
2.87
2.45
2.45
1.50
1.50
1.50

Figure 2.2: Example of an uDGP instance and solution with 4 atoms.

By comparing Figures 2.1 and 2.2, one can observe the increased complexity of the uDGP in contrast to the aDGP, since the same structure must be determined with less input information. The uDGP class is particularly challenging because the graph structure and the graph realization both need to be determined at the same time.

Depending on the application, the embedding space for the Distance Geometry Problem can be very general. Due to the central focus of this study on molecular structures, our approach will revolve around three-dimensional space, with the value of K set as 3.

Chapter 3

Applications of the Distance Geometry Problem

In this chapter, we elaborate further into the main applications of the Distance Geometry Problem (DGP). Section 3.1 detailshow the DGP can be a tool in determining the structural characteristics of molecules and proteins, and Section 3.2 extends the discussion for nanoparticles. Section 3.3 gives a brief overview of other applications outside the field of chemistry.

3.1 Determination of molecular and protein structures

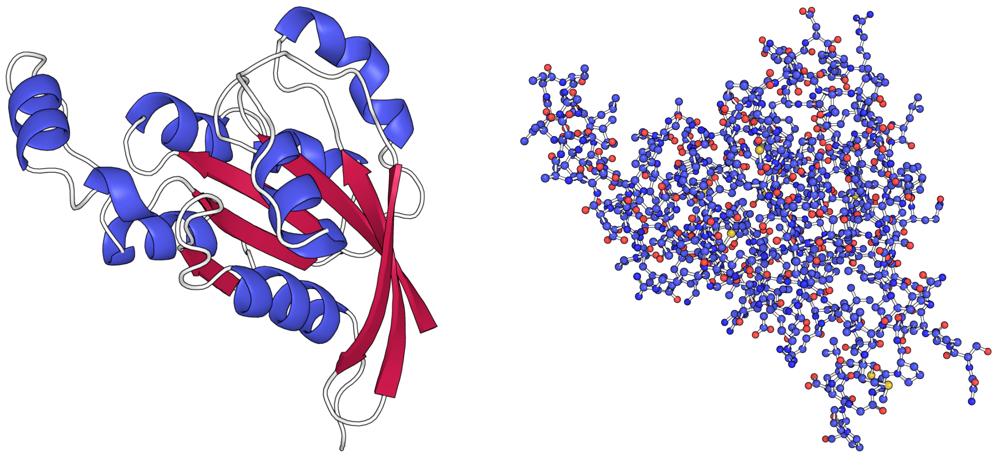
Experimental methods can be employed to determine the atomic distances within molecules and proteins, enabling the realization of their three-dimensional structures as a solution to a Distance Geometry Problem. The most common approach is to derive the list of distances from Nuclear magnetic resonance spectroscopy (NMR) data [21–24]. Recently, various alternative methodologies for measuring distances in biological and inorganic materials have emerged, holding significant potential for ongoing advancements in this field [25, 26].

In an NMR experiment, a sample is positioned within a magnetic field, and the signal is generated by the excitation of the sample's nuclei using radio waves, inducing nuclear magnetic resonance. This resonance is subsequently detected using sensitive radio receivers. The intramolecular magnetic field surrounding an atom in a molecule alters the resonance frequency. Since these magnetic fields are unique or highly characteristic to individual compounds, the data obtained allows access to intricate details regarding the electronic structure of the molecule and its specific functional groups [27].

NMR experiments can also be used to calculate interatomic distances because of the Nuclear Overhauser Effect (NOE). This effect describes the change in integrated intensity of one NMR resonance when another is saturated through irradiation with a radio frequency field. This change in resonance intensity arises from the proximity of the nucleus to those directly affected by the radio frequency perturbation and enables the determination of the distance between two nuclei.

Due to the significant errors in the distances calculated in NMR experiments exploiting the NOE, those are usually treated as intervals, rather than precise values. Hence, when formulated as an instance of the DGP, they become restraints rather than constraints.

After the distances are determined, a substantial amount of experimental effort is dedicated to determining the specific pair of nuclei associated with each distance extracted from NMR data. This process enables the formulation of the problem as an aDGP [13]. Since the information that is actually given by NMR experiments consists of a list of distance values that are only subsequently assigned to atom pairs, the problem can be formulated directly as an uDGP [19]. This poses a challenge, however, due to the large size of those instances, as shown in Figure 3.1, and the additional complexity of the uDGP when compared to the aDGP.



(a) Ribbon diagram

(b) Individual atoms

Figure 3.1: Three-dimensional structure of the small human rac1 protein [4].

Figure 3.1 depicts the configuration of the human rac1 protein, which is a relatively small protein [4]. Usual protein structures are comprised of thousands of atoms. As the number of distance values increases, the complexity of the uDGP escalates significantly, posing a substantial challenge when attempting to determine protein structures using this method. DUXBURY *et al.* [19] proposed the first methods to realize the structure of proteins using the uDGP inside a heuristic.

3.2 Determination of nanoparticle structures

Nanostructure-related distance geometry problems arise across a diverse spectrum of materials, encompassing intricate molecules, nanoparticles, polymers, as well as non-crystalline elements embedded within crystalline matrices, among others. In the determination of nanostructures, achieving high resolution is necessary. This is because the behavior of nanostructured materials and intricate molecules is exceptionally responsive to even small alterations in interatomic distances. Thus, precision in nanostructure determination is crucial for comprehending and crafting materials effectively [13].

The pair-distribution function (PDF) method serves as a versatile and easily accessible approach for investigating the local atomic structure of nanoparticles. PDF results can be derived from X-ray, neutron, or electron total scattering data, and in many instances, data collection can be conducted efficiently [28].

The PDF results can be further processed into a radial distribution function, RDF. An example of a PDF and its associated RDF are shown in Figures 3.3 and 3.4, respectively. The experimental interatomic distances are obtained from the positions of peak maxima and shoulders of the RDF, and their multiplicities are set proportionally to the peak areas. From this list of distances, the structure can be reconstructed by using global optimization methods like the uDGP.

Figure 1.2 in Section 1.1 illustrates different types of nanoparticles. The number of atoms in the structure of those particles can vary from a few hundred to thousands. A common example of a small nanoparticle is fullerene, C_{60}, shown in Figure 3.2.

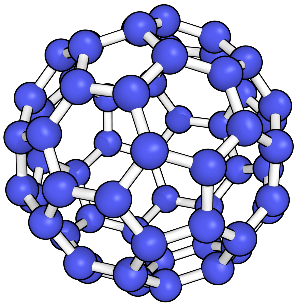


Figure 3.2: Ball-and-stick representation of the structure of fullerene, C_{60}.

Nanoparticles typically contain fewer atoms compared to proteins, and the more precise distance data from PDF when compared to NMR makes the determination of nanoparticle structure an easier problem than the protein analog. Consequently, employing the uDGP for structural determination of nanoparticles is a more common application.

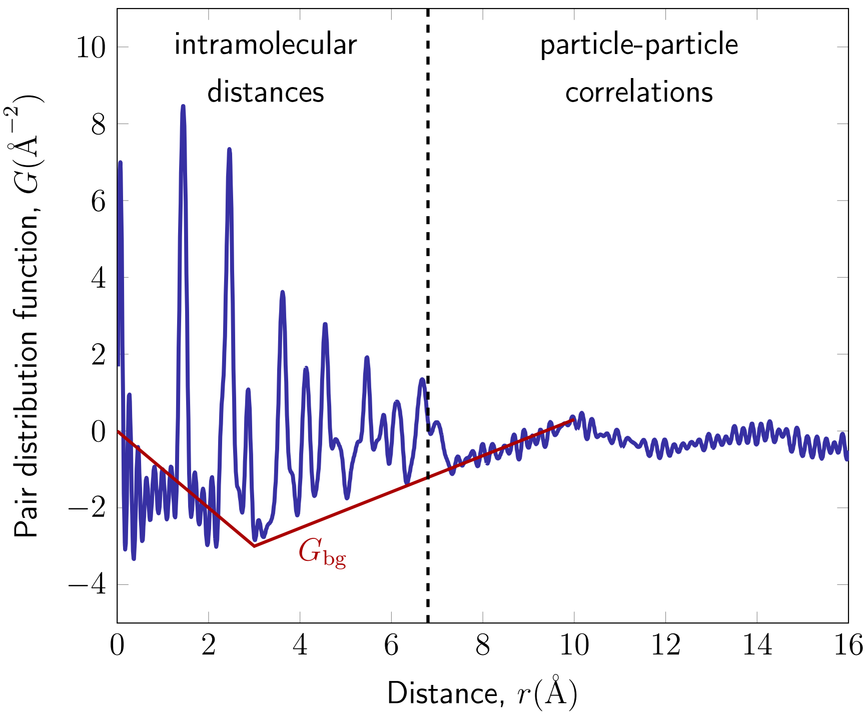


Figure 3.3: Experimental pair distribution function, PDF, from solid C_{60} as a function of distance. The red line shows background arising from interparticle correlations [5].

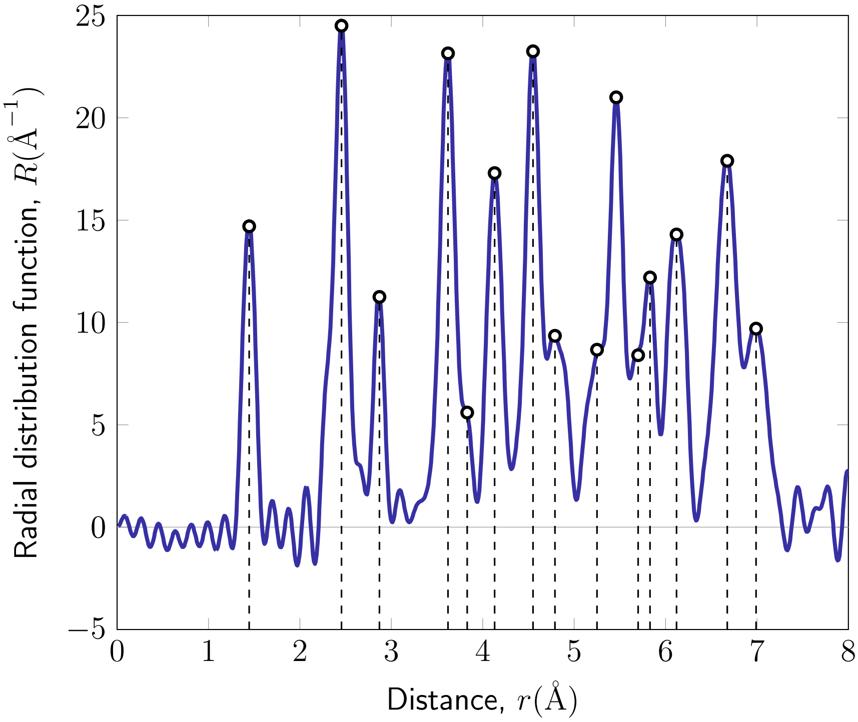


Figure 3.4: The background-subtracted data from the solid C_{60} PDF in the form of the radial distribution function [5].

3.3 Other applications

The Distance Geometry Problem has various well-established applications, including wireless networks, statics, dimensionality reduction, and robotics. LIBERTI *et al.* [8] provide an extensive list of applications for DGP. Here, we summarize a selection of some of the most relevant ones.

In wireless networks, mobile sensors often determine their pairwise distances by monitoring their communication energy consumption. These distance measurements are subsequently utilized to compute the precise positions of each sensor within the network.

Statics, on the other hand, focuses on analyzing the equilibrium of rigid structures, predominantly those of human construction, such as buildings and bridges, when subjected to external forces. A well-known model for such structures is the bar-and-joint framework, which is essentially a weighted graph. The central challenge in this context is to determine whether a given graph, with a specified distance function along its edges, exhibits rigidity or flexibility. Additionally, there is the related task of determining whether a given graph effectively models a rigid structure, independently of the specific distance function applied.

In the case of dimensionality reduction, the objective is to discover a projection within the plane or space that visually aligns the graph as closely as possible with its higher-dimensional representation.

In robotics, the primary concern revolves around understanding the movement of a robotic arm or a system of robotic arms within a given space to execute specific tasks. This involves knowledge of known distances, such as the distances from a joint to its neighboring joints. The central problem lies in assigning coordinate values to the position vector of the farthest joint.

Chapter 4

Methods for solving the uDGP

In this chapter, we explore the main methods in the literature for solving the uDGP in the context of the realization of the structure of nanoparticles and proteins. Section 4.1 presents the heuristic methods which were the initial techniques employed for solving molecular conformation problems. Section 4.2 discusses mathematical programming methods, which constitute the fundamental approach within our proposed methodology.

4.1 Heuristic methods

The first methods which employed the uDGP for the realization of chemical structures focused their studies on nanoparticles. Since nanostructures usually have fewer atoms than proteins, and more accurate experimental distance data is available, their instances are easier.

JUHÁS *et al.* [5] proposed the Liga method in 2006 and GUJARATHI *et al.* [29] proposed the TRIBOND method in 2014. Those were the first generation methodologies to solve the uDGP applied to chemical structure problems [19]. Both methods are heuristics that adopt a build-up approach and rely on the availability of adequate distance constraints to guarantee a singular and unambiguous solution at each stage of the procedure.

4.1.1 Liga

The Liga method consists of a stochastic algorithm which grows large clusters by adding atoms to a population of high-quality subclusters. This algorithm incorporates a strategy for backtracking and updating populations of high-quality clusters at each size, which is inspired by promotion and relegation in sport—such as occurs in European soccer leagues like La Liga in Spain

The Liga algorithm can be described as follows [5]:

A. **Build-up procedure**. Start with a single atom, and place the second atom at a randomly selected distance from the target list. Find the third position by constructing a triangle using two target distances. Add additional atoms are by constructing 4-vertex pyramids, while attempting to use only the allowed distances from the target list.

There are many small clusters that use allowed target lengths, but are inconsistent with the target structure. Growth from these incorrect clusters eventually leads to an increase in the cost function, and the algorithm then has to backtrack to repair the faulty part of the cluster.

B. **Backtracking procedure**. Backtracking is carried out by first evaluating the individual atom contributions to the total error and removing the *worst* atoms according to a stochastic procedure where the probability an atom is removed is proportional to its associated error contribution.

4.1.2 TRIBOND

The TRIBOND method consists of a deterministic algorithm. The theoretical foundations of the method are based on rigidity theory, which enables derivation of a polynomial bound on its efficiency. The algorithm consists of finding a core, which is shown in Figure 4.1 by setting the smallest bond as the *base bond* and then test all the bond combinations using the triangle inequality to generate feasible triangle pairs. Subsequently, in the build-up procedure, more atoms are added to the core.

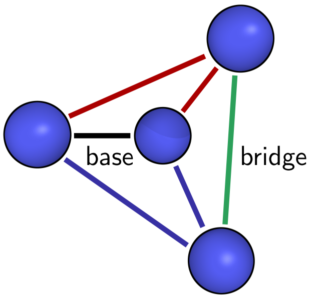


Figure 4.1: An example of a core in the TRIBOND method. In two dimensions, it consists of four points. The horizontal bond is the base, in black, the bonds below it, in blue, make up the base triangle while those above it, in red, make up the top triangle. The vertical bond is the bridge, in green.

The TRIBOND algorithm can be described as follows [29]:

A. **Core finding procedure**.

1. Choose the shortest bond as the base bond and a window (subset) of $W = 6$ smallest entries in the distance list for the core finding search.
2. Iterate over all triangles constructed with the triangle inequality that have the same base bond using distances in the window W.
3. Search over all pairs of feasible triangles generated above and calculate the bridge bond. Using a binary search, test if there is an unused distance that matches the bridge bond. If such a value is found, we have a core. Remove the edges used from the distance list and exit to the buildup procedure.

B. **Build-up procedure**.

1. Search over all sets of two edges from the distance list to find a set compatible with the base triangle in the existing structure. Search over the distance list to test the bridge bond.
2. If successful, remove from the distance list the edges that are used in connecting the newly added node. If not all atoms have been places, return to the previous step and resume the search.
3. If no compatible set can be found, find a new core and restart.

Core finding requires a search across all potential base and top triangles, whereas buildup only demands a search through top triangles, since the base triangle is a predetermined component of the structure. As a result, buildup necessitates considerably fewer computations than core finding. This procedure is outlined in Figure 4.2.

4.2 Mathematical programming methods

In 2020, DUXBURY *et al.* [19] introduced mathematical programming formulations for the uDGP in the context of molecular and biomolecular structure determination. Alongside presenting theoretical findings connected to these formulations, they also introduced a novel heuristic approach for addressing this problem.

Due to the properties of the assignment function g outlined in Definition 5, binary variables denoted as $a_{i,j}^k$ are introduced such that

$$a_{i,j}^k = 1 \iff \text{distance } d_k \text{ is assigned to the pair } (i,j) \in V \times V \tag{4.1}$$

4.2.1 Model M1

Considering vertices $v_1, \dots, v_n \in V$ and distance values $d_1, \dots, d_m$, the uDGP can be modeled as:

$$\min \quad \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\sum_{k=1}^{m} a_{i,j}^k \left(\|x_i - x_j\|^2 - d_k^2 \right)^2 \right) \tag{4.2}$$

subject to:

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{i,j}^k \leq 1 \qquad k = 1, 2, \dots, m \tag{4.3}$$

$$\sum_{k=1}^{m} a_{i,j}^k = 1 \qquad i = 1, \dots n-1, \;\; j = i+1, \dots n \tag{4.4}$$

where $x_i \in \mathbb{R}^3$, $a_{i,j}^k \in \{0, 1\}$.

The relationship between a uDGP solution and a solution to Model M1 is stated in Theorem 1 [19].

Theorem 1. *A pair (g, x) is a solution for an uDGP instance associated to a graph $G = (V, E)$, with $|V| = n$, $|E| = m$, $g : 1, \dots, m \to V \times V$, and $x : V \to \mathbb{R}^3$, if and only if (x, a) is a global optimal solution to Model M1.*

4.2.2 Model M1R

Model M1 can only solve instances of up to five atoms in a reasonable time. One of the reasons for the slow performance of the model is the huge number of binary variables $a_{i,j}^k$. To reduce number of binary variables in Model M1, a formulation with only continuous variables was introduced, inspired by the *Solid Isotropic Material with Penalization* (SIMP) method [30, 31].

$$\min \quad t - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\sum_{k=1}^{m} \left(a_{i,j}^k \right)^2 \right) \tag{4.5}$$

subject to:

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\sum_{k=1}^{m} a_{i,j}^{k}\Big(\|x_i - x_j\|^2 - d_k^2\Big)^2\right) = t \tag{4.6}$$

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_{i,j}^{k} \leq 1 \quad k = 1, 2, \ldots, m \tag{4.7}$$

$$\sum_{k=1}^{m} a_{i,j}^{k} = 1 \quad i = 1, \ldots n-1, \;\; j = i+1, \ldots n \tag{4.8}$$

where $t \geq 0$, $x_i \in \mathbb{R}^3$, $0 \leq a_{i,j}^{k} \leq 1$.

The relationship between a uDGP solution and a solution to Model M1R is stated in Theorem 2 [19].

Theorem 2. *A pair (g, x) is a solution for an uDGP instance associated to a graph $G = (V, E)$, with $|V| = n$, $|E| = m$, $g : 1, \ldots, m \to V \times V$, and $x : V \to \mathbb{R}^3$, if and only if (x, a) is a global optimal solution to Model M1R for some value of t with globally optimal objective function value equal to $-m$.*

4.2.3 NEW-TRIBOND (NT)

Model M1R demonstrates superior capacity in solving larger instances when compared to Model M1. Nevertheless, when confronted with instances featuring hundreds of atoms, it falls short. To address such instances, a heuristic inspired by the TRIBOND method, leveraging the foundation of Model M1R, can be employed.

The first step is to find a *core*, positions in $\mathbb{R}^3$ for five vertices with ten associated distances provided from the list of distance values, solving Model M1R considering just five points, and then increase its size by adding one vertex position at a time solving a modification of Model M1R, where four random points (already fixed) are used to find the next position:

A. **Core finding procedure**.

Find a core $x_1, \ldots, x_5 \in \mathbb{R}^3$ solving the problem

$$\min \quad t - \sum_{i=1}^{4}\sum_{j=i+1}^{5}\left(\sum_{k=1}^{m}\left(a_{i,j}^{k}\right)^2\right) \tag{4.9}$$

subject to:

$$\sum_{i=1}^{4}\sum_{j=i+1}^{5}\left(\sum_{k=1}^{m} a_{i,j}^{k}\left(\|x_i - x_j\|^2 - d_k^2\right)^2\right) = t \tag{4.10}$$

$$\sum_{i=1}^{4}\sum_{j=i+1}^{5} a_{i,j}^{k} \leq 1 \quad k = 1, 2, \ldots, m \tag{4.11}$$

$$\sum_{k=1}^{m} a_{i,j}^{k} = 1 \quad i = 1, \ldots 4, \;\; j = i+1, \ldots 5 \tag{4.12}$$

where $t \geq 0$, $x_i \in \mathbb{R}^3$, $0 \leq a_{i,j}^k \leq 1$.

B. **Build-up procedure**.

1. For $i = 6, \ldots, n$, solve the problem

$$\min \quad t - \sum_{j \in J}\left(\sum_{k=1}^{m_i}\left(a_{i,j}^{k}\right)^2\right) \tag{4.13}$$

subject to:

$$\sum_{j \in J}\left(\sum_{k=1}^{m_i} a_{i,j}^{k}\left(\|x_i - x_j\|^2 - d_k^2\right)^2\right) = t \tag{4.14}$$

$$\sum_{j \in J} a_{i,j}^{k} \leq 1 \qquad k = 1, 2, \ldots, m_i \tag{4.15}$$

$$\sum_{k=1}^{m} a_{i,j}^{k} = 1 \qquad j \in J, \ldots n \tag{4.16}$$

where $t \geq 0$ and $0 \leq a_{i,j}^k \leq 1$. Here, $x_i \in \mathbb{R}^3$ is the position to be determined. Here the set J includes the indices of all of the already fixed points $x_j \in \mathbb{R}^3$, $j \in J \subset \{1, \ldots, i-1\}$, and m_i is the number of available distances.

2. If a set of compatible distances cannot be found for some $i = 6, \ldots, n$, find a new core (return to Step 1) and restart.

The importance of a core in Step A is to allow, with high probability [29], to start correctly the reconstruction of the molecular structure. After finding a core, the geometric idea of Step B is to intersect fours spheres [32] (centered at points y_j), which gives one point if there are consistent distance values (radii of the spheres) from the list of distances. This algorithm follows the same procedure outlined in Figure 4.2.

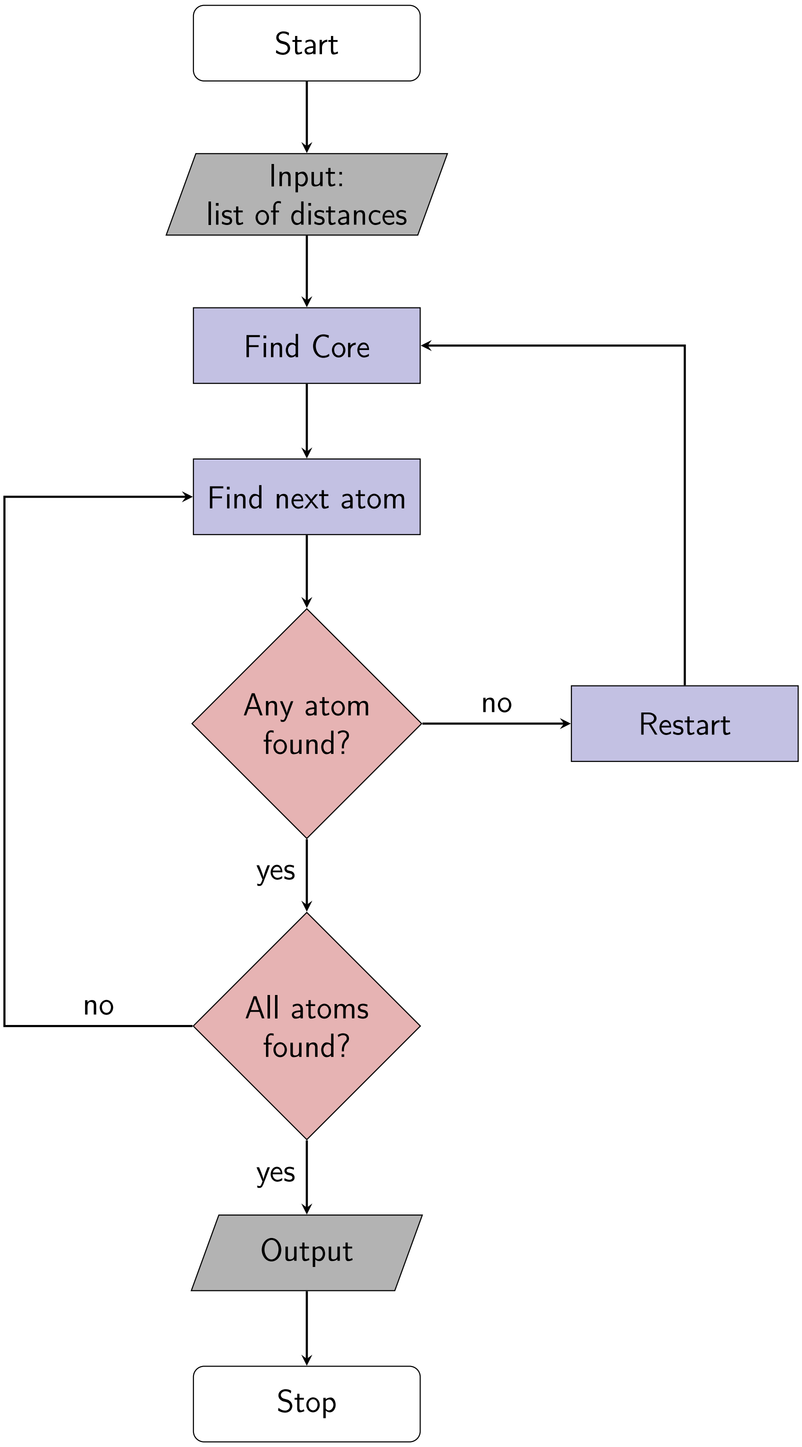


Figure 4.2: The TRIBOND algorithm.

Chapter 5

Proposed Methods

In this chapter, the proposed methods for solving the Unassigned Distance Geometry Problem (uDGP) will be discussed. To begin, Section 5.1 will introduce our new mathematical programming formulations for uDGP, building upon the concepts outlined in Section 4.2. Following that, in Section 5.3, we will elaborate on a novel heuristic that harnesses these newly developed formulations.

5.1 New mathematical programming models

The core concept behind our approach involves reconfiguring the objective function in Equation 4.2 in a manner that is becomes convex. We propose the following objective function:

$$\min \quad \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\sum_{k=1}^{m} a_{i,j}^{k} \Big| \|x_i - x_j\| - d_k \Big| \right) \tag{5.1}$$

5.1.1 Model M2

Considering our new objective function, which can be rewritten as:

$$\min \quad \sum_{k=1}^{m} y_k \tag{5.2}$$

subject to the original constraints 4.3 and 4.4:

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{i,j}^{k} \leq 1 \qquad k = 1, 2, \ldots, m \tag{5.3}$$

$$\sum_{k=1}^{m} a_{i,j}^{k} = 1 \qquad i = 1, \ldots n-1, \;\; j = i+1, \ldots n \tag{5.4}$$

and,

$$y_k \geq \alpha_k, \quad y_k \geq -\alpha_k \tag{5.5}$$

$$t_{i,j}^2 = \|x_i - x_j\|^2 \tag{5.6}$$

$$-(1 - a_{i,j}^k)D + t_{i,j} \leq z_{i,j}^k \leq t_{i,j} + (1 - a_{i,j}^k)D \tag{5.7}$$

$$-a_{i,j}^k D \leq z_{i,j}^k \leq a_{i,j}^k D \tag{5.8}$$

$$-(1 - a_{i,j}^k)D + (d_k + \alpha_k) \leq z_{i,j}^k \leq (d_k + \alpha_k) + (1 - a_{i,j}^k)D \tag{5.9}$$

where $y_k,\ t_{ij},\ z_{ijk} \geq 0,\ x_i,\ \alpha_k \in \mathbb{R},\ a_{i,j}^k \in \{0,1\}$ for $i = 1, \dots n-1,\ j = 1 + 1, \dots n,\ k = 1, 2, \dots, m$, and $D = \max\{d_k\}$:

The relationship between a uDGP solution and a solution to Model M1R is stated in Theorem 3.

Theorem 3. *A pair (g, x) is a solution for an uDGP instance associated to a graph $G = (V, E)$, with $|V| = n$, $|E| = m$, $g : 1, \dots, m \to V \times V$, and $x : V \to \mathbb{R}^3$, if and only if (x, a) is a global optimal solution to Model M2 for some values of (y, α, z, t).*

Theorem 3 follows trivially from an adaptation of Theorem 1.

5.1.2 Model M2C

Another approach to diminish the large number of binary variables in instances with more atoms involves considering the multiplicity of the input distances. This can be seen by considering the C_{60} molecule as illustrated in Figure 5.1, where there are only 21 different interatomic distances of the total 1770.

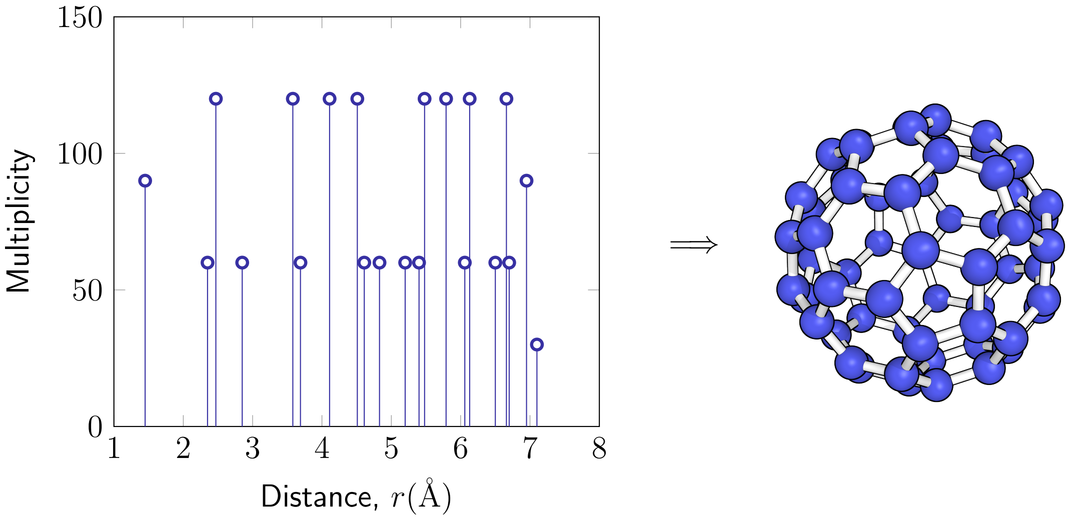


Figure 5.1: Fullerene, C_{60}, that has a degenerate distance list with a total of 1770 interatomic distances, but only 21 unique.

Thus, it is possible to modify Model M2 reducing the number of integer variables by considering the distance multiplicities. Considering that each distance d_k has multiplicity c_k, the original constraints 4.3 and 4.4 can be rewritten as:

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{i,j}^k \leq c_k \qquad k = 1, 2, \ldots, m \tag{5.10}$$

$$\sum_{k=1}^{m} a_{i,j}^k = 1 \qquad i = 1, \ldots n-1, \;\; j = i+1, \ldots n \tag{5.11}$$

The objective function and all the other constraints in M2 remain unchanged.

5.2 Model M2R

Unfortunately, the continuous relaxation of the objective function of Models M2 and M2C is not convex. Inspired by the increased performance of Model M1R in relation to Model M1, we have also proposed a modification of Model M2 with objective function:

$$\min \quad \sum_{k=1}^{m} y_k + \sum_{i=1}^{n-1} \sum_{j=1+1}^{n} \left(t_{ij}^2 - \|x_i - x_j\|^2 \right) \tag{5.12}$$

Subject to constraints 5.5, 5.7, 5.8, 5.9, 5.10 and 5.11 of Model M2C, and with constraint 5.6 being replaced by:

$$t_{i,j}^2 \geq \|x_i - x_j\|^2 \tag{5.13}$$

for $i = 1, \ldots n-1, \;\; j = 1+1, \ldots n$. The continuous relaxation of this new Model M2R has a non-convex objective function, and its set of constraints is convex. From constraint 5.13 we can say that any local optimum of will imply:

$$t_{i,j}^2 = \|x_i - x_j\|^2 \tag{5.14}$$

for $i = 1, \ldots n-1, \;\; j = 1+1, \ldots n$.

5.3 New heuristic (NT2)

Likewise for M1 and M1R, Model M2C can solve larger instances when compared to Model M2. However, it also cannot solve instances containing dozens of atoms.

To address these challenging instances, a similar approach as in Heuristic NT can be employed. The Heuristic NEW-TRIBOND-2 (NT2) employs the same principle as NEW-TRIBOND, with the distinction that Model M2C is employed at each step:

A. **Core finding procedure**.
Find a core $x_1, \dots, x_5 \in \mathbb{R}^3$ solving the problem

$$\min \quad \sum_{k=1}^{m_i} y_k \tag{5.15}$$

subject to:

$$\sum_{i=1}^{4} \sum_{j=i+1}^{5} a_{i,j}^k \leq c_k \qquad k = 1, 2, \dots, m \tag{5.16}$$

$$\sum_{k=1}^{m} a_{i,j}^k = 1 \qquad i = 1, \dots 4, \;\; j = i+1, \dots 5 \tag{5.17}$$

And constraints 5.5-5.6 for $i = 1, 2, \dots 5, \;\; j = 1, 2, \dots 5,$

B. **Build-up procedure**.

1. For $i = 6, \dots, n$, solve the problem

$$\min \quad \sum_{k=1}^{m_i} y_k \tag{5.18}$$

subject to:

$$\sum_{i=1}^{4} \sum_{j=i+1}^{5} a_{i,j}^k \leq c_k^i \qquad k = 1, 2, \dots, m \tag{5.19}$$

$$\sum_{k=1}^{m_i} a_{i,j}^k = 1 \qquad i = 1, \dots 4, \;\; j = i+1, \dots 5 \tag{5.20}$$

And constraints 5.5-5.6 where i is the index of the position to be determined, $x_i \in \mathbb{R}^3$, for $j \in J \subset \{1, \dots, i-1\}$, and $k \in 1, \dots, m_i$. Here, J is a random set with four indices related to already fixed points and m_i is the number of available distances with their respective available multiplicity c_k^i.

2. If a set of compatible distances cannot be found for some $i = 6, \dots, n$, find a new core (return to Step 1) and restart.

This algorithm also follows the same procedure outlined in Figure 4.2.

Chapter 6

Instances

This chapter describes the instances used to validate our proposed approaches for solving the Unassigned Distance Geometry Problem (uDGP). In the context of this work, an instance is regarded as the input data for all the mathematical programming models for the uDGP discussed previously. Two types of instances were studied: Lennard-Jones clusters and Lavor Instances.

In Chapter 3 we stated that the main application for the uDGP is the realization of the three-dimensional structure of nanoparticles and proteins. Section 6.1 details the Lennard-Jones clusters were used to assess the model's efficacy in reconstructing nanostructures, while Section 6.2.2 elaborates on the Lavor Instances, employed to evaluate the performance of the new methods in reconstructing molecules and proteins.

6.1 Lennard-Jones Clusters

The n-atom Lennard-Jones cluster, LJ-n, is the ground state configuration of n atoms assuming a Lennard-Jones pair potential acting between all the atoms. These systems occupy a distinct and crucial role in evaluating models for nanostructure investigation, which is primarily attributed to the simplicity in modeling weak forces they exhibit.

Another important aspect is the convenient generation of physical realizations of these systems in the form of rare gas clusters. Through techniques such as electron diffractometry [33] and mass spectrometry [34, 35], the experimental measurement of these rare gas clusters enables the validation of computational global optimization outcomes against real-world data. Consequently, Lennard-Jones clusters have become the standard benchmark for new optimization methods for the study of nanoparticle structures [36–38].

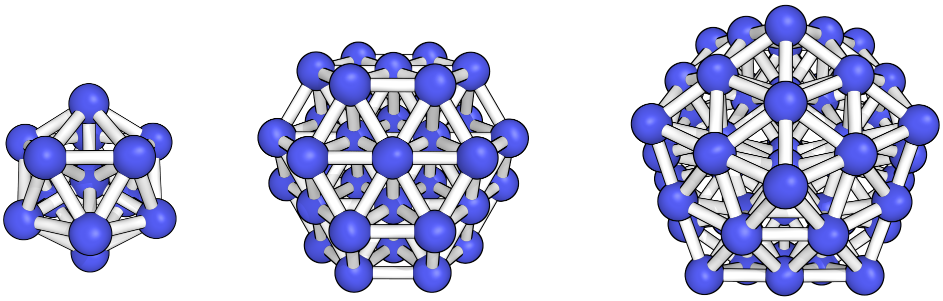
6.1.1 Lennard-Jones Cluster instance generation

The three-dimensional structure of the cluster is determined by minimizing the sum:

$$f_v = \sum_{(i,j)\in E} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right). \tag{6.1}$$

where f_v is the Lennard-Jones potential, E is the set of all atom pairs (i, j) and A_{ij} and B_{ij} are constants defined for each pair.

Hoare *et al.* [6] and Northby [39] have successfully determined the globally optimal structures for LJ-n systems across different n values.



(a) $n = 13$ (b) $n = 38$ (c) $n = 55$

Figure 6.1: Ball-and-stick representation of the three-dimensional structures of Lennard-Jones clusters of different sizes [6, 7].

Utilizing these coordinates, the distances between every atom pair are computed and subsequently employed as input for the uDGP.

6.2 Lavor Instances

The Lavor Instances [40] are designed to resemble the geometry of molecular and protein frameworks, offering a crucial tool for evaluating the effectiveness of models for investigating molecular and protein structures.

6.2.1 Lavor Instances background model

The Lavor Instances are based on the model proposed by Phillips *et al.* [41]. In this model, a molecule is conceived as a sequence of N atoms, each possessing Cartesian coordinates denoted by $x_1, \ldots, x_N$ in the three-dimensional space $\mathbb{R}^3$. For any consecutive pair of atoms i and j, the bond length r_{ij} represents the Euclidean distance between them. When considering three consecutive atoms, i, j, and k, the bond angle θ_{ik} reflects the geometric angle formed by the third atom in relation to the line defined by the preceding two. Similarly, for a sequence of four consecutive atoms, i, j, k, and l, the angle ω_{il}, called the torsion angle, quantifies the rotation between the planes formed by the atoms i, j, k and j, k, l.

Phillips also defines the following sets to facilitate further description:

- M_1 is the set of pairs of consecutive atoms (i, j).
- M_2 is the set of atom pairs (i, k) separated by two covalent bonds.
- M_3 is the set of atom pairs (i, l) separated by three covalent bonds.
- M_4 is the set of atom pairs (i, j) separated by more than two covalent bonds.

The three-dimensional structure of a molecule is determined by minimizing the sum of the following terms:

$$\begin{aligned} f_d &= \sum_{(i,j)\in M_1} c^r_{ij} \left(r_{ij} - r^0_{ij}\right)^2, \\ f_a &= \sum_{(i,k)\in M_2} c^\theta_{ik} \left(\theta_{ik} - \theta^0_{ik}\right)^2, \\ f_\omega &= \sum_{(i,l)\in M_3} c^\omega_{il} \left(1 + \cos(n_{il}\omega_{il} - \omega^0_{il})\right)^2, \\ f_v &= \sum_{(i,j)\in M_4} \left(\frac{A_{ij}}{r^{12}_{ij}} - \frac{B_{ij}}{r^6_{ij}}\right). \end{aligned} \tag{6.2}$$

where f_d, f_a and f_ω are the potentials corresponding to bond lengths, bond angles, and torsion angles, respectively. The constant c^r_{ij} is the bond stretching force constant, c^θ_{ik} is the angle bending force constant, and c^ω_{il} is the torsion force constant. The constants r^0_{ij} and θ^0_{ik} represent the equilibrium values for bond length and bond angle, respectively. The constant n_{il} defines the number of minima involved and $\omega^0_{il}$is the phase angle that defines the position of the minima.

Similarly to the case of Lennard-Jones clusters, the term f_v is the Lennard-Jones potential, where A_{ij} and B_{ij} are constants defined by each atom pair (i, j).

In this model, the bond lengths and bond angles are defined as $r_{ij} = 152.6\,\text{pm}$ for all $(i, j) \in M_1$ and $\theta_{ik} = 109.5^\circ$ for all $(i, k) \in M_2$, respectively. Additionally, $c^\omega_{il} = 1$, $n_{il} = 3$, and $w^0_{il} = 0$ for all $(i, l) \in M_3$, which results in three distinct *preferred* torsion angles: 60°, 180°, and 300°. By employing these parameters, we can calculate atomic distances and create instances for the uDGP.

6.2.2 Lavor Instance generation

Considering bond lengths $r_{ij} = 152.6\,\text{pm}$ and bond angles $\theta_{ik} = 109.5^\circ$ fixed, the three-dimensional structure of a molecule can be completely determined by its torsion angles, which are randomly selected from the set $\{60^\circ, 180^\circ, 300^\circ\}$.

To generate distances for the uDGP input, we first obtain Cartesian coordinates for each atom of the chain (x_{n1}, x_{n2}, x_{n3}), using the following matrices:

$$\begin{bmatrix} x_{n1} \\ x_{n2} \\ x_{n3} \\ 1 \end{bmatrix} = B_1 B_2 \dots B_n \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad n = 1, 2, \dots, N \tag{6.3}$$

where B_1 is the 4×4 identity matrix,

$$B_2 = \begin{bmatrix} -1 & 0 & 0 & -r_{12} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad B_3 = \begin{bmatrix} -\cos\theta_{13} & -\sin\theta_{13} & 0 & -r_{23}\cos\theta_{13} \\ \sin\theta_{13} & -\cos\theta_{13} & 0 & r_{23}\sin\theta_{13} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

and, for $i = 4, \dots, N$,

$$B_i = \begin{bmatrix} -\cos\theta_{(i-2)i} & -\sin\theta_{(i-2)i} & 0 & -r_{(i-1)i}\cos\theta_{(i-2)i} \\ \sin\theta_{(i-2)i}\cos\omega_{(i-3)i} & -\cos\theta_{(i-2)i}\cos\omega_{(i-3)i} & -\sin\omega_{(i-3)i} & r_{(i-1)i}\sin\theta_{(i-2)i}\cos\omega_{(i-3)i} \\ \sin\theta_{(i-2)i}\sin\omega_{(i-3)i} & -\cos\theta_{(i-2)i}\sin\omega_{(i-3)i} & \cos\omega_{(i-3)i} & r_{(i-1)i}\sin\theta_{(i-2)i}\sin\omega_{(i-3)i} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Figure 6.2 shows an instance generated using this procedure.

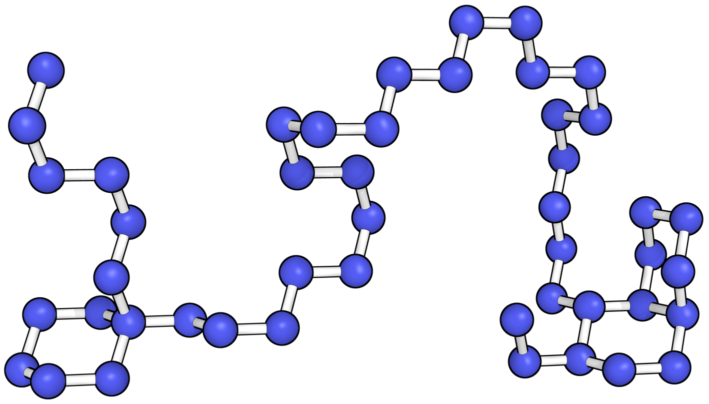


Figure 6.2: Example of a Lavor instance with 47 atoms.

Once these coordinates are generated, the distances between each pair of atoms are calculated to serve as input for the uDGP. In Section 3.1 we stated that because the distance information from NMR experiments has significant errors, they are treated as restraints rather than constraints. In this work, however, no uncertainties will be incorporated to instances, and the distance data will be precise.

Chapter 7

Results and Discussion

This chapter discusses all the experimental infrastructure to validate our new approach to solve the Unassigned Distance Geometry Problem (uDGP). Section 7.1 details the experimental setup of the experiments and the software used to solve the mathematical programming models, and Section 7.2 explores the experiments done, and the results obtained. Section 7.3 discusses the results.

7.1 Experimental Setup

This section presents the experimental setup used for the experiments. All experiments were conducted using a server equipped with an Intel® Core™ i9-10885H CPU @ 2.40 GHz with 8 cores and hyperthreading disabled.

Gurobi Optimizer is a prescriptive analytics platform and a decision-making technology developed by Gurobi Optimization. The Gurobi Optimizer is a solver, since it uses mathematical optimization to calculate the answer to a problem [42]. The Gurobi 10.0.2 build v10.0.2rc0 (linux64) solver was used to obtain the solution for the mathematical programming models in all experiments.

The experiments were conducted using the Python 3.11 programming language employing the `pyomo` package, an open-source software package that supports a diverse set of optimization capabilities for formulating, solving, and analyzing optimization models [43, 44].

7.2 Experiments and Discussion

This section explores and discusses the experiments done to validate our new mathematical programming formulation and heuristic to solve the uDGP.

7.2.1 Mathematical programming models

To assess the performance of Models M2 and M2C, both were employed in the realization of Lennard-Jones clusters of varying sizes. To benchmark our findings, we employed method M1R, to solve the same instances. The outcomes are presented in Table 7.1.

Table 7.1: Performance comparison of mathematical programming models in solving Lennard-Jones cluster instances for the uDGP of different sizes and a time limit of 1000 seconds. Ten instances were solved for each size.

		CPU time (s)			
Size, n	Success rate	Average	Standard deviation	Minimum	Maximum
Model M1R					
4	100%	0.0997	0.0016	0.0986	0.102
5	100%	1.56	0.0358	1.53	1.62
7	0%				
Model M2					
4	100%	0.0903	0.0122	0.0717	0.106
5	100%	0.535	0.0128	0.521	0.550
6	100%	8.06	0.477	7.38	8.61
7	0%				
Model M2C					
4	100%	0.0696	0.008 74	0.0608	0.0801
5	100%	0.0832	0.0474	0.795	0.913
6	100%	2.38	0.0952	2.25	2.48
7	100%	325	9.47	309	333
8	0%				
Model M2R					
4	100%	0.223	0.0354	0.184	0.276
5	100%	10.5	0.517	9.81	11.2
7	0%				

For all models, instances of larger size than those shown in Table 7.1 could not be solved within the time limit of 1000 seconds. The short time limit of 1000 seconds was set just to compare the capability of the models in solving small instances. Real applications could be solved with a larger time limit.

Table 7.1 reveals a notable enhancement in performance for Model M2 when solving various instances. Specifically, for instances with 4 atoms, the performance gain is modest. However, for 5 atoms, there is a substantially improved performance, with model M2 solving instances in approximately one-third of the time taken by Model M1R. The performance improvement becomes even more striking when dealing with instances larger than 6 atoms, as Model M2 successfully found a correct solution in just 8 seconds, while the benchmark Model M1R failed to find any solution within 1000 seconds.

Model M2C exhibited superior performance compared to Model M2, highlighting the impact of reducing the number of binary variables on the time needed to encounter a solution. Across all instance sizes, Model M2C consistently achieved faster solution times, even successfully finding solutions for instances with 7 atoms within the specified time limit. Unfortunately, the relaxed Model M2R showed poor performance, taking a much longer time to solve small instances.

Similar experiments were conducted for Lavor instances, and the results can be found in Table 7.2.

Table 7.2: Performance comparison of mathematical programming models in solving Lavor instances for the uDGP of different sizes and a time limit of 1000 seconds. Ten different instances were solved for each size.

		CPU time (s)			
Size, n	Success rate	Average	Standard deviation	Minimum	Maximum
Model M1R					
4	100%	0.335	0.003 74	0.330	0.340
5	100%	47.5	32.2	12.7	81.3
6	0%				
Model M2					
4	100%	0.237	0.0177	0.223	0.267
5	100%	10.8	6.60	2.23	17.6
6	60%	255	124	39.9	490
7	0%				
Model M2C					
4	100%	0.104	0.0218	0.0799	0.139
5	100%	2.12	0.702	0.909	2.61
6	80%	47.2	23.1	17.4	73.8
7	0%				
Model M2R					
4	100%	0.834	0.226	0.609	1.14

The results obtained while evaluating the models' performance in solving Lavor instances closely resembled those found for Lennard-Jones cluster instances. In both cases, both Models M2 and M2C outperformed the benchmark Model M1R, with M2C exhibiting faster performance owing to the reduced number of binary variables. Notably, for instances with smaller sizes, solving a Lavor instance typically required more time compared to a Lennard-Jones cluster of equivalent size. This difference is likely attributed to the distinct nature of the instances in this scenario, and it's worth noting that the standard deviation was significantly higher, underscoring the significant influence of geometry on the models' performance. Here we also see that Model M2R takes longer to find a correct solution, even when compared to the benchmark Model M1R.

7.2.2 Heuristics

We assessed our new Heuristic, NT2, which utilizes Model M2C, against the benchmark NT. As expected, the new heuristic performed faster than the benchmark, as the key distinction lies in the fact that the new heuristic leverages the faster model M2C. For the heuristics, the time limit was set to 5000 seconds. Again, the short time limit was set just to compare the capability of the models. Real applications could be solved with a larger time limit.

Table 7.3 showcases the results of a comparison of the performance in solving Lavor instances of different sizes, confirming reduction in computational times for Model NT2.

Table 7.3: Performance comparison of heuristics in solving Lavor instances for the uDGP of different sizes and a time limit of 5000 seconds. Ten different instances were solved for each size.

		CPU time (s)			
Size, n	Success rate	Average	Standard deviation	Minimum	Maximum
Model NT					
6	100%	23.7	11.8	6.26	42.8
7	100%	39.6	16.0	5.62	69.5
8	100%	57.1	22.2	17.4	84.8
9	100%	139	47.5	59.6	185
10	100%	297	128	53.2	443
20	100%	628	203	252	909
50	60%	2310	651	1380	3590
Model NT2					
6	100%	2.02	0.822	0.945	2.94
7	100%	4.57	2.60	1.71	8.59
8	100%	11.6	5.18	2.21	19.3
9	100%	25.2	18.9	2.90	49.2
10	100%	162	85.8	41.3	308
20	100%	501	269	81.0	900
50	80%	1450	609	339	2180

Similar experiments were conducted for Lennard-Jones cluster instances using Model NT2. The results can be found in Table 7.4.

Table 7.4: Performance comparison of heuristics in solving Lennard-Jones cluster instances for the uDGP of different sizes and a time limit of 5000 seconds. Ten instances were solved for each size.

		CPU time (s)			
Size, n	Success rate	Average	Standard deviation	Minimum	Maximum
Model NT2					
6	100%	0.166	0.137	0.0320	0.448
7	100%	0.681	0.271	0.399	1.41
8	100%	5.90	2.88	2.91	12.9
9	100%	22.1	9.76	1.09	35.8
10	100%	107	74.7	5.42	252
20	100%	523	446	39.6	1110
50	100%	1690	736	523	1950

The solution of the uDGP using the NEW-TRIBOND type heuristics results in a notably high standard deviation, as evident from Tables 7.3 and 7.4. This variance occurs because there is no assurance that the initially encountered core is correct. At each step, if an atom is placed incorrectly, the building-up procedure must reset and search for a new core. Consequently, the same instance can entail a significant or minimal number of resets, resulting in widely varying computational times.

The comparison between Tables 7.3 and 7.4 reveals that, for smaller instances, the time needed to find a solution is shorter for Lennard-Jones instances than for Lavor instances. As the instance size increases, the time difference diminishes. This phenomenon occurs because the time required to find a core in a Lavor instance increases at a slower rate than the time to find a core in a Lennard-Jones cluster instance as the instance size grows, as illustrated in Figure 7.1.

The smaller increase in time required to find a core as the size of the instance increases for Lavor instances is likely attributed to the presence of more recurring patterns in Lavor instances as they expand in size. These recurring patterns can facilitate the search for a core and contribute to the more gradual increase in computational time compared to Lennard-Jones instances.

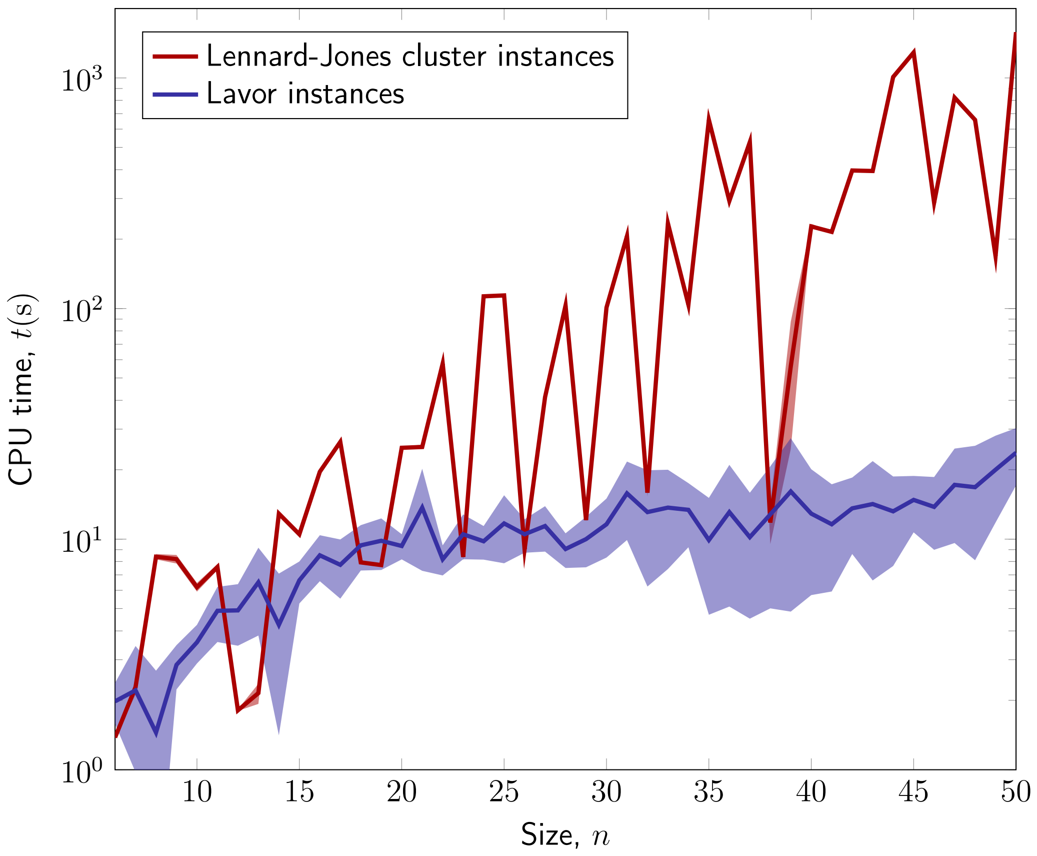


Figure 7.1: Variation of the computational time required to find a five atom core with instance size for Lennard-Jones cluster and Lavor instances.

7.2.3 Heuristics with core guessing

A set of experiments was conducted utilizing the NT2 heuristic, but with a slight change: instead of solving Model M2C to find a core, cores were guessed based on the inherent characteristics of the instance. This approach aimed to explore the potential of applying chemical knowledge to construct structures by using known structural subunits, such as aromatic rings, as building blocks. This method bypasses the need for the most time-consuming step: finding the initial core. This approach could lead to more efficient and faster solutions in cases where domain-specific knowledge can inform the core construction process.

Table 7.5 presents the results of experiments conducted on Lavor instances, where the initial core was guessed instead of being determined through computation.

In these experiments, the initial core was guessed by generating a smaller Lavor instance with 5 atoms. A comparison between Tables 7.5 and 7.3 clearly demonstrates that this approach significantly reduces the time needed to find a solution for the uDGP. This method holds promise because, in many real-world applications, there is often some prior chemical knowledge about the structure based on the nature of the molecule that needs its structure determined. For instance, in the case

Table 7.5: Performance comparison of heuristics in solving Lavor instances for the uDGP of different sizes and a time limit of 1000 seconds, starting from a core guess. Ten instances were solved for each size.

		CPU time (s)			
Size, n	Success rate	Average	Standard deviation	Minimum	Maximum
Model NT2					
6	100%	0.0714	0.0491	0.024	0.138
7	100%	0.203	0.184	0.0450	0.503
8	100%	3.06	2.10	0.749	6.52
9	100%	8.36	6.96	1.36	21.6
10	100%	17.9	10.7	1.78	29.5
20	100%	45.1	16.6	15.1	63.0
50	100%	335	244	98.6	765

of protein structure determination, commonly occurring amino acid structures can serve as initial core guesses, streamlining the subsequent build-up procedure. This approach can be valuable in situations where domain-specific knowledge can inform and accelerate the structure determination process.

7.3 Discussion

As expected, both of our novel models: M2 and its variant M2C, were able to solve the Unassigned Distance Geometry Problem, uDGP and showed good performance when compared to a benchmark model: Model M1R [19].

The new heuristic NT2 also exhibited better performance than the benchmark NT due to its implementation with the faster Model M2C. We also demonstrated the capability of incorporating of chemical knowledge, which can substantially decrease the time required to find the solutions of a Distance Geometry Problem.

Chapter 8

Conclusion

This work presented new mathematical programming models, M2 and its variant M2C, to solve the Unassigned Distance Geometry Problem, uDGP. Both results were compared to a benchmark model: Model M1R [19]. Our results demonstrate that both M2 and M2C outperform M1R across instances that resemble the structure of molecules and nanoparticles, showing the potential of our new models in real applications.

Furthermore, our investigation extends to the development and application of a novel heuristic, NT2, which capitalizes on the faster Model M2C. This new Heuristic also proved to be able to solve uDGP instances in less time than the benchmark method NT [19].

In addition, we demonstrated how the incorporation of chemical knowledge pertaining to the structures under investigation can be integrated into this novel heuristic, resulting in even faster solutions.

As future work, we intend to delve into a detailed examination of the building-up procedure within the heuristics. The objective is to develop strategies that minimize the need to restart the entire process whenever an atom is positioned incorrectly. This will lead to consistent and predictable solution times.

Additional research directions encompass the consideration of errors in experimental distance data. This investigation seeks to enhance the robustness of our reconstruction process by accounting for uncertainties in the input data.

References

[1] ZAHN, R., LIU, A., LÜHRS, T., et al. "NMR solution structure of the human prion protein", *Proceedings of the National Academy of Sciences*, v. 97, n. 1, pp. 145–150, 2000.

[2] SPAGNOLLI, G., RIGOLI, M., ORIOLI, S., et al. "Full atomistic model of prion structure and conversion", *PLoS pathogens*, v. 15, n. 7, pp. e1007864, 2019.

[3] YAN, X., SEDYKH, A., WANG, W., et al. "Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations", *Nature communications*, v. 11, n. 1, pp. 2519, 2020.

[4] HIRSHBERG, M., STOCKLEY, R. W., DODSON, G., et al. "The crystal structure of human rac1, a member of the rho-family complexed with a GTP analogue", *Nature structural biology*, v. 4, n. 2, pp. 147–152, 1997.

[5] JUHÁS, P., CHERBA, D., DUXBURY, P., et al. "Ab initio determination of solid-state nanostructure", *Nature*, v. 440, n. 7084, pp. 655–658, 2006.

[6] HOARE, M., PAL, P. "Physical cluster mechanics: statistical thermodynamics and nucleation theory for monatomic systems", *Advances in Physics*, v. 24, n. 5, pp. 645–678, 1975.

[7] DOYE, J. P., WALES, D. J., MILLER, M. A. "Thermodynamics and the global optimization of Lennard-Jones clusters", *The Journal of Chemical Physics*, v. 109, n. 19, pp. 8143–8153, 1998.

[8] LIBERTI, L., LAVOR, C., MACULAN, N., et al. "Euclidean distance geometry and applications", *SIAM review*, v. 56, n. 1, pp. 3–69, 2014.

[9] LANSBURY, P. T., CAUGHEY, B. "The double life of the prion protein", *Current Biology*, v. 6, n. 8, pp. 914–916, 1996.

[10] LAVOR, C., LIBERTI, L., DONALD, B., et al. "Minimal NMR distance information for rigidity of protein graphs", *Discrete Applied Mathematics*, v. 256, pp. 91–104, 2019.

[11] WOOLFSON, M. M. *An introduction to X-ray crystallography.* Cambridge University Press, 1997.

[12] HARISH, V., TEWARI, D., GAUR, M., et al. "Review on nanoparticles and nanostructured materials: Bioimaging, biosensing, drug delivery, tissue engineering, antimicrobial, and agro-food applications", *Nanomaterials*, v. 12, n. 3, pp. 457, 2022.

[13] BILLINGE, S. J., DUXBURY, P. M., GONÇALVES, D. S., et al. "Assigned and unassigned distance geometry: applications to biological molecules and nanostructures", *4OR*, v. 14, pp. 337–376, 2016.

[14] BILLINGE, S. J., DUXBURY, P. M., GONÇALVES, D. S., et al. "Recent results on assigned and unassigned distance geometry with applications to protein molecules and nanostructures", *Annals of Operations Research*, v. 271, pp. 161–203, 2018.

[15] BILLINGE, S. J., LEVIN, I. "The problem with determining atomic structure at the nanoscale", *science*, v. 316, n. 5824, pp. 561–565, 2007.

[16] MENGER, K. "Untersuchungen über allgemeine Metrik", *Mathematische Annalen*, v. 100, n. 1, pp. 75–163, 1928.

[17] BLUMENTHAL, L. M. "Theory and applications of distance geometry", *Chelsea Publishing Company*, 1970.

[18] DUXBURY, P. M., GRANLUND, L., GUJARATHI, S., et al. "The unassigned distance geometry problem", *Discrete Applied Mathematics*, v. 204, pp. 117–132, 2016.

[19] DUXBURY, P., LAVOR, C., LIBERTI, L., et al. "Unassigned distance geometry and molecular conformation problems", *Journal of Global Optimization*, pp. 1–10, 2022.

[20] SAXE, J. B. "Embeddability of weighted graphs in k-space is strongly NP-hard". In: *17th Allerton Conf. Commun. Control Comput., 1979*, pp. 480–489, 1979.

[21] ALMEIDA, F. C., MORAES, A. H., GOMES-NETO, F. "An overview on protein structure determination by NMR: historical and future perspectives of the use of distance geometry methods", *Distance Geometry: Theory, Methods, and Applications*, pp. 377–412, 2012.

[22] WUETHRICH, K. "The development of nuclear magnetic resonance spectroscopy as a technique for protein structure determination", *Accounts of chemical research*, v. 22, n. 1, pp. 36–44, 1989.

[23] HENDRICKSON, B. "The molecule problem: Exploiting structure in global optimization", *SIAM Journal on Optimization*, v. 5, n. 4, pp. 835–857, 1995.

[24] NILGES, M., O'DONOGHUE, S. I. "Ambiguous NOEs and automated NOE assignment", *Progress in nuclear magnetic resonance spectroscopy*, v. 32, n. 2, pp. 107–139, 1998.

[25] GUERRY, P., HERRMANN, T. "Advances in automated NMR protein structure determination", *Quarterly reviews of biophysics*, v. 44, n. 3, pp. 257–309, 2011.

[26] BOUCHEVREAU, B., MARTINEAU, C., MELLOT-DRAZNIEKS, C., et al. "An NMR-Driven Crystallography Strategy to Overcome the Computability Limit of Powder Structure Determination: A Layered Aluminophosphate Case", *Chemistry–A European Journal*, v. 19, n. 16, pp. 5009–5013, 2013.

[27] SLICHTER, C. P. *Principles of magnetic resonance.* Springer Series in Solid-State Sciences. 2 ed. , Springer, sep 1980.

[28] EGAMI, T., BILLINGE, S. J. *Underneath the Bragg peaks: structural analysis of complex materials.* Elsevier, 2003.

[29] GUJARATHI, S., FARROW, C., GLOSSER, C., et al. "Ab-initio reconstruction of complex Euclidean networks in two dimensions", *Physical Review E*, v. 89, n. 5, pp. 053311, 2014.

[30] BENDSOE, M. P., SIGMUND, O. *Topology optimization: theory, methods, and applications.* Springer Science & Business Media, 2003.

[31] MARTINEZ, J. "A note on the theoretical convergence properties of the SIMP method", *Structural and Multidisciplinary Optimization*, v. 29, pp. 319–323, 2005.

[32] MALLIAVIN, T. E., MUCHERINO, A., LAVOR, C., et al. "Systematic exploration of protein conformational space using a distance geometry approach", *Journal of Chemical Information and Modeling*, v. 59, n. 10, pp. 4486–4503, 2019.

[33] FARGES, J., DE FERAUDY, M., RAOULT, B., et al. "Noncrystalline structure of argon clusters. II. Multilayer icosahedral structure of Ar N clusters $50 < N < 750$", *The Journal of chemical physics*, v. 84, n. 6, pp. 3491–3501, 1986.

[34] ECHT, O., SATTLER, K., RECKNAGEL, E. "Magic numbers for sphere packings: experimental verification in free xenon clusters", *Physical Review Letters*, v. 47, n. 16, pp. 1121, 1981.

[35] HARRIS, I., KIDWELL, R., NORTHBY, J. "Structure of charged argon clusters formed in a free jet expansion", *Physical review letters*, v. 53, n. 25, pp. 2390, 1984.

[36] DAVEN, D., TIT, N., MORRIS, J., et al. "Structural optimization of Lennard-Jones clusters by a genetic algorithm", *Chemical physics letters*, v. 256, n. 1-2, pp. 195–200, 1996.

[37] WALES, D. J., SCHERAGA, H. A. "Global optimization of clusters, crystals, and biomolecules", *Science*, v. 285, n. 5432, pp. 1368–1372, 1999.

[38] CAI, W., SHAO, X. "A fast annealing evolutionary algorithm for global optimization", *Journal of computational chemistry*, v. 23, n. 4, pp. 427–435, 2002.

[39] NORTHBY, J. "Structure and binding of Lennard-Jones clusters: $13 \leq N \leq 147$", *The Journal of chemical physics*, v. 87, n. 10, pp. 6166–6177, 1987.

[40] LAVOR, C. "On generating instances for the molecular distance geometry problem", *Global Optimization: from Theory to Implementation*, pp. 405–414, 2006.

[41] PHILLIPS, A. T., ROSEN, J. B., WALKE, V. H. "Molecular structure determination by convex, global underestimation of local energy minima." *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*, v. 23, pp. 181–198, 1995.

[42] GUROBI OPTIMIZATION, LLC. "Gurobi Optimizer Reference Manual". 2023. Disponível em: <`https://www.gurobi.com`>.

[43] HART, W. E., WATSON, J.-P., WOODRUFF, D. L. "Pyomo: modeling and solving mathematical programs in Python", *Mathematical Programming Computation*, v. 3, n. 3, pp. 219–260, 2011.

[44] BYNUM, M. L., HACKEBEIL, G. A., HART, W. E., et al. *Pyomo–optimization modeling in python*, v. 67. Third ed. , Springer Science & Business Media, 2021.