MARCOS FELIPE MARTINS SILVA

GAMING AND DATA MINING IN THE LEARNING AND EVALUATION OF CONTACTS IN BIOLOGICAL COMPLEXES

Belo Horizonte

Janeiro de 2018

MARCOS FELIPE MARTINS SILVA

GAMING AND DATA MINING IN THE LEARNING AND EVALUATION OF CONTACTS IN BIOLOGICAL COMPLEXES

Dissertação apresentada ao Programa de Pós-Graduação em computer science do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em computer science.

Orientadora: Raquel Cardoso de Melo Minardi
Coorientadora: Cristiane Neri Nobre

Belo Horizonte

Janeiro de 2018

MARCOS FELIPE MARTINS SILVA

GAMING AND DATA MINING IN THE LEARNING AND EVALUATION OF CONTACTS IN BIOLOGICAL COMPLEXES

Dissertation presented to the Graduate Program in computer science of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Master in computer science.

Advisor: Raquel Cardoso de Melo Minardi
Co-Advisor: Cristiane Neri Nobre

Belo Horizonte

January 2018

Ficha catalográfica elaborada pela Biblioteca do ICEx - UFMG

Silva, Marcos Felipe Martins

S586g Gaming and data mining in the learning and: evaluation of contacts in biological complexes / Marcos Felipe Martins Silva. — Belo Horizonte, 2018.
xviii, 82 f. : il. ; 29 cm.

Dissertação (mestrado) - Universidade Federal de Minas Gerais – Departamento de Ciência da Computação.

Orientadora: Raquel Cardoso de Melo Minardi
Coorientadora: Cristiane Neri Nobre

1. Computação – Teses. 2. Mineração de dados (Computação) 3. Jogos eletrônicos. 4. Forças intermoleculares. 5. Bioinformática. I. Orientadora. II. Coorientadora III. Título.

CDU 519.6*72 (043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Gaming and data mining in the learning and evaluation of contacts in biological complexes

MARCOS FELIPE MARTINS SILVA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROFA. RAQUEL CARDOSO DE MELO MINARDI - Orientadora
Departamento de Ciência da Computação - UFMG

PROFA. CRISTIANE NERI NOBRE - Coorientadora
Departamento de Ciência da Computação - PUC/MG

PROFA. GISELE LOBO PAPPA
Departamento de Ciência da Computação - UFMG

PROF. LUCAS BLEICHER
Instituto de Ciências Biológicas - UFMG

Belo Horizonte, 23 de Janeiro de 2018.

Acknowledgments

I would like to thank my advisor Raquel Minardi and Co-advisor Cristiane Nobre, for continually and convincingly conducted a spirit of adventure respecting research and scholarship. Also, for the excitement regarding academic life. Without their guidance and persistent advice, this dissertation would not have been possible.

I also would like to thank all my friends from LBS for all the essential help. Lucianna for revising the present dissertation and Alexandre for the acquired knowledge concerning the biochemistry field. Among these friends, I give special thanks to Larissa and Mivian for all moments we have shared together during the last seven years.

I must thank my entire family for understanding my moments of absence at many times of fellowship. Especially my grandmother, for always giving me food! As a member of this family, I include Fred Dutra, for always supporting me in the most challenging moments faced during the academic life.

Thanks also to CAPES for the financial support to carry out the research presented here.

Last, but not least, I would like to give many thanks to God for putting the people mentioned above in my life.

"Quando eu disse ao caroço de laranja que dentro dele dormia um laranjal inteirinho, ele me olhou estupidamente incrédulo.
When I told the orange lump that an entire orange grove was sleeping inside it, it looked at me stupidly incredulous."
(Hermógenes)

Abstract

Molecules of life (carbohydrates, lipids, nucleic acids, and proteins) can change their conformation in the space frequently due to non-covalent bonds (or contacts) between their molecules. The comprehension of patterns created through the contacts of atoms and amino acids has been used as an aid in solving a range of problems in bioinformatics such as protein conformation, the functional similarities between proteins, structural alignment, thermodynamic stability prediction, prediction of protein structures, drug design, and so forth. Several paradigms of calculation of contacts have been developed. However, all models present advantages and disadvantages. The model based only on distance present the problem of not guaranteeing that just the first layer of neighbor atoms are connected by edges, and occlusions may occur (when there is an intervening atom in-between). Concerning geometry, some techniques determine contacts in proteins through Delaunay and Voronoi tesselations. These methods present an issue where two atoms could establish an interaction, but the unique triangulation did not identify them as connected. Besides, the state-of-the-art lacks visualization techniques of non-covalent interactions, what would highly help in the identification of occlusions. An efficient manner to make the visualization of non-covalent interaction is through the use of games due to its benefits such as pleasure, stimulation, creativity, and enthusiasm. Also, several games in education have been an aid of understanding open problems in biochemistry. Hence, the present dissertation proposes a methodology for the construction and evaluation of a manually curated database for the classification of contacts between atoms of amino acids in protein chains through the use of digital games. To evaluate the built database with the state-of-the-art, data mining methods were used to find patterns in the classification of contacts. We also counted on the participation of inexperienced players in biochemistry, and we have significant evidence that they learned about molecular non-covalent bonds as they played.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

When two molecules are close to each other, two phenomena are most likely to happen: 1) they may react or 2) they may interact. A chemical reaction requires the rupture of chemical bonds and/or the formation of other bonds. A chemical interaction means that the molecules attract or repel each other, without breaking or forming new chemical bonds. Such interactions are often called non-covalent interactions, or contacts. The energies involved in such types of interactions are often much smaller than covalent bonds (when there is the sharing of electrons). A contact differs from a covalent bond in that it occurs when two molecules are close, but it does not involve the sharing of electrons. According to Berg and collaborators [1], "non-covalent molecular interactions are the essence of the dance of life." The reason for such affirmation is that molecules of life (proteins, carbohydrates, lipids, and nucleic acids) can change their conformation in the space frequently due to non-covalent bonds between their molecules.

The technological advance has made it possible to perform experiments promptly and with reasonable cost, substantially due to the development of increasingly efficient algorithms, as well as the increase of computational resources. This advance was of great importance for the growth of biology, especially with the creation of the genome project whose objective was to determine the DNA sequence of the entire euchromatic human genome [2]. Consequently, biological data of various molecules are available in public databases on the Web. However, such data are little explored if compared to their quantity. The advance of technology, combined with the advancement of biology, created a new area of knowledge: bioinformatics; whose purpose is to develop models and algorithms for understanding biological data [3].

Concerning the data from proteins, we highlight the Protein Data Bank (PDB) which is a database that stores data of proteins and nucleic acids (in the case of DNA-

protein or RNA-protein interactions) [4]. As a PDB file shows the three-dimensional structure of proteins as x, y and z coordinates of the amino acids; a pair of adjacent residues is considered to be in contact if the distance between their specific atoms is less than a distance threshold. Thus, the comprehension of patterns created through the contact of atoms and residues has been used as an aid in solving a range of problems in bioinformatics. For instance: protein conformation [5], the functional similarities between proteins [6], structural alignment [7], thermodynamic stability prediction [8], prediction of protein structures [9], drug design, among other contributions. Therefore, the understanding of such intermolecular forces is of high relevance if we are to understand the role of chemical systems at the molecular level.

Several paradigms of calculation of contacts have been developed. The calculations may be based on the Euclidian distance of atoms, geometry, angle constraints, or a combination of them [10]. All models present advantages and disadvantages. The model based only on distance presents the problem of establishing a high number of contacts, which may also be incorrect. Furthermore, it is not guaranteed that just the first layer of neighbor atoms are connected by edges, and occlusions may occur. Two atoms are in contact if there is no intervening atom in-between (i.e., if they are not occluded by others). Regarding geometry, some techniques determine contacts in proteins through Delaunay and Voronoi tesselations [11, 12]. These methods present an issue when two atoms could establish an interaction, but the unique triangulation did not identify them as connected [10]. There are also algorithms that, besides distance, still consider the angle formed between the atoms of the interaction. Of which we highlight the Piccolo database [13].

Piccolo database is the state-of-the-art of identified protein-protein interactions database [13]. It counts with more than 260 million interacting atom pairs from more than 38,000 protein complexes. 12 chemical contacts are characterized by the authors in this study, from those we highlight Hydrogen bonds (HB), Hydrophobic contacts (HP), ionic (IO), and aromatic stacking contacts (AR). However, as the methodology used in [13] lacks visualization of the contacts, it does not verify cases of occlusion. In addition to defining the types of contacts, the literature lacks a characterization of what is not an interaction. Given the problems discussed above, a manually curated database that explores the visualization of non-covalent bonds would be an alternative method in the classification of contacts. Moreover, such a technique could be used to verify the correctness of the interactions that are said to be proximal in the Piccolo database [13]. In the present study, we classify these proximal interactions as non-contact, since these "contacts" do not follow the criteria of HB, HP, IO, and AR contacts.

An efficient manner to make the visualization of non-covalent interaction is

through the use of games. In fact, the literature has shown that digital game-based learning (DGBL) have had great success in aiding the resolution of open problems in biochemistry through crowdsourcing, such as [14] [15]. What differs a DGBL from an ordinary game, is its educational purpose. Furthermore, many benefits have been stated - such as pleasure, motivation, stimulation, creativity, enthusiasm and so forth - in the use of games in the most diverse areas of knowledge [16, 17, 18].

1.1 Problem description

The lack of a visualization of contacts already characterized in the literature makes it difficult to classify these contacts correctly. Furthermore, the literature so far has not provided a basis for non-covalent interactions between residue atoms in protein chains that had been evaluated through crowdsourcing. This investigation supports the formulation of the problem that this study aims to clarify.

It is possible to formulate the problem presented in this dissertation through the following questions: 1) Is it possible to construct a manually curated database of non-covalent interactions related to hydrogen bonds, hydrophobic, ionic and aromatic interactions? 2) Is it possible to characterize the non-contact, besides the contacts already foreseen in the literature?

1.2 General objective

The present dissertation proposes a methodology for the construction and evaluation of a manually curated database for the classification of contacts between atoms of amino acids in protein chains. Four distinct interactions are here considered: HB, HP, IO, and AR. Harnessing the advantages shown by the development of Digital Game-Based Learning (DGBL) as one of the proposes to support the resolution of problems in science, we present the conception of Proteingo[1]. Proteingo is a game developed to build a crowdsourced database of protein-protein contacts. We also counted on the participation of inexperienced players in biochemistry, and we have significant evidence that they learned about molecular non-covalent bonds as they played.

[1]Available on: `http://bioinfo.dcc.ufmg.br/proteingo/`

1.3 Specific goals

In order to fulfill the proposed general objective, this work had the following specific objectives:

- To develop an online game to automate and make the process of user contact curation more enjoyable.
- To evaluate the proposed game according to the criteria of motivation, user experience, and learning.
- To characterize the non-covalent and non-contact interactions according to the concepts presented by PICCOLO.
- To characterize non-covalent and non-contact interactions according to users' interactions from the developed game.
- To identify which features are most relevant for the classification of contacts in molecular networks through data mining procedures.
- To carry out a comparative analysis of the models generated by PICCOLO and users to evaluate the quality of the base that was manually curated.

1.4 Relevance of the research

Understanding the classification of contacts between atoms of residues in different chains of proteins can contribute to the several areas of knowledge (for instance, biology and computing). As mentioned earlier, non-covalent bonds are strictly and commonly involved in folding and stabilizing protein structures, but the principles governing protein contacts are not fully understood [19]. For instance, protein-protein interactions are being increasingly examined as a potential help for drug-design techniques. Drug-design is the inventive process of finding new medications based on the knowledge of a biological target (such as proteins). Frequently, drug design relies on computer modeling techniques, from which we highlight machine learning (ML) methods. ML is more efficient, allowing larger-scale predictions than other biology laboratory techniques (such as docking), and thus examining a more significant number of promising candidates for further experimental screening [20].

Therefore, investigating computational methods capable of providing a better visualization of the non-bonded interactions, can make visible the unobservable scientific phenomena. The visualization methods united to data mining techniques for

evaluating a manually curated database of non-bonded interactions would contribute to responding many biology issues concerning molecular systems. Regarding the area of computer science, it could use the models and algorithms discussed here to solve classification problems in general. In this way, a serious game (SG) that unites visual information of contacts, with a data mining technique to evaluate the created database, is the primary justification that motivates the development of this research.

1.5 Contributions

This work contributes to scientific research in the following aspects:

- Offering of a method for automatic data collection of users through games.
- Evaluating the learning of players throughout the plays.
- Providing a manually cured database capable of hitting contacts with a high-quality rating.
- Providing a methodology capable of classifying non-contacts.

1.6 Work overview

This dissertation is divided into the following structures.

Chapter 2 presents the background to understand the central ideas in the current study. The concepts of proteins and the non-covalent bonds (hydrogen bond, hydrophobic, ionic and aromatic interactions) are described. Also, the data mining techniques used as well as the main work related to learning through the use of DGBLs are discussed. This chapter further presents several works developed as a possible solution of problems in the area of science.

Chapter 3 shows the methodology suggested to construct and evaluate the database of interactions created through Proteingo game.

Chapter 4 reports and discusses the primary outcomes of the present dissertation. First, the results regarding the players learning are displayed. Then, the evaluation of the proposed database is analyzed and discussed.

Chapter 5 concludes the dissertation and suggest some future works, either in the learning and database evaluation experiments.

Chapter 2

Background Concepts and Related Works

Here we describe all the central concepts needed to understand our work. First, Section Proteins and Molecular Interactions defines the formation of proteins and further information about the non-covalent bonds analyzed in this study. Also, the definition of Digital Game-Based Learning (DGBL) is presented in Section 2.2. We finalize this chapter by talking about the related works which developed games in the biology and chemistry fields to analyze users' learning and DGBLs as an aid to deal with open problems as well.

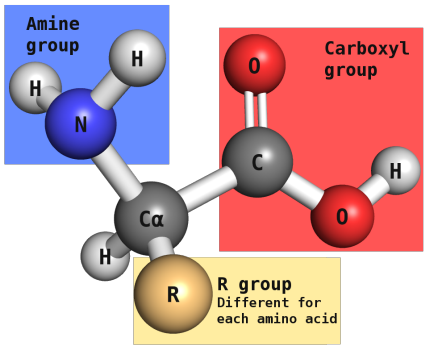
2.1 Proteins and Molecular Interactions

Proteins are responsible for controlling almost all cellular processes. They are one of the most abundant biological macromolecules, being present in all cells. Besides, proteins are the molecular instruments by which genetic information is expressed. Proteins are polymers formed by one or more chains of amino acids which connect with each other through peptide bonds. The rearrangement of these amino acids can generate diverse types of proteins such as enzymes (e.g. Hydrolases catalyze the hydrolysis of several bonds); hormones (e.g. Insulin controls the rate of blood glucose); antibodies (e.g. Immunoglobulin attacks to foreign substances in the body's immune system.); transporters (e.g. Hemoglobin is responsible for the transport of oxygen); and a range of other substances with distinct biological activities [21].

The arrangement of 20 different types of amino acids (the most commonly found in living beings) happens through covalent bonds in the formation of proteins. The general structure of the amino acids involves an amine group (NH2) and a carboxyl

group (COOH), both attached to the α carbon atom. The α carbon is also connected to a hydrogen and a side chain (known as R group), which determines the identity of a particular residue (amino acids in polypeptide chains are called residues). Figure 2.1 exhibits the generic structure of an amino acid, and Table 2.1 shows the 20 types of amino acids with their key features.

Figure 2.1: **Amino acid general structure**



Source: Author's elaboration.

2.1.1 Hydrogen Bond (HB)

A hydrogen bond is a weak sort of bond between two polar groups (dipole-dipole attraction). For a hydrogen bond to happen, a hydrogen donor and an acceptor must be present. The donor is the atom to which the hydrogen atom is covalently connected, and is frequently a strongly electronegative atom such as nitrogen (N) and oxygen (O). The adjacent electronegative atom which has a lone electron pair is referred to as the acceptor in the hydrogen bond interaction [23].

Piccolo database [13] considers the algorithm developed by McDonald and Thornton [24] to calculate HB interactions. This algorithm first places the missing hydrogen atoms. Then, a query to the table, generated by the algorithm, tells whether a hydrogen bond is valid based on the distance between the atoms and their respective types (donor and acceptor) [13]. Figure 2.2 (a) shows an example of a hydrogen bond present in the Proteingo game.

2.1.2 Hydrophobic Interaction (HP)

The hydrophobic effect ("water-fearing") is the recognized propensity of nonpolar substances to accrete in aqueous solution and reject water molecules [21]. Hydrophobic interactions are important in keeping a protein stable and biologically active because

Table 2.1: **Standard amino acids and properties.** The charge of the amino acids is based on a 7.4 pH solution

Amino acid	Abreviation	Side Chain Polarity	Charge
Alanine	Ala	Nonpolar	Neutral
Arginine	Arg	Polar	Positive
Asparagine	Asn	Polar	Neutral
Aspartic acid	Asp	Polar	Negative
Cysteine	Cys	Nonpolar	Neutral
Glutamic acid	Glu	Polar	Negative
Glutamine	Gln	Polar	Neutral
Glycine	Gly	Nonpolar	Neutral
Histidine	His	Polar	Positive
Isoleucine	Ile	Nonpolar	Neutral
Leucine	Leu	Nonpolar	Neutral
Lysine	Lys	Polar	Positive
Methionine	Met	Nonpolar	Neutral
Phenylalanine	Phe	Nonpolar	Neutral
Proline	Pro	Nonpolar	Neutral
Serine	Ser	Polar	Neutral
Threonine	Thr	Polar	Neutral
Tryptophan	Trp	Nonpolar	Neutral
Tyrosine	Tyr	Polar	Neutral
Valine	Val	Nonpolar	Neutral

Fonte: Elaborado pelo autor. Adaptado de Cooper and Hausman [22].

it gives the protein the ability to reduce the surface area, and diminish the undesirable interactions with the aqueous environment. The authors of Piccolo database [13] consider as hydrophobic interactions all of those where both atoms are labeled as hydrophobic, according to a pre-established table, and the inter-atomic distance is less than 5Å. Figure 2.2 (b) illustrates a hydrophobic interaction.

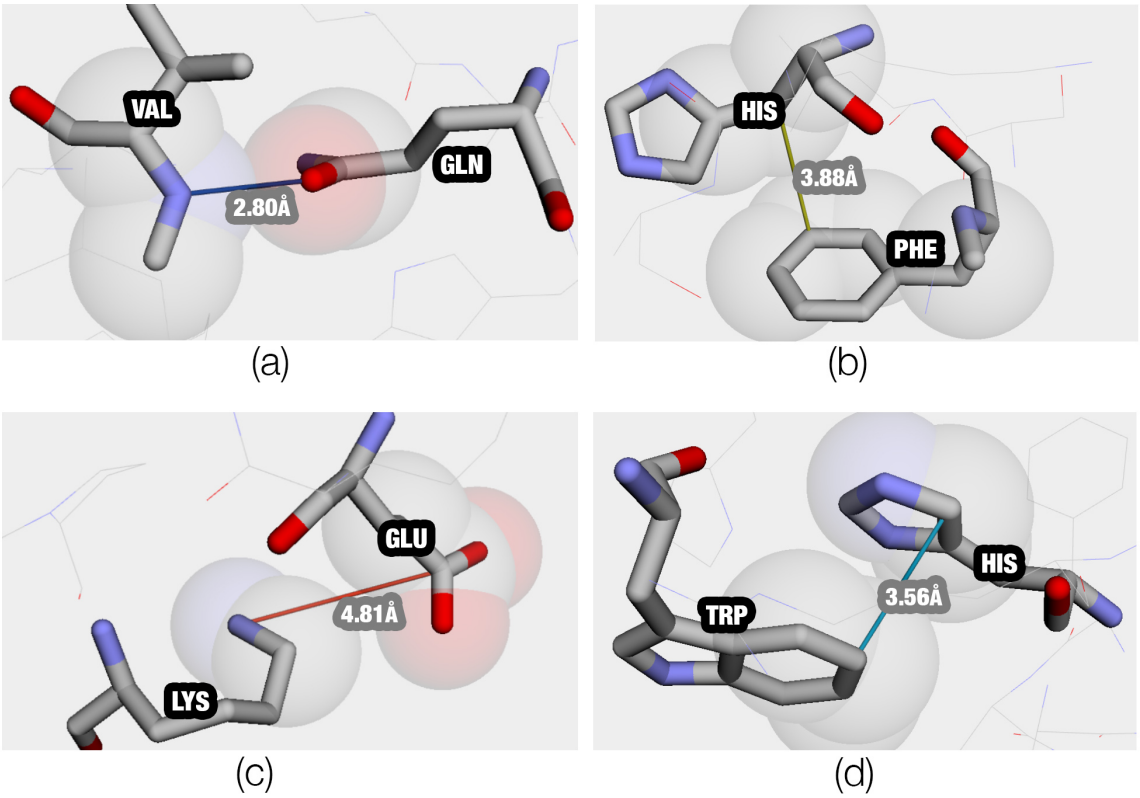
2.1.3 Ionic Interaction (IO)

Proteins have charged groups that can unite to each other or other biomolecules. Ionic interactions comprise the attraction of ions or molecules with long-lasting charges of opposing signals. For example, the carboxyl groups of the Aspartic and Glutamic residues (negatively charged) can be attracted to the free amino groups of residues Lysine and Arginine (positively charged). Figure 2.2 (c) shows an example of ionic interactions considered in the proteingo game.

Ionic interactions with charged groups that are typically close to the linear se-

quence of the polypeptide chain may take part in folding steps. Also, ionic interactions are among the forces that sustain the protein molecule and enable an enzyme to identify and connect to its substrate. The adequate manner of calculating interactions would be to use quantum chemical approaches to solving the Coulomb equation separately for each nucleus. However, this is impractical for large biological systems. A more straightforward way is to consider the formal charges on the protein alone (whether an electron was lost or gained), which was adopted by Bickerton et. al. [13].

Figure 2.2: **Examples of noncovalent interactions present in the game.** (a) **Hydrogen Bond:** the nitrogen of valine (blue stick) and the oxygen of glutamine (red stick) are in contact through a hydrogen bond. (b) **Hydrophobic Interaction:** two nonpolar carbons (of histidine and phenylalanine, respectively) are in hydrophobic interaction. (c) **Ionic Interaction:** a nitrogen of lysine (blue stick) is ionically interacting with a negatively charged carbon of glutamic acid. (d) **Aromatic Stacking Interaction:** the tryptophan and histidine aromatic rings are in a "face-to-face" stack interaction.



Source: Author's elaboration.

2.1.4 Aromatic Stacking Interaction (AR)

Histidine, Phenylalanine, Tryptophan, and Tyrosine have rings which can form aromatic stacking interactions. The aromatic rings have double bonds in resonance, acquired by the cyclical movements that the electrons make in the orbitals. In the orbital within the plane of the ring, electrons create partially positive charges in the plane of the ring and the electrons in the π orbitals, located in the upper and lower part of the plane of the ring, develop a partially negative charge in these regions. The presence of these partial charges allows aromatic rings to associate with each other. This phenomenon is called aromatic stacking and is illustrated in Figure 2.2 (d).

Aromatic stacking interactions play an essential part in defining the structure, property, and activity of many complexes [25]. Aromatic interactions are determined when two criteria are satisfied. Firstly, when a pair of aromatic atoms is within the suitable distance threshold, then the centroids of the two parent planar ring arrangements are calculated. If the centroids are within the threshold distance as well, the contact is considered aromatic. Schreye and collaborators [26] divided the procedure to subclassify aromatic contacts as being "face-to-face", "edge-to-face" or "displaced edge to face". Thus, the same method was used in Bickerton et. al. [13].

2.2 Data Mining

For a better understanding of Data Mining, one must know the context where it is embedded in the procedure of Knowledge Discovery in Databases (KDD). KDD, according to Fayyad and collaborators [27], can be divided into five steps: data selection, data preprocessing, data transformation, data mining, and evaluation.

Data selection and preprocessing: The data selection phase is where specific items in a database are selected. At this stage the data set, belonging to a domain, containing all possible variables (also called features or attributes) and records (also called cases or observations) that will be part of the analysis. Preprocessing is required for data cleaning. For instance, in the selection of amino acid data from PDB (Protein Data Bank) protein chains, the retrieved data comes with information that is not relevant to the analysis (such as the header of the files and details of the experiments). Therefore, cleaning is necessary to maintain only what is relevant [27].

Data transformation: After being selected, cleaned and preprocessed, the data needs to be stored appropriately and formatted so that the algorithms can be

applied. Furthermore, at this stage, if necessary, it is possible to obtain missing data through the transformation or combination of others, or just deleting them [27].

Data mining: All stages of the process are essential to the success of the process. However, it is the Data Mining (DM) stage that receives the most prominent step in the literature. The two primary aims of DM in practice tend to be prediction and description [27]. Prediction is a statement about an unknown event that might happen based on attributes in the database. Description is related to finding patterns in the data that are human-interpretable. To achieve the prediction and description requests, several DM tasks can be used. Of which we highlight classification and regression. Classification is a process of categorizing through a learning function. Classification is considered an instance of supervised learning (a task of inferring a function from labeled training data) and several algorithms are listed in the literature [28]: Support Vector Machines, Naive Bayes, Decision Trees (DT), Neural Networks, and so forth. DTs generates trees which can be read as rules. Such rules have a simple representational form, which makes the inferred model easy to understand [29], and so was used in the present study. Regression is a set of statistical processes for estimating the relationships among variables. Several regression techniques can be used depending on the data to be analyzed: Linear Regression, Logistic Regression, Binomial Logistic Regression, Multinomial Logistic Regression (MLR), and so on. We highlight the use of MLR in the present study as a technique for feature selection. Although several techniques are used for feature selection approach (ranking, greedy algorithms, genetic algorithms, and so forth.) [30], MLR has had excellent results in the field. In addition to verifying correlations between variables, MLR still verifies which attributes ought to be used to predict a nominal output [31].

Evaluation: DM brings with it a series of ideas and techniques for a wide variety of fields. Statisticians, Artificial Intelligence (IA) researchers, and database administrators use different techniques to interpret and evaluate the results obtained with mining to reach the same end: information [27, 28, 30, 31].

2.2.1 Decision Trees

C4.5 is an algorithm used to create a decision tree developed by Ross Quinlan [32]. The C4.5 tree attempts to recursively split the data set into subsets by evaluating the Information Gain (IG). The IG is the difference in entropy that results from choosing

a feature for splitting the data. The variable with the highest IG is used in every step. The training process stops when the resulting nodes contain instances of a single classe or if no feature can be found to calculate the information gain. Algorithm 1 shows the pseudocode of the C4.5 algorithm, and Equations 2.1 and 2.2 show the entropy and IG calculation, respectively.

Algorithm 1: General algorithm for building decision trees

```
C4.5 Algorithm ();
Input  : set of attributes
Output: Decision tree
if all samples of the set belong to the same category then
    create a leaf node for the decision tree and chooses the actual category
else
    if none of the features provides any IG or
    previously unseen instances then
        create a decision node above using the expected value
    end
end
for each feature f do
    find the IG ratio from splitting on f
    f_best = Best feature with the highest IG
    for each subset obtained by splitting on f_best do
        add those nodes as children of the node
    end
end
```

$$Entropy\ h(f) = -\sum p(f)\log p(f) \tag{2.1}$$

$$IG\ I(f, f\_best) = h(f) - h(f|f\_best) \tag{2.2}$$

where $h(f)$ represents the entropy before the general division of the system (before division), and $h(f|f\_best)$ the entropy of the system since the division was performed in f_best.

The DT present in WEKA environment [33] is called J48, which is an open source Java implementation of the C4.5 algorithm. The user is supposed to set the pruning confidence parameter (C) and the minimum number of instances per leaf (M). Knowing that pruning is a technique to prevent overfitting, the lower its value, the more general the model is. Weka has a default of 0.25 and 2 for C and M parameters, respectively.

2.2.2 Multinomial Logistic Regression (MLR)

The multinomial logistic regression (MLR) is applied to predict a nominal dependent variable with one or more independent variables. It is considered an extension of the binomial logistic regression (LR) to enable a dependent variable with more than two levels. As with other types of regression, MLR may have independent continuous and discrete variables and may also hold interactions between independent variables to predict the dependent variable. It is essential to keep in mind that what is wanted is to verify how much the independent variables explain the dependent variable, that is, whether to measure its power of influence on the dependent variable. As in a LR the variable to be predicted is dichotomous (0 or 1), LR estimates the posterior probabilities by using the formula shown in Equation 2.3 as seen in [34].

$$P(Y_i = 1|P_i) = \frac{\exp(\alpha + \sum_{j=1}^{p} \beta_j p_{i_j})}{1 + \exp(\alpha + \sum_{j=1}^{p} \beta_j p_{i_j})} \tag{2.3}$$

where $P(Y_i = 1|P_i)$ is the posterior probability of the $i-th$ instance be 1 given P_i. $P_i = (p_{i_1}, p_{i_2}, \cdots, p_{i_p})$ is a probability vector for the $i-th$ instance, of which each component p_{i_j} is the probability estimate of the component predictor j on the $i-th$ residue pair. According to a Quasi-Newton optimization, the constants α and $\beta j (j = 1, 2, \cdots, p)$ are the coefficients of the regression whose values can be estimated with the given training set.

For an MLR, the equations generate probabilities to predict whether a category is above/below the reference category (known as a dummy variable) and $k-1$ equations are generated, where k is the number of categories. Several statistical programs have automatic statistical tests for the significance of the variables given a CI confidence interval (usually $CI = 95\%$).

2.3 Digital Game Based Learning (DGBL)

Several authors define Digital Game-Based Learning (DGBL), serious games or games for education as being an environment in which the acquisition of knowledge and skills are derived from the improvement of game content as well as its gameplay [35, 36, 37, 38]. It can be achieved through an interaction based on a set of rules and directed towards a clear objective (challenge) [39, 40]. Also, DGBLs resemble Problem Based Learning, where scenarios of specific problems (challenges) are placed within a framework of the play. As a consequence, learners reach a sense of accomplishment [41, 42]. The feeling of accomplishment can be achieved through feedback, such as a

score; allowing players to know their progress during the game [40]. It is important to emphasize that we are considering as DGBLs all digital games that can be played through technological platforms such as computers/laptops, cell phones, and tablets. DGBLs must have been developed with the purpose of aiding the growth and better means of propagating knowledge in fields of education.

Santos and colleagues [43] state that educational games offer attractive activities, which increase players' interest in the content. In addition to generating interest in the subject, science-related games also support the beliefs of students in their own abilities [44]. This feature is fundamental when it is stated that knowledge must be built to be properly assimilated [42]. Based on this, gamification becomes increasingly common in the learning of students [45, 46]. Moreover, in the review of Vandercruysse and collaborators [18], the authors investigated the major learning effects of serious games, and under which conditions a game is useful for learning. Elements such as enjoyability, curiosity, fantasy, challenge, motivation, interactivity, and competition are used to engage players in learning through games. These components match those proposed by Lepper and Malone [39] in the investigation of the correlation between motivation and learning, when in the use of games. In this manner, it seems that these principles are fundamental in the development of DGBLs, providing students with a more pleasant learning alternative.

2.3.1 DGBLs to Support Science Education

Due to the biochemical content focus of the DGBL explored in the present work, we point out findings that have surveyed the use of games to support the learning of biological and chemical subjects. As DGBLs requires some technological environment/device, studies which involve the use of board games - as "Discovering the Cell" in [47] and "The Recal Case" in [48] were not taken into account in this section.

Annetta, Cheng, and Holmes [49] investigated the use of a Multiplayer Educational Gaming Application (MEGA) for teaching genetics to biology students in high school. The results of the referred study indicated that students were deeply engaged in playing MEGA. However, statistical results stated that the understanding of genetics was not reached by the students who played the game when compared to peers involved in more traditional instruction. The authors still emphasized that further studies should be conducted to investigate whether innovative technologies captivate students in learning more than traditional print materials.

In [50] the authors developed "Chairs!" a game-based mobile application for teaching and learning organic chemistry. In this study, the authors carried out an

experiment with high-school and college students, and the results were optimistic regarding student engagement, motivation, and learning outcomes. With the high-school students, the concept of cyclohexane was taught through traditional lecture, and the students were divided into two clusters: some of them played "Chairs!" and the remaining did not. The students were then submitted to a 4-point drawing quiz, and the first group (who played the game) outperformed the second one in the quiz. A brief lecture on the concept was given to college students, and in an informal survey following the exam, 90% of the students stated that the game had helped them understand the ring flip of cyclohexane. Nevertheless, further experiments are needed to evaluate the game as a learning tool.

Silva Júnior and collaborators [51] developed an educational computer game that allows students to review stereochemistry concepts in an entertaining manner: the "Stereogam". The game was tested by both students and teachers (43 and 202, respectively) of Introductory Organic Chemistry. The participants responded to an online questionnaire with ten statements to evaluate the game on a 10-points Likert scale. Again, the game seemed to be motivating, entertaining, user-friendly and dynamic. As a consequence, it could be used in the process of stereochemistry learning. Nonetheless, experiments with statistical analysis to prove the effectiveness of the game in student learning was not presented in the study.

2.3.2 DGBLs to aid the resolution of problems in science

As we pointed out earlier, we are interested in DGBLs that were developed with the intention of gathering more knowledge about open problems in biology and chemistry. Thus, here we present four studies in the area of genetics, proteomics, and chemistry.

EteRNA is a straightforward puzzle game developed through the synergy of two universities: Carnegie Mellon University and Standford University [52]. EteRNA unites an interactive interface for modeling biomolecules with a remote experimental laboratory. It presents a web-based interface which challenges participants to design and rank RNA sequences that will fold into a target structure when tested in the real laboratory. In this virtual Lab, the players use their skills to design RNAs that could be used as an aid of future life-saving therapies. The developers highlight three main features of the game. First, the game prevents many forms of data manipulation. Second, the platform allows rapid tests of reproducibility. Finally, EteRNA requires rigorous adherence to the scientific method where it demands a nontrivial prediction or hypothesis before each experiment. The challenges are placed in the game according to the levels

reached by the players, and they are rewarded with points according to the puzzle's difficulty.

Kawrykow and collaborators [53] designed a game named Phylo to allow players to perform multiple sequence alignment (MSA) without requiring significant knowledge in the biology field. The authors transformed the MSA problem (which is proved to be NP-hard) into a casual colored blocks game that takes advantages of the human ability to interpret visual patterns a rapid manner. The blocks represent DNA sequences, and so, have four different colors (one for each nucleotide). The game starts with a pair of sequences to align, and, depending on the choice of the player, may include up to ten separate sequences that will be aligned together. Each block displayed can be moved horizontally, if necessary pushing its neighbors, but can never be traded with another block. Phylo goals are to move the blocks to find an arrangement that maximizes conservation across columns while the number of gaps is minimized. The developers also added a timer to the game. Each level must be completed within a specified time limit (variation of the challenge).

To spread awareness of protein fold complexity, Cooper et al. [14] developed an online game named Foldit. This uses crowdsourcing to predict a protein 3D conformation through the use of manipulation tools and user-friendly versions of algorithms already presented in the literature. The authors affirm that the game is convenient to a wide variety of people, even the ones who are not expert on the subject. Foldit displays improperly folded protein conformations for a pre-established time, during which players interactively reshape them in the direction they believe will drive them to reach the highest score in the game. The score represents the negative Rosetta energy method (further information in [54]). One of the advantages of the game is that the plays may be performed by a group of people (as a team) or individually. This feature could help students to understand the protein conformation problem further, as well as to learn about the contacts that take place in the folding process (such as hydrophobic, hydrophilic, and hydrogen bonds interactions).

From the studies presented earlier, we can conclude that citizen science approach ought to be applied to help in solving problems. Even though the mentioned studies have excellent teaching potential, they have not been submitted to any evaluation of the learning of the players throughout the game. Besides, they do not present any data mining technique to find patterns within the data generated by the players, which is the aim of the present study.

Chapter 3

Material and Methods

The development and evaluation of Proteingo followed the structure shown in Figure 3.1. The section *Requirements and project* describes the main steps of the Proteingo drafting project, including the use of Piccolo database. *Game Development* is the section where we cover in detail the creation and key features of Proteingo. The *Evaluation of the game* section details how this study evaluates motivation, user experience (UX), and user's learning. The section *Proteingo database* characterization relates the building of Proteingo database and its characterization. Finally, *PTS and UTS evaluation* aims to evaluate Piccolo's answers (PTS) and the majority vote of Proteingo users (UTS) to assess whether both models can be used to classify contacts. Further, this section checks the agreement of those referred to specialists in the area of contact analysis between residue atoms in protein chains. Also, a proposal of characterization of Non-Contacts is presented according to the PTS and UTS models.

3.1 Requirements and project

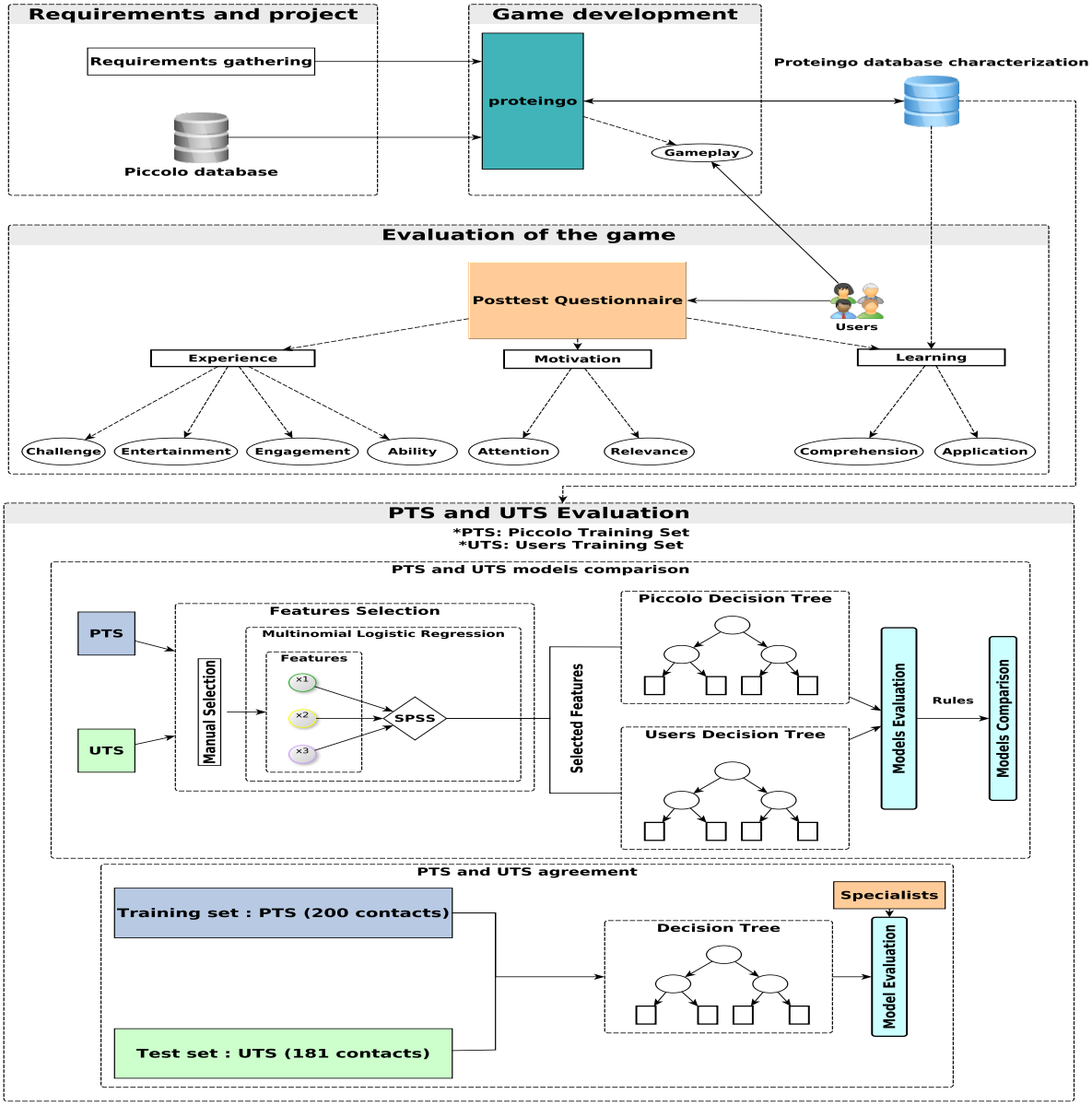
In the present section, we will cover all the steps and principles taken into account when creating the Proteingo game. Essential techniques in the area of Human-Computer Interaction (HCI), as well as the database of contacts that gave origin to the game are detailed in this passage.

3.1.1 Requirements gathering

The first step in building software is the requirements gathering. It identifies the needs of users for the best creation of the software. We counted on a group of people with in-depth knowledge in the development of web-based software, specialists

in the field of biochemistry, and also experts in the area of HCI. Hence, numerous meetings were held to elaborate Proteingo, taking into account the needs for a better visualization of the interactions that occur between atoms of protein residues in different chains. Our goal has always been to create a game that is attractive, simple and understandable to users.

Figure 3.1: **Methodology proposed for the development and evaluation of Proteingo. PTS** stands for Piccolo Training Database and **UTS** is short for Users Training Database.



Source: Author's elaboration.

An examination of the literature on design principles for game creation was con-

ducted prior to Proteingo elaboration. We then use several basic principles as found in [55]: *acceptability*, *accessibility*, *simplicity*, and *flexibility*. *Acceptability* happens when one makes use of elements in games which are accepted by a large group of people. An essential feature of *accessibility* is the availability of the game in the most diverse platforms: mobiles, laptops/desktops, and tablets. The more straightforward a game is, the more flourishing is the user experience at it. Therefore, a minimum number of elements should be used in the interface of the game for greater *simplicity*. Thus, it is expected that the player will not have to worry about unnecessary information on the screen and, consequently, will be more focused on the game itself. Finally, *flexibility* gives the user greater control over the game. The mere fact that a player decides to stop or summarizes the game at any time can positively contribute to their experience.

3.1.2 Piccolo database

Piccolo is a structurally characterized database of protein-protein interactions described at atomic level. The established interactions are calculated as follows: first, a radial cutoff is used to identify the atoms within 6.05Å of distance since it is the maximum length of a water-mediated hydrogen bond. After that, the atomic pairs are annotated with a particular type of bond depending on their atomic types, distance, and the angle between them. The authors considered 12 types of interactions for constructing the database, including Van der Waals interactions, hydrogen bonds, and hydrophobic interactions. Besides, the Piccolo database has more than 260 million pairs of atoms interacting in more than 38,000 complexes [13]. Table 3.1 shows the distance and angle criteria adopted by Piccolo in classifying the contacts of hydrogen bond (HB), hydrophobic (HP), ionic (IO), and aromatic stacking (AR), which are the non-covalent bonds considered in Proteingo.

Table 3.1: **Contact angulation and distance criteria**

Interaction	Distance criteria	Angulation criteria
HB	$d(a_i, a_j)$ <3,9 Å	$\theta(a_d,a_h,a_a)$ >90°
	$d(a_h, a_a)$ <2.5 Å	$\theta(a_d,a_a,a_{a-ant})$ >90° $\theta(a_h,a_a,a_{a-ant})$ >90°
HP	$d(a_i, a_j)$ <5 Å	-
IO	$d(a_i, a_j)$ <6 Å	-
AR	$d(a_i, a_j)$ <6 Å	-

Contact between atoms i and j based on their properties. Letter **d** is the distance between them and θ the angle. $\theta(a_1,a_2,a_{3t})$ represents the angle between a_2, a_1 and a_3; a_d = donnor; a_a = acceptor; a_h = hydrogen; a_{a-ant} = atom which is the antecedent to acceptor (Adapted from Bickerton, Higueruelo and Blundell [13].)

Source: Author's elaboration.

3.2 Game Development

All procedures for building the Proteingo game, as well as the features of the stated game, are present in this section. We detail all aspects of Proteingo's gameplay and also how the players' performance is assessed in the game (through its accuracy and level of play).

3.2.1 Proteingo

Proteingo is a game that aims to build a manually curated database of non-covalent bonds between atoms of proteins residues. We also let users who are not experienced in evaluating contacts play Proteingo, and we have great evidence that they learned while playing. This is important once it is known that many students have difficulty in understanding the biochemical processes that occur at atomic level. One of the leading causes of the problem is the fact that many textbooks present visual examples that are difficult for students to absorb. Proteingo, on the other hand, shows interatomic interactions through a straightforward and intuitive interface.

Proteingo is an online multiplatform game. Hence, users only need an electronic device such as mobile phone, notebook/computer or tablet connected to the internet. The game was developed using the interpreted language JavaScript that, besides being the primary language for programming in web browsers, also allows the control of the behaviors of the navigators in codes sent in HTML pages. 3Dmol and D3 (Data Driven Documents) were the main libraries used for the creation of Proteingo gaming fashion. The 3Dmol library enables the rotation, translation, and zoom of the user analyzed interaction [56]. While the D3 library offers more data dynamics, and better user interaction with game [57].

3.2.2 Gameplay

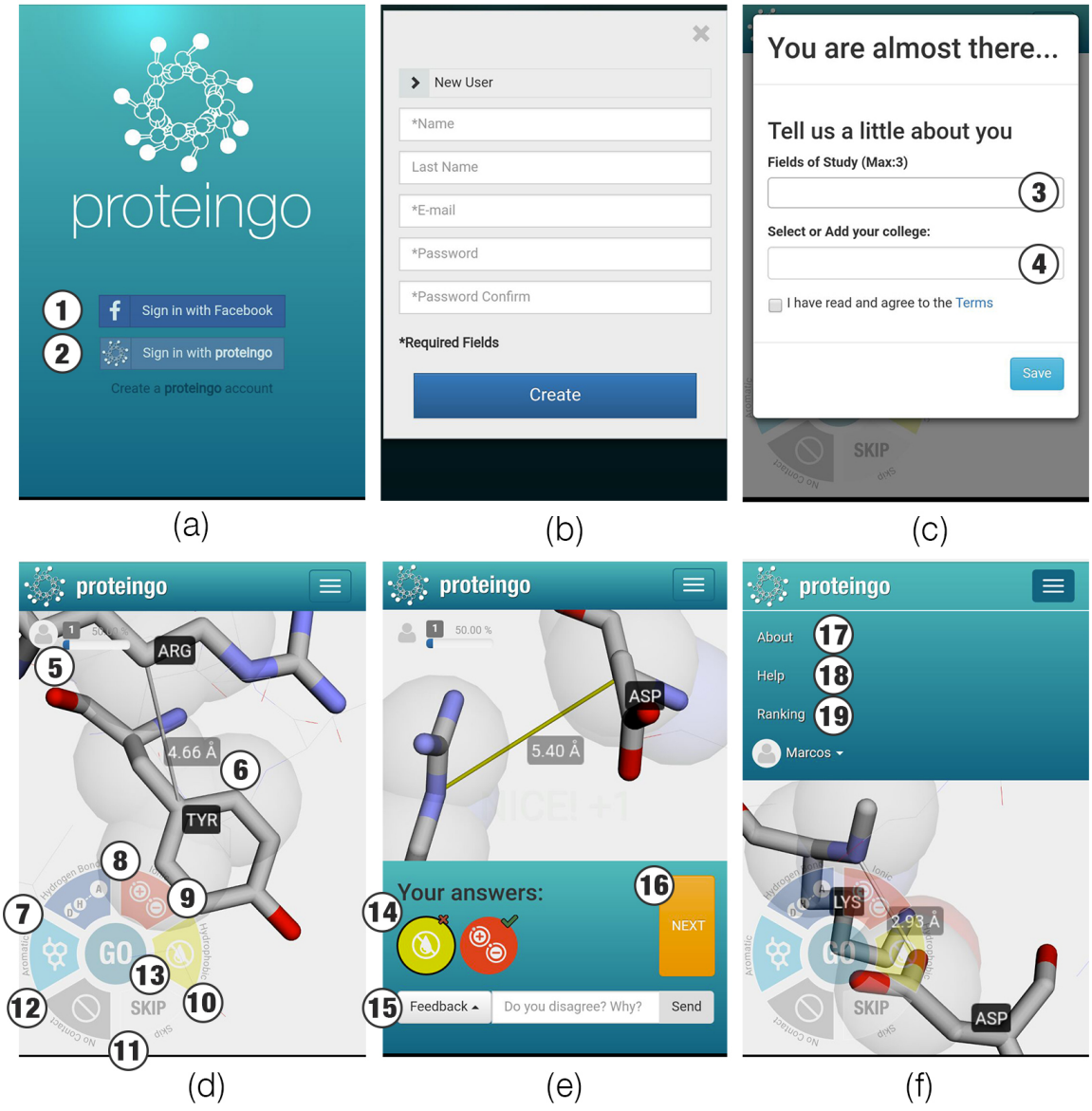
Initially, a Proteingo user interacts with the game by logging in the platform. Hence, a Proteingo account needs to be created. The user can log into the system either by using his/her Facebook account or creating a new user through the Proteingo platform as illustrated in Figure 3.2 (a). When creating a new user, a screen same as Figure 3.2 (b) is shown, and the user must fill the required fields with his/her data (name, email, and password). In the next screen, the system needs the field of study and college from the user (Figure 3.2 (c)). The field of study is necessary for discriminating the users who are more acquainted with the chemical non-covalent bonds analyzed in the game.

When logged in, a screen as shown in Figure 3.2 (d) is displayed. The user can rotate, translate and zoom in and out on the 3D image that illustrates the atoms of residues in contact. It is relevant to highlight that the interactions are randomly displayed for users. For instance, Figure 3.2 (d) shows the possible contact between two carbon atoms of arginine and tyrosine. There is a 4.66 Angstroms edge bonding the two atoms of carbon. The user is supposed to select the (right) interaction for the displayed image by choosing one of the options on the button, which is located at the left-bottom corner of the screen. The button counts with six options to be selected: ionic interaction, hydrophobic interaction, skip, non-contact, aromatic interaction and hydrogen bond (numbers 7 - 12 in Figure 3.2 (d)). After selecting the (right) option, the user should click on "GO" located in the center of the circle at the left-bottom corner of the screen (number 13 in Figure 3.2 (d)). Knowing that feedback can positively influence user's behavior and performance in educational games, the messages "Nice!" or "Miss!" pop up on the screen regarding the selection of a correct or an incorrect answer, respectively. It is important to mention that feedback messages are according to Piccolo's responses. The scoring system is uncomplicated: for each hit, a Proteingo player earns a point.

A pair of residues may perform more than one interaction. Figure 3.2 (e) illustrates a case where an interaction with a "red x" and the other with a "green mark" represent a correct and an incorrect response from the same pair of residues, respectively (number 14 in Figure 3.2 (e)). The user may select each one of the interactions to send us feedback, in a case of disagreement with the response given by the game (number 15 in Figure 3.2 (e)). This feedback is essential to identify possible errors coming from the Piccolo database [13]. The user may provide feedback either by writing it in the text box or by choosing one of the previous options on the menu "Feedback." If one desires to analyze the next interaction, one should click on the "Next" button as shown in number 16 in Figure 3.2 (e).

Proteingo offers further information in the "Menu" option located on the right-top corner of the screen (see Figure 3.2 (f)). We provide a "help" menu that shows simple tutorials on how to play Proteingo. The user can also find further information about the teamwork of Proteingo by clicking on the "About" option. When selecting "Ranking" in the menu option, the player still can check his/her position in the ranking. The more the user plays, the higher they get in the ranking. A user's progress is displayed in the top-left corner of the screen, together with his/her Facebook profile photo. The progress is based on the accuracy and the level the user achieves. Equations 3.1 and 3.2 show how the accuracy and the level are calculated in the game.

Figure 3.2: **Proteingo screenshot illustrating functionalities and visualizations.** Button to log in the game by using the user Facebook account (number 1). In case the user desires to log in the game by using a Proteingo account (number 2), the user should click on "Create a Proteingo account" and provide the data as in (b). Proteingo requires the area of study of the user (number 3); the field of study should not exceed three. After that, the user has to provide his/her institution (number 4) and accept our terms to play Proteingo. The main screen shows the user's progress (accuracy and level) at the top-left corner (number 5). Interactive game screen showing the residues (arginine and tyrosine in the example) and their interaction(s) (number 6). Button for choosing the aromatic interaction (number 7). Button for choosing the hydrogen bond interaction (number 8). Button for choosing the ionic interaction (number 9). Button for choosing the hydrophobic interaction (number 10). skip option (number 11). Non-contact (number 12). After choosing one of the options on the circle button, the user should press "Go" (number 13). A review of the last interaction(s) played by the user is shown as in (e); a "red x" and a "green mark" represents the misses and hits, respectively (number 14). A feedback must be provided by the user in a case of disagreement with the response given by the game (number 15). Button to analyze the next interaction(s) (number 16). Description about the teamwork of Proteingo (number 17). How to play Proteingo (number 18). Option to check the user's position in the "Ranking" as well as the scores of other users (number 19).



Source: Author's elaboration.

$$Accuracy = \frac{Hits}{Total\ number\ of\ plays} \tag{3.1}$$

where Hits stands for the number of correct interactions the user got. Total number of plays is the total number of interactions the user has analyzed.

$$\begin{cases} Level_1 = 10 \\ Level_n = Level_{n-1} + n * 10 \end{cases} \tag{3.2}$$

Initially, every user starts at $Level_0$. To reach $Level_1$ a total of 10 interactions must be played by the user. Following that, to reach $Level_2$ the user must play 30 interactions once $Level_2 = Level_1 + 2 * 10 \rightarrow Level_2 = 10 + 20 = 30$.

3.3 Evaluation of the game

The game was evaluated according to the criteria of motivation, user experience (UX), and their learning of the analyzed contacts while playing. All of these criteria are detailed in this section. Also, we have created a posttest questionnaire that evaluates these Proteingo evaluation criteria. The questions of the posttest questionnaire and its objectives are outlined in the present passage as well.

3.3.1 Posttest questionnaire

Players responded to the questionnaire from February through June 2017. Since it was available online[1], such users had the flexibility to respond wherever they intended. The posttest questionnaire was divided into three sections. The first one had the purpose of assembling information about the users' level of education, and previous knowledge of chemical interactions. The second section aimed to evaluate the Proteingo from its user's view considering the aspects of *Motivation*, *User experience (UX)* and *Learning*. The statements and possible answers to both sections of the posttest questionnaire are shown in Table 3.2.

Motivation refers to the inner impulse that takes a user to a particular action. It is easily found authors in the literature emphasizing that motivation influence the learning performance of students in a more indirect way; once motivation maintains them more focused on the task [18, 40, 41]. This element assumes that two main factors influence the user's motivation for learning: 1) *Attention*: relates to users' congruent responses to instructional stimuli. The intention is to capture user's attention and

[1]The questionnaire is available at https://goo.gl/forms/05Q8J6xHaMgt60132.

curiosity, at a satisfactory level, at the beginning and throughout the process of learning [58] [59]. A simple question should keep in mind in this factor: Do I want to use it?, and 2) *Relevance:* cooperates with users in associating their previous experience with didactic materials. The user also needs to comprehend that an educational proposal is consistent with his goals. He/She should be able to connect the content of learning with their professional or academic future tasks [58] [59]. A good question to answer here would be: Do I need to use it?

Table 3.2: Posttest Questionnaire Summary

User Demographics		
Education	Undergraduate, M.Sc., Ph.D., Postdoc	
Field of Study	Exact and Earth Science; Biological Sciences; Biostatistics; Biophysics; Bioinformatics; Biochemistry; Biotechnology; Botanical; Pharmacology; Genetics; Immunology; Microbiology; Engineerings; Health Sciences; Agriculture Sciences; Applied Social Sciences; Human Sciences; Linguistics, Arts, and Literature; Others.	
Previous knowledge	No. It was my first contact with the concept. Yes, but I did not remember anything. Yes, and I remembered some concepts. Yes. I consider myself someone with a lot of experience in the area.	
Motivation, User Experience (UX), and Learning		
Question	**Statement**	**Class**
Q1	The content presented by the game is important to me.	Motivation ⇒ Relevance
Q2	The content presented by the game is relevant to my interests.	Motivation ⇒ Relevance
Q3	The content presented by the game is challenging. Images are neither too easy nor too difficult.	UX ⇒ Challenge
Q4	I got deeply involved in the game.	UX ⇒ Engagement
Q5	The game was enjoyable.	UX ⇒ Entertainment
Q6	The game was fun.	UX ⇒ Entertainment
Q7	I believe that the game could be very useful as a didactic-pedagogical resource in the classroom.	Learning ⇒ Application
Q8	I was able to understand the concepts of hydrogen bonds, ionic, hydrophobic and aromatic interactions through the game in surprising and unexpected ways.	Learning ⇒ Comprehension
Q9	The game interface is attractive.	Motivation ⇒ Attention
Q10	I understood the messages that the game sent me, about the correct and incorrect answers.	UX ⇒ Ability
Q11	I would play this game again.	UX ⇒ Entertainment
Criteria for Choosing an Interaction as Hydrogen Bond, Hydrophobic, Ionic, Aromatoc, or Non-Contact		
Criterion	**Statement**	
Random	I did not learn a rule of how the interactions happen. In general, I was randomly playing.	
Color	Colors of the atoms involved.	
Distance	The distance of the atoms involved.	
Charge	Charges of atoms involved.	
Polarity	The Polarity of the residues involved.	
Chemical Element	Chemical elements (atomic type) involved.	
Covalent Bond	Covalent bonds of the atoms involved.	
Amino Acid Types	Types of amino acid residues involved.	
Angle	Angles between the atoms involved.	
No Occlusion	No occlusion (no atom between two connected atoms)	
Others	Others.	

User Demographics: Data about the user's background and knowledge of the chemical interactions in the game. **Motivation, User Experience (UX), and Learning:** All questions were evaluated by the respondents on a 5-point Likert scale: 1-completely disagree, 2-disagree, 3-not agree nor disagree, 4-agree, 5-completely agree. **Criteria for choosing the interactions in the Proteingo game:** Users were supposed to choose all the criteria they thought was right. The "Others" option was carefully analyzed in order to evaluate its biochemical sense.

Source: Personal collection.

Even though there is no unanimity or a unified interpretation of User Experience (UX), some attempts have been stated. It involves a person's feelings about the use of a particular product, system or service as seen on ISO 9241-210 [60]. On the Proteingo user experience, we analyzed three critical features of experience in a game: *1) Challenge:* A game must be challenging enough, and appropriate to the skill level of the player. Keeping variations in the degree of difficulty could maintain a proper rhythm to control the speed with which new details or challenges are revealed. The outcome of achieving the goal must be unpredictable as well. However, the variations of challenges should be planned so as not to cause any fatigue or tension on the user [41, 61]; *2) Entertainment:* As stated by Vandercruysse and authors [18] entertainment is often seen as one of the most important reasons to implement educational games. When playing is something memorable for the user, there will be an intensely positive experience, and s/he tends to return to the game and recommend it to others. Therefore, Users become more enthusiastic when the learning process is also entertaining, and *3) Engagement:* it is usual for players to report that his/her engagement in a specific game is so deep that they spend many hours playing utterly unaware of what is around. A good game should keep a user emotionally involved to provide positive distractions and forgetfulness of his/her daily concerns.

As for *Learning*, often users learn in an unplanned and unexpected ways. For example, while they play a game, the learning process can occur through mistakes made by the player or the systematized trial and error method. Thus, to this byproduct coming from the realization of other activities is given the name of incidental learning [62]. Incidental learning is likely to occur when the user is persuaded to perform a task regardless of the type of task that may be [41, 63]. The user's learning outcome was studied in two manners in this work. First, we studied their development in the Proteingo game. Their accuracy (showed in Equation 3.1) over the number of plays gave us valuable insights into their learning process. Later on, the last questions in the posttest questionnaire confirmed the hypothesis respecting their learning outcome.

To respond to each question above, the participants were supposed to choose one of the following options (according to a 5-point scale): 5 - Strongly Agree, 4 - Agree, 3 - Neither agree nor disagree, 2 - Disagree or 1 - Strongly disagree. These questions about Motivation, User Experience, and Learning are presented in Table 3.2.

Finally, the last part of the questionnaire collected information about the users' knowledge on the definitions of interactions presented in Proteingo. That is, the users had to choose the options that were mostly linked to the definition of hydrogen bonds, hydrophobic, ionic, and aromatic interactions, as well as "Non-Contact" option (see Table 3.2).

3.4 Proteingo database characterization

We randomly selected 200 contacts of each type (HB, HP, IO, AR, and NC) from the Piccolo database [13]. Hence, a total of 1,000 distinct contacts were chosen. Sometimes, a contact can be classified into two categories (hydrogen bond and ionic interaction or hydrophobic and aromatic interaction). To avoid a possible muddle with users or in the database, these contacts were excluded. All user information is safely stored in the Proteingo database, and these data are used to express the learning factor of each user in this study. Furthermore, these data were used to extract knowledge regarding the analyzed contacts.

We also obtained 13 features from the contacts taken from the Piccolo base. These features' description and domain are shown in Table 3.3.

Table 3.3: **Criteria used as attributes in the classification of the contacts.**

Features	Description	Domain
p1_resname, p2_resname	Residues that are part of the interaction are obtained.p1_resname and p2_resname represent values (amino acid names) of the firstand second residue of the interaction, respectively.	{Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile, Leu, Lys, Met,Phe, Pro, Ser, Thr, Trp, Tyr, Val}
polarity_res1, polarity_res2	refers to the polarity of the first residue of the interaction, while polarity_res2 denotes the polarity of the second residue. The polarity of the side chain was obtained according to [22].	{polar, nonpolar}
type_res1, type_res2	type_res1 and type_res2 are relative to the properties of residues 1 and 2, respectively. The types can assume five values: aliphatic, aromatic, neutral, negative and positive. The properties mentioned are according to [21].	{aliphatic, aromatic, neutral, negative, positive}
p1_atname, p2_atname	These are the names assumed by the atoms of the interaction. While p1_atname identifies the name of the first atom of the contact, p2_atname identifies the name of the second atom.	{CB, CD, CD1, CD2, CE, CE1, CE2, CE3, CG, CG1, CG2,CH2, CZ, CZ2, CZ3, N, ND1, ND2, NE, NE1, NE2, NH1, NH2,NZ, O, OD1, OD2, OE1, OE2, OG, OG1, OH, SD}
charge_at1, charge_at2	The charges of the atoms were obtained according to the definition of [13]. In this way, atoms likely to be more positive receive 1, more negative atoms are scored as -1. All the remaining atoms receive 0 because they are neither positive nor negative. The charges of the first and second atoms of the interaction are charge_at1 and charge_at2, respectively.	{-1, 0, 1}
valence_at1, valence_at2	The number of covalent bonds the atoms make is also one of our characteristics. The residual topology file generated by the CHARMM²software was used to calculate the valences of the first and second atoms in contact (valence_at1 and valence_at2). It is important to highlight that histidine on,AMBER may be assigned into three manners: HIE (hydrogen in nitrogen Eta: neutral), HID (hydrogen in nitrogen delta: neutral) and HIP (in both nitrogen: positive) [64]. In our experiment, histidine is protonated as HIP.	{1, 2, 3, 4}
distance	The distance (in Å) between the atoms involved in the interaction.	{2.13 - 6,0 Å}

All the afforementioned criteria are used to predict whether the interaction is a hydrogen bond (HB), hydrophobic (HP), Ionic (IO), or aromatic stacking (AR). The colored green lines represent the characteristics selected through manual selection and MLR.

Source: Personal collection.

After applying the Multinomial Logistic Regression (MLR), only the charge, valence and distance were statistically significant for the classification. It is the reason these features are colored with green in Table 3.3.

3.5 PTS and UTS evaluation

Since data mining is a process of discovering hidden valuable knowledge by analyzing large amounts of data, we ran an experiment to find rules (through decision trees) that best classify the hydrogen bond (HB), hydrophobic (HP), ionic (IO), and aromatic contacts (AR). We evaluate PTS and UTS through data mining process. In order to do so, we used a plugin developed by Stiglic and collaborators [65] to the WEKA platform [28]. The plugin's name is VTJ48 (Visually Tuned J48) and is an adaptation of the original J48 (old C4.5) in WEKA. VTJ48 automatically optimizes the parameters C and M of the decision tree algorithm. Therefore, two decision trees were created: a model constructed according to Piccolo's responses to the interactions and another tree according to the users' responses (as can be seen in Table 4.5).

3.5.1 PTS and UTS models comparison

In this experiment, the feature selection technique was applied through the Multinomial Logistic Regression (MLR). As our variable is the analyzed contact, and our features are both numeric and categorical, we use SPSS software[3] to handle the MLR in our analysis. Another reason for using MLR is the fact that it does not assume normality, linearity, or homoscedasticity (constant variance of the experimental errors for different observations) [66]. Table 3.3 presents the attributes submitted to the MLR analysis.

Features that have high variability are likely to impact the generated models in the data mining process negatively. Thus, in the method we named *Manual Selection* we decided to remove the features with high variability, verifying how significant they are through the MLR process. The features highlighted in green in Table 3.3 (charge, valence, and distance) identify the attributes that remained in our bases (both PTS and UTS) for the accomplishment of this experiment. It was done to identify the rules that are alike between the two bases, as well as the generated rules that are different; with the aim of verifying whether PTS and UTS are both suitable for the classification of HB, HP, IO, and AR contacts. The models were compared through the rules generated

[2]Available on `https://www.charmm.org`.
[3]available at https://www.ibm.com/analytics/us/en/technology/spss/

by Piccolo and users trees (PTS, UTS). The rules comparison technique is a convenient formalism for expressing knowledge discovery in expert systems [67].

3.5.2 PTS and UTS models agreement

This experiment intends to verify the compliance of the responses given by the players (UTS) with the responses given by Piccolo (PTS). In this way, the Piccolo's database was used to train our decision tree, and the base with the answers of the users was used to test the generated model. A total of 800 instances were used in training (200 interactions of each type: HB, HP, IO, and AR); and 781 instances were used as the test in this experiment (199 HB, 195 HP, 182 IO, 205 AR). Noting that all the contacts in the test database had their labels marked according to the majority vote of the players.

3.5.3 Non-contact characterization

A difference of the present work, when compared to the study developed by Piccolo [13], is the characterization of what we call "Non-Contact" (NC). In the Proteingo game, an NC option should be chosen when a contact was not an HB, HP, IO, nor an AR. That is, this choice should be taken by exclusion. Our intention in this characterization is to differentiate a valid contact (be it HB, HP, IO, or AR) from what it is not a valid contact. For this, the same data mining procedures (already described) were applied in this experiment. First, we created two models with Piccolo's responses and the majority of users vote according to what they considered to be a contact (IC) and non-contact (NC). Then, the rules generated by both PTS and UTS models were compared to verify similarities and differences in the characterization of NCs.

3.5.4 Validation and Evaluation Metrics

To obtain a model with better generalization capacity, cross-validation [68] was used in the present work. The cross-validation method consists of dividing the total set of data into k mutually exclusive subsets of the same size, and from this, a subset is used for testing, and the remaining $k-1$ are used for training the model (its accuracy is calculated). This process is performed k times by circularly alternating the test subset [69]. In the present work, we consider a $k = 10$ in the realization of the experiments. The precision, sensitivity, specificity, f-measure, and efficiency metrics were used for the evaluation of the models generated in the training phase with the bases of Piccolo

and Users [29, 69]. These are briefly described below (consider TP = True Positives, TN = True Negatives, FP = False Positives and FN = False Negatives):

Precision: The percentage of instances of a correctly-assigned class among all those that were classified in that class. Eq. (3.3) shows how the precision calculation is performed in our experiments.

$$Precision = \frac{TP}{TP + FP} \tag{3.3}$$

Sensitivity: It is the ratio of true positives, that is, the ability of the system to correctly predict specific interaction that was classified as belonging to the correct class. The sensitivity calculation is shown in Eq. (3.4).

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.4}$$

Specificity: Identifies the proportion of true negatives, i.e., the system's ability to correctly predict a particular interaction that does not belong to the class being analyzed. The calculation of the specificity is shown in Eq. (3.5).

$$Specificity = \frac{TN}{TN + FP} \tag{3.5}$$

F-measure: Also called as F1 score, is the harmonic average of the precision and sensitivity, where an F1 score is its best value at 1 (perfect precision and recall) and worst at 0. Eq. (3.6) shows the calculation of f-measure.

$$F - measure = \frac{2 * TP}{2 * TP + FP + FN} \tag{3.6}$$

Efficiency: This metric is the arithmetic mean of both sensitivity and specificity. In practice, sensitivity and specificity vary in inverse directions. When a method is very sensitive to positive, it tends to generate many false positives and vice versa. Hence, a perfect decision method (100% sensitivity and 100% specificity) is rarely achieved, and a balance between the two must be achieved. Eq. (3.7) shows how efficiency calculation is performed in our experiments.

$$Efficiency = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2} \tag{3.7}$$

Chapter 4

Results and Discussion

The primary results of the evaluation of the game, as well as the evaluation of the models generated by Piccolo (PTS) and Proteingo users (UTS), are shown in this chapter. First, we present the outcomes regarding user's motivation, experience (UX), and learning according to user's performance in the game together with their responses to posttest questionnaire. The evaluation of the manually curated database is presented in the second section of this chapter. To simplify PTS and UTS models (features selection), a multinomial logistic regression was performed. A decision tree was the data mining technique utilized to compare the models as well as to evaluate them.

4.1 Proteingo evaluation outcomes

Since we depended on the users' answers to the posttest questionnaire, we were able to work with data from only 27 players in the learning analysis. Among these, 19 are graduate students (M. Sc., Ph.D., and PostDoc). Respondents could choose up to 3 fields of study. Not surprisingly, the majority of users were in the areas of bioinformatics (77.8%), biochemistry (14.8%), genetics (18.5%), and biophysics (7.4%). 13 of the 27 respondents are also from the area of exact sciences, and only one belongs to the field of health sciences. According to the contacts presented in the game, 77.8% of the players remembered of some concepts and 22.2% considered themselves experts on the subject. 22.2% of users (a total of 6 people) did not know or remember the non-covalent bonds analyzed. User demographics are summarized in Table 4.1.

Table 4.1: **Participant demographics**

	N	%
Education		
Undergraduate	8	29.6
M.Sc.	5	18.5
Ph.D.	11	40.7
Postdoc	3	11.1
Field of Study *		
Biophysics	2	7.4
Bioinformatics	20	74.1
Biochemistry	4	14.8
Biological Sciences	1	3.7
Health Sciences	1	3.7
Exact and Earth Sciences	13	48.2
Genetics	5	18.5
Chemistry	1	3.7
Games and Technologies	1	3.7
Biotechnology	2	7.4
Engineering	1	3.7
Human Sciences	1	3.7
Pharmacology	1	3.7
Previous knowledge about the contacts considered in the game		
No. It was my first contact with the concept. (**Poor Experience**)	3	11.1
Yes, but I did not remember anything. (**Middle-Poor Experience**)	3	11.1
Yes, and I remembered some concepts. (**Middle-Great Experience**)	15	55.6
Yes. I consider myself someone with a lot of experience in the area. (**Great Experience**)	6	22.2

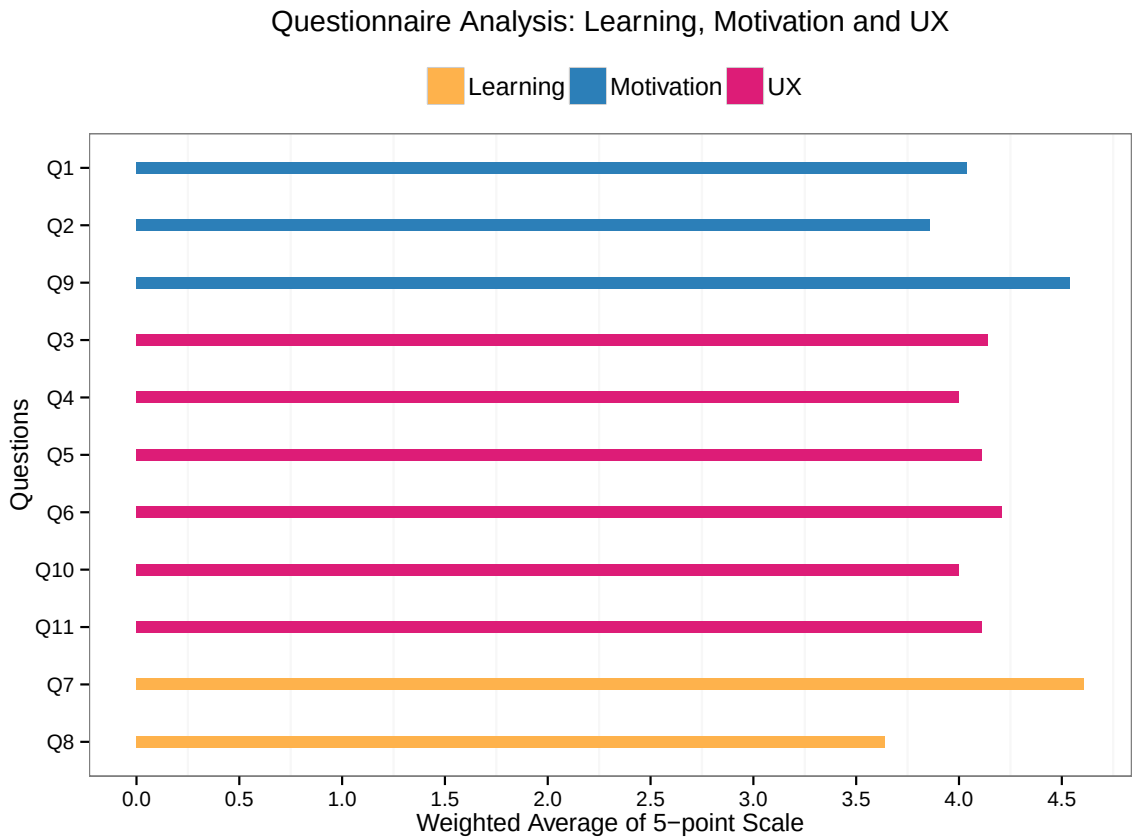
*Each user had the right to choose up to three fields of study. This is the reason for the sum of percentages be greater than 100%

Source: Personal collection.

The post-test questionnaire assessed the motivation, experience as well as its user knowledge, according to the questions presented in Table 3.2, page 25. Figure 4.1 illustrates the weighted average values of the 11 questions assessed on a scale ranging from 1 (strongly disagree) to 5 (strongly agree). The formula used to calculate the weighted mean is the sum of the products of the number of respondents, who selected each response and their respective weight divided by the total number of those surveyed. Note that, except for Q8, all others received high marks - above 3.75 points on average.

Concerning users' motivation, 93.5% of users said that Proteingo has a compelling interface. It shows us that one of the definitions of DGBLs (that visuals of the game should attract user's attention [36]) was reached in the game. In the relevance aspect of the game, only 9.7% of players said that the molecular interactions addressed were not relevant to their interests. So, we say that overall, users were motivated to play the Proteingo game. Figure 4.1 shows that all UX-related questions obtained a weighted

Figure 4.1: **Learning, Motivation and UX Analysis**. The chart shows the weighted average values of the 11 questions (see Table 3.2, page 25) evaluated on a 5-point scale



Source: Personal collection.

mean score higher than 3.5. It shows that not all users were deeply involved in the game (only 64.5% of Players). One of the players in our experiment stated: "*I did not play a lot, and I felt frustrated because I was getting a lot of incorrect responses. The reason might be that I work with DNA/RNA, and do not know anything about proteins*". 83.9% of the players affirmed that the challenge present in the game was adequately purposed (Q3). Images were not too easy and not too hard. 4 out of 31 users remained neutral regarding Q3. Relating to the ability of the users, 80.6% understood the feedback given by the game about correct and incorrect answers. As stated by Ebner [41], this result also certifies all the premises that good usability leads to correct intuitive manipulation.

Although serious games are designed not only for pure entertainment [16], fun still seems to arouse the player's interest in the game [61]. The results on fun and entertainment were not different in the Proteingo game. 71% of respondents stated

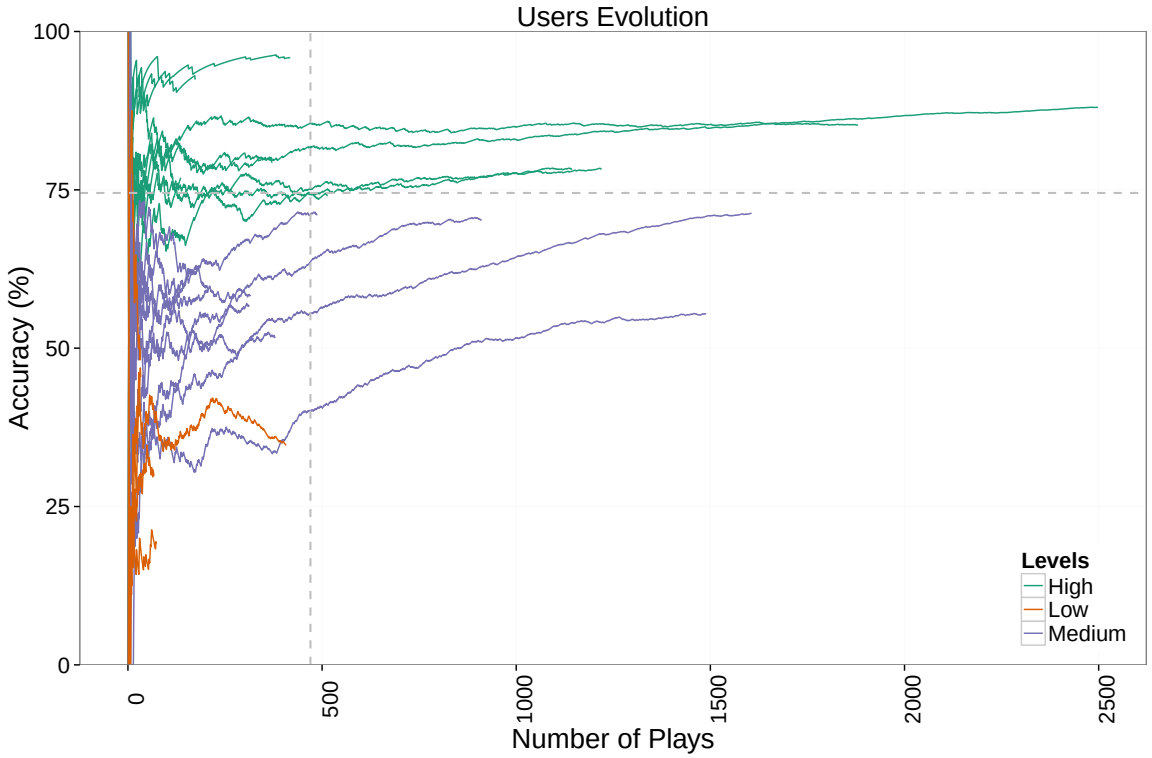
that the game was pleasurable and 80.6% said they had fun while playing. Moreover, 27 of the 31 (87.1%) players would play Proteingo again.

Evidence that the Proteingo game has relevance in the incidental learning of molecular interactions can also be seen in Figure 4.1. Q8 received a weighted average of more than 3.5 on a scale ranging from 1 to 5. Thus, most players learned more about hydrogen bonds (HB), hydrophobic (HP), ionic (IO), and aromatic (AR) contacts while playing. Regarding Q7, 96.3% of the players believe that the game ought to be used as a didactic-pedagogical resource in the classroom. One of the respondents still said: "*It (the Proteingo game) was interesting and helped me in a bioinformatics discipline that I took during undergraduate school*". Still, others believe that Proteingo can be improved to teach several other molecular interactions: "*I had such an excellent experience with the Proteingo game, and I believe the game has a great potential in the field of education. I wish I had a game like that to help me with the organic chemistry in high school*".

Summary: Except for Q8 (I was able to understand the concepts of HB, HP, IO, and AR contacts through the game in surprising and unexpected ways) all questions regarding user's learning, motivation, and experience received mean values higher than 3.75 on a 5-point Likert-scale. Many players believe that the game could be beneficial as a didactic-pedagogical resource in the classroom (Q7 average > 4). 87.1% of the players would play Proteingo again. 71% and 80.6% of the players found the game pleasurable and fun, respectively.

Figure 4.2 shows the evolution of users according to their accuracy in the game. The evolution of the users was measured according to the accuracy reached by the number of plays. Users were divided into three levels: for final accuracy less than 50%, such users were classified in the "Low" category (orange line in the chart). When a user's final accuracy is greater than or equal to 50% and less than the median of accuracy (72.73%), this player is included in the "Medium" category (blue line in the chart). The "High" category was assigned to all users who achieved a final accuracy higher than the median (green line in the graph). The gray dotted horizontal line indicates the value of the median accuracy. The median number of plays is given by the gray vertical dotted line in the chart.

Figure 4.2: **Users Evolution**. **High-level:** Final accuracy >= 72.73% (median); **Medium-level:** 50% <= Final accuracy < 72.73% (median); **Low-level:** Final accuracy < 50%.



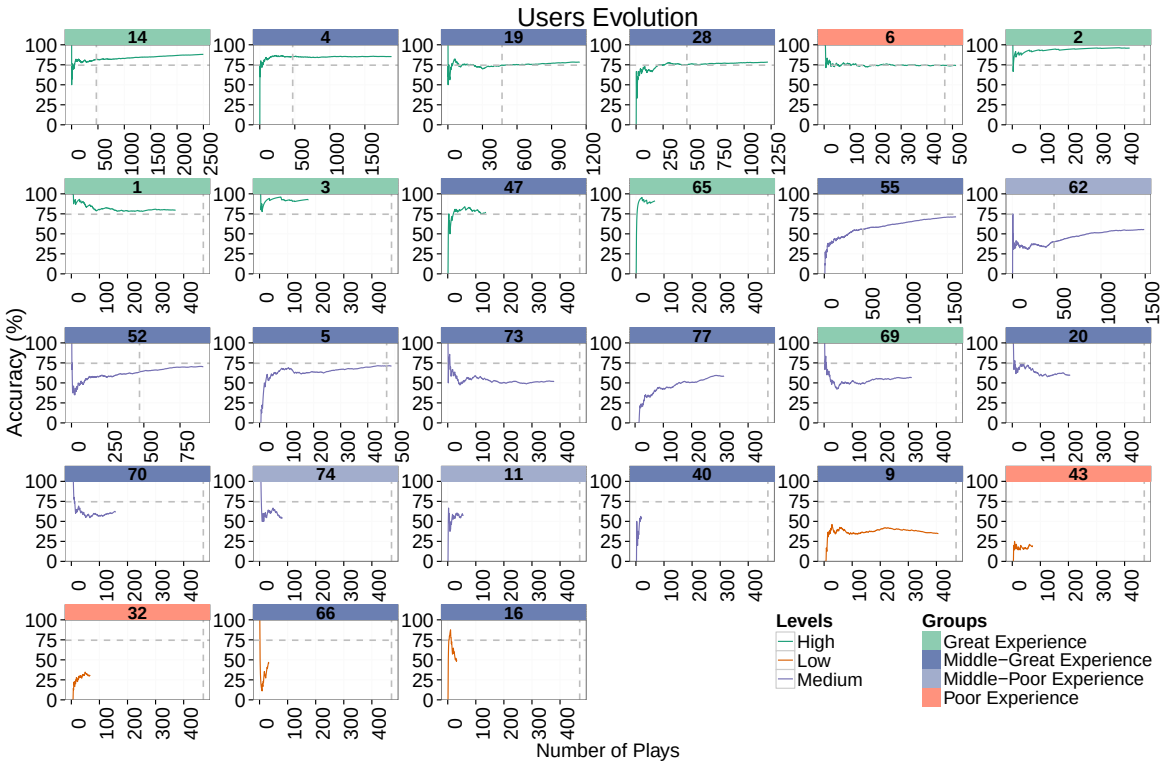
Source: Personal collection.

It is possible to notice from the plot in Figure 4.2 that the majority of the users turned out satisfactorily since their accuracy increases along the plays. The users who played the most are at the High level, and only three players performed more than 1500 plays; one of them did not reach an accuracy more significant than the median, remaining in the Medium category. The evolution curves of the High-level users slightly changed throughout the plays. In contrast, there is a significant difference in the curves of all Medium-level users. A vast majority of users in the Medium category have hit more than 100 plays. Due to these facts, it is suspected that Medium-level players have benefited the most from the game. None of the Low-level users reached the median number of plays (424 plays). We have only one Low-level user who has played more than 400 times. However, his final accuracy did not exceed 50%.

Figure 4.3 is derived from Figure 4.2 and shows the progress of each one out of the 27 users in the experiment. The levels in the chart in Figure 4.3 is calculated as stated in the paragraph above. The groups are separated as follows: **Poor Experience** are the users who had the contact with the concept for the first time;

Middle-Poor Experience included the users who know the concepts but did not remember anything about them; **Middle-Great Experience** categorize users who recalled some of the concepts in the game; and finally, users in the **Great Experience** group are those who consider themselves as experts in the noncovalent interactions shown in the present study. It is also shown in Table 4.1.

Figure 4.3: **Users Evolution - Individual.** Users were sorted in the order of High, Medium and Low, respectively. Then the users were sorted into each level according to their number of plays. The strip colors vary according to the prior knowledge that the respondents put in the post-test questionnaire (related to the interactions in the Proteingo game) as shown in Table 4.1.



Source: Personal collection.

Figure 4.3 also shows that, except for player 69, all users who consider themselves experienced in the area (Great Experience group), are classified into the High level. One possible explanation for the inclusion of user 69 at the Medium level would be his/her familiarity with the game. It is noticeable that the accuracy of this user decreased in the first 50 plays. However, that user's performance improved after that number; at which time the player 69 may have become more acquainted with the Proteingo game.

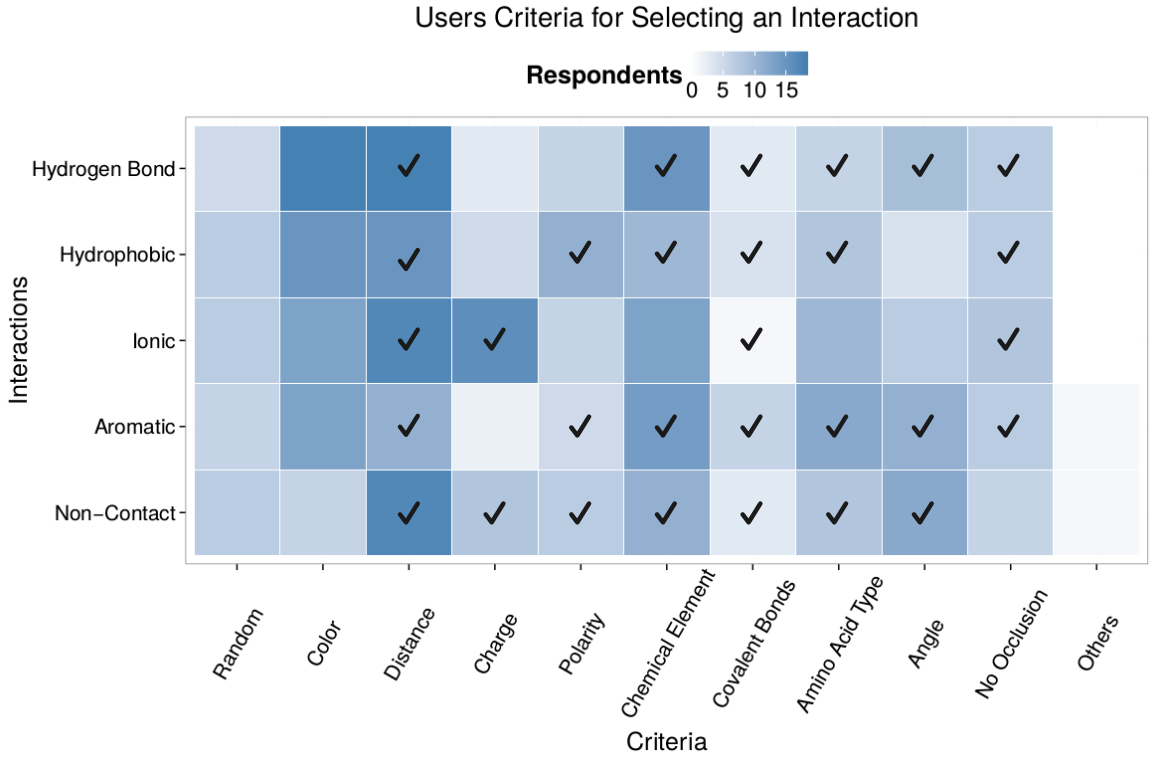
Another abnormality observed in Figure 4.3 is the inclusion of player 6 in the High level, since this player considers him/herself inexperienced in the area (Poor Experience). The accuracy of this user remained practically constant throughout the plays. Player 9's field of study is bioinformatics; perhaps his/her knowledge about the interactions analyzed may not be insignificant as it seems. For players 43 and 32 in the Poor Experience group, their inclusion in the Low level seems appropriate; although, there is a significant increase in the accuracy of player 32 throughout the plays. Both players 43 and 32 did not exceed 100 plays, consequently did not reach the median number of plays of 424. The Middle-Great Experience group were the majority and corresponded to a total of 55.6%. Of these, 26.7% are in the High-level, 53.3% are in the Medium-level, and the other 20% were categorized in the Low-level. We suspect that the Middle-Great Experience players took significant advantage of the game when they have a considerable difference in their initial and final accuracy (e.g., player 77). Of all the players in the Middle-Poor Experience group, we believe that player 62 has learned better; once s/he has effected approximately 1500 plays. In sum, we believe that the Poor Experience and Middle-Poor Experience groups categorized in the Medium-level have benefited the most from the Proteingo game in the learning process.

Summary: In sum, the majority of the users turned out satisfactorily since their accuracy increases along the plays. It seems that the Poor Experience and Middle-Poor Experience groups categorized in the Medium-level have benefited the most from the Proteingo game in the learning process. Of all the players in the Middle-Poor Experience group, we believe that player 62 has learned better; once s/he has effected approximately 1,500 plays.

The last five questions had the objective of verifying the criteria that the players used in the choices of HB, HP, IO, AR contacts and non-contact (NC) as well. The first criterion in the posttest questionnaire was intended to encompass users who played randomly. Thus, such users would not have to choose any other criteria given in each question. The guidelines presented in the posttest questionnaire are shown in Table 3.2, page 25, and user responses are illustrated in the heat map of Figure 4.4.

The criteria for the formation of a hydrogen bond are: the distance between the atoms, atomic type, covalent bonds that the atoms are forming, types of the residues in contact can also be considered, the angulation formed between the atoms involved and the absence of occlusion. Of these, the distance was the most relevant criterion that users noticed (55.6%), followed by atomic type (48.1%), angle (25.9%) and absence of occlusion (25.9%). 4 of the 27 respondents (22.2%) confessed that they played

Figure 4.4: **Users criteria for each interaction in the Proteingo game.** Check markers indicate which criteria are expected according to each interaction.



Source: Personal collection.

randomly (Random), and 16 players (59.3 %) used the colors of the atoms to choose a contact as hydrogen bond (color). The charge and polarity criteria represent the responses of 11.1% and 18.5% of users, respectively.

Polarity, atomic type, covalent bonds of the atoms, amino acid types, and no occlusion should be the criteria for selecting the hydrophobic contact in Proteingo. However, only 29.6% of the users chose the criterion polarity, followed by chemical element (33.3%), covalent bonds (11.1%), amino acid types (22.2%) and no occlusion (25.9%). 48.1% of the users selected that contact based on the colors and distance between the atoms (color and distance). Additionally, 18.5% mistakenly chose criterion charge, and 11.1% used angle regarding hydrophobic contacts. 6 out of 27 respondents randomly played.

The ionic contact is related to the distance, the charge of the atoms, and the absence of occlusions as well. Once more, the majority of users correctly chose criteria distance and charge (51.9% and 44.4%, respectively). 20 out of 27 players did not consider the occlusion. Criteria random, color, polarity, chemical element, covalent

bonds, amino acid type and angle, represent 22.2%, 40.7%, 18.5%, 48.1%, 3.7%, 33.3% and 22.2% of the responses, respectively.

Excluding random, color, and charge, all remaining criteria should be selected regarding the aromatic contact. Thus, more than 18.8% of respondents chose at least one of the right rules. 5 (18.5%) users said that they did not learn to identify the interaction (random), and 11 (40.7%) stated that the color helped in choosing the right option. Three respondents still add an important criterion when selecting an interaction as aromatic: the presence of aromatic rings. This fact may have contributed to the high accuracy of this contact in the game.

The characterization of non-contacts could be done by exclusion. When none of the criteria identified one of the interactions as mentioned earlier, the players were supposed to select the non-contact option in the game. Thus, criteria starting at a distance and ending in angle should be taken into consideration in this case. One of the users stated by choosing criterion others: "*It took me a long time to realize that this option existed. When I noticed it, I started to select the non-contact option for those who did not follow the pattern of other interactions*".

In general, respondents chose the color as essential in all interactions (see Figure 4.4), especially in the hydrogen bond interaction. The color is not at all wrong when the coloring of the atoms comes from the chemical elements (atomic types). Another highlighted criterion is distance, showing users perceived that a distance threshold is relevant to distinguish one interaction from another. Players also noted that the charge of the atoms is necessary to assign an interaction as ionic; so is the polarity in the hydrophobic interaction. The covalent bonds criterion is also critical in all interactions, and it was well assigned by the majority of users. The angle between two interacting atoms is not shown in the Proteingo game. It might be the reason for the lack of this criterion regarding hydrogen bond, aromatic stacking, as well as the non-contact case. In sum, it seems users learned how to identify the patterns followed by each interaction.

Summary: 55.6% of the users considered the distance as the most relevant criterion in HB interactions, and 2.2% randomly played. 6 out of 27 respondents randomly played regarding HP interactions, and only 29.6% of the users chose the criterion polarity as relevant to choose HP option in the game. 51.9% and 44.4% correctly chose criteria charge and distance, respectively to form IO contacts. 3 out of 27 respondents added the presence of aromatic rings when selecting an interaction as AR. This fact could be the reason for high accuracy of AR contact in the game. In sum, respondents chose color criterion as essential in all interactions, especially in HB contacts.

We performed a Wilcoxon signed-rank test to verify the differences between the first and last ten plays of the users according to each interaction present in the Proteingo game. The results of the experiment together with results of descriptive statistics are shown in the Table 4.2. Hydrogen Bond and Ionic interactions stand out (p-values in bold). Thus, with a significance level of 0.05, it can be stated that the first and last ten plays of the users are different. It is also observed (see Table 4.2) that both means increased (from 5.92 to 6.80 in the hydrogen bonds, from 4.86 to 5.96 in the ionic interactions). Then, we can conclude that the average accuracy of the users increased in their final moves, giving strong indications that the hydrogen bonds and ionic contacts were the most learned by the Proteingo players.

Table 4.2 still shows that both, averages and medians increased in the comparison of the first ten with the last ten plays of the users (except for the aromatic contact). The AR contacts were the most successful among users, with a final average of 8.33 and a median of 9 successful contacts. The identification of non-contacts was the most difficult (final mean of only 4.30 and final median of 5.00 as seen in Table 4.2). HP contacts were well-established, different from IO (mean 4.83 to 5.96 and median 4.50 to 5.50). A change in the Proteingo game (perhaps the addition of tips along the game) could contribute to the identification of IO contacts and non-contact.

Table 4.2: **Users 10 Initial and Final Plays per Interaction**

Interaction	Before (mean)	After (mean)	Before (median)	After (median)	P-value
Hydrogen Bond	5.92	6.80	7.00	8.00	*0.04118*
Hydrophobic	6.35	7.17	8.00	8.00	0.35500
Ionic	4.83	5.96	4.50	5.50	*0.03373*
Aromatic	8.33	8.33	9.00	9.00	0.88460
Non-Contact	3.91	4.30	4.00	5.00	0.90900

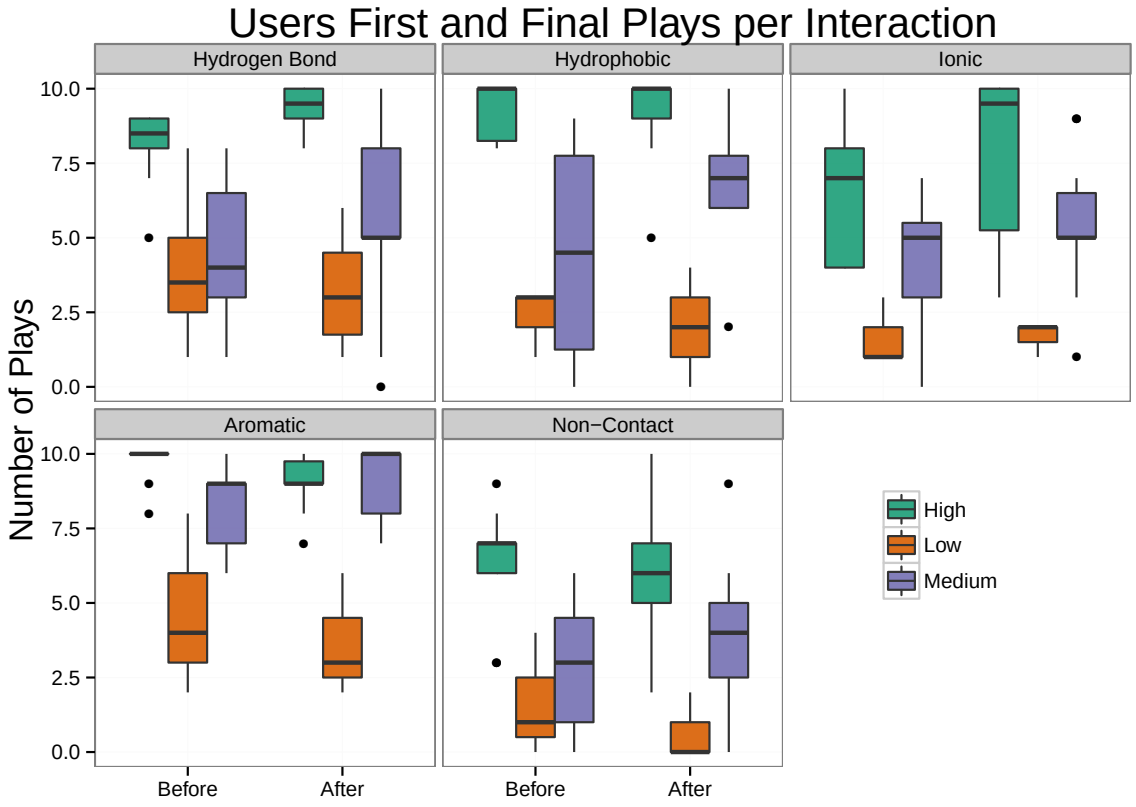
Wilcoxon signed-rank test to check the average differences of the first 10 plays (Before) with the last 10 plays (After) per interaction. For those who performed exactly 10 plays, we took these plays as the ten first and last plays for them.

Source: Personal collection.

The boxplot of Figure 4.5 illustrates the development of users according to the first 10 (Before) and last ten plays (After). The low-level users did not have a significant improvement in the initial and final plays, especially in the non-contact analysis, where 50% of these users hit less than three interactions. The High-level users declined in plays involving aromatic and non-contact. Nevertheless, the medians of correct answers for these users remained high in the analysis of the aromatic contacts; apparently the same can be observed in all other contacts. We once again believe that the Medium-level users have benefited the most from the Proteingo game. Since, even in the identification of non-contact (apparently the most difficult to reach), Medium-level users had a median hit that rose from approximately three contacts to about four contacts in the first and last ten plays, respectively.

Among the ten players who evaluated Q8, (see Table 3.2 page 25) and Figure 4.1) on a scale of 1-3 (completely disagree to neutral), 40% of them are high-level users and are included in the Middle-Great Experience and Great Experience groups. It may be the reason why they did not feel they were learning because they already were experts relating to the interactions analyzed. The other 60% are low and medium-level users, and none of them belongs to the Great Experience group. Of these six users, we highlight user 62 who performed more than 1,400 plays, increasing his accuracy and reinforcing the incidental learning of the non-bonded interactions in the game. Players 70, 74, 11, 43, 32 and 16 did not perform more than 100 plays. Hence, we believed they did not have the opportunity to learn all the contacts in the game, and consequently evaluated Q8 with low grades.

Figure 4.5: **Users 10 Initial and Final Plays per Interaction.** Before corresponds to the first 10 plays and After corresponds to the last 10 plays of the users. This figure illustrates boxplots according to user levels: High, Medium, and Low. 0 means that all contacts were incorrectly classified, and 10 means that all contacts were correctly identified.



Source: Personal collection.

Summary: With a significance level of 0.05, it can be stated that the first and last ten plays of the users are different regarding HB and IO interactions. AR interaction was the interaction most correctly assigned by the players. On the other hand, NC was the most difficult one (final mean of 4.30 and final median of 5.00 in 10 plays). The low-level users did not have a significant improvement in the initial and final plays. Players who did not perform more than 100 plays: 70, 74, 11, 43, 32 and 16. Player 62 effected more than 1400 plays, and his/her increase of accuracy reinforces the incidental learning of the interactions.

4.2 Data mining outcomes

This section presents the results of the evaluation and comparison of the PTS (Piccolo Training Set) and UTS (Users Training Set) models. The features selection through Multinomial Logistic Regression (MLR) is shown, and the rules generated by the decision trees of the models are compared. PTS and UTS performances according to the evaluation metrics are also compared. In addition, an experiment was carried out to verify the agreement of the Piccolo, users, and specialists.

4.2.1 PTS and UTS models evaluation and comparison outcomes

The features selection through the MLR is described in this section. Besides dimensionality reduction of the models, feature selection may also enhance generalization by reducing overfitting in the training step (it can create a more generic final model). Thus, we had a total of 13 attributes for the classification of contacts, but the MLR showed that only 5 of these attributes are significant for the PTS and UTS models.

Table 4.3 shows the likelihood ratio tests for all attributes used in our dataset which were submitted for the feature selection. We can see from Table 4.3 that only the attribute distance in the UTS is statistically significant to predict the analyzed interactions (p-value < 0.05). On the other hand, the remaining features are not statistically significant in this first analysis (all have a significance value of 1.000). The attributes regarding the polarity and types of the residues (which are independent variables) are perfectly collinear. It is the reason why we see several "dots" in the Sig. column of PTS and UTS models. Therefore, except for the variable distance, there is no statistical evidence that the remaining variables matter to the classification of HB, HP, IO, and AR interactions.

It is known that variables that have a high degree of variability may impair the classification of instances. Among the features used in our work, variables that can assume the largest number of values are presented in the following order: p1_atname/p2_atname (variability of 33), p1_resname/p2_resname (variability 20), and type_res1/type_res2 (variability of five). Hence, such attributes were removed from our analysis. Of the remaining variables, polarity_res1 and polarity_res2 were not statistically significant for the models (p-values of 0.138 and 0.972 for polarity_res1 and polarity_res2, respectively, in PTS; and polarity_res1 p-value = 1 and polarity_res2 p-value = 0.994 in UTS model). Hence, the polarities of both residues were also removed. After the performed features selection, the remaining variables with their level of significance are shown in Table 4.4 The goodness-of-fit reached a statistically

significant result for Pearson's significance. Knowing that a p-value < 0.05 indicates that the model did not fit the data well, our level of significance for PTS and UTS are both 1,000, and is, therefore, not statistically significant. Based on this measure, we can conclude that the model fits the data well [70]. Also, the model fitting information tells us whether any of the coefficients are statistically significant. Thus, it is expected that the level of significance of this measure is less than 0.05. In our case, we have a p-value < 0.05 for both Piccolo and UTS. It means that the model significantly predicts the dependent variable better than the intercept-only model (Null model) alone [70].

Table 4.3: **MLR - Likelihood Ratio Tests**

Feature	**PTS Sig.**	**UTS Sig.**
valence_at1	1.000	1.000
valence_at2	1.000	1.000
distance	0.136	*0.000*
p1_resname	1.000	1.000
polarity_res1	.	.
type_res1	.	.
p1_atname	1.000	1.000
charge_at1	1.000	1.000
p2_resname	1.000	1.000
polarity_res2	.	.
type_res2	.	.
p2_atname	1.000	1.000
charge_at2	1.000	1.000

The null hypothesis is that all parameters of that effect are 0. The "dots" in the Sig. column denotes a high correlation between the attributes regarding polarity and type of residue.

Source: Personal collection.

Table 4.4: **MLR - Likelihood Ratio Tests (Simplified)**

Feature	PTS	UTS
	Sig.	Sig.
valence_at1	0.000	0.000
valence_at2	0.000	0.000
distance	0.000	0.000
charge_at1	0.000	0.000
charge_at2	0.000	0.000

The null hypothesis is that all parameters of that effect are 0.

Source: Personal collection.

The datasets with the selected features in the MLR test (distance, charge, and polarity) were then submitted to the VTJ48 algorithm. The total instances of each contact for both PTS and UTS are displayed in Table 4.5. As for the optimization of parameters, both PTS and UTS obtained the best accuracy results with a $C = 0.5$ (pruning confidence) and a $M = 2$ (minimum number of instances per leaf). Both models were validated by 10-fold cross-validation and reached an average accuracy of 98.33% and 97.87% contacts for PTS and UTS, respectively. Tables 4.6, 4.7, 4.8, and 4.9 show the comparison of the rules generated by the PTS and UTS decision tree for the HB, HP, IO, and AR contacts, respectively. The colored cells indicate the rules shared by PTS and UTS models.

Table 4.5: **Number of instances used in each step of the data mining process**

Interaction	PTS	UTS	Test set
HB	180	180	20
HP	180	175	20
IO	180	164	20
AR	180	185	20
NC	180	125	20

Source: Personal collection.

Table 4.6 shows that both PTS and UTS generated a total of four rules in the definition of hydrogen bond contacts. Rule 1 classifies 95.56% of the HB contacts in PTS and 94.44% in UTS, being the rule with more relevance in the classification of HB according to the models. For Rule 1, UTS was more specific when it checks contacts that occur at a distance less than or equal to 3.28Å. Although HB interactions occur

at a short distance, this rule for both models does not match the rule adopted by the Piccolo, which is 3.8Å. In this case, the rule created by the PTS model is more precise when it checks the valence of the second atom as being less than or equal to three. PTS's Rule 1 is emphasizing that one of the atoms must necessarily be a nitrogen or oxygen (since they do not assume a valence higher than three). Rule 2 seems incomplete, once for the PTS it takes into account only the valence of the second atom and charge of only the first atom in the interaction. Once again, UTS is further specific when it considers only contacts that happens with a maximum distance of 3.28Å, and only the valence of the first atom matters. Both models have one rule in common (Rule 3). However, it does not classify more than three instances in both models. This rule would correctly classify a hydrogen bond that may occur at a distance greater than 3.46Å. For the valencies of the atoms must be less than three (i.e., N and O), and being the charge of the first negative atom, the charge of the second atom will necessarily be neutral, although the model has said that such charge can also be negative (charge_at2 < 0). Rule 4 of the PTS model is similar to Rule 3, but it is more precise when it specifies that the second atom of the contact must be neutral, being the first one negatively charged. Rule 4 of the UTS model has classified only a single instance and is irrelevant in this case.

Table 4.7 shows the rules for the classification of hydrophobic contacts. In this case, PTS generated only two rules, and UTS generated three rules to classify such an interaction. Rule 5 classifies 143 instances of HP-type, while Rule 6 does not have more than 31 instances in the leaves. Rule 5 is almost complete, missing only the valence and charges of the second atom in contact. Still, atoms with valence greater than three can only be carbons, and charges smaller than or equal to 0, include the neutral atoms (i.e., carbons generally). In Rule 6, when the first atom has valence less than or equal to three (some carbons of the backbone) the second atom must also be a carbon (valence higher than three). Besides, Rule 6 also exclude atoms negatively charged (charge_at1 > -1). UTS still generated a rule to identify the HP contacts that occur in distances less than or equal to 3.46Å (Rule 7). Even though the valence_at1 > 2 (this includes carbons, oxygens, and nitrogens), the rule verifies whether the charge of the first atom is less than or equal to 0 (which include neutral atoms such as carbons). However, Rule 7 classify only two instances in UTS.

Table 4.6: **HB Rules**

PTS Rules	UTS Rules
Rule 1: IF distance <= 3.46 and valence_at2 <= 3 THEN HB Leaf: 172 instances	**Rule 1**: IF distance <= 3.46 and distance <= 3.28 THEN HB Leaf: 170 instances
Rule 2: IF distance <= 3.46 and valence_at2 >3 and charge_at1 >-1 THEN HB Leaf: 3 instances	**Rule 2**: IF distance <= 3.46 and distance <= 3.28 and valence_at1 <= 2 THEN HB Leaf: 5 instances
Rule 3: IF distance >3.46 and valence_at1 <= 3 and valence_at2 <= 3 and charge_at2 <= -1 and charge_at1 <= 0 THEN HB Leaf: 3 instances	**Rule 3**: IF distance >3.46 and valence_at1 <= 3 and valence_at2 <= 3 and charge_at2 <= -1 and charge_at1 <= 0 and THEN HB Leaf: 2 instances
Rule 4: IF distance >3.46 and valence_at1 <= 3 and valence_at2 <= 3 and charge_at2 >-1 and charge_at1 <= -1 and charge_at2 <= 0 THEN HB Leaf: 2 instances	**Rule 4**: IF distance <3.46 and distance >3.28 and charge_at1 >0 THEN HB Leaf: 1 instance

Source: Personal collection.

Table 4.7: **HP Rules**

PTS Rules	UTS Rules
Rule 5: IF distance >3.46 and valence_at1 >3 and charge_at1 <= 0 THEN HP Leaf: 143 instances	**Rule 5**: distance >3.46 and valence_at1 >3 and charge_at1 <= 0 THEN HP Leaf: 143 instances
Rule 6: IF distance >3.46 and valence_at1 <= 3 and valence_at2 >3 and charge_at1 >-1 THEN HP Leaf: 31 instances	**Rule 6:** distance >3.46 and valence_at1 <= 3 and valence_at2 >3 and charge_at1 >-1 THEN HP Leaf: 29 instances
-	**Rule 7**: distance <= 3.46 and distance >3.28 and valence_at1 >2 and charge at1 <= 0 THEN HP Leaf = 2 instances

Source: Personal collection.

The rules to identify IO contacts are exemplified in Table 4.8. PTS generated a total of five rules while UTS generated four rules. Being Rules 7, 9, and 10 in PTS the same Rules 8, 10, and 11 in UTS, respectively. Since ionic interactions occur more often between nitrogens and oxygens than between carbons, Rule 7 (PTS) and 8 (UTS) fits right well to identify IO contacts. It first verifies the distance between the atoms. Then the two atoms must be N or O and the first and second atoms must be negatively and positively charged. The difference between PTS's Rule 8 and UTS's Rule 9, is that the first rule specifies that the second atom in contact should be positive when the first one is negatively charged. Rules 9 and 10 (PTS and UTS, respectively) are incomplete once it checks only the first atoms of the interaction to classify the IO interactions. Rules 10 and 11 of PTS and UTS, respectively, are also incomplete because they do not take into account the charge of the second atom of the interaction. PTS's Rule 11 identify the contacts that occur in distances less than or equal to 3.46Å, but it only classifies three instances.

Table 4.8: **IO Rules**

PTS Rules	UTS Rules
Rule 7: IF distance >3.46 and valence_at1 <= 3 and valence_at2 <= 3 and charge_at2 <-1 and charge_at1 >0 and THEN IO Leaf: 63 instances	**Rule 8**: IF distance >3.46 and valence_at1 <= 3 and valence_at2 <= 3 and charge_at2 <= -1 and charge_at1 >0 and THEN IO Leaf: 58 instances
Rule 8: IF distance >3.46 and valence_at1 <= 3 and valence_at2 <= 3 and charge_at2 >-1 and charge_at1 <= -1 and charge_at2 >0 THEN IO Leaf: 59 instances	**Rule 9**: IF distance >3.46 and valence_at1 <= 3 and valence_at2 <= 3 and charge_at2 >-1 and charge_at1 <= -1 and THEN IO Leaf: 49 instances
Rule 9: IF distance >3.46 and valence_at1 >3 and charge_at1 >0 THEN IO Leaf: 33 instances	**Rule 10**: IF distance >3.46 and valence_at1 >3 and charge_at1 >0 THEN IO Leaf: 33 instances
Rule 10: IF distance >3.46 and valence_at1 <= 3 and valence_at2 >3 and charge_at1 <= -1 THEN IO Leaf: 31 instances	**Rule 11**: IF distance >3.46 and valence_at1 <= 3 and valence_at2 >3 and charge_at1 <= -1 THEN IO Leaf: 21 instances
Rule 11: IF distance <= 3.46 and valence_at2 >3 and charge_at1<= -1 THEN IO Leaf: 3 instances	-

Source: Personal collection.

Finally, both models have the same rule to predict aromatic interactions (see Table 4.9.) Since AR interactions occur only between carbons, Rule 12 somewhat includes nitrogens and oxygens in their rank (valence_at1 and valence_at2 <= 3), which would be incorrect. So if this rule considered only atoms with valence greater than three, it would be cohesive. Rule 12 also considers the charges of the atoms in contact as being either neutral or positive. The last rule would be right if it considered only neutral atoms (carbons) in AR contacts.

Table 4.9: **AR Rules**

PTS Rules	**UTS Rules**
Rule 12:	**Rule 12**:
IF distance >3.46 and	IF distance >3.46 and
valence_at1 <= 3 and	valence_at1 <= 3 and
valence_at2 <= 3 and	valence_at2 <= 3 and
charge_at2 >-1 and	charge_at2 >-1 and
charge_at1 >-1	charge_at1 >-1
THEN AR	THEN AR
Leaf: 179 instances	Leaf: 182 instances

Source: Personal collection.

Table 4.10 shows the contingency matrix for the held tests in the VTJ48 algorithm regarding PTS set. The two HB contacts predicted as IO followed the Rule 1 (see Table 4.6), and they have a distance of 3.75Å and 3.88Å between the atoms in contact. The two HB interactions that were classified as AR followed the rules generated by the folds (and do not appear in the rules of the final tree). The first rule is: 1) if the distance between atoms greater than 3.39Å, 2) if the charges of the first and second atoms higher than -1 (neutral or positive), and 3) if the valence of the atoms less than or equal to 3 $\rightarrow$ this interaction should be classified as AR. The second rule says: 1) if the distance > 3.41Å, 2) the charges of the first and second atoms> -1 (neutral or positive), and 3) valence of atoms <= 3 $\rightarrow$ such a contact also should be classified as AR. Two out of 180 HP interactions were predicted as HB because they also followed Rule 1 shown in Table 4.6. The remaining 4 HP contacts classified in the AR class followed Rule 12 exemplified in Table 4.6. Rule 1 from Table 4.6 also misclassified two IO and AR contacts. It happened because of the distance between the atoms in contact (3.12Å and 3.35Å, respectively).

Table 4.10: **Contingency Table - PTS and UTS models**

PTS				
Actual	**Predicted**			
	HB	HP	IO	AR
HB	176	0	2	2
HP	2	174	0	4
IO	1	0	179	0
AR	1	0	0	179

UTS				
Actual	**Predicted**			
	HB	HP	IO	AR
HB	174	0	4	2
HP	2	172	1	0
IO	3	0	160	1
AR	1	2	0	182

Source: Personal collection.

Respecting to UTS model, its contingency matrix is also shown in Table 4.10. Rule 9 (see in Table 4.8 that the system wrongly placed two HB and one HP interactions as belonging to IO class. Once again the rules among the folds wrongly classified other two HB interactions as being of IO-type. This rule dictates: 1) if the distance $>$ 3.46Å, 2) the valence_at2 <= 3, and 3) the charge_at2 <= -1, then this interaction is of type IO. Two rules among the folds also misclassified two HB contacts in the AR class. The first rule is: 1) if the distance $>$ 3.39Å, 2) the charges of the first and second atoms greater than -1, and 3) the valence of the atoms less than or equal to three $\rightarrow$ this interaction should be classified as AR. The second rule says: 1) if distance $>$ 3.41Å, 2) the charges of the first and second atoms higher than -1 (neutral or positive), and 3) valence of atoms <= 3 $\rightarrow$ such contact also should be classified as AR. Rule 1 (see Table 4.6) also misclassified five contacts (two HP, two IO, and one AR) as being hydrogen bonded. Finally, Rule 2 and 12 shown in Table 4.6 misclassified two IO interactions each.

Table 4.11 shows the precision, sensitivity, specificity, F-measure, and efficiency in the classification of the contacts of type HB, HP, IO, and AR. The average of these metrics is also displayed in Table Table 4.11. All the metric averages of both models (PTS and UTS) have reached a value higher than 97.7%, showing that both models are useful classifiers for assigning contacts. The HP-type contacts in PTS have reached 100% accuracy and specificity, showing that the model is excellent in distinguishing between interactions that are HP-like contacts from the type that is not HP. The UTS

model, on the other hand, obtained the lowest precision and sensitivity value (97.7% both). Because both models had the lowest precision and sensitivity values for HB interaction, the F-measure in both models was the lowest for this type of interaction (98.1% and 97.7%, for PTS and UTS, respectively). Regarding the efficiency of the models, we can see from Table 4.11 that the AR interactions reached the highest value in both PTS (99.2%) and UTS (98.9%). However, in general, both models are efficient in both sensitivity and specificity (PTS average efficiency of 98.7% and UTS average of 98.5%). By the evaluation results of the models, we can conclude that the model generated by the majority vote of the users is as good as the model generated according to Piccolo, in the classification of the contacts.

Table 4.11: **PTS and UTS validation and evaluation**

PTS Model Evaluation					
Interaction	**Metrics**				
	Precision	**Sensitivity**	**Specificity**	**F-measure**	**Efficiency**
HB	97.8%	97.8%	99.3%	97.8%	98.5%
HP	100.0%	96.7%	100.0%	98.3%	98.3%
IO	98.9%	99.4%	99.6%	99.2%	99.5%
AR	96.8%	99.4%	98.9%	98.1%	99.2%
Average	**98.1%**	**98.1%**	**99.4%**	**98.1%**	**98.7%**
UTS Model Evaluation					
Interaction	**Metrics**				
	Precision	**Sensitivity**	**Specificity**	**F-measure**	**Efficiency**
HB	96.7%	96.7%	98.9%	96.7%	97.8%
HP	98.9%	98.3%	99.5%	98.6%	98.9%
IO	97.0%	97.6%	99.1%	97.3%	98.3%
AR	98.4%	98.4%	99.4%	98.4%	98.9%
Average	**97.7%**	**97.7%**	**99.2%**	**97.7%**	**98.5%**

Source: Personal collection

Summary: We reduced the number of features from 13 to 5 after applying the MLR technique to our PTS and UTS bases. Both PTS and UTS had an average accuracy of 98.33% and 97.87%, respectively. PTS and UTS have only one rule in common regarding HB interactions. However, it does not classify more than three instances in both models. The models have two rules in common regarding the classification of HP-type contacts. Furthermore these rules together correctly classify 174 and 172 cases for PTS and UTS, respectively. Three rules are the same for the models in the prediction of IO interactions. All of these three are significant in the classification of cases in PTS and UTS. Both models have just one rule to identify AR interactions, and PTS correctly classifies 99.4%, while UTS classifies 98.4% of the instances. In sum, UTS is as good as PTS model in the classification of contacts.

4.2.2 Piccolo and users agreement outcomes

As described in the Materials and Methods chapter, the proposed experiment intended to evaluate how much the database obtained by the users' plays was similar to the Piccolo database. This experiment is important to discriminate the interactions correctly assigned by Piccolo and by the players. Thus, to verify users' agreement with Piccolo we generated the model with Piccolo's response base and tested the model with the users' answers base (majority vote). A total of 800 contacts were used in training (200 interactions of each type: HB, HP, IO, and AR); and 781 contacts were used to test the generated model (199 HB, 195 HP, 182 IO, 205 AR). 770 out of a total of 781 instances of the user base were correctly classified by the Piccolo-generated model (accuracy of 98.5%). This fact indicates that the Proteingo players' plays matched more than mismatched with the interactions classified by Piccolo. Table 4.12 shows the contingency matrix of the interactions classification.

Table 4.12: **Contingency table- PTS training and UTS validation.** A total of 200 instances was used in training and 181 instances as a test of the generated model. The model misclassified a total of 11 contacts (indicated in bold).

Actual	**Predicted**			
	HB	HP	IO	AR
HB	194	0	**4**	0
HP	**2**	192	**1**	0
IO	0	0	181	**1**
AR	**1**	**2**	0	202

Source: Personal collection.

From the contingency matrix, it is possible to notice that few interactions were in disagreement (an error less than 2%). Four contacts labeled as HB were misclassified as belonging to IO class. It is known that one of the differences between HB and IO contacts is the distance between the atoms of the interaction. Hence, the smaller the distance between them, the higher the chance of being an HB contact. Also, if the distance between atoms is greater than 3.41Å, the second atom is more positively charged, and the first atom is negative, this contact is also misclassified as IO, which happened with two of the four HB contacts. The model correctly classified 98.62% of the HP contacts, two being classified as HB and one as IO (see Table 4.12). The atoms of the first two interactions were 3.41Å and 3.31Å distant from each other, and so were misclassified as HB-type. For the only IO contact predicted as belonging to the AR class, the algorithm once again failed because of a deceptive rule generated by the decision tree. Finally, only three out of 205 AR contacts were erroneously classified by the algorithm; one as belonging to HB and two to HP. For the one predicted as HB, the problem was the distance between the atoms (which is 3.35Å, and therefore, less than to 3.41Å). Although the last two AR contacts have been misclassified as HP, it makes sense because the atoms of both contacts are carbons and the distance is favorable for the occurrence of HP.

A group of six specialists in the field of bioinformatics and molecular biochemistry analyzed the 11 misclassified contacts in the test phase. According to the majority vote of these specialists, we have cases in which Piccolo agreed with them, cases in which answers given by the users matched the specialists, and even cases in which Piccolo, users, and experts agreed on a specific contact, but the system misclassified them. Figure 4.6 illustrates a case of an agreement between the specialists and Piccolo. It is possible to observe that most users rated this contact erroneously as HB, which is not possible due to the carbon atom occluding this interaction. In the case shown in Figure 4.7, the contact between the Lys-NZ and the GLU-OE1 is of the HB type according to the majority of users, but Piccolo classifies it as being IO; in this case, the response of the specialists runs against the majority of the Proteingo players. Finally, a case of the system misclassification is shown in Figure 4.8, where Piccolo, users, and specialists classify the interaction between the CB of Tyr and the CZ2 of Trp as HP, but one of the rules generated by the model labeled it as HB type.

Figure 4.6: **Interaction 890.** This is a case where Specialists and Piccolo agree on a particular contact.[1]

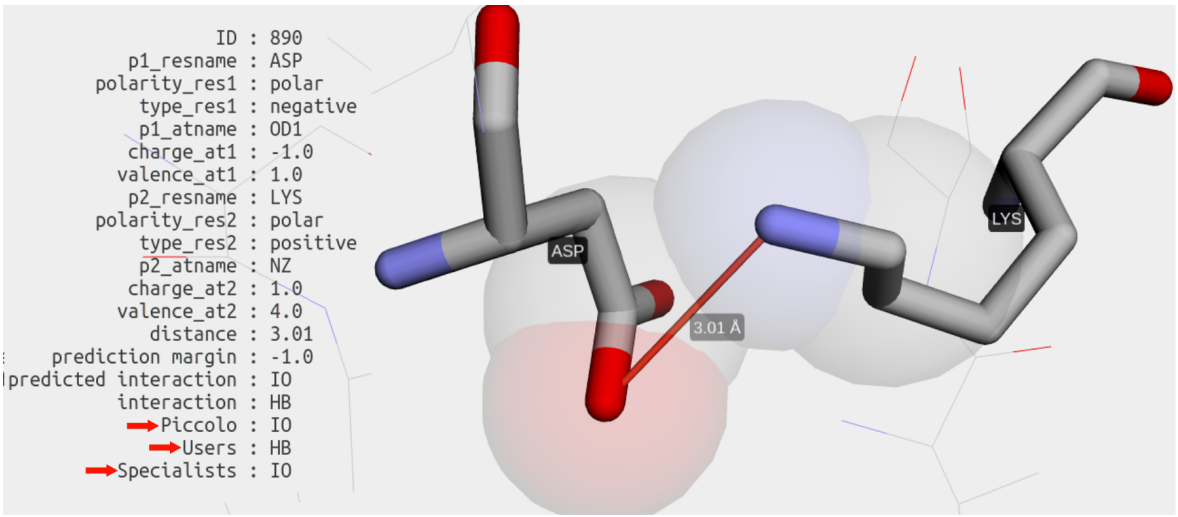
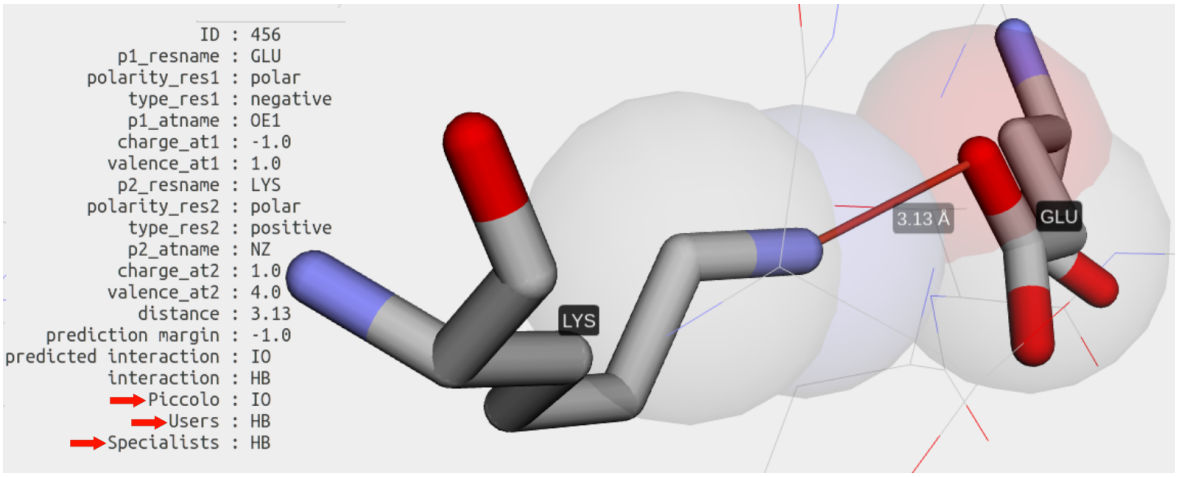


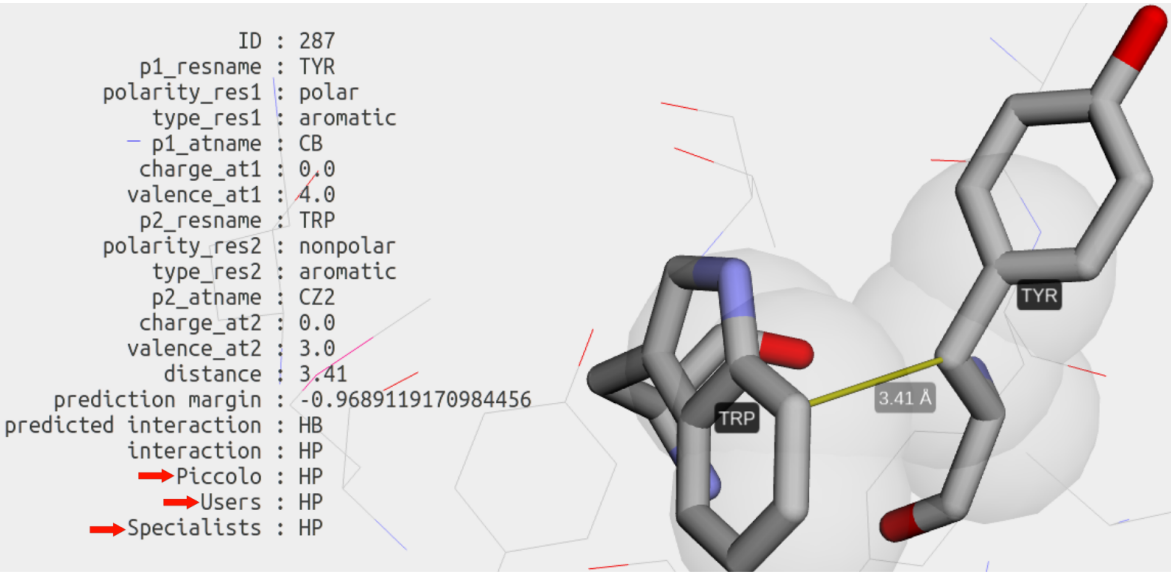
Figure 4.7: **Interaction 456.** This is a case where Specialists and Proteingo users agree on a particular contact.[2]



[1]The 3D visualization of this interaction is available on `http://bioinfo.dcc.ufmg.br/Proteingo/?id=890`.

[2]The 3D visualization of this interaction is available on `http://bioinfo.dcc.ufmg.br/Proteingo/?id=456`.

Figure 4.8: **Interaction 287.** This is a case where Specialists, Proteingo users, and Piccolo agree on a particular contact. However, the model misclassified this interaction.[3]



Regarding the model evaluation metrics, Table 4.13 shows the precision, sensitivity, specificity, F-measure, and efficiency in the classification of HB, HP, IO, and AR type contacts. The average value for each type of contact is also shown in this table. It is possible to notice that on average all metrics have reached a value higher than 98%. Therefore, the model has a high capacity to identify the true positives (sensitivity average=98.4%), and also a high potential to indicate the true negatives (average specificity=99.5%). Besides, the system presents 98.4%, 98.5%, and 99% of Precision, F-measure, and efficiency, respectively. Therefore, it can be assumed that the model has a high classification capacity. From the 11 contacts misclassified in the test step, seven were correctly assigned by UTS, taking into account the majority vote of the specialists. Thus, the manually curated database, built through the users' interactions with Proteingo, was able to identify contacts that were mistakenly classified by Piccolo.

[3]The 3D visualization of this interaction is available on `http://bioinfo.dcc.ufmg.br/Proteingo/?id=287`.

Summary: Only 12 out of 181 were misclassified when training with Piccolo database and testing with users database (error less than 2%). According to 6 specialists, Piccolo hit 4 out of the 11 interactions. Users and specialists agreed on 7 interactions. Finally, the system misclassified a total of 2 interactions. In sum, the manually cured through database built with the users' plays was able to identify contacts that were mistakenly classified by Piccolo.

Table 4.13: **PTS and UTS models validation and evaluation**

Interaction	Metrics				
	Precision	**Sensitivity**	**Specificity**	**F-measure**	**Efficiency**
HB	98.5%	97.5%	99.5%	98.0%	98.5%
HP	99.0%	98.5%	99.5%	98.7%	99.0%
IO	96.8%	99.5%	99.0%	98.1%	99.2%
AR	99.5%	98.5%	99.8%	99.0%	99.2%
Average	**98.4%**	**98.5%**	**99.5%**	**98.5%**	**99.0%**

Source: Personal collection.

4.3 Non-Contact data mining characterization outcomes

The analysis of the Non-Contact (NC) was held separately from the other contacts because it is not a type of contact as the HB, HP, IO, and AR. NC should be chosen on Proteingo when the analyzed contact does not follow any of the criteria for the mentioned contacts. However, NC was submitted to the same experiments of evaluation and comparison as the contacts mentioned earlier.

4.3.1 Non-Contact - Piccolo and users models comparison

This experiment has the purpose of comparing PTS and UTS models according to the decision trees respecting to what is a contact (IC) and what is not (NC). The metrics used to evaluate PTS and UTS models' performance is illustrated in Figure 4.9. We can see from the Figure 4.9 that both PTS and UTS models presented a low achievement in the classification of NCs, reaching only a sensitivity of 12.2% in PTS and 11.2% in UTS (before applying the multinomial logistic regression). The only metric with high rate in the prediction of NCs in both models is the specificity (97.8%

and 99.5% in PTS and UTS, respectively). The maximum values reached by the models in the IC classifications are the values of sensitivity, precision, and F-measure (97.4%, 81.6%, and 99.5% for PTS; and 99.5%, 87.0%, and 92.8% for UTS, respectively). The average performance of all metrics before feature selection approach did not exceed 67.6% in PTS and 82.4% in UTS. Besides, only 22 out of 180 NC instances were correctly classified in PTS, and only 14 out of 125 NC were correctly classified in UTS. Thus, we can see that the high rate of classification of IC class caused the average accuracy reached by the models to be high before MLR.

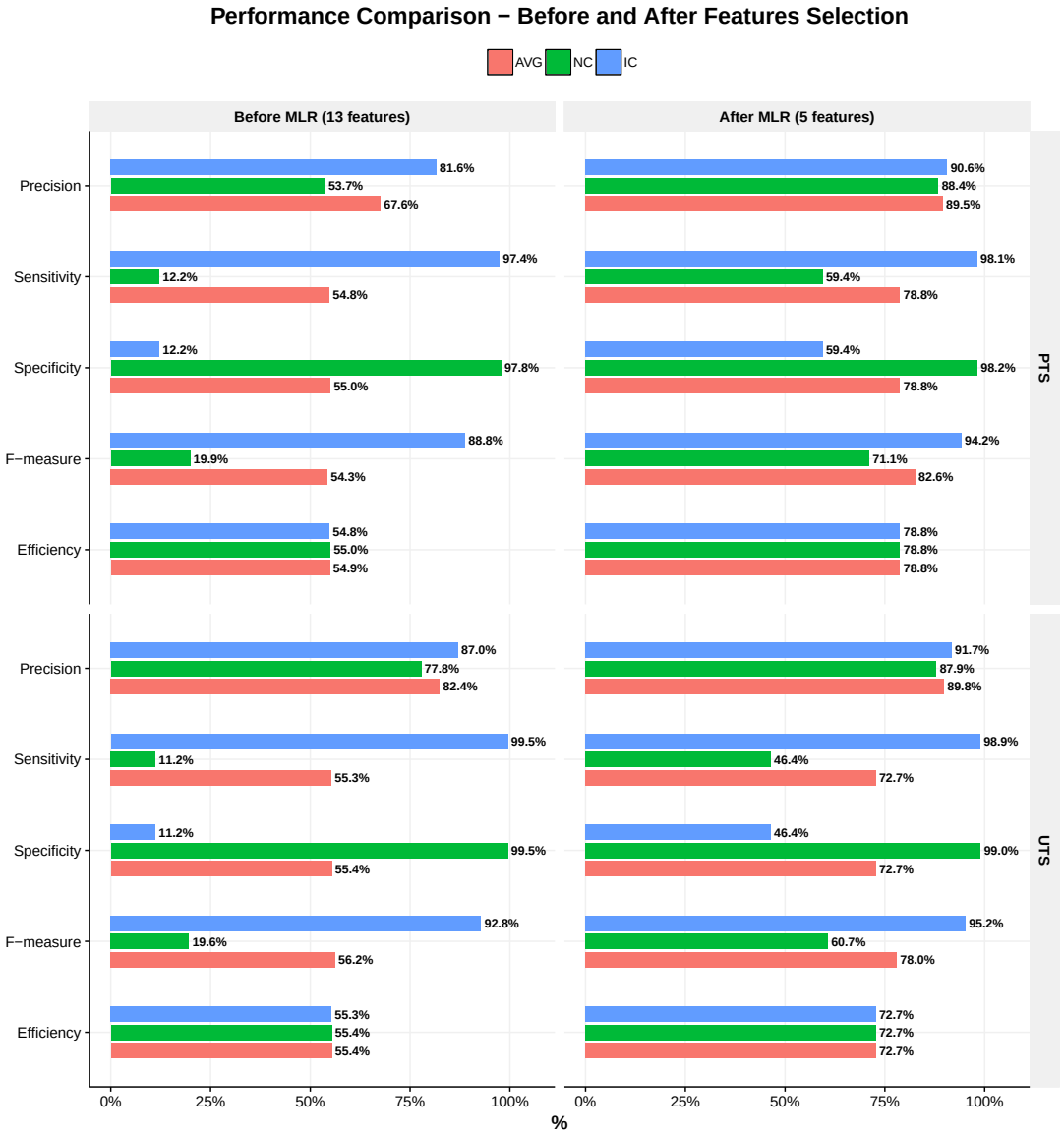


Figure 4.9: **Performance PTS and UTS Models Comparison According to Their Metrics. IC** stands for Is-Contact and NC is short for Non-Contact. **AVG** is the average of the metric between IC and NC

The features selected by the MLR approach (charge, valence, and distance) took part in the present experiment as well. The average accuracy in PTS rose from 80.3% to 90.3%, and from 86.9% to 91.4% in UTS model. Figure 4.9 also shows the performance metrics of the models in the present experiment. It is noticeable that all metrics had their values increased after eliminating some attributes that were not significant in the classification process. It is important to highlight that the improvement regarding the NC class in both models did not have a significant harmful impact on the performance of IC classification. In this experiment, the system correctly classified 107 out of 180 instances for the PTS (sensitivity of 59.4 %) model. Also, 58 out of 125 NC instances of UTS database were correctly predicted by the system (sensitivity of 46.4%). Although the performance of both models has raised after implementing the feature selection technique, its classification for non-contact instances should still be improved. The rules generated by PTS are shown in Tables x and y in Appendix B (IC and NC, respectively). Tables a and b (Appendix B) show the rules generated in the classification of non-contacts by the UTS model. The models, in this case, had no common rule (the opposite of what happened in the classification of hydrogen, hydrophobic, ionic and aromatic bonding contacts). Therefore, Although the performance of both models has raised after implementing the feature selection technique, its classification for non-contact instances should still be improved; so should the evaluation of other features to generate more humanly understandable rules.

Summary: The average performance of all metrics before feature selection approach did not exceed 67.6% in PTS and 82.4% in UTS. The average accuracy in PTS rose from 80.3% to 90.3%, and from 86.9% to 91.4% in UTS model after the feature selection with MLR. Furthermore, The average performance of all metrics was higher than 72.7%. Although the performance of both models has raised with the feature selection technique, its classification for non-contact instances should be improved.

4.3.2 Non-Contact - Piccolo and users agreement

A total of 1,000 interactions took part in the training phase of this experiment. Since 800 instances represent what Piccolo refers to as a contact (IC is the short form for Is Contact), and the remaining 200 represent the non-contact (NC). The test phase counted on 961 instances, where 827 were assigned as being an active interaction by the majority of the players (above 50% of the votes), the remaining 134 contacts were labeled as not being an interaction in the game. A binary classification was performed through the VTJ48 algorithm in WEKA environment, where we trained the decision

tree with the 1000 interactions according to Piccolo, and tested the referred model with the instances voted by the users throughout the plays. The model misclassified a total of 71 interactions. Table 4.14 shows the contingency matrix for this experiment and Table 4.15 exhibit the performance of the classifier according to the metrics of precision, sensitivity, specificity, f-measure, and efficiency.

Table 4.14: **NC Contingency Table - PTS and UTS agreement**

Actual	**Predicted**	
	IC	NC
IC	797	30
NC	41	93

Table 4.15: **NC Models Evaluation**

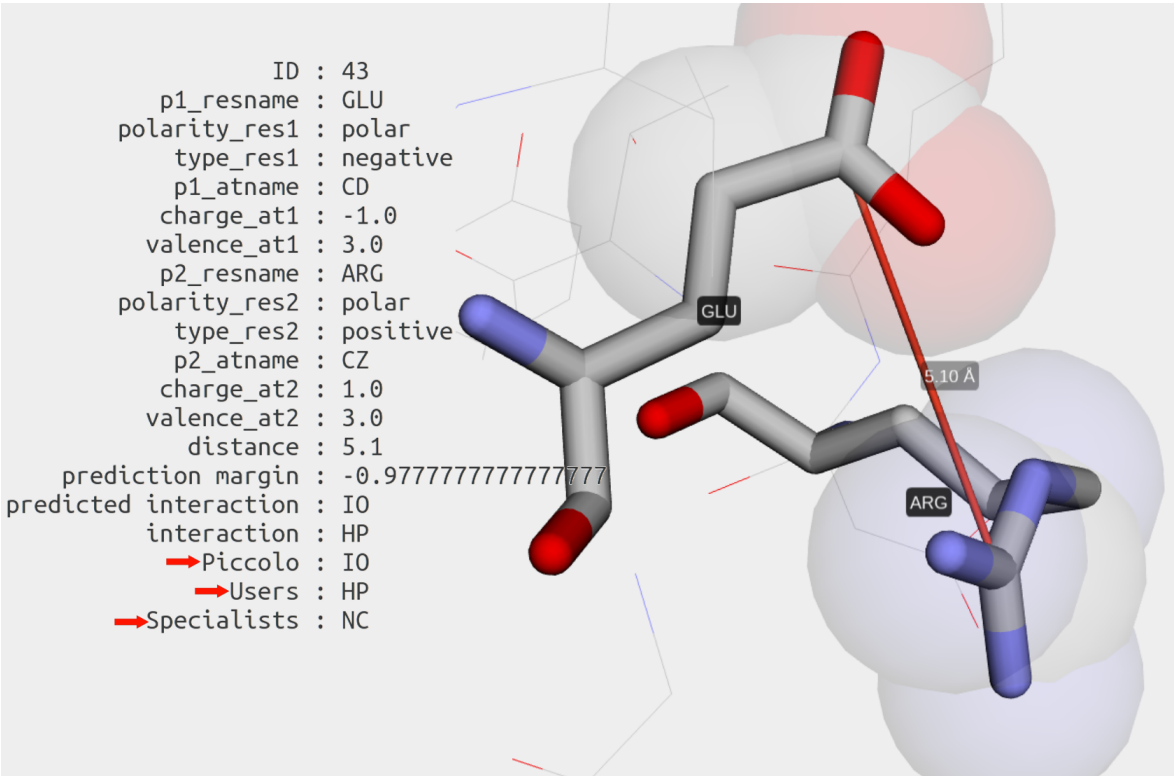
Class	**Metrics**				
	Precision	**Sensitivity**	**Specificity**	**F-measure**	**Efficiency**
IC	95.1%	96.4%	69.4%	95.7%	82.9%
NC	75.6%	69.4%	96.5%	72.4%	83.0%
Average	**85.4%**	**82.9%**	**83.0%**	**84.1%**	**82.9%**

From Table 4.14 we can notice that only 30 IC interactions were misclassified as NC. On the other hand, 93 instances of NC were correctly classified. This low rate of prediction respecting to NC explains the lower values of performance in Table 4.15. IC class were mostly agreed between Piccolo and Users answeres, as we see a precision of 95.1%, the sensitivity of 96.4%, and consequently, an F-measure of 95.7%. However, due to a large number of misclassified NC instances (41out of 134), the specificity related to the IC class reached a low value of 69.4%, and consequently, reached 82.% of efficiency. Even though the metrics regarding NC class are low (precision=85.4%, sensitivity=82.9%, F-measure=84.1%, and efficiency=82.9%), it still reached a specificity of 83.0%. It happened because only 30 instances of IC were misclassified as being NC. Therefore, users and Piccolo had a higher rate of agreement with what is a contact (IC), when compared to what is a Non-Contact. What also happened in identifying non-contacts by players.

Still, two cases where Piccolo signals the contacts as being ionic, most Proteingo users disagreed. Figure 4.10 illustrates the example of a possible ionic contact between the glutamate delta carbon and the arginine zeta carbon. In this case, the users classified this contact as being hydrophobic (which would not be incorrect, were the oxygen and nitrogen of said residues occluding such contact). For the contact shown in Figure

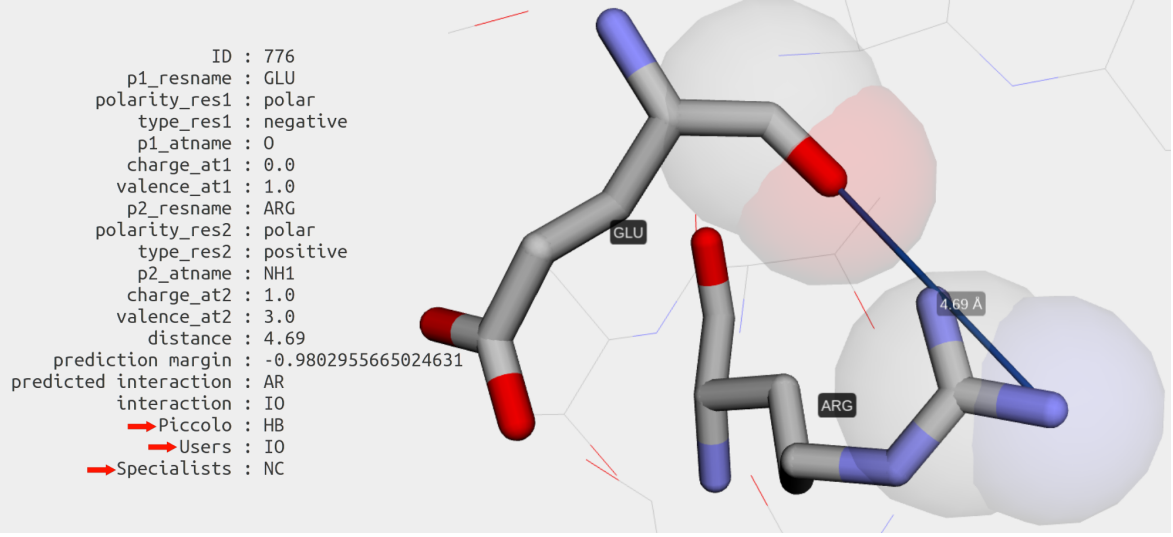
4.10, most experts rated it as a non-contact because of this occlusion. Furthermore, Figure 4.11 shows another case of occlusion not evaluated by Piccolo. Possibly, the oxygen of glutamate is in bonding the NH1 of arginine (according to Piccolo). However, the Proteingo players disagreed with the specialists that the referred contact is not a valid contact due to the nitrogen of the arginine occluding such interaction.

Figure 4.10: **Interaction 43.** This is a case where Specialists assigned a contact as Non-Contact.[4]



[4]The 3D visualization of this interaction is available on `http://bioinfo.dcc.ufmg.br/Proteingo/?id=43`.

Figure 4.11: **Interaction 776.** This is a case where Specialists assigned a contact as Non-Contact.[5]



Summary: The model misclassified a total of 71 out of 961 interactions. Due to a large number of misclassified NC instances (41 out of 134), the specificity related to the IC class reached a low value of 69.4%, and consequently, reached 82.9% of efficiency. Due to the high discordance with two contacts in the experiment, two mistakenly assigned contacts could be discovered.

[5]The 3D visualization of this interaction is available on `http://bioinfo.dcc.ufmg.br/Proteingo/?id=776`.

Chapter 5

Final Thoughts, Limitations, and Future Studies

5.1 Piccolo and user models evaluation (including non-contact)

As the main proposal of this dissertation is to propose a methodology for the construction and evaluation of a manually cured database for the classification of contacts between amino acids of protein chains, we reached the following conclusions. Proteingo was effective in creating the base of hydrogen bonding, hydrophobic, ionic, and aromatic bonds. Considering that the base created with the vote of the majority of the users had a performance as good as or even better than the one previously created by Piccolo. Regarding the characterization of non-contacts, we conclude that much can still be done aiming at its better classification. But we have already shown that the selection of characteristics through multinomial logistic regression was crucial for increasing the metrics of model evaluation in the identification of non-contacts.

Since the foundation of Proteingo has only 1,000 contacts, it would be of great importance to analyze a more significant number of contacts. Also, other experts, preferably who have not played Proteingo, could examine part of what was separated for testing to validate the PTS and UTS models and thus conclude whether, in a real test, UTS is as good as PTS in the classification of non-covalent bonds between atoms of residues in protein chains.

5.2 Proteingo evaluation

Respecting student's motivation, experience, and learning of molecular bonds that take place in protein-protein interactions. It can be concluded that players were highly motivated to play the game mainly because of its attractive interface and the manner which the interactions are displayed. As suspected, the majority of users affirmed that they had fun while playing proteingo; a feature that is widely approached by all sort of games [41, 49, 61]. From the respondents of the posttest questionnaire, it can also be established that players deeply acclaim the use of proteingo into the classroom as a tool for learning molecular bonds. About users learning, our experiment concludes that the users of the middle-great and middle-poor experience groups were the most benefited by the creation of the game. Furthermore, we find that the users were able to note the main criteria for choosing the hydrogen bonds and ionic interactions. Although the analysis of the differences of means of the first and last plays of the users concerning the hydrophobic interactions, aromatic stacks and non-contact was not significant; the mean and median hits were higher in the final plays of all users. Therefore DGBLs are recommended to retain students engaged and increase their interest in the learning of science.

One of the limitations found in our analysis is the lack of a control group to show the effectiveness of learning along the plays compared to traditional ways of learning (through lectures or from textbooks). Consequently, further validation of our results is needed. Another limitation is the sample in the present study. As the experiment was administered out of a classroom environment, not all players were willing to respond to the posttest questionnaire. Resulting in a considerable sample of only 27 users. Once reviewing the current limitations, a line for future work could be focused on developing further experiments into classrooms, as well as increasing the sample size and including a higher number of low experience users. Another improvement would be the use of this method to support the prospection of protein-protein interactions through the feedback given by the players.

Bibliography

[1] Jeremy M Berg, John L Tymoczko, and Lubert Stryer. Biochemistry. 5th. *New York: WH Freeman*, 38(894):76, 2002.

[2] Periannan Senapathy, Marvin B. Shapiro, and Nomi L. Harris. [16] splice junctions, branch point sites, and exons: Sequence statistics, identification, and applications to genome project. volume 183 of *Methods in Enzymology*, pages 252 – 278. Academic Press, 1990.

[3] Joel B Hagen. The origins of bioinformatics. *Nature Reviews Genetics*, 1(3):231–236, 2000.

[4] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235, 2000.

[5] Oleg B Ptitsyn and Kai-Li H Ting. Non-functional conserved residues in globins and their possible role as a folding nucleus1. *Journal of Molecular Biology*, 291(3):671 – 682, 1999.

[6] Andreas Martin Lisewski and Olivier Lichtarge. Rapid detection of similarity in protein structure and function through contact metric distances. *Nucleic Acids Research*, 34(22):e152–e152, 2006.

[7] Liisa Holm and Chris Sander. Protein structure comparison by alignment of distance matrices. *Journal of molecular biology*, 233(1):123–138, 1993.

[8] Frederic Middlebrook Richards. Protein stability: still an unsolved problem. *Cellular and Molecular Life Sciences CMLS*, 53(10):790–802, 1997.

[9] Ram Samudrala and John Moult. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of molecular biology*, 275(5):895–916, 1998.

[10] Carlos H. Da Silveira, Douglas E V Pires, Raquel C. Minardi, Cristina Ribeiro, Caio J M Veloso, Julio C D Lopes, Wagner Meira, Goran Neshich, Carlos H I Ramos, Raul Habesch, and Marcelo M. Santoro. Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins: Structure, Function and Bioinformatics*, 74(3):727–743, 2009.

[11] Frederic M Richards. The interpretation of protein structures: total volume, group volume distributions and packing density. *Journal of molecular biology*, 82(1):1–14, 1974.

[12] JL Finney. Volume occupation, environment and accessibility in proteins. the problem of the protein surface. *Journal of molecular biology*, 96(4):721–732, 1975.

[13] George R. Bickerton, Alicia P. Higueruelo, and Tom L. Blundell. Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: the piccolo database. *BMC Bioinformatics*, 12(1):313, 2011.

[14] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popovic, and Foldit Players. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760, August 2010.

[15] Jean-Claude Bradley, Robert J. Lancashire, Andrew SID Lang, and Antony J. Williams. The spectral game: leveraging open data and crowdsourcing for education. *Journal of Cheminformatics*, 1(1):9, 2009.

[16] Cristina Alonso-Fernandez, Antonio Calvo-Morata, Manuel Freire, Ivan Martinez-Ortiz, and Baltasar Fernandez-Manjon. Systematizing game learning analytics for serious games. In *2017 IEEE Global Engineering Education Conference (EDUCON)*, pages 1106–1113, April 2017.

[17] Juan Antonio Caballero-Hernández, Manuel Palomo-Duarte, and Juan Manuel Dodero. Skill assessment in learning experiences based on serious games: A systematic mapping study. *Computers & Education*, 113:42 – 60, 2017.

[18] Sylke Vandercruysse, Mieke Vandewaetere, and Geraldine Clarebout. Game-based learning: A review on the effectiveness of educational games. In *Handbook of Research on Serious Games as Educational, Business and Research Tools*, pages 628–647. IGI Global, 2012.

[19] Jeremy Esque, Christophe Oguey, and Alexandre G de Brevern. Comparative analysis of threshold and tessellation methods for determining protein contacts. *Journal of chemical information and modeling*, 51(2):493–507, 2011.

[20] Hao Ding, Ichigaku Takigawa, Hiroshi Mamitsuka, and Shanfeng Zhu. Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Briefings in Bioinformatics*, 15(5):734–747, 2014.

[21] D. L. Nelson and M. M. Cox. *Amino Acids, Peptides and Proteins*. New York : W.H. Freeman and Company, 2017.

[22] Geoffrey M Cooper and Robert E Hausman. *The Cell: A Molecular Approach 7th Edition*. Sinauer Associates Oxford University Press, 2017.

[23] Gautam R Desiraju and Thomas Steiner. *The weak hydrogen bond: in structural chemistry and biology*, volume 9. Oxford University Press on Demand, 2001.

[24] Ian K. McDonald and Janet M. Thornton. Satisfying hydrogen bonding potential in proteins. *Journal of Molecular Biology*, 238(5):777 – 793, 1994.

[25] Jungwun Hwang, Ping Li, William R. Carroll, Mark D. Smith, Perry J. Pellechia, and Ken D. Shimizu. Additivity of substituent effects in aromatic stacking interactions. *Journal of the American Chemical Society*, 136(40):14060–14067, 2014. PMID: 25238590.

[26] Adrian Schreyer and Tom Blundell. Credo: A protein–ligand interaction database for drug discovery. *Chemical Biology & Drug Design*, 73(2):157–167, 2009.

[27] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. Advances in knowledge discovery and data mining. chapter From Data Mining to Knowledge Discovery: An Overview, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.

[28] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[29] Steven J Darnell, David Page, and Julie C Mitchell. An automated decision-tree approach to predicting protein interaction hot spots. *Proteins: Structure, Function, and Bioinformatics*, 68(4):813–823, 2007.

[30] Mariana Tasca, Alexandre Plastino, Celso Ribeiro, and Bianca Zadrozny. A fast and effective strategy for feature selection in high-dimensional datasets. *Journal of Information and Data Management*, 7(2):155, 2017.

[31] Mahesh Pal. Multinomial logistic regression-based feature selection for hyperspectral data. *International Journal of Applied Earth Observation and Geoinformation*, 14(1):214 – 220, 2012.

[32] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[33] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[34] Yang Yang, Ryan N Lichtenwalter, Nitesh V Chawla, Y Yang, R N Lichtenwalter, and N V Chawla. Evaluating link prediction methods. *Knowledge and Information Systems*, 45:751–782, 2015.

[35] Meihua Qian and Karen R Clark. Game-based learning and 21st century skills: A review of recent research. *Computers in Human Behavior*, 63:50–58, 2016.

[36] Chih Wei Chao, Lei Chang, An-Chi Cheng, and Ting-Ting Wu. *Exploration on the Effectiveness of Learning, Interest, and Attitude of the Integration of Review System of History Based on Mobile Game and Forgetting Curve*, pages 34–42. Springer International Publishing, Cham, 2017.

[37] Marc Prensky. Digital game-based learning. *Computers in Entertainment (CIE)*, 1(1):21–21, 2003.

[38] John Kirriemuir and Angela McFarlane. Literature review in games and learning. 2004.

[39] Mark R Lepper and Thomas W Malone. Intrinsic motivation and instructional effectiveness in computer-based education. *Aptitude, learning, and instruction*, 3:255–286, 1987.

[40] Pieter Wouters and Herre van Oostendorp. *Overview of Instructional Techniques to Facilitate Learning and Motivation of Serious Games*, pages 1–16. Springer International Publishing, Cham, 2017.

[41] Martin Ebner and Andreas Holzinger. Successful implementation of user-centered game based learning in higher education: An example from civil engineering. *Computers & Education*, 49(3):873 – 890, 2007.

[42] Leonard A Annetta, Meng-Tzu Cheng, and Shawn Holmes. Assessing twenty-first century skills through a teacher created video game for high school biology students. *Research in Science & Technological Education*, 28(2):101–114, 2010.

[43] Wilk Oliveira Santos and Clovis Gomes Silva Junior. An introduction to educational games virtualization. pages 1403–1408. SBC - Proceedings of SBGames, September 2016.

[44] Troy D Sadler, William L Romine, Parker E Stuart, and Dominike Merle-Johnson. Game-based curricula in biology classes: Differential effects among varying academic levels. *Journal of Research in Science Teaching*, 50(4):479–499, 2013.

[45] Juho Hamari, David J Shernoff, Elizabeth Rowe, Brianno Coller, Jodi Asbell-Clarke, and Teon Edwards. Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior*, 54:170–179, 2016.

[46] Geoffrey T Crisp. Assessment in next generation learning spaces. In *The future of learning and teaching in next generation learning spaces*, pages 85–100. Emerald Group Publishing Limited, 2014.

[47] Carolina N. Spiegel, Gutemberg G. Alves, Tâ da S. Cardona, Leandra M. C. Melim, Mauricio R. M. Luz, Tania C. Araújo-Jorge, and Andrea Henriques-Pons. Discovering the cell: an educational game about cell and molecular biology. *Journal of Biological Education*, 43(1):27–36, 2008.

[48] Marina Casanoves, Zoel Salvadó, Ángel González, Cristina Valls, and Maria Teresa Novo. Learning genetics through a scientific inquiry game. *Journal of Biological Education*, 51(2):99–106, 2017.

[49] Leonard A Annetta, James Minogue, Shawn Y Holmes, and Meng-Tzu Cheng. Investigating the impact of video games on high school students' engagement and learning about genetics. *Computers & Education*, 53(1):74–85, 2009.

[50] Julia Winter, Michael Wentzel, and Sonia Ahluwalia. Chairs!: A mobile game for organic chemistry students to learn the ring flip of cyclohexane. *Journal of Chemical Education*, 93(9):1657–1659, 2016.

[51] José Nunes da Silva Júnior, Mary Anne Sousa Lima, João Victor Xerez Moreira, Francisco Serra Oliveira Alexandre, Diego Macedo de Almeida, Maria da Conceição Ferreira de Oliveira, and Antonio José Melo Leite Junior. Stereogame:

An interactive computer game that engages students in reviewing stereochemistry concepts. *Journal of Chemical Education*, 94(2):248–250, 2017.

[52] Jeehyung Lee, Wipapat Kladwang, Minjae Lee, Daniel Cantu, Martin Azizyan, Hanjoo Kim, Alex Limpaecher, Snehal Gaikwad, Sungroh Yoon, Adrien Treuille, Rhiju Das, and EteRNA Participants. Rna design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences*, 111(6):2122–2127, 2014.

[53] Alexander Kawrykow, Gary Roumanis, Alfred Kam, Daniel Kwak, Clarence Leung, Chu Wu, Eleyine Zarour, Phylo players, Luis Sarmenta, Mathieu Blanchette, and Jérôme Waldispühl. Phylo: A citizen science approach for improving multiple sequence alignment. *PLOS ONE*, 7(3):1–9, 03 2012.

[54] Carol A Rohl, Charlie EM Strauss, Kira MS Misura, and David Baker. Protein structure prediction using rosetta. *Methods in enzymology*, 383:66–93, 2004.

[55] Annakaisa Kultima. Casual game design values. In *Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era*, MindTrek '09, pages 58–65, New York, NY, USA, 2009. ACM.

[56] Nicholas Rego and David Koes. 3dmol. js: molecular visualization with webgl. *Bioinformatics*, 31(8):1322–1324, 2015.

[57] Mike Bostock. D3. js-data-driven documents (2016). *URL: https://d3js. org*, 2016.

[58] John M Keller. *Motivational design for learning and performance: The ARCS model approach*. Springer Science & Business Media, 2009.

[59] Wenhao Huang, Wenyeh Huang, Heidi Diefes-Dux, and Peter K. Imbrie. A preliminary validation of attention, relevance, confidence and satisfaction model-based instructional material motivational survey in a computer-based tutorial setting. *British Journal of Educational Technology*, 37(2):243–259, 2006.

[60] ISO DIS. 9241-210: 2010. ergonomics of human system interaction-part 210: Human-centred design for interactive systems. *International Standardization Organization (ISO). Switzerland*, 2009.

[61] J. Derek Lomas, Kenneth Koedinger, Nirmal Patel, Sharan Shodhan, Nikhil Poonwala, and Jodi L. Forlizzi. Is difficulty overrated?: The effects of choice, novelty and suspense on intrinsic motivation in educational games. In *Proceedings of the*

2017 CHI Conference on Human Factors in Computing Systems, CHI '17, pages 1028–1039, New York, NY, USA, 2017. ACM.

[62] Victoria J Marsick and Karen E Watkins. Informal and incidental learning. *New directions for adult and continuing education*, 2001(89):25–34, 2001.

[63] Toru Fujimoto and Christopher Michael Yap. The analysis of incidental learning in the affinity spaces of a smartphone game "neko atsume". In *International Symposium on Emerging Technologies for Education*, pages 3–13. Springer, 2016.

[64] Robert B Best, Xiao Zhu, Jihyun Shim, Pedro EM Lopes, Jeetain Mittal, Michael Feig, and Alexander D MacKerell Jr. Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone φ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of chemical theory and computation*, 8(9):3257–3273, 2013.

[65] Gregor Stiglic, Simon Kocbek, Igor Pernek, and Peter Kokol. Comprehensive decision tree models in bioinformatics. *PloS one*, 7(3):e33812, 2012.

[66] Jon Starkweather and Amanda Kay Moske. Multinomial logistic regression. *Consulted page at September 10th: http://www. unt. edu/rss/class/Jon/Benchmarks/MLR_JDS_Aug2011. pdf*, 29:2825–2830, 2011.

[67] J Ross Quinlan. Generating production rules from decision trees. In *ijcai*, volume 87, pages 304–307, 1987.

[68] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. pages 1137–1143, 1995.

[69] Marcos Felipe Martins Silva, Larissa Fernandes Leijoto, and Cristiane Neri Nobre. Algorithms analysis in adjusting the svm parameters: An approach in the prediction of protein function. *Applied Artificial Intelligence*, pages 1–16, 2017.

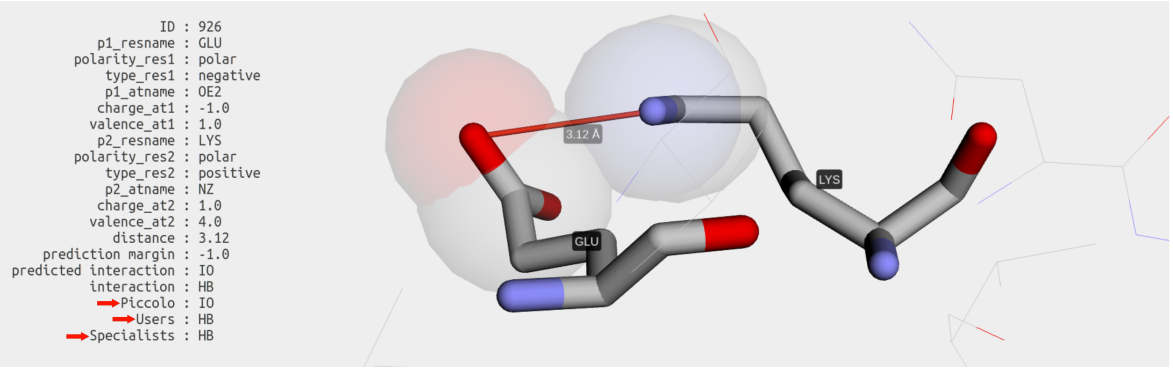
[70] Yap Hang Chan. Biostatistics 305. multinomial logistic regression. *Singapore medical journal*, 46(6):259, 2005.

Chapter 6

Appendix A

6.1 Cases where Proteingo users agree with the specialists

Figure 6.1: **Interaction 926.** This is a case where Specialists and Proteingo users agree on a particular contact.[1]



[1]The 3D visualization of this interaction is available on `http://bioinfo.dcc.ufmg.br/Proteingo/?id=926`

Figure 6.2: **Interaction 997.** This is a case where Specialists and Proteingo users agree on a particular contact.[2]

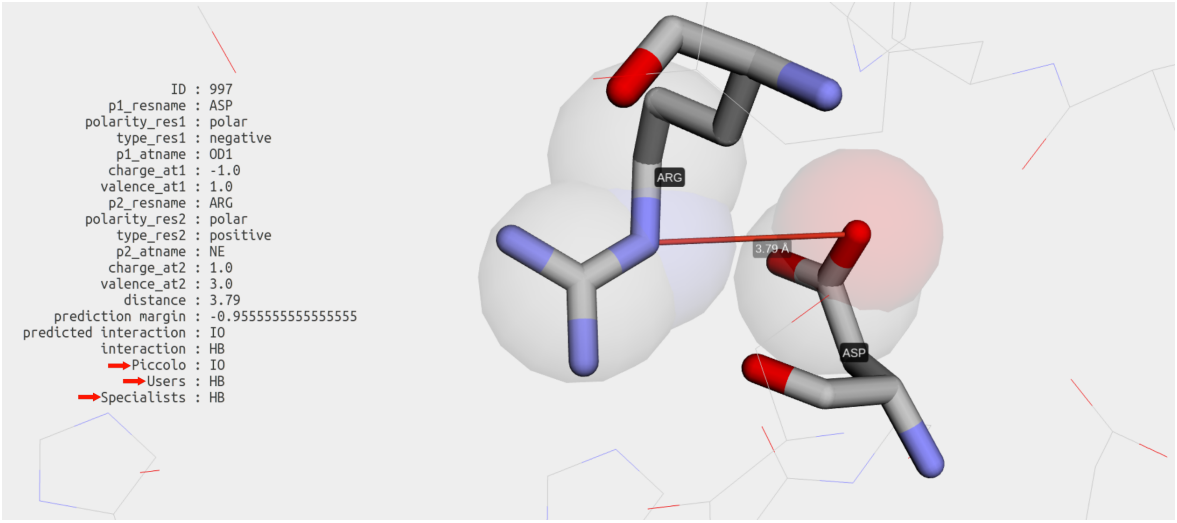
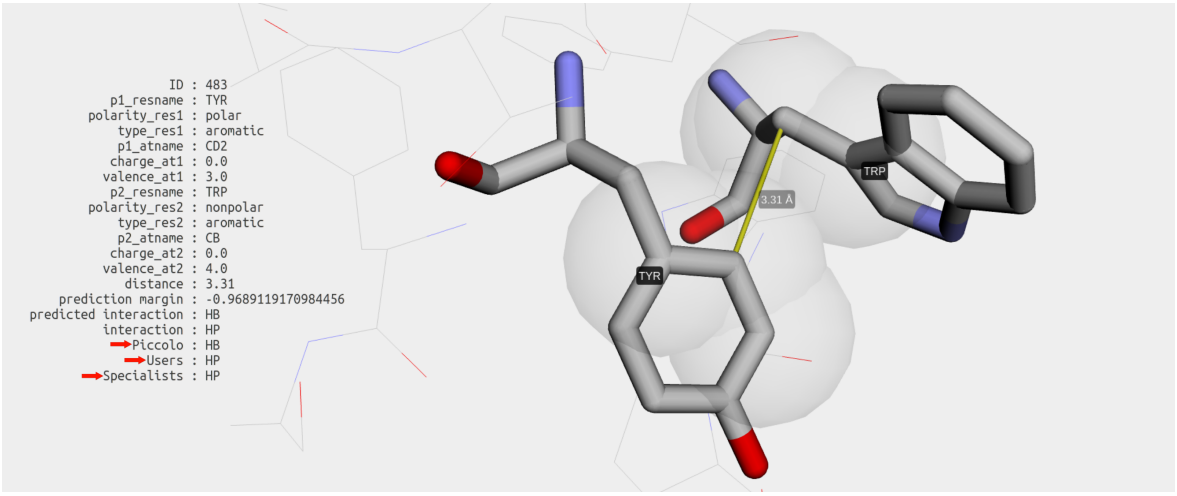


Figure 6.3: **Interaction 483.** This is a case where Specialists and Proteingo users agree on a particular contact.[3]



[2]The 3D visualization of this interaction is available on `http://bioinfo.dcc.ufmg.br/Proteingo/?id=997`

6.2 Cases where Piccolo agree with the specialists

Figure 6.4: **Interaction 673.** This is a case where Specialists and Piccolo agree on a particular contact.[4]

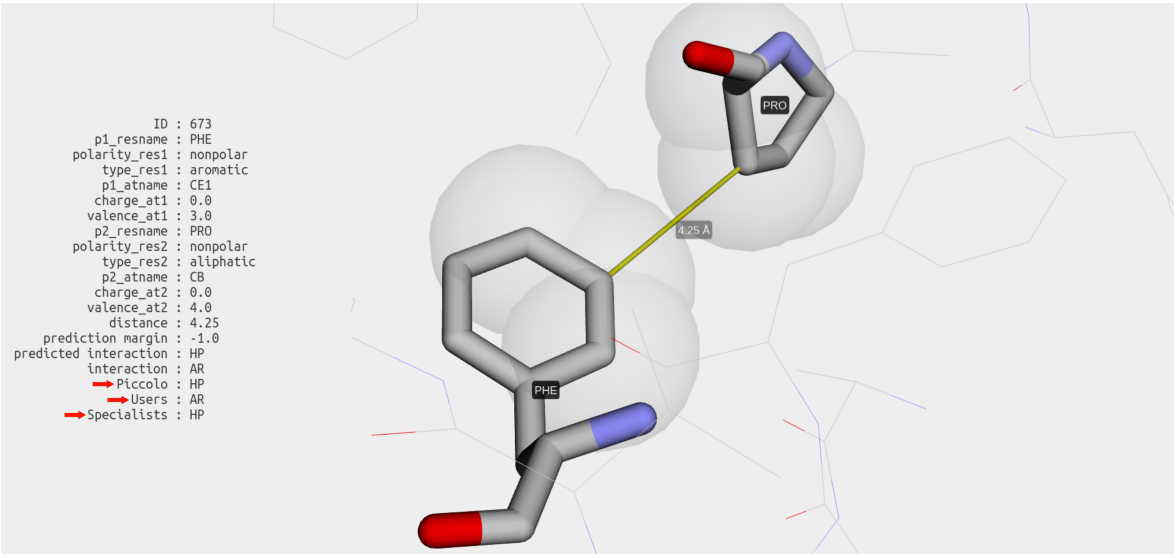
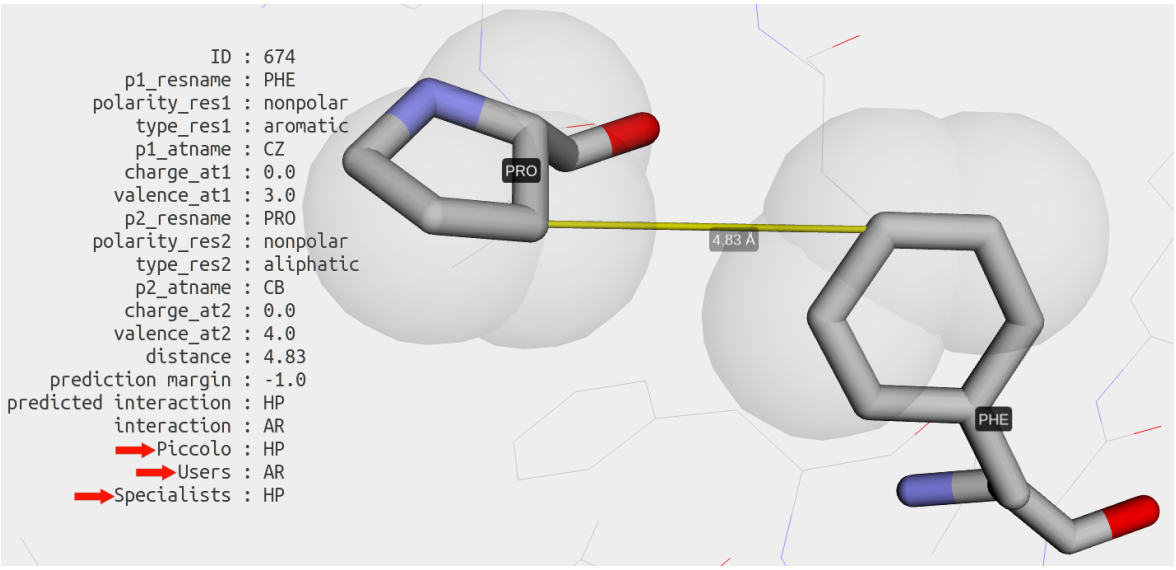


Figure 6.5: **Interaction 674.** This is a case where Specialists and Piccolo agree on a particular contact.[5]

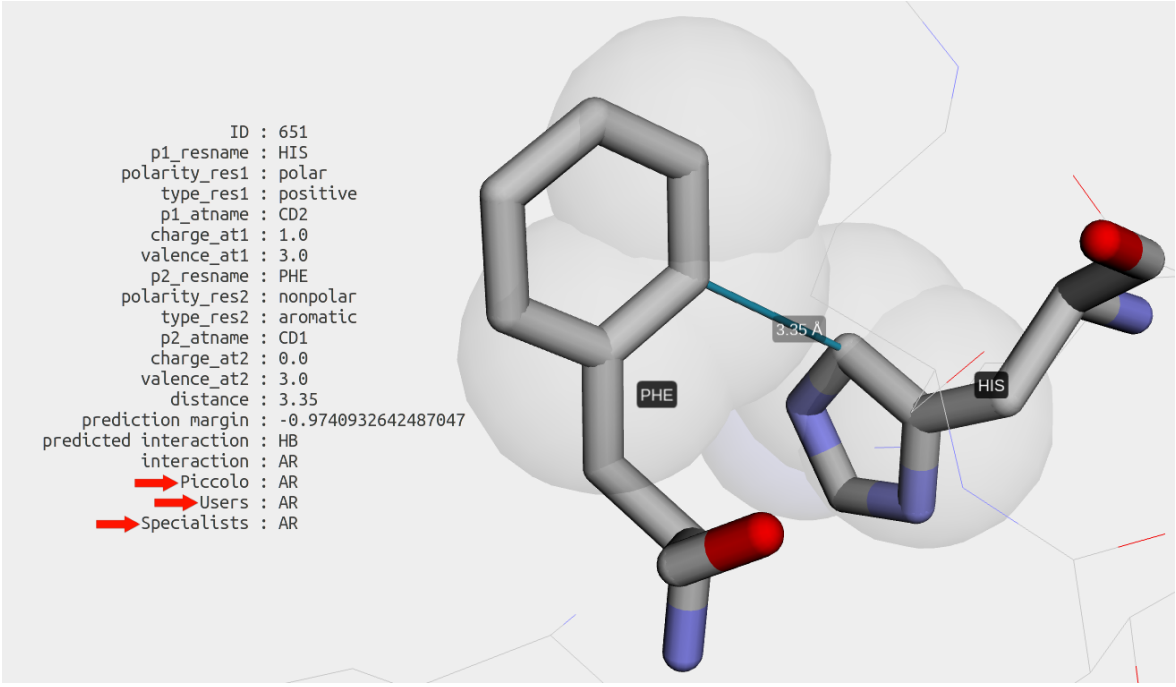


[3]The 3D visualization of this interaction is available on `http://bioinfo.dcc.ufmg.br/Proteingo/?id=483`

[4]The 3D visualization of this interaction is available on `http://bioinfo.dcc.ufmg.br/Proteingo/?id=673`

6.3 Case where the contact was misclassified by the system

Figure 6.6: **Interaction 651.** This is a case where Specialists, Proteingo users, and Piccolo agree on a particular contact. However, the model misclassified this interaction.[6]



[5]The 3D visualization of this interaction is available on `http://bioinfo.dcc.ufmg.br/Proteingo/?id=674`

[6]The 3D visualization of this interaction is available on `http://bioinfo.dcc.ufmg.br/Proteingo/?id=651`

Chapter 7

Appendix B

Table 7.1: **Rules created by PTS to categorize IC class**

IC PTS Rules		
Rule 1: IF distance <= 3.19 THEN IC Leaf = 162 instances	Rule 2: IF distance >3.19 and valence_at1 <= 2 and charge_at2 <= 0 and valence_at2 <= 2 and distance <= 4.14 THEN IC Leaf = 13 instances	Rule 3: IF distance >3.19 and valence_at1 <= 2 and charge_at2 >0 and charge_at1 <= -1 THEN IC Leaf = 52 instances
Rule 4: IF distance >3.19 and valence_at1 >2 and distance <= 4.82 THEN IC Leaf = 236 instances	Rule 5: IF distance >3.19 and valence_at1 >2 and distance >4.82 and valence_at2 <= 3 and valence_at1 <= 3 and valence_at2 <= 2 and charge_at2 <= -1 THEN IC Leaf = 23 instances	Rule 6: IF distance >3.19 and valence_at1 >2 and distance >4.82 and valence_at2 <= 3 and valence_at1 <= 3 and valence_at2 >2 THEN IC Leaf = 175 instances
Rule 7: IF distance >3.19 and valence_at1 >2 and distance >4.82 and valence_at2 <= 3 and valence_at1 >3 and charge_at1 <= 0 distance <= 5.03 THEN IC Leaf = 7 instances	Rule 8: IF distance >3.19 and valence_at1 >2 and distance >4.82 and valence_at2 <= 3 and valence_at1 >3 and charge_at1 >0 THEN IC Leaf = 21 instances	Rule 9: IF distance >3.19 and valence_at1 >2 and distance >4.82 and valence_at2 >3 and distance <= 4.99 THEN IC Leaf = 23 instances

Table 7.2: **Rules created by UTS to categorize IC class**

IC UTS Rules			
Rule 1: IF distance <= 4.85 and valence_at1 <= 2 and valence_at2 <= 3 THEN IC Leaf = 155 instances	Rule 2: IF distance <= 4.85 and valence_at1 <= 2 and valence_at2 >3 and charge_at2 >0 THEN IC Leaf = 11 instances	Rule 3: IF distance <= 4.85 and valence_at1 >2 and THEN IC Leaf = 284 instances	Rule 4: IF distance >4.85 and valence_at1 <= 2 and charge_at2 >0 and valence_at1 <= 1 THEN IC Leaf = 30 instances
Rule 5: IF distance >4.85 and valence_at1 >2 and valence_at2 <= 3 and valence_at2 <= 2 and valence_at2 <= 1 and charge_at1 >0 THEN IC Leaf = 37 instances	Rule 6: IF distance >4.85 and valence_at1 >2 and valence_at2 >3 and charge_at1 <= 0 and charge_at1 <= -1 and charge_at2 >0 THEN IC Leaf = 3 instances	Rule 7: IF distance >4.85 and valence_at1 >2 and valence_at2 >3 and charge_at1 <= 0 and charge_at1 >-1 THEN IC Leaf = 36 instances	-

Table 7.3: **Rules created by PTS to categorize NC class**

NC PTS Rules			
Rule 8: IF distance <= 4.85 and valence_at1 <= 2 and valence_at2 >3 and charge_at2 <= 0 THEN NC Leaf = 7 instances	Rule 9: IF distance,>4.85 and valence_at1 <= 2 and charge_at2 <= 0 THEN NC Leaf = 29 instances	Rule 10: IF distance >4.85 and valence_at1 <= 2 and charge_at2 >0 and valence_at1 >1 THEN NC Leaf = 3 instances	Rule 11: IF distance >4.85 and valence_at1 >2 and valence_at2 <= 3 and valence_at2 <= 2 and valence_at2 <= 1 and charge_at1 <= 0 THEN NC Leaf = 14 instances
Rule 12: IF distance >4.85 and valence_at1 >2 and valence_at2 <= 3 and valence_at2 >,1 THEN NC Leaf = 6 instances	Rule 13: IF distance >4.85 and valence_at1 >2 and valence_at2 <= 3 and valence_at2 >2 THEN NC Leaf = 190 instances	Rule 14: IF distance >4.85 and valence_at1 >2 and valence_at2 >3 and charge_at1 <= 0 and charge_at1 <= -1 and charge_at2 <= 0 THEN NC Leaf = 3 instances	Rule 15: IF distance >4.85 and valence_at1 >2 and valence_at2 >3 and charge_at1 >0 THEN NC Leaf = 11 instances

Table 7.4: **Rules created by UTS to categorize NC class**

NC UTS Rules		
Rule 10: IF distance >3.19 and valence_at1 <= 2 and charge_at2 <= 0 and valence_at2 <= 2 and distance >4.14 THEN NC Leaf = 14 instances	Rule 11: IF distance >3.19 and valence_at1 <= 2 and charge_at2 <= 0 and valence_at2 >2 THEN NC Leaf = 34 instances	Rule 12: IF distance >3.19 and valence_at1 <= 2 and charge_at2 >0 and charge_at1 >-1 THEN NC Leaf = 8 instances
Rule 13: IF distance >3.19 and valence_at1 >2 and distance >4.82 and valence_at2 <= 3 and valence_at1 <= 3 and valence_at2 <= 2 and charge_at2 >-1 THEN NC Leaf = 9 instances	Rule 14: IF distance >3.19 and valence_at1 >2 and distance >4.82 and valence_at2 <= 3 and valence_at1 >3 and charge_at1 <= 0 distance >5.03 THEN NC Leaf = 27 instances	Rule 15: IF distance >3.19 and valence_at1 >2 and distance >4.82 and valence_at2 >3 and distance >4.99 THEN NC Leaf = 34 instances