

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS

VIVIANE CORRÊA SANTOS

**DEVELOPING SELECTIVE CRUZAIN INHIBITORS THROUGH STRUCTURE-
BASED TECHNIQUES**

2017
BELO HORIZONTE

Viviane Corrêa Santos

Master's Thesis

**DEVELOPING SELECTIVE CRUZAIN INHIBITORS THROUGH STRUCTURE-
BASED TECHNIQUES**

**Thesis presented to the Graduate Program
in Biochemistry and Immunology at the Biological
Sciences Institute, Universidade Federal de Minas
Gerais, as a requirement to obtain the Master degree
in Biochemistry and Immunology.
Research Line: Biotechnology, vaccines,
diagnosis tests and chemotherapy.
Advisor: Rafaela Salgado Ferreira**

2017

Belo Horizonte

ACKNOWLEDGEMENTS

I shall thank to Rafaela Ferreira, my advisor, for accepting me in her group in UFMG and leading me through these two years. Liza Felicori, Lucas Bleicher and Carlos Salas, other professors in the lab also should be mentioned. My colleagues, Rafael Vieira, Glaécia, Elany, Núbia, Landry and Filipe were always solicitous and provided me very nice conversations and good lunch companionship. The other lab colleagues also deserve my grateful though because they were always opened to help, listen, and are good congress companies.

Birgit and Milene helped me with Python scripts I would never be able to write. Luan and Thiago Bruno shared with me their knowledge about computers and saved me a lot. Professor Klaus gave me good advices during the execution of this work.

Turma de Bases also should be remembered here due all the good time we were together in paper discussion and seminar presentation. My LGB friends also should be mentioned for providing me with ears, arms, laughs and brains, special thanks to André, Bruno Repolês, Ceres and Polly.

My dear brothers and sisters from GC JA's do Caminho supported me a lot in these period, together with Édén and Pierre. Other people who provided me with love and care were Ariane, Carmen, Carol, Lidi and Raíza (husbands and child included). My pharmacists Luciana, Michele, Marina and Tati also deserve my thanks.

And as the responsible for all this, I shall say many, many thanks to my family. Papai e mamãe always supported and encouraged me to follow this path and Bellinha was always by my side listening to me and helping me.

Work in our lab was funded by CNPq, CAPES and FAPEMIG.

“For from him and through him and to him are all things. To him be the glory forever! Amen.” Romans 11:36

RESUMO

Cruzaína é a principal cisteína protease do *Trypanosoma cruzi*, sendo um alvo validado para o desenvolvimento de fármacos nesse parasito. O *T. cruzi* é o agente etiológico da doença Chagas, uma doença negligenciada da América Latina. O tratamento dessa doença hoje no Brasil é feito apenas com o Benznidazol, um medicamento com eficácia comprovada apenas na forma aguda da doença e que apresenta uma série de efeitos colaterais que diminuem a adesão ao tratamento. Sendo assim, o desenvolvimento de medicamentos alternativos é imprescindível e o presente trabalho se propõe a fazer uma contribuição nesse sentido. É proposta uma triagem virtual com docking molecular na busca de possíveis inibidores de cruzaína. O diferencial desse trabalho é a inclusão da ideia de seletividade já na etapa de triagem de moléculas. Isso foi incorporado ao se realizar o docking não apenas contra cruzaína, mas também contra duas enzimas homólogas em humanos, catepsinas L e B. Assim, foram selecionados oito compostos que a partir da triagem e de uma análise visual possuem o potencial de serem inibidores de cruzaína seletivos quando comparados a pelo menos uma dessas enzimas. As perspectivas desse trabalho incluem a realização de ensaios enzimáticos para determinar a atividade e a potência das moléculas contra as três enzimas, e havendo um ligante, propor modificações para melhorar sua atividade e seletividade.

Palavras – chave: cruzaína, docking, triagem virtual

ABSTRACT

Cruzain is the major *Trypanosoma cruzi* cysteine proteinase, a validated target for drug development against this parasite. *T. cruzi* is the etiological agent of Chagas Disease, a pathological condition from Latin America, considered a neglected disease. Benznidazole is the medicine currently used in the treatment of this disease in Brazil. Its efficacy is proven only in the acute phase of the disease, presenting many side effects which contribute to the treatment abandonment by the patients. Therefore, it is very important to develop alternative drugs, and this work aims to contribute in this sense. To do so, a virtual screening strategy employing molecular docking was proposed to search for possible cruzain inhibitors. A differential aspect of this work was the inclusion of the selectivity idea already in the screening step. This has been incorporated by docking molecules against cruzain, but also against the human homologous enzymes cathepsin L and B. Thereby, after the virtual screening protocol and visual inspection of top scoring compounds, eight hits that might be selective for cruzain were selected. Perspectives include the evaluation of these molecules in enzymatic assays and, if a hit is confirmed, design analogues with improved activity and selectivity.

Key – words: cruzain, docking, virtual screening

SUMMARY

Acknowledgements.....	3
Resumo	4
Abstract.....	5
Summary	6
Abbreviation List	8
Figure List	9
Table List.....	12
Appendix List	13
1. Introduction	14
1.1. Chagas Disease	14
1.2. Proteases	16
1.3. Structure-Based Drug Design (SBDD)	18
1.4. Docking	19
1.5. Structure-Based Pharmacophore	20
1.6. Virtual screening workflow validation.....	25
1.7. Fingerprints definition	28
2. Justification	30
3. Objectives	30
3.1 Specific Objectives	30
4. Materials and methods.....	31
4.1. Construction of a Validation Database	31
4.2. Structure-Based Pharmacophore	31
4.3. Docking	32
4.4 Evaluation of Virtual Screening Methods	35
4.5 Virtual Screening Database and Workflow	35
5. Results and discussion	36
5.1 Construction of Validation Databases	36
5.2 Structure-Based Pharmacophore.....	41

5.3 Docking	45
5.4 Comparison of the Virtual Screening Performance for Different Grids	49
5.5 Virtual Screening	59
5.6 Compound Selection for In Vitro Assays	61
6. Conclusions	71
7. References	72
8. Appendix.....	79

ABBREVIATION LIST

AUC - Area under the curve

BEDROC - Boltzmann-enhanced discrimination of receiver operating characteristic

DUDE - Database of Useful (docking) decoys — Enhanced

EF - Enrichment Factor

HTVS - High Throughput Virtual Screening

logP - Logarithm of partition coefficient

MMFF – Merck Molecular Force Field

MW - Molecular weight

NMR - Nuclear magnetic resonance

pAUC - Partial area under curve

PDB - Protein Data Bank

RIE - Robust initial enhancement

ROC - Receiver Operating Characteristics

SAR – Structure-activity relationships

SP - Standard Precision

Tc - Tanimoto coefficient

VS – Virtual Screening

XP - Extra Precision

FIGURE LIST

Figure 1 - Nomenclature of proteases subsites according to Schechter and Berger, 1967. Adapted from https://prosper.erc.monash.edu.au/methodology.html	16
Figure 2 - Illustration of a cysteine protease proregion. In this example, the cruzain sequence is shown (GenBank AAAB41119.1.1). Yellow: signal peptide; Green: propeptide; Black: active domain; Blue: C-terminal portion. Information retrieved from Uniprot (UniProtKB - P25779)	17
Figure 3 - Pharmacophore generation with LigandScout omits topological information from the molecule and translates this into an abstract representation – the pharmacophore features. A: Ligand-based pharmacophore generation. B: Structure-based pharmacophore generation. Adapted from (Wolber, Dornhofer and Langer, 2006).	22
Figure 4 - Electron density from 2Fo – Fc maps of crystals with different resolutions. Adapted from (Vuorinen and Schuster, 2015).	25
Figure 5 - ROC curve example. Blue diagonal curve illustrates a model with AUC = 0.5, red curve an ideal model with AUC = 1 and the green curve a model with performance between random and ideal. Adapted from Vuorinen and Schuster, 2015.	27
Figure 6 - Hypothetical fingerprint representation. Figure from (Cereto-Massagué et al. 2015).	28
Figure 7 - A hashed topological fingerprint hypothetical example. Figure from (Cereto-Massagué et al. 2015).	29
Figure 8 - Ten most chemically diverse cruzain ligands according to Morgan fingerprints.....	40
Figure 9 - Electron density from 2Fo – Fc maps of ligands from 4KLB, 2.62Å resolution (A) and 3KKU, 1.28Å resolution (B). Maps were contoured at 1 sigma.	42
Figure 10 - The best structure pharmacophoric model generated by LigandScout. On the left, pharmacophore features with their respective residues, on the right, its ROC curve.	45
Figure 11 - Comparison between cruzain, CatL and CatB subsites. Cruzain is pink, CatL is cyan and CatB is magenta. In the table residues are colored according to this scheme: red - negatively charged; white - hydrophilic non-charged; yellow - hydrophobic; green - cysteine.	47

Figure 12 - GLU 208 charges calculated by PropKa. A) In the presence of an apolar GLU 208 is predicted to be neutral (PDB ID: 1F2C); B) In the presence of a polar group GLU 208 is predicted to be charged (PDB ID: 1AIM).	48
Figure 13 - EF10% values obtained in the grids for Glide HTVS and SP docking.	51
Figure 14 - Compounds retrieved in the top 10% of HTVS docking against cruzain in all the Grids.	52
Figure 15 - Compounds retrieved in the top 10% of SP docking in all the Grids.	52
Figure 16 - Compound 59 predicted poses among the four grids did not exhibit a high variation. The only group with a significant difference was the benzyl, it was exposed to the solvent (Grid 2 to 4) or occupying S3 pocket (Grid 1).....	53
Figure 17 - Protonation states variation in compound 40 does not interfere in pose prediction in docking HTVS	54
Figure 18 – Pose prediction of compound 60 in docking SP is influenced by the protonation states of the residues in the active site.....	55
Figure 19 - Example of ZINC compounds MMFFs force field was not able to process during ligand preparation.....	59
Figure 20 – VS workflow. Zinc molecules were submitted to a hierarchical VS starting with HTVS docking against cruzain. Molecules were ranked according to their calculated ΔG by the docking algorithm. Top 10% of ranked molecules were submitted to SP docking against cruzain. Then, compounds which filled at least one interaction from the interaction filter were submitted to SP docking against human cathepsins L and B. Molecules were ranked according to their docking scores (calculated ΔG) and the bottom 10% were subjected to XP docking against the three enzymes. Molecules were visually inspected and selected to purchase and in vitro assays.	61
Figure 21 - Predicted poses and chemical structure of ZINC97114414, a putative selective cruzain inhibitor, likely inactive against CatB.	63
Figure 22 - Predicted poses and chemical structure of ZINC81113908, a supposed selective cruzain inhibitor over CatB.	64
Figure 23 - Predicted poses and chemical structure of ZINC95480744, a supposed selective cruzain inhibitor over CatL.....	65
Figure 24 - Predicted poses and chemical structure of ZINC55314924, a supposed selective cruzain inhibitor over CatL and CatB.	66
Figure 25 - Predicted poses and chemical structure of ZINC05173978, a supposed selective cruzain inhibitor over CatB.	67

Figure 26 - Predicted poses and chemical structure of ZINC71859319, a supposed selective cruzain inhibitor over CatL.....	68
Figure 27 - Predicted poses and chemical structure of ZINC81063926, a supposed selective cruzain inhibitor over CatL and CatB.	69
Figure 28 - Predicted poses and chemical structure of ZINC83820332, a supposed selective cruzain inhibitor over CatB.	70

TABLE LIST

Table 1 - LigandScout recognized pharmacophore features and representations (Wolber & Langer 2005; Vuorinen & Schuster 2015; LigandScout3 online manual)..	22
Table 2 - PDB IDs used in docking-based virtual screening	34
Table 3 – Cruzain active compound number, ZINC code, reference, IC ₅₀ and/or Ki values.....	37
Table 4 - PDB ID of cruzain crystals employed in Structure-based pharmacophore modeling, their authors and resolution.	44
Table 5 - Atoms and respective interactions selected for the interaction filter. Interactions were analyzed with Napoli server (http://www.napoli.dcc.ufmg.br/). Values in the column % ligands were calculated as the number of ligands making this type of interaction, divided per total ligands analyzed for that protein multiplied per 100.....	48
Table 6 - Protonation states of HIS 162, GLU208 and CYS 25 in each grid considered for docking calculations against cruzain	49
Table 7 - Enrichment metrics obtained for the cruzain validation database, employing different docking methods and Grids	50
Table 8 - Enrichment metrics calculated to docking against cruzain with actives and inactive molecules	57
Table 9 - Enrichment metrics obtained in docking studies against CatL	58

APPENDIX LIST

Appendix 1 - CatL known inhibitors employed for docking evaluation.79

1. INTRODUCTION

Virtual screening of compound libraries have been contributing to drug development in the past years, and it is claimed to optimize time and costs of the process (Kar and Roy, 2013). Neglected tropical diseases can be particularly aided by this approach since the research on this field is mainly supported by public financial agencies, receiving less investment than other research topics that attract private funding.

The aim of the present work is to search for new scaffolds for inhibitors of cruzain, an enzyme from *Trypanosoma cruzi*, the etiological agent of Chagas Disease, a neglected tropical disease. Developing inhibitors for this enzyme is a challenge since it is homologous to human cathepsins, possibly leading to selective problems. Thus, it is proposed to rationalize an approach to develop selective inhibitors over human enzymes by applying a structure-based virtual screening with docking.

The present work can help to increase the comprehension about selectivity among cysteine proteases and usefulness of structure-based computational methodologies in these studies.

1.1. Chagas Disease

Chagas disease, discovered in 1909, infects about 6 and 7 million people (WHO 2016), with 12,000 deaths estimated per year and 70 million people under infection risk worldwide (WHO, 2015). This is an endemic disease in Latin American countries except Caribbean, and emerging cases have been appearing globally as a result of recent migration patterns (Albajar-Vinas and Jannin, 2011; Bern *et al.*, 2011; Connors *et al.*, 2016).

Patients affected by this disease complain not just about the physical aspects of it, but also social consequences. In literature is reported that 92.3% of chagasic people who were submitted to pre-hiring physical exams have failed and 8.9% declare their condition was the reason to be fired from their jobs (Guariento, Camilo and Camargo, 1999). Consequently, an effective treatment would improve the life quality of these patients physical and mentally.

More than one hundred years after its description by Carlos Chagas, there is still no medicine able to cure Chagas Disease without significant side effects. The available pharmacotherapy is based on two trypanocidal drugs, nifurtimox and benznidazole, which are mostly effective in the acute phase. However, in Brazil only benznidazole is licensed for therapy (Ministério da Saúde, 2010). Despite the relative success in the acute phase (75%) (Britto *et al.*, 2001), no solid evidence was seen about benznidazole efficiency in the treatment of chronic cardiac Chagas Disease in a systematic review (Reyes and Vallejo, 2011). Another systematic review was able to show that the treatment of asymptomatic chronic patients can decrease the parasite load, although it is reported these results presented some inconsistency (Villar *et al.*, 2014). Furthermore, despite decreasing the parasite load, it is described that benznidazole treatment of chronic chagasic patients was not able to prevent heart complications (Morillo *et al.*, 2015).

The Brazilian Agency on Health Regulation (Ministério da Saúde) establishes that the treatment of adults in the acute phase must be done with the administration of 5 to 7 mg/kg/day benznidazole for 60 days and in the asymptomatic chronic phase 3 mg/kg/day during 90 days. This treatment is long and displays several side effects such as exanthema, pruritus and/or allergic dermatitis (35 to 68%); nausea and vomiting, abdominal pain and/or weight loss (25% to 27%) (Ministério da Saúde, 2010). This condition contributes to the patient abandonment of treatment when the disease symptoms begin to disappear. Consequently, this can lead to parasite drug resistance.

These data reinforce the need of more effective trypanocidal drugs, as well as earlier diagnosis and treatment onset. There is evidence of increased effectiveness when the treatment starts earlier (Britto *et al.*, 2001). Therefore, several efforts to develop new and more effective antichagasic drugs against different molecular targets such as nitroreductase type I, topoisomerase, cruzain and trans-sialidase are currently described in literature (Bermudez *et al.*, 2016). Some drugs, as primaquine and carbidium can reduce parasitemia in model animals and patients, but are unable to lead to the cure (Rassi and Marcondes de Rezende, 2012).

A validated target in *Trypanosoma cruzi* (*T. cruzi*) is cruzain, the main cysteine protease in this parasite. Many inhibitors of this target have been proposed so far, although there

are some selectivity problems with them. Thus, there is a lot of structural and activity data about this enzyme which can be used to develop this project (Martinez-Mayorga *et al.*, 2015).

1.2. Proteases

Proteases (or peptidases) catalyze the hydrolysis of amide linkages in peptides and proteins. They are classified as endoproteases when this reaction happens inside the polypeptide chain and exoprotease when they cut the amino or carboxy terminal portion of the peptide (Castro *et al.*, 2011). In the case of cysteine proteases, this reaction is mediated by a cysteine residue and a histidine that polarizes the cysteine and enable it to make a nucleophilic attack in the carbonyl carbon of a susceptible peptide bond (Keillor and Brown, 1992).

Schechter and Berger defined residues located in the carboxy terminal side of the scissile peptide bond as primed (P') and the residues in the amino terminal as non-primed (P) (Schechter and Berger, 1967). Residues P and P' interact with their complementary subsites in the enzyme named S and S' (Figure 1).

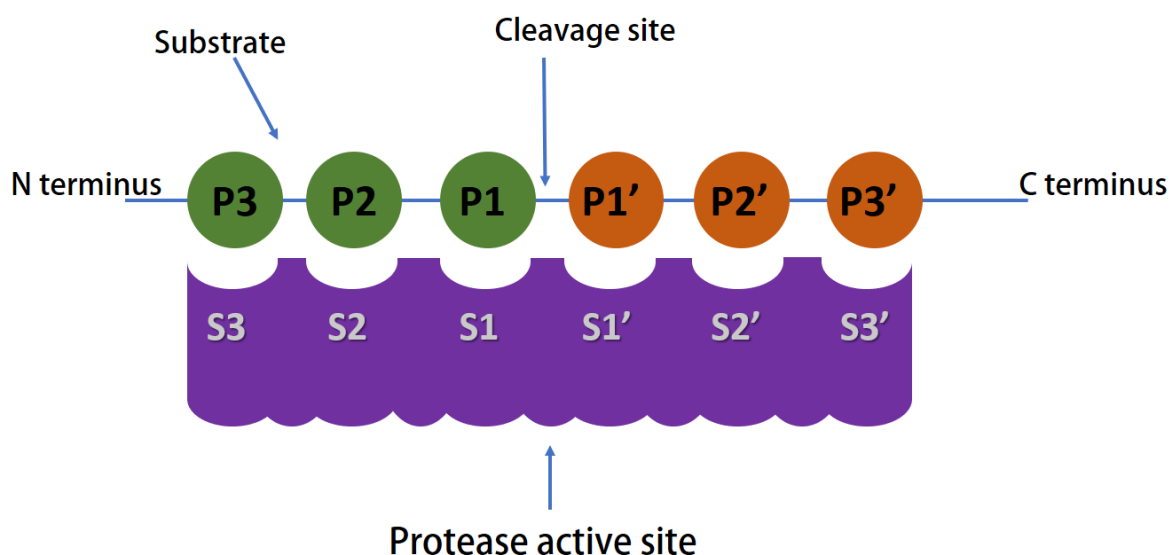


Figure 1 - Nomenclature of proteases subsites according to Schechter and Berger, 1967. Adapted from <https://prosper.erc.monash.edu.au/methodology.html>

Cruzain alongside with human cathepsins L and B (CatL and CatB, respectively) is a papain-like cysteine peptidase. They are not synthesized in their active form, but as

zymogen. They have a N-terminal proregion which is important for proper protein folding and prevents enzyme denaturation in neutral to alkaline pHs. This region is also important in the transport to the lysosome and membrane association (**Error! Reference source not found.**). In acidic pH the enzyme catalyzes its own activation by removing the propeptide (Tao *et al.*, 1994; Carmona *et al.*, 1996; Nägler *et al.*, 1997). Cruzain has also a C-terminal region which is the immunodominant domain (Martínez *et al.*, 1993). The fold of these proteins has two subdomains with the active site in the middle of them. One of them presents a bunch of helices and the catalytic cysteine and the other shows a beta-sheet with the catalytic histidine.

MSGWARALSAAVLVVMACLVPAATASLHAEETLASQFAEFKQKHGRVYGSAA
 EEFRLSVFRENFLARLHAAANPHATFGVTAFSDLTREEFRSRYHNGAAHFAAA
 QERARVPVNVEVVGAPAAVDWRARGAVTAVKDQGGCGSCWAFSAIGNVECC
 WFLAGHPLTNLSEQMLVSCDKTDSGCGGGLMNNAFEWIVQENNGAVYTEDS
 YPYASGEGISPPCTTSGHTVGATITGHVELPQDEAQIAAWLAVNGPVAVAVDAS
 SWMTYTGGVMTSCVSEQLDHGVLLVGYNDSAAVPYWVIKNSWTTQWGEDG
 YIRIAKGSNQCLVKEEASSAVVGPGPPTPEPTTTTTTSAPGPSYFVQMSCTDA
 ACIVGCENVTLPTGQCLLTSGVSAIVTCGAETLTEEVFLTSTHCSGPSVRSSVPLN
 KCNRLLRGSVEFFCGSSSSGRLADVDRQRRYQPYHSRHRRL

Figure 2 - Illustration of a cysteine protease proregion. In this example, the cruzain sequence is shown (GenBank AAAB41119.1.1). Yellow: signal peptide; Green: propeptide; Black: active domain; Blue: C-terminal portion. Information retrieved from Uniprot (UniProtKB - P25779)

Cruzain is expressed in the three *T. cruzi* forms: epimastigote (Cazzulo *et al.*, 1990), trypomastigote (Sant'Anna *et al.*, 2008) and amastigote (McGrath *et al.*, 1995). It is located in the parasite cell surface (Doyle *et al.*, 2011), flagellar pocket and other lysosome-related organelles (Sant'Anna *et al.*, 2008). Beyond that, it is secreted by trypomastigotes to the medium to digest host proteins (Yokoyama-Yasunaka *et al.*, 1994). It is also related to parasite cell differentiation (Franke De Cazzulo *et al.*, 1994), host cell invasion (Scharfstein *et al.*, 2000; Aparicio, Scharfstein and Lima, 2004) and immune response evasion (Bontempi and Cazzulo, 1990). These evidences reveal cruzain as a validated target for Chagas Disease drug design. Furthermore, it was

described a cure in murine model treated with cruzain inhibitor in lethal infection (Engel *et al.*, 1998; Cazzulo, Stoka and Turk, 2001).

Despite being a good drug target to Chagas Disease, cruzain has a very similar active site to human CatL and B. However, it is known that small differences in sub-pockets may allow to achieve selectivity (Kuhn *et al.*, 2016). One example is the Fatty Acid Binding Proteins (FABPs), in which the difference in one sub-pocket was decisive to develop a selective inhibitor for FABP4/5 over FABP3. Roche's researches performed a screen and discovered a compound with activity for FABP4 that was neither active for FABP5 nor selective over FABP3. Based on structural comparisons they realized that FABP3 has a smaller pocket than the other two enzymes due to the presence of three LEU while the others have ILE, VAL and CYS at the equivalent positions. So, the scientists increased the size of the portion of the molecule interacting in this pocket and could improve FABP4/5 activities and reduced FABP3 activity. Another successful case is the odanacatib, a selective inhibitor for human cathepsin K which is in phase 3 clinical trials (Gauthier *et al.*, 2008).

Attempts at developing inhibitors against cruzain have been successfully performed in recent decades. Some examples scaffolds displaying inhibitory activity ranging from 0.02 to 1000 μM are thiosemicarbazones, tetrahydropyran, vinyl sulfones and pyrimidines (Du *et al.*, 2002; Siles *et al.*, 2006; Zanatta *et al.*, 2008; Martinez-Mayorga *et al.*, 2015). However, these studies usually don't evaluate selectivity over human cathepsins.

1.3. Structure-Based Drug Design (SBDD)

This knowledge about known scaffolds and the basis of the interaction between a small molecule and a macromolecule can be very useful in drug design. Analyzing data from x-ray crystallography and nuclear magnetic resonance (NMR) is very helpful to achieve this purpose. In the absence of experimental data, comparative modeling can be applied. Through structural data, it is possible to investigate the steric and electronic principles that rule the interaction, subsequently trying to transpose it to other molecules to obtain ligands with higher binding affinities.

SBDD is a cyclic procedure that starts with the analysis of the target and the use of computational programs to identify potential ligands. Afterward, they can be purchased or synthesized and experimental assays are performed. Once the activity of one or more molecules is confirmed, the complex macromolecule-ligand must have its structure solved. From this point, it is possible to propose modifications in the ligand or search analog compounds on virtual databases seeking for molecules with higher affinity (Ferreira *et al.*, 2015).

A method widely used in SBDD design is virtual screening (VS), in which large compounds databases are screened against the target molecule and those that are predicted to be active are tested *in vitro*. Well established techniques used in VS include molecular docking, structure-based pharmacophores and molecular dynamics simulations.

1.4. Docking

Docking is a structure-based computational technique for prediction of the binding mode between two molecules. In our case, the interactions analyzed will be between enzymes and small molecules. Moreover, it can make a quantitative prediction of the energy involved in these binding modes, providing not just a binding mode, but also a ranked list based on the scores obtained from docking (Ferreira *et al.*, 2015).

The docking software employed in this study was Glide (Grid-based Ligand Docking with Energetics, version 6.8, Schrödinger, LLC, New York, NY, 2015). It makes as close as possible an exhaustive search in the ligands conformation, to provide a prediction similar to the reality. Glide produces a set of ligand conformations next to the minima. These conformations are docked in different places in the macromolecule to find the ligands with poses more energetic favorable. Then, the ligand is minimized using OPLS-AA force field and the lowest energy poses obtained have their torsional minima analyzed. After predicting binding poses, Glide ranks the molecules according to their predicted binding affinities, based on GlideScore and Emodel scoring functions. GlideScore is an empirically-based scoring function with many terms, some of which are force field contributions, and others penalize, or reward special motifs identified in the predicted pose. To select the best pose of a ligand, Glide uses Emodel, a combination of GlideScore, ligand-receptor interaction energy and ligand strain energy.

In the end, the best scoring poses from different ligands are ranked according to their GlideScore (Friesner *et al.*, 2004).

Glide has three pose prediction levels: HTVS, SP and XP, with increasing accuracy and computational cost of calculations. HTVS or High Throughput Virtual Screening is used for rapid conformational search in a set which has many ligands. SP or Standard Precision applies a scoring function that enables the program to find molecules that are likely to bind the receptor, although the pose prediction is not precise and can have significant imperfections. Its objective is to minimize false negatives. XP or Extra-precision seeks to minimize the false positive poses with a scoring function that applies severe penalties in the poses when charged and polar groups are not adequately exposed to solvent and rewards features such as special hydrogen bond motifs and hydrophobic interactions within enclosed pockets (Friesner *et al.*, 2004).

Another difference in Glide XP is the sampling methodology, which evaluates more conformations than Glide SP. If at least one predicted pose has a key fragment anchored in the active site, Glide performs what is called growing algorithm. The key fragment is generally a ring or other rigid fragment that presents various positions sampled and clustered. Afterward, a representative member of each cluster has its side chains grown from the anchor fragment. Molecules with severe steric clashes are discarded, and the retrieved are minimized using an energy function from Glide. Following they are ranked according to Emodel, explicit water molecules are added to top-ranked molecules and penalties are calculated. Any side chains that are responsible for a penalty are grown and minimized again. Then, a single pose is selected based on a weighted scoring function that combines protein-ligand Coulomb and van der Waals interaction energies, terms that favor binding affinity and the penalty terms (Friesner *et al.*, 2006). These can help the program to predict accurately the ligand binding pose with the drawback of the high computational cost of these calculations.

1.5. Structure-Based Pharmacophore

Another technique widely used in drug discovery is based on the set of features present in a molecule that are essential for its biological activity, the pharmacophore. These features are related to atoms electronic aspects and their position on the tridimensional

space. Therefore they represent a particular binding mode of a molecule – or a set of molecules (Güner, 2002; Wolber and Langer, 2005). Pharmacophores are very useful in drug virtual screening because they allow the identification of new scaffolds based on consensus information obtained from experimentally tested molecules. One advantage of pharmacophore modeling is the low computational cost. That is the reason why many groups apply this technique as a filter in hierarchical virtual screening before methods that can predict binding affinities but also present a high computational cost. One example of its application can be found in work published by Shah and colleagues. They described the identification of 21 falcipain-2 inhibitors by submitting an extensive database to a hierarchical virtual screening that started with a structure-based pharmacophore filter (Shah *et al.*, 2011).

Pharmacophore modeling is always interesting to help us to comprehend better the target studied and the known binding molecules. Still, it is important to settle down some concepts about pharmacophore modeling and its uses in virtual screening.

Each pharmacophore modeling software translates topological information into an abstract representation (Figure 3) recognizing particular features.

Table 1 summarizes features from LigandScout, a software commonly used in pharmacophore generation, and their respective representation. It identifies hydrogen bonds, charge transfers, lipophilic and aromatic interactions, metal and covalent binding.

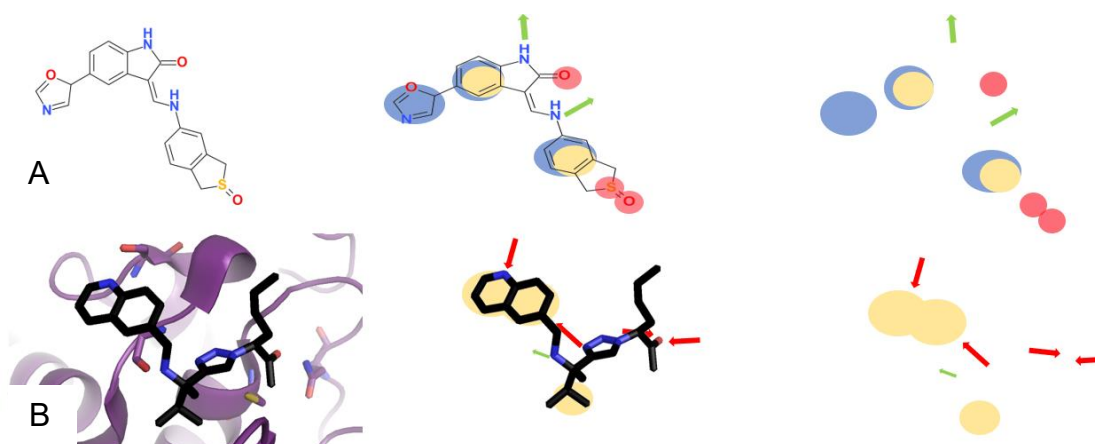











Figure 3 - Pharmacophore generation with LigandScout omits topological information from the molecule and translates this into an abstract representation – the pharmacophore features. A: Ligand-based pharmacophore generation. B: Structure-based pharmacophore generation. Adapted from (Wolber, Dornhofer and Langer, 2006).

Table 1 - LigandScout recognized pharmacophore features and representations (Wolber & Langer 2005; Vuorinen & Schuster 2015; LigandScout3 online manual)

Pharmacophore feature	LigandScout definition
	Hydrogen bond acceptor – an atom with negative partial charge distant from 2.5 until 3.8Å of a partially positive hydrogen. The angle between the acceptor and the donor atoms should be 180°; the bond is considered broken when the angle differs by 34°.
	Hydrogen bond donor – the heavy atom of these functional groups: nonacidic hydroxyls, thiols, acetylenic hydrogens and NHs ^a that is found between 2.5 and 3.8Å from an acceptor group. The angle between the acceptor and the donor atoms should be 180°; the bond is considered broken when the angle differs by 34°.
	Positive ionizable area – atoms or groups that are usually protonated at pH 7.4.
	Negative ionizable area – atoms or groups that are typically deprotonated at pH 7.4.

Pharmacophore feature	LigandScout definition
	Hydrophobic area - a sphere located in the center of a lipophilic group that is located between 1 to 5 Å from the receptor.
	Aromatic ring – aromatic group of the ligand which can interact with an aromatic ring or positive group from the environment.
 a	Metal binding – groups able to interact with magnesium, zinc or iron from proteins.
	Exclusion volume - areas in the macromolecule that cannot be assessed by the ligand due to steric clashes.
	Residue binding point - a region where there is a covalent bond between ligand and protein.

^aexcept tetrazoles and trifluoromethyl hydrogens

Modeling a pharmacophore requires the use of data retrieved from a public database to obtain either target structures or small molecules structures and activity data. The Protein Data Bank (PDB) is the most used database to search for target structures. Nevertheless, when these data are used to obtain Structure-based pharmacophore, it is important to have in mind that the focus of this database is the protein, not the ligand. The PDB file, for instance, was created to describe protein atoms types, not all the others that can appear in the ligands, and some PDB entries have wrong ligand topology description. Only part of the ligand graph information is defined in the PDB file, while bond types and atom hybridization states are missing (Wolber and Kosara, 2006).

Pharmacophore modeling can be divided into two approaches, ligand and Structure-based. When the receptor structure is not available, but there is information about ligands, the ligand-based methodology can be very useful. Furthermore, in structure-based modeling information retrieved from X-ray

crystallography and NMR data or even comparative modeling is used to translate ligand-target interactions to pharmacophore features (Vuorinen and Schuster, 2015).

Depending on the selected approach, some standard procedures must be followed. Ligand-based pharmacophore modeling requires two sets of molecules. The training set contains molecules upon which the model will be built, while the test set consists of molecules that will be used to validate the model. It is interesting to have these sets with a high chemical diversity to try to generate a pharmacophore hypothesis that would be able to recognize molecules as structurally diverse as possible (Vuorinen and Schuster, 2015).

If high-quality 3D coordinates are available from the complex macromolecule-small molecule, it is intuitive trying to study how the ligand is complementary to the binding site. Besides docking, structure-based pharmacophore can be very useful in that sense. LigandScout may firstly interpret ligand coordinates from PDB file to build a structure-based pharmacophore model. Therefore, it is important to check the electron density maps of these structures, in particular on the ligand, and try to use structures with the best maps as possible. In Figure 4 three PDBs can be observed. The first one has a high resolution of 0.85Å, enabling to locate each atom properly. The second PDB has an electron density map with medium resolution, so atoms locations are estimated. The third PDB shows a low-resolution structure where only the general shape of proteins is estimated. Since the models are going to be based on this information, it is essential to observe these characteristics to guarantee the quality of the model and the following work.

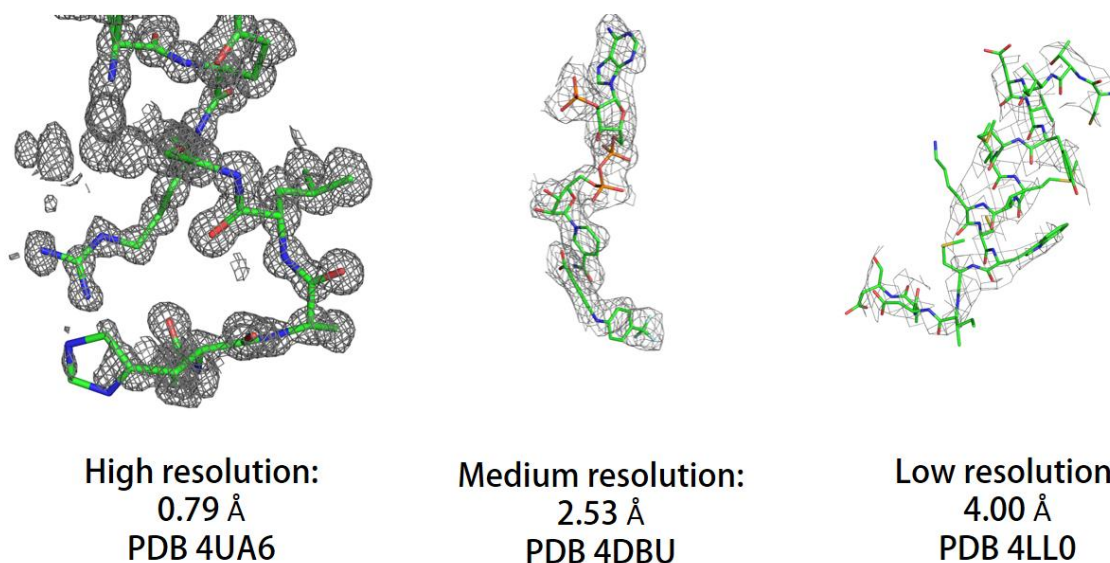


Figure 4 - Electron density from $2Fo - Fc$ maps of crystals with different resolutions. Adapted from (Vuorinen and Schuster, 2015).

The first step in ligand interpretation is the analysis of its topology, which enables the software to build a molecular graph with rings and untyped bonds. Since the PDB file does not contain information on the hybridization states of the atoms, the program must determine this first, to subsequently derive bond types. Finally, LigandScout can analyze ligand-protein interactions to recognize the chemical features that will be used in the model construction (Wolber and Kosara, 2006).

Regardless of the approach employed, once a satisfying pharmacophore model is defined compounds are searched against the pharmacophore query. LigandScout creates a pharmacophore model for each conformer molecule in the database and uses these models in the search. Compounds that fit the query model are considered as hits. Otherwise, the molecules are suggested to be inactive (Markt, Schuster and Langer, 2011).

1.6. Virtual screening workflow validation

An *in silico* method chosen for virtual screening should: be selective (find the active ligands in the middle of a pool of inactive molecules) and be able to find new chemical scaffolds for the desired target. A suitable method must also be able to help the researchers to prioritize the real actives for *in vitro* assays, since only a small number

of compounds are experimentally evaluated. To address this issue, it is very useful to generate a validation database prior to designing the VS workflow.

This database should contain structurally diverse active molecules and inactive ones structurally related to the actives. This is crucial since using inactive molecules which are very different from the actives might lead to artificial enrichment metrics (Markt, Schuster and Langer, 2011). However, a frequent issue is the fact that negative results are not widely published. Also, when a commercial compound database is evaluated, the compounds tested are as diverse as possible. In consequence, using computationally generated decoys when evaluating a model might be a way to overcome these potential issues.

Decoys are molecules that have physicochemical properties similar to the known active ligands and are topologically different. They can be generated on web-based servers as DUD-E (Mysinger *et al.*, 2012). The server calculates the following properties based on an active compound: molecular weight (MW), the logarithm of the partition coefficient (logP – reflects hydrophobicity of a small molecule), the number of rotatable bonds, hydrogen bond donors and acceptors. Then, molecules with similar parameters are retrieved from the database. The selected molecules are compared to the query, and the most different ones are returned to the user. In effect, it is useful to have 45 decoys for each active ligand to mimic the reality (Mysinger *et al.*, 2012).

Some statistical evaluations can be used to check a model performance in virtual screening: sensitivity, specificity, enrichment factor (EF), area under the receiver operating characteristics (ROC) curve (AUC) and partial area under the ROC curve (pAUC) (Braga and Andrade, 2013):

$$\text{Sensitivity} = \frac{\text{corrected classified positives}}{\text{total true positives}} \quad (1)$$

$$\text{Specificity} = \frac{\text{corrected classified negatives}}{\text{total true negatives}} \quad (2)$$

ROC curves are obtained by plotting the sensitivity versus 1-specificity, while enrichment plots contain %actives versus %screened molecules. As Figure 5 shows, the performance of a model might be evaluated based on its AUC values. AUC = 0 means the model only finds false positives, AUC = 0.5 means the model finds and rank actives and inactive molecules by chance and AUC = 1 means the model ranks all true positives before finding a false positive. An AUC = 1 is almost impossible to be achieved. So, the desired value is something as close as possible to 1 and higher than 0.5.

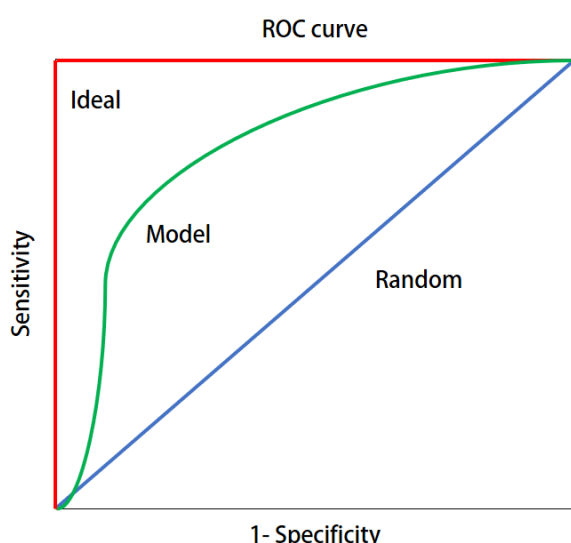


Figure 5 - ROC curve example. Blue diagonal curve illustrates a model with AUC = 0.5, red curve an ideal model with AUC = 1 and the green curve a model with performance between random and ideal. Adapted from Vuorinen and Schuster, 2015.

Enrichment factors (EF) are used to evaluate how much the methods enriches true positives within a given percentage of the database, when compared to random ordering. EF are calculated by determining the fraction of true positives selected in the top X% of the ranked list of screened compounds and dividing this value by the fraction of total true positives in the total molecules tested.

$$EF^{X\%} = \frac{\text{true positives selected}^{X\%} / \text{total molecules selected}^{X\%}}{\text{total positives} / \text{total molecules tested}} \quad (3)$$

When running a virtual screening the goal is to find models not only able to find real actives, but also to rank them higher than the inactive molecules and within the very top of the database. However, evaluating the AUC of the entire ROC curve does not provide a clear picture of whether the VS algorithm ranks the actives at the top or at the bottom of the score list. Thus, the pAUC, the AUC of the top X% of evaluated compounds ordered in a score list, is applied to help the evaluation of the top compounds ranked by a model (Braga and Andrade, 2013).

1.7. Fingerprints definition

The validation performed with active molecules and decoys rely on the need of knowing the similarity of molecules. Cheminformatics methods like fingerprints are well established and widely used to address the chemical diversity of a set of compounds. In the present work, fingerprints were calculated to compare molecules and group the most similar ones. The fingerprint is a string representation of a molecule structure and properties in which a binary pattern is used to describe and compare molecules. In Figure 6 there is a hypothetical fingerprint representation, in which the molecule was divided into substructures and, once they matched with their bit representation, they were marked as 1, and the other bits that were not represented in the molecule were kept as 0.

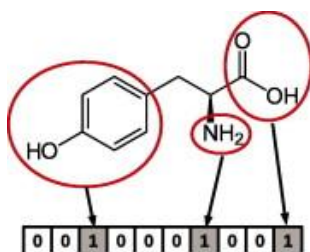


Figure 6 - Hypothetical fingerprint representation. Figure from (Cereto-Massagué et al. 2015).

Fingerprints are widely used to compare molecules, especially in similarity searches. Generally, with a given query, a search in a database for similar molecules is done by calculating a similarity coefficient in a pairwise comparison. The mainly used is the Tanimoto coefficient (Tc), which is calculated according to the following equation:

$$Tc = \frac{c}{a + b - c} (4)$$

Where a is the number of bits set to 1 in molecule A, b is the number of bits set to 1 in molecule B and c is the number of bits set to 1 common to both molecules. Tc values range from 0 to 1. 0 means no similarity, while 1, means very similar or identical (Cereto-massagué *et al.*, 2015).

There are different types of fingerprints, and the focus of the present work is circular fingerprints, a kind of hashed topological fingerprint, illustrated in Figure 7. Hashed topological fingerprints work by selecting an atom and analyzing all the fragments generated by following a path until a predetermined number of bonds. This process is done to all the atoms in the molecule. A given bit can correspond to more than one feature, which is called bit collision. Moreover, circular fingerprints do not analyze the path in a molecule, but the environment of each atom until a given radius. The ECFP4 fingerprint is a particular type of circular fingerprint used in this thesis (Rogers and Hahn, 2010).

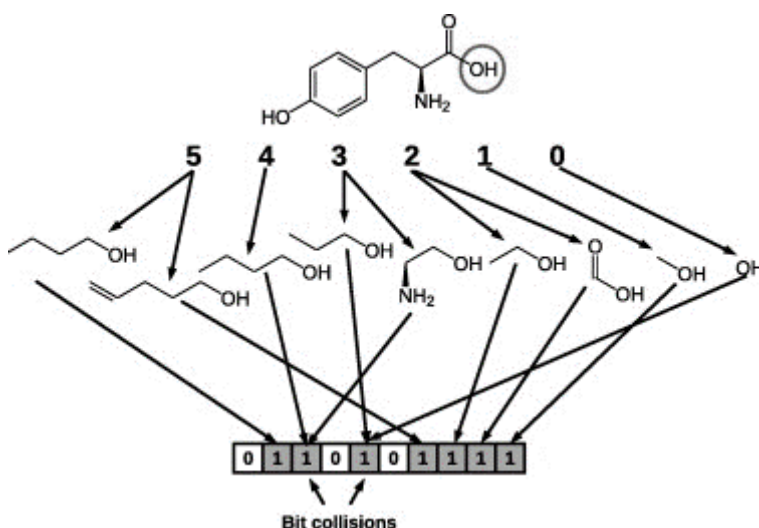


Figure 7 - A hashed topological fingerprint hypothetical example. Figure from (Cereto-Massagué *et al.* 2015).

2. JUSTIFICATION

Considering the epidemiological importance of Chagas disease and the lack of safe and efficient trypanocidal drugs, it is important to development new antichagasic medicines. Cruzain is a validated target in *T. cruzi*, with many studies describing the importance of this enzyme in parasite development, cell infection and survival. Many structural and activity data have been reported in literature making this a good target to structure-based approaches.

Moreover, drug development can take advantage of computational methods, as they can make the process cheaper and faster. Structure-based pharmacophores and molecular docking are two established methods in drug discovery with several successful reported cases, which were chosen in this work to aid in the discovery of selective cruzain inhibitors.

3. OBJECTIVES

To propose potential selective cruzain inhibitors through a validated virtual screening approach.

3.1 Specific Objectives

- I. To compare the performance of structure-based pharmacophore and molecular docking techniques in retrospective virtual screening of a library containing cruzain ligands and decoys
- II. To rationalize a workflow to address the selectivity issue in virtual screening
- III. To propose new cruzain inhibitors by virtually screening the ZINC database

4. MATERIALS AND METHODS

Two approaches were evaluated to propose a suitable VS workflow to identify potential selective cruzain inhibitors. A structure-based pharmacophore and a docking approach were considered.

4.1. Construction of a Validation Database

Prior to establishing the virtual screening workflow, we tested and validated some methods. In doing so is important to use sets of known cruzain, human CatL and CatB inhibitors and to analyze if the methods can correctly predict their activities. A search in the BindingDB database (Gilson et al. 2015) and in the literature was carried out looking for molecules with IC_{50} or K_i below 100 μ M. Besides, decoys were generated on DUD-E (Mysinger *et al.*, 2012) to evaluate the specificity of the tested methods. On the DUD-E website (dude.docking.org) the user can upload the SMILES of the ligand and subsequently the server generates the appropriate protonation states in pH ranging from 6 to 8. Then, it calculates the molecular properties for each protonation state (molecular weight, estimated water-octanol partition coefficient, rotatable bonds, hydrogen bond acceptors and donors, plus net charge) and finds in the ZINC database molecules that match these properties. This search is performed with a protocol that adapts to local chemical space by narrowing or widening windows in seven steps around these properties. The aim is to find from 3,000 to 9,000 putative decoys and to calculate their ECFP4 fingerprints to select the 25% most different molecules. After removing possible duplicates, 50 decoys for each ligand are picked randomly. In the case of cruzain, experimental HTS data was available in the PubChem BioAssay database (Wang *et al.*, 2017). Therefore, we also download a subset of experimentally confirmed inactive molecules to evaluate VS performance also considering this set.

4.2. Structure-Based Pharmacophore

Modeling a Structure-based pharmacophore requires a 3D structure of the macromolecular receptor and a known ligand. Crystallographic structures of complexes between cruzain and competitive inhibitors were searched in PDB, and 2Fo-Fc maps at 1.0 sigma of these crystals were evaluated using Coot software (Emsley *et al.*, 2010), specifically in the region corresponding to the ligand. Among complexes for which ligand binding modes were unequivocally defined based on their

electron density, clustering was performed with LigandScout, to define representative structures for pharmacophore generation. Twenty-five conformations were generated per compound, and a pharmacophore was calculated for each conformation. Afterward, a similarity score was calculated based on these pharmacophores. An average of these scores was calculated, and ligands were clustered based on the distance to this average score. Compounds representing each cluster were employed in the generation and evaluation of the model, as described below in Item 4.4.

4.3. Docking

Docking with Glide (version 6.8, Schrödinger, LLC, New York, NY, 2015) was employed to predict binding modes and for virtual screening. Glide has two scoring functions, GlideScore and GlideScore XP. GlideScore is an empirically based scoring function and can be written as follows:

$$\Delta G_{binding} = C_{lipo-lipo} \sum f(r_{lr}) + C_{hbond-neut-neut} \sum g(\Delta r)h(\Delta\alpha) + C_{hbond-neut-charged} \sum g(\Delta r)h(\Delta\alpha) + C_{hbond-charged-charged} \sum g(\Delta r)h(\Delta\alpha) + C_{max-metal-ion} \sum f(r_{lm}) + C_{rotb}H_{rotb} + C_{polar-phob}V_{polar-phob} + C_{coul}E_{coul} + C_{vdW}E_{vdW} + solvation terms \quad (5)$$

Where l refers to ligands' atoms, r to receptor atoms and $f(r_{lr})$ is a linear function of the interatomic distance. The first term in the eq 5 is related to hydrophobic effect, that happens when a lipophilic atom from ligand interacts with a lipophilic atom from receptor, releasing water molecules from the active site and leading to a decrease in the free energy of the system. If these water molecules have a low movement, once released, they induce an entropy gain, which is also favorable to the free energy. The hydrogen bond terms (2nd to 4th) are separated according to the nature of the atoms involved, if both are neutral, one neutral and the other charged or both charged. The $C_{max-metal-ion} \sum f(r_{lm})$ term measures metal-ligand interactions, the $C_{polar-phob}V_{polar-phob}$ rewards occurrences in which a polar atom that is not involved in hydrogen bond is in a lipophilic region. $C_{coul}E_{coul}$ and $C_{vdW}E_{vdW}$ terms compute Coulomb and van der Waals interactions energies. The solvation terms are calculated in the competitive ligand poses by adding explicit water molecules to the system and

measuring the exposition of many groups to these molecules. Friesner et al claim these terms to be helpful in reducing false positive results (Friesner *et al.*, 2004, 2006).

XP Glide scoring function is also empirically based, its description is following:

$XP\ GlideScore = E_{coul} + E_{bind} + E_{penalty}$ (6), where:

$$E_{bind} = E_{hyd_enclosure} + E_{hb_nn_motif} + E_{hb_cc_motif} + E_{PI} + E_{hb_pair} + E_{phobic_pair} \quad (7)$$

and

$$E_{penalty} = E_{desolv} + E_{ligand_strain} \quad (8)$$

E_{hb_pair} and E_{phobic_pair} are the same hydrogen bond and lipophilic pair terms described above. $E_{hyd_enclosure}$ is the hydrophobic enclosure score, differently from E_{phobic_pair} , it considers not a pair of atoms in ligand and in the receptor, but a group of connected lipophilic atoms. $E_{hb_nn_motif}$ are the special neutral-neutral hydrogen-bonding motifs. They are identified in positions in the active site where the water molecules form a hydrogen bond to protein in such a way it is difficult for them to perform additional ones. So, this interaction has some geometrical constraints due the environment where they happen, that generally is a hydrophobic protein region which surrounds the water molecule in two faces. The donor or acceptor atom must be a ring atom, except nitrogen. $E_{hb_cc_motif}$ refers to special charged-charged hydrogen-bond motifs that occurs, for example, when a positive ligand group binds to a weakly solvated negative protein group, or a ligand CO_2^- group binds to multiple positive groups in the protein that are close to each other. E_{PI} rewards pi-stacking and pi-cation interactions.

However, there are other naïve terms included that reward halogen atoms in hydrophobic regions and an empirical correction enhancing the binding affinity of smaller ligands over to larger ones (Friesner *et al.*, 2006). Some terms penalize the pose, like E_{desolv} that refers to water scoring and it is calculated in the docking step in which explicit water molecules are added. When a polar or charged group in the ligand is not properly solvated, a desolvation penalty is applied. E_{ligand_strain} are the contact penalties, which penalize only bad internal contacts such as some ligands geometries retrieved from spectroscopy data exhibiting high strain energies (Friesner *et al.*, 2006).

4.3.1 Subsites definition

Prior doing docking studies we wanted to know better the residues composing the subsites of the enzymes we would study. To do so, we aligned crystal structures of cruzain (PDB ID: 20Z2), cathepsin L (PDB ID: 2XU1) and cathepsin B (PDB ID: 1GMV). These crystals are from complexes of the enzymes with peptidic ligands. Then, we analyzed the residues interacting with the crystal's ligands and summarized them (Figure 11).

4.3.1 Compound preparation

Compounds were prepared with LigPrep (LigPrep version 3.5, Schrödinger, LLC, New York, NY, 2015) using Merck Molecular Force Field (MMFF). Ionization states were generated with Epik at the following pHs: 5.5 ± 2 (docking against cruzain and human cathepsin L) and 6.0 ± 2 (docking against human cathepsin B).

4.3.2 Protein preparation

Proteins were prepared with Protein Preparation Wizard in Maestro (Maestro version 10.3, Schrödinger, LLC, New York, NY, 2015). Hydrogens were added with PropKa using the following pHs: 5.5 for cruzain and CatL and 6.0 for CatB. The PDBs employed in this study are summarized in Table 2:

Table 2 - PDB IDs used in docking-based virtual screening

PDB ID	Authors	Protein	Resolution (Å)
3KKU	Ferreira, R.S. et al.	Cruzain	1.28
1MHW	Chowdhury, S. et al.	Cathepsin L	1.9
3AI8	Renko, M. et al.	Cathepsin B	2.11

Grids were generated to each prepared protein with centroids set on the catalytic cysteine (CYS 25) and dimensions of 10x10x10 Å. SER 24, CYS 25 and SER 64 side chains were considered rotatable in cruzain. SER 24 and CYS 25 side chains were

kept rotatable in CatL. SER 28 and CYS 29 side chains were considered as rotatable in CatB. All other residues were kept rigid during docking.

4.4 Evaluation of Virtual Screening Methods

LigandScout was the program applied for pharmacophore modeling, while Glide was employed for docking. Both programs generate ROC curves and calculate AUC, pAUC EF values at 1, 5 and 10% database screen. Statistical analysis with the R package pROC were performed to address the significance of observed differences in ROC curves and AUC values. DeLong's unpaired test was carried out to compare AUC values and Venkatraman's test to compare ROC curves shape (Robin *et al.*, 2011).

4.5 Virtual Screening Database and Workflow

The VS was performed using ZINC database, a library of commercially available compounds containing more than 35 million 3D structures divided into subsets (Irwin and Shoichet, 2005; Irwin *et al.*, 2012). The subset employed in the present work was Leads Now molecules, which contains 3,687,621 molecules with a molecular weight between 250 and 350 g/mol; xlogP under 3.5 and less than seven rotatable bonds. The molecules from this subset were clustered and filtered based on a Tc = 0.9. To do so, the molecules were ranked according to their molecular weight, ascending. Thus, a compound is selected and next one would be selected if it differs from the first one by a Tc of 0.9 using calculated fingerprints.

A hierarchical virtual screening was performed. Firstly, the ZINC molecules were docked with Glide docking HTVS against cruzain. Molecules on the top ten percent ranking were docked again against cruzain by SP docking. Next, only molecules involved in the main interactions with the receptor were subsequently docked against CatL and B by SP docking. These interactions were defined by analyzing PDB complexes of these enzymes. From this new ranking, bottom ten percent of molecules were docked on the XP mode against the three enzymes. Compounds were visually inspected and selected to be purchased for *in vitro* assays.

5. RESULTS AND DISCUSSION

5.1 Construction of Validation Databases

Before performing VS studies, compound databases were built for validation and comparison of the performance of possible VS strategies. Databases specific for each protein target have been constructed, containing: experimentally validated competitive inhibitors, retrieved from BindingDB; computationally generated decoys, obtained through the DUD-E website; and, in the case of cruzain, experimentally validated inactive compounds retrieved from PubChem.

Cruzain active ligands considered in the analysis were from five published papers (Du et al. 2000; Huang et al. 2003; Ferreira et al. 2010; Rogers et al. 2012; Ferreira et al. 2014) and in-house compounds, totalizing 66 molecules (

Table 3). All of them are competitive and non-covalent binders from several classes as aryl ureas, ketones and benzimidazoles. The potency (measured by IC_{50} and K_i) of these compounds ranged between 0.004 and 100 μ M, being in most cases under 25 μ M. Ligands Morgan fingerprints were calculated with RDKit (www.rdkit.org) to illustrate their chemical diversity. Then, they were clustered and the ten most diverse scaffolds are shown in **Error! Reference source not found.**

Table 3 – Cruzain active compound number, ZINC code, reference, IC₅₀ and/or Ki values.

Compound number	ZINC code	Reference	IC ₅₀ (μM)	Ki (μM)
1	ZINC36962557	Biocomp		2.0
2	ZINC00314954	Biocomp		3.0
3	ZINC10434589	Biocomp	27.0	
4	Du_2000_D12	Du et al. 2000	<10.0	
5	ZINC5223994	Du et al. 2000	4.8	
6	ZINC1038200	Du et al. 2000	3.1	
7	ZINC1033017	Du et al. 2000	<10.0	
8	ZINC3106209	Du et al. 2000	2.7	
9	Du_2000_D17	Du et al. 2000	<10.0	
10	ZINC2161657	Du et al. 2000	<10.0	
11	ZINC1040170	Du et al. 2000	3.7	
12	Du_2000_D22	Du et al. 2000	<10.0	
13	ZINC1035011	Du et al. 2000	10.0	
14	Du_2000_D45	Du et al. 2000	3.0	
15	ZINC1043567	Du et al. 2000	<10.0	
16	ZINC1047389	Du et al. 2000	<10.0	
17	ZINC1042930	Du et al. 2000	1.2	
18	ZINC2161654	Du et al. 2000	<10.0	
19	Du_2000_D23	Du et al. 2000	1.9	
20	ZINC1026484	Du et al. 2000	2.9	
21	ZINC2148801	Du et al. 2000	<10.0	
22	Du_2000_D61	Du et al. 2000	<10.0	
23	Du_2000_D34	Du et al. 2000	6.9	
24	ZINC20191037	Ferreira et al. 2010		2.0
25	ZINC03242874	Ferreira et al. 2010	0.3	

Compound number	ZINC code	Reference	IC ₅₀ (μM)	K _i (μM)
26	ZINC03363866	Ferreira et al. 2010		6.0
27	ZINC05212600	Ferreira et al. 2010	0.5	
28	ZINC08693977	Ferreira et al. 2010		0.8
29	ZINC09580294	Ferreira et al. 2010	1.0	
30	ZINC03282619	Ferreira et al. 2010	38.0	
31	ZINC01852276	Ferreira et al. 2010	0.7	
32	ZINC05061372	Ferreira et al. 2010	18.0	
33	ZINC03363859	Ferreira et al. 2010	7.0	6.0
34	ZINC02236859	Ferreira et al. 2010		2.0
35	ZINC00943080	Ferreira et al. 2010		2.0
36	ZINC8691187	Ferreira et al. 2010	1.0	2.0
37	ZINC02652325	Ferreira et al. 2010	3.0	
38	ZINC00002334	Ferreira et al. 2014	0.2	
39	ZINC13824869	Ferreira et al. 2014	0.5	
40	ZINC13824883	Ferreira et al. 2014	0.6	
41	ZINC13824781	Ferreira et al. 2014	0.6	
42	ZINC13824822	Ferreira et al. 2014	0.8	
43	ZINC00003658	Ferreira et al. 2014	2.7	
44	ZINC00011550	Ferreira et al. 2014	1.6	
45	ZINC13824833	Ferreira et al. 2014	3.0	
46	ZINC00002336	Ferreira et al. 2014	3.0	
47	ZINC13824805	Ferreira et al. 2014	4.1	
48	ZINC13824835	Ferreira et al. 2014	5.2	
49	ZINC13824857	Ferreira et al. 2014	13.5	
50	ZINC13824824	Ferreira et al. 2014	10.9	
51	ZINC13824804	Ferreira et al. 2014	38.4	
52	ZINC13824839	Ferreira et al. 2014	5.3	

Compound number	ZINC code	Reference	IC ₅₀ (μM)	K _i (μM)
53	ZINC13824780	Ferreira et al. 2014	9.9	
54	ZINC13824851	Ferreira et al. 2014	8.2	
55	ZINC13824867	Ferreira et al. 2014	12.7	
56	ZINC13824793	Ferreira et al. 2014	23.9	
57	ZINC13824797	Ferreira et al. 2014	13.2	
58	ZINC13824826	Ferreira et al. 2014	77.5	
59	huang_2003_1	Huang et al. 2003		0.1
60	huang_2003_3	Huang et al. 2003		0.004
61	huang_2003_2	Huang et al. 2003		0.06
62	huang_2003_4	Huang et al. 2003		>10.0
63	huang_2003_5	Huang et al. 2003		>10.0
64	huang_2003_6	Huang et al. 2003		>5.0
65	ZINC01694053	Rogers et al. 2012	66.0	
66	ZINC85548285	Rogers et al. 2012	16.0	

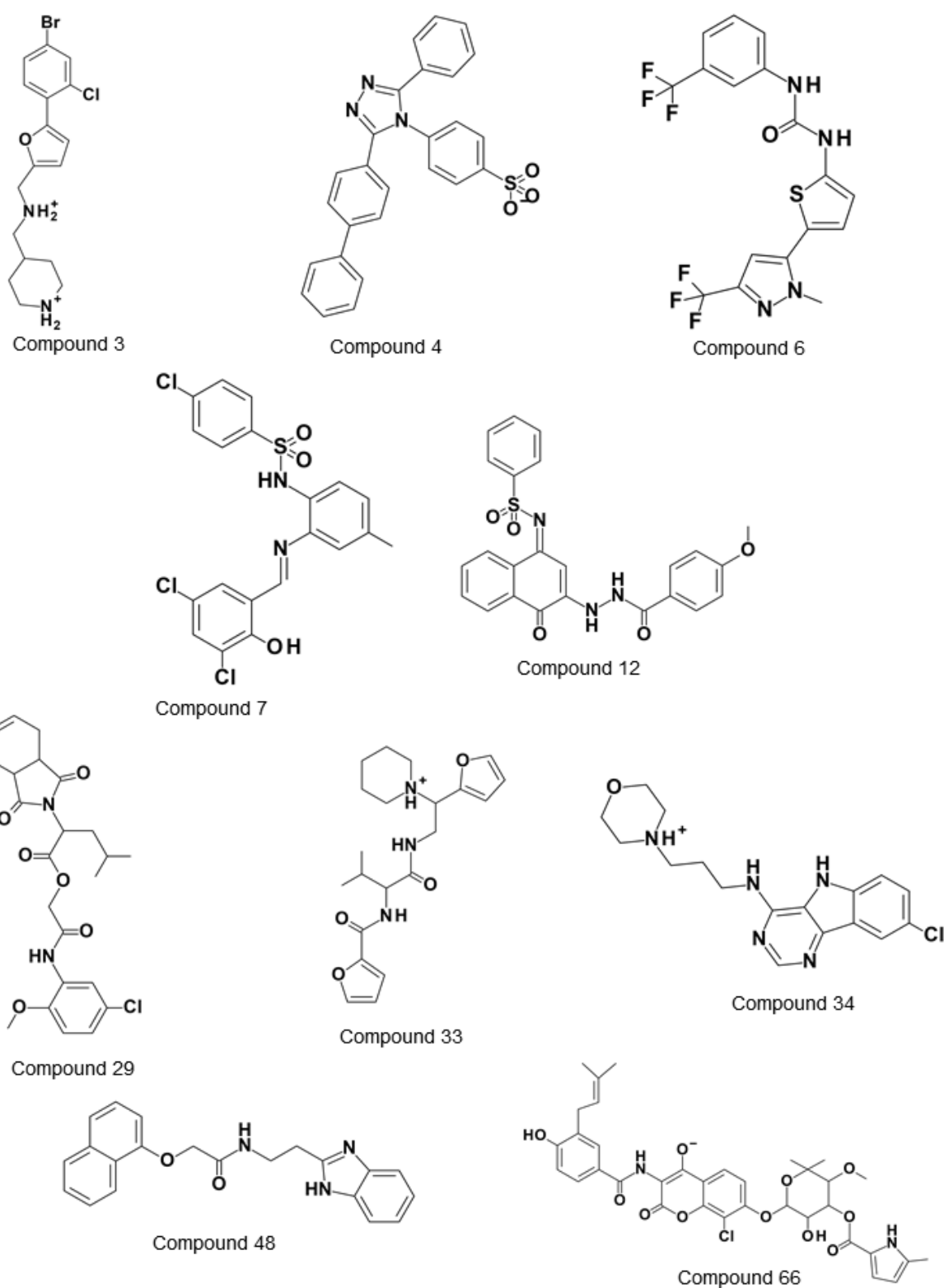


Figure 8 - Ten most chemically diverse cruzain ligands according to Morgan fingerprints.

Molecules displaying similar chemical properties to cruzain ligands were obtained in DUD-E, generating 3,772 decoys. Experimentally validated inactive compounds

against cruzain were obtained from PubChem, referring to the assay AID: 1478, from NIH Chemical Genomics Center [NCGC]. This set contains 197,846 molecules screened against cruzain in the presence of 0.01% Triton X-100, in concentrations ranging between 57.5 μ M and 3.7 nM (Jadhav *et al.*, 2010). Inactive molecules were downloaded (195,907 molecules), among which 1,980 compounds were randomly picked.

CatL active ligands applied in the analysis were from five published papers: Yamashita *et al.* 1999; Chowdhury *et al.* 2002; Marquis *et al.* 2005; Chowdhury *et al.* 2008 and Marques *et al.* 2012, totalizing 68 compounds (Appendix 1). All of them are competitive and non-covalent binders from several classes as peptidomimetics, azepanone-based, acridones and quinolinones. The potency of these compounds ranged between 0.0002 and 100 μ M. 2,766 decoys were generated with DUD-E.

After a literature search, no competitive, non-covalent binders of CatB were found. Therefore docking performance against this target was not evaluated.

5.2 Structure-Based Pharmacophore

Our purpose was to use a structure-based pharmacophore model as a pre-filter to docking in a virtual screening to identify cruzain inhibitors. However, the first objective was to evaluate whether it was possible to generate a model that was general enough to identify as many chemical classes as possible and at the same time, be selective.

Modeling a structure-based pharmacophore begins with searching for structures of the target complexed with a ligand. A search on PDB for crystal structures of cruzain-ligand complexes revealed 24 hits. It is important to obtain a high-quality model and assure if these structures have a good electronic density in the ligand region, so electronic density maps were analyzed on Coot. In Figure 9 two ligands can be observed. The first one (A) is inadequate for pharmacophore modeling, as the electron density map has low resolution and it cannot be seen where the atoms are properly located. It was discarded from the analysis together with another four complexes that did not have satisfactory electron densities. Other ten complexes were also discarded as there were no maps available. On the other hand, electron density for the second ligand (Figure 9B) allows clear determination of atom coordinates, making this a good structure for

pharmacophore modeling. Some of these high-quality crystals contain similar ligands, generated in the context of Structure-Activity Relationship (SAR) studies. In these, with the purpose of finding which part of a molecule is necessary to trigger its biological activity, researchers introduce several modifications on a scaffold. Thus, the molecules of these studies are very similar, so they were clustered, and a representative complex was chosen based on the crystal resolution. Six clusters were generated, and their representative complexes have resolutions ranging from 1.1 to 2.0 Å (Table 4).

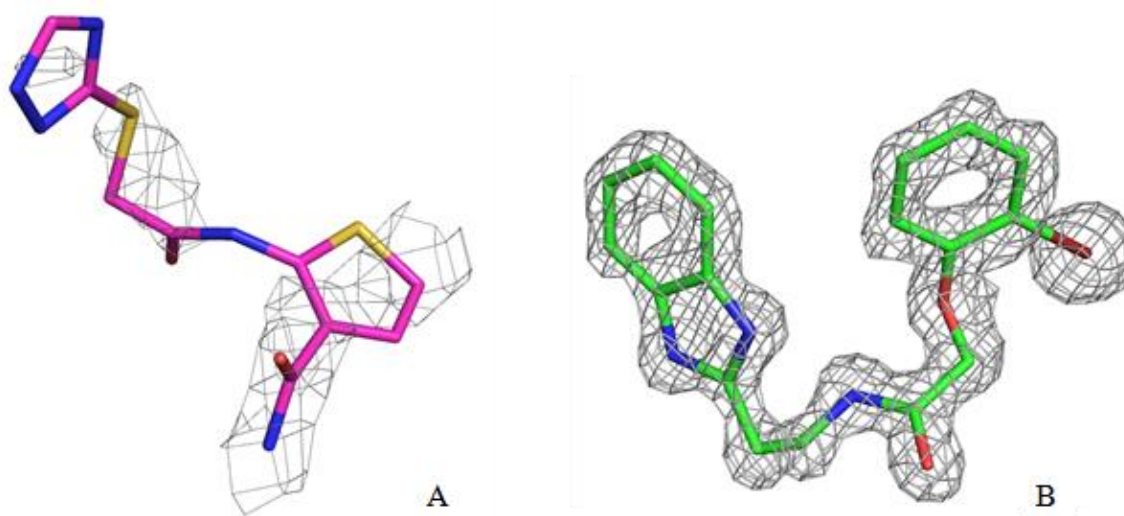
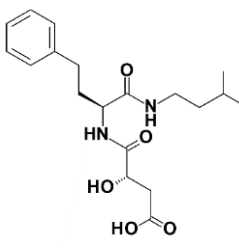
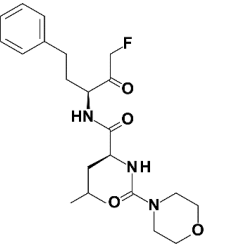
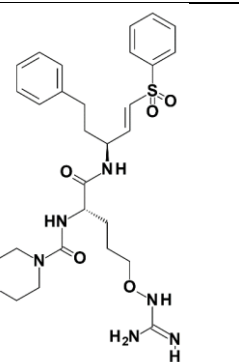
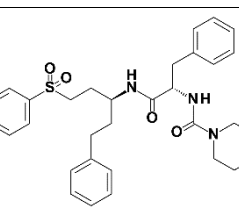
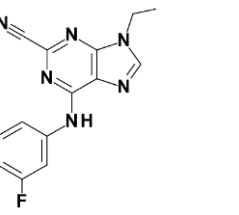
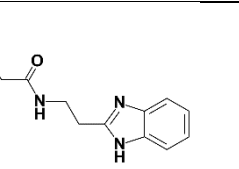


Figure 9 - Electron density from 2Fo – Fc maps of ligands from 4KLB, 2.62Å resolution (A) and 3KKU, 1.28Å resolution (B). Maps were contoured at 1 sigma.

The first step to building a pharmacophore is to define which features are important. Thus, to recognize true binders, some features in each pharmacophore were deleted (initially once by time), then the generated model was evaluated by observing the AUC and pAUC values obtained. This process was repeated until all features in a pharmacophore model were tested. After this, features that were not shown to be important were deleted in pairs and this new simplified model was evaluated again. By employing this methodology, 55 pharmacophores were generated and assessed with the active ligands and decoys set. AUC values obtained ranged from 0.50 to 0.65. The best model obtained was a simplification of the pharmacophore from 1EWP PDB structure. From its graphical representation, it can be observed that its features correspond to a hydrogen bond acceptor and a donor, which interact to GLY66. There is another donor feature interacting with GLY163 and an acceptor interacting with

ASP161. The hydrophobic area of the pharmacophore is close to ALA136, LEU67 and MET68 (Figure 10). The corresponding ROC curve was considered the best obtained because of its AUC value (0.65) and mainly due to the pAUC 10% of 0.90, meaning that in the top 10% almost exclusively active ligands were found. On the other hand, sixteen out of seventeen of these ligands are benzimidazoles derivatives, revealing low diversity of the hits recapitulated. Since the goal is to find new scaffolds to drug design, this is a drawback of this pharmacophore model.

Table 4 - PDB ID of cruzain crystals employed in structure-based pharmacophore modeling, their authors and resolution.

PDB ID and ligand structure	Authors	Resolution (Å)
<p>1EWM</p> 	Brinen, L.S. et al.	2.0
<p>1EWP</p> 	Gillmor, S.A.	1.75
<p>4PI3</p> 	Tochowicz, A., McKerrow, J.H.	1.1
<p>2OZ2</p> 	Rickert, M., Brinen, L.	1.95
<p>3I06</p> 	Ferreira, R.S., et al.	1.27
<p>3KKU</p> 	Ferreira, R.S., et al.	1.28

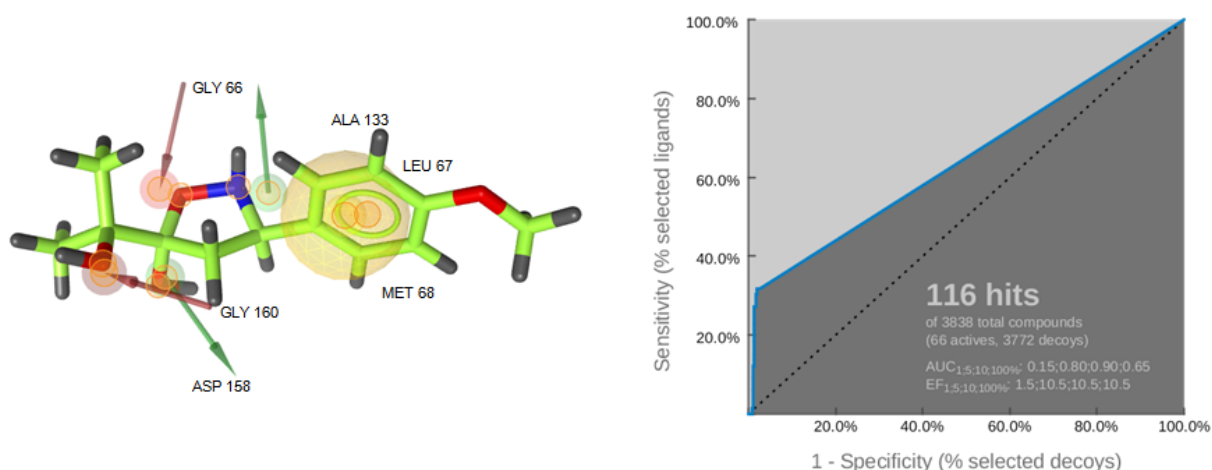


Figure 10 - The best structure pharmacophoric model generated by LigandScout. On the left, pharmacophore features with their respective residues, on the right, its ROC curve.

5.3 Docking

Before starting a docking study, it is important to study the active site of the target protein. Cruzain and the human cathepsins L and B are papain-like cysteine proteases which have very similar active sites. However, some differences can be exploited in drug design to achieve selectivity. In the S3 pocket, for example, cruzain has an ASP 60, an acidic residue, while CatL, contains an ASN, a polar neutral residue at the corresponding position (Figure 11). Another interesting difference is the ASN 70 in cruzain which is a GLU in CatL and TYR in CatB, which may allow the charge and volume of residues to be explored to reach selectivity.

The S2 pocket is a well-described pocket among these proteases, and its selectivity is addressed to the presence of a GLU 208 in cruzain and CatB, while CatL has an ALA. Because of this, cruzain and CatB can receive both hydrophobic and positively charged groups in this pocket, while CatL receives mainly hydrophobic groups, as the side chain of LEU and VAL (Castro *et al.*, 2011). Moreover, the major differences are between cruzain and CatB, while cruzain has a LEU 67, the human enzyme has a TYR 75, a bigger residue. MET 68 in cruzain is substituted to PRO 76 in CatB, a less flexible residue. ASP 161, a negatively charged residue in the parasite protease is a GLY 198,

a hydrophobic residue in CatB. Moreover, LEU 160 in cruzain is GLY 197 and MET 161 in CatB and L, respectively.

In the S1 and S1' pockets, these enzymes share a high sequence and conformational similarity, as Figure 11 illustrates. However, polar interactions with the ASP present in cruzain (ASP 161) and CatL can be exploited due to its absence in CatB, which contains a GLY at the same position.

Interesting interactions could be selected to be explored to increase selectivity towards cruzain when comparing PDB complexes of enzyme-ligand of cruzain and CatL. CatB has only one PDB complexed with a small molecule, and the electronic density map in the ligand is incomplete. These interactions are summarized in Table 5. Half of them are hydrogen bonds with polar atoms of the side chains; half are hydrophobic contacts. There are two hydrogen bonds in cruzain (SER 61 and MET 145) which cannot be observed in CatL, as in the human proteinase these residues are mutated to hydrophobic ones (Figure 11). The hydrophobic interactions selected are important to both proteases. This information was incorporated to the VS adding an interaction filter to the workflow.

An interesting observation made is that PropKa predicts two protonation states to GLU 208 in cruzain according to the nature of the group interacting with it. Figure 12 shows the PDB 1F2C, in which the ligand has a lipophilic ring and the calculated state of the GLU 208 is protonated, and pKa is 5.82. On the other hand, in PDB 1AIM the ligand has a hydroxyl and the GLU 208 is deprotonated with a pKa of 5.30. Without any ligand interacting the calculated pKa is 5.53, very close to the experimental pH. Although in this case, the program predicts it to be neutral, the protonation state is uncertain. Charge variation might be related to the group interacting with the residue. Thus, it is reasonable to perform docking with both protonation states as the charges are constant during docking. Doing so also allows to observe whether GLU 208 protonation state interferes in the results.

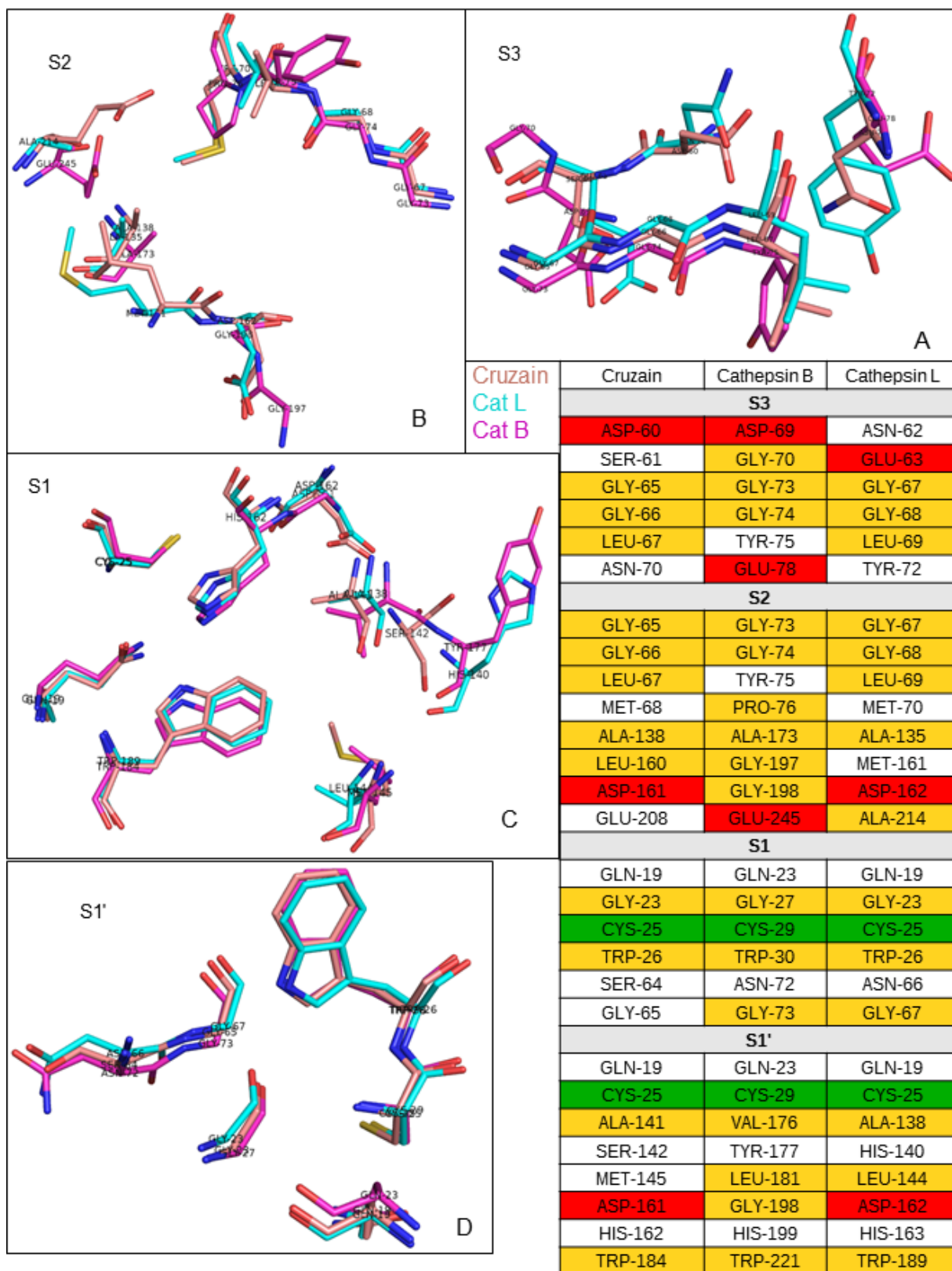


Figure 11 - Comparison between cruzain, CatL and CatB subsites. Cruzain is pink, CatL is cyan and CatB is magenta. In the table residues are colored according to this scheme: red - negatively charged; white - hydrophilic non-charged; yellow - hydrophobic; green - cysteine.

Table 5 - Atoms and respective interactions selected for the interaction filter. Interactions were analyzed with Napoli server (<http://www.napoli.dcc.ufmg.br/>). Values in the column % **ligands** were calculated as the number of ligands making this type of interaction, divided per total ligands analyzed for that protein multiplied per 100.

Residue	Atom	Type of interaction	% ligands in cruzain	% ligands in CatL
ASP161	OD1/2(-1)	Hydrogen bond ¹	6.7%	6.3%
SER61	OG	Hydrogen bond ¹	2.7%	-
MET145	SD	Hydrogen bond ¹	40%	-
LEU67		Hydrophobic ²	86.7%	100%
MET68		Hydrophobic ²	73.3%	93.8%
LEU160		Hydrophobic ²	73.3%	81.25%
ALA141		Hydrophobic ²	33.3%	12.5%

- This interaction cannot occur in the protein because the residue is different.

¹ Hydrogen bond is defined with the following parameters: minimum angle, 120°; maximum distance between acceptors and donors, 3.9Å; maximum distance between the hydrogen and the acceptor, 2.5 Å.

² Hydrophobic contacts distance is between 2.0 and 4.5Å.

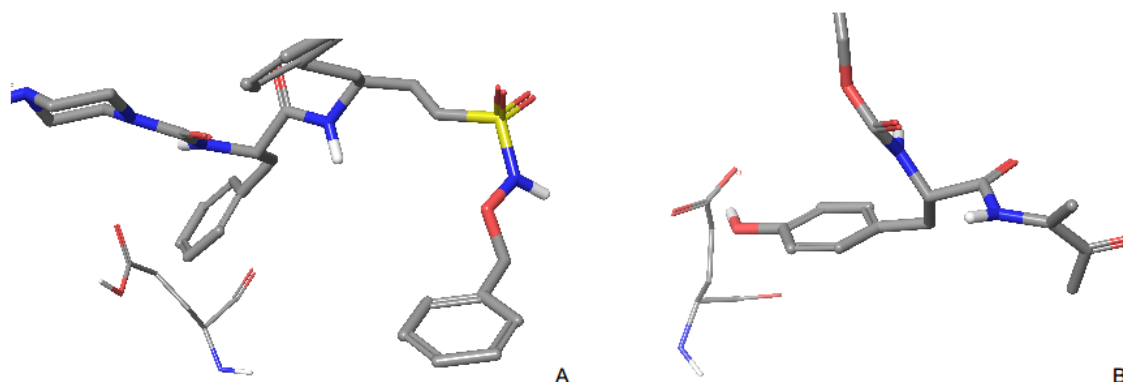


Figure 12 - GLU 208 charges calculated by PropKa. A) In the presence of an apolar GLU 208 is predicted to be neutral (PDB ID: 1F2C); B) In the presence of a polar group GLU 208 is predicted to be charged (PDB ID: 1AIM).

Furthermore, when analyzing previous docking studies from different groups, CYS 25 and HIS 162 protonation states are variable. One work set both residues as uncharged (Wiggers *et al.*, 2013), other two used a higher dipole in CYS 25 (Ferreira *et al.*, 2009, 2010); and other formed a thiolate-imidazolium pair (Rogers *et al.*, 2012). Based on

this lack of theoretical consensus, we decided to analyze whether varying the protonation states of these residues would interfere in docking enrichment metrics. Thus, four cruzain grids were prepared to be evaluated varying the protonation states of CYS 25, HIS 162 and GLU 208 residues as shown in Table 6:

Table 6 - Protonation states of HIS 162, GLU208 and CYS 25 in each grid considered for docking calculations against cruzain

	Grid 1	Grid 2	Grid 3	Grid 4
Residue	State	State	State	State
HIS 162	Neutral (HIE)	Charged (HIP)	Charged (HIP)	Neutral (HIE)
GLU 208	Neutral	Neutral	Charged	Charged
CYS 25	Neutral	Charged	Charged	Neutral

5.4 Comparison of the Virtual Screening Performance for Different Grids

5.4.1 Evaluation with cruzain inhibitors and decoys

Aiming to analyze if the various grids proposed (Table 6) would alter the docking results, the four cruzain grids were employed for docking the validation datasets of active ligands and decoys, employing both HTVS and SP Glide. Table 7 brings the enrichment metrics calculated and shows that the AUC values ranged between 0.580 (Grid 3 HTVS) and 0.704 (Grid 1 SP). For Grids 1, 3 and 4, the AUC from SP was higher than HTVS (comparing 0.704 to 0.616 in Grid 1; 0.636 to 0.580 in Grid 3; 0.652 to 0.612 in Grid 4). Thus, it is reasonable to conclude that docking SP discriminated better between active ligands and decoys. The highest AUC values were from SP Grid 1 and 4, 0.704 and 0.652 respectively, the grids with both CYS 25 and HIS 162 neutral.

AUC values for themselves are not enough to attest if the curves are different. A slight difference can be significant in some cases, while a big one may not. Delong's Test

verifies whether the AUC value obtained from two ROC curves is significant (Delong, Delong and Clarke-Pearson, 1988). When the p-value is under 0.05, the AUC values from the curves are significantly different. All the curves were compared with a random curve with 0.5 AUC. The only curve whose AUC was not different from 0.5, was from Grid 3 HTVS (p-value = 0.06692). Venkatraman's test was performed to verify if the ROC curves shape were significantly distinct from the form of a random ROC curve. All the ROC curves obtained were different from a random curve.

Table 7 - Enrichment metrics obtained for the cruzain validation database, employing different docking methods and Grids

Grid	Method	AUC	p-value (Delong's)	p-value (Venkatraman's)	EF 1%	EF 2%	EF 10%
Grid 1	HTVS	0.616	0.00323	0.003	9.1	6.1	2.6
	SP	0.704	1.68x10 ⁻¹¹	< 2.2x10 ⁻¹⁶	9.1	6.9	2.4
	XP	0.615	0.001427	0.0025	7.6	5.3	2.0
Grid 2	HTVS	0.623	0.004196	0.0035	6.0	4.5	1.5
	SP	0.615	0.00718	0.005	7.5	3.8	2.1
Grid 3	HTVS	0.580	0.06692	0.049	11.0	5.3	1.8
	SP	0.636	0.001483	< 2.2x10 ⁻¹⁶	6.0	3.0	2.9
Grid 4	HTVS	0.612	0.01259	0.0145	6.0	3.8	1.8
	SP	0.652	0.0002817	< 2.2x10 ⁻¹⁶	9.0	5.3	2.3

The main intention behind a VS is to reduce the quantity of experimentally tested compounds, so this method is used to help to choose the molecules to be prioritized for *in vitro* assays. Therefore, the objective is not just to discriminate real positives from real negatives, but also to rank the actives at the very top. EF is the metrics to be looked in this sense.

HTVS and SP docking were used in the intermediate steps of VS with their top 10% ranked molecules being submitted to the following step. Hence, EF10% values were analyzed to compare the different grids performance (Figure 13). In Grid 2, Grid 3 and Grid 4 the SP docking rank was more enriched with actives ligands in top 10%. This data reveals that the enrichment differences obtained with the grids were minimal. In HTVS docking the higher value was from Grid 1 and in SP docking, from Grid 3.

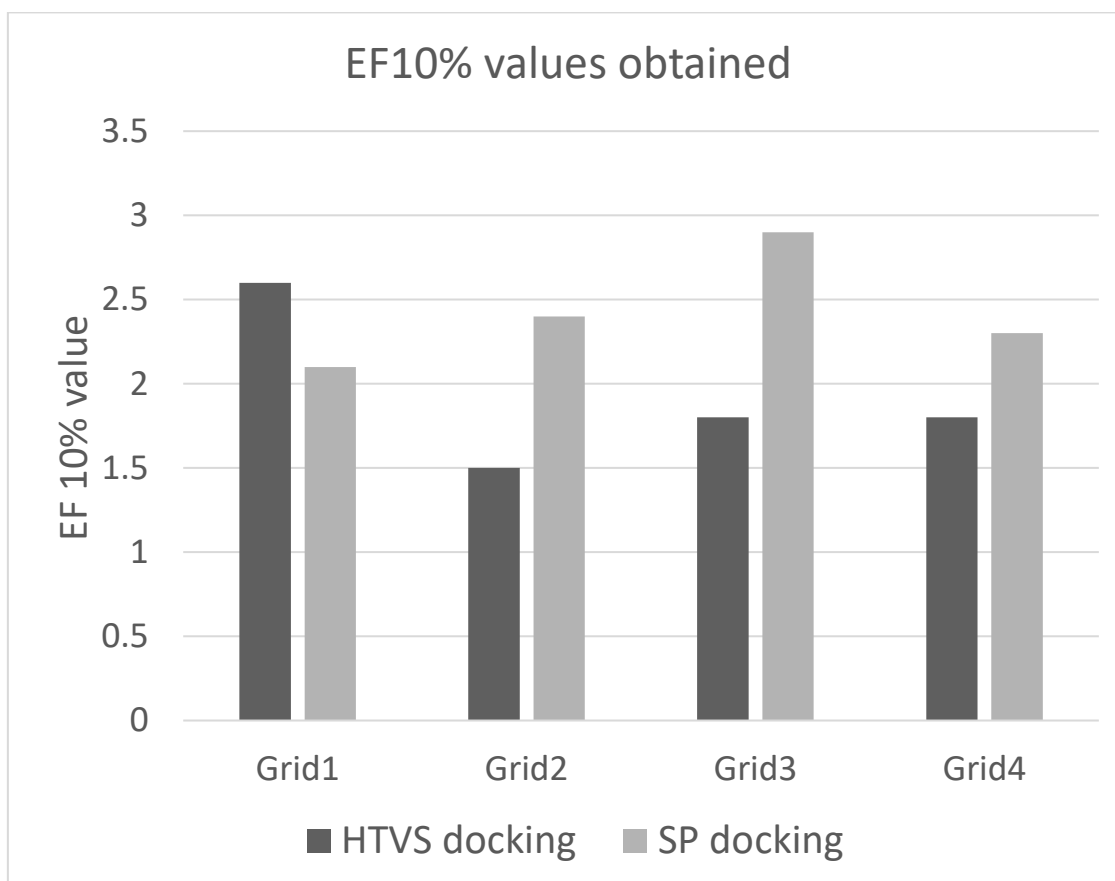


Figure 13 - EF10% values obtained in the grids for Glide HTVS and SP docking.

We wanted to check also whether varying residues protonation states could interfere in the docked ligands. Ligands ranked in the top 10% in each docking were analyzed to answer this question. In HTVS docking 10 to 14 active ligands (15 to 21% of total actives) were retrieved in the top 10%. Among these, four ligands appeared in all ranked lists, all from Huang et al., 2003: **59**, **62**, **63** and **64** (Figure 14). Therefore, we observed that in HTVS docking the different protonation states of protein residues can alter the nature of docked ligands.

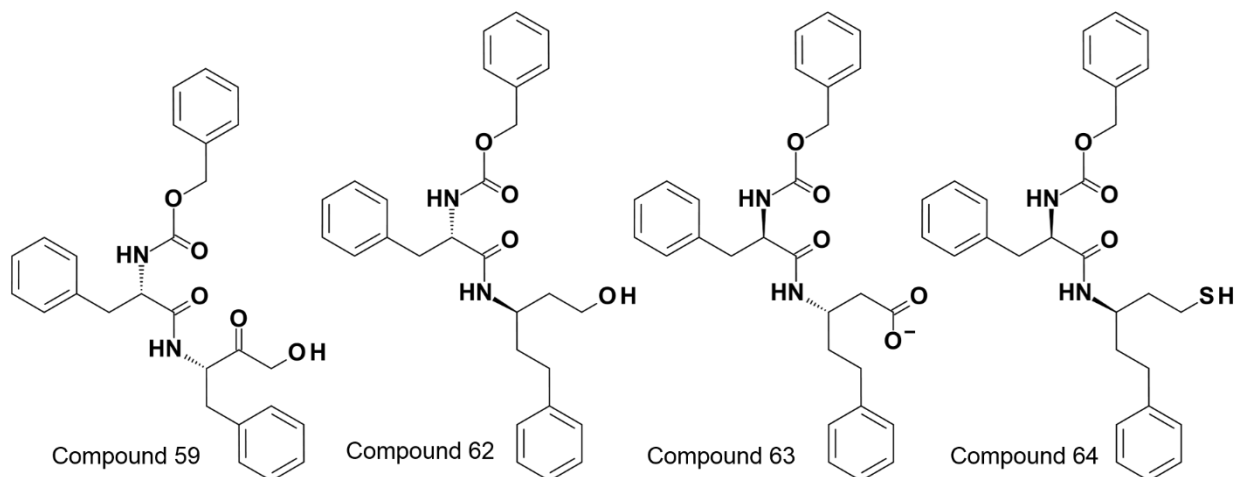


Figure 14 - Compounds retrieved in the top 10% of HTVS docking against cruzain in all the Grids.

When analyzing Glide SP results, the number of active ligands recovered in the top 10% of the screen varies from 14 to 19 (21 to 28% of all actives). Looking at the compounds that appear in all grids, the number is the double when compared with HTVS, including the four from Figure 14 (**59**, **62**, **63**, and **64**) plus **31**, **60**, **61** and **66** (Figure 15).

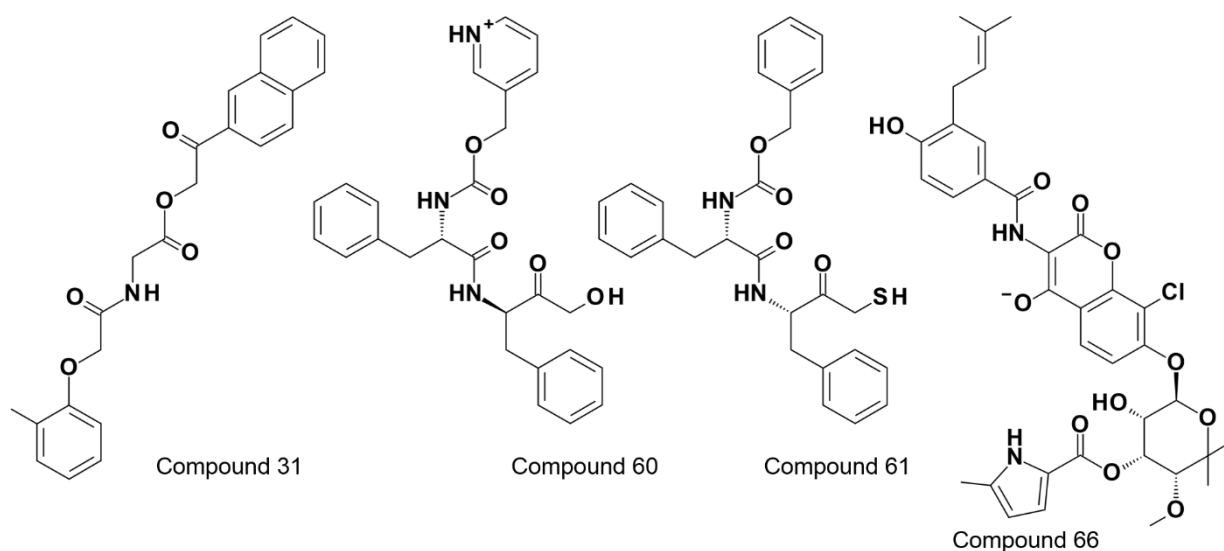


Figure 15 - Compounds retrieved in the top 10% of SP docking in all the Grids.

Even if a compound is highly ranked in several grids, the pose predicted may differ among them. Therefore, we also evaluated whether poses for the same compound

retrieved from different grids were consistent. Some compounds have similar poses predicted among the grids. Compound **59** in HTVS docking was chosen to illustrate this. It has a conserved hydrogen bond pattern. The only group varying its conformation is one of the benzyl groups, which is highly exposed to solvent. It is attached to the rest of the molecule by a flexible bond, and only in Grid 1 its predicted to interact with the S3 pocket (Figure 16).

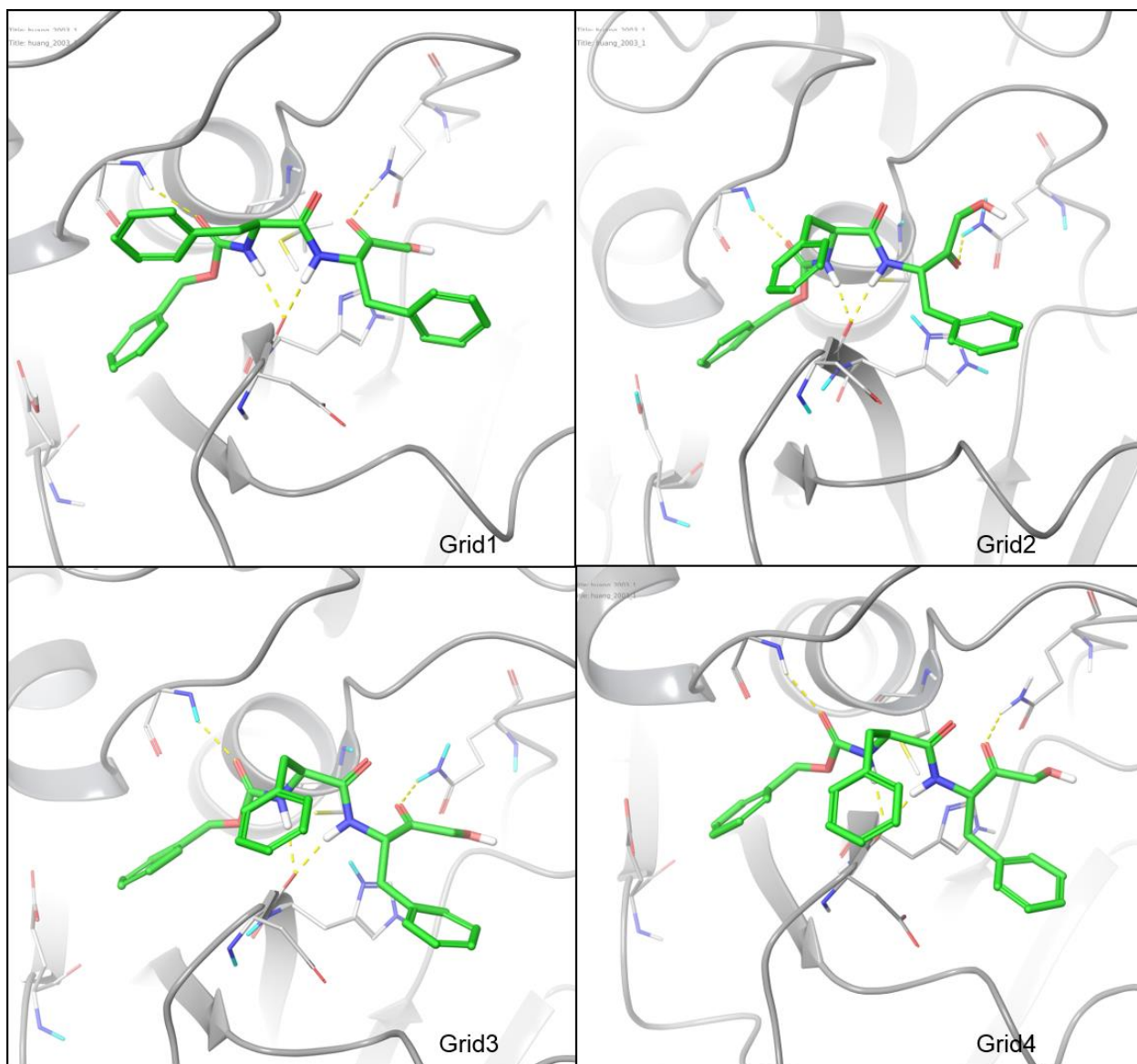


Figure 16 - Compound 59 predicted poses among the four grids did not exhibit a high variation. The only group with a significant difference was the benzyl, it was exposed to the solvent (Grid 2 to 4) or occupying S3 pocket (Grid 1).

Some compounds contain groups with multiple possible protonation states in the pH range of the docking against cruzain (5.5 ± 2.0). Aiming to investigate if the different ligand protonation states could interfere in the predicted pose, some compounds were analyzed. Compound **40** was the only compound whose best-scored pose predicted had its protonation state varied in HTVS docking. All other molecules had the same calculated charge in the top 10 percentage screen or, do not have ionizable groups. In compound **40** the benzimidazole ring is protonated in Grid 1 and deprotonated in Grid 4. Despite this difference, the poses predicted were very similar, with the benzimidazole ring positioned in S1 (Figure 17).

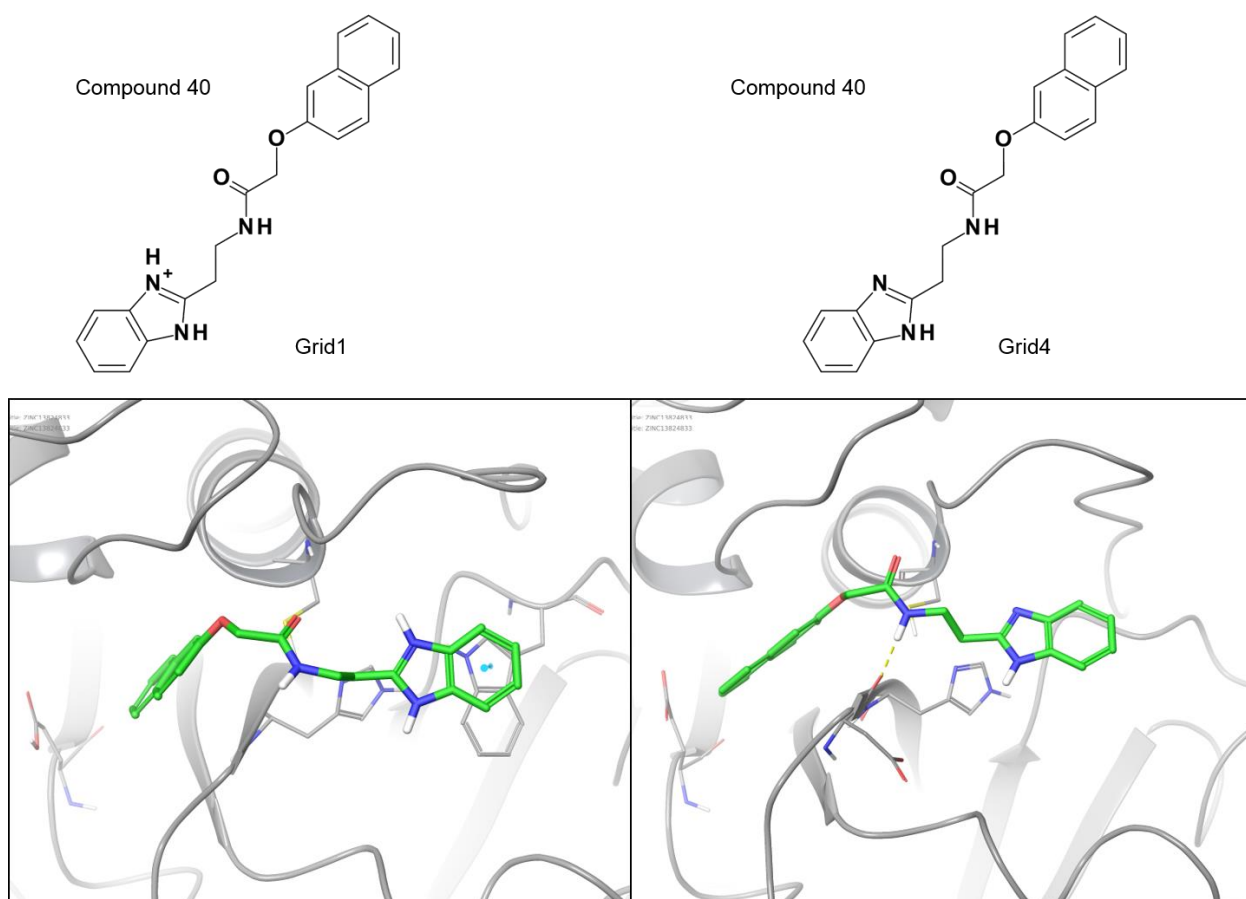


Figure 17 - Protonation states variation in compound 40 does not interfere in pose prediction in docking HTVS

A case in which the charges of the molecule and the receptor influenced in the predicted pose was observed in SP docking. Compound **60** has a pyridine ring that can be charged or not at docking conditions ($\text{pH } 5.5 \pm 2.0$), as the calculated pK_a is

4.7 (ACE and JChem acidity and basicity calculator, available in <https://epoch.uky.edu/ace/public/pKa.jsp>). In Grid 1 to 3 the calculated charge of the best-predicted pose is neutral, but in Grid 4, it is positively charged and positioned towards the negatively charged GLU 208 (Figure 18). The charged nitrogen and the oxianion are 3.7Å apart and therefore involved in a salt bridge interaction.

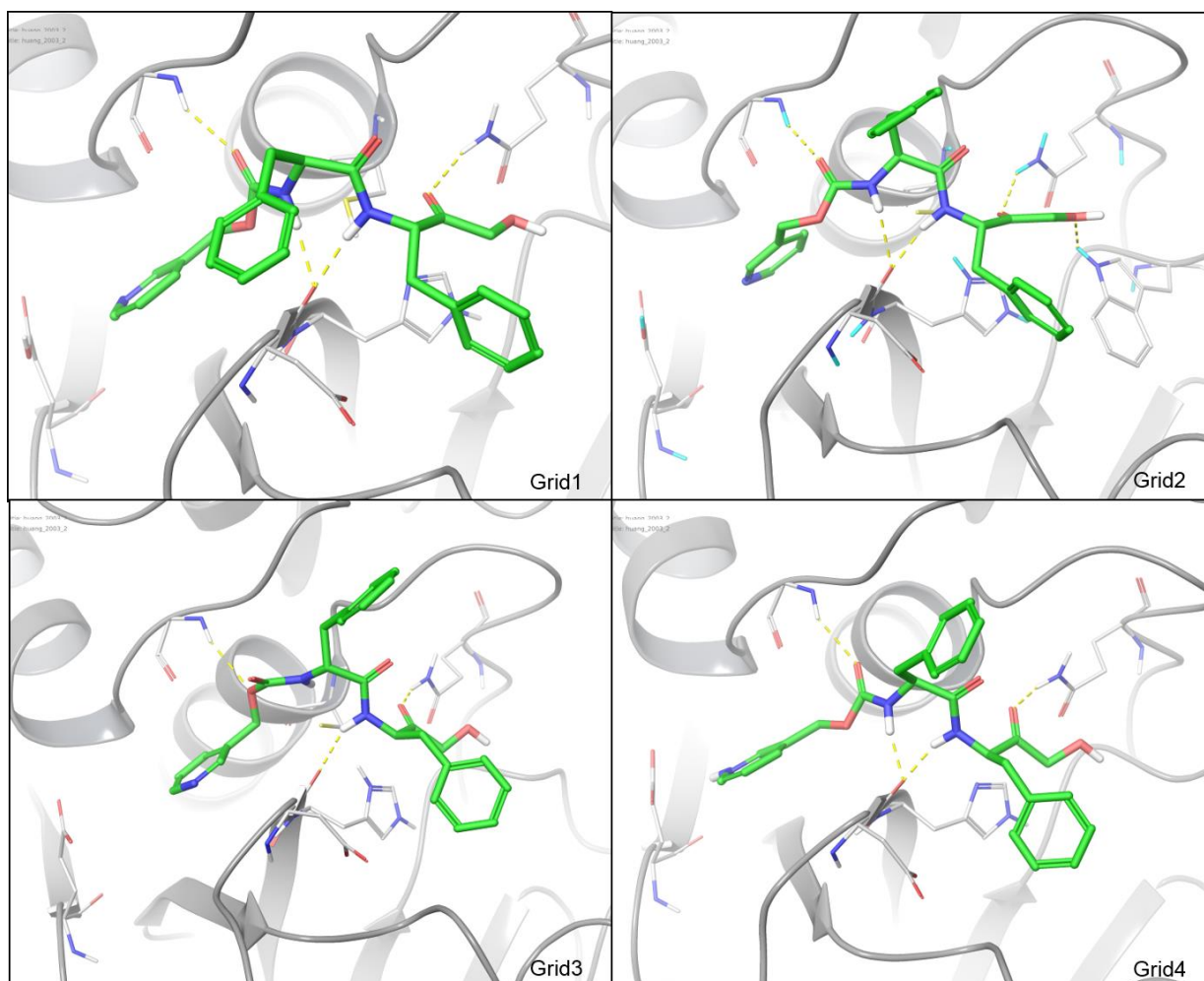


Figure 18 – Pose prediction of compound 60 in docking SP is influenced by the protonation states of the residues in the active site.

Docking XP calculations need considerably more CPU time than SP and HTVS docking, so, only Grid 1 was evaluated in this method. Its AUC value was 0.615, lower than the values obtained to HTVS and SP, although the difference is marginal when compared to HTVS. These results agree with a previous study called CASF-2013 (Comparative Assessment of Scoring Functions) (Li *et al.*, 2014). Using a benchmark docking scoring functions were evaluated in terms called Scoring Power Test, Ranking

Power Test, Docking power test and Screening Power Test. The overall conclusion is that Glide scoring functions (SP and XP) are not useful in predicting binding affinities or even to rank the molecules properly. Still, SP had a better performance than XP docking.

Concerning the ability to identify the native binding pose of a ligand, both methods were well evaluated. If only the best-predicted pose is considered, the success rate is 79% to SP and 75% to XP (Li *et al.*, 2014). Moreover, they had a good performance in finding the true binders among random molecules and ranking them in the top. EF1% of SP was 19.54, while in XP was 16.81. EF10% of SP was 4.14, while in XP was 4.07 (Li *et al.*, 2014).

Based on these analyses it is plausible to conclude that no Grid had a superior enrichment performance. Despite this, different compounds were ranked among the grids. Thus, it could be worthwhile to use more than one grid to enhance the possibility of selecting a real active for *in vitro* assays. Moreover, this could help to identify scaffolds that would not be identified if only one grid was used.

Desiring to improve the chances of obtaining cruzain competitive inhibitors, both charged and neutral GLU 208 were utilized in the VS. Between the grids with the neutral GLU 208, Grid 1 had a considerably higher EF10% value in HTVS docking. So, Grid 1 was selected to the VS. In the case of the charged GLU208, the EF10% from docking SP were appreciably higher in Grid 3, which was chosen for VS. Finally, the compound selection from the virtual screening was performed looking at the best-predicted poses, keeping in mind that their ranking position is not necessarily related to its binding affinities.

5.4.2 Cruzain evaluation with actives and experimentally inactive molecules

Besides evaluating the Grids with computationally generated decoys, in the cruzain case, it was also checked how docking would deal with real inactive molecules. Table 8 summarizes the results. AUC values ranged from 0.651 to 0.756, higher than those obtained with decoys. A possible explanation is that the decoys were obtained assuring to be chemically similar to the actives, while the inactive molecules were picked randomly. It is “easier” for docking to identify actives in a set with high chemical diversity

than in a group with a narrow chemical diversity. Comparing the Grids with GLU 208 neutral (Grid 1 and 2), the highest AUC to HTVS is from Grid 1 (0.715), and in SP the values are very close (0.712 for Grid 1 and 0.716 for Grid 2). Looking at the EF10%, the higher enrichment in HTVS is from Grid 1 (4.2 against 3.2) and in SP also (4.8 against 4.2).

Comparing the grids with GLU 208 charged, the highest AUC value from HTVS was obtained by Grid 4 (0.709 against 0.651), and in SP, again Grid 4 had the highest value (0.756 against 0.749). When EF10% is analyzed in HTVS, Grid 4 has the highest value (3.8 against 2.7). In SP, Grid 3 has the highest value, 4.5 against 4.1. Therefore, work with experimental inactive molecules confirmed the results discussed with decoys.

Table 8 - Enrichment metrics calculated to docking against cruzain with actives and inactive molecules

		AUC	p-value (DeLong's)	p-value (Venkatraman's)	EF 1%	EF 2%	EF 10%
Grid 1	HTVS	0.715	3.00×10^{-8}	$< 2.2 \times 10^{-16}$	11	7.5	4.2
	SP	0.712	3.14×10^{-6}	$< 2.2 \times 10^{-16}$	12	9.7	4.8
Grid 2	HTVS	0.693	1.722×10^{-5}	$< 2.2 \times 10^{-16}$	12	9.8	3.2
	SP	0.716	7.412×10^{-8}	$< 2.2 \times 10^{-16}$	14	8.2	4.2
Grid 3	HTVS	0.651	0.001188	5e-04	15	8.3	2.7
	SP	0.749	2.049×10^{-8}	$< 2.2 \times 10^{-16}$	12	12	4.5
Grid 4	HTVS	0.709	1.503×10^{-6}	$< 2.2 \times 10^{-16}$	12	8.3	3.8
	SP	0.756	1.121×10^{-10}	$< 2.2 \times 10^{-16}$	17	13	4.1

5.4.3 Enrichment results for docking the CatL ligand database

Docking performance was also evaluated for CatL database of active molecules and decoys. Protonation states of CYS 25 and HIS 163 were kept as predicted by PropKa: neutral and positively charged, respectively. This protein has an ALA 214 in the corresponding position of GLU 208 in cruzain.

68 active molecules and 2,766 decoys were evaluated in these analyses. Table 9 summarizes the results. All the AUC values and ROC curves were significantly different from a random ROC curve. The AUC value obtained in docking HTVS was lower than SP (0.748 and 0.776, respectively). Enrichment factors from SP were also higher than the values obtained for HTVS.

Table 9 - Enrichment metrics obtained in docking studies against CatL

	AUC	p-value (DeLong's)	p-value (Venkatraman's)	EF 1%	EF 2%	EF 10%
HTVS	0.748	1.137×10^{-10}	$< 2.2 \times 10^{-16}$	2.9	4.4	4.4
SP	0.776	1.137×10^{-10}	$< 2.2 \times 10^{-16}$	17	16	5

Based on these values it can be concluded that Glide docking can distinguish well between actives and decoys in the tested dataset. Thus, it is possible that in the virtual screening the top of ranking would be enriched with active ligands. On the other hand, the bottom of the ranked list might be full of molecules that are not able to inhibit CatL.

Based on these results, we decided to address the selectivity issue in the present VS by selecting compounds that are at the same time in the top of cruzain docking results and in the bottom of CatL and CatB docking results.

5.5 Virtual Screening

5.5.1 Virtual screening database preparation

The database for VS, was the ZINC Leads Now molecules (www.zinc.docking.org), which contains 3,687,621 molecules with a molecular weight between 250 and 350 g/mol; xlogP under 3.5 and less than seven rotatable bonds. Then, were selected a subset of molecules clustered and filtered based on a $T_c = 0.9$. To apply this filter, first the molecules are ranked according to molecular weight, in an ascending fashion. Then, a compound is selected if it differs from the previously chosen by the T_c of 0.9 using fingerprints.

372,632 molecules were downloaded from ZINC and after preparation with LigPrep 321,811 molecules were retrieved. This number is smaller than the input because MMFFs force field used by LigPrep does not have all the atom types found in the ligands file, so these ligands were not prepared and were removed from the virtual screening. Figure 19 illustrates some non-processed molecules. As it can be observed, they contain groups as a carbocation, a positively charged sulfur in an aromatic ring and a phosphorus making five bonds. These groups are not interesting in drug development, so discarding them did not represent a problem. Some of them might be a result of errors in ZINC database or misinterpretation of molecules graphs from Maestro.

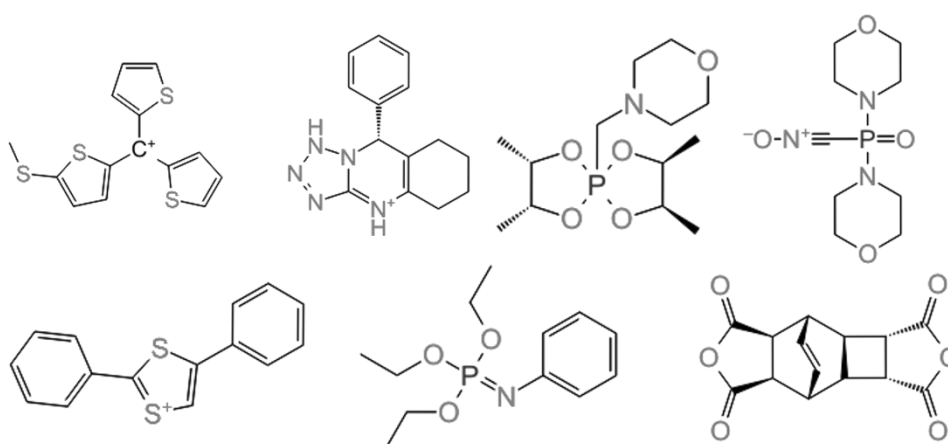


Figure 19 - Example of ZINC compounds MMFFs force field was not able to process during ligand preparation.

5.5.2 Virtual Screening Workflow

A hierarchical virtual screening was performed with two grids in parallel, Grid 1, with GLU 208, HIS 162 and CYS 25 neutral, and with Grid 3, which has these residues charged. The first step was to dock ZINC compounds against cruzain with Glide HTVS. The second step was to select top ten percent ranked molecules (54,639 molecules in Grid 1 and 42,466 molecules in Grid 3) and dock them against cruzain once more, but this time using docking SP. The third step was to submit them to the interaction filter (Table 5). If they satisfied at least one condition, they were kept, otherwise they were discarded. After this step, 5,461 molecules were retrieved from SP docking results with Grid 1 and 3,312 molecules with Grid 3.

Then, these compounds were docked against CatL and B with docking SP. Next, the bottom ten percent of ranked molecules were docked in the mode XP against the three enzymes (489 compounds from Grid 1 and 578 compounds from Grid 3). Finally, molecules were visually inspected and selected to be purchased and evaluated in *in vitro* assays. Figure 20 summarizes the VS workflow employed.

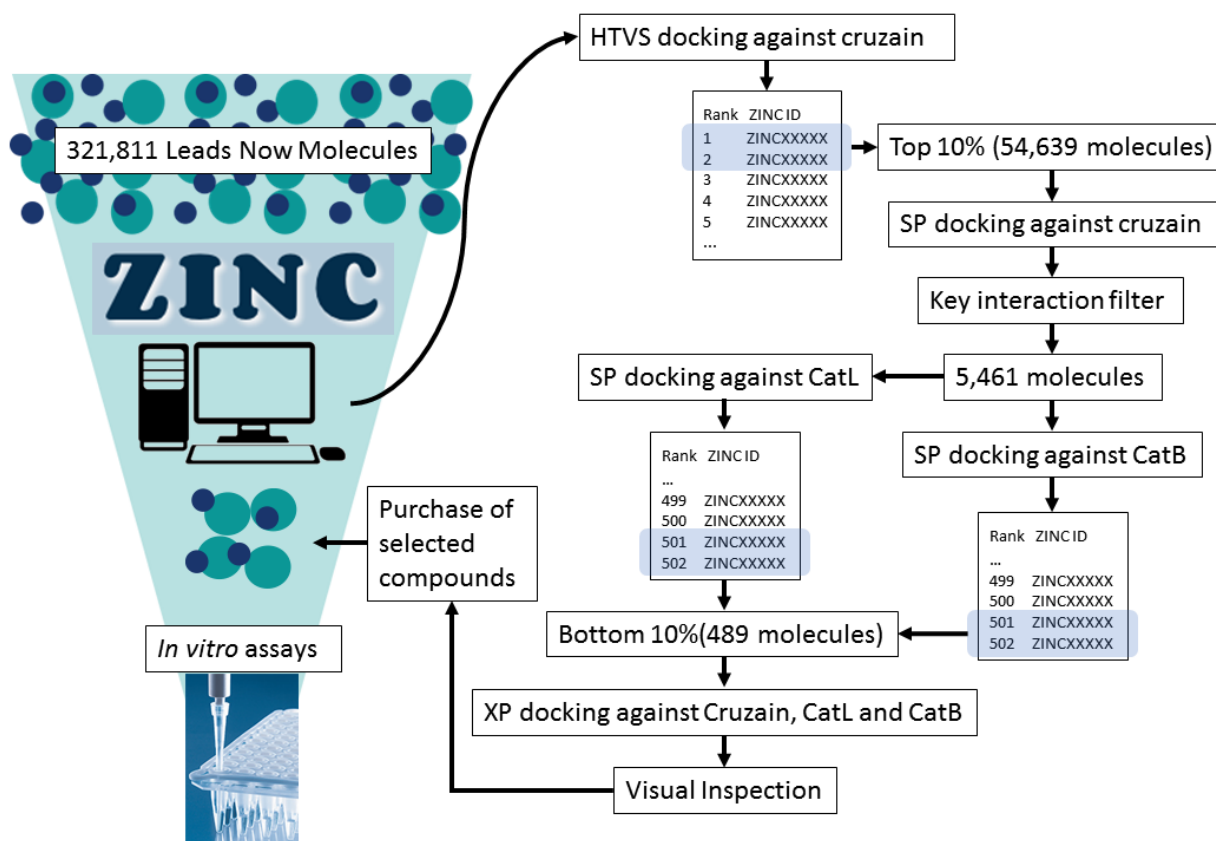


Figure 20 – VS workflow. Zinc molecules were submitted to a hierarchical VS starting with HTVS docking against cruzain. Molecules were ranked according to their calculated ΔG by the docking algorithm. Top 10% of ranked molecules were submitted to SP docking against cruzain. Then, compounds which filled at least one interaction from the interaction filter were submitted to SP docking against human cathepsins L and B. Molecules were ranked according to their docking scores (calculated ΔG) and the bottom 10% were subjected to XP docking against the three enzymes. Molecules were visually inspected and selected to purchase and *in vitro* assays.

5.6 Compound Selection for *In Vitro* Assays

Compound selection for *in vitro* assays was based not solely on compound ranks, but mainly on a visual inspection of their predicted poses. In this inspection it was verified how well the cruzain S2 pocket was occupied, the hydrogen bonding pattern and overall chemical complementarity. Additionally, compounds displaying highly reactive group were discarded, since non-covalent binders were the focus of this research. Finally, the chemical diversity and hits commercial availability selected were considered.

All selected compounds occupy S2 pocket in cruzain, and many of them establish a hydrogen bond with the oxygen present in the ASP 161 backbone. This interaction was also conserved in the predicted poses against CatB, except by the fact the ASP 161 is a GLY 122 in this enzyme. On compounds well ranked by Grid 3 salt bridge interactions with CYS 25 were observed, while in top hits from Grid 1 this interaction was observed with the charged ASP 161. Pi-stacking interactions were observed only in the CatB.

Eight compounds were selected, two of them are thought to be selective cruzain inhibitors over CatL and CatB. All the others are supposed to be cruzain inhibitors selective over one of the human cathepsins. Top-ranked cruzain molecules were inspected first and had their predicted poses compared to the human enzymes. Then, CatL and CatB poorly ranked molecules were analyzed and had their predicted poses compared with cruzain. The following discussion refers to molecules selected by the first approach.

Compound ZINC97114414 (Figure 21) has its polar groups turned towards solvent and makes a salt bridge between the positively charged nitrogen of methyl thiazole with the negatively charged sulfur of CYS 25. In cruzain and CatL the methyl imidazole pyridine ring is well suited to S2, but in CatB, this ring occupies S3 pocket. Cruzain active site is more similar to CatL active site than to CatB active site. That might be the reason why the predicted pose in CatB is considerably different from the others. Its S2 pocket is not so tight as in other two enzymes, and the ligand is well fitted in S1 and S1' pocket, being stabilized by the electrostatic interactions. In CatB, the five-member heterocycle makes a pi-stacking with HIS 199 (HIS162 in cruzain). Moreover, the molecule is involved in a hydrogen bond with GLU 122 in CatB (in the other enzymes this residue is absent). There is one conserved interaction in all the poses predicted. It is a hydrogen bond with ASP 161 (cruzain)/ ASP 162 (CatL)/ GLY 198 (CatB). Based on these poses and the ranking position among the grids, this molecule is supposed to be a selective cruzain hit over CatB.

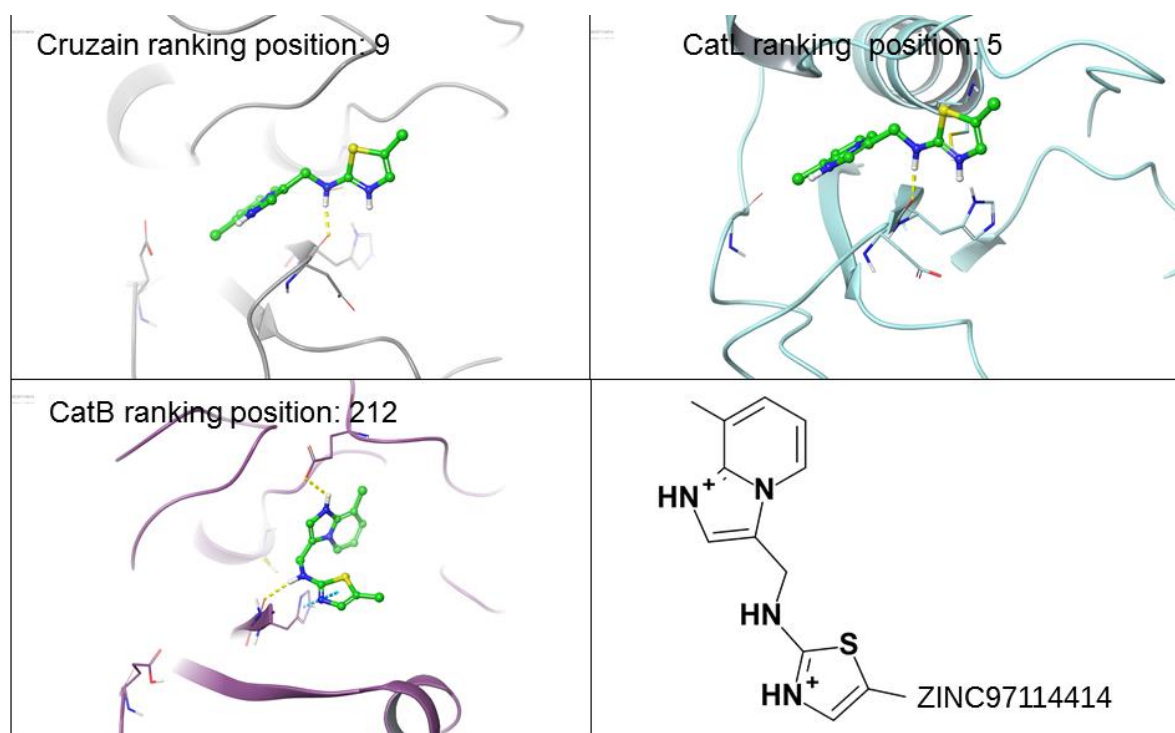


Figure 21 - Predicted poses and chemical structure of ZINC97114414, a putative selective cruzain inhibitor, likely inactive against CatB.

Another supposed selective cruzain hit over CatB is ZINC81113908 (Figure 22). It has a well-stabilized pose in cruzain, establishing two conserved hydrogen bonds with the

backbone of GLY 66 and ASP 161 and two salt bridges between the charged catalytic cysteine and the positively charged groups from the rings. In CatL and CatB the poses are very similar, except by a pi-stacking predicted in CatB with HIS 199. However, as the active site is shallower in CatB, it is supposed to be a weaker binder in this enzyme than in the others.

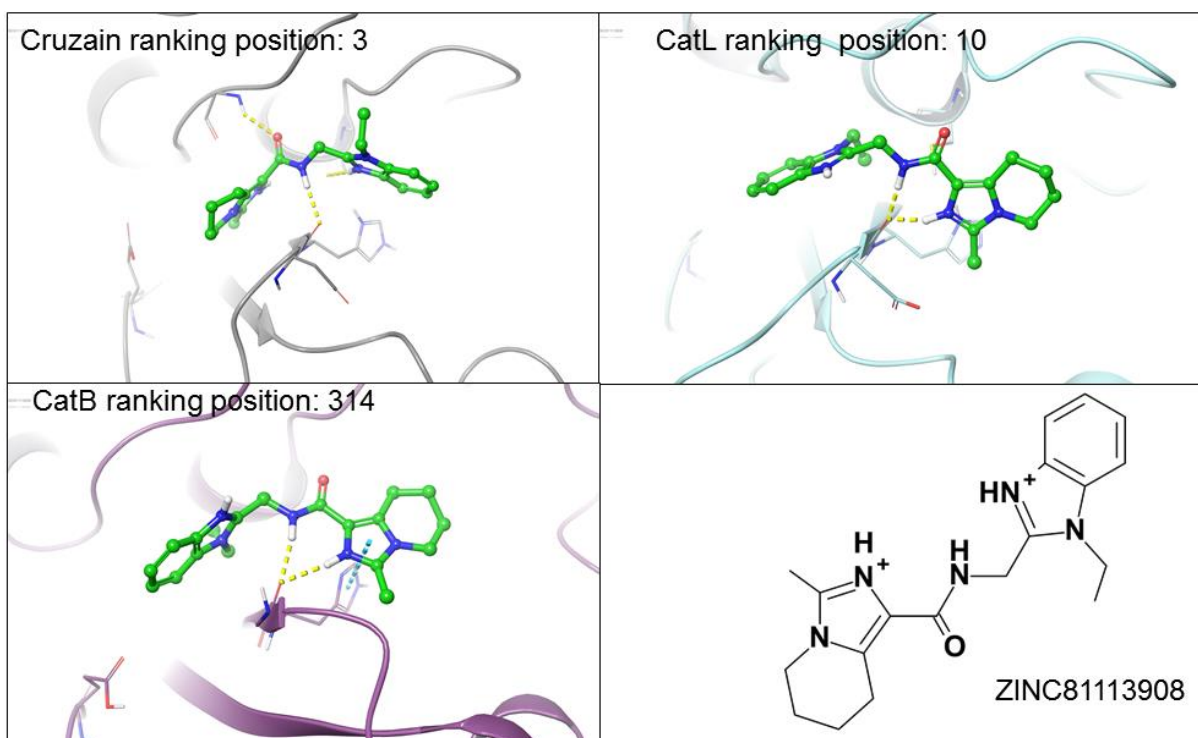


Figure 22 - Predicted poses and chemical structure of ZINC81113908, a supposed selective cruzain inhibitor over CatB.

A putative selective cruzain inhibitor over CatL is ZINC95480744, with predicted poses in CatB and cruzain occupying from S2 to S1. In CatL in which the S2 pocket is not so closed as in cruzain, the ligand is not predicted to occupy this pocket. There is no hydrogen bond predicted, and there is a great portion of ligand exposed to solvent in CatL making this pose less stable (Figure 23). On the other hand, in cruzain, there is a hydrogen bond with ASP 161 backbone and in CatB, besides the corresponding interaction, there are two others. One with GLY 74 and other with GLU 122.

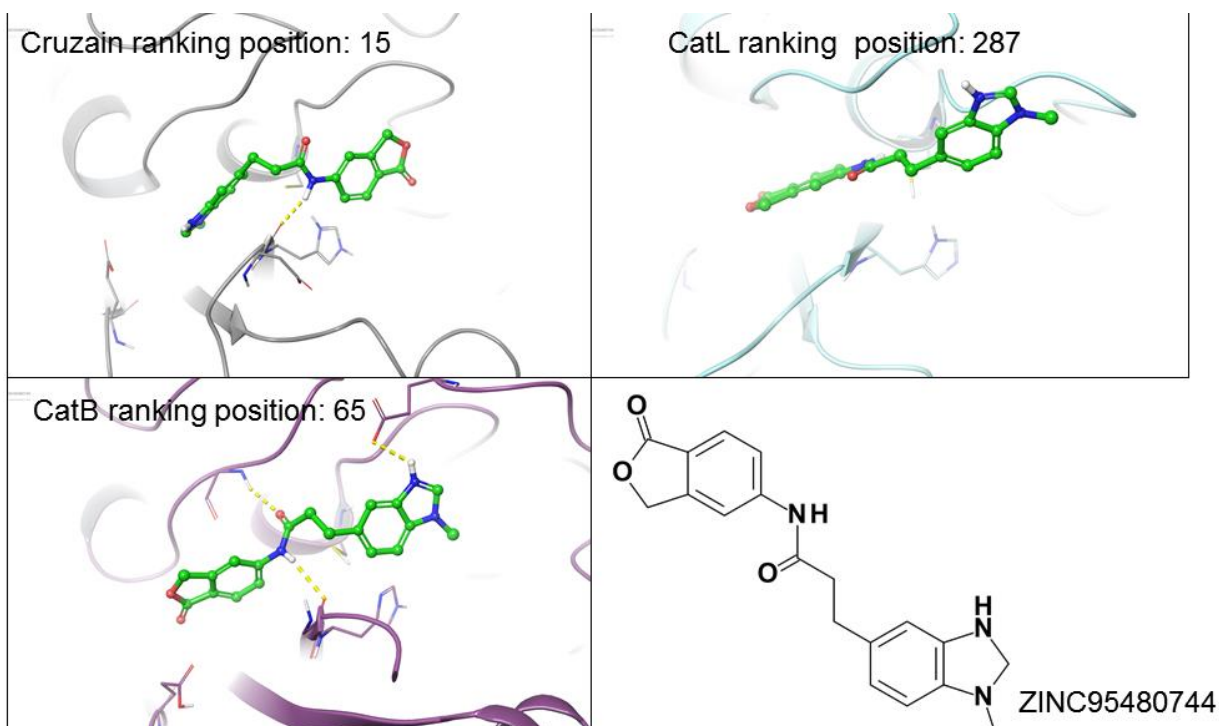


Figure 23 - Predicted poses and chemical structure of ZINC95480744, a supposed selective cruzain inhibitor over CatL.

ZINC55314924 is expected to be a selective cruzain inhibitor over CatL and CatB (Figure 24). Its predicted pose in cruzain has all the polar groups involved in hydrogen bonds or polar contacts. In CatL only the nitrogen atoms are not exposed to the solvent, the nitrogen from the ring is involved in a salt bridge with ASP 162, and the other is making a hydrogen bond with the backbone of this same residue. In CatB there is the conserved hydrogen bond with GLY 198. As the ligand pose predicted in cruzain is more stable than in the others, it is proposed that this compound might have a higher activity against cruzain.

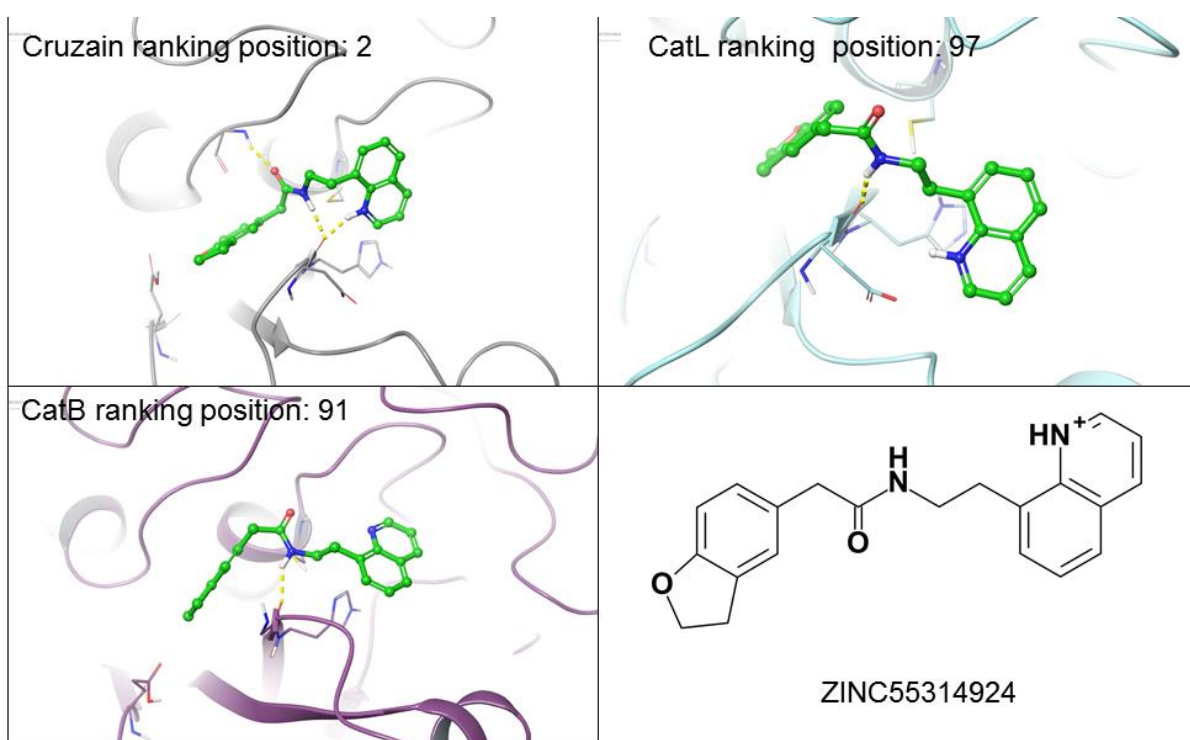


Figure 24 - Predicted poses and chemical structure of ZINC55314924, a supposed selective cruzain inhibitor over CatL and CatB.

Compound ZINC05173978 is predicted to be a cruzain selective inhibitor over CatB. The pose in cruzain and CatL is well suited in S2 and S1', with benzene in S2 and thiazole in S1' making a salt bridge interaction with ASP 161 in the cruzain. The other poses also have the benzene in the S2, but in the CatB the thiazole is making a salt bridge interaction with GLU 122, in a pocket very exposed to the solvent so that might be a weaker binder to this enzyme (Figure 25).

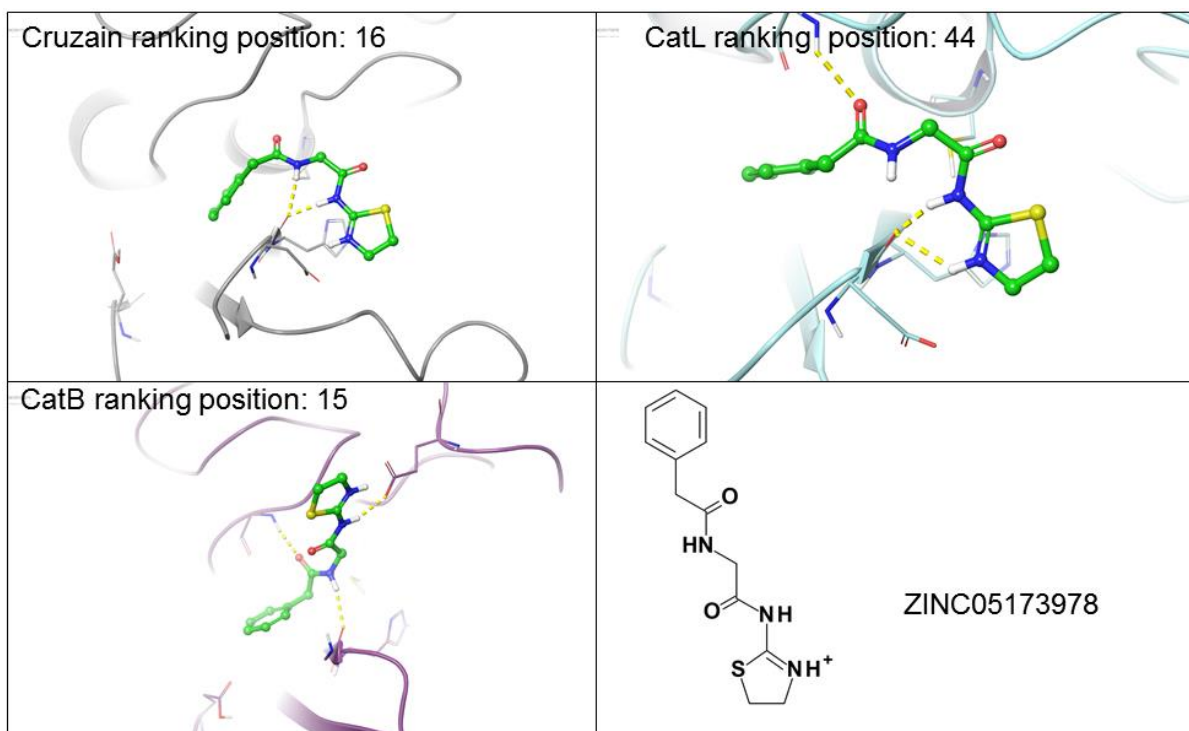


Figure 25 - Predicted poses and chemical structure of ZINC05173978, a supposed selective cruzain inhibitor over CatB.

Until this point, the compounds described were selected because they were well ranked against cruzain. Molecule ZINC71859319, however, was chosen because it is predicted to be a weak CatL binder (Figure 26). The molecule is positioned in the S2 and S3 pockets, the latter a pocket very exposed to solvent, and it is only stabilized by a single hydrogen bond. In cruzain, on the other hand, despite its bad ranking position (302), the pose is plausible, occupying S2 and S1 pockets and making a hydrogen bond with ASP 161. In CatB there is an excellent predicted pose, occupying S2 and S1' making key hydrogen bonds with GLY 74 and GLY 198. Based on these analyses is supposed that this ligand is a cruzain inhibitor selective over CatL.

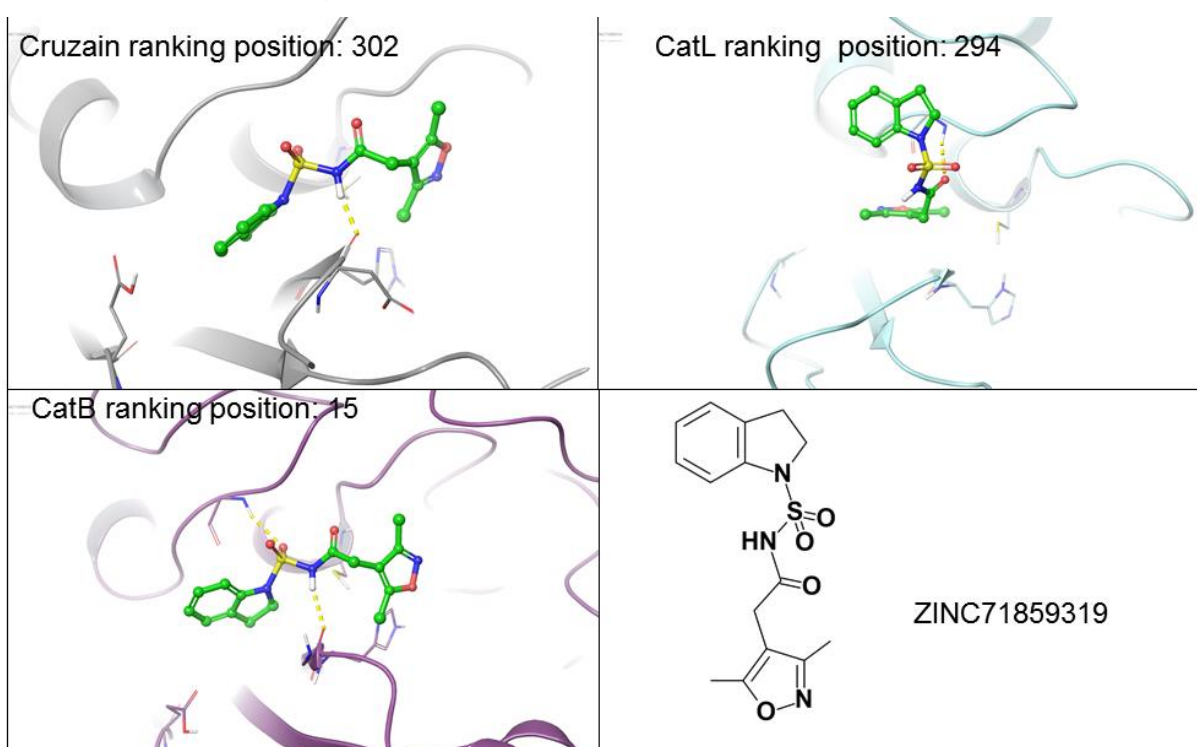


Figure 26 - Predicted poses and chemical structure of ZINC71859319, a supposed selective cruzain inhibitor over CatL.

ZINC81063926 is expected to be a selective cruzain inhibitor over both human cathepsins. The predicted pose in cruzain is very well fitted in the active site. Moreover, it is stabilized by hydrogen bonds, in a fashion that most polar groups are involved in these interactions. In CatL and B the ring with three polar atoms is positioned in S2, a very hydrophobic pocket, without any hydrogen bond with these atoms to stabilize this binding mode. Therefore, this is a poorly scoring pose and probably a weak CatL and B binder (Figure 27).

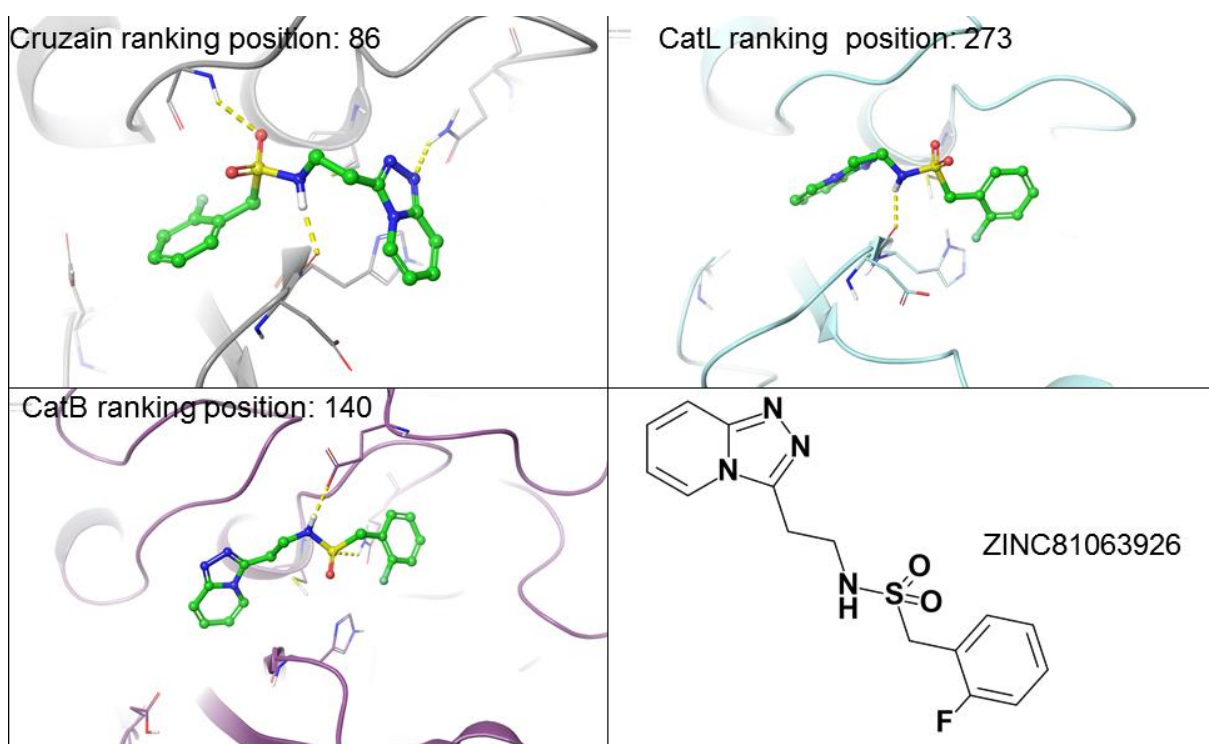


Figure 27 - Predicted poses and chemical structure of ZINC81063926, a supposed selective cruzain inhibitor over CatL and CatB.

ZINC83820332 was chosen based on its poor predicted pose against CatB (Figure 28). Its CatB pose is positioned from S1 pocket onwards, a very shallow region in this enzyme, with many polar groups exposed to solvent. In cruzain, the molecule fits the S2 pocket and make a hydrogen bond with GLN 19 stabilizing the other ring. In CatL, the S2 pocket is also occupied, and the other ring is stabilized by a hydrogen bond with the backbone of ASP 162.

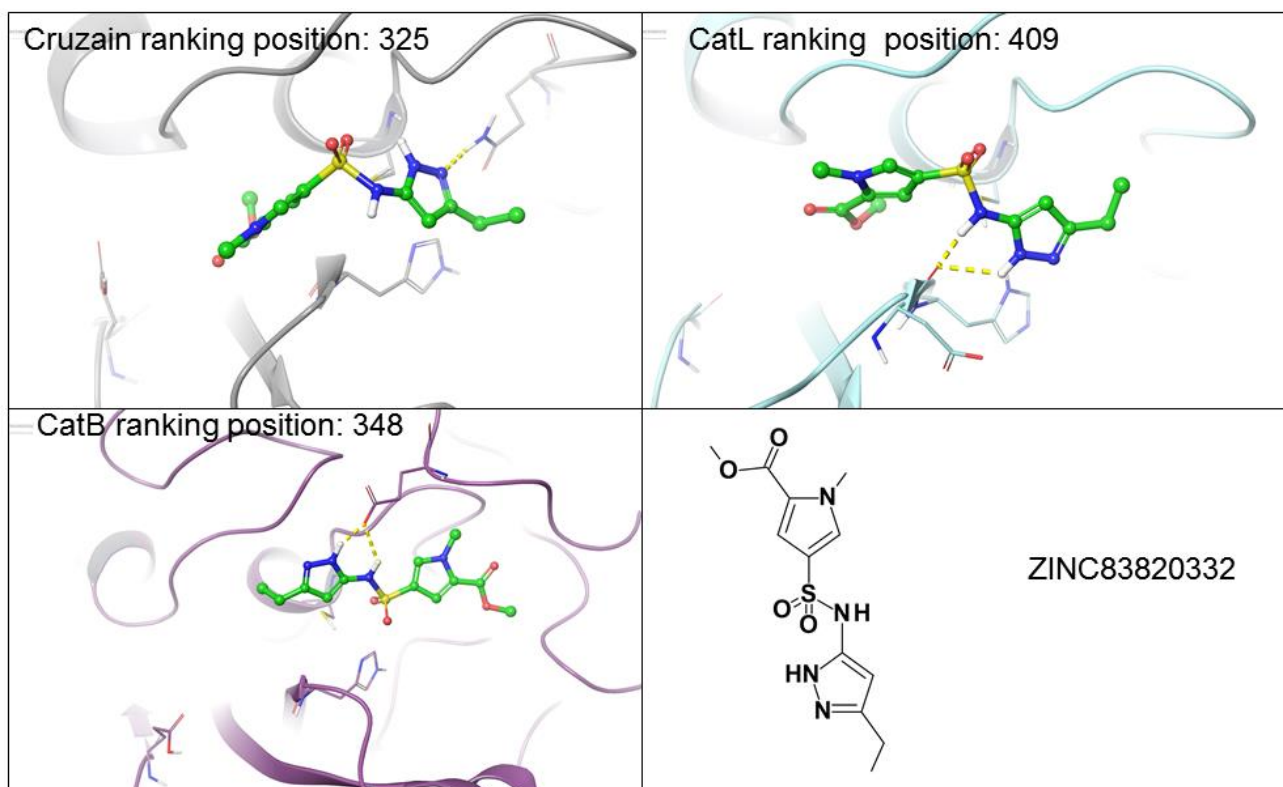


Figure 28 - Predicted poses and chemical structure of ZINC83820332, a supposed selective cruzain inhibitor over CatB.

6. CONCLUSIONS

Proposing a robust virtual screen workflow to find a cruzain selective inhibitor is a challenge, as its human homologous have a very similar active site. A structure-based VS was proposed, and two methods were tested: structure-based pharmacophores and molecular docking.

The objective in working with pharmacophores was to use this as a pre-filter in a hierarchical virtual screening to find new scaffolds of cruzain inhibitors. However, we were not able to generate a general pharmacophoric hypothesis; our best model was able to recognize only benzimidazole derivatives. Despite this result, it is important to keep in mind that in the future, with more ligands being described, a better model might be achieved. Also, another option could be the generation of many pharmacophoric models, one for each chemical class characterized in the literature so far, and evaluate if any new scaffolds can be found with this approach.

Docking was the selected method to the VS screening as it could better discriminate between ligands and decoys and enrich the top of screens with a high chemical diversity of ligands. The evaluation done revealed that using grids with different protonation states might increase the chemical diversity of retrieved compounds in the top of docking rank. Two grids were used in this VS, of one with CYS 25, GLU 208 and HIS 162 neutral and other with these residues charged.

After the VS, eight cruzain hits were proposed. Two of them (ZINC55314924 and ZINC81063926) are supposed to be selective over both CatL and CatB, and the others are selective over one of them. Next steps are to make the *in vitro* assay against these enzymes and, if obtaining a selective inhibitor, work to enhance its activity. Even if we are not able to find a selective inhibitor among this compounds, this should not be a huge problem, since the knowledge about the structure of these enzymes can be useful to propose modification the inhibitors to try to achieve selectivity.

7. REFERENCES

- Albajar-Vinas, P. and Jannin, J. (2011) 'The hidden Chagas disease burden in Europe', *Euro Surveill*, 16(38), p. pii=19975. Available at: <http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19975%0AArticle>.
- Aparicio, I. M., Scharfstein, J. and Lima, A. P. C. A. (2004) 'A new cruzipain-mediated pathway of human cell invasion by *Trypanosoma cruzi* requires trypomastigote membranes.', *Infection and immunity*, 72(10), pp. 5892–902. doi: 10.1128/IAI.72.10.5892-5902.2004.
- Bermudez, J. *et al.* (2016) 'Current drug therapy and pharmaceutical challenges for Chagas disease', *Acta Tropica*. Elsevier B.V., 156, pp. 1–16. doi: 10.1016/j.actatropica.2015.12.017.
- Bern, C. *et al.* (2011) 'Trypanosoma cruzi and chagas' disease in the united states', *Clinical Microbiology Reviews*, 24(4), pp. 655–681. doi: 10.1128/CMR.00005-11.
- Bontempi, E. and Cazzulo, J. J. (1990) 'Digestion of human immunoglobulin G by the major cysteine proteinase (cruzipain) from *Trypanosoma cruzi*', *FEMS Microbiology Letters*, 70(3), pp. 337–341. doi: 10.1016/S0378-1097(05)80019-9.
- Braga, R. C. and Andrade, C. H. (2013) 'Assessing the performance of 3D pharmacophore models in virtual screening: how good are they?', *Current topics in medicinal chemistry*, 13(9), pp. 1127–38. doi: 10.2174/1568026611313090010.
- Britto, C. *et al.* (2001) 'Parasite Persistence in Treated Chagasic Patients Revealed by Xenodiagnosis and Polymerase Chain Reaction', *Memorias do Instituto Oswaldo Cruz*, 96(6), pp. 823–826. doi: 10.1590/S0074-02762001000600014.
- Carmona, E. *et al.* (1996) 'Potency and selectivity of the cathepsin L propeptide as an inhibitor of cysteine proteases', *Biochemistry*, 35(25), pp. 8149–8157. doi: 10.1021/bi952736s.
- Castro, H. C. *et al.* (2011) 'Looking at the proteases from a simple perspective', *Journal of Molecular Recognition*, 24(2), pp. 165–181. doi: 10.1002/jmr.1091.
- Cazzulo, J.-J., Stoka, V. and Turk, V. (2001) 'The major cysteine proteinase of *Trypanosoma cruzi*: a valid target for chemotherapy of Chagas disease.', *Current pharmaceutical design*, 7(12), pp. 1143–1156. doi: 10.2174/1381612013397528.
- Cazzulo, J. J. *et al.* (1990) 'Some kinetic properties of a cysteine proteinase (cruzipain) from *Trypanosoma cruzi*', *Biochimica et Biophysica Acta - Protein Structure and*

- Molecular Enzymology*, 1037(2), pp. 186–191. doi: 10.1016/0167-4838(90)90166-D.
- Cereto-massagué, A. *et al.* (2015) 'Molecular fingerprint similarity search in virtual screening', 71, pp. 58–63. doi: 10.1016/j.ymeth.2014.08.005.
- Chowdhury, S. F. *et al.* (2002) 'Design of noncovalent inhibitors of human cathepsin L. From the 96-residue proregion to optimized tripeptides.', *Journal of medicinal chemistry*, 45(24), pp. 5321–9. doi: 10.1021/jm020238t.
- Chowdhury, S. F. *et al.* (2008) 'Exploring inhibitor binding at the S' subsites of cathepsin L.', *Journal of medicinal chemistry*, 51(5), pp. 1361–1368. doi: 10.1021/jm701190v.
- Conners, E. E. *et al.* (2016) 'A global systematic review of Chagas disease prevalence among migrants', *Acta Tropica*. Elsevier B.V., 156, pp. 68–78. doi: 10.1016/j.actatropica.2016.01.002.
- DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. (1988) 'Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach Author (s): Elizabeth R . DeLong , David M . DeLong and Daniel L . Clarke-Pearson Published by: International Biometric Society Stable', *Biometrics*, 44(3), pp. 837–845.
- Doyle, P. S. *et al.* (2011) 'The trypanosoma cruzi protease cruzain mediates immune evasion', *PLoS Pathogens*, 7(9), pp. 1–11. doi: 10.1371/journal.ppat.1002139.
- Du, X. *et al.* (2000) 'Aryl ureas represent a new class of anti-trypanosomal agents', *Chemistry & Biology*, 7(9), pp. 733–742. doi: 10.1016/S1074-5521(00)00018-1.
- Du, X. *et al.* (2002) 'Synthesis and structure-activity relationship study of potent trypanocidal thio semicarbazone inhibitors of the trypanosomal cysteine protease cruzain', *Journal of Medicinal Chemistry*, 45(13), pp. 2695–2707. doi: 10.1021/jm010459j.
- Emsley, P. *et al.* (2010) 'Features and development of Coot', *Acta Crystallographica Section D: Biological Crystallography*. International Union of Crystallography, 66(4), pp. 486–501. doi: 10.1107/S0907444910007493.
- Engel, J. C. *et al.* (1998) 'Cysteine protease inhibitors cure an experimental Trypanosoma cruzi infection.', *The Journal of experimental medicine*, 188(4), pp. 725–34. doi: 10.1084/jem.188.4.725.
- Ferreira, L. *et al.* (2015) *Molecular Docking and Structure-Based Drug Design Strategies*, *Molecules*. doi: 10.3390/molecules200713384.
- Ferreira, R. S. *et al.* (2009) 'Divergent modes of enzyme inhibition in a homologous structure-activity series', *Journal of Medicinal Chemistry*, 52(16), pp. 5005–5008. doi:

10.1021/jm9009229.

Ferreira, R. S. *et al.* (2010) 'Complementarity between a docking and a high-throughput screen in discovering new cruzain inhibitors', *Journal of Medicinal Chemistry*, 53(13), pp. 4891–4905. doi: 10.1021/jm100488w.

Ferreira, R. S. *et al.* (2014) 'Synthesis, biological evaluation, and structure-activity relationships of potent noncovalent and nonpeptidic cruzain inhibitors as anti-Trypanosoma cruzi agents', *Journal of Medicinal Chemistry*, 57(6), pp. 2380–2392. doi: 10.1021/jm401709b.

Franke De Cazzulo, B. M. *et al.* (1994) 'Effects of proteinase inhibitors on the growth and differentiation of Trypanosoma cruzi', *FEMS Microbiology Letters*, 124, pp. 81–86. doi: 10.1111/j.1574-6968.1994.tb07265.x.

Friesner, R. a. *et al.* (2004) 'Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy', *Journal of Medicinal Chemistry*, 47(7), pp. 1739–1749. doi: 10.1021/jm0306430.

Friesner, R. a. *et al.* (2006) 'Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes', *Journal of Medicinal Chemistry*, 49(21), pp. 6177–6196. doi: 10.1021/jm051256o.

Gauthier, J. Y. *et al.* (2008) 'The discovery of odanacatib (MK-0822), a selective inhibitor of cathepsin K', *Bioorganic and Medicinal Chemistry Letters*, 18(3), pp. 923–928. doi: 10.1016/j.bmcl.2007.12.047.

Gilson, M. K. *et al.* (2015) 'BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology.', *Nucleic acids research*, 44(October 2015), p. gkv1072. doi: 10.1093/nar/gkv1072.

Guariento, M. E., Camilo, M. V. F. and Camargo, A. M. A. (1999) 'Working conditions of Chagas' disease patients in a large Brazilian city Situação trabalhista do portador de doença de Chagas crônica, em um grande centro urbano', *Cadernos de Saúde Pública*, 15(2), pp. 381–386. Available at: http://www.scielo.br/scielo.php?script=sci%7B_%7Darttext%7B&%7Dpid=S0102-311X1999000200022%7B&%7Dlang=pt.

Güner, O. F. (2002) 'History and Evolution of the Pharmacophore Concept in Computer-Aided Drug Design History and Evolution of the Pharmacophore Concept in Computer-Aided', *Current Topics in Medicinal Chemistry*, 2, pp. 1321–1332. doi: 10.2174/1568026023392940.

Huang, L., Brinen, L. S. and Ellman, J. a. (2003) 'Crystal structures of reversible

ketone-Based inhibitors of the cysteine protease cruzain', *Bioorganic and Medicinal Chemistry*, 11(1), pp. 21–29. doi: 10.1016/S0968-0896(02)00427-3.

Irwin, J. J. *et al.* (2012) 'ZINC: A Free Tool to Discover Chemistry for Biology', *J. Chem. Inf. Model*, 52, pp. 1757–1768. doi: dx.doi.org/10.1021/ci3001277.

Irwin, J. J. and Shoichet, B. K. (2005) 'for Virtual Screening', *J Chem Inf Model*, 45(1), pp. 177–182.

Jadhav, A. *et al.* (2010) 'Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease', *Journal of Medicinal Chemistry*, 53(1), pp. 37–51. doi: 10.1021/jm901070c.

Kar, S. and Roy, K. (2013) 'How far can virtual screening take us in drug discovery?', pp. 245–261.

Keillor, W. and Brown, R. (1992) 'Simple Demonstration of a Bell-Shaped pH / Rate Profile', *J. Am. Chem. Soc.*, 114, pp. 7983–7989.

Kuhn, B. *et al.* (2016) 'A Real-World Perspective on Molecular Design', *Journal of Medicinal Chemistry*, p. acs.jmedchem.5b01875. doi: 10.1021/acs.jmedchem.5b01875.

Li, Y. *et al.* (2014) 'Comparative assessment of scoring functions on an updated benchmark: 2. evaluation methods and general results', *Journal of Chemical Information and Modeling*, 54(6), pp. 1717–1736. doi: 10.1021/ci500081m.

Markt, P., Schuster, D. and Langer, T. (2011) 'Pharmacophore Models for Virtual Screening', in Sotriffer, C. (ed.) *Virtual Screening: Principles, Challenges, and Practical Guidelines*. Weinheim, Germany: WILEY-VCH Verlag GmbH & Co. KGaA, pp. 115–152. doi: 10.1002/9783527633326.ch5.

Marques, E. F. *et al.* (2012) 'Evaluation of synthetic acridones and 4-quinolinones as potent inhibitors of cathepsins L and V.', *European journal of medicinal chemistry*, 54, pp. 10–21. doi: 10.1016/j.ejmech.2012.04.002.

Marquis, R. W. *et al.* (2005) 'Azepanone-based inhibitors of human cathepsin L.', *Journal of medicinal chemistry*, 48(22), pp. 6870–8. doi: 10.1021/jm0502079.

Martinez-Mayorga, K. *et al.* (2015) 'Cruzain inhibitors: efforts made, current leads and a structural outlook of new hits', *Drug Discovery Today*. Elsevier Ltd, 0(0), pp. 1–9. doi: 10.1016/j.drudis.2015.02.004.

Martínez, J. *et al.* (1993) 'The reactivity of sera from chagasic patients against different fragments of cruzipain, the major cysteine proteinase from *Trypanosoma cruzi*, suggests the presence of defined antigenic and catalytic domains', *Immunology*

Letters, 35(2), pp. 191–196. doi: 10.1016/0165-2478(93)90090-O.

McGrath, M. E. *et al.* (1995) 'The crystal structure of cruzain: a therapeutic target for Chagas' disease.', *Journal of molecular biology*, 247(2), pp. 251–259. doi: 10.1006/jmbi.1994.0137.

Ministério da Saúde (2010) *Formulário terapêutico nacional 2010: Rename 2010*. 2ª edição, *Série B. Textos Básicos de Saúde*. 2ª edição. Brasília: Ministério da Saúde. Available at: http://bvsms.saude.gov.br/bvs/publicacoes/formulario_terapeutico_nacional_2010.pdf

Morillo, C. A. *et al.* (2015) 'Randomized Trial of Benznidazole for Chronic Chagas' Cardiomyopathy', *New England Journal of Medicine*, 373(14), pp. 1295–1306. doi: 10.1056/NEJMoa1507574.

Mysinger, M. M. *et al.* (2012) 'Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking', *Journal of Medicinal Chemistry*, 55(14), pp. 6582–6594. doi: 10.1021/jm300687e.

Nägler, D. K. *et al.* (1997) 'Major Increase in Endopeptidase Activity of Human Cathepsin B upon Removal of Occluding Loop Contacts †', *Biochemistry*, 36, pp. 12608–12615.

Rassi, A. and Marcondes de Rezende, J. (2012) *American trypanosomiasis (Chagas disease)*., *Infectious disease clinics of North America*. doi: 10.1016/j.idc.2012.03.002.

Reyes, P. a and Vallejo, M. (2011) 'Trypanocidal drugs for late stage, symptomatic Chagas disease (*Trypanosoma cruzi* infection).', *Cochrane database of systematic reviews (Online)*, (4), p. CD004102. doi: 10.1002/14651858.CD004102.pub2.

Robin, X. *et al.* (2011) 'pROC : an open-source package for R and S + to analyze and compare ROC curves', *BMC Bioinformatics*, 12(77). Available at: <http://www.biomedcentral.com/1471-2105/12/77>.

Rogers, D. and Hahn, M. (2010) 'Extended-connectivity fingerprints', *J Chem Inf Model.*, 50(5), pp. 742–754. doi: 10.1021/ci100050t.

Rogers, K. E. *et al.* (2012) 'Novel Cruzain Inhibitors for the Treatment of Chagas ' Disease', (20), pp. 398–405. doi: 10.1111/j.1747-0285.2012.01416.x.

Sant'Anna, C. *et al.* (2008) 'All *Trypanosoma cruzi* developmental forms present lysosome-related organelles', *Histochemistry and Cell Biology*, 130(6), pp. 1187–1198. doi: 10.1007/s00418-008-0486-8.

Scharfstein, J. *et al.* (2000) 'Host cell invasion by *Trypanosoma cruzi* is potentiated by

activation of bradykinin B(2) receptors.', *The Journal of experimental medicine*, 192(9), pp. 1289–1300. doi: 10.1084/jem.192.9.1289.

Schechter, I. and Berger, A. (1967) 'On the size of the active site in proteases I. Papain', *Biochemical and Biophysical Research Communications*, 27(3), pp. 157–162. doi: 10.1016/0006-291X(67)90299-9.

Shah, F. *et al.* (2011) 'Identification of novel malarial cysteine protease inhibitors using structure-based virtual screening of a focused cysteine protease inhibitor library', *J Chem Inf Model*, 51(4), pp. 852–864. doi: 10.1021/ci200029y.

Siles, R. *et al.* (2006) 'Design, synthesis, and biochemical evaluation of novel cruzain inhibitors with potential application in the treatment of Chagas' disease.', *Bioorganic & medicinal chemistry letters*, 16(16), pp. 4405–4409. doi: 10.1016/j.bmcl.2006.05.041.

Tao, K. *et al.* (1994) 'The proregion of cathepsin L is required for proper folding, stability, and ER exit.', *Archives of Biochemistry and Biophysics*, 311(1), pp. 19–27. doi: 10.1006/abbi.1994.1203.

Villar, J. C. *et al.* (2014) 'Trypanocidal drugs for chronic asymptomatic Trypanosoma cruzi infection (Review) Trypanocidal drugs for chronic asymptomatic Trypanosoma cruzi infection (Review) Trypanocidal drugs for chronic asymptomatic Trypanosoma cruzi infection', *Cochrane Database of Systematic Reviews*, (5). doi: 10.1002/14651858.CD003463.pub2.

Vuorinen, A. and Schuster, D. (2015) 'Methods for generating and applying pharmacophore models as virtual screening filters and for bioactivity profiling', *Methods*. Elsevier Inc., 71, pp. 113–134. doi: 10.1016/j.ymeth.2014.10.013.

Wang, Y. *et al.* (2017) 'PubChem BioAssay: 2017 update', *Nucleic Acids Research*, 45(D1), pp. D955–D963. doi: 10.1093/nar/gkw1118.

WHO (2015) 'Chagas disease in Latin America: an epidemiological update based on 2010 estimates', *Weekly Epidemiological Record*, (6), pp. 33–44.

Wiggers, H. J. *et al.* (2013) 'Non-peptidic Cruzain Inhibitors with Trypanocidal Activity Discovered by Virtual Screening and In Vitro Assay', 7(8). doi: 10.1371/journal.pntd.0002370.

Wolber, G., Dornhofer, A. A. and Langer, T. (2006) 'Efficient overlay of small organic molecules using 3D pharmacophores', *J Comput Aided Mol Des*, 20, pp. 773–788. doi: 10.1007/s10822-006-9078-7.

Wolber, G. and Kosara, R. (2006) 'Pharmacophores from Macromolecular Complexes with LigandScout', in Langer, T. and Hoffmann, R. D. (eds) *Pharmacophores and*

Pharmacophore Searches. Weinheim: WILEY-VCH Verlag GmbH & Co. KGaA, pp. 131–150.

Wolber, G. and Langer, T. (2005) 'LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters', *Journal of Chemical Information and Modeling*, 45(1), pp. 160–169. doi: 10.1021/ci049885e.

Yamashita, D. S. *et al.* (1999) 'Solid-Phase Synthesis of a Combinatorial Array of 1,3-Bis (acylamino) -2-butanones , Inhibitors of the Cysteine Proteases Cathepsins K and L', (Scheme 1), pp. 207–215.

Yokoyama-Yasunaka, J. K. U. *et al.* (1994) 'Trypanosoma cruzi: Identification of proteinases in shed components of trypomastigote forms', *Acta Tropica*, 57(4), pp. 307–315. doi: 10.1016/0001-706X(94)90076-0.

Zanatta, N. *et al.* (2008) 'Convergent synthesis and cruzain inhibitory activity of novel 2-(N'-benzylidenehydrazino)-4-trifluoromethyl-pyrimidines', *Bioorganic and Medicinal Chemistry*. Elsevier Ltd, 16(24), pp. 10236–10243. doi: 10.1016/j.bmc.2008.10.052.

8. APPENDIX

Appendix 1 - CatL known inhibitors employed for docking evaluation.

ZINC / ChEMBL ID	Reference	IC₅₀ (μM)	K_i (μM)
ZINC96142574	Yamashita et al., 1999		>1
ZINC96142576	Yamashita et al., 1999		>1
ZINC96142578	Yamashita et al., 1999		>1
ZINC96142580	Yamashita et al., 1999		0.23
ZINC96142582	Yamashita et al., 1999		0.12
ZINC96142584	Yamashita et al., 1999		>1
ZINC96142586	Yamashita et al., 1999		0.15
ZINC96142588	Yamashita et al., 1999		0.053
ZINC96142590	Yamashita et al., 1999		>1
ZINC96142592	Yamashita et al., 1999		0.13
ZINC96142594	Yamashita et al., 1999		0.036
ZINC96142596	Yamashita et al., 1999		>1
ZINC96142598	Yamashita et al., 1999		0.13
ZINC96142600	Yamashita et al., 1999		0.11
ZINC96142602	Yamashita et al., 1999		>1
ZINC96142604	Yamashita et al., 1999		0.09

ZINC / ChEMBL ID	Reference	IC50 (μM)	Ki (μM)
ZINC96142606	Yamashita et al., 1999		0.018
ZINC96142608	Yamashita et al., 1999		>1
ChEMBL148544	Chowdhury et al., 2002		0.490
ChEMBL148943	Chowdhury et al., 2002		0.210
ChEMBL149738	Chowdhury et al., 2002		6.6
ChEMBL341609	Chowdhury et al., 2002		0.021
ChEMBL415539	Chowdhury et al., 2002		0.019
ZINC27562733	Chowdhury et al., 2002		0.930
ZINC27563201	Chowdhury et al., 2002		>100
ZINC27563212	Chowdhury et al., 2002		0.24
ZINC27624073	Chowdhury et al., 2002		0.21
ZINC39290220	Chowdhury et al., 2002		0.27
ZINC39290221	Chowdhury et al., 2002		0.045
ZINC39290222	Chowdhury et al., 2002		0.016
ZINC39290223	Chowdhury et al., 2002		3.9
ZINC95542088	Chowdhury et al., 2002		0.27
ZINC95542462	Chowdhury et al., 2002		74
ZINC95588063	Chowdhury et al., 2002		0.067
ZINC95588077	Chowdhury et al., 2002		38
ZINC95612041	Chowdhury et al., 2002		0.16
ZINC95612249	Chowdhury et al., 2002		0.2

ZINC / ChEMBL ID	Reference	IC50 (μ M)	Ki (μ M)
ZINC95614747	Chowdhury et al., 2002		0.17
ChEMBL271747	Marquis et al., 2005		0.008
ChEMBL270269	Marquis et al., 2005		0.009
ChEMBL407515	Marquis et al., 2005		0.001
ZINC29135725	Marquis et al., 2005		0.0004
ZINC29135841	Marquis et al., 2005		0.0005
ZINC29136045	Marquis et al., 2005		0.002
ZINC29136295	Marquis et al., 2005		0.005
ZINC58638412	Marquis et al., 2005		0.0006
ZINC85549974	Marquis et al., 2005		0.0002
ZINC95540444	Chowdhury et al., 2008		0.112
ZINC96308785	Chowdhury et al., 2008		0.023
ZINC96308786	Chowdhury et al., 2008		0.511
ZINC96308787	Chowdhury et al., 2008		10.7
ZINC04430352	Chowdhury et al., 2008		6
ZINC04579248	Chowdhury et al., 2008		0.021
ZINC84651994	Chowdhury et al., 2008		0.045
ZINC84652016	Chowdhury et al., 2008		0.019
ZINC84652064	Chowdhury et al., 2008		0.46
ZINC84652767	Chowdhury et al., 2008		0.464
ZINC84672029	Chowdhury et al., 2008		0.155

ZINC / ChEMBL ID	Reference	IC50 (μM)	Ki (μM)
ZINC84712659	Chowdhury et al., 2008		0.024
ChEMBL196023	Chowdhury et al., 2008		57.5
ChEMBL197958	Marques et al., 2012	0.007	
ChEMBL382286	Marques et al., 2012	0.002	
ChEMBL194643	Marques et al., 2012	1.1	
ZINC03813507	Marques et al., 2012	19.3	
ZINC03934226	Marques et al., 2012	13	
ZINC13731099	Marques et al., 2012	1.5	
ZINC27212266	Marques et al., 2012	0.5	
ZINC27524329	Marques et al., 2012	3.9	