# Cerca d'inhibidors amb ML

Carla Feliu[a]

[a] *University of Vic - Central University of Catalonia, , Vic, , ,*

**Abstract**

Example abstract for the astronomy and computing journal. Here you provide a brief summary of the research and the results. RESUM DE TOT

*Keywords:* keyword 1, keyword 2, keyword 3, keyword 4

## 1. Introduction

Proteases, also known as peptid hydrolases, are enzymes capable of catalysing hydrolytic reactions that degrade protein molecules to peptides, and finally to free amino acids. They also carry out proteolytic reactions, and regulate various enzymatic cascades that form part of metabolic cycles. [1] The role they play is key in carrying out vital biological processes, such as the regulation of various cellular processes as well as differentiation, gene expression and cell death. Protests are a very broad group of enzymes that are differentiated by, among other things, the type of amino acid residue in their catalytic site. [1]

Cystein proteases (also known as thiol proteases), are one of the types of proteases that exist, and are characterised by catalysing the breakdown of proteins by cleavage of peptide bonds using a cysteine nucleophile tilo. These proteases are classified into two types according to their location within the organism; Catepsins, located in the lysosome, and Calpains, located in the cytosol. [2]

Although these proteases have a key role in vital processes, when there is overexposure or dysregulation in pathological conditions, they can contribute to the development of many diseases. In several studies, it has been shown that cathepsins, located in the lysosome, are related to tumour progression. [3] [4]. These cystein proteases allow cancer cells to attack nearby tissues, blood and lymphatic vessels and metastasise to peripheral tissues. [2] [5] For this reason, the search for inhibitors of cystein proteases has become an objective in medical and pharmacological research, in order to develop new therapeutic opportunities. [6]

The interest of cystein proteases as targets is related to the fact that by working with their inhibition, it is possible to find therapies for many of the diseases they are involved in. They are also interesting targets due to the reactive cisterna active site, as the high nucleophilicity of the cisterna thiol, under physiological conditions, provides an ideal anchoring site for small electrophilic molecules. [7]

The inhibition process is defined by the effectiveness of the binding of a protein to the selected targets. This depends on the reaction equilibrium between the functional groups of the drug and the active site residues of the protein. This balance is determined by the type of interaction that is created, and can be either covalent or non-covalent interactions. Covalent bonds are usually formed by the interaction of certain amino acids, including the nucleophilic cisterna and a reactive functional group of the ligand. They are more effective due to their long permanence in the active site, high binding affinity and high selectivity and specificity, resulting in complete inhibition of the target compared to the reversible effect of non-covalent drugs. [8]

The drug development process is a very elaborate process, which can involve a lot of time and resources. It begins with the search for and characterisation of a possible biological target for a specific disease or enzyme, and then the process of creating the most suitable therapeutic compound begins. Before reaching drug development, many molecular properties must be optimised to identify and validate the target, as this is one of the most important steps in drug development. [9] Historically, this search was guided by the calculations and intuition of expert scientists, but with the great advances in new technologies and the increase in resources, it has been possible to apply new methods that are much faster and more reliable. [10]

One of the most used methods currently for this drug development, and which has allowed a great advance in this field of medical research, making the process less expensive and more efficient, is machine learning (ML). ML is a branch of artificial intelligence (AI) that applies computer algorithms with the ability to learn themselves from dirty, raw data, to then perform a specific task. Within ML methods, the most widely used and expanding is deep learning (DL). It is inspired by the information processing patterns of the human brain and is designed using multiple layers of algorithms (artificial neural networks, ANNs), each of which makes an interpretation of the data it has received. [11]

Artificial neural networks, inspired by the human brain, are created to simulate a network of model neurons on a computer. By applying algorithms that mimic the processes of real neurons, models are generated that learn to solve many types of problems. The networks are composed of artificial neuron units, capable of transmitting signals to each other.(what are artificial networks ). Each reception of a signal in the neurons is associated with a weight that varies as learning progresses, which can increase or decrease the strength of the signal that will then be sent to the other neurons. These signals pass the multilayers of artificial neurons until they reach the final layer, where the result of these connections is the prediction of the algorithm. [12]

The chemprop Architecture is a neural network built for the prediction of molecular properties that includes a directed message passing neural network (D-MPNN) module for molecular feature extraction and a feed-forward neural network (FNN) for property prediction. This model takes molecular SMILES as input and extracts all the atomic and bonding characteristics of the molecules to generate a single vector of characteristics which, by entering it into the FNN, can predict the activity of the candidate molecules. [13]

Given the growing interest in cysteine proteases as targets for therapeutic treatments and the availabikity of new machine learning tools for these drug discovery processes, three questions are raised in this work.

- Is it possible to identify covalent targets that can serve as inhibitors for cysteine proteases using ANN?
- What is the best way to train a model using ANN?
- Is it possible to identify molecular patterns or specific molecules that can help us in finding cysteine protease inhibitors in two completely different databases?

We will use a ML model, ANN, applying a known dataset of compounds that interact with cysteine proteases. This dataset will be used to train our model, allowing us to make predictions from two different databases, that lead us to solve the proposed questions.

Specific objectives:

1. Identify a database with known targets of cysteine proteases that includes inhibitory activity, which will serve as the training data of our ML model.
2. Download the database and extract the (Simplified Molecular Input Line Entry Sistem)SMILES information. Prepare the required dataframe structure for working with the ML algorithm, obtaining various features for each molecule and its inhibitory activity (IC50).
3. Train the chemprop model using various approaches, evaluating which one makes the model more stable for our predictions.
4. Identify suitable databases to obtain the molecules for our predictions. Similar to the training dataframe, construct the structure necessary to make the predictions with our training model.
5. Analyze the data generated from the predictions made by the ML model, checking the correlation of the data and their similarity to understand if the predictions obtained are suitable for our questions raised.
6. Detect patterns of molecules that could be optimal for working in the discovery of cystein protease inhibitors.

## 2. Methods

### *2.1. Datasets*

Per realitzar aquest estudi, hem treballat en tres bases de dades diferents per extreure la informació necessària per realitzar el nostre estudi. L'extracció de les dades s'ha fet depenent de cada base de dades i es detallara mes endavant.

### *2.1.1. ChEMBL*

Chembl es una base de dades de petites molècules bioactives similars a fàrmacs, que contenen propietats calculades i bioactivitats. Les dades són extretes de literatura científica i participen en el descobriment de fàrmacs. [14]

El nostre interès en aquesta base de dades es tracta d'obtenir informació de molècules que puguin interaccionar amb cystein proteases per tal d'obtenir un punt de referència sobre l'activitat d'interacció que ens permeti entrenar el nostre model de ML. Per accedir a la informació que ens interessa per el nostre estudi, hem filtrat dins la base de dades de ChEMBL amb el terme "cystein proteasa" i hem extret els 129 target ID resultants. Un cop obtinguts els targets de cystein proteasa, amb el mòdul requests de python hem obtingut informació de totes les molècules que tenen interacció amb cada un dels lligands, junt amb la seva activitat d'interacció, el valor IC50, que hem utilitzat per poder entrenar el model ML.

### *2.1.2. ChemDiv*

ChemDiv és una organització d'investigació reconeguda mundialment en solucions de descobriment de fàrmacs. Tenen identificades diverses llibreries específiques dirigides a abordar diverses dianes, dominis proteics, processos cel·lulars... [15]

D'aquesta base de dades hem obtingut la llibreria "Cysteine Targeted Covalent Library", amb 39,301 compostos que tenen warheads específics per reaccionar amb cisteïna. L'objectiu d'utilitzar aquesta llibreria és, aplicant el nostre model entrenat, trobar possibles compostos que tinguin una activitat d'interacció significativa amb la cisteïna per trobar un patró que ens pugui servir per entendre quin tipu de molecula seria un bon inhibidor.

### *2.1.3. Zinc*

ZINC

| Train | Initial dim. | Data curated | Norm. |
|---|---|---|---|
| Train 1 | 19,262 | 19,124 | No |
| Train 3 | 19,262 | 11700 | Z-Score |
| Train 4 | 19,262 | 11700 | MinMaxScale |

Table 1: Train results

### *2.2. Data preparation and curation*

Per dur a terme el procés d'entrenament de models de Machine learning de manera efectiva, és molt important tenir un conjunt de dades adequat i degudament preparat. Per tant, es requereix d'una preparació prèvia. Aquesta estructura de dades ha de tenir una primera columna que contingui els SMILES de les molècules, features que descriguin les molècules, i finalment, en una ultima fila l'activitat d'interacció d'aquestes molècules, el valor IC50. La preparació de dades inclou també la realització d'un anàlisis exhaustiu de l'estructura de les dades, la qual cosa resulta fonamental per entendre a fons les dades, i així dur a terme un anàlisis de resultats més precís.

### *2.2.1. Train*

## 3. Results

RESULTATS DE LA CARLA

### *3.1. Subsection title*

## 4. Discussion

## 5. Conclusions

## Acknowledgements

Thanks to ...

## Appendix A. Appendix title 1

## Appendix B. Appendix title 2

## Appendix C. Bibliography

## References

[1] Oscar L. Ramos and F. Xavier Malcata. *Food-grade enzymes*. Elsevier, 1 2019. ISBN 9780444640475. doi: 10.1016/B978-0-12-809633-8.09173-1.

[2] Satya P. Gupta and Sayan Dutta Gupta. *Cancer-leading proteases: An introduction*. Elsevier, 1 2020. ISBN 9780128181683. doi: 10.1016/B978-0-12-818168-3.00001-2.

[3] Izabela Berdowska. Cysteine proteases as disease markers, 4 2004. ISSN 00098981.

[4] Mona Mostafa Mohamed and Bonnie F. Sloane. Cysteine cathepsins: Multifunctional enzymes in cancer. volume 6, pages 764–775, 10 2006. doi: 10.1038/nrc1949.

[5] Vasilena Gocheva, Wei Zeng, Danxia Ke, David Klimstra, Thomas Reinheckel, Christoph Peters, Douglas Hanahan, and Johanna A. Joyce. Distinct roles for cysteine cathepsin genes in multistage tumorigenesis. *Genes and Development*, 20:543–556, 3 2006. ISSN 08909369. doi: 10.1101/gad.1407406.

[6] Johanna A Joyce, Amos Baruch, Kareem Chehade, Nicole Meyer-Morse, Enrico Giraudo, Fong-Ying Tsai, Doron C Greenbaum, Jeffrey H Hager, Matthew Bogyo, and Douglas Hanahan. Cathepsin cysteine proteases are effectors of invasive growth and angiogenesis during multistage tumorigenesis.

[7] Aaron J. Maurais and Eranthie Weerapana. Reactive-cysteine profiling for drug discovery, 6 2019. ISSN 18790402.

[8] Aimen Aljoundi, Imane Bjij, Ahmed El Rashedy, and Mahmoud E.S. Soliman. Covalent versus non-covalent enzyme inhibition: Which route should we take? a justification of the good and bad from molecular modelling perspective, 4 2020. ISSN 18758355.

[9] J. P. Hughes, S. S. Rees, S. B. Kalindjian, and K. L. Philpott. Principles of early drug discovery, 3 2011. ISSN 00071188.

[10] Patrick Hop, Brandon Allgood, and Jessen Yu. Geometric deep learning autonomously learns chemical features that outperform those engineered by domain experts. *Molecular Pharmaceutics*, 15: 4371–4377, 10 2018. ISSN 15438392. doi: 10.1021/acs.molpharmaceut.7b01144.

[11] Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8, 12 2021. ISSN 21961115. doi: 10.1186/s40537-021-00444-8.

[12] Alan Talevi, Juan Francisco Morales, Gregory Hather, Jagdeep T. Podichetty, Sarah Kim, Peter C. Bloomingdale, Samuel Kim, Jackson Burton, Joshua D. Brown, Almut G. Winterstein, Stephan Schmidt, Jensen Kael White, and Daniela J. Conrado. Machine learning in drug discovery and development part 1: A primer. *CPT: Pharmacometrics and Systems Pharmacology*, 9:129–142, 3 2020. ISSN 21638306. doi: 10.1002/psp4.12491.

[13] Liying Wang, Zhongtian Yu, Shiwei Wang, Zheng Guo, Qi Sun, and Luhua Lai. Discovery of novel sars-cov-2 3cl protease covalent inhibitors using deep learning-based screen. *European Journal of Medicinal Chemistry*, 244, 12 2022. ISSN 17683254. doi: 10.1016/j.ejmech.2022.114803.

[14] About - chembl interface documentation, . URL `https://chembl.gitbook.io/chembl-interface-documentati`

[15] Focused and targeted libraries - chemdiv inc., . URL `https://www.chemdiv.com/catalog/focused-and-targeted-l`