UAB

Universitat Autònoma de Barcelona

AI application for azophotoswitches' optimization with pharmacological interest

Author : Sergio Castañeiras Morales
Supervisor : Miquel Moreno Ferrer
Co-Supervisor : Àngels Gonzalez Lafont

Final Degree Project
Bachelor's Degree in Chemistry

2024-2025

"The dumbest people I know are those who know it all."

Malcolm S. Forbes

Abstract

We explore AI-based algorithms capabilities, specifically the Random Forest model, to predict a drug's inhibition potential for cyclooxygenase-2 (COX-2), a key protein linked to cancer. Using molecular descriptors extracted from the ChEMBL database, AI models are trained to identify patterns correlating with inhibition potential. The study validates AI's effectiveness in drug discovery and molecular analysis, proving its potential as a powerful tool in computational chemistry research.

Keywords: *COX-2, cyclooxygenase-2, Random Forest, Machine Learning, AI, Artificial Intelligence.*

I Introduction

The impact of Artificial Intelligence (AI) on science has been nothing but an outstanding breakthrough, with few comparable predecessors. The rapid advancements in AI have transformed numerous scientific fields[1,2] including computational chemistry. Nowadays, one of the main goals of computational chemistry is to predict certain properties of unstudied substances with minimal experimental costs.[3] Traditional approaches in chemistry often rely on complex laboratory techniques, which, while effective, can be time-consuming, expensive and resource-intensive. On the other hand, AI algorithms have already proved exceptional predictive capabilities in countless fields, and computational chemistry is no exception. AI provides an alternative by offering highly accurate predictions based on existing data, optimising research processes, and accelerating scientific discovery.

This project aims to implement artificial intelligence in computational chemistry, concretely, using AI-based algorithms to predict a drug's inhibition potential[4] for a given protein. To achieve this, we make use of the ChEMBL database,[5] a vast repository of bioactive molecules with drug-like properties. We extract all known molecular data with a documented inhibition potential for the target protein, creating a comprehensive dataset. The chemical descriptors of each molecule in the database are then computed using AlvaDesk[6,7] software. Around 10^4 descriptors are calculated,[8] which comprehend from the elemental molecular weight to the complex equipotential electronic surface, providing critical information about each compound's behaviour. The resulting dataset is subsequently used to train AI models, enabling them to predict the inhibition potential of unknown compounds. Finally, we evaluate the reliability of each model by testing it against real experimental data.

It is important to emphasise the central hypothesis of this project: *There exists a combination (or combinations) of chemical descriptors that are directly correlated with the inhibition of the protein.* While this idea may seem fundamental, it remains unproven due to the complexity of molecular interactions and the vast number of possible descriptor combinations. Despite significant progress in computational chemistry, identifying the exact descriptors that govern inhibition potential has been a persistent challenge. The lack of an ultimate proof underscores the need for advanced computational techniques. By analysing large datasets, AI can detect hidden correlations that may not be immediately apparent through traditional statistical methods.

At this stage, we focus on cyclooxygenase-2 (COX-2), a protein well known for its strong association with cancer development and inflammatory diseases.[9] COX-2 plays a crucial role in the biosynthesis of prostaglandins, which me-

diate inflammation and pain. Overexpression of COX-2 has been linked to various types of cancer, making it a prime target for drug development. COX-2 inhibitors, such as Celecoxib (Def. 2) and Rofecoxib (Def. 3), have been widely studied for their therapeutic potential. The scientific community has devoted an extensive research to COX-2, even before the rise of AI, due to its biomedical significance.[10] By applying AI models to COX-2, we assess their compatibility with the latest research findings, demonstrating AI's potential as a powerful tool in computational chemistry research. Our approach not only validates AI's effectiveness in predicting inhibition potential but also provides insights into the underlying molecular mechanisms governing COX-2 interactions.

The AI algorithm used in this study is the Random Forest algorithm,[11] a powerful ensemble learning method that generates multiple decision trees and combines their outputs to improve prediction accuracy. This approach is particularly well-suited for computational chemistry due to its ability to handle large datasets, manage complex relationships between variables, and reduce overfitting. The Random Forest algorithm operates by constructing numerous random decision trees, each trained on different subsets of the dataset. The final prediction is obtained by averaging the outputs of all trees, ensuring robust and reliable results.

Moreover, the choice of the Random Forest algorithm is motivated by the presence of decision trees in various chemistry-related fields. In spectroscopy, for instance, decision trees are used in group theory to classify molecular symmetry. Similarly, in analytical chemistry, decision trees assist in substance separation techniques, while in organic chemistry, they are used to model reaction pathways.

This study aims to bridge the gap between artificial intelligence and computational chemistry, proving AI's potential to revolutionise drug discovery and molecular research. The ability to predict inhibition potential with high accuracy can accelerate the development of new pharmaceuticals, reduce reliance on costly laboratory experiments, and contribute to a more efficient drug screening process. Furthermore, identifying key molecular descriptors correlated with inhibition could lead to a deeper understanding of chemical interactions, opening new avenues for research in medicinal chemistry and bioinformatics.

II Methodology

The source code is all stored in the *AI application for azophotoswitches optimization with pharmacological interest* GitHub repository.[12]

The target protein's ID is set at *CHEMBL230* corresponding to the COX-2 ID in the ChEMBL database. Utilising *requests* python package[13] a query URL is sent asking for all molecules with a know IC_{50} value (Def. 1) with a limit of 1000 entries per request. The process is iterated until all data is extracted leading a total of 7979 molecules. Hence the datasheet is processed in pandas dataframes[14] and encrypted into binary feather files to optimise reading-writting speed. By removing entries with the same canonical smiles a total of 5112 molecules remain. Among this entries well known drugs such as Celecoxib (Def. 2), Rofecoxib (Def. 3) or even Ibuprofen can be found. However the IC_{50} molecules range is comprehended from 10^{-3} to 10^{8} nM, a counterproductive range for the AI training procedure. A hard-coded range is filtered discarding all molecules outside the given range, for the most part of the analysis this range is set at $[0, 200]$ nM [1] which reduces the dataset to 1438 entries (i.e. molecules).

[1] this IC_{50} working range is the standard in this kind of studies.[15]

With the AlvaDesk-python[7] facility, the chemical descriptors (i.e., the chemical fingerprint) of each molecule are computed, providing a total of 5800 descriptors per molecule. Still, by deleting molecular descriptors with null values 2917 remain. Here, the Pearson correlation coefficient between each chemical descriptor and the IC_{50} value is computed, providing insight into the direct relationship between IC_{50} and the descriptors.

At this stage, the average IC_{50} is calculated, and the neighborhood size corresponding to the percentage defined by the hard-coded variable *percentageErased* is removed. This allows us to distinguish between *highly active molecules* and *least active molecules*, those with lower and higher IC_{50} values, respectively. Subsequently, each set of substances is randomly divided into two datasets: a *training set* and a *testing set*, following the proportion specified by the hard-coded variable *testSizeProportion*. This procedure is illustrated in Figure (1).

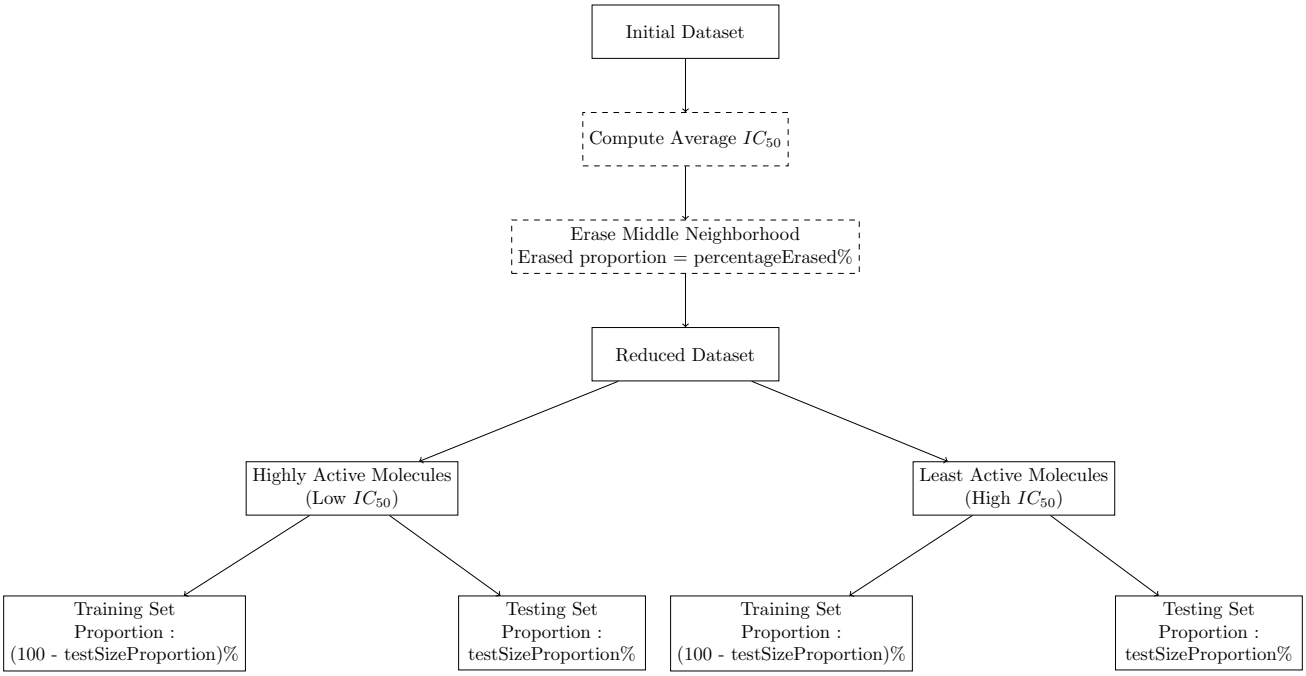

Figure 1: Scheme of the splitting procedure.

Afterward, a Random Forest algorithm is trained using the Sticky Learn[16] Python package, which is supported by Microsoft and Google among others. This algorithm generates a large number of random decision trees (determined by the hard-coded variable *numberOfTrees*), which are trained with the training sets. Later, these models are evaluated by predicting the IC_{50} values of the *testing sets*.

Using the results, the *True Positive Rate* (Def. 5), *True Negative Rate* (Def. 6), *Classification Accuracy* (Def. 7), and *Matthews Correlation Coefficient* (Def. 8) are computed. Based on these computational results, the variables *percentageErased*, *testSizeProportion*, and *numberOfTrees* are manually adjusted to obtain the best results.

References

[1] Baek, M.; et al. *Signal Transduction and Targeted Therapy* **2023**, *8*, 1–10.

[2] Singh, S.; Kumar, R.; Payra, S.; Singh, S. K. *Cureus* **2023**, *15*, e44359.

[3] Tunyasuvunakool, K.; et al. *Nature Structural and Molecular Biology* **2022**, *29*, 1155–1163.

[4] Swinney, D. C. In *Chapter 18 - Molecular Mechanism of Action (MMoA) in Drug Discovery*; Macor, J. E., Ed.; Annual Reports in Medicinal Chemistry; Academic Press, 2011; Vol. 46; pp 301–317.

[5] Zdrazil, B. et al. *Nucleic Acids Research* **2024**, *52*, D1180–D1192.

[6] Mauri, A. In *Ecotoxicological QSARs*; Roy, K., Ed.; Springer US: New York, NY, 2020; pp 801–820.

[7] Mauri, A.; Bertola, M. *International Journal of Molecular Sciences* **2022**, *23*.

[8] Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*, 2nd ed.; Wiley-VCH: Weinheim, Germany, 2009.

[9] National Cancer Institute Definition of COX-2 - NCI Dictionary of Cancer Terms. 2024; `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cox-2`, Accessed: 2024-10-09.

[10] Davies, N. M.; Jamali, F. *Pharmacology & Therapeutics* **2000**, *89*, 133–155.

[11] Breiman, L. *Machine Learning* **2001**, *45*, 5–32.

[12] Morales, S. C. AI Application for Azophotoswitches Optimization with Pharmacological Interest. 2025; `https://github.com/SirSergi0/Repository---AI-application-for-azophotoswitc`

[13] Reitz, K.; Chalasani, A. Requests: HTTP for Humans. `https://pypi.org/project/requests/`, 2023; Python package.

[14] McKinney, W. pandas: A Foundational Python Library for Data Analysis. 2023; `https://pandas.pydata.org`, Version 1.5.3.

[15] Khan, H. A.; Jabeen, I. *Frontiers in Pharmacology* **2022**, *13*.

[16] Pedregosa, F. et al. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

III Abbreviation index

COX-2 : cyclooxygenase-2

AI : Artificial Intelligence

ML : Machine Learning

RD : Random Forest

ID : Identification

IC_{50}: Half maximal inhibitory concentration

IC_{90}: 90 percent inhibitory concentration

IC_{99}: 99 percent inhibitory concentration

TP: True Positive

FP: False Positive

TN: True Negative

FN: False Negative

A Relevant definitions

Definition 1. *IC_{50}: Half maximal inhibitory concentration assigned to the drug concentration required for a 50% inhibition a protein. Other quantities such as IC_{90} or IC_{99} are also*

commonly used. However, IC_{90} is generally approximated as 10 times the IC_{50} concentration in virtue of experimental observations.[4] *For this project, we aim to identify substances with the lowest possible IC_{50}, as our goal is to minimize the presence of foreign substances in the living organism.*

Definition 2. *Celecoxib:* [2] *drug known to be a selective COX-2 inhibitor (currently is not* highly selective *respect to newer drugs), see Figure (2). It IC_{50} value is* 120 *nM.*

Definition 3. *Rofecoxib:* [3] *drug known to be a selective COX-2 inhibitor, see Figure (3). It IC_{50} value is* 180 *nM.*

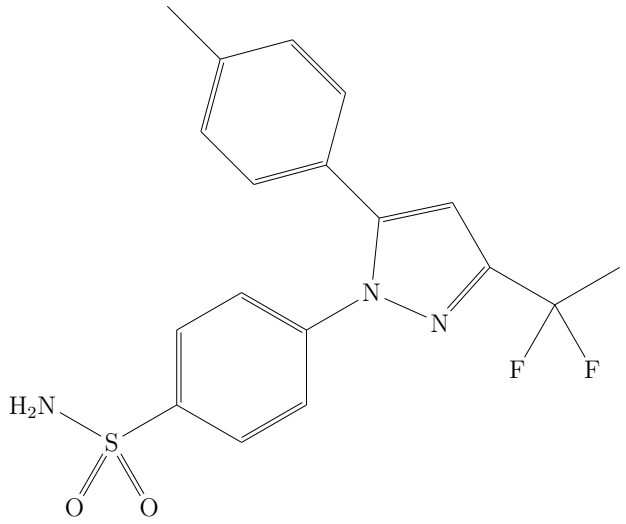

Figure 2: Chemical graph of Celecoxib.

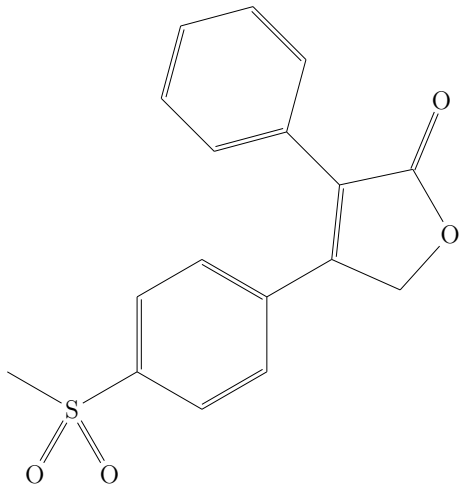

Figure 3: Chemical graph of Rofecoxib.

Definition 4. *Pearson correlation coefficient: Given set of pairs of data $\{(x_i, y_i)\}_{i=1}^n$ the pearson correlation factor r_{xy} is defined as,*

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}^2)}\sqrt{\sum_{i=1}^n (y_i - \bar{y}^2)}}, \quad (1)$$

where $\bar{x}$ and $\bar{y}$ stand for the average value of $x_{i\,i=1}^n$ and $y_{i\,i=1}^n$ respectively. Note that $r_{xy} \in [-1, 1]$. Therefore the sign of r_{xy} is tightly related to the sign of alinear regression, more precisely if $x > 0$, "y" generally[4] *increases when "x" increases, as well as if $x < 0$, "y" decreases when "x" increases.*

Definition 5. *True Positive Rate: quantity related to a Machine Learning Model's sensitivity defined as:*

$$\frac{TP}{TP + FN} \quad (2)$$

where TP, FP, TN, FN stands for "True Positive", "False Positive", "True Negative" & "False Negative" respectively.

Definition 6. *True Negative Rate: quantity related to a Machine Learning Model's specificity*

[2] UPAC name: 4-[5-(4-Methylphenyl)-3-(trifluoromethyl)pyrazol-1-yl]benzenesulfonamide

[3] UPAC name: 4-(4-methylsulfonylphenyl)-3-phenyl-5H-furan-2-one

[4] We would like to remark that the word "generally" stands for "the majority of the cases", since "generally" is commonly interpreted as a non-scientific/non-objective word

defined as:

$$\frac{TN}{TN + FN} \tag{3}$$

where TP, FP, TN, FN stands for "True Positive", "False Positive", "True Negative" & "False Negative" respectively.

Definition 7. *Classification Accuracy: quantity related to a Machine Learning Model's efectiveness defined as:*

$$\frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

where TP, FP, TN, FN stands for "True Positive", "False Positive", "True Negative" & "False Negative" respectively.

Definition 8. *Matthews Correlation Coefficient: quantity related to a Machine Learning Model's prediction capacity defined as:*

$$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{5}$$

where TP, FP, TN, FN stands for "True Positive", "False Positive", "True Negative" & "False Negative" respectively. A Matthews Correlation Coefficient equal to 1 stands for a perfect prediction a Matthews Correlation Coefficient equal to 0 indicates that the classifier's predictions are no better than random guessing, and a Matthews Correlation Coefficient equal to -1 stand for a total disagreement between predictions and actual outcomes.