UAB
Universitat Autònoma de Barcelona

Facultat de Ciències

Treball de fi de grau

TFG2425_045 AI Application for Azophotoswitches Optimization with Pharmacological Interest

Direcció:

Dr. Miquel Moreno Ferrer

Dr. Àngels González Lafont

Alumne:

Sergio Castañeiras Morales

NIU:

1598456

Juny 2025

Treball de fi de grau realitzat al Departament de Quimica i presentat a la Facultat de Ciancies de la Universitat Autnòma de Barcelona per a l'obtenció del Grau en Quimica

“The dumbest people I know are those who know it all.”

Malcolm S. Forbes

Resum analític

L'intel·ligència artificial es presenta com una de les revolucions del segle XXI. En particular, el sector de la química computacional està sent profundament sacsejat per aquesta revolució. Aprofitant la inèrcia i l'interès creixent en aquest camp, aquest treball pretén aplicar diferents models d'intel·ligència artificial en l'estudi d'una proteïna d'especial interès per a la nostra salut, la Ciclooxigenasa-2 (COX-2).

La *prostaglandina-endoperoxid sintasa 2* (PTGS2), també coneguda com COX-2, és una proteïna que, en circumstàncies normals, acostuma a romandre inactiva [1], llevat de la seva expressió durant processos inflamatoris. Així mateix, la manca de retorn a nivells baixos d'expressió després de la inflamació ha estat relacionada amb l'aparició de diferents formes de càncer [2]. Aquest fet ha convertit la COX-2 en objecte d'estudi de nombroses investigacions científiques [3], fet que la fa un punt de partida idoni per al desenvolupament d'algoritmes de Machine Learning (ML), ja que disposa d'una gran quantitat de dades per entrenar els models i contrastar els resultats.

L'objectiu principal del projecte és el desenvolupament d'un programari generador d'IAs capaces de predir la concentració d'inhibició al 50% (IC_{50}) per a la COX-2[1] amb la màxima precisió possible. Per fer-ho, s'extreuen totes les dades de molècules conegudes amb un potencial d'inhibició establert per a la COX-2. Després d'un filtratge configurable per l'usuari, es calculen 5.900 descriptors químics per a cadascuna de les entrades amb el programari AlvaDesk [5][6]. Seguidament, una part de les dades s'utilitza per entrenar models de Random Forest (RF) [7], mentre que la resta es reserva per validar la precisió de les prediccions.

Cal remarcar la principal hipòtesi que sustenta aquest procés i el projecte en general: *Existeix una combinació (o diverses combinacions) de descriptors químics directament relacionada amb el potencial d'inhibició de la proteïna*. Malgrat que aquesta afirmació pugui semblar natural, el cost computacional associat és immens. Tot i així, la precisió de les prediccions dels models apunta a la validesa d'aquesta hipòtesi, si bé continua essent una conjectura per manca d'una prova definitiva.

Finalment, es fan servir models per predir l'IC_{50} de 50 *azophotoswitches* amb dades d'energia de Gibbs d'acoblament proporcionades pel grup MolBioMed [8]. L'anàlisi mostra una correlació clara entre ambdues quantitats, reforçant la hipòtesi del projecte.

[1] En realitat, el programari funciona per a qualsevol proteïna amb entrada a la base de dades de ChEMBL [4], malgrat que l'objecte d'estudi és la COX-2.

Resumen analítico

La inteligencia artificial se presenta como una de las revoluciones del siglo XXI. En particular, el sector de la química computacional está siendo profundamente sacudido por esta revolución. Aprovechando la inercia y el interés creciente en este campo, este trabajo pretende aplicar diferentes modelos de inteligencia artificial en el estudio de una proteína de especial interés para nuestra salud, la Ciclooxigenasa-2 (COX-2).

La prostaglandina-endoperóxido sintasa 2 (PTGS2), también conocida como COX-2, es una proteína que, en circunstancias normales, suele permanecer inactiva [1], excepto por su expresión durante procesos inflamatorios. Asimismo, la falta de retorno a niveles bajos de expresión después de la inflamación ha sido relacionada con la aparición de diferentes formas de cáncer [2]. Este hecho ha convertido a la COX-2 en objeto de estudio de numerosas investigaciones científicas [3], lo que la convierte en un punto de partida idóneo para el desarrollo de algoritmos de Machine Learning (ML), ya que dispone de una gran cantidad de datos para entrenar los modelos y contrastar los resultados.

El objetivo principal del proyecto es desarrollar un software capaz de generar AIs que predigan la concentración de inhibición al 50% (IC_{50}) para la COX-2[2] con la máxima precisión. Para ello, se extraen datos de moléculas con potencial inhibidor conocido, se aplican filtros configurables y se calculan 5.900 descriptores químicos por entrada mediante AlvaDesk [5][6]. Parte de los datos se usa para entrenar modelos de Random Forest (RF) [7], y el resto, para validar su precisión.

Cabe remarcar la principal hipótesis que sustenta este proceso y el proyecto en general: *Existe una combinación (o varias combinaciones) de descriptores químicos directamente relacionada con el potencial de inhibición de la proteína*. Aunque esta afirmación pueda parecer natural, el coste computacional asociado es inmenso. Aun así, la precisión de las predicciones de los modelos apunta a la validez de esta hipótesis, si bien sigue siendo una conjetura por falta de una prueba definitiva.

Finalmente, los modelos se utilizan para predecir el IC_{50} de 50 *azophotoswitches*, de los cuales se tienen datos sobre la energía libre de acoplamiento proporcionados por el grupo de investigación MolBioMed [8]. El análisis estadístico de las predicciones refleja una clara correlación entre ambas cantidades, lo que refuerza la hipótesis del proyecto.

[2] Aunque el software es aplicable a cualquier proteína con entrada en la base de datos ChEMBL [4], el estudio se centra en la COX-2.

Analytical abstract

Artificial intelligence is emerging as one of the revolutions of the 21st century. In particular, the field of computational chemistry is being profoundly shaken by this revolution. Taking advantage of the momentum and growing interest in this field, this work aims to apply different artificial intelligence models to the study of a protein of special interest to our health, Cyclooxygenase-2 (COX-2).

The prostaglandin-endoperoxide synthase 2 (PTGS2), also known as COX-2, is a protein that, under normal circumstances, tends to remain inactive [1], except for its expression during inflammatory processes. Likewise, the failure to return to low expression levels after inflammation has been linked to the onset of various forms of cancer [2]. This fact has made COX-2 the subject of numerous scientific investigations [3], making it an ideal starting point for the development of Machine Learning (ML) algorithms, as it provides a large amount of data for training models and validating results.

The main objective of the project is to develop software capable of generating AIs that can predict the 50% inhibition concentration (IC_{50}) for COX-2[3] with the highest possible accuracy. To achieve this, all known molecular data with an established inhibition potential for COX-2 are extracted. After a user-configurable filtering process, 5,900 chemical descriptors are calculated for each entry using the AlvaDesk software [5][6]. Subsequently, part of the data is used to train Random Forest (RF) models [7], while the rest is reserved to validate the accuracy of the predictions.

It is important to highlight the main hypothesis that underpins this process and the project as a whole: *There exists a combination (or multiple combinations) of chemical descriptors that are directly related to the inhibition potential of the protein*. While this statement may seem intuitive, the computational cost associated with it is immense. Nevertheless, the accuracy of the model predictions supports the validity of this hypothesis, although it remains a conjecture due to the lack of definitive proof.

Finally, models are used to predict the IC_{50} of 50 *azophotoswitches* based on Gibbs free energy of binding data provided by the MolBioMed research group [8]. The analysis shows a clear correlation between both quantities, supporting the project's hypothesis.

[3] Actually, the software works for any protein with an entry in the ChEMBL database [4], although the study focuses on COX-2.

Contents

1 List of abbreviations

AI	Artificial intelligence.
COX-2	Cyclooxygenase-2.
FN	False Negative.
FP	False Posive.
IC_{50}	Half Maximal Inhibitory Concentration.
IC_{90}	90 Percent Inhibitory Concentration.
IC_{99}	99 Percent Inhibitory Concentration.
ID	Identificator.
ML	Machine Learning.
NSAID	Non-Steroidal Anti-Inflammatory Drug.
PTGS2	Prostaglandin-endoperoxide synthase 2.
RF	Random Forest.
SMILES	Simplified Molecular Input Line Entry System.
TN	True Negative.
TP	True Posive.

List of Figures

List of Tables

2 Introduction

The impact of Artificial intelligence (AI) on science has been nothing short of a groundbreaking revolution, with few comparable precedents. The rapid advancements in AI have transformed numerous scientific fields[9][10], including computational chemistry. Today, one of the primary goals of computational chemistry is to predict the properties of unstudied substances while minimizing experimental costs. Traditional approaches in chemistry often rely on complex laboratory techniques, which, while effective, can be time-consuming, expensive, and resource-intensive. In contrast, computational chemistry offers a wide range of methods capable of predicting a molecule's properties with reasonable accuracy. However, when AI comes into play, predictions have demonstrated an almost surgical precision.

Perhaps one of the most representative events showcasing the enormous impact of AI on chemistry is the 2024 Nobel Prize in Chemistry. The winners, David Baker[11], along with Demis Hassabis and John Jumper[12], were not traditionally trained chemists. Instead, their expertise lies in AI algorithms and Machine Learning (ML) methods applied to protein research. This milestone, among others, triggered a surge of chemistry researchers diving into the world of AI, seeking applications for their respective fields. Today, the thrilling progress in computational chemistry has been further reinforced by these cutting-edge tools[13], and the rapid pace of development keeps the scientific community eagerly anticipating future applications in fields such as medicine, materials science, and beyond. In this project we aim to apply the new AI and ML algorithms to our object of study, the Prostaglandin-endoperoxide synthase 2 (PTGS2) also known as COX-2, depicted in Figure (1), a protein tightly linked to the onset of numerous cancers forms[14].

Although significant advancements have been made, cancer still accounts for over 8 million deaths per year worldwide, and the scientific and medical communities remain far from achieving its complete eradication. Inflammation is one of the hallmarks of carcinogenesis, in fact, various cancer therapies target inflammation as a means of preventing and reducing cancer occurrences. When a tissue is damaged, inflammation protects the organism from infections caused by external pathogens, a key function of the immune system to prevent the presence of invaders in the body. During the inflammatory process, cells proliferate under the command of the immune system to replace the damaged cells of the affected tissue. However, if this cell reproduction continues beyond the healing

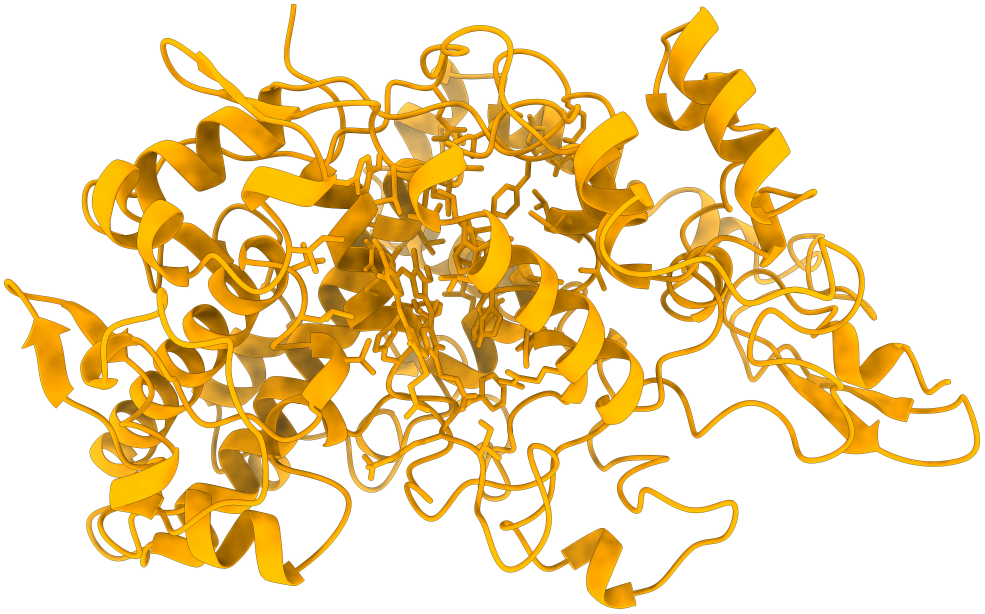
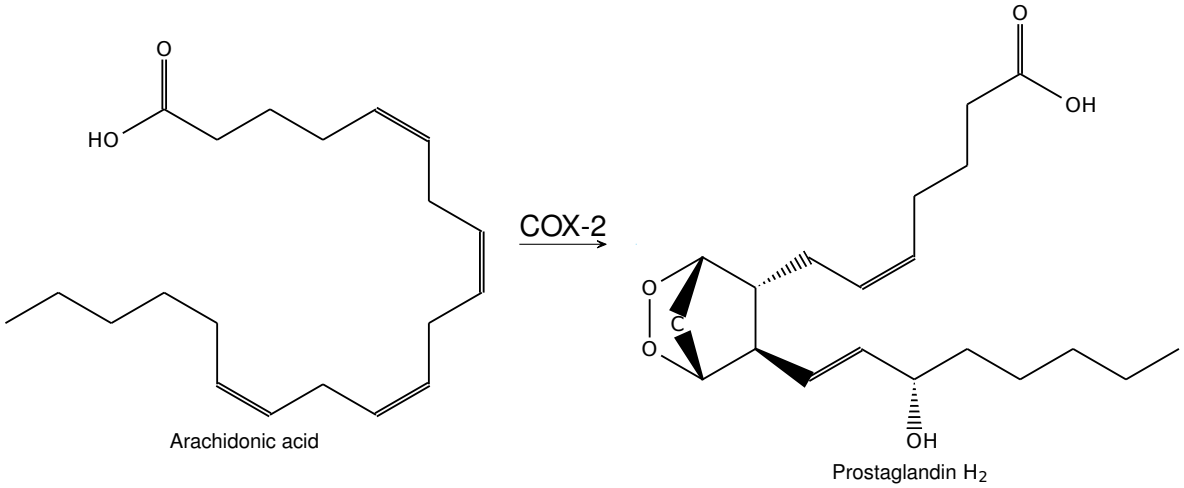


Figure 1: Structural representation of the COX-2 generated by the ChimeraX software[15].

of the damaged tissue, it can potentially lead to cancer, contradicting the initial healing purpose of the inflammatory process. In some cases, inflammation can become chronic, leading to tumour development and uncontrolled cell proliferation. As a result, a wide range of drug prototypes have been designed to suppress inflammation. However, many of these drugs have been linked to severe side effects, including immunosuppression, cardiovascular risks, and gastrointestinal complications. Consequently, the administration of these drugs is often contraindicative, and the search for a more effective and safer treatment remains ongoing. Plenty of the research in this field is mainly focused on the pro-inflamatory enzyme COX-2, one of the main commanders in the inflammation process and responsible to convert the arachidonic acid to prostaglandin H_2 (Scheme (1)). Subsequently, a therapy based on the chemical inhibition of the COX-2 with no side effects has been one of the research lines in cancer treatment leading to a considerable amount of experiments and data.

The increasing interest in COX-2 inhibitors has granted the scientific community with an extensive database of molecular inhibition potentials for this protein. In this project, we focus on the Half Maximal Inhibitory Concentration (IC_{50}), a standard metric representing the concentration of a drug required to inhibit 50% of a target protein's activity. Related measures include 90 Percent Inhibitory Concentration (IC_{90}) and 99 Percent Inhibitory Concentration (IC_{99}), which correspond to 90% and 99% inhibition, respectively. The

lower the IC_{50} value of a molecule, the lower the concentration needed to inhibit COX-2, indicating higher efficiency. This factor is crucial in drug design, as a lower required dosage minimizes the presence of foreign substances in the body, thereby reducing the risk of adverse effects[4].



Scheme 1: Reaction catalysed by the COX-2 from arachidonic acid to prostaglandin H_2.

Determining an experimental IC_{50} value can be both costly and time-consuming[5]. On the other hand, AI provides an alternative by offering highly accurate predictions based on existing data, optimising research processes, and accelerating scientific discovery. In this scenario this project aims to implement artificial intelligence in computational chemistry, concretely, using AI-based algorithms to predict a drug's inhibition potential[16] for a given protein with a relativelly good accuracy [6]. To achieve this, we make use of the ChEMBL database[4], 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 (Definition (2)) [7] of each molecule in the database are then computed using AlvaDesk[5][6] software. Around 5000 descriptors are calculated[17], 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

[4]Naturally, multiple other factors influence drug side effects.

[5]Usually this parameter is computed throughout the Cheng Prusoff Equation (Definition (5)) with experimental data.

[6]The accuracy of the ML models is discussed on the Results and Discussion section (5).

[7]All the definitions are stored in the Appendix (A) *Relevant definitions*. We recommend its consultance for further information.

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.

The AI algorithm used is a in this study is a ML model known as the Random Forest (RF) algorithm[7], a powerful ensemble learning method that generates multiple decision trees (Definition 8) 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.

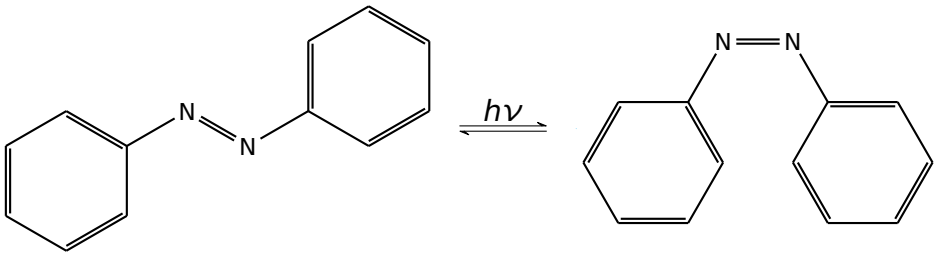
Moreover, the choice of the Random Forest (RF) algorithm is motivated by the presence of decision trees in various chemistry-related fields. In spectroscopy, for instance, decision trees are used in group theory to classify molecular symmetry. Similarly, in analytical chemistry, decision trees assist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

By applying AI models to COX-2, we assess their compatibility with the latest research findings[3], 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. Additionaly, this study aims to bridge the gap between AI and computational chemistry, reinforcing the 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.

Currently, the pharmacological therapies for the COX-2 inhibition are based on Non-Steroidal Anti-Inflammatory Drug (NSAID)[2] such as the popular ibuprofen or aspirin. However they have proved to be related to cardiovascular diseases and are contraindicated for people with more than 50 years or people with gastrointestinal problems, among others. Still some alternative therapies related with NSAIDs are being explored, in particular, one of the most revolutionary ideas is the application of azophotoswitches in drug design.

We define a molecule as a *photoswitch* if it contains a bond that undergoes a configurational transformation upon photoexcitation. In particular, we take *azophotoswitches* as a class of photoswitches in which the photo-sensitive bond responsible for isomerization is an azo group (i.e. N=N). Azophotoswitches represent the most common and widely studied type of photoswitches due to their robust and tunable photochemical properties. For instance, consider the case of *(cis/trans)-N,1-diphenylmethanimine*,



Scheme 2: (cis/trans)-N,1-diphenylmethanimine conversion as an example of an azophotoswitch.

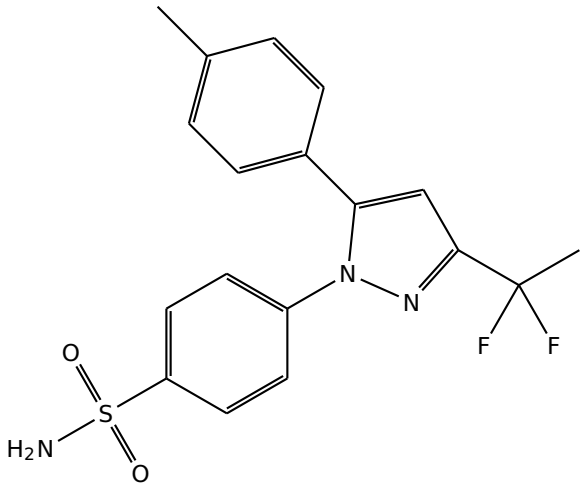
The photo-induced cis/trans isomerization leads to distinct bioactivities for each configuration. This variation arises from the stereochemical constraints required for a molecule to bind to a target protein. Typically, a protein's binding site has a specific shape, and only molecules whose conformation matches this shape can interact effectively. In the context of an azophotoswitch, one isomer may fit precisely into the binding pocket, leading to strong interactions, while the other may not. Accordingly, we refer to the isomer that interacts most effectively with the protein as the active configuration, and the other as the inactive configuration.[8] [9]

[8] In the context of this project, interaction refers to inhibition of the protein's activity.

[9] Generally, the cis-isomer is considered the active configuration, while the trans-isomer is considered inactive. However, exceptions exist.

The primary application of azophotoswitches in drug design is the administration of the inactive isomer, which is assumed to be non-toxic to the organism. Later, the active configuration is generated through selective photo-excitation at the target site. This strategy minimizes drug activity in unintended tissues, thereby reducing side effects. As a result, higher dosages may be administered safely by localizing the therapeutic effect to the desired area. These therapies are still in the experimental stage and remain primarily within research and development.

Among the most promising drug candidates for COX-2 inhibition are Celecoxib (Definition 3) depicted in Scheme (3) and Rofecoxib (Definition 4).[10] At the MolBioMed research group [8], researchers are investigating azophotoswitches as potential COX-2 inhibitors, using molecular structures inspired by Celecoxib. Various computational analyses have been conducted, with particular attention to the binding free energy ($\Delta G_{\text{binding}}$).[11]



Scheme 3: Chemical graph of Celecoxib.

As an application of our ML-based predictive models, we aim to estimate the IC_{50} values of azophotoswitch prototypes.[12] We will then compare these predictions with the computed $\Delta G_{\text{binding}}$ values, as both metrics are indicative of inhibition efficacy. While a direct linear correlation is not expected, we hypothesize that lower IC_{50} values should correspond to lower (more negative) $\Delta G_{\text{binding}}$ values, reflecting stronger binding affinities. These relationships will be explored in detail in Section 5.

All in all, this project demonstrates the application of AI in drug design, promoting its integration into chemical computations as an initial step prior to experimental investiga-

[10] Naturally, these compounds are often referred to as *coxib drugs* due to the COX-2.
[11] Additional results are provided in Appendix C.
[12] These values are predictions, as no experimental or simulated data are currently available.

tions. Such tools can help avoid time-consuming and costly unproductive experiments, assisting both theoretical and experimental chemists by providing fast computational insights into molecular properties, thus offering valuable intuition before proceeding to laboratory work.

3 Objectives

This project aims to develop a fully functional software tool capable of accurately predicting a molecule's inhibition potential (IC_{50}), leveraging existing data from the ChEMBL database [4]. The software is thoroughly documented and available in the GitHub repository: *AI Application for Azophotoswitches Optimization with Pharmacological Interest* [18][19].

The project specifically focuses on the COX-2 protein. Thanks to the availability of a large IC_{50} dataset for this target, the model's predictive reliability can be effectively validated. The prediction process is powered by an AI tool designed to deliver high-quality results, taking into account relevant molecular descriptors. The software also serves to cross-check computational results previously obtained by the MolBioMed research group [8]. The application centres on azophotoswitches, molecules of particular pharmacological interest as discussed in Section 2. Additionally, a molecular docking study is performed on one of the most promising azophotoswitch candidates.

In conclusion, this study seeks to bridge the gap between Artificial Intelligence and computational chemistry, demonstrating AI's potential to revolutionize drug discovery and molecular research. Accurate predictions of inhibition potential could accelerate the development of new pharmaceuticals, reduce the need for costly laboratory experiments, and streamline the drug screening process. Furthermore, identifying key molecular descriptors linked to inhibition may provide valuable insights into chemical interactions, opening new directions for research in medicinal chemistry and bioinformatics.

4 Methodology

The source code is entirely stored in the *AI application for azophotoswitches optimization with pharmacological interest* GitHub repository[18] and it has been entirely written by Sergio Castañeiras Morales for this project[19]. The consultace of this repository is highly recommended in order to comprehended the insights and the workflow of the project. By the time this project is presented, it remains as an open-source repository of the MolBioMed research group [8].

The target protein's ID is set at *CHEMBL230* corresponding to the COX-2 ID in the ChEMBL database[4]. Utilising *requests* python package[20] a query URL is sent asking for all molecules with a know IC_{50} value (Definition 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 data-sheet is processed in pandas dataframes [21] (Definition 14) and encrypted into binary feather files to optimise reading-writing speed. By removing entries with the same canonical SMILES (Definition 9) a total of 5112 molecules remain. Among these entries well known drugs such as Celecoxib (Definition 3), Rofecoxib (Definition 4) 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. Since we are interested in testing azophotoswitches with presumably low IC_{50}, training ML models with data of the order of 10^{8} or 10^{3} nM can be counterproductive since the model might misinterpret the data.[13]. Consecutivelly, 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 [14] which reduces the dataset to 1438 entries, i.e. molecules.

With the AlvaDesk-python [6] facility, the chemical descriptors (i.e., the chemical fingerprint) of each molecule are computed, providing a total of 5800 descriptors per molecule, although only 2917 chemical descriptors yield to no *null values*. This is because some chemical descriptors need from the presence of a certain atom or active group, for example chemical descriptor related to the presence of a carboxylic acid, or chemical descriptor related triple bound among other possibilities. The computations of these chemical descriptors appear as *null values*. Here, the Pearson correlation coefficient (Definition 6) between each chemical descriptor and the IC_{50} value is computed, providing insight into the direct relationship between IC_{50} and the descriptors. This rela-

[13] Typically the range of the azophotoswitches IC_{50} is arround the Celecoxib's IC_{50}, i.e. arround 120 nM

[14] this IC_{50} working range is the standard in these kind of studies [13].

tionship will be discussed in the Results and Discussion Section (5).

At this stage, the average IC_{50} is calculated, and the neighbourhood 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. By doing so, it is possible to compute the classification accuracy statistics of the model from the cluster association of each prediction. Thus we can compare each prediction of each molecule with experimental data. For instance, for an experimentally proved highly active molecule if the model predicts this molecule to be highly active (has a low IC_{50} prediction) we consider this prediction as True Posive (TP). Similarly we define a True Negative (TN), False Posive (FP) and False Negative (FN). This procedure is a standard for evaluating a ML model reliability.

One can see a clear analogy to medical testing. Consider a subject who undergoes a COVID-19 test to determine whether they are infected or not. Analogously, we test whether a molecule is a potential COX-2 inhibitor using our ML model. There are four possible outcomes: the subject is infected (or not), and the test is positive (or negative). If the subject is infected and the test result is positive, it is considered a TP. Similarly, if a molecule is a highly active inhibitor and the model predicts it as such, it is also classified as a TP. This analogy extends to the definitions of TN, FP, and FN.

Subsequently, each set of substances is randomly divided into two datasets: a *training set* for training the RF algorithm and a *testing set* for testing the statistics and reliability of the RF model, following the proportion specified by the hard-coded variable *testSizeProportion*. This procedure is illustrated in Figure (2).

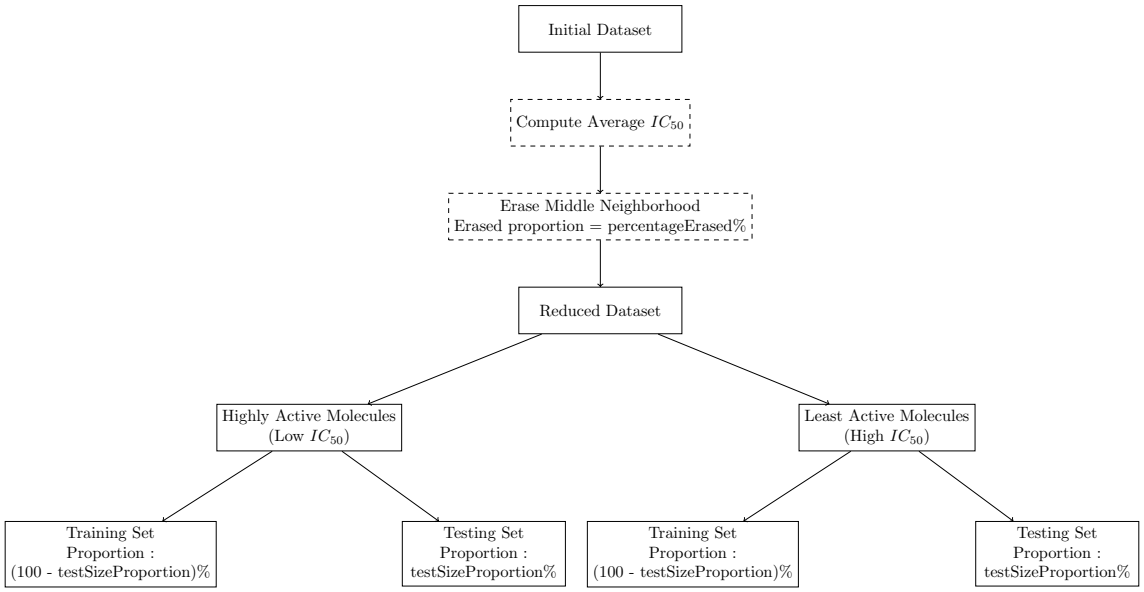


Figure 2: Splitting and processing data's scheme.

Afterwards, a Random Forest (RF) algorithm is trained using the Sticky Learn [22] Python package, which is supported by Microsoft and Google among others. Although the project's primary focus is not the detailed workings of Machine Learning (ML) algorithms but its applications in computational chemistry, it is natural to ponder the essence of what the machine is doing. The name "Random Forest" is directly tied to the analogy between a tree and a decision tree.

As described in Definition (8), the Random Forest algorithm generates a collection of random decision trees. The number of trees in the forest is a configurable parameter of the algorithm. More specifically, the Random Forest algorithm produces a set of decision trees, each trained on a random subset of the training data. The structure of each tree adjusts during training to make predictions that closely approximate the true target values. The final output of the Random Forest is typically an aggregate of the predictions from all the individual trees.

One notable characteristic of this method is its inherent randomness. Each Random Forest model is unique, even if configured with the same parameters and initialized under identical conditions. However, experimental evidence suggests this randomness does not significantly impact the results. If two Random Forest algorithms are trained on the same dataset using identical parameters, their outputs will converge, even if the internal structures of their decision trees differ.

The RF algorithm generates a large number of random decision trees (as defined by

the hard-coded variable *numberOfTrees*), which are trained on subsets of the training data. Typically, increasing the number of trees enhances the model's predictive accuracy. However, this comes at the cost of greater computational complexity. Importantly, prediction accuracy cannot be arbitrarily improved simply by increasing the number of trees. The model's performance is fundamentally limited by the quality and quantity of the training data. If the data lacks from information about the underlying protein-related phenomena, higher accuracy cannot be achieved regardless of how many trees are used.

In general, the relationship between data quality/quantity and the number of trees can be summarized as the following

$$\uparrow \text{Data} + \uparrow \text{Trees} \Rightarrow \uparrow \text{Prediction Accuracy}$$

$$\downarrow \text{Data} + \downarrow \text{Trees} \Rightarrow \downarrow \text{Prediction Accuracy}$$

However, increasing either component in isolation, i.e. adding more data or more trees, does not necessarily lead to better accuracy. For the available dataset in this work, we observed training with 500 trees yields similar results respect the ones obtained by training with 300 trees.

Subsequently, these models are evaluated by predicting the IC_{50} values of the *testing sets*. Using the predictions, we compute the *True Positive Rate*[15] (Definition 10), *True Negative Rate* (Definition 11), *Classification Accuracy* (Definition 12), and *Matthews Correlation Coefficient* (Definition 13). Based on these metrics, the variables *percentageErased*, *testSizeProportion*, and *numberOfTrees* are manually adjusted to optimize model performance.

Finally, the RF model with the best performance[16] is used to predict the IC_{50} values of the azophotoswitch prototypes listed in Appendix B. This process begins with computing the chemical descriptors using AlvaDesc software [5]. Irrelevant descriptors are then removed, retaining only those used during training. By introducing the selected descriptors into the trained RF model, the predicted IC_{50} values are obtained and subsequently analyzed.

Finally, a docking study was performed using GOLD2022 [23] by the MolBioMed research group ([8]). This software employs a genetic algorithm, along with molecular

[15]Which is related to how well performs a ML model when predicting a molecule to be highly active.

[16]The criteria for determining the model with "better statistics" is discussed in the Results section (Section 5.3) since this term references AI statistics rather than the usual statistics.

modelling techniques, to predict the optimal binding pose of the ligand within the COX-2 binding site. The genetic nature of the algorithm allows for a broad exploration of the conformational and positional space between the ligand and the protein. By evaluating and ranking ligand interactions, such as hydrogen bonding, van der Waals forces, and metal coordination, unstable or energetically unfavourable conformations are discarded, resulting in a set of plausible, energetically stable binding orientations. The docking calculations were carried out within a 26 Å radius sphere centred at Arg513, comprehending the totality of the binding site.

5 Results and Discussion

5.1 Available Data

Figure (3) presents the IC_{50} values extracted from the ChEMBL database under target ID *CHEMBL230*, limited to the interval $[0, 200]$ nM. Figure (3) clearly shows that multiples of 10 are significantly overrepresented compared to other values. Upon analyzing data from additional proteins, we can confirm this is not an exception but rather the norm. This suggests experimental chemists tend to round their results to the nearest multiple of 10, introducing a systematic bias. As a result, a certain degree of inherent error will always be present in the dataset, and this limitation must be acknowledged and accounted for in subsequent analyses. This observation forms a key foundation for our later results.

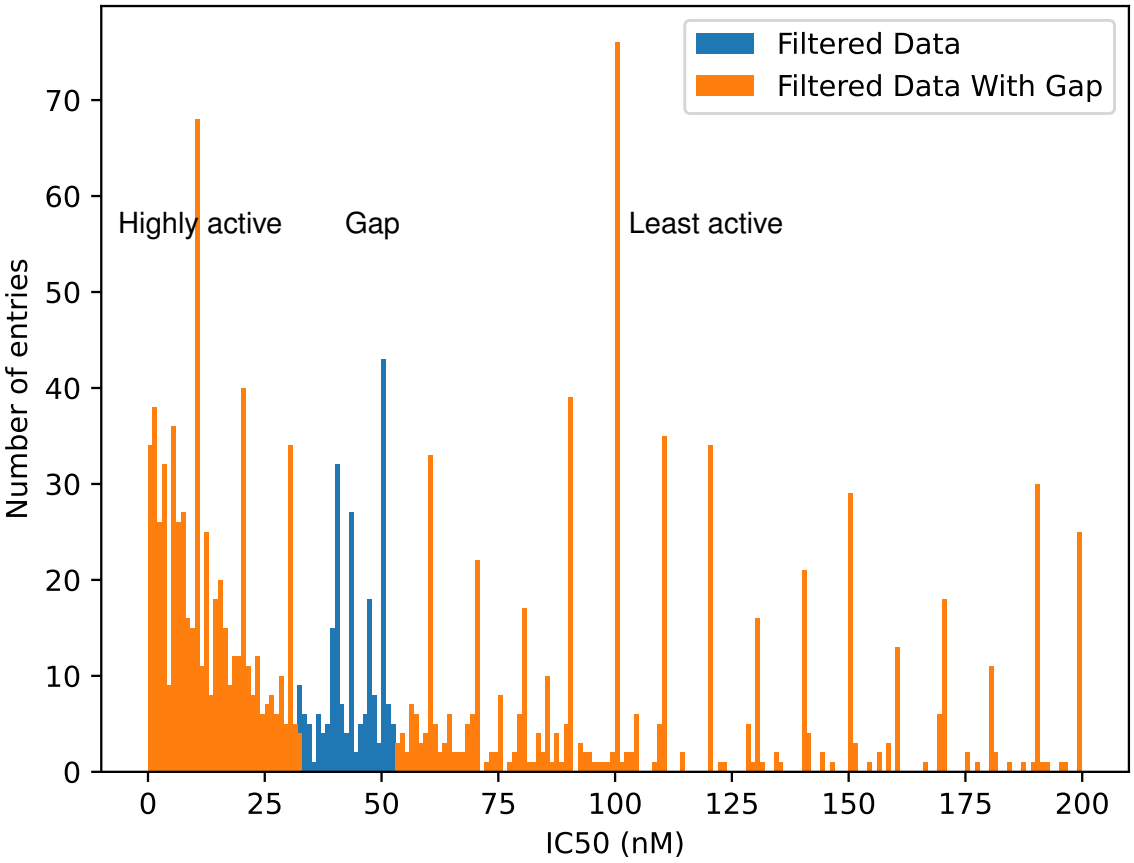


Figure 3: IC_{50} values from our dataset and the creation of a 20% gap.

Figure (3) also illustrates the process of generating a gap in the data. Given a dataset, the average IC_{50} is first computed, approximately 47 nM in this case. A proportion of the data, defined by the hard-coded variable *percentageErased*, is then removed to create

a gap centred around the average value. In Figure (3), the full dataset is represented in blue and orange, while the blue region, centred at 47 nM, has been deleted to introduce the gap. In this specific example, the gap accounts for 20% of the total data.

5.2 Correlation of Individual descriptors with IC_{50}

Afterwards, chemical descriptors are calculated using the AlvaDesc software. Subsequently, we compute the Pearson Correlation Coefficients[17] as defined in Definition (6), measuring the linear relationship between each descriptor and the corresponding IC_{50} value for every molecule. The results are visualized in Figure (4).

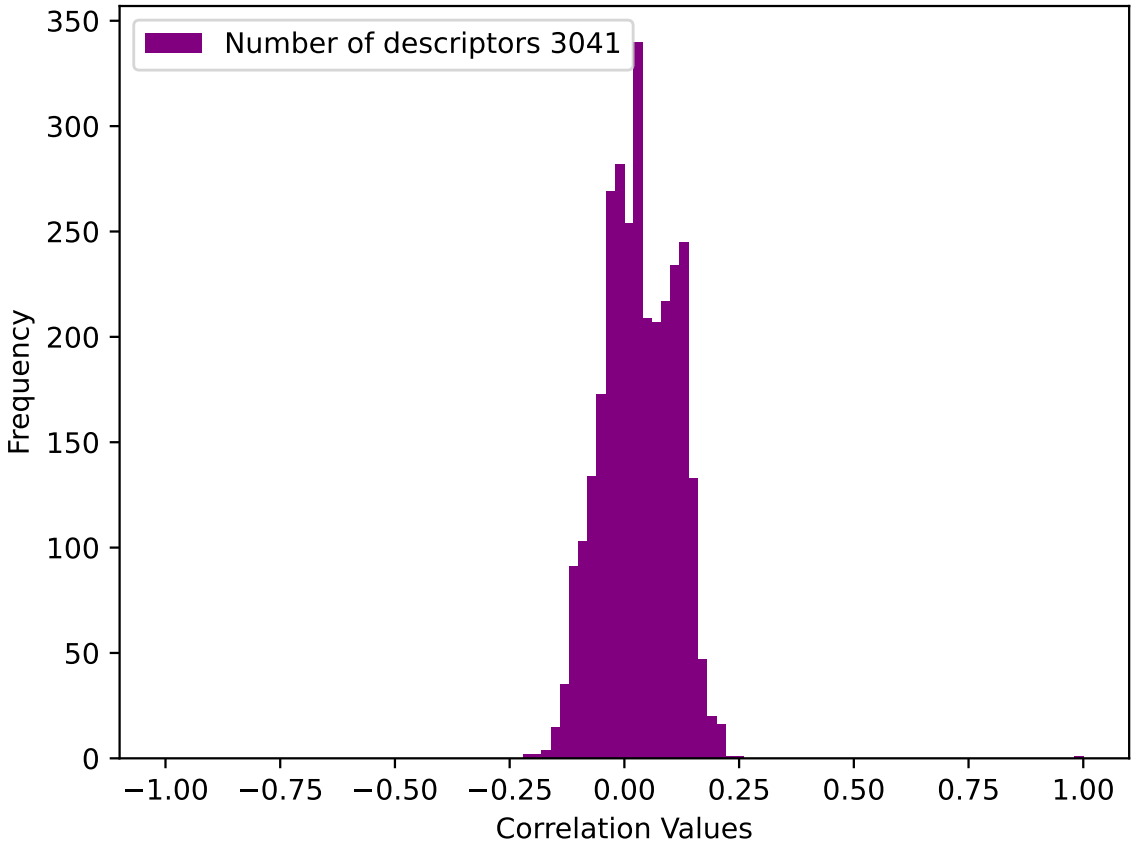


Figure 4: Pearson correlation coefficient for each chemical descriptor.

The distribution of correlation values resembles a Gaussian bell curve centred around zero, as expected. This suggests that there is no dominant or inherently preferred chemical descriptor strongly associated with IC_{50} variation for this protein. While this may seem to conflict with the project's objective, it actually reinforces the importance of using

[17]The Pearson Correlation Coefficient is sometimes denoted as r^2, though strictly speaking r^2 refers to the coefficient of determination, which is the square of the correlation coefficient.

ensemble methods.

This result is entirely expected. If a universally preferred descriptor existed, the use of AI methods would be largely unnecessary, and significant advancements would likely have already been achieved through simpler statistical models. Nonetheless, our central hypothesis remains valid: "*there exists at least one combination of chemical descriptors that correlates directly with the inhibition of the target protein*". The observed distribution of correlation values does not contradict this hypothesis, rather, it highlights the need for models that can identify and exploit subtle multivariable relationships.

It might be tempting to assume that discarding descriptors with the lowest r^2 values would improve model performance. This idea will be further explored in later sections.

5.3 Pearson Correlation Coefficient and Mean Squared Error

Once the models have been generated, i.e., once the relevant parameters are set, we proceed with generating predictions. At this stage, we compute the Pearson Correlation Coefficient (Definition 6) and the Mean Squared Error (Definition 7) between the predicted and experimental IC_{50} values.

Figure (5) presents two plots showing how both the Pearson Correlation Coefficient and the Mean Squared Error vary as a function of the erased percentage. These metrics provide insight into the model's performance under different levels of data availability.

From the discussion in Section 5.1, we understand the dataset suffers from inherent precision issues, particularly at higher IC_{50} values. Additionally, the precision of several measurements remains unknown. Since our models do not incorporate uncertainty estimates for each individual data point, it is unrealistic to expect r^2 and mean squared error values of exactly 1 and 0, respectively. Some degree of variance is inevitable due to the nature of the data. As such, discarding more than 50% of the dataset, regardless of what statistical metrics may suggest, would be a critical error, as it would lead to substantial information loss. Hence, we consider this statistics as not representative of the hole model's accuracy. Still, the computed values of r^2 and mean squared error remain useful indicators of the model's performance up to a certain threshold.

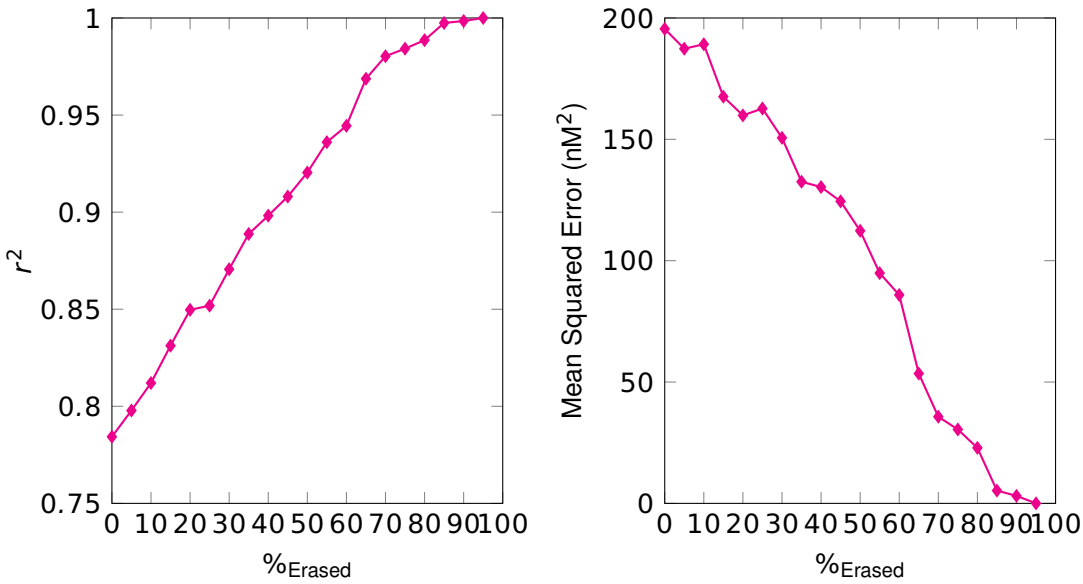


Figure 5: Plots of $\%_{\text{Erased}}$ vs r^2 and $\%_{\text{Erased}}$ vs Mean Squared Error. For a maximum IC_{50} = 200 nM and a minimum of 0 nM, taking into account all the chemical descriptors, the 20.0% of the remaining data is saved for testing and the 80.0% for training. The training is done with 200 trees.

In this work, we arbitrarily define this threshold at $r^2 = 0.90$. This hard-coded value is a conventional choice, established to determine the point beyond which further improvements in r^2 become statistically irrelevant, given the underlying noise in the data. This convention is loosely inspired by standards in analytical chemistry,[18] where correlations above 0.90 are often deemed acceptable. However, it is important to emphasize this threshold is subjective and should not be treated as an absolute rule.

Moreover, we can hypothesize creating a larger gap in the data facilitates the model's ability to distinguish between highly active and weakly active inhibitors. This intuition is reflected in the boundary cases illustrated in Figure (5). As the erased percentage approaches 100%, the r^2 value tends to zero and the mean squared error approaches one. This indicates that, in such extreme cases, the RF model is able to classify only the most distinct molecules, those with IC_{50} values near the extremes of the interval. However, this is not a desirable scenario, as a significant portion of intermediate data is lost, limiting the model's capacity to generalize across the full spectrum of inhibitor activity.

More reliable indicators of model's performance are provided by the *True Positive Rate* (Definition 10), *True Negative Rate* (Definition 11), *Classification Accuracy* (Definition

[18] Particularly in analytical chemistry, two datasets are typically considered significantly correlated if their r^2 value exceeds 0.90.

12), and *Matthews Correlation Coefficient* (Definition 13), the standard metrics used to evaluate the performance of a ML model. Generally speaking, we define a method A to have "better statistics" than other method B if the computed quantities: *True Positive Rate*, *True Negative Rate*, *Classification Accuracy*, and *Matthews Correlation Coefficient*, for A take higher values than the same quantities for B.[19] Specifically, if this procedure for the method A and B lead to incoherences, i.e. some statistics of A are better than B while other are worst, we take the Matthews Correlation Factor as the most reliable statistical outcome.[20]

5.4 Azophotoswitches Analysis

Figure (6) shows the predicted IC_{50} values of azophotoswitch compounds as a function of their binding free energy ($\Delta G_{binding}$), as provided by the MolBioMed research group [8]. Detailed prediction results can be found in Appendix (C), specifically in Tables (9) and 10. The training conditions for the Random Forest (RF) model, along with a comprehensive statistical analysis of its predictions, are presented in Table (11).

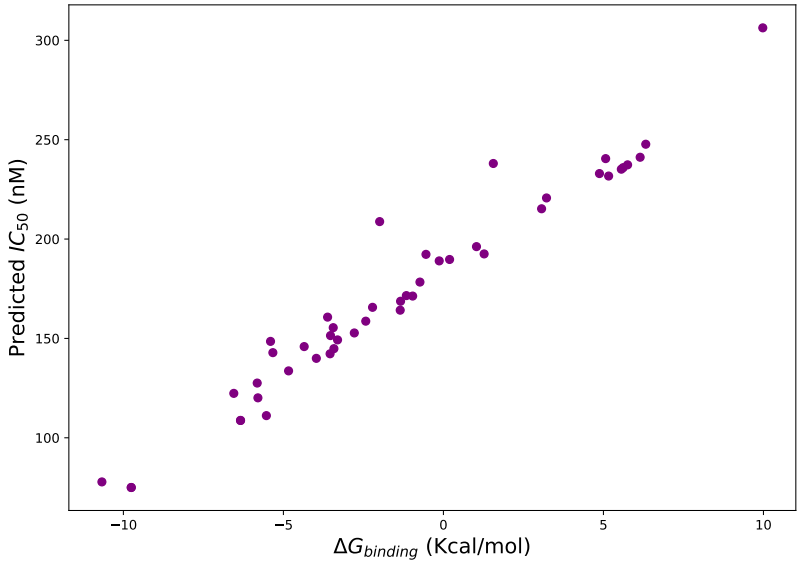


Figure 6: Predicted IC_{50} values for azophotoswitches from Appendix B. The x-axis represents $\Delta G_{binding}$, and the y-axis represents the predicted IC_{50}.

Furthermore, the performance metrics of the RF model for this predictions are summarized in Table (1) that yield the results of Tables (9) and (10) from the appendix (C)

[19] Since all quantities are highly bounded by 1 we might also affirm, the closer to 1 the statistics values are the more reliable the method is.

[20] This is the usual procedure in the analysis of ML predictions[13].

Tables of results. All key statistical indicators exceed the 0.90 threshold. Notably, the model achieves a *True Negative Rate* of 0.99 (Definition (11)), indicating that it correctly identifies non-active inhibitors 99% of the time. The *Classification Accuracy* (Definition (12)) and *Matthews Correlation Coefficient* both approach 1, reinforcing the reliability of this particular RF model in distinguishing between active and inactive azophotoswitches.

Table 1: Conditions and statistics for the machine's learning models in the computation of the IC_{50} from Tables 9 and 10.

Metric	**Value**
Mean Squared Error	1202.70
r^2	0.90
True Positive	994
False Positive	8
True Negative	1015
False Negative	101
True Positive Rate	0.90
True Negative Rate	0.99
Classification Accuracy	0.95
Matthews Correlation Coefficient	0.90

At this stage, it is important to note that the model is currently unable to distinguish between the active and inactive conformations of a molecule, i.e, between the *cis* and *trans* isomers. This limitation is not due to the model architecture or the way chemical descriptors are computed, but rather to a deficiency in the training data. The ChEMBL database [4] does not explicitly include *cis*/*trans* comparable molecular configurations. In other words, if cis/trans isomerization appears we can only find one of the two conformations in the database, typically the one with higher inhibition potential. Consequently, the model cannot learn to differentiate them since this distinction is not encoded in the dataset. As a result, the model provides IC_{50} predictions corresponding to the conformation it has implicitly learned, the more active configuration, which in our context is the *cis* form.

Nevertheless, our hypothesis is supported. In Figure 6, we observe a clear correlation between the predicted IC_{50} values and the corresponding binding free energies ($\Delta G_{binding}$), consistent with our initial intuition that molecules with more negative free energy values tend to exhibit lower IC_{50} values, and are, therefore, more active inhibitors.

The broader hypothesis that a combination of chemical descriptors exists which correlates directly with the inhibitory power of a molecule toward a specific protein also appears to be valid. However, identifying this combination manually remains a significant challenge for chemists. At first glance, one might believe that since this information is encoded within the RF models, analyzing or "decrypting" the models could reveal valuable biochemical insights into the binding behavior of COX-2,[21] or similar targets.

Yet this decryption process is computationally infeasible. The internal logic of a RF model is akin to verifying a password. It can efficiently determine whether an input is valid (e.g., predict an IC_{50} from a molecule's descriptors or checking a username and a password), but reverse-engineering (e.g., determining which chemical descriptors most strongly influence inhibition or cracking a user password) is significantly more complex. In essence, it is easy for the model to make predictions, but very difficult to extract interpretable rules or patterns from its internal structure.

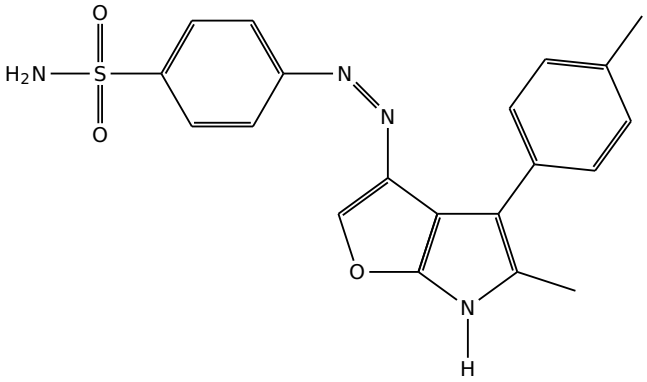
One might consider a brute-force approach, varying one descriptor at a time while holding all others constant, and observing the effect on the predicted IC_{50}. However, this strategy quickly runs into chemical contradictions and incoherences. For example, altering the average bond length of a molecule could make it unrealistically large or physically unstable, resulting in a structure that is chemically invalid although this hypothetic molecule might perfectly fit in the COX-2 binding site. Similarly, descriptors that are inherently discrete, such as the number of carbon atoms, could yield nonsensical results. A model might indicate that 25.5 carbon atoms yields optimal inhibition, which is clearly an unphysical outcome. In addition, arbitrarily generated sets of descriptor values may not correspond to any real molecule, making the predictions scientifically meaningless.

It is important to emphasize that these limitations do not undermine the value of AI in computational chemistry. Rather, they highlight the boundaries of current methods and the importance of careful dataset design, descriptor selection, model interpretation and understanding the machine's workflow.

[21] Or any other protein, since the method is designed to be generalizable to any protein entry available in the ChEMBL database [4].

5.5 Docking

Following the prediction phase, molecular docking is performed by the MolBioMed research group [8] on one of the most promising azophotoswitch candidates. Specifically, we select the molecule labeled as "*13.7*" in appendix B *Tables of azophotoswitches*, which is depicted in Scheme(4). This molecule exhibited a predicted $\Delta G_{binding}$ of -10.6 kcal/mol and a IC_{50} prediction of 77 nM, i.e. yielding to a lower IC_{50} prediction than the experimental Celecoxib's IC_{50} (120 nM). An image of the final frame from the docking simulation is shown in Figure (7).



Scheme 4: Template for Celecoxib azo-derivatives highly active molecule.

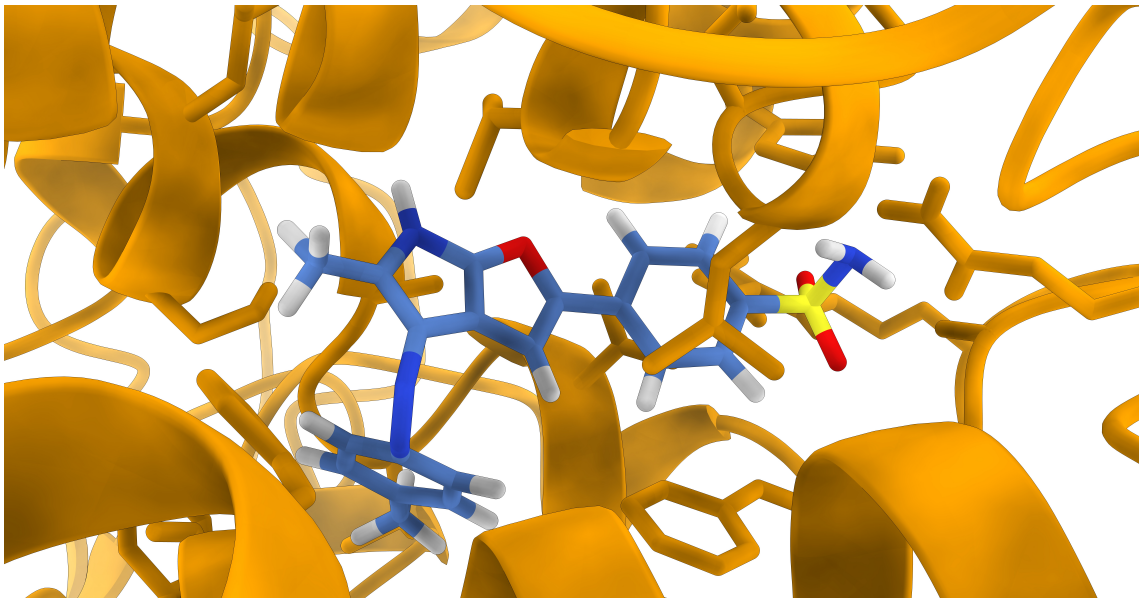


Figure 7: Structural ilustration of the molecule identified as 13.7 and COX-2 docking generated by the ChimeraX software[15].

We observe that the azophotoswitch in the cis- configuration fits well within the COX-2 binding cavity, adopting a conformation similar to that of its counterpart, the Celecoxib

(Definition 3). Although the trans- isomer can also occupy the binding site with minor steric hindrance, thanks to the large size of the COX-2 cavity, which can accommodate various molecules, it is unlikely to access the site in practice due to stereochemical constraints. In particular, the critical differences between the cis- and trans- configurations manifest during the entry into the protein cavity, rather than within the cavity itself. Specifically, the extended structure of the trans- isomer causes steric clashes with the outer regions of COX-2, thereby preventing it from entering the binding pocket.

It is also important to note that the molecule, in both cis/trans isomeric forms, could potentially interact with unintended proteins. Such off-target effects could lead to significant side effects if the compound is administered as a drug. Therefore, further investigation is needed to evaluate its selectivity and pharmacological safety.

6 Conclusions

In this work, we have demonstrated the potential of Artificial Intelligence in the fields of theoretical and computational chemistry. We developed a fully functional machine learning software capable of predicting IC_{50} values for any protein listed in the ChEMBL database. Moreover, the software automatically computes performance statistics for each ML model based on a set of hard-coded parameters that define the training dataset. This flexibility enables users to experiment with different variable configurations to optimize predictive performance.

Despite their usefulness, the models exhibit several limitations. They are entirely dependent on the training dataset, therefore, any lack of information or approximations in the data can lead to systematic errors in their predictions. Furthermore, the models are inherently unable to generate new knowledge. For instance, the distinction between cis and trans conformations is not represented in the ChEMBL COX-2 database. As a result, our models are unable to provide different IC_{50} predictions for these two configurations. Nonetheless, we have demonstrated that the predictions are reliable within a certain threshold, as determined by the statistical analysis of each case.

In addition, we obtained results that reinforce our hypothesis: *There exists a combination (or multiple combinations) of chemical descriptors that are directly related to the inhibition potential of a protein*. Since we observe that the AI is capable of effectively predicting IC_{50} values for certain molecules, we infer that a correlation must exist between the chemical descriptors and the IC_{50}. Naturally, this correlation is complex, as there is no direct relationship between any single chemical descriptor and the IC_{50}, as shown in the results from Section 5.2. Instead, the correlation exhibits a structure far more intricate than a simple linear relationship between descriptor and IC_{50}. This complexity can be tackled using AI algorithms, which are specifically designed to handle such dependencies, although extracting explicit chemical insights from these models remains a challenge due to the nature of their internal architectures, as discussed in Section 5.4.

Additionally, our results confirm the hypothesized correlation between predicted COX-2's IC_{50} values and the experimentally or computationally derived binding free energy ($\Delta G_{binding}$). This relationship aligns with our chemical intuition and reinforces the validity of our methodology. To further validate our approach, we performed a molecular docking study of one of the most promising azophotoswitch candidates. The docking results were

consistent with model predictions, providing additional support for the application of AI techniques in rational drug design.

7 Bibliography

(1) Kase, S.; Saito, W.; Ohno, S.; Ishida, S. *Retina* **2010**, *30*, 719–723, DOI: 10.1097/iae.0b013e3181c59698.

(2) National Cancer Institute Definition of COX-2 - NCI Dictionary of Cancer Terms, Accessed: 2024-10-09, 2024.

(3) Davies, N. M.; Jamali, F. *Pharmacology & Therapeutics* **2000**, *89*, 133–155, DOI: 10.1016/S0163-7258(00)00100-0.

(4) Zdrazil, B. et al. *Nucleic Acids Research* **2024**, *52*, D1180–D1192, DOI: 10.1093/nar/gkad1004.

(5) Mauri, A. In *Ecotoxicological QSARs*, Roy, K., Ed.; Springer US: New York, NY, 2020, pp 801–820, DOI: 10.1007/978-1-0716-0150-1_32.

(6) Mauri, A.; Bertola, M. *International Journal of Molecular Sciences* **2022**, *23*, DOI: 10.3390/ijms232112882.

(7) Breiman, L. *Machine Learning* **2001**, *45*, 5–32, DOI: 10.1023/A:1010933404324.

(8) Group, M. R. MolBioMed | Molecular Biomedicine – Computational Modelling for Inflammation Research, https://webs.uab.cat/molbiomed/en/, Accessed: 2025-05-28, 2025.

(9) Baek, M.; et al. *Signal Transduction and Targeted Therapy* **2023**, *8*, 1–10, DOI: 10.1038/s41392-023-01381-z.

(10) Singh, S.; Kumar, R.; Payra, S.; Singh, S. K. *Cureus* **2023**, *15*, e44359, DOI: 10.7759/cureus.44359.

(11) Bale, J. B. et al. *Nature* **2016**, *500*, 705–710, DOI: 10.1038/nature18010.

(12) Jumper, J. et al. *Nature* **2021**, *596*, 583–589, DOI: 10.1038/s41586-021-03819-2.

(13) Khan, H. A.; Jabeen, I. *Frontiers in Pharmacology* **2022**, *13*, DOI: 10.3389/fphar.2022.825741.

(14) Hashemi Goradel, N.; Najafi, M.; Salehi, E.; Farhood, B.; Mortezaee, K. *Journal of Cellular Physiology* **2019**, *234*, 5683–5699, DOI: 10.1002/jcp.27411.

(15) Meng, E. C.; Goddard, T. D.; Pettersen, E. F.; Couch, G. S.; Pearson, Z. J.; Morris, J. H.; Ferrin, T. E. *Protein Science* **2023**, *32*, e4792, DOI: 10.1002/pro.4792.

(16) Swinney, D. C. In Macor, J. E., Ed.; Annual Reports in Medicinal Chemistry, Vol. 46; Academic Press: 2011, pp 301–317, DOI: https://doi.org/10.1016/B978-0-12-386009-5.00009-6.

(17) Gasteiger, J.; Engel, T., *Molecular Descriptors for Chemoinformatics*; Wiley-VCH: Weinheim, Germany, 2003.

(18) Castaneiras Morales, S. AI Application for Azophotoswitches Optimization with Pharmacological Interest, https://github.com/SirSergi0/Repository---AI-application-for-azophotoswitches-optimization-with-pharmacological-interest, 2025.

(19) Castañeiras Morales, S. **2025**, DOI: 10.5281/zenodo.15546442.

(20) Reitz, K.; Chalasani, A. Requests: HTTP for Humans, https://pypi.org/project/requests/, version 2.31.0, Python package, 2023.

(21) McKinney, W. pandas: A Foundational Python Library for Data Analysis, Version 1.5.3, 2023.

(22) Pedregosa, F. et al. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(23) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *Journal of Molecular Biology* **1997**, *267*, 727–748, DOI: 10.1006/jmbi.1996.0897.

(24) Weininger, D. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36, DOI: 10.1021/ci00057a005.

8 Acknowledgements

I am grateful to Dr. Josep Maria Lluch, Dr. Miquel Moreno Ferrer and Dra. Àngels González Lafont for allowing me to be one of thei pupils and for presenting this project. I am also deeply grateful to PhD Àlex Pérez Sánchez and PhD Pedro Martínez Zaragoza for sharing their ideas and fruitful conversations.

I would like to acknowledge my mother and sister for their unconditional support, and my father for the values he taught me. Additionally, I would like to thank my friends Francisco Montaño, David Muñoz and Manel Martin for listening to my monologues about my excitement for the COX-2 and azophotoswitches.

Molecular graphics and analyses were performed with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases.

Appendices

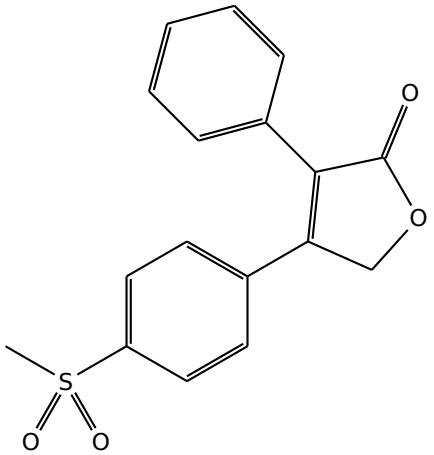
A Relevant definitions

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[16]. 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. *Molecular descriptor: "A molecular descriptor is the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment."[17]*

Definition 3. *Celecoxib: [22] drug known to be a selective COX-2 inhibitor, see Scheme (3). Its IC_{50} value is 120 nM.*

Definition 4. *Rofecoxib: [23] drug known to be a selective COX-2 inhibitor, see Scheme (5). Its IC_{50} value is 180 nM.*



Scheme 5: Chemical graph of Rofecoxib.

[22] UPAC name: 4-[5-(4-Methylphenyl)-3-(trifluoromethyl)pyrazol-1-yl]benzenesulfonamide
[23] UPAC name: 4-(4-methylsulfonylphenyl)-3-phenyl-5H-furan-2-one

Definition 5. *Cheng Prusoff equation: standard equation used for the experimental computation of the IC50.*

$$K_i = \frac{IC_{50}}{1 + \frac{[S]}{K_m}}$$

where K_i is the binding affinity, $[S]$ is the substrate concentration, K_m is the Michaelis constant and IC_{50} the half maximal inhibitory concentration.

Definition 6. *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,*

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

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 a linear regression. More precisely if $x > 0$, "y" generally[24] increases when "x" increases, as well as if $x < 0$, "y" decreases when "x" increases.

Definition 7. *Mean Squared Error: Given set of pairs of data $\{(x_i, y_i)\}_{i=1}^n$ the Mean Squared Error is the quantity defined as,*

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2 \tag{2}$$

The name of this quantity is self-descriptive, since $X_i - Y_i$ is the error associated to the i-th prediction and the MSE is the mean of the squares of this error.

Definition 8. *A decision tree is a classification algorithm based on a series of ordered "`if`" statements. The algorithm begins at the top of the tree, where a question is posed to the data. Depending on the answer, the data follows different branches, each corresponding to a subsequent question. This process is repeated at each node until the data reaches the bottom of the tree, where the path it has followed determines the classification of the given data.*

This kind of algorithm is vividly present in the chemical landscape, for instance in the spectroscopy realm the determination of a molecule's symmetry group is provided by a

[24] 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

decision tree depicted in Figure (8).

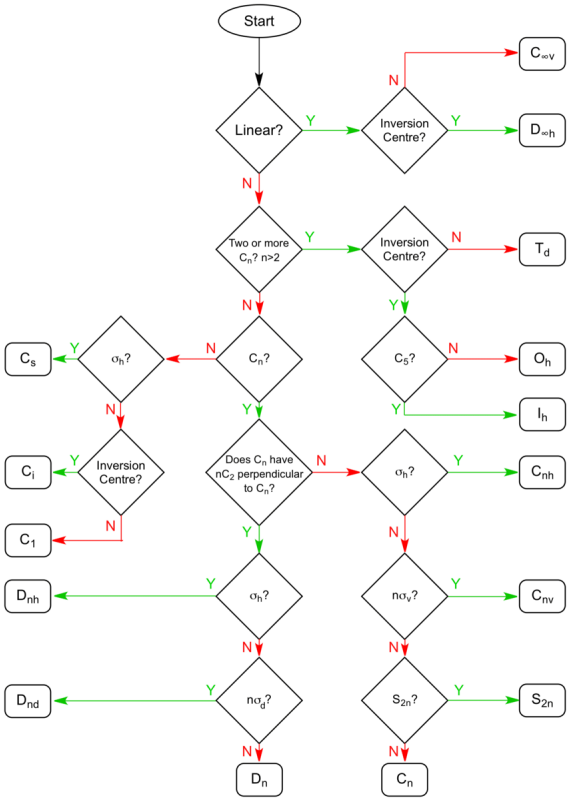


Figure 8: Decision tree for determining the point group of a molecule

Definition 9. *Canonical Simplified Molecular Input Line Entry System (SMILES): A text-based notation that encodes a molecule's structure as a linear string, containing all the essential information needed to reconstruct its 3D representation [24]. SMILES files typically use the extensions* `.smi` *or* `.smiles`*.*

Definition 10. *True Positive Rate: quantity related to a Machine Learning Model's sensitivity defined as:*

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

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

Definition 11. *True Negative Rate: quantity related to a Machine Learning Model's specificity defined as:*

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

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

Definition 12. *Classification Accuracy: quantity related to a Machine Learning Model's efectiveness defined as:*

$$\frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

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

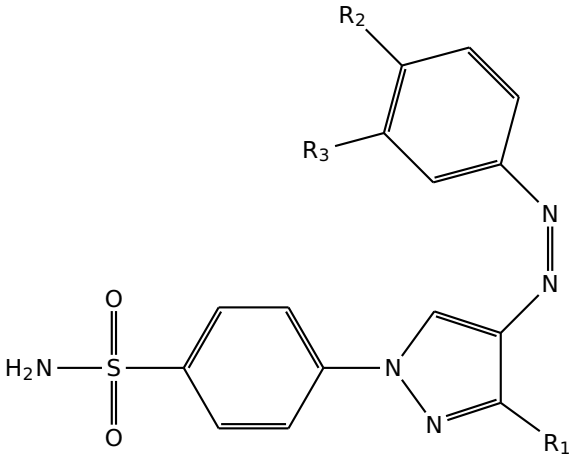
Definition 13. *Matthews Correlation Coefficient: quantity related to a Machine Learning Model's prediction capacity defined as:*

$$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{6}$$

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 the 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.

Definition 14. *Pandas dataframes: type of data structure in Python characterized by a two-dimensional, size-mutable, and heterogeneous tabular data structure. The class dataframes class is provided by the Pandas library[21]. It resambles to a table in a database or an Excel spreadsheet.*

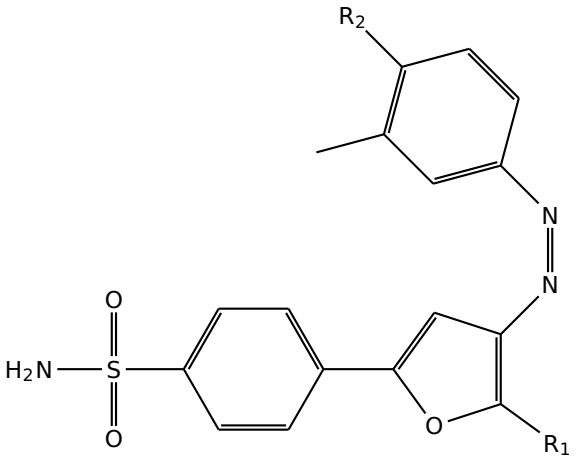
B Tables of azophotoswitches



Scheme 6: Template for Celecoxib's azo-derivates with pyrazole as heterocycle.

Table 2: Table of potential photoswitches derivated from Celecoxib's azo-derivates with pyrazole as heterocycle

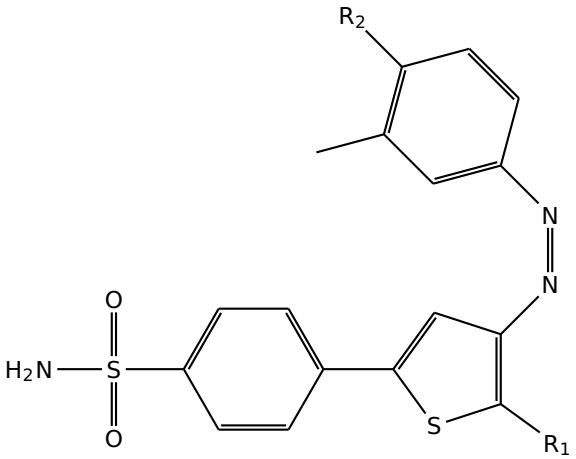
Identifier	R_1	R_2	R_3
6.1	CF_3	CH_2CH_3	H
6.2	CF_3	CH_2CH_3	F
6.3	CF_3	CH_3	F
6.4	CF_3	OCH_3	H
6.5	CF_3	OCH_3	F
6.6	CF_3	CH_3	H
6.7	H	CH_3	H
6.8	F	CH_3	H
6.9	Cl	CH_3	H
6.10	Br	CH_3	H
6.11	CH_3	CH_3	H
6.12	H	CH_3	F
6.13	F	CH_3	F
6.14	Cl	CH_3	F
6.15	Br	CH_3	F
6.16	CH_3	CH_3	F



Scheme 7: Template for Celecoxib azo-derivatives with furan as a heterocycle.

Table 3: Table of potential photoswitches derivated from Celecoxib's azo-derivates with furan as heterocycle.

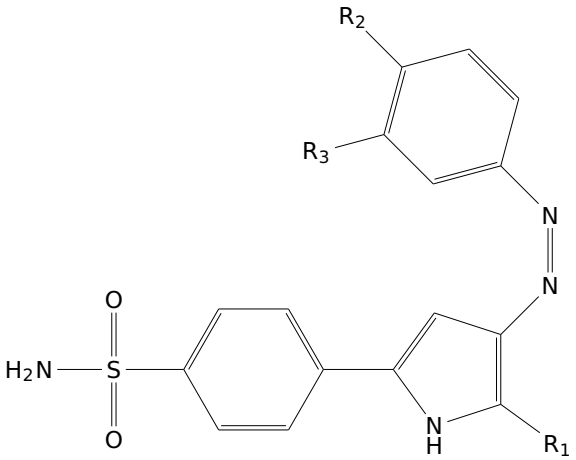
Identifier	R_1	R_2
7.1	CF_3	H
7.2	H	H
7.3	F	H
7.4	Cl	H
7.5	Br	H
7.6	CH_3	H
7.7	CF_3	F
7.8	H	F
7.9	F	F
7.10	Cl	F
7.11	Br	F
7.12	CH_3	F



Scheme 8: Template for Celecoxib azo-derivatives with thiophene as a heterocycle.

Table 4: Table of potential photoswitches derivated from Celecoxib's azo-derivates with thiophene as heterocycle.

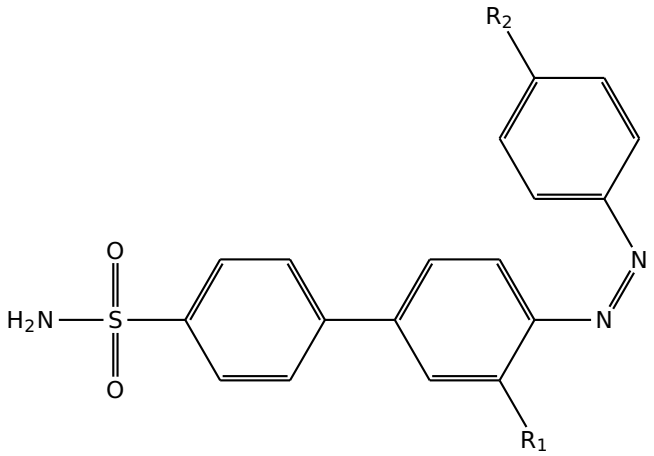
Identifier	R_1	R_2
8.1	F	H
8.2	H	F
8.3	Cl	F



Scheme 9: Template for Celecoxib azo-derivatives with pyrrole as a heterocycle.

Table 5: Table of potential photoswitches derivated from Celecoxib's azo-derivates with pyrrole as heterocycle.

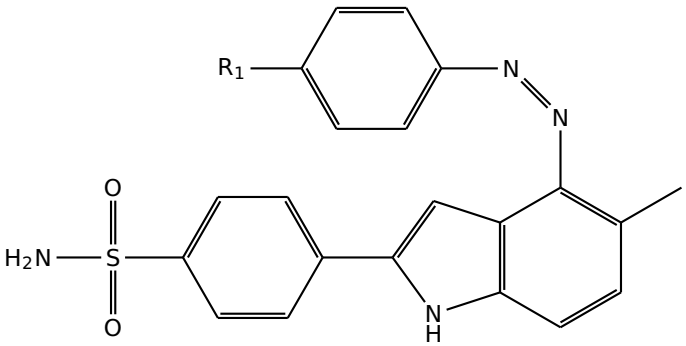
Identifier	R_1	R_2	R_3
9.1	CF_3	CH_3	H
9.2	Cl	CH_3	F



Scheme 10: Template for Celecoxib azo-derivatives with benzene in place of the original heterocycle.

Table 6: Table of potential photoswitches derivated from Celecoxib azo-derivatives with benzene in place of the original heterocycle.

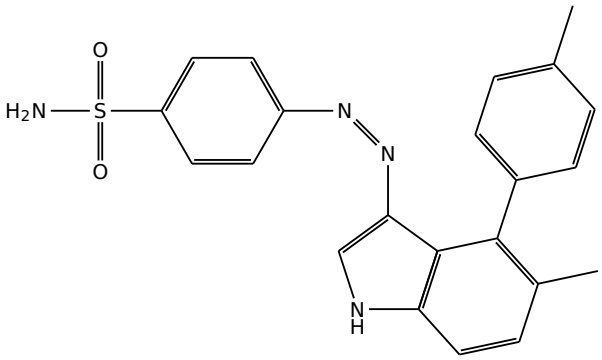
Identifier	R_1	R_2
10.1	CF_3	CH_2CH_3
10.2	CF_3	NCH_3COCH_3
10.3	CF_3	$NHCH_3$
10.4	CF_3	OCH_3
10.5	Cl	CH_3



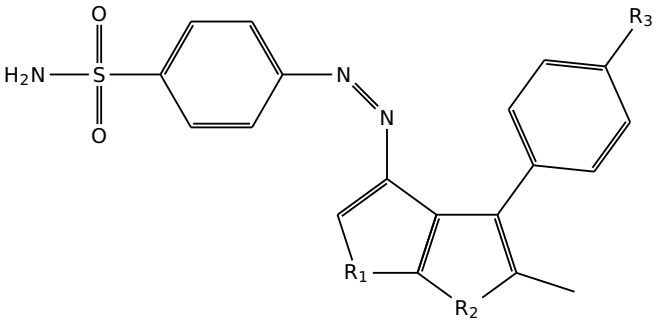
Scheme 11: Template for Celecoxib azo-derivatives with indole ring as a heterocycle.

Table 7: Table of potential photoswitches derivated from Celecoxib azo-derivatives with indole ring as a heterocycle.

Identifier	R_1
11.1	H
11.2	F



Scheme 12: Template for Celecoxib azo-derivatives with indole ring as a heterocycle.



Scheme 13: Template for Celecoxib azo-derivatives with two rings of five members joint as a heterocycle.

Table 8: Table of potential photoswitches derivated from Celecoxib azo-derivatives with two rings of five members joint as a heterocycle.

Identifier	R_1	R_2	R_3
13.1	NH	NH	H
13.2	NH	O	H
13.3	O	NH	H
13.4	O	O	H
13.5	NH	NH	CH_3
13.6	NH	O	CH_3
13.7	O	NH	CH_3
13.8	O	O	CH_3

C Tables of results

Notice some molecules data do not appear in Table 9. This is due to inconclusive results of the $\Delta G_{binding}$ computations.

Table 9: Results for the $\Delta G_{binding}$ and IC_{50}. The conditions and statistics under which this computations have been done are stored in Table (11).

Type	Identifier	R_1	R_2	R_3	Δ $G_{binding}$ (kcal/mol)	Predicted IC_{50} (nM)
Pyrazole	6.1	CF_3	CH_2CH_3	H	-4.35	145
Pyrazole	6.2	CF_3	CH_2CH_3	F	-3.62	160
Pyrazole	6.3	CF_3	CH_3	F	-0.54	192
Pyrazole	6.4	CF_3	OCH_3	H	9.98	306
Pyrazole	6.5	CF_3	OCH_3	F	1.56	238
Pyrazole	6.6	CF_3	CH_3	H	-5.32	142
Pyrazole	6.8	F	CH_3	H	5.07	240
Pyrazole	6.9	Cl	CH_3	H	-1.16	171
Pyrazole	6.10	Br	CH_3	H	6.32	247
Pyrazole	6.11	CH_3	CH_3	H	3.07	215
Pyrazole	6.12	H	CH_3	F	1.03	196
Pyrazole	6.13	F	CH_3	F	3.22	220
Pyrazole	6.14	Cl	CH_3	F	5.16	231
Pyrazole	6.15	Br	CH_3	F	5.61	235
Pyrazole	6.16	CH_3	CH_3	F	-3.30	149
Furan	7.1	CF_3	H		-0.96	171
Furan	7.2	H	H		0.19	189
Furan	7.3	F	H		-4.84	133
Furan	7.4	Cl	H		-2.42	158
Furan	7.5	Br	H		-3.42	144
Furan	7.6	CH_3	H		-3.97	139
Furan	7.7	CF_3	F		-5.40	148
Furan	7.8	H	F		-6.55	122
Furan	7.9	F	F		-0.13	189

Table 10: Results for the $\Delta G_{binding}$ and IC_{50}. The conditions and statistics under which this computations have been done are stored in Table (11).

Type	Identifier	R_1	R_2	R_3	$\Delta G_{binding}$ (kcal/mol)	Predicted IC_{50} (nM)
Furan	7.10	Cl	F		-5.81	127
Furan	7.11	Br	F		6.15	241
Thiophene	8.1	F	H		1.27	192
Thiophene	8.2	H	F		-2.78	152
Thiophene	8.3	Cl	F		-5.79	120
Pyrrole	9.1	CF_3	CH_3	H	-3.53	151
Pyrrole	9.2	Cl	CH_3	F	-1.35	164
Benzene	10.1	CF_3	CH_2CH_3		-1.34	168
Benzene	10.2	CF_3	NCH_3COCH_3		-1.99	208
Benzene	10.3	CF_3	$NHCH_3$		4.88	232
Benzene	10.4	CF_3	OCH_3		-3.44	155
Indole	11.1	H			-9.75	74
Indole	11.2	F			-6.34	108
Indole	12				-5.53	111
TwoRings	13.2	NH	O	H	-2.21	165
TwoRings	13.3	O	NH	H	-0.73	178
TwoRings	13.6	NH	O	CH_3	5.59	235
TwoRings	13.7	O	NH	CH_3	-10.67	77
TwoRings	13.8	O	O	CH_3	-3.54	142

Table 11: Conditions and statistics for the machine's learning models for the computation of the IC_{50} from Tables (9) and (10).

Metric	Value
NumberOfTrees	250
ErasedPercentatge	0.0%
SplittingProportion	10.0% for testing
minimumCorrelationFactor	0.0
Number of descriptors	3040
Mean Squared Error	1202.70
r^2	0.90
True Positive	994
False Positive	8
True Negative	1015
False Negative	101
True Positive Rate	0.90
True Negative Rate	0.99
ClassificationAcuracy	0.95
MatthewsCorrelationFactor	0.90