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 Gonzalez 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 *prostaglandin-endoperoxide sinthasa 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*, 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, els models es fan servir per predir l'IC_{50} de 50 *azophotoswitches* dels quals es tenen dades sobre l'energia lliure d'acoblament proporcionades pel Departament de Química Física de la UAB [8]. L'anàlisi estadístic de les prediccions reflecteix una clara correlació entre ambdues quantitats, fet que reforça 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 *prostaglandin-endoperoxide sinthasa 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*, ya que dispone de una gran cantidad de datos para entrenar los modelos y contrastar los resultados.

El objetivo principal del proyecto es el desarrollo de un software generador de IAs capaces de predecir la concentración de inhibición al 50% (IC_{50}) para la COX-2[2] con la máxima precisión posible. Para ello, se extraen todos los datos de moléculas conocidas con un potencial de inhibición establecido para la COX-2. Tras un filtrado configurable por el usuario, se calculan 5.900 descriptores químicos para cada una de las entradas con el software AlvaDesk [5][6]. Seguidamente, una parte de los datos se utiliza para entrenar modelos de *Random Forest* (RF) [7], mientras que el resto se reserva para validar la precisión de las predicciones.

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 Departamento de Química Física de la UAB [8]. El análisis estadístico de las predicciones refleja una clara correlación entre ambas cantidades, lo que refuerza la hipótesis del proyecto.

[2] En realidad, el software funciona para cualquier proteína con entrada en la base de datos de ChEMBL [4], aunque el objeto de estudio es 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* 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, the models are used to predict the IC_{50} of 50 *azophotoswitches*, for which data on the free binding energy have been provided by the Physical Chemistry Unit at UAB [8]. The statistical analysis of the predictions shows a clear correlation between both quantities, which supports the project's hypothesis.

[3] In reality, 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.
TN	True Negative.
TP	True Posive.

List of Figures

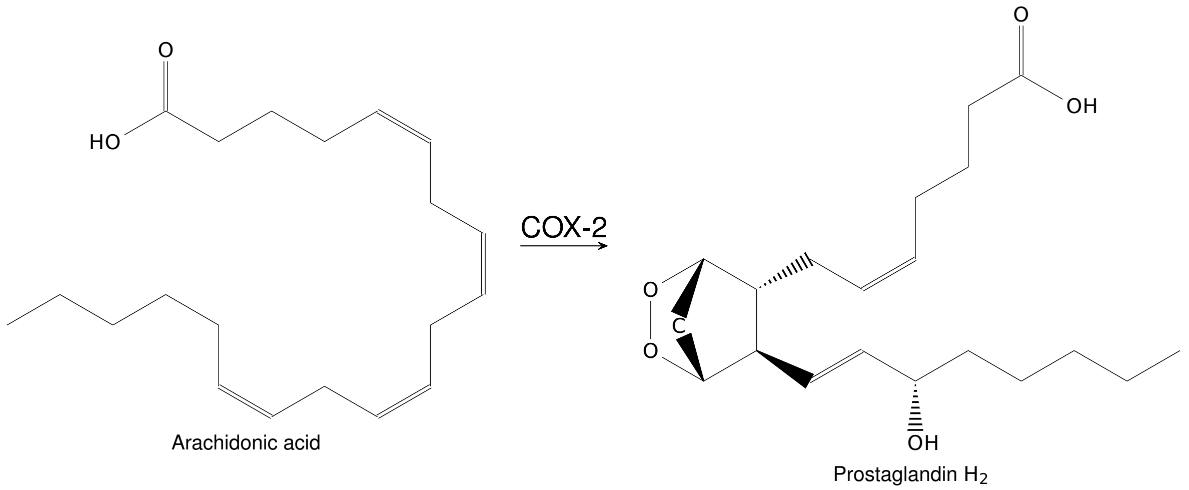
List of Tables

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 Cyclooxygenase-2 (COX-2), 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 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 tumor 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 contrindicative, and the search for a more effective and safer treatment remains ongoing. Plently 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 responsable 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.



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

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].

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

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[15] 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 of each molecule in the database are then computed using AlvaDesk[5][6] software. Around 5000 descriptors are calculated[16], 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.

The AI algorithm used is a in this study is a ML model known as the Random Forest (RF) algorithm[7][8], 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.

[5] Usually this parameter is computed throughout the Cheng Prusoff Equation (Def. (4)) with experimental data.

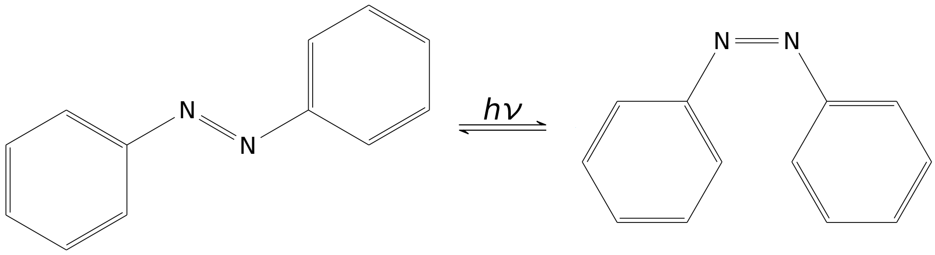
[6] The accuracy of the ML models is discussed on the Results and Discussion section (5).

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)[17] 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 an azophotoswitch if it contains an azo bond (N=N) that is sensitive to configurational transformations upon photo-excitation. For instance, let us consider the case of (E/Z)-N,1-diphenylmethanimine,



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

The photo-induced E/Z 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.[7] [8]

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 2) and Rofecoxib (Definition 3).[9] At the Physical Chemistry Unit of the UAB [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 ($\Delta G_{\text{binding}}$).[10]

As an application of our ML-based predictive models, we aim to estimate the IC_{50} values of azophotoswitch prototypes.[11] 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.

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

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

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

[9] These compounds are often referred to as coxib drugs due to the common suffix -coxib.

[10] Additional results are provided in Appendix C.

[11] These values are predictions, as no experimental or simulated data are currently available.

medicinal chemistry and bioinformatics.

3 Objectives

4 Methodology

The source code is all stored in the *AI application for azophotoswitches optimization with pharmacological interest* GitHub repository[18].

The target protein's ID is set at *CHEMBL230* corresponding to the COX-2 ID in the ChEMBL database. Utilising *requests* python package[19] 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[20] 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. 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.[12]. 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 [13] 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. Still, some chemical descriptor need from the presence of a certain atom, for instance a chemical descriptor related to presence of a metal in the molecule, or active gorup, for example chemical descriptor related to the presence of a carboxilic acid, or chemical descriptor related triple bound among other possiblities. The computation of this chemical descriptors appear as *null values*. By deleting molecular descriptors with null values 2917 chemical descriptors remain. Here, the Pearson correlation coefficient (Def. (5)) between each chemical descriptor and the IC_{50} value is computed, providing insight into the direct relationship between IC_{50} and the descriptors. This relationship will be discussed in the Results and Discussion Section (5).

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. By doing so, it is possible to compute the classification

[12] Typically the range of the azophotoswitches IC$_{50}$ is arround the Celecoxib's IC$_{50}$, i.e. arround 120 nM

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

accuracy statistics of the model from the cluster association of each prediction. Thus we can compare each prediction of each molecules to experimental data, i.e. the model predicts a molecule to be highly active molecule to a highly active molecule this will be denoted as a True Posive (TP). Similarly we can define a True Negative (TN), False Posive (FP) and False Negative (FN) . This procedure is an standard for evaluating a ML model realiability.

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 realiability of the RF model, following the proportion specified by the hard-coded variable *testSizeProportion*. This procedure is illustrated in Figure (1).

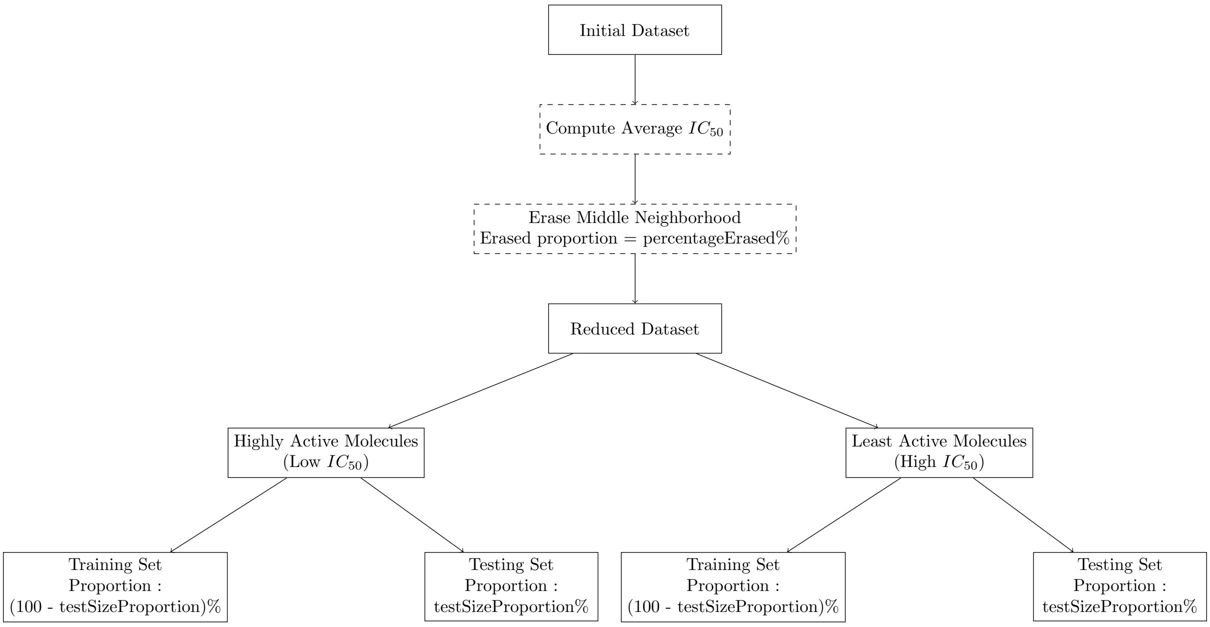


Figure 1: Splitting and processing data's scheme.

Afterward, a Random Forest algorithm is trained using the Sticky Learn [21] 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. Typically the more trees the model is trained with the best accuracy in the predictions, nevertheless the computational cost for each model training also increases with the number of trees. It is key to remark that the predictions' accuracy can not be arbitrarily improved by only increasing the number of trees of the RF model. The predictions' accuracy is bonded to the data the model is trained with, thus if the data does not provide enough information about the insight of the protein, more accurate predictions can not be archived independently of the number of trees of the model. As a general criteria we can affirm that,

$$\uparrow \text{Data}+\uparrow \text{Number of trees} \implies \uparrow \text{Predictions' accuracy}$$
$$\downarrow \text{Data}+\downarrow \text{Number of trees} \implies \downarrow \text{Predictions' accuracy}$$

but more data or more number of trees independently will not lead to better accuracy,

$$\uparrow \text{Number of trees} \not\uparrow \text{Predictions' accuracy}$$
$$\uparrow \text{Data} \not\uparrow \text{Predictions' accuracy}$$

Later, these models are evaluated by predicting the IC_{50} values of the *testing sets*. Using the results, the *True Positive Rate* (Def. 6), *True Negative Rate* (Def. 7), *Classification Accuracy* (Def. 8), and *Matthews Correlation Coefficient* (Def. 9) are computed. Based on these computational results, the variables *percentageErased*, *testSizeProportion*, and *numberOfTrees* are manually adjusted to obtain the best results.

Finally, the RF algorithm with best statistics[14] are used to predict the IC_{50} of the azophotoswitch prototipes classified in Appendix (B). This procedure begins with the computation of the chemical descriptors with AlvaDesk software [5], then the pertinent chemical descriptor are erased in order to match the ones th

[14] The ruling criteria used to classify one model to have "better statistics" than other is discussed in section (5)

5 Results and Discussion

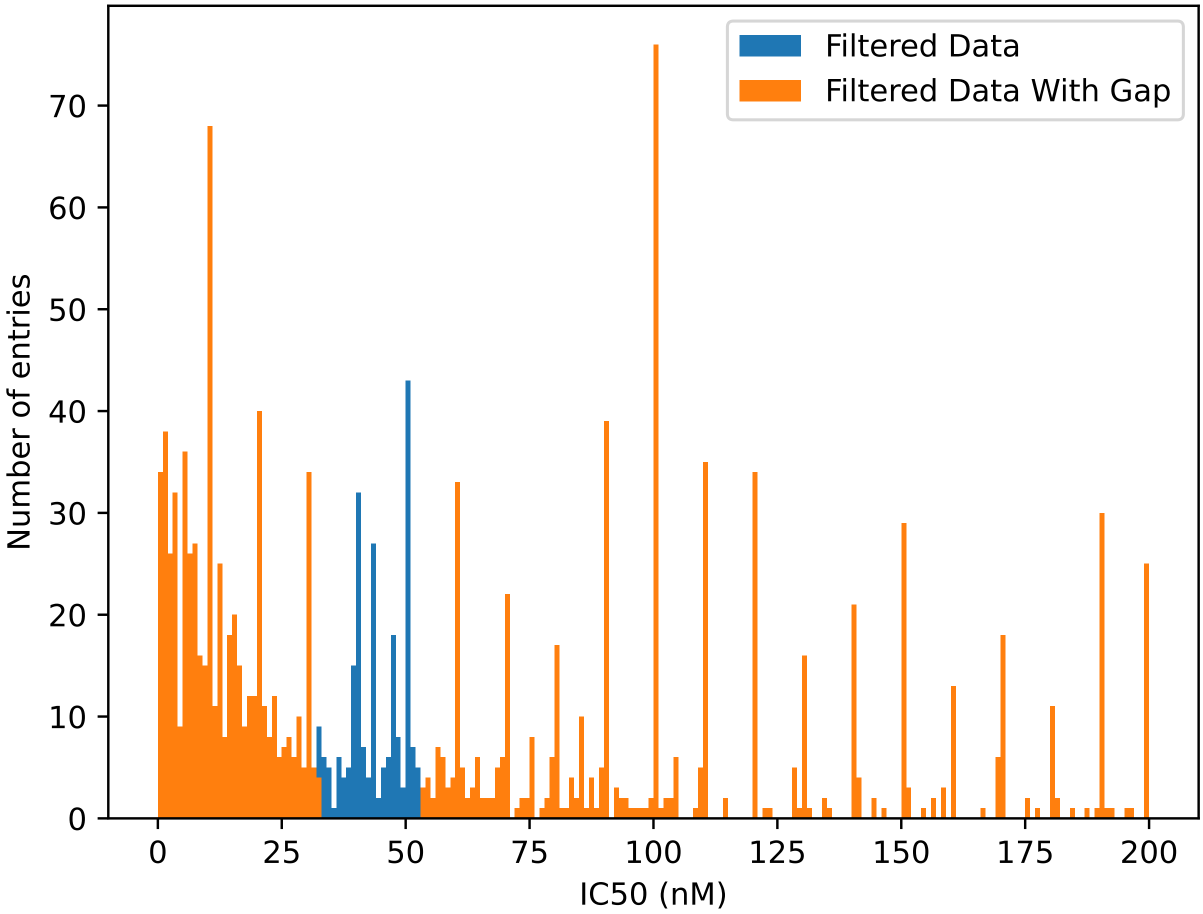


Figure 2: IC_{50} values

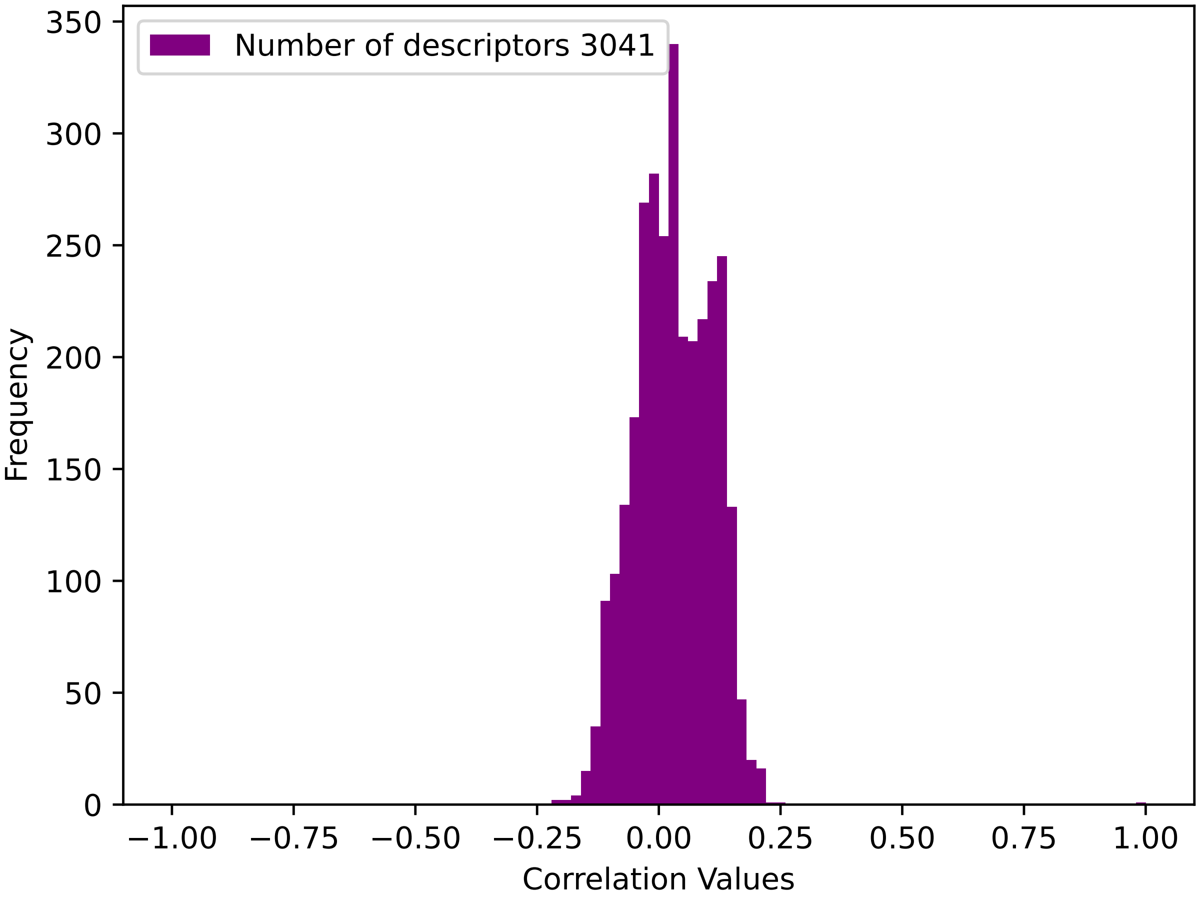


Figure 3: Pearson correlation coefficient for each chemical descriptor.

In Figure (4) we present the results for the predicted IC_{50} in terms of the $\Delta G_{binding}$ provided by the Physical Chemistry Unit of the UAB. The exact results are stored in Appendix C in Tables (9) and (10), additionall the conditions under which the Random Forest (RF) algorithm has been trained and the statistics analysis of its predictions is stored in Table (11).

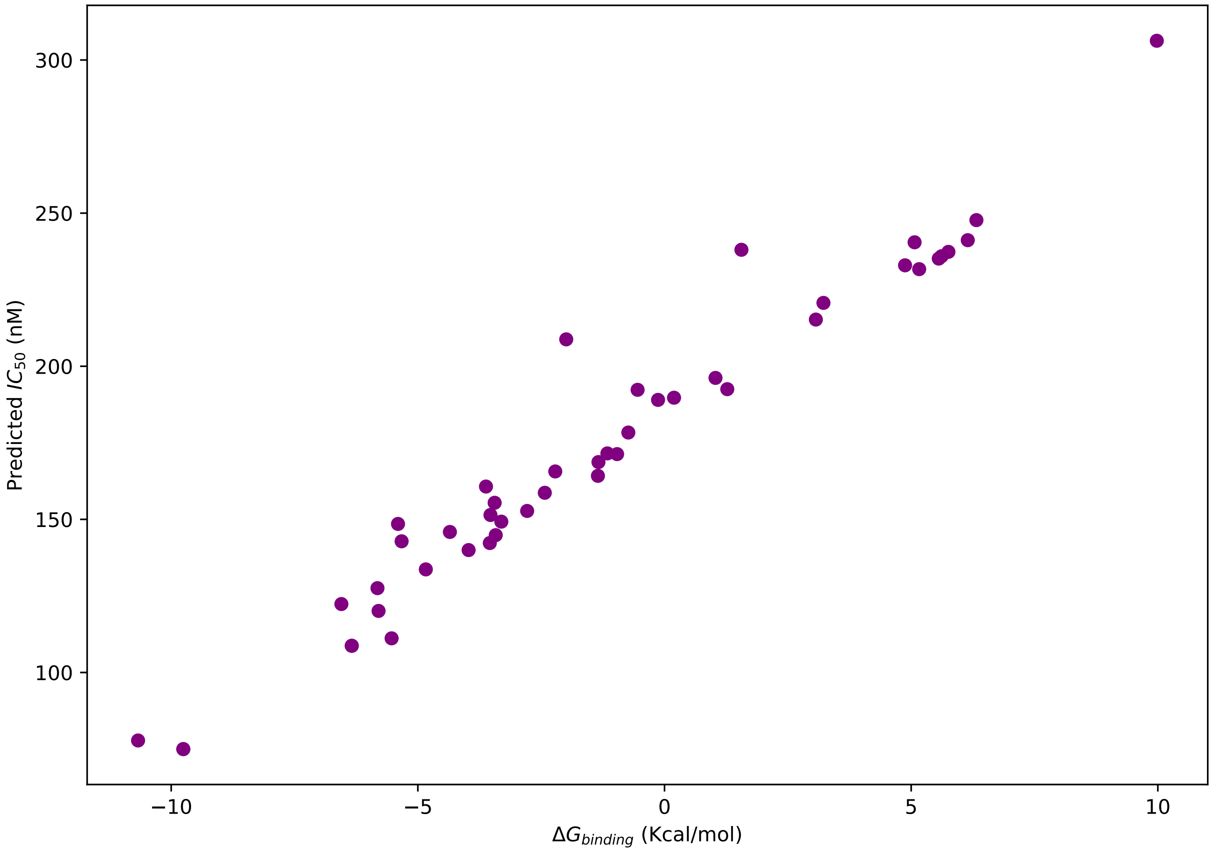


Figure 4: Graphic of the azophotoswitches from Appendix B where the x-axis represents the $\Delta G_{binding}$ and the y-axis the computed IC_{50}.

Furthermore, we may also show the statistics and reliability of this results in Table (1).

Table 1: Statistics for the computation of the results from tables (9) and (10).

Mean Squared Error	1202.70
R-squared	0.89
ClassificationAcuracy	0.94
MatthewsCorrelationFactor	0.87

6 Conclusions

7 Bibliography

(1) Kase, S.; Saito, W.; Ohno, S.; Ishida, S. *Retina* **2010**, *30*, 719–723.

(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.

(4) Zdrazil, B. et al. *Nucleic Acids Research* **2024**, *52*, D1180–D1192.

(5) Mauri, A. In *Ecotoxicological QSARs*, Roy, K., Ed.; Springer US: New York, NY, 2020, pp 801–820.

(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.

(8) Computational Chemistry Department Computational Chemistry Department, Universitat Autònoma de Barcelona, Website of the Computational Chemistry Department, UAB, 2025.

(9) Baek, M.; et al. *Signal Transduction and Targeted Therapy* **2023**, *8*, 1–10.

(10) Singh, S.; Kumar, R.; Payra, S.; Singh, S. K. *Cureus* **2023**, *15*, e44359.

(11) Bale, J. B. et al. *Nature* **2016**, *500*, 705–710.

(12) Jumper, J. et al. *Nature* **2021**, *596*, 583–589.

(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.

(15) Swinney, D. C. In Macor, J. E., Ed.; Annual Reports in Medicinal Chemistry, Vol. 46; Academic Press: 2011, pp 301–317.

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

(17) Dictionary, O. E. Non-Steroidal Anti-Inflammatory Drug, Online.

(18) Morales, S. C. AI Application for Azophotoswitches Optimization with Pharmacological Interest, 2025.

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

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

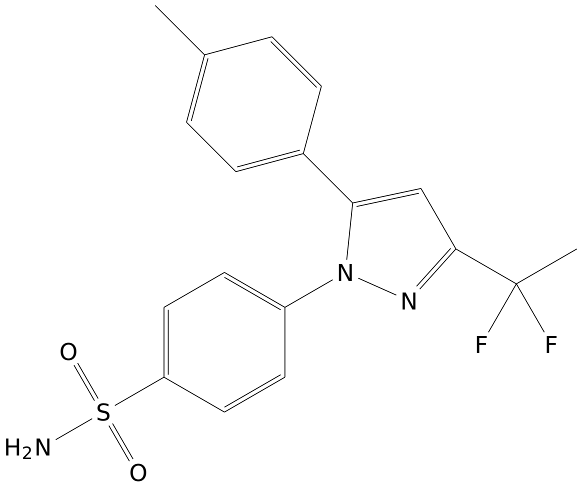
(21) Pedregosa, F. et al. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

A Rellevant 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[15]. 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:* [15] *drug known to be a selective COX-2 inhibitor (currently is not* highly selective *respect to newer drugs), see Scheme (3). It IC_{50} value is 120 nM.*

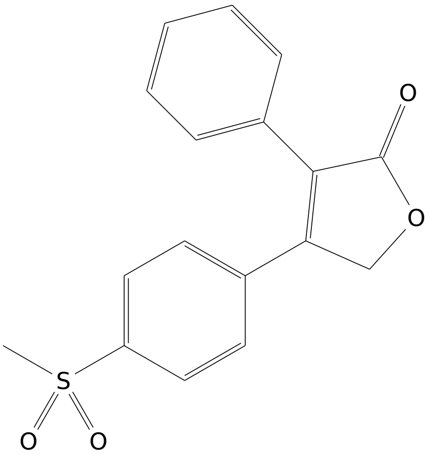
Definition 3. *Rofecoxib:* [16] *drug known to be a selective COX-2 inhibitor, see Scheme (4). It IC_{50} value is 180 nM.*



Scheme 3: Chemical graph of Celecoxib.

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

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



Scheme 4: Chemical graph of Rofecoxib.

Definition 4. *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 5. *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 alinear regression, more precisely if $x > 0$, "y" generally[17] increases when "x" increases, as well as if $x < 0$, "y" decreases when "x" increases.

Definition 6. *True Positive Rate: quantity related to a Machine Learning Model's sensitivity defined as:*

$$\frac{TP}{TP + FN} \tag{2}$$

[17] 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

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

Definition 7. *True Negative Rate: quantity related to a Machine Learning Model's specificity 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 8. *Classification Accuracy: quantity related to a Machine Learning Model's efectiveness defined as:*

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

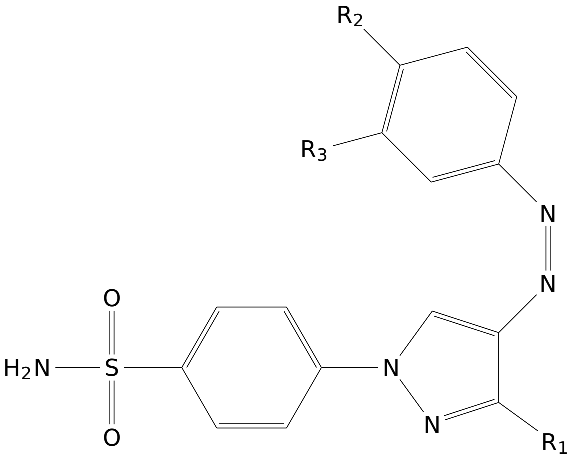
where TP, FP, TN, FN stands for "True Positive", "False Positive", "True Negative" & "False Negative" respectively.

Definition 9. *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 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.

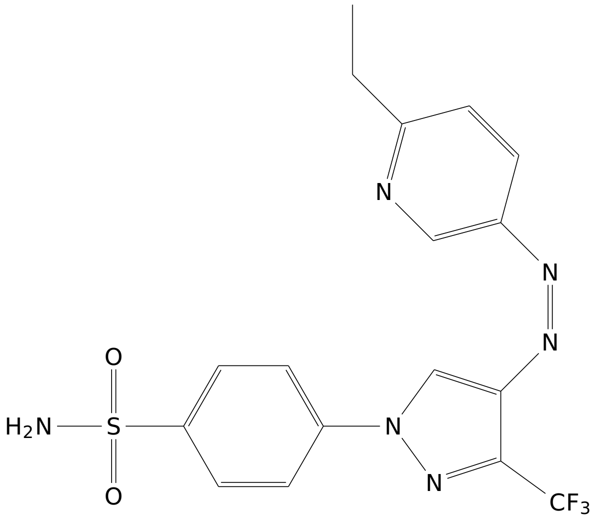
B Tables of azophotoswitches



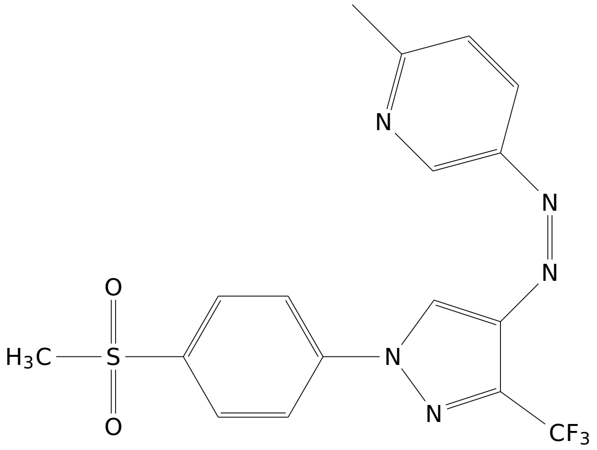
Scheme 5: Template for Celecoxib's azo-derivates with pyrazole as heterocycle.

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

Identifier	R_1	R_2	R_3
5.1	CF_3	CH_2CH_3	H
5.2	CF_3	CH_2CH_3	F
5.3	CF_3	CH_3	F
5.4	CF_3	OCH_3	H
5.5	CF_3	OCH_3	F
5.6	CF_3	CH_3	H
5.7	H	CH_3	H
5.8	F	CH_3	H
5.9	Cl	CH_3	H
5.10	Br	CH_3	H
5.11	CH_3	CH_3	H
5.12	H	CH_3	F
5.13	F	CH_3	F
5.14	Cl	CH_3	F
5.15	Br	CH_3	F
5.16	CH_3	CH_3	F

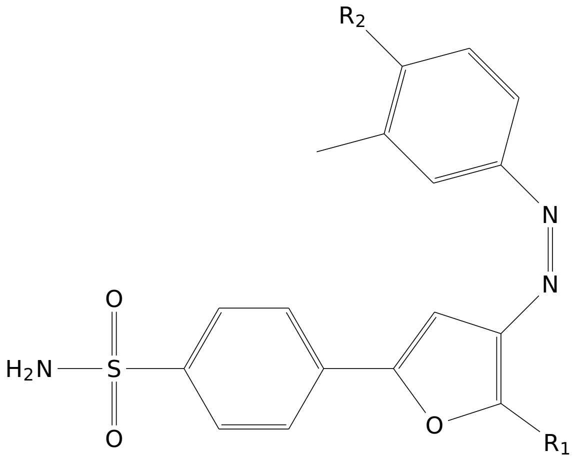


Pyridine derivative



SO_2CH_3 group derivative

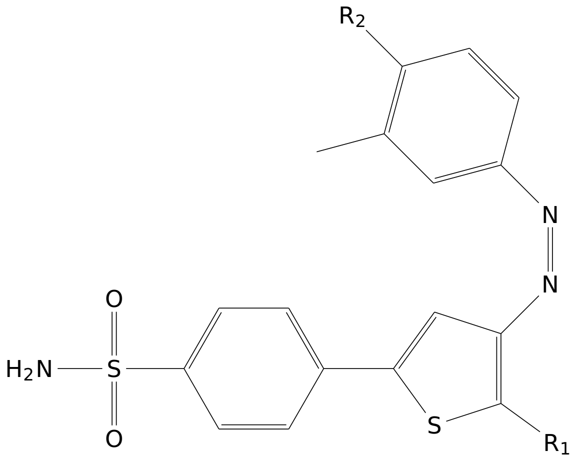
Scheme 6: Scheme for Celecoxib azo-derivatives based on pyridine and SO_2CH_3 groups.



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.

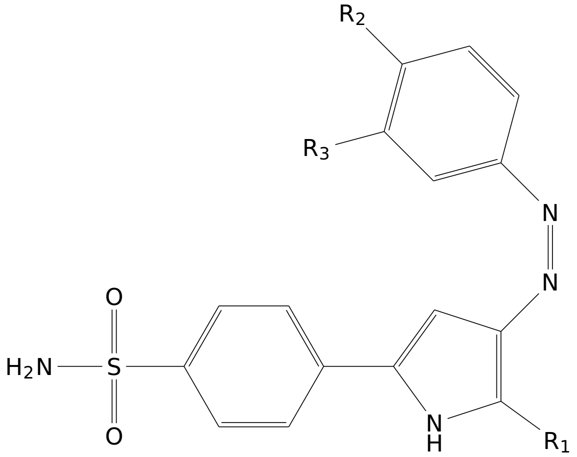
Identifier	R_1	R_2
7.1	CF_3	H
7.2	H	H
7.3	F	H
7.4	Cl	H
7.5	Br	H
7.6	CH_3	H
7.7	CF_3	F
7.8	H	F
7.9	F	F
7.10	Cl	F
7.11	Br	F
7.12	CH_3	F



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.

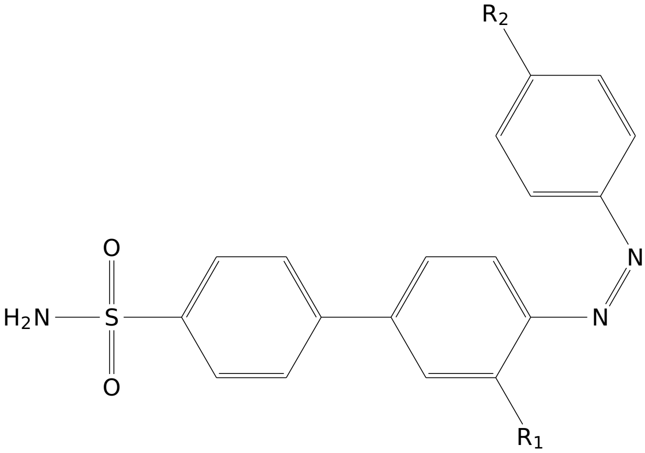
Identifier	R_1	R_2
8.1	F	H
8.2	H	F
8.3	Cl	F



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.

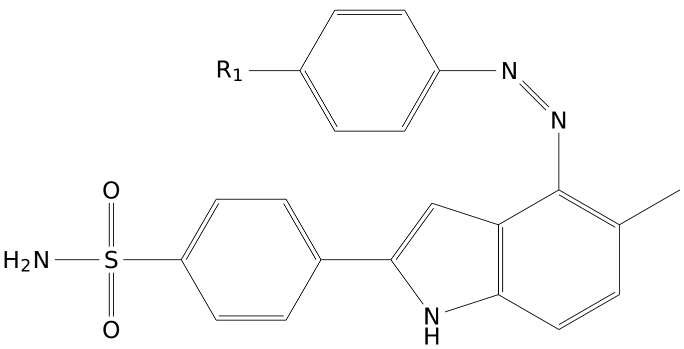
Identifier	R_1	R_2	R_3
9.1	CF_3	CH_3	H
9.2	Cl	CH_3	F



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.

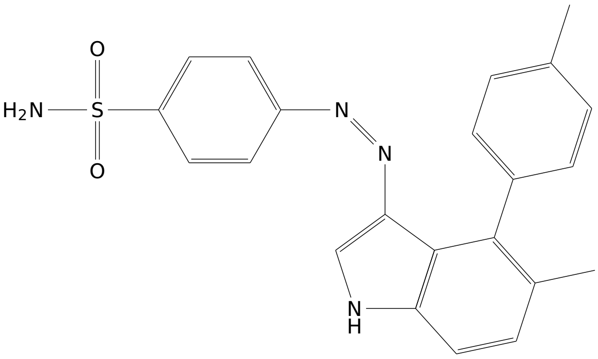
Identifier	R_1	R_2
10.1	CF_3	CH_2CH_3
10.2	CF_3	NCH_3COCH_3
10.3	CF_3	$NHCH_3$
10.4	CF_3	OCH_3
10.5	Cl	CH_3



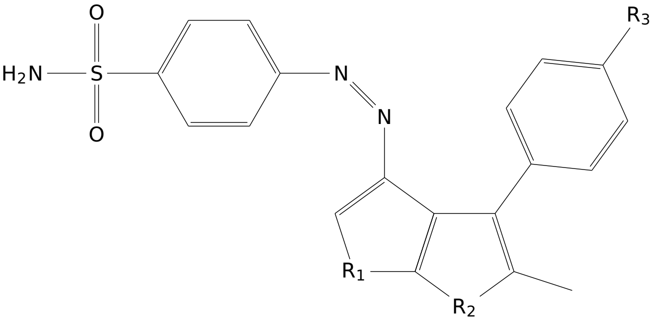
Scheme 11: Template for Celecoxib azo-derivatives with indole ring as a heterocycle.

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 Table of results

Table 9: Results for the $\Delta G_{binding}$ and IC_{50}. The conditions and estatistics 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	5.1	CF_3	CH_2CH_3	H	-4.3505	145
Pyrazole	5.2	CF_3	CH_2CH_3	F	-3.619	160
Pyrazole	5.3	CF_3	CH_3	F	-0.5452	192
Pyrazole	5.4	CF_3	OCH_3	H	9.9793	306
Pyrazole	5.5	CF_3	OCH_3	F	1.5591	238
Pyrazole	5.6	CF_3	CH_3	H	-5.3292	142
Pyrazole	5.8	F	CH_3	H	5.0694	240
Pyrazole	5.9	Cl	CH_3	H	-1.1579	171
Pyrazole	5.10	Br	CH_3	H	6.3225	247
Pyrazole	5.11	CH_3	CH_3	H	3.0677	215
Pyrazole	5.12	H	CH_3	F	1.0334	196
Pyrazole	5.13	F	CH_3	F	3.2216	220
Pyrazole	5.14	Cl	CH_3	F	5.1623	231
Pyrazole	5.15	Br	CH_3	F	5.6163	235
Pyrazole	5.16	CH_3	CH_3	F	-3.3079	149
Furan	7.1	CF_3	H		-0.9611	171
Furan	7.2	H	H		0.193	189
Furan	7.3	F	H		-4.8383	133
Furan	7.4	Cl	H		-2.4267	158
Furan	7.5	Br	H		-3.4217	144
Furan	7.6	CH_3	H		-3.9711	139
Furan	7.7	CF_3	F		-5.4022	148
Furan	7.8	H	F		-6.55	122
Furan	7.9	F	F		-0.1311	189
Furan	7.10	Cl	F		-5.8185	127
Furan	7.11	Br	F		6.1474	241

Table 10: Results for the $\Delta G_{binding}$ and IC_{50}. The conditions and estatistics 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)
Thiophene	8.1	F	H		1.2724	192
Thiophene	8.2	H	F		-2.7842	152
Thiophene	8.3	Cl	F		-5.7964	120
Pyrrole	9.1	CF_3	CH_3	H	-3.5279	151
Pyrrole	9.2	Cl	CH_3	F	-1.3499	164
Benzene	10.1	CF_3	CH_2CH_3		-1.339	168
Benzene	10.2	CF_3	NCH_3COCH_3		-1.9922	208
Benzene	10.3	CF_3	$NHCH_3$		4.8759	232
Benzene	10.4	CF_3	OCH_3		-3.4432	155
Indole	11.1	H			-9.7543	74
Indole	11.2	F			-6.3397	108
Indole	11.3				-5.532	111
TwoRings	13.2	NH	O	H	-2.2146	165
TwoRings	13.3	O	NH	H	-0.7334	178
TwoRings	13.6	NH	O	CH_3	5.5592	235
TwoRings	13.7	O	NH	CH_3	-10.6712	77
TwoRings	13.8	O	O	CH_3	-3.5413	142

Table 11: Conditions and statistics for the computation of the results from tables (9) and (10).

NumberOfTrees	250
Erased Percentatge	0.0%
Splitting proportion	10.0% for testing
minimumCorrelationFactor	0.0
Number of descriptors	3040
Mean Squared Error	1202.70
R-squared	0.89
True Positive	994
False Positive	8
True Negative	1015
False Negative	101
True Positive Rate	0.91
True Negative Rate	0.99
ClassificationAcuracy	0.94
MatthewsCorrelationFactor	0.87