Universitat Autònoma de Barcelona

AI application for azophotoswitches' optimization with pharmacological interest

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

Final Degree Project
Bachelor's Degree in Chemistry

2024-2025

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

Malcolm S. Forbes

Abstract

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

I Introduction

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

This project aims to implement artificial intelligence in computational chemistry, concretely, using AI-based algorithms to predict a drug's inhibition potential for a given protein. To achieve this, we make use of the ChEMBL database, 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 software. Around 10^4 descriptors are calculated, which go 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 emphasize 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 analyzing large datasets, AI can detect hidden correlations that may not be immediately apparent through traditional statistical methods.

At this stage, we focus on cyclooxygenase-2 (COX-2), a protein well known for its strong association with cancer development and inflammatory diseases. COX-2 plays a crucial role in the biosynthesis of prostaglandins, which mediate inflammation and pain. Overexpression of COX-2 has been linked to various types of cancer, making it a prime target for drug development. COX-2 inhibitors, such as celecoxib and rofecoxib, have been widely studied for their therapeutic potential. The scientific community has devoted an extensive research to COX-2, even before the rise of AI, due to its biomedical significance. By applying AI models to COX-2, we assess their compatibility with the latest research findings, demonstrating AI's potential as a powerful tool in computational chemistry research. Our approach not only validates AI's effectiveness in predicting inhibition potential but also provides insights into the underlying molecular mechanisms governing COX-2 interactions.

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

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

This study aims to bridge the gap between artificial intelligence and computational chemistry, demonstrating AI's potential to revolutionize 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.