Universitat Autònoma de Barcelona

Notes: 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

1 Definitions

Definition 1. *IC_{50}: Half maximal inhibitory concentration "IC_{50} is the concentration of drug required for 50% inhibition. IC_{50} is an operational term dependent on the assay conditions. IC_{90} or IC_{99} is sometimes used when complete inhibition is required. Calculation of the fractional occupancy shows that IC_{90} concentration is approximately 10-fold greater than the IC_{50} concentration assuming one-site binding at equilibrium with a Hill coefficient of 1."[3]*

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."[2]*

2 Observations

Observation 1. *Machine Leaning (ML) and Artificial Intelligence (AI) are not the same concept, in fact, a ML models are a subset of AIs. The key relies in word* Learning*, generally speaking, an AI does not need to actually learn from a set of data. It can be set up within a decision tree such that logically responses with the proper answer following its criteria.*

However a Machine Leaning Method is a kind of AI that learns *from data and evolves with the provided data.*

Observation 2. *Mathematically speaking, Machine Learning methods are no different from an optimisation problem [1].*

Generally speaking, all ML methods seek to find a criterion to determine the quantity of a certain property of some data, which we lack knowledge of and will call the target*. In order to do so, they rely on a dataset where the sought property is already known and which establishes relationships between this property and other features of the data. Then, based on these relationships and the properties of the* target*, the model searches for the position where the* target *fits within the data and determines the quantity of the property, taking this placement into account within the known data.*

For instance, we could think of standard calibration with linear regression as an ultra-simplification of a Machine Learning problem. We have a set of data (such as salt concentration in a solution) $\{x_i\}_{i=0}^{n}$, where we know the quantity of a certain property (such as the solution's conductivity) for each entry x_j, denoted as $\{f(x_i)\}_{i=0}^{n}$. In this example, our ML method assumes that the data and the property are linearly related and that the distance between data and the property is equivalently defined as $d(x_i, x_j) = \sqrt{x_i^2 + x_j^2}$. Hence, based on the assumption of linearity, the best possible relationship is a straight line, and the method looks for the line that minimises the distance between the line and the data.

Then, given a target *(for instance, a certain concentration), about which we initially lack knowledge, the method is capable of quantitatively computing the theoretical property (for instance, its conductivity) by using the fitted line representing the relationship. Consequently, this can be helpful in determining the theoretical concentration of salt needed to achieve a certain conductivity without having actual data for that concentration.*

The learning *part (the most important part) comes from the fact that the model (the line) learns from the data. As we train our system (i.e., perform linear regression) with increasingly larger datasets, the accuracy of our model's predictions improves. Thus, we can conclude that the system is* learning *from the data, which qualifies it as a Machine Learning model.*

Observation 3. *Since we are working with molecules, our Machine Learning model will require certain parameters to determine the relationship with the property.*

In this case, our property *(i.e., $\{f(x_i)\}_{i=0}^{n}$) will be the inhibition of a specific protein (COX-2), and more specifically, it will be IC_{50}, IC_{90}, or IC_{99} (1).*

Moreover, our parameters *(i.e., $\{x_i\}_{i=0}^{n}$) will be the molecular descriptors, which must be **carefully chosen**. This problem introduces additional complexities compared to the linear regression case:*

- *We will be working with approximately 10^3 descriptors, making this a 10^3-dimensional problem.*
- *We want the system to automatically discard non-relevant data, as some molecules could be exceptions due to external factors.*

- *We do not assume linearity.*
- *Not all descriptors have the same distances or weights, as we will want some of them to be more relevant than others.*

Interestingly enough, despite these complexities, in some way, we can say that this project is not entirely different from computing a linear regression...

References

[1] Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. *Mathematics for Machine Learning*. Cambridge University Press, 2020.

[2] J. Gasteiger and T. Engel. *Molecular Descriptors for Chemoinformatics*. Wiley-VCH, Weinheim, Germany, 2003.

[3] David C. Swinney. Chapter 18 - molecular mechanism of action (mmoa) in drug discovery. volume 46 of *Annual Reports in Medicinal Chemistry*, pages 301–317. Academic Press, 2011.