UAB

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.”[5]*
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]*

Definition 3. *Constitutional Indices Descriptors: Kind of descriptor based on the constitutional composure of a molecule. They describe the basic composition of the molecule without considering its geometry or connectivity of it.*

Definition 4. *Ring Descriptors: Kind of descriptor that describes the ring systems of our molecule. Quantifies properties such as the ring connectivity, type of ring (odd/even ring), aromaticy & saturation.*

Definition 5. *Topological Indices: These descriptors are based on the connectivity of atoms in a molecule without considering three-dimensional coordinates in the context of chemical graph theory. They help to describe the molecular graph (atoms as vertices, bonds as edges) using metrics like the Wiener index or the Zagreb index.*

Definition 6. *Wiener index: Specific topological descriptor based defined as the sum of the lengths of the shortest paths between all possible vertices in the chemical graph. Mathematically speaking, if we have a molecule with a set of $\{A_i\}_{i=0}^n$ atoms and we denote $d_B(A_i, A_j)$ the bond distance between atoms A_i and A_j, we define the Wiener index of this molecule as,*

$$Ind_{Wiener} = \sum_{i=0}^{n} \sum_{j>i}^{n} d_B(A_i, A_j) \tag{1}$$

For instance, if we take the n-Butane & Isobutane chemical graph from the Figure (1) the Wiener index would be,

$$d_B(C_1, C_2) + d_B(C_1, C_3) + d_B(C_1, C_4) \tag{2}$$
$$+d_B(C_2, C_3) + d_B(C_2, C_4) + d_B(C_3, C_4) \tag{3}$$

Though this quantity is computed equally for both, n-Butane & Isobutane, it results different due to their connectivity. For the n-Butane:

$$1 + 2 + 3 + 1 + 2 + 1 = 10 \tag{4}$$

While for the Isobutane:

$$1 + 1 + 1 + 2 + 2 + 2 = 9 \tag{5}$$

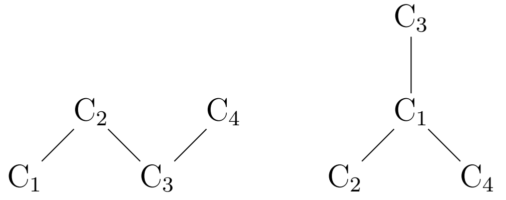

Figure 1: n-Butane's & Isobutane's chemical graphs.

Definition 7. *Zagreb index: ”For a (molecular) graph, the first Zagreb index M_1 is equal to the sum of squares of the degrees of vertices, and the second Zagreb index M_2 is equal to the sum of the products of the degrees of pairs of adjacent vertices.” [7]*

Definition 8. *Walk and Path Counts Descriptors: Kind of descriptor focused on counting specific paths or walks (connected sequences of atoms) in the molecular graph. For instance we can get a Walk and Path Counts Descriptor quantifying the number of Eulerian paths in a molecule's graph. They give insights into the molecule's connectivity and complexity.*

Definition 9. *Connectivity Indices: Connectivity Indices are a class of molecular descriptors that quantify the connectivity or bonding patterns between atoms in a molecule based on its topology (i.e., the molecular graph). These indices capture the degree to which atoms are connected, considering how atoms are bonded to each other rather than their physical 3D arrangement. They are typically derived from graph theory, where atoms are represented as vertices and bonds as edges.[4]*

Definition 10. *Randic Index: Specific connectivity index defined as the following,*

$$\sum_{\textit{all bonds}} \frac{1}{\sqrt{d_i d_j}} \tag{6}$$

where d_i and d_j denote the number of bonds of atoms i and j connected by a bond.[3]

Definition 11. *Information Indices: Kind of descriptor based on the information theory, where the molecular structure is represented as a distribution of information (or entropy). They measure molecular complexity and diversity.*

Definition 12. *2D-Matrix Based Descriptors: Kind of descriptor based in the matricidal representation of a molecule[6].*

Definition 13. *Adjacency Matrix: Matricidal representation of a molecule. If we have $\{x\}_{i=1}^{n}$ atoms that form our molecule, labeled from 1 to n, the adjacent matrix A will contain as the element $a_{i,j}$ 1 if the atom i and j are connected and 0 otherwise. (We can also naturally add the possibilities of $a_{i,j} = 2, 3$ for double and triple bonds).*

For instant in the Figure (1) the adjacent matrix for the n-Butane is

$$\textit{n-Butane } AM = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{7}$$

Trivially we observe that the main diagonal will always be filled with 0s and the matrices will also be doubly symmetrical (respect both diagonals) due to the bounds transitivity and reflectivities properties.

1. 2D Autocorrelations: These descriptors calculate correlations between atomic properties at different distances in the molecular graph. They capture how molecular properties change across the structure.

2. Burden Eigenvalues: These descriptors are derived from the eigenvalues of matrices related to the structure, such as the Burden matrix, which takes into account atomic properties like electronegativity and partial charges.

3. P_VSA-like Descriptors: P_VSA stands for "Property-weighted Van der Waals Surface Area." These descriptors quantify the surface area of a molecule associated with specific properties like polarizability or charge.

4. ETA Indices (Extended Topochemical Atom): ETA indices capture aspects of molecular topology in conjunction with atomic properties. They are often used in the analysis of electronic properties or biological activities.

5. Edge Adjacency Indices: These descriptors are derived from the adjacency of bonds (edges) in the molecular graph.

They measure the proximity of bonds in the structure and can reflect molecular complexity.

6. Fractional Group Counts: These descriptors count specific chemical groups or fragments within the molecule, providing information about the functional groups present.

7. Atom-Centered Fragments: These descriptors focus on the environment of individual atoms within a molecule. They describe the types and arrangements of atoms surrounding each atom in the structure.

8. Atom-Type E-State Indices: The E-State (Electrotopological State) descriptors combine information about the electronic environment and topology of atoms. They reflect the potential reactivity of atoms in the molecule.

9. Pharmacophore Descriptors: These describe the spatial arrangement of features critical for molecular recognition by biological targets, such as hydrogen bond donors/acceptors, hydrophobic regions, or aromatic rings.

10. 2D Atom Pairs: These descriptors capture the relationships between pairs of atoms at a specific distance (in terms of bonds) from each other. They describe the distribution of atom pairs within the molecular structure.

11. Charge Descriptors: These describe the charge distribution within the molecule, including total charge, partial atomic charges, and the dipole moment.

12. Molecular Properties: These include physical and chemical properties of the molecule such as molecular weight, boiling point, or solubility.

13. Drug-Like Indices: These descriptors quantify how closely the molecule follows the rules for drug-likeness, such as Lipinski's Rule of Five, which includes factors like molecular weight, hydrophobicity, and hydrogen bonding.

14. MDE Descriptors (Molecular Distance Edge descriptors): These describe the distances between atoms in a molecule based on the molecular graph. They are useful for capturing molecular shape and spatial relationships.

15. Chirality Descriptors: These quantify the chirality (handedness) of molecules. They reflect the 3D arrangement of atoms around chiral centers, which can influence how molecules interact with biological systems.

Definition 14. *QSAR: "Quantitative Structure–Activity Relationships are the finnal result of the process that starts with a suitable description of molecular structures and ends with some inference, hypothesis, and prediction on the behaviour of molecules in environmental, biological, and physico-chemical systems in analysis.*

QSARs are based on the assumption that the structure of a molecule (for example, its geometric, steric, and electronic properties) must contain features responsible for its physical, chemical, and biological properties and on the ability to capture these features into one or more numerical descriptors. By QSAR models, the biological activity (or property, reactivity, etc.) of a new designed or untested chemical can be inferred from the molecular structure of similar compounds whose activities (properties, reactivities, etc.) have already been assessed."[6]

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) $\{x_i\}_{i=0}^n$, we also 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.*
- *Some descriptors will be continuous (for instance, bond distances) while others will remain discrete (for instance, topological indexes) by construction which implies that some dimensions need to be treated differently to others.*

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

Observation 4. *Taking into account definitions (6) and (7), we can also define a norm, denoted by $\mathcal{N}(A, B)$, which refers to the norm between molecules A and B. For instance, this can be expressed as $\mathcal{N}(A, B) = |Ind_A - Ind_B|$, where Ind_A and Ind_B represent the respective indices (Wiener, Zagreb, etc.) of the molecules*[1]*. Using this norm, we can generate a metric space that tells us how close two molecules are, based on the similarity of their indices.*

In this way, we can apply the concept of "linear regression" and compute the relationship between the indices and the IC_{50}. This refers to one dimension, corresponding to a specific topological index. We can then extend this to consider all remaining indices (approximately 10^3), and find the optimal position in this 10^3-dimensional space for our resulting molecule.

[1] Actually, any kind of norm would work, as long as it depends on the indices.

3 Reflections

Reflection 1. *During the first meeting with Dra.González, Dr. Moreno & Dr. Lluch a certain theme was brought up. "Wouldn't be great to actually have the wave function of the molecules?" This way we could avoid the use of descriptors and QSARs and have more accurate results.*

Though it was a joke, it might not be a utopia someday. I believe quantum computers will be capable of this task and maybe a hundred years from today, this project could be doable working directly with the waves functions (not analytically obviously). How knows? Time will tell us.

Reflection 2. *Fun fact, [7] is a mathematical article, completely apart from the chemical world but somehow we can manage to use it to describe molecules whatsoever. I believe applying knowledge from other sectors is key to evolve and reach even greater results.*

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] Ivan Gutman, Boris Furtula, and Vladimir Katanić. Randić index and information. *AKCE International Journal of Graphs and Combinatorics*, 15(3):307–312, 2018.

[4] Milan Randic. Characterization of molecular branching. *Journal of the American Chemical Society*, 97(23):6609–6615, 1975.

[5] 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.

[6] Roberto Todeschini and Viviana Consonni. *Molecular Descriptors for Chemoinformatics*. Wiley-VCH, Weinheim, Germany, 2nd edition, 2009.

[7] Kexiang Xu. The zagreb indices of graphs with a given clique number. *Applied Mathematics Letters*, 24(6):1026–1030, 2011.