UNIVERSITÀ DEGLI STUDI DI MILANO

Department of Pharmaceutical Sciences

DOCTORATE SCHOOL IN CHEMICAL SCIENCES AND TECHNOLOGIES

Chimica del farmaco

(XXVIII Cycle)
CHIM/08

In silico approaches in drug design and development: applications to rational ligand design and metabolism prediction

Dott. ssa Angelica Mazzolari

R10087

Tutor: Prof. Giulio Vistoli
Coordinator: Prof. Marco De Amici

Academic Year 2014/2015

“Nobody ever figures out what life is all about, and it doesn't matter.
Explore the world.
Nearly everything is really interesting if you go into it deeply enough.”

Richard Feynman

Acknowledgments

I would like to thank my supervisor Prof. Giulio Vistoli and Prof. Alessandro Pedretti for their guidance and support during my PhD.
I would also like to express my gratitude to Prof. Bernard Testa for giving me the opportunity to work on the interesting project about metabolism prediction.
My thanks also go to all the colleagues and students who took part in our research group during these years, especially Dr Matteo Lo Monte, for his scientific and emotional support, and Dr Elisa Varè for the contribution to the Purinergic project.

I would also like to thank my supervisor at the Department of Chemistry, University of Cambridge, Dr Andreas Bender for allowing me to spend the third year as a research visitor in his group, and his appreciated advice.
My grateful thanks are also extended to all the colleagues of the Bender group for providing such an enjoyable working environment, and for the interesting discussions. Among them, I would like to make a special mention of Avid Afzal for his teaching and valuable contribution to the PCM project.

Correspondence to previous publications

Parts of this thesis are taken from previously published papers:

- A. Pedretti, A. Mazzolari, C. Ricci, G. Vistoli.
  "Enhancing the Reliability of GPCR Models by Accounting for Flexibility of their Pro-Containing Helices: The Case of the Human mAChR1 Receptor"
  Molecular Informatics 34(4):216–27.
  (http://doi.wiley.com/10.1002/minf.201400159).
  This paper has been included in Chapter 5.2;

- F. Del Bello, A. Bonifazi, W. Quaglia, A. Mazzolari, E. Barocelli,
  S. Bertoni,R. Matucci, M. Nesi, A. Piergentili, G. Vistoli.
  "Mode of Interaction of 1,4-Dioxane Agonists at the M2 and M3 Muscarinic Receptor Orthosteric Sites."
  Bioorganic & medicinal chemistry letters 24(15):3255–59.
  (http://www.ncbi.nlm.nih.gov/pubmed/24980056).
  Part of this paper has been included in Chapter 5.3.

Contents

Chapter 1 – Introduction

1.1 The role of computational methods in drug discovery

In the last decades, the applications of computational methods in medicinal chemistry have experienced significant changes which have incredibly expanded their approaches, and more importantly their objectives.
Indeed, computational methods were initially focused on the rational design of novel improved bioactive molecules. Using both ligand-based and structure-based approaches, the in silico studies were primarily focused on parameterizing the ligand-receptor binding processes and were substantially aimed at optimizing the ligand structure in terms of affinity, potency and selectivity. Usually, these studies involved a few dozen of compounds and their primary objective was to extract qualitative or (better) quantitative relationships to guide the rational design of novel (hopefully improved) derivatives. Not surprisingly, the most utilized and simplest way to define such kind of computational studies was "drug design" because all their objectives were limited to ligand design and optimization.

Over time, computational applications felt a rich extension of their objectives and one of the clearest examples in this context is represented by the ever increasing applications of the *in silico* tools to optimize the ADME/Tox profile of the novel compounds. In fact, the development of computational models to predict ADME/Tox parameters is surely older and dates back to the 70s, but only on the last decades these approaches are applied in a comprehensive and synergistic way and, more importantly, during the early stage of the discovery process, to focus the attention only on the most promising derivatives, so limiting the remarkable attrition caused by unsuitable pharmacokinetic profiles. In particular, the last two decades have seen an incredibly increasing effort spent in developing reliable approaches for metabolism prediction, especially as an inappropriate metabolic

profile is probably the most impeding attrition cause in drug discovery and development.

In parallel, computational methods are finding successful applications in the research phase which precedes the ligand design and which is focused on a detailed validation and characterization of the biological target. In this context, in silico approaches can be conveniently exploited, for example, to elucidate the mechanism of action at an atomic level, to explore the role of secondary/allosteric binding sites and to deorphanize new receptors or new binding pockets. Also here, the protein modeling techniques are clearly older, but recently they found new avenues with the incredibly increasing number of potential targets coming from the "omics" disciplines. On these grounds, we can notice that nowadays computational approaches can support the drug discovery pipeline in all its phases, starting from the target validation, to hit selections, until lead optimization.

Moreover, the above mentioned "omics" disciplines dictated also a remarkable change in the commonly used *in silico* approaches, or at least in their founding philosophy. Indeed and as said above, old analyses usually involved a limited number of derivatives and only one (or at most very few) targets, whereas a typical current study can include thousands compounds potentially interacting with many targets. This change of paradigm required the development of novel computational approaches able to manage such an incredibly huge amount of data. When considering the biological (proteins and genes) data analyses, these approaches are often referred by using the term of bioinformatics, the development of which paralleled in the last few years the advances in all omics techniques. Similarly, the huge amount of chemical data coming from these kinds of studies required the development of similar computational approaches which thus are defined with the overall term of chemoinformatics.

1.2 Aim of this PhD project

During my PhD studies, I had the opportunity to make experience of the three principal application areas of computational methods in drug discovery, as I have mentioned before, and to divide my efforts between the hit selection phase, as well as the target validation one and, finally, the lead optimization and traditional drug-design studies.

The first and main object of my project regards the field of metabolism prediction, and is based upon the meta-analysis and the corresponding metabolic database, called MetaSar, manually collected in a collaborative project involving our research group. The aim of this database, the compilation of which required a great deal of effort and a long work, is the building of a global predictive algorithm, and is perfectly in line with the actual trend in drug discovery which gives increasing importance to the metabolic fate of new leads, and thus recommends an earlier screening of the hit compounds based on their pharmacokinetic profiles.

Among the numerous possible secondary applications of MetaSar, one in particular, the Proteochemometric modelling (PCM), consists in a predictive technique that is at the forefront of the latest modelling techniques, as it perfectly fits the growing request of new solutions to deal with big data. In this context, MetaSar represents an alternative and still appropriate source of data which enables the extension of the fields of application of this predictive technique to a new avenue, represented by metabolism prediction. As this methodology is relatively novel and based on innovative approaches, Chapter 2 is entirely dedicated to a description of the basics of PCM, while the computational details of our application to the glucuronidation reactions' prediction and the obtained results are reported in Chapter 3.2

Always involving the almost unexplored field of glucuronidation reactions on which PCM is focused, Chapter 3.3 reports also a second application of

computational approaches, represented by homology modelling techniques and MD simulations. In detail, the study involves the main isoform of glucuronidation enzymes in humans, UGT2B7, and uses Steered Molecular Dynamics (SMD) simulations to gain deeper insights on the reaction mechanism of this enzyme, as well as to optimize its homology model.

A proper example of detailed validation and characterization of the biological target is given by the study about purinergic receptors, reported in Chapter 4, which is aimed at improving the knowledge about this promising and recently discovered biological target. In particular, the research regards the identification and characterisation of potential allosteric binding pockets for the already reported inhibitors, and involves not only standard docking-based modelling techniques but also applications of new approaches for binding site predictions (SPILLO-PBSS).

The canonical computational methods are instead exploited in Chapter 5, which regards the muscarinic receptors. Beside the use of already well-known ligand-based and structure-based approaches aimed at optimizing the ligand design, we applied also a novel method for target optimization, which explore the structural flexibility of the modelled GPCR structure by generating the so-called conformational chimeras.

In conclusion, the overall aim of the present research project is to explore the different fields of the modelling studies by using different and innovative techniques, from the first examples to the more recent approaches, to contribute to draw the images of the frontiers in the world of chemoinformatics.

Chapter 2 – Proteochemometric modelling

2.1 Introduction

This introductive chapter is aimed at presenting the methodological bases of an innovative computational technique, the Proteochemometric modelling (PCM), that we utilized in the prediction of the metabolism by UGT enzymes (see Chapter 3.4). The other here reported computational studies involved rather conventional approaches, and, as such, they do not deserve detailed methodological introductions. Therefore, their computational details will be separately described in each chapter.

In this section, we are including two computational strategies which can be classified in a hybrid position between traditionally ligand-based and structure-based approaches. They are mostly oriented towards the ligand space, but, with a different importance, they are also focused on the target space. They are both part of the wide world of Chemogenomics, and interdisciplinary field which lies in the interface of biology, chemistry, and informatics. Chemogenomics defines the systematic screening of the whole chemical universe against the whole targets universe, with the final aim to identify novel drugs as well as novel targets[1] (Figure 2.1).

The ultimate goal can never be achieved but the realistic purpose is that, along this systematic screening, we can increase the chances in the research of compounds with significant targets.

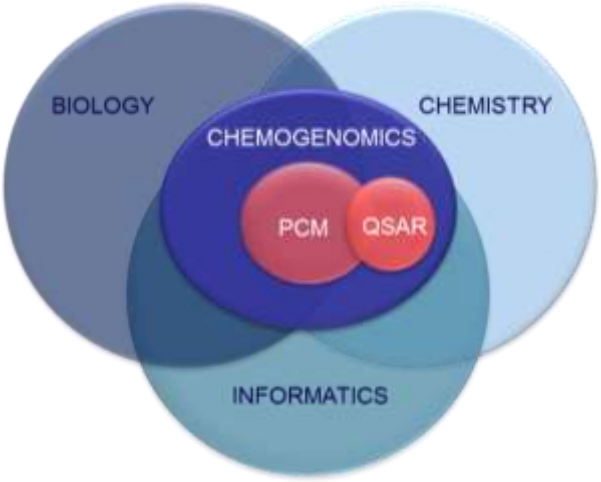


Figure 2.1: PCM and QSAR in the Chemogenomics world.

2.2 Quantitative structure-activity relationships (QSAR)

"Quantitative structure-activity relationships" (QSAR) is a conventional modelling technique that interprets the quantitative relationship between the bioactivity of a specific molecule and its structure. By using statistical approaches, it models the interactions of a group of compounds to a single target, in order to extrapolate the bioactivity of novel compounds on that specific target. This approach is founded on the principle of congenericity, by which chemicals sharing similar properties also share similar targets[2].

As QSAR considers just a single target, it has some limitations and drawbacks. First of all, to build the predictive model, it requires a good amount of data

available for ligands of that specific target, and this is a rare situation, especially when dealing with recently identified targets. Furthermore, it has an unsatisfactory ability to extrapolate into novel area of chemical space: this means that it is difficult to identify new classes of ligands or new binding modes outside the training set. Moreover, as traditional QSAR models are built using only compounds features, it is not even completely able to describe the interactions between that specific target and its ligands[3]. Indeed, the ligand-target interaction depends on the reciprocal affinity between chemicals and protein and to better describe this interaction it is necessary to take in consideration also the features of the binding site[4]. To overcome this drawback, some QSAR studies include also features depending on the specific target, such as docking scores or "cross-terms" of ligand and target features [5]. This is the case in which the technique advances towards the structure-based approaches.

2.3 Proteochemometric modelling (PCM)

Proteochemometric modelling (PCM) is a computational technique that describes both small molecules and targets and combines description from the ligand and the biological side of the system within a single predictive algorithm, in order to completely model the compound-target interaction space. Thanks to the introduction of target descriptors in its data matrices, PCM can be regarded as an extension of the classical QSAR modelling and can actually overcome the limitations of that conventional technique. It permits to extrapolate the bioactivity of novel compounds on both known targets and yet untested one[3].

Generally, the term "*target*" refers to protein, but it can also refers to protein-complexes or to gene expression levels of particular cell lines and, conversely, to a specific part of the protein target, such as the binding site, enabling distinctions

between diverse binding pockets on the same protein, as well as different binding modes or protein conformations[5].

The relevance of considering more targets is also consistent with the recent overcoming of the so-called "one drug one target" paradigm and the demonstration that drugs exerts their therapeutic effects by modulating about six targets[6]. Thus, to exhaustively understand the effect of a given drug in a biological system, we cannot avoid describing the activity of that drug on all its targets.

Moreover, including more targets in the input matrices of the model, it is possible to describe a different selectivity of these targets regarding their ligands. The capability to virtually screen selective compounds that are solely active on a single member of a subfamily of targets is of particular and increasing interest in the contest of drug discovery.

The interesting advantage of PCM is that the model is able to simultaneously describe the interaction between a set of compounds and a set of targets without losing the capability to describe the interaction of an individual compound on an individual target inside the dataset. Of course, in order to build the best PCM model, it is important to have the maximum number of bioactivity data of multiple compounds on multiple targets.

In summary, while QSAR and PCM are based on similar assumptions, PCM can benefits from additional information in model training, leading to additional results. It permits to outperform conventional QSAR methods and enlarges the principle of congenericity: not only compounds with similar features share similar targets, but also targets with similar compounds share similar features (Figure 2.2).

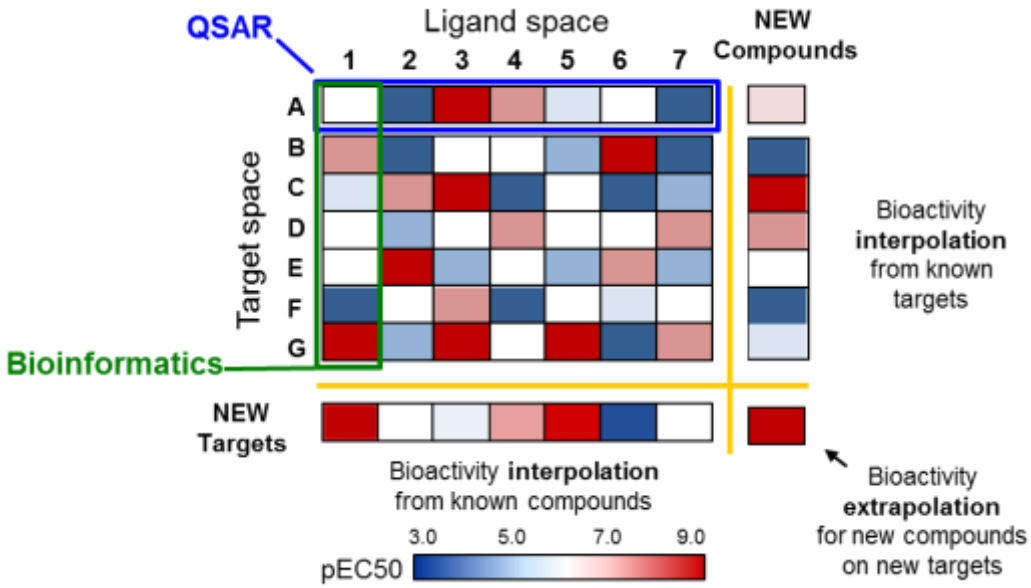


Figure 2.2: Heat map of ligand target interaction. The interaction between a set of ligand against a set of targets is represented in a matrix. The binding activity is calculated as pEC_{50} and transposed in colours according to the legend. Each matrix cell contains the binding activity of a given compound on a given target. QSAR models deal with the similarity between ligands and predict their activity on one target. On the contrary, traditional bioinformatics techniques deal with similarity between targets. PCM models are able to contemporary deal with both ligands and targets similarity, and thus they are capable of bioactivity interpolation from known targets to new compounds or from known compounds to new targets, and also of bioactivity extrapolation for new compounds on new targets.

The first example of a predictive model including descriptors of several proteins and their ligand has been developed by a research group at the Uppsala University, in 2001, and the name "Proteochemometric" has been coined by the same group. It concerned the interaction between chimeric peptides and chimeric Melanocortin Receptor[7]. Nowadays, although PCM is a relatively new technique, it has already been successfully applied to a wide variety of drug targets.

Among these, the family of G protein coupled receptor (GPCR) has been studied by many research groups. Weill and Rognan[8] created on global GPCR model combining several models on class A GPCR and they used it to identify the natural receptor ligands in a dataset of 200000 decoys. Bock *et al.*[9] built another global

model able to identify selective ligands for an orphan GPCR receptor among a dataset of 1.9 million data points. Another interesting application is in the field of viral proteins, such as HIV Protease, Dengue virus NS3 Protease, Influenza virus A and B Neuramidase[10], in consideration of the high similarity between the mutants, even more evident in comparison with the average similarity between multiple GPCR families. This is also true for enzyme mutants, and another important topic subjected to PCM is the Kinase superfamily and the corresponding inhibitors compounds.

PCM comprises a vast variety of computational tools and can implement different Machine Learning models. By exploring the whole range of opportunities it offers, and using well curated data so to take full advantage from the technique, it is possible to achieve interesting progresses in drug discovery and explore new avenues. Among these, we can mention the generation of multitarget or multispecies models, the "deorphanization" process of novel emerging targets, the integration of phenotypic and toxicity data in predictive models, the design of new personalized medicine for antiviral or anticancer therapy based on genotypic information[11].

Two kinds of PCM models can be implemented, the first based on a regression and the second on a classification algorithm. Regression model involves estimating or predicting of a response and the output variable are continuous values. Conversely, classification aims at identifying group memberships and the output variables are class labels.

In this Thesis, a Proteochemometric technique was applied to MetaSar substrates and UGTs enzymes resulting in the first PCM model about the regioselectivity of glucuronidation reactions. The study led to the generation of a classification algorithm able to predict if a given molecule can be a UGTs substrate or not.

2.4 Chemical descriptors

A chemical descriptor is a numerical value which describes the properties of the molecule in order to establish relationships between its structure and its functions. Both QSAR and PCM traditionally use them and a large collection of ligand descriptors is now available, all aiming at both describing the ligand itself, and putting in evidence the differences to one another.
A brief overview of chemical descriptors currently used in QSAR and PCM model will be given, with a particular attention on those applied by the studies described in this thesis.
A first classification of chemical descriptors can be done on the basis of the data type, which can be integer values, real numbers or a Boolean variables, when indicating whether a specific functional group is present in the ligand or not.
Another common classification of chemical descriptors relies on the dimensionality of the molecular representation from which they are calculated. According to this criterion, a ligand descriptor can have zero, one, two, or three dimensions (0D, 1D, 2D, 3D)[1].

OD descriptors are merely derived from the chemical formula, indicating for example the occurrence of a particular functional group. 1D descriptors are rudimental physicochemical descriptors, such as the atom counts, the bond counts and the molecular weight, and are generally simple, computationally very fast and easy to interpret. This means also that they are not informative enough to provide a satisfactory discriminating power and are particularly affected by possible overfitting of the data. Thus, they should be used always in combination with other, more informative, descriptors.

2D descriptors are computed from the graphical two-dimensional representation of the molecule caught by the SMILES, which allows a precise definition of the

atomic connectivity. The connectivity is the bonds' pattern connecting the atoms of a given molecule, and, from it, a molecule can be easily dissected into constitutive fragments. This kind of dissection allows the calculation of all physicochemical descriptors, which have a constitutive/additive character by which their values can be computed as a sum of fragmental increments. These physicochemical descriptors include, among the others, the Virtual LogP, which is a calculation of the octanol-water partition coefficient, namely a measure of the molecular lipophilicity. Topological molecular descriptors are calculated based on the molecular graphs (in a graph, the atoms form the nodes and the bonds correspond to the edges), and include bond properties, atomic properties and inner atom distances between the functional groups. In the 2D Circular Fingerprints and in other kind of fingerprints, the molecular 2D structure is transformed into a numerical string.

2D descriptors are frequently used both in traditional QSAR model and in PCM, since they can catch fairly well the properties of the chemicals and the differences among them, in order to establish a direct relationship between structure and function. They also have the advantage to be quickly calculated and modelled, even when dealing with large datasets[12].

3D descriptors are derived from the three-dimensional representation of the chemicals. They overcome the typical limitations due to the insufficiently informative two-dimensional characterization, while clearly having other drawbacks. The calculation process can be time-consuming and it does not derive directly from the three-dimensional representation of the molecule[13]. Indeed, alignment-based 3D compound descriptors require superimposition in 3D space of the chemicals in their bioactive conformation, a step that can introduce more noise than functional information and requires the active conformation of the ligand to be known. This problem is overcome by the Grid Independent descriptors (GRINDs), which are calculated starting from a set of molecular interaction fields. These

descriptors are computed in such a way as to remain highly relevant for describing biological properties of chemicals while being alignment-independent, chemically interpretable and easy to derive[14].

2.4.1 Extended-Connectivity Fingerprints

Extended-Connectivity Fingerprints (ECFPs) are an important class of topological 2D fingerprints for molecular characterization, developed at the "University Of San Diego", California, in 2000. They were originally designed to assist the structure searching in chemical databases, and later largely applied to a wide range of applications, including similarity searching and structure-activity relationships. The first application of ECFPs was in the area of high-throughput screening (HTS), as a tool to evaluate the results, analysing false positive and false negative hits. Furthermore, ECFPs were frequently applied in ligand-based virtual screening studies, in order to distinguish between actives and inactives. Among their most relevant advantages, we can mention the wide range of different molecular features that they are able to catch, the capacity to represent either the presence or the absence of functionalities - both crucial for analysing molecular activity- and the rapidity of computational time required to calculate them[15]. Another crucial peculiarity of ECFPs is that they don't use a set of substructure-based keys, as happened for many other kinds of fingerprints, such as MACCS. Since they are not defined *a priori*, they can represent novel structural classes.

The generation process of ECFPs consists of three subsequent steps. First of all, all the non-hydrogen atoms of the input molecule are considered individually and an initial integer identifier is assigned to each of them, according to various atom properties (e.g., atomic number, connection, count, etc.). The set of such local atom properties is an important configuration parameter that can be variously

customized. After that, these initial integer identifiers are modified combining them with identifiers that capture the neighbourhood around each atom, with circular subsequent iterations having an increasing diameter, until a value that can be customized. Finally, in the third step, eventual duplicates are removed, defining as a duplicate an identifier of an equivalent atomic neighbourhood.

This process results in the first kind of ECFPs representation: the list of Integer Identifiers. This representation consists in strings of different length made up by a list of integer numbers, each of them describing a specific circular atomic neighbourhood in the molecule. This varying-length string can also be interpreted as a virtual bit string, in which each position accounts for the presence or the absence of a specific substructural feature (Figure 2.3).

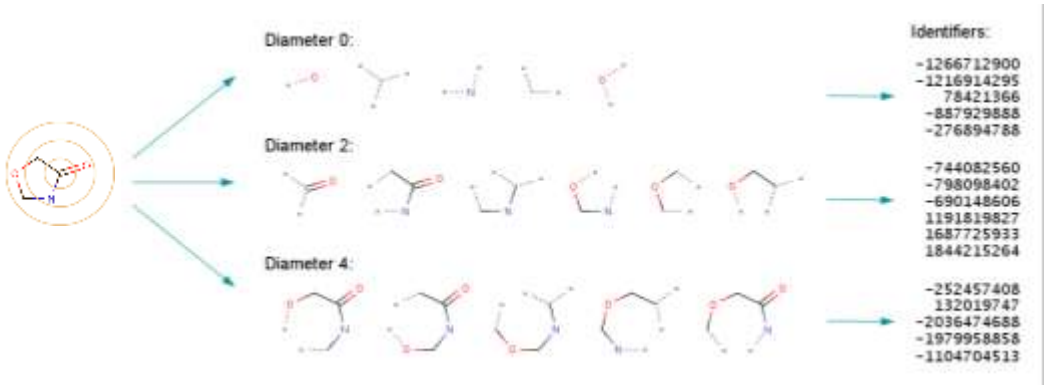


Figure 2.3: Illustration of the generation of integer identifiers by effect of iterative updating for a selected atom in a sample molecule. (Reproduced from docs.chemaxon.com)

A second kind of ECFPs representation is the Fixed-length bit string, which is obtained by "folding" the underlying virtual bit string into a much shorter bit string of specified length. This binary representation has numerous advantages that render it the first choice in QSAR and PCM models: compared to the identifier list, the

binary string reduces the required storage and, most of all, its specific length, which is another parameter customizable, allows their use by Machine Learning techniques and by similarity analysis. On the other hand, more than one different substructural feature could be represented in the same bit position, resulting in collisions and loss of pieces of information. Furthermore, the interpretability of the bits is partially compromised and the process to go back to the specific circular atom neighbourhood which generates the specific bit is not direct (Figure 2.4).

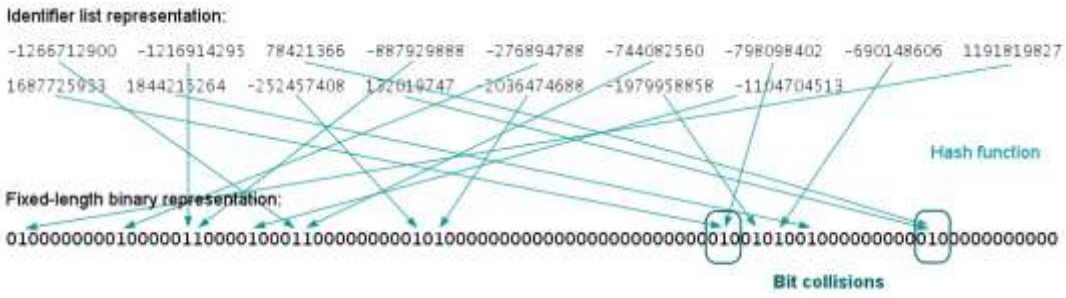


Figure 2.4: Generation of the fixed-length bit string from the list of Integer Identifier. (Reproduced from docs.chemaxon.com)

2.4.2 EigenSpectrum Shape Fingerprints

EigenSpectrum Shape Fingerprints (ESshape3) are a quite new kind of fingerprints belonging to the class of the shape-based descriptors. Instead of describing molecules as a set of atoms in a two-dimensional space, as the other 2D fingerprints do, they consider molecules into a three-dimensional contest, as having a volume and a surface. Since the drug-target interaction rely to the shape complementariness of the ligand to the active site, this kind of descriptors permit to add relevant information to the model[16].

ESsape3D are implemented in the Molecular Operating Environment software (MOE)[17], and are fixed-length string of integer values. The generation process starts with the individuation of the heavy atoms in the input molecule and the calculation of the Euclidean distance matrix for each pair of them[18]. The matrix affords a pattern characteristic of the molecular shape, which is thereafter encoded as a fingerprint[19]. The process consists in the calculation of the eigenvalues from the distance matrix, and then the signed square roots of these eigenvalues are smoothened by a Gaussian function and stored in a histogram with 122 bins containing values between -30 and +30.

However, the accuracy of shape-based descriptors is limited because of the lack of electrostatic properties, which are relevant as well as the steric properties to characterize the ligand-protein complementary.

2.4.3 3DAPfp and 3DXfp

The research group of Professor Jean-Luis Raymond, from the "University of Berne", designed a new type of 3D-atom pair fingerprints, called 3DAPfp and 3DXfp[20]. These descriptors belong to the class of topological descriptors and are an extension in the three-dimensional space of the two-dimensional topological atom pair fingerprints APfp and Xfp, developed by the same group.

APfp and Xfp count the number of atom pairs at increasing topological distances, through the shortest path. They are designed following a concept originally reported by Carhart *et al*[21] and are computed from 2D structure only. Nevertheless, they are found to correlate with molecular shape and able to encode 3D-features of molecules, in various enrichment studies[22].

The new 3D descriptors are a simplification of the 3D-atom pair fingerprints developed by Sheridan *et al*[23], whom calculation is more time consuming. 3DAPfp

and 3DXfp are designed to represent the actual 3D-shape more closely than the 2D corresponding one to distinguish among different stereoisomers of the same molecule. 3DAPfp is a 16-bit descriptor that treats all heavy atoms equally, while 3DXfp is a 80-bit descriptor that considers different atom categories, such as hydrophobic atoms, H-bond donors, H-bond acceptors and others.

2.5 Protein descriptors

A protein descriptor can be just one descriptor referred to the whole protein, characterizing one of its general properties, as well as the presence of a specific domain, or a list of descriptors referring to the sequence of amino acids. In both cases, considering the large dimension of the protein in comparison to the ligand, a selection of a subset of the protein, usually the binding pocket would be recommended. The collection of available protein descriptors is wide and can be grouped into different subclasses, among which the following descriptors are the ones typically used in PCM studies.

The first protein descriptors exploited in PCM models are the binary protein descriptors, where each column corresponds to a specific domain or to a key residue in the protein, and each binary value encodes the presence or the absence of that particular element. These descriptors are computationally fast and, in comparison with sequential descriptors, they often have more satisfactory performances, but also less interpretable capabilities and more limits in extrapolation [3].

While binary descriptors encoded only one-dimensional properties of the protein, some 3D protein descriptors have been developed by different research groups, both for the full protein[24], and for a specific region[25]. The performances of this kind

of descriptors in PCM are quite good but their use is limited to those cases in which a diverse dataset of protein 3D-structures is available.

Sequential protein descriptors, on the contrary, can be computed for almost all drug targets, since they only require the amino acid sequence. They consist in lists of descriptors encoding for the properties (mostly physicochemical properties) of each residue in the protein. Among them, the alignment-dependent sequence descriptors are the most commonly used in PCM models. This descriptors require the sequences of all the targets in the model to be aligned and, therefore, models using them are limited to proteins with a sufficient similarity in sequence or in structure[5].
The alignment-dependent sequence descriptors are computed using a large property matrix that describes all the individual amino acids, and performing a process of dimensionality reducing, usually by using the Principal Component Analysis (PCA).

2.5.1 PCA and alignment-dependent sequence descriptors

PCA is a technique used in the statistical field to simplify the description of the analysed objects, by reducing the number of their representative variables. The input is a matrix with rows corresponding to the objects and columns corresponding to the variables, called ‘original variables’, which describe the objects.
For the protein descriptors, the objects are the amino acids and the variables are more often physicochemical properties, but also topological or electrostatic ones. The aim is to extrapolate the variables showing the largest variances, assuming that, in PCM, as in all predictive approaches such as the Machine Learning methods, variables with largest variance are the most significant ones, since they are able to highlight the differences between the objects. Each variable in the

matrix is a vector described by a number of dimensions equal to the number of objects in the matrix. In order to catch the reciprocal influences of the variables, a linear transformation is applied to the matrix, by rotating the vectors. The output is a new matrix, including a new set of variables, called 'latent variables', ordered by the degree of variance. The plot which describes the reducing of the variance (from the first to the last latent variables) is very steep at the beginning and then it reaches a plateau. Each latent variable is obtained by combining some original variable, and is conventionally referred to the most important variable in terms of percentage.

Thus, if the original variables are different physicochemical properties (Pr_1, Pr_2, Pr_3... etc.), the obtained latent variables (PC_1, PC_2, PC_3... etc.) can still be considered as some specific physicochemical properties, based on how much each Pr_i contributes to the definition of PC_i.

To reduce the number of the latent variables, the best ranked variables are taken into account, discarding the other ones. This is done with different approaches, such as the choice a priori of the amount of total variance to be described ('cumulative percentage of variance explained' method) or the observation of the plot which describes the decreasing trend of the variance ('screen plot' method).

Table 2.1 shows the list of the amino acid descriptor sets used in this study and their main features.

For example, the Z-scales (5) are based on physicochemical descriptors and derived by PCA. After the dimensionality reducing, they take into account only 5 properties, which are the lipophilicity, the size, the polarity/charge, the electronegativity and the electrophilicity, with an amount of explained variance equal to 87%. This means that each amino acid is described by five values. For other protein descriptors, there are different amino acids, different original variables, different latent variables and a different amount of explained variance.

As shown in Table 2.1, Z-scales(3), Z-scales(5), T-scales and ST-scales descriptors cover a number of amino acids larger than the number of natural ones, which are

Descriptor set	Type	Derived by	n. of comp.	Variance explained	AAs covered
BLOSUM	Physicochemical and substitution matrix	VARIMAX	10	n/a	20
FASGAI	Physicochemical	Factor Analysis	6	84%	20
MSWHIM	3D electrostatic potential	PCA	3	61%	20
ProtFP (PCA3)	Physicochemical	PCA	3	75%	20
ProtFP (PCA5)	Physicochemical	PCA	5	83%	20
ProtFP (PCA8)	Physicochemical	PCA	8	92%	20
ProtFP (Feature)	Feature based	Hashing	n/a	n/a	20
ST-scales	Topological	PCA	5	91%	167
T-scales	Topological	PCA	8	72%	135
VHSE	Physicochemical	PCA	8	77%	20
Z-scales (3)	Physicochemical	PCA	3	n/a	87
Z-scales (5)	Physicochemical	PCA	5	87%	87

Table 2.1: Amino acid descriptor sets. Not available is abbreviated as n/a.

relevant in bioactivity models. Therefore, the PCA applied might select information that is not the most significant one to describe the space we are interested in, resulting in a lack of resolution.

There is a large arsenal of different descriptors and it is possible to group them on the basis of the type of variables (physicochemical, topological, electrostatic...), and of the technique used for dimensionality reducing. When considering only the first two principal components, descriptor sets can be clustered according to the approach by which they are derived, while when increasing the numbers of considered components, the clustering pattern shifts significantly. This indicates that including more principal components changes the descriptor behaviour, although the added principal components typically describe less variance than the first two components[26].

2.6 Machine Learning and Random Forest

Machine Learning has been described by Arthur Samuel in 1959 as the "*Field of study that gives computers the ability to learn without being explicitly programmed*". The focus of Machine Learning is the construction of algorithms capable of learning from data. In the context of medicinal chemistry, this refers to the capacity to build models able to predict properties of molecules learning from a set of input features.

Machine Learning is commonly divided into unsupervised and supervised learning approaches[1]. The unsupervised methods do not give any output, and their goal is to describe association and patterns among a set of input variables. An example can be a PCA applied to PCM data to reduce their dimensions. The supervised methods, on the contrary, aim at predicting the value of an outcome variable based on a collection of input variables. This is the case of QSAR or PCM models, in which the outcome variable is a measure of bioactivity, and the input variables are protein and ligand descriptors.

PCM relies on many different Machine Learning algorithms, including both linear and non-linear methods, and Random Forrest (RF) is one of the most efficient approaches.

RF is a highly versatile and all-purpose Machine Learning method, developed in the nineteens. It belongs to the large class of Machine Learning algorithms called "*ensemble methods*", which use multiple learning algorithms to obtain predictive performances higher than those reachable by exploiting separately the constituent learning algorithms. In other words, it builds many models, each model providing an independent prediction, and combines these predictions resulting in the final model. In the case of RF, the models are classification or regression decision trees.

The input data are collected into a matrix including a set of samples and of features or attributes. Each sample has different values for each feature, leading to a given

prediction, or label. The "*decision tree*" method essentially uses "if-then" statements to cluster the data and to obtain a final prediction. To evaluate the models, the "*decision tree*" methods generally use the so called mean decrease impurity (MDI, normally using the Gini impurity function), which roughly corresponds to the abundance of randomly incorrect selections within a predictive model. This means that in a suitably performed decision tree procedure, the random incorrect predictions should progressively decrease during the simulation and the so obtained final models can be ranked according to their resulting mean decrease impurity.

Hence, the "*decision tree*" method begins with the so-called "*root*", which corresponds to the best predictor, namely the feature that allows the greatest partition of the data in terms of "impurity" of the prediction. After the first partition, the tree is built along the subsets of "impure" data by adding nodes, which represent questions about the other features, and branches, which represent the outcome of the decisions. This process continues until all the groups are devoid of "impure" data and lead to a unique final prediction, which can be obtained by following the so generated decision tree. As said above, the level of "impurity" at each node gives a measure of the confidence of the final prediction. Figure 2.5 depicts a scheme of how Random Forest works.

The Random Forest algorithm automatically generates bunch of random decision trees, based on many subsets of the datasets. The final prediction comes from a process of averaging of all the collected predictions. It is interesting to note that most single predictions, around 99.9%, are not helpful for the final result, because they cancel each other out and only a minority of the single predictions (about 0.1%) contributes to the final result.

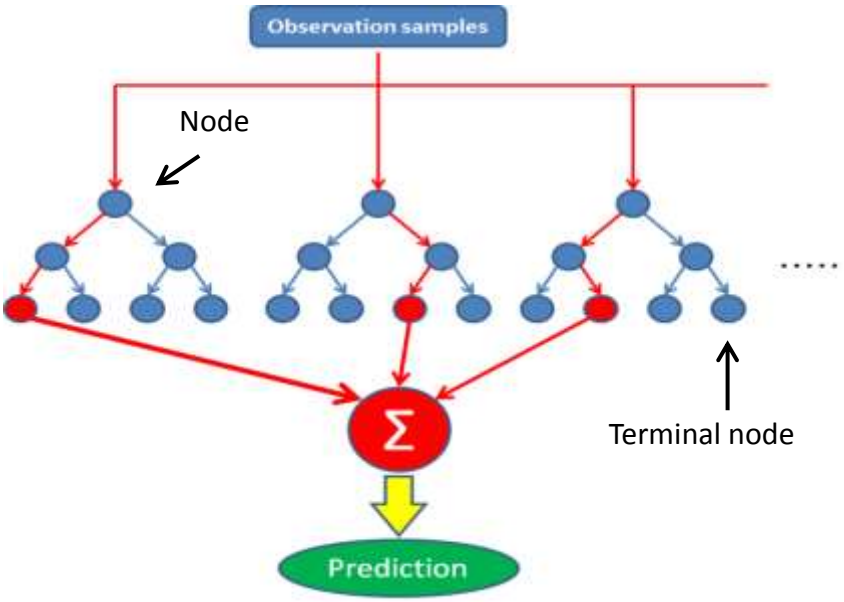


Figure 2.5: Random Forrest flow chart

2.7 Selection of data and building of the dataset

All the predictive models, including QSAR and PCM ones, are built on datasets and the first step involves the generation of a good database. Indeed, it has been proved that the quality of the descriptors in a dataset has much greater influence on the prediction then the nature of the model optimization techniques. The accuracy of the experimental data is a key prerequisite to calculate good descriptors[27]. The collection of data is done by searching from the available resources, which can include extremely large online databases, or small databases manually collected directly from the literature. Regardless of the utilized approach, the collected data should always undergo a manual curation, which can remove inadvertent mistakes or undesired features leading to substantial increase in the model predictivity.

This procedure consists in many "cleaning steps". First, inorganic compounds, counter-ions, organometallics, and molecules with rare elements should be removed, depending on the capability of the modelling tools to handle them. Secondly, some specific chemotypes as tautomers should be standardized, since they could be represented with different structural patterns, leading to different descriptors. Finally, the ionization state of all ionisable molecules should be defined by considering specific pH conditions.

Starting from a curated collection of data, it is possible to build the dataset by calculating properties and arranging them in a matrix. To be used in the model generation, the input dataset is modified to be compatible with the chosen Machine Learning technique. This process includes a dataset cleaning phase followed by some steps of data pre-processing, described in the following sections[28]. At the end of this process, the size of the input dataset is highly reduced.

2.7.1 Dataset cleaning

The cleaning and the plausibility checking of the dataset is a crucial step to avoid failures of the model in the following steps. First of all, the missing values in the features have to be checked and replaced by the mean of all values for that feature or, more likely, if the dataset is large, simply removed. Then, the dataset has to be checked for outliers, the features values of which are far away from the others. The outliers could be due to a mistake in data or can be seen as suggestive of an unusual behaviour, which might have a predictive power, even though the most common procedure when dealing with outliers is to remove the corresponding features. Indeed, a broad normal distribution for most features, as well as a wide range of the target values, are generally considered as the best situations for training a predictive model[29]. Moreover, at this stage, it is not possible to guess which

features offer the richest information about the target values, and make a decision on that basis. This information is though later derived by the predictive model itself.

2.7.2 Removing "Near Zero Variance" features

"Near Zero Variance" features are features whose values along the column show a very low amount of variance. Since "Variance" is a measure of how a set of values is spread out, columns which are found to have a near zero variance, contain nearly identical values. In other words, this situation corresponds to a column reporting values which differ very slightly from their average value. Including these columns in the model does not add any information and can, in the worst case, causes the model to fail. Therefore, these features, as well as those with zero variance, which include always a constant value, need to be removed. This procedure has some customizable parameters, which can be set to adapt the model to specific situations. One of those is the "frequency-cut-off", which corresponds to the ratio between the frequency of the most common value and the frequency of the second most common value. The most used frequency-cut-off is equal to 30/1. A second parameter is the "unique cut-off", which represents the percentage of unique values out of the number of total samples. Generally, a cut-off of 90% is chosen[30]. Another rather obvious parameter to control the deleting of features with "near zero variance" is the threshold of variance below which the columns will be discarded.

2.7.3 Feature normalization

Features with highly different scales can compromise the performances of the model, resulting in unrealistic predictions. By standardizing the input features, all features are weighted equally in their representation and similarly their role in the model is normalized. As a down-side, the variables are no longer in their original units and lose interpretability. The procedure to standardize the data consists in "centring" and "scaling" the input features. First of all, the mean and the standard deviation for each compound feature column are calculated. Then, the mean of a given feature is subtracted to every individual feature value to obtain the centring, and the resulting values are divided by the standard deviation of that feature to perform the scaling step.

2.7.4 Removing correlated features

Correlated feature are features which, while describing different characteristics, are influenced by some common properties and tend to be interrelated. While correlated features can impair the predictive power of correlative equations, the do not affect accuracy of classifications *per se*. Increasing the number of features typically increases classification accuracy, but this holds true until we are not *under-sampled* relative to the large number of features. Taking correlated features into account does not add any additional information to the model and causes a longer training time for the model. For this reason, the dataset analysis to search and delete the correlated features is usually recommended, as the last step of the pre-processing phase.

2.8 Model building

The key idea in the model building is that the prediction is trained by using the input database, which is a list of sample features with the corresponding labels, in a process that allows the predictive algorithm to learn from data. In other words, the model offers a plausible explanation of the relationships between data points. After that, the Machine Learning is able to exploit the learnt information to make a prediction about the labels of new samples. These labels are usually measures of bioactivity, for regression models, and binary labels, for classification models. A regression model is a function which maps objects to a real-valued outcome variable, whereas a classification model is designed to classify objects into two or more discrete classes[1].

The building of the model has to be performed fulfilling different aspects. Indeed, many parameters can compromise the predictive power of the model, and the accuracy of the original design is of crucial importance to obtain the best predictive performances. Very often, the ligand target interaction matrix is not complete, and the large variability both on the chemical and on the target sides can lead to wrong predictions. Therefore, the model building has to be based on a procedure of validation, chosen with respect to the size and the variability of the input dataset and to the kind of required prediction[31]. This ability can be tested on a subset of data appositively discarded from the dataset and assessed by comparing the predicted labels with the known values. The difference between the known values and the predicted labels in the test set determines the accuracy and the precision of the model.

That said, the ideal validation for any computational model, and also the only one that can be really true, is the prospective validation. The prospective validation uses new samples as the test set and the experimental determination of their activity to

assess model performance. Obviously, this is a not often feasible condition, and other kinds of validation are applied.

2.8.1 External and internal validation

The external validation consists in subdividing the input dataset in two subsets with different sizes, usually about 70% and 30%, respectively. The model is trained on the larger subset, called "training set", and then tested on the smaller one, the "test set", predicting the labels of the samples. In this case, because the subdivision is done randomly, it is likely that all targets and compounds are present in both the training and the test sets. This hypothesis depends on the number of different compounds and targets. This method is usually used to assess whether a reliable model can be generated for the dataset[32].

Very often, this method is completed by exploiting the so-called "*Nested Cross Validation*" (NCV), in which two validation loops are nested. The outer loop, which corresponds to the external validation, is run just after the inner loop is completed and has provided the best parameters. The inner loop is used to optimize the values of the Machine Learning parameters through the "internal validation". Among different types of internal validation, the "*k*-fold cross-validation" represents the current state of the art. This procedure consists in subdividing the training set into *k* equal subsets. Then, a model is trained on *k-1* subsets and tested on the remaining subset. This process is repeated *k* times, each time changing the composition of the test set, until all k subsets are selected as the test set. In each round, the best values of the parameters are selected, obtaining a set of best values, which change according to the variance of the different training set. Therefore, the k-fold cross-validation does not provide an absolute set of best parameters[33], but more likely a q^2 value and cross-validated RMSE[34].

2.8.2 Leave One Target Out

The "Leave One Target Out" method (LOTO) consists in excluding all the bioactivity data or the classification labels about one target out from the training set, and testing on it the predictive performances of the model. This corresponds to an external validation in which the test set is formed by all the data points belonging to one target. This procedure is repeated until all the targets are left out from the training set, whose composition is thus changing at all iterations. Differently from the external validation, in this case, the model is trained on a training set which does not include at all the target of the test set.

2.8.3 Leave One Compound Out

The "*Leave One Compound Out method*" (LOCO), similarly to the LOTO validation, excludes all the bioactivity data or the classification labels about one compound from the training set, using it to test the predictive performances of the model. This corresponds to an external validation in which the test set is formed by all the data-points belonging to one compound. This procedure is repeated until all the compounds are left out from the training set, the composition of which is thus changing at all iterations. Differently from the external validation, in this case, the model is trained on a training set which does not include any compounds of the test set. This situation perfectly corresponds to the scenario where a PCM model is applied to the selection of novel interesting chemicals. This is also the more likely situation in drug discovery, and the objective of this study.

If the number of compounds in the training dataset is large, the compounds can be grouped in cluster, performing the "*Leave One Compound Cluster Out*" method (LOCCO).

2.9 Model Evaluation

The efficiency of the predictive model is assessed by comparing the predicted values for the test set with the experimental ones. The evaluation of these performances depends on whether the modelling task is a classification or a regression. In a classification model, a common method for describing its performance is the Confusion Matrix (Figure 2.6).

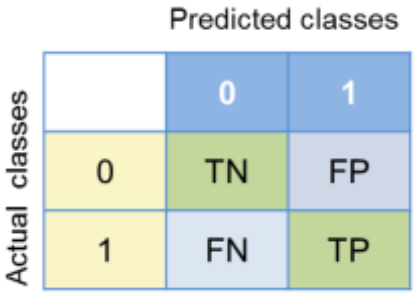


Figure 2.6: Confusion Matrix scheme for a binary classification

The Confusion Matrix for a classic binary model is a simple cross-tabulation of the observed labels: the actual and the predicted ones. The two classes are typically transformed in binary values, where 0 corresponds to the "negative" class and 1 to the "positive" class. In this way, "true positive" (TP) and "true negative" (TN) are the well predicted labels of both classes, and "false positive" (FP) and "false negative" (FN) are similarly the incorrect predictions in both classes. Looking at the matrix, the green diagonal cells denote cases where the classes are correctly predicted, while the other off-diagonal cells indicate the number of errors for both classes.

From the Confusion Matrix values, it is possible to calculate other metrics for the model performance.

The "error rate" is the percentage of a test set that is misclassified, while the "accuracy" is the percentage of a test set that is correctly classified.

$$error\ \ rate = FP + FN$$
$$accuracy = TP + TN$$

These metrics reflect the agreement between the observed and predicted classes and offer the most straightforward interpretation, but are not free of disadvantages. First, accuracy cannot distinguish between the types of errors which are made. Therefore, in applications where certain errors are to be considered more serious than others, weighting those errors is advisable.

Second, error rate and accuracy do not consider the frequencies or the percentages of each class. "*Recall*", also called "*sensitivity*", is the ratio of true positives on the total amount of actual positives. This measure is particularly interesting when one class is considered as the event of interest. In this case, recall corresponds to the rate that the event of interest is correctly predicted, for all samples having that event. In other words, it is the ability of the model to successfully identify the samples having the event of interest. The sensitivity is sometimes also called the "*true positive rate*" since it measures the accuracy in the event population.

$$recall = \frac{TP}{TP + FN}$$

In contrast, the "*specificity*" is the ratio of true negative on the total amount of real negative. In other words, it is defined as the ability of the model to correctly identify the samples not having the event of interest. The "false-positive rate" is defined as described below.

$$specificity = \frac{TN}{TN + FP}$$

$$false\ positive\ rate = 1 - specificity$$

Intuitively, sensitivity and specificity are inversely proportional, meaning that when a model increases its sensitivity, likely it will show a loss of specificity, since more samples are being predicted as events. For each situation, it is possible to establish a typical trade-off between sensitivity and specificity. This is particularly appropriate when dealing with types of errors with different penalties.

The "*precision*" is the ratio of the true positive on the total amount of the predicted positive.

$$precision = \frac{TP}{TP + FP}$$

Usually, precision and recall scores are not analysed individually. Indeed, either values for one measure are compared for a fixed level at the other measure (e.g. precision at a recall level of 0.75) or both are combined into a single parameter. As examples for metrics that are combinations of precision and recall, we can mention the F_1-score and the Matthews Correlation Coefficient (MCC).
The former corresponds to the harmonic mean of precision and recall, in which both parameters are evenly weighted.

$$F_1 score = 2\frac{precision \cdot recall}{precision + recall}$$

The latter, introduced by biochemist Brian W. Matthews in 1975, can be seen as a correlation coefficient between the experimental and predicted binary classifications. It takes into account all the values in the Confusion Matrix and it is generally regarded as a balanced measure which can be used even when the classes show very different sizes. It returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 corresponds to a random prediction and −1 indicates a complete disagreement between predicted and experiential classification.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$

Recall and Inverse Recall, namely true positive rate and false positive rate, are frequently plotted in the ROC curves which provide a well-established way to graphically represent the predictive performances of a given model.

Chapter 3 – Prediction of xenobiotic metabolism

3.1 Metabolism

The term “metabolism” comes from the Greek word “μεταβολή”, which means “change”. This change refers to the complex of chemical and physical transformations taking place within the cells of leaving organisms, to keep them alive. The changes result in assimilation of new materials, growth of cells, maintenance of homeostasis, production of life-sustaining energy and elimination of waste material.

Considering these changes in more detail, they can be seen a series of chemical reactions occurring on small molecules, defined as “substrates”, to make energy available and transform them into products, easily eliminated from the organism.

Going even more in-depth, we will recognize as mainly responsible for these changes the activity of the enzymes, which take a crucial role in metabolism. In effect, metabolic enzymes allow organisms to drive desirable reactions that require energy and will not occur spontaneously, by coupling them to spontaneous transformations, which release energy. Enzymes basically act as catalysts that help the reactions to proceed more rapidly, and the metabolic processes depend upon their activity. At the same time, enzymes cannot work unless the body is kept at a consistent temperature, therefore homeostasis must be maintained within the cells to allow the reactions required by metabolism to take place. This assumption closes the circle of interdependence between metabolism and homeostasis, on whose foundations is based the whole of life (Figure 3.1).

To try to look over the principles of metabolism and to figure out the rules on which they are founded is an ambitious porpoise, which always represents an extremely interesting topic for scientists. To put effort into the field of metabolism prediction, in particular, means to be involved in drug discovery from a comprehensive point of view, since all new drugs need to be screened based on

their pharmacokinetic profile, in order to maximize their bioavailability and, as a consequence, their efficacy.

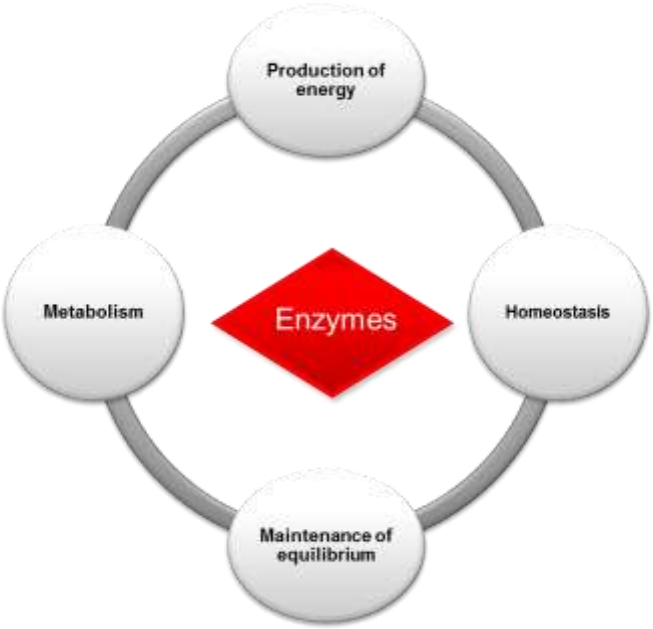


Figure 3.1: Role of enzymes in organisms. Enzymatic metabolism, thanks to the production of energy, contributes to the maintenance of homeostasis. Homeostasis, thanks to the maintenance of the internal equilibrium in the organism, allows the enzymatic metabolism.

3.1.1 Xenobiotic metabolism

The metabolic system has evolved as the main line of defence against what is considered "foreign to life": exogenous substances which do not possess nutritional or biochemical roles and are broadly defined as "xenobiotics". They include natural and synthetic pollutants, cosmetics, food additives, toxins coming from the secondary metabolism of bacteria and fungi, drugs and more generally, all synthetic chemicals. They are transformed into more hydrophilic metabolites which are usually (but not necessarily) less active or totally inactive and are always more readily excreted.

Focusing on drugs, the metabolic biotransformation alters the physicochemical profile of the starting molecules, resulting in an alteration of their physiological, pharmacological and toxicological profiles. This can mediate either a deactivation and detoxification of the original drug, or its activation - in the case of prodrugs - up to a conversion into a toxic effect. However, for about 75% of all drugs, metabolism represents the major clearance pathway[35].
The metabolic system is characterized by a complex organization, based on many different classes of enzymes, whose activities are meticulously coordinated and classifiable in two main phases: the functionalization and the conjugation. The enzyme expression patterns are highly adaptable and their substrates specificities are extremely variable among different species, as well as different tissues and organs in the same organism. Moreover, many other individual factors concur to define distinctions in the metabolic process, such as gender differences, genetic polymorphism, age, diseases, lifestyle, diet, intestinal flora, medications[36].

As a consequence, even though the new technologies and the available knowledge are rapidly advancing, an accurate prediction of drug metabolism remains extremely challenging.

3.1.2 Metabolism prediction

An exhaustive comprehension of the metabolic process at a molecular level is at the basis of a successful research in drug discovery. An unsuitable metabolic profile of lead compounds represents an important cause of failures and attrition in drug development[37]. An accurate knowledge of the metabolic properties of the hit compounds can lead to the optimization of the stability of new drugs.
Therefore, these last years have seen a continuous development of experimental methods for investigating the metabolic fate of the hit compounds. Among them,

High-Throughput methods (HTS) for metabolic screening are conveniently applied for large chemical libraries and in the early phases of drug development[38]. Unfortunately, these experimental approaches are often too demanding with respect to scientific equipment, human expertise, cost and time. Therefore, the interest around computational tools for metabolism prediction is increasing, since the virtual strategies offer a higher throughput at a lower cost, and can be applied to large chemical libraries at the early beginning of the discovery process[39]. Nevertheless, it is crucial to realize that there is an important potential synergy between experimental and theoretical approaches and, probably, the right way to best investigate drug metabolism can be found in the integration of the both areas[40].

The various *in silico* predictive approaches reported in literature can be subdivided into two main classes[41]. Local methods utilize classic ligand-based or structure-based approaches (e.g., QSAR models, pharmacophore mapping or molecular docking), to predict the outcome of a single metabolic reaction[42], whereas global methods use meta-analyses and expert systems to interrogate metabolic databases and extrapolate general rules or metabolic networks, in order to predict the entire metabolic pathway of a new molecule[36,43].

In spite of the mentioned advantages, computational approaches are affected by some drawbacks, which restrict their broad and truly successful application. In more details, local methods require to know a priori which metabolic reactions a given substrate can undergo, and which specific isozymes are involved in these reactions. Furthermore, structure-based approaches also necessitate the 3D-structure of the enzymes involved in the reaction. When all necessary data are available, local methods can provide accurate predictions, but it is evident that such requirements markedly reduce their general applicability. On the other hand, the global methods can suffer from inaccuracies hidden in their metabolic datasets, which are usually collected interrogating large online databases, such as PubChem, ZINC, ChEMBL, Drug Bank, WOMBAT.

Mistaken data in the metabolic databases markedly affect the reliability of the derived global predictions. In particular, the level of curation of the dataset has much greater influence on the prediction then the nature of the model optimization techniques[27]. What is now lacking in the available global methods is a good level of critical review of the retrieved data; a condition that can be filled only by a meticulous manual curation, which cannot be automatized.

3.1.3 MetaSar

The here presented project concerning the *in silico* metabolism prediction involved the development of a novel global method which aim at overcoming the inaccuracy drawback affecting other global methods. On these grounds, our approach is based on a meta-analysis and the resulting database, called MetaSar. MetaSar is an extension of a database previously published by Testa *et al*[44]., in which all the metabolic reactions are manually collected by critically reviewing the recent specialized literature. This accurate procedure is actually a continuous work in progress in our research group.

The selected specialized journals are "*Chemical Research in Toxicology*", "*Drug Metabolism Disposition*" and "*Xenobiotica*", and more than 1200 papers have been analysed in the period from 2004 to 2012. Currently, MetaSar comprises 1730 metabolic substrates, on which the metabolic reactions are annotated to obtain around 10 000 different metabolites.

Some specific rules are always followed during the collection of data. The focus is on drugs and other xenobiotics, excluding endogenous compounds; both studies in humans or in mammalians are taken in consideration, with distinctions weather they are carried out *in vivo*, in cellular system, or at a subcellular or enzymatic level; both metabolic trees and single enzyme studies are analysed. Each substrate

is considered separately, avoiding duplicates. Regio-isomers and stereoisomers are treated as distinct substrates or metabolites, to underline eventual substrate or product selectivity. Most importantly, all selected papers have to report convincing analytical conditions, and all the collected data are critically reviewed. In MetaSar, all the substrates are reported with annotated the sites of metabolism and the specification of the class of reaction for each of them. Moreover, in the database, other levels of classification are included: the enzyme super-family involved in the reactions; the metabolic generation to which each metabolite belongs (1^{st}, 2^{nd}, 3^{rd} or plus); the annotations of which metabolites are pharmacologically active, reactive and/or toxic.

Besides the MetaSar curation, the metabolism prediction involved the preparation of the three-dimensional structures for some important metabolic enzymes, both derived from resolved crystals and homology modelling. This research field started with some predictive studies on the enzymes involved in the hydrolytic metabolism (CES1, CES2, hBChE, PON1 and PON2) to arrive to the homology model of the human UGT2B7, whose optimization is presented in the following sections of this chapter.

My personal contribution regarding MetaSar project involved the update of a previously existing classification of reactions to a new one, including now 21 classes and 101 subclasses of reactions. Moreover, I contributed to the enrichment of the database, by adding 3D structures of new substrates and metabolites. Finally, I personally curated the preparation of the majority of the enzymes structure included.

3.1.4 Possible applications of MetaSar

The MetaSar project was initially undertaken to collect a database suitable to develop global approaches, the main objective of which is the generation of predictive algorithms able to model the entire metabolic pathway of new molecules. Beside such a key project, MetaSar can find other interesting applications.

MetaSar can be an exhaustive source of data for metabolism-driven virtual screening studies, aimed at validating predictive models of single metabolic reactions. Indeed, the MetaSar substrates can be seen as an appropriate dataset for enrichment studies, since they include a minority of "active" substrates on a specific target (enzyme family), while the other part can be seen as a good collection of "inactive" decoys. This approach can be partly questionable, since some molecules can be considered as putative non-substrates simply because the corresponding metabolite was not searched in the reviewed study. Nevertheless, for the most studied metabolic reactions/enzymes (such as CYP450 or UGTs), or for those reactions/enzymes requiring substrates with well-defined reactive groups (e.g., hydrolysable functions for hydrolases), MetaSar can be seen as an effective database for virtual screening campaigns, and indeed it was successfully utilized to validate predictive models of the metabolism by Human Carboxylesterases.

Furthermore, MetaSar is also a well curated source of data for PCM studies aimed at modelling the regioselectivity of the metabolic reactions occurring to a specific class of enzymes. We present in the following chapters a first example of this application, regarding the prediction of glucuronidation reactions.

3.2 UDP-Glucuronosyltransferase

UDP-glucuronosyltransferases (UGT) are recognized as the most important non-P450 enzymes due to their relevant contribution to metabolism of endogenous and exogenous chemicals.

They are responsible for the majority of the phase II metabolic reactions, and they are the most important conjugation enzymes in terms of number and diversity of substrates.

UGTs are implicated in the last and crucial step of metabolism: the conversion of sufficiently polar molecules, drugs and other xenobiotics, or their first metabolites, into more readily excreted hydrophilic products, inducing the definitive interruption of their biological activity. Considering that the current trend in drug development strategy is to focus on new chemical entities with a lower lipophilicity, therefore ready for phase II metabolism, the non-P450 enzymes become the most prominent players to the clearance of drug candidates. Accordingly, there is an ever increasing quest for a better understanding of UGT, both at a functional and molecular level, since they can represent a key issue in drug development[45].

3.2.1 Historical notes

The first compound ever characterized as a sugar conjugate has been the Euxanthic acid, isolated in 1855 in the urine of cows fed mango leaves, and responsible for the colour of the famous dye Indian yellow. After that, almost 100 years have elapsed before the enzymatic nature of glucuronidation has been established and in 1953 the process of formation of Euxanthic acid was elucidated since the Uridine

Diphosphoglucuronic acid (UDPGA) was identified as the co-factor in the UGT activity.

In the period 1950–80, our knowledge about glucuronidation has been largely derived from *in vitro* studies. Cell fractionation experiments demonstrated that UGTs were integral membrane proteins residing in the Endoplasmic Reticulum and nuclear envelope. UGT enzymes were found in many tissues and organs, but predominantly in the liver and intestinal tract. Tissue homogenates and purified enzyme preparations have been used to elucidate many of their key features, such as the influence of the membrane environment for an optimal activity, their catalytic properties and capacity to react with a large number of lipophilic chemicals, their differential inducibility by compounds, such as barbiturates and polycyclic aromatic hydrocarbons, and their property of latency (i.e., the requirement for disruption of the membrane bilayer to achieve a maximal activity). The same studies also provided some progresses toward understanding the multiplicity of the UGT family and the substrate preferences of individual UGT isozyme.

In the 1980s, thanks to the advent of the UGT cloning and expression, a substantial progress was made in elucidating UGT multiplicity, substrate preference, and structure–function relationships and in determining the mechanisms underlying the regulation of the expression of the *UGT* gene. In particular, the first cloning studies were carried out in rodents and provided the first amino acid sequence of a UGT, establishing that there were several UGT isoforms encoded by a superfamily of genes. After that, many cDNA expression systems have been developed with the aim of defining the substrate specificities of each UGT isozyme, and very soon we assisted to the cDNA cloning of human UGTs, quickly followed by the identification of all human UGTs by the human genome project[46].

The nascent UGT polypeptide measures about 530 amino acids and contains a signal peptide, which is involved in integrating the protein into the endoplasmic reticulum, and a highly hydrophobic stretch of 17 amino acids near the C-terminus which most likely traverses the lipid bilayer.
The mature protein is orientated on the luminal side of the Endoplasmic Reticulum, where catalysis occurs. It forms two domains: the N-terminal domain, which is involved in substrate recognition, and the more conserved C-terminal domain, which binds the co-factor, the UDP glucuronic acid. Due to this particular localization and in order to induce a maximal glucuronidation activity, the rupture of the membrane has to be performed by using detergents with controlled low concentrations or other mechanical procedures such as sonication.

Nowadays, UGTs and their metabolites have been identified in a wide range of vertebrate species, including humans, other primates, other mammalian species, birds, and fishes, as well as in non-vertebrates, plants, and bacteria. The research has focused on vertebrate species, leading to the isolation of multiple UGT isoforms and to the identification of an extensive range of lipophilic chemicals that are metabolized by these enzymes. On one hand, each enzyme has its own distinct set of substrates; on the other hand, many compounds are recognized by more than one UGT enzymes, revealing a considerable degree of overlapping between the UGT substrate specificity. This redundancy in the glucuronidation system reduces the impact of genetic and regulatory aberrations[47].

3.2.2 UGT nomenclature and gene organization

The UGTs are members of a broader superfamily of UDP-glycosyltransferases, enzymes that transfer glycosylic groups to lipophilic substances from a variety of UDP sugars. All UDP-glycosyltransferases, including UGTs, share a 44 amino acid

characteristic sequence signature in their C-terminal domain that appears to be involved in the binding of the UDP moiety of the nucleotide sugar.

The gene superfamily contains four UGT families: UGT1, UGT2, UGT3 and UGT8[48]. Among these, the first two families are the most important and share the same UDP sugar specificity, using preferentially UDP glucuronic acid, and less frequently other sugars, including UDP glucose and UDP xylose. The UDP sugar specificity of the UGT3 family is still unknown, whereas UGT8A1 utilizes UDP galactose as the sugar donor. While the enzymes of the families UGT1 and UGT2 play a notable role in detoxifying both endogenous and exogenous chemicals, other UGTs have a specific biosynthetic role, as exemplified by UGT8A1 which is involved in the synthesis of cell-membrane components[49].

Our interest is focused on the two major families of mammalian UGTs, UGT1 and UGT2. In general, members within each family share more than 45% of sequence identity but are less than 45% when compared to UGTs of the other family. The UGT2 family has been further divided into subfamilies, UGT2A and UGT2B, the members of each sharing more than 70% of amino acid identity.
To date, humans are known to possess 19 functional UGT isozymes, including 9 members of the UGT1 family (known as *UGT1A* genes), 3 members of UGT2A subfamily and 7 members of the UGT2B subfamily. The dendogram depicting the sequence similarity of human UGTs is shown in Figure 3.2.

All the human *UGT1A* genes are found on a single locus on chromosome 2q37, spanning approximately 200 kb. This locus contains 13 unique exons, which encode for the N-terminal domains of 13 potential UGT1A forms, and 5 exons that are shared by all the full-length UGT1A transcripts. As a result, the UGT1A enzymes possess an unique N-terminal region, which provides functional diversity

and substrate selectivity, while the C-terminal domain is identical for each enzyme, and is involved in the binding of the sugar donor, namely the UDP glucuronic acid.

On the opposite, the *UGT2* genes are encoded by almost all discreet genes, with the only known exceptions of *UGT2A1* and *UGT2A2*, which share exons in a manner similar to the *UGT1A* locus. All the *UGT2* genes are located on chromosome 4 at position 4q13. The *UGT2B* genes consist of six exons and share similar intron/exon boundaries, although intron lengths vary between genes. Despite being separate genes, the C-terminal domains of the UGT2B forms, which bind the UDP glucuronic acid, are still highly conserved, both within the UGT2B family and, to a lesser extent, with the UGT1A subfamily[50].

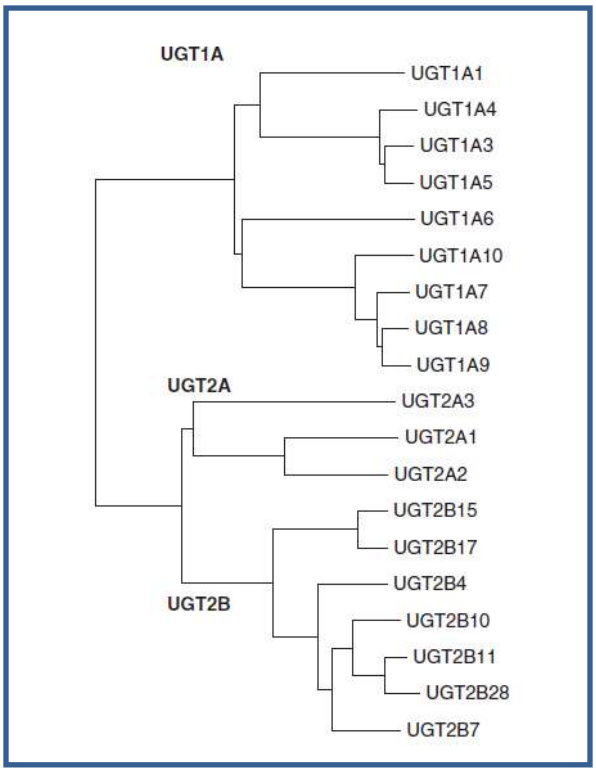


Figure 3.2: Dendrogam of UGTs subfamilies.

3.2.3 UGT structure

UDP-glucuronosyltransferases belong to one of the main families of the large glycosyltransferase superfamily, the GT-B fold[51]. These enzymes catalyse the formation of glyosidic bonds using generic sugar donors and comprise proteins conserved across several species. Enzymes belonging to plants and bacteria have been crystallized, and, despite a low percentage of identity within the superfamily, the structural similarity is good enough to delineate the main features of the mammalian UGTs. These last enzymes have not been completely resolved yet, due to an additional transmembrane segment that complicates the crystallization process making its experimental conditions more delicate[52,53].

The UGT's structure consists of two domains: the N-terminal domain, responsible for the binding of the substrate, which accepts the glycoside moiety, and the C-terminal domain, which binds the cofactor, the donor of the glycoside group. The active site is, indeed, formed by a deep cleft at the interface of these two domains. Each domain is characterized by a Rossmann fold motif, including six parallel β-sheets separated by seven α-helices. Two important modifications distinguish the mammalian UGTs structures from the corresponding plants and bacterial crystallized enzymes: (1) the above mentioned hydrophobic C-terminal region, that links the enzyme to the membrane of the Endoplasmatic Reticulum, and (2) a part of the C-terminal region that is crossing back to approach the N-terminal domain.

In 2007, Miley *et al.* published the crystal structure of the C-terminal domain of UGT2B7 that, up to now, remains the only one ever resolved. The cofactor is not included in the crystal, but some insights can be deduced from a superimposition of the 2B7CT structure and the plant glucosyltransferase VvGT1. The key interactions stabilizing the binding with the cofactor can be schematized as follows: (a) Asp398 and Gln399, which elicit H-bonds with the glucuronic acid moiety, (b) Asn378 and

His374, which interact with the phosphate groups, (c) Glu382, whose backbone atoms form a H-bond with the ribose moiety and (d) Trp356 and Gln359, which interact with the uracil base[54].

The N-terminal domain of UGT2B7 can be obtained through homology modelling using some plants crystals as the templates, even though it is less conserved than the C-terminal region, and so its modelling cannot be an easy task. Interesting insights about the residues which form the substrate binding site arise by the comparison with the templates. Two important residues in the mechanism of action have been identified, namely His35 and Asp151. They are highly conserved in vertebrates, probably mediate the nucleophilic attack of the substrate on the cofactor, promoting the deprotonation of the substrate. This event is not always necessary and UGT1A4, for example, has a Proline (Pro40) instead of the histidine and it still able to catalyse glucuronidation reactions on primary, secondary, and tertiary amines[54]. Furthermore, replacement of Pro40 in UGT1A4 with a histidine residue greatly reduced the enzyme activity on secondary and tertiary amines but enhanced its capacity to react with phenols and carboxylic acids[55]. Similarly, UGT2B10 has a leucine instead of the histidine, consonant with its ability to N-glucuronidate cotinine and nicotine, despite showing very low rates[56]. These studies emphasize that His35 plays a key role in determining substrate selectivity in UGT2B7, and the same holds true for the corresponding residues in the other human UGTs.

3.2.4 UGT catalytic mechanism

UGTs catalyse the covalent linkage of the glucuronic acid from the high energy UDPGA cofactor on lipophilic substrates, to form D-glucuronides. The cofactor is produced endogenously by oxidation of UDP-α-D-glucose. This process is very abundant in the adult human body (about 5 grams daily synthesized), confirming the high capacity of this metabolic route[57]. The list of electron-rich functional groups which are potential substrates of the reaction is rather wide as summarized in Figure 3.3.

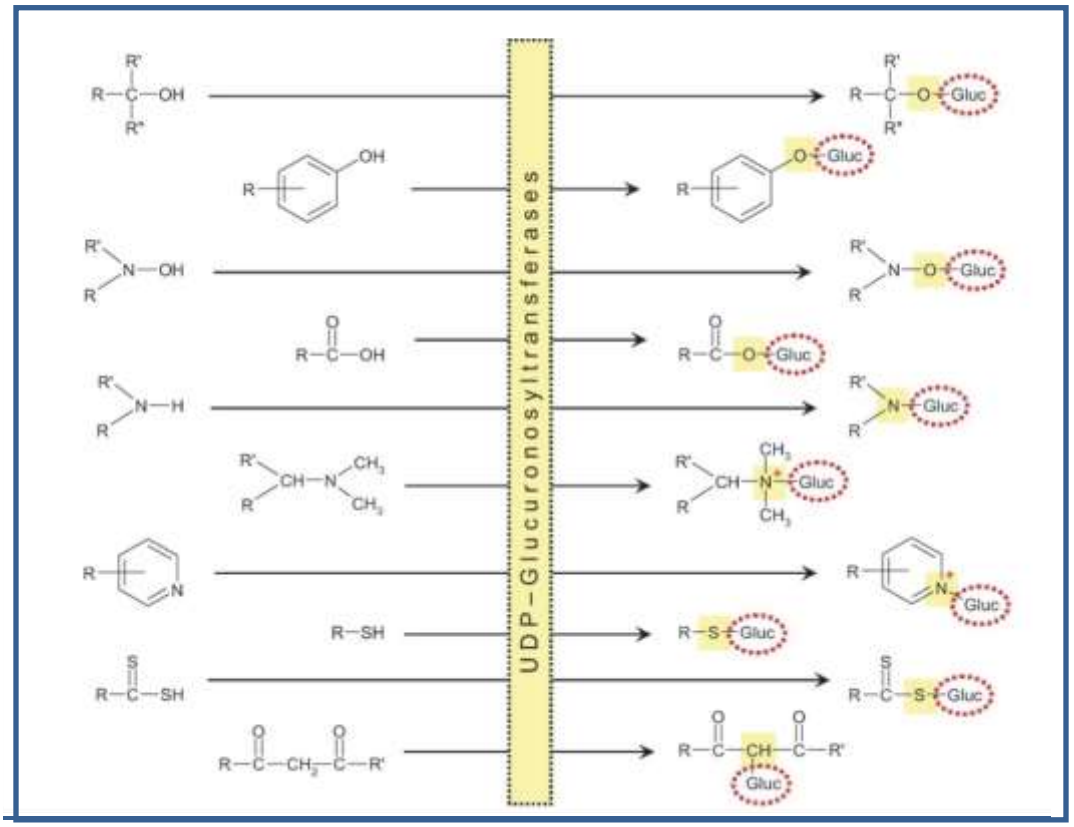


Figure 3.3: Summary of the functional groups being potential substrates for UGTs.
(Reproduced from [58])

The reaction is a second order nucleophilic substitution (SN_2), and the scheme of the reaction is reported in Figure 3.3.

The nucleophilic group on the substrate is often activated by proton abstraction with a general base in the UGT catalytic site. Then, the electron-rich group on the substrate attacks the anomeric carbon atom of UDPGA and induces the release of UDP. The degree of activation by proton abstraction varies depending on the stability of the charged product of the reaction[53]. For example, deprotonation of the hydroxyl groups is required because the formation of the positively charged oxonium ion has a large energy barrier. In contrast, deprotonation of amino groups is often not necessary as charged amines are relatively stable. The glucuronidation reaction relies on a single displacement mechanism, and requires that the attacking nucleophilic group and the leaving UDP molecule should be arranged on either side of the anomeric carbon atom and approximately in line. The nucleophilic substitution made by UGTs induces the inversion of the configurations of the anomeric carbon atom from α-UDPGA to β-glucuronides[53]. Their acidity is an important feature of glucuronides: the pK_a of glucuronic acid is 3.0 and that of O-glucuronides in the range of 2.9-3.1, implying nearly complete ionization at physiological pH range. The N-glucuronides show a special feature since, as a consequence of the loss of deprotonation, they have a permanent positive charge in addition to the negative charge of the carboxylate, namely they are zwitterions[57].

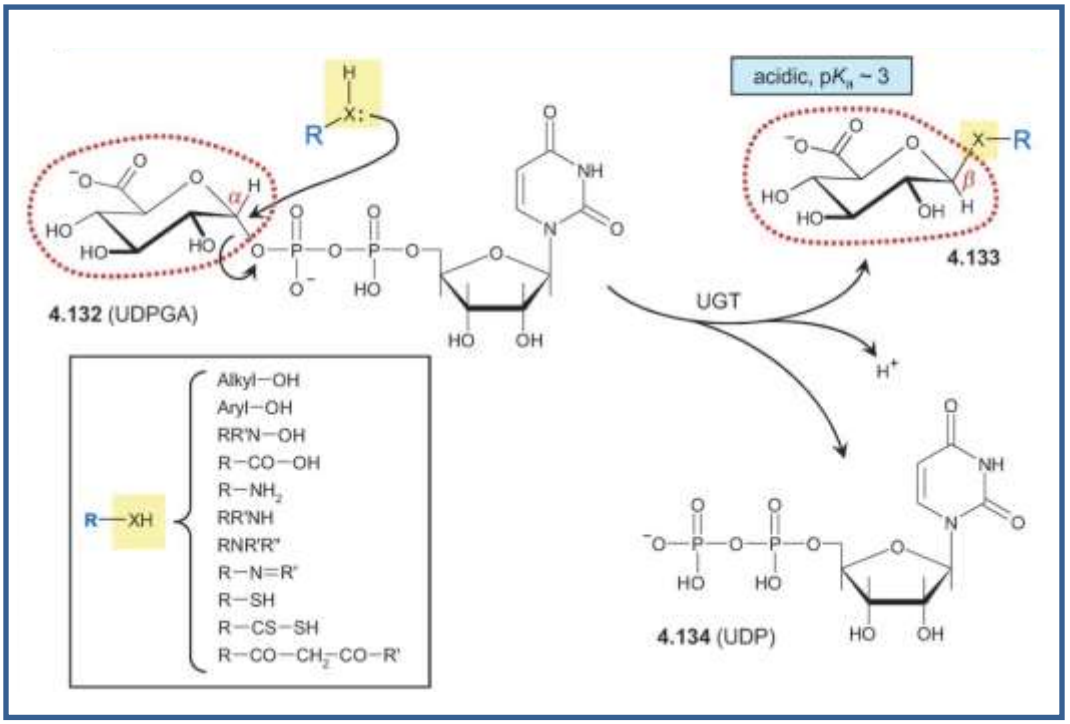


Figure 3.4: UGTs catalytic mechanism of action. (Reproduced from [58])

3.2.5 UGT substrate selectivity

A large number of structurally heterogeneous metabolites are metabolized by glucuronidation, including endogenous compounds, such as a variety of androgens (testosterone, andrsterone, epiandrosterone), estrogens (bestradiol, estriol) and gestagens (17a-hydroxyprogesterone), biliary acids (lithocholic acid, deoxycolic acid, chenodeoxycholic acid)[58], bilirubin, thyroid hormone, the heme breakdown product, neurotransmitters, fatty acids, and eicosanoids, as well as drugs and dietary chemicals. Considering that there are only 18 functional isoforms of human UGT, each isoform can likely glucuronidate different classes of chemicals. Furthermore, a single chemical undergoes glucuronidation by more than one isoform.

The determination of the substrate selectivity of each UGT proceeds along with the discovery of new classes of metabolites and takes advantage from recombination studies involving mammalian or insect cells. However, the establishment of the accurate substrate profile for each UGT isozyme remains difficult, mainly because of differences in the utilized recombinant systems for expressing UGTs and in the assay sensitivities and methodologies[59]. In addition, many computational studies indicate that UGTs exhibit distinct, but overlapping, substrate selectivity[59,60].

Most UGT enzymes (with the exception of UGT1A4) are able to metabolize small alcohols and phenols but, when increasing the structural complexity of substrates, the selectivity appears to increase, due to steric, hydrophobic, and electronic factors. For example, bilirubin is glucuronidated only by UGT1A1, trifluoperazine by UGT1A4, and zidovudine by UGT2B7[59], and many hydroxysteroids exhibit distinct enzyme selectivity[61].

Indeed, selective substrates have now been identified for most UGT enzymes, and this information can be useful for reaction phenotyping. One of the best techniques to investigate the selectivity is the use of "probe" substrates, when they have been identified (see Figure 3.5). The inhibition of metabolite formation obtained by an enzyme selective inhibitor on enzymes of human liver microsomes or hepatocytes provides strong evidence for the involvement of that enzyme in the metabolic pathway. However, few UGT enzyme selective inhibitors have been hitherto identified. For example, hecogenin is a highly selective inhibitor of UGT1A4[62], and fluconazole shows moderate selectivity as an inhibitor of UGT2B7[63].

UGT	Catalytic capacity	Preferred substrates	Enzyme-selective 'probe' substrates for liver UGTs
1A1		Bilirubin Phenolic groups on large molecules	Etoposide
1A3		Carboxylic acids Primary and secondary amines	Chenodeoxycholic acid *R*-lorazepam F-1α, 23S, 25(OH)$_3$ vitamin D3
1A4		Primary, secondary, and tertiary amines	Trifluoperazine
1A5	Low	Phenols	
1A6		Phenolic groups on small molecules	Serotonin
1A7		Hydroxyl groups on large molecules	
1A8		Hydroxyl groups on large molecules	
1A9		Hydroxyl groups on large molecules C-glucuronidation of sulfinpyrazone and phenylbutazone Carboxylic acids	Propofol Sulfinpyrazone
1A10		Hydroxyl groups on large molecules (Heterocyclic amines)	
2A1		Phenolic odourants (e.g., caracrol, eugenol)	
2A2		Unknown	
2A3	Low	Unknown	
2B4	Low	Hydroxyl groups on bile acids, Steroids	None
2B7		Hydroxyl groups on bile acids, Steroids (mainly the 3-OH of androsterone and 3α,17β androstane diol) Many drugs, especially opioids and carboxylic acids	Zidovudine Morphine (6-OH)
2B10	Low	Primary amines (nicotine, cotinine)	None
2B11	Low	Fatty acids	None
2B15		Androgen OH groups (mainly 17-OH groups), dihydrotestosterone	*S*-lorazepam *S*-oxazepam
2B17		Androgen OH groups (mainly 17-OH groups), dihydrotestosterone and testosterone	None
2B28	Low	Steroid OH groups	None

Figure 3.5: Catalytic capacity and preferred substrates for human UGTs
(Reproduced from Mackenzie *et al.*[64])

3.2.6 UGT role in toxicity and clinical significance

The glucuronidation metabolic reactions are primarily a detoxification pathway and a protective mechanism to prevent the accumulation of highly hydrophobic chemicals. Indeed, UGT enzymes modulate toxicity of drugs and other chemicals because the addition of a glucuronic acid moiety to a given compound enhances its polarity and thus its rate of elimination from the body and as a trend decreases its toxicity.

Nevertheless, the possibility to have metabolites more toxic or biologically active than their parent compounds is also true for glucuronides. A first example is represented by steroid alpha-D-ring glucuronides, which are more cholestatic than beta-D-ring glucuronides[65]. A second example is given by the acyl glucuronides of drugs that can react with cell constituents forming adducts which induce apoptosis[66]. These drugs include clofibric acid, benoxaprofen, bezafibrate, and probenecid, which cause DNA nicking through their acyl-glucuronide metabolites[67]. Even though glucuronidation as a rule abolishes the biological activity of metabolized drugs, there are relevant examples of drugs the glucuronides of which maintain or enhance their bioactivity. For example, the morphine-6-glucuronide[68] is an analgesic markedly more potent than morphine, and the retinoic glucuronide, acts as a chemopreventive agent in breast cancer and is more potent than the parent retinoic acid[69].

Also other negative events, such as adverse drug reactions, altered drug efficacy, and outcomes of organ transplantation, are associated with alterations in UGT function or variability in their expression. There are several examples of relationships between the outcome of pharmacological treatments and the patient's genotypes for UGT enzymes.

One of the best-known examples regards the association between Irinotecan, a widely exploited anticancer prodrug, primarily used to treat colorectal cancer, and UGT1A1. The metabolism of Irinotecan is firstly regulated by carboxylesterases enzymes, which transform the prodrug in its therapeutically active metabolite 7-ethyl-10-hydroxycamptothecin (SN-38). SN-38 has a narrow therapeutic window, since over-dosing can cause life-threatening toxicities including diarrhoea and neutropenia, and its deactivation and elimination is mostly controlled by glucuronidation[70]. In particular, UGT1A1, UGT1A9, and UGT1A7 have been proposed to be the major catalysts of SN- 38 glucuronide formation[71]. Indeed, genetic variations of the underling isoforms that lead to a decrease of SN-38

glucuronidation have been associated with altered treatment outcomes. The association of Irinotecan-mediated toxicity with the UGT1A1 allele *UGT1A1*28*[72,73], in particular, is so clearly established that the American Food and Drug Administration added recommendations for testing of the UGT1A1 genotype of patients prior to irinotecan treatment.

Another example involving UGT1A1 isoform is given by the association between Indinavir and the development of severe hyperbilirubinemia in Thai HIV patients carrying the *UGT1A1*6* or *UGT1A1*6* plus *UGT1A1*28* genotypes, but not *UGT1A1*28* alone[74]. In addition, inhibition of UGT1A1 by Indinavir results in an additive effect in patients with already impaired bilirubin glucuronidation activity. A similar gene–environment interaction might be predicted for other inhibitors of the UGT1A1 activity.

Other examples where UGT genotype may affect the outcome of pharmacological treatments involve other isoforms. One of them regards UGT1A6 enzymes, which play a role in the metabolism of Aspirin. In particular, some evidences associated the low-activity of *UGT1A6* with a greater protective effect induced by aspirin on the risk of developing colorectal adenoma, while individuals carrying the wild-type of *UGT1A6* do not benefit from the protective effect[75]. However, a conflicting study showed that low-activity *UGT1A6* genotypes are protective against colorectal adenoma recurrence irrespective of aspirin intake[76]. Thus, the relationship between *UGT1A6* genotype, aspirin, and colorectal adenoma remains controversial, particularly given the apparently minor role of the UGT1A6 isoform in the salicylic acid glucuronidation.

Another example involves UGT1A9, and possibly UGT2B7, which are responsible for the glucuronidation of mycophenolic acid, an immunosuppressant characterised by considerable inter-individual variations in its pharmacokinetics[77]. In healthy volunteers, the UGT1A9*3 and the UGT2B7*2 alleles have been associated with

alterations in mycophenolic acid exposure, enterohepatic recycling, and production of the toxic acylglucuronide metabolite[46].

Similarly, UGT2B15 is an important enzyme for the metabolism of the benzodiazepines Oxazepam and Lorazepam, used as anxiolytic and hypnosedative. The *UGT2B15*2* variant appears to be associated with lower glucuronidation of Oxazepam in human liver and lower clearance of Lorazepam in healthy volunteers[78,79], thus influencing their dosage and their pharmacological effects [78].

Finally, the *UGT2B17* gene has been found connected to the transplant-related mortality in recipients of hematopoietic stem cells. In detail, the UGT2B17 protein seems to be immunogenic in individuals that are genetically devoid of the UGT2B17 gene, and may be responsible for a heightened risk of complications in recipients given transplants from donors mismatched for UGT2B17[80].

3.2.7 UGT: the current situation and future perspectives

Considering the large importance of glucuronidation reactions in the metabolic process, the key aspects of the activity of the UGT enzymes, some of them still unclear, require an in-depth research. The lack of crystal structures for mammalian UGTs represents a crucial drawback, and its resolution remains a high priority in the field. In the meanwhile, the exploitation of reliable homology models for the mammalian UGTs, as generated using the available plant and bacterial templates, remains the best option.

Many aspects need to be further investigated; first, how UGT enzymes' activity is organized and the precise mechanism of their catalysis. In detail, the role of UGTs oligomerization and their complexes with other drug-metabolizing enzymes and possibly transporters should be clarify, in order to determine whether it is necessary to deliver substrates and sugar nucleotides to the active site, and then to remove the

glucuronidated metabolites. It is important to define which other proteins are parts of these complexes and how their formation is modulated. Moreover, the impact of phosphorylation and other post-translational modifications on glucuronidation capacity should be elucidated, as well as how the UGT proteins are integrated in the endoplasmic reticulum environment.

The identification of selective inhibitors for each UGT remains a crucial requisite for reaction phenotyping. The use of probe substrates can assist the *in vitro*/*in vivo* correlations, but the selectivity of the underling probes is sometimes not enough to assure a high binding rate to the UGT of interest, compromising their use in assays (e.g., serotonin as a probe for UGT1A6). Another strategy to characterize the distribution and profile of UGTs in human tissues is based on antibodies which show specificity for each UGT, as recently demonstrated in the kidney[81].

A complex system of co-activators and repressors determines the levels of UGTs in tissues and organs during development and in response to hormones and other external stimuli. A more extensive study of this system and the UGT regulation can clarify the relationships between changes in UGT expression and risk of chemical toxicity and diseases, and can suggest how to manipulate these metabolizing enzymes for therapeutic applications.

Finally, considering that the role of UGTs in the pathway of uptake, metabolism, and egress of drugs from cells is intrinsically interconnected with the role of the other metabolic enzymes, only a better understanding of the entire metabolic process and the coordination between the various metabolic pathways can lead to significant progresses in drug discovery.

3.3 Modelling studies on UGT2B7 catalytic site

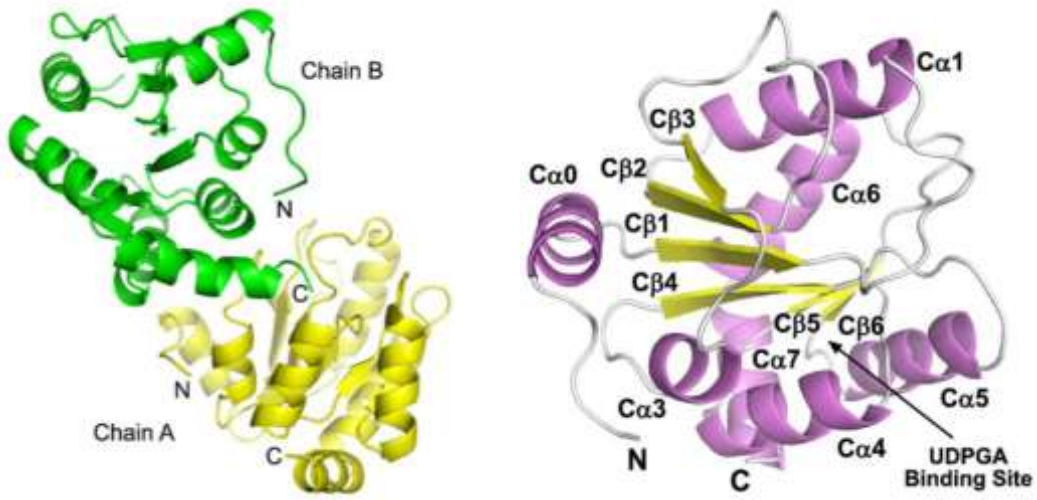
UGB2B7 is the most important UGT isoform as it is responsible for the glucuronidation of 35% of clinically used drugs[45,82]. Among the main examples of metabolized drugs, we can mention opioids, including morphine, codeine, buprenorphine and naloxone[83,84], anti-cancer agents[85,86], gemfibrozil[87], valproic acid and other carboxylic acid containing drugs[88], and anti-viral drugs including zidovudine[83]and efavirenz. Moreover, UGT2B7 is also found to be implicated in metabolism and detoxification of anti-inflammatory agents, as S-Naproxen and other non-steroidal drugs[88,89], as well as the mineralocorticoid aldosterone and other C19 and C21 hydroxy-steroids[90,61]. In addition, UGT2B7 metabolizes endogenous compounds, including bile acids, fatty acids, and steroids[45].

For these reasons, the research around glucuronidation enzymes, up to now, has been mainly focused on human UGT2B7, unanimously considered the most interesting enzyme within UGT family[45]. A better understanding of its structural features and catalytic mechanism can represent the right starting point for a correct interpretation of the drug glucuronidation processes.

3.3.1 UGT2B7 structure: state of the art

UGT2B7, as all human UGTs, belongs to the GT1 family and is predicted to adopt a GT-B fold. The GT-B structural organization consists in two Rossman-like domains, associated to form a catalytic cleft at their interface. The C-terminal domain contributes to the majority of the contacts with the donor cofactor, whereas the N-terminal domain is responsible for the main interactions with the acceptor substrate. Nevertheless, a strong cross-talk between both domains is recognized to be important to determine the final shape of the binding site[91].

In 2007, Miley *et al.* published the first and the only ever resolved structure of a mammalian UGT: the 1.8-Å resolution apo crystal structure of the UDPGA binding domain of human UGT2B7 (see Figure 3.6)[54]. The crystal structure includes an asymmetric dimer of the UGT2B7 C-terminal domain, consisting in two nearly identical molecules. Each molecule comprises the residues 285-472 but lacks in the final C-terminal region, which has a transmembrane segment between the residues 493-509, anchoring the enzyme to the luminal side of the Endoplasmic Reticulum. The globular domain consists in a β-sheet core formed by six individual strands, surrounded by seven α-helices.



(a) (b)

Figure 3.6: Overall structure of UGT2B7 C-terminal domain. (a) Asymmetric dimer of 2B7CT. (b) Ribbon cartoon of 2B7CT with labelled secondary structure elements.
(Reproduced from Miley *et al.*[54])

The cofactor is not included in the crystal, but some insights about his binding site can be derived from a superimposition of the 2B7CT structure and the plant flavonoid glucosyltransferases VvGT1 and UGT71G1[52,92]. These sequences reveal a high level of structural homology, despite a lower sequence identity (~.19%). In particular, the putative UDPGA binding site sequence of UGT2B7 is remarkably similar to the UDP-glucose binding site of VvGT1. A more complete structure-based sequences alignment of representative GT1 family enzymes is shown in Figure 33.7.

These observations suggested that the cofactor can be modelled in the UGT2B7 predicted binding site using the VvGT1 crystal as the template. Moreover, the sequence alignment shows that UDPGA binding site is highly conserved also among human UGTs, suggesting a common binding mode for all the isoforms.

The mechanism of action, already hypothesized by studies with selective inhibitors[93] and site-directed mutagenesis[54], finds a confirmation also in the structure-based sequence alignment. It consists in a serine hydrolase-like catalytic triad, in which Asp151 stabilized the protonation of His35, which, in turn, activates the electron rich substrate by proton abstraction, here holding the role of serine in the hydrolase enzymes. The substrate then attacks the anomeric carbon of the cofactor by a second order nucleophilic substitution, leading to the conjugation reaction.

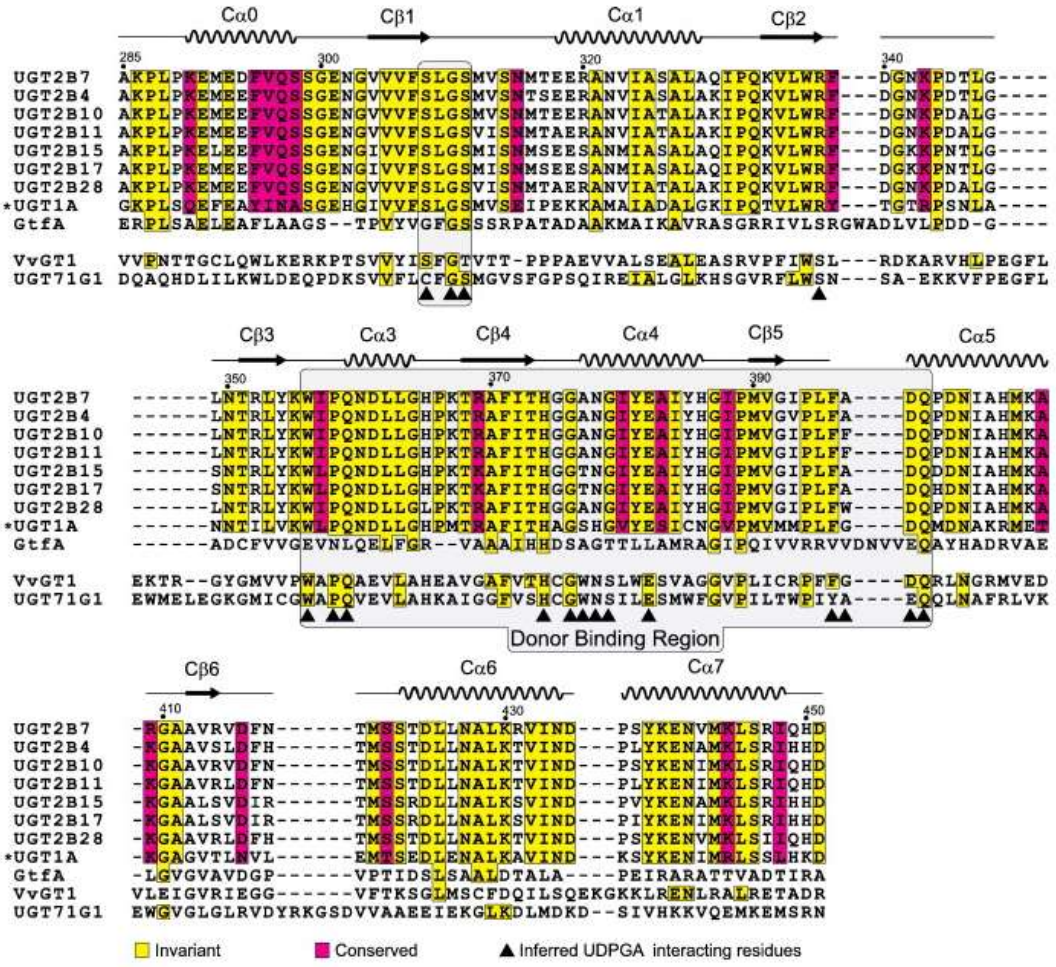


Figure 33.7: Structure-based sequence alignment of GT1 family enzymes. All unique C-terminal domain sequences from human UGTs were aligned to the crystallized 2B7CT sequence: six 2B subfamily UGTs and all 1A subfamily UGTs, which share an identical C-terminal domain. In the alignment are also comprised a representative GT from bacteria, GtfA, and two plant flavonoid GTs, VvGT1 and UGT71G1. The secondary structure of 2B7CT is depicted above its sequence. The conservation mapping is highlighted, and the regions important for donor ligand binding are comprised in a shaded box. The inferred important residues for interaction with UDPGA are specifically marked. (Reproduced from Miley *et al.*[54])

In 2011, Lewis *et al.* generated a homology model of the whole human UGT2B7 structure, using the FUGUE and ORCHESTRAR pieces of software[94]. The crystal templates used in the study are the UDPGA binding domain of human UGT2B7 obtained by Miley *et al.* (PDB Id: 2O6L), for the C-terminal domain, and some crystals of grape UDP-glucose flavonoid 3-O glucosyltransferase (PDB Id: 2C1X, 2C1Z, 2C9Z)[52], plus barrel medic UDP-glucose flavonoid glucosyltransferase UGT71G1 (PDB Id: 2ACV and 2ACW)[92], for the N-terminal domain.

The protein consists in eleven α-helices, designated as A-K, and nine β-sheets, numerated from 1 to 9, with four additional helices which called A', B', C' and F'. Two main domains can be recognized: the N-terminal and the C-terminal domains, responsible for the catalytic activity, which are hinged together by I, J and K helices, and are joined by the A-A' loop. A third, smaller domain, separated to the catalytic domains, encompasses the hydrophobic B'-C loop region (Figure 3.8)[94].

This homology model represented the starting point of our computational study concerning the UGT2B7 three-dimensional structure, which was undertaken with a view to optimizing the catalytic binding site, including the cofactor, as well as to investigating the correct position of the substrate, and its path to enter into an exit from the binding pocket.

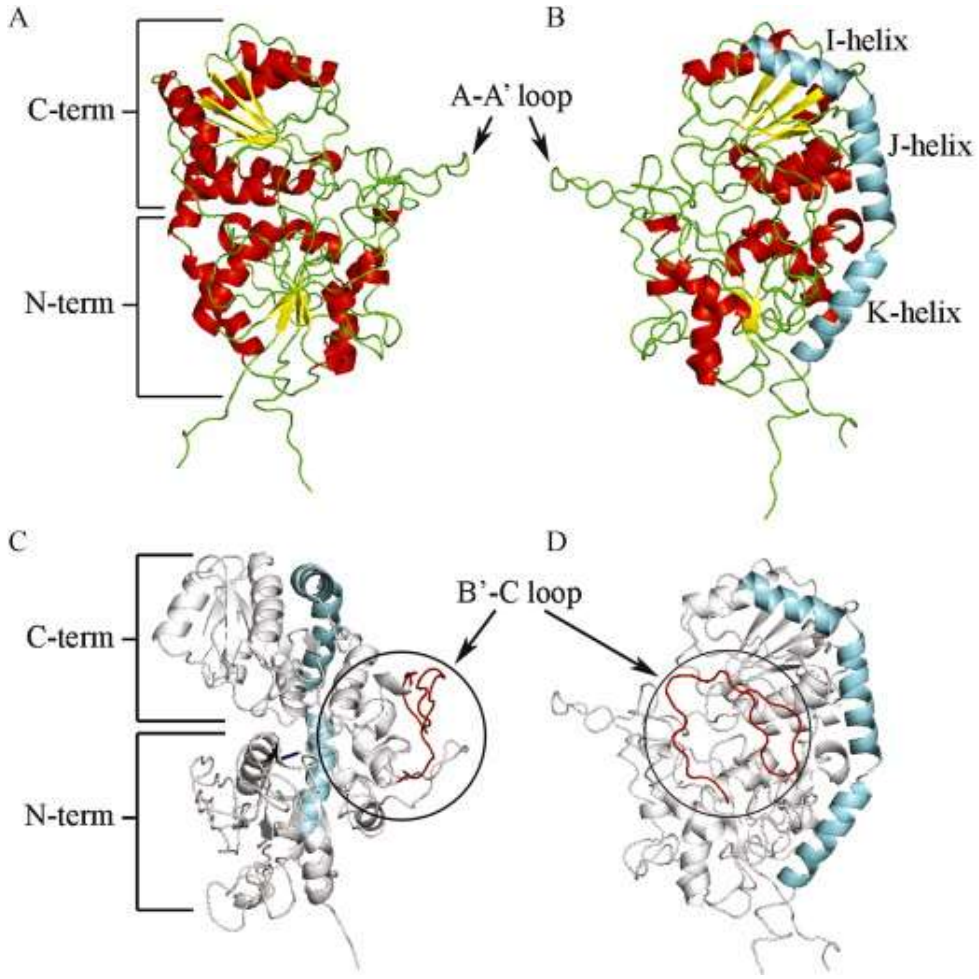


Figure 3.8: The tertiary organization of the UGT2B7 homology model. (Panel A and B) The catalytic N-terminal and C-terminal domains are coupled via helices I, J and K, and the A-A' loop. (Panel C and D) A third smaller domain, in the circle, separated to the catalytic domains, encompasses hydrophobic residues of the B'-C loop. (Reproduced from Lewis *et al.*[94])

3.3.2 Computational detail

Our computational studies involved the UGT2B7 homology model, which was kindly provided to us by Lewis. The obtained model was lacking in bound ligands (both cofactor and substrate), therefore the first step of our study was aimed at optimizing the catalytic site by docking UDPGA and Naproxen, chosen as exemplificative substrate. The completed structure underwent to SMD simulations to better assess the stability of the computed poses as well as to investigate the different pathways of the substrates.

In detail, the optimization of the catalytic site in the homology model by Lewis *et al.*[94] was based on the resolved crystal structure of the plant flavonoid glucosyltransferases VvGT1 (PDB Id: 2C1Z)[52]. This crystal contains the cofactor uridine-5'-diphosphate-2-deoxy-2-fluoro-alpha-D-glucose, the conformation of which was considered as a good template for the pose of the UDPGA cofactor within the UGT2B7 binding pocket. The VvGT1 binding site also includes the substrate 3,5,7-trihydroxy-2-(4-hydroxyphenyl)-4H-chromen-a-one (Kaempherol), which was similarly used as the template to accommodate the UGT substrate Naproxen. (Figure 3.9-A)

The overall studies consisted in three main steps: the modelling of the UGT2B7-UDPGA complex, the docking of the substrate, and, finally, the analysis of the ligand's pathways, by Steered Molecular Dynamics simulations (SMD).

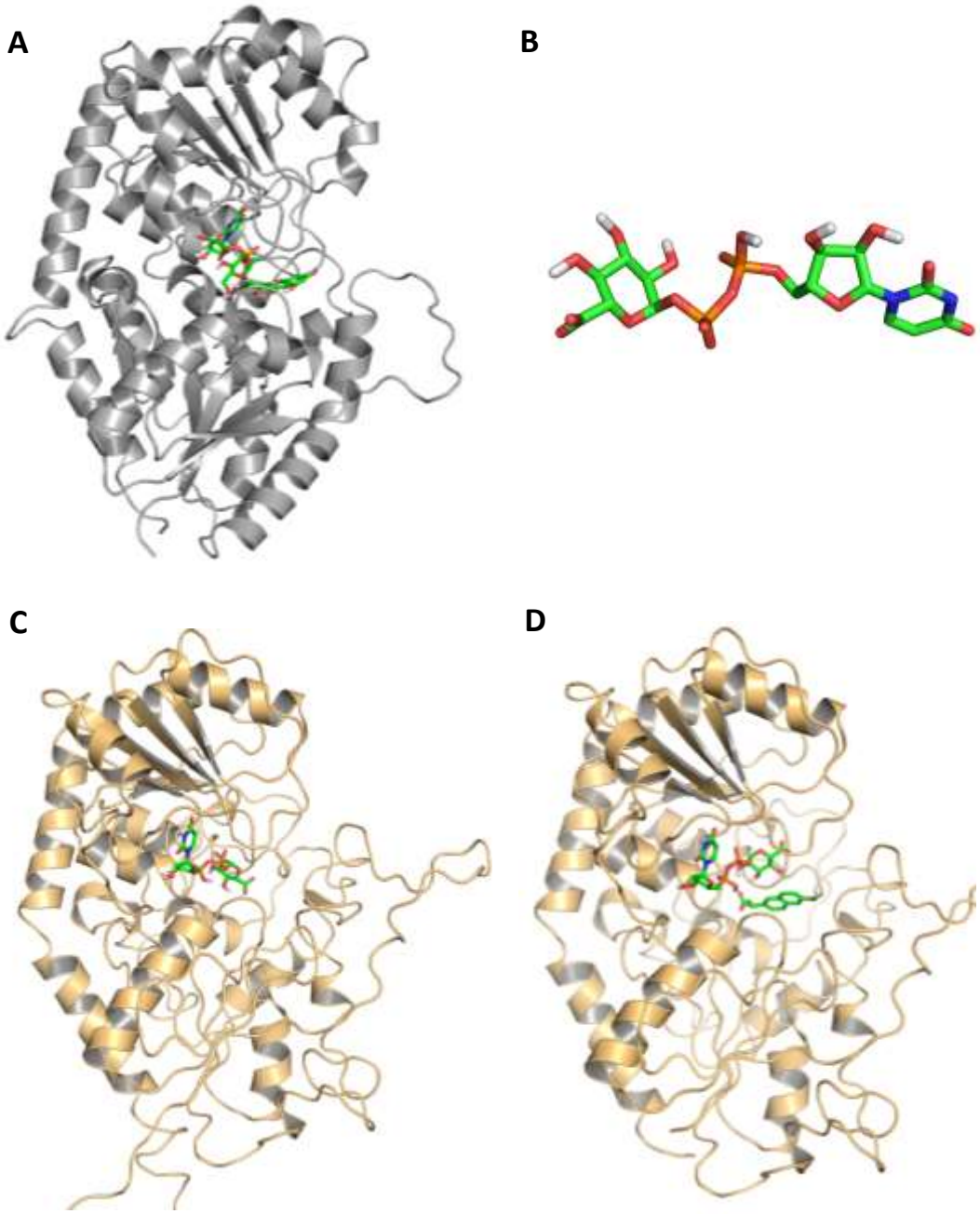


Figure 3.9: Modelling studies of the UGT2B7. (A) Crystal structure of VvGT1 with its cofactor and Kaempherol in the binding pocket. **(B)** UDPGA structure extracted from crystal 2Y0C and optimized. **(C)** UGT2B7 model structure with UDPGA. **(D)** UGT2B7 model structure with UDPGA and Naproxen.

3.3.2.1 Modelling of UGT2B7-UDPGA complex

The molecule of UDPGA was extracted from the crystal of Burkholderia Cepacia Udp-Glucose Dehydrogenase (2Y0C)[95], standardized according to physiological pH to a double negatively charged ligand, and optimized by the semi-empirical quantum chemistry program MOPAC[96]. (Figure 3.9-B) The obtained structure was docked in the catalytic pocket with the docking software PLANTS[97] and the best conformation out the 100 generated was selected as the best superimposing to the cofactor of VvGT1 enzyme, in the crystal 2C1Z[52].

The complex was optimized by a first energetic minimization, in order to find a lower intermolecular energy, coupled to a Molecular Dynamic simulation. In detail, the minimization was performed with NAMD software[98], setting the protein in a water cluster, considering free of moving all the residues included in a 12 Å radius sphere centred on the cofactor, fixing with constraints the backbone of the others, and running 50 000 steps. The Dynamic simulation was performed with NAMD, for 1 nanosecond and at a temperature equal to 300 K. The frame with the UDPGA conformation best overlapping the shape of the cofactor in 2C1Z crystal was selected and minimized (Figure 3.9-C).

3.3.2.2 Modelling of UGT2B7-UDPGA-Naproxen complexes

Two complexes were prepared starting from the UGT2B7-UDPGA complex. The complex with the Naproxene substrate was prepared superimposing the UGT2B7-UDPGA minimized complex to 2C1Z crystal and adding a molecule of Naproxen by aligning it to Kaempherol molecule present in 2C1Z binding site. The obtained complex was optimized with a minimization performed by NAMD with a sphere of free residues measuring 12 Å radius centred on the cofactor. The

substrate was then extracted and re-docked with Plants software to verify the reliability of the complex (Figure 3.9-D).

The UGT2B7-UDP-GANaproxen, namely the ternary complex involving the enzymatic product, was prepared by manually transforming the optimized UGT2B7-UDPGA-Naproxen and was minimized by keeping fixed all atoms outside a 12 Å radius sphere around the bound ligands.

3.3.2.3 Steered Molecular Dynamics

Due to their net negative charge equal to -5, the two optimized complexes were neutralized by adding 5 sodium ions using the SODIUM tool[99]. The neutralized complexes were then inserted into a 50 Å radius spherical box of water molecules so as to generate hydrated complexes containing about 10000 solvent molecules. The so obtained systems were finally minimized to optimize the relative position of solvents and ions, and underwent the following SMD simulations.

The two prepared complexes underwent 1.5 ns all-atoms SMD simulations with the following characteristics: (a) 60 Å radius spherical boundary conditions were applied to stabilize the simulation space; (b) Newton's equation was integrated using the r-RESPA method (every 4 fs for long-range electrostatic forces, 2 fs for short-range non bonded forces, and 1 fs for bonded forces); (c) the temperature was maintained at 300 ± 10 K by the Langevin's algorithm; (d) to the selected atoms the spring constant equal to 5 kcal/mol/$Å^2$ was applied with a pulling velocity of 0.003 nm/ps (e) Lennard-Jones (L-J) interactions were calculated with a cut-off of 10 Å and the pair list was updated every 20 iterations; (e) a frame was memorized every 10 ps, thus generating 150 frames; and (f) no constraints were imposed to the systems.

The simulations were carried out in two phases: an initial period of heating from 0 K to 300 K over 10000 iterations and the monitored phase of 1.5 ns.

3.3.3 Computational results

3.3.3.1 Analysis of the UGT2B7-UDPGA-Naproxen ternary complex

Figure 3.10-A shows the key interactions stabilizing the computed pose for the UDPGA cofactor, which can be summarized, as follows. The phosphate groups are involved in ionic interactions with Arg338 reinforced by H-bonds with His374, while the hydroxyl groups of the ribose ring are engaged in clear H-bonds with Gln359 and Glu382. The uridyl moiety is tightly inserted between the side-chains of Gln359 and Arg338 with which it can stabilize π–π stacking reinforced by charge transfer interactions. Finally the glucuronic acid substructure approaches the substrates and stabilizes H-bonds with Gln399. Interestingly, its carboxyl group is not involved in significant contacts apart from some weak H-bonds with the backbone atoms of Met312 and Val313 and, more generally, the glucuronic acid appears to be not engaged in strong interactions. This finding can be explained considering that this substructure has to maintain a certain degree of flexibility to better approach the substrate, thus promoting the enzymatic reaction.

Similarly, Figure 3.10-B shows the key interactions stabilizing the putative pose of the Naproxen substrate, which appears to be engaged in a rich network of p-p stacking and hydrophobic interactions with the surrounding residues such as Tyr33, Phe105, Trp106 and Phe174. The carboxyl group approaches the catalytic residue His35 and stabilizes H-bonds with Ser34 and Asn374.

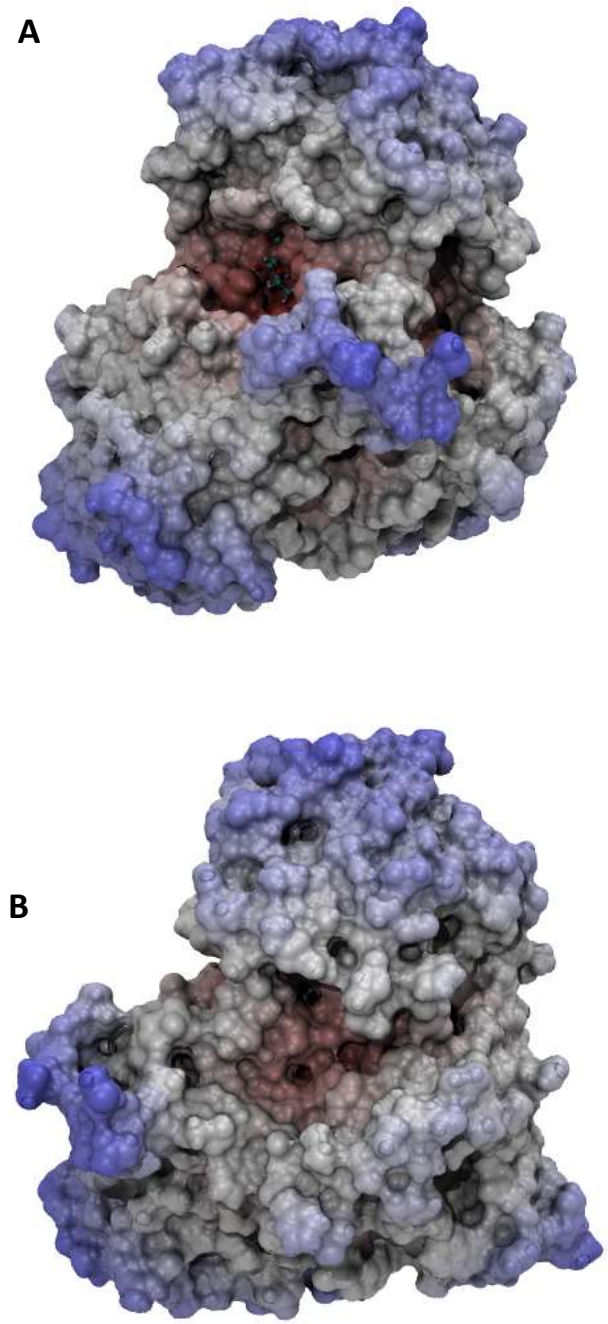


Figure 3.10: Modelling studies of the UGT2B7. (A) Principal and lateral door in UGT2B7 model. **(B)** Rotation of the complex of 90° to better show the lateral door.

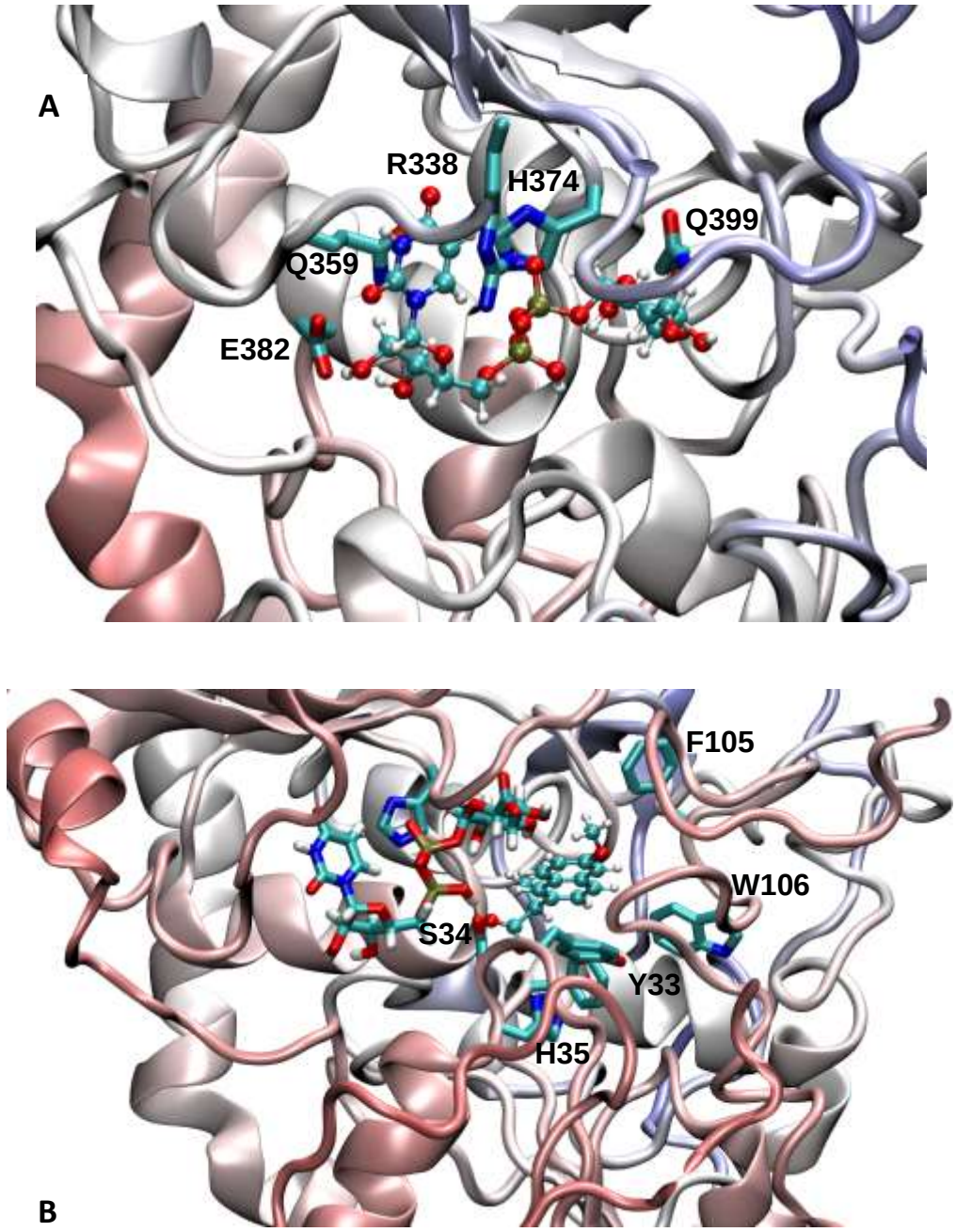


Figure 3.11: Modelling studies of the UGT2B7. (A) Main residues stabilizing the UDPGA cofactor in the complex. **(B)** Main residues stabilizing Naproxen in the complex.

3.3.3.1 Analysis of the substrate's pathways by SMD

As a preamble, Figure 3.10 shows the overall structure of the above described ternary complex and highlights the two main doors through which substrate and cofactor can reach the catalytic cavity. As outlined in Figure 3.10, these two entrances can be defined as principal and lateral doors. One may reasonably suppose that the cofactor can enter through the former, also considering its final pose within the catalytic pocket as shown by Figure 3.10-C, while the exact role of the lateral door is still unclear. When considering that the substrate reaches the catalytic site after the cofactor is already bound to it (see below), the first objective of the performed SMD runs was to assess whether the substrate can however enter through the principal door, even though its pathway would be obstructed by the bound cofactor, or the lateral door is actually the entrance by which substrates can reach the catalytic pocket.
Hence, the SMD simulations compared the energy profiles as computed when moving the substrate through the two possible doors. For completeness also two intermediate pathways were simulated, thus exploring all possible channels characterizing the modelled UGT2B7 structure.

Figure 3.12 compares the pull force profiles for the substrate undocking as computed for the principal and lateral doors, and shows that the latter is clearly favored. This is also confirmed by Table 3.1, which reports the pull force maximum and average values for all simulated pathways, and reveals that the pathway through the lateral door shows a pull force average which is nearly half that through the principal door. The intermediate pathways reveal even worse pull force profiles, especially to be concern the average values, thus indicating that they are in fact unrealistic solutions.

Notably, when focusing the attention on the last part of the SMD runs (namely the last 0.5 ns) a contrasting trend is observed in the average values, since the principal door shows an average value which is largely lower than that of the lateral door (52.48 pN vs 130.48 pN). This finding can have two different explanations.

The first consideration is that the key difference between the two compared pathways is represented by the presence of the cofactor, which impedes the substrate ingress along the pathway of the principal door. Without this, the channel corresponding to the principal door is clearly wider and thus can be easily passed through as emphasized by the last part of the MD runs. This means that, without the cofactor, the substrate ingress through the principal door would be surely favored, but this case is unrealistic. Indeed, several recent studies confirmed that the enzyme involves the formation of a compulsory order ternary-complex in which the cofactor binds first even in presence of factors (such as BSA), which can impair (or modulate) the catalytic efficiency of the UGT enzymes. As a matter of facts, the planned SMD runs involved neither the substrate movements without cofactor, as they are unrealistic, nor the analysis of the cofactor pathway, since it should reasonably pass through the principal door, which represents the favored path for the free enzyme.

The second explanation for the observed discrepancies in the last part of the SMD simulations can involve the geometrical arrangement of the rim of the lateral door, which is not finely optimized as that of the principal door. In this case, the performed simulations had also the indirect effect of optimizing the folding of this lateral channel allowing a proper ingress for every substrate.

Pathway	average	maximum
Principal door	195.63	1042.61
Intermediate2	497.72	1063.28
Intermediate3	290.40	983.76
Lateral door	116.21	687.47

Table 3.1: Pull force for the substrate undocking. The table reports the pull force maximum and average values for the substrate undocking as computed for the principal, lateral and two intermediate doors (all values are expressed in pN).

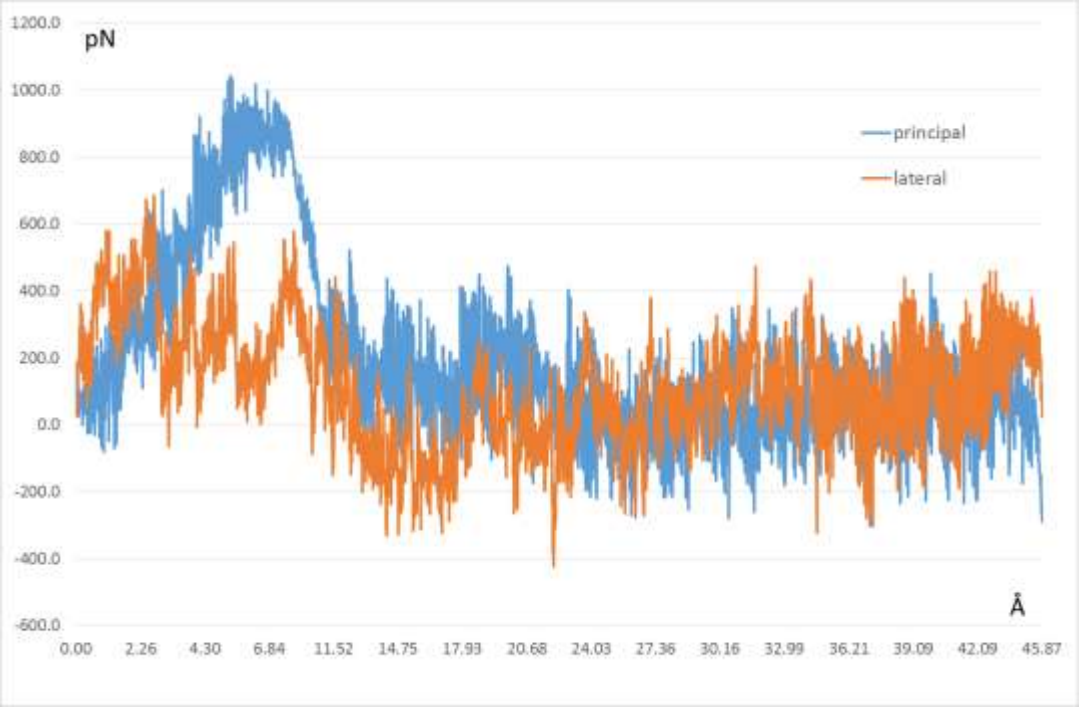


Figure 3.12: Pull force profiles for the substrate undocking. The principal and the lateral doors are colored in blue and orange respectively.

3.3.3.1 Analysis of the product's pathways by SMD

When considering the results obtained for the substrate's pathways, SMD simulations on the ternary complex containing the enzymatic product and the aglyconic UDP cofactor were similarly performed, with a view to revealing the product's pathways.

As depicted in Figure 3.13, three SMD simulations were performed by considering the pathway through the lateral door as well as that through the principal door as simulated with and without the bound UDP cofactor. Indeed and differently from what was previously discussed for the substrate pathways, here the sequential order with which UPD and substrate leave the catalytic pocket is unknown. Nonetheless, the computed pull force profiles allow for some relevant considerations.

First, the product egress through the lateral door shows the worst profile with high force values during all the simulation (average = 528.74 pN, maximum = 1175.41 pN). These results may indicate that this pathway is reasonably unrealistic and the channel is completely unsuitable for product undocking. As already seen but here markedly more pronounced, the last part of the SMD run shows a remarkable increase of the pull force which suggests that the product is substantially unable to pass through the cavity rim.

Second, the pathway through the principal door without UDP (average = 217.51 pN, maximum = 821.77 pN) shows a notably more favored pull force profile compared to the same path with UDP (average = 337.33 pN, maximum = 1202.94 pN). Even though, this difference cannot offer a clear indication concerning the order with which UDP and product leave the enzyme, it leads to the hypothesis that (as already seen for the binding), the UDP leaves first, followed by the product both of them passing through the principal door.

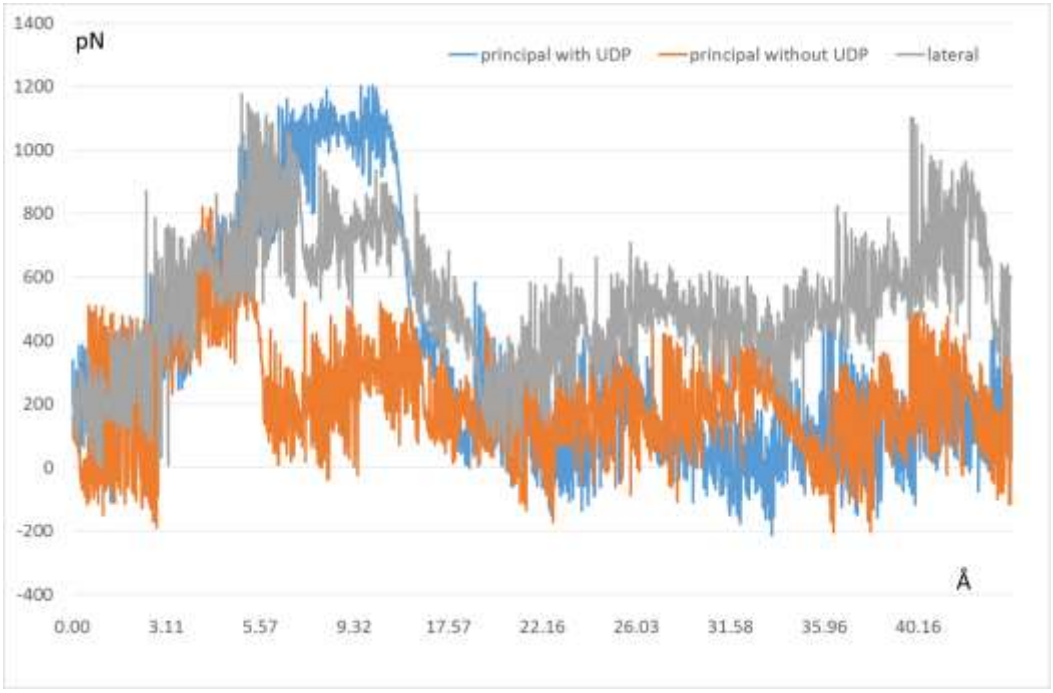


Figure 3.13: Pull force profiles for the product undocking. The principal and the lateral doors are colored in blue and orange respectively.

In summary, the performed SMD simulations allow a complete binding/unbinding mechanism for the UGT reaction to be reasonably supposed. In detail such a mechanism can be schematized as follows:

- first, the UDPGA cofactor binds entering through the principal door;
- second, the substrate binds with the bound cofactor entering through the lateral door;
- third, the transformed UDP cofactor leaves reasonably first egressing through the principal door (even though as detailed above the cofactor movements were not investigated), thus the UDP egressing through the lateral door cannot be excluded a priori;
- fourth, the enzymatic product leaves for last egressing through the principal door.

3.4 PCM

In this study, a Proteochemometric technique was applied to MetaSar substrates and UGTs enzymes resulting in the first PCM model concerning the regioselectivity of glucuronidation reactions. A simple classification algorithm was built to predict if a given molecule can be a UGTs substrate or not.

The input dataset was assembled starting from three sources: the structures of compounds collected within MetaSar database, including substrates as well as "non-substrates" of UGTs enzymes, the annotation of the reactions, namely the occurrence or not occurrence of glucuronidation reaction for a given molecule, and the sequences of the human UGT enzymes belonging to families A and B. The first two sources were taken from MetaSar. In details, we referred to the version of the database updated at January 2015. With regard to the targets, 18 UGTs sequences were taken in consideration and retrieved from the website Uniprot[100].

Many different models were tested, changing the Machine Learning technique, its parameters, the combination of descriptors, the number of compounds and targets.

For simplicity's sake, the attention will be focused only on the best so obtained results.

3.4.1 Chemical descriptors

The study involved all the 1730 molecules included in MetaSar. The cleaning phase of data was rather simple, because the original database is manually compiled, and thus already well curated. First, the molecules were checked for duplicates and cases of homonymy. Then, some substrates were discarded because they include counter ions and others because they were identified as outliers in terms of molecular size. In particular, molecules having less than 8 atoms were

considered too small and compounds with molecular mass higher than 1000 u.m.a. were considered too big. Finally, the remaining 1700 molecules were standardized by considering their most probable ionization state at physiological pH using a specific script implemented in the VEGA ZZ[101] software.

The compound's three-dimensional structure was optimized by a specific feature of the Molecular Operating Environment software (MOE) version 2012.10[17], which builds reasonable and minimized conformations starting from a database of 3D-structures. During this procedure, the stereochemical configuration of the molecules was preserved. Indeed, more than one stereoisomer for a single compound is often reported in the database and glucuronidation reactions seem to be influenced by configuration.

Then, some descriptors were calculated directly by MOE, while other descriptors required different pieces of software. For this second case, the 3D-structures of the compounds were exported from MOE as smiles containing the stereochemical configuration, and then converted into an "sdf" file containing atomic coordinates and connectivity by using MarvinViewer version 15.2.9.0[102], a software supplied by ChemAxon Software Company[103].

In detail, the ESshape3D descriptors were calculated by MOE[17], obtaining a matrix with fingerprints measuring 122 integer values. Together with this kind of shape fingerprints, all the other fingerprint descriptors implemented in MOE were also calculated. The presence of duplicates was checked with a code written in R version 3.1.3 specifically for this aim, running it by using RStudio Version 0.98.1103[104]. Actually, ESshape3D were found to be the only MOE fingerprints able to distinguish between all the stereoisomers included in our dataset.

Some physicochemical descriptors were calculated by MOE[17] and a complete list of those included in the models is reported in Appendix 1, with a brief explanation of their characteristics.

The ECFP were calculated from the "sdf" file by using the software Molecular Descriptor Generator (GenerateMD) version 15.3.2.0 supplied by ChemAxon[103]. Specifically, the folded form of these fingerprints was calculated, so generating a matrix including the ECFP descriptors for all compounds and measuring 1024 binary values. As expected and accordingly to the well-known ECFP incapability to handle stereoisomers, the checking for duplicates revealed the presence of 299 molecules with the same string of binary values.

The 3DAPfp and 3DXfp were calculated exploiting the "sdf" file and the Java commands supplied by Awale *et al.*[20] by using Java Runtime Environment (JRE) version 1.8.0_51[105] and dependencies from the ChemAxon package. The output files were matrices of fingerprints for all the compounds measuring respectively 16 and 80 integer values. The use of these fingerprints was especially oriented to obtain descriptors able to distinguish stereoisomers. Disappointingly, the checking for duplicates revealed the presence of an extremely high number of molecules having the same string of fingerprints. For this reason, these descriptors were not included in the models.

3.4.2 Protein descriptors

The protein descriptors chosen for this study belong to the class of alignment dependent sequence descriptors. The sequences were downloaded from UniProt[100] and the total number of the considered enzymes was reduced from 19 to 18 because the UGT2A2 protein does not have a distinct sequence but is considered as a

"synonyms" of UGT1A1. According to the information reported in the PubMed-UniGene[106], UGT2A2 seems to be an isoform of UGT1A1, whose sequence is still unknown.

The list of the considered proteins with their UniProt ID is summarized in the following Table 3.2.

Gene names	Uniprot id	N. amino acids	Identity %
UGT2B7	P16662	529	100
UGT2B10	P36537	528	88
UGT2B4	P06133	528	86
UGT2B11	O75310	529	86
UGT2B28	Q9BY64	529	85
UGT2B15	P54855	530	78
UGT2B17	O75795	530	77
UGT2A1	Q9Y4X1	527	60
UGT2A3	Q6UWM9	527	59
UGT1A1	P22309	533	42
UGT1A7	Q9HAW7	530	42
UGT1A8	Q9HAW9	530	42
UGT1A9	O60656	530	42
UGT1A10	Q9HAW8	530	42
UGT1A3	P35503	534	41
UGT1A5	P35504	534	41
UGT1A4	P22310	534	40

Table 3.2: UGT proteins analysed in the study. For each enzyme name, the Uniprot id is reported, together with the number of residues in the whole sequence and the percentage of identity referring to UGT2B7. The colour ramp encodes for the degree of similarity.

The sequence alignments were performed by using the webserver Clustal Omega[107]. As evidenced by Table 2, the identity percentages of UGTs to UGB2B7 protein, which is the core of our study, can be subdivided into two classes of importance. The proteins belonging to the subfamily 2B exhibit the highest similarity, and the key models were built on them. The protein belonging to the subfamilies 2A and 1A show a lower similarity, and were introduced to the model afterwards to verify if they can improve the reliability of the predictions. For this reason, two alignments were performed, one with only the 7 proteins with higher identity percentage, and another with the all the 18 sequences.

To improve the resolution of the model, the sequences corresponding to domains not involved in the ligand binding can be discrded[5]. The selection of the residues belonging to the binding site was based on the previously obtained UGT2B7 homology model. The corresponding residues on other UGT sequences were then derived from the alignment. The selected residues were those comprised in the following list of sub-sequences, referred to the UGT2B7 sequence having UniProt ID P16662:

24-41 97-127 147-159 271-289 304-340 352-420

All the alignment-dependent physicochemical descriptors were calculated by using a code written in R version 3.1.3 specifically for that aim[104].

3.4.3 Dataset building

The files of the input dataset were suitably assembled by using a code written in Python[108] and ran by PyCharm Community Edition version 4.0.4[109] starting from three sources: two files containing the descriptors for the compounds and the proteins and a third file containing compound names annotated with labels indicating whether they are UGT substrates or not. The total number of the simulated compounds was equal to 1700, among which 338 were UGT substrates and 1362 were non-substrates. In contrast, the number of proteins in the different models ranged from 7, to 18, to 2. All the possible "chemical-protein" combinations were included and, for each combination, the data in the input matrix included its name, its label and the related descriptors of both chemical and protein.

3.4.4 The code

The code to perform the model was written in Python and ran by PyCharm Community Edition version 4.0.4[109]. The code is organized in functions, one for each phase of the model generation, which can be schematized as follows: (a) the dataset reading, (b) the data pre-processing, (c) the splitting of data, (d) the generation and validation of the model through the NCV method, with the corresponding evaluation of the so obtained results, and (e) the validation of the model through the LOCO method, with the corresponding evaluation of the results. Thanks to this organization, the code can be used to perform different models, just by calling the specific required functions. Each part of the code was written by assembling online available functions imported by two open source BSD-licensed libraries, namely Pandas[110] and Scikit-learn[111]for the Python programming language. The former provides high-performance tools for data manipulation and

analysis. The latter features various classifications, regression and clustering algorithms, including RF.

3.4.5 Dataset pre-processing

The pre-processing of the dataset involved three consequent steps, leading to a gradual decreasing of the features number.

The first step is the removing of "near zero variance" descriptors, and exploits the "*VarianceThreshold*" function from the Scikit-learn sub library called "*feature_selection*". To better define the threshold of the minimum variance value required for keeping a given feature, the descriptors were subdivided according to the type of values and three different sections were written to perform this pre-processing step. In detail, the ECFP descriptors, which involve Boolean variables, were treated as a Bernoulli distribution, the ESshape3d fingerprints, which include discrete variables, were treated as a multinomial distribution, and physicochemical properties as well as protein descriptors, which correspond to continuous variables, were treated as a normal distribution. The variance threshold was set as a customizable parameter, and different values were tested to find the best performing one.

For the ECFP descriptors, the best results were obtained using a threshold of minimum variance equal to 0.0475, which corresponds to a binary feature that is composed for 95% by the same value. With this threshold, the ECFP columns were reduced from 1024 to just 152 features.

For all the other features, a threshold of minimum variance equal to 0.8 was found to be the most efficient one. Looking at the distribution of variances reported in Table 3.3, this threshold permitted about 70% of the non-binary features to be deleted, while maintaining about 30% of them.

Count	1.24E+03
Min	0.00E+00
Max	1.40E+09
Mean	1.25E+07
25%	4.53E-13
50%	6.17E-12
75%	9.24E-01

Table 3.3: Distribution of variances in all the non-binary features. The analysis involved the ESshape3D fingerprints, the pyhsicochemical properties and the protein Zscales3 descriptor

The second step is the normalization of the data, which was applied to all the "not-binary" features exploiting the "*preprocessing.scale*" function of Scikit-learn. Each feature value was subjected to two mathematical operations: the subtraction of the mean of that feature, obtaining the so called "*centring*", and the division by the standard deviation of that feature, called "*scaling*". Finally, the remaining features were checked for the presence of correlated features.

Finally, in the third step, the remaining features were checked for the presence of correlated features.

3.4.6 Model building

Random Forest was performed by using the specific function "*RandomForestClassifier*" imported from the Scikit-learn sublibrary called "*ensemble*". The customized parameters were:

- "*max_features*", which refers to the maximum number of features that are taken in consideration to build the model. Depending on the input dataset, the internal validation set this parameter equal to 80, 100 or 200.
- "*n_estimators*", which is the number of decision trees built by the model. Depending on the input dataset, the internal validation set this parameter equal to 100 or 200.
- "*class_weight*", which assigns to the two classes in the dataset a different weight value to solve issues of unbalanced results, as those deriving from unbalanced datasets. The balanced weight corresponds to a value of 0.5 for both classes. All the possible combinations of different weights were tested in an iterative way, but the model predictive power did not increase significantly with any of them.

Two methods for model validation were tested with Random Forest algorithms, namely the Nested Cross Validation (NCV) and the Leave one compound out (LOCO).

The NCV method consisted in an external validation (outer loop), in which the dataset is subdivided into training and test sets, which is nested with an internal validation (inner loop), in which the customizable parameters are tested within the training set. For what concern the external validation, the splitting of the dataset corresponded to the 70% of the database for the training set and the 30% for the test set. Since all UGT enzymes are considered equally able to metabolize their substrates in the model, the data splitting was preceded by a preliminary data clustering according to the compound name, so to avoid biasing conditions

deriving by the presence of the same compound with the same label both in training and test sets. For the internal validation, the "*k*-fold cross-validation" method was chosen, by setting *k* equal to 5 and performing the selections of parameters based on the F_1 score. The so selected best parameters were then employed to predict the labels of the test set in the external validation.
The evaluation of the NCV analyses was performed on the basis of Precision, Recall, F1 score and Matthews Correlation Coefficient.

The LOCO method consisted in the training of the model on the whole dataset except for one compound, with all its combinations with the enzymes properties. The model was trained using the same best parameters selected by the previously performed NCV on the same dataset. The evaluation of the model was done on the basis of the Precision, Recall and F_1 score.

3.4.7 NCV results

The Random Forest classifier was tested on many different input datasets, changing the descriptors for the compounds and the proteins, the number of the targets, and the model parameters, in order to find the best classification algorithm. The following sections illustrate the composition of the input dataset, the best values for the customizable parameters, as selected by the "k-fold cross-validation", and the resulting performances obtained by the external validation. These results are reported in terms of "precision", "recall" and "F_1 score", with distinct values for the class "0", corresponding to "non-substrates", and the class "1", corresponding to "substrates", and an average value for the two classes together. The Matthews Correlation Coefficient (MCC) is also reported, as an overall measure of the predictive power of the model.

3.4.7.1 Selection of the right combination of compound features

The first objective involved the definition of the best combination of descriptors to include into the input dataset.

A set of models was developed maintaining fixed all the conditions with the only exception of the type of the compounds' descriptors.

Model n. 1 - Descriptors: ECFP, ESshape3D, Zscales_3.

Composition of dataset	
Compounds	1700
Targets	7
Training set	8330
Test set	3570
Datapoints	11900
Features before processing	2137
Features after removing near zero variance	385
Features after removing correlated descriptors	269

Parameters	
Max features	100
Estimators	200
Class weight	0.5 / 0.5

Results			
	Precision	Recall	F1 Score
0	0.86	0.97	0.91
1	0.74	0.38	0.50
Average	0.83	0.85	0.82
MCC	0.46		

Model n.2 - Descriptors: ECFP, ESshape3D, Physicochemical, Zscales_3.

Composition of dataset	
Compounds	1700
Targets	7
Training set	8330
Test set	3570
Datapoints	11900
Features before processing	2267
Features after removing near zero variance	472
Features after removing correlated descriptors	317

Parameters	
Max features	100
Estimators	200
Class weight	0.5 / 0.5

Results			
	Precision	Recall	F1 Score
0	0.86	0.96	0.91
1	0.72	0.39	0.51
Average	0.83	0.84	0.82
MCC	0.45		

Figure 3.14: Tables and results of the model n. 1 and 2.

The reported results for the first two models set out what can be considered the best predictions that can be obtained when training the model on the whole dataset. The performances of all following models were, thus, compared with these first results. Both models were built using 7 targets (i.e., the 2B subfamily of the UGT enzymes), and the only difference between them consists in the introduction of the compounds' physicochemical properties in the second model.

As can be seen in the Figure 3.14, the resulting performances are broadly similar, underling that in this case the physicochemical properties are not crucial for the prediction. The most striking result to emerge from the data is that the predictive performances of both models are outstanding for the "non-substrates" class, in which about 97% of compounds are correctly predicted, but not satisfactory enough for the "substrates" class, successfully predicted only in the 39% of cases. Clearly, the attention is here focused on the recall results rather than on the precision ones, since the former is effectively a measure of the correctly predicted compounds among the entire set of compounds belonging to a certain class. Even though the average recall value is satisfying, the unbalanced dataset is the main weakness of the model, and the overcoming of this drawback was the primary objective of the following studies.

Model n.3 - Descriptors: ECFP, Zscales_3.

Composition of dataset	
Compounds	1700
Targets	7
Training set	8330
Test set	3570
Datapoints	11900
Features before processing	2015
Features after removing near zero variance	284
Features after removing correlated descriptors	203

Parameters	
Max features	80
Estimators	200
Class weight	0.5 / 0.5

Results			
	Precision	Recall	F1 Score
0	0.85	0.94	0.90
1	0.63	0.36	0.46
Average	0.80	0.82	0.81
MCC	0.38		

Model n.4 - Descriptors: ESshape3D, Zscales_3.

Composition of dataset	
Compounds	1700
Targets	7
Training set	8330
Test set	3570
Datapoints	11900
Features before processing	1112
Features after removing near zero variance	373
Features after removing correlated descriptors	132

Parameters	
Max features	100
Estimators	200
Class weight	0.5 / 0.5

Results			
	Precision	Recall	F1 Score
0	0.80	0.87	0.83
1	0.26	0.18	0.21
Average	0.69	0.73	0.71
MCC	0.06		

Figure 3.15: Tables and results of the model n. 3 and 4.

After verifying the marginal importance of the physicochemical properties, two models (n. 3 and n. 4) were developed to establish the relative relevance of the computed fingerprints (Figure 3.15).

In the model number 3, only the ECFP fingerprints were used. The performance of the model is very near to that of the model n. 2, thus confirming the considerable ability of these descriptors in catching the molecule structure. However, we can't exclude that these correct predictions are, in part, the result of a bias introduced in the model by the use of compounds descriptors unable to distinguish among all the molecules in the dataset. Indeed, considering the high number of duplicates reported in the similarity matrix build with Tanimoto distance based on ECFP, the probability of having the same compound-target combination both in the training and in the test set is high.

In the model n. 4, only the ESshape3D fingerprints were used. On the opposite, the results showed an important decrease of accuracy, in particular in predicting the substrate class. This trend is reflected also in the MCC value. This means that ESshape3D, even though crucial to achieve a complete distinction among all compounds, prove to be insufficient to provide a complete molecular description, and, as a consequence, a correct prediction. This weakness is even more evident for the unbalanced class, already affected by worse performances due to its insufficient presence in the training set.

3.4.7.2 The issue of unbalanced data

A dataset is called unbalanced if it contains many more samples within one class than within the other one, or within the remaining classes[112]. The issue of unbalanced data intrinsically characterizes MetaSar, since the number of substrates for a specific enzyme family is clearly a minority compared to the total number of compounds in the database. Therefore, the capacity to suitably address this problem is of particular importance, especially considering that MetaSar can be a fertile source of data for other PCM studies on other enzyme classes.
As already described, the number of substrates in our dataset is one fifth of the number of non-substrates, namely 338 against 1362. Two techniques were then performed in order to solve the problem of the unbalanced dataset.

The first attempted strategy exploited the "*class weight*" parameter of the function "*RandomForestClassifier*". This parameter allows assigning a different weight to the two classes in the model, and by default is set to 0.5 for both classes, so as to assign the same weight to them. By assigning a heavier weight to the less represented class in the training set, it should be possible to compensate the unbalanced data.

Model n.5 - Descriptors: ECFP, ESshape3D, Physicochemical, Zscales_3.

Composition of dataset	
Compounds	1700
Targets	7
Training set	8330
Test set	3570
Datapoints	11900
Features before processing	2267
Features after removing near zero variance	472
Features after removing correlated descriptors	317

Parameters	
Max features	100
Estimators	100
Class weight	0.2 / 0.8

Results			
	Precision	Recall	F1 Score
0	0.85	0.96	0.90
1	0.70	0.37	0.48
Average	0.82	0.84	0.82
MCC	0.43		

Figure 3.16: Tables and results of the model n. 5.

According to the instructions[113], in the model n. 5, different couples of relative weights were associated to the classes of compounds[113]. The reported results in Figure 3.16 showed that, among the possibilities, the "k-fold cross-validation" selected the combination of the class weights equal to 0.2 / 0.8, which corresponds to a weight of 0.2 for "non-substrates" and 0.8 for substrates, as expected. Nevertheless, the performances of the model were substantially superimposable with those of the model n. 2, taken as the benchmark for our comparisons, and, in particular, the recall for the substrate's class did not increase.

Model n.6 - Descriptors: ECFP, ESshape3D, Physicochemical, Zscales_3.

Composition of dataset	
Compounds	1700
Targets	7
Training set	8330
Test set	3570
Datapoints	11900
Features before processing	2267
Features after removing near zero variance	472
Features after removing correlated descriptors	317

Parameters	
Max features	80
Estimators	200
Class weight	0.8 / 0.2

Results			
	Precision	Recall	F1 Score
0	0.85	0.94	0.89
1	0.60	0.37	0.46
Average	0.80	0.82	0.80
MCC	0.37		

Figure 3.17: Tables and results of the model n. 6.

In the model n. 6, the class weights were set as 0.8 for non-substrates and 0.2 for substrates, which correspond to the opposite of the right correction to solve the problem of unbalanced data. The aim of this model was to confirm that the "class weight" strategy is ineffective for our purpose As seen in Figure 3.17, the results were not significantly worse than those of the model n. 2, confirming the hypothesis.

In conclusion, we can emphasize the complete inefficiency of this approach for our intent. This result is in line with other on-line reported examples of inadequate returns of this method[114].

Model n.7 - Descriptors: ECFP, ESshape3D, Physicochemical, Zscales_3.

Composition of dataset	
Compounds	676
Targets	7
Training set	3312
Test set	1420
Datapoints	4732
Features before processing	2136
Features after removing near zero variance	360
Features after removing correlated descriptors	233

Parameters	
Max features	100
Estimators	200
Class weight	0.5 / 0.5

Results			
	Precision	Recall	F1 Score
0	0.86	0.77	0.81
1	0.72	0.82	0.76
Average	0.80	0.79	0.79
MCC	0.58		

Figure 3.18: Tables and results of the model n. 7.

The second strategy attempted in order to overcome the issue of the unbalanced dataset involved the so called "*random under-sampling*" srtategy[112].

According to this procedure and in the dataset used for the model n. 7, an amount of data points concerning the more abundant non-substrate's class was randomly selected and deleted so to render the number of non-substrates equal to that of substrates. The resulting balanced dataset consisted of 338 compounds for both classes. As depicted in Figure 3.18, the performances of the resulting model were strongly influenced by the applied strategy. Indeed, the recall values underline that the substrate's class is reasonably well predicted, even better than the non-substrate's one. This trend is also confirmed by the increasing of the MCC value, even though, as a consequence of the decreasing performances for the non-substrate's class, the averages for the precision and recall values are lower than those of the model n. 2.

Model n.8 - Descriptors: ECFP, ESshape3D, Zscales_3.

Composition of dataset	
Compounds	676
Targets	7
Training set	3312
Test set	1420
Datapoints	4732
Features before processing	2267
Features after removing near zero variance	447
Features after removing correlated descriptors	281

Parameters	
Max features	80
Estimators	20
Class weight	0.5 / 0.5

Results			
	Precision	Recall	F1 Score
0	0.83	0.72	0.77
1	0.66	0.78	0.72
Average	0.76	0.75	0.75
MCC	0.50		

Figure 3.19: Tables and results of the model n. 8.

In the model n. 8, the same under-sampling strategy was applied, with the only difference consisting in the absence of physicochemical properties within the dataset features. This analysis was useful to confirm that even using balanced classes the model did not take significant advantages from including these descriptors. As shown in Figure 3.19, the performances were slightly worse than those of the model n. 7.

The "random under-sampling" reveled to be an efficient method to correctly address unbalanced dataset, when using this kind of learning algorithms. However, the approach is not free of weakness and the main drawback is that it discards potentially useful information contained in the deleted datapoints[115]. Moreover, it basically reduces the size of the dataset, leading, in many cases, to a general decreasing in predictive accuracy.

The "oversampling" strategy, reported in some reviews[116,115], was here not applied because it very often induces disadvantageous overfitting situations, due to the replication of already existing compounds in the minority class.

3.4.7.3 Relevance of the included targets

The first models were performed using 7 targets, namely all the human UGTs belonging to the 2B subfamily. Introducing more targets or reducing their number has two major consequences: the modification of the number of the data points, which influences the shape of the input matrix, and the variation of the overall similarity within the protein space. Two analyses were set up to investigate the relationships between the number of the considered targets and the predictive power of the model.

Model n.9 - Descriptors: ECFP, ESshape3D, Physicochemical, Zscales_3.

Composition of dataset	
Compounds	1700
Targets	18
Training set	21420
Test set	9180
Datapoints	30600
Features before processing	2267
Features after removing near zero variance	784
Features after removing correlated descriptors	571

Parameters	
Max features	100
Estimators	100
Class weight	0.5 / 0.5

Results			
	Precision	Recall	F1 Score
0	0.85	0.97	0.91
1	0.76	0.33	0.46
Average	0.83	0.84	0.82
MCC	0.43		

Figure 3.20: Tables and results of the model n. 9.

The model n. 9 replicated the conditions of the second model, with the only difference consisting in a higher number of targets included in the dataset. Indeed, the protein descriptors were calculated by considering the alignment of all 18 enzymes, thus adding both the 1A and 2A subfamilies of the UGT enzymes. This led to a decrease of similarity within the whole set of sequences, as reported in Figure 3.20. As a result, the total number of data points is widely larger and this has impacted the larger number of features kept after pre-processing. This also reflects an increasing of the learning time.

When introducing additional proteins and despite the increasing of diversity between the included UGT sequences, the model performances got worse only moderately, the class 1 Recall being dropped from 0.39 to 0.33. This suggested that, in these conditions, an identity percentage among sequences around to 40% is still suitable to build reasonable models.

Model n.10 - Descriptors: ECFP, ESshape3D, Physicochemical, Zscales_3.

Composition of dataset	
Compounds	1700
Targets	2
Training set	2380
Test set	1020
Datapoints	3400
Features before processing	2267
Features after removing near zero variance	376
Features after removing correlated descriptors	268

Parameters	
Max features	80
Estimators	100
Class weight	0.5 / 0.5

Results			
	Precision	Recall	F1 Score
0	0.86	0.97	0.91
1	0.74	0.37	0.49
Average	0.83	0.84	0.82
MCC	0.45		

Figure 3.21: Tables and results of the model n. 10

In the model n.10, only two targets were taken in consideration. In detail, to maximize the similarity and according to Table 3.2, UGT2B7 and UGT2B10 were selected, having 88% of identity.

The first obtained result was a sensible decrease of the learning time, while maintaining relatively good performances, as shown in Figure 3.21.

3.4.7.4 Selection of the best protein descriptor

A set of models was generated to verify the different performances of the calculated alignment-dependent physicochemical descriptors for the protein targets. In detail, ten models were built using the same conditions of the model n. 2 and changing the protein descriptors. The results are reported in Table 3.4, where the compiled values correspond to the averages of the results as obtained by running the model three times.

As shown in the table and better clarified by the corresponding plots (see Figure 3.22-Figure 3.25), all the computed scores follow a clear trend. The performance is always outstanding for the prevailing class of compounds, and indeed the non-substrates exhibit F_1 scores not lower than 0.90. Regarding the substrate's class, while the precision scores are quite satisfactory, falling in the range between 0.69 and 0.78, the Recall scores are always under the random threshold. However, among the tested features, the four following descriptors were able to provide better results, both for recall of the substrates' class and for Matthews Correlation Coefficient.

- Zscales_3 and Zscales_5 are generated by a Principal Component Analysis based on physicochemical properties. More in detail, PCA mainly captures lipophilicity, size and polarity/charge for Zscales_3, while the fourth and the fifth component of Zscales_5 are more difficult to interpret relating to properties such as electronegativity, heat of formation, electrophilicity and hardness[26]. The database source for these descriptors includes many non-natural amino acids as shown in Table 2.1, but this does not seem to affect their capability to gather most of the key information for our study.
- ProtFP_3 and ProtFP_8, termed "Protein Fingerprints", are generated by a PCA based on a selection of different amino acidic physicochemical and biochemical properties extracted from the AAindex database[117]. The main difference in comparison with the previous descriptors is that they are focused only on the 20

natural amino acids. Moreover, they select the descriptors with the largest variance using a recursive elimination process which starts with the full set of descriptors.

In conclusion, the analysis of different protein descriptors did not highlight features which were able to significantly improve the predictive power of the model but confirmed that the descriptors used for all previous models, namely Zscales_3, are the best performing ones. This allowed us to use the same protein descriptors also in the LOCO validation step.

Target descriptor	Class 0 Precision	Class 1 Precision	Average Precision	Class 0 Recall	Class 1 Recall	Average Recall	Class 0 F_1 Score	Class 1 F_1 Score	Average F_1 Score	Matthews Correlation Coefficient
ProtFP_feature	0.85	0.72	0.82	0.97	0.34	0.84	0.90	0.46	0.81	0.44
MSWHIM	0.85	0.75	0.83	0.97	0.36	0.84	0.91	0.49	0.82	0.44
ProtFP_3	0.85	0.73	0.83	0.97	0.36	0.84	0.91	0.48	0.82	0.44
Zscales_3	0.86	0.72	0.83	0.96	0.39	0.84	0.91	0.51	0.82	0.45
ProtFP_5	0.86	0.71	0.83	0.96	0.40	0.84	0.91	0.51	0.83	0.45
Tscales	0.86	0.78	0.84	0.97	0.38	0.85	0.91	0.51	0.83	0.41
Zscales_5	0.86	0.75	0.84	0.97	0.39	0.85	0.91	0.51	0.83	0.46
FASGAI	0.85	0.69	0.82	0.96	0.35	0.83	0.90	0.47	0.81	0.41
ProtFP_8	0.86	0.74	0.84	0.96	0.40	0.85	0.91	0.52	0.83	0.47
VHSE	0.85	0.68	0.82	0.96	0.36	0.83	0.90	0.47	0.81	0.41
STscales	0.85	0.73	0.83	0.97	0.36	0.84	0.91	0.48	0.82	0.44

Table 3.4: Protein descriptors performances. All models are built on ECFP fingerprints, ESshape3D fingerprints and physicochemical properties, and include 7 targets.

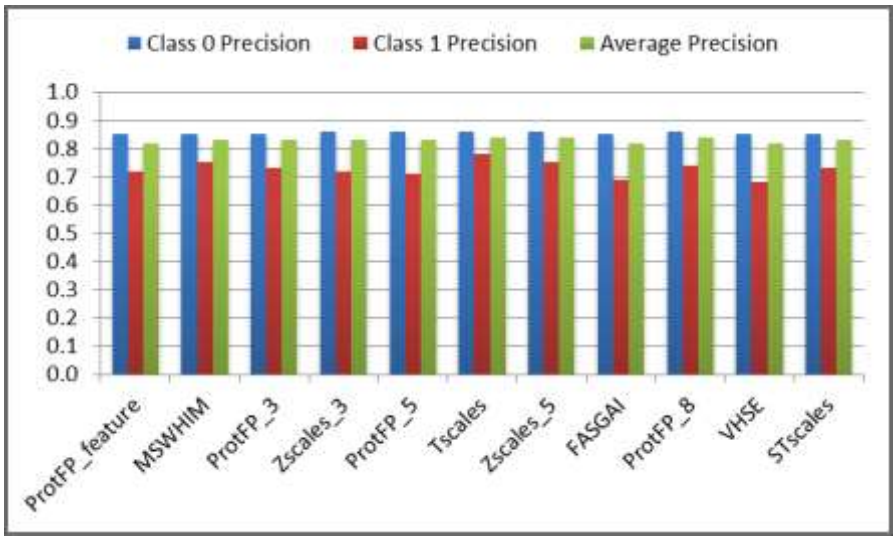


Figure 3.22: Bar plot of precision for each protein descriptor.

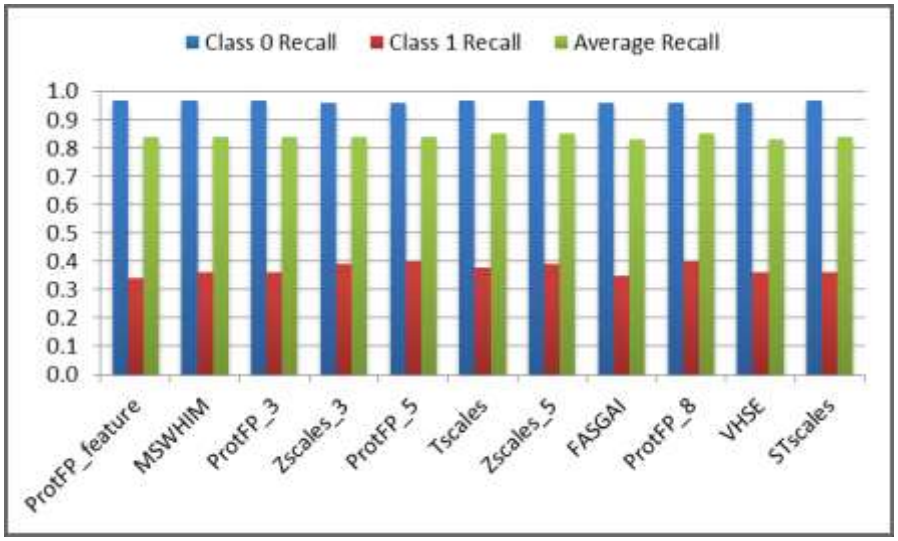


Figure 3.23: Bar plot of recall for each protein descriptor.

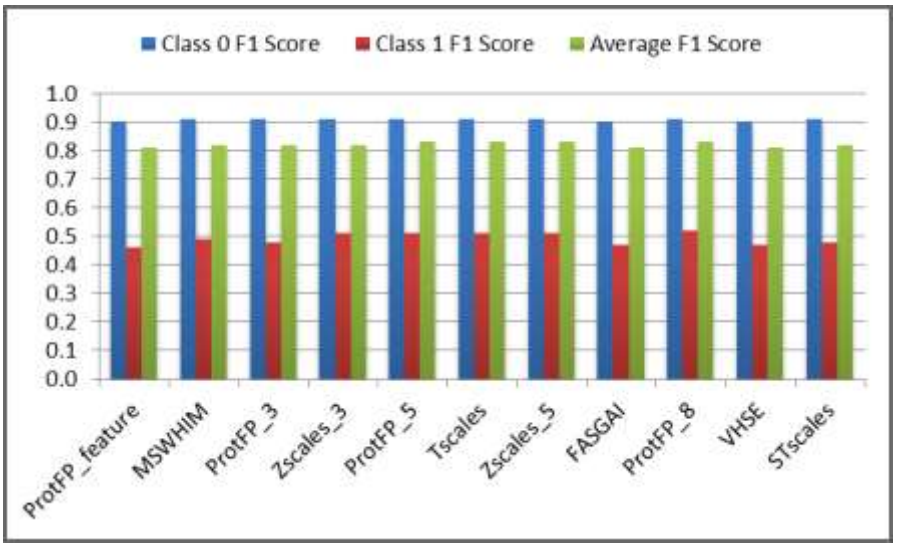


Figure 3.24: Bar plot of F1 score for each protein descriptor.

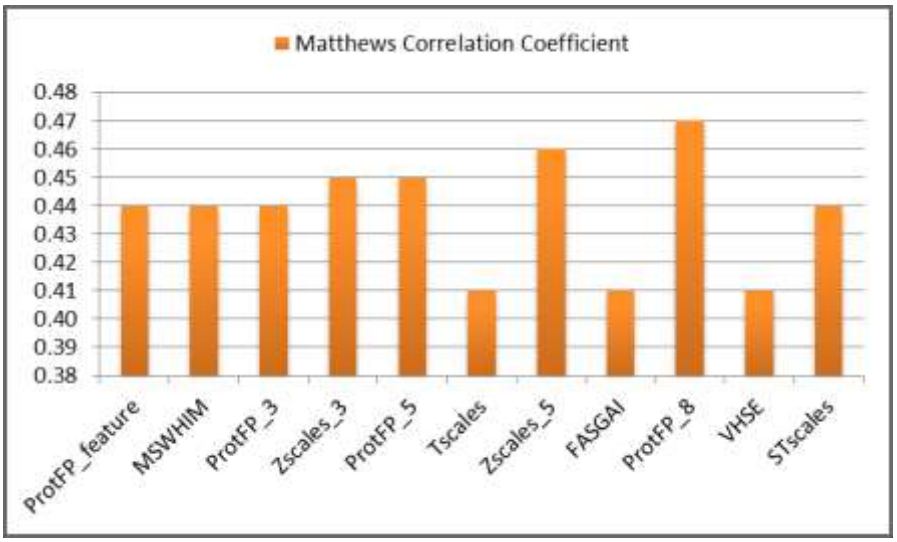


Figure 3.25: Bar plot of MCC score for each protein descriptor.

3.4.7.5 NCV results overall observations

Interesting overall observations about NCV results arise from a comparison of the performances obtained with the ten generated models. To this purpose, we will consider the recall and the precision scores separately.

Figure 3.26 summarizes the main differences between the performances of the NCV models as given by the recall results. We first focused our attention on this measure, as it represents the sensitivity of the model, namely the number of correctly predicted instances out of the total number of the instances belonging to a given class.

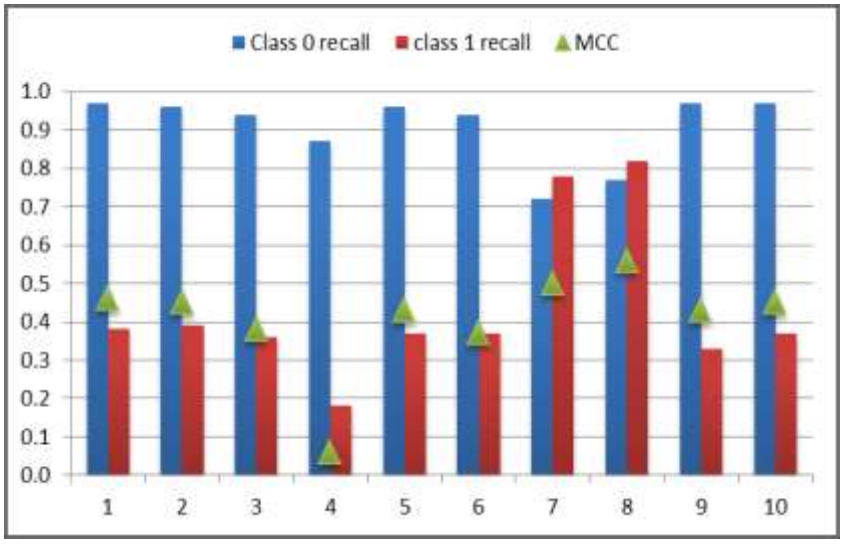


Figure 3.26: Class 0 Recall, Class 1 Recall and Matthews Correlation Coefficient obtained with 10 NCV models.

The recall scores of the non-substrates' class (class 0) and the substrates' class (class 1) as obtained by the 10 considered NCV models are reported as bar plots, in order to reveal the differences in the predictive performances between the

prevailing class and the minority one. As shown in Figure 3.26, the worse results are provided by the model n. 4, in which the ECFP descriptors were excluded from the input dataset. Nevertheless, all the other considered models present unsatisfactory recall values for the substrates, between 0.33 and 0.39, substantially under the random.

The Matthews Correlation Coefficient (MCC) is reported as dot plot and describes the overall performances of the ten models, as it comprises the weighted average between the two Recall scores. Furthermore, for an easy comparison between the MCC results and the recall scores, it is important to consider that the former values are in the range of -1/+1, while the latter are in the range of 0/+1. Therefore, a MCC score around 0.40 can be seen as a truly satisfactory performance.

The obtained results also emphasize that the only strategy which proved successful in balancing the results involved the "random under-sampling" procedure, as applied in models n. 7 and n. 8. While this technique appears to lack in elegance, since it entails data discarding, it represents an efficient method to increase the absolute number of well predicted substrates. This is noticeable when considering the above below confusion matrices which are provided as the output of the models (see Figure 3.27).

Model n. 2

PREDICTED

TRUE	0	1	tot
0	2699	111	2860
1	433	**277**	**710**
tot	3132	388	3570

Results			
	Prec.	Recall	F1 S.
0	0.86	0.96	0.91
1	0.72	0.39	0.51
Ave.	0.83	0.84	0.82
MCC		0.45	

Model n. 7

PREDICTED

TRUE	0	1	tot
0	549	161	710
1	129	**581**	**710**
tot	749	665	1420

Results			
	Prec.	Recall	F1 S.
0	0.86	0.77	0.81
1	0.72	0.82	0.76
Ave.	0.80	0.79	0.79
MCC		0.58	

Model n. 11

PREDICTED

TRUE	0	1	tot
0	67	109	176
1	43	**667**	**710**
tot	110	776	886

Results			
	Prec.	Recall	F1 S.
0	0.65	0.38	0.48
1	0.84	0.94	0.88
Ave.	0.79	0.81	0.79
MCC		0.40	

Figure 3.27: Confusion matrices and results reports of model n. 2, 7 and 11.

The under-sampling procedure does not affect the l number of substrates in the test set, which is 30% of total number of the substrates, thus allowing a comparison among the original models and the balanced ones. In details, the number of corresponding datapoints for a model with 7 targets is 710. This number does not change even when deleting other "non-substrate" datapoints, so generating a "re-unbalanced" dataset, which exactly reproduces the opposite proportion between the two classes in the original dataset. Such a "re-unbalanced" dataset was then utilized to generate the last model n. 11, which was developed maintaining all the other parameters identical to those of the other considered models.

The reported confusion matrices reveal that the true positive rate (namely the correctly predicted substrates) is related to their abundance in the dataset. The number of well predicted substrates rises from 277 in the model n. 2, to 581 in the model n. 7, up to 667 in the model n. 11. At the same time, the true negative rate (namely the correctly predicted non-substrates) drastically decreases due to their deletion when balancing and "re-unbalancing" the dataset.

Moreover Figure 3.26 shows that model n. 2 and model n. 11 display opposite performances, with the recall score for the minority class around 0.40 and that for the majority class around 0.7. Nevertheless, the performance of the latter is slightly worse, as a result of the decrease of the total number of instances in the dataset. Indeed the Matthews Correlation Coefficients decrease from 0.45 to 0.40. In case of prediction on an external test, the balanced and the "re-unbalanced" datasets can be useful as a double check of the results after using the full dataset.

Figure 3.28 summarizes the main differences between the performances of the NCV models, as revealed by precision results. Indeed, a model which simply predicts all instances as active, achieving 100% recall, does not provide trustworthy predictions. To completely evaluate the predictive power of a model, the false positive rate has to be taken into account, and this is possible by considering the precision score. Precision indeed evaluates the number of correctly predicted instances, for a given class, out of the total number of instances predicted to belong to that class, namely the sum of true and false positive.

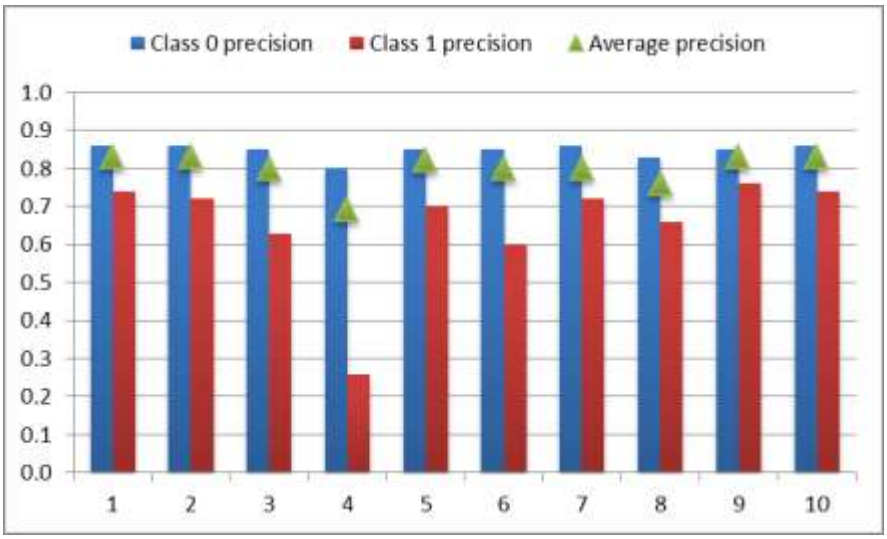


Figure 3.28: Class 0 precision, Class 1 precision and average precision obtained with 10 NCV models.

The precision scores of the non-substrates' class and the substrates' class are reported as bar plots, in order to reveal potential differences in the related predictive accuracy, and the average of precision for both classes is reported as dot plot, as a measure of the overall precision performances.

As shown in Figure 3.28, the worse results are provided, also in this case, by the model n. 4, in which the ECFP descriptors were excluded from the input dataset. Differently from the recall performances, the results of the other models are sufficiently balanced, underlying that the effect of the unbalanced dataset is weaker when weighting the results on the false positive rate.

The first two models, already considered as the referring results, are the best balanced ones, along with the model n.8, built on the whole set of targets, which reaches scores equal to 0.85 and 0.76 for the non-substrates' and the substrates' classes, respectively.

Therefore, for what concerns the precision performances, the random under-sampling procedure is not necessary, and even not useful, because when applied, it does not improve the results (see Figure 3.28).

3.4.8 LOCO results

The NCV identified the model n. 2 as the best one in terms of combination of features in the input file. The same dataset was then used to validate the model by the procedure of "Leave One Compound Out".

According to this procedure, a group of data points referring to the same compound is left out, then the model is trained on the remaining data points and finally is tested on the previously discarded instances. The model can correctly predict all the labels of the discarded group, none of them or just some of them. As seen for the NCV results, the predictions are reported in a confusion matrix and the LOCO results can be expressed in terms of precision, recall and F_1 scores. The analysis is repeated for all the groups included in the dataset and, in the end, the list of the corresponding predictive performances is generated.

Some observations are possible about precision and recall values. First of all, differently from what happened when performing NCV, they are referred just to

one class, which is the class of the compound left out. Moreover, for this reason, the precision value is always either 0 or 1, while the recall value can be also a continuous number comprised between 0 and 1, when only some data points are correctly predicted.

3.4.8.1 Recall LOCO results

The first LOCO analysis was conducted to verify the number of correctly predicted compounds, as returned by the Recall values. The obtained results are reported in Figure 3.29.

Model n. 12

		PREDICTED		
		0	1	tot
TRUE	0	1312	50	1362
	1	119	139	338
	tot	1431	189	1700

Results	
	Recall
0	0.96
1	0.41
Total	0.85

Figure 3.29: Confusion matrix and results of the LOCO analysis performed on the same dataset used in model n. 2 for NCV analysis.

In many cases, LOCO results are better than NCV results just as a consequence of the larger size of the training set. In other cases, LOCO results can be worse and reveal biases hidden in NCV models, due to the presence of the same compound both in the training and the test sets. Moreover, in LOCO models all data points are

predicted in the same analyses, revealing the overall predictive performances and, sometimes, resulting in worst results.
In our model, as it is shown in the reported results, the overall LOCO recall values are comparable to the ones obtained by NCV. This finding confirms that the strategy consisting in the previous grouping of data according to the compound name, before the splitting, proved successful. Moreover, the results reveal that the different number of instances in the training set is not particularly relevant, and that the model is robust enough to maintain similar predictive powers, even when trained on the whole dataset.

3.4.8.2 Probability LOCO results

Using the specific function of Scikit-Learn called "*predict_proba*", it is possible to print the probability associated with the prediction of each instance. More in detail, Random Forest assigns to each test instance two different probabilities, which are associated to the predicted label "0" (namely non-substrates) and to the predicted label "1" (namely substrates), respectively. Then, the algorithm chooses as the final predicted label the one with the higher probability. Therefore, each final predicted label has a probability within the range 0.51 to 1. This probability can be considered as a measure of the reliability of the final prediction: for correctly predicted labels, the higher is the probability, the stronger is the predictive power of the model; conversely, for incorrectly predicted labels, the higher is the probability, the weaker is the model performance.

This probability measure can be also utilized to specifically evaluate the probability associated only with the correct prediction of each test instance. In this case, the probability takes continuous values in the whole range between 0 and 1. The results of this analysis are illustrated in the box-plot in Figure 3.30.

In the plot, the compounds are subdivided according to their true label, which corresponds to their classification into substrates or non-substrates based on experimental results. Then, the probabilities associated to a correct prediction are reported for the compounds of both classes. The most striking observation to emerge from the data comparison is the different distribution of the probabilities for the compounds predicted to be substrate (class 1, blue bar) and non-substrates (class 0, green bar).

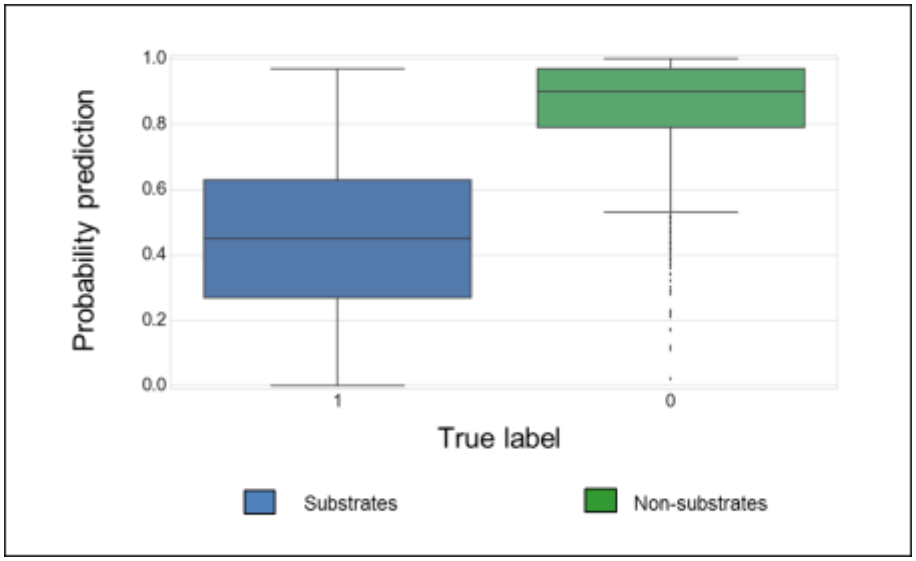


Figure 3.30: Box plot of probabilities of correct prediction for both classes compounds.

As depicted in Figure 3.30, the probabilities associated with the predicted substrates are spread out within the whole range between 0 and 1, and the median of data is below 0.50. This reflects the low recall value for the prediction of the substrate's class. On the opposite, the majority of the probabilities associated with the predicted non-substrates is in the top half of the range, with a median higher than 0.90, as expected from the outstanding performances for this class.

3.4.9 Applicability Domain

The LOCO validation method offers also the opportunity for an Applicability Domain (AD) analysis. The definition of the AD of the model is always a crucial step to increase the quality of the model itself[118]. A model will yield reliable predictions when its assumptions are valid and unreliable predictions when they are violated[119]. Therefore, it is important to define the space where model predictions are reliable.

One of the possible approaches to applicability domain estimation is based on a similarity analysis among the training set: a compound will have a reliable prediction if it is enough similar to the ones used by the algorithm in the learning phase[120]. The similarity can be calculated according to many criteria and the performance of the model is plotted against the whole range of similarity in the training set.

Here, we present a first attempt to define the applicability domain of the model n. 2. It contains an important approximation, since the performance of the model takes advantage from three compound descriptors (ECFP, ESshape3D and physicochemical properties), while the similarity is here measured just according to one of them, the ESshape3D descriptor. This choice can be justified considering that the ESshape3D fingerprints account for the 3D structure of molecules and thus they are able to distinguish between different stereoisomers.

Since these descriptors take discrete values, the similarity matrix, comprising the whole set of molecules, is computed as the Euclidian distance between each pair of compounds. The distance between a given test compound and its first neighbour is calculated from the matrix and then associated to the recall score for that compound. The distances are clustered according to their values and the percentages of correct predictions for each distance cluster are reported in the plot.

The analysis considers the predictions as a whole, without distinguishing between the two classes.

The results of the analysis are shown in the Figure 3.31. A clear trend is shown in the dot plot: the performances of prediction tend to decrease when the first neighbour distance increases, which can be considered as a measure of the similarity between a test compound and the whole set of training compounds. This is consistent with what expected: the predictive power of the model is directly correlated with the similarity of the test molecule to the ones in the training set. The applicability domain evaluation can be utilised to foresee the reliability of the prediction for a new compound by measuring its first neighbour distance with the molecules in the training set.

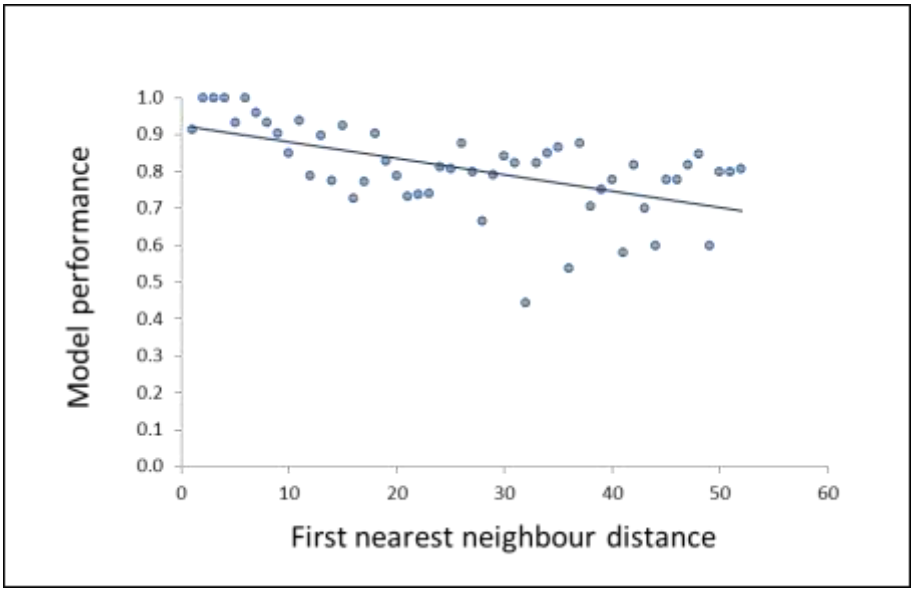


Figure 3.31: **Applicability Domain analysis.** The performances of the model get worse with the increasing of first nearest neighbour distance.

3.4.10 Conclusions

This study involved the generation of a classification model able to predict the regioselectivity of the glucuronidation reactions. The objective was reached by exploiting data included in the metabolic reaction database MetaSar, which has been manually curated and critically reviewed by our research group. The predictive technique used was PCM, here applied for the first time in the field of metabolism prediction.

The global predictive accuracy of the model is encouraging, even though the recall results are unbalanced, as a direct consequence of an input dataset including two unbalanced classes of compounds. To increase the recall performances regarding the minority class, the under-sampling procedure has been successfully applied. Therefore, in case of prediction on an external test, the balanced and the "re-unbalanced" dataset can be used as a double check of results, after the use of the full dataset.
In detail, we can identify two interesting models. The first is the model n. 2, which is built on the whole dataset, and gives satisfying results for three out of the four measures, namely the precisions for both classes and the recall for only the non-substrates' class. The second is the model n. 7, which is instead built on a subset of the available data, but reaches outstanding results for both classes, measured by precision as well as recall scores.

The model can be improved to predict 8 different subclasses of glucuronidation reactions reported in the MetaSar reactions classification, or to predict selectivity among different UGTs isoforms. Moreover, the same predictive technique can be applied to the other metabolic classes of reactions. This can lead to a collection of models able to singly predict the regioselectivity of the specific metabolic reaction, as well as, on the whole, the entire metabolic pathway occurring to new substrates.

Chapter 4 – Purinergic Receptors

4.1 Introduction

The adenosine 5'-triphosphate, well known as ATP, is the ubiquitous energy currency of all living organisms. Its role in the transport of energy within the cell has been recognized for many years. Thanks to its high phosphate-transfer potential, ATP is involved in a broad variety of biological processes, such as energy metabolism, active transport, biosynthetic reactions, motility and cell division.

Moreover, ATP can influence many biological processes also in the extracellular environment. It can be released from the cytoplasm of several cell types as a consequence of damage, and it is physiologically released by exocytosis from platelets and neurons. The role as a signaling molecule is held by interacting with specific receptors on the surface of different cells, before being quickly metabolized by ectonucleotidases.
The field of purinergic signaling has grown steadily over the last forty years, since Holton confirmed the liberation of ATP on antidromic simulation of sensory nerves[121] (in 1959), and later Burnstock and co-workers discovered that ATP is the principal transmitter of some of the "non-adrenergic, non-cholinergic" nerves[122]. Now, the purinergic receptors are found to be expressed throughout the human body, including the nervous, the cardiovascular and the immune systems. ATP as a signaling molecule is thus implicated in a wide range of physiological functions, such as synaptic transmission, smooth muscles contraction, taste perception, inflammation and nociception. So that, purinergic receptors hold great interest as new therapeutic targets for inflammatory, cardiovascular and neuronal disease.

In the present Chapter, the attention will be focused on the $P2X_3$ receptor, with the primary objective of going through its mechanism of action and, in particular, its allosteric modulation. To better frame the performed computational studies, an

introductive description of the main features of the purinergic receptors will be given in the next sections.

4.1.1 Purinergic receptor classification

The purinergic receptors are membrane proteins activated by ATP and its metabolites. In detail, two main classes can be recognized: P1, which are generally known as adenosine receptors, and P2, which respond to extracellular ATP.
Focusing on P2 receptors, two main families can be recognized, on the basis of the type of protein: P2Y receptors, which are G protein-coupled receptors, and P2X receptors, which are ligand-gated ion channels.
P2X receptors are homo- or hetero-trimers resulting by the combinations of different monomers. Up to know, seven subunits have been cloned and characterized (1-7). Each monomer consists of 379-472 amino acids, with the exception represented by the $P2X_{7,}$ which includes 595 amino acids, due to the increased length of its C-terminus.

4.1.2 Molecular architecture

P2X receptors are trimeric ion channels that exist in two main conformations: the closed conformation (apo), which corresponds to the inactive form, and the open conformation (holo), in which the receptor is activated by the binding of ATP.
The architecture of the trimer started to be clarified thanks to the resolution of two important x-ray structures: the zebrafish $P2X_4$ channel in closed state (PDB Id 3HV9)[123] and the later resolution of the same protein in its open state (PDB Id 4DW1)[124], in which the receptor is co-crystallized with three ATP molecules. Together with this second structure, a third crystal structure of the apo state having

a better resolution was published (PDB Id 4DW0)[124]. These resolved structures confirmed many of the functional experiments carried out since the mid-1990s. In particular, they revealed the molecular architecture of the trimeric cation channel and its biological assembly, and they elucidated the ATP binding mode and the ionic flow through the channel.

4.1.2.1 The monomer structure

The resolved purinergic receptor structures represent a model for the whole P2X receptors family, and key information regarding the monomer of the human P2X3 can be derived by the alignment of its sequence with that of the zebrafish P2X4 monomer, reported in Figure 4.1. The alignment shows the most conserved regions and the sequence correspondences in the crucial parts of the receptors, such as the ATP binding site.

In the zebrafish resolved crystals, each monomer brings to mind the shape of a dolphin, and shares a common topology, characterized by the following regions (Figure 4.2):

- two transmembrane helices, TM1 and TM2, which form the cation permeable channel, akin to the tail of the dolphin;
- two short intracellular termini;
- a large extracellular domain, rich of glycosilated residues and disulfide bonds, corresponding to the head, the upper body, the two flippers, the dorsal fin and the lower body of the dolphin. This region also includes the binding site for ATP, competitive antagonists and modulatory metal ions.

```
P2X4_zf       1 MSESVGCCDSVSQCFFDYYTSKILIIRSKKVGTLNRFTQALVIAYVIGYV     50
                    :.|...    ||.|.|:|.::::|..:|.:||..|.|:|:|.:|:|
P2X3_h        1 ----MNCISD----FFTYETTKSVVVKSWTIGIINRVVQLLIISYFVGWV     42

P2X4_zf      51 CVYNKGYQDTDTVL-SSVTTKVKGIALTNTSELGERIWDVADYIIPPQED     99
                .::.|.||..||.: |||.|||||..|     ...|:.||:||:.|||..
P2X3         43 FLHEKAYQVRDTAIESSVVTKVKGSGL-----YANRVMDVSDYVTPPQGT     87

P2X4_zf     100 GSFFVLTNMIITTNQTQSKCAENPTPASTCTSHRDCKRGFNDARGDGVRT    149
                ..|.::|.||:|.||.|..|.|: .....|.|...|  |.....|.|:.|
P2X3_h       88 SVFVIITKMIVTENQMQGFCPES-EEKYRCVSDSQC--GPERLPGGGILT    134

P2X4_zf     150 GRCVSYSASVKTCEVLSWCPLEKIVDPPNPPLLADAENFTVLIKNNIRYP    199
                ||||:||:.::|||:..|||.|  ||....|::.:|||||:.|||:||:|
P2X3_h      135 GRCVNYSSVLRTCEIQGWCPTE--VDTVETPIMMEAENFTIFIKNSIRFP    182

P2X4_zf     200 KFNFNKRNILPNINSSYLTHCVFSRKTDPDCPIFRLGDIVGEAEEDFQIM    249
                .|||.|.|:|||:.:..:..|.|....||.|||.|:||:|..|.:||..:
P2X3_h      183 LFNFEKGNLLPNLTARDMKTCRFHPDKDPFCPILRVGDVVKFAGQDFAKL    232

P2X4_zf     250 AVHGGVMGVQIRWDCDLDMPQSWCVPRYTFRRLDNKDPDNNVAPGYNFRF    299
                |..|||:|::|.|.||||.....|:|:|:|.|||:....::|:|||||||
P2X3_h      233 ARTGGVLGIKIGWVCDLDKAWDQCIPKYSFTRLDSVSEKSSVSPGYNFRF    282

P2X4_zf     300 AKYYKNSDGTETRTLIKGYGIRFDVMVFGQAGKFNIIPTLLNIGAGLALL    349
                |||||..:|:|.|||:|.:|||||:|:|.||||||||::::..|....:
P2X3_h      283 AKYYKMENGSEYRTLLKAFGIRFDVLVYGNAGKFNIIPTIISSVAAFTSV    332

P2X4_zf     350 GLVNVICDWIVLTFMKRKQHYKEQKYTYVDDFGL--------LHNEDK--    389
                |:..|:||.|:|.|:|....||.:|:..|::..|        ::..|:
P2X3_h      333 GVGTVLCDIILLNFLKGADQYKAKKFEEVNETTLKIAALTNPVYPSDQTT    382

P2X4_zf     390 ---------------    389

P2X3_h      383 AEKQSTDSGAFSIGH    397
```

Figure 4.1: P2X4zf and P2X3h alignment. The TM domain are highlighted in yellow, the N-terminal part in light blue, the C-terminal in green, the residues implicated in the ATP binding in red, the residues involved in disulfide bonds in violet . The pipes (|) indicate the conserved residues, while the colons (:) indicate the homologues one.

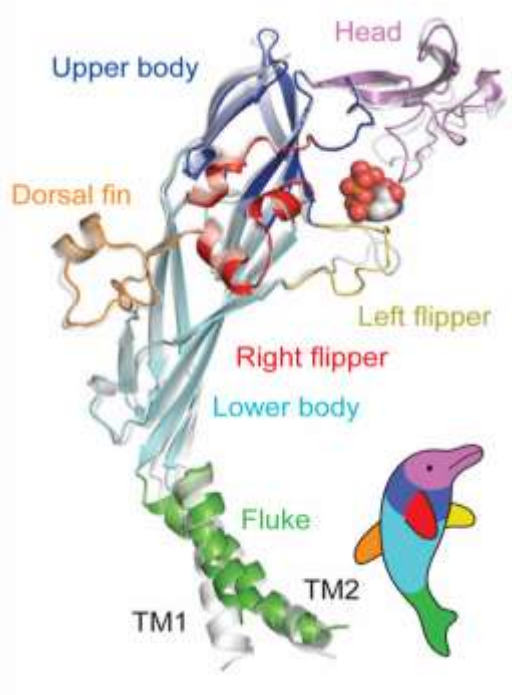


Figure 4.2: the dolphin-like shape of the P2X monomer.
(Reproduced from Hattori *et al*.[124])

In detail, the central architecture of the extracellular body domain is characterized by a transthyretin-like β-sandwich motif. This segment appears to be rigid, and perhaps even resistant to conformational changes, because the two β-sheets of the sandwich are knit together by extensive contacts[123]. The stability of the monomer is also kept by disulfide bonds present in the extracellular ectodomain, which involve the following residues:

- 258-267: connecting the ends of two antiparallel β-strands in the lower part of the dolphin's body;
- 214-224: in the dorsal fin;
- 113-164; 124-147 and 130-158 located in the head region[125].

4.1.2.2 The trimer structure: apo and holo conformations

The purinergic receptors are composed of three monomers that can build homomeric or heteromeric trimers, organized in a chalice-like shape. Inside the receptor, three vestibules can be identified whose amplitude differs between the apo and the holo state, due to the structural rearrangement which happens during the channel activation (Figure 4.3-C).

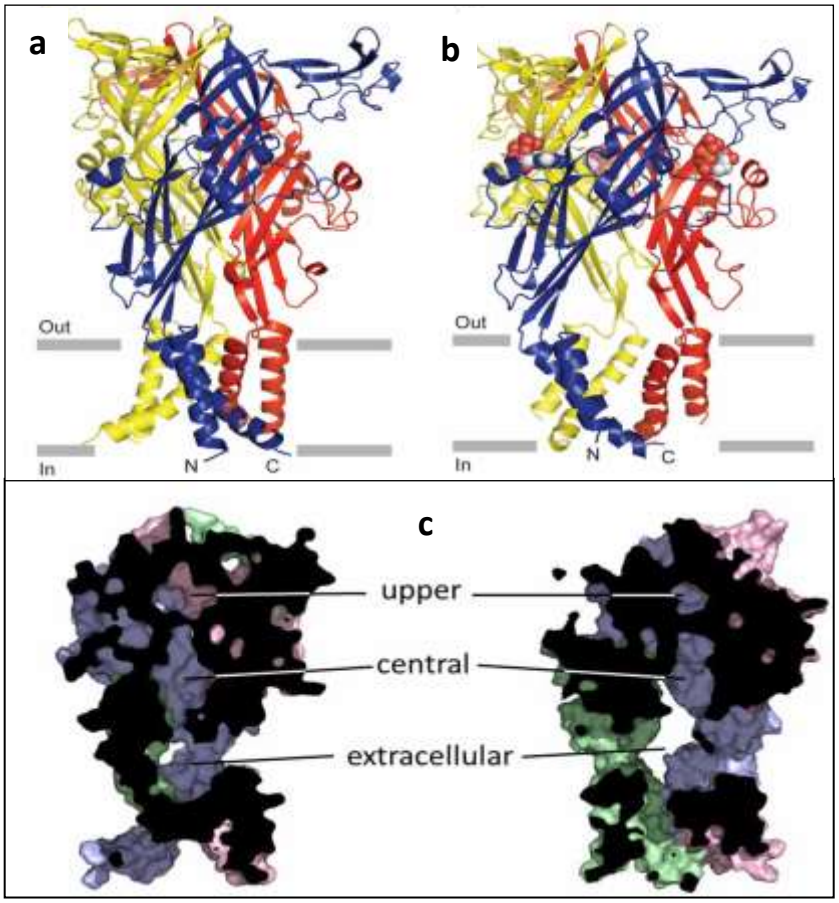


Figure 4.3: The trimer structures of the P2X4 zebrafish receptor. On the left, the apo conformation **(a)**, on the right, the holo conformation **(b)**. In the bottom of the figure the three internal vestibules of both conformations are depicted. (Reproduced from Hattori *et al.*[124])

The apo conformation is the inactive structure, with the channel in a closed state (Figure 4.3-a). The shape of the TM region is like an hourglass and is formed by three pairs of helices, two from each of the three subunits. Considering a single monomer, the transmembrane helices are oriented approximately in antiparallel way each other and form an angle of about 45° with the membrane normal. The inner TM2 segments cross each other about in the middle of the membrane length. Therefore, the mutual arrangement of the TM segments constitutes the closed resting ion channel, where the ionic permeation pathway is occluded.
The extracellular domain of each dolphin wraps around one neighbor with a right-handed twist, establishing extensive interactions. The major interfaces that characterize this conformation are body to body, head to body and left flipper to dorsal fin. The upper region of the body domain contacts the other subunits, whereas there are no contacts at the base of the extracellular domain, proximal to TM segments. The body to body contacts involve a set of highly conserved external residues and are responsible for the structural rigidity of this region of the receptor. On the opposite, the residues in the left flipper and the dorsal fin are less conserved and these regions are characterized by a greater flexibility.

The holo conformation is the active structure, with the channel in an open state (Figure 4.3-b). The overall protein structure of the ATP-complexed subunit is similar to that of the apo one, as demonstrated by the low root mean squared deviation between the two overlapped resolved structures (as computed considering the Cα atoms, about 1.8 Å).
The substantial conformational changes associated with the ATP binding are localized at the interfacial regions adjacent to the ATP pocket in the extracellular domain, and within the ion conducting TM domain. Therefore, the lower body domain of the holo monomer does not superimpose well to that of the apo monomer, because of an outward flexing of the body domain resulting by the ATP binding.

This movement directly expands the lower extracellular vestibule, increasing the central separation between the monomers. Consequently, it leads to the iris-like motion of the lower body domain that elicits the separation of the TM helices, so opening the pore.

4.1.2.3 The ATP binding site

The ATP binding site is located in the extracellular domain, at the interface between two monomers. Therefore, in the whole trimer, there are three equivalent ATP binding sites, one at each of the three pairs of monomeric interfaces. In detail, the ATP biding pocket is cradled by the head domain, the upper body and the left flipper of the chain "A", contacting the lower body and the dorsal fin of the chain "B", of two adjacent monomers (Figure 4.4-a,b,c).

The ATP binding site is rich in positively charged residues, which establish ionic interactions with the negatively charged ATP structure (see Figure 4.4-d). The ATP molecule, when bound to the trimer, adopts a particular U-shape conformation in which the β- and the γ-phosphates are folded toward the adenine ring, and the base-sugar complex is in an "anti" conformation. Thanks to the resolved structure and the mutagenesis studies, the ATP binding mode is now completely clarified and the intermolecular contacts stabilizing the ligand-protein complex are identified.
The three negatively charged phosphates establish salt-bridges and H-bond interactions with Lys70 and Lys72, located in the center of the "U". Asn296 and Lys316 from chain "A" elicit additional contacts with the β-phosphate group, while Lys72, Arg298 and Lys316 participate in interactions with the γ phosphate group.The adenine base, which is deeply buried in the pocket, is involved in three H-bonds stabilized by the side chain of Thr189 and the backbone carbonyl oxygen

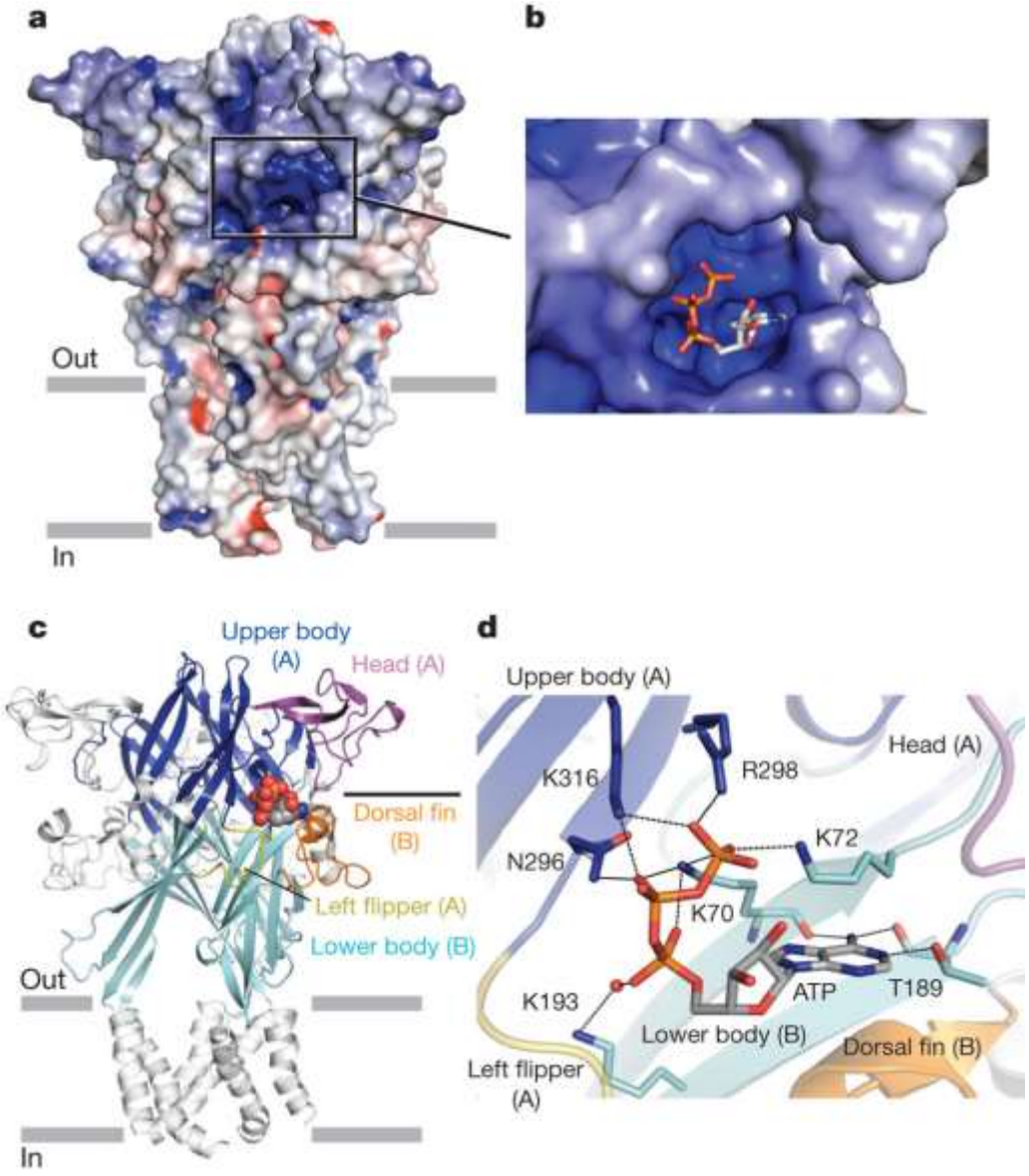


Figure 4.4: ATP binding site in holo trimer structure (PDB Id 4DW0) (a,b) An electrostatic potential surface of ΔP2X4-C contoured from −10 kT (red) to +10 kT (blue) (dielectric constant: 80), and its close-up view. **(c)** The regions forming the ATP-binding pocket. The ATP molecule is shown in sphere representation. **(d)** Close-up view of the ATP-binding site. The oxygen atom from the glycerol molecule is shown in sphere representation. Black dashed lines indicate hydrogen bonding (<3.3 Å). (Reproduced from Hattori *et al.*[124])

atoms of Lys70 and Thr189. Leu191 in the lower body and Ile132 in the dorsal fin establish also hydrophobic interactions with the adenine base.
Finally, the ribose ring of ATP is contacted only by Leu217 in the dorsal fin through hydrophobic interactions, while the O2 and O3 oxygen atoms of the ribose are solvent accessible.

4.1.2.4 The pore conformation

In the closed state, looking from the extracellular surface, Leu340 and Asn341 define the extracellular boundary of the ion channel gate, with the hydrophobic side chains of every Leu340 occluding the pore. On the opposite side of the membrane, Ala347 and Ala346 define the cytoplasmic gate. Moreover, the center of the gate is occupied by Ala344, which corresponds to the closest contact between the TM2 helices[123].
In the open pore, the same residues are positioned far from the central threefold axes, allowing the cation to pass inside the cell. Interestingly, there are no hydrophilic residues in the middle of the pore. Therefore, water molecules coordinated to permeating cations likely interact with the backbone atoms[124].

The pore can become permeable also to large organic cations, such as N-metyl-D-glucamine (NMDG), thanks to the well-known “pore-dilatation” phenomenon, which causes an enlargement of the pore up to about 7.3 Å of diameter. According to the two-voltage clamps studies by Hattori *at al.*, the pore dilatation appears upon to 5-minute application of saturating ATP. They found that the evoked current remains constant, suggesting that the resolved structure represents a non-pore dilated open state.

Thanks to the 4DW0 holo resolved structure, also the pathway by which the hydrated ions enter and exit through the channel has been clarified. This pathway involves the lateral fenestrations nearly above the membrane pore. Once the ions pass through the fenestrations, the highly acidic central vestibule attracts cations and repels anions.

4.1.3 The mechanism of channel activation mediated by ATP

The comparison of the extracellular regions in the apo and the holo crystals elucidates how ATP binding leads to channel activation. A clear scheme of the movements occurring is depicted in Figure 4.5.

First, when binding within the inter-subunit cleft, ATP promotes closure between the head and the dorsal fin domains, inducing the movement of the dorsal fin domain up toward the head domain. Simultaneously, ATP pushes out the left flipper from the pocket. Because the dorsal fin and the left flipper are structurally coupled to the lower body domain, this movement promotes a concomitant outward flexing of the lower body domain that substantially expands the extracellular vestibule, increasing the separation by about 10 Å. During the flexing, the domain behaves as a rigid body and each subunit rotates by about 8° around an axis located in the upper body. Finally, the movement reaches TM1 and TM2 domains, which flex and rotate, opening the ion channel pore in an iris-like way. In particular, TM1 and TM2 undergo a rotation of 10° and 55° counterclockwise and increase their tilt angle by about 8° and 2°, respectively. The consequence of the iris-like movement is the expansion of about 3 Å of the central pore.
In the ATP bound state, the interactions between the helices are broken, and the open pore conformation is stabilized by new contacts between subunits involving Leu346 and Ile353.

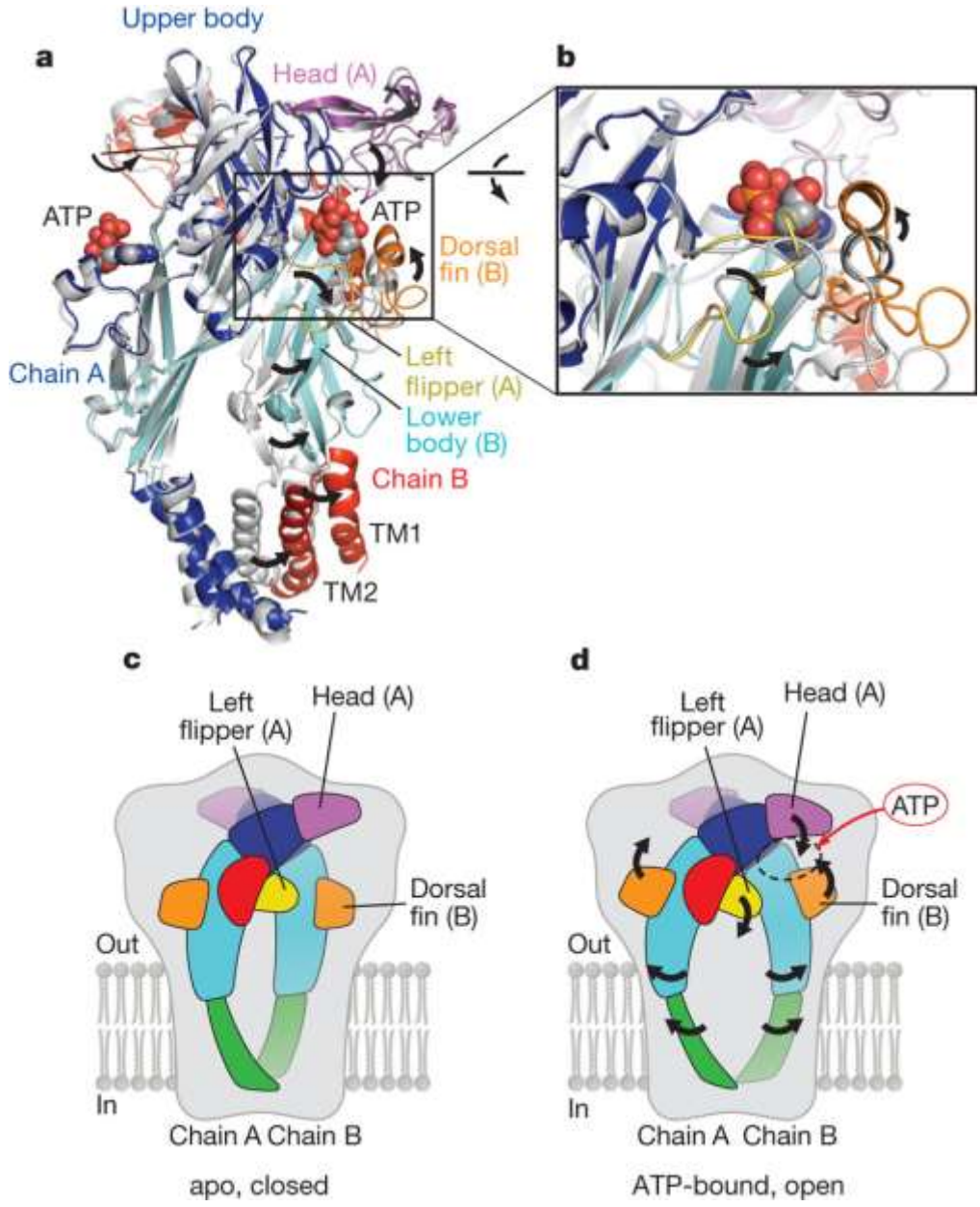


Figure 4.5: ATP mechanism of activation. (a) The holo structure (coloured) is superimposed to the apo (in grey). Only two subunits are shown. The rotation axis describes the superposition of the apo subunit onto the ATP-bound one. **(b)** Close-up view of the conformational changes resulting from ATP binding. **(c, d)** Cartoon model of the ATP-dependent activation mechanism. The black arrows denote the movement from the apo closed state (c) to the ATP-bound open state (d). (Reproduced from Hattori *et al.*[124])

4.1.4 Multiple allosteric conformational states in P2X receptors

The P2X receptors can be classified as "allosteric proteins" with reference to their capability to interconvert among several conformational states (Figure 4.6). In normal conditions, the ATP binding drives the receptor from the resting closed channel state (R), to an active open channel state (O_1), selective for small cations.

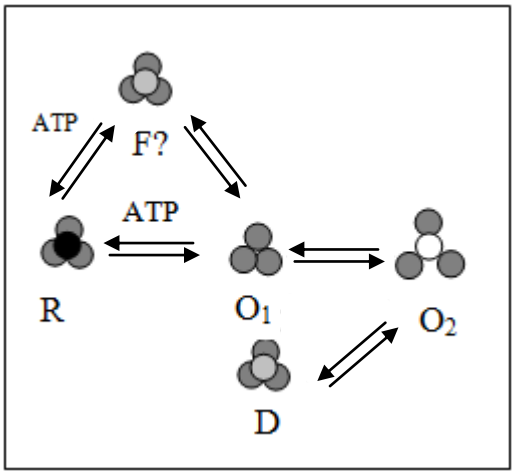


Figure 4.6: Multiple allosteric conformational model.
(Reproduced from Iiang *et al*. [126])

However, as before mentioned, a subset of P2X receptors, including $P2X_2$, $P2X_4$, $P2X_7$ and $P2X_{2/3}$, displays an eventual second open state (O_2) induced by a prolonged application of ATP[127]. In this other conductive state, known as pore dilation, the pore becomes progressively permeable to larger cations, such as NMDG and propidium dyes. During the pore dilation, structural rearrangements occur in both the pore and the cytosolic regions of the protein, but the amino acids involved in this conformational change are yet unknown[126].

Generally, sustained ATP applications lead the receptor to a desensitized closed channel state (D), which is refractory to further activation. The P2X receptor class presents very different desensitization kinetics: the $P2X_1$ and $P2X_3$ display fast and nearly complete desensitization within 2 seconds of application of ATP, while, $P2X_2$, $P2X_4$ and $P2X_7$ show nearly no desensitization when expressed in HEK293 cells[128].

Moreover, recent studies supports also the existence of an intermediate closed channel state (F) that would precede the open state and follow the resting one, as already proposed for the pentameric Nicotin receptors[128].n

4.1.5 Physiological function of the P2X receptors

All P2X receptors are characterized by a simple mechanism involving the opening of the central pore in response to ATP activation. The consequent permeation of cations, especially Ca^{2+}, leads to the depolarization of the cells and, at the same time, to the generation of the downstream calcium signaling[126]. Then, the variety of different effects elicited by these receptors is mainly due to their different localization, expression and modulation (Table 4.1).

Among the most observed pathways, we can include the following examples: P2X1 receptors are involved in the vas deferens contraction and the male infertility, as demonstrated by studies using the first knock out mice[121]; $P2X_3$ and $P2X_{2/3}$ receptors are implicated in the nociception, being expressed in particular in terminal sensory nerves, where they detect ATP released from peripheral tissue or visceral organs[121]; $P2X_4$ receptors are expressed in the microglia and they are recognized to play a role in mediating neuropathic pain[121]; $P2X_5$ receptors are found in sensory neurons, where they are involved in sensing muscle ischemia; $P2X_7$ receptors play a substantial role in inflammation and immunity; finally $P2X_6$

receptors are likely expressed as heteromeric receptors with $P2X_2$ and $P2X_4$ in motoneurons of the spinal cord, with a still almost unknown role.

RECEPTOR	MAIN DISTRIBUTION
$P2X_1$	Smooth muscle, platelets, cerebellum, dorsal horn spinal neuron
$P2X_2$	Smooth muscle, CNS, retina, chromaffin cells, autonomic and sensory ganglia
$P2X_3$	Sensory neurones, nucleus tractus solitarii, some sympathetic neurones
$P2X_4$	CNS, testis, colon
$P2X_5$	Proliferating cells in skin, gut, bladder, thymus, spinal cord
$P2X_6$	CNS, motor neurones in spinal cord
$P2X_7$	Apoptotic cells in immune system, pancreas, skin etc.

Table 4.1: P2X receptors main distribution

$P2X_3$ receptors, on which the present chapter is focused, are predominately and selectively localized on small to medium diameter C- and Aδ-fiber of primary afferent neurons (PAN), within the dorsal root ganglion and the cranial sensory ganglia. They are also present on the respective peripheral nerve terminals in tissues comprising skin, joints and hollow organs.

PAN can be grouped in more types which are characterized by differential morphological properties, speed of conduction, molecular markers and receptor patterns on their surface. These diversities among sensorial neuron types reflect functional differences: they recover a wide range of physio-pathological roles from

low threshold (non-nociceptive), proprioceptive, mechanosensitive, and thermosensitive detection, to high threshold fibers (nociceptive).

Although the expression of $P2X_3$ receptors appears to be sparse elsewhere, there are reports indicating its localization in epithelial cell population, for example within the urinary bladder, and on brainstem neuron's dendrites. In any case, thanks to these fibers, the signal induced by the painful stimulation in periphery can reach the central ganglion through the PAN system. ATP can be released by various cells as a result of tissue inflammation, injury or stress, as well as visceral organ distension and stimulate these local nociceptors[129].

Because of its specific and limited location, this receptor offers unique opportunity to investigate sensory and nociceptive mechanisms; moreover, its inhibition could provoke a lower likelihood of adverse effects in brain, gastrointestinal or cardiovascular tissues, effects that remain limiting factors for many existing painkillers.

4.1.6 $P2X_3$ antagonist: the current situation and future perspectives

P2X3 antagonism has a uniquely broad range of activities across visceral, inflammatory and neuropathic models. The identification of potent and selective inhibitors is an interesting goal in drug discovery, with many fertile potential applications.

Prior to 2000, there were no reports of "drug-like" small molecules that selectively antagonized the activation of $P2X_3$ receptors by ATP. The existing antagonists were large poly-anionic molecules with little specificity (Suramin, PPADS) or nucleotides (TNP-ATP), neither of which provided ideal starting points for medicinal optimization[130].

In 2002, more promising leads were reported and patented by Abbott Laboratories, in particular A-317491, which offered sub-micromolar potency, competitive and

selective antagonism at $P2X_3$ and $P2X_{2/3}$ receptors, and which presented properties that could offer applicability for in vivo studies in sensory models[131] (Figure 4.7). However, all these antagonists were poly-anionic molecules, and despite having a good *in vivo* plasma half-life, they were almost completely bound to proteins. For this reason, they had essentially no permeability from enteric into systemic compartments. After all, the removal of the acidic functions led to loss of activity, and, despite considerable efforts, this class of competitive ATP antagonists could not be chemically optimized in developable small molecule candidates.

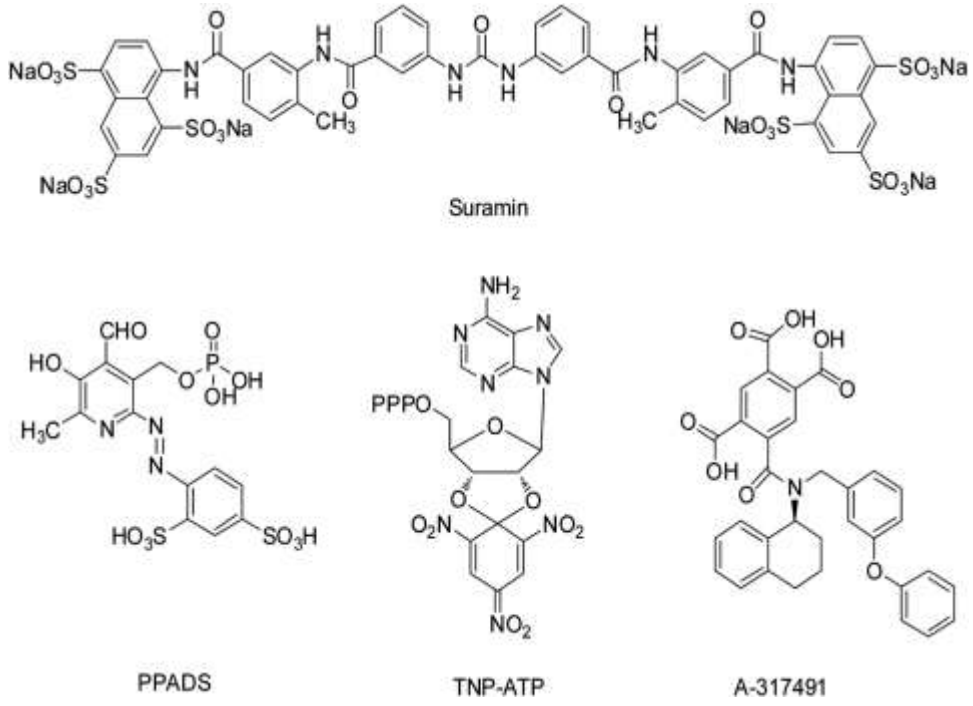
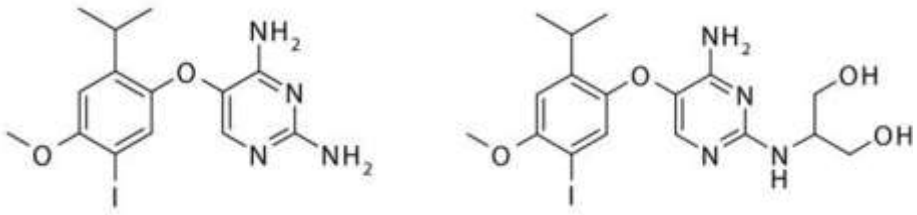


Figure 4.7: Non-drug-like purinergic antagonist

Later on, from 2004 onwards, Roche Pharmaceuticals patented several novel classes of compounds with potential drug-like features, and new scaffolds of optimized competitive antagonists for $P2X_3$ receptors were identified[132] (Figure 4.8).



AF-353 (RO-4) **AF-906 (RO-51)**

Figure 4.8: Potent orally bioavailable $P2X_3$ and $P2X_{2/3}$ antagonists

Likewise, other companies, including Evotec AG, Astra Zeneca, Merck and Shionogi, identified potentially developable molecular scaffold.

So far, despite the efforts of the pharmaceutical companies, only one candidate $P2X_3$ antagonist is progressing into human studies: the orally available aryloxy-pirimidine-diamine molecule AF-219, developed by Roche[133]. Its inhibitory potency (IC_{50}) has been reported about 30 nM versus the recombinant $hP2X_3$ and 100-250 nM at the $hP2X_{2/3}$ receptor. To date, AF-219 has being tested for four different therapeutic indications, which are osteoarthritic joint pain, BPS/interstitial cystitis, chronic cough and asthmatic syndrome. For the first three indications, the Phase II of the clinical trials has already been completed, while the studies on asthma have completed the Phase I.

The results of the clinical trials on the chronic cough were disclosed in the European Respiratory Society Annual Congress in September 2013, where it was

shown a small pilot study in patients with considerable cough burden. 24 patients were treated for 2 weeks, with a single high daily dose of AF-219 compared with placebo. The molecule markedly reduced cough frequency in daytime of 75%. This was the first promising result of clinical use for a purinergic antagonist[133]. These findings also augur well for other symptoms of airways diseases, that are implicate by afferent hyperaxcitability.

Competition binding and intracellular calcium flux experiments showed that Both AF-353 and AF-219 inhibit activation by ATP in a non-competitive manner. These studies clearly suggested the presence of one or more allosteric binding sites in the trimeric structure of P2X3 receptors, which might be the principal targets of the existing and the future ATP competitors. Therefore, the identification of a potential allosteric binding site is an essential and compelling requisite to support the rational development of purinergic inhibitors.

4.2 Binding site identification techniques

4.2.1 Introduction

The identification and the full characterization of the protein binding sites is a demanding problem in the field of computer aided drug design. The complete definition of the potential pockets within a given protein structure is a crucial challenge in drug discovery, and its achievement can lead to a better understanding of the molecular interactions, and to an important improvement in the quality of the rational drug design.

In the last years, the classic key-lock model for the ligand-receptor interaction evolved along with the discovery of the dynamic aspects of proteins, which consist in conformational changes that range from small side-chains adjustments to large domain motions. Furthermore, protein-ligand interactions rely not only on the steric complementarity, but also on the physicochemical compatibility. The approaches to search binding sites should be able to deal with these aspects, and to overcome the difficulty to find procedures that can be used universally for all proteins.

On this ground, we assisted to the development of different new computational methods aimed at obtaining maps of the possible interactions and identifying the features of the binding sites. The structure based methods are one of the current solutions: they include approaches based on the geometrical space, on the energy space and on the analysis of the interacting key residues[134]. The first approach detects the possible binding site by searching for void volumes, while the energy based methods calculates the interaction energies with some representative probes

to identify the binding pocket. In both cases, the analysis is supported by the genomic characterization of the protein.

Here, we present three different pieces of software which implement diverse approaches to face the problem:

- FPocket: a classic geometric approach;
- SPILLO-PBSS: an innovative structure-based approach;
- PELE: a Monte Carlo-based technique.

4.2.2 FPocket

Fpocket is an open source software which exploits a classic geometric approach, based on the individuation of the void volumes within the protein[135]. Through this method, a clear map of the protein cavities and their structural characterization can be obtained. The identified pockets are then ranked according to their reliability in terms of capacity to accommodate small compounds.

The FPocket approach is based upon the use of alpha spheres (Figure 4.9). An alpha sphere is defined as a sphere which contacts four atoms on its surface and contains no internal atoms. In proteins, these spheres can be large, when considering the surface of the whole structure, small, for the core of the protein, and medium, in correspondence with cavities and clefts. Therefore, a given pocket can be located by calculating the position and the corresponding radius of the medium alpha spheres[135].
The algorithm's workflow can be summarized in three steps: first the whole ensemble of alpha spheres is calculated on the protein structure, then the spheres

are clustered together, finally the clustered pockets are ranked according to their ability to accept small molecules, so selecting the most interesting pockets.

The ranking criteria of pockets do not reflect druggability, since the small molecule might be a drug or a sugar, a cofactor or a coactivator[135]. Moreover, the search is not based on a specific class of compounds and the method reveals to be not suitable for describing ligand selectivity. For these reasons, the algorithm often returns a large number of potential pockets, as in our case, and the selection of the most reliable binding sites can be problematic and too demanding.

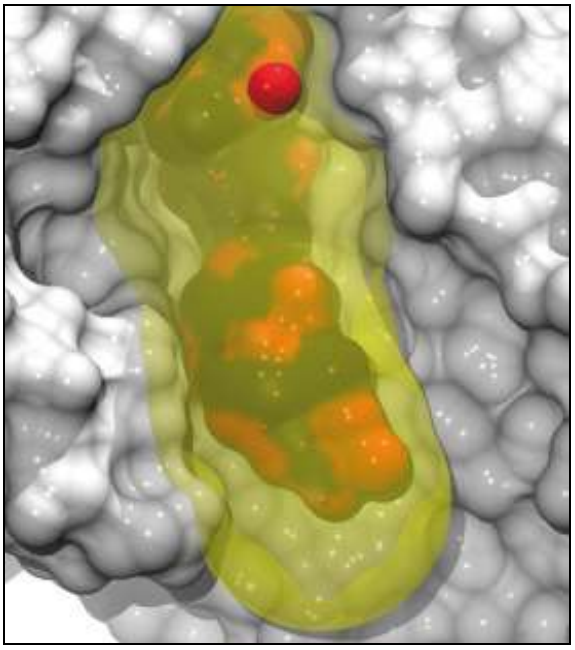


Figure 4.9: Example of pocket output
(Reproduced from Le Guilloux *et al.*[135])

4.2.3 SPILLO-PBSS: protein binding site searcher

According to the concept of protein druggability, all the protein binding sites, despite their structural differences, share key properties that allow them to form stable complexes with the respective ligands. On this basis, different approaches were developed to identify potential binding sites within the protein 3D-structures.
An important limitation of the currently available structural approaches for the binding site prediction is that they are not able to find the right binding sites on protein 3D-structures unless they are already in a suitable conformation for the binding event: in fact, if the proteins are in a distorted conformational state, involving, for example, a closed conformation of the binding site, the software fail to identify the right binding sites.

SPILLO-Potential Binding Site Searcher[136] (SPILLO PBSS) is an innovative software which overcomes the limitations of the other tools for the binding site prediction, detecting potential binding sites within protein 3D-structures, even when these are highly distorted compared to a suitable binding conformation.
SPILLO-PBSS is based on the assumption that 'similar binding site can bind the same ligand in a similar way', and it uses a well-known reference binding site (RBS) to search the target proteins for potential binding sites (PBSs) similar to the RBS (Figure 4.10).
In detail, the RBS is obtained from the three-dimensional structure of a given protein bound to its ligand, by an analysis of the complex geometrical structure and physicochemical properties. In particular, the residues included in the RBS are those surrounding the ligand (i.e. having at least one atom within a predefined distance from the ligand), and for each of them a weight (a coefficient) representing the specific relevance in stabilizing the interaction (classical non-bonded electrostatic and Lennard-Jones potential energy) with the ligand is calculated. The

weights can be also assigned manually by the user when given residues are already known to have relevant roles in the complex stability.

Once obtained the RBS, the software creates a reduced representation of the system, in which all the residues are described by spheres ad vectors. This simplified model, together with a specially designed geometric tolerance, allows SPILLO-PBSS to implicitly take into account protein flexibility without recurring to explicit simulations and to easily detect the binding sites even if they underwent to conformational distortions or similar phenomenon.

The next step is the search of PBSs throughout the whole target proteins (not only on the surface), by a comparison of the RBS to the different regions in the proteins. To this aim, the target protein is put inside a virtual grid with cubic cells and the RBS is iteratively and systematically translated to all grid nodes, and for each of them it is rotated around the three Cartesian axes. For each position of the template, the residues of the examined protein are individuated and analysed to see if they can correspond to those present in the RBS. The scoring function takes also into account binding sites with different amino acid composition with respect to the RBS, but with similar physicochemical properties. In this way, a set of potentially relevant binding sites are collected on a list in decreasing order according to their similarity to the program input (to the RBS).

As already mentioned, the geometric tolerance plays a fundamental role in the PBSs individuation, because the software allows that the residues on the target proteins can be disposed in a different way compared to the reference. Furthermore the tolerance is also expected to compensate the inaccuracies introduced during the simplification steps.

Finally, the scoring function of the identified PBSs is expressed as a percentage, so that the maximum value of 100% correspond to the ideal case of complete agreement between RBS and PBS.

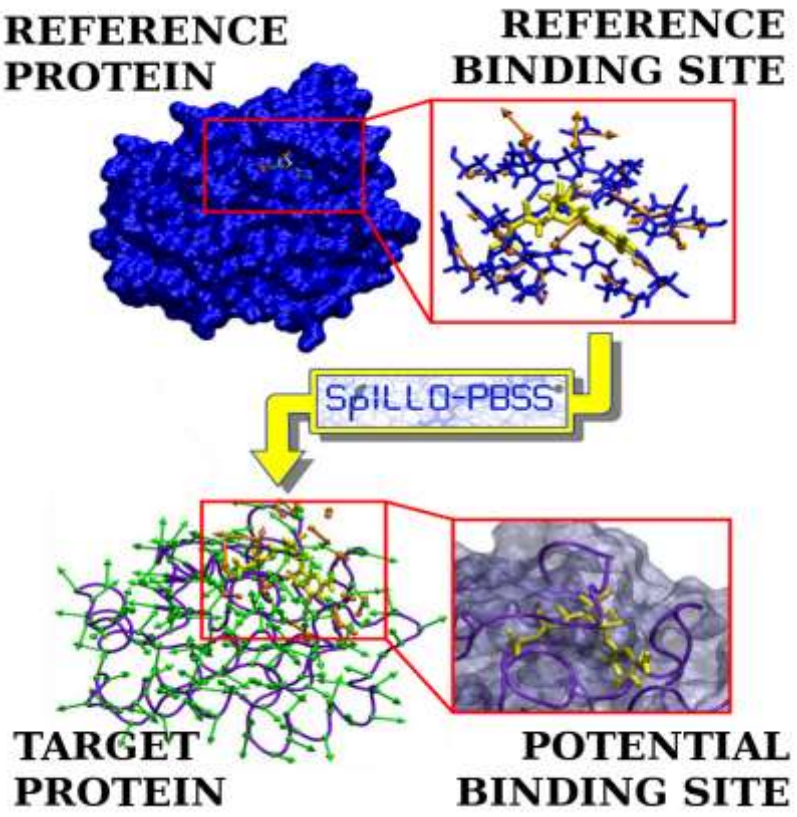


Figure 4.10: SPILLO-PBSS workflow
(Reproduced from Di Domizio *et al.*[136])

4.2.4 PELE: Protein Energy Landscape Exploration

PELE (Protein Energy Landscape Exploration)[137] is a Monte Carlo based technique that combines protein structure prediction algorithms to explore all-atom energy landscapes. It was first developed to map ligand entrance and exit from proteins binding site, but its application was then expanded to other similar issues, such as ligand diffusion in the binding site, induced fit docking, evaluation of ligand-protein binding energy, and overall protein dynamics. It is always characterised by a reduced computational cost if compared to canonical MD simulations.

Pele performs an unconstrained ligand exploration or a binding site search. In the present study, it was indeed exploited to identify binding site arrangements, by using a method that considers the conformational flexibility of the protein.

The software workflow (Figure 4.11) is organized in consecutive steps which are iteratively executed to explore the protein landscape. First, after an initial estimation of the starting energy, the system is subjected to a local perturbation, a process that can involve the protein backbone, the side chains, and the ligand. The second step involves the conformation optimization of the side chains, by using a specific algorithm and a rotamer library. Finally, the perturbed regions are minimized to generate a rearrangement of the backbone in response to the initial system alteration. The overall process was designed on the assumption that first the protein side chains act as sensors that trigger the protein and ligand perturbations, and then the backbone atoms follow the side chains perturbations.

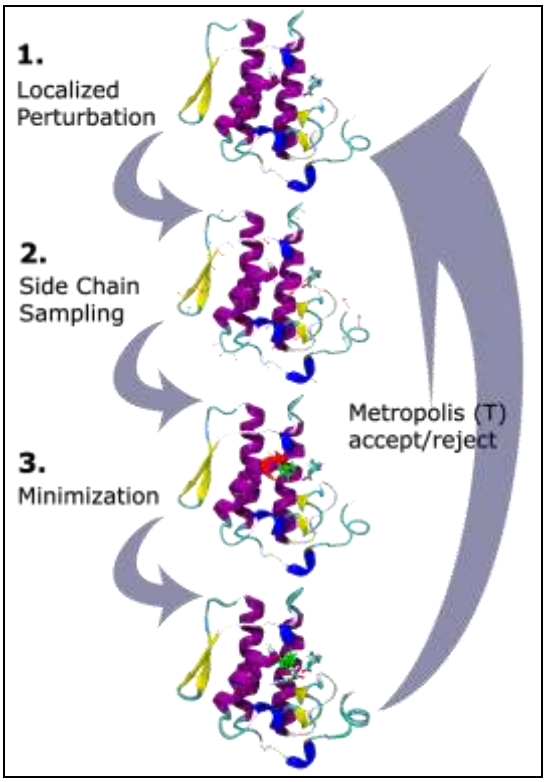


Figure 4.11: PELE workflow
(Reproduced from Madadakar-Sobhani *et al*[138].)

These three steps compose a move which can be accepted or rejected, based on a Monte Carlo algorithm, which provides an efficient method for sampling the conformational space.

In detail, the software input consists in a file containing the protein and the ligand located to a certain distance from the macromolecule surface[138]. Among the calculation parameters that can be set, there are the number of CPUs used, which corresponds to the number of different trajectories created, the number of Monte Carlo steps and the wall clock time limits (maximum time provided for the simulation).

Once the simulation is completed, the software returns a trajectory file with different parameters, calculated at each step of the simulation:

- SASA: percentage that show the exposed surface area of the ligand;
- Binding energy: computed between the ligand and protein;
- Native RMSD: root mean square deviation calculated between the protein and a given native protein;
- Ligand RMSD: root mean square deviation of the ligand in respect to the starting position;
- Point/atoms distance: distance calculated between a ligand atom and a point/ atom of the protein

Moreover, these data are visualized in a table and in a plot, which are continuously updated during the simulation (see for an example Figure 4.12).

The trajectory file shows the movement of the ligand around and inside the protein structure, and the most stable complexes can be selected looking at the simulation frames reporting the lowest binding energy values. By using this approach, we were able to identify an interesting potential binding site on the purinergic receptors structure (see Chapter 4.3.10)

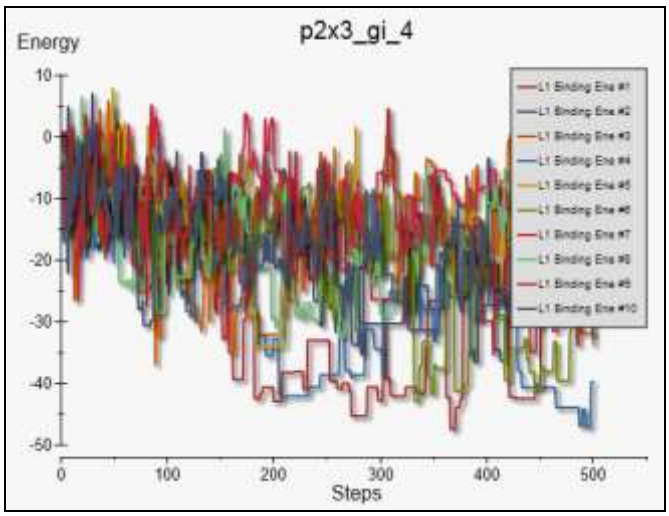


Figure 4.12: Pele output file: Binding energy of ten trajectories around P2X3 receptor.

4.3 Modelling of human $P2X_3$ receptor allosteric inhibition.

4.3.1 Setting the scene

The field of purinergic signalling has grown constantly over the last forty years, so that the scientific community is now well aware of many characterizing features of this system. Both academic laboratories and pharmaceutical companies started projects and experimental trials to better explore the purinergic physiological path, in order to exploit the druggability potential of the purinergic receptors. $P2X_3$ receptors, in particular, are identified as pain-related membrane proteins, and are considered as new targets for the treatment of painful conditions in acute and chronical diseases.
In the Chapter 4.1 we reported the main recent efforts carried out by Roche and Afferent Pharmaceuticals in order to identify and test $P2X_3$ antagonists, which can represent a valid alternative to the painkillers already present in the pharmaceutical market.
For the time being, many aspects of the purinergic signalling remain unclear. One of the most challenging is the complete definition of the actual molecular mechanism of action of $P2X_3$ antagonists. Indeed, although the not-competitive behaviour of ATP inhibitors was demonstrated by functional *in vitro* studies, the allosteric binding site and the mode of action of the antagonists are not yet clarified.

The aim of the present study was the modelling of the inhibition activity of purinergic antagonists through the identification and validation of a potential allosteric binding site on the trimeric structure of $P2X_3$ receptors. The project

involved the homology modelling of the trimer structure, the identification of the potential allosteric binding sites by different approaches, and the validation of the defined pockets by a virtual screening approach coupled to the enrichment factor analysis and by a QSAR analysis.

4.3.2 Modelling and optimization of hP2X$_3$ receptor 3D-structures

4.3.2.1 Set-up of the protein templates

The first step of the project involved the homology modelling of the human P2X$_3$ receptor, both in the apo and in the holo forms, by using the resolved zebrafish P2X$_4$ receptor structures as the templates. In particular, the crystal structure of the apo trimer, namely the PDB Id 4DW0[124] structure, was used as the template for the closed model of the ion channel, while the resolved structure of the holo trimer, namely the PDB Id 4DW1[124] structure, served as the template for the ATP bound, open model.

To this end, the three-dimensional structures of the templates were assembled starting from the resolved monomers included in the crystals, by exploiting the reported transformation matrices. In detail, the three sets of transformation coordinates were applied to the single monomer and the three differently oriented subunits were assembled to build the trimeric structures. To optimize the inter-subunit contacts in the so obtained structures and to avoid high-energy arrangements, an energy minimization was performed by using NAMD[98]. The atomic partial charges were computed by the Gasteiger-Marsili method, while the potential and the atom types were assigned using the CHARMM 22 force field. In order to preserve the resolved structure of the templates, the protein backbone atoms were kept fixed. For the holo trimer, the three ATP molecules included in the binding sites were free to move.

4.3.2.2 Modelling methods

Both the models were obtained by using the modelling software Modeller 9.10[139]and applying the same modelling protocol, as follow.

The software firstly calculated the alignment of the template monomer's sequences with the monomer's sequence of the human $P2X_3$ receptor. The alignment of the subunit 4DW0-A is reported for example in Figure 4.13.

```
aln.pos         10        20        30        40        50        60
4DW0A     ----------------------GTLNRFTQALVIAYVIGYVFVYNKGYQDTDTVL-SSVTTKVKGIA
P2X3      MNCISDFFTYETTKSVVVKSWTIGIINRVVQLLIISYFVGWVFLHEKAYQVRDTAIESSVVTKVKGSG
 _consrvd                       *  **  * * * *  * **   * **  **   *** *****

 _aln.p   70        80        90       100       110       120       130
4DW0A     LTKTSELGERIWDVADYIIPPQEDGSFFVLTNMIITTNQTQSKCAENPTPASTCTSHRDCKRGFNDAR
P2X3      LYA-----NRVMDVSDYVTPPQGTSVFVIITKMIVTENQMQGFCPESEE-KYRCVSDSQCG--PERLP
 _consrvd *        *  ** **  ***    *   * ** * ** *  * *       * *   *

 _aln.pos  140       150       160       170       180       190       200
4DW0A     GDGVRTGRCVSYSASVKTCEVLSWCPLEKIVDPPNPPLLADAERFTVLIKNNIRYPKFNFNKRNILPN
P2X3      GGGILTGRCVNYSSVLRTCEIQGWCPTE-V-DTVETPIMMEAENFTIFIKNSIRFPLFNFEKGNLLPN
 _consrvd * *  ***** **    ***   *** *   *    *    ** **  *** ** * *** * * ***

 _aln.pos    210       220       230       240       250       260       270
4DW0A     INSSYLTHCVFSRKTDPDCPIFRLGDIVGEAEEDFQIMAVRGGVMGVQIRWDCDLDMPQSWCVPRYTF
P2X3      LTARDMKTCRFHPDKDPFCPILRVGDVVKFAGQDFAKLARTGGVLGIKIGWVCDLDKAWDQCIPKYSF
 _consrvd         * *    ** *** * ** *  *  **   *  *** *  * * ****     * * * *

 _aln.pos      280       290       300       310       320       330       340
4DW0A     RRLDNKDPDNNVAPGYNFRFAKYYKNSDGTETRTLIKGYGIRFDVMVFGQAGKFNIIPTLLNIGAGLA
P2X3      TRLDSVSEKSSVSPGYNFRFAKYYKMENGSEYRTLLKAFGIRFDVLVYGNAGKFNIIPTIISSVAAFT
 _consrvd  ***       * ************   * * *** *  ****** * * *********     *

 _aln.pos        350       360       370       380       390       400
4DW0A     LLGLVNVICDWIVL-------------------------------------------------------
P2X3      SVGVGTVLCDIILLNFLKGADQYKAKKFEEVNETTLKIAALTNPVYPSDQTTAEKQSTDSGAFSIGH
 _consrvd   *   * ** * *
```

Figure 4.13: Modeller alignment of hP2X3 and zfP2X4 (4DW0A) monomer's sequences. The symbol (*) indicates the conserved residues.

Then, two sets of five different models for the human monomers were generated, one for the apo structure and another for the holo one. The selection of the best models was made based upon the scores provided by Modeller 9.10 (namely DOPE

and GA341) and the Ramachandran Plots, as reported in Figure 4.14. In both cases, the best selected model was the model n. 1, which reports the highest values for all the considered scores.

Apo Model	DOPE score	GA341 score	RP favoured residues (%)	RP highly fav. residues (%)
1	**-33507.25**	**1.000**	**89.92**	**65.74**
2	-33021.60	1.000	89.92	65.24
3	-32994.92	1.000	89.18	65.20
4	-33296.24	1.000	89.69	65.68
5	-32467.81	1.000	89.54	64.74

Holo Model	DOPE score	GA341 score	RP favoured residues (%)	RP highly fav. residues (%)
1	**-32230.53**	**1.000**	**91.18**	**64.74**
2	-32189.07	1.000	91.18	64.24
3	-31941.93	1.000	89.67	64.01
4	-31928.46	1.000	90.68	64.50
5	-31812.09	1.000	88.66	64.74

Figure 4.14: Scores values of the sets of five models for apo and holo monomers.

Finally, the selected monomer structures were assembled to generate the trimers, using the previously prepared trimeric templates by superimposing each modelled monomer to the corresponding monomer template. During this procedure, the unsatisfactorily superimposed segments were manually modified to be coherent to the template folding, and consequently optimized through a minimization performed by NAMD (10000 iterations, dielectric constant = 1). Since a perfect superimposition was clearly impossible, the matching atoms guiding the

overlapping were differently weighted and the priority was given to the regions close to the ATP binding site. Then, the trimeric structures of the human apo and holo $P2X_3$ receptors were generated by applying to the same transformation matrix of the corresponding templates.

The obtained trimeric models were optimized by carrying out a minimization (fixed backbone atoms, 30 000 steps, free ATP molecules), followed by a 10 ns MD study performed by NAMD. In detail, firstly, the trimers were embedded in a box of water (85 Å x 85 Å x 85 Å) containing 16250 solvent molecules. The systems were then minimized to optimize the relative position of the solvent molecules and then were subjected to MD runs with the following characteristics: (a) periodic boundary conditions (95 Å x 95 Å x 95 Å) were applied to stabilize the simulation space; (b) Newton's equation was integrated using the r-RESPA method (every 4 fs for long-range electrostatic forces, 2 fs for short-range non bonded forces, and 1 fs for bonded forces); (c) the long-range electrostatic potential was computed by the Particle Mesh Ewald summation method (80 × 80 × 80 grid points) (d) the temperature was maintained at 300 ± 10 K by Langevin's algorithm; (e) Lennard-Jones (L-J) interactions were calculated with a cut-off of 10 Å and the pair list was updated every 20 iterations; (e) a frame was memorized every 10 ps, thus generating 1000 frames; and (f) no constraints were imposed to the systems. The simulations were carried out in two phases: an initial period of heating from 0 K to 300 K over 100000 iterations (100 ps, i.e. 0.3 K/ps) and the monitored phase of 10 ns.

The lowest energy structures was then extracted by the MD runs and finally optimized after removing water molecules and neutralizing ions.

4.3.2.3 Modelling results

As a preamble, Figure 4.15 shows the optimized monomeric and trimeric structures of the templates for the apo and holo state, obtained as before described.

Figure 4.16 depicts a comparison between the superimposition of a single modelled monomer before and after its optimization. The pictures underline the importance of the refinement performed on the models generated by standard homology modelling technique, in order to obtain truly reliable final models.
The trimeric structures underwent a structural validation performed by PROCHECK[140], which calculates all structural parameters of a protein model and gives indications of its reliability. In particular, the attention was focused on: Ramachandran plot, bond lengths, angles distortions and planar group distortions. The obtained trimeric models are depicted in Figure 4.17 .

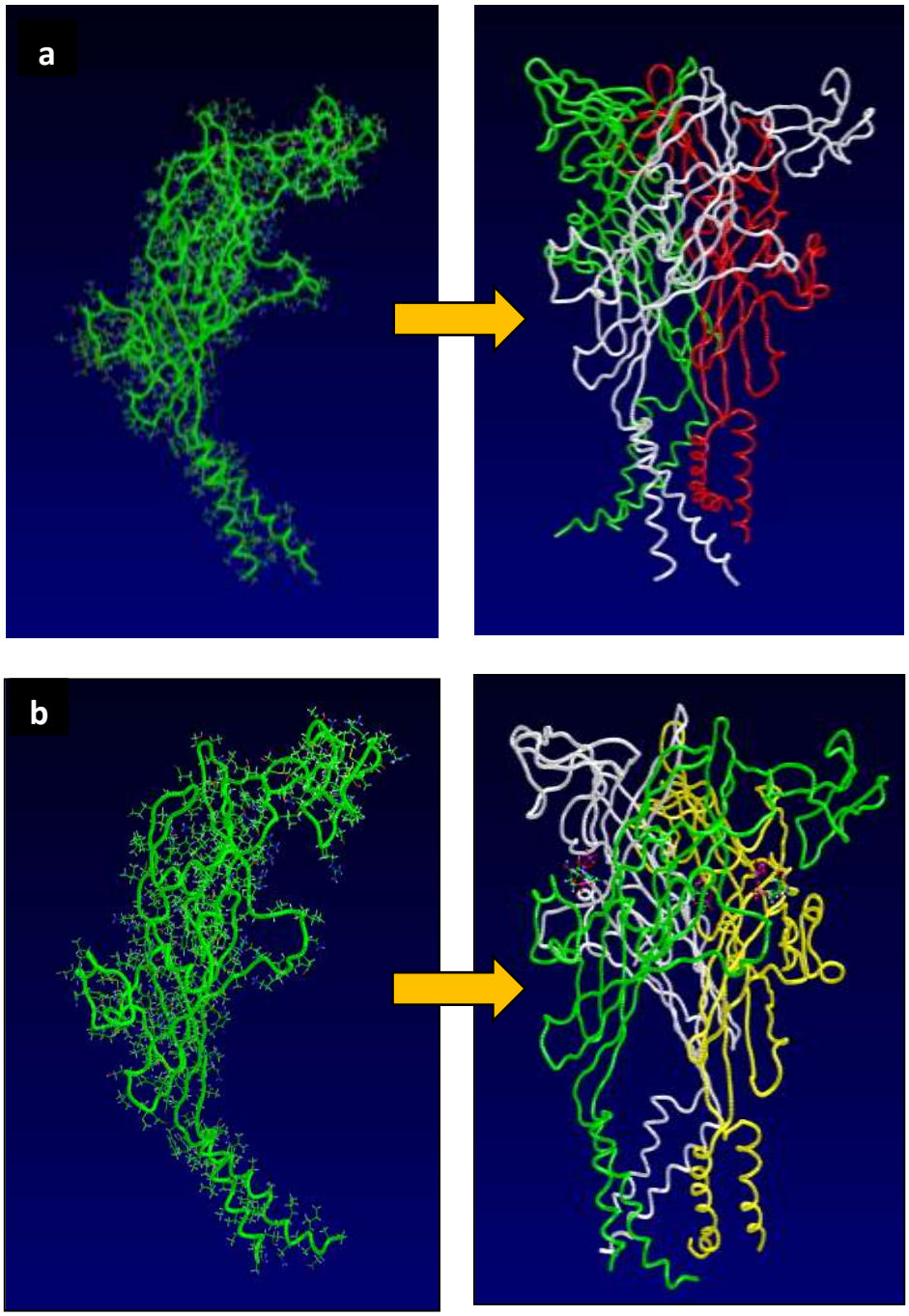


Figure 4.15: Optimized trimer templates. (a) Apo structure. **(b)** Holo structure. On the left the monomers, on the right the trimers. In the holo trimer, are depicted also the three ATP molecules included in the binding sites.

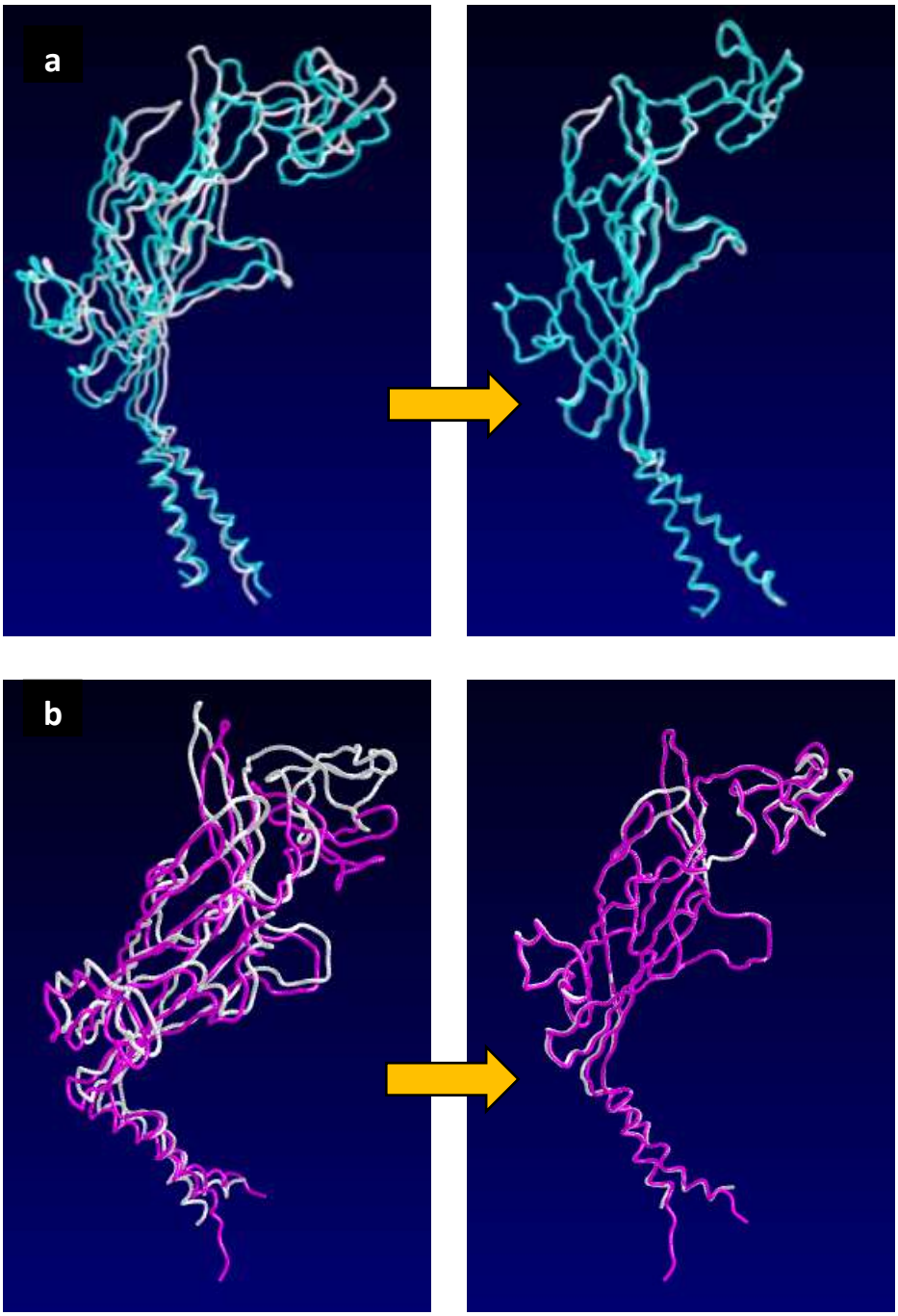


Figure 4.16: Optimization of the superimposition for the monomer structure. The crystals are coloured and the models are while. On the left is reported the superimposition before the optimization, on the right the result after the optimization. (a) apo trimer, (b) holo trimer

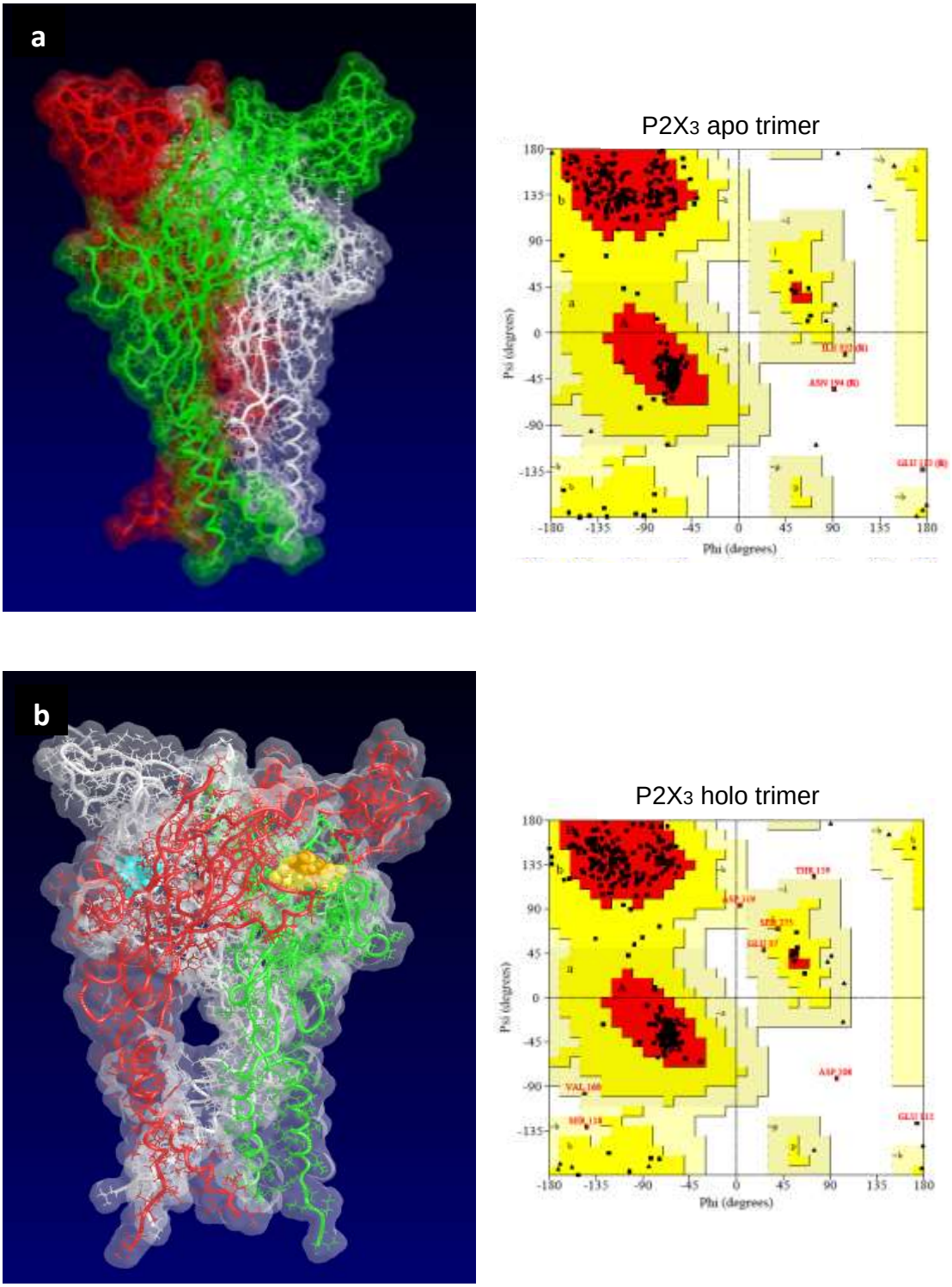


Figure 4.17: $P2X_3$ models and relative Ramachandran plots. (a) apo trimer, (b) holo trimer

4.3.3 Set-up of the selected purinergic inhibitors

The second step of the project involved the collection of allosteric ligands for the $P2X_3$ purinergic receptor by focusing the attention on allosteric inhibitors with a value of IC_{50} in the range of nanomolar activity. In detail, we collected molecules belonging to two structural families of purinergic ligands:

- the diaminopyrimidines (DAPs)[132,141,142,143], which are unselective for $P2X_3$ being also active on $P2X_{2/3}$. These compounds came from a series of Roche molecules developed as structural analogues of Trimethoprim, a well-known inhibitor of the bacterial dihydrofolate reductase (DHFR).

- the arylamide derivatives (SAAs)[144], which are selective for $P2X_3$ receptor (pIC_{50} for $P2X_{2/3}$ <5). Their parent compound has been identified thanks to a high throughput screening campaign using the rat $P2X_3$ recombinant receptor at Roche.

The dataset of purinergic ligands consisted in 24 compounds belonging to the DAPs family (deriving from 21 compounds by generating all possible stereoisomers), and 6 compounds belonging to the SAAs family. The complete list of these compounds is reported in Appendix 2, with the relative pIC_{50} values experimentally obtained using the FLIPR method on recombinant $P2X_3$.

To be compatible with physiological condition (pH = 7.4), SAA molecules were simulated in their ionization state by protonating the tertiary ammine group of the piperazinic ring. On the other hand, DAP molecules can present the piperazine ring either protonated or not, and both protonation states of these compounds were taken in consideration.

Each ligand included in the study underwent the same protocol of preparation, based on a MonteCarlo procedure which generated 1000 minimized geometries by randomly rotating the flexible torsions. The so generated lowest energy structure underwent the following docking simulations.

4.3.4 Generation of decoy datasets

The first strategy adopted to assess the reliability of a potential allosteric binding pocket was based on a virtual screening study coupled with an enrichment factor analysis. Therefore, to perform such analysis, two different decoy databases were built, both including 30 active molecules, namely the purinergic inhibitors previously collected, and 2970 presumably inactive compounds.

Both decoy databases are characterized by a random percentage of active molecules equal to 1 (this means that choosing randomly 100 molecules, 1 active molecule is expected to be found). The inactive molecules were collected within the Directory of Useful Decoys (DUD)[145], taking in consideration the size and the charge parameters. In detail, when considering that DAP derivatives can be protonated or not, the two decoy sets had the following characteristics:

- DECOY_1: The DAP molecules were considered in the protonated state, according to the condition of Trimethoprim when it is bound to its target, namely the Diyhdropholate reductase (see Figure 4.18). Therefore, all the active molecules have one positive charge and, consequently, the whole set of inactive molecules was collected choosing compounds with one positive charge, searching from a database of inactive compounds on acetylcholinesterase.
- DECOY_2: The DAP molecules were considered in the neutral state, therefore also a proportional part of the inactive molecules were chosen with a neutral charge, searching from a database of inactive compounds on Hivrt (HIV-retrotranscriptase) and on Inha (enoyl alyl carrier protein).

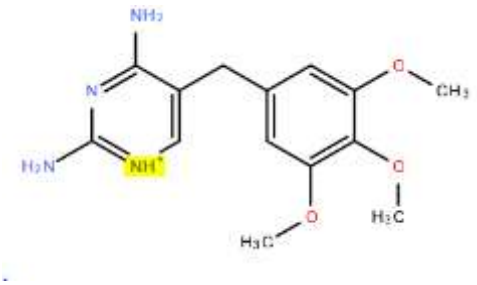


Figure 4.18: Trimetoprim protonation state

4.3.4.1 Reliability of the collected datasets

The potentially unbiased composition of the two decoy datasets was assessed by performing a preliminary virtual screening study using a different biological target, namely the recently resolved muscarinic AChM2 receptor structure (PDB Id 3UON), also objected of modelling studies which are reported in Chapter 5. Docking simulations involved both datasets and, as mentioned above, were performed by using PLANTS and focusing the search in a 10 Å radius sphere around the key residue Asp103.

In all performed simulations and for both datasets no significant enrichment factors were obtained, thus suggesting that the decoy sets do not have biases which can weaken the following analysis. For example, Figure 4.19 shows the results depicted by the above mentioned bar plot for the PLP95 score on dataset 1

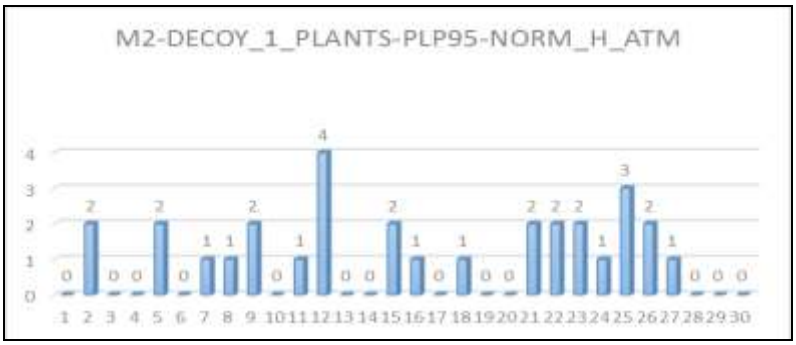


Figure 4.19: Bar plot of enrichment analysis performed on M2 receptor. The docking study uses the DECOY set number 1 and the scoring function Plp95, normalized on the number of heavy atoms.

4.3.5 Docking simulations strategies

Apart from the blind docking analysis, which was performed by applying specific conditions as detailed below, all docking simulations carried out for the here described virtual screening campaigns were performed by using the PLANTS[146] software, which calculates reliable poses by a colony optimization algorithm. Three score functions were taken in consideration, namely CHEMPLP, PLP and PLP95[147], by focusing the search within a 10 Å radius sphere around key residues characterizing the explored binding pocket. For each ligand, one pose was generated with speed equal to 1. The ligand was considered as flexible during the simulations, while the protein atoms were kept fixed. Each post-docking optimization procedure was applied to the so computed poses.

For the enrichment studies, the obtained results were analysed by subdividing the entire ranking in 30 clusters composed of 100 compounds, and graphically reporting how many active molecules are found in each cluster. Such a new approach to evaluate the reliability of a virtual screening campaign will be better detailed in the next chapter.

4.3.6 Strategies for the allosteric binding site search

Since in literature there are no information about the position of the putative allosteric pocket on the human $P2X_3$ receptor, the identification of some potential binding sites was a challenging and demanding issue. Different approaches were applied, based on different conceptual strategy; some of them proved successful, suggesting the presence of more than one binding site, others revealed to be failures.

In the results' subchapters, we will report the outcome of four strategies which are indeed representative of three major approaches for binding site search, based on docking calculations, on pocket detection, and on ligand pathway simulations by MonteCarlo procedure, respectively.

- Blind docking
- Fpocket[135]
- SPILLO-PBSS[136]
- PELE[137]

4.3.7 Blind docking

The first strategy applied to identify e potential allosteric binding site on the human $P2X_3$ receptor exploited the blind docking approach. This method consists in performing a molecular docking study by considering a single ligand on the entire receptor structure, so generating multiple poses (form 100 to 500). The resulting potential pockets are selected by considering the most populated ones among the docking poses.

The blind docking was performed by PLANTS[146] as integrated in the software VEGA ZZ on both the apo and the holo $P2X_3$ models. Three score functions were taken in consideration, namely CHEMPLP, PLP and PLP95. To perform this docking analysis, the receptor structures was subdivided into three overlapped regions, because the software does not allow a docking sphere with a maximum radius greater than 60 Å to be considered, while the receptor would require a sphere of about 108 Å.

500 poses were generated in five different studies, carried out by simulating three representative compounds: Trimethoprim, in both charged and neutral form, the most active DAP compound, namely the one with id: 13t_20 (see Appendix 2), in both charged and neutral form, the most active SAA compound, namely the one with id: 16_4 (see Appendix 2), only in the charged form.

For each $P2X_3$ model, the poses obtained with all the docking studies were merged on the same structure, to better identify the most populated pockets (Figure 4.20). 13 binding sites, both on the apo and the holo models were tested by docking simulations and enrichments studies by using both decoy sets as previously collected.

Disappointingly, none of the examined pockets were able to discriminate between the active molecules and the other ones included in the databases.

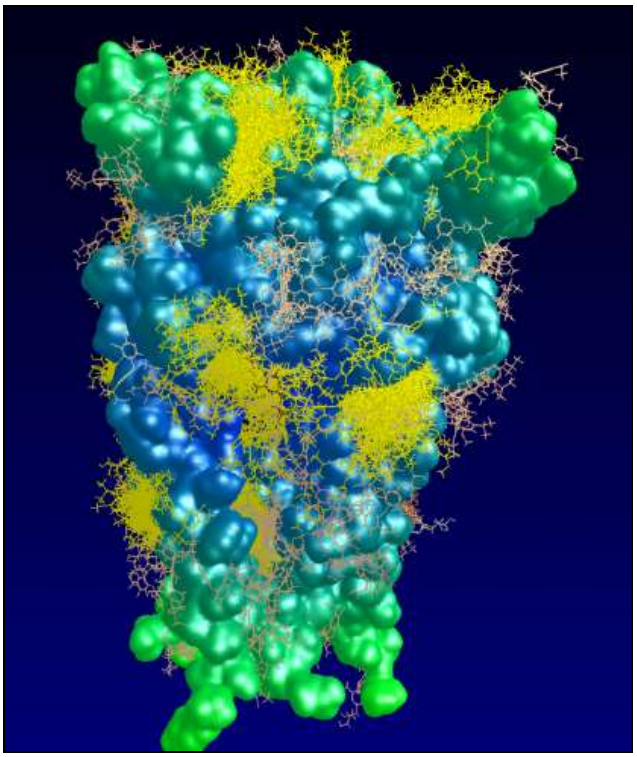


Figure 4.20: Example of cluster of poses derived from one of the blind docking studies.

4.3.8 Fpocket

The first applied strategy involved the search of reliable binding pockets based on a simple geometrical analysis as performed by Fpocket. This open source package for pocket detection can individuate voids and cavities in a macromolecule and, for each of the identified pocket, the software provided also a detailed physicochemical characterization and a rank according to their reliability.

The weakness of the method consists in the lacking of a criterion which considers the specific ligand features, resulting in the identification of a huge number of potential pockets. In detail, among them, the first ranked was the ATP binding pocket, followed by a long list of potentially allosteric pockets, all interestingly positioned at the interface between two or three monomers.

The number of identified pockets was too high to allow their exhaustive examination by enrichment factor studies. However, the results can be exploited as an additional confirmation for the pockets found by other approaches. In particular, Fpocket ranked in second position the same pocket identified by the SPILLO-PBSS approach, both in the apo and in the holo model, increasing the reliability of the prediction (see below, Figure 4.23).

4.3.9 SPILLO-PBSS

SPILLO-PBSS is software able to detect potential binding sites within protein 3D-structure, even when they are highly distorted and the binding site does not assume a suitable conformation. Differently from Fpocket package, SPILLO-PBSS searches pockets based on a specific ligand and uses, as a reference, a complex of that ligand with another protein (the so-called "reference protein").

Since none of the $P2X_3$ antagonists was co-crystallized with any protein, the pivotal intuition that allows us to use the software was the opportunity to exploit the structurally analogue Trimethoprim as a surrogate.

In detail, the DHFR inhibitor was co-crystallized with two classes of proteins:

- Dihydropholate Reductase: PDB entries 2W3A, 2W3V, 2W9S, 3FL9, 3NO0H, 3TQ8,4G8Z and 4KM2;
- Pteridine reductase 1: PDB entry 2BFM.

4.3.9.1 Dihydropholate Reductase

Concerning the Dihydropholate reductases, the resolved structures PDB id 4KM2 and 3NOH were selected according to the best resolution and the absence of any other co-crystallized molecule in the trimethoprim (TOP) binding site.

Both the Dihydropholate reductases resolved structures underwent the same preparing steps applied to the zebrafish templates (see chapter 4.3.2.1). The reference binding sites (RBS) were generated on both reference structures, considering the ligand in its charged state.

SPILLO-PBSS provided as output a list of the potential binding sites (PBS) of Trimethoprim in the apo as well as in the holo model, ranked according to the similarity to the RBS. The analysis of the accessibility by the ligand of the best

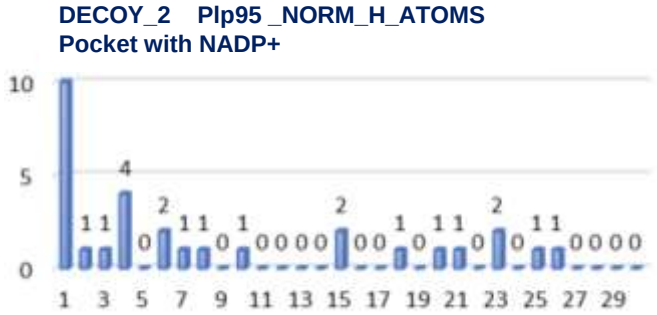
ranked identified pockets leads to discard all of them. The potential binding sites proved unrealistic since they were symmetrically located in the central body of each monomer, within structurally rigid and inaccessible regions.

4.3.9.2 Pteridine reductase

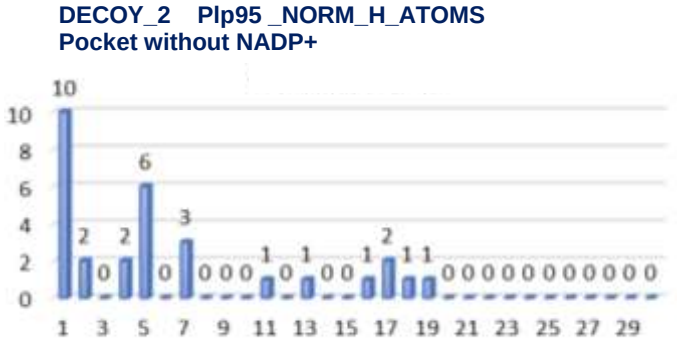
The resolved structure of the Pteridine reductase is a multimeric model, and a single monomer was taken into consideration. The monomer included a molecule of Trimethoprim and a molecule of the $NADP^+$, which is considered the natural cofactor of the enzyme. The 3D-structure was optimized by using the same protocol already utilized for the Dihydropholate reductases.

4.3.9.2.1 Identification and optimization of the pocket

As a preliminary study, an enrichment factor analysis was performed directly on the TOP binding site of the Pteridine reductase protein. To verify the importance of the cofactor in the pocket, two docking simulations were carried out, one including NADP+ and another without it. Both the decoy sets were used. Both docking simulations with and without cofactor gave encouraging results, the best results being obtained by using the score PLP95 normalized on the heavy atoms and the dataset 2. As detailed in Figure 4.21, these simulations afforded in the TOP 1% enrichment factors of 30.00 and 10.00 for the pockets with and without the $NADP^+$, respectively.



Top %	Mol	Att	EF
1.00	30	9	30.00
2.00	60	10	16.67
5.00	150	11	7.33
10.00	300	12	4.00
20.00	600	18	3.00



Top %	Mol	Att	EF
1.00	30	3	10.00
2.00	60	8	13.33
5.00	150	11	7.33
10.00	300	12	4.00
20.00	600	20	3.33

Figure 4.21: Enrichment factor analysis of the TOP binding site on Pteridine reductase.

Before the generation of the RBS, the Trimethoprim pocket on Pteridine reductase was optimized.

First, since the better results were obtained with the pocket containing NADP+, but considering that SPILLO-PBSS can only handle amino acidic residues and cannot parametrize the cofactor, its stabilizing contribution was approximate by replacing it with an amino acid. More in detail, complex between Trimethoprim and $NADP^+$ was analysed, revealing two main kinds of interaction:

- a π–π interaction between the ethero-aromatic cycle of the NADP+ and the diamminopyrimidinic moiety of Trimethoprim;
- potential hydrogen bonds.

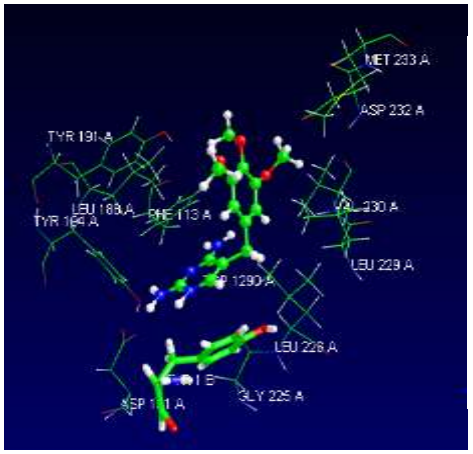
On these bases, tyrosine appeared to be the most suitable amino acid to replace the $NADP^+$, since it is able to stabilize both interactions. Therefore, a molecule of tyrosine was manually inserted within the Pteridine reductase binding site, superimposing it on the cofactor and the obtained ternary complex was finally minimized to optimize the arrangement of the "new" cofactor.

Moreover, the charge of Trimethoprim was investigated by a re-docking study in which both ionization states were tested, and the best complexes, according to all score functions, were always found with the molecule in the protonated state (see Table 4.2).

SCORE	CHARGED	NOT CHARGED
Plp95	**-123,58**	-94,15
Plp	**-86,28**	-69,63
ChemPlp	**-81,84**	-63,5

Table 4.2: re-docking scores of TOP in the binding site on Pteridine reductase.

Finally, the so optimized pocket was submitted to SPILLO-PBSS for the generation of the RBS, as reported in Figure 4.22, which also compiles the key stabilizing residues ranked by relative relevance.



ASP A 181	52.931
ASP A 232	10.660
PHE A 113	8.333
GLY A 225	5.818
VAL A 230	4.740
TYR B 1	4.137
LEU A 188	3.833
LEU A 229	3.568
TYR A 191	2.215
LEU A 226	1.775
MET A 233	1.718
TYR A 194	0.270

Figure 4.22: Pteridine reductase RBS

The software found three symmetrical pockets in both the apo and the holo trimers and all pockets were located at the interface between two monomers, in the top of the extracellular part of the receptor (see Figure 4.23).

Since the three pockets were symmetrical and almost identical and there were not significant differences between the two receptor states, the following steps of optimization and validation were carried out on a single binding site, selected within the receptor structure in its closed state.

The optimization of the binding site involved two iterative induced fit procedures performed with a view to rendering the pocket able to accommodate the entire datasets. The first one took advantage of the re-docking of Trimethoprim in the binding site, while the second induced fit procedure exploited the DAP inhibitor **13W_20** (see Appendix 2), chosen as the best performing one in the previous docking study on the pocket resulting from the first induced fit. This second procedure improved the stability of the ligand-target complexes, and after it all the docked inhibitors assumed a binding mode in satisfactory agreement with the RBS generated by SPILLO-PBSS.

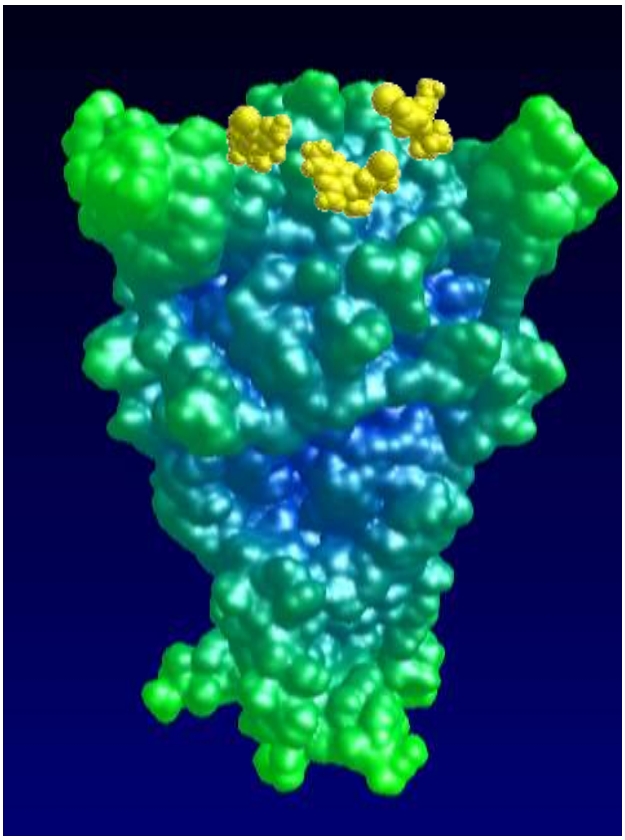
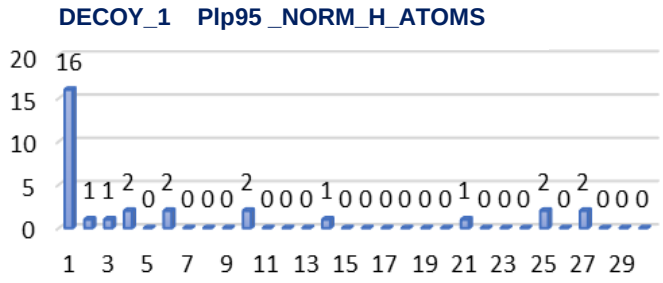


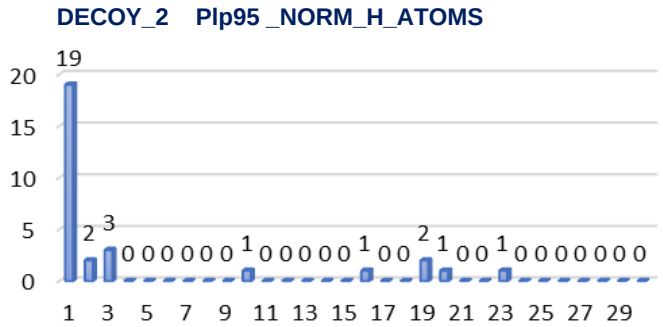
Figure 4.23: Potential allosteric binding site identified by SPILLO-PBSS on the apo model of the human $P2X_3$ receptor.

4.3.9.2.2 Validation of the pocket by virtual screening study

Once the binding site was optimized, its reliability was tested by performing virtual screening studies using both datasets and the three scoring functions implemented by PLANTS (i.e., CHEMPLP, PLP and PLP95). The best obtained results for the enrichment factor analysis are reported in Figure 4.24.



Top %	Mol	Att	EF
1.00	30	8	26.67
2.00	60	10	16.67
5.00	150	17	11.33
10.00	300	18	6.00
20.00	600	22	3.67



Top %	Mol	Att	EF
1.00	30	9	30.00
2.00	60	18	30.00
5.00	150	20	13.33
10.00	300	24	8.00
20.00	600	24	4.00

Figure 4.24: Enrichment factor analysis of the optimized pocket identify by SPILLO-PBSS.

The score function Plp95 normalized on the heavy atoms ranked in the top 1% and top 2% respectively 9 and 18 active molecules, so reaching an outstanding enrichment factor of 30.00. In particular, using both datasets, the enrichment of top clusters was obtained thanks to the high scores of the DAP molecules, which were well discriminated by the pocket. On the opposite, the SAA molecules are not ranked in the best positions, thus suggesting that this model do not successfully describe their interactions with the receptor.

An analysis was carried out on the complexes formed by the DAP molecules to identify the key residues for the interactions in the binding site. To this purpose, the RBS of each complex has been calculated by SPILLO-PBSS, obtaining the list of the residues in the pocket with the relative contribution to the overall interaction.

Four key residues were identified, namely Asp79, Thr82, Leu298 and Tyr70, all belonging to the same chain.
Table 4.3 reports the scores assigned by SPILLO-PBSS to the key residues for each complex with DAP compounds.

Figure 4.25 depicts an example of inhibitor in the allosteric pocket. The above mentioned key residues are involved in the following interactions:

- the carboxylate group of Asp79 establishes an ionic interaction with the positively charged nitrogen of the Trimethoprim heterocycle;
- the hydroxyl group of Thr82 establishes a H-bond with the amino substituent of the heterocycle;
- Leu289 and Tyr70 surround the ligand, stabilizing hydrophobic and π-π interactions.

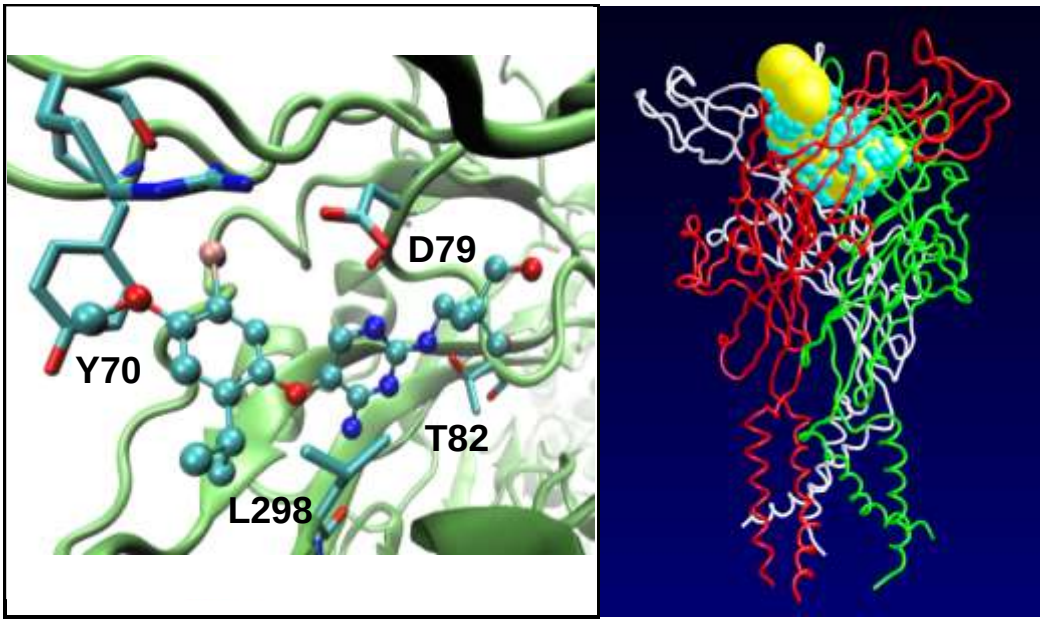


Figure 4.25: Allosteric pocket identified by SPILLO-PBSS. On the left the interaction map, on the right the same pocket as it was identified by Fpocket approach. The cavity is coloured in yellow, the atoms around in in blue.

ID	ASP79	LEU298	TYR70	THR82
c_6	54.549	9.171	< 5	5.506
13a_20	45.991	9.560	6.889	7.244
25_19	54.906	9.111	5.993	< 5
a_6	56.917	9.700	< 5	< 5
c_13	50.787	9.793	< 5	6.639
39_19	51.967	7.161	7.431	< 5
30_19	54.859	9.587	6.768	5.732
31_19	53.257	9.399	6.182	5.540
13e_20	42.138	8.496	5.840	6.701
13o_20	41.350	8.406	5.873	8.569
13r_S_20	47.067	9.066	5.673	10.215
13r_R_20	42.848	8.554	5.930	8.419
13q_20	40.939	9.799	6.317	7.662
13m_20	43.686	8.403	6.480	8.391
13d_20	38.926	8.349	5.460	11.233
13w_20	40.810	8.176	5.324	7.040
32_19	51.967	7.161	7.431	< 5
13s_S_20	44.990	7.588	5.082	9.273
13s_R_20	43.322	7.309	5.616	6.672
13p_S_20	15.825	11.052	6.953	5.674
28_19	50.437	9.803	< 5	6.984
13p_R_20	40.939	9.799	6.317	7.662
13u_20	41.667	7.236	< 5	8.252
13t_20	44.850	8.058	6.077	7.898

Table 4.3: SPILLO-PBSS pocket interaction profile. The reported values are percentages.

4.3.9.2.3 Validation of the pocket by QSAR study

The reliability of the binding site was also assessed by a QSAR analysis using a specific script implemented in VEGA ZZ software which performs combinatorial regression analyses by exhaustively combining all computed descriptors. For this analysis, only the 21 DAP derivatives were taken in consideration, and for the undefined chiral molecules both stereoisomers were simulated by averaging the corresponding descriptors. The ligand-based features included in the analysis were represented by constitutive and structural descriptors as well as physicochemical properties. Moreover, a docking study of the compounds in the binding pocket was performed by the docking algorithm PLANTS and a set of docking scores were calculated on the minimized obtained complexes and included among the compound features exploited by the QSAR analysis. The complete list of the ligand-based and docking-based descriptors taken in consideration is reported in Appendix 3.

The best predictive equations are reported in Table 4.4. The compound activity, measured by the value of pIC_{50} is well correlated with the docking score *CHEMPLP*, found in almost all reported equations. H-bonding seems to play a fundamental role in the stabilization of the ligand-target interaction and the compound features *HbTot* and *HbDon* are positively correlated with the activity measure. The polarity of the molecules (*Dipole* and H-bond terms) is also an essential element to define the affinity for the receptor, as well as the flexibility (*FlexTorsion* and *Torsion* terms) which presumably account for entropic factors.

Figure 4.26 reports the plot obtained by predicting the inhibitors' activity using the best equation, highlighted in yellow.

Mols	Vars	Outliers	R^2	Q^2	SE	F	Equation (pIC_{50}=)
21	3	None	0.63	0.41	0.254	9.49	1.9576 - 0.0486 CHEMPLP + 0.0629 Dipole + 0.1404 HbDon
21	2	None	0.59	0.44	0.259	12.93	2.6060 - 0.0503 CHEMPLP + 0.0531 Dipole
20	3	1 (32_19)	0.71	0.51	0.230	12.88	1.5560 - 0.0528 CHEMPLP + 0.0646 Dipole + 0.1339 HbDon
20	2	1 (32_19)	0.67	0.53	0.236	17.54	2.1660 - 0.0545 CHEMPLP + 0.0554 Dipole
19	3	2 (32_19) (13w_20)	0.76	0.62	0.212	16.20	1.6347 - 0.0533 CHEMPLP + 0.0395 Dipole + 0.1607 HbDon
19	3	2 (32_19) (13w_20)	0.76	0.53	0.213	15.94	3.0003 - 0.0589 CHEMPLP - 0.3185 Gyrrad + 0.0642 hbtot
19	3	2 (32_19) (13w_20)	0.75	0.59	0.217	15.19	3.1735 - 0.0394 CHEMPLP + 0.0310 Flex Torsions + 0.1573 Hb Don
19	3	2 (32_19) (13w_20)	0.75	0.58	0.220	14.62	2.7800 - 0.0456 CHEMPLP - 0.0002 PLP95 + 0.1480 HbDon
19	2	2 (32_19) (13w_20)	0.75	0.63	0.213	23.40	2.7785 - 0.0458 CHEMPLP + 0.1485 HbDon
19	2	2 (32_19) (13w_20)	0.73	0.55	0.221	21.24	3.2308 - 0.0444 CHEMPLP + 0.0488 HbTot
18	2	3 (32_19) (13w_20) (13o_20)	0.77	0.59	0.211	24.78	2.8862 - 0.0460 CHEMPLP + 0.0335 Torsions
18	**2**	**3 (32_19) (13w_20) (13o_20)**	**0.77**	**0.61**	**0.122**	**24.53**	**2.9827 - 0.0479 CHEMPLP + 0.0416 HbTot**

Table 4.4: Best correlative predicted equations for validation of SPILLO-PBSS pocket.
Mols = number of molecules; Vars = number of variables in the correlative equation; Outliers = number of molecules excluded to improve the relationship; R^2 = correlation coefficient; Q^2 = predictive power by leave-one-out cross-validation; SE = standard deviation of the errors; F = Fisher statistical coefficient; Equation (pIC_{50} =) = correlative equation.

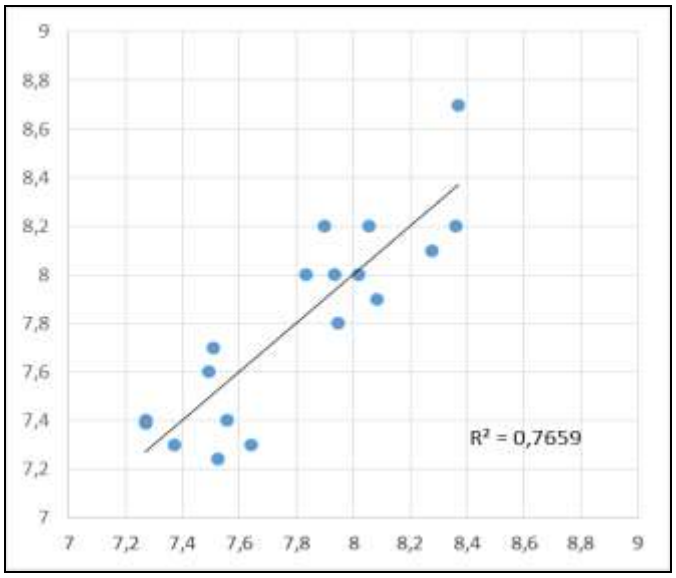


Figure 4.26: Plot of the linear equation yellow highlighted.
The experimental IC_{50} are reported on the X axe,
the predicted IC_{50} are reported on the Y axe.

4.3.10 PELE

PELE[148] is an on-line server able to perform an unconstrained search for binding sites within a protein structure. It simulates the movement of a given ligand on the receptor, considering both partners in a flexible way, thus enabling the ligand to explore regions which would be inaccessible when the interaction partners were considered fixed. The opportunity to take in consideration the conformational flexibility of the protein is the specific advantage which suggested the use of this approach to search the allosteric pockets within the purinergic receptors

4.3.10.1 Identification and optimization of the pocket

The selected compound to be used as a probe was Trimethoprim and not a specific inhibitor, because PELE needs a correct parametrization of the ligand atoms and so can recognize only the ligands already present in the Protein Data Bank.

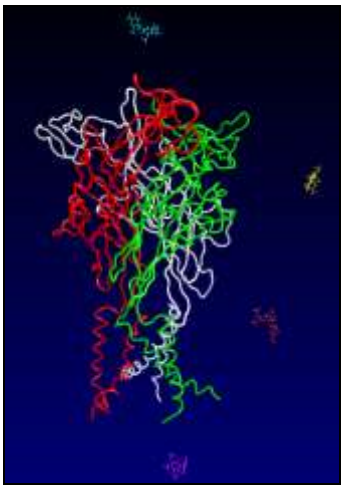


Figure 4.27: Starting positions of Trimethoprim simulations around $P2X_3$ receptor

Four different simulations were performed by using always the $P2X_3$ model in its apo state and by placing the probe about 30 Å distant from the receptor in different starting positions (Figure 4.27). The simulation parameters were set to obtain four simulations 24 hours long, including 600 steps, and using 11 CPU to obtain 10 trajectories for each simulation.

The 40 obtained trajectories were analyzed based on the binding energy calculated by PELE (see Figure 4.28) and the lowest energy frame was selected (-58.00 Kcal/mol).

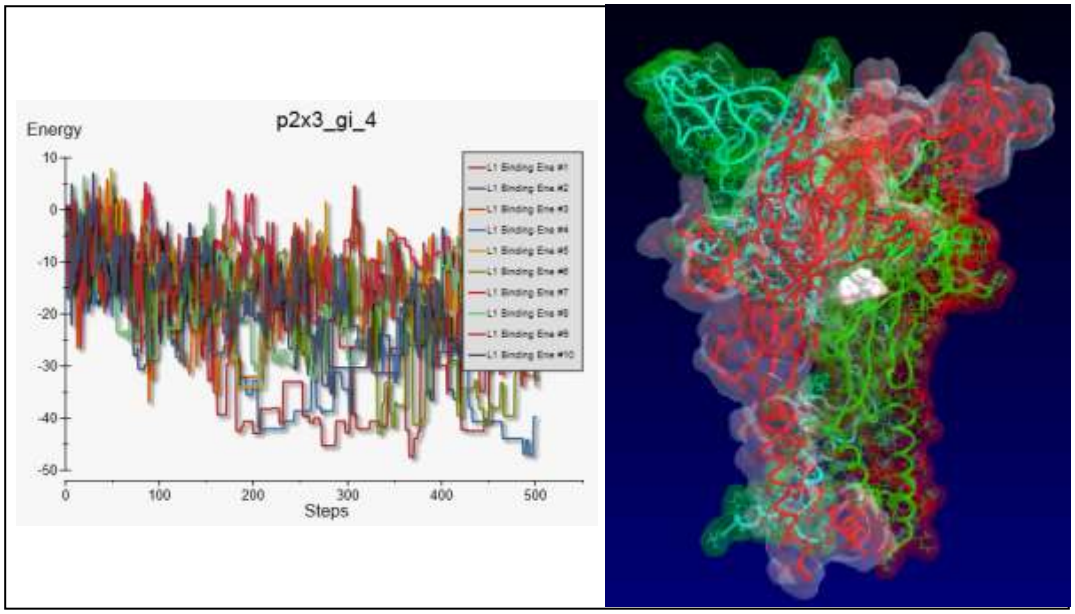


Figure 4.28: PELE output. On the left, an example of binding energy plot. Each trajectory has a different colour. On the right, the best selected frame. Trimethoprim is highlighted in white.

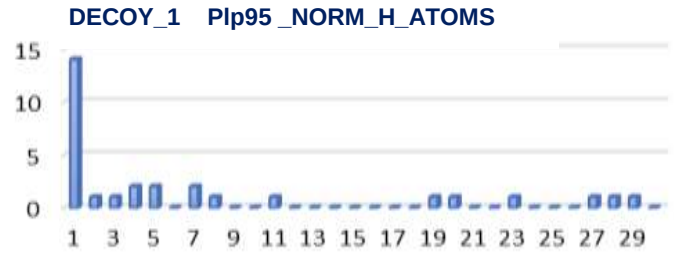
The complex corresponding to the best frame was optimized by an energy minimization, keeping free all the atoms in a sphere of 12 Å around the bound Trimethoprim. The selected pocket is located in the body domain below the ATP binding site, in a superficial and easily accessible region of the trimer, at the interface between two monomers.

The optimization of the pocket was completed with an induced fit procedure. A preliminary docking study was performed with the whole set of inhibitors and then the complex with the higher docking score (**13o_20**, see Appendix 2) was selected for the induced fitting procedure.

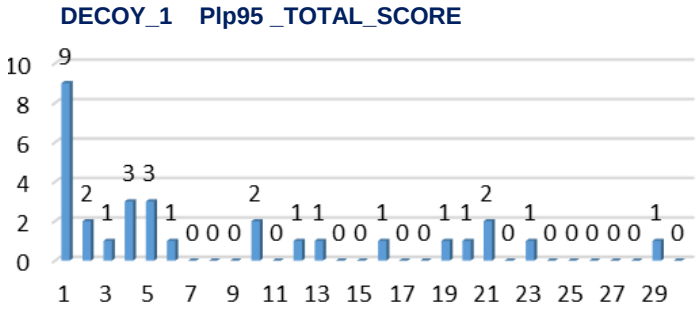
4.3.10.2 Validation of the pocket by virtual screening study

The reliability of the so optimized binding site was verified by a virtual screening study, following the same procedure used for the pocket identified by SPILLO_PBSS. The best EF was reached by using the decoy set number 1 and the PLANT scoring function Plp95, both as total score and as score normalized on the number of heavy atoms (Figure 4.29).

Interestingly and in contrast to what was observed for the allosteric pocket identified by SPILLO-PBSS, SAA derivatives are ranked in the best positions of the Plp95 total score. This evidence gives insights about the presence of more allosteric binding site, one of which can be specific for SAA compounds. However, their poses are not all perfectly located inside the pocket, therefore further investigation are required.



Top %	Mol	Att	EF
1.00	30	8	27.58
2.00	60	10	16.94
5.00	150	15	10.06
10.00	300	16	5.35
20.00	600	20	3.34



Top %	Mol	Att	EF
1.00	30	6	20.68
2.00	60	7	11.86
5.00	150	10	6.71
10.00	300	12	4.01
20.00	600	19	3.17

Figure 4.29: Enrichment factor analysis of the optimized pocket identify by PELE.

An analysis to identify the key residues for the interactions in the binding site was carried out similarly to what was done for the SPILLO_PBSS pocket. In this case, the RBS were computed with the DAP molecules in the neutral state, because they provided better docking scores. The key residues were Leu265, Ser267 and Asp266 in the chain A; Lys176, Glu187, Leu191, Ser178, Asn177 and Lys188, in the chain B. The results are shown in Table 4.5.

Mol Id	LYS 176	LEU 265	GLU 187	SER 267	LEU 191	SER 178	ASN 177	GLY 189	ASN 190	ASP 266	ILE 179	LYS 188
c_6	8.36	12,23	11,51	8,82	8,76	7,72	7,81	9,71		6,29		10,34
c_13_6	20.77	9,33	11,25	10,26	7,88		5,96	7,55				7,74
a_6		16,35	14,08	10,12		5,46	110,56	9,03		8,52	5,60	10,95
39_19		16,57	12,22		12,00		12,43			7,01	5,38	11,81
32_19	12.13	10,49	8,35	8,58	7,14	7,74	8,62	7,52	5,81	6,61		5,31
31_19	11.00	10,38	8,99	9,05	7,64	7,25	8,04		6,16	5,82		5,69
30_19	11.30	10,22	9,18	8,96	7,72	7,15	7,83	8,04	6,15	5,92		5,65
28_19	11.98	10,33	10,12	8,94	7,78	7,31	5,65	8,44	5,64			5,34
25_19	10.47	10,69	7,85	8,23	6,89	7,25	8,21	7,31	5,54	6,21		5,46
13w_20				8,33			10,85		6,04	11,54		9,93
13u_20	7.42	10,33	12,17	5,17		8,84	5,77					
13t_20	9.92	10,57		7,45	6,00		9,64	7,64		6,92	7,25	5,57
13s_S_20	8.87	9,55	6,67	6,65	5,12	10,22	5,73	5,58				
13s_R_20	10.29	10,91	8,82	6,37	5,65	9,73	6,38	5,32		5,23		
13r_S_20	11.52	12,30	11,08	6,75	6,47	10,43	6,44	5,27		6,05		
13r_R_20	9,26	9,95	9,22	5,82	5,76	7,75	6,33					
13q_20	9,71	10,96		6,58	5,64	8,24	7,06	5,74			5,31	
13p_S_20	8,75	9,80	8,36	6,39	6,28	5,49	6,14	5,75		5,33		5,68
13p_R_20	8,42	9,99	13,53	6,59	5,42	7,02	5,98	5,52				
13o_20	9,91	10,27	9,09	7,36	5,47	11,67	6,64	6,16				
13m_20	9,52	9,59	6,97	6,28	6,62	6,11	6,05	5,97			5,38	
13e_20		18,23	6,71	8,50			10,54			8,11	11,57	9,74
13d_20	9,37	9,06	9,76	5,37	6,01	10,53			6,10	5,20	5,81	
13a_20	10,84	10,37	8,16	7,60	5,84	7,37	7,05	6,58	5,00	5,50		5,70

Table 4.5: PELE pocket interaction profile. The reported values are percentage**s.**

4.3.10.3 Validation of the pocket by QSAR study

Similarly to the analysis performed on the SPILOO-PBSS pocket, the reliability of this binding site was also assessed by QSAR studies which included 21 DAP derivatives and the list of descriptors reported in Appendix 3.

The analysis of the best obtained correlative equations (Table 4.6) pointed out that the activity of the inhibitors can be here predicted using combinations of these different descriptors: PLP95 normalized per heavy atoms, lipole, binding energy and HMScore computed by X-Score and PSA. In particular, we can observe that the activity is:

- well correlated with the X-Score (both binding energy and HMScore) and the PLP95 score (as is or normalized on heavy atoms);
- proportional to the lipole value thus suggesting that the cavity prefers ligands in which apolar and polar moieties are structurally separated;
- proportional to both PSA and the number of H-Bonds, both descriptors emphasizing the key role of this kind of contacts in stabilizing the computed complexes.

Figures Figure 4.30 and Figure 4.31 report the plots obtained by predicting the inhibitors' activity using the best equations which include two and three variables, (highlighted in yellow and in green, respectively).

Mols	Vars	Outliers	R^2	Q^2	SE	F	Equation (pIC_{50}=)
21	3	None	0.60	0.41	0.306	8.56	3.3821 - 0.0452 PLP95 - 0.0455 Atoms - 0.0070 Electr
21	2	None	0.57	0.41	0.311	11.86	2.3795 + 0.7029 HMScore + 0.1574 Lipole
20	3	1 (28_19)	0.79	0.69	0.224	20.14	0.5461 + 1.1431 HMScore - 0.3581 Gyrrad + 0.2215 Lipole
20	3	1 (28_19)	0.78	0.67	0.225	19.27	1.1830 + 0.7713 HMScore + 0.1771 Lipole + 0.1463 HbDon
20	2	1 (28_19)	0.75	0.66	0.238	25.43	1.4146 + 0.8285 HMScore + 0.1753 Lipole
20	2	1 (28_19)	0.73	0.56	0.247	22.91	1.4025 + 1.0412 HMScore - 0.1699 VirtualLogP
20	2	1 (28_19)	0.72	0.59	0.251	22.09	1.4882 + 0.8458 HMScore + 0.1435 HbDon
19	3	2 (28_19) (39_19)	0.90	0.80	0.160	43.03	-2.3833 - 1.4320 Binding energy + 0.0290 PSA - 0.1189 Atoms
19	3	2 (28_19) (39_19)	0.90	0.80	0.161	42.94	-2.3816 + 1.9529 Average + 0.0291 PSA - 0.1193 Atoms
19	2	2 (28_19) (39_19)	0.78	0.62	0.226	28.34	1.4233 + 1.1117 HMScore - 0.2950 VirtualLogP
19	3	2 (28_19) (13t_20)	0.84	0.72	0.174	26.55	3.2159 + 1.1380 HMScore - 0.4471 Vdiam + 0.1449 Lipole
19	3	2 (28_19) (13t_20)	0.83	0.73	0.184	23.73	0.6799 + 0.7993 HMScore + 0.1169 Lipole - 0.1969 PLP95_NORM_H_ATM
19	2	2(28_19) (13t_20)	0.79	0.72	0.194	30.99	0.4366 + 0.8478 hms score - 0.2714 PLP95_NORM_H_ATM

Table 4.6: Best correlative predicted equations for validation of PELE pocket.

-

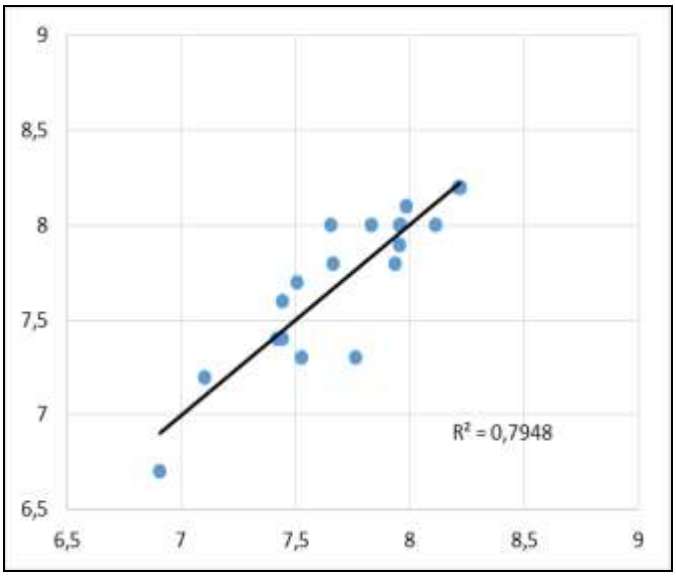


Figure 4.30: Plot of the linear equation yellow highlighted.
The experimental IC_{50} are reported on the X axe,
the predicted IC_{50} are reported on the Y axe.

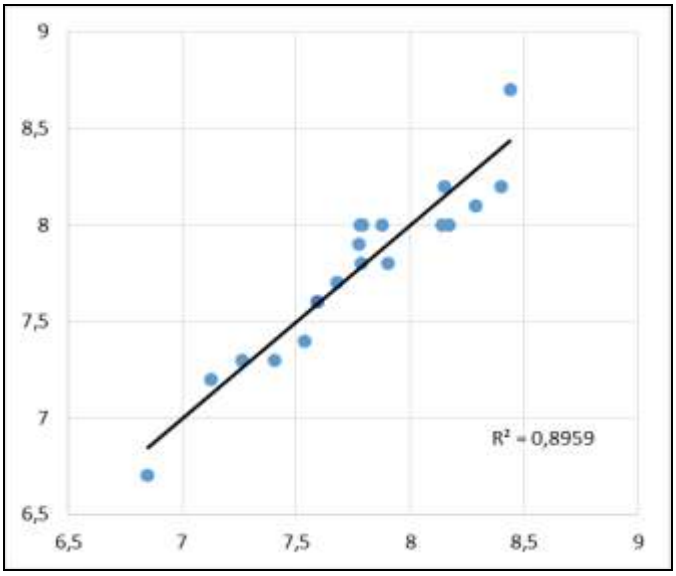


Figure 4.31: Plot of the linear equation green highlighted.
The experimental IC_{50} are reported on the X axe,
the predicted IC_{50} are reported on the Y axe.

4.3.11 Conclusions

The aim of the present study was to identify the potential allosteric binding sites within the structure of the human $P2X_3$ purinergic receptor. Several classes of compounds showing non-competitive antagonistic activities are reported in literature, but the binding pocket and the mechanism of action is yet unknown.

The study required the homology modelling of both the apo and the holo conformations of the human $P2X_3$ receptor, followed by the selection of highly active inhibitors, useful to validate the identified pockets.

Among different approaches, two in particular were successful: SPILLO-PBSS and PELE. The former identified three symmetrical pockets in the highest extracellular part of the receptor; the latter revealed a potential binding site in a lower region of the body domain, between two monomers. Both the potential pockets were validated by virtual screening approaches coupled to QSAR analyses.

It was also possible to observe a different behaviour for the two classes of considered compounds: in particular the DAP derivatives show good performances on both the pockets, while the SAA compounds seem likely to prefer only the binding site identifies by PELE, suggesting the possibility that the allosteric modulation of purinergic receptors is defined by multiple regions and different binding modes on the trimer structure.

We consider this study a first promising attempt to obtain reliable interaction models of the allosteric inhibition of $P2X_3$ receptors' activity, with the aim to gain insights for the rational drug design of novel purinergic ligands.

Chapter 5 – Muscarinic receptors

5.1 Introduction

In the last few years, several experimental advances have resulted in an incredible increase of the number of resolved GPCR structures[149]. With regard to class A, they involve rather heterogeneous receptors, including aminergic, nucleoside binding, peptide binding, and lipid binding GPCRs.
Although the inactive states are the most frequently resolved conformations, the GPCR structures include active, inactive and intermediate states, the comparison of which allows their activation mechanism to be investigated at molecular level[150]. Moreover, a systematic comparison of all resolved GPCR structures allows the common features of all binding sites as well as the key interactions stabilizing the GPCR folding to be revealed[151].

Such an abundance of resolved GPCR structures can clearly support homology modelling of the unresolved GPCRs[152], and in this context muscarinic receptors represent an invaluable arena for some key reasons which can be summarized as follows.
First, two muscarinic receptor subtypes have been recently resolved (i.e. mAChR2 and mAChR3) and this could allow truly reliable homology models to be generated for the other unresolved subtypes, especially considering the high homology between them. Additionally and as seen in the first below reported study, the availability of such highly homologous resolved structures can be utilized as a benchmark to assess the reliability of innovative modeling strategies.
Second, the mAChR2 subtype was resolved in its inactive more open state (namely co-crystallized with an antagonist, QNB) as well as in its active more closed state (namely in complex with a potent agonist, iperoxo) and this allows the differences

between the complexes generated by the two receptor states to be analyzed as seen in the second part of this Chapter.
Third, muscarinic receptors possess both orthosteric and allosteric binding sites and both pockets are well studied and there is a plethora of known ligands with various biological activity including bitopic ligands. Notably, a mAChR2 in its active state and in complex with an allosteric ligand was resolved thus providing an invaluable tool to investigate also the allosteric recognition at an atomic level as explored in the third part of this Chapter.

Fourth, since most known muscarinic agents are completely unselective there is a constant quest for improved ligands, the selectivity of which can be also gained by exploiting the allosteric site. Hence and although muscarinic receptors are probably the longest studied GPCRs, there is a continuous interest in these receptors and, due to the above mentioned reasons, the computational approaches can play a key role in the rational design of improved ligands.

Thus and as mentioned above, this chapter presents three different structure-based computational studies focused on the muscarinic receptors.
In detail, the first part described a standalone modeling study aimed at investigating the role of Pro-containing transmembrane helices to generate more reliable homology model.
The second study was carried out in collaboration with Prof. Piergentili and Prof. Quaglia from the University of Camerino and involved a docking analysis on a set of dioxane-based muscarinic agonists considering for the mAChR2 both open and closed states. Its key objective was to investigate whether docking results obtained by using highly reliable resolved structures can conveniently account even for very specific ligand features as exemplified here by the conformational equilibria characterizing the dioxane ring.

The last study, performed in collaboration with Prof. Romanelli from the University of Florence, concerned docking simulations on a set of bitopic ligands which are structurally related to carbachol. This study represents a clear example about how docking studies on muscarinic receptors can simultaneously involve both orthosteric and allosteric binding sites.

5.2 The role of Pro-containing helices in GPCR homology modeling

5.2.1 Setting the scene

Even though the above mentioned richness of resolved GPCR structures can support homology modelling studies, two different situations can be however recognized. In the first case, homology studies involve a receptor which belongs to a subfamily, a member of which has already been resolved and therefore modelling can greatly benefit from such a highly homologous template allowing truly reliable models to be easily generated. In the second and, unfortunately, more common situation, homology studies involve a GPCR which does not possess a highly homologous resolved congener and thus the generated models, despite using the best available templates, unavoidably show less reliability compared to that of the models obtained in the first fortunate situation.

Besides the availability of homologous templates, a deeper knowledge of their dynamic behaviour can further support the GPCR homology modeling. In this regard, great deal of attention has been paid to the conformational switches which modulate receptor activation[153].

Among them, several biophysical studies emphasized the key role of the Pro-containing transmembrane helices which can vary their bending acting as hinges to modulate the width of the TM bundle and, thus, receptor activation[154]. In a recent study, we showed that the structural effects of the Pro-containing transmembrane helices can be simulated by generating the so-called conformational chimeras, namely GPCR models in which the possible conformations of the Pro-containing

helices are exhaustively combined. These chimeras were found to account for receptor flexibility and were applied to explore ligand recognition by the human Cys-LTR1 receptor[155].
By accounting for the flexibility of their binding sites, such an approach should also be able to enhance the reliability and the predictive power of the modelled GPCR structures and such an effect should be particularly beneficial in the above described unfortunate situation, namely when the lack of highly homologous templates prevents the generation of truly reliable homology models.

On these grounds, the present study was undertaken with a view to investigating whether the reliability of GPCR models can be enhanced by simulating the bending of their Pro-containing TM helices. Specifically, the study involved the modeling of a muscarinic receptor subtype (mAChR1) which can be generated by using highly homologous templates, since the mAChR2 and mAChR3 subtypes have recently been resolved in complex with agonist and antagonist[156]. Such a reasonably reliable model was then compared with a homology model obtained starting from a less homologous non-muscarinic GPCR template, and optimized by varying the bending of its Pro-containing transmembrane helices.

The human mAChR1 subtype was here selected due to its relevant medicinal role, since selective modulation of this muscarinic receptor was found to be effective in cognitive models of Alzheimer's disease[157], and antipsychotic models of schizophrenia[158]. Hence, selective agonists (also including allosteric and bitopic ligands) are currently being pursued for the above mentioned therapeutic effects, avoiding the cholinergic adverse events induced by an unselective activation of all muscarinic receptor subtypes, which have hitherto limited the use of the available mAChR1 agonists in neurodegenerative disorders[159].
The evaluation of the so obtained models (and chimeras) was carried out by virtual screening campaigns where the reliability of the models was assessed by

computing the enrichment factors as well as a novel parameter based on the distribution of the active compounds within the entire screened database. The possibility of combining the scores of more chimeras was also evaluated by using a specially developed maximizing consensus algorithm.

5.2.2 Computational Methods

5.2.2.1 Generation of the starting models

The primary sequence of the human mAChR1 was retrieved by the Uniprot database (Entry Id: P11229, Entry name: ACM1_HUMAN). As discussed above, two homology models were initially generated. The first model (hereinafter named $M1_{M3}$) was built by using the highly homologous rat mAChR3 structure as the template (PDB Id: 4DAJ, Identity = 52.8%; similarity = 91.2%)[160], while the second model (hereinafter named $M1_{\beta2}$) was generated by using a less homologous non-muscarinic template. The human β2 adrenergic receptor was chosen as the second template due to the good homology with mAChR1 (PDB Id: 2RH1, Identity = 22.1%; similarity = 62.2%) and the fact that both GPCRs belong to the family of human aminergic receptors[161].

In detail, both models were generated by Modeller9.10 using the default parameters and generating 20 models for each run[162]. Among the generated models, the best structures were selected according to the scores computed by Modeller9.10 (i.e. DOPE and GA341) as well as to the percentage of residues falling in the allowed regions of the Ramachandran Plot. The satisfactory structural quality of the two selected models was then assessed by (a) the agreement with the predicted secondary structure from the sequence alignment, as obtained using ClustalX (see Figure S2, Supporting Information); (b) the lack of not predicted gaps; (c) the

remarkable percentage of residues falling in the allowed regions of the Ramachandran Plot (93.0 % and 91.8 %, for $M1_{M3}$ and $M1_{\beta 2}$, respectively) and of the χ-space (96.0% and 95.4%, for $M1_{M3}$ and $M1_{\beta 2}$, respectively).

The selected models were then completed by adding hydrogen atoms and to remain compatible with physiological pH, Asp, Glu, Lys and Arg residues were considered in their ionized form while His and Cys were maintained neutral by default. Such a standard protonation state should not influence the following docking results since the mAChR1binding site (as defined by a 12Å radius sphere around Asp105) does not include ionizable residues apart from the mentioned Aps105 which has to be stably ionized to interact with ligands and Cys407 which is in its neutral state as suggested by its very high pK value as predicted by PropKa[163].

The completed models were carefully checked to avoid unphysical occurrences such as cis peptide bonds, wrong configurations, improper bond lengths, non-planar aromatic rings or colliding side-chains. Finally, the mAChR1 models were optimized by a minimization made up by two phases: a first minimization without constraints until RMS = 0.1 kcal $mol^{-1}Å^{-1}$ and then a second minimization with backbone fixed until RMS = 0.01 kcal $mol^{-1}Å^{-1}$ to preserve the predicted structures. As said before, both modelled mAChR1 structures were utilized in the following docking simulations and the $M1_{\beta 2}$ model was also exploited to generate the corresponding chimeras by varying the conformation of its Pro-containing helices.

5.2.2.2 Generation of conformational chimeras

By breaking an intra-helical H-bond, a proline residue, when inserted in a TM helix, increases its flexibility and assumes two major conformations. In the first, the proline causes the helix axis to bend by about 20° while in the second the helix maintains a straight axis annulling the proline bending effect. When viewing the receptor from the extracellular side, such helices are usually bent outwards, thus increasing the overall opening of the TM bundle[164]. Since the TM region generally includes more Pro-containing helices, the effect of these flexible segments can significantly vary the opening of the TM bundle which indeed can shift from a cylindrical to a calix shape.

The mAChR1 receptor includes four Pro-containing TM helices, namely TM4 (Pro159), TM5 (Pro200), TM6 (Pro380) and TM7 (Pro415). Since the study involved the generation of all possible chimeras as obtained by exhaustively combining the two conformations of the four Pro-containing helices, the number of the so modelled $mAChR_1$ chimeras was equal to 16 (2^4). By considering that the major difference between straight and bent helix conformations involves the ψ angle, the four Pro-containing TM helices underwent a systematic search in which 360 conformers were generated for each helix by systematically rotating the proline's ψ angle (1 conformer/1 degree). The obtained conformers were minimized to discard high-energy interactions. Similarly to what was obtained in the previous study, all Pro-containing helices show two major conformations. The first is a negative synclinal geometry ($\psi \cong 320$) which is typical for left-handed helices and corresponds to the straight conformation, whereas in the second the ψ angle assumes a synperiplanar conformation ($\psi \cong 0$), inducing the corresponding bending in the helix axis.

Starting from the so generated conformations, the 16 chimeras were assembled by systematically combining the straight and bent conformations of each Pro-

containing helix. In detail, the model assembly was performed by superimposing the backbone atoms of a helix conformation with those of the correspondent segment in the $M1_{\beta 2}$ model and manually connecting the adjacent segments using the VEGA suite of programs[165]. After a careful scrutiny of the obtained structures to avoid unphysical conditions, the sixteen models underwent a local minimization until RMS = 0.05 kcal $mol^{-1}Å^{-1}$, where all atoms were kept fixed except for those included within a 10.0 Å sphere around the manually connected bonds (at the helix ends). Finally, the models were minimized with backbone fixed until RMS = 0.01 kcal $mol^{-1}Å^{-1}$ to preserve the folding of the assembled structures. All described minimizations were performed using the conjugate gradient algorithm by Namd2.9[166] with the force field CHARMM and the Gasteiger's atomic charges.

The chimeras are termed according to the nomenclature proposed in the previous study[155]. Briefly, their name defines in sequence the conformation of the four Pro-containing helices (namely, TM4, TM5, TM6 and TM7) using the code "S" for straight conformations and "B" for bent conformations. For example, SSSS indicates the fully close chimera (i.e. all straight helices) while BBBS denotes the chimera with all bent helices apart from TM7.

5.2.2.3 Dataset collection

As reported in Figure 5.1 and Table 5.1, a representative dataset of 35 mAChR1 agonists were collected from the literature[167,168,169,170,171,172,173,174,80,175,176,177].

The dataset shows a good heterogeneity as assessed by a Tanimoto distance average equal to 0.18 and spans a rather wide range of activity of 5.55 logarithmic units. The ligands were simulated in their ionized forms since they are involved in molecular recognition. The conformational behaviour of the compounds was investigated by a clustered MonteCarlo procedure (as implemented in VEGA) which generated 1000 conformers by randomly rotating the rotors. For each ligand, the so obtained lowest energy structure was then exploited in the following docking simulations.

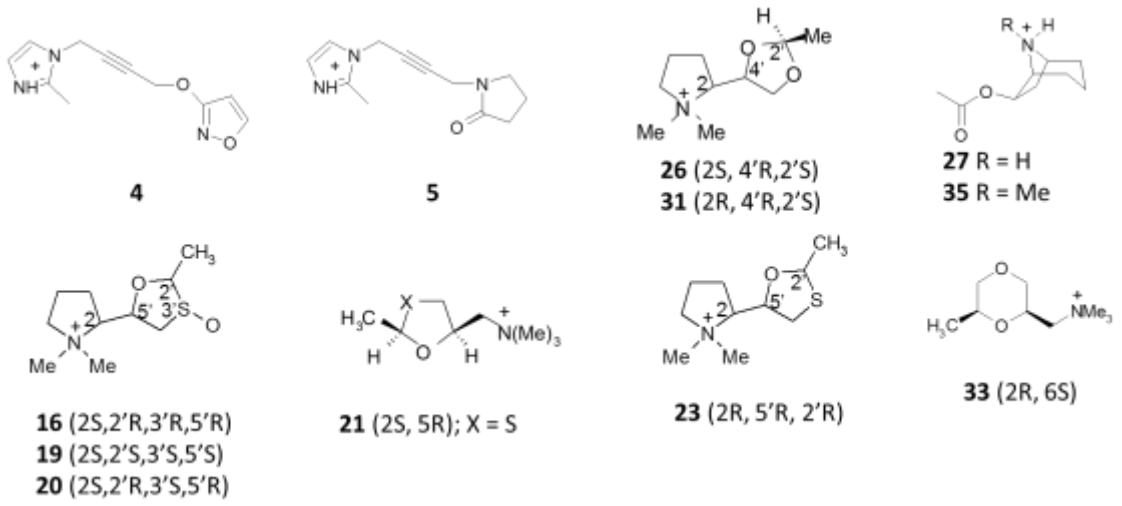


Figure 5.1: mAChR1 agonists. Molecular structure of the lesser known mAChR1 agonists included in the dataset

Cpd	Common name (when available)	pK_{M1} (nM)	Ref.	Cpd	Common name (when available)	pK_{M1} (nM)	Ref.
1	Oxotremorine	8.95	19	**19**		5.08	19
2	Xanomeline	8.10	19	**20**		5.04	19
3	WAY-132983	7.75	20	**21**		4.96	19
4		7.04	19	**22**	Carbachol	4.92	19
5		6.92	19	**23**		4.90	19
6	Talsaclidine	6.86	25	**24**	Muscarine	4.74	19
7	Sabcomeline	6.64	19	**25**	Oxotremorine-M	4.69	19
8	Tiopilocarpine	6.53	26	**26**		4.67	19
9	Arecaidinepropargyl ester	6.53	24	**27**		4.67	19
10	RS-86	6.31	27	**28**	Methylfurmethide	4.55	30
11	LY593093	6.21	21	**29**	Areocolina	4.53	19
12	CI-1017	6.17	22	**30**	Methacholine	4.52	28
13	77-LH-28-1	6.00	23	**31**		4.28	19
14	Alvameline (Lu25-109)	5.72	29	**32**	Furmethide	4.11	30
15	AC-42	5.52	23	**33**		4.09	19
16		5.23	19	**34**	Bethanechol	4.01	28
17	Pilocarpine	5.18	19	**35**		3.55	19
18	McN-A-343	5.10	28				

Table 5.1: mAChR1 agonists. Set of 35 known mAChR1 agonists taken from literature

By combining the selection features of the ChemBridge website[178] with those of the database explorer of the VEGA suite of programs, a set of decoys to be used in virtual screening campaigns was collected. The selected compounds had to fulfil the following criteria: (1) the molecular charge equal to +1 due to the presence of only one positively charged group (quaternary ammonium or ionizable amino group) belonging to a linear or cyclic moiety; (2) the presence of at least one H-bond acceptor group; (3) the molecular weights within the same range of the active

ligands (i.e. from 159 Da to 378 Da). In this way 2465 decoy molecules were identified and collected in a database together with the 35 active ligands, so as to have a complete dataset of 2500 compounds, where the active ones represent the 1.4 %.

The decoy molecules were prepared by using an automatic script of the VEGAZZ suite of programs which for each compound performs the following tasks: (i) generating the 3D structure; (ii) adding the hydrogen atoms; (iii) assigning atom types and atomic charges according to Gasteiger method; (iv) for ionizable molecules, selecting the predominant form at physiological pH; (v) for undefined chiral molecules, generating all possible stereoisomers; (vi) minimizing the so obtained molecules combining steepest descent and conjugate gradient algorithms.

5.2.2.4 Docking simulations and virtual screening

Docking simulations involved: (1) the more reliable $M1_{M3}$ model; (2) the less reliable $M1_{\beta 2}$ model; (3) all sixteen chimeras. Docking simulations were carried out using PLANTS, which finds plausible ligand poses by ant colony optimization algorithms (ACO)[147]. For all docking simulations, PLANTS was used with default settings and without geometric constraints. The search was focused on a 10.0 Å radius sphere around Asp105 thus encompassing the entire binding cavity. To avoid excessively time-demanding calculations which would be unsuitable for screening of large databases, the ligands were considered as flexible, while protein atoms were kept fixed. For each campaign, speed 1 was used and 1 pose was generated for each ligand.

Virtual screening campaigns can be subdivided into two parts. The first involved the two starting models (namely $M1_{\beta 2}$ and $M1_{M3}$) and the simulations were

repeated by scoring the poses with all three implemented score functions (ChemPlp, Plp and Plp95) with a view to revealing the most predictive scoring function. The second part involved the 16 generated chimeras and the simulations utilized only the most predictive Plp95 score.

In order to compare the predictive power of each chimera, the so generated scores were analysed by using a script which analyses the distribution of the active compounds throughout the entire ranking. In detail, the script subdivides the ranking into an user-defined number of bins (by default 100 bins) and counts how many active molecules are contained in each bin. Such an approach helps the analysis and comparison of virtual screening campaigns by providing both a graphical output (see Figure 5.3) and a metric based on the so obtained distributions. In detail, this novel parameter corresponds to the skewness of the distribution which is a measure of its asymmetry and is computed by Eq. 1 as the third standardized moment where x_i is the number of active compounds within the bin i and n is the total number of bins. In a satisfactory distribution the parameter has to be the largest possible positive number.

$$Skewness = \frac{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{n}}{\left(\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}\right)^{3/2}}$$

(Eq. 1)

5.2.2.5 Consensus algorithm

There are several applications for consensus algorithms in computational chemistry. When applied to docking simulations, they are based on the concept that an optimal combination of more docking scores always performs better than even the best performing score[179]. In correlative studies, linear correlations are thus developed by regression techniques involving more docking scores and aimed at maximizing the predictive power of the so obtained equations. Several studies have reported that a proper combination of different scoring methods can perform better than individual scoring functions in virtual screening, as well[180]. Several consensus approaches (rank-by-score, rank-by-rank, and rank-by-vote strategies) have been proposed even though maximizing algorithms have been scantly exploited to enhance virtual screening performances. For example, several general algorithms have been reported to maximize the Area under ROC Curve even though applications to virtual screening are not yet described (see for example ref [181]).

On these grounds and considering the possibility of combining the scores of more chimeras, we developed a fitting consensus algorithm which maximizes the enrichment factors by combining more scoring functions. As generalized by Eq. 2, the consensus algorithm allows a new ranking function to be obtained, resulting from the linear combination of two or more docking scores. Notice that the included *S* values can be both docking scoring functions and ligand-based descriptors.

$$Consensus = k_1S_1 + k_2S_2 + k_3S_3 + \cdots + k_nS_n + k_0 \qquad \text{(Eq. 2)}$$

The script calculates the *k* coefficients ($k_0 \ldots k_n$) by using the gradient-free Hooke-Jeeves algorithm[182] and, to avoid local maxima, random sampling is also applied. In detail and for each trial, random coefficients are initially assigned to each term of the linear combination and then optimized iteratively by maximizing the sum of the ranking positions of the active compounds in the list (as sorted from worst to best scores). This procedure is repeated for the selected number of the random trials and at the end of the calculation the best performing equation is shown.

The algorithm was implemented in two scripts. In the first, the user has to indicate two or more descriptors to be utilized to maximize the enrichment factors, while in the second script the user has to define only the number of descriptors to be included in the consensus equation and the script automatically searches for the best consensus equation by exhaustively combining the set of descriptors defined by the user. Due to the large number of combinations that can be generated, the second script was parallelized to exploit modern multi-core CPUs. The scripts are written in C language as implemented in the script engine of the VEGA ZZ software. They read the input data (biological activities plus ligand-based and/or docking-based descriptors) from CSV files and include a graphical interface which permits a user-friendly setting of all required parameters.

5.2.3 Results and Discussion

5.2.3.1 Overview of the generated chimeras

Figure 5.2 reports the superimposition of the 16 generated chimeras and shows in the right panel the regions with the greatest structural differences. Understandably, marked differences are seen in the two regions including the Pro-containing helices, namely the TM4-EL2-TM5 region (Ala142-Trp203) and the TM6-EL3-TM7 region (Leu367-Cys421). In both regions, one segment (i.e., TM7 and TM4) has the proline in the middle and shows clear differences between the simulated chimeras, whereas the other TM segment (i.e. TM5 and TM6) includes the proline residue in a more lateral position and thus induces more restricted conformational differences. In all chimeras, bending of the Pro-containing helices induces significant effects on the conformation of the two comprised extracellular loops (EL2 and EL3), the mobility of which slightly impacts also on the arrangement of the other two extracellular segments (i.e. NT and EL1) as seen in Figure 5.2.

Table 5.2 reports some relevant geometrical and physicochemical properties for the modelled chimeras including the void volume of the binding cavity as computed by FPocket [183].

The reported values suggest that bending of the Pro-containing helices induces an increase of solvent accessible surface area, an effect clearly understandable considering that the helix bending induces the overall opening of the TM bundle. When decomposing the surface area in polar and apolar components, one may note that the helix bending induces a substantial increase of the apolar surface, while the polar area remains roughly constant regardless of the bending. Such a behaviour can be explained considering the abundance of hydrophobic residues in both the external surface and in the accessible surface of the binding cavity the exposure of

which is necessarily increased by the TM bundle widening. When considering the apolar nature of the membrane environment, one may suppose that the reported increase of the apolar surface should not have destabilizing effects.

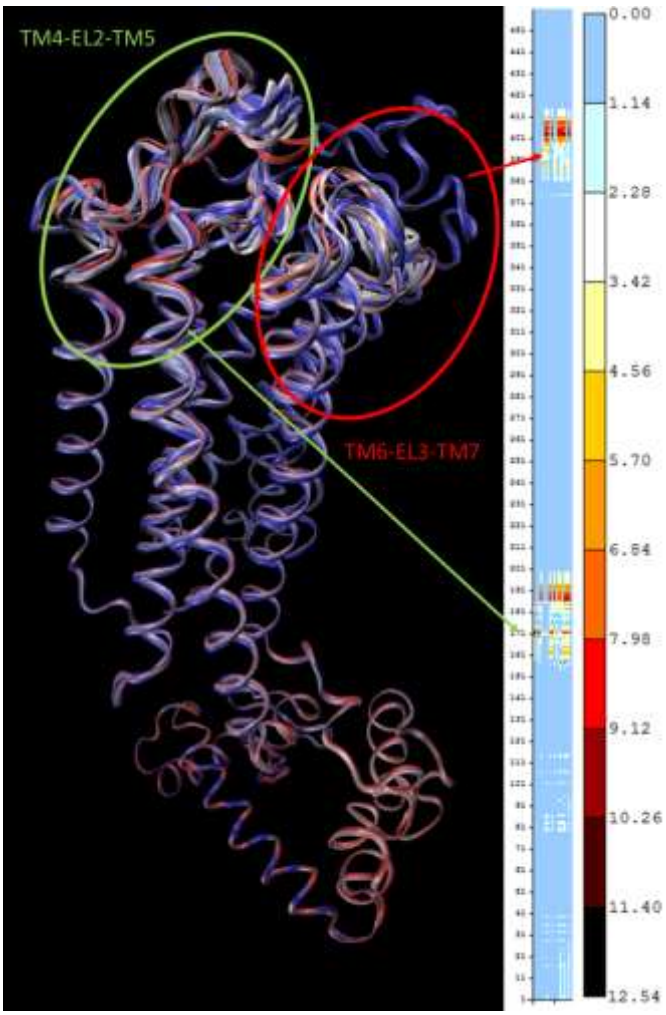


Figure 5.2: Superimposition of the 16 generated chimeras. It is evidencing the regions with the greatest structural differences as detailed in the right panel which maps the corresponding RMSD values.

Table 5.2 also shows that the opening of the TM bundle induces a clear enlargement of the binding cavities as described by their void volumes. In detail, the computed volumes allow the chimeras to be subdivided into three groups depending on the accessibility of their binding sites.

The first group (volume < 3000 A3, 5 structures) is composed of the closed chimeras, the binding sites of which are completely surrounded by residues and rather inaccessible from the outside; $M1_{\beta 2}$ belongs to this closed group.

The second group (3000 A^3 < volume < 6000 A^3, 6 structures) is composed of the intermediate chimeras, the binding sites of which are not so enlarged but open and accessible from the outside; $M1_{M3}$ belongs to this intermediate group. The last group (volume > 6000 A^3, 5 structures) is composed of the open chimeras, the binding sites of which are wide-open and accessible from the outside.

The volumes of the cavities reveal that bending of the TM6-EL3-TM7 region plays the most significant role in determining the opening of the TM bundle, TM7 being the most effective helix.

Finally, the RMSD values reported in Table 5.2 suggest that the arrangement of the TM bundle of $M1_{\beta 2}$ resembles that of the fully close SSSS chimera, while that of $M1_{M3}$ appears to be similar to the arrangement of the SBSS chimera. This finding emphasizes that simulation of the helical bending should not be only seen as a way to reproduce the conformation of the more reliable $M1_{M3}$ model but also as an approach to extensively simulate the receptor flexibility with a view to optimizing the ligand recognition.

Model	PSA (Å^2)	ASA (Å^2)	Area (Å^2)	Cavity Volume[a] (Å^3)	RMSD vs $M1_{M3}$[b] (Å)	RMSD vs $M1_{\beta2}$[b] (Å)
$M1_{b2}$	16861	32222	49083	2328	1.03	0.00
$M1_{M3}$	16805	33301	50106	7401	0.00	1.03
SSSS	16861	31226	48787	1534	0.80	0.45
BSSS	16705	32400	49105	4649	0.61	1.23
SBSS	16685	32809	49494	1668	1.10	1.49
SSBS	16837	33220	50057	5845	1.56	1.23
SSSB	16641	32664	49305	5181	1.49	1.87
SSBB	16655	32696	49352	7717	1.55	1.92
BBSS	16672	32877	49550	1822	1.24	1.31
SBBS	16716	32838	49554	5074	1.53	1.45
BSSB	16632	32743	49375	2534	1.17	1.58
BSBS	16821	33217	50038	4332	0.79	0.92
SBSB	16454	34235	50689	6719	1.74	2.08
SBBB	16604	34499	51103	4860	1.79	2.19
BSBB	16623	34719	51343	7363	1.56	1.57
BBSB	16563	34422	50984	6494	1.85	1.92
BBBS	16718	34904	51621	2412	1.30	1.37
BBBB	16433	34338	50771	8165	1.84	1.93

Table 5.2: Geometrical properties of the generated mAChR1 models. a) Void volumes computed by FPocket b) RMSD values as computed by considering the backbone atoms only.

5.2.3.2 Preliminary correlative analysis

Despite being unrelated to following virtual screening campaigns, a preliminary study involved the analysis of the linear correlations between the Plp95 scores (which were found to be the best performing ones, see below) and the pKi affinity (as computed by their r^2 values, see Table 5.3).

Such an analysis revealed some interesting correlations, although the exploited docking scores are directly derived from docking simulations without any complex's refinement. Understandably, the $M1_{M3}$ scores perform markedly better than the $M1_{\beta2}$ scores, while the computed r^2 values reveal significant differences among the considered chimeras, thus suggesting that the generated chimeras represent significantly different conformations compared to the starting homology model.

Notably, the docking scores of some chimeras (e.g. SBSB) reveal promising correlations with the affinity values while showing a conformation significantly different compared with that of $M1_{M3}$ (as encoded by the RMSD values, see Table 5.2). In other words, the capacity to approximate the $M1_{M3}$ model is not the key parameter when assessing the predictive power of a given chimera. Again, the five best performing chimeras (namely with an r^2 value greater than 0.4) do not show common features neither with regard to the helix's conformation nor with regard to the cavity volume.

This suggests that there is not a structural element which determines the chimera's performance, but the simulated chimeras represent relevant and distinct receptor states involved in the molecular recognition process and as such they may find fruitful applications in provisional studies regardless of their similarity to the starting homology models.

Model	r^2 Plp95 vs pK	EF 1%	EF 2%	EF 5%	EF 10%	EF 20%	EF mean	bottom 50%	Skew ness	AUC ROC
$M1_{b2}$	0.31	0.00	2.87	3.44	3.71	2.86	2.58	20.00	1.80	0.74
$M1_{M3}$	0.47	2.87	2.87	3.45	4.58	3.00	3.35	5.71	3.18	0.81
SSSS	0.42	2.87	2.87	4.02	4.86	3.29	3.58	2.86	2.78	0.82
BSSS	0.47	0.00	1.44	4.6	4.00	2.57	2.52	5.71	1.86	0.78
SBSS	0.35	0.00	0.00	1.15	2.86	3.15	1.43	8.57	2.04	0.77
SSBS	0.28	0.00	5.74	5.17	4.86	3.29	3.81	2.86	2.45	0.83
SSSB	0.31	0.00	2.87	4.6	4.29	3.00	2.95	17.14	1.84	0.76
SSBB	0.32	0.00	2.87	4.6	4.29	2.72	2.90	17.14	2.24	0.76
BBSS	0.21	0.00	4.31	4.6	4.86	3.43	3.44	8.57	2.64	0.82
SBBS	0.36	0.00	0.00	1.15	2.86	3.00	1.40	8.57	2.20	0.77
BSSB	0.35	5.74	4.31	5.17	4.58	3.15	4.59	14.29	2.73	0.78
BSBS	0.36	0.00	1.44	5.17	4.29	3.15	2.81	5.71	2.57	0.8
SBSB	0.42	0.00	0.00	3.45	2.57	2.57	1.72	17.14	2.07	0.74
SBBB	0.41	0.00	4.31	3.45	3.15	2.29	2.64	11.43	2.24	0.74
BSBB	0.40	0.00	4.31	4.6	4.00	2.86	3.15	8.57	1.85	0.77
BBSB	0.29	2.87	2.87	2.87	4.00	2.72	3.07	14.29	2.62	0.76
BBBS	0.11	0.00	2.87	4.02	4.29	3.29	2.89	8.57	2.20	0.8
BBBB	0.31	2.87	1.44	4.02	4.00	2.86	3.04	14.29	2.20	0.76

Table 5.3: Figures of merit for the performed virtual screening campaigns plus the r^2 parameter for the correlation between the Plp95 docking scores and the experimental affinity values.

5.2.3.3 Virtual screening campaigns

As a preamble, it should be observed that the docking scores normalized per the ligand's molecular weight perform vastly better than the total scores as witnessed by the overall enrichment mean of the normalized scores, which is about twice that of the total scores (2.88 vs 1.48, see Table 5.5). This result can be explained by considering that the normalized values indirectly account for the unavoidable differences between decoy set and active ligands[184]. Hence, most of the reported virtual screening results are focused on the normalized scores only.

The preliminary analysis concerns the choice of the most effective docking score from among the three functions computed by PLANTS (ChemPlp, Plp and Plp95) using the $M1_{\beta 2}$ and $M1_{M3}$ models as reference structures. Table 5.4 compares the enrichment factors obtained by using the normalized scores for the three implemented functions, and reveals that the Plp95 score produces significantly better results in both the M1 models. Consequently, the following virtual screening campaigns are performed by using the Plp95 function only.

Table 5.3 compares the results obtained by virtual screening campaigns on each generated chimera and reports the enrichment factors as computed for different thresholds (1%, 2%, 5%, 10% and 20%), as well as the corresponding enrichment means. As expected, the $M1_{M3}$ model performs remarkably better than the $M1_{\beta 2}$ model. The greatest enrichment is observed in both models in the top 10%, while only the $M1_{M3}$ model is able to place active compounds in the top 1%.

Also with regard to virtual screening, the simulated chimeras show significantly different results even though there is no agreement between the above discussed r^2 values and the chimeras offering the highest enrichment factors. As examples, the chimera producing the highest enrichment mean (BSSB) gives average r^2 value and that showing the second highest enrichment mean (SSBS) affords a very poor correlation.

For completeness, Table 5.5 compiles the enrichment factors as obtained by using the total Plp95 scores and as mentioned above shows that these scores perform vastly worse than the normalized values. Curiously, most chimeras shows EF top 1% values greater than 0 probably due to 1 or 2 very large active molecules properly evaluated but then the following enrichment factors dramatically drop.

The computed enrichment factors reveal that the most performing chimeras have cavities with intermediate volumes thus confirming that they show a satisfactory selectivity while possessing a suitable capacity to accommodate even large ligands. However, even the fully closed chimera shows remarkable results that can be explained by considering that this chimera is able to properly recognize small active ligands such as carbachole and metacholine. By contrast, chimeras

Model	score	EF 1%	EF 2%	EF 5%	EF 10%	EF 20%	EF mean
$M1_{\beta2}$	Plp95	0.00	2.87	3.44	3.71	2.87	2.58
$M1_{\beta2}$	ChemPlp	0.00	0.00	3.44	2.87	2.29	1.72
$M1_{\beta2}$	Plp	0.00	0.00	3.44	3.44	2.29	1.83
$M1_{M3}$	Plp95	2.87	2.87	3.44	4.58	3.00	3.35
$M1_{M3}$	ChemPlp	0.00	1.15	3.44	2.87	2.29	1.95
$M1_{M3}$	Plp	0.00	1.15	2.27	2.87	2.29	1.72

Table 5.4: Figures of merit for the preliminary virtual screening analyses performed on the $M1_{M3}$ and $M1_{\beta2}$ models by testing all three implemented score functions.

Model	EF 1%	EF 2%	EF 5%	EF 10%	EF 20%	EF mean
SSSS	0,00	1,43	0,57	0,86	1,29	0,83
BSSS	2,87	1,43	2,29	1,43	1,14	1,83
SBSS	2,87	2,87	2,29	1,43	1,14	2,12
SSBS	0,00	0,00	1,15	1,43	0,86	0,69
SSSB	2,87	2,87	1,15	0,86	0,43	1,64
SSBB	2,87	1,43	0,57	0,57	0,43	1,17
BBSS	0,00	2,87	2,29	1,71	2,00	1,77
SBBS	5,74	4,30	2,87	1,43	1,14	3,10
BSSB	2,87	1,43	0,57	0,57	0,29	1,15
BSBS	2,87	4,30	2,29	1,43	1,43	2,46
SBSB	2,87	1,43	0,57	0,57	0,43	1,17
SBBB	0,00	0,00	0,57	1,14	1,00	0,54
BSBB	2,87	1,43	1,15	0,57	0,86	1,38
BBSB	2,87	1,43	0,57	0,57	0,43	1,17
BBBS	2,87	1,43	1,72	0,86	0,86	1,55
BBBB	2,87	1,43	0,57	0,29	0,57	1,15

Table 5.5: Enrichment factors for the generated chimeras as generated by using the total Plp95 scores.

5.2.3.4 Consensus Functions

The last part of this analysis utilizes the above described script which combines more docking scores to maximize the corresponding enrichment factors.

Table 5.6 lists the best performing equations as obtained by maximizing the EF top 1% values. The first two equations (Eqs. 2 and 3) confirm the fruitful opportunity to combine more chimeras to improve the enrichment factors. Such an improvement appears to be significant even using the scores of only two different chimeras and becomes impressive when combining three chimeras. Even though the script could combine more scores (in theory even all sixteen chimeras), the developed equations include at most three variables to avoid overfitting conditions.

Notably, Eq. 3 includes chimeras with intermediate cavities thus confirming the general applicability of these in-between chimeras.

On the other hand, Eq. 4 includes chimeras differing in cavity volume probably because in this way they can recognize ligands proportionally differing in their size and shape. Stated differently, a single chimera recognizes heterogeneous ligands with difficulty and chimeras with intermediate cavities may represent an acceptable compromise to do this. However, vastly better results are obtained by combining different chimeras which are optimized to recognize small, intermediate and large ligands.

The last two equations (Eqs. 5 and 6) combine docking scores with ligand-based descriptors and reveal very remarkable enrichment factors which emphasize the synergistic effect of combining heterogeneous parameters[185]. In detail, the study exploited a list of 20 well-known physicochemical and stereo-electronic parameters as computed by VEGA and semi-empirical calculations (see Appendix 4). Although consensus functions may be further enhanced by considering a richer set of ligand-based descriptors, such a preliminary analysis was aimed at giving a hint about the beneficial effects of combining diverse descriptors without stressing the study with an exaggeratedly rich list of descriptors.

The best performing equation (Eq. 4) includes two ligand-based descriptors and the score of one chimera while Eq. 5 includes the docking scores of two chimeras. The included ligand-based descriptors underline the key role of ligand size and polarity, while the chimeras included in Eq. 5 confirm the positive role of combining cavities of different capaciousness for suitably accommodating small and large ligands.

N.	Equation	EF 1%	EF 2%	EF 5%	EF 10%
2	1.00 Plp95_SBBS - 0.77 Plp95_SBSB	7.94	5.29	6.35	3.95
3	1.00 Plp95_SSBS + 0.69 Plp95_SSBB - 1.62 Plp95_SBBB	26.45	14.55	6.35	4.74
4	1.00 Plp95_SSBS + 1.21 Radius of gyration + 1.42 HbDon	52.91	30.42	14.28	7.65
5	1.00 Plp95_SSSS - 0.25 Plp95_BBBB + 0.24 HbDon	42.35	27.77	13.75	7.37

Table 5.6: Enrichment factors for the optimized equations as obtained by combining the scores of more chimeras with ligand-based descriptors.

5.2.4 Discussion

The above reported virtual screening campaigns were analysed by computing the corresponding enrichment factors, which indeed appear to be meaningful when comparing virtual screening campaigns performed by using the same dataset and different conformations of the same receptor (as in this study)[186]. Besides the arbitrariness of the defined thresholds, the use of enrichment factors has often been questioned, since they are focused on the top-ranking molecules and unavoidably ignore what happens in the remaining part of the ranking. Thus, Table 5.6 also compiles the percentage values of active ligands found in the lower half of the ranking. However, this parameter describes the abundance of false negatives but is unable to convey a clear picture concerning the distribution of active compounds throughout the ranking.

Hence, Figure 5.3 shows the distributions for some illustrative models of the active compounds by subdividing the ranking into 100 bins (each composed of 25 molecules). Ideally, the active compounds should be included in the first bins, gradually decrease in number in the following bins and be absent from the last bins. By contrast, a virtual screening which evenly distributes the active compounds throughout the ranking should be considered as ineffective regardless of how many active ligands are placed in the first bins.

Figure 5.3 compares the virtual screening campaigns as performed on $M1_{\beta 2}$ (Fig. 5.3-A) and $M1_{M3}$ (Fig. 5.3-B) models and emphasizes that the latter, besides having markedly higher enrichment factors, also shows a more satisfactory distribution as confirmed by the percentage of active compounds in the bottom 50%. In detail, Figure 5.3-B shows a significant concentration of the active compounds in the top-ranked 8 bins, while Figure 5.3-A shows an unsatisfactorily flat distribution of the active compounds until the 75^{th} bin.

The histograms can also be useful in analysing the modelled chimeras. Figure 5.3 reports the two chimeras affording the highest enrichment factors (BSSB, Fig. 5.3-C and SSBS, Fig. 5.3-D) and emphasizes that the former, while showing slightly higher enrichment factors, displays a clearly worse distribution which is reflected in the reported abundance in the bottom 50%.

Besides a graphical support for the analysis of virtual screening campaigns, the proposed distributions can also be exploited to derive a novel metric, namely the skewness, which quantitatively describes how much the active compounds are focused on the first bins and thus which can be exploited to tackle the so-called "early recognition problem".

Table 5.3 compiles the skewness values for the performed virtual screening campaigns along with the AUC values of the ROC curves computed for easy comparison. Even though the chimeras with the highest enrichment means also show the largest skewness and AUC values, the reported skewness values reveal fair correlations with both the enrichment means ($r = 0.54$) and AUC values ($r = 0.65$), thus emphasizing that they encode for additional information regarding both the complete reliability of a screening study and, more importantly, its ability to focus the active compounds on the best bins.

Although skewness and AUC values show a similar correlation with the enrichment means, Figure 5.4 shows that the correlations of AUC values enhance when the number of the analysed top-ranked compounds is increased reaching their maximum with the Top 20% enrichment factors, while skewness values reveal their best relation with the Top 1% enrichment factor and then the relations roughly decrease with larger top-ranked enrichment factors. Notably, the same difference is noticeable for the correlations with the bottom 50% values. These results confirm the well-known incapacity of ROC curves in early recognition and emphasize the fruitfulness of skewness as a simple metric to conveniently tackle this problem.

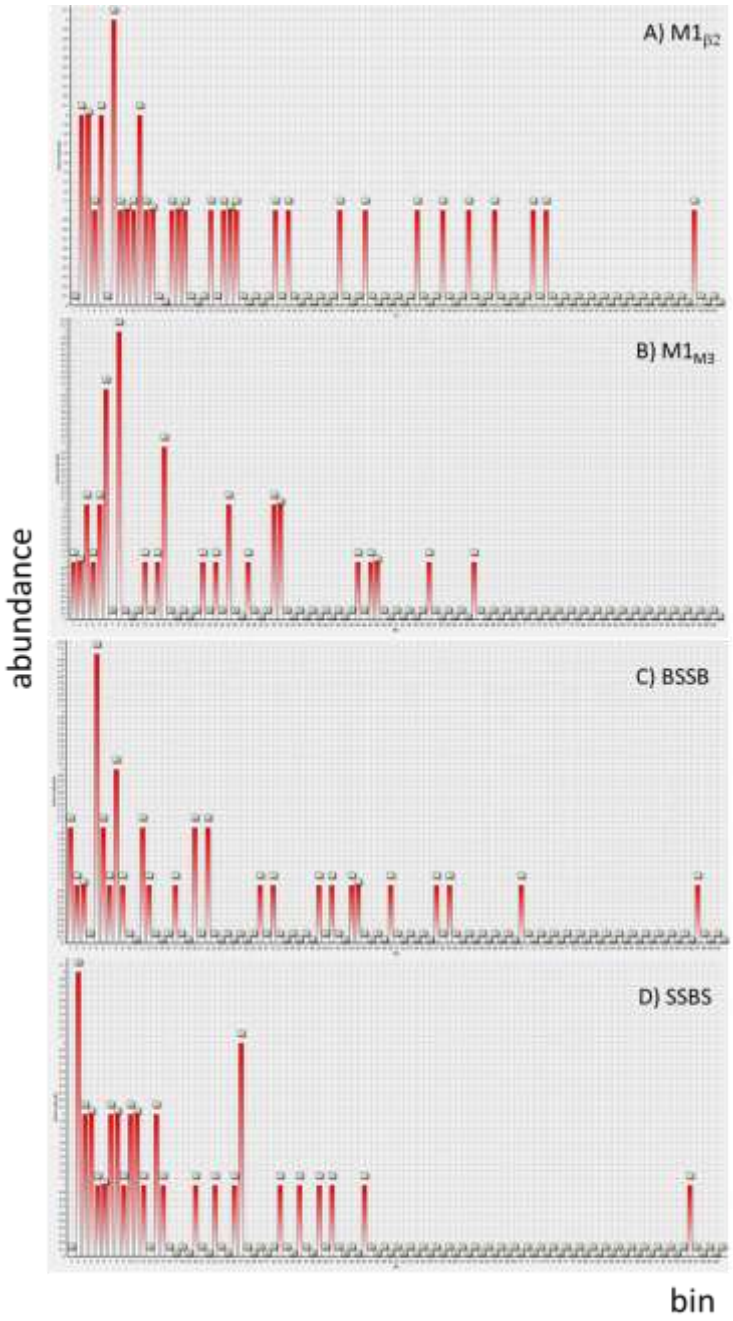


Figure 5.3: Histogram distribution of the active molecules in the overall ranking as computed for the $M1_{M3}$ **(3A)**, $M1_{M3}$ **(3B)**, BSSB **(3C)** and SSBS **(3D)** models.

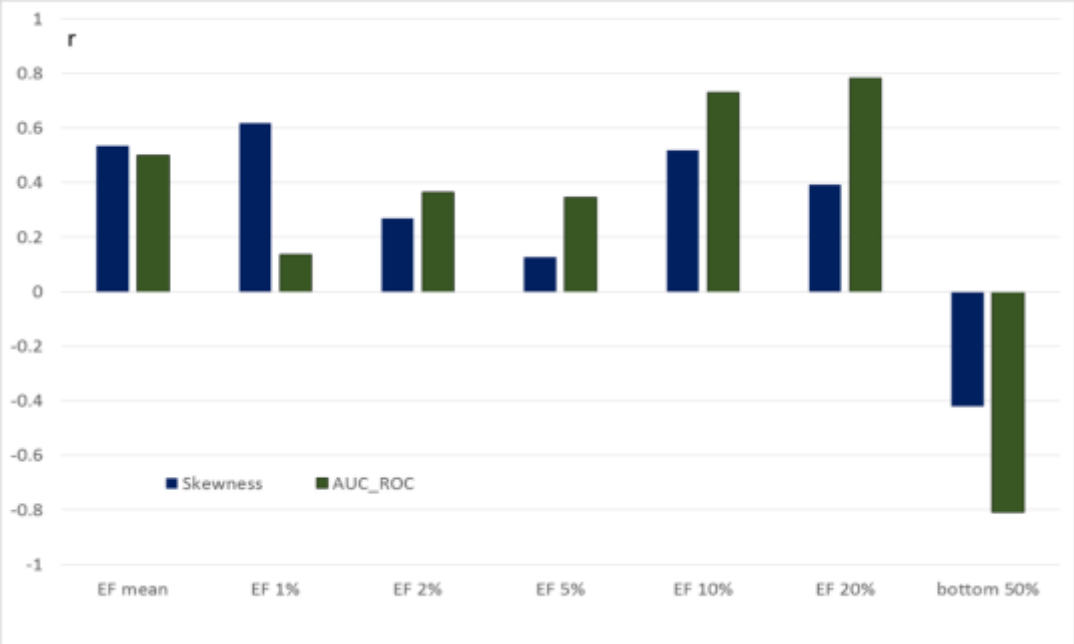


Figure 5.4: Correlations (expressed by their Pearson coefficient) between enrichment factors and skewness values (blue bars) and AUC values (green bars).

5.2.5 Conclusions

As stated above, the primary objective of the study was to investigate the role of the Pro-containing helices in the GPCR conformational behaviour, by assessing the capacity of the simulated chimeras to enhance the predictive power of the corresponding homology models. The obtained virtual screening results emphasize that the chimeras are able to reach and surpass the predictive power of the more reliable $M1_{M3}$ model, thus confirming the fruitful potentialities of such an approach in provisional studies. Nevertheless, the obtained results underline that the chimeras cannot be seen as a way to reproduce the $M1_{M3}$ model, but their marked predictive efficacy can be ascribed to the capacity to represent relevant and distinct conformational states involved in the molecular recognition. On average, the chimeras showing the highest predictive power are endowed with binding pocket of intermediate size, and this is in agreement with experimental lines of evidence showing that the agonists bound GPCR structures appear to be intermediate structures which do not correspond to a fully-active state[187].

The best performing chimeras are not the same in correlative study (as seen in their r^2 values) and virtual screening, but this finding is not very surprising and can find at least two possible explanations. On one hand, correlative analysis and virtual screening pursue different objectives; in the former, docking scores should parameterize the differences among the active ligands, while in the latter they should discriminate between active and inactive compounds. On the other hand, a good correlation can be developed only when the docking scores properly account for all considered compounds, and few badly predicted molecules are enough to pull down the statistics of a given equation. In contrast, virtual screening campaigns do not require that all active compounds are correctly top-ranked and the enrichment factors tolerate better few outliers (as seen in the bottom 50%). This concept is well clarified by chimeras with very restricted cavities (e.g. BBSS),

which properly recognize a set of small active ligands thus providing remarkable enrichment factors, even though their incapacity to harbour also the larger active compounds worsens their r^2 value.

Furthermore, the study describes and applies two new tools included as scripts in the VEGA suite of programs, and which can greatly support the virtual screening studies. The first tool allows both the analysis of the distribution of the active compounds throughout the ranking and the calculation of a parameter that corresponds to the skewness of the distribution, and which accounts for the abundance of the active compounds in the first selected bins thus successfully tackling the early recognition problem. The second and more powerful tool allows the enrichment factors to be maximized by developing fitting equations that linearly combine more docking scores (or ligand-based descriptors). Even though this script should be validated by other tests, the obtained results confirm its relevant potentialities and give a glimpse of the manifold applications it can have.

5.3 Interaction features of 1,4-dioxane agonists at the Orthosteric Sites of Muscarinic Receptors

5.3.1 Setting the scene

Numerous ACh analogues have been synthesized in an attempt to obtain therapeutically better muscarinic agents[188] (see Figure 5.5). Among them, the cis-N,N,N-trimethyl-(2-methyl-1,3-dioxolan-4-yl)methanaminiumiodide[189], compound **1,** emerged as a potent agonist. Esatomic nuclei are also compatible with mAChR activity as recently demonstrated by 1,4-dioxane compounds which are effective muscarinic agonists[190]. Among these compounds, **2a,** the higher homologue of **1**, shows affinity and potency values similar to those of compound **1** (

Table **5.7**).

According to Ing's rule of five, for maximal mAChR agonist activity, there should be no more than five atoms between the quaternary nitrogen atom and the terminal hydrogen of the acetyl mimicking chain. Moreover, it has been demonstrated that acetic acid esters of quaternary ammonium alcohols of greater length than choline have decreased mAChR activity, indicating that no more than two carbon atoms are tolerated between the nitrogen and the ether oxygen atoms[191].

The methyl group in position 2 of the 1,3-dioxolane nucleus of compound **1** has been demonstrated to be essential for the activation of the muscarinic receptors of the guinea pig ileum, since compound **13**, its desmethyl analogue, is practically inactive in this tissue[192]. Indeed, only in the case of the 1,3-dioxolane 1 Ing's rule is respected. Instead, compound **14**, the desmethyl analogue of **2a**, has been reported to be a potent muscarinic agonist[193], but so far a complete pharmacological study at all mAChR subtypes has not been performed yet. Therefore, to verify whether the

methyl group of 1,4-dioxane **2a** is essential for its potent mAChR activity and to understand the mode of interaction of 2a with the muscarinic receptors, compound **14** was re-prepared and pharmacologically characterized.

In a structure-activity relationship (SAR) study, in which the methyl group of **2a** has alternatively or simultaneously been inserted in positions 5 and 6 of the 1,4-dioxane nucleus in all combinations (compounds 3–12, Figure 5.5)[194], we demonstrated that the presence of one methyl group in both positions 5 and 6 with a trans stereochemical relationship with each other (diastereomers 4 and 5) or the geminal dimethylation in position 6 (compound **8**) favoured the selective activation of the M3 receptor subtype[194]. To extend such a SAR study the methyl group in position 6 of the 1,4-dioxane nucleus of **2a** was moved to positions 2 and 3 (compounds **15** and **16a**,**b**, respectively).

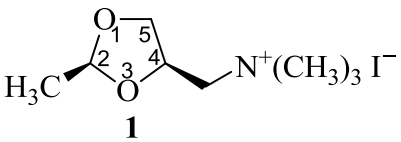
To get information about the recognition process of the 1,4-dioxane agonists, a retrospective computational study was performed involving the known derivatives **2**-**14** and the novel derivatives **15** and **16** as compiled in Figure 5.5 and

Table **5.7**. Considering that the observed affinities can be due to the ligand capacity to assume an optimal bioactive conformation as well as to elicit the key interactions, the study involved two different approaches.

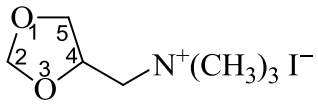
In the first, a ligand-based correlative study was performed to reveal how the substituents can influence the puckering of the 1,4-dioxane ring and whether the so evidenced conformational differences can, in turn, affect the ligand binding to the mAChRs.

To investigate the preferred binding modes for the examined 1,4-dioxanes, a second part of the study involved docking simulations on the recently resolved M_2 and M_3 receptors[195,160], which should allow the diverse affinity profiles to be rationalized.

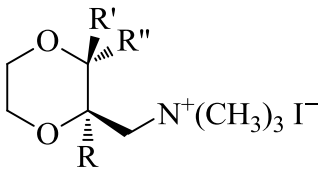
Moreover, docking analyses involved the M_2 receptor in both open and closed states to reveal how the ligand binding is influenced by receptor activation.



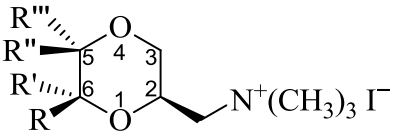
1



13



14: R = R' = R" = H
15: R = CH_3, R' = R" = H
16a: R = H, R' = CH_3, R" = H
16b: R = R' = H, R" = CH_3



2a, R = CH_3, R' = R" = R"' = H
2b, R = R" = R"' = H, R' = CH_3
3a, R = R' = R"' = H, R" = CH_3
3b, R = R' = R" = H, R"' = CH_3
4, R = R"' = CH_3, R' = R" = H
5, R = R"' = H, R' = R" = CH_3
6, R = R" = H, R' = R"' = CH_3
7, R = R" = CH_3, R' = R"' = H
8, R = R' = CH_3, R" = R"' = H
9, R = R' = H, R" = R"' = CH_3
10a, R = R" = R"' = CH_3, R' = H
10b, R = H, R' = R" = R"' = CH_3
11a, R = R' = R" = CH_3, R"' = H
11b, R = R' = R"' = CH_3, R" = H
12, R = R' = R" = R"' = CH_3

Figure 5.5: Chemical structures of compounds 1-16.

Comp.	M_2	M_3	Δ Energy	Dipole Twist chair	Dipole Chair	Dipole average	Δ Dipole	Log P_{MLP} Chair	AD Electr Chair M_2 open	AD Electr Chair M_3	AD VdW Twist M_2 closed
2a	5.80	5.12	2.67	6.80	8.05	7.43	1.25	-1.62	-0.66	-0.66	-6.54
2b	5.31	4.45	1.81	7.17	7.17	7.17	0.00	-2.03	-0.53	-0.65	-6.24
3a	4.95	4.18	1.57	7.95	7.08	7.52	0.86	-2.10	-0.59	-0.68	-6.50
3b	4.76	4.35	1.49	7.62	7.69	7.65	0.07	-1.59	-0.51	-0.62	-6.46
4	4.86	4.96	2.57	8.24	8.63	8.43	0.39	-1.56	-0.53	-0.5	-7.31
5	4.78	<4	2.30	8.61	8.24	8.43	0.37	-1.61	-0.52	-0.62	-7.31
6	4.05	<4	2.76	8.29	8.60	8.45	0.31	-1.56	-0.51	-0.54	-7.26
7	4.70	<4	2.46	8.28	8.38	8.33	0.10	-1.56	-0.62	-0.66	-7.14
8	4.74	4.67	2.88	7.59	7.88	7.73	0.29	-0.73	-0.60	-0.7	-7.02
9	4.02	4.11	1.45	9.31	9.21	9.26	0.10	-0.01	-0.52	-0.67	-6.87
10a	4.22	4.49	1.75	9.50	9.66	9.58	0.16	-0.08	-0.51	-0.56	-7.53
10b	4.14	4.33	1.78	9.53	9.58	9.56	0.04	-0.07	-0.49	-0.61	-7.63
11a	4.56	4.79	2.47	8.77	8.87	8.82	0.10	-0.28	-0.58	-0.66	-7.94
11b	4.47	4.98	3.15	8.45	9.17	8.81	0.72	-0.26	-0.57	-0.56	-7.65
12	4.33	4.39	2.94	9.58	9.95	9.76	0.37	1.22	-0.59	-0.58	-7.96
14	5.76	4.57	2.43	6.18	6.21	6.21	0.03	-2.10	-0.60	-0.49	-6.67
15	<4	<4	2.25	6.49	6.84	6.84	0.35	-1.65	-0.59	-0.42	-6.42
16a	<4	<4	2.64	6.96	6.97	6.97	0.01	-1.56	-0.64	-0.44	-6.55
16b	<4	<4	2.55	6.85	6.15	6.50	0.70	-0.92	-0.56	-0.35	-6.21

Table 5.7: Compounds, affinity values and descriptors used in the correlative analyses. ΔEnergy and the AutoDock (AD) scores are expressed in Kcal/mol, while dipole moments are expressed in Debye

5.3.2 Computational methods

The simulated compounds were built by the VEGA suite of programs manually generating the favoured conformations (i.e. chair and twist-boat geometries) which were then optimized by PM7 semi-empirical calculations as implemented in MOPAC2012. The so minimized structures were used to compute the conformational energies and the dipole moments as compiled in Table 5.7 and were utilized in the following docking simulations.

For each ligand, docking calculations involved separately chair and twist-boat geometries and were performed by using the recently resolved M_2 in open and closed states and M_3 structures (PDB Id: 3UON, 4MQS and 4DAJ, respectively). After removing the bound antagonist, the two structures were completed by adding hydrogen atoms, and the side-chains of Arg, Lys, Glu, and Asp were ionized to remain compatible with physiological pH values, while His and Cys residues were considered neutral by default. The structures so obtained were minimized, keeping the backbone fixed to preserve the experimental folding. The docking search was focused in a 12 Å radius sphere around the co-crystallized ligand. In detail, the resolution of the grid was 60×60×60 points with a grid spacing of 0.450 Å and each ligand was docked into this grid by the Lamarckian algorithm as implemented in AutoDock. The genetic-based algorithm ran 20 simulations per substrate with 2000000 energy evaluations and a maximum number of generations of 27000. The crossover rate was increased to 0.8, and the number of individuals in each population to 150. All other parameters were left at the AutoDock default settings[196]. The selected complexes were finally minimized keeping fixed all atoms outside a 12 Å radius sphere around the docked ligand and then used to re-calculated the AutoDock scores.

5.3.3 Results

5.3.3.1 Ligand-based analyses: the role of ring conformations

The conformational analysis of 1,4 dioxane ring showed that the more extended chair conformation is the preferred one for all considered derivatives even though the energy barrier separating the chair and the 2,5 twist-boat geometries can significantly vary depending on the arrangement of the methyl groups as compiled in

Table **5.7**. Briefly, a bird's eye view of these energy gaps suggests that the methyl groups in position 5 reduce the energy barrier below 2.0 Kcal/mol, whereas the methyl groups in position 6 increase this energy barrier which shows its maximum values (around 3.0 Kcal) for the trimethyl analogue **11b** and tetramethyl analogue **12**. Also for all novel derivatives, the chair geometry is the favored one with an energy barrier (about 2.5 Kcal/mol) in line with the previous data. This may suggest that the poor affinity of methyl derivatives **15** and **16a,b** cannot be ascribed to distorted ring puckering, but it should be due to their incapacity to stabilize the required key contacts within the receptor binding sites because of the methyl hindrance (as examined below).

With a view to investigating the influence of ring conformations on receptor binding, a correlative study involved all 1,4-dioxane derivatives by considering the above described energy gaps as well as some relevant physicochemical properties as computed for both chair and twist-boat geometries and the relative averages and differences (as compiled in

Table **5.7**). In analogy to docking simulations (see below), the correlative study was focused on M_2 and M_3 mAChRs only.

As a preamble, it should be noted that there is no correlation between energy gaps and ligand affinities for the M_2 subtype, whereas there is a significant direct correlation between them for the M_3 subtype ($r^2 = 0.55$). Such a difference may suggest that the M_2 receptor can conveniently recognize both geometries, while the M_3 receptor recognizes preferentially (or almost exclusively, as suggested by docking results) the chair conformations and thus the affinity is linearly correlated to the relative stability of these latter. Such a hypothesis is in line with previous computational studies emphasizing the greater flexibility of the M_2 binding site[197,198].

This hypothesis is further corroborated by the correlative study since Table 5.8 shows that the best predictive equation derived for the M_2 subtype (Eq. 1) involves the dipole averages and the corresponding difference as independent variables, while similar equations obtained using physicochemical properties as computed for chair or twist-boat conformations separately performed worst. On the other hand, Table 5.8 shows that the best relationship for the M_3 subtype is derived by using the already discussed energy gaps plus the virtual $\log P_{MLP}$ as computed for the chair geometries (Eq. 2), while the average values or the twist-boat properties afforded worst results. Overall, these results may further confirm that both ring conformations are recognized by the M_2 binding site and thus the average values best encode the ligand properties, while the M_3 subtype strongly prefers the energetically favoured chair conformations. Based on these preliminary correlative results, the following docking simulations involved the M_2 and M_3 subtypes considering both chair and twist-boat geometries.

Eq.	Equation	Statistics
1	pK_{M2} = 8.38 - 0.35 Dipole Average + 0.45 ΔDipole	n = 16; **r^2 = 0.78**; q^2 = 0.59; SE = 0.27; F = 23.43; p < 0.001
2	pK_{M3} = 344 + 0.45 ΔEnergy - 013 LogP_MLP_Chair	n = 13; **r^2 = 0.71**; q^2 = 0.55; SE = 0.19; F = 12.37; p < 0.01
3	pK_{M2} = 6.25 - 0.41 Dipole Average - 3.28 Electr Chair Open	n = 16; **r^2 = 0.81**; q^2 = 0.68; SE = 0.25; F = 26.95; p < 0.001
4	pK_{M2} = 8.34 - 0.55 Dipole Average – 0.14 vdW/HB Twist Close	n = 16; **r^2 = 0.74**; q^2 = 0.64; SE = 0.29; F = 19.31; p < 0.001
5	pK_{M3} = 3.87 + 0.41 ΔEnergy - 0.36 Electr Chair	n = 13; **r^2 = 0.60**; q^2 = 0.47; SE = 0.23; F = 10.56; p < 0.01

Table 5.8: Best relationships developed in the study

5.3.3.2 Docking studies on M2 receptor: comparison between open and closed states

Regarding to the M_2 subtype in its open inactive state, Figure 5.6 compares the key interactions stabilized by **2a** in its chair (Figure 5.6-A) and twist-boat (Figure 5.6-B) geometry. Remarkably, the key polar contacts are conserved in both complexes and involve: (1) ion-pair between the ligand ammonium head and Asp103; (2) charge transfer interactions involving the ammonium head and a set of surrounding aromatic residues (Trp400, Tyr403, Tyr426, Tyr430, plus Tyr104 not shown for

reasons of clarity); (3) H-bonds between O1 and Ser107 as well as between O4 and Asn404.

The two complexes mainly differ in the accessibility of the ligand oxygen atoms, which influences the stability of the mentioned H-bonds, the chair conformation allowing a closer approach of Ser107 and Asn404.

A second evident difference between the complexes involves the arrangement of 6-methyl group which better approaches Ala194 when the dioxane ring assumes twist-boat conformation, even though the polar contacts appear to be here largely predominant as suggested by the following docking based correlative studies. Indeed, the best obtained equation (Eq. 3, Table 5.8) combines the dipole average (already used by Eq. 1) with electrostatic score as computed by AutoDock for the complexes with dioxane in chair conformation. This emphasizes that the chair conformations and the polar interactions play a key role in ligand recognition, even though the M_2 binding site might accommodate both ring geometries.

While the M_2 receptor in its open state can suitably recognize both dioxane conformations but prefers the chair geometry (see above), docking simulations suggest that the closed active M_2 state stabilizes significantly better complexes when interacting with the ligand in its twist-chair conformation, a results that can be ascribed to the markedly narrower binding cavity of the M_2 active state which accommodates more conveniently the less hindered twist-chair geometry.

Figure 5.6 compares the putative complexes for the derivative **2a** in its chair (Figure 5.6-C) and twist-chair geometry (Figure 5.6-D) and evidence in both complexes a set of interactions almost superimposable to that observed in the M_2 open state. Nevertheless, Figure 5.6-C and 5.6-D reveal a significant difference between the two complexes since only the ligand assuming twist-chair geometry is able to elicit the key H-bond with Asn404, while the clashes exerted by the methyl group prevent a proper approaching of Asn404 when adopting the more hindered chair geometry.

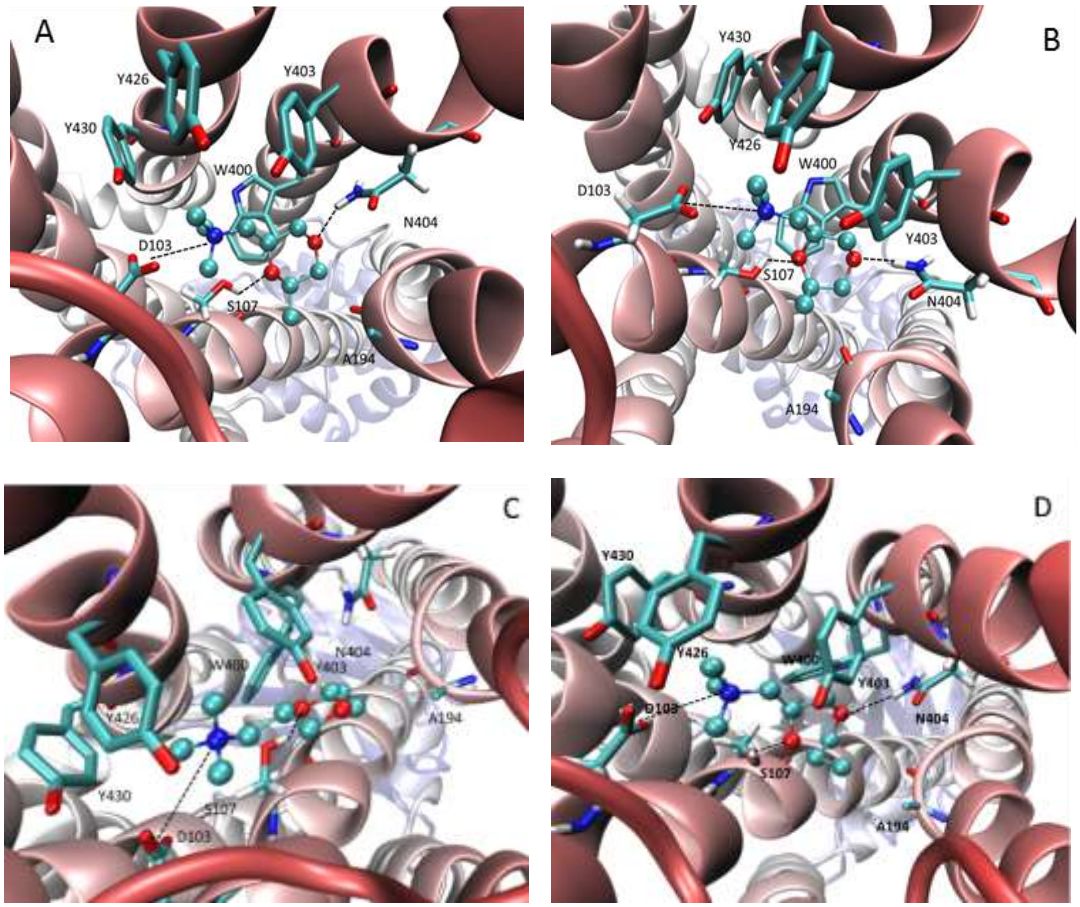


Figure 5.6: Main interactions stabilizing the compound 2a in its chair **(A)** and twist-boat **(B)** conformation within the M_2 binding site in its open state as well as in its chair **(C)** and twist-boat **(D)** conformation within the M_2 binding site in its closed state

The role of twist-chair geometries in ligand recognition by the M_2 closed state is corroborated by Eq. 4 (Table 5.8) which, despite showing worst statistics than Eq. 3, allows some considerations to be made.

First, Eq. 4 includes a score obtained by docking ligands in their twist-chair geometry, while docking scores derived by chair conformations did not afford satisfactory results.

Second, the comparison of Eq. 3 and Eq. 4 confirms the role of dipole moment and evidences a notable difference since Eq. 4 includes a score encoding for H-bonds

and apolar contacts instead of an electrostatic score function as seen in Eq. 3. This may mean that well-defined ionic interactions are required for triggering receptor activation, while the binding to closed state is rather influenced by a precise fitting of the ligand within the narrower binding site stabilized by apolar contacts.

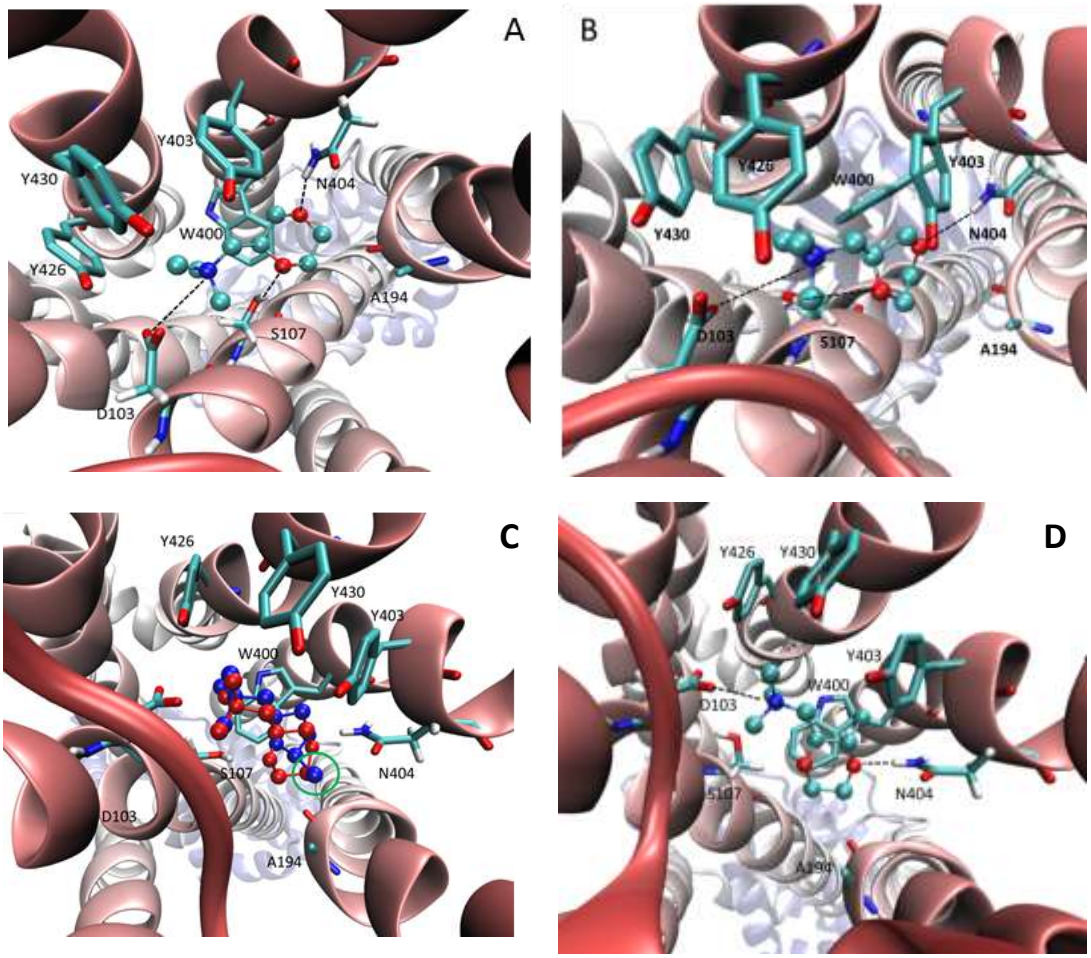


Figure 5.7: Main interactions stabilizing 14 in its chair conformation within the M_2 binding site in its open state **(A)** as well as in its twist-chair conformation within the M_2 binding site in its closed state **(B)**. Comparison of the docked poses for compounds 1 (in blue) and 14 (in red) within the M_2 binding site in its open state **(C)** emphasizing (see the green circle) the remarkable overlapping between the methyl group of 1 and the 5-methylene group of 14. Main interactions stabilizing 16b in its chair conformation within the M2 binding site in its open state **(D)**.

Concerning the here reported novel derivatives, Figure 5.7-A shows the complex as obtained for the most affine derivative **14** in its chair conformation within the M_2 binding site in its open state and evidences a set of key interactions in line with those reported in Figure 5.6-A. Similarly, Figure 5.7-B reports **14** in twist-chair conformation within the M_2 binding site in the closed state and shows interactions almost identical to those displayed by Figure 5.6-D.

In both cases, the major difference involves the arrangement of the dioxane ring because the absence of the methyl group allows the ring to be accommodated more superficially in a pose where it can maintain the crucial H-bonds with Ser107 and Asn404 and its carbon skeleton can conveniently approach Ala194, thus replacing the hydrophobic contacts stabilized by methyl groups in the previous analogues. When focusing the attention on the open state, the capacity of the dioxane ring of **14** to replace the methyl substituents eliciting hydrophobic interactions with Ala194 and the surrounding residues is confirmed by the superimposition of the poses obtained for **1** and **14** (see Figure 5.7-C for M_2 in the open state, **1** in blue and **14** in red) and can explain the ability of **14** to obey Ing's rule. Figure 5.7-C shows indeed the precise overlapping between the methyl substituent of **1** and the 5-methylene group of the dioxane ring of **14**, as evidenced by the green circle, thus confirming that the unsubstituted dioxane ring can elicit hydrophobic interactions similar to those afforded by the methyl dioxolanes regardless of ring conformation and receptor activation (as seen for **1**).

As illustrated in Figure 5.7-D for **16b**, docking results can also explain the negligible affinity of methyl derivatives **15** and **16a**,**b**, since even in the M_2 open state the methyl group bumps against Trp400 plus Tyr403 which constrain the ligands in a more lateral pose where they lose some key contacts such as that with Ser107, thus justifying their lack of affinity.

5.3.3.3 Docking studies on the M_3 receptor in its open state

When docking both ligand geometries, completely different results are obtained for the M_3 subtype which is considered in its open state (the only conformation hitherto resolved). Indeed, while chair and twist-boat conformations are stably retained within the M_2 binding site even after complex minimization, the minimization of most twist-boat conformations within the M_3 binding site leads to their conversion to chair geometries, thus indicating that the twist-boat conformations cannot be involved in ligand recognition by the M_3 subtype. This difference can be rationalized by analyzing the minimized complex for the derivative **2a** in its chair geometry as shown in Figure 5.8-A. The key stabilizing interactions involve: (1) ion-pair between the ligand ammonium head and Asp147; (2) charge transfer interactions involving the ligand ammonium head and a set of surrounding aromatic residues (Trp503, Tyr506, Tyr529, Tyr533, plus Tyr148 not shown for reasons of clarity); (3) H-bonds between O1 and Ser151 as well as O4 and Asn507.

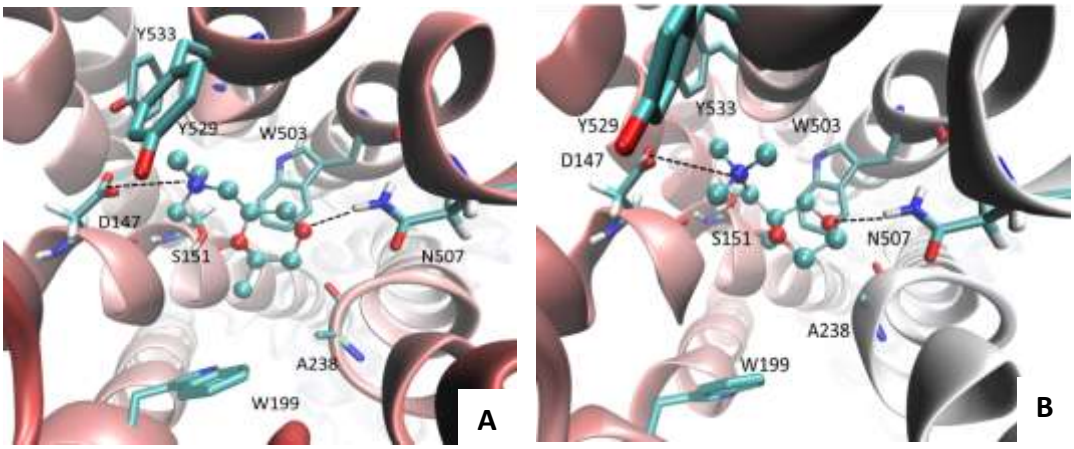


Figure 5.8: Main interactions stabilizing the compound 2a **(A)** and 15 **(B)** in their allowed chair conformations within the M_3 binding site.

The comparison of this complex with the corresponding one for the M_2 subtype reveals some remarkable differences, which can be summarized as follows: (1) the aromatic residues mostly contact the ammonium head in the M_2 subtype while they completely surround the ligand in the M_3 subtype, thus rendering the binding site more constrained and less flexible; (2) the H-bond between the conserved Ser151 and the O1 atom appears to be weakened when compared to the corresponding contact within the M_2 binding cavity; on one hand this underlines the greater relevance of the other H-bond involving the conserved Asn507 residue; on the other hand, this suggests that the surrounding tyrosine residues can be involved in additional H-bonds with the dioxane oxygen atoms but they can be conveniently approached only by the more accessible chair geometry; (3) similarly to what was observed for the M_2 subtype, the methyl groups contact the conserved Ala238, but here these hydrophobic interactions are markedly influenced by the closeness of Trp199 which constrains this hydrophobic subpocket which can be suitably contacted only by the more extended chair geometries; notably, these steric requirements influence ligand conformation and receptor recognition regardless of position and configuration of the methyl groups. Indeed, only the *trans* diastereoisomers **2b** and **3b** are able to retain a distorted twist-boat conformation within the M_3 binding cavity.

The following correlative study confirms the key role of chair conformations since the best derived equation (Eq. 4, Table 5.8) includes the energy gaps between ring geometries and the electrostatic scores as computed by AutoDock. This result further underlines how the M_3 affinity depends on the relative stability of the chair conformations and their ability to stabilize polar contacts within the M_3 binding site.

Docking simulations on the M_3 subtype afford results in line with the previous ones and, as evidenced by the retrospective study, only chair conformations can be accommodated within its binding site. In detail, docking results show that the

methyl derivatives **15** and **16a**,**b** are unable to be conveniently harboured due to the methyl group which bumps against Trp503 and Ser151, thus hampering the H-bond stabilized by this latter with the ligand O1 atom (as exemplified by Figure 5.8-B for **15**, Supporting Information).

Similarly to what was observed for the M_2 binding site, Figure 5.9 compares the computed poses of compounds **1** (red) and **14** (blue) within the binding cavity of M_3 and shows an interaction pattern which is very similar to that observed for M_2. One may note the remarkable overlapping between the methyl substituent of **1** and the 5-methylene group of the dioxane ring of **14**, as evidenced by the green circle, and indeed both ligands stabilize apolar contacts with Ala235 and Ala238. This observation confirms that the dioxane ring can elicit hydrophobic interactions similar to those stabilized by methyl dioxolanes (as seen for **1**), thus explaining why dioxane derivatives show similar affinity values regardless of the methyl substituent.

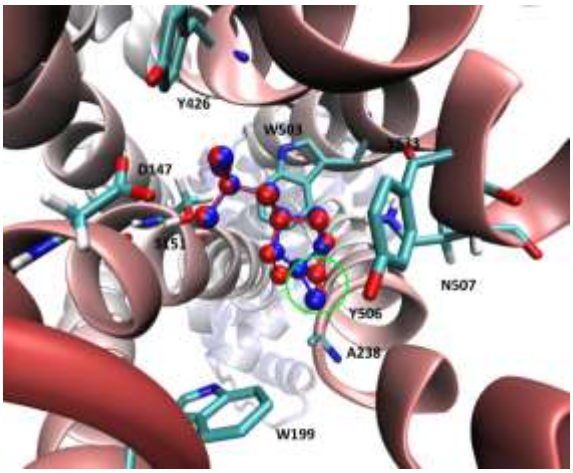


Figure 5.9: Comparison of the docked poses for compounds 1 (in blue) and 14 (in red) within the M_3 binding site emphasizing (see the green circle) the remarkable overlapping between the methyl group of 1 and the 5-methylene group of 14.

5.3.4 Conclusions

The present study completes a SAR study of 1,4-dioxane muscarinic agonists, and investigates their binding capacities by docking simulations as performed using the recently resolved human M_2 and rat M_3 crystal structures. On one hand, it should be observed that, to the best of our knowledge, this is one of the first studies which exploit such resolved GPCR structures to rationalize the affinity values of muscarinic ligands. On the other hand, one may imagine that the availability of resolved receptor structures can afford more reliable docking results, accounting also for very specific ligand features as here exemplified by the puckering of the dioxane ring.

Indeed, the obtained docking results were able to reveal the key role played by often disregarded conformational profile of the 1,4-dioxane ring, thus underlining the already reported flexibility differences between the orthosteric binding sites of the M_2 and M_3 subtypes. Notably, the conformational profile of the 1,4-dioxane ring was found to play a key role also to second receptor activation and to optimize the ligand interactions as seen in the M_2 closed state. Moreover, the so obtained results evidence that the 1,4-dioxane compounds interact with muscarinic receptors in their orthosteric site similarly to their 1,3-dioxolane analogues. In the case of compound **14**, the methylene group in position 5 of the 1,4-dioxane nucleus occupies the same lipophilic pocket as the methyl group of the potent muscarinic compounds **1** and **2a**, thus explaining why **14** has a good affinity and obeys Ing's rule.

While considering the evidenced differences between the binding modes of the M_2 and M_3 subtypes, docking results overall emphasize the remarkable similarity between the two binding pockets, thus justifying the difficulty of designing selective agonists as indeed confirmed by the here analysed ligands. Nevertheless, one may argue that the possibility of utilizing the resolved muscarinic structures

might provide ever deeper insights allowing the design of new agonists selectively targeting M_2 and M_3 mAChR subtypes.

5.4 Bitopic modulators of muscarinic receptors: a modelling study

5.4.1 Setting the scene

As mentioned in the Introduction, the main problem in the development of muscarinic agents is the high homology in their orthosteric binding site. This has hampered the development of drugs whose activity was associated to many adverse effects to have a concrete therapeutic role, arising from lack of selectivity[199,200]. The discovery of allosteric modulators has opened a new avenue in the development of selective drugs, since these compounds interact with binding sites which are less conserved within the five muscarinic receptors, thus allowing a selective modulation of only one subtypelindse[201,202].

New intriguing perspectives have been associated also to the discovery of dualsteric (bitopic) ligands, i.e. divalent compounds in which two pharmacophoric units, connected by a suitable spacer, are able to interact at the same time both with the orthosteric site and the allosteric binding areas, thus exploiting the favourable characteristics of both sites[203,204,205,206]. Originally divalent ligands have been designed to study receptor dimerization[206], but in many instances the length of the linker allowed only the bridging of two neighboring interaction sites on the same protein. In principle, bitopic ligands may interact with the othosteric or the allosteric binding site, or with both, within a monomeric receptor; very recently it has been shown a bitopic interaction also within a dimeric receptor[207].

Several muscarinic homodivalent ligands, i.e. divalent compounds carrying two identical pharmacophoric units, have also been disclosed, such as the M2 selective

antagonist methoctramine[208,209] or the dimers of agonists such as xanomeline[210] or arecaidine propargyl ester (APE)[211]. These compounds displayed different range of potency, affinity and intrinsic activity, depending on the pharmacophoric structure and on the linker, as the pharmacophoric doubling not always resulted in an increased affinity or potency. In addition, homo or heterobivalent ligands carrying at least one agonist unit were not always endowed with receptor activation properties (see, for instance, refs[212,213])

On these bases the effect of homodimerization on carbachol, the well-known cholinergic agonist was here investigated. The idea underlying this approach concerned the possibility to bind two different sites in the same receptor, either orthosteric and allosteric, or two orthosteric sites in a dimeric receptor. Although carbachol itself does not display allosteric properties, the allosteric site of muscarinic receptors can bind compounds carrying choline residues as exemplified by gallamine. Therefore, a series of compounds where the two agonist units are symmetrically connected through a methylene chain of variable length were designed, linking the carbamic nitrogen atoms and considering both tertiary amines and ammonium derivatives. Ongoing biological analyses revealed that the tested compounds are able to occupy both orthosteric and allosteric binding sites, showing affinity values which increase with the linker length. Notably, all compounds lose the agonistic activity and show an antagonistic profile in all tested receptors. Again, the compounds, while remaining substantially unselective, show a clear preference for the M1/M3 subtypes, and this feature is completely different compared to the parent compound carbachol which, albeit unselective, is characterized by a known preference for M2/M4 subtypes.

5.4.3 Computational details

Docking simulations involved the recently resolved structures of the hM_2 subtype in complex with both the agonist iperoxo and the allosteric modulator LY2119620 (PDB Id: 4MQT), as well as in complex with the antagonists QNB (PDB Id: 3UON) The choice of the first hM_2 structure is justified by the bound ligands which should assure that both binding cavities (allosteric and orthosteric sites) are finely optimized for ligand recognition. The second hM_2 structure was selected to simulate receptor conditions similar to those experienced during the kinetic experiments.

Again, docking studies involved the hM_1 (Entry Id: P11229, Entry name: ACM1_HUMAN) homology model as generated by using the first hM_2 structure as the template. Briefly, the homology modelling was performed by Modeller 9.10 using the default parameters; among the 20 generated models, the best structure was selected according to the computed scores (i.e. DOPE and GA341), as well as to the percentage of residues falling in the allowed regions of the Ramachandran (91.2 %) and chi plots (95.8%). The completed model was carefully checked to avoid unphysical occurrences such as cis peptide bonds, wrong configurations, improper bond lengths, non-planar aromatic rings or colliding side-chains.

The so obtained hM_1 and hM_2 structures were then completed by adding hydrogen atoms and to remain compatible with physiological pH, Asp, Glu, Lys and Arg residues were considered in their ionized form while His and Cys were maintained neutral by default. Finally, the structures were optimized by a minimization made up by two phases: a first minimization without constraints until RMS = 0.1 kcal mol-1Å-1 and then a second minimization with backbone fixed until RMS = 0.01 kcal mol-1Å-1 to preserve their folding.

All ligands were simulated in their protonated state, since this is involved in receptor recognition. The conformational profile was investigated by MonteCarlo

simulations (as implemented in the VEGA program), which produced 1000 minimized conformations by randomly rotating the rotatable bonds, and the so computed lowest energy structure underwent docking simulations.

Docking simulations were carried out using PLANTS with default settings and without geometric constraints. The search within the first hM_2 structure and the hM_1 homology model was focused on a region obtained combining a 10.0 Å radius sphere around Asp105 (hM_1) or Asp103 (hM_2) plus a 10.0 Å radius sphere around Tyr179 (hM_1) or Tyr177 (hM_2), thus completely encompassing both binding cavities. In contrast, the search within the second hM_2 structure was focused on a region obtained combining a 10.0 Å radius sphere around Tyr177, keeping the bound QNB antagonist. For each ligand, speed 1 was used and 10 poses were generated and scored using the ChemPlp function. The so obtained best complexes were finally optimized by a minimization keeping all atoms fixed apart from those included within a 10.0 Å radius sphere around the bound ligand.

5.4.3 Results

5.4.3.1 Docking results on hM_2 in its closed state

To rationalize the results of binding and functional experiments, docking studies were performed on the recently resolved hM_2 structure in complex with the agonist iperoxo and the allosteric modulator LY2119620 , for both ammounium derivatives and ammines in the protonated form; since both classes interact in a similar way, only the methiodides are discussed here.

In detail, the short-chain derivatives (n = 3 or 5) can assume two distinct binding modes. In the first pose (as shown in Figure 5.10-A, n=3), the ligands are completely accommodated in the orthosteric cavity, where they stabilize the following set of interactions: (i) one ammonium head is engaged in the key ion-pairing with Asp103, reinforced by a set of charge transfer interactions with surrounding aromatic residues (e.g., Tyr80, Trp99, Tyr426, Tyr430); (ii) the two carbamate moieties elicit clear H-bonds with Tyr104, Ser107 and Tyr403 and (iii) the second ammonium head approaches Asn108, Phe195, Trp400, and Asn404.

In the second binding mode, the ligands assume a central pose by which they contact both the orthosteric and the allosteric sites. Specifically, both ammonium heads are engaged in ion-pairs, one with Asp103 in the orthosteric cavity and the other with Glu172 and Glu175 in the allosteric site, the latter being further reinforced by charge transfer interactions with Tyr177. The two carbamate moieties elicit again H-bond interactions, but with Tyr426 and Asn419 in the orthosteric and the allosteric sites, respectively. For the short-chain derivatives, the first binding mode is markedly favoured in terms of both calculated docking scores and relative abundance among the computed poses.

When extending the linker, the first binding mode becomes progressively less favoured to disappear for n = 6 due to obvious steric hindrance. As exemplified in

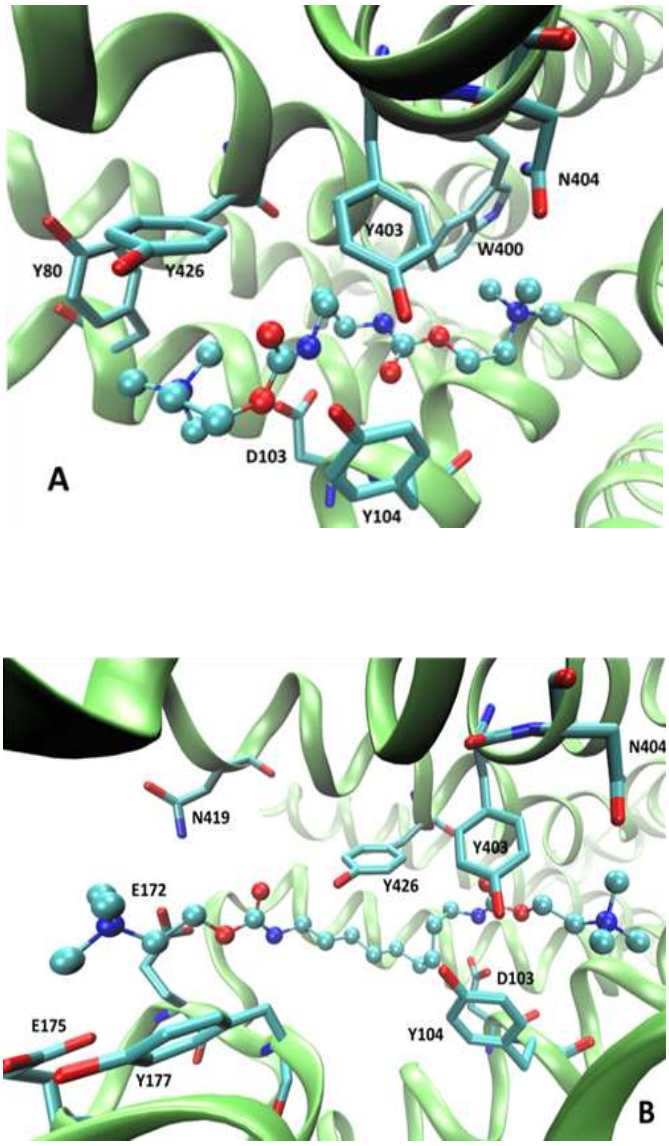


Figure 5.10: Docking poses of selected ligands into the hM_1 and hM_2 receptors. A: n =3 within the orthosteric cavity of hM_2. B: n = 9 docked between the orthosteric and allosteric sites of hM_2.

Figure 5.10-B, the long-chain analogues show only the second binding mode and assume a deeper pose compared to that shown by short-chain derivatives, by which the dimethylaminoethylcarbamate moiety within the orthosteric cavity contacts Asn404, while retaining the already mentioned interactions with Asp103 plus the surrounding aromatic residues. The type of interactions established by the second carbachol unit, which are accommodated in the allosteric site, are similar to those already seen for the short-chain analogues, consisting in ion pairing between the ammonium moiety and Glu172 and Glu175, reinforced by a set of H-bonds with Tyr83, Tyr177, Asn419. However, it is worth noting that all derivatives, when assuming this second binding mode, are unable to insert the carbachol unit into the orthosteric cavity in a pose comparable to that shown by CCh. As a matter of fact, the linker constrains the dimethylaminoethylcarbamate moiety in an inverted arrangement which prevents the carbamate group to contact Asn404. In such a second binding mode the linker is inserted in a constrained channel lined by aromatic residues.

Taken together, these results can provide an explanation for the antagonistic activity of the long-chain derivatives, since they do achieve a rich interaction pattern while being unable to elicit the key contacts usually established by the agonists within the orthosteric site.

5.4.3.2 Docking results on the hM_1 homology model

Since the compounds show higher affinity for the hM_1 receptor, compared to hM_2 subtype, docking simulations were also performed on an hM_1 homology model as generated by using the above mentioned recently resolved hM_2 structure as the template. Also on this subtype, docking studies suggest significant differences on the binding mode of the compounds, depending on the length of the linker.

In detail, the short-chain derivatives tend to remain in the allosteric site where one carbachol moiety elicits ionic interactions with Glu397 plus a set of H-bonds with Tyr85, Tyr179, Gln177, while the second carbachol group approaches the orthosteric site without reaching it. On the contrary, as shown in Figure 5.11-A, the long-chain derivatives assume a central pose by which they occupy both the orthosteric and the allosteric sites. Such a pose brings to mind that previously described for the hM_2 subtype, although these ligands appear to be more fittingly accommodated in the hM_1 receptor. In fact, while the key contacts within the allosteric cavity are similar in both subtypes, the carbachol moiety within the orthosteric site assumes an optimal orientation being able to elicit the same key interactions established by CCh, namely the ammonium head with Asp105 plus the surrounding aromatic residues and the carbamate moiety with Asn382. Moreover, the alkyl linker appears to be suitably inserted in a tight channel completely lined by aliphatic residues, which can form a rich set of apolar contacts with the ligand.

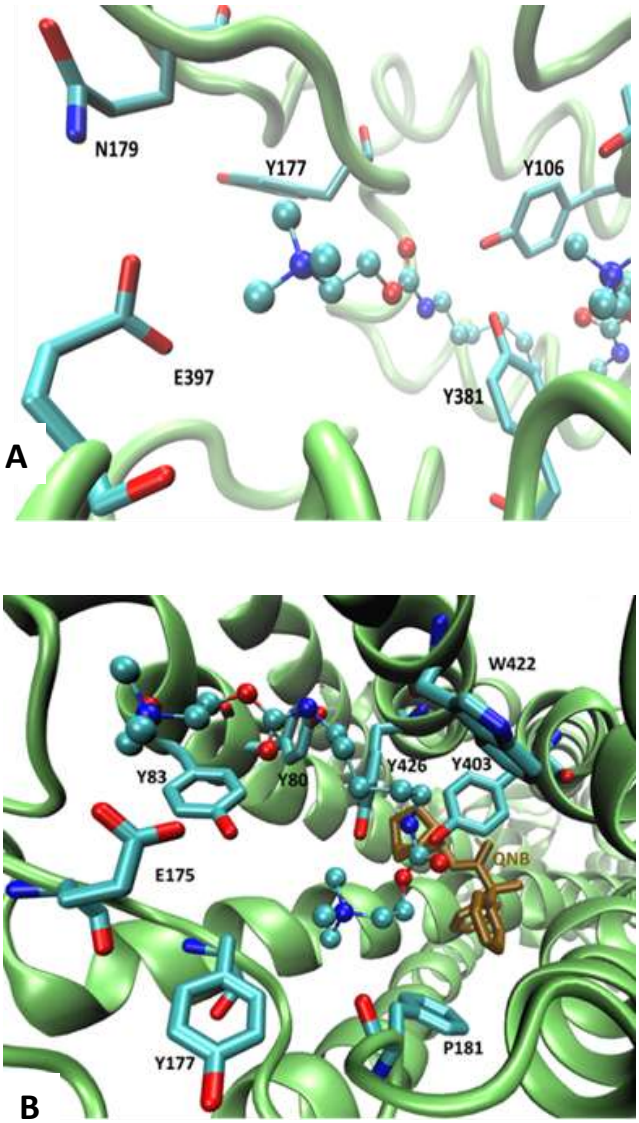


Figure 5.11: Docking poses of selected ligands into the hM_1 and hM_2 receptors. A: n = 9, docked between the orthosteric and allosteric sites of hM1. B: n = 9 into the allosteric site of the QNB-occupied hM2 receptor.

5.4.3.3 Docking results on hM_2 in its open state and in complex with QNB

Docking simulations were repeated by using the M_2 resolved structure in complex with the QNB antagonist in order to mimic an experimental condition comparable with the kinetic binding studies. The obtained results reveal that all compounds with two carbachol units are conveniently accommodated in the allosteric binding cavity regardless of the length of the linker. Thus, Figure 5.11-B depicts the putative complex for n= 5. The ligand is completely harboured in the allosteric site where it can elicit a rich set of polar contacts.

In detail, (i) one ammonium head is engaged in the key ion-pairing with Glu175 reinforced by charge transfer interactions with surrounding aromatic residues (e.g., Tyr83 and Tyr88); (ii) the second ammonium head elicits only charge transfer interactions with Tyr177 and Phe181; (iii) both carbamate moieties are involved in several H-bonds with Tyr80, Thr187, Tyr403, Asn419, Thr423, and Tyr426; (iv) the linker elicits hydrophobic contacts with Trp422. All examined ligands show a very similar binding mode, permitted by their folding degree which increases with the length of the linker. Notably, some tyrosine residues which normally belong to the orthosteric site (i.e. Tyr403 and Tyr426) are here slightly shifted due to the presence of the antagonist and can participate to the interaction in the allosteric site, thus suggesting that these aromatic residues act as a watershed to divide the two considered binding sites.

5.4.4 Conclusions

As mentioned above, preliminary biological data revealed that affinity increases with the linker's length, even though the increase is not smooth, and a "jump" in the affinity values can be seen going from the compounds with $n \leq 5$ to those with $n \geq 7$. Although the beneficial effect of the linker length can be rationalized by

considering a direct contribution of this moiety, probably by hydrophobic interactions, such a discontinuous behaviour may be explained by a change in the binding mode of the compounds, which allows a better fit within the receptor, as if, for instance, a gap could be bridged between the binding sites accommodating the two pharmacophoric units.

Indeed, docking simulations, performed on the recently resolved hM_2 structure, revealed significant differences between the simulated compounds depending on the length of their linker, suggesting a change in binding mode for n = 6. While short compounds ($n \leq 5$) may interact within the orthosteric binding site or between the orthosteric and allosteric sites, by increasing the length of the linker only the bitopic interaction is possible due to steric hindrance.

The comparison of the docking results obtained for hM_1 and hM_2 allows some considerations which correlate with the outcomes of equilibrium binding studies. First, in both receptors the binding mode depends on the length of the linker and only the long-chain analogues are able to occupy both binding sites. Second, the short-chain derivatives reveal significant differences between the two receptor subtypes: in fact, when docked on hM_2 receptor they tend to be completely harboured within the orthosteric site while on the hM_1 subtype they remain in the allosteric cavity. This different behaviour is easily explained by considering that the hM_2 orthosteric pocket is larger and more flexible than the one in hM_1, as demonstrated by previous studies. Third, while assuming comparable pose, the long-chain derivatives are predicted to elicit a more favourable pattern of interactions on hM_1 receptors compared to hM_2. This difference is due to the contacts established by the carbachol moiety within the orthosteric site as well as to the hydrophobic interactions stabilized by the linker.

Appendices

Appendix 1: Physicochemical properties

MOE CODE	DESCRIPTION
AM1_dipole	The dipole moment calculated using the AM1 Hamiltonian
AM1_E	The total SCF energy (kcal/mol) calculated using the AM1 Hamiltonian
AM1_Eele	The electronic energy (kcal/mol) calculated using the AM1 Hamiltonian
AM1_HF	The heat of formation (kcal/mol) calculated using the AM1 Hamiltonian
AM1_HOMO	The energy (eV) of the Highest Occupied Molecular Orbital calculated using the AM1 Hamiltonian
AM1_IP	The ionization potential (kcal/mol) calculated using the AM1 Hamiltonian
AM1_LUMO	The energy (eV) of the Lowest Unoccupied Molecular Orbital calculated using the AM1 Hamiltonian
ASA	Water accessible surface area calculated using a radius of 1.4 A for the water molecule. A polyhedral representation is used for each atom in calculating the surface area.
ASA_H	Water accessible surface area of all hydrophobic (|qi|<0.2) atoms.
ASA_P	Water accessible surface area of all polar (|qi|>=0.2) atoms.
DASA	Absolute value of the difference between ASA+ and ASA-.
DCASA	Absolute value of the difference between CASA+ and CASA-.
dens	Mass density: molecular weight divided by van der Waals volume as calculated in the vol descriptor.
dipole	Dipole moment calculated from the partial charges of the molecule.
E	Value of the potential energy. The state of all term enable flags will be honored (in addition to the term weights). This means that the current potential setup accurately reflects what will be calculated.
E_ang	Angle bends potential energy. In the Potential Setup panel, the term enable (Bonded) flag is ignored, but the term weight is applied.
E_ele	Electrostatic component of the potential energy. In the Potential Setup panel, the term enable flag is ignored, but the term weight is applied.

E_nb	Value of the potential energy with all bonded terms disabled. The state of the non-bonded term enable flags will be honored (in addition to the term weights).
E_oop	Out-of-plane potential energy. In the Potential Setup panel, the term enable (Bonded) flag is ignored, but the term weight is applied.
E_sol	Solvation energy.
E_stb	Bond stretch-bend cross-term potential energy.
E_str	Bond stretch potential energy.
E_strain	Local strain energy: the current energy minus the value of the energy at a near local minimum. The current energy is calculated as for the E descriptor. The local minimum energy is the value of the E descriptor after first performing an energy minimization.
E_tor	Torsion (proper and improper) potential energy.
E_vdw	Van der Waals component of the potential energy.
FASA_H	Fractional ASA_H calculated as ASA_H / ASA.
FASA_P	Fractional ASA_P calculated as ASA_P / ASA.
glob	Globularity or inverse condition number (smallest eigenvalue divided by the largest eigenvalue) of the covariance matrix of atomic coordinates. A value of 1 indicates a perfect sphere while a value of 0 indicates a two- or one-dimensional object.
MNDO_dipole	The dipole moment calculated using the MNDO Hamiltonian.
MNDO_E	The total SCF energy (kcal/mol) calculated using the MNDO Hamiltonian.
MNDO_Eele	The electronic energy (kcal/mol) calculated using the MNDO Hamiltonian.
MNDO_HF	The heat of formation (kcal/mol) calculated using the MNDO Hamiltonian.
MNDO_HOMO	The energy (eV) of the Highest Occupied Molecular Orbital calculated using the MNDO Hamiltonian.
MNDO_IP	The ionization potential (kcal/mol) calculated using the MNDO Hamiltonian.
MNDO_LUMO	The energy (eV) of the Lowest Unoccupied Molecular Orbital calculated using the MNDO Hamiltonian.
npr1	Normalized PMI ratio pmi1/pmi3.
npr2	Normalized PMI ratio pmi2/pmi3.
PM3_dipole	The dipole moment calculated using the PM3 Hamiltonian.
PM3_E	The total SCF energy (kcal/mol) calculated using the PM3 Hamiltonian.

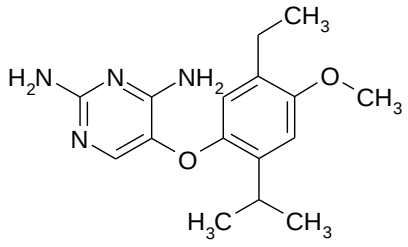
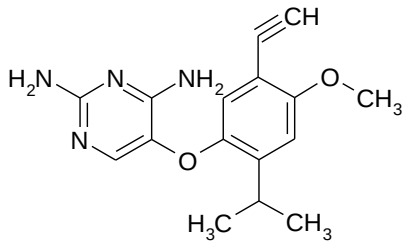
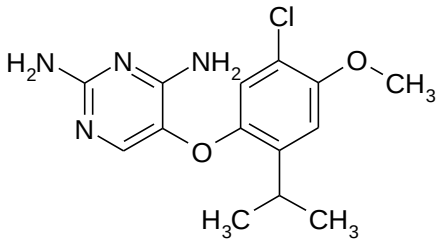
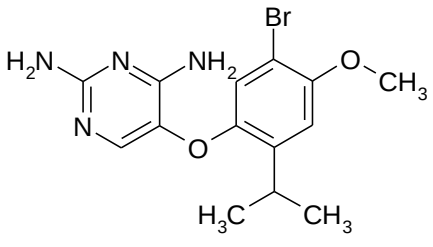
PM3_Eele	The electronic energy (kcal/mol) calculated using the PM3 Hamiltonian.
PM3_HF	The heat of formation (kcal/mol) calculated using the PM3 Hamiltonian.
PM3_HOMO	The energy (eV) of the Highest Occupied Molecular Orbital calculated using the PM3 Hamiltonian.
PM3_IP	The ionization potential (kcal/mol) calculated using the PM3 Hamiltonian.
PM3_LUMO	The energy (eV) of the Lowest Unoccupied Molecular Orbital calculated using the PM3 Hamiltonian.
pmi	Principal moment of inertia.
pmi1	First diagonal element of diagonalized moment of inertia tensor.
pmi2	Second diagonal element of diagonalized moment of inertia tensor.
pmi3	Third diagonal element of diagonalized moment of inertia tensor.
rgyr	Radius of gyration.
std_dim1	Standard dimension 1: the square root of the largest eigenvalue of the covariance matrix of the atomic coordinates. A standard dimension is equivalent to the standard deviation along a principal component axis.
std_dim2	Standard dimension 2: the square root of the second largest eigenvalue of the covariance matrix of the atomic coordinates. A standard dimension is equivalent to the standard deviation along a principal component axis.
std_dim3	Standard dimension 3: the square root of the third largest eigenvalue of the covariance matrix of the atomic coordinates. A standard dimension is equivalent to the standard deviation along a principal component axis.
vol	Van der Waals volume calculated using a grid approximation (spacing 0.75 A).
VSA	Van der Waals surface area. A polyhedral representation is used for each atom in calculating the surface area.
vsurf_A	Amphiphilic moment.
vsurf_CP	Critical packing parameter.
vsurf_CW1	Capacity factor (1).
vsurf_CW2	Capacity factor (2).
vsurf_CW3	Capacity factor (3).
vsurf_CW4	Capacity factor (4).
vsurf_CW5	Capacity factor (5).
vsurf_CW6	Capacity factor (6).
vsurf_CW7	Capacity factor (7).

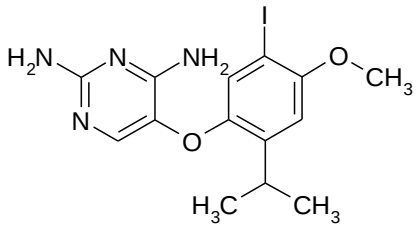
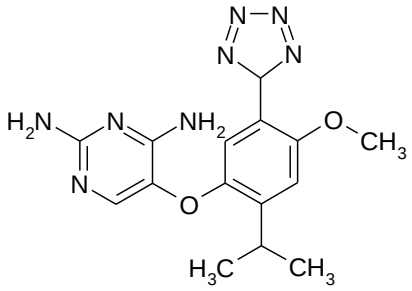
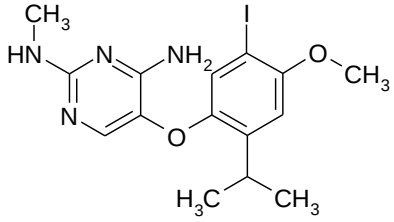
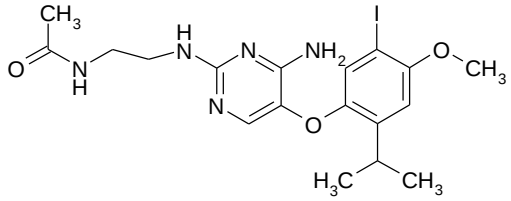
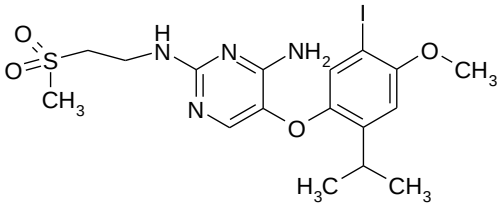
vsurf_CW8	Capacity factor (8).
vsurf_D1	Hydrophobic volume (1).
vsurf_D2	Hydrophobic volume (2).
vsurf_D3	Hydrophobic volume (3).
vsurf_D4	Hydrophobic volume (4).
vsurf_D5	Hydrophobic volume (5).
vsurf_D6	Hydrophobic volume (6).
vsurf_D7	Hydrophobic volume (7).
vsurf_D8	Hydrophobic volume (8).
vsurf_DD12	Contact distances of vsurf_DDmin (12).
vsurf_DD13	Contact distances of vsurf_DDmin (13).
vsurf_DD23	Contact distances of vsurf_DDmin (23).
vsurf_DW12	Contact distances of vsurf_EWmin (12).
vsurf_DW13	Contact distances of vsurf_EWmin (13).
vsurf_DW23	Contact distances of vsurf_EWmin (23).
vsurf_EDmin 1	Lowest hydrophobic energy (1).
vsurf_EDmin 2	Lowest hydrophobic energy (2).
vsurf_EDmin 3	Lowest hydrophobic energy (3).
vsurf_EWmin 1	Lowest hydrophilic energy (1).
vsurf_EWmin 2	Lowest hydrophilic energy (2).
vsurf_EWmin 3	Lowest hydrophilic energy (3).
vsurf_G	Surface globularity.
vsurf_HB1	H-bond donor capacity (1).
vsurf_HB2	H-bond donor capacity (2).
vsurf_HB3	H-bond donor capacity (3).
vsurf_HB4	H-bond donor capacity (4).
vsurf_HB5	H-bond donor capacity (5).
vsurf_HB6	H-bond donor capacity (6).
vsurf_HB7	H-bond donor capacity (7).
vsurf_HB8	H-bond donor capacity (8).
vsurf_HL1	Hydrophilic-Lipophilic (1).
vsurf_HL2	Hydrophilic-Lipophilic (2).

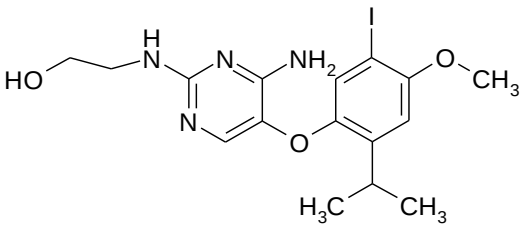
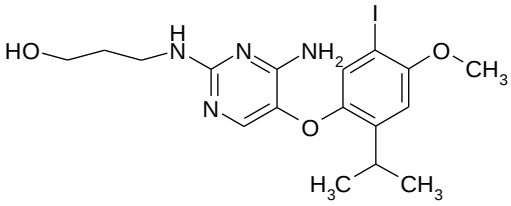
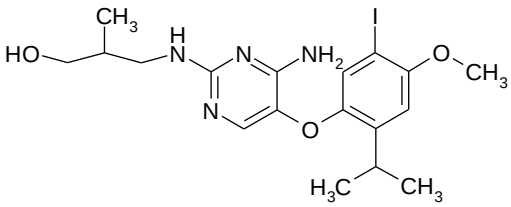
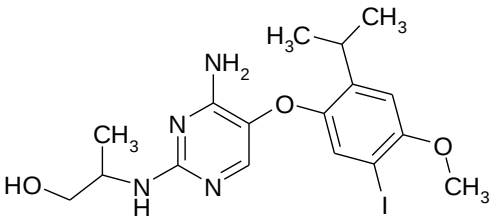
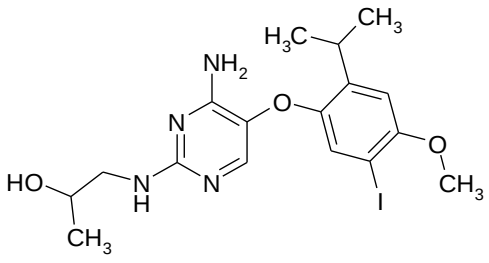
vsurf_ID1	Hydrophobic integy moment (1).
vsurf_ID2	Hydrophobic integy moment (2).
vsurf_ID3	Hydrophobic integy moment (3).
vsurf_ID4	Hydrophobic integy moment (4).
vsurf_ID5	Hydrophobic integy moment (5).
vsurf_ID6	Hydrophobic integy moment (6).
vsurf_ID7	Hydrophobic integy moment (7).
vsurf_ID8	Hydrophobic integy moment (8).
vsurf_IW1	Hydrophilic integy moment (1).
vsurf_IW2	Hydrophilic integy moment (2).
vsurf_IW3	Hydrophilic integy moment (3).
vsurf_IW4	Hydrophilic integy moment (4).
vsurf_IW5	Hydrophilic integy moment (5).
vsurf_IW6	Hydrophilic integy moment (6).
vsurf_IW7	Hydrophilic integy moment (7).
vsurf_IW8	Hydrophilic integy moment (8).
vsurf_R	Surface rugosity.
vsurf_S	Interaction field surface area.
vsurf_V	Interaction field volume.
vsurf_W1	Hydrophilic volume (1).
vsurf_W2	Hydrophilic volume (2).
vsurf_W3	Hydrophilic volume (3).
vsurf_W4	Hydrophilic volume (4).
vsurf_W5	Hydrophilic volume (5).
vsurf_W6	Hydrophilic volume (6).
vsurf_W7	Hydrophilic volume (7).
vsurf_W8	Hydrophilic volume (8).
vsurf_Wp1	Polar volume (1).
vsurf_Wp2	Polar volume (2).
vsurf_Wp3	Polar volume (3).
vsurf_Wp4	Polar volume (4).
vsurf_Wp5	Polar volume (5).
vsurf_Wp6	Polar volume (6).
vsurf_Wp7	Polar volume (7).
vsurf_Wp8	Polar volume (8).
ASA+	Water accessible surface area of all atoms with positive partial charge (strictly greater than 0).
ASA-	Water accessible surface area of all atoms with negative partial

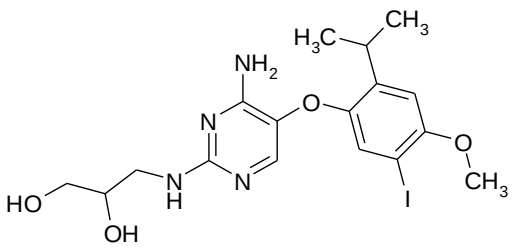
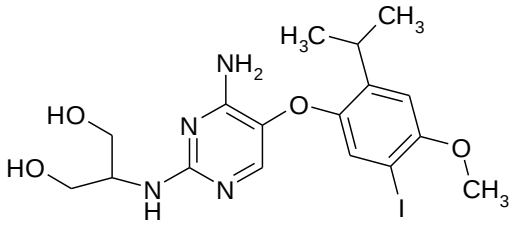
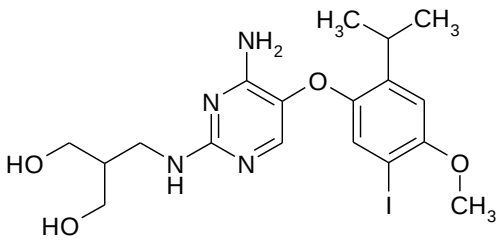
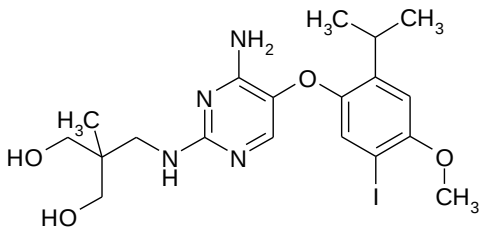
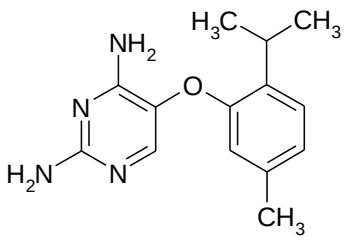
	charge (strictly less than 0).
CASA+	Positive charge weighted surface area, ASA+ times max { qi > 0 } [Stanton 1990].
CASA-	Negative charge weighted surface area, ASA- times max { qi < 0 } [Stanton 1990].
FASA+	Fractional ASA+ calculated as ASA+ / ASA.
FASA-	Fractional ASA- calculated as ASA- / ASA.
FCASA+	Fractional CASA+ calculated as CASA+ / ASA.
FCASA-	Fractional CASA- calculated as CASA- / ASA.

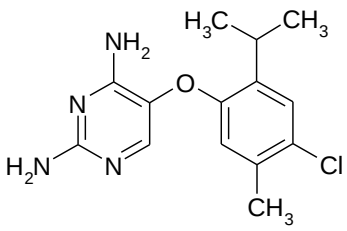
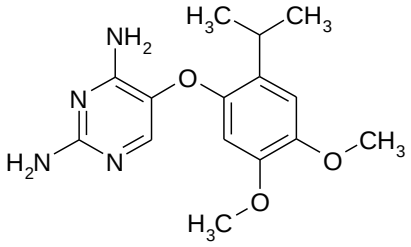
Appendix 2: Purinergic inhibitors

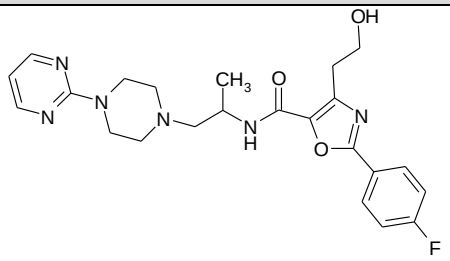
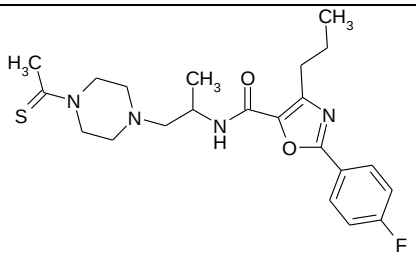
ID	Diaminopyrimidines (DAP)	$P2X_3$ pIC_{50}	$P2X_{2/3}$ pIC_{50}
25_19		7.3	6.8
28_19		8.2	7.9
30_19		7.6	7.0
31_19		7.7	7.1

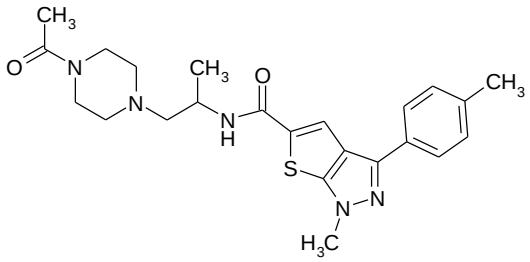
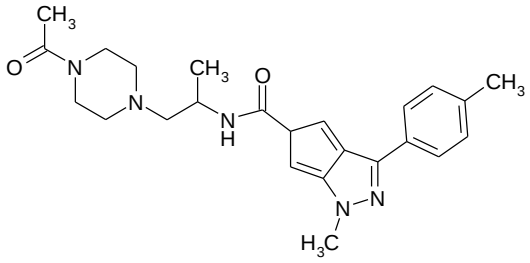
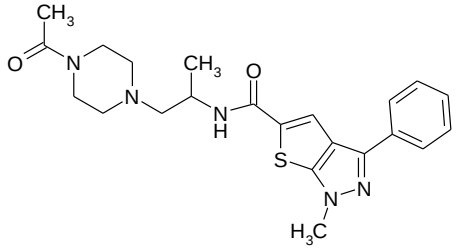
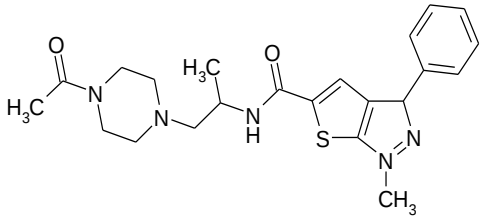
32_19		8.0	7.1
39_19		7.4	6.5
13a_20		7.3	6.7
13d_20		8.0	7.4
13e_20		7.8	7.4

13m_20		8.0	7.8
13o_20		7.8	7.2
13p_20 (R/S)		8.2	7.8
13q_20		8.0	7.5
13r_20 (R/S)		7.9	7.6

13s_20 (R/S)		8.1	7.5
13t_20		8.7	8.3
13u_20		8.2	7.5
13w_20		8.0	7.3
a_6		7.39±0,06	6.68±0,02

c_6		7.24±0,18	6.37±0.12
c_13		7.4	?

ID	Arylamidic derivates (SAA)	$P2X_3$ pIC_{50}
9a_4		7.3
10c_4		7.0

7_4		6.7
11g_4		7.4
11h_4		7.4
16_4 or RO-85		7.5

Appendix 3: Compound descriptors for QSAR models

DESCRIPTOR	DESCRIPTION
COSTITUTIVE	
Angles	Number of bond angles
Atoms	Number of atoms
Bonds	Number of bonds
ChiralAtm	Number of chiral atoms
EzBonds	Number of atoms with E/Z geometric isomerism
FlexTorsion	Number of flexible torsions
HbAcc	Number of H-bond acceptors atoms
HbDon	Number of H-bond donors atoms
HbTot	Number of H-bond acceptors and donors atoms (HbAcc + HbDon)
Impropers	Number of improper angles (out of plane or pyramidal angles)
Torsion	Number of torsions
STRUCTURAL	
Gyrrad	Gyration radius in Å
Mass	Mass in Daltons
MassMi	Monoisotopic mass in Daltons
Ovality	Ovality of the molecule (if it's one, the molecule has a spherical shape)
PSA	Polar surface area in $Å^2$
SAS	Solvent accessible surface in $Å^2$
ASA	Apolar surface area in $Å^2$
Surface	Surface area in $Å^2$
Sdiam	Surface diameter (diameter of the sphere having the same surface of the molecule)
Vdiam	Volume diameter (diameter of the sphere having the same volume of the molecule)
Volume	Molecular volume in A^3
PHYSICO-CHEMICAL	
Charge	Total charge

Dipole	Dipole moment
Lipole	Lipophilic moment (it indicates the lipophilic distribution of the molecule)
VlogP	Bernard Testa's Virtual LogP
DOCKING SCORES	
PLANTS docking score	PLP95 score before the complex minimization
CHEMPLP	CHEMPLP score after the complex minimization
PLP	PLP score after the complex minimization
PLP95	PLP95 score after the complex minimization
CHEMPLP_NORM_H_ATM	CHEMPLP normalized by the number of heavy atoms
PLP_NORM_H_ATM	PLP normalized by the number of heavy atoms
PLP95_NORM_H_ATM	PLP95 normalized by the number of heavy atoms
MLP_{inS}	Molecular Lipophilicity Potential Interaction Score: it indicates the hydro/lipophilic complementarity between the ligand and the receptor
MLP_{inS2}	Like MLP_{inS} but take in account the square distance between the interacting atoms
MLP_{inS3}	Like MLP_{inS} but take in account the cubic distance between the interacting atoms
MLP_{inS}	Like MLP_{inS} but uses the Fermi equation to establish the effect of the distance between the interacting atoms
VdW + Hbond + Desolv	Interaction energy calculated by AutoDock 4 as sum the contributes of Van der Waals, hydrogen bonds and desolvatation energy
CHARMM R6-R12	Non-bond interaction energy calculated by CHARMM 22 force field
Electr	Coulomb electrostatic interaction energy
ElectrDD	Distance-dependent electrostatic interaction energy
Ki	Ki of the minimized complex calculated by AutoDock 4
APBS	Electrostatic binding energy calculated using the Poisson-Botzman model implemented in APBS software
HMScore	Hydrophobic match score calculated by X-Score software
HSScore	Steric score calculated by X-Score software

HPScore	Hydrophobic pair score calculated by X-Score software
Average	Average score calculated by X-Score ((HPScore + HMScore + HSScore) / 3)
Binding Energy	X-score binding energy

Appendix 4: Compound descriptors for consensus functions

Type	SW	Name	Description
Ligand-based descriptors	VEGA	Atoms	number of atoms
		HeavyAtoms	number of heavy atoms
		Gyrrad	radius of gyration
		HbAcc	number of H-bond acceptor atoms
		HbDon	number of H-bond donor atoms
		HbTot	number of H-bonding atoms
		Rotors	number of flexible dihedral angles
		Impropers	number of unsaturations taken from the improper angles
		Lipole	lipophilicity moment
		Mass	molecular weight
		Ovality	molecular ovality
		Psa	polar surface area
		Asa	apolar surface area
		VirtualLogP	logP calculated by the MLP approach
	MOPAC	Heat of formation Dipole Homo energy Lumo energy Cosmo area Cosmo volume	parameters obtained by PM7 semi-empirical optimization
Docking scores	PLANTS	Plp95 $Plp95_{Normalized}$	docking scores calculated by PLANTS

(SW stands for the utilized software).

Bibliography

1. Strömbergsson H. *Chemogenomics : Models of Protein-Ligand Interaction Space*.; 2009. doi:citeulike-article-id:7989621.

2. Bender A, Glen RC. Molecular similarity: a key technique in molecular informatics. *Org Biomol Chem*. 2004;2(22):3204-3218. doi:10.1039/b409813g.

3. van Westen GJP, Wegner JK, IJzerman AP, van Vlijmen HWT, Bender a. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Medchemcomm*. 2011;2(1):16. doi:10.1039/c0md00165a.

4. Kauvar LM. Affinity fingerprinting. *Biotechnology (N Y)*. 1995;13(9):965-966. doi:10.1038/nbt0995-965.

5. Cortés-Ciriano I, Ain QU, Subramanian V, et al. Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. *Med Chem Commun*. 2015;6(1):24-50. doi:10.1039/C4MD00216D.

6. Mestres J, Gregori-Puigjané E, Valverde S, Solé R V. The topology of drug-target interaction networks: implicit dependence on drug properties and target families. *Mol Biosyst*. 2009;5(9):1051-1057. doi:10.1039/b905821b.

7. Prusis P, Muceniece R, Andersson P, Post C, Lundstedt T, Wikberg JES. PLS modeling of chimeric MS04/MSH-peptide and MC1/MC3-receptor interactions reveals a novel method for the analysis of ligand-receptor interactions. *Biochim Biophys Acta - Protein Struct Mol Enzymol*. 2001;1544(1-2):350-357. doi:10.1016/S0167-4838(00)00249-1.

8. Weill N, Rognan D. Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: Application to G protein-coupled receptors and their ligands. *J Chem Inf Model*. 2009;49(4):1049-1062. doi:10.1021/ci800447g.

9. Bock JR, Gough DA. Virtual screen for ligands of orphan G protein-coupled receptors. *J Chem Inf Model*. 2005;45(5):1402-1414. doi:10.1021/ci050006d.

10. Strömbergsson H, Kryshtafovych A, Prusis P, et al. Generalized modeling of enzyme-ligand interactions using proteochemometrics and local protein substructures. *Proteins Struct Funct Genet*. 2006;65(3):568-579. doi:10.1002/prot.21163.

11. Cortés-Ciriano I, van Westen GJP, Bouvier G, et al. Improved large-scale

prediction of growth inhibition patterns using the NCI60 cancer cell line panel. *Bioinformatics*. 2015. doi:10.1093/bioinformatics/btv529.

12. Rognan D. Chemogenomic approaches to rational drug design. *Br J Pharmacol*. 2007;152(1):38-52. doi:10.1038/sj.bjp.0707307.

13. Kubinyi H, Hamprecht FA, Mietzner T. Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. *J Med Chem*. 1998;41(14):2553-2564. doi:10.1021/jm970732a.

14. Pastor M, Cruciani G, McLay I, Pickett S, Clementi S. GRid-INdependent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. *J Med Chem*. 2000;43(17):3233-3243. doi:10.1021/jm000941m.

15. Rogers D, Hahn M. Extended-connectivity fingerprints. *JCheInformModel*. 2010;50(5):742-754. doi:10.1021/ci100050t.

16. Laeeq S, Sirbaiya AK, Siddiqui HH, Mehdi S, Zaidi H. An overview of the computer aided drug designing. 2014;3(5):963-994.

17. MOE. https://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm.

18. Berenger F, Voet A, Lee XY, Zhang KYJ. A rotation-translation invariant molecular descriptor of partial charges and its use in ligand-based virtual screening. *J Cheminform*. 2014;6(1). doi:10.1186/1758-2946-6-23.

19. Ballester PJ, Richards WG. Ultrafast shape recognition for similarity search in molecular databases. *Proc R Soc A Math Phys Eng Sci*. 2007;463(2081):1307-1321. doi:10.1098/rspa.2007.1823.

20. Awale M, Jin X, Reymond J-L. Stereoselective virtual screening of the ZINC database using atom pair 3D-fingerprints. *J Cheminform*. 2015;7(1):1-15. doi:10.1186/s13321-014-0051-5.

21. Carhart RE, Smith DH, Venkataraghavan R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J Chem Inf Comput Sci*. 1985;25(2):64-73. doi:10.1021/ci00046a002.

22. Awale M, Reymond JL. Atom pair 2D-fingerprints perceive 3D-molecular shape and pharmacophores for very fast virtual screening of ZINC and GDB-17. *J Chem Inf Model*. 2014;54:1892-1907. doi:10.1021/ci500232g.

23. Sheridan RP, Miller MD, Underwood DJ, Kearsley SK. Chemical Similarity Using Geometric Atom Pair Descriptors. *J Chem Inf Model*. 1996;36(1):128-136. doi:10.1021/ci950275b.

24. Van Westen GJP, Wegner JK, Bender A, IJzerman AP, Van Vlijmen HWT. Mining protein dynamics from sets of crystal structures using "consensus structures." *Protein Sci*. 2010;19(4):742-752. doi:10.1002/pro.350.

25. Hvidsten TR, Kryshtafovych A, Komorowski J, Fidelis K. A novel approach to fold

recognition using sequence-derived properties from sets of structurally similar local fragments of proteins. In: *Bioinformatics*. Vol 19. ; 2003. doi:10.1093/bioinformatics/btg1064.

26. van Westen GJ, Swier RF, Wegner JK, Ijzerman AP, van Vlijmen HW, Bender A. Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets. *J Cheminf*. 2013;5(1):41. doi:10.1186/1758-2946-5-41.

27. Fourches D, Muratov E, Tropsha a. Trust but verify: on the importance of chemical structure curation in chemoinformatics and QSAR modeling research. *J Chem Inf Model*. 2010;50(7):1189-1204.

28. Tropsha a, Gramatica P, Gombar VK. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *Qsar Comb Sci*. 2003;22(1):69-77. doi:10.1002/qsar.200390007.

29. Eklund M, Norinder U, Boyer S, Carlsson L. Choosing feature selection and learning algorithms in QSAR. *J Chem Inf Model*. 2014;54(3):837-843. doi:10.1021/ci400573c.

30. Wegner JK, Fröhlich H, Zell A. Feature selection for descriptor based classification models. 1. Theory and GA-SEC algorithm. *J Chem Inf Comput Sci*. 2004;44(3):921-930. doi:10.1021/ci0342324.

31. Cortes-Ciriano I, Van Westen GJP, Lenselink EB, Murrell DS, Bender A, Malliavin T. Proteochemometric modeling in a Bayesian framework. *J Cheminform*. 2014;6(1):1-16. doi:10.1186/1758-2946-6-35.

32. Park Y, Marcotte EM. Flaws in evaluation schemes for pair-input computational predictions. *Nat Methods*. 2012;9(12):1134-1136. doi:10.1038/nmeth.2259.

33. Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J Cheminform*. 2014;6(1):1-15. doi:10.1186/1758-2946-6-10.

34. Golbraikh A, Tropsha A. Beware of q^2! *J Mol Graph Model*. 2002;20(4):269-276. http://www.ncbi.nlm.nih.gov/pubmed/11858635.

35. Di L. The role of drug metabolizing enzymes in clearance. *Expert Opin Drug Metab Toxicol*. 2014;10(3):379-393. doi:10.1517/17425255.2014.876006.

36. Testa B, Balmat AL, Long A, Judson P. Predicting drug metabolism - An evaluation of the expert system METEOR. *Chem Biodivers*. 2005;2(7):872-885. doi:10.1002/cbdv.200590064.

37. Kola I, Landis J. Opinion: Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov*. 2004;3(8):711-716. doi:10.1038/nrd1470.

38. Wunberg T, Hendrix M, Hillisch A, et al. Improving the hit-to-lead process: data-driven assessment of drug-like and lead-like screening hits. *Drug Discov Today*. 2006;11(3-4):175-180. doi:10.1016/S1359-6446(05)03700-1.

39. Gleeson MP, Hersey A, Hannongbua S. In-silico ADME models: a general assessment of their utility in drug discovery applications. *Curr Top Med Chem*. 2011;11(4):358-381. doi:10.2174/156802611794480927.

40. Kirchmair J, Göller AH, Lang D, et al. Predicting drug metabolism: experiment and/or computation? *Nat Rev Drug Discov*. 2015;(April). doi:10.1038/nrd4581.

41. Testa B, Krämer SD. The Biochemistry of Drug Metabolism – An Introduction. *Chem Biodivers*. 2006;3(10):1053-1101. doi:10.1002/cbdv.200690111.

42. Kirchmair J, Williamson MJ, Tyzack JD, et al. Computational Prediction of Metabolism: Sites, Products, SAR, P450 Enzyme Dynamics, and Mechanisms. *J Chem Inf Model*. 2012;52(3):617-648. doi:10.1021/ci200542m.

43. Faust K, Croes D, van Helden J. Prediction of metabolic pathways from genome-scale metabolic networks. *BioSystems*. 2011;105(2):109-121. doi:10.1016/j.biosystems.2011.05.004.

44. Testa B, Pedretti A, Vistoli G. Reactions and enzymes in the metabolism of drugs and other xenobiotics. *Drug Discov Today*. 2012;17(11-12):549-560. doi:10.1016/j.drudis.2012.01.017.

45. Oda S, Fukami T, Yokoi T, Nakajima M. A comprehensive review of UDP-glucuronosyltransferase and esterases for drug development. *Drug Metab Pharmacokinet*. 2015;30(1):30-51. doi:10.1016/j.dmpk.2014.12.001.

46. Lévesque E, Turgeon D, Carrier JS, Montminy V, Beaulieu M, Bélanger a. Isolation and characterization of the UGT2B28 cDNA encoding a novel human steroid conjugating UDP-glucuronosyltransferase. *Biochemistry*. 2001;40(13):3869-3881. doi:10.1021/bi002607y.

47. Mackenzie PI, Gonzalez FJ, Owens IS. Cloning and characterization of DNA complementary to rat liver UDP-glucuronosyltransferase mRNA. *J Biol Chem*. 1984;259(19):12153-12160.

48. Mackenzie PI, Owens IS, Burchell B, et al. The UDP glycosyltransferase gene superfamily: recommended nomenclature update based on evolutionary divergence. *Pharmacogenetics*. 1997;7(4):255-269. doi:10.1097/00008571-199708000-00001.

49. Stoffel W, Bosio A. Myelin glycolipids and their functions. *Curr Opin Neurobiol*. 1997;7(5):654-661. doi:10.1016/S0959-4388(97)80085-2.

50. Mackenzie PI, Bock KW, Burchell B, et al. Nomenclature update for the mammalian UDP glycosyltransferase (UGT) gene superfamily. *Pharmacogenet Genomics*. 2005;15(10):677-685. doi:10.1097/01.fpc.0000173483.13689.56.

51. Lairson LL, Henrissat B, Davies GJ, Withers SG. Glycosyltransferases: structures, functions, and mechanisms. *Annu Rev Biochem*. 2008;77(February):521-555. doi:10.1146/annurev.biochem.76.061005.092322.

52. Offen W, Martinez-Fleites C, Yang M, et al. Structure of a flavonoid glucosyltransferase reveals the basis for plant natural product modification. *EMBO*

J. 2006;25(6):1396-1405. doi:10.1038/sj.emboj.7600970.

53. Brazier-Hicks M, Offen WA, Gershater MC, et al. Characterization and engineering of the bifunctional N- and O-glucosyltransferase involved in xenobiotic metabolism in plants. *Proc Natl Acad Sci U S A*. 2007;104(51):20238-20243. doi:10.1073/pnas.0706421104.

54. Miley MJ, Zielinska AK, Keenan JE, Bratton SM, Radominska-Pandya A, Redinbo MR. Crystal Structure of the Cofactor-Binding Domain of the Human Phase II Drug-Metabolism Enzyme UDP-Glucuronosyltransferase 2B7. *J Mol Biol*. 2007;369(2):498-511. doi:10.1016/j.jmb.2007.03.066.

55. Kerdpin O, Mackenzie PI, Bowalgaha K, Finel M, Miners JO. Influence of N-terminal domain histidine and proline residues on the substrate selectivities of human UDP-glucuronosyltransferase 1A1, 1A6, 1A9, 2B7, and 2B10. *Drug Metab Dispos*. 2009;37(9):1948-1955. doi:10.1124/dmd.109.028225.

56. Kato Y, Izukawa T, Oda S, et al. Human UDP-glucuronosyltransferase (ugt) 2b10 in drug n-glucuronidation: Substrate screening and comparison with UGT1A3 and UGT1A4. *Drug Metab Dispos*. 2013;41(7):1389-1397. doi:10.1124/dmd.113.051565.

57. Testa B, Krämer SD. The biochemistry of drug metabolism - An introduction: Part 4. Reactions of conjugation and their enzymes. *Chem Biodivers*. 2008;5(11):2171-2336. doi:10.1002/cbdv.200890199.

58. Testa B, D. Kramer S. *The Biochemistry of Drug Metabolism: Volume 2: Conjugations, Consequences of Metabolism, Influencing Factors (v. 2)*. Wiley-VCH; 2010.

59. Miners JO, Knights KM, Houston JB, Mackenzie PI. In vitro-in vivo correlation for drugs and other compounds eliminated by glucuronidation in humans: pitfalls and promises. *Biochem Pharmacol*. 2006;71(11):1531-1539. doi:10.1016/j.bcp.2005.12.019.

60. Miners JO, Smith PA, Sorich MJ, McKinnon RA, Mackenzie PI. Predicting human drug glucuronidation parameters: application of in vitro and in silico modeling approaches. *Annu Rev Pharmacol Toxicol*. 2004;44:1-25. doi:10.1146/annurev.pharmtox.44.101802.121546.

61. Bowalgaha K, Elliot DJ, Mackenzie PI, Knights KM, Miners JO. The Glucuronidation of ⌬ 4 -3-Keto C19- and C21-Hydroxysteroids by Human Liver Microsomal and Recombinant Hydroxyprogesterone Are Selective Substrates for UGT2B7 ABSTRACT : *Pharmacology*. 2007;35(3):363-370. doi:10.1124/dmd.106.013052.2B15.

62. Uchaipichat V, Mackenzie PI, Elliot DJ, Miners JO. Selectivity of substrate (trifluoperazine) and inhibitor (amitriptyline, androsterone, canrenoic acid, hecogenin, phenylbutazone, quinidine, quinine, and sulfinpyrazone) "probes" for human udp-glucuronosyltransferases. *Drug Metab Dispos*. 2006;34(3):449-456.

doi:10.1124/dmd.105.007369.

63. Uchaipichat V, Winner LK, Mackenzie PI, Elliot DJ, Williams JA, Miners JO. Quantitative prediction of in vivo inhibitory interactions involving glucuronidated drugs from in vitro data: the effect of fluconazole on zidovudine glucuronidation. *Br J Clin Pharmacol*. 2006;61(4):427-439. doi:10.1111/j.1365-2125.2006.02588.x.

64. P I Mackenzie, D A Gardner-Stephen and JOM. *Comprehensive Toxicology*. Vol 4. second. (Elsevier, ed.).; 2010.

65. Vore M, Hadd H, Slikker W. Ethynylestradiol-17 beta D-ring glucuronide conjugates are potent cholestatic agents in the rat. *Life Sci*. 1983;32(26):2989-2993. http://www.ncbi.nlm.nih.gov/pubmed/6865644. Accessed November 27, 2015.

66. Boelsterli UA. Xenobiotic acyl glucuronides and acyl CoA thioesters as protein-reactive metabolites with the potential to cause idiosyncratic drug reactions. *Curr Drug Metab*. 2002;3(4):439-450. http://www.ncbi.nlm.nih.gov/pubmed/12093359. Accessed November 27, 2015.

67. Southwood HT, DeGraaf YC, Mackenzie PI, Miners JO, Burcham PC, Sallustio BC. Carboxylic acid drug-induced DNA nicking in HEK293 cells expressing human UDP-glucuronosyltransferases: role of acyl glucuronide metabolites and glycation pathways. *Chem Res Toxicol*. 2007;20(10):1520-1527. doi:10.1021/tx700188x.

68. Penson RT, Joel SP, Bakhshi K, Clark SJ, Langford RM, Slevin ML. Randomized placebo-controlled trial of the activity of the morphine glucuronides. *Clin Pharmacol Ther*. 2000;68(6):667-676. doi:10.1067/mcp.2000.111934.

69. Curley RW, Abou-Issa H, Panigot MJ, Repa JJ, Clagett-Dame M, Alshafie G. Chemopreventive activities of C-glucuronide/glycoside analogs of retinoid-O-glucuronides against breast cancer development and growth. *Anticancer Res*. 16(2):757-763. http://www.ncbi.nlm.nih.gov/pubmed/8687125. Accessed November 27, 2015.

70. Garcia-Carbonero R, Supko JG. Current perspectives on the clinical experience, pharmacology, and continued development of the camptothecins. *Clin Cancer Res*. 2002;8(3):641-661. http://www.ncbi.nlm.nih.gov/pubmed/11895891. Accessed November 27, 2015.

71. Gagné J-F, Montminy V, Belanger P, Journault K, Gaucher G, Guillemette C. Common human UGT1A polymorphisms and the altered metabolism of irinotecan active metabolite 7-ethyl-10-hydroxycamptothecin (SN-38). *Mol Pharmacol*. 2002;62(3):608-617. http://www.ncbi.nlm.nih.gov/pubmed/12181437. Accessed November 27, 2015.

72. Ando Y, Saka H, Ando M, et al. Polymorphisms of UDP-glucuronosyltransferase gene and irinotecan toxicity: a pharmacogenetic analysis. *Cancer Res*. 2000;60(24):6921-6926. http://www.ncbi.nlm.nih.gov/pubmed/11156391.

73. Minami H, Sai K, Saeki M, et al. Irinotecan pharmacokinetics/pharmacodynamics

and UGT1A genetic polymorphisms in Japanese: roles of UGT1A1*6 and *28. *Pharmacogenet Genomics*. 2007;17(7):497-504. doi:10.1097/FPC.0b013e328014341f.

74. Udomuksorn W, Elliot DJ, Lewis BC, Mackenzie PI, Yoovathaworn K, Miners JO. Influence of mutations associated with Gilbert and Crigler-Najjar type II syndromes on the glucuronidation kinetics of bilirubin and other UDP-glucuronosyltransferase 1A substrates. *Pharmacogenet Genomics*. 2007;17(12):1017-1029. doi:10.1097/FPC.0b013e328256b1b6.

75. Bigler J, Whitton J, Lampe JW, Fosdick L, Bostick RM, Potter JD. CYP2C9 and UGT1A6 genotypes modulate the protective effect of aspirin on colon adenoma risk. *Cancer Res*. 2001;61(9):3566-3569. http://www.ncbi.nlm.nih.gov/pubmed/11325819. Accessed November 28, 2015.

76. Hubner RA, Muir KR, Liu J-F, et al. Genetic variants of UGT1A6 influence risk of colorectal adenoma recurrence. *Clin Cancer Res*. 2006;12(21):6585-6589. doi:10.1158/1078-0432.CCR-06-0903.

77. Picard N, Ratanasavanh D, Prémaud A, Le Meur Y, Marquet P. Identification of the UDP-glucuronosyltransferase isoforms involved in mycophenolic acid phase II metabolism. *Drug Metab Dispos*. 2005;33(1):139-146. doi:10.1124/dmd.104.001651.

78. Chung J-Y, Cho J-Y, Yu K-S, et al. Effect of the UGT2B15 genotype on the pharmacokinetics, pharmacodynamics, and drug interactions of intravenous lorazepam in healthy volunteers. *Clin Pharmacol Ther*. 2005;77(6):486-494. doi:10.1016/j.clpt.2005.02.006.

79. Court MH, Hao Q, Krishnaswamy S, et al. UDP-glucuronosyltransferase (UGT) 2B15 pharmacogenetics: UGT2B15 D85Y genotype and gender are major determinants of oxazepam glucuronidation by human liver. *J Pharmacol Exp Ther*. 2004;310(2):656-665. doi:10.1124/jpet.104.067660.

80. Wanibuchi F, Konishi T, Harada M, et al. Pharmacological studies on novel muscarinic agonists, 1-oxa-8-azaspiro[4.5]decane derivatives, YM796 and YM954. *Eur J Pharmacol*. 1990;187(3):479-486. http://www.ncbi.nlm.nih.gov/pubmed/1963596. Accessed November 26, 2015.

81. Gaganis P, Miners JO, Brennan JS, Thomas A, Knights KM. Human renal cortical and medullary UDP-glucuronosyltransferases (UGTs): immunohistochemical localization of UGT2B7 and UGT1A enzymes and kinetic characterization of S-naproxen glucuronidation. *J Pharmacol Exp Ther*. 2007;323(2):422-430. doi:10.1124/jpet.107.128603.

82. Williams JA, Hyland R, Jones BC, et al. Drug-drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (AUCi/AUC) ratios. *Drug Metab Dispos*. 2004;32(11):1201-1208. doi:10.1124/dmd.104.000794.

83. Court MH, Krishnaswamy S, Hao Q, et al. Evaluation of 3-azido-3-deoxythymidine, morphine, and codeine as probe substrates for udp-glucuronosyltransferase 2B7 (UGT2B7) in human liver microsomes: Specificity and influence of the UGT2B7*2 polymorphism. *Drug Metab Dispos*. 2003;31(9):1125-1133. doi:10.1124/dmd.31.9.1125.

84. Raungrut P, Uchaipichat V, Elliot DJ, Janchawee B, Somogyi A a, Miners JO. In vitro-in vivo extrapolation predicts drug-drug interactions arising from inhibition of codeine glucuronidation by dextropropoxyphene, fluconazole, ketoconazole, and methadone in humans. *J Pharmacol Exp Ther*. 2010;334(2):609-618. doi:10.1124/jpet.110.167916.

85. Miners JO, Valente L, Lillywhite KJ, et al. Preclinical prediction of factors influencing the elimination of 5,6-dimethylxanthenone-4-acetic acid, a new anticancer drug. *CANCER Res*. 1997;57(2):284-289.

86. Innocenti F, Iyer L, Ramírez J, Green MD, Ratain MJ. Epirubicin glucuronidation is catalyzed by human UDP-glucuronosyltransferase 2B7. *Drug Metab Dispos*. 2001;29(5):686-692.

87. Mano Y, Usui T, Kamimura H. The UDP-glucuronosyltransferase 2B7 isozyme is responsible for gemfibrozil glucuronidation in the human liver. *Drug Metab Dispos*. 2007;35(11):2040-2044. doi:10.1124/dmd.107.017269.

88. Jin C, Miners JO, Lillywhite KJ, Mackenzie PI. Complementary deoxyribonucleic acid cloning and expression of a human liver uridine diphosphate-glucuronosyltransferase glucuronidating carboxylic acid-containing drugs. *J Pharmacol Exp Ther*. 1993;264(1):475-479. http://www.ncbi.nlm.nih.gov/pubmed/8423545.

89. Bowalgaha K, Elliot DJ, Mackenzie PI, Knights KM, Swedmark S, Miners JO. S-naproxen and desmethylnaproxen glucuronidation by human liver microsomes and recombinant human UDP-glucuronosyltransferases (UGT): Role of UGT2B7 in the elimination of naproxen. *Br J Clin Pharmacol*. 2005;60(4):423-433. doi:10.1111/j.1365-2125.2005.02446.x.

90. Jin CJ, Mackenzie PI, Miners JO. The regio- and stereo-selectivity of C19 and C21 hydroxysteroid glucuronidation by UGT2B7 and UGT2B11. *Arch Biochem Biophys*. 1997;341(2):207-211. doi:10.1006/abbi.1997.9949.

91. Breton C, Šnajdrová L, Jeanneau C, Koča J, Imberty A. Structures and mechanisms of glycosyltransferases. *Glycobiology*. 2006;16(2):29-37. doi:10.1093/glycob/cwj016.

92. Shao H, He X, Achnine L, Blount JW, Dixon RA, Wang X. Crystal structures of a multifunctional triterpene/flavonoid glycosyltransferase from Medicago truncatula. *Plant Cell*. 2005;17(11):3141-3154. doi:10.1105/tpc.105.035055.

93. Chau N, Elliot DJ, Lewis BC, et al. Morphine glucuronidation and glucosidation represent complementary metabolic pathways that are both catalyzed by UDP-

glucuronosyltransferase 2B7: kinetic, inhibition, and molecular modeling studies. *J Pharmacol Exp Ther*. 2014;349(1):126-137. doi:10.1124/jpet.113.212258.

94. Lewis BC, Mackenzie PI, Miners JO. Homodimerization of UDP-glucuronosyltransferase 2B7 (UGT2B7) and identification of a putative dimerization domain by protein homology modeling. *Biochem Pharmacol*. 2011;82(12):1-8. doi:10.1016/j.bcp.2011.09.007.

95. Rocha J, Popescu AO, Borges P, et al. Structure of Burkholderia cepacia UDP-Glucose Dehydrogenase (UGD) BceC and Role of Tyr10 in Final Hydrolysis of UGD Thioester Intermediate. *J Bacteriol*. 2011;193(15):3978-3987. doi:10.1128/JB.01076-10.

96. MOPAC. http://openmopac.net/.

97. Plants. http://www.tcd.uni-konstanz.de/research/plants.php.

98. NAMD. http://www.ks.uiuc.edu/Research/namd/.

99. Md tools. http://www.ks.uiuc.edu/Development /MDTools/sodium/.

100. Uniprot. http://www.uniprot.org/.

101. VEGA ZZ. http://nova.disfarm.unimi.it/vegazz.

102. JChem Suite. https://www.chemaxon.com/download/jchem-suite/.

103. Chemaxon Software Company. https://www.chemaxon.com/.

104. RStudio. https://www.rstudio.com/.

105. Java. https://java.com/it/.

106. Pubmed. http://www.ncbi.nlm.nih.gov/pubmed.

107. Clustal Omega. http://www.clustal.org/omega/.

108. Python. https://www.python.org/.

109. pycharm. https://www.jetbrains.com/pycharm/.

110. Pandas. http://pandas.pydata.org/index.html.

111. scikit-learn. http://scikit-learn.org/stable/.

112. Ganganwar V. An overview of classification algorithms for imbalanced datasets. *Int J Emerg Technol Adv Eng*. 2012;2(4):42-47. http://www.ijetae.com/files/Volume2Issue4/IJETAE_0412_07.pdf.

113. scikit learn RFC. http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

114. class weight warning. https://github.com/scikit-learn/scikit-learn/issues/3928.

115. Longadge R, Dongre S, Malik L. Class Imbalance Problem in Data Mining : Review. 2013;2(1).

116. Galar M, Fern A, Barrenechea E, Bustince H. Hybrid-Based Approaches. 2011:1-22.

117. Kawashima S, Ogata H, Kanehisa M. AAindex: Amino Acid Index Database. *Nucleic Acids Res*. 1999;27(1):368-369.

118. Sahigara F. Defining the Applicability Domain of QSAR models : An overview. *Mol Descriptors - Free online Resour*. 2007;Tutorial 7:1-6.

119. Safety CP. REVIEW OF METHODS FOR ASSESSING THE APPLICABILTY DOMAINS OF SARS AND QSARS Review of methods for applicability domain estimation. 1853;2004(September 2004):1-5.

120. Nikolova-Jeliazkova N, Jaworska J. An approach to determining applicability domains for QSAR group contribution models: An analysis of SRC KOWWIN. *ATLA Altern to Lab Anim*. 2005;33(5):461-470.

121. Holton P. The liberation of adenosine triphosphate on antidromic stimulation of sensory nerves. *J Physiol*. 1959;145(3):494-504. doi:10.1113/jphysiol.1959.sp006157.

122. Burnstock G. Purinergic nerves. *Pharmacol Rev*. 1972;24(3):509-581. http://www.ncbi.nlm.nih.gov/pubmed/4404211.

123. Kawate T, Michel JC, Birdsong WT, Gouaux E. Crystal structure of the ATP-gated P2X(4) ion channel in the closed state. *Nature*. 2009;460(7255):592-598. doi:10.1038/nature08198.

124. Hattori M, Gouaux E. Molecular mechanism of ATP binding and ion channel activation in P2X receptors. *Nature*. 2012;485(7397):207-212. doi:10.1038/nature11010.

125. Browne LE. Structure of P2X receptors. *Wiley Interdiscip Rev Membr Transp Signal*. 2012;1(1):56-69. doi:10.1002/wmts.24.

126. Jiang R, Taly A, Grutter T. Moving through the gate in ATP-activated P2X receptors. *Trends Biochem Sci*. 2013;38(1):20-29. doi:10.1016/j.tibs.2012.10.006.

127. Chaumont S, Jiang L-H, Penna A, North RA, Rassendren F. Identification of a trafficking motif involved in the stabilization and polarization of P2X receptors. *J Biol Chem*. 2004;279(28):29628-29638. doi:10.1074/jbc.M403940200.

128. North RA. Molecular Physiology of P2X Receptors. *Physiol Rev*. 2002;82(4):1013-1067. doi:10.1152/physrev.00015.2002.

129. Ford AP. In pursuit of P2X3 antagonists: novel therapeutics for chronic pain and afferent sensitization. *Purinergic Signal*. 2012;8(S1):3-26. doi:10.1007/s11302-011-9271-6.

130. Coddou C, Yan Z, Obsil T, Huidobro-Toro JP, Stojilkovic SS. Activation and Regulation of Purinergic P2X Receptor Channels. *Pharmacol Rev*. 2011;63(3):641-683. doi:10.1124/pr.110.003129.

131. Wu G, Whiteside GT, Lee G, et al. A-317491, a selective P2X3/P2X2/3 receptor antagonist, reverses inflammatory mechanical hyperalgesia through action at peripheral receptors in rats. *Eur J Pharmacol*. 2004;504(1-2):45-53. doi:10.1016/j.ejphar.2004.09.056.

132. Gever JR, Soto R, Henningsen RA, et al. AF-353, a novel, potent and orally bioavailable P2X3/P2X2/3 receptor antagonist. *Br J Pharmacol*. 2010;160(6):1387-1398. doi:10.1111/j.1476-5381.2010.00796.x.

133. Abdulqawi R, Dockry R, Holt K, et al. P2X3 receptor antagonist (AF-219) in refractory chronic cough: a randomised, double-blind, placebo-controlled phase 2 study. *Lancet (London, England)*. 2015;385(9974):1198-1205. doi:10.1016/S0140-6736(14)61255-1.

134. Fauman EB, Rai BK, Huang ES. Structure-based druggability assessment--identifying suitable targets for small molecule therapeutics. *Curr Opin Chem Biol*. 2011;15(4):463-468. doi:10.1016/j.cbpa.2011.05.020.

135. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*. 2009;10(1):168. doi:10.1186/1471-2105-10-168.

136. Di Domizio A, Vitriolo A, Vistoli G, Pedretti A. SPILLO-PBSS: detecting hidden binding sites within protein 3D-structures through a flexible structure-based approach. *J Comput Chem*. 2014;35(27):2005-2017. doi:10.1002/jcc.23714.

137. Borrelli KW, Vitalis A, Alcantara R, Guallar V. PELE: Protein Energy Landscape Exploration. A Novel Monte Carlo Based Technique. *J Chem Theory Comput*. 2005;1(6):1304-1311. doi:10.1021/ct0501811.

138. Madadkar-Sobhani A, Guallar V. PELE web server: atomistic study of biomolecular systems at your fingertips. *Nucleic Acids Res*. 2013;41(Web Server issue):W322-W328. doi:10.1093/nar/gkt454.

139. Modeller. https://salilab.org/modeller/.

140. PROCHECK. http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/.

141. Ballini E, Virginio C, Medhurst SJ, et al. Characterization of three diaminopyrimidines as potent and selective antagonists of P2X3 and P2X2/3 receptors with in vivo efficacy in a pain model. *Br J Pharmacol*. 2011;163(6):1315-1325. doi:10.1111/j.1476-5381.2011.01322.x.

142. Carter DS, Alam M, Cai H, et al. Identification and SAR of novel diaminopyrimidines. Part 1: The discovery of RO-4, a dual P2X(3)/P2X(2/3) antagonist for the treatment of pain. *Bioorg Med Chem Lett*. 2009;19(6):1628-1631. doi:10.1016/j.bmcl.2009.02.003.

143. Jahangir A, Alam M, Carter DS, et al. Identification and SAR of novel diaminopyrimidines. Part 2: The discovery of RO-51, a potent and selective, dual P2X(3)/P2X(2/3) antagonist for the treatment of pain. *Bioorg Med Chem Lett*.

2009;19(6):1632-1635. doi:10.1016/j.bmcl.2009.01.097.

144. Ford AP, Undem BJ. The therapeutic promise of ATP antagonism at P2X3 receptors in respiratory and urological disorders. *Front Cell Neurosci*. 2013;7:267. doi:10.3389/fncel.2013.00267.

145. DUD. http://dud.docking.org/.

146. PLANTS. http://www.tcd.uni-konstanz.de/research/plants.php.

147. Korb O, Stützle T, Exner TE. Empirical scoring functions for advanced protein-ligand docking with PLANTS. *J Chem Inf Model*. 2009;49(1):84-96. doi:10.1021/ci800298z.

148. PELE online server. https://pele.bsc.es/.

149. Topiol S. X-ray structural information of GPCRs in drug design: what are the limitations and where do we go? *Expert Opin Drug Discov*. 2013;8(6):607-620. doi:10.1517/17460441.2013.783815.

150. Tehan BG, Bortolato A, Blaney FE, Weir MP, Mason JS. Unifying family A GPCR theories of activation. *Pharmacol Ther*. 2014;143(1):51-60. doi:10.1016/j.pharmthera.2014.02.004.

151. Venkatakrishnan AJ, Deupi X, Lebon G, Tate CG, Schertler GF, Babu MM. Molecular signatures of G-protein-coupled receptors. *Nature*. 2013;494(7436):185-194. doi:10.1038/nature11896.

152. Costanzi S. Modeling G protein-coupled receptors and their interactions with ligands. *Curr Opin Struct Biol*. 2013;23(2):185-190. doi:10.1016/j.sbi.2013.01.008.

153. Deupi X. Relevance of rhodopsin studies for GPCR activation. *Biochim Biophys Acta*. 2014;1837(5):674-682. doi:10.1016/j.bbabio.2013.09.002.

154. Deupi X. Quantification of structural distortions in the transmembrane helices of GPCRs. *Methods Mol Biol*. 2012;914:219-235. doi:10.1007/978-1-62703-023-6_13.

155. Bettinelli I, Graziani D, Marconi C, Pedretti A, Vistoli G. The approach of conformational chimeras to model the role of proline-containing helices on GPCR mobility: the fertile case of Cys-LTR1. *ChemMedChem*. 2011;6(7):1217-1227. doi:10.1002/cmdc.201100037.

156. Kruse AC, Hu J, Kobilka BK, Wess J. Muscarinic acetylcholine receptor X-ray structures: potential implications for drug development. *Curr Opin Pharmacol*. 2014;16:24-30. doi:10.1016/j.coph.2014.02.006.

157. Fisher A. Cholinergic modulation of amyloid precursor protein processing with emphasis on M1 muscarinic receptor: perspectives and challenges in treatment of Alzheimer's disease. *J Neurochem*. 2012;120 Suppl :22-33. doi:10.1111/j.1471-4159.2011.07507.x.

158. Dean B. Selective activation of muscarinic acetylcholine receptors for the treatment

of schizophrenia. *Curr Pharm Biotechnol*. 2012;13(8):1563-1571. http://www.ncbi.nlm.nih.gov/pubmed/22283751. Accessed November 26, 2015.

159. Davie BJ, Christopoulos A, Scammells PJ. Development of M1 mAChR allosteric and bitopic ligands: prospective therapeutics for the treatment of cognitive deficits. *ACS Chem Neurosci*. 2013;4(7):1026-1048. doi:10.1021/cn400086m.

160. Kruse AC, Hu J, Pan AC, et al. Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature*. 2012;482(7386):552-556. doi:10.1038/nature10867.

161. Cherezov V, Rosenbaum DM, Hanson MA, et al. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science*. 2007;318(5854):1258-1265. doi:10.1126/science.1150577.

162. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*. 2000;29:291-325. doi:10.1146/annurev.biophys.29.1.291.

163. Bas DC, Rogers DM, Jensen JH. Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins*. 2008;73(3):765-783. doi:10.1002/prot.22102.

164. Wilman HR, Shi J, Deane CM. Helix kinks are equally prevalent in soluble and membrane proteins. *Proteins*. 2014;82(9):1960-1970. doi:10.1002/prot.24550.

165. Pedretti A, Villa L, Vistoli G. VEGA: a versatile program to convert, handle and visualize molecular structure on Windows-based PCs. *J Mol Graph Model*. 2002;21(1):47-49. http://www.ncbi.nlm.nih.gov/pubmed/12413030. Accessed November 26, 2015.

166. Phillips JC, Braun R, Wang W, et al. Scalable molecular dynamics with NAMD. *J Comput Chem*. 2005;26(16):1781-1802. doi:10.1002/jcc.20289.

167. Vistoli G, Pedretti A, Dei S, Scapecchi S, Marconi C, Romanelli MN. Docking analyses on human muscarinic receptors: unveiling the subtypes peculiarities in agonists binding. *Bioorg Med Chem*. 2008;16(6):3049-3058. doi:10.1016/j.bmc.2007.12.036.

168. Sullivan NR, Leventhal L, Harrison J, et al. Pharmacological characterization of the muscarinic agonist (3R,4R)-3-(3-hexylsulfanyl-pyrazin-2-yloxy)-1-aza-bicyclo[2.2.1]heptane (WAY-132983) in in vitro and in vivo models of chronic pain. *J Pharmacol Exp Ther*. 2007;322(3):1294-1304. doi:10.1124/jpet.106.118604.

169. Watt ML, Schober DA, Hitchcock S, et al. Pharmacological characterization of LY593093, an M1 muscarinic acetylcholine receptor-selective partial orthosteric agonist. *J Pharmacol Exp Ther*. 2011;338(2):622-632. doi:10.1124/jpet.111.182063.

170. Tecle H, Barrett SD, Lauffer DJ, et al. Design and synthesis of m1-selective muscarinic agonists: (R)-(-)-(Z)-1-Azabicyclo[2.2.1]heptan-3-one, O-(3-(3'-

methoxyphenyl)-2-propynyl)oxime maleate (CI-1017), a functionally m1-selective muscarinic agonist. *J Med Chem*. 1998;41(14):2524-2536. doi:10.1021/jm960683m.

171. Avlani VA, Langmead CJ, Guida E, et al. Orthosteric and allosteric modes of interaction of novel selective agonists of the M1 muscarinic acetylcholine receptor. *Mol Pharmacol*. 2010;78(1):94-104. doi:10.1124/mol.110.064345.

172. Verspohl EJ, Tacke R, Mutschler E, Lambrecht G. Muscarinic receptor subtypes in rat pancreatic islets: binding and functional studies. *Eur J Pharmacol*. 1990;178(3):303-311. http://www.ncbi.nlm.nih.gov/pubmed/2187704. Accessed November 26, 2015.

173. Wienrich M, Meier D, Ensinger HA, et al. Pharmacodynamic profile of the M1 agonist talsaclidine in animals and man. *Life Sci*. 2001;68(22-23):2593-2600. http://www.ncbi.nlm.nih.gov/pubmed/11392631. Accessed November 26, 2015.

174. Fisher A. M1 muscarinic agonists target major hallmarks of Alzheimer's disease--an update. *Curr Alzheimer Res*. 2007;4(5):577-580. http://www.ncbi.nlm.nih.gov/pubmed/18220527. Accessed November 26, 2015.

175. Mei L, Lai J, Yamamura HI, Roeske WR. Pharmacologic comparison of selected agonists for the M1 muscarinic receptor in transfected murine fibroblast cells (B82). *J Pharmacol Exp Ther*. 1991;256(2):689-694. http://www.ncbi.nlm.nih.gov/pubmed/1704434. Accessed November 26, 2015.

176. Sánchez C, Arnt J, Didriksen M, Dragsted N, Moltzen Lenz S, Matz J. In vivo muscarinic cholinergic mediated effects of Lu 25-109, a M1 agonist and M2/M3 antagonist in vitro. *Psychopharmacology (Berl)*. 1998;137(3):233-240. http://www.ncbi.nlm.nih.gov/pubmed/9683000. Accessed November 26, 2015.

177. Michal P, El-Fakahany EE, Dolezal V. Muscarinic M2 receptors directly activate Gq/11 and Gs G-proteins. *J Pharmacol Exp Ther*. 2007;320(2):607-614. doi:10.1124/jpet.106.114314.

178. ChemBridge website. www.hit2lead.com.

179. Kukol A. Consensus virtual screening approaches to predict protein ligands. *Eur J Med Chem*. 2011;46(9):4661-4664. doi:10.1016/j.ejmech.2011.05.026.

180. Plewczynski D, Łaźniewski M, von Grotthuss M, Rychlewski L, Ginalski K. VoteDock: consensus docking method for prediction of protein-ligand interactions. *J Comput Chem*. 2011;32(4):568-581. doi:10.1002/jcc.21642.

181. Guvenir HA, Kurtcephe M. Ranking Instances by Maximizing the Area under ROC Curve. *IEEE Trans Knowl Data Eng*. 2013;25(10):2356-2366. doi:10.1109/TKDE.2012.214.

182. Hooke R, Jeeves TA. `` Direct Search'' Solution of Numerical and Statistical Problems. *J ACM*. 1961;8(2):212-229. doi:10.1145/321062.321069.

183. Schmidtke P, Le Guilloux V, Maupetit J, Tufféry P. fpocket: online tools for

protein ensemble pocket detection and tracking. *Nucleic Acids Res*. 2010;38(Web Server issue):W582-W589. doi:10.1093/nar/gkq383.

184. Wallach I, Jaitly N, Nguyen K, Schapira M, Lilien R. Normalizing molecular docking rankings using virtually generated decoys. *J Chem Inf Model*. 2011;51(8):1817-1830. doi:10.1021/ci200175h.

185. Wilson GL, Lill MA. Integrating structure-based and ligand-based approaches for computational drug design. *Future Med Chem*. 2011;3(6):735-750. doi:10.4155/fmc.11.18.

186. Truchon J-F, Bayly CI. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J Chem Inf Model*. 47(2):488-508. doi:10.1021/ci600426e.

187. Deupi X, Standfuss J, Schertler G. Conserved activation pathways in G-protein-coupled receptors. *Biochem Soc Trans*. 2012;40(2):383-388. doi:10.1042/BST20120001.

188. Broadley KJ, Kelly DR. Muscarinic Receptor Agonists and Antagonists. *Molecules*. 2001;6(3):142-193. doi:10.3390/60300142.

189. Ridley HF, Chatterjee SS, Moran JF, Triggle DJ. Studies on the cholinergic receptor. IV. The synthesis and muscarinic activity of 3,7-dimethyl-2,4-dioxo-7-azaspiro[3.4]octane methiodide. *J Med Chem*. 1969;12(5):931-933. http://www.ncbi.nlm.nih.gov/pubmed/5812223. Accessed November 27, 2015.

190. Piergentili A, Quaglia W, Giannella M, et al. Dioxane and oxathiane nuclei: suitable substructures for muscarinic agonists. *Bioorg Med Chem*. 2007;15(2):886-896. doi:10.1016/j.bmc.2006.10.040.

191. Takayanagi I, Harada M, Koike K. A difference in receptor mechanisms for muscarinic full and partial agonists. *Jpn J Pharmacol*. 1991;56(1):23-31. http://www.ncbi.nlm.nih.gov/pubmed/1880983. Accessed November 27, 2015.

192. Triggle DJ, Belleau B. STUDIES ON THE CHEMICAL BASIS FOR CHOLINOMIMETIC AND CHOLINOLYTIC ACTIVITY: PART I. THE SYNTHESIS AND CONFIGURATION OF QUATERNARY SALTS IN THE 1,3-DIOXOLANE AND OXAZOLINE SERIES. *Can J Chem*. 1962;40(6):1201-1215. doi:10.1139/v62-183.

193. Fourneau JP, Rybak B. Mise au point d'une nouvelle technique de titration sur coeur ouvert de mammifere des effets muscariniques a partir d'une nouvelle serie de composes parasympathomimetiques. *Life Sci*. 1962;1(5):185-193. doi:10.1016/0024-3205(62)90016-4.

194. Piergentili A, Quaglia W, Del Bello F, et al. Properly substituted 1,4-dioxane nucleus favours the selective M3 muscarinic receptor activation. *Bioorg Med Chem*. 2009;17(24):8174-8185. doi:10.1016/j.bmc.2009.10.027.

195. Haga K, Kruse AC, Asada H, et al. Structure of the human M2 muscarinic

acetylcholine receptor bound to an antagonist. *Nature*. 2012;482(7386):547-551. doi:10.1038/nature10753.

196. Morris GM, Goodsell DS, Halliday RS, et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem*. 1998;19(14):1639-1662. doi:10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B.

197. Pedretti A, Vistoli G, Marconi C, Testa B. Muscarinic receptors: A comparative analysis of structural features and binding modes through homology modelling and molecular docking. *Chem Biodivers*. 2006;3(5):481-501. doi:10.1002/cbdv.200690052.

198. Vistoli G, Pedretti A, Testa B, Matucci R. The conformational and property space of acetylcholine bound to muscarinic receptors: an entropy component accounts for the subtype selectivity of acetylcholine. *Arch Biochem Biophys*. 2007;464(1):112-121. doi:10.1016/j.abb.2007.04.022.

199. Bodick NC, Offen WW, Levey AI, et al. Effects of xanomeline, a selective muscarinic receptor agonist, on cognitive function and behavioral symptoms in Alzheimer disease. *Arch Neurol*. 1997;54(4):465-473. http://www.ncbi.nlm.nih.gov/pubmed/9109749. Accessed November 27, 2015.

200. Mirza NR, Peters D, Sparks RG. Xanomeline and the antipsychotic potential of muscarinic receptor subtype selective agonists. *CNS Drug Rev*. 2003;9(2):159-186. http://www.ncbi.nlm.nih.gov/pubmed/12847557. Accessed November 27, 2015.

201. Conn PJ, Jones CK, Lindsley CW. Subtype-selective allosteric modulators of muscarinic receptors for the treatment of CNS disorders. *Trends Pharmacol Sci*. 2009;30(3):148-155. doi:10.1016/j.tips.2008.12.002.

202. Melancon BJ, Hopkins CR, Wood MR, et al. Allosteric modulation of seven transmembrane spanning receptors: theory, practice, and opportunities for central nervous system drug discovery. *J Med Chem*. 2012;55(4):1445-1464. doi:10.1021/jm201139r.

203. Mohr K, Schmitz J, Schrage R, Tränkle C, Holzgrabe U. Molecular alliance-from orthosteric and allosteric ligands to dualsteric/bitopic agonists at G protein coupled receptors. *Angew Chem Int Ed Engl*. 2013;52(2):508-516. doi:10.1002/anie.201205315.

204. Mohr K, Tränkle C, Kostenis E, Barocelli E, De Amici M, Holzgrabe U. Rational design of dualsteric GPCR ligands: quests and promise. *Br J Pharmacol*. 2010;159(5):997-1008. doi:10.1111/j.1476-5381.2009.00601.x.

205. Valant C, Robert Lane J, Sexton PM, Christopoulos A. The best of both worlds? Bitopic orthosteric/allosteric ligands of g protein-coupled receptors. *Annu Rev Pharmacol Toxicol*. 2012;52:153-178. doi:10.1146/annurev-pharmtox-010611-134514.

206. Portoghese PS. From Models to Molecules: Opioid Receptor Dimers, Bivalent

Ligands, and Selective Opioid Receptor Probes †. *J Med Chem*. 2001;44(14):2259-2269. doi:10.1021/jm010158+.

207. Steinfeld T, Mammen M, Smith JAM, Wilson RD, Jasper JR. A novel multivalent ligand that bridges the allosteric and orthosteric binding sites of the M2 muscarinic receptor. *Mol Pharmacol*. 2007;72(2):291-302. doi:10.1124/mol.106.033746.

208. Bock A, Merten N, Schrage R, et al. The allosteric vestibule of a seven transmembrane helical receptor controls G-protein coupling. *Nat Commun*. 2012;3:1044. doi:10.1038/ncomms2028.

209. Lane JR, Donthamsetti P, Shonberg J, et al. A new mechanism of allostery in a G protein–coupled receptor dimer. *Nat Chem Biol*. 2014;10(9):745-752. doi:10.1038/nchembio.1593.

210. Melchiorre C, Cassinelli A, Quaglia W. Differential blockade of muscarinic receptor subtypes by polymethylene tetraamines. Novel class of selective antagonists of cardiac M-2 muscarinic receptors. *J Med Chem*. 1987;30(1):201-204. http://www.ncbi.nlm.nih.gov/pubmed/3806594. Accessed November 27, 2015.

211. Christopoulos A, Grant MK, Ayoubzadeh N, et al. Synthesis and pharmacological evaluation of dimeric muscarinic acetylcholine receptor agonists. *J Pharmacol Exp Ther*. 2001;298(3):1260-1268. http://www.ncbi.nlm.nih.gov/pubmed/11504829. Accessed November 27, 2015.

212. Rajeswaran WG, Cao Y, Huang XP, et al. Design, synthesis, and biological characterization of bivalent 1-methyl-1,2,5,6-tetrahydropyridyl-1,2,5-thiadiazole derivatives as selective muscarinic agonists. *J Med Chem*. 2001;44(26):4563-4576. http://www.ncbi.nlm.nih.gov/pubmed/11741475. Accessed November 27, 2015.

213. Moser U, Gubitz C, Galvan M, Immel-Sehr A, Lambrecht G, Mutshcler E. Aliphatic and heterocyclic analogues of arecaidine propargyl ester. Structure-activity relationships of mono- and bivalent ligands at muscarinic M1 (M4), M2 and M3 receptor subtypes. *Arzneimittelforschung*. 1995;45(4):449-455. http://www.ncbi.nlm.nih.gov/pubmed/7779140.