UNIVERSITÀ
DI SIENA
1240

University of Siena – Department of Medical Biotechnologies

Doctorate in Genetics, Oncology and Clinical Medicine (GenOMeC)

XXXV cycle (2019-2022)

Coordinator: Prof. Ariani Francesca

Molecular Dynamics simulation of the hERG channel assisting Precision Medicine in Channelopathies

Scientific disciplinary sector:

ING-INF/06 – ELECTRONIC AND INFORMATICS BIOENGINEERING

Supervisor	PhD Student
Prof. Furini Simone	**Pettini Francesco**

Academic Year 2021/2022

INDEX

Summary

In the last two decades, a revolution in biology has shifted the traditional reductive approach to a bottom-up study of virtual models. This discipline, known as System Biology, integrates information coming from individual components, in order to predict the functioning of biological systems, with the idea that complex systems are made up of many independent components that can interact within well-structured networks changing over time, and that the functional properties of biological systems emerge as a consequence of interactions among their components. This paradigm shift is enabled by rapid advancements in technologies providing high-throughput instruments able to analyse in detail biological processes at the single molecule and single cell scale. The vast amount of data produced by these experimental techniques asks for adequate methods of analyses. The present dissertation focues on structural based methods for simulating the functioning of biological molecules, and in particular on the role of Molecular Dynamics simulations. The advantage of Molecular Dynamics simulations is that it is based on physical description of the systems, and consequently it might offer an atomistic description of the process under investigation. The first chapter of this thesis will provide an introduction on the role of Molecular Genetics and Biology in Medicine, also considering new challenges for the prediction of protein interactions and for development of Precision Medicine. In the Second Chapters, Molecular Dynamics simulations will be discussed, with an emphasis on the methods for data analysis adopted in the research projects presented in the second part of the thesis. The third Chapter will be present the main research project produced during my PhD: the study of inactivation and drug binding in the hERG potassium channel. Side project and parallel collaborations are briefly discussed in the fourth Chapter, followed by concluding remarks.

PART I – *Introduction*

Chapter 1 - Molecular Genetics and Biology in Medicine

1.1. Genomics and bioinformatics contribution to Precision Medicine

We are living the era of omics science, a revolution in biology that has directed the traditional reductive approach to a *bottom-up* study of biological systems, known as Systems Biology. The new approach has been responsible for some of the most important developments in the science of human health. It is a holistic approach to deciphering the complexity of biological systems that starts from the understanding that living organisms are more than the sum of their parts. It is collaborative, integrating many scientific disciplines – biology, computer science, engineering, bioinformatics, physics, and others – to predict how these systems change over time and under varying conditions, and to develop solutions to the world's most pressing health and environmental issues (Oltvai & Barabási, 2002; Sauer et al., 2007; Tavassoly et al., 2018) (**Figure 1.1**).

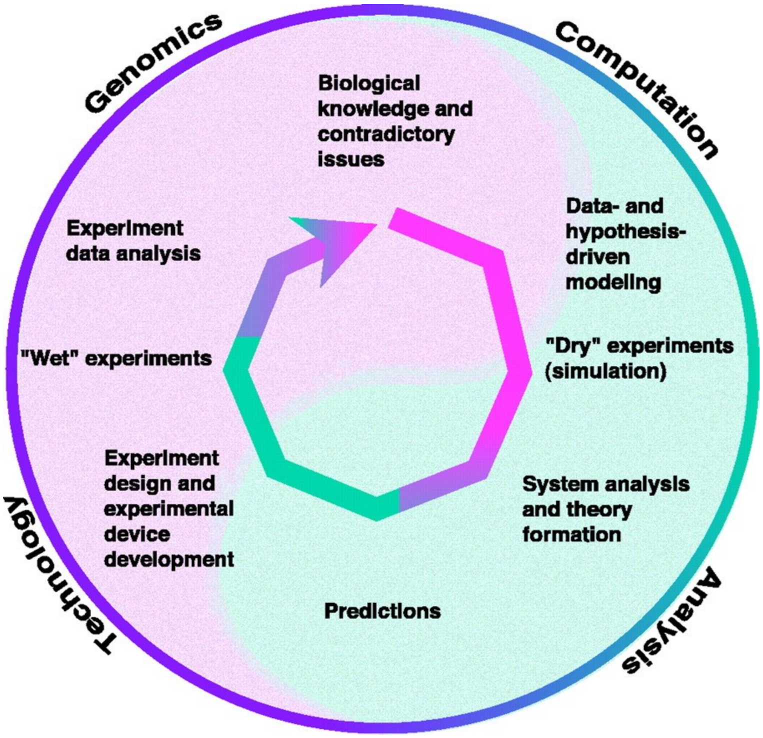
Genomics
Computation
Analysis
Technology
Biological knowledge and contradictory issues
Data- and hypothesis-driven modeling
"Dry" experiments (simulation)
System analysis and theory formation
Predictions
Experiment design and experimental device development
"Wet" experiments
Experiment data analysis

Figure 1.1 System Biology.

A fundamental tenet of systems biology is that solving challenging biological problems requires the development of new technologies in order to explore new dimensions of data space, and that new data types require novel analytical tools. This virtuous cycle of biology driving technology driving computation can exist only in a cross-disciplinary environment where biologists, chemists, computer scientists, engineers, mathematicians, physicists, physicians, contribute to tackle grand challenges (Hood et al., 2004). The Human Genome Project (Lander et al., 2001) could be considered a significant example of the result that can be achieved through the integration of new technologies in a multi-disciplinary open-source environment. The achievement of these extraordinary goals represents the starting point of an innovative process that is continuously developing. Deciphering the human genome, simultaneously to the complete examinations of the whole genomes of other numerous species (including numerous microorganisms, yeast, plants, and animals) with differing degrees of complexity, generated new techniques of data analyses (Xia, 2013), and several databases storing distinct genomes are now accessible, to allow for comparative study and reconstruction of evolutionary connections between single genes and genomes. Despite the researchers' original emphasis on the human genome's codifying region (< 3%), the examination expanded to the remainder area with the objective of discovering the significance of DNA sequences that are difficult to define as "useless" just because they are not codifying. Currently, genomic studies are focused on comparing the various genes, identifying sequences associated with a particular phenotype, and comprehending the function of the genes and their products (transcripts, proteins, and metabolites), as well as the complex interactions between them at multiple levels. Comparative analysis is possible at several stages: genomic, chromosomic, sub-chromosomic, and genic. For example, taking into consideration the genome of two closely related species, such as murine and human, and evaluating DNA sequences of single chromosomes, emerged that the order of the genes is conserved across the two species in many chromosomic areas. The identification of chromosomic regions in which the order and type of genes are preserved between two species enables fascinating comparisons between single genes. When two genes that derive from the same ancestral gene are compared (orthologous genes), it is possible to determine the presence of differences in their intron and exon composition, their regulatory sequences, and their codified product. There are, however, genes with sequence similarity that are not orthologous (in general, they have a different structure and function in the two organisms under consideration) and genes derived from the duplication of a particular gene in one specie and subsequent diversification from one specie to the next (paralogous genes). The idea of sequence alignment is that a given DNA sequence will undergo random mutations over time; two identical sequences will thus diverge over time, barring the presence of certain constraints, such as selection against mutations (modifications to a sequence that codes for a specific product may be harmful to the individual). As a result, one can expect some "phylogenetic conservation" in the case of functionally relevant sequences, while strong and fast divergence is expected for non-functional sequences. Comparing the genomes of distinct organisms is a necessary tool for reconstructing both the phylogenesis and the convolution of the evolutionary history. Apart from establishing similarity (and occasionally identity) between codifying sequences in two distinct species, the principle of evolutionary conservation of the sequence leads the search for significant, but non-codifying, sequences present in the genome. If functional annotations are required, they should be based on conserved phylogenetic sequences. While an organism's genome contains all the instructions for making proteins necessary for survival, interaction with the environment, and reproduction, protein expression allows individual cells to have distinct characteristics and functions. In fact, cells not only contain instructions for protein modification, but also information on the conditions in which amino acids must be synthesised. This information is encoded by complex regulatory and control mechanisms. An integrated analysis of these molecules (DNA, mRNA, and proteins) allows for new and solid experimental foundations for Systems Biology: gene sequences, how they are expressed and translated into proteins, the nature of these proteins and their function encode all information required for cell function and are therefore the events of interest. Genetic, genomic, and proteomic research aims to reveal the mechanisms that govern cellular processes to refine existing tools for early disease detection, prevention, and treatment development. Similarly, the development of pharmaceutical genomics and genetics is changing the way we treat diseases with a greater social impact, focusing on the development of new classes of drugs that can target biochemical pathways, or be personalized to an individual's genetic profile (Pirmohamed, 2001). A crucial step in all these scientific and technological advances has been the development of computerised systems for managing, analysing, and using heterogeneous and high-dimensional data generated by high-throughput tools, due to drastically reduce the time to transfer molecular biology and genomics information to practical clinical application. This progress steered the treatment of patients towards an approach called Precision Medicine (PM), due to the underlying concept of tailoring individual health according to medical model that use molecular profiling technologies for finding the right therapeutic strategy for the right person at the right time, and for maximizing the benefit-to-risk ratio (Stone, 2016). In particular, "The Precision Medicine Initiative", by the "National Institute of Health", reported (NIH, 2022):

"Precision medicine is an emerging approach for disease prevention and treatment that takes into account people's individual variations in genes, environment, and lifestyle."

Among the main goals of PM are to encourage research into a wide range of diseases, analysing pharmacogenomics, and finally create biological markers capable of objectively describing and accurately signalling disease risk. In this way, it will be possible to link genetic and environmental factors to a wide range of health outcomes. Notably, the goals of PM are not focused on a single patient, but rather on the categorization of individuals into subpopulations or patient groups. As reported from the U.S. National Research Council, PM rests on a "new taxonomy for human disease based on molecular biology" giving rise to the concept of "stratification". The term "stratification" refers to the ability to cluster patients in groups basing on specific biological features that can be found through molecular and biochemical diagnostic tests. To achieve stratification, a vast amount of relevant data must be considered. Even tiny or routine aspects cannot be ignored; age, ethnicity, and ancestral population membership, as well as geographic and social context and conditions should all be considered. In this setting, identifying biomarkers is required to identify patient subgroups. Biomarkers are biological indicators, with a unique molecular, anatomic, physiological, or biochemical characteristic, that can be detected and accurately evaluated in a variety of ways. From this perspective, in clinical field, biomarkers have a wide range of uses in diagnostic, prognostic, risk assessment, and predictive objectives, among others (Bahcall, 2015; Iyengar et al., 2015; NIH, 2022) (**Figure 1.2**).

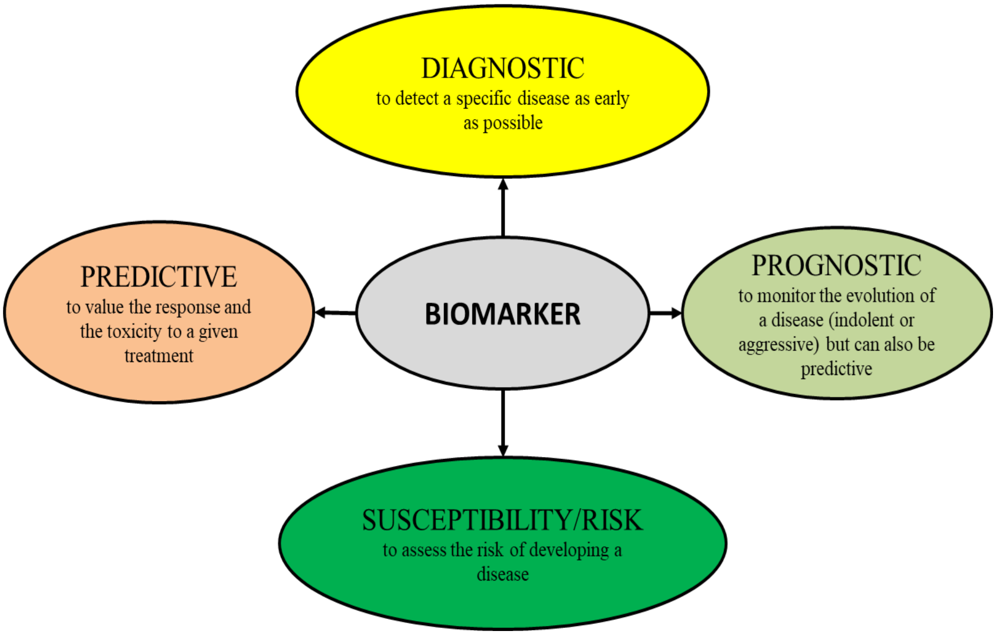
DIAGNOSTIC
to detect a specific disease as early as possible
PREDICTIVE
to value the response and the toxicity to a given treatment
BIOMARKER
PROGNOSTIC
to monitor the evolution of a disease (indolent or aggressive) but can also be predictive
SUSCEPTIBILITY/RISK
to assess the risk of developing a disease

Figure 1.2 Biomarker application in clinical field.

All these biomarker functions are possible because of genomics research that identifies and correlates certain indicators with clinical illness. However, genomic technologies may have limitations in some cases, such as false negative findings due to the high-dimensionality for data with respect to the number of samples usually available. Then, further approaches such as phenotypic studies, imaging, and functional in vivo studies, as well as scaling up population-based genome sequencing and combining it with clinical data, are necessary to enhance biomarker discovery for a particular clinical illness. Additionally, because it is ideal to identify biomarkers concurrently with drug research and development (a lengthy, expensive, and prone to failure process), a new discipline called Pharmacogenomic has lately gained increasing interest. This field of study is concerned with identifying genetic variations that change the pharmacokinetics or pharmacodynamics of drugs, by altering the target, or by perturbing a biological process (Bahcall, 2015). The identification of biomarkers remains a key element, especially in rare diseases. PM accomplishes this goal by establishing patient registries, utilising massive volumes of data to uncover possible correlations, and including patients as active participants. The associated process is complex and is distributed into four stages: discovery, development, validation, and application.

Metadata involved in this process are collected in the Precision Medicine Ecosystem, a virtual "ecosystem" of several databases, in which new information is constantly stored, based on the previously mentioned criteria of stratification (Iyengar et al., 2015).

1.2. Multiscale modelling

The development of multiscale models of cellular functions, starting from its molecular constituents, represents one of the most promising sectors of Systems Biology. Although the genomes of many organisms have been completely sequenced, we are not yet able to associate a role with many genes and/or proteins. The bioinformatic analyses of the homology of sequences or structures of a molecular element with unknown function to sequences, or structures, with already known function, is useful for directing research, restricting the field of investigation to a limited number of elements. This approach, however, does not suggest whether a certain function is accomplished through the interaction between certain molecules. Mathematical modelling and numerical simulations are powerful tools to address this question. Multiscale models of cell functioning allow, in fact, to integrate current knowledge on interactions between molecules to predict the effects that emerge from such interactions at the cellular level. To exemplify the transition from the atomic to the macromolecular scale, this thesis focuses on the study of the relationship between structure and function of a channel protein, using Molecular Dynamics (MD) simulations and related *in silico* techniques. MD simulations was performed because starting from a generic model of the physics driving interatomic interactions, MD simulations may disclose the locations of all the atoms at femtosecond temporal precision (Karplus & McCammon, 2002). In fact, there are several applications where MD simulations provide predictive information useful for investigating complex biomolecular processes or assisting experimental data (**Figure 1.3**).

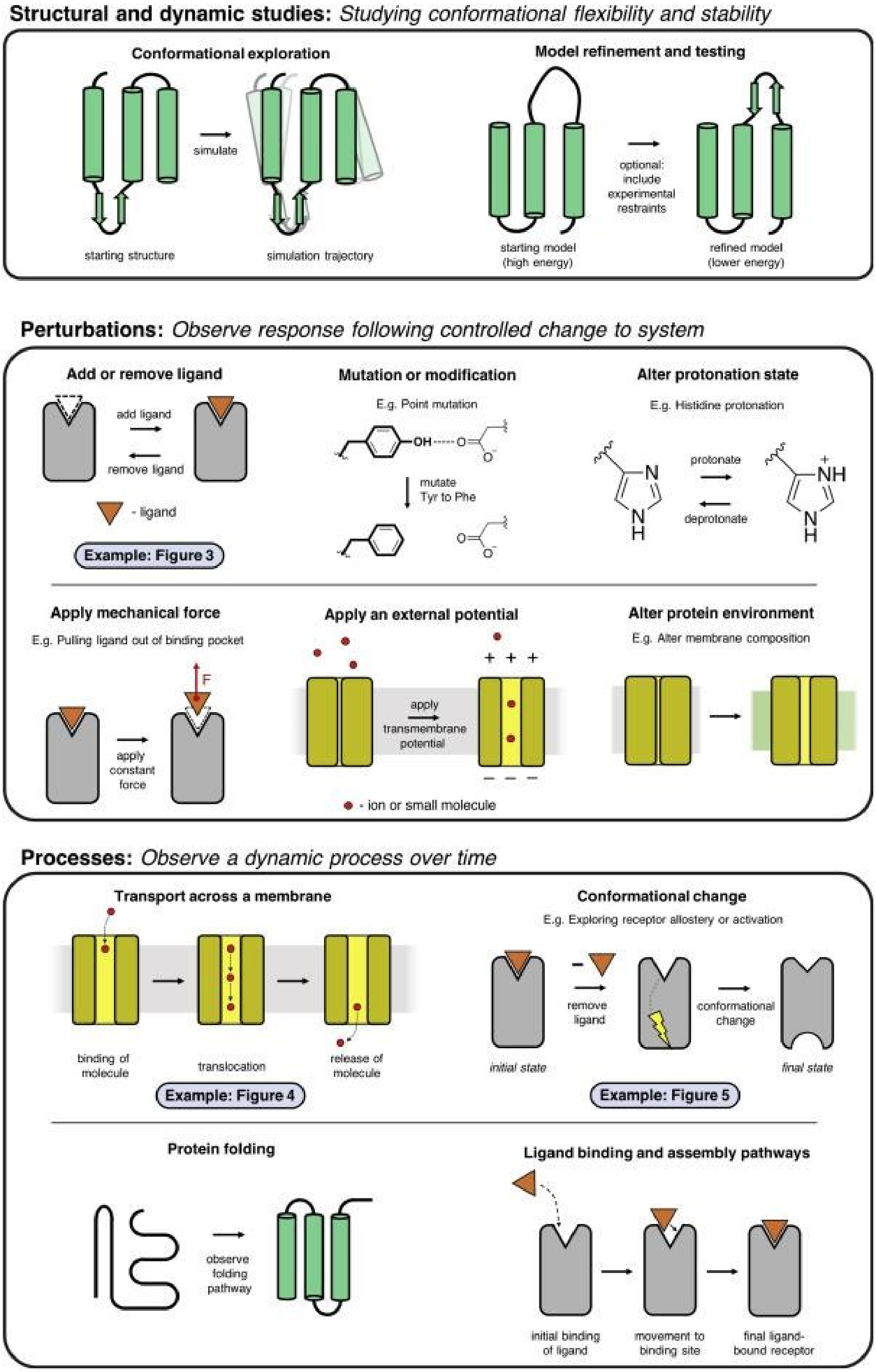
Structural and dynamic studies: Studying conformational flexibility and stability
Conformational exploration
simulate
starting structure
simulation trajectory
Model refinement and testing
optional: include experimental restraints
starting model (high energy)
refined model (lower energy)
Perturbations: Observe response following controlled change to system
Add or remove ligand
add ligand
remove ligand
- ligand
Example: Figure 3
Mutation or modification
E.g. Point mutation
OH
O
O⁻
mutate Tyr to Phe
Alter protonation state
E.g. Histidine protonation
N
NH
protonate
deprotonate
+
NH
Apply mechanical force
E.g. Pulling ligand out of binding pocket
F
apply constant force
Apply an external potential
+ + +
apply transmembrane potential
− − −
● - ion or small molecule
Alter protein environment
E.g. Alter membrane composition
Processes: Observe a dynamic process over time
Transport across a membrane
binding of molecule
translocation
release of molecule
Example: Figure 4
Conformational change
E.g. Exploring receptor allostery or activation
remove ligand
conformational change
initial state
final state
Example: Figure 5
Protein folding
observe folding pathway
Ligand binding and assembly pathways
initial binding of ligand
movement to binding site
final ligand-bound receptor

Figure 1.3. Applications of Molecular Dynamics simulations [*reprinted from Hollingsworth and Dror, 2018*].

Firstly, simulations can be employed to refine the accuracy of 3D structures of macromolecular entities or even investigating conformational flexibility and stability of various regions of a biomolecule, both if starting from an *ab initio* approach or not. In fact, crystal structures may suffer from purification and crystallization artifacts (Niedzialkowska et al., 2016), that can be corrected by performing a simulation starting from the initial coordinates and letting the structure to relax to a more favourable energetic (Burg et al., 2015). X-ray crystallography, for example, is commonly refined by an MD-based simulated annealing protocol that fits multiple models to the experimental data preserving the physical properties of the structure (Hao et al., 2015). Another approach is regularly in use to solve atomistic molecular models from low-resolution cryo-EM density maps, specifically when high-resolution configurations of distinct elements of a complex are independently available (Scapin et al., 2018). MD simulations are also employed to refine ensembles from NMR experiments (Lindorff-Larsen et al., 2005). In a similar attempt, MD simulation, as other specific structure-based methods, can be used to assess the binding pose of ligands when cryo-EM produce conformational space with ambiguous density (Koehl, 2018). Alternatively, MD simulation can be used to observe response following some controlled change to the system. Perturbation involves the addition of a bias, and often a partial restrains; generally, they request several replicates in both the perturbed and unperturbed systems to identify consistent differences in the results. A typical application of this approach is the identification of protein-ligand interaction profile from an experimentally determined structure or from a homology model. Other applications might be related to observe response of the system following mutation into the protonation state of an acidic or basic amino acid or adding some post-translational modification, as phosphorylation (Fields et al., 2017; Groban et al., 2006; Liu et al., 2015). In last instance, MD simulation are frequently used to observe a biomolecular dynamic process over time, such in ligand binding, ligand- or voltage-induced conformational change, protein folding, or membrane transport (Hollingsworth & Dror, 2018) (**Figure 1.3**).

The cellular function of all molecules-whether protein or otherwise- ultimately depends on interactions (a molecule that does not interact with any other component is irrelevant to the system). MD can study interactions because they are mediated by mechanisms and forces that also control conformational transitions. The binding to specific sites (allosteric) can modulate enzymatic activity in other parts of the protein, regulating its conformation and therefore the binding properties to other ligands. The study of interactions finds immediate biotechnological application in pharmacogenomics. In traditional approaches, new drugs are developed by random and multiple modifications of existing molecules and subsequent selection of active moieties through the experimental characterization of the interaction with specific receptors or target molecules. Thanks to the study of protein structures and MD simulations, it is possible to greatly cut time and costs by creating libraries of modified ligands *in silico* and predicting their interactions. In the case of drug-target interactions, MD are used to support the study of the molecular aspect of the binding of known drugs. In Drug Discovery, potential new drug target interactions (DTI) are predicted by other methods, ranging from ligand/receptor-based methods (Cheng et al., 2007; Wang et al., 2013) to gene ontology-based (Mutowo et al., 2016), text-mining-based methods (Zhu et al., 2005), and reverse virtual screening techniques (reverse-docking) (A. Lee et al., 2016; Wang et al., 2018), which are currently being developed. Docking calculations are commonly employed in receptor-based approaches, but they require 3D structures of target proteins, which are not always accessible. However, ligand-based techniques perform poorly when the number of known ligands is minimal, because they forecast DTIs based on the similarity of candidate to known target protein ligands. The fundamental limitation of both gene ontology and text mining techniques appears to be related to the query term. This is compounded by the frequent usage of duplicate drugs and target protein names. Moreover, since text-mining is confined to current information (i.e., published content), discovering new knowledge is difficult (Ruch, 2017).

1.3. Artificial Intelligence methods

Recent *in silico* approaches introduced Artificial Intelligence (AI) to address the limitations that involve large combinatorial spaces or nonlinear processes featuring structure-based techniques. AI refers to a procedure that creates an artificial system with a specific level of intelligence and employs computer software and hardware to replicate intelligent behaviours in a computer-simulated environment. AI-related algorithms can benchmark datasets with structural information provided in Protein Data Bank (PDB) (Berman et al., 2000; Westbrook et al., 2003) adding more information to validate structure-based methods via scoring functions and docking techniques, or to refine prediction in the absence of structural data. Learning techniques used in *in silico* studies might be classified as: supervised learning, unsupervised learning, semi-supervised learning, active learning, reinforcement learning, transfer learning, and multitask learning, with each class having its own benefits and limitations (X. Yang et al., 2019), also in relation with the particular task at hand. For example, in peptide-MHC binding (**Chapter 4.2**) prediction is more suitable to handle with Machine Learning (ML) modelling framework implemented by supervised algorithm such as NetMHCpan, that is a popular tool trained integrating datasets with all the binding affinity information and/or data retrieved from eluted ligands (Reynisson et al., 2020). These benchmark datasets are not only used to train models and check how well they work with standardised data, but also to be compared with state-of-the-art methods that are already in use to find the best performance. Taking up the previous DTIs example, there are three large datasets called BindingDB, Davis, and Kinase Inhibitor BioActivity that provide these binding affinities for interaction strength (Huang et al., 2021). In all three datasets, there are large-scale biochemical tests to see how well the kinase inhibitors work, because this protein family has more biological activity and is important for cancer cells communication (Tatar & Taskin Tok, 2019). As in other life sciences field, predictors based on sequence-data are data-driven, using AI and Deep Learning (DL) techniques for regression task rather than classification. AI and statistical analysis approaches are usually based on features or similarities. Known DTIs chemical descriptors for drugs and the descriptors for their targets are used by feature-based approaches to make feature vectors. Similarity-based AI techniques, on the other hand, employ the *guilt-by-association* criterion, which suppose that similar drugs tend to interact with similar targets and similar targets are targeted by similar drugs. Sometimes DL algorithms show better performance when compared to AI predictors (Thafar et al., 2019). Artificial Neural Networks (ANN) can be used for both supervised and unsupervised learning. They differ from each other in two main aspects. The first is related to the representation of input data by specific features, such as Simplified Molecular Input Line Entry System, Ligand Maximum Common Substructure, and Extended Connectivity Fingerprint in drug-target binding affinities prediction. The second is involving the different ANN types on which is implemented the system architecture (Krig, 2016). An ANN model comprises a series of connected layers; every node in the network performs a linear transformation of nodes in the previous layer and feeds it through a non-linear activation function. The shape of the layers, and the choice of activation function are static hyperparameters of the network's topology, whereas the weight of the linear transformations is tuned during training. A few of the most common ANN types include Feedforward, Radial Basis Function, Multilayer Perceptron, Recurrent, and Convolutional neural networks. Feedforward and Convolutional neural networks have been employed in the algorithms that predict drug-target binding affinities (Thafar et al., 2019).

Despite the last decade's rising interest in AI -based computational models developed for the research environment, the persistence of several limitations constrained performance. Fortunately, today open-science initiative spanning new vision of unifying resource to systematically access and evaluate AI methods across the entire range of therapeutics, as emphasized from the new-born Therapeutics Data Commons (TDC) (Huang et al., 2021):

"TDC supports the development of novel ML theory and methods, with a strong bent towards developing the mathematical foundations of which ML algorithms are most suitable for drug discovery applications and why."

Chapter 2 – Molecular Dynamics (MD) simulations

2.1 Molecular Dynamics and Force Fields

In Molecular Dynamics, molecules are described as a set of elementary particles, usually corresponding to atoms, and forces are calculated using simple functional forms. Dynamic is simulated by numerical integration of the Newton's equation of motion:

$$\frac{dp}{dt} = F \qquad Equation\ 2.1$$

where p is the array of momenta of all particles and F is the array of forces. In classical atomistic MD simulations, one or more atoms are represented as a unique particle with fixed mass and charge. The acceleration of each atom in the system at each timesteps is determined by considering the coordinates and velocity that the atom had in the steps before. The functional form and the corresponding parameters used to compute the interatomic forces are collectively known as the Force Field (FF). Parameters of the FF are identified by a combination of quantum mechanical calculations and fitting to experimental data (Frenkel et al., 1997). A simplified scheme of an algorithm for MD simulations is reported in **Figure 2.1**. Most of the software and FFs in use today for proteins and other organic molecules are based on pioneering studies starting in the 70's, when the evolution of experimental and parameterization techniques boosted FF developments in organic chemistry. The ECEPP potentials (Hagler et al., 1974; Momany et al., 1975), and the Consistent FF (Warshel & Lifson, 1970), represented important milestones in the field. In the 1980's, were first released some of the FFs still in use today, including CHemistry at HARvard Molecular Mechanics (CHARMM) (Brooks et al., 1983), Assisted Model Building with Energy Refinement (AMBER) (Cornell et al., 1995) and Optimized Potentials for Liquid Simulations All-Atoms (OPLS) (Jorgensen & Tirado-Rives, 1988). These FF are characterized by similar functional forms (**Equation 2.2**). Pragmatically, the potential energy of a molecule is described in terms of contributions from bonded (**Equation 2.3 and Figure 2.2**) and non-bonded interactions (**Equation 2.4**) (Frenkel et al., 1997):

$$E_{TOT} = E_{nonbonded} + E_{bonded} \qquad Equation\ 2.2$$

$$E_{nonbonded} = \sum\nolimits_{\substack{nonbonds \\ pairs\ ij}} \varepsilon_{ij} \left[\left(\frac{R_{min,ij}}{r_{ij}} \right)^{12} - 2 * \left(\frac{R_{min,ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{r_{ij}} \qquad Equation\ 2.3$$

$$E_{bonded} = \sum_{bonds} K_b (b - b_0)^2 + \sum_{angles} K_\theta (\theta - \theta_0)^2 K_\theta (\theta - \theta_0)^2$$

$$+ \sum_{dihedrals} K_\chi [1 + \cos(n\chi - \sigma)] \qquad Equation\ 2.4$$

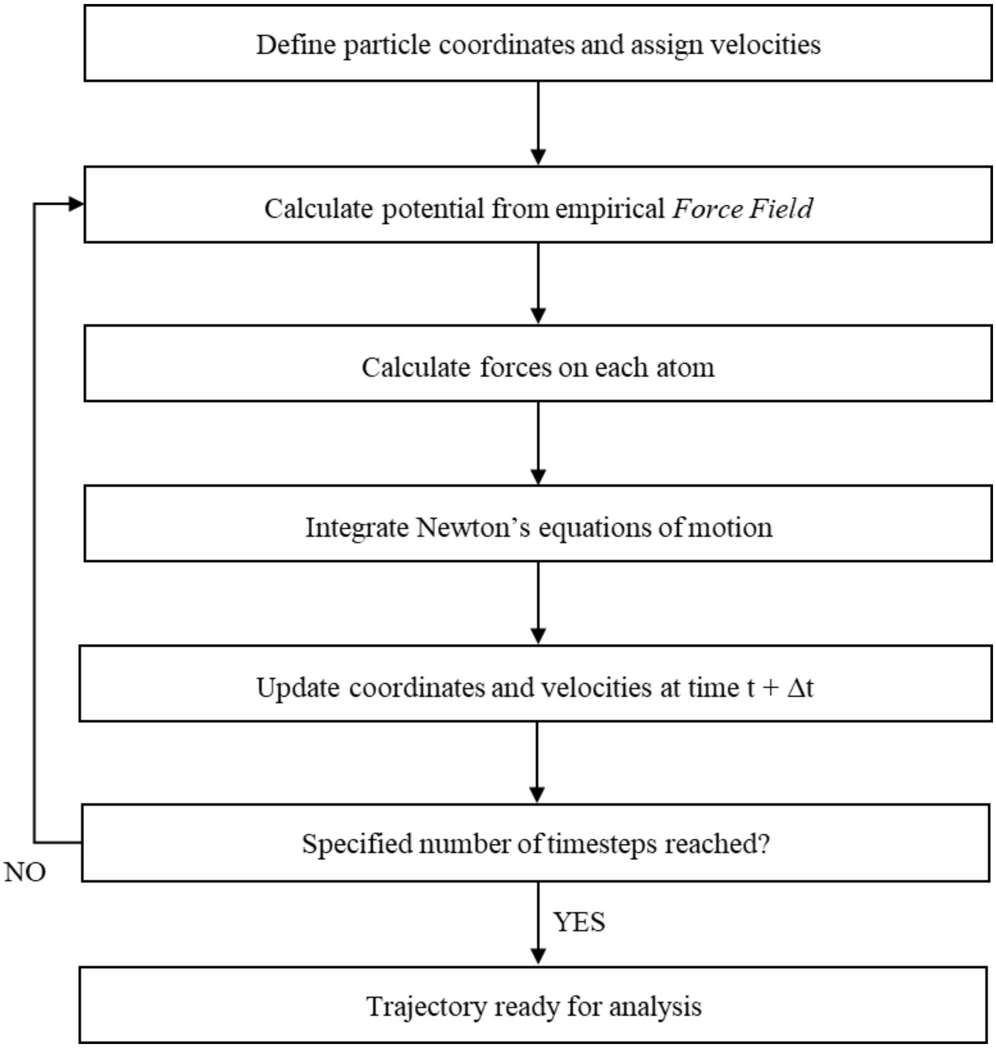
Define particle coordinates and assign velocities
Calculate potential from empirical *Force Field*
Calculate forces on each atom
Integrate Newton's equations of motion
Update coordinates and velocities at time t + Δt
Specified number of timesteps reached?
NO
YES
Trajectory ready for analysis

Figure 2.1. Overview of MDs algorithm.

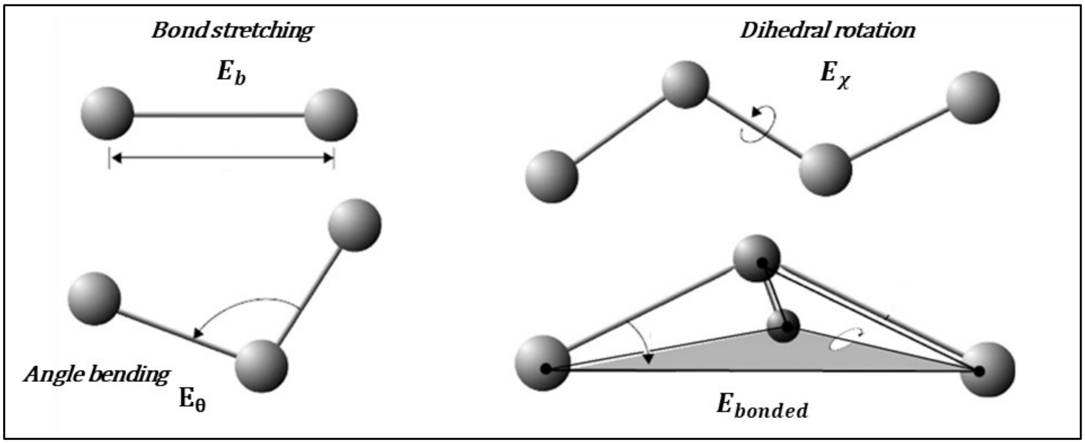

Figure 2.2. Schematic representation of the bonded terms contribution to a molecular mechanics' force field.

The first term in **Equation 2.3** defines the interaction between pairs of bonded atoms (stretch contributions) with a simple harmonic function that controls the length of covalent bonds *b* with respect to the reference value b_0. Estimates of b_0 can be obtained from X-ray diffraction experiments, while the spring constant may be estimated from infrared or Raman spectra. The harmonic potential is a good approximation for bond deformations smaller than 10% from the reference value. The use of harmonic functions implies that bonds are "hard-coded" in the initial system setup, and consequently, no change in the bonding structure can be simulated. This is one of the main limitations of FF-based MD simulations compared to *ab initio* methods. Occasionally, some other functional forms (in particular, the Morse's potential) might be employed to improve accuracy. Nevertheless, because of the higher computation cost of computing Morse's potentials, and the good approximation provided by harmonic functions for the simulation of biological systems, most of the existing FFs use the simpler harmonic functions.

In **Equation 2.3**, the angle bending contribution is also expressed by harmonic potentials. θ is the angle formed by the two bond vectors of the triplet of bonded atoms; K_θ and θ_0 are the parameters describing the force constant and equilibrium value of the angle. K_θ can be estimated, experimentally or theoretically, from vibrational analysis of the molecule. Compared to the force constants associated with bond stretching, the ones associated with angle bending are usually lower. The third term in **Equation 2.3** is the dihedral or torsional potential. It describes the energetic contributions associated with rotation of dihedral angles defined by each group of 4 bonded atoms. Torsional energy is usually estimated by trigonometric functions. In the case of **Equation 2.3**, χ is the torsional angle, σ is the phase, *n* specifies the number of minima or maxima between 0 and 2π, and $K\chi$ determines the peak of the potential energy barrier. $K\chi$ is usually obtained starting from *ab initio* calculations that are successively refined by fitting to experimental data.

The non-bonded terms (**Equation 2.4)** are, at least theoretically, calculated between any pair of atoms *i* and *j*, with position vectors, r_i and *rj.* Non-bonded interactions are due to two terms, respectively represented by the Lennard-Jones and electrostatic potentials. Classical 12-6 Lennard-Jones potentials are typically used to represent van der Waals interactions:

$$V_{LJ}(r) = \frac{C_{ij}^{12}}{r^{12}} - \frac{C_{ij}^{6}}{r^{6}} \qquad \text{Equation 2.5}$$

where $\frac{C_{ij}^{12}}{r^{12}}$ represent the Pauli repulsive term due to the short-range repulsive forces, while $\frac{C_{ij}^{6}}{r^{6}}$ provides the London attractive contribution due to dispersion force. In a widely used form of the Lennard-Jones potential, C^{12} and C^{6} values are replaced by σ and ε,

$$V_{LJ}(r) = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right] \qquad \text{Equation 2.6}$$

Where σ represents the distance that nullifies the potential, and ε is the energy value at the minimum point of the function. At long distances, the negative part $4\varepsilon_{ij}\left(\frac{\sigma_{ij}}{r}\right)^{6}$ prevails, while the positive part $4\varepsilon_{ij}\left(\frac{\sigma_{ij}}{r}\right)^{12}$ is predominant at short distances (**Figure 2.3)**. Values of σ and ε depend on the type of interacting atoms. Composition rules can be applied to define the parameters for the interaction between each pair of atoms starting from single-atom values. There are different types of combination rules; in Lorentz-Berthelot rule, a widely used one, the parameters for the pair of atoms i,j are obtained as:

$$\sigma_{ij} = \frac{\sigma_{ii} + \sigma_{jj}}{2} \qquad \text{Equation 2.7}$$

$$\varepsilon_{ij} = \sqrt{\varepsilon_{ii}\varepsilon_{jj}} \qquad \text{Equation 2.8}$$

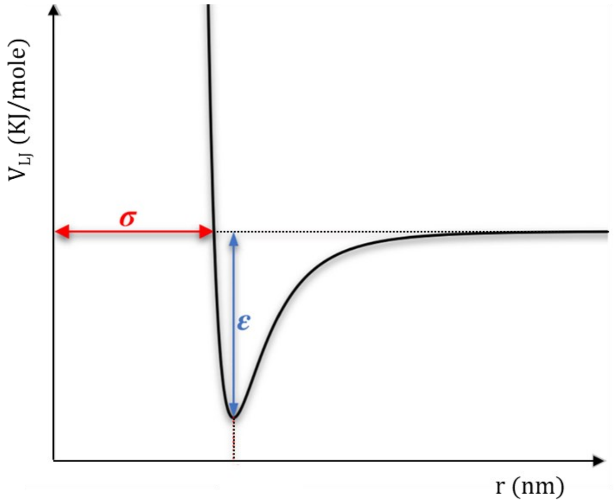

Figure 2.3. Lennard-Jones curve with parameters.

Non-bonding interactions between atoms covalently bonded or separated by two bonds are generally not included. In the case, on the other hand, of interactions between atoms separated by three bonds, the parameters of the LJ functions might be suitably reduced to avoid the strong repulsions that would occur at short distances. In Lennard-Jones potentials, the attractive term is due to the polarizability of the atoms which gives rise to an induced dipole-dipole interaction. The electronic cloud of the two atoms is organised on average to align the two dipoles, in order to minimise the electrostatic energy of polarisation. Based on the combination rules, **Equation 2.7** and **Equation 2.8**, it is possible to define non-bonded Lennard-Jones potentials among any pair of atoms, once the values for pairs of identical atom-types are defined. Based on the combination rule **Equation 2.7**, σ and ε are defined by purely atomic quantities. The Van der Waals radii as well as the ε parameter for the atoms inserted in a protein depend, as just mentioned, on the chemical context. For example, a hydrogen atom bonded to an oxygen will have a smaller σ than that associated with a hydrogen atom bonded to an aliphatic carbon. This is because the hydrogen of the hydroxyl has on average fewer electrons around it than a hydrogens atom bonded to a carbon (*sp^2*) which is a less electronegative atom than oxygen. The electrostatic potential in a dielectric model is estimated using Coulomb's law, as shown in **Equation 2.9** in vectorial form.

$$F = \frac{q_i q_j}{4\pi\varepsilon_0}\frac{r_i - r_j}{\left|r_i - r_j\right|^3} = \frac{q_i q_j}{4\pi\varepsilon_0}\frac{\hat{r}_{ij}}{\left|r_{ij}\right|^2} \qquad \mathit{Equation}\ 2.9$$

Where ε_0 is the dielectric permittivity in vacuum, which is approximately equal to $8*10^{-12}$ in the International System of Units (SI system), q_i and q_j are the partial charges of the atoms i and j, and r_{ij} is the distance between the atoms. Partial atomic charges arise due to the differences in the electronegativity of the atoms. For example, in the water molecule there are on average more electrons on oxygen than on hydrogens: oxygen in water, due to its greater electronegativity, has around it "on average" 8.8 electrons instead of 8 of the isolated atoms, while hydrogens have an "average" of 0.4 in place of 1 for the isolated atom. Experimental thermodynamic data, or methods based on quantum mechanics (i.e., variations that include electronegativity equalisation methods), can be used to derive partial charges for small molecules.

Nowadays, while FFs are quickly evolved and several versions of them are available for biomolecular MD simulation studies (**Figure 2.4**), CHARMM, AMBER and OPLS still remains the most used by researchers (Zerze et al., 2019).

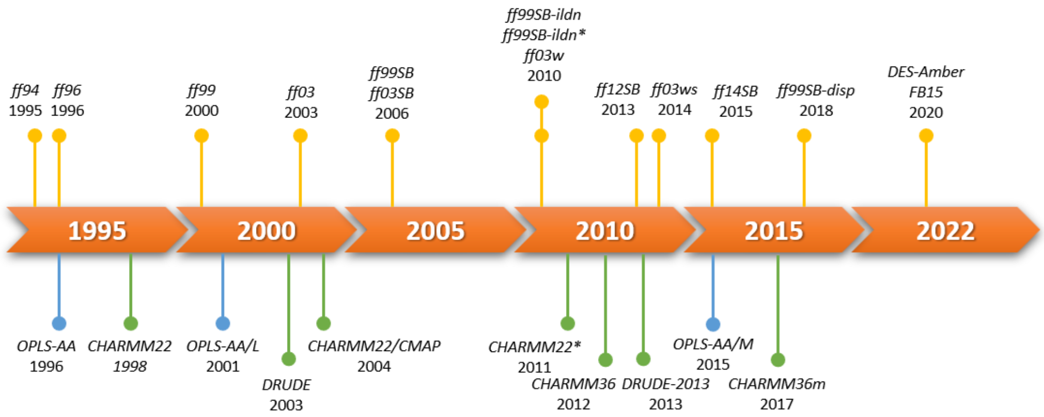

Figure 2.4. Overview of the improvements in the most popular all-atom force fields.

Evolution of the AMBER force field did not introduce new terms in the general equation discussed above, as shown in **Equation 2.10**, except for the optional electrostatic description of hydrogen bonds.

$$
\begin{aligned}
E_{TOT} = & \sum_{bonds} k_b(b-b_0)^2 + \sum_{angles} k_\theta(\theta-\theta_0)^2 \\
& + \sum_{dihedrals} K_\chi[1 + cos\, cos\,(n\chi-\sigma)\,] \\
& + \sum_{\substack{nonbonds \\ pairs\, ij}} \varepsilon_{ij}\left[\left(\frac{R_{min,ij}}{r_{ij}}\right)^{12} - 2*\left(\frac{R_{min,ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \\
& + \sum_{H-bonds}\left[\frac{R_{min,ij}}{r_{ij}^{12}} - \frac{R_{min,ij}}{r_{ij}^{10}}\right] \qquad Equation\ 2.10
\end{aligned}
$$

Pivotal improvements of AMBER in recent years regarded conformational preferences for typical secondary structures of amino acids by somewhat non intuitive parameterization of protein backbone dihedral angles. In Amber, each dihedral profile is defined by a set of four atoms. The set of atoms used to define φ and ψ for glycine is as expected, following φ and ψ along the main chain, while, for other amino acids that have a side chain, an additional set of dihedrals also influences rotation around the φ/ψ bonds connecting the C^α atom to the amide C and N atoms. This extra set of terms corresponds to dihedral angles branched out to the C^β carbon (Hornak et al., 2006). Refinement of these parameters was possible thanks to an increase in the resolution of experimental data.

The newest CHARMM36m version and the complementary CHARMM General Force Field (CGenFF) have also been widely optimized for proteins, lipids, and drug-like ligands (Huang et al., 2017). As shown by **Equation 2.11,** differences from the general energy function include Urey-Bradley terms improper terms, which were introduced in order to refine the representation of *in-plane* deformations and *out-of-plane* bending modes. Moreover, since CHARMM22, the 2D dihedral energy grid correction (CMAP) was introduced to improve the description of both backbone and sidechain dihedral angles (Mackerell et al., 2004; MacKerell et al., 2004).

$$E_{TOT} = \sum_{bonds} k_b(b - b_0)^2 + \sum_{UB} k_{UB}(S - S_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} K_\chi[1 + cos\, cos\,(n\chi - \sigma)\,] + \sum_{impropers} k_{imp}(\varphi - \varphi_0)^2 + \sum_{\substack{nonbonds\\ pairs\, ij}} \varepsilon_{ij}\left[\left(\frac{R_{min,ij}}{r_{ij}}\right)^{12} - \left(\frac{R_{min,ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\varepsilon_0 r_{ij}} \qquad Equation\ 2.11$$

where k_{UB} and k_{imp} are force constants; *S* is the 1,3 distance; ϕ is the improper angle. As the previous two force fields, AMBER and CHARMM, OPLS also used a united atom representation initially and later moved to an all-atom representation. It is implemented directly with AMBER force field parameters for the bonded interactions terms, initially with the 1984 (UA version) and with 1986 Amber force field later (AA version), and uses neutral charge groups, like CHARMM. The OPLS-AA potential form is:

$$E_{TOT} = \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} K_\chi[1 + cos\,(n\chi - \sigma)\,] + \sum_{impropers} k_{imp}(\varphi - \varphi_0)^2 + \sum_{\substack{nonbonds\\ pairs\, ij}} f_{ij}\left\{4\varepsilon_{ij}\left[\left(\frac{R_{min,ij}}{r_{ij}}\right)^{12} - \left(\frac{R_{min,ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\varepsilon_0 r_{ij}}\right\} \qquad Equation\ 2.12$$

The non-bonding interactions contain the f_{ij} factor ($f_{ij} = 0$ if atoms *i* and *j* are within the same molecule and separated by less than three bonds; $f_{ij} = 0.5$ if the two atoms are within the same molecule and separated by exactly three bonds; $f_{ij} = 1$ if atoms are either not in the same molecule, or they are separated by more than three bonds).

The previously described FFs have been successful in simulations of globular proteins and short peptides (Beauchamp et al., 2012) and are mature enough that protein folding simulations of small single domain proteins are feasible (Nguyen et al., 2014). More recently, however, detailed experimental data have revealed that these FFs have deficiencies in simulating intrinsically disordered proteins (Henriques et al., 2015; Rauscher et al., 2015), protein folding equilibria and their dependence on temperature (Lindorff-Larsen et al., 2005), and correctly identifying protein folding pathways/intermediates (McKiernan et al., 2017). The residue-specific FF (RSFF1) (Jiang et al., 2014), based on conformational local free-energy distributions of the 20 amino acids from the OPLS-AA FF, introduced single specific parameters from AMBER FF to refine the stability of both α-helix and β-sheet of residues for obtaining folding enthalpies and entropies in reasonably good agreement with available experimental results. These adjustments resulted in the RSFF2 FF (Zhou et al., 2015). Beside the development of this alternative approach to improve accuracy of FF, there are some few specialised FFs in common use today for sugars, nucleic acids, or protein folding, that remain anchored to an extended-atom model, in which hydrogen atoms can be ignored to varying degrees, with the non-bonding parameters of the atom coupled to the hydrogen adjusted accordingly. While, in some cases, these FFs will explicitly include polar hydrogen atoms helping treatment of hydrogen bonding, certain interactions (i.e, π-π stacking) will be poorly treated leading to inaccurate simulations (Burley & Petsko, 1985). This is the case of the latest 53A5 and 53A6 versions of GROMOS (GROningen MOlecular Simulation) (van Gunsteren & Berendsen, 1987), that are widely used FFs for simulations of protein folding, or other protein structural changes. They are parameterized to reproduce free energies of solvation, using two different sets of partial atomic charges in polar and apolar solvents; consequently, it is advantageous only when considering a homogeneous environment due to the ambiguity in setting the charges of the molecule.

Another fundamental aspect of FF parameterization is chemical perception: the process by which molecular simulation software recognizes the chemistry of a molecule with the goal of assigning appropriate FF parameters for that molecule. Up to now this has been done indirectly by first assigning predefined atom types to the molecule and then using these atom types to assign parameters. An alternative route is the use of direct chemical perception that automatically recognizes the entirety - or at least a large fragment - of the molecule and assigns parameters accordingly. Such an approach could substantially reduce the number of unique atom types and parameters necessary for biomolecular simulations, especially those involving small molecule ligands. Currently this approach is being pursued by Mobley et al. in the new SMIRNOFF format (Mobley et al., 2018), which uses the SMIRKS chemical query language to define molecular structure and topology. The use of a chemical query language enables the parameterization engine to identify molecular substructures and assign parameters directly to them, thereby rendering large numbers of pre-defined atom types unnecessary (Nerenberg & Head-Gordon, 2018).

In the next few years, FFs are expected to face some major challenges. Many challenges relate to accurate estimation of thermodynamic quantities of interest (Nerenberg & Head-Gordon, 2018). Real physical systems polarise substantially when placed in a high-dielectric medium such as water, or even when a strongly charged system approaches a neutral body in the gas phase. The atomic charges are often increased in current biomolecular force fields to create molecular or fragment dipole moments that are roughly 10–20 percent greater than those seen in the gas phase, which just serves to average out this polarisation. Given that the frequently used non polarizable water models contain such charges, increased charges are required to accurately explain the bulk properties of liquid water and to achieve a proper balance between solvent-solvent, solute-solvent, and solute-solute interactions. The dielectric environment, and the polarisation response that results from it, can differ greatly across a biomolecular system. For instance, it may range from the nearly gas-phase environment of a nonpolar pocket in the protein interior to a nearly bulk-water environment at the protein surface. This heterogeneity limits the accuracy of the mean field approximation used by classical force fields. In addition, including polarizability in the gas phase environment can have a strong effect on the energy of intramolecular interactions, as Caldwell and Kollman showed in a seminal study on aromatic–cation interactions (Caldwell & Kollman, 1995). In such situations, ‘one size’, in atomic charge, does not ‘fit all’ and yet every widely used FF makes this approximation. Polarizable force fields, in contrast, allow the charge distribution to vary with the demands of the local environment (Halgren & Damm, 2001). These were introduced over 20 years ago, but recent improvements in CPU/GPU parallelism have made them available for use in more complex systems. They show good agreement with experimental solvation and free binding energy (Albaugh et al., 2016; Bradshaw & Essex, 2016) and, in principle, provide the flexibility needed to capture the subtle energy landscapes of folded proteins, IDPs and protein folding intermediates (Nerenberg & Head-Gordon, 2018). Several methods exist for modelling induced polarisation, such as the classical Drude oscillator model. In the Drude-2013 polarizable FF, which is derived from the CHARMM additive force field, the electronic degrees of freedom are modelled by charge particles attached to the nucleus of their central atom by a harmonic spring (Lamoureux, 2003). The Drude model has the benefit of preserving the simple particle-particle Coulomb electrostatic interaction used in non-polarizable force fields, making its implementation in MD engines easier than alternative polarisation models. For instance, the MD engine NAMD conducts Drude oscillator integration by using a new dual Langevin thermostat to freeze the Drude oscillators while keeping the heated degrees of freedom at the proper temperature (Jiang et al., 2011). The Drude polarizable force field requires some extensions to the CHARMM force field (Lamoureux, 2003). The main difference between Drude oscillators and normal spring bonds is that they have a zero-equilibrium length. A maximal bond length parameter is optionally added to the Drude oscillators, beyond which a quartic restraining potential is additionally applied. An anisotropic spring term provides for out-of-plane forces between a polarised atom and its covalently bonded neighbour with two additional covalently bonded neighbours (similar in structure to an improper bond). Thole's screened Coulomb correction is computed between pairs of Drude oscillators that would normally be excluded from non-bonded contact, as well as between non-excluded, nonbonded pairs of Drude oscillators that are within a defined cut-off distance (Thole, 1981; van Duijnen & Swart, 1998). The use of off-centred massless interaction sites, known as “lone pairs” (LPs), to escape the restrictions of centrosymmetric-based Coulomb interactions is also incorporated in the Drude force field (Harder et al., 2006). Each LP site's coordinate is built from three "guide" atoms. The forces estimated on the massless LP must be transmitted to the guide atoms while maintaining total force and torque. The location of the LP is updated based on the three guide atoms, as well as additional geometry parameters that offer displacement and in-plane and out-of-plane angles (Harder et al., 2006) after an integration step of velocities and positions. The implementation of algorithms for simulations with the Drude-2013 force field in CHARMM, NAMD, OpenMM, and GROMACS, in conjunction with available input generation servers such as the "Drude Prepper" in the CHARMM-GUI and automated parameterization in GAAMP, enables widespread use throughout the theoretical chemistry community (Lemkul et al., 2016).

2.2 Simulation Protocol

Before the production run, there are usually three phases necessary to set up the system and equilibrate it. In the first phase, the system is incorporated in a bulk liquid, and through the use of a periodic boundary conditions (PBC) a theoretically infinite system is simulated by using unit cells (Frenkel et al., 1997). In PBC the unit cell is surrounded by translated copies in all directions to approximate an infinitely large system. When one molecule diffuses across the boundary of the simulation box it reappears on the opposite side. Each molecule interacts with its neighbours even though they may be on opposite sides of the simulation box. This approach replaces the artefacts present when simulating isolated systems in vacuum with PBC artefacts, which are in general much less severe (Braun et al., 2019). There are several possible shapes for space-filling unit cells. Some periodic cells, like the rhombic dodecahedron and the truncated octahedron are better suited to the study of approximately spherical macromolecules in solution, since fewer solvent molecules are required to fill the box given a minimum distance between macromolecular images. At the same time, rhombic dodecahedra and truncated octahedra are the least symmetric unit cells of all types of periodic boxes. The unit cells adopted in PBC can be defined by three basis vectors, named in the following equations a, b, and c (Frenkel et al., 1997). The box vectors must satisfy the following conditions:

$$a_y = a_z = b_z = 0 \qquad Equation\ 2.13$$

$$a_x > 0,\ b_y > 0,\ c_z > 0 \qquad Equation\ 2.14$$

$$|b_x| \leq \frac{1}{2}a_x,\ |c_x| \leq \frac{1}{2}a_x,\ |c_y| \leq \frac{1}{2}b_y \qquad Equation\ 2.15$$

Equation 2.13 can be satisfied by rotating the box, while inequalities **Equation 2.14** and **Equation 2.15** can be satisfied by adding and subtracting box vectors (Deiters, 2013). It is also possible to simulate without periodic boundary conditions, but it is usually more efficient to simulate an isolated cluster of molecules in a large periodic box, since searching for pairs of atoms that are within a certain cut-off radius of each other can only be used in a periodic system. According to the minimum image convention, the cut-off radius used to truncate non-bonded interactions may not be greater than half the length of the shortest box vector (**Equation 2.16**), since several images would otherwise be contained inside the force's cut-off distance (Deiters, 2013).

$$R_c < \frac{1}{2}min(\|a\|, \|b\|, \|c\|) \qquad Equation\ 2.16$$

This constraint alone is insufficient when a macromolecule, such as a protein, is investigated in solution, since, in theory, a single solvent molecule should not be able to observe both sides of the macromolecule (de Souza & Ornstein, 1997). Accordingly, each box vector must be longer than the length of the macromolecule facing that edge plus twice the cut-off radius R_c. However, it is typical to adopt a compromise in this area and lower the solvent layer's size to reduce the computational cost.

In the second phase of a classical simulation protocol, the potential energy of the system is minimized. The aim of this minimization step is to find a local energy minimum in molecular conformations by identifying a path that, through the variation of molecular freedom levels, leads, with the least possible number of calculations, to the nearest local minimum (Mackay et al., 1989). Algorithms for minimization often locate the minimum closest to the starting structure, as remote minima separated from the initial configuration by energetic barriers are difficult to identify, since their investigation entails a rise in the gradient of the function's energy. Several minimization algorithms have been developed, with some of the most common in MD simulations being Steepest Descent and Conjugate Gradient (Curry, 1944; Hestenes & Stiefel, 1952). In most cases, the minimization step is only needed to remove possible steric clashes that could cause instabilities in case of numerical integration of the equations of motion, and consequently any algorithm capable of finding a local minimum of the potential energy can be used (Frenkel et al., 1997).

Once the system has reached a local energy minimum, and steric clashes are removed, it is possible to start the following phase, usually referred to as Equilibration. The name refers to the fact that the aim of this phase is to move the system towards an equilibrium state, at the desired thermodynamic conductions (for instance of temperature and density) (Nosé & Klein, 1986; Rice & Sewell, 2008). The equilibration of the pressure is usually more complicated than the equilibration of the temperature, with fluctuations in the pressure value with respect to the reference value being common. Therefore, all properties calculated from pressure, such as interfacial tension, could be more very difficult to estimate. One of the most common practices is to carry out the Equilibration of the system in two stages: in the first, an NVT (Canonical ensemble: constant number of molecules, volume, and temperature) simulation is carried out to bring the temperature to the desired value; in the second, an NPT (Isothermal-Isobaric ensemble: constant number of molecules, pressure and temperature) simulation is carried out to balance the density of the system. During the equilibration, the solute molecules (e.g. proteins or other macromolecules of interest) might be (at least initially) restrained to their experimental positions by inserting energy constraints. In this way, all the solvent molecules adapt to their positions to bring the quantities listed above to equilibrium. In order to obtain the desired value of pressure and temperature, a barostat and a thermostat are used respectively. Thermostat operates as a simulated heat bath, keeping the average temperature at a set value, for instance using **Equation 2.17**, while barostat can be applied to keep the pressure constant by adapting the volume, for instance according to **Equation 2.18** (Hoover, 1985).

$$\lambda_T = \sqrt{1 + \frac{\Delta t}{\tau_{bath}}\left(\frac{T_{bath}}{T} - 1\right)} \qquad Equation\ 2.17$$

$$\lambda_P = 1 + \frac{k\Delta T}{\tau_{pressurebath}}(p_{bath} - p) \qquad Equation\ 2.18$$

Unlike thermostats, a barostat always applies to the entire system. Among the common methods, the Berendsen friction thermostat (Berendsen et al., 1984) and the Parrinello-Rhaman barostat (Parrinello & Rahman, 1981), are described by **Equations 2.19** and **Equation 2.20**:

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau} \qquad Equation\ 2.19$$

where T_0 and T are respectively the reference temperature and the current temperature of the system, while τ is a time constant that depends on the total heat capacity of the system and on the total number of degrees of freedom;

$$\frac{db^2}{dt^2} = \frac{V(p - p_{ref})}{Wb'} \qquad Equation\ 2.20$$

where the volume of the box is denoted V, and W is a matrix that determines the strength of the coupling. The values p and p_{ref} are the current and reference pressures, respectively.

2.3 Analyses of MD trajectories

Analysing MDs results is difficult for several reasons. A typical simulation tracks hundreds of thousands of atoms positions and velocities across billions of time steps, generating a large amount of data. In several circumstances, it is complicated to identify the most relevant and physiologically significant information, such as calculating interaction energy between a ligand and a protein to understand a functional mechanism of a pathway. Thus, to maximise the insights obtained from simulations, they must be interpreted considering all available experimental evidence (and, often, related systems as well). The analysis requires a good balance of visual and quantitative analysis. Most simulation projects benefit from building bespoke analysis programmes or scripts, which is made easier by numerous analysis software frameworks. Alongside this, a paradigm shift has begun to evolve, which involves the simulation of many short trajectories in parallel instead of a single long trajectory. Given that much of the calculation is wasted with systems trapped in an energy minimum waiting for rare events, the idea is to start many simulations in parallel of the same system, and then, as soon as a simulation escapes the minimum, stop the remaining trajectories converging in the new state. The new paradigm has found its maximum expression in the use of discrete state and of discrete time stochastic master equation models. The construction of these models involves: a discretization of the trajectory in low-dimensional set of collective variables describing the essential position and velocities of all atoms, and the calculation of the transition probability matrix with respect to the lag-time, chosen so that the transition probability depends only on the current state and not on the entire trajectory. The resulting matrix approximates the dynamic behaviour of the system, allowing to extrapolate the time scales of the slower processes and the probability distribution at equilibrium, from which it is possible to evaluate thermodynamics and kinetics data. From a qualitative point of view, this type of analysis allows building multi-stage models of the process under examination, identifying metastable states of the system (Husic & Pande, 2018; Kitao, 2022; Wu et al., 2017).

Data featurization can be performed using several methods ranging from counting the number of atoms in a particular space, such as the number of water molecules in the binding site of a channel protein, to the set of collective variables describing the essential position and velocities of all atoms. The Root Mean Square Deviation (RMSD) is the most often used expression for measuring the structural similarity between two molecular conformations (Kabsch, 1976). It is the average atom displacement from a reference structure, generally the initial frame of the simulation or an experimental structure (Damm & Carlson, 2006). It can be calculated for any type and subset of atoms, i.e. Cα or all heavy atoms of the entire protein, Cα atoms of all residues in a specific subset (e.g. the transmembrane helices, binding pocket, or a loop), by:

$$RMSD(t) = \sqrt{\frac{1}{M}\sum_{i=1}^{N} \|r_i(t) - r_i(0)\|^2} \qquad Equation\ 2.21$$

where $M = \sum m_i$, $with\ m_i$ being the atomic mass of each atom of the ensemble; N the total number of atoms, vector $r_i(0)$ is the reference position of the i-th particle, vector $r_i(t)$ is the position of the i-th particle at time t. RMSD is computed in two steps: alignment and optimum superposition. Aligning two conformations involves matching equivalent atoms, while optimal superposition corresponds to finding rotation and translation of one structure to minimize the weighted sum of the squares of the distances between equivalent atoms in the two structures (Coutsias & Wester, 2019). The RMSD is a valuable tool for analysing the structure's time-dependent movements. It is widely used to determine if a structure is stable or if it is deviating from the starting coordinates. Usually, a drift from its original coordinates is considered as an indication that the simulation is not equilibrated (Kabsch, 1976). When a simulation is equilibrated, or rather when the system moves around a stable average conformation, it makes sense to calculate the fluctuations of each subset of atoms relative to the average structure of the simulation, called Root Mean Square Fluctuation (RMSF) (Pitera, 2014):

$$RMSF_i = \sqrt{\frac{1}{M}\sum_{i=1}^{N} \|r_i(t) - r_i\|^2} \qquad Equation\ 2.22$$

where r_i is the average position of the i-th atom.

RMSD or RMSF calculations for proteins often include the rigid-body alignment of the structures in each frame of the simulation to reference coordinates. High RMSDs or RMSFs may suggest that the entire structure is not equilibrated, or they may represent just massive displacements of a tiny structural subgroup within an overall stiff structure, respectively (Pitera, 2014). It becomes increasingly typical, as the number of structures investigated by MD simulations grows, to find high RMSDs that are due to substantial fluctuations in structural subsets that do not represent the structural fluctuations of the macromolecule. Because finding an optimal superimposition is an ambiguous task with several solutions each improving a specific parameter, all superimposition dependent approaches are restrained by the ambiguity of the problem. Methods that are not dependent on superimposition, such as contact-based measurements, are not affected by this issue. Additionally, concentrating on local similarities can help to prevent these problems. A local similarity score can be read as a sum of similarity scores for all sections of a protein or, alternatively, it can be used to focus on a single portion of the protein, such as, for example, a ligand binding pocket, while disregarding the rest of the protein's structure.

After the featurization, the following task is usually to identify relevant combination of features that could accurately describe different states of the simulated system. A common method to reduce the number of dimensions in this step is Principal Component Analysis (PCA) (Jolliffe & Cadima, 2016). This mathematical method is used to reduce the number of dimensions by projecting the coordinates on a linear subspace of the largest-amplitude motions. Dimensionality-reduction of a data set naturally comes at the expense of accuracy, but with the advantage of trading accuracy for simplicity. The idea of PCA is to simply reduce the number of variables of a data set, while preserving as much information as possible, following a four steps protocol:

- Standardise the range of the continuous initial features thus all the variables will be transformed to the same scale.
- Compute the covariance matrix aimed to identify a correlation among the variables of the input data set.
- Compute the eigenvectors and eigenvalues of the covariance matrix and order them in descending order. to identify the principal components in order of significance. The first coordinates of this new space are those having the most significant variances, or fluctuations.
- Project the data along the principal components' axes; the aim is to use the feature vector formed using the eigenvectors of the covariance matrix, to reorient the data from the

original axes to the ones represented by the principal components (hence the name Principal Components Analysis).

In MDs, PCA has been used successfully to improve the accuracy of Markov State Models (MSM), even though there are no universal assurance that large-amplitude movements are linked with slow transitions (Kitao, 2022). As the aim is usually to identify states of the simulated systems that correctly describe slow state transitions, it would be helpful to directly use a metric that offers a good indicator of the slow processes and that allows for a proper approximation of the eigenfunctions to be achieved with a moderate number of clusters. With respect to this problem, Time-Lagged Independent Component Analysis (TICA) is one of the most useful improvements to cluster the slow transitions (Kitao, 2022; Molgedey & Schuster, 1994; Pérez-Hernández et al., 2013; Schwantes & Pande, 2013). TICA defines a linear transform of some (usually high-dimensional) set of input coordinates to some (usually low-dimensional) set of output coordinates. The transformation is chosen such that amongst all linear transforms, TICA maximizes the autocorrelation of transformed coordinates at a certain lag time τ. In other words, TICA finds a subspace of slow coordinates, or a subspace of good reaction coordinates, when the input coordinates come from MD. Given a sequence of multivariate data X_t, it computes the mean-free covariance and time-lagged covariance matrix:

$$C_0 = (X_t - \mu)^T diag(\omega)(X_t - \mu) \qquad Equation\ 2.23$$

$$C_\tau = (X_t - \mu)^T diag(\omega)(X_t + \tau - \mu) \qquad Equation\ 2.24$$

where ω is a vector of weights for each time step. The weights are all equal to one by default, although other weights are possible (Wu et al., 2017). Subsequently, the eigenvalue problem is solved by:

$$C_\tau r_i = C_0 \lambda_i r_i \qquad Equation\ 2.25$$

where r_i are the independent components and λ_i are their respective normalised time-autocorrelations.

The eigenvalues of the transition matrix of a MSM are related to the relaxation timescale by:

$$t_i = -\frac{\tau}{\ln|\lambda_i|} \qquad Equation\ 2.26$$

The input data is projected onto the dominating independent components when performed as a dimension reduction approach. Because the eigenvalues λ_i are indications of the slowness of their respective processes, these prominent components correspond to the slowest processes in the data. As shown in **Section 3.4.2**, TICA was crucial to discern the most significant clusters of the trajectory in simulations of potassium channels.

2.4 Estimate of free energies

The free energy is often considered to be the most important quantity in thermodynamics, as the difference in free energy between two states is what determines their relative probability. This also determines whether a process is energetically favourable so that work is obtainable from it, or unfavourable so that work needs to be done for the process to take place. The free energy of a system is directly related to its partition function. In the NVT and NPT ensembles respectively the Helmholtz free energy (A) and the Gibbs free energy (G) are used:

$$A(T) = U - TS \qquad \textit{Equation } 2.28$$

$$G(p,T) = A + pV = U - TS + pV = H + pV \qquad \textit{Equation } 2.29$$

where S is entropy of the system, T is temperature, H is the enthalpy, U is the internal energy, p is the pressure, and V the volume.

These thermodynamic variables are the most complicated to calculate in numerical simulations due to difficulties in determining the dimensionality of a system. Unlike what happens for the analyses inherent to the mechanical variables of the system, where it makes sense to use a partitioned approach on each frame to extrapolate the overall picture of the dynamics of a system, in thermodynamics it is possible to work only on the set of microstates of the system by reconstructing the energy profile. Indeed, the entropy is defined by:

$$S = k_B \ln \Omega \qquad \textit{Equation } 2.30$$

Where Ω is the number of microstates compatible to the observed thermodynamic macrostate, and k_B is the Boltzmann constant. MD simulations of a molecular system have a propensity to explore restricted regions of the configurational space, which cause a poor estimation of the number of accessible microstates. Many methods have been developed to improve free energies calculation, distinguished by different trade-offs between accuracy and computational costs (Hall et al., 2020; King et al., 2021). This dissertation is focused on the prediction of the biological activity of a ligand (small-molecules and drugs), which is related to its affinity for a protein target; so, in the next sub-sections it will be discussed two of the commonly used scoring approaches, based on the post-processing of snapshots from MD by a function that employs a physical or statistical potential.

2.4.1 Implicit solvent end-point free energy methods

Among all ensemble-based simulation approaches, the Molecular Mechanics Poisson-Boltzmann Surface Area (MMPBSA) end-point free energy methods is recognized as an effective and reliable tool for modelling binding events, such as protein-ligand binding interactions (Gohlke & Klebe, 2002; Kollman et al., 2000). The development of techniques for calculating binding free energies has been a primary focus of molecular simulations. Indeed, both the free energy perturbation and thermodynamic integration methods, which are theoretically rigorous but computationally demanding, were introduced considerably earlier than the MMPBSA approach. These approaches have been proved to be more accurate than the MMPBSA method for basic and small systems when they can be implemented reliably. Their applicability to complex biomolecular systems, on the other hand, is hampered by the high computational costs. The MMPBSA approach makes many critical approximations that enable it to be used as an efficient and approximation for free energy calculations (Swanson et al., 2004). The PBSA model is used to estimate the contribution of solvation to the free energy using a continuum solvent model (Genheden et al., 2011; Guimarães & Mathiowetz, 2010). Additionally, the technique approximates the contributions of enthalpy and entropy to the free energy independently (Genheden & Ryde, 2012; Swanson et al., 2004). The MMPBSA technique is commonly used in ligand-receptor recognition studies, such as small molecule screening, although large inter-biomolecular recognitions have also been reported, as DNA-protein interactions (Srinivasan et al., 1998). The binding free energy of the bound ligand-receptor complex in an aqueous solvent can be approximated as (Foloppe & Hubbard, 2006):

$$\Delta G_{bind,aq} = \Delta H - T\Delta S \approx \Delta E_{MM} + \Delta G_{bind,solv} - T\Delta S \qquad Equation\ 2.31$$

$$\Delta E_{MM} = \Delta E_{covalent} + \Delta E_{elec} + \Delta E_{vdW} \qquad Equation\ 2.32$$

$$\Delta E_{covalent} = \Delta E_{bond} + \Delta E_{angle} + \Delta E_{torsion} \qquad Equation\ 2.34$$

$$\Delta G_{bind,solv} = \Delta G_{polar} + \Delta G_{non-polar} \qquad Equation\ 2.35$$

where ΔE_{MM}, $T\Delta S$, $\Delta G_{bind,solv}$, represent the gas-phase molecular mechanical energy change, the conformational entropy change, and the solvation free energy change upon binding, respectively. All of these changes are computed via ensemble averaging over a large set of sampled conformations. ΔE_{MM} includes three terms calculated using molecular mechanics (MM): the covalent energy change ($\Delta E_{covalent}$), the electrostatic energy change (ΔE_{elec}), and the van der Waals energy change (ΔE_{vdW}). $\Delta E_{covalent}$ consists of changes in the bond terms (ΔE_{bond}), the angle terms (ΔE_{angle}), and the torsion terms ($\Delta E_{torsion}$). The solvation free energy change ($\Delta G_{bind,solv}$) is usually separated into polar and non-polar contributions (ΔG_{polar} and $\Delta G_{non-polar}$).

In binding affinity studies with the MMPBSA method, the MD simulations are almost always conducted in an explicit solvent model to obtain the most accurate snapshots and the multi-trajectory approach is preferred to the single-trajectory one (Wang et al., 2017). In case vacuum simulations are directly compared to simulations in solvent, the energy contributions coming from solvent-solvent interactions and the fluctuations in total energy would be an order of magnitude larger than binding energy, entailing an excessive amount of time to converge. According to the thermodynamic cycle in **Figure 2.5**, the energy estimates are usually obtained by (Hou et al., 2011):

$$\Delta G_{bind,solv} = \Delta G_{bind,vacuum} + \Delta G_{solv,complex} \\ - \left(\Delta G_{solv,ligand} + \Delta G_{solv,receptor}\right) \qquad Equation\ 2.36$$

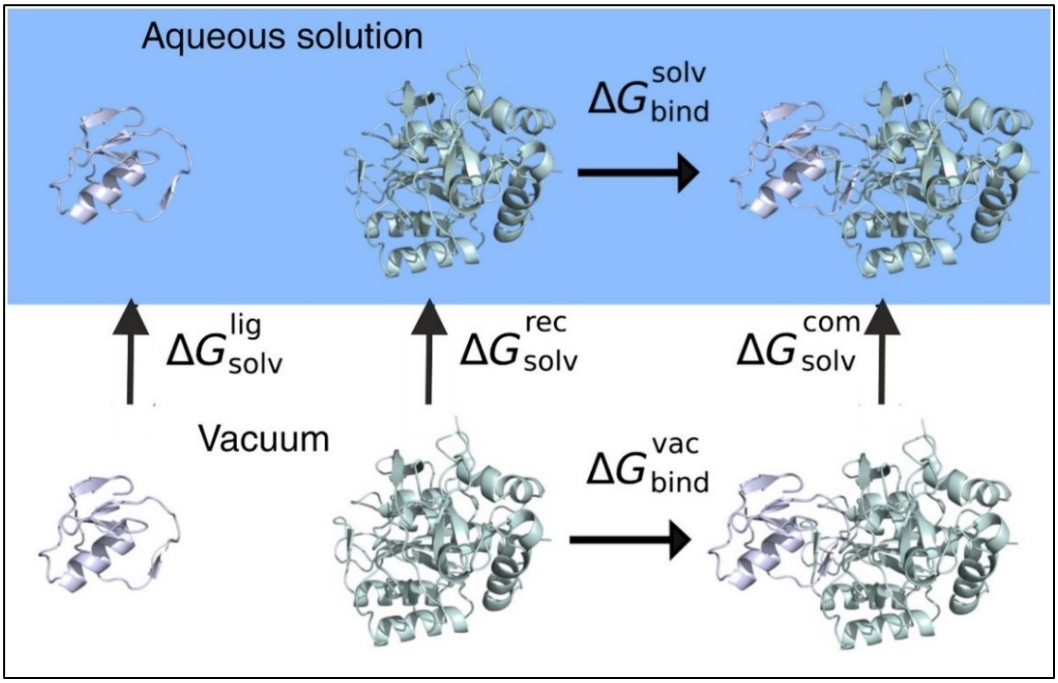

Figure 2.5 Evaluation of protein–protein complexes based on a continuum solvent model during MM-PBSA calculations. The binding process consists of an interaction contribution indicated in the lower panel (interaction energy is calculated as difference in the vacuum energies of the complex and the separate partners). The transfer of the partners and the complex into the aqueous environment (upper panel) adds a solvation contribution (also calculated as difference between complex and partner contributions). The solvation part consists typically of separate cavity terms and van der Waals interaction with the solvent plus an electrostatic reaction field (solvation) based on solving the finite difference PB equation numerically. *Reprinted from (Siebenmorgen & Zacharias, 2020)*.

The different contributions for the solvation free energies above, can be calculated by solving the linearised Poisson Boltzmann or Generalized Born equation for each of the three states and adding an empirical term for hydrophobic contributions:

$$\Delta G_{solv} = G_{elec,\varepsilon=80} - G_{elec,\varepsilon=1} + \Delta G_{hydrophobic} \qquad Equation\ 2.37$$

where $\Delta G_{bind,vacuum}$ is obtained by calculating the average interaction energy between receptor and ligand and taking the entropy change upon binding into account if necessary:

$$\Delta G_{vacuum} = \Delta E_{MM} - T\Delta S \qquad Equation\ 2.38$$

The entropy term can be estimated by performing normal mode analysis on the three species, but in practice entropy contribution can be neglected if only a comparison of states of similar entropy is desired. This is because computations involving normal mode analysis are computationally costly and frequently have a huge margin of error that creates a considerable amount of uncertainty in the outcome. In fact, all these terms are computed on each single frame of the MD simulations and those single values are averaged and an estimate of the binding free energy of the complex can be obtained. Both of the multi-trajectory and single-trajectory strategies, which can be employed in the MM-PBSA as well as the complementary MM-GBSA technique, include approximations, and their performance varies strongly with the tested systems (Tuccinardi, 2021; Wang et al., 2017).

2.4.2 Free energy estimates based on collective variables

The standard atomistic MD methods do not adequately sample the ensemble of possible configuration, leading to inaccurate estimates of probabilities, and free energies, along collective variables. Consequently, a diverse range of methods have been developed to enhance sampling of biomolecular simulations. The bottleneck of MD simulations is correlated with the presence of high-energy barriers separating distinct structures/conformations, resulting in transitions between them appearing as rare events in the simulation. The addition of bias potentials to the Hamiltonian of the systems, i.e., lowering the energy barrier to increase the sample transition zones, is a straightforward and effective method of speeding up the exploration of the configuration space. Umbrella Sampling (US), adaptive biassing force method, and MetaDynamics (MetaD), are examples of such methods. These approaches make use of specified collective variables to efficiently accelerate sampling during the simulations (Y. I. Yang et al., 2019).

CV, $s(r)$, are defined as low-dimensional functions of the atomistic coordinate r of the system, which are designed to describe the slower motions in the process of interest. The CV $s(r)$ has an equilibrium distribution $p_0(s)$ and a free energy $A(s)$.

$$p_0(s) = \int dr\delta[s - s(r)]p_0(r) = \langle \delta[s - s(r)] \rangle \qquad Equation\ 2.39$$

where $p_0(r) = e^{\frac{-\beta U(r)}{\int dre^{-\beta U(r)}}}$ is the Boltzmann distribution of the system with potential energy *U(r).*

$$A(s) = -\frac{1}{\beta} log[p_0(s)] \qquad Equation\ 2.40$$

And $\beta = \frac{1}{k_B T}$ is the inverse temperature associated with the Boltzmann constant k_B.

In CV-based sampling methods, to overcome the energy barriers that divide two or more configurations in a defined region, a bias potential $V(r)$ along the CV $s(r)$ is introduced into the system. The free energy $A(s)$ may be determined in this case by:

$$A(s) = -\frac{1}{\beta} log[p(s)] - V(s) \qquad Equation\ 2.41$$

where $p(s)$ is the sampled distribution of collective variables $s(r)$ from simulation.

In this thesis, US was used to estimate the Potential of Mean Force (PMF) (Torrie & Valleau, 1977). The PMF is the driving force in stochastic dynamics models. It provides a free energy profile along a preferred coordinate, such as the distance between two atoms or the torsional angle of a bond in a molecule. The PMF describes the average force of all possible configurations of a given system (the ensemble average of the force) on particles of interest. When the system is in a solvent, the PMF incorporates solvent effects as well as the intrinsic interaction between the two particles. US allows the calculation of PMF along an arbitrary CV, $s(r)$, that describes the extent of the transformation under analysis. The technique consists in splitting the reaction coordinate domain in intervals (or windows) and separately sample each window with the addition of a biasing potential that allows to overcome possible energy barriers. In practice, the biasing potential restrains the reaction coordinate to the window centre, forcing the system to explore the centre neighbourhood. The name "Umbrella" derives from the fact that the restraining potential has a parabolic functional form, like a spring potential with the window centre as equilibrium distance. Thus, applying an adequate restraining potential, every point of the reaction coordinate domain can be explored, also states not spontaneously explored due to the presence of energy barriers. The higher the energy needed to restrain the system to a specific region, the higher the free energy corresponding to those configurations. The PMF profile is reconstructed from the probability distribution of the reaction coordinate, that is the practical result of an US simulation (Roux, 1995) (**Figure 2.6**).

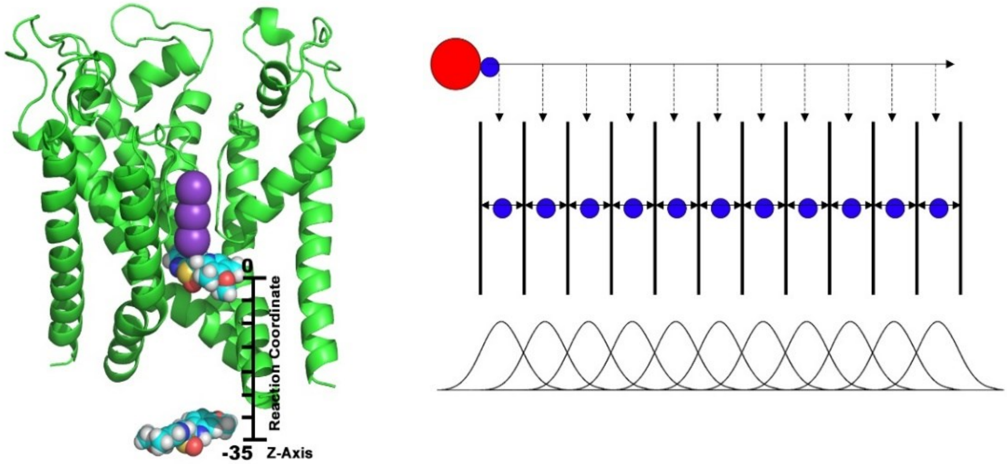

Figure 2.6 On the left, an example of US simulation with harmonic potential to restrain the ligand along Z-axis; on the right, a scheme of the approach used to recreate the PMF profile.

The calculation of the PMF is not straightforward as in plain MD, because the resulting distribution is not canonical due to the bias presence. Various methods have been developed to unbias and recombine the results of US simulations; the Weighted Histogram Analysis Method (WHAM) of Kumar et al is the most adopted for Free Energy calculations (Kumar et al., 1992). Combining US simulations and WHAM, PMF can be computed (Souaille & Roux, 2001). It is worthy to notice that in principle a N-dimensional PMF can be obtained, but no more than two degrees of freedom are feasible because the number of sampling windows increases exponentially with the number of reaction coordinates.

PART II – *Research Projects*

Chapter 3 - MD simulations of the hERG potassium channel

3.1 Ion channels: the hERG K^+ channel

Ion channels are pore-forming transmembrane proteins that allow ions to flow across the plasma membrane according to electro-chemical gradients. The rate of transport can be as high as 10^6 ions per second, or greater. When a channel is in the closed state, its permeability for all kinds of ions and water is roughly zero, whereas in the open state they conduct electrical current by allowing specific types of ions to pass through them, and thus, across the plasma membrane of the cell (selective permeability). Transition between the open (conductive) and the closed (non-conducting) conformation is called gating. A variety of cellular changes can lead to gating, depending on the ion channel, including voltage changes across the cell membrane (voltage-gated ion channels), chemicals interacting with the ion channel (ligand-gated ion channels), changes of temperature, elongation or deformation of the cell membrane (mechanosensitive ion channels), addition of a phosphate group to the ion channel (phosphorylation) and interaction with other molecules in the cell (i.e., G-proteins). The rate at which one of these gating processes occurs in response to these triggers is known as gating kinetics. The intensity and direction of the ionic movement across the pore are governed by electrochemical gradients. In normal physiological conditions, sodium, calcium, and chloride ions tend to enter the cell, while potassium ions tend to exit. Selective permeability depends on the structural and electrostatic characteristics of the pore. The capacity to adjust ion flow because of ion channel development may have offered an evolutionary advantage by allowing single-celled organisms to control their volume in response to environmental changes. Ion channels have evolved throughout time to perform many processes in excitable and non-excitable cells. The first quantitative description of the role of membrane currents in signalling dates to the Hodgkin and Huxley investigations on nerve transmission (HODGKIN & HUXLEY, 1952). Ion channels are associated with several physiological functions in all types of cells, such as cardiac, skeletal, and smooth muscle contraction, epithelial transport of nutrients and ions, T-cell activation, and pancreatic beta-cell insulin release. Furthermore, ion channels can be employed in a variety of technical applications (sensing of organic molecules, DNA sequencing) leading great interest to elucidate the molecular determinants of channel functions.

The focus of this thesis is the hERG channel. In 1994, Warmke and Ganetzky firstly identified it as the 11th member of the voltage-gated potassium channels family (Warmke & Ganetzky, 1994). The functional channel is a homo-tetramer with each subunit containing six α-helices transmembrane domains, named as S1 to S6. S1 to S4 contribute to the voltage sensor domain, while the S5 and S6 segments, along with the intervening pore-loop (P-loop), contribute to the pore domain. The central cavity presents an atypically small central volume, surrounded by four deep hydrophobic pockets for potassium ions. The region responsible for the selective conduction of potassium ions, Selectivity Filter (SF), is located at the extracellular side of the central cavity. In addition to the membrane-spanning region, the hERG protein contains large intracellular amino and carboxy terminal domains. The N-termini contains a Per-Arnt-Sim (PAS) domain that defines the ether-a-go-go subfamily of Kv channels. The PAS domain modulates the deactivation of the channel following membrane depolarization. The C-termini of the channel contains a cyclic nucleotide binding domain (CNBD), which has been linked to mutations affecting trafficking (Wang & MacKinnon, 2017) (**Figure 3.1**).

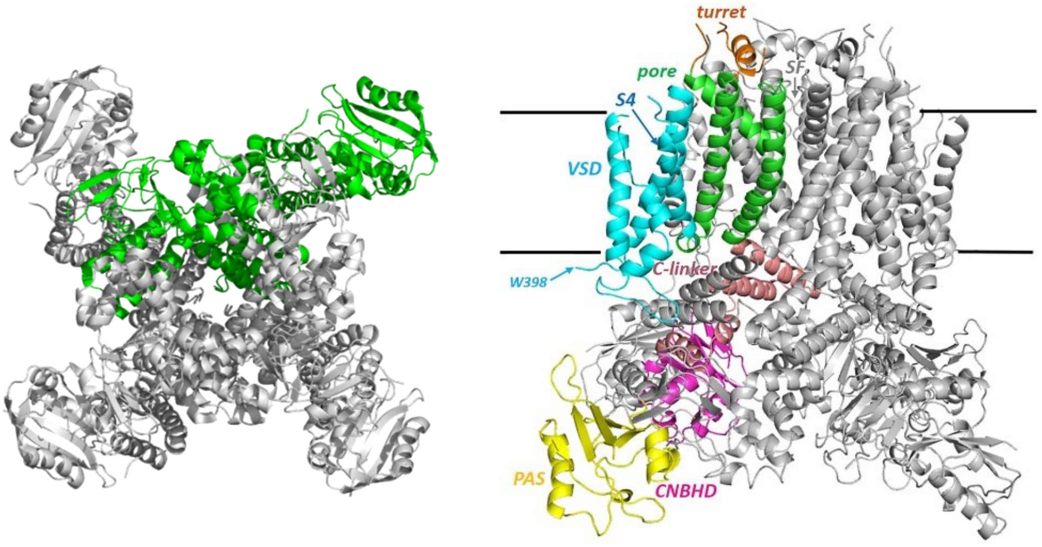

Figure 3.1. Structural representation of the hERG potassium channel. *Adapted from (Butler et al., 2019).*

hERG has peculiar molecular features that distinguish it from other members of the Kv family. Firstly, the amino acid sequence TVGYG of the SF is a conserved pattern among the vast majority of potassium channels, and it is responsible for the selectivity function. In hERG, the threonine and the tyrosine of this motif are respectively replaced by serine and phenylalanine (SVGFG) (Perrin et al., 2008) (**Figure 3.2**).

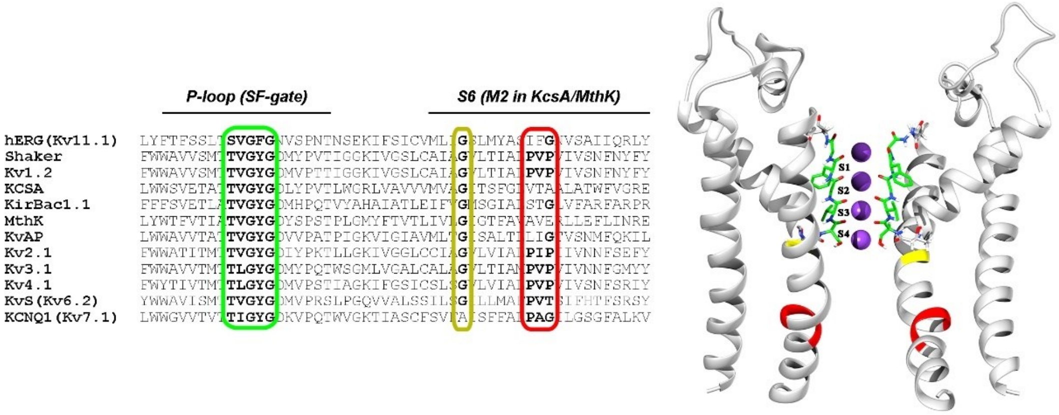
P-loop (SF-gate)
S6 (M2 in KcsA/MthK)
hERG(Kv11.1) LYFTFSSLTSVGFGNVSPNTNSEKIFSICVMLIGSLMYASIFGNVSAIIQRLY
Shaker FWWAVVSMTTVGYGDMYPVTIGGKIVGSLCAIAGVLTIALPVPVIVSNFNYFY
Kv1.2 FWWAVVSMTTVGYGDMVPTTIGGKIVGSLCAIAGVLTIALPVPVIVSNFNYFY
KCSA LWWSVETATTVGYGDLYPVTLWGRLVAVVVMVAGITSFGLVTAALATWFVGRE
KirBac1.1 FFFSVETLATVGYGDMHPQTVYAHAIATLEIFVGMSGIALSTGLVFARFARPR
MthK LYWTFVTIATVGYGDYSPSTPLGMYFTVTLIVLGIGTFAVAVERLLEFLINRE
KvAP LWWAVVTATTVGYGDVVPATPIGKVIGIAVMLTGISALTLLIGTVSNMFQKIL
Kv2.1 FWWATITMTTVGYGDIYPKTLLGKIVGGLCCIAGVLVIALPIPIIVNNFSEFY
Kv3.1 FWWAVVTMTTLGYGDMYPQTWSGMLVGALCALAGVLTIAMPVPVIVNNFGMYY
Kv4.1 FWYTIVTMTTLGYGDMVPSTIAGKIFGSICSLSGVLVIALPVPVIVSNFSRIY
KvS(Kv6.2) YWWAVISMTTVGYGDMVPRSLPGQVVALSSILSGILLMAFPVTSIFHTFSRSY
KCNQ1(Kv7.1) LWWGVVTVTTIGYGDKVPQTWVGKTIASCFSVFAISFFALPAGILGSGFALKV
S1
S2
S3
S4

Figure 3.2. Topology of potassium channels. Sequence alignment of the pore domain (PD) (S5–S6) which contains the TVGYG signature sequence (green) among the member of Kv channel. Highlighted in yellow is the highly conserved glycine residue in the middle of the inner pore helix. The PXP motif, absent in hERG is highlighted in red.

The aromatic residue, F, might contribute to the channel unusual susceptibility to block by a diverse range of drugs. Secondly, these promiscuous drug interactions might also be related to the lack of the S6 proline-X-proline (PXP) motif necessary to restrict the inner cavity size in other Kv channels and to accept limited drug molecule sizes. Finally, the S5-P linker is a large segment that assumes an amphipathic helical arrangement in membrane mimetic sodium dodecyl sulphate (SDS) micelles, that is believed to affect channel inactivation (Perrin et al., 2008; Torres et al., 2003; Wang & MacKinnon, 2017; Wulff et al., 2009; Zhou et al., 2011).

3.2 Gating and inactivation of hERG

Kv are characterised by three main conformational states: closed, open, and inactivated (**Figure 3.3 -A**). Variations in the kinetics and voltage dependence of the gating and inactivating processes give rise to phenotypic diversity among the different Kv channel subtypes. The Kv11.1 subfamily is characterized by its rapid onset of C-type inactivation at depolarized potentials (on the order of ms to tens of ms), followed by recovery from inactivation during the repolarization phase (**Figure 3.3-B**). This property, combined with a slow channel gating (on the order of hundreds of ms to s), results in an inward rectified current, that is crucial for maintaining a prolonged plateau phase of the cardiac action potential (Li et al., 1996; Perry et al., 2015; Sanguinetti et al., 1995) **(Figure 3.3-C)**.

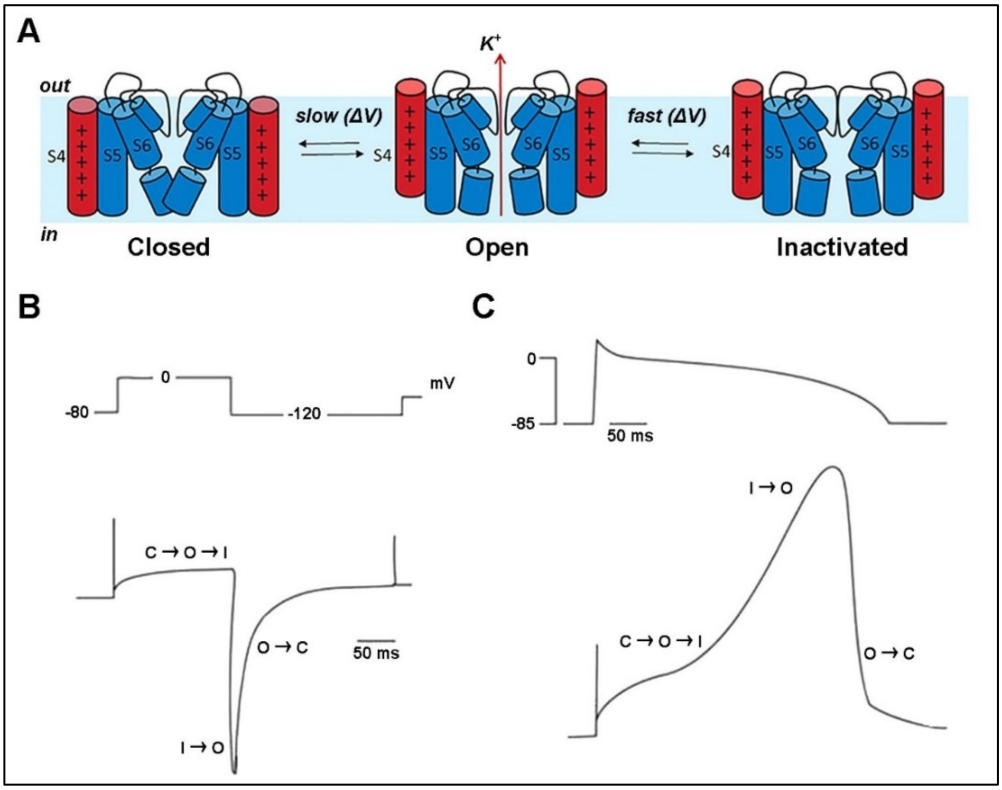

Figure 3.3. Gating of hERG. [**A]** Three main conformational states: Closed (C), Open (O), and Inactive (I). [**B**] Typical traces showing kinetics: slow activation and deactivation, coupled with rapid inactivation. [**C**] Currents recorded during cardiac ventricular AP.

Just as importantly, the rapid and voltage-dependent recovery from inactivation during the terminal phase of cardiac repolarization coupled to slow deactivation during the early diastolic period confers upon these channels an important role in the suppression of ectopic beats during the late repolarization phase. Indeed, in patients with reduced hERG K^+ channel activity, e.g., due to drug-block, the reduced hERG K^+ current results in longer action potentials (AP) as well as lower current response to premature beats. The surface electrocardiogram (ECG) represents the summed activity of all the cells in the heart with the major deflections being the P-wave (representing atrial depolarization), the QRS complex (representing ventricular depolarization) and the T-wave (representing ventricular repolarization). The duration of the interval from the start of the QRS complex to the end of the T-wave (QT interval) is usually ~400ms (at a heart rate of 60 beats per minute). Patients with reduced hERG K^+ channel activity have prolonged QT intervals on their surface electrocardiogram and an increased risk of developing ventricular arrhythmias initiated by ectopic beats. They are prone to develop a particular arrhythmia called "Torsades-de-Pointes" (TdP) (Raschi et al., 2008; Vandenberg et al., 2012) (**Figure 3.4**).

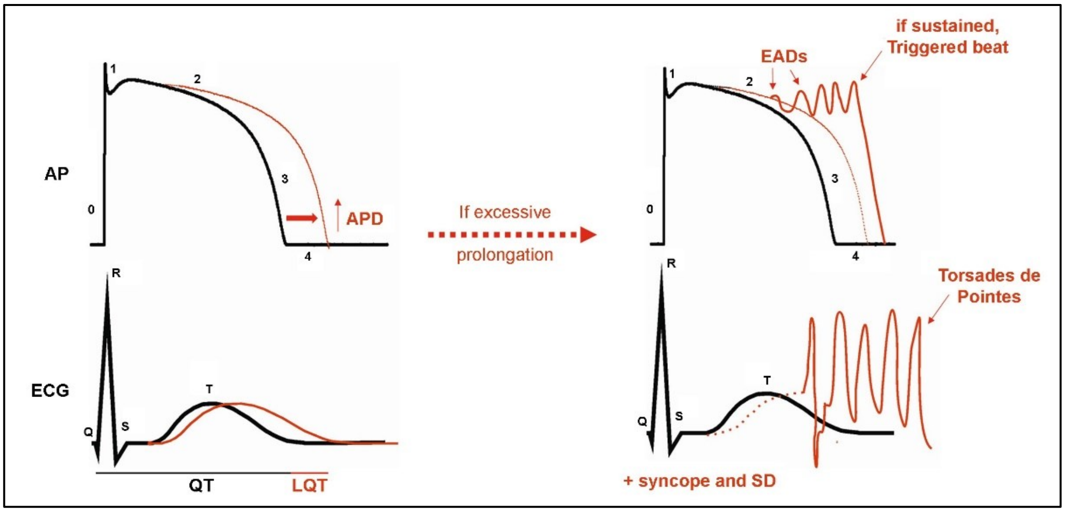

Figure 3.4. Diagram linking main causes of LQTS and associated arrhythmia with their effects on biological markers of the human cardiac ventricular repolarization. *Reprinted from (Vandenberg et al., 2012).*

In silico kinetic modelling of ion channel activity provides a formal quantitative mechanism for testing hypotheses about how channels work. The biophysical accuracy of Kv11.1 models in reproducing gating kinetics is particularly important for several reasons. Simpler models may be able to approximate the overall properties of the current but are less effective at reproducing more complex time- and voltage-dependent effects. Over the past 25 years, models of hERG behaviour have been developed from Hodgkin-Huxley type descriptions (Rockman et al., 2002), through quite simple linear Markov schemes (Wang et al., 1997), to more complex Markov descriptions containing multiple closed, open, and inactive states (Kiehn et al., 1999; Lu et al., 2001) as well as subunit cooperativity (Piper et al., 2003). The simplest, and perhaps most used, model of voltage-gated ion channel behaviour is based on the formalism introduced by Hodgkin and Huxley 60 years ago (HODGKIN & HUXLEY, 1952). In these formulations, the current passed through the channel is simply a function of two gating variables, that describes activation and inactivation. In 1997, Wang et al. developed a simple linear Markov scheme based on experimental data obtained from hERG currents expressed in oocytes (Wang et al., 1997). In contrast to the Hodgkin-Huxley type models, Markov schemes allow incorporation of multiple open, closed, and inactive states, with transitions between each individual state determined by voltage-dependent rate constants. Since the authors were able to fit their measured forward and reverse rate constants for inactivation with first-order voltage-dependent models, it was possible to include a single pre-open closed state in the linear gating scheme. Two years later, Kiehn et al. published an upgrade of this model that included inactivation from the final pre-open closed state (Kiehn et al., 1999). The authors proposed that this transition was necessary to explain the presence of channel openings upon repolarization when no channel openings had occurred during the preceding depolarization step and the voltage-dependence of the magnitude of the transient peak upon depolarization. The authors evaluated several Markov model structures by fitting to macroscopic Kv11.1 currents and showed that a closed to inactive transition was necessary to accurately reproduce the transient peak observed experimentally upon depolarization. The model, however, was only used to simulate a few simple voltage protocols. The "closed-state inactivation model" was not parameterized fully or used to simulate more physiologically relevant action potential waveforms until two years later when an adaptation of the model proposed by Kiehn et al. was used in an *in silico* evaluation of long QT syndrome and in evaluating the effects of premature stimulation on

Kv11.1 gating. Their data provided evidence that hERG encodes repolarizing K^+ current (I_{Kr}) in cardiomyocytes. Recently, a similar model of gating current kinetics was used to describe hERG channel voltage sensor relaxation. In this model, two independent transitions per subunit are followed by a voltage-dependent concerted transition to the activated state, and a subsequent voltage-independent transition into the relaxed state. The model recapitulates the main features of hERG gating currents including voltage sensor mode-shift behaviour. Moreover, acceleration of the rate out of the relaxed state to mimic the destabilisation of relaxation observed at low pH, selectively abolished mode-shift behaviour without other gating consequences, recapitulating the experimentally observed voltage sensor behaviour. From this model, acceleration of de-relaxation, or exit from relaxed state, was sufficient to reduce voltage sensor mode-shift and supported the hypothesis that destabilisation of the relaxed state of the voltage sensor may drive voltage sensor return leading to accelerated deactivation. One limitation of this model is the absence of description of ionic activation and inactivation gating of the channel, which are needed to develop a more complete model of gating transitions of the hERG channel voltage sensor in conjunction with pore gating during voltage sensor stabilisation and relaxation. This might involve an approach used previously in Shaker and Kv1.2 channels to construct models that describe transitions of voltage sensor domains corresponding to those of the pore, which may be applicable for adaptation into a hERG scheme (Shi et al., 2020).

hERG unique kinetic features make it an important channel in the repolarization phase of cardiac action potentials. Consequently, the pharmacology of Kv11.1 has become a topic of interest following the discovery that this channel is the molecular target of numerous compounds related to drug-related arrhythmias. Understanding the molecular basis for promiscuous drug block of hERG would be enormously beneficial in efforts to pre-screen drugs for hERG liability in drug development programs, and to reduce adverse effects in otherwise-useful drugs through targeted chemical modification. Likewise, insight into the molecular basis for hERG's anomalous gating properties, particularly the mechanisms of rapid onset and recovery from inactivation, should greatly facilitate development of therapeutic interventions for LQTS.

3.3 Atomistic models of hERG

Cryo-Electron Microscopy (Cryo-EM) was firstly utilized to analyse small and periodic collections of proteins in the mid-1970s. Owing to advances in software development, cryo-EM structure resolution improved steadily from sub-nanometre to near atomic resolution by the late 2000s. Cryo-EM not only establish a revolution in resolution quality but can also capture protein's conformational states. In 2017, Wang and MacKinnon released the first cryo-EM structure of the hERG channel (Wang & MacKinnon, 2017). Although some missing residues remains in the extracellular loop regions (from His578 to Arg582, and from Asn598 to Leu602), this model (PDB entry 5VA2) offered the opportunity to study the dynamics of the hERG channel at the atomic scale, and it was used for the analyses presented in this thesis. Since the interest was in studying the binding of drugs to the central cavity, and the dynamics of the SF, only the pore region of the channel, from residue Tyr545 to residue Tyr667, was included in the atomic models (Wang & MacKinnon, 2017). MODELLER tool was used to modelling of the chirality and initial positions of missing residues. Ions were manually added at site S0, S2 and S4 (see **Figure 3.2** for the definition of the binding sites). Restrain on the SF derive from a previous study by Furini et al. in the homologue KcsA channel in the open state (PDB entry 5VK6) (Furini & Domene, 2020; Li et al., 2018). The lipid membrane was a mixture of 1-palmitoyl-2-oleoyl-glycero-3-phosphocholine (POPC) and 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphate (POPA), with a ratio of 3-POPC:1-POPA. The axis of the channel was aligned with the z-axis of the simulation box. The system was solvated using TIP3P water molecules (~17.000 molecules) and 200 mM of KCl were added. The ff14sb version of the AMBER force field was used, in combination with ion parameters by Joung and Cheatham for the TIP3P water model (Joung & Cheatham, 2009). Van der Waals interactions were truncated at 9 Å. Standard AMBER scaling of 1-4 interactions was applied. Long-range electrostatic interactions were calculated with the Particle Mesh Ewald method using a grid spacing of 1.0 Å. The SETTLE algorithm was used to restrain bonds with the hydrogen atom (Essmann et al., 1995; Tuckerman et al., 1992). The temperature was controlled at 310 K by coupling to a Langevin thermostat with a damping coefficient of 1 ps^{-1}. A pressure of 1 atm was maintained by coupling the system to a Nose–Hoover Langevin piston, with a damping constant of 25 ps and a period of 50 ps (Feller et al., 1995). NAMD2.12 was used for all the simulations (Phillips et al., 2005). The equilibration protocol consisted of 10.000 steps of energy minimization, followed by 15 ns in the NPT ensemble with timestep equal to 1 fs and 70 ns in the NPT ensemble with timestep equal to 2 fs. During the equilibration protocol, restraints on protein and lipid atoms were gradually reduced to zero. The atomic coordinates at the end of this equilibration protocol were used to define the starting configurations for further studies of the thesis (**Figure 3.5**).

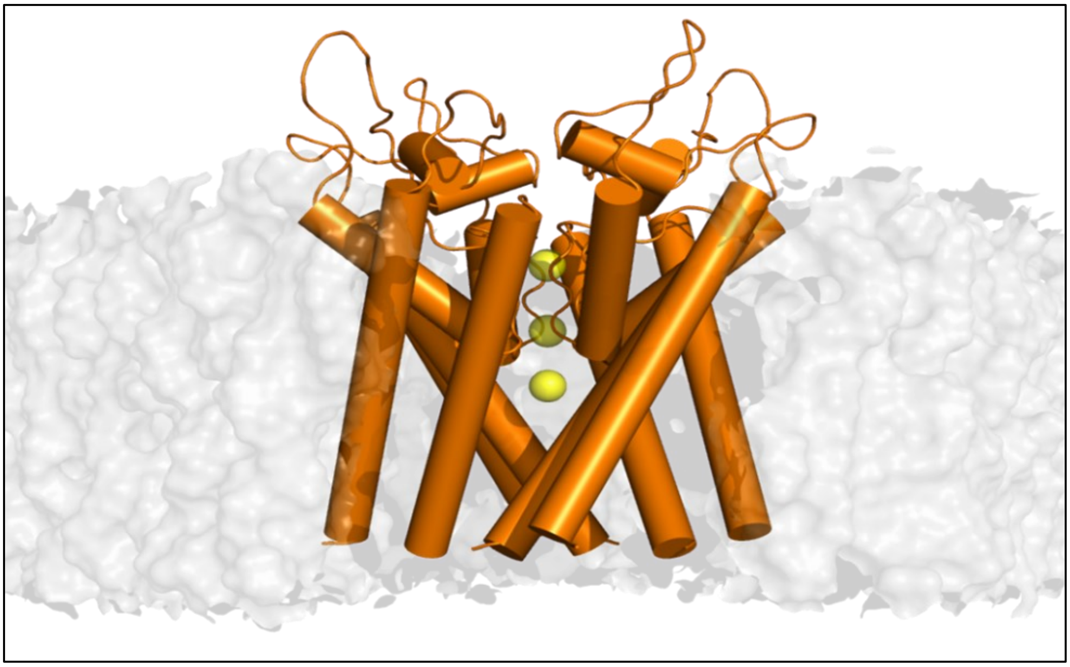

Figure 3.5 The atomic model of hERG used for the atomistic MD simulations.

3.4 *In silico* modelling of blockade by drug binding to the internal cavity

The emerging picture is that drug binding to Kv11.1 may be a highly dynamic process with multiple drugs and channel conformations involved in the binding of any given drug (Vandenberg et al., 2012). The development of accurate computer models is critical to deciphering how the interaction or kinetics of a drug molecule are affected by the gating state of the channel. As a result, a better understanding of the molecular basis of hERG channel gating will facilitate computational drug design and discovery. In addition, the description of the physicochemical properties of the drug binding site will complement the pharmacophore models of the drugs (Asai et al., 2021). Toward this goal, several studies with site-directed mutagenesis and voltage clamp analysis of mutant channels try to elucidate which is the key role of specific residues at the polar surface area. In 2004, Fernandez et al. summarized evidence that emphasize the role of two aromatic residues, Tyr652 and Phe656, in determine the sensitivity to hERG blockers (Fernandez et al., 2004). These eight aromatic residues (2 for each subunit), located in the S6 domain facing the central cavity of the channel, combined to the lack of the PXP motif of hERG prevent bending of α-helix and the consequent alteration of the shape of the central cavity with respect to other potassium channels. Residues Thr623, Ser624, and Val625 close to the intracellular entrance to the SF domain results involved in drug binding as well as the two aromatic residues Tyr625 and Phe656. Whereas polar residues at the base of the pore helix are highly conserved among members of the Kv channel family, the aromatic residues on the S6 helices are not (Shealy et al., 2003). Early structural modelling of the hERG pore region based on the closed-state of the KcsA channel highlighted that the gating-dependent positioning of these residues relative to the inner cavity (particularly with inactivation) may be critical for high-affinity binding and could explain the higher potency of compounds for hERG compared with EAG block (Mitcheson et al., 2000). In addition, the pharmacophore models extended these findings to show that hydrophobic volume of Phe-656 and aromaticity of Tyr-652 determine the sensitivity of hERG to block by structurally diverse drugs known to cause acquired LQTS (Chen et al., 2002; Ficker et al., 2001; Herzberg et al., 1998). Finally, computational approaches proposed in this dissertation confirmed unusual features of the inner cavity associated to the different transition state of the gating, providing also important insight in the molecular recognition of drugs. These studies are described in the next sections.

3.4.1 Proton Pump Inhibitors Directly Block hERG

The pharmacology of Kv11.1 has become a subject of intense interest following the discovery that this channel is the molecular target for the majority of compounds associated with drug-induced arrhythmias. The relationship between Kv11.1 channel block, QTc prolongation, and TdP is crucial in order to identify the proarrhythmic risk of new and existing drugs. Proton Pump Inhibitors (PPIs) represent one of the best-selling class of drugs in the market, with millions of chronic users worldwide (Strand et al., 2017). Omeprazole, pantoprazole, lansoprazole, and esomeprazole, representing the commonly prescribed molecules, are responsible for adverse events due to long-term overutilization (Patterson Burdsall et al., 2013). The potential cardiovascular harmfulness of PPIs is increasingly recognized, including a higher risk of malignant arrhythmias (Manolis et al., 2020). In particular, the 4 PPIs mentioned above are presently listed by the AriZona Center for Education and Research on Therapeutics (Woosley et al., 2020) as drugs with conditional risk of TdP. It is currently believed that PPI treatment can increase the LQTS/TdP risk only indirectly, by lowering magnesium levels (Chrysant, 2019). Although the exact mechanisms underlying PPI-induced hypomagnesemia are not completely elucidated, evidence points to gastrointestinal and renal magnesium losses (Famularo et al., 2013). Recent data suggest that PPIs may increase the TdP risk also beyond hypomagnesemia induction, due to a direct interference on the electrophysiological properties of cardiac myocyte. Lorberbaum et al. demonstrated that the PPI lansoprazole inhibited the hERG function and increased the risk of QTc prolongation when used in combination with the antibiotic ceftriaxone (Lorberbaum et al., 2016). Accordingly, cases of LQTS/TdP during PPI treatment but in the absence of low magnesium levels have been published (Lazzerini et al., 2017). Moreover, although PPI-associated hypomagnesemia was found to be a common finding in a cohort of 48 TdP patients, in most of the PPI-treated subjects (≈60%), TdP developed in the presence of normal magnesium levels (Lazzerini et al., 2018). Defining whether PPIs can directly promote QTc prolongation, regardless of hypomagnesemia, has important clinical implications. In fact, it is currently recommended that in LQTS/TdP patients, PPI treatment is discontinued only if hypomagnesemia, resistant to replacement therapy, occurs. Thus, a significant number of subjects may continue to be unnecessarily exposed to a risk for TdP occurrence/recurrence. Accordingly, a Sweden register-based cohort study found that gastric acid secretion inhibitors were used in 32% of 410 TdP cases (Danielsson et al., 2020). By combining electrophysiology, MD simulations, and population data, in this project it was evaluated whether PPIs can:

1. inhibit the hERG current in an in vitro cellular model,
2. directly bind to hERG and enter the intracellular cavity of the channel by using MDs,
3. independently increase the risk of QTc prolongation in a large cohort of US veterans.

I was responsible for the second step of this project: the analysis of the possible hERG blockage by PPIs using MD simulations.

As discussed in **Chapter 2**, drug-binding processes usually exceed the time scale currently accessible by all-atom MD simulations. Consequently, enhanced techniques are more suitable to accelerate sampling and to optimise the usage of computational resources. In this study US was employed for estimating free-energies of hERG in complex with PPIs, because they accelerate the sampling by "flattening" the energy barriers along the pre-defined collective variables. The distributions of the collective variables collected from the set of independent simulations were combined by the WHAM, to estimate PMF profiles. Here, US simulation was performed using the atom-based model of hERG described in **Section 3.3**. Compounds are available in the DrugBank database with accession numbers DB00338 (Omeprazole), DB00213 (Pantoprazole), DB00448 (Lansoprazole) in the form of 2D structures (Wishart et al., 2018) (**Figure 3.6-A**). The three-dimensional structure used for the simulations does not explicitly describe electrons, and consequently it cannot distinguish omeprazole from its S-isomer esomeprazole. The compounds were parameterized using the Antechamber software, adding atomic charges, atom types and bond types according to the General AMBER Force Field (GAFF) (Case et al., 2005). The drug-specific parameters not included in the GAFF force field were estimated using ParmCheck and stored in a specific library. Next, ligands were docked into the extracellular space of the solvated and neutralised hERG channel system, removing all water molecules within a cut-off of 2 Å. We repeated the procedure of parameterization with ParmCheck to the ligand-protein complexes, adding the library of the drug-specific parameters. Then, an equilibration protocol consisting of 10.000 steps of energy minimization and 2 ns in the NPT ensemble was performed. The atomic coordinates at the end of this equilibration protocol were used to define the starting configurations for the three atomic systems used to analyse the binding of the three PPIs to the internal cavity of the hERG channel. The same equilibration protocol and simulation parameters described in **Section 3.3** was used. The entrance of the drug into the channel cavity was described using as reaction coordinate (ξ) the distance along the z-axis between the centre of mass of the drug and the centre of mass of the ring of carbonyl oxygens at the intracellular side of the SF (residues Ser624) (**Figure 3.6-B**).

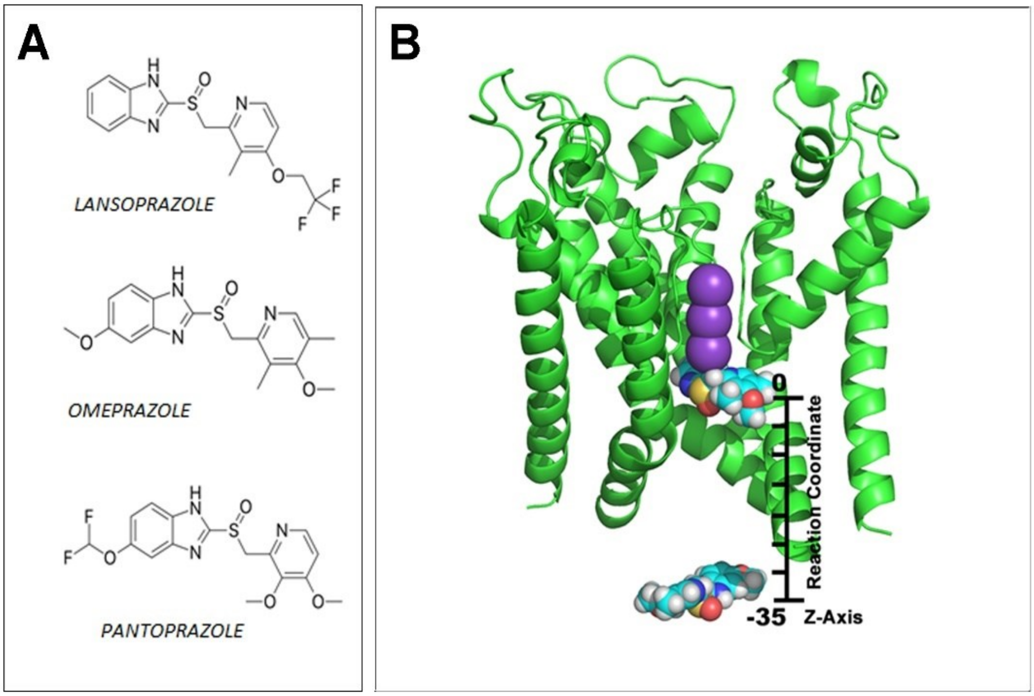

Figure 3.6. **[A]** 2D structure of the three PPI used in the USs; **[B]** schematic representation of US simulation using z-axis as RC.

Preliminary set of simulations, each 1 ns long, was collected moving the centre of the harmonic potential that act on the reaction coordinate ξ from the intracellular compartment (ξ = -35 Å) to the intracellular side of the SF (ξ = -6 Å) in steps of 1 Å. The simulation with the harmonic potential centred at -35 Å was initialised from the previously equilibrated atomic system with the drug in bulk solution. Then, the final snapshot of this simulation acted as starting configuration for the simulation with harmonic potential centred at -34 Å, and the same strategy was applied for the successive simulations. The aim of this initial set of simulations was only to create starting configurations of the drug at different positions along the axis of the channel. Afterward, these configurations were used to initialise a second set of simulations, with the aim to obtain a more exhaustive sampling of the drug movements inside the channel cavity. Each umbrella sampling simulation was extended until the reaction coordinate was at equilibrium (minimum length of each simulation was 20 ns). To check if the reaction coordinate was at equilibrium, the trajectory was divided into three segments, and the average of the reaction coordinate was compared between the last two segments. The simulations were extended until the average of the reaction coordinate in the last two segments of the trajectory was closer than 10^{-3} Å. The PMF was calculated by the WHAM algorithm using the last 2/3 of the simulated trajectories.

Results indicate the binding energies were in the 8-10 kcal/mol range for all the compounds, despite their specific pose in the pore, confirming the hypothesis that the inhibitory effects on hERG-current might be due to a direct, steric interference with the channel function. From PMFs profiles it emerges that Omeprazole preferentially bound at the intracellular side of the cavity, where a swallow free-energy minimum exists, while Pantoprazole and Lansoprazole were characterised by well-defined energy minima profile in proximity of the SF. In support of the electrophysiological findings, simulation studies indicated that all the three PPIs evaluated can bind hERG by entering the channel cavity, thereby strongly supporting the hypothesis that the inhibitory effects on hERG-current might be due to a direct, steric interference with the channel function. Moreover, the evidence that each compound shows a specific binding pose, provides a molecular basis possibly accounting for the different potency of channel inhibition observed in the electrophysiological study. Notably, as this region is structurally stable in the different functional states of the channel, the binding pose of pantoprazole and lansoprazole might favour the trapping of the drugs at their binding sites when hERG is in the closed state, giving rise a more robust and durable blockade (**Figure 3.7**).

The MD results qualitatively agree with experimental data. The whole-cell patch-clamp recordings performed in hERG human embryonic kidney 293 cells demonstrated that all PPIs tested were able to significantly inhibit the hERG current in a concentration-dependent manner, although with different potency. Notably, the concentrations effectively reducing the current were in a clinically relevant range (10–100 µmol/L), reached during routine therapy with PPIs (alone or, as it frequently occurs, in combination with other drugs or diseases slowing PPI metabolism). At these concentrations, pantoprazole was the most potent inhibitor (≈35%–85% current decrease), followed by lansoprazole (≈20%–50%) and then omeprazole/esomeprazole (≈10%–30%). Based on these mechanistic data, it was evaluated whether such electrophysiological effects translate in the clinical setting, by assessing the impact of PPI treatment on the QTc in a sample size of almost 4000 US veterans, including 1289 PPI users. In this cohort, PPI-treated subjects exhibited a significantly longer mean QTc, as well as a higher QTc prolongation prevalence when compared with PPI-untreated subjects. Consequently, stepwise regression analysis of the cohort indicate that the PPI therapy can per se promote QTc lengthening, regardless of other concomitant QT-prolonging risk factors, including hypomagnesemia, thereby, confirming that the direct electrophysiological effects of these drugs observed in vitro and in silico have a clinically relevant impact in a large population of individuals.

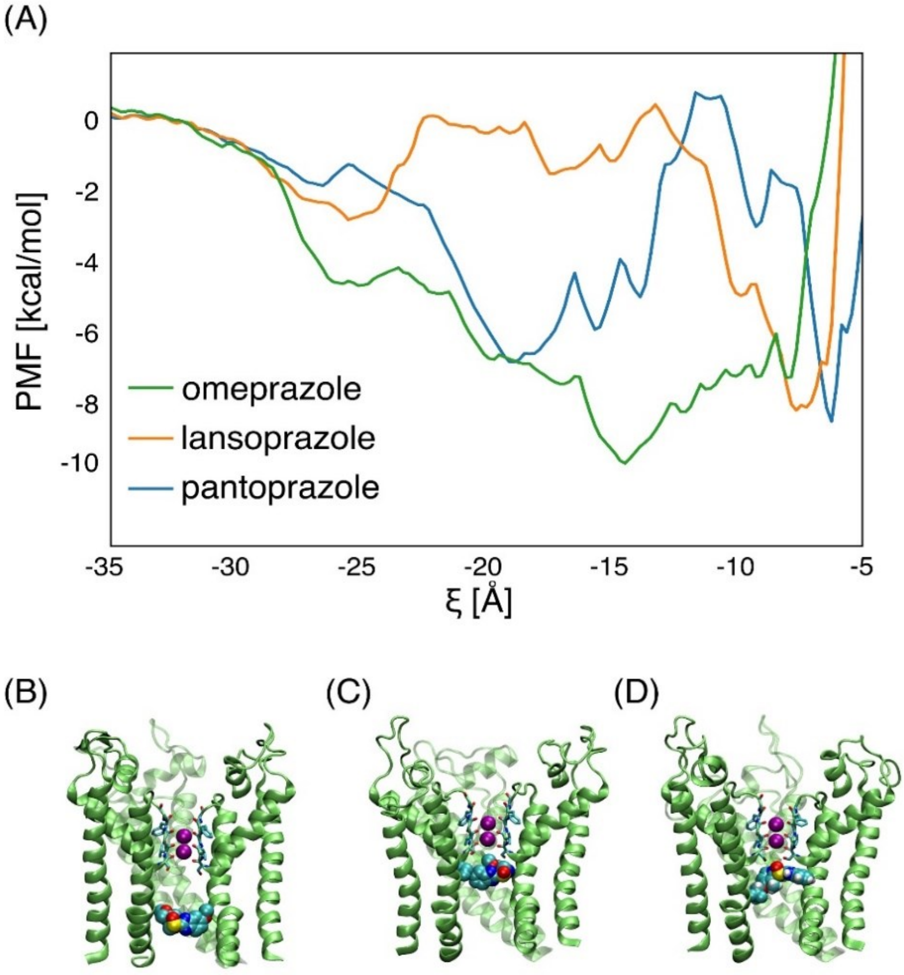

Figure 3.7. [A] PMF profiles calculated for USs. **[B]** Omeprazole, **[C]** Pantoprazole, and **[D]** Lansoprazole in the corresponding PMF minimum (VDW representation). *Reprinted from (Lazzerini et al., 2021).*

The workflow adopted in this comparative study is summarized in **Figure 3.8**.

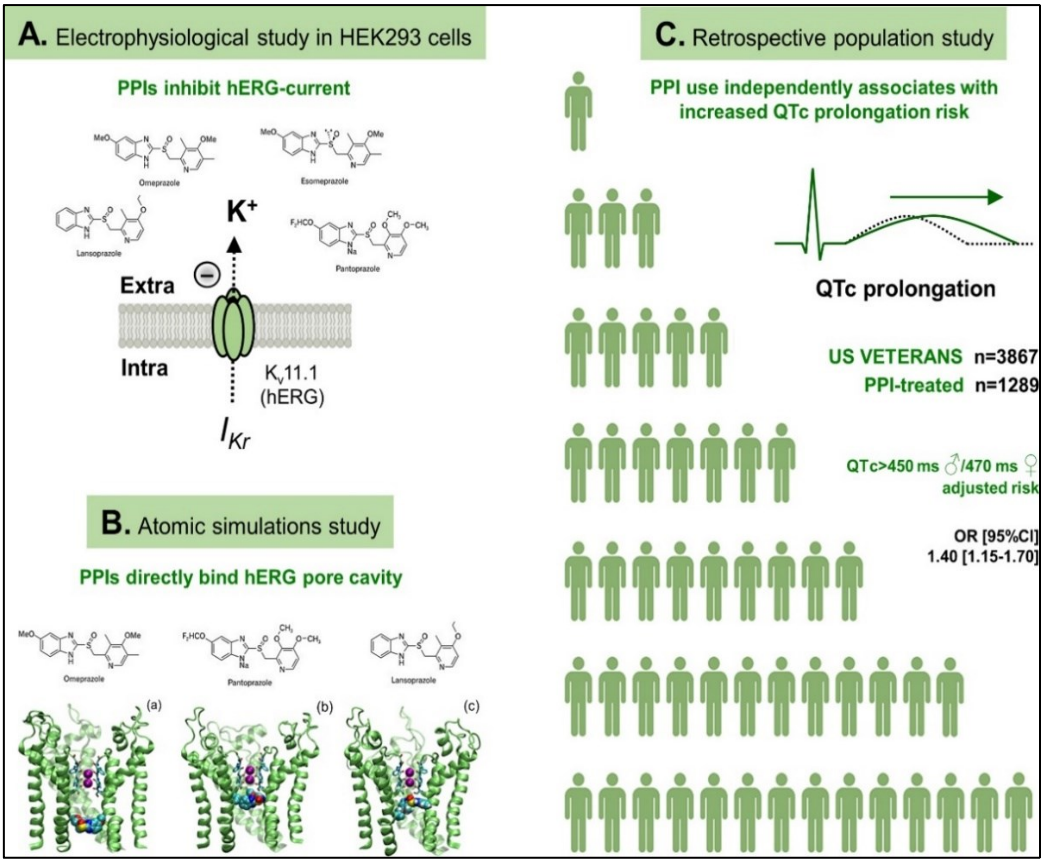
A. Electrophysiological study in HEK293 cells
PPIs inhibit hERG-current
Omeprazole
Esomeprazole
Lansoprazole
Pantoprazole
K+
Extra
Intra
Kv11.1 (hERG)
IKr
B. Atomic simulations study
PPIs directly bind hERG pore cavity
Omeprazole
Pantoprazole
Lansoprazole
(a)
(b)
(c)
C. Retrospective population study
PPI use independently associates with increased QTc prolongation risk
QTc prolongation
US VETERANS n=3867
PPI-treated n=1289
QTc>450 ms ♂/470 ms ♀ adjusted risk
OR [95%CI] 1.40 [1.15-1.70]

Figure 3.8. Graphical abstract of the workflow used for this project including electrophysiology, molecular dynamics simulations, and population data. *Reprinted from (Lazzerini et al., 2021).*

3.4.2 MD simulations of C-type inactivation

As mentioned in **Section 3.2**, the role of hERG in acquired and inherited cardiac arrythmias justifies the profound interest in characterising the atomic mechanism of its C-type inactivation, also to provide useful information about the pharmacological properties of this channel, with potential implications on drug discovery. This process involves time-dependent transition between two metastable conformations: the open-conductive (O/O) and the stable open deep C-type inactivated (O/I) states. Together with the structures associated with the closed-inactivated (C/I) and the closed conductive (C/O) states they recapitulate four distinct kinetic states of the gating cycle in the pore domain. Most of the current knowledge about the atomic mechanisms of the inactivating event has resulted from studies of the bacterial KcsA potassium channel. Evidence from functional measurements (Cordero-Morales et al., 2007), X-Ray Crystallography (Cuello et al., 2010), NMR spectroscopy (Weingarth et al., 2014) and MD simulations (Li et al., 2018), support hypothesis that C-type inactivation of the KcsA channel is due to a progressive constriction of the SF in the region corresponding to binding site S2. The constriction in S2 blocks the entry of further potassium ions, preventing ion conduction. Others voltage-gated K+ channels exhibit different mechanism of inactivation, i.e., C-type inactivation of the Shaker channel is attributed to an opening of the extracellular side of the filter (Reddi et al., 2021) (**Figure 3.9**).

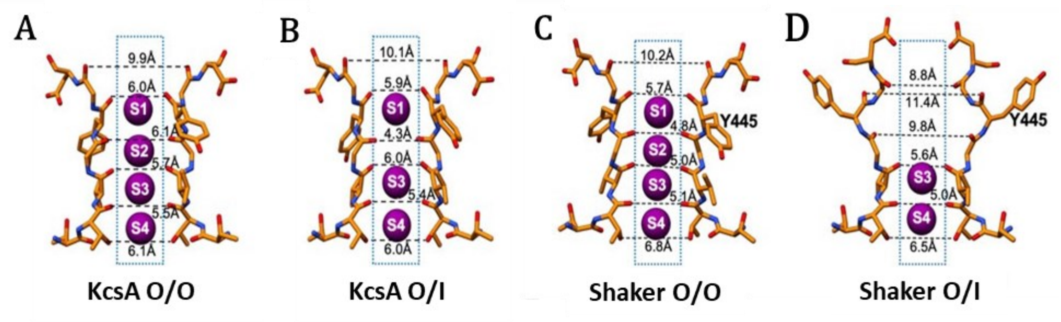

Figure 3.9. Distances between carbonyl O atoms of residues in SF in (**A**) KcsA open-conducting (PDB 3B5F), (**B**) KcsA open-inactivated (PDB 3F7Y), (**C**) Shaker open-conducting (PDB 7SIP), (**D**) Shaker open-inactivated (PDB 7SJI). Only two subunits of SF are displayed for clarity with licorice representation. K^+ ions (purple spheres) are shown at S1-S4 binding sites.

In the case of the hERG channel, there are no direct experimental data about the atomic structure of the C-type inactivated state, as the cryo-EM structure presented the SF in the canonical conductive state. In this context, MD simulations are a powerful tool to complement the experimental data. The analysis of C-type inactivation by MD simulations is not expected to reveal all the transition steps involved in the process. Inactivation rate of hERG – while being an extremely fast process in the context of ion channel kinetics – it is still far beyond the time scale accessible by MD simulations. However, by comparing the dynamics of channels with different inactivation properties, it should be possible to unveil dynamic events related to the early steps of C-type inactivation. To these purposes, we simulated, together with the wild-type hERG channel, the altered inactivating F627Y (hERG-F627Y) and N629D (hERG-N629D) mutants. These mutants were selected because the substituted amino acids modify respectively the polarity or charge of the native S0 binding site altering C-type inactivation:

1. hERG-F627Y is a fast-inactivating mutant discovered by Guo et al. after investigating the molecular determinants of hERG channels in cocaine-hERG interactions using site-targeted mutations and patch-clamp method (Guo et al., 2006).
2. hERG-N629D is a non-inactivating mutant of hERG that is stable in the conductive conformation; it was reported by Lees-Miller et al. as the first LQTS K^+ channel mutation that exhibits gain of function (Lees-Miller et al., 2000).

Methods involved a first phase of production data by MD simulations replicas using the model of the pore domain of the hERG channel described in **Section 3.3**. Models of hERG-N629D and hERG-F627Y were built by manually replacing the mutated residues in the initial wild-type model using CHIMERA software (Pettersen et al., 2004). The same equilibration protocol and simulation parameters described in **Section 3.3** was used for the three channel models. Eight independent replicas of 1 microsecond each were simulated for each channel model. In the second phase, trajectories were analysed by two different approaches to identify representative conformations and to compare channels with different inactivation properties, as described in the following sub-sections.

Clustering using ion occupancy state of the SF

Input featurization for the clustering analysis was performed calculating the occupancy states of binding sites S0-S4 by K^+. Counting of ions was achieved by **Equation 3.1**:

$$C_i = \sum_{j \in \mathcal{A}} \frac{1 - \left(\frac{d_{i,j}}{d_c}\right)^6}{1 - \left(\frac{d_{i,j}}{d_c}\right)^{12}} \qquad Equation\ 3.1$$

which calculates the coordination number C_i at the position defined by index i from contributions by atoms in selection $\mathcal{A}$. $d_{i,j}$ is the distance between atom j and the position defined by index i, and d_c is a cut-off distance after which the contribution of atom j to the coordination number decreases to zero as dictated by the 6, 12 exponents. In the case of K^+, all the potassium ions in the system were considered in $\mathcal{A}$; the index i identified the centre of the binding site, $i \in$ [S0, S1, S2, S3, S4], defined as the average position of the eight oxygen atoms delineating the site, and the cut-off distance was assumed equal to 1.4 Å, which corresponds to approximately half of the binding site length along the channel axis. In this way, the coordination number tends to 1 when a K^+ ion is close to the centre of the corresponding binding site, and it approaches 0 if a binding site is empty. Outputs were used to cluster independently each MDs data. It was employed K-means algorithm with Euclidean distance and the Ward linkage criterion, setting the optimal number of clusters equals to the maximum value of the Silhouette score. The usage of ion occupancy as clustering features is motivated by the relationship between the presence of ions at specific binding sites and the conformation of the SF. Next, clustering was evaluated by computing **Equation 3.1** to estimate the coordination number of oxygen atoms in binding sites S0-S4 (in this case, $\mathcal{A}$ equal to any oxygen atoms in the system and d_c=3.2 Å). The number of coordinating oxygen atoms could be considered an index of structural integrity of binding sites S0-S4, with a value higher than 5 corresponding to a binding site in the conductive state (Furini & Domene, 2011). In addition, it was calculated the number of K^+ ions and water molecules in the intracellular cavity, with the same approach starting from a position 2 Å below the lower boundary of S4 in the intracellular direction and extending 20 Å toward the intracellular compartment. About potassium ions, it was considered the maximum values of C_i to confirm their presence in the cavity regardless of the exact ion positions. While the presence of water molecules was evaluated by the minimum values of C_i, which can be used to identify possible regions with lack of hydration along the axis of the channel. Results show optimal number of clusters was 10, 8 and 4 respectively for hERG-N629D, hERG-WT, and hERG-F627Y (**a**, **b** and **c** in **Figure 3.10**).

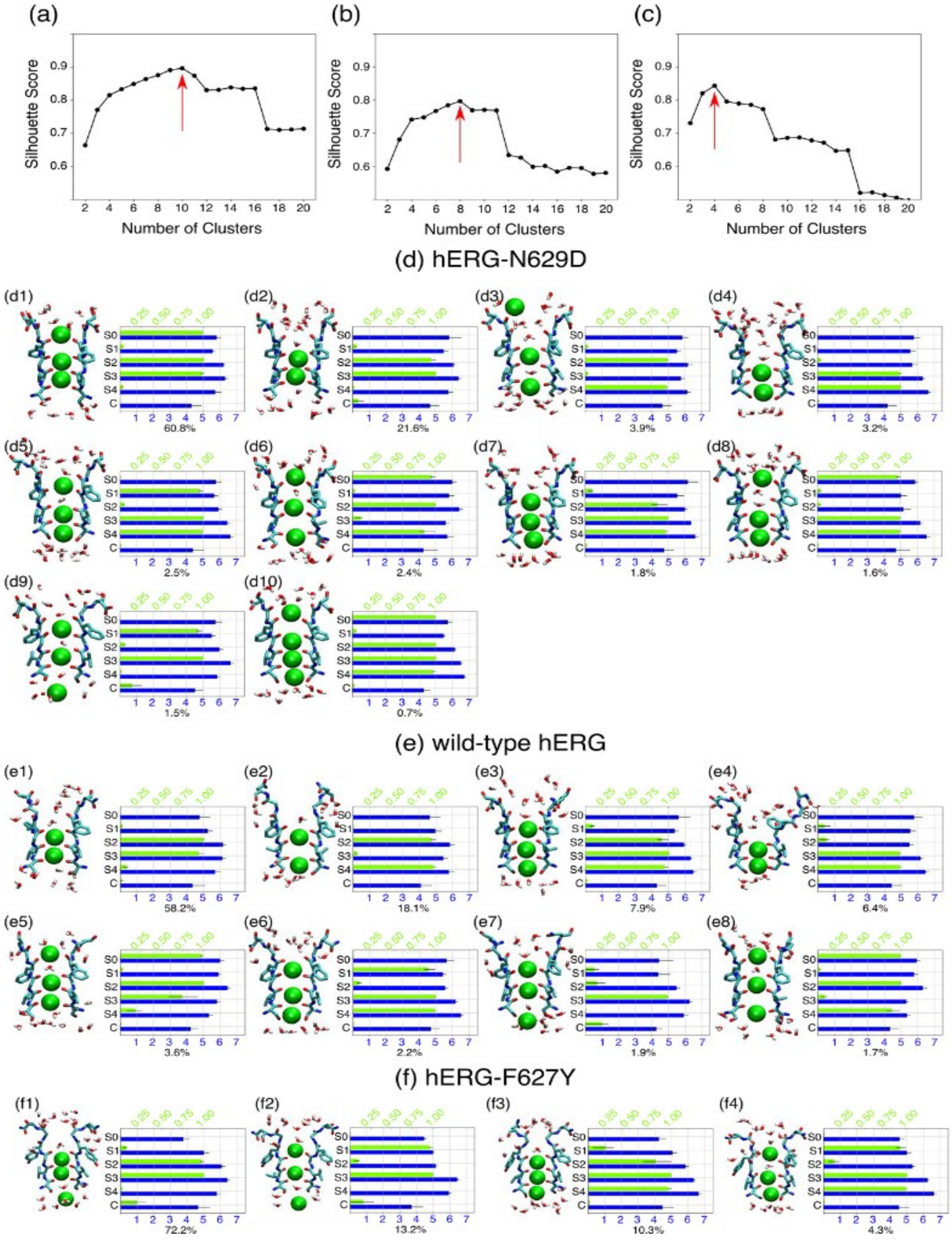

Figure 3.10. The silhouette score as a function of the number of clusters is shown for hERG-N629D **(a)**, hERG-WT **(b)**, and hERG-F627Y **(c)**. From **(d)** to **(f)** are described the results of calculation of the number of oxygen atoms and K^+ occupancy of S0-S4 and the cavity (blue and green bars, respectively), with the probability of each cluster. A representative snapshot of S0-S4 is shown illustrating K^+ ions (green spheres), water molecules and residues of the SF (licorice). *Reprinted from (Pettini et al., 2022).*

The set of ion configurations explored by hERG-WT and hERG-N629D resembles the one described in MDs of other K$^+$ channels, highlighting that all the most populated cluster share occupancy in the contiguous binding sites S2 and S3 by ions. The major difference was found in the geometry of binding sites S0 and S1, which can be attributed to the role of the presence of a ring of negative charges by Asp-629 to stabilize the opening at the outer entrance of the SF. This observation is confirmed by the evaluation of structural integrity of binding sites S0-S4: all clusters of hERG-N629D are characterized by more than five oxygen atoms in each of all binding site, while 3 cluster of the hERG-WT (about a total of 78.2% conformations) count fewer number of coordinating oxygens in S0-S1 (**e1**, **e2**, and **e7** in **Figure 3.10**). The loss of oxygen atoms at a 3.2 Å distance from the centre of S0 and S1 is also observed in all the four clusters of hERG-F627Y. The conductive structure of binding site S0 and S1 was not sampled in any of the eight trajectories of hERG-F627Y. Therefore, the different distribution of ions in the three model systems agrees with expectations to found similarity between hERG-WT and hERG-N629D, but clear differences with respect to hERG-F627Y. Analysis of the intracellular cavity by ions and water suggest a general lack of stable binding of K$^+$ ions and a remarkable constriction of the gate in the region occupied by the aromatic side chains of residues Tyr-652 for those clusters with average ion occupancy of the cavity closer to one at S4 (**d9**, **e7**, **f1**, and **f2** in **Figure 3.10**). Configurations that belong to these specific clusters correspond to channel structures with a partially dehydrated intracellular cavity with might reduce the conductance of the channel. In none of the simulations of the three model systems, a constriction of binding site S2 was observed, at odds with previous simulations using the CHARMM force field (Miranda et al., 2020). Instead, widening events were observed in S0-S1. These widening events were more likely in the rapidly inactivating hERG-F627Y channel than in the wild-type channel, and they were totally absent in the non-inactivating hERG-N629D channel. These results support the hypothesis that inactivation of hERG is more similar to what observed in the Shaker potassium channel (widening of the extracellular portion of the SF) than what observed in the KcsA channels (constriction of binding site S2).

Clustering by TICA projections

TICA is a dimensionality reduction technique, which was described in **Section 2.3**, aimed to identify high-autocorrelation linear combinations of the input features. This technique was used as an alternative to the clustering presented in the previous sub-section to reveal metastable states in MD trajectories of the three model systems of the hERG channel: the wild-type model, hERG-F627Y and hERG-N629D. The input features for the TICA analyses were:

- occupancy states of binding sites S0-S4 and of the cavity of the gate by K^+, water molecules, and coordination oxygens,
- distances among heavy atoms of the cavity of the gate,
- distances among coordination oxygens of binding sites S0-S4.

From projection of the first ten TICA components it was shown marginal changes for time intervals above 2 ns up to the first two TICA components, indicating three predominant density regions. Therefore, it was used t = 2 ns as lag time for constructing the transition matrix for clustering analysis. The calculation of inertia curve, in agreement with the mapped TICA signal, show that it was possible to restrict clustering analysis to three conformations. K-Means algorithm was set with a random initiation of centres among the features and with a stride of 10 frames, to choose centroids that minimise the inertia (**Figure 3.11**).

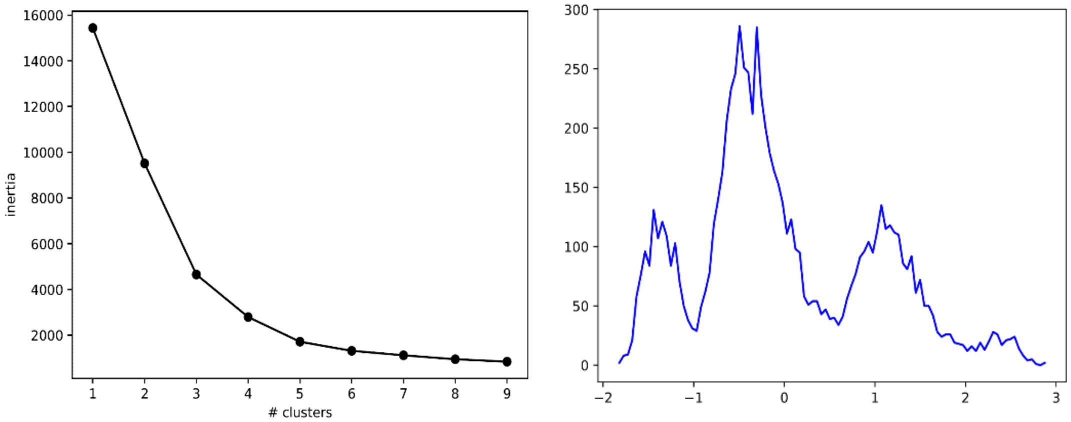

Figure 3.11. (**A**) Inertia curve of MD trajectories and (**B**) TICA signal mapped onto the first two component projections of the hERG-WT.

For each cluster, it was calculated separately:

- medium counts (with corresponding Standard Deviation) of water molecules in a cut-off of 3.2 Å from the centre of mass of residues, indexing from T634 to Y667, which form all the α-helix delimiting the intracellular cavity (pink dots in **Figure 3.12**),

- medium counts of potassium ions in a cut-off of 1.4 Å from the centre of mass of residues from T634 Y667, which form all the α-helix delimiting the intracellular cavity (green dots in **Figure 3.12**),
- minimum and maximum values of the distance between alpha carbon of residues from I647 to Y667, which form part the α-helix surrounding the intracellular cavity, about opposing chains of the tetramers (respectively blue and red squares in **Figure 3.12**),
- minimum and maximum values of the distance between coordination oxygen atoms of residues delimiting binding sites S0-S4 by ions (index from I647 to Y667) about opposing chains of the tetramers (respectively blue and red dots in **Figure 3.12**).

Results of TICA projections confirms that the three model systems differs with respect to the occupancy state of the binding site S0-S4 by K^+ ions and water molecules. In addition, these results suggest that the different distribution of water molecules in the intracellular cavity have an important role to refine the discretization step. Cluster 0 of hERG-N629D and Cluster 1 of hERG-WT confirm that S2 and S3 are always occupied by ions, and that water molecules tend to condense in the intracellular cavity in the region below S4. Moreover, the mutation of Asp629 favours the presence of ions in S0. We observed a different situation in Cluster 1 of hERG-N629D and Cluster 2 of hERG-WT, in which the probability to find an ion is higher at S3 and S4 and a gap in hydration is observed at the center of the cavity. This significant difference may be explained by the increase in structure compactness of the intracellular portion of hERG and the reduced distances among heavy atoms of the cavity of the gate. In fact, distances among heavy atoms of the cavity of the gate are reduced at the middle region, so the occupancy of the cavity by water molecules decreases. The more populated clusters show a very similar trend, with the difference that also in this case the mutation Asp629 causes the permanence of a potassium ion in S0. As far as hERG-F627Y is concerned, it is not possible to observe major differences between the two most populated clusters; instead, Cluster 1 confirms the relationship between hydration of the intracellular cavity and the presence of K^+ ions in the contiguous S3 and S4 binding sites.

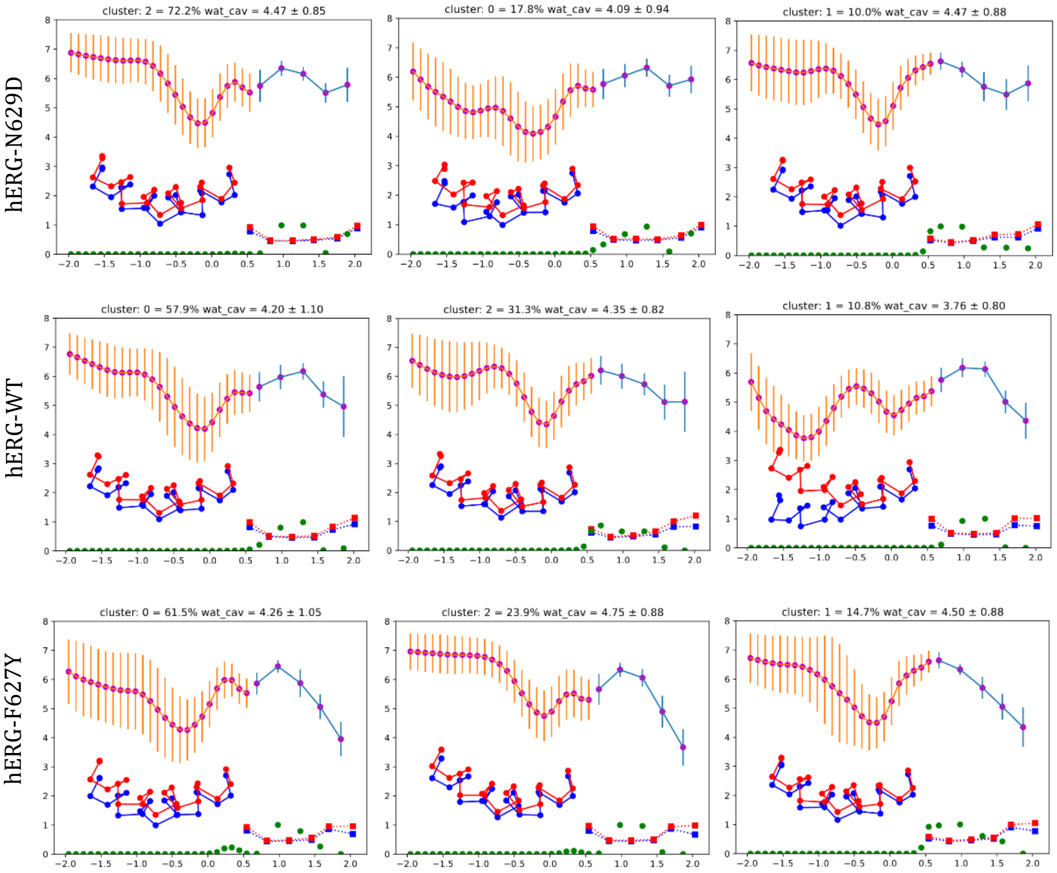

Figure 3.12. Clustering results after K-Means algorithm of the three models. The X axis is in nm and represents the distance on the z-axis between the centres of mass of the triad of the first (+2.0) and last (-2.0) amino acid that make up the α-helix delimiting cavity of the gate. The Y axis is in Å if it represents the distance between two atoms of the opposing chains, otherwise it is a real number if it denotes the count of water molecules and K^+ ions.

In conclusion, MDs suggest that fast C-type inactivation requires geometrical reorientation of residues delimiting binding sites S0 and S1, and not closure of the selectivity filter, since no closure events of the SF was sampled in the three model systems over a cumulative simulation time of 24 µs. The extent of these structural changes runs in parallel to the degree of C-type inactivation, with hERG-F627Y > hERG-WT > hERG-N629D. The hypothesis that early steps of inactivation are consistent with an initial widening of the SF is sustained by experimental data, confirming the robustness of employing MD simulations as a complementary tool to experimental analyses to help revealing the details of C-type inactivation (Domene & Furini, 2009; Gang & Zhang, 2006). Moreover, since C-type inactivation is dictated by the subtle atomic interactions of the SF and ions, water molecules, and protein residues, it is not surprising that the details of C-type inactivation may be channel dependent. The effect of inactivation on the properties of the cavity might have an impact on the drug recognition mechanism.

3.4.3 Drug binding profile associated to inactivation

In 2021, Asai et al. determined the cryo-EM structures of the hERG channel in the presence and absence of Astemizole, a well-known potassium channel inhibitor that increases the risk of potentially fatal arrhythmia (Asai et al., 2021). Although the quality of EM densities for astemizole was limited, structures validated the open state model determined by Mackinnon and colleagues and provided insights into the binding sites of hERG inhibitors, which have been predicted by various other studies but have not yet been characterized from the actual 3D structure (Wang & MacKinnon, 2017). The orientation of astemizole within the hERG, and its relationship to the amino acid residues in the vicinity of this inhibitor, may explain the promiscuous bind interaction of drugs to the hERG channel. On the other hand, the pathways by which inhibitors can access hERG and the relationship between inactivation and drug binding are still largely unknown. Experimental data revealed that C-type inactivation and the pharmacological properties are linked (Mitcheson et al., 2000). However, to the best of our knowledge, in computational analysis of drug binding events, the open-conductive structure of the channel is always adopted. Since the previous analyses identified metastable states that might be involved in the early stages of inactivation, we decided to investigate by docking calculations how (and if) these structures exhibited different binding properties to well-known hERG blockers. The inhibitor-binding site of astemizole, reported by Asai et al, was compared with the three clusters. Meanwhile, the same prediction was performed on other two drugs to investigate how the binding-affinity is dependent to the interactions with Ser624, Tyr652 and Phe656.

For these analyses, we adopted Dofetilide and Moxifloxacin:

- Dofetilide is a methane-sulfonanilide compound, analogue of MK-499, about what is reported many experimental studies of its hERG inhibition effect (Stansfeld et al., 2007). In addition, it offers a three times bigger volume of topological polar surface area (121,57 Å) than astemizole, with a more flexible folding (11 rotatable bonds) and the possibility to establish more hydrophobic interactions with residues of the cavity (7 hydrogen bond acceptors and 2 hydrogen bond donors). These chemical properties combined with astemizole similar high binding affinity make dofetilide an excellent candidate for our study of binding sites in the intracellular cavity of the channel;
- Moxifloxacin is a fluoroquinolone class of antibacterial, widely prescribed for the treatment of infections. Kang et al. reported clinically relevant positive voltage dependence of blockade of the hERG channel (Kang et al., 2001). However, data showed that only high concentration impact to inhibit potassium trafficking of hERG channel. Moxifloxacin binds to the open state of the channel and, to a lesser extent, the inactivated state, and drug binding occurs at the aromatic residue Tyr652 but not Phe656 in the inner cavity of the channel, making it an interesting case-study. Alexandrou et al. reported that mutagenesis of the S6 helix residue Phe656 to Ala failed to eliminate or reduce the Moxifloxacin-mediated block whereas mutation of Tyr652 to Ala reduced Moxifloxacin block by ~66% (Alexandrou et al., 2006).

Briefly, this study is aimed at:

1) testing the hypothesis that drug binding features are dependent on the channel metastable states observed in MD trajectories;
2) Validating *in silico* pharmacophore prediction of astemizole with Asai et al. structural information, by comparing with conformers of moxifloxacin and dofetilide after molecular docking simulations;
3) Investigating the contribution of known interacting residues (Ser624, Tyr652 and Phe656) to the blockade, and which is the best pose that drugs can assumes in the inner cavity.

The protocol of this study involves a phase of production data, in which configuration of the drug binding protein was designated by the AutoDock Vina tool after water molecules and phospholipids removal (Eberhardt et al., 2021). It was performed a docking with rigid bonds and a sufficiently large cubic box to include the investigated residues in all the clusters, so as to more easily compare the results. The binding cavity was identified by selecting an area of 17 Å positioned at 15 Å from the center of the SF along z-plane. The selected area covered aromatic (Tyr652, Phe565) and polar (Tyr623, Ser624, Val625, Ser649 and Gly648) residues located on the pore helix and lining the inner cavity.

Vina calculates the energetic score by **Equation 3.2**:

$$\Delta G = \sum_{i<j} f_{t_i t_j}(r_{ij}) \qquad Equation\ 4.2$$

where the summation is over all the pairs of atoms that can move relative to each other, normally excluding 1–4 interactions. Each atom i is assigned a type t_i, and a symmetric set of interaction functions $f_{t_i t_j}$ of the interatomic distancer_{ij} is defined.

In order to validate the input structures for conformers analysis it is essential comparing energetic score with predicted IC_{50} (pIC_{50}) values. The best pose of each frame was used to measure:

- minimum radius distance (R_{min}) between the drug and closest residues known to be directly involved in the inhibition mechanism, Ser624 (R_{min}^{S624}) and Tyr652 (R_{min}^{Y652}) (Table 3.1 and Figure 3.13). This measurement estimates where the drug ranks in relation to the SF for each cluster.
- Distance between the two far atoms of the molecular structure of compounds (R_{min}^{drug}), able to establish interactions with residues of the protein (Table 3.1 and Figure 3.13). This data does not provide us with direct information on the spatial relationships of the elements that characterize a 3D pharmacophore, because it does not take into account the angles of rotation around single bonds and/or inversion of atomic centres: it only indicates whether the structural conformation opens proportionally to the 'increase in the value of the distance between the two atoms. However, in the case of a ligand-receptor interaction characterized by the establishment of a dense network of hydrogen bonds or a large number of other strong interactions, it is very likely that the ligand binds to the receptor in a conformation that corresponds to a low energy conformation for the molecule in

vacuum. Therefore, by comparing the average values of R_{min}^{drug} with the minimum energy resulting from the calculation of the docking simulation, in each cluster, it is possible to have an idea of the degree of distension that the probable bioactive conformation can assume: if the value of Vina Score decreases as R_{min}^{drug} increases, the bioactive conformation could assume a spatial distribution bound to a main axis, limiting the number of interactions that can be established between the four channel chains, and vice versa.

According to the rationale of this analysis, the position of the drug with respect to the distance between the two residues involved in the binding and the probable spatial conformation of the drug in the cavity can be useful to validate at a computational level the different value of pIC_{50} obtained from the experimental data. Furthermore, the subsequent analysis of the interaction profile can highlight how much the inner region of the cavity is involved compared to the SF in the inhibition of hERG, i.e., if Tyr656 and Phe652 show a higher number of bond donor atoms for the drug with respect to residues close to the SF it is reasonable to expect the intracellular region of the cavity to be essential for the inhibition mechanism in the sampled metastable state. Non-covalent interaction analysis of the most representative conformation of each cluster was performed with Protein–Ligand Interaction Profiler (PLIP) web server tool (Adasme et al., 2021).

Results from the Vina scoring function confirms astemizole and dofetilide are strong inhibitors, while moxifloxacin is a weak inhibitor, in accordance with the predicted IC_{50}, reported by Cavalli et al. and Munawara et al. (Cavalli et al., 2012; Munawar et al., 2018): the mean value of ΔG in all frames decrease when pIC_{50} increase (**Table 3.1**). Frequencies distribution of R_{min}^{S624}, R_{min}^{Y652}, and R_{min}^{drug}, confirm that significant differences exist in the orientation and structural relaxation of the drugs among the clusters. This agrees with the difference in cavity and SF observed among clusters, and it provides evidence for proceeding to analyse the interaction profile of the most representative conformation of each metastable state. Indeed, frequencies distribution of R_{min}^{S624} and R_{min}^{Y652} compared with ΔG indicate that the binding affinity of astemizole and moxifloxacin is strictly related to the interactions established with Ser624 and Tyr652, while the binding affinity of dofetilide is correlated to variation of the R_{min}^{drug}. This observation agrees with the different chemical properties of dofetilide with respect to astemizole and moxifloxacin (**Figure 3.13**).

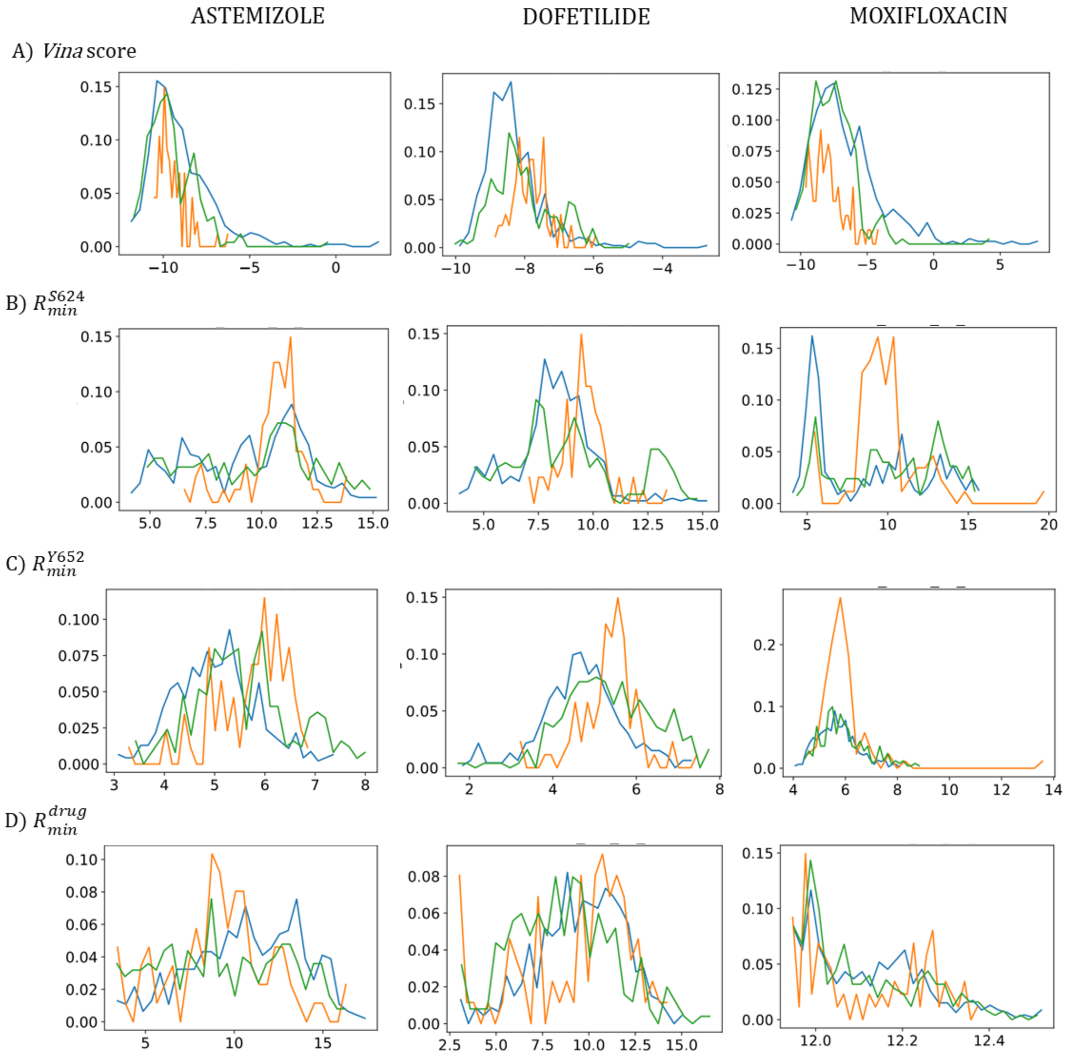

Figure 3.13. Frequency distribution of (**A**) Vina Score (**B**) R_{min}^{S624}, (**C**) R_{min}^{Y652}, and (**D**) R_{min}^{drug}; Cluster 0 is coloured in blue, Cluster 1 in orange, Cluster 2 in green.

Classes	**Vina Score** (Kcal/mol)	$\mathbf{R_{min}^{S624}}$ (Å)	$\mathbf{R_{min}^{Y652}}$ (Å)	$\mathbf{R_{min}^{drug}}$ (Å)
ASTEMIZOLE (pIC_{50}= 9,0 (IC_{50}=1x10^{-9} M)) (Zhou et al., 1999)				
All	-9,60	10,15	5,25	10,25
Cluster 0	-9,50	9,69	5,02	10,77
Cluster 1	-9,80	10,76	5,93	9,36
Cluster 2	-9,70	10,15	5,45	9,39
DOFETILIDE (pIC_{50}= 7,9 and pK_i = 8,2 (K_i = 6,4x10^{-9} M)) (Singleton et al., 2007)				
All	-8,30	8,55	4,94	9,46
Cluster 0	-8,50	8,30	4,68	9,81
Cluster 1	-7,90	9,52	5,42	10,30
Cluster 2	-8,20	8,61	5,26	8,66
MOXIFLOXACIN (pIC_{50}= 3,78 (IC_{50}=1,65x10^{-5} M) (Abi-Gerges et al., 2011)				
All	-7,50	9,38	5,72	12,09
Cluster 0	-7,20	8,41	5,64	12,10
Cluster 1	-8,20	9,55	5,79	12,09
Cluster 2	-7,70	10,18	5,83	12,06

Table 3.1 Median of measured Vina Score, R_{min}^{S624} , R_{min}^{Y652}, and R_{min}^{drug}, related to all frames and specific cluster.

In the following paragraphs, the results of PLIP analyses for the three compounds are reported.

Astemizole. In the most populated cluster, astemizole binds perpendicular to the pore, where it can establish interactions with residues surrounding the cavity (specifically hydrogen bonds with Ser624 and Phe656 side chains, and hydrophobic interactions with oxygen atoms of Tyr652, Ala653 and Phe656). A similar binding pose was observed in snapshots of the second most populated cluster, that also allow π-π stacking interaction with aromatic rings of Tyr652. Instead, in the less populated cluster, corresponding to structures with cavity radius smaller than in the previous two clusters, the drug is located parallel to the cavity making only interactions with Tyr652 (**Table 3.2** and **Figure 3.14**).

Residues	Occurrence			R_{min}		
	Clust 0	*Clust 1*	*Clust 2*	*Clust 0*	*Clust 1*	*Clust 2*
Hydrophobic Interactions						
Y652	4	4	2	3,30	3,65	3,73
A653	none	none	1	none	none	3,73
F656	2	1	2	3,86	3,62	3,60
H-Bonds						
S624	1	none	none	3,69	none	none
S660	none	none	1	none	none	2,80
π-π Stacking						
Y652	none	none	1	none	none	4,96

Table 3.2. Protein-Ligand Interaction Profile of the most representative pose of Astemizole docked in the hERG cavity.

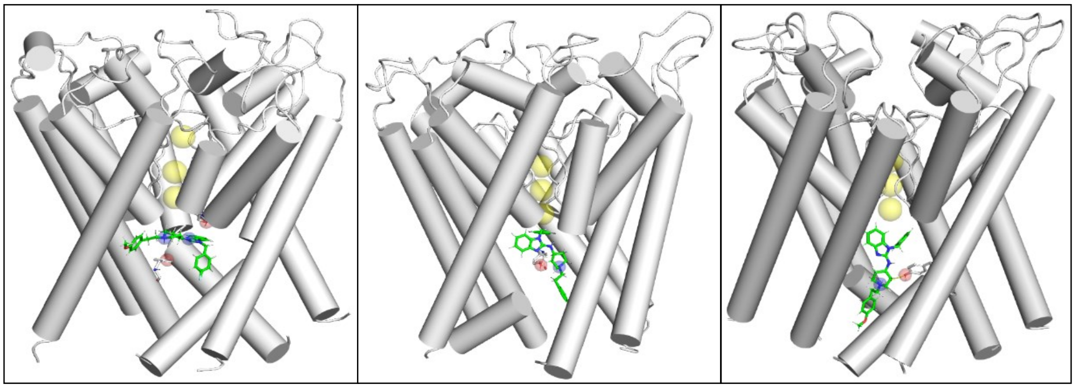

Figure 3.14. Most representative pose of Astemizole docked in each cluster.

Dofetilide. In the most populated cluster, dofetilide binds parallel to the pore folded in a "C" shape conformation with the two donor Oxygen atoms equidistant to Ser624 and Tyr652, such as highlighted from pharmacophore model (yellow colour in the **Table 3.3**). It can establish interactions with residues surrounding the cavity: hydrogen bonds with Ser624 and Ala653, hydrogen bonds with Ser624 and Tyr652, and π-π stacking with aromatic rings of Tyr652. In Cluster 2, the pose of the compound is superimposable to the pose of Cluster 0, but, due to different orientation of the terminal carboxyl, it cannot establish interactions with Ser624. Instead, in Cluster 1, the central scaffold of the drug responsible for H-bonds interactions is located closed to the Ser624, Tyr652 and Ser660 (**Figure 3.15**).

Residues	**Occurrence**			$\boldsymbol{R_{min}}$		
	Clust 0	*Clust 1*	*Clust 2*	*Clust 0*	*Clust 1*	*Clust 2*
Hydrophobic Interactions						
Y652	5	2	4	3,52	3,6	3,59
A653	1	None	1	3,74	none	3,57
H-Bonds						
S624	2	2	none	2,85	2,96	none
Y652	3	3	1	2,85	3,15	2,97
S660	none	1	none	none	3,49	none
π-π Stacking						
Y652	1	none	none	5,29	none	none

Table 3.3. Protein-Ligand Interaction Profile of the most representative pose of Dofetilide docked in the hERG cavity.

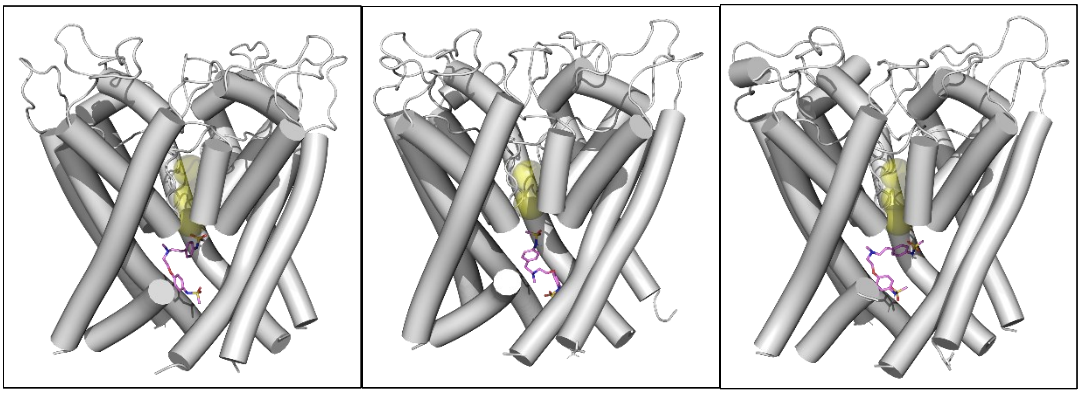

Figure 3.15. Most representative pose of Dofetilide docked in each cluster.

Stansfeld et al. reported experimental results that show a similar binding profile between dofetilide and astemizole (Stansfeld et al., 2007). In the present study, molecular docking predictions of astemizole and dofetilide do not report similar interactions profile in any of the most representative pose of each cluster, even if the binding affinity score match with the experimentally measured pIC_{50}. These discrepancies could be explained by the fact that the standard protocol of Molecular docking refinement phase considered just the best 10 poses, due to save time-consuming performance. Thus, it is plausible to have a docking prediction in accordance with experimental data just increasing exhaustiveness of the prediction parameters, because it is possible to cluster more poses. However, the prevalence of possible hydrophobic interactions and hydrogen bonds that dofetilide can established with Tyr652 respect to Phe656 underline the importance of this interaction and dissociation from SF binding site for the sensitivity to the blockade, as reported by Gomez-Varela (2006) (Gómez-Varela et al., 2006).

Moxifloxacin. The pose of the drug in Cluster 0 a Cluster 2 is similar: the central scaffold of moxifloxacin intercepts the Z-plane at about 45°. As shown in **Figure 3.16,** the pose differed for the docked positioning in the box: the compound docked in the most populated cluster is closer to the selectivity filter allowing more interactions with amino acids surrounding the interior part of the cavity, while in Cluster 2 the drug is located closer to the entrance permitting possible interaction with Tyr652, Phe656, and Asn658 (**Table 3.4**). Instead, in Cluster 1, moxifloxacin assumes a stretched pose perpendicular to the XY-plane with the external carbonyl rings oriented towards the SF (**Figure 3.16**).

Residues	**Occurrence**			$\boldsymbol{R_{min}}$		
	Clust 0	*Clust 1*	*Clust 2*	*Clust 0*	*Clust 1*	*Clust 2*
Hydrophobic Interactions						
Y652	5	4	3	2,80	3,57	3,14
A653	none	1	none	none	3,72	none
F656	3	none	3	3,48	none	3,52
H-Bonds						
S624	2	1	none	3,42	4,02	none
Y652	1	1	none	3,12	3,43	none
F656	1	none	1	3,08	none	3,17
π-π Stacking						
Y652	1	2	none	4,22	4,61	none

Table 3.4. Protein-Ligand Interaction Profile of the most representative pose of Moxifloxacin docked in the hERG cavity.

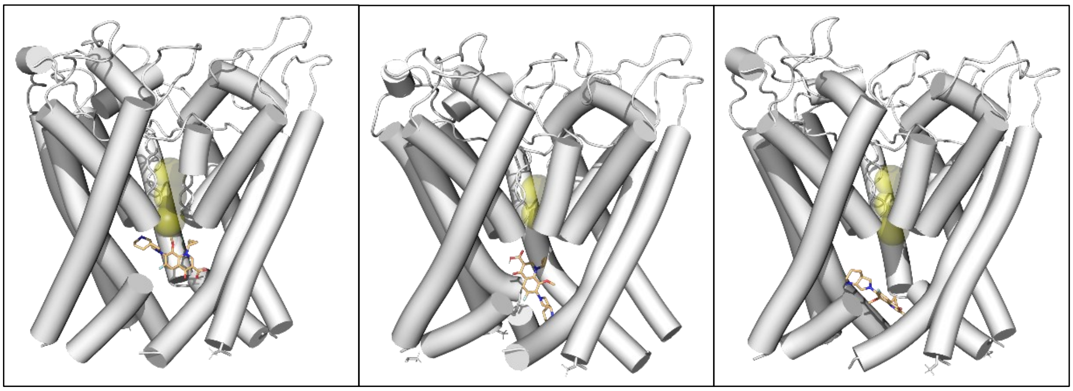

Figure 3.16. Most representative pose of Moxifloxacin docked in each cluster.

The most populated cluster of moxifloxacin is more suitable to represent the state closest to inactivation by drug binding, because:

- binding pose of the drug move parallel to the z-plane, leading to establish the higher number of strength interactions with Y652, and adjacent residues, versus the minimum steric ingombrance;
- R_{min}^{S624} and R_{min}^{Y652} increase according to decrease of the number of possible interactions with Ser-624 and Phe-656 (**Table 3.1**).

As dofetilide prediction, molecular docking of moxifloxacin reported higher number of probable contacts with Tyr652, concordant with an interaction in the channel inner cavity, resembling the mechanism of hERG blockade observed by Alexandrou et al. (Alexandrou et al., 2006).

Chapter 4 – Side projects

4.1 DNA-binding to LacI repressor protein

DNA interactions with proteins are necessary for many of its functions: DNA-binding proteins have a central role in all aspects of genetic activity within an organism, such as transcription, packaging, rearrangement, replication, and repair. It is therefore essential to investigate the nature of DNA-protein complexes to understand how these cellular processes take place. For example, the expression degree of genes is regulated by a broad number of proteins, named transcription factors, which recognize and bind specific DNA-sequences (Luscombe et al., 2000). DNA-Protein binding is mediated by many factors such hydrogen bonds (H-bonds), Van der Waals (VdW) contacts, DNA shape, protonation states, flexibility, and many others (Hogan & Austin, 1987; Luscombe et al., 2001); while the stability of DNA-Protein complexes is linked to DNA–backbone interactions, proteins recognize specific DNA sequence by forming bonds between amino-acid side chains and DNA bases (Luscombe et al., 2001; Rohs et al., 2010; Rohs et al., 2009). Consequently, mutations occurring in DNA-binding proteins that alter the physical and chemical properties of the binding interfaces may influence binding specificity and affinity (Luscombe & Thornton, 2002; Treisman et al., 1989). There has been an increasing interest in the role that DNA-binding proteins may play in medicine and biology, and it has been shown that alteration in DNA-Protein binding affinity is involved in heart and neurological diseases, as well as cancer. Hence, understanding their molecular effects is crucial for deciphering disease origins and pursuing treatment (Chahrour et al., 2008).

DNA-Protein interactions are of mainly two types: specific or non-specific interactions (Ganji et al., 2016; Hudson & Ortlund, 2014). In general, proteins interact with the major groove of B-DNA, because it exposes more functional groups that identify a base pair (Bewley et al., 1998). Recent single-molecule experiments showed that DNA binding proteins undergo rapid rebinding in order to bind in the correct orientation for recognizing the target site (Redding & Greene, 2013). DNA complexes with structural proteins, within chromosomes, are perfect examples of non-specific DNA-protein interactions. These proteins organise the DNA into a compact structure called chromatin. In eukaryotes, this structure involves DNA binding to a complex of small basic proteins called histones. The histones form a disk-shaped complex called a nucleosome, which contains two complete turns of double-stranded DNA packed around its surface (Dame, 2005). These interactions are formed through basic residues in the histones making ionic bonds to the acidic sugar-phosphate backbone of the DNA, and are, thus, widely independent of the base sequence (Harteis & Schneider, 2014). Chemical alteration of these basic amino acid residues includes epigenetic mutations, such as methylation, acetylation, and phosphorylation (Javaid & Choi, 2017). These chemical modifications alter the strength of the interaction between the DNA and the histones, making the Accessible Surface Area (ASA) at the interface accessible to transcription factors and changing the rate of transcription (Workman & Kingston, 1998). Other non-specific DNA-binding proteins in chromatin include the high-mobility group (HMG) proteins, which fold or distort DNA. Biophysical studies highlighted that these architectural HMG proteins bind, fold and loop DNA to perform their biological functions (Reeves, 2010). These proteins are important in bending arrays of nucleosomes and arranging them into the larger structures that form chromosomes. In contrast, other proteins bind to specific DNA sequences. Each transcription factor recognizes one specific set of DNA sequences, and activates or inhibits the transcription of genes linked to these sequences (Frietze & Farnham, 2011). The transcription factors do this in two ways: binding the RNA polymerase, directly or through other mediator proteins, or alternatively, binding enzymes that modify the histones at the promoter. This alters the accessibility of the DNA template to the polymerase. Consequently, changes in the activity of one type of transcription factor can interest lots of genes (Grove & Walhout, 2008). So, these proteins are the principal targets of the signalling pathways that control responses to environmental changes or cellular differentiation (Duronio & Xiong, 2013). The specificity of these transcription factors interactions with DNA come from the proteins making multiple contacts with the DNA sequence, until they recognize their binding site. Descriptions of DNA-binding proteins considering sequence-specificity, and competitive and cooperative binding of proteins of different types are usually conducted with the help of computational methods, to establish a collaboration among the major -omics tools aimed to perform a more accurate and less time-consuming process (Rastogi et al., 2018). The purpose of the present research was to explore how residue mutations might impact on the nonspecific protein-DNA interactions, and if it was possible to identify general rules for modulating the movement of proteins along nonspecific DNA, inspired by a previous bioinformatic analysis (Gardini et al., 2017). The analysis protocol was based on fully atomistic molecular dynamics simulation and MM-PBSA energetic analysis to quantify the free-energy landscapes that underlie the dissociation kinetics of the lacI repressor in dimeric and monomeric conformation. The atomic structures of the protein-DNA complexes were based respectively on an altered specificity mutant of the *lac* repressor headpiece that mimics the *gal* repressor (available in RCSB database at PDB ID 2BJC) (Kopke Salinas et al., 2005) (**Figure 4.1-A**) and on the lacI repressor complexed to a nonspecific B-DNA template (available in RCSB database at PDB ID 1OSL) (Kalodimos et al., 2004) (**Figure 4.1-B**).

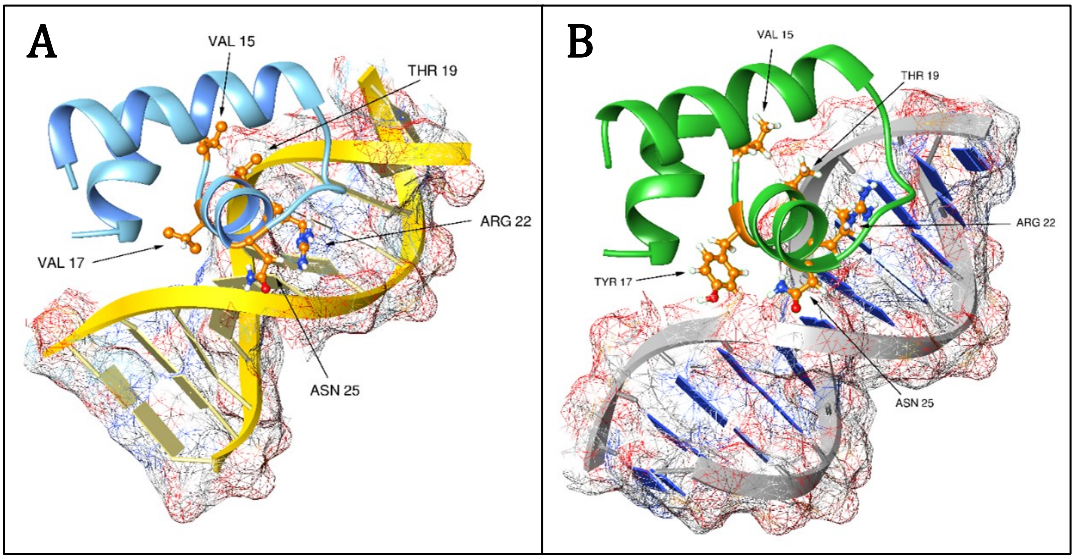

Figure 4.1. [**A**] Cartoon representation of the atomic model of the *lacI* repressor complexed to its specific DNA (2BJC) and [**B]** to a non-specific DNA (1OSL). Residues selected to mutations are displayed, and labelled, in ball and stick (orange).

The headpiece 62 (HP62) molecules of alpha chains were chosen for proteins and their bended double-helix portion for DNA fragment (5'-CGATAAGATAT-3' of 1OSL; 5′-GAATTGTGAGC-3′ of 2BJC). We selected 5 mutants for the repressor complexed to specific DNA and 9 mutants for the LacI monomer 2BJC. All residues mutated are centred in the α-helix interacting with the major groove of DNA (**Figure 4.1** and **Table 4.1** for the list) and were selected by considering three main parameters of the side chains able to alter the recognition affinity to DNA: size, VdW, and electrostatic properties.

The standard single-trajectory MM/PBSA protocol was employed to estimate binding entropy of each system. Dynamics of dimeric assembly were studied following the same ensemble with same energetic parameters and Force Field but using the 3-trajectories variant approach. In this case, it is crucial to consider the free energies of the complex in a bounded state or the single entity in the unbounded state, due to the presence of S-S bond in Cys52 interface that could lead to estimation accuracy. The initial atomic coordinates of the two simulated systems and parameters of the ensemble were generated using CHARMM-GUI for the Gromacs MD engine (Jo et al., 2008; J. Lee et al., 2016). The protein-DNA complexes were solvated by ~12500 TIP3P water molecules and 0.15 M NaCl was used to neutralize the electrical charge. The 36m version of the Charmm force field was used, in combination with ion parameters by Joung and Cheatham. Van der Waals interactions were truncated at 12 Å. Long-range electrostatic interactions were treated using the Particle Mesh Ewald method with a real-space cut-off of 12 Å, and a grid spacing of 1,2 Å. Newton's equations of atomic motion were integrated by the Verlet algorithm with 2 fs time steps. The LINCS algorithm was used to constraint bonds and angles with the hydrogen atom. The temperature was controlled at 300 K by coupling to a velocity-rescaling scheme thermostat with a damping coefficient of 1 ps^{-1}, due to ensure a more efficient kinetic energy distribution. Pressure of 1 atm was maintained by coupling the system to a Berendsen thermostat, with a damping constant of 2 ps. The equilibration protocol consisted of 50.000 steps of energy minimization by steepest descent algorithm, with an initial force step size of 0,1 Å, followed by a total of 30 ns in the NPT ensemble with timestep equal to 1 fs and 10 ns in the NPT ensemble with timestep equal to 2 fs. During the equilibration protocol, position restraints on protein and DNA atoms were gradually reduced to zero, except for the backbone of the two nucleotides at the 5' and 3' extremities which were maintained also in production trajectories. All the systems were subjected to a step of production, 1 μs long, in NPT condition with a time step equal to 2 fs each. Conformational snapshots were saved for each trajectory at 100 ps intervals, leading to a database of 150000 snapshots which represents 150 Gb of data including solvent. Convergence assessment and structural analyses were performed by integrated tools of Gromacs. The initial 200 ns of the trajectories were treated as an extended equilibration period and only the remaining 800 ns of simulations were analysed. Energetic components were estimated with the approximate post-processing end-state method available in the *g_mmpbsa* python script (Kumari et al., 2014). Solvation energies and forces were determined with respect to a homogeneous medium with a dielectric constant of 1. Ionic strength was set to 0,15 M, radius of positive charged ions was set to 0,95 Å, radius of negative charged was set to 1,81 Å, and solvent probe radius was set to 1,4 Å. The linearized PB equation was solved using grid spacing of 0,5 Å, internal and external dielectric constants of 8 and 80, respectively. The non-polar solvation free energy calculation is calculated from the solvent accessible surface area using the traditional one component method. In this approach the surface tension, γ, was set to 0,00542 kcal mol^{-1} Å^{-2}) and the offset to 0,92 kcal mol^{-1}.

Results from the MMPBSA calculation are summarized in **Table 4.1**. As expected, 2BJC mutations of internal amino acids in the alpha-helix show a progressive loss of affinity for DNA, which can be attributed to the concomitance of two factors: one of a structural type and one of a functional type. Hydrophobic index and occurrence of amino acids with negatively charged side chain can be used as tool to modulate protein mobility along DNA chains. Therefore, substitutions of Asparagine and Threonine with Glycine, a non-polar amino acid with aliphatic side chain, may represent a very feasible way to decrease affinity to DNA, as confirmed from positive values of ΔΔG. From a functional point of view, the range of specification may be attributable to the variation in non-bonding interactions with water molecules. Consequently, to these two factors the V17Y-A18Q to be the negative control on the other mutants, which appear to have a progressively lower affinity for DNA as it moves away from the alpha-helix. In the case of 1OSL, both the WT mutant and the T19G-N25G mutant, used as negative controls, have been shown to assume an unexpected state that invalidates the initial hypothesis: the WT seems to be the clearly more unstable complex of all. Furthermore, the residues known to be responsible for the interaction between DNA and protein do not show significant differences in terms of binding energy. These discrepancies might be related to a rude quality of the sampling and/or to inaccurancies of the Force Field.

LacI-monomer	ΔG *(KCal/mol)*	ΔG – ΔGWT *(KCal/mol)*
2BJC		
Wildtype	-31,84	-
V17Y_A18Q	-46,28	-14,44
V17Y_A18Q_T19G	-25,38	6,46
V17Y_A18Q_N22G	-21,30	10,54
V17Y_A18Q_N25G	-9,12	22,72
V17Y_A18Q_T19G_N25G	-23,95	7,89
1OSL		
Wildtype	-5,02	-
T19G_N25G	-35,57	-30,55
N25R	-26,79	-21,77
V15R	-27,34	-22,32
V15G	-22,33	-17,31
T17G	-23,01	-17,99
T19G	-22,73	-17,71
T19R	-10,91	-5,89
N22G	-7,42	-2,4
N25G	-23,41	-18,39

Table 4.1 Energetic results from MMPBSA calculation.

4.2 Peptide-MHC complexes in Sars-Cov2

The pandemic COVID-19 (COronaVIrus Disease 2019) is caused by the SARS-CoV-2 virus. The World Health Organisation (WHO) on March 3rd, 2020, reported that the lethality rate of the virus is 3,4%. One of the main characteristics of the disease is the high variability of symptoms in the population. Most affected patients have mild symptoms, while others develop acute pneumonia and require mechanical ventilation. Approximately 20% of cases require hospitalisation and 5% of them the intensive care unit. Patients who require respiratory assistance are often the elderly and/or are affected by previous pathologies, but this is not sufficient to explain the huge variability in response. The host genetic component, together with age and gender, leads to a different immune response or permissiveness to the virus, according to some studies (Brodin, 2021). Indeed, it has been reported that genetic factors, such as the type of Human Leukocyte Antigens (HLA), can affect the progression and the severity of the disease (Amoroso et al., 2021). The HLA locus is a polymorphic gene complex present on chromosome 6 (6p21.3), that encodes approximately 27,000 variant surface molecules, grouped into class I and class II, that bind peptides derived from different sources, determining resistance to infectious diseases and susceptibility to autoimmunity (Dendrou et al., 2018; Klein & Sato, 2000a, 2000b). Therefore, HLAs are essential for immune responses to viral infections, as they present the pathogen antigens to CD8$^+$ and CD4$^+$ T cells, promoting the elimination of infected cells and the production of antibodies (Dendrou et al., 2018). In addition, some HLA class I and II molecules can interact with specific natural killer (NK) cell receptors. The combination of HLA alleles is crucial during the immune response and can make an individual more susceptible to a specific disease (Choo, 2007; Klein & Sato, 2000a, 2000b), such as reported in virus-related diseases (SARS, influenza, HIV infection, hepatitis, cytomegalovirus (CMV) and Herpes simplex Virus 1 (HSV-1)) (Hu et al., 2014; McAulay et al., 2007; Singh et al., 2007). Although Genome Wide Association Studies (GWAS) have not elucidated the role of HLA variability in susceptibility to SARS-CoV-2, targeted analyses have identified specific HLA variants associated with severity of COVID-19 in different populations (Augusto & Hollenbach, 2022; Francis et al., 2022; Nguyen et al., 2020; Parker et al., 2021). Population-based studies across Italy reported that HLA diversity could impact on disease severity and susceptibility (Amoroso et al., 2021; Correale et al., 2020; Guerini et al., 2022; Novelli et al., 2020; Pisanti et al., 2020). Among them, Pisanti and colleagues reported that the HLA-A*01:01-B*08:01-C*07:01-DRB1*03:01 haplotype shows a significant positive correlation with the incidence and mortality of COVID-19, while HLA-A*02:01-B*18:01-C*07:01-DRB1*11:04 has been shown to confer protection against serious illness. Littera and colleagues, from the study on the Sardinian population, showed that the three-loci haplotype HLA-A*30:02, B*14:02, C*08:02 is more frequent in patients with COVID-19 (Littera et al., 2021). Finally, Zhang and colleagues recently demonstrated overexpression of the HLA-B*18:01:01:01 and B*44:03:01:01 gene in human lung epithelial cells infected with the virus. The affinity of SARS-CoV-2 peptides for HLA molecules varies between polymorphic HLA alleles, potentially influencing antigen presentation and the strength of the immune response (Blackwell et al., 2009; Matzaraki et al., 2017). Several *in silico* studies showed that the various HLA alleles bind to peptides derived from SARS-CoV-2 nucleocapsid with different affinity. Successful presentation of peptides depends on their effective binding to HLA molecules via hydrogen bonds and salt bridge interactions, allowing for high affinity with a broader specificity. The autoimmune response to self-molecules is also mediated by the different binding affinity of peptides derived from them and individual HLA antigens. There are numerous autoimmune diseases that are associated with HLA class II alleles (Liu et al., 2021; Naito & Okada, 2022). The combination of HLA molecules therefore defines the repertoire of both non self and self-peptides that can activate an adaptive immune response. Since the polymorphism of HLA genes is extreme, each individual expresses numerous different HLA molecules, each of which defines the repertoire of peptides that are presented with the greatest affinity. Ultimately, it is the combination of these that defines susceptibility to certain autoimmune diseases or the response to viral infections. Only when the HLA molecule plays a major role can an association be clearly identified between it and a certain disease, autoimmune or infectious. This is one of the reasons that explain why the numerous studies that have examined the HLA association with COVID-19 have often led to inconclusive, or difficult to replicate, results (Deb et al., 2022). In addition, it should be evaluated different population ancestries and clinical outcomes, with HLA frequencies that can vary significantly. Furthermore, the cases examined were not always of sufficient number, or HLA genotyping was not performed for all loci, or the disease phenotype was not always well defined.

This project aimed to investigate the impact of HLA polymorphism on COVID-19 severity in a cohort of 1,978 SARS-Cov-2 infected subjects with different disease severity belonging to the Italian GEN-COVID Multicenter study, a network of more than 40 Italian Hospitals (https://sites.google.com/dbm.unisi.it/gen-covid). This cohort has several advantages. First, it collects both severe and not severe cases allowing the use of infected asymptomatic subjects, as control, instead of the general population. This advantage leads to cleaner association. Second, it collects a relatively homogenous population: white ethnicity based in Italy. Third, the HLA haplotypes are well characterised in Italy including the ancestral ones. Besides, this study taken advantage of approach with haplotype versus single alleles association: it was reasoned that the HLA locus should be seen as a single functional entity and the global HLA response in terms of balance between response toward spike making, strong host defence, and response toward self, making weak host defence should be relevant for COVID-19 severity. Additionally, this study was aimed to evaluate how differential peptide specificities of HLA allotypes can affect immune response by alteration of antibody production. It was considered that higher antibodies affinity against Spike protein can decrease infection and disease severity, while increased antibodies recognizing an autoantigen, such as anti-interferon-α (anti-IFN-α), can impact negatively on infection clearance and contribute to severity (Bastard et al., 2020; Bastard et al., 2022) (**Figure 4.2**). Evaluation of the HLA allotypes and spike/IFN-α peptides affinity binding was performed by *in silico* prediction.

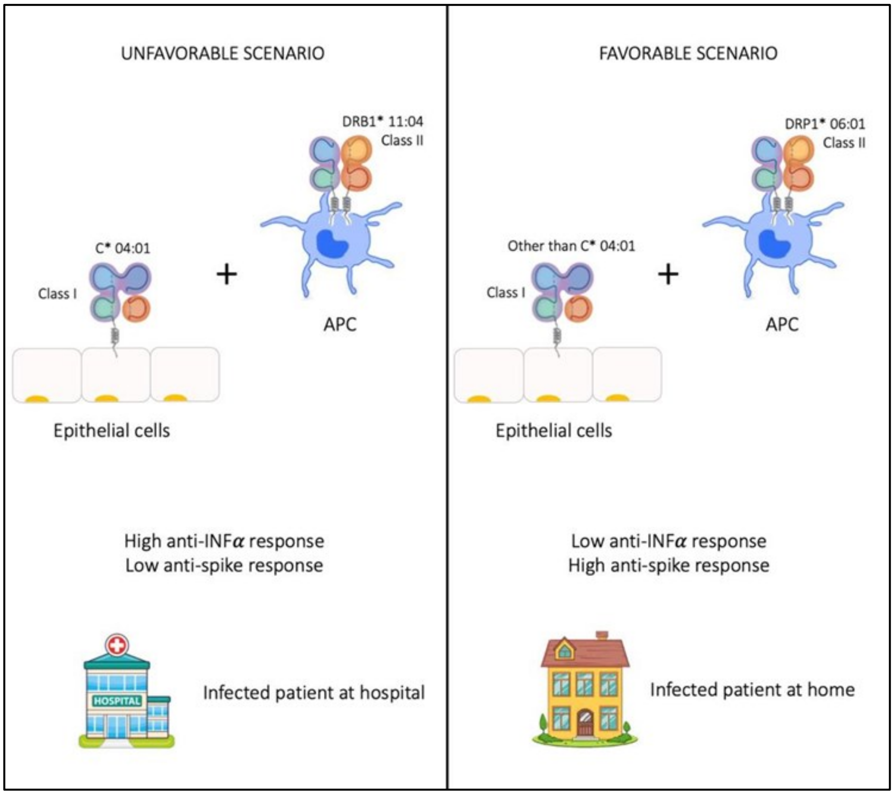

Figure 4.2. Graphical abstract of the study describing the rational of the project.

Methods can be grouped as follows:

1. Experimental steps performed by the GEN-COVID Multicenter of Siena: typing of both *HLA* class I (-*A*, -*B*, -*C*, -*E*, -*F*, -*G*) and class II genes (-*DRB1*, -*DQA1*, -*DQB1*, -*DPA1*, -*DPB1*) from the cohort; plasma detection of IFN-α specific antibodies from hospitalized patients.

2. Statistical association study was performed in three steps:
   - Preliminary association analysis of the cohort divided according to hospitalization status: 403 patients with very mild symptoms, that could be treat from home (non-hospitalized), and 1575 patients who need any hospital assistance (hospitalized). Clinical and demographic feature (severity, sex, age) were included in the frequency's evaluation by logistic regression.
   - Association study of HLA polymorphism with COVID-19 hospitalization, using Bonferroni correction on p-value after logistic regression. Regarding HLA class II, which are expressed as heterodimers with specific alleles encoding the respective α and β subunits, it was also analysed the two-locus haplotypes.
   - Association study with Italian population. Given the high variance of HLA across human populations, driven in part by their impact on infectious disease susceptibility, the 20 most frequent haplotypes in the Italian population were tested in the cohort.
3. Binding affinity prediction by sequence-based tools. I performed prediction analysis between HLA molecules and SARS-CoV-2 spike protein. HLA allotypes were grouped according to the number of different peptides they can present: strong, weak or none. In both of *in silico* predictions, it was used netMHCpan v4.1, setting a <0.5% rank for strong binders and <2% rank for weak binders, and netMHCIIpan v4.0, setting a <0.5% rank for strong binders and <5% rank for weak binders. The spike protein of SARS-CoV-2 alpha (B.1.1.7) virus strain was used (UniProtKB P0DTC2); for HLA class I, binding affinities were estimated for peptides fragments of 8-11 amino acids in length, and for HLA class II the peptides were 15 amino acids in length. The parallel investigation of HLA allotypes having high affinity to IFN-α involved similar predictive protocol. IFN-α subunit protein sequences were retrieved from the GenBank NCBI reference database: EAW58609, EAW58611, EAW58621, EAW58615, EAW58613, EAW58620, EAW58610, EAW58619, AAI04160, EAW58618, EAW58617, EAW58623. A final evaluation of the HLA allele frequencies between patients with and without IFN-α specific antibodies completed the work. HLA-DRB1 was analysed alone due to lack of variability of HLA-DRA allotype.

Outline of the binding affinity prediction analysis was:

- estimate the affinity for all SARS-CoV-2 spike protein related peptides in the NetMHCpan v4.1 dataset to any HLA allele, such as shown in **Table 5.2** for MHC type I of patient with identifier 'COV1'. The same procedure was performed for MHC type II by using netMHCIIpan v4.0 dataset.
- calculate the total number of binding peptides to its HLA allele in each patient. The rational for counting the total number of peptides was the probability to present a peptide is higher when more alleles can bind to that peptide.

EPITOPE Patient *'COV1'*	**HLA-A** Allele1 *A*01:01*	**HLA-A** Allele2 *A*31:01*	**HLA-B** Allele1 *B*51:01*	**HLA-B** Allele2 *B*52:01*	**HLA-C** Allele1 *C*05:01*	**HLA-C** Allele2 *C*12:02*
LADAGFIKQY	0,08	18,38	12,93	19,66	1,40	0,70
QTGKIADYNY	0,35	16,63	51,57	48,33	44,00	22,89
ILDITPCSF	0,70	19,08	6,42	2,85	0,07	1,72
SQSIIAYTM	6,49	8,06	6,49	0,34	3,24	0,82
CYFPLQSY	7,93	11,22	18,10	18,75	29,40	4,64
VQPTESIVRF	7,95	14,63	1,07	1,28	4,43	3,48
QTNSPRRAR	8,68	0,04	44,38	41,77	24,26	7,32
...	...	...	...	...	...	...

Table 4.2. Schematic description of the binding affinity score in patients with identifier 'COV1' related to HLA-A, -B and -C allotypes for each allele. In grey are strong binders, in gold are weak binders.

From HLA typing analysis in the cohort, 94 HLA alleles resulted having frequency greater than 2%. After screening association, logistic regression association analysis with HLA alleles candidates indicated HLA class I alleles showing no significant differences after p-value correction, while HLA-DPB1*13:01 was enriched in non-hospitalized patients, when sex and age >60 years old were covariates. Non-classical HLA class I antigens (HLA-E, G, F) were excluded from next analysis because they were founded no related to the severity of the disease. Next, the two-locus haplotypes that carry HLA-DPB1*13:01 with

HLA-DPA1 alleles shown HLA-DPA1*02:01 and HLA-DPA1*01:03 were more frequent in non-hospitalized than hospitalized patients. Also, these two heterodimers predicted to bind a variety of spike peptides, being classified as a strong binder. Interestingly, also DPA1*02:02-HLA-DPB1*13:01, the other allotype heterodimers observed in GEN-COVID-19 cohort, have strong predicted spike peptide binding, highlighting the potential role of DPB1*13:01 in peptide presentation and preventing disease severity.

About differential binding profile, HLA allotypes were divided into weak/strong presenters on a class-by-class basis, using as threshold the median of the number of high-affinity peptides in each class: 30 for HLA-A, -B, -C; 9 for HLA*DRB1. The threshold for IFN-α was set at five peptides in total from the 12 protein subtypes evaluated, for both HLA class I and II. Some HLA allotypes were exclusively strong binders for IFN-α or Spike peptides (**Table 4.3**).

Allotypes	**Strong** Spike	**Weak** Spike	**No** Spike	**Strong** IFN-α	**Weak** IFN-α	**No** IFN-α
*A*01:01*	X			X		
*A*02:01*		X		X#		
*A*03:01*	X			X		
*A*24:02*	X			X		
*A*29:02*	X			X		
*A*30:01*		X		X#		
*A*30:02*	X			X		
*A*33:01*		X		X#		
*B*07:02*		X		X#		
*B*08:01*		X		X#		
*B*13:02*	X$					X
*B*14:02*		X		X#		
*B*15:01*	X			X		
*B*15:17*	X			X		
*B*18:01*		X		X#		
*B*35:01*	X			X		
*B*35:02*	X$					X
*B*35:03*		X		X#		
*B*44:03*		X		X#		
*B*57:01*		X		X#		
*C*03:03*	X			X		

*C*04:01*		X		X$^{\#}$		
*C*05:01*		X		X$^{\#}$		
*C*06:02*	X			X		
*C*07:01*	X			X		
*C*07:02*	X			X		
*C*08:02*		X		X$^{\#}$		
*C*12:03*	X			X		
*C*16:01*	X$^{\$}$					X
*31*01:01*		X		X$^{\#}$		
*31*01:02*			X		X	
*31*03:01*	X$^{\$}$				X	
*31*07:01*	X			X		
*DRB1*11:01*			X		X	
*DRB1*11:03*		X		X$^{\#}$		
*DRB1*11:04*			X	X$^{\#}$		
*RB1*12:01*		X			X	
*RB1*13:02*	X$^{\$}$				X	
*RB1*14:01*		X			X	
*RB1*15:01*	X$^{\$}$				X	

Table 4.3. HLA-A, -B, -C, and -DRB1 allotypes in Italian top-20 haplotypes and SARS-CoV-2 Spike and IFN-α peptides affinity binding prediction analysis. [X$^{\$}$ HLA allotypes stronger spike binders and weak of absent IFN-α peptides binders; X$^{\#}$ HLA allotypes strong IFN-α peptides binders and weak of absent spike peptides binders].

Results from association studies with the pre-pandemic cohort by Rendine et al. indicates frequencies of the 20 most common HLA haplotypes in the COVID-19 cohort diverge from the Italian population (Rendine et al., 2012). Specifically, 6 haplotypes (ranked as 1, 7, 10, 14, 15 and 16) enriched at higher frequencies ($p_c < 0,05$) than in COVID-19 patients. **Table 5.4** compare the binding predictions for both spike and IFN-α proteins considering the HLA allotypes present in those haplotypes.

Rank	Italian# F (%)	All patients F (%)	Peptides	HLA-A allotype	HLA-B allotype	HLA-C allotype	HLA-DRB1 allotype
Haplotypes enriched in pre-pandemic population							
1	91 (4.7%)	90 (2.3%)	Spike	*A*01:01*	*B*08:01*	*C*07:01*	*DRB1*03:01*
			IFN-α	*A*01:01*	*B*08:01*	*C*07:01*	*DRB1*03:01*
7	23 (1.2%)	5 (0.1%)	Spike	*A*24:02*	*B*15:01*	*C*03:03*	*DRB1*11:03*
			IFN-α	*A*24:02*	*B*15:01*	*C*03:03*	*DRB1*11:03*
10	17 (0.9%)	6 (0.2%)	Spike	*A*01:01*	*B*57:01*	*C*06:02*	*DRB1*01:01*
			IFN-α	*A*01:01*	*B*57:01*	*C*06:02*	*DRB1*01:01*
14	14 (0.7%)	4 (0.1%)	Spike	*A*01:01*	*B*15:17*	*C*07:01*	*DRB1*13:02*
			IFN-α	*A*01:01*	*B*15:17*	*C*07:01*	*DRB1*13:02*
15	15 (0.7%)	7 (0.2%)	Spike	*A*02:01*	*B*35:01*	*C*04:01*	*DRB1*01:01*
			IFN-α	*A*02:01*	*B*35:01*	*C*04:01*	*DRB1*01:01*
16	16 (0.7%)	5 (0.1%)	Spike	*A*03:01*	*B*07:02*	*C*07:02*	*DRB1*15:01*
			IFN-α	*A*03:01*	*B*07:02*	*C*07:02*	*DRB1*15:01*
Haplotypes enriched in Covid-19 cohort							
5	23 (1.2%)	62 (1.6%)	Spike	*A*02:01*	*B*18:01*	*C*07:01*	*DRB1*11:04*
			IFN-α	*A*02:01*	*B*18:01*	*C*07:01*	*DRB1*11:04*
17	17 (0.7%)	34 (0.9%)	Spike	*A*24:02*	*B*35:02*	*C*04:01*	*DRB1*11:04*
			IFN-α	*A*24:02*	*B*35:02*	*C*04:01*	*DRB1*11:04*

Table 4.4. HLA allotypes and spike and IFN-α peptides presenting prediction from haplotypes enriched in COVID-19 cohort or pre-pandemic population. In grey strong peptides binders, weak binders are underlined, and in yellow are the allotypes that do not bind any peptides in high affinity. [*Italian population HLA haplotypes frequencies were obtained from (Rendine et al., 2012)*]

In conclusion, after screening HLA alleles frequencies, outcome of this study suggests HLA-DPB1*13:01 was driving independent associations with protection of infected individuals from severe COVID-19, such as described in other virus infection (Ou et al., 2021). Results from binding affinity prediction of HLA-DBP1*13:01 paired to HLA-DPA1 support this thesis, due can strongly bind viral spike protein peptides which can be recognized by CD4$^+$ T cells (Anczurowski & Hirano, 2018). However, the pair DPB1*13:01-DPA1*02:01 is known to not affect CD4$^+$ T cell response levels on SARS-CoV-2 infection (Hyun et al., 2021). This lack of activation on CD4$^+$ T cells by DPB1*13:01 could be related to low mRNA DPB1 expression levels in virus infection (Ou et al., 2021), and low protein expressed by CD14$^+$ monocytes infected by SARS-CoV-2. Perhaps, DPB1*13:01 allotype could have distinct role in SARS-CoV-2 protein peptides presenting, not via CD4$^+$T cells, modulating immune response to less severe disease outcome. Besides, data from association of GEN-COVID cohort and Italian top-20 HLA haplotypes showed production of autoantibodies to IFN-α is correlated with some HLA class II allele (DRB1*11:04) that it seems to be stratified in good or bad presenters. This suggests that in COVID-19 patients with predisposing HLA*DR alleles that bind with higher affinity to IFN-α derived peptides a strong activation and expansion of CD4$^+$ T cells occurred, and that this specific subset could efficiently modulate help B cell to produce high levels if anti IFN-α autoantibodies.

4.3 Phenotype/Genotype investigation in Alkaptonuria

To extend PM to an ultra-rare condition such as Alkaptonuria (AKU, OMIM: 203500), it is critical to gather as much information about each patient as possible, without disregarding seemingly insignificant aspects, to obtain a first patient stratification. AKU is an autosomal recessive aminoacidopathy of the phenylalanine/tyrosine metabolism. It was first described as a heritable entity by Sir Archibald Garrod in 1902; a disease from which he later formulated the concept of inborn errors of metabolism (Garrod, 1908). It was estimated a prevalence of 1 case per 250.000–1.000.000 births (Phornphutkul et al., 2002) in the majority of ethnic groups and 1233 cases worldwide (Zatkova et al., 2020). AKU is caused by mutations in the Homogentisate 1,2-dioxygenase (*HGD*) gene, which leads to an enzyme deficiency, resulting in a deposition of Homogentisic acid (HGA) especially in connective tissues (La Du et al., 1958). The active form of the HGD enzyme is a highly complex hexamer (Titus et al., 2000) with a poor tolerance for mutations, including missense variations (about 65% of all known AKU substitutions), which might impair protein folding stability and thereby affect HGA accumulation (Nemethova et al., 2016). In 2000, Rodiguez et al. performed a correlation analysis between the most prevalent AKU-causing missense variations (G161R, M368V, and A122V) with the manifestation of varying quantities of unmetabolized HGA, resulting in varying serum and urine levels and, subsequently, varying disease severity. The three mutants displayed considerably lower activity than the wild-type enzyme, ranging from 1% for G161R to 37% and 34% for A122V and M368V, respectively (Rodríguez et al., 2000). It is possible that variations in the residual catalytic activity of the HGD protein due to various polymorphisms might be reflected in this alteration. In this perspective, a genotype–phenotype correlation study was done on 33 individuals, and a minor but statistically significant difference in urine HGA excretion was seen between patients carrying variations with 1% and >30% residual HGD activity (Ascher et al., 2019). The HGA in excess is mostly removed by urine, while the remaining part leads to the formation of an ochronotic pigment deposited in cartilage. Ochronosis plays a critical role in the early stages of arthropathy, lowering patients' quality of life and generating acute pain and locomotor deficits (Milch, 1961). Additionally, recent studies from Millucci et al. have shown that HGA overexpression causes oxidative stress and chronic inflammation in AKU (Millucci, Ghezzi, Bernardini, et al., 2014; Millucci, Ghezzi, Paccagnini, et al., 2014; Millucci et al., 2012). Moreover, they have classified AKU as a secondary amyloidosis, characterised by the condensation of serum amyloid A (SAA) fibres, a circulating protein produced at elevated levels (100–1000 times the normal plasmatic level of about 4-6 mg/L) during chronic inflammation, making SAA a sensitive biomarker of inflammation (Gabay & Kushner, 1999). The presence of ochronotic pigment and amyloid fibres in many AKU samples (i.e. cartilage, salivary glands, chondrocytes, and synoviocytes) confirms the increased plasma levels seen in AKU patients (Braconi et al., 2016; Braconi et al., 2018). Another indicator of persistent inflammation is chitotriosidase (CHIT1), a chitinase that is predominantly expressed in differentiated and polarised macrophages (Cho et al., 2014). The serum concentration of CHIT1 correlates with the progression and severity of numerous conditions (e.g., sarcoidosis, rheumatoid arthritis), suggesting that CHIT1 may act as an AKU biomarker []. Thus, in addition to inflammation, individuals with AKU experience considerable oxidative stress as a result of elevated systemic levels of HGA and its metabolites. Braconi et al. (Braconi et al., 2016) examined proteome changes in AKU samples from six individuals, which revealed intriguing parallels to other rheumatic disorders. In this regard, the Protein Thiolation Index (PTI) defines and summarises the oxidative condition of AKU patients in an intriguing manner (Cicaloni et al., 2019). One of the primary challenges in doing clinical research on AKU is the absence of a consistent technique for assessing illness severity and response to therapy, which is confounded by the wide range of AKU symptoms seen in different individuals (Vilboux et al., 2009). A reliable way to monitor patients' clinical condition and overall health status is the use in clinical practice and research of measures of Quality of Life (QoL). In the study of Braconi et al. was shown that, in a rare and multisystemic disease like AKU, QoL scores help to identify health needs and to evaluate the impact of the disease (Braconi et al., 2018). So, the presence of a correlation between QoL and the clinical data deposited in an AKU-dedicated digital platform, such as ApreciseKUre database (www.bio.unisi.it/aprecisekure/), could be helpful in shading light on AKU complexity and in discovery of new biomarkers.

In this project, it was performed computational methods (such as ML) to achieve a first AKU patient stratification based on phenotypic and genotypic data in a typical precision medicine perspective, retrieved from ApreciseKUre. The workflow can be summarised in four steps, in which I performed the last one (**Figure 4.3**):

1. Data pre-processing and preliminary statistical analysis. It performed a preliminary analysis based on Pearson Correlation Coefficient to evaluate the relationship between pairs of clinical data, biochemical parameters and QoL scores.
2. Unsupervised Clustering. We applied both K-means and Hierarchical Clustering to stratify the AKU population into subgroups with similar features. The experiment was conducted using three different stratifications, i.e., setting (i) K=2, (ii) K=3 and (iii) K=4 to obtain two, three and four clusters respectively. The resulting clusters are grouped according to the severity of the AKU disease, by considering age, the levels of oxidative stress, inflammation, and amyloidosis biomarkers and QoL scores.
3. Statistical Clusters Evaluation. To evaluate if the clusters were significantly identifying sub-groups of individuals, we applied the Kruskall–Wallis (KW) ranking non-parametric test. Additionally, we computed the Silhouette Score with the aim to test the consistency within elements which have been assigned to the same cluster.
4. Genotype–phenotype correlation. Once AKU stratification and cluster validation were performed, it was investigated the HGD mutation distribution across the

obtained clusters, paying attention to G161R, M368V and A122V (representing about 44% of AKU patients' mutation in ApreciseKUre). Specifically, G161R mutation, responsible for a dramatic reduction of HGD activity, occurred in higher percentages in the most phenotypically severe clusters. On the contrary, for M368V and A122V mutations, in which enzymatic activity of HGD is conserved for more than 30%, the trend shows a higher percentage in less severe phenotypic sub-groups.

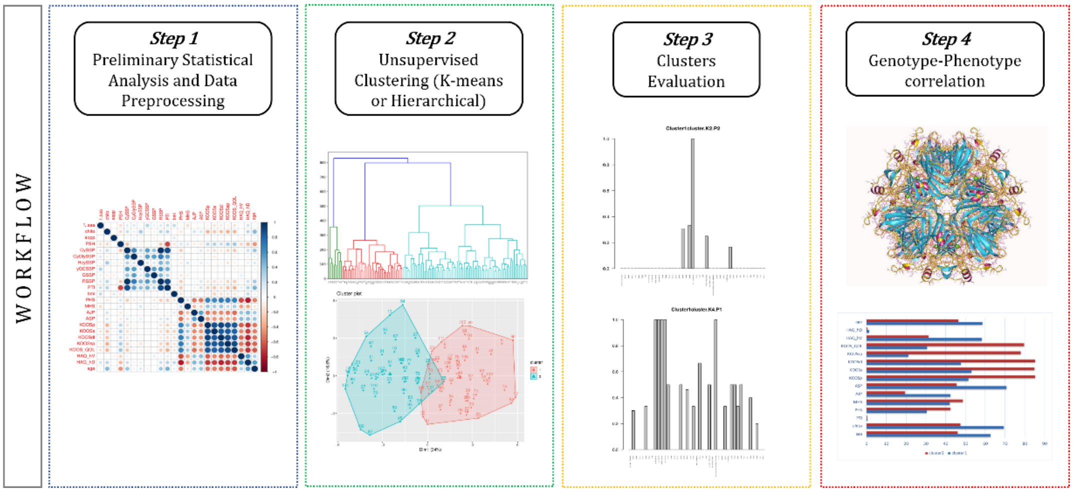
WORKFLOW
Step 1
Preliminary Statistical Analysis and Data Preprocessing
Step 2
Unsupervised Clustering (K-means or Hierarchical)
Step 3
Clusters Evaluation
Step 4
Genotype-Phenotype correlation

Figure 4.3 Workflow applied to achieve AKU patient stratification based on phenotypic and genotypic data. *Adapted from (Spiga et al., 2020).*

Interestingly, in the first step, important biomarkers of chronic inflammation and amyloidosis like CHIT1 and SAA do not result strongly correlated with disease severity differently from PTI, which instead is correlated with KOOS scores and age. For both the clustering methods, we found that for K=2, the most severe phenotype seems to be the cluster number 1, cluster number 2 for K=3 and cluster number 4 for K=4. Statistical cluster evaluation by KW ranking non-parametric test and Silhouette Score corroborated the application of the K-means algorithm for K=2 and K=3, and hierarchical clustering for K=4 for AKU patient stratification. Starting from this point, it was possible to detect the most/less severe subgroups based on demographics, QoL scores and biochemical markers. Specifically, for K = 2, cluster 1 turns out to group AKU patients with most severe symptoms and QoL scores, older age and higher levels of biomarkers of oxidative stress, chronic inflammation and amyloidosis. For K = 3, clusters 2 and 3 comprehend older patients (especially in cluster 2). In cluster 3 there are higher level of SAA and PTI, whereas higher values for CHIT1 are in cluster 2. Patients with less severe symptoms are present in cluster 1, on the contrary patients with the worst QoL score are all included in cluster 2, which turns out to be the most severe one. For K = 4, older patients with more severe symptoms and higher levels of CHIT1, SAA and PTI are stratified in cluster 4, whereas in cluster 2 are grouped younger individuals with less severe AKU manifestations. To sum up the genotype results, this study shows that the mutations G161R, A122V, and M368V are always present in all the stratifications and in greater quantities than the others mutation, being the most frequent mutations in Europe. However, the G161R mutation, despite being present in all the clusters, is mostly represent in the phenotypically more serious subgroups according to a low enzymatic specific activity of 1% w-t. The A122V and P230S mutations, also present in all the clusters, are more represent in the phenotypically less severe subgroup. In these cases, the A122V mutation results agree with the experiments conducted in vitro for the measurement of the specific enzymatic activity (33.5% w-t), whilst the P230S specific enzymatic activity (4% w-t) is lower than expected (Rodríguez et al., 2000). Analogously, mutations D153G and F227S are always present in the phenotypically most serious subgroups. This agrees with the specific activity of mutation F227S (0.1 % w-t), whilst mutation D153G shows higher activity than expected (32.7% w-t). Mutations R225H and W97C are present only in the less severe phenotype subgroups. We only retrieved specific activity (0.1% w-t) for the R225H mutation. Finally, stop mutations are only found in phenotypically serious subgroups because they significantly destroy enzyme activity. Besides, the combination of a ML to analyse and re-interpret data shows the potential direct benefits for patient care and treatments, highlighting the necessity of patient databases for rare diseases, like ApreciseKUre. This approach can be turned into a best practice model also for other rare diseases and can be useful for overcoming the obstacles in small dataset management and analysis. Phenotype and genotype distribution of this results are reported in the following sections.

4.3.1 Phenotype

K=2 - In cluster 1 are stratified older patients showing high severity of AKU disease and with higher level of SAA, Chitotriosidase and PTI. The values of aopp, PSH, CySSP, CyGlySSP, HcySSP, yGCSSP, GSSP, RSSP are similar between the 2 clusters, so that they have little influence on stratification. Moreover, in the first cluster all the KOOS scores are low, which indicates that patients in cluster 1 have greater knee problems than those in cluster 2. Similarly, HAQ_hV, HAQ_hD have higher scores in cluster 1, indicating worse arthritic conditions. PHS (physical health status) and MHS (mental health status), have lower values in cluster 1, indicating worse physical and mental conditions. Analogously, cluster 1 shows higher scores of AKUSSI, indicating greater severity of the disease.

K=3 - In the three clusters stratification, cluster 2 and 3 contain older patients with higher level of SAA, Chitotriosidase and PTI. Patients with less severe symptoms are present in cluster 1. Significantly higher values of SAA can be observed in cluster 3 with respect to the other two clusters. Similarly, to the two-cluster stratification, also in this case the values of aopp, PSH, CySSP, CyGlySSP, HcySSP, yGCSSP, GSSP, RSSP are similar for all of the clusters, and therefore have little influence on stratification. In cluster 1 the values of KOOS are higher when compared to the other two clusters. Therefore, patients in cluster 2 and 3 present greater problems at the level of knees. The same trend can be observed in the HAQ_hV, HAQ_hD scores, which are higher in clusters 2 and 3 with respect to cluster 1 and thus the correspondent patients are in worse conditions. PHS (physical health status) and MHS (mental health status), have lower values in the cluster 2 and 3, indicating worse physical and mental conditions. As far as the AKUSSI values are concerned, cluster 2 shows higher scores, and, consequently, patients present greater severity of the disease.

K=4 - In the stratification with four clusters, the analysis of the phenotype shows that older patients are stratified in cluster 4. In cluster 3 higher level of SAA can be observed. As observed in the previous stratifications, also in this case the values of aopp, PSH, CySSP, CyGlySSP, HcySSP, yGCSSP, GSSP, RSSP are similar for all the clusters, and therefore have little influence on the stratification. In cluster 1 and 3 are present much higher values of KOOS, compared to the other 2 clusters. Therefore, cluster 1 and 4 present patients with greater problems at the knee. The same trend is found in the HAQ_hV, HAQ_hD scores, which are higher in clusters 1 and 4, therefore the correspondent patients are in worse conditions. PHS (physical health status) and MHS (mental health status), have lower values in clusters 1 and 4, indicating worse physical and mental conditions of the patients. Cluster 1 and 4 show higher AKUSSI scores, consequently present patients with greater severity of the disease.

Parameters	K=2		K=3		K=4	
	Statistic	*FDR*	*Statistic*	*FDR*	*Statistic*	*FDR*
Age	35,39	6,48E-09	36,08	2,58E-08	145,31	7,09E-08
AJP	31,32	4,78E-08	29,49	2,74E-06	19,32	5,94E-06
ASP	19,45	1,66E-05	21,31	0,00027	32,85	0,0010
CHIT1	20,129	1,24E-05	22,26	2,35E-05	32,43	7,29E-07
hapVAS	29,75	2,49E-11	26,84	7,98E-06	13,33	0,00033
HAQ-DI	46,46	9,81E-08	36,70	1,34E-08	60,18	2,09E-09
KOOS_QOL	71,43	1,72E-16	63,36	8,36E-14	55,61	1,72E-13
KOOSdl	58,23	9,32E-14	51,01	2,88E-11	528,80	1,61E-12
KOOSp	61,02	2,72E-14	51,19	2,88E-11	16,77	1,44E-12
KOOSs	54,12	6,47E-13	49,20	6,21E-11	40,15	1,36E-11
KOOSsp	74,21	5,62E-17	64,76	5,20E-14	45,13	1,10E-13
MHS	87,93	0,0040	15,36	0,00061	65,49	0,0050
PHS	48,60	9,44E-12	38,22	3,19E-08	37,53	8,47E-07
PTI	23,68	2,10E-06	24,55	3,53E-05	60,69	2,16E-08
SAA	981,52	0,0024	17,08	7,88E-07	67,05	6,38E-07

Table 4.5 Statistically significant values adjusted with multiple test (FDR < 0.05), related to the Kruskall–Wallis (KW) ranking of the biomarkers and QoL scores. *Adapted from (Spiga et al., 2020).*

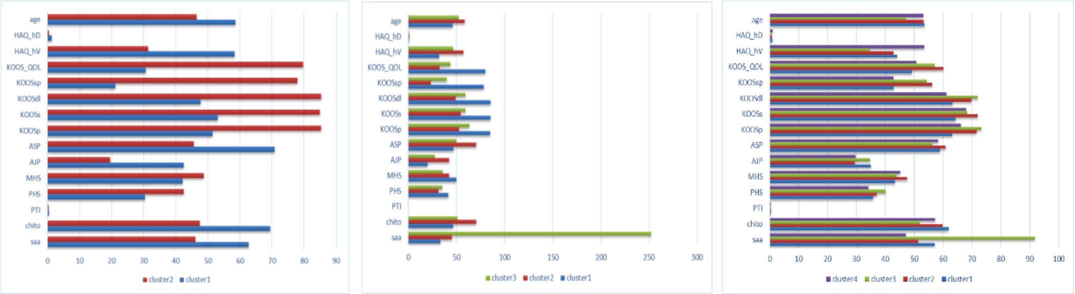

Figure 4.4 Phenotype results of K-means clustering. *Adapted from (Spiga et al., 2020).*

4.3.2 Genotype

K=2 - Based on the previous phenotypic observations, the G161R, A122V, and M368V mutations are present at high extent in both clusters, i.e., both in patients with severe and in less severe AKU disease (**Table 4.6**). Furthermore, also the P230S, I216T and V300G mutations are present in both clusters even if in minor abundance. It is worth noticing that, in this first stratification represented by two clusters, mutations D153G and F227S are present only in the cluster of the most serious patients. While the R225H and W97C mutations are present only in the cluster of less severe patients. A further result revealed from this first stratification, is the presence of stop mutations, specifically R321* and W60*, only in the cluster of the most serious patients.

K=3 - Based on the previous phenotypic analysis, the stratification with K=3 groups in cluster 1 patients with less severe disease, in cluster 2 the most serious patients and in cluster 3 the patients with intermediate biomarkers values. Also, in this case the mutations G161R, A122V, and M368V, are present in all the clusters and in high amount (**Table 4.6**). P230S, I216T and V300G mutations are also found in all the three clusters, but in smaller quantities. Similarly to the previous stratification, also in this case the mutations D153G and F227S are present only in the cluster of the most serious patients, while the R225H and W97C mutations are present only in the cluster with the less severe patients. Moreover, the R321* and W60* stop mutations are present only in the cluster with elevated disease biomarkers.

K=4 - Based on the previous phenotypic analysis the stratification with K=4 (4 clusters) presents the most severe AKU patients in clusters 1 and 4. Cluster 2 patients have the lowest biomarker values, while cluster 3 presents patients with intermediate values. The same trend of the previous stratifications is found. Which shows the G161R, A122V, and M368V mutation in all the clusters and in much greater quantity (**Table 4.6**). Also, in this stratification, mutations D153G and F227S are present only in the cluster of the most serious patients. While the R225H and W97C mutations are present only in cluster 2, which groups less severe patients. Also, in this case the stop mutations W60* and R321* are in cluster 4 and cluster 1, respectively.

	Allele 1			**Allele 2**		
K-means = 2	G161R	M368V	A122V	G161R	M368V	A122V
Cluster 1	60	40	20	60	70	30
Cluster 2	40	60	80	40	30	70
K-means = 3	G161R	M368V	A122V	G161R	M368V	A122V
Cluster 1	30	60	80	30	40	70
Cluster 2	50	30	20	40	60	30
Cluster 3	10	10	0	30	0	0
Hierarchical = 4	G161R	M368V	A122V	G161R	M368V	A122V
Cluster 1	25	30	55	10	25	60
Cluster 2	25	60	45	20	25	30
Cluster 3	15	10	0	35	0	0
Cluster 4	35	0	0	35	50	10

Table 4.6. Three most abundant missense mutations (G161R, M368 and A122V) and their normalized mutation percentages (order of magnitude) in each cluster. *Adapted from (Spiga et al., 2020).*

Conclusions

The research activity of my PhD was mainly focused on understanding of the molecular basis of heredited and drug-induced channelopathies by MD simulations in the Kv11.1 channel. The first project employed Umbrella Sampling simulations to study the inhibitory effects on hERG-current by three known PPIs. Results indicated that the binding energies were in the 8-10 kcal/mol range for all the compounds, despite their specific pose in the pore, confirming the hypothesis that their inhibitory effects on hERG-current might be due to a direct, steric interference with the channel function. From PMFs profiles it emerged that Omeprazole preferentially bound at the intracellular side of the cavity, where a swallow free-energy minimum exists, while Pantoprazole and Lansoprazole were characterized by well-defined energy minima profile in proximity of the selectivity filter of the channel. The simulation results were in agreement with electrophysiological findings. Moreover, the evidence that each compound shows a specific binding pose, provides a molecular basis possibly accounting for the different potency of channel inhibition observed experimentally. The second phase of the study concerning hERG was aimed to identify possible metastable states related to hERG C-type inactivation, and to estimate how the ensemble of metastable states might impact on drug-binding. Results suggested that fast C-type inactivation requires geometrical reorientation of residues delimiting binding sites S0 and S1, and not a closure of the selectivity filter, as previously suggested. Docking calculations confirmed that significant differences exist in the orientation of the drugs among the various metastable states identified both for inactivating and non-inactivating mutants of the hERG channel.

Molecular Dynamics simulations are certainly affected by severe limitations, as the limited time scales accessible by simulated trajectories, and the inaccuracies of the adopted Force Fields. However, results on the Kv11.1 channels presented in this thesis reveal how these atomistic simulations can assist for the analysis of biological processes. Indeed, MD kinetic modelling of ion channel activity provides a formal, quantitative mechanism for testing hypotheses about channel function. These models can then be used to investigate the relationship between molecular defects and whole-organ phenotypes. Recent advances in computational hardware have enabled organ-level simulations involving more complex biophysically accurate cellular models on practical timescales. However, to properly comprehend and eventually overcome the complexity of heart's electrical system and its related safety-pharmacology challenge, it is anticipated that more reliable structural investigations would be required. MD simulations are a powerful strategy to investigate biological processes at the atomic scale, and to estimate parameters that could be useful in models on higher temporal and spatial scales to investigate biological processes in a classical System Biology approach.

References

Abi-Gerges, N., Holkham, H., Jones, E. M., Pollard, C. E., Valentin, J. P., & Robertson, G. A. (2011). hERG subunit composition determines differential drug sensitivity. *Br J Pharmacol*, *164*(2b), 419-432. https://doi.org/10.1111/j.1476-5381.2011.01378.x

Adasme, M. F., Linnemann, K. L., Bolz, S. N., Kaiser, F., Salentin, S., Haupt, V. J., & Schroeder, M. (2021). PLIP 2021: expanding the scope of the protein-ligand interaction profiler to DNA and RNA. *Nucleic Acids Res*, *49*(W1), W530-W534. https://doi.org/10.1093/nar/gkab294

Albaugh, A., Boateng, H. A., Bradshaw, R. T., Demerdash, O. N., Dziedzic, J., Mao, Y., . . . Head-Gordon, T. (2016). Advanced Potential Energy Surfaces for Molecular Simulation. *The Journal of Physical Chemistry B*, *120*(37), 9811-9832. https://doi.org/10.1021/acs.jpcb.6b06414

Alexandrou, A. J., Duncan, R. S., Sullivan, A., Hancox, J. C., Leishman, D. J., Witchel, H. J., & Leaney, J. L. (2006). Mechanism of hERG K+ channel blockade by the fluoroquinolone antibiotic moxifloxacin. *Br J Pharmacol*, *147*(8), 905-916. https://doi.org/10.1038/sj.bjp.0706678

Amoroso, A., Magistroni, P., Vespasiano, F., Bella, A., Bellino, S., Puoti, F., . . . Centers, I. N. o. R. T. C. (2021). HLA and AB0 Polymorphisms May Influence SARS-CoV-2 Infection and COVID-19 Severity. *Transplantation*, *105*(1), 193-200. https://doi.org/10.1097/TP.0000000000003507

Anczurowski, M., & Hirano, N. (2018). Mechanisms of HLA-DP Antigen Processing and Presentation Revisited. *Trends Immunol*, *39*(12), 960-964. https://doi.org/10.1016/j.it.2018.10.008

Asai, T., Adachi, N., Moriya, T., Oki, H., Maru, T., Kawasaki, M., . . . Murata, T. (2021). Cryo-EM Structure of K. *Structure*, *29*(3), 203-212.e204. https://doi.org/10.1016/j.str.2020.12.007

Ascher, D. B., Spiga, O., Sekelska, M., Pires, D. E. V., Bernini, A., Tiezzi, M., . . . Zatkova, A. (2019). Homogentisate 1,2-dioxygenase (HGD) gene variants, their analysis and genotype–phenotype correlations in the largest cohort of patients with AKU. *European Journal of Human Genetics*. https://doi.org/10.1038/s41431-019-0354-0

Augusto, D. G., & Hollenbach, J. A. (2022). HLA variation and antigen presentation in COVID-19 and SARS-CoV-2 infection. *Curr Opin Immunol*, *76*, 102178. https://doi.org/10.1016/j.coi.2022.102178

Bahcall, O. (2015). Precision medicine. *Nature*, *526*(7573), 335. https://doi.org/10.1038/526335a

Bastard, P., Rosen, L. B., Zhang, Q., Michailidis, E., Hoffmann, H. H., Zhang, Y., . . . Effort, C. H. G. (2020). Autoantibodies against type I IFNs in patients with life-threatening COVID-19. *Science*, *370*(6515). https://doi.org/10.1126/science.abd4585

Bastard, P., Zhang, Q., Zhang, S. Y., Jouanguy, E., & Casanova, J. L. (2022). Type I interferons and SARS-CoV-2: from cells to organisms. *Curr Opin Immunol*, *74*, 172-182. https://doi.org/10.1016/j.coi.2022.01.003

Beauchamp, K. A., Lin, Y. S., Das, R., & Pande, V. S. (2012). Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *J Chem Theory Comput*, *8*(4), 1409-1414. https://doi.org/10.1021/ct2007814

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, *81*(8), 3684-3690. https://doi.org/10.1063/1.448118

Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., Weissig, H., & Westbrook, J. (2000). The Protein Data Bank and the challenge of structural genomics. *Nat Struct Biol*, *7 Suppl*, 957-959. https://doi.org/10.1038/80734

Bewley, C. A., Gronenborn, A. M., & Clore, G. M. (1998). Minor groove-binding architectural proteins: structure, function, and DNA recognition. *Annu Rev Biophys Biomol Struct*, *27*, 105-131. https://doi.org/10.1146/annurev.biophys.27.1.105

Blackwell, J. M., Jamieson, S. E., & Burgner, D. (2009). HLA and infectious diseases. *Clin Microbiol Rev*, *22*(2), 370-385, Table of Contents. https://doi.org/10.1128/CMR.00048-08

Braconi, D., Bernardini, G., Paffetti, A., Millucci, L., Geminiani, M., Laschi, M., . . . Santucci, A. (2016). Comparative proteomics in alkaptonuria provides insights into inflammation and oxidative stress. *International Journal of Biochemistry and Cell Biology*. https://doi.org/10.1016/j.biocel.2016.08.016

Braconi, D., Giustarini, D., Marzocchi, B., Peruzzi, L., Margollicci, M., Rossi, R., . . . Santucci, A. (2018). Inflammatory and oxidative stress biomarkers in alkaptonuria:

data from the DevelopAKUre project. *Osteoarthritis and Cartilage*. https://doi.org/10.1016/j.joca.2018.05.017

Bradshaw, R. T., & Essex, J. W. (2016). Evaluating Parametrization Protocols for Hydration Free Energy Calculations with the AMOEBA Polarizable Force Field. *J Chem Theory Comput*, *12*(8), 3871-3883. https://doi.org/10.1021/acs.jctc.6b00276

Braun, E., Gilmer, J., Mayes, H. B., Mobley, D. L., Monroe, J. I., Prasad, S., & Zuckerman, D. M. (2019). Best Practices for Foundations in Molecular Simulations [Article v1.0]. *Living J Comput Mol Sci*, *1*(1). https://doi.org/10.33011/livecoms.1.1.5957

Brodin, P. (2021). Immune determinants of COVID-19 disease presentation and severity. *Nat Med*, *27*(1), 28-33. https://doi.org/10.1038/s41591-020-01202-8

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. In: Journal of Computational Chemistry.

Burg, J. S., Ingram, J. R., Venkatakrishnan, A. J., Jude, K. M., Dukkipati, A., Feinberg, E. N., . . . Garcia, K. C. (2015). Structural biology. Structural basis for chemokine recognition and activation of a viral G protein-coupled receptor. *Science*, *347*(6226), 1113-1117. https://doi.org/10.1126/science.aaa5026

Burley, S. K., & Petsko, G. A. (1985). Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science*, *229*(4708), 23-28. https://doi.org/10.1126/science.3892686

Butler, A., Helliwell, M. V., Zhang, Y., Hancox, J. C., & Dempsey, C. E. (2019). An Update on the Structure of hERG. *Front Pharmacol*, *10*, 1572. https://doi.org/10.3389/fphar.2019.01572

Caldwell, J. W., & Kollman, P. A. (1995). Cation-.pi. Interactions: Nonadditive Effects Are Critical in Their Accurate Representation. *Journal of the American Chemical Society*, *117*(14), 4177-4178. https://doi.org/10.1021/ja00119a037

Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., . . . Woods, R. J. (2005). The Amber biomolecular simulation programs. *J Comput Chem*, *26*(16), 1668-1688. https://doi.org/10.1002/jcc.20290

Cavalli, A., Buonfiglio, R., Ianni, C., Masetti, M., Ceccarini, L., Caves, R., . . . Recanatini, M. (2012). Computational design and discovery of "minimally structured" hERG blockers. *J Med Chem*, *55*(8), 4010-4014. https://doi.org/10.1021/jm201194q

Chahrour, M., Jung, S. Y., Shaw, C., Zhou, X., Wong, S. T., Qin, J., & Zoghbi, H. Y. (2008). MeCP2, a key contributor to neurological disease, activates and represses transcription. *Science*, *320*(5880), 1224-1229. https://doi.org/10.1126/science.1153252

Chen, J., Seebohm, G., & Sanguinetti, M. C. (2002). Position of aromatic residues in the S6 domain, not inactivation, dictates cisapride sensitivity of HERG and eag potassium channels. *Proc Natl Acad Sci U S A*, *99*(19), 12461-12466. https://doi.org/10.1073/pnas.192367299

Cheng, A. C., Coleman, R. G., Smyth, K. T., Cao, Q., Soulard, P., Caffrey, D. R., . . . Huang, E. S. (2007). Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol*, *25*(1), 71-75. https://doi.org/10.1038/nbt1273

Cho, S. J., Weiden, M. D., & Lee, C. G. (2014). Chitotriosidase in the pathogenesis of inflammation, interstitial lung diseases and COPD. *Allergy, Asthma and Immunology Research*. https://doi.org/10.4168/aair.2015.7.1.14

Choo, S. Y. (2007). The HLA system: genetics, immunology, clinical testing, and clinical implications. *Yonsei Med J*, *48*(1), 11-23. https://doi.org/10.3349/ymj.2007.48.1.11

Chrysant, S. G. (2019). Proton pump inhibitor-induced hypomagnesemia complicated with serious cardiac arrhythmias. *Expert Rev Cardiovasc Ther*, *17*(5), 345-351. https://doi.org/10.1080/14779072.2019.1615446

Cicaloni, V., Spiga, O., Dimitri, G. M., Maiocchi, R., Millucci, L., Giustarini, D., . . . Santucci, A. (2019). Interactive alkaptonuria database: investigating clinical data to improve patient care in a rare disease. *FASEB Journal*. https://doi.org/10.1096/fj.201901529R

Cordero-Morales, J. F., Jogini, V., Lewis, A., Vásquez, V., Cortes, D. M., Roux, B., & Perozo, E. (2007). Molecular driving forces determining potassium channel slow inactivation. *Nat Struct Mol Biol*, *14*(11), 1062-1069. https://doi.org/10.1038/nsmb1309

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., . . . Kollman, P. A. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, *117*(19), 5179-5197. https://doi.org/10.1021/ja00124a002

Correale, P., Mutti, L., Pentimalli, F., Baglio, G., Saladino, R. E., Sileri, P., & Giordano, A. (2020). HLA-B*44 and C*01 Prevalence Correlates with Covid19 Spreading across Italy. *Int J Mol Sci*, *21*(15). https://doi.org/10.3390/ijms21155205

Coutsias, E. A., & Wester, M. J. (2019). RMSD and Symmetry. *J Comput Chem*, *40*(15), 1496-1508. https://doi.org/10.1002/jcc.25802

Cuello, L. G., Cortes, D. M., Jogini, V., Sompornpisut, A., & Perozo, E. (2010). A molecular mechanism for proton-dependent gating in KcsA. *FEBS Lett*, *584*(6), 1126-1132. https://doi.org/10.1016/j.febslet.2010.02.003

Curry, H. B. (1944). The method of steepest descent for non-linear minimization problems. *Quarterly of Applied Mathematics*, *2*, 258-261.

Dame, R. T. (2005). The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin. *Mol Microbiol*, *56*(4), 858-870. https://doi.org/10.1111/j.1365-2958.2005.04598.x

Damm, K. L., & Carlson, H. A. (2006). Gaussian-weighted RMSD superposition of proteins: a structural comparison for flexible proteins and predicted protein structures. *Biophys J*, *90*(12), 4558-4573. https://doi.org/10.1529/biophysj.105.066654

Danielsson, B., Collin, J., Nyman, A., Bergendal, A., Borg, N., State, M., . . . Fastbom, J. (2020). Drug use and torsades de pointes cardiac arrhythmias in Sweden: a nationwide register-based cohort study. *BMJ Open*, *10*(3), e034560. https://doi.org/10.1136/bmjopen-2019-034560

de Souza, O. N., & Ornstein, R. L. (1997). Effect of periodic box size on aqueous molecular dynamics simulation of a DNA dodecamer with particle-mesh Ewald method. *Biophys J*, *72*(6), 2395-2397. https://doi.org/10.1016/s0006-3495(97)78884-2

Deb, P., Zannat, K. E., Talukder, S., Bhuiyan, A. H., Jilani, M. S. A., & Saif-Ur-Rahman, K. M. (2022). Association of HLA gene polymorphism with susceptibility, severity, and mortality of COVID-19: A systematic review. *HLA*, *99*(4), 281-312. https://doi.org/10.1111/tan.14560

Deiters, U. K. (2013). Efficient Coding of the Minimum Image Convention. *Zeitschrift für Physikalische Chemie*, *227*(2-3), 345-352. https://doi.org/doi:10.1524/zpch.2013.0311

Dendrou, C. A., Petersen, J., Rossjohn, J., & Fugger, L. (2018). HLA variation and disease. *Nat Rev Immunol*, *18*(5), 325-339. https://doi.org/10.1038/nri.2017.143

Domene, C., & Furini, S. (2009). Dynamics, energetics, and selectivity of the low-K+ KcsA channel structure. *J Mol Biol*, *389*(3), 637-645. https://doi.org/10.1016/j.jmb.2009.04.038

Duronio, R. J., & Xiong, Y. (2013). Signaling pathways that control cell proliferation. *Cold Spring Harb Perspect Biol*, *5*(3), a008904. https://doi.org/10.1101/cshperspect.a008904

Eberhardt, J., Santos-Martins, D., Tillack, A. F., & Forli, S. (2021). AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J Chem Inf Model*, *61*(8), 3891-3898. https://doi.org/10.1021/acs.jcim.1c00203

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., & Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, *103*(19), 8577-8593. https://doi.org/10.1063/1.470117

Famularo, G., Gasbarrone, L., & Minisola, G. (2013). Hypomagnesemia and proton-pump inhibitors. *Expert Opin Drug Saf*, *12*(5), 709-716. https://doi.org/10.1517/14740338.2013.809062

Feller, S. E., Zhang, Y., Pastor, R. W., & Brooks, B. R. (1995). Constant pressure molecular dynamics simulation: The Langevin piston method. *The Journal of Chemical Physics*, *103*(11), 4613-4621. https://doi.org/10.1063/1.470648

Fernandez, D., Ghanta, A., Kauffman, G. W., & Sanguinetti, M. C. (2004). Physicochemical features of the HERG channel drug binding site. *J Biol Chem*, *279*(11), 10120-10127. https://doi.org/10.1074/jbc.M310683200

Ficker, E., Jarolimek, W., & Brown, A. M. (2001). Molecular determinants of inactivation and dofetilide block in ether a-go-go (EAG) channels and EAG-related K(+) channels. *Mol Pharmacol*, *60*(6), 1343-1348. https://doi.org/10.1124/mol.60.6.1343

Fields, J. B., Németh-Cahalan, K. L., Freites, J. A., Vorontsova, I., Hall, J. E., & Tobias, D. J. (2017). Calmodulin Gates Aquaporin 0 Permeability through a Positively Charged Cytoplasmic Loop. *J Biol Chem*, *292*(1), 185-195. https://doi.org/10.1074/jbc.M116.743724

Foloppe, N., & Hubbard, R. (2006). Towards predictive ligand design with free-energy based computational methods? *Curr Med Chem*, *13*(29), 3583-3608. https://doi.org/10.2174/092986706779026165

Francis, J. M., Leistritz-Edwards, D., Dunn, A., Tarr, C., Lehman, J., Dempsey, C., . . . Team, M. C.-C. a. P. (2022). Allelic variation in class I HLA determines CD8. *Sci Immunol*, *7*(67), eabk3070. https://doi.org/10.1126/sciimmunol.abk3070

Frenkel, D., Smit, B., & Ratner, M. A. (1997). Understanding Molecular Simulation: From Algorithms to Applications. *Physics Today*, *50*(7), 66-66. https://doi.org/10.1063/1.881812

Frietze, S., & Farnham, P. J. (2011). Transcription factor effector domains. *Subcell Biochem*, *52*, 261-277. https://doi.org/10.1007/978-90-481-9069-0_12

Furini, S., & Domene, C. (2011). Selectivity and permeation of alkali metal ions in K+-channels. *J Mol Biol*, *409*(5), 867-878. https://doi.org/10.1016/j.jmb.2011.04.043

Furini, S., & Domene, C. (2020). Critical Assessment of Common Force Fields for Molecular Dynamics Simulations of Potassium Channels. *J Chem Theory Comput*, *16*(11), 7148-7159. https://doi.org/10.1021/acs.jctc.0c00331

Gabay, C., & Kushner, I. (1999). Acute-phase proteins and other systemic responses to inflammation. In.

Gang, H., & Zhang, S. (2006). Na+ permeation and block of hERG potassium channels. *J Gen Physiol*, *128*(1), 55-71. https://doi.org/10.1085/jgp.200609500

Ganji, M., Docter, M., Le Grice, S. F., & Abbondanzieri, E. A. (2016). DNA binding proteins explore multiple local configurations during docking via rapid rebinding. *Nucleic Acids Res*, *44*(17), 8376-8384. https://doi.org/10.1093/nar/gkw666

Gardini, S., Furini, S., Santucci, A., & Niccolai, N. (2017). A structural bioinformatics investigation on protein-DNA complexes delineates their modes of interaction. *Mol Biosyst*, *13*(5), 1010-1017. https://doi.org/10.1039/c7mb00071e

Garrod, A. E. (1908). The Croonian Lectures ON INBORN ERRORS OF METABOLISM. *The Lancet*. https://doi.org/10.1016/S0140-6736(01)78482-6

Genheden, S., Mikulskis, P., Hu, L., Kongsted, J., Söderhjelm, P., & Ryde, U. (2011). Accurate predictions of nonpolar solvation free energies require explicit consideration of binding-site hydration. *J Am Chem Soc*, *133*(33), 13081-13092. https://doi.org/10.1021/ja202972m

Genheden, S., & Ryde, U. (2012). Will molecular dynamics simulations of proteins ever reach equilibrium? *Phys Chem Chem Phys*, *14*(24), 8662-8677. https://doi.org/10.1039/c2cp23961b

Gohlke, H., & Klebe, G. (2002). Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew Chem Int Ed Engl*, *41*(15), 2644-2676. https://doi.org/10.1002/1521-3773(20020802)41:15<2644::AID-ANIE2644>3.0.CO;2-O

Groban, E. S., Narayanan, A., & Jacobson, M. P. (2006). Conformational changes in protein loops and helices induced by post-translational phosphorylation. *PLoS Comput Biol*, *2*(4), e32. https://doi.org/10.1371/journal.pcbi.0020032

Grove, C. A., & Walhout, A. J. (2008). Transcription factor functionality and transcription regulatory networks. *Mol Biosyst*, *4*(4), 309-314. https://doi.org/10.1039/b715909a

Guerini, F. R., Bolognesi, E., Lax, A., Bianchi, L. N. C., Caronni, A., Zanzottera, M., . . . Clerici, M. (2022). HLA Allele Frequencies and Association with Severity of COVID-19 Infection in Northern Italian Patients. *Cells*, *11*(11). https://doi.org/10.3390/cells11111792

Guimarães, C. R., & Mathiowetz, A. M. (2010). Addressing limitations with the MM-GB/SA scoring procedure using the WaterMap method and free energy perturbation calculations. *J Chem Inf Model*, *50*(4), 547-559. https://doi.org/10.1021/ci900497d

Guo, J., Gang, H., & Zhang, S. (2006). Molecular determinants of cocaine block of human ether-á-go-go-related gene potassium channels. *J Pharmacol Exp Ther*, *317*(2), 865-874. https://doi.org/10.1124/jpet.105.098103

Gómez-Varela, D., Contreras-Jurado, C., Furini, S., García-Ferreiro, R., Stühmer, W., & Pardo, L. A. (2006). Different relevance of inactivation and F468 residue in the mechanisms of hEag1 channel blockage by astemizole, imipramine and dofetilide. *FEBS Lett*, *580*(21), 5059-5066. https://doi.org/10.1016/j.febslet.2006.08.030

Hagler, A. T., Huler, E., & Lifson, S. (1974). Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *Journal of the American Chemical Society*, *96*(17), 5319-5327. https://doi.org/10.1021/ja00824a004

Halgren, T. A., & Damm, W. (2001). Polarizable force fields. *Current Opinion in Structural Biology*, *11*(2), 236-242. https://doi.org/https://doi.org/10.1016/S0959-440X(00)00196-2

Hall, R., Dixon, T., & Dickson, A. (2020). On Calculating Free Energy Differences Using Ensembles of Transition Paths. *Front Mol Biosci*, *7*, 106. https://doi.org/10.3389/fmolb.2020.00106

Hao, G. F., Xu, W. F., Yang, S. G., & Yang, G. F. (2015). Multiple Simulated Annealing-Molecular Dynamics (MSA-MD) for Conformational Space Search of Peptide and Miniprotein. *Sci Rep*, *5*, 15568. https://doi.org/10.1038/srep15568

Harder, E., Anisimov, V. M., Vorobyov, I. V., Lopes, P. E., Noskov, S. Y., MacKerell, A. D., & Roux, B. (2006). Atomic Level Anisotropy in the Electrostatic Modeling of Lone Pairs for a Polarizable Force Field Based on the Classical Drude Oscillator. *J Chem Theory Comput*, *2*(6), 1587-1597. https://doi.org/10.1021/ct600180x

Harteis, S., & Schneider, S. (2014). Making the bend: DNA tertiary structure and protein-DNA interactions. *Int J Mol Sci*, *15*(7), 12335-12363. https://doi.org/10.3390/ijms150712335

Henriques, J., Cragnell, C., & Skepö, M. (2015). Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment. *J Chem Theory Comput*, *11*(7), 3420-3431. https://doi.org/10.1021/ct501178z

Herzberg, I. M., Trudeau, M. C., & Robertson, G. A. (1998). Transfer of rapid inactivation and sensitivity to the class III antiarrhythmic drug E-4031 from HERG to M-eag channels. *J Physiol*, *511 (Pt 1)*(Pt 1), 3-14. https://doi.org/10.1111/j.1469-7793.1998.003bi.x

Hestenes, M. R., & Stiefel, E. (1952). Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, *49*, 409-435.

HODGKIN, A. L., & HUXLEY, A. F. (1952). Propagation of electrical signals along giant nerve fibers. *Proc R Soc Lond B Biol Sci*, *140*(899), 177-183. https://doi.org/10.1098/rspb.1952.0054

Hogan, M. E., & Austin, R. H. (1987). Importance of DNA stiffness in protein-DNA binding specificity. *Nature*, *329*(6136), 263-266. https://doi.org/10.1038/329263a0

Hollingsworth, S. A., & Dror, R. O. (2018). Molecular Dynamics Simulation for All. *Neuron*, *99*(6), 1129-1143. https://doi.org/10.1016/j.neuron.2018.08.011

Hood, L., Heath, J. R., Phelps, M. E., & Lin, B. (2004). Systems biology and new technologies enable predictive and preventative medicine. *Science*, *306*(5696), 640-643. https://doi.org/10.1126/science.1104635

Hoover, W. G. (1985). Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A*, *31*(3), 1695-1697.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., & Simmerling, C. (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, *65*(3), 712-725. https://doi.org/10.1002/prot.21123

Hou, T., Wang, J., Li, Y., & Wang, W. (2011). Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J Chem Inf Model*, *51*(1), 69-82. https://doi.org/10.1021/ci100275a

Hu, Z., Yang, J., Xiong, G., Shi, H., Yuan, Y., Fan, L., & Wang, Y. (2014). HLA-DPB1 Variant Effect on Hepatitis B Virus Clearance and Liver Cirrhosis Development Among Southwest Chinese Population. *Hepat Mon*, *14*(8), e19747. https://doi.org/10.5812/hepatmon.19747

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., . . . MacKerell, A. D. (2017). CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods*, *14*(1), 71-73. https://doi.org/10.1038/nmeth.4067

Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., . . . Zitnik, M. (2021). Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Developments. In: Proceedings of Neural Information Processing Systems, NeurIPS Datasets and Benchmarks.

Hudson, W. H., & Ortlund, E. A. (2014). The structure, function and evolution of proteins that bind DNA and RNA. *Nat Rev Mol Cell Biol*, *15*(11), 749-760. https://doi.org/10.1038/nrm3884

Husic, B. E., & Pande, V. S. (2018). Markov State Models: From an Art to a Science. *J Am Chem Soc*, *140*(7), 2386-2396. https://doi.org/10.1021/jacs.7b12191

Hyun, Y. S., Lee, Y. H., Jo, H. A., Baek, I. C., Kim, S. M., Sohn, H. J., & Kim, T. G. (2021). Comprehensive Analysis of CD4. *Front Immunol*, *12*, 774491. https://doi.org/10.3389/fimmu.2021.774491

Iyengar, R., Altman, R. B., Troyanskya, O., & FitzGerald, G. A. (2015). MEDICINE. Personalization in practice. *Science*, *350*(6258), 282-283. https://doi.org/10.1126/science.aad5204

Javaid, N., & Choi, S. (2017). Acetylation- and Methylation-Related Epigenetic Proteins in the Context of Their Targets. *Genes (Basel)*, *8*(8). https://doi.org/10.3390/genes8080196

Jiang, F., Zhou, C. Y., & Wu, Y. D. (2014). Residue-specific force field based on the protein coil library. RSFF1: modification of OPLS-AA/L. *J Phys Chem B*, *118*(25), 6983-6998. https://doi.org/10.1021/jp5017449

Jiang, W., Hardy, D. J., Phillips, J. C., Mackerell, A. D., Schulten, K., & Roux, B. (2011). High-performance scalable molecular dynamics simulations of a polarizable force field based on classical Drude oscillators in NAMD. *J Phys Chem Lett*, *2*(2), 87-92. https://doi.org/10.1021/jz101461d

Jo, S., Kim, T., Iyer, V. G., & Im, W. (2008). CHARMM-GUI: a web-based graphical user interface for CHARMM. *J Comput Chem*, *29*(11), 1859-1865. https://doi.org/10.1002/jcc.20945

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Jorgensen, W. L., & Tirado-Rives, J. (1988). The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc*, *110*(6), 1657-1666. https://doi.org/10.1021/ja00214a001

Joung, I. S., & Cheatham, T. E. (2009). Molecular dynamics simulations of the dynamic and energetic properties of alkali and halide ions using water-model-specific ion parameters. *J Phys Chem B*, *113*(40), 13279-13290. https://doi.org/10.1021/jp902584c

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, *32*(5), 922-923. https://doi.org/doi:10.1107/S0567739476001873

Kalodimos, C. G., Biris, N., Bonvin, A. M., Levandoski, M. M., Guennuegues, M., Boelens, R., & Kaptein, R. (2004). Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science*, *305*(5682), 386-389. https://doi.org/10.1126/science.1097064

Kang, J., Wang, L., Chen, X. L., Triggle, D. J., & Rampe, D. (2001). Interactions of a series of fluoroquinolone antibacterial drugs with the human cardiac K+ channel HERG. *Mol Pharmacol*, *59*(1), 122-126. https://doi.org/10.1124/mol.59.1.122

Karplus, M., & McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nat Struct Biol*, *9*(9), 646-652. https://doi.org/10.1038/nsb0902-646

Kiehn, J., Lacerda, A. E., & Brown, A. M. (1999). Pathways of HERG inactivation. *Am J Physiol*, *277*(1), H199-210. https://doi.org/10.1152/ajpheart.1999.277.1.H199

King, E., Aitchison, E., Li, H., & Luo, R. (2021). Recent Developments in Free Energy Calculations for Drug Discovery. *Front Mol Biosci*, *8*, 712085. https://doi.org/10.3389/fmolb.2021.712085

Kitao, A. (2022). Principal Component Analysis and Related Methods for Investigating the Dynamics of Biological Macromolecules. *J*, *5*(2), 298-317.

Klein, J., & Sato, A. (2000a). The HLA system. First of two parts. *N Engl J Med*, *343*(10), 702-709. https://doi.org/10.1056/NEJM200009073431006

Klein, J., & Sato, A. (2000b). The HLA system. Second of two parts. *N Engl J Med*, *343*(11), 782-786. https://doi.org/10.1056/NEJM200009143431106

Koehl, P. (2018). Large Eigenvalue Problems in Coarse-Grained Dynamic Analyses of Supramolecular Systems. *J Chem Theory Comput*, *14*(7), 3903-3919. https://doi.org/10.1021/acs.jctc.8b00338

Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., . . . Cheatham, T. E. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res*, *33*(12), 889-897. https://doi.org/10.1021/ar000033j

Kopke Salinas, R., Folkers, G. E., Bonvin, A. M., Das, D., Boelens, R., & Kaptein, R. (2005). Altered specificity in DNA binding by the lac repressor: a mutant lac headpiece that mimics the gal repressor. *Chembiochem*, *6*(9), 1628-1637. https://doi.org/10.1002/cbic.200500049

Krig, S. (2016). Feature Learning and Deep Learning Architecture Survey. In S. Krig (Ed.), *Computer Vision Metrics: Textbook Edition* (pp. 375-514). Springer International Publishing. https://doi.org/10.1007/978-3-319-33762-3_10

Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., & Kollman, P. A. (1992). THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, *13*.

Kumari, R., Kumar, R., Lynn, A., & Consortium, O. S. D. D. (2014). g_mmpbsa--a GROMACS tool for high-throughput MM-PBSA calculations. *J Chem Inf Model*, *54*(7), 1951-1962. https://doi.org/10.1021/ci500020m

La Du, B. N., Zannoni, V. G., Laster, L., & Seegmiller, J. E. (1958). The nature of the defect in tyrosine metabolism in alcaptonuria. *The Journal of biological chemistry*, *230*(1), 251-260.

Lamoureux, G. a. M. A. D. a. R. B. t. (2003). A simple polarizable model of water based on classical Drude oscillators. *The Journal of Chemical Physics*, *119*(10), 5185-5197. https://doi.org/10.1063/1.1598191

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., . . . Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860-921. https://doi.org/10.1038/35057062

Lazzerini, P. E., Bertolozzi, I., Finizola, F., Acampa, M., Natale, M., Vanni, F., . . . Capecchi, P. L. (2018). Proton Pump Inhibitors and Serum Magnesium Levels in Patients With Torsades de Pointes. *Front Pharmacol*, *9*, 363. https://doi.org/10.3389/fphar.2018.00363

Lazzerini, P. E., Bertolozzi, I., Rossi, M., Capecchi, P. L., & Laghi-Pasini, F. (2017). Combination Therapy With Ceftriaxone and Lansoprazole, Acquired Long QT Syndrome, and Torsades de Pointes Risk. *J Am Coll Cardiol*, *69*(14), 1876-1877. https://doi.org/10.1016/j.jacc.2016.11.090

Lazzerini, P. E., Cartocci, A., Qu, Y. S., Saponara, S., Furini, S., Fusi, F., . . . Boutjdir, M. (2021). Proton Pump Inhibitors Directly Block hERG-Potassium Channel and Independently Increase the Risk of QTc Prolongation in a Large Cohort of US Veterans. *Circ Arrhythm Electrophysiol*, *14*(7), e010042. https://doi.org/10.1161/CIRCEP.121.010042

Lee, A., Lee, K., & Kim, D. (2016). Using reverse docking for target identification and its applications for drug discovery. *Expert Opin Drug Discov*, *11*(7), 707-715. https://doi.org/10.1080/17460441.2016.1190706

Lee, J., Cheng, X., Swails, J. M., Yeom, M. S., Eastman, P. K., Lemkul, J. A., . . . Im, W. (2016). CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J Chem Theory Comput*, *12*(1), 405-413. https://doi.org/10.1021/acs.jctc.5b00935

Lees-Miller, J. P., Duan, Y., Teng, G. Q., Thorstad, K., & Duff, H. J. (2000). Novel gain-of-function mechanism in K(+) channel-related long-QT syndrome: altered gating and selectivity in the HERG1 N629D mutant. *Circ Res*, *86*(5), 507-513. https://doi.org/10.1161/01.res.86.5.507

Lemkul, J. A., Huang, J., Roux, B., & MacKerell, A. D., Jr. (2016). An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chemical Reviews*, *116*(9), 4983-5013. https://doi.org/10.1021/acs.chemrev.5b00505

Li, G. R., Feng, J., Yue, L., Carrier, M., & Nattel, S. (1996). Evidence for two components of delayed rectifier K+ current in human ventricular myocytes. *Circ Res*, *78*(4), 689-696. https://doi.org/10.1161/01.res.78.4.689

Li, J., Ostmeyer, J., Cuello, L. G., Perozo, E., & Roux, B. (2018). Rapid constriction of the selectivity filter underlies C-type inactivation in the KcsA potassium channel. *J Gen Physiol*, *150*(10), 1408-1420. https://doi.org/10.1085/jgp.201812082

Lindorff-Larsen, K., Best, R. B., Depristo, M. A., Dobson, C. M., & Vendruscolo, M. (2005). Simultaneous determination of protein structure and dynamics. *Nature*, *433*(7022), 128-132. https://doi.org/10.1038/nature03199

Littera, R., Chessa, L., Deidda, S., Angioni, G., Campagna, M., Lai, S., . . . Perra, A. (2021). Natural killer-cell immunoglobulin-like receptors trigger differences in immune response to SARS-CoV-2 infection. *PLoS One*, *16*(8), e0255608. https://doi.org/10.1371/journal.pone.0255608

Liu, B., Shao, Y., & Fu, R. (2021). Current research status of HLA in immune-related diseases. *Immun Inflamm Dis*, *9*(2), 340-350. https://doi.org/10.1002/iid3.416

Liu, Y., Ke, M., & Gong, H. (2015). Protonation of Glu(135) Facilitates the Outward-to-Inward Structural Transition of Fucose Transporter. *Biophys J*, *109*(3), 542-551. https://doi.org/10.1016/j.bpj.2015.06.037

Lorberbaum, T., Sampson, K. J., Chang, J. B., Iyer, V., Woosley, R. L., Kass, R. S., & Tatonetti, N. P. (2016). Coupling Data Mining and Laboratory Experiments to Discover Drug Interactions Causing QT Prolongation. *J Am Coll Cardiol*, *68*(16), 1756-1764. https://doi.org/10.1016/j.jacc.2016.07.761

Lu, Y., Mahaut-Smith, M. P., Varghese, A., Huang, C. L., Kemp, P. R., & Vandenberg, J. I. (2001). Effects of premature stimulation on HERG K(+) channels. *J Physiol*, *537*(Pt 3), 843-851. https://doi.org/10.1111/j.1469-7793.2001.00843.x

Luscombe, N. M., Austin, S. E., Berman, H. M., & Thornton, J. M. (2000). An overview of the structures of protein-DNA complexes. *Genome Biol*, *1*(1), REVIEWS001. https://doi.org/10.1186/gb-2000-1-1-reviews001

Luscombe, N. M., Laskowski, R. A., & Thornton, J. M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res*, *29*(13), 2860-2874. https://doi.org/10.1093/nar/29.13.2860

Luscombe, N. M., & Thornton, J. M. (2002). Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J Mol Biol*, *320*(5), 991-1009. https://doi.org/10.1016/s0022-2836(02)00571-5

Mackay, D. H. J., Cross, A. J., & Hagler, A. T. (1989). The Role of Energy Minimization in Simulation Strategies of Biomolecular Systems. In (pp. 317-358). Springer US.

Mackerell, A. D., Feig, M., & Brooks, C. L. (2004). Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem*, *25*(11), 1400-1415. https://doi.org/10.1002/jcc.20065

MacKerell, A. D., Feig, M., & Brooks, C. L. (2004). Improved treatment of the protein backbone in empirical force fields. *J Am Chem Soc*, *126*(3), 698-699. https://doi.org/10.1021/ja036959e

Manolis, A. A., Manolis, T. A., Melita, H., Katsiki, N., & Manolis, A. S. (2020). Proton pump inhibitors and cardiovascular adverse effects: Real or surreal worries? *Eur J Intern Med*, *72*, 15-26. https://doi.org/10.1016/j.ejim.2019.11.017

Matzaraki, V., Kumar, V., Wijmenga, C., & Zhernakova, A. (2017). The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol*, *18*(1), 76. https://doi.org/10.1186/s13059-017-1207-1

McAulay, K. A., Higgins, C. D., Macsween, K. F., Lake, A., Jarrett, R. F., Robertson, F. L., . . . Crawford, D. H. (2007). HLA class I polymorphisms are associated with development of infectious mononucleosis upon primary EBV infection. *J Clin Invest*, *117*(10), 3042-3048. https://doi.org/10.1172/JCI32377

McKiernan, K. A., Husic, B. E., & Pande, V. S. (2017). Modeling the mechanism of CLN025 beta-hairpin formation. *J Chem Phys*, *147*(10), 104107. https://doi.org/10.1063/1.4993207

Milch, R. A. (1961). Studies of Alcaptonuria: A Genetic Study of 58 Cases Occurring in Eight Generations of Seven Inter-related Dominican Kindreds. *Arthritis & Rheumatism*. https://doi.org/10.1002/art.1780040202

Millucci, L., Ghezzi, L., Bernardini, G., Braconi, D., Lupetti, P., Perfetto, F., . . . Santucci, A. (2014). Diagnosis of secondary amyloidosis in alkaptonuria. *Diagnostic Pathology*. https://doi.org/10.1186/s13000-014-0185-9

Millucci, L., Ghezzi, L., Paccagnini, E., Giorgetti, G., Viti, C., Braconi, D., . . . Santucci, A. (2014). Amyloidosis, inflammation, and oxidative stress in the heart of an alkaptonuric patient. *Mediators of Inflammation*. https://doi.org/10.1155/2014/258471

Millucci, L., Spreafico, A., Tinti, L., Braconi, D., Ghezzi, L., Paccagnini, E., . . . Santucci, A. (2012). Alkaptonuria is a novel human secondary amyloidogenic disease. *Biochimica et Biophysica Acta - Molecular Basis of Disease*. https://doi.org/10.1016/j.bbadis.2012.07.011

Miranda, W. E., DeMarco, K. R., Guo, J., Duff, H. J., Vorobyov, I., Clancy, C. E., & Noskov, S. Y. (2020). Selectivity filter modalities and rapid inactivation of the hERG1 channel. *Proc Natl Acad Sci U S A*, *117*(6), 2795-2804. https://doi.org/10.1073/pnas.1909196117

Mitcheson, J. S., Chen, J., Lin, M., Culberson, C., & Sanguinetti, M. C. (2000). A structural basis for drug-induced long QT syndrome. *Proc Natl Acad Sci U S A*, *97*(22), 12329-12333. https://doi.org/10.1073/pnas.210244497

Mobley, D. L., Bannan, C. C., Rizzi, A., Bayly, C. I., Chodera, J. D., Lim, V. T., . . . Eastman, P. K. (2018). Escaping Atom Types in Force Fields Using Direct Chemical Perception. *Journal of Chemical Theory and Computation*, *14*(11), 6076-6092. https://doi.org/10.1021/acs.jctc.8b00640

Molgedey, L., & Schuster, H. G. (1994). Separation of a mixture of independent signals using time delayed correlations. *Phys Rev Lett*, *72*(23), 3634-3637. https://doi.org/10.1103/PhysRevLett.72.3634

Momany, F. A., McGuire, R. F., Burgess, A. W., & Scheraga, H. A. (1975). Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. *The Journal of Physical Chemistry*, *79*(22), 2361-2381. https://doi.org/10.1021/j100589a006

Munawar, S., Windley, M. J., Tse, E. G., Todd, M. H., Hill, A. P., Vandenberg, J. I., & Jabeen, I. (2018). Experimentally Validated Pharmacoinformatics Approach to Predict hERG Inhibition Potential of New Chemical Entities. *Front Pharmacol*, *9*, 1035. https://doi.org/10.3389/fphar.2018.01035

Mutowo, P., Bento, A. P., Dedman, N., Gaulton, A., Hersey, A., Lomax, J., & Overington, J. P. (2016). A drug target slim: using gene ontology and gene ontology annotations to navigate protein-ligand target space in ChEMBL. *J Biomed Semantics*, *7*(1), 59. https://doi.org/10.1186/s13326-016-0102-0

Naito, T., & Okada, Y. (2022). HLA imputation and its application to genetic and molecular fine-mapping of the MHC region in autoimmune diseases. *Semin Immunopathol*, *44*(1), 15-28. https://doi.org/10.1007/s00281-021-00901-9

Nemethova, M., Radvanszky, J., Kadasi, L., Ascher, D. B., Pires, D. E. V., Blundell, T. L., . . . Zatkova, A. (2016). Twelve novel HGD gene variants identified in 99 alkaptonuria patients: Focus on 'black bone disease' in Italy. *European Journal of Human Genetics*. https://doi.org/10.1038/ejhg.2015.60

Nerenberg, P. S., & Head-Gordon, T. (2018). New developments in force fields for biomolecular simulations. *Curr Opin Struct Biol*, *49*, 129-138. https://doi.org/10.1016/j.sbi.2018.02.002

Nguyen, A., David, J. K., Maden, S. K., Wood, M. A., Weeder, B. R., Nellore, A., & Thompson, R. F. (2020). Human Leukocyte Antigen Susceptibility Map for Severe Acute Respiratory Syndrome Coronavirus 2. *J Virol*, *94*(13). https://doi.org/10.1128/JVI.00510-20

Nguyen, H., Maier, J., Huang, H., Perrone, V., & Simmerling, C. (2014). Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *J Am Chem Soc*, *136*(40), 13959-13962. https://doi.org/10.1021/ja5032776

Niedzialkowska, E., Gasiorowska, O., Handing, K. B., Majorek, K. A., Porebski, P. J., Shabalin, I. G., . . . Minor, W. (2016). Protein purification and crystallization artifacts: The tale usually not told. *Protein Sci*, *25*(3), 720-733. https://doi.org/10.1002/pro.2861

NIH, N. I. o. H.-. (2022). *Precision Medicine Initiative (PMI) Cohort Program*. https://allofus.nih.gov/

Nosé, S., & Klein, M. L. (1986). Constant-temperature-constant-pressure molecular-dynamics calculations for molecular solids: Application to solid nitrogen at high pressure. *Phys Rev B Condens Matter*, *33*(1), 339-342. https://doi.org/10.1103/physrevb.33.339

Novelli, A., Andreani, M., Biancolella, M., Liberatoscioli, L., Passarelli, C., Colona, V. L., . . . Locatelli, F. (2020). HLA allele frequencies and susceptibility to COVID-19 in a group of 99 Italian patients. *HLA*, *96*(5), 610-614. https://doi.org/10.1111/tan.14047

Oltvai, Z. N., & Barabási, A. L. (2002). Systems biology. Life's complexity pyramid. *Science*, *298*(5594), 763-764. https://doi.org/10.1126/science.1078563

Ou, G., Liu, X., Xu, H., Ji, X., & Wang, J. (2021). Variation and expression of HLA-DPB1 gene in HBV infection. *Immunogenetics*, *73*(3), 253-261. https://doi.org/10.1007/s00251-021-01213-w

Parker, R., Partridge, T., Wormald, C., Kawahara, R., Stalls, V., Aggelakopoulou, M., . . . Ternette, N. (2021). Mapping the SARS-CoV-2 spike glycoprotein-derived peptidome presented by HLA class II on dendritic cells. *Cell Rep*, *35*(8), 109179. https://doi.org/10.1016/j.celrep.2021.109179

Parrinello, M., & Rahman, A. (1981). Polymorphic transitions in single crystals: A new molecular dynamics method. *52:12 C1 - Research Org.: Univ. of Trieste (Italy)*.

Patterson Burdsall, D., Flores, H. C., Krueger, J., Garretson, S., Gorbien, M. J., Iacch, A., . . . Homa, T. (2013). Use of proton pump inhibitors with lack of diagnostic indications in 22 Midwestern US skilled nursing facilities. *J Am Med Dir Assoc*, *14*(6), 429-432. https://doi.org/10.1016/j.jamda.2013.01.021

Perrin, M. J., Subbiah, R. N., Vandenberg, J. I., & Hill, A. P. (2008). Human ether-a-go-go related gene (hERG) K+ channels: function and dysfunction. *Prog Biophys Mol Biol*, *98*(2-3), 137-148. https://doi.org/10.1016/j.pbiomolbio.2008.10.006

Perry, M. D., Ng, C. A., Mann, S. A., Sadrieh, A., Imtiaz, M., Hill, A. P., & Vandenberg, J. I. (2015). Getting to the heart of hERG K(+) channel gating. *J Physiol*, *593*(12), 2575-2585. https://doi.org/10.1113/JP270095

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem*, *25*(13), 1605-1612. https://doi.org/10.1002/jcc.20084

Pettini, F., Domene, C., & Furini, S. (2022). Early Steps in C-Type Inactivation of the hERG Potassium Channel. *J Chem Inf Model*. https://doi.org/10.1021/acs.jcim.2c01028

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., . . . Schulten, K. (2005). Scalable molecular dynamics with NAMD. *J Comput Chem*, *26*(16), 1781-1802. https://doi.org/10.1002/jcc.20289

Phornphutkul, C., Introne, W. J., Perry, M. B., Bernardini, I., Murphey, M. D., Fitzpatrick, D. L., . . . Gahl, W. A. (2002). Natural history of alkaptonuria. *New England Journal of Medicine*. https://doi.org/10.1056/NEJMoa021736

Piper, D. R., Varghese, A., Sanguinetti, M. C., & Tristani-Firouzi, M. (2003). Gating currents associated with intramembrane charge displacement in HERG potassium channels. *Proc Natl Acad Sci U S A*, *100*(18), 10534-10539. https://doi.org/10.1073/pnas.1832721100

Pirmohamed, M. (2001). Pharmacogenetics and pharmacogenomics. *Br J Clin Pharmacol*, *52*(4), 345-347. https://doi.org/10.1046/j.0306-5251.2001.01498.x

Pisanti, S., Deelen, J., Gallina, A. M., Caputo, M., Citro, M., Abate, M., . . . Martinelli, R. (2020). Correlation of the two most frequent HLA haplotypes in the Italian population to the differential regional incidence of Covid-19. *J Transl Med*, *18*(1), 352. https://doi.org/10.1186/s12967-020-02515-5

Pitera, J. W. (2014). Expected distributions of root-mean-square positional deviations in proteins. *J Phys Chem B*, *118*(24), 6526-6530. https://doi.org/10.1021/jp412776d

Pérez-Hernández, G., Paul, F., Giorgino, T., De Fabritiis, G., & Noé, F. (2013). Identification of slow molecular order parameters for Markov model construction. *J Chem Phys*, *139*(1), 015102. https://doi.org/10.1063/1.4811489

Raschi, E., Vasina, V., Poluzzi, E., & De Ponti, F. (2008). The hERG K+ channel: target and antitarget strategies in drug development. *Pharmacol Res*, *57*(3), 181-195. https://doi.org/10.1016/j.phrs.2008.01.009

Rastogi, C., Rube, H. T., Kribelbauer, J. F., Crocker, J., Loker, R. E., Martini, G. D., . . . Bussemaker, H. J. (2018). Accurate and sensitive quantification of protein-DNA binding affinity. *Proceedings of the National Academy of Sciences*, *115*(16), E3692-E3701. https://doi.org/10.1073/pnas.1714376115

Rauscher, S., Gapsys, V., Gajda, M. J., Zweckstetter, M., de Groot, B. L., & Grubmüller, H. (2015). Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J Chem Theory Comput*, *11*(11), 5513-5524. https://doi.org/10.1021/acs.jctc.5b00736

Reddi, R., Matulef, K., Riederer, E. A., Whorton, M. R., & Valiyaveetil, F. I. (2021). Structural basis for C-type inactivation in a Shaker family voltage gated K<sup>+</sup> channel. *bioRxiv*, 2021.2010.2001.462615. https://doi.org/10.1101/2021.10.01.462615

Redding, S., & Greene, E. C. (2013). How do proteins locate specific targets in DNA? *Chem Phys Lett*, *570*. https://doi.org/10.1016/j.cplett.2013.03.035

Reeves, R. (2010). Nuclear functions of the HMG proteins. *Biochim Biophys Acta*, *1799*(1-2), 3-14. https://doi.org/10.1016/j.bbagrm.2009.09.001

Rendine, S., Ferrero, N. M., Sacchi, N., Costa, C., Pollichieni, S., & Amoroso, A. (2012). Estimation of human leukocyte antigen class I and class II high-resolution allele and haplotype frequencies in the Italian population and comparison with other European populations. *Hum Immunol*, *73*(4), 399-404. https://doi.org/10.1016/j.humimm.2012.01.005

Reynisson, B., Alvarez, B., Paul, S., Peters, B., & Nielsen, M. (2020). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res*, *48*(W1), W449-W454. https://doi.org/10.1093/nar/gkaa379

Rice, B. M., & Sewell, T. D. (2008). Equilibrium Molecular Dynamics Simulations. In (pp. 255-290). Springer Berlin Heidelberg.

Rockman, H. A., Koch, W. J., & Lefkowitz, R. J. (2002). Seven-transmembrane-spanning receptors and heart function. *Nature*, *415*(6868), 206-212. https://doi.org/10.1038/415206a

Rodríguez, J. M., Timm, D. E., Titus, G. P., Beltrán-Valero De Bernabé, D., Criado, O., Mueller, H. A., . . . Peñalva, M. A. (2000). Structural and functional analysis of mutations in alkaptonuria. *Hum Mol Genet*, *9*(15), 2341-2350. https://doi.org/10.1093/oxfordjournals.hmg.a018927

Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., & Mann, R. S. (2010). Origins of specificity in protein-DNA recognition. *Annu Rev Biochem*, *79*, 233-269. https://doi.org/10.1146/annurev-biochem-060408-091030

Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S., & Honig, B. (2009). The role of DNA shape in protein-DNA recognition. *Nature*, *461*(7268), 1248-1253. https://doi.org/10.1038/nature08473

Roux, B. (1995). The calculation of the potential of mean force using computer simulations. *91*(1), 275-282.

Ruch, P. (2017). Text Mining to Support Gene Ontology Curation and Vice Versa. *Methods Mol Biol*, *1446*, 69-84. https://doi.org/10.1007/978-1-4939-3743-1_6

Sanguinetti, M. C., Jiang, C., Curran, M. E., & Keating, M. T. (1995). A mechanistic link between an inherited and an acquired cardiac arrhythmia: HERG encodes the IKr potassium channel. *Cell*, *81*(2), 299-307. https://doi.org/10.1016/0092-8674(95)90340-2

Sauer, U., Heinemann, M., & Zamboni, N. (2007). Genetics. Getting closer to the whole picture. *Science*, *316*(5824), 550-551. https://doi.org/10.1126/science.1142502

Scapin, G., Potter, C. S., & Carragher, B. (2018). Cryo-EM for Small Molecules Discovery, Design, Understanding, and Application. *Cell Chem Biol*, *25*(11), 1318-1325. https://doi.org/10.1016/j.chembiol.2018.07.006

Schwantes, C. R., & Pande, V. S. (2013). Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J Chem Theory Comput*, *9*(4), 2000-2009. https://doi.org/10.1021/ct300878a

Shealy, R. T., Murphy, A. D., Ramarathnam, R., Jakobsson, E., & Subramaniam, S. (2003). Sequence-function analysis of the K+-selective family of ion channels using a comprehensive alignment and the KcsA channel structure. *Biophys J*, *84*(5), 2929-2942. https://doi.org/10.1016/s0006-3495(03)70020-4

Shi, Y. P., Thouta, S., & Claydon, T. W. (2020). Modulation of hERG K. *Front Pharmacol*, *11*, 139. https://doi.org/10.3389/fphar.2020.00139

Siebenmorgen, T., & Zacharias, M. (2020). Computational prediction of protein–protein binding affinities. *WIREs Computational Molecular Science*, *10*(3), e1448. https://doi.org/https://doi.org/10.1002/wcms.1448

Singh, R., Kaul, R., Kaul, A., & Khan, K. (2007). A comparative review of HLA associations with hepatitis B and C viral infections across global populations. *World J Gastroenterol*, *13*(12), 1770-1787. https://doi.org/10.3748/wjg.v13.i12.1770

Singleton, D. H., Boyd, H., Steidl-Nichols, J. V., Deacon, M., Groot, M. J., Price, D., . . . Boyd, J. G. (2007). Fluorescently labeled analogues of dofetilide as high-affinity fluorescence polarization ligands for the human ether-a-go-go-related gene (hERG) channel. *J Med Chem*, *50*(13), 2931-2941. https://doi.org/10.1021/jm0700565

Souaille, M., & Roux, B. t. (2001). Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *135*(1), 40-57.

Spiga, O., Cicaloni, V., Fiorini, C., Trezza, A., Visibelli, A., Millucci, L., . . . Santucci, A. (2020). Machine learning application for development of a data-driven predictive model able to investigate quality of life scores in a rare disease. *Orphanet Journal of Rare Diseases*. https://doi.org/10.1186/s13023-020-1305-0

Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A., & Case, D. A. (1998). Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate−DNA Helices. *Journal of the American Chemical Society*, *120*(37), 9401-9409. https://doi.org/10.1021/ja981844+

Stansfeld, P. J., Gedeck, P., Gosling, M., Cox, B., Mitcheson, J. S., & Sutcliffe, M. J. (2007). Drug block of the hERG potassium channel: insight from modeling. *Proteins*, *68*(2), 568-580. https://doi.org/10.1002/prot.21400

Stone, A. (2016). Precision medicine: Health care tailored to you. *The White House Blog*.

Strand, D. S., Kim, D., & Peura, D. A. (2017). 25 Years of Proton Pump Inhibitors: A Comprehensive Review. *Gut Liver*, *11*(1), 27-37. https://doi.org/10.5009/gnl15502

Swanson, J. M., Henchman, R. H., & McCammon, J. A. (2004). Revisiting free energy calculations: a theoretical connection to MM/PBSA and direct calculation of the association free energy. *Biophys J*, *86*(1 Pt 1), 67-74. https://doi.org/10.1016/S0006-3495(04)74084-9

Tatar, G., & Taskin Tok, T. (2019). Structure prediction of eukaryotic elongation factor-2 kinase and identification of the binding mechanisms of its inhibitors: Homology modeling, Molecular Docking and Molecular Dynamics Simulation. *J Biomol Struct Dyn*, 1-16. https://doi.org/10.1080/07391102.2019.1592024

Tavassoly, I., Goldfarb, J., & Iyengar, R. (2018). Systems biology primer: the basic methods and approaches. *Essays Biochem*, *62*(4), 487-500. https://doi.org/10.1042/EBC20180003

Thafar, M., Raies, A. B., Albaradei, S., Essack, M., & Bajic, V. B. (2019). Comparison Study of Computational Prediction Tools for Drug-Target Binding Affinities. *Front Chem*, *7*, 782. https://doi.org/10.3389/fchem.2019.00782

Thole, B. T. (1981). Molecular polarizabilities calculated with a modified dipole interaction. *Chemical Physics*, *59*(3), 341-350. https://doi.org/https://doi.org/10.1016/0301-0104(81)85176-2

Titus, G. P., Mueller, H. A., Burgner, J., Rodriguez De Córdoba, S., Peñalva, M. A., & Timm, D. E. (2000). Crystal structure of human homogentisate dioxygenase. *Nature Structural Biology*. https://doi.org/10.1038/76756

Torres, A. M., Bansal, P. S., Sunde, M., Clarke, C. E., Bursill, J. A., Smith, D. J., . . . Vandenberg, J. I. (2003). Structure of the HERG K+ channel S5P extracellular linker: role of an amphipathic alpha-helix in C-type inactivation. *J Biol Chem*, *278*(43), 42136-42148. https://doi.org/10.1074/jbc.M212824200

Torrie, G. M., & Valleau, J. P. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, *23*(2), 187-199. https://doi.org/https://doi.org/10.1016/0021-9991(77)90121-8

Treisman, J., Gönczy, P., Vashishtha, M., Harris, E., & Desplan, C. (1989). A single amino acid can determine the DNA binding specificity of homeodomain proteins. *Cell*, *59*(3), 553-562. https://doi.org/10.1016/0092-8674(89)90038-x

Tuccinardi, T. (2021). What is the current value of MM/PBSA and MM/GBSA methods in drug discovery? *Expert Opinion on Drug Discovery*, *16*(11), 1233-1237. https://doi.org/10.1080/17460441.2021.1942836

Tuckerman, M., Berne, B. J., & Martyna, G. J. (1992). Reversible multiple time scale molecular dynamics. *The Journal of Chemical Physics*, *97*(3), 1990-2001. https://doi.org/10.1063/1.463137

van Duijnen, P. T., & Swart, M. (1998). Molecular and Atomic Polarizabilities: Thole's Model Revisited. *The Journal of Physical Chemistry A*, *102*(14), 2399-2407. https://doi.org/10.1021/jp980221f

van Gunsteren, W. F., & Berendsen, H. J. C. (1987). *Groningen Molecular Simulation (GROMOS) Library Manual*.

Vandenberg, J. I., Perry, M. D., Perrin, M. J., Mann, S. A., Ke, Y., & Hill, A. P. (2012). hERG K(+) channels: structure, function, and clinical significance. *Physiol Rev*, *92*(3), 1393-1478. https://doi.org/10.1152/physrev.00036.2011

Vilboux, T., Kayser, M., Introne, W., Suwannarat, P., Bernardini, I., Fischer, R., . . . Gahl, W. A. (2009). Mutation spectrum of homogentisic acid oxidase (HGD) in alkaptonuria. In.

Wang, C., Greene, D., Xiao, L., Qi, R., & Luo, R. (2017). Recent Developments and Applications of the MMPBSA Method. *Front Mol Biosci*, *4*, 87. https://doi.org/10.3389/fmolb.2017.00087

Wang, F., Yang, W., & Hu, X. (2018). Discovery of High Affinity Receptors for Dityrosine through Inverse Virtual Screening and Docking and Molecular Dynamics. *Int J Mol Sci*, *20*(1). https://doi.org/10.3390/ijms20010115

Wang, K., Sun, J., Zhou, S., Wan, C., Qin, S., Li, C., . . . Yang, L. (2013). Prediction of drug-target interactions for drug repositioning only based on genomic expression similarity. *PLoS Comput Biol*, *9*(11), e1003315. https://doi.org/10.1371/journal.pcbi.1003315

Wang, S., Liu, S., Morales, M. J., Strauss, H. C., & Rasmusson, R. L. (1997). A quantitative analysis of the activation and inactivation kinetics of HERG expressed in Xenopus oocytes. *J Physiol*, *502 (Pt 1)*(Pt 1), 45-60. https://doi.org/10.1111/j.1469-7793.1997.045bl.x

Wang, W., & MacKinnon, R. (2017). Cryo-EM Structure of the Open Human Ether-à-go-go-Related K. *Cell*, *169*(3), 422-430.e410. https://doi.org/10.1016/j.cell.2017.03.048

Warmke, J. W., & Ganetzky, B. (1994). A family of potassium channel genes related to eag in Drosophila and mammals. *Proc Natl Acad Sci U S A*, *91*(8), 3438-3442. https://doi.org/10.1073/pnas.91.8.3438

Warshel, A., & Lifson, S. (1970). Consistent Force Field Calculations. II. Crystal Structures, Sublimation Energies, Molecular and Lattice Vibrations, Molecular Conformations, and Enthalpies of Alkanes. *The Journal of Chemical Physics*, *53*(2), 582-594. https://doi.org/10.1063/1.1674031

Weingarth, M., van der Cruijsen, E. A., Ostmeyer, J., Lievestro, S., Roux, B., & Baldus, M. (2014). Quantitative analysis of the water occupancy around the selectivity filter of a K+ channel in different gating modes. *J Am Chem Soc*, *136*(5), 2000-2007. https://doi.org/10.1021/ja411450y

Westbrook, J., Feng, Z., Burkhardt, K., & Berman, H. M. (2003). Validation of protein structures for protein data bank. *Methods Enzymol*, *374*, 370-385. https://doi.org/10.1016/S0076-6879(03)74017-8

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., . . . Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*, *46*(D1), D1074-D1082. https://doi.org/10.1093/nar/gkx1037

Woosley, R., Heise, C., & Romero, K. (2020). *QTdrugs List*

Workman, J. L., & Kingston, R. E. (1998). Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu Rev Biochem*, *67*, 545-579. https://doi.org/10.1146/annurev.biochem.67.1.545

Wu, H., Nüske, F., Paul, F., Klus, S., Koltai, P., & Noé, F. (2017). Variational Koopman models: Slow collective variables and molecular kinetics from short off-equilibrium simulations. *J Chem Phys*, *146*(15), 154104. https://doi.org/10.1063/1.4979344

Wulff, H., Castle, N. A., & Pardo, L. A. (2009). Voltage-gated potassium channels as therapeutic targets. *Nat Rev Drug Discov*, *8*(12), 982-1001. https://doi.org/10.1038/nrd2983

Xia, X. (2013). *Comparative Genomics* (1 ed.). Springer Berlin, Heidelberg. https://doi.org/https://doi.org/10.1007/978-3-642-37146-2

Yang, X., Wang, Y., Byrne, R., Schneider, G., & Yang, S. (2019). Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem Rev*, *119*(18), 10520-10594. https://doi.org/10.1021/acs.chemrev.8b00728

Yang, Y. I., Shao, Q., Zhang, J., Yang, L., & Gao, Y. Q. (2019). Enhanced sampling in molecular dynamics. *J Chem Phys*, *151*(7), 070902. https://doi.org/10.1063/1.5109531

Zatkova, A., Ranganath, L., & Kadasi, L. (2020). Alkaptonuria: Current perspectives. In.

Zerze, G. H., Zheng, W., Best, R. B., & Mittal, J. (2019). Evolution of All-Atom Protein Force Fields to Improve Local and Global Properties. *J Phys Chem Lett*, *10*(9), 2227-2234. https://doi.org/10.1021/acs.jpclett.9b00850

Zhou, C. Y., Jiang, F., & Wu, Y. D. (2015). Residue-specific force field based on protein coil library. RSFF2: modification of AMBER ff99SB. *J Phys Chem B*, *119*(3), 1035-1047. https://doi.org/10.1021/jp5064676

Zhou, P. Z., Babcock, J., Liu, L. Q., Li, M., & Gao, Z. B. (2011). Activation of human ether-a-go-go related gene (hERG) potassium channels by small molecules. *Acta Pharmacol Sin*, *32*(6), 781-788. https://doi.org/10.1038/aps.2011.70

Zhou, Z., Vorperian, V. R., Gong, Q., Zhang, S., & January, C. T. (1999). Block of HERG potassium channels by the antihistamine astemizole and its metabolites desmethylastemizole and norastemizole. *J Cardiovasc Electrophysiol*, *10*(6), 836-843. https://doi.org/10.1111/j.1540-8167.1999.tb00264.x

Zhu, S., Okuno, Y., Tsujimoto, G., & Mamitsuka, H. (2005). A probabilistic model for mining implicit 'chemical compound-gene' relations from literature. *Bioinformatics*, *21 Suppl 2*, ii245-251. https://doi.org/10.1093/bioinformatics/bti1141

List of publications

Pettini, F., Domene, C., & Furini, S. (2023). Early Steps in C-Type Inactivation of the hERG Potassium Channel. *Journal of Chemical Information and Modeling*. https://doi.org/10.1021/acs.jcim.2c01028

Lazzerini, P. E., Cartocci, A., Qu, Y. S., Saponara, S., Furini, S., Fusi, F., . . . Boutjdir, M. (2021). Proton Pump Inhibitors Directly Block hERG-Potassium Channel and Independently Increase the Risk of QTc Prolongation in a Large Cohort of US Veteran. *Circulation: Arrhythmia and Electrophysiology*. https://doi.org/10.1161/CIRCEP.121.010042

Pettini, F., Visibelli, A., Cicaloni, V., Iovinelli, D., & Spiga, O. (2021). Multi-Omics Model Applied to Cancer Genetics. *International Journal of Molecular Sciences*. https://doi.org/10.3390/ijms22115751

Spiga, O., Cicaloni, V., Fiorini, C., Trezza, A., Visibelli, A., Millucci, L., . . . Santucci, A. (2021). Machine learning application for development of a data-driven predictive model able to investigate quality of life scores in a rare disease. *Orphanet Journal of Rare Diseases*. https://doi.org/10.1186/s13023-020-1305-0

Cicaloni, V., Trezza, A., Pettini, F., & Spiga, O. (2019) Applications of in Silico Methods for Design and Development of Drugs Targeting Protein-Protein Interactions. *Current Topics in Medicinal Chemistry.* https://doi.org/10.2174/1568026619666190304153901

Spiga, O., Gardini, S., Rossi, N., Cicaloni, V., Pettini, F., Niccolai, N., & Santucci, A. (2018). *Structural investigation of Rett-inducing MeCP2 mutations*. Genes & diseases. https://doi.org/10.1016/j.gendis.2018.09.005

Acknowledgements

Ringrazio il Professore Simone Furini per le competenze professionali che mi ha trasmesso, spronandomi sempre a non demordere durante questo percorso.

Ringrazio i miei genitori per il supporto incondizionato nella vita di tutti i giorni e in tutte le mie scelte accademiche.

Ringrazio la mia splendida compagna che mi ha affiancato (e sopportato) durante questa difficile fase della vita, donandomi infinito amore e comprensione.

Dedico questa tesi a tutte le persone che mi hanno privato di momenti felici,
augurandogli di calcare le mie orme.