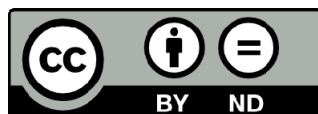




UNIVERSITAT DE
BARCELONA

**Development and application
of computer-aided strategies
for virtual screening and hit-to-lead**

Carles Perez Lopez

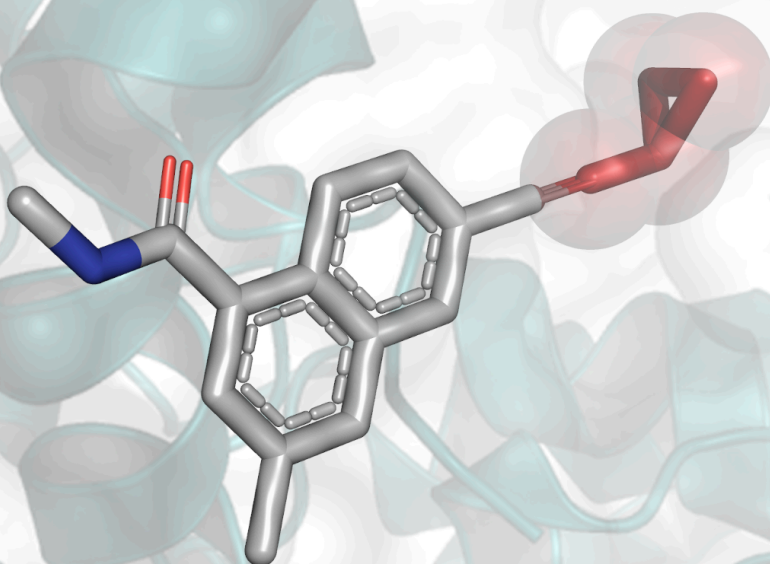


Aquesta tesi doctoral està subjecta a la llicència *Reconeixement- SenseObraDerivada 4.0.
Espanya de Creative Commons.*

Esta tesis doctoral está sujeta a la licencia *Reconocimiento - SinObraDerivada 4.0.
España de Creative Commons.*

This doctoral thesis is licensed under the *Creative Commons Attribution-NoDerivatives 4.0.
Spain License.*

Development and application of computer-aided strategies for virtual screening and hit-to-lead



Carles Pérez López

Development and application of computer-aided strategies for virtual screening and hit-to-lead

Universitat de Barcelona - Facultat de Biologia

Programa de doctorat en Biomedicina

Línia d'investigació en Bioinformàtica

Desenvolupada a Electronic and Atomic Protein Modeling (EAPM) group del
Barcelona Supercomputing Center (BSC).

Memòria presentada per **Carles Perez Lopez** per optar al grau de doctor per la
Universitat de Barcelona



Dirigida per:

Prof. Victor Guallar Tasies (BSC)

Tutor:

Prof. Josep Lluís Gelpí Buchaca (UB)

Barcelona, 2022



UNIVERSITAT DE
BARCELONA

To all my loved ones

Acknowledgments

Per començar vull donar les gràcies al meu director, en Victor Guallar, per haver-me donat l'oportunitat i la confiança necessària per dur a terme tots els projectes. Gràcies a ell he après a fer recerca, a entendre les simulacions moleculars i a aplicar-les al disseny de fàrmacs. Tanmateix, m'ha ensenyat a comunicar, a col·laborar amb els meus companys, i a mantenir sempre encesa la llum de la curiositat. També vull agrair al meu tutor, Josep Lluís Gelpí, la seva gran ajuda i dedicació en tot el que ha fet referència al programa de doctorat.

No hauria arribat al final d'aquest llarg camí sense companys i companyes increïbles, i per això, els agraeixo de tot cor haver format part d'aquest viatge. Donar gràcies especialment a Dani Soler per totes les seves lliçons de programació i intel·ligència artificial, a Joan Gilabert i a Martí Municoy per totes les idees i discussions, i a Sergi Rodà, Pep Amengual i Gerard Santiago pels seus útils consells i reflexions. Sobretot, donar les gràcies a l'Ignasi Puch. Ha sigut un plaer poder col·laborar amb ell, argumentar els resultats i aprendre junts. Sense dubte, una gran part d'aquesta tesi no hauria sigut possible sense la seva ajuda. Estic molt agraït també a tota la resta de companys que han format part del grup: Alexis, Alberto, Ferran, Isaac, Ruben, Jelisa, Laura, Ana, Anna, Masoud, Oliver, Marc, Marina, i un llarg etcètera. Agrair també a Nostrum Biodiscovery la seva col·laboració en els projectes i tot el que he pogut aprendre dels seus professionals. Entre d'altres, donar gràcies especials a la Suwipa i la Lucía. També vull agrair a Robert Soliva per tot el que ens va ajudar durant el desenvolupament de FragPELE. Tanmateix, moltes gràcies a Almirall per haver pogut formar part de SilicoDerm.

Fora de l'àmbit laboral també he rebut el suport dels amics i el caliu de la família. Moltes gràcies a tots als que hi sou, i també, als que malauradament, ens heu deixat. I sobretot, dedicar les últimes paraules a la meva parella, Ona, per estar sempre al meu costat, donant-me suport i estimar-me més dia rere dia. T'estimo.

Abstract

The incremental application of supercomputers to offer solutions to complex problems has motivated the usage of computational modeling tools in drug design pipelines. Specifically, small low-affinity compounds are modified by including multiple decorators in the hit-to-lead phase to obtain more potent compounds. Techniques such as docking provide quick answers on classifying millions of candidates, differentiating active from inactive, but their accuracies tend to drop when ranking ligand's potencies. Expensive methods such as FEP are more precise; however, the time consumption limitate their application in hit-to-lead campaigns.

This thesis aims to implement and test novel methodologies in a mid computation and accuracy term, focused on facing hit-to-lead stages. We developed FragPELE, a novel ligand growing method integrated with PELE, an unconventional Monte Carlo sampling algorithm. FragPELE introduces a new concept of progressively expanding a small atom-sized moiety of atoms (fragment) within PELE simulations, adapting the protein binding site to the newly grown R-group. Structural and scoring benchmarks remarked accurate geometrical predictions and correlation with relative free energies, with a reasonable consumption of time and resources. Besides, we combined FragPELE with the recently developed aquaPELE algorithm to expand fragments on hydrated binding sites. Results stressed improved accuracies when introducing the mixed implicit/explicit solvent models integrated within aquaPELE.

Additionally, we participated in a collaborative project with Almirall. We assessed our FragPELE tool in two prospective hit-to-lead studies. One of them ended with synthesizing an improved version of the initial hit. On the other, the method showed good predictive power in classifying non-terminal R-groups on 27 new compounds (not reported in the literature). Finally, we optimized virtual screening pipelines by integrating machine learning analysis with simulated data, training, testing, and validating the designed classification models with external experimental sets. From 785 compounds, Almirall purchased 23 based on our results. Two of them showed inhibition of the target, one in the nM range of activity.

List of Contents

Acknowledgments	3
Abstract	5
List of Contents	7
Abbreviations	11
Chapter 1. Introduction	13
Motivation: the cost of a new drug	13
Drug design basics	13
Biomolecular target: proteins	13
Protein-ligand binding	16
Binding sites	16
Binding kinetics	16
Inhibition	17
Relation between kinetics and binding free energies	20
Early drug discovery	20
Molecular modeling	23
Molecular mechanics and force fields	23
Virtual Screening methods	27
Docking	27
Molecular Dynamics	29
Monte Carlo Simulations	30
PELE	32
AquaPELE	34
Estimating binding free energies	36
Docking scores	37
End-point methods	37
Alchemical methods	38

Machine learning	40
Ligand growing	43
Chapter 2. Objectives	47
Chapter 3. FragPELE: dynamic ligand growing within a binding site	49
FragPELE algorithm	49
The method	49
Structural validation	56
Self-growing	56
Cross-growing	58
Cryptic sub-pockets	64
Growing bulky R-groups	68
Protein motion	70
Growing and scoring	71
Growing on hydrated sites	74
Structural validation with explicit waters	75
Methods	75
Results	81
Growing and scoring on hydrated systems	86
Methods	86
Results	90
Chapter 4. SilicoDerm: an industrial drug discovery project	97
Precedents	97
FragPELE on Kinase 1	98
Methods	98
Results	101
Study of Kinase 2	104
Machine learning on optimizing virtual screening pipelines	105
Methods	105

Results	113
Hit-to-lead with FragPELE	117
Methods	117
Results	119
Chapter 5. Discussion	123
Total flexibility to predict binding poses with FragPELE	123
Binding affinities and scoring	124
Effect of explicit waters	125
FragPELE: growing and sampling	127
Importance of automatizing in drug design studies	128
Prospective hit-to-lead with FragPELE	129
Optimizing virtual screening pipelines through machine learning algorithms	131
The humanity behind machine learning	132
Strengths, limitations and opportunities of the current techniques	134
Chapter 6. Conclusions	139
References	141
Appendices	161
Appendix A	161
Appendix B	167
Appendix C	171
Appendix D	177
Appendix E	181
Appendix F	187

Abbreviations

$\Delta\Delta G$	Relative free energy of binding
ΔG	Gibbs free energy
ANM	Anisotropic Network Model
BS	Binding site
CADD	Computer-aided drug discovery
CPU	Central processing unit
DL	Deep learning
EDD	Early drug discovery
EGFR	Epidermal growth factor receptor
FBDD	Fragment-based drug design
FEP	Free energy perturbation
FFP	Force field parameters
GPU	Graphical processing unit
GS	Growing step
H2L	Hit-to-lead
HBs	Hydrogen bonds
HTS	High throughput screening
HTVS	High throughput virtual screening
KNN	K nearest neighbor
MC	Monte Carlo
MD	Molecular dynamics
ML	Machine learning
MM	Molecular mechanics
MW	Molecular weight

PCNA	Proliferating cells nuclear antigen
QSAR	Quantitative structure-activity relationship
QM	Quantum mechanics
RMSD	Root mean square deviation
VDW	Van der Waals

Chapter 1. Introduction

1. Motivation: the cost of a new drug

Drug discovery is an extraordinarily costly and lengthy process. Recent studies have shown that the mean cost of developing a new drug is estimated to be 1.3 billion US\$ (Wouters et al., 2020) and, in terms of time, the average is around 12 years (Van Norman, 2016). These are the general rules for well-characterized diseases with a reasonably known mechanism. Still, if we move to rare disorders, the numbers are heartbreaking (approximately four years longer) (Burton et al., 2021). Accordingly with the United States government, since the application of computer-aided drug design (CADD), costs have been reduced by around 130 million US\$ and by a year of research time (“Report to Congressional Requesters: New Drug Development—Science, Business, Regulatory, And Intellectual Property Issues Cited as Hampering Drug Development Efforts,” 2007). This fact clearly shows the potency of using computers to enhance drug discovery. Thus, any development or improvement in this field is becoming a crucial turning point in designing new treatments.

2. Drug design basics

2.1. Biomolecular target: proteins

The central dogma of molecular biology states, “DNA encodes RNA, RNA encodes protein” (Crick, 1970), meaning that the flow of genetic information starts from the DNA, passes through the RNA, and ends with proteins. Therefore, proteins are the biomolecules in charge of applying the information encoded in the genes. They are involved in a massive variety of roles such as structural scaffolds, hormonal signaling, catalyzing digestive reactions, muscle contraction and relaxation, transportation, metals storing, immunological functions, regulate gene expression, among others (Petsko & Ringe, 2004).

The building blocks to assemble proteins are amino acids. As their name suggests, they are organic molecules formed by amino ($-\text{NH}_2$) and carboxylic ($-\text{COOH}$) functional groups and a specific side-chain (R-group) (*Figure 1.1*) that provide their particular characteristics (A. B. Hughes, 2009).

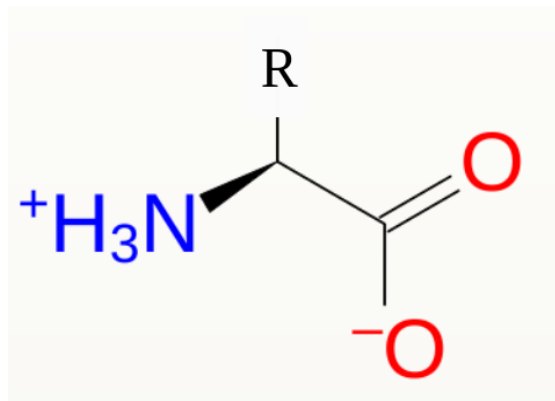


Figure 1.1. Chemical structure of amino acid (zwitterion state). R = side-chain.

Structurally, proteins are distributed in four levels, as represented on the right side of *Figure 1.2*. The primary structure is the linear filament or sequence of amino acids, encoded by the RNA to synthesize the 21 amino acid types (including selenocysteine) (*Figure 1.2- left side*). After the attachment of several amino acids, the filament loses stability and folds, generating the *secondary structure*. Here two regular folding patterns exist. When the chain coils around an axis in the clockwise direction, it causes *α -helices*; if it is planar, they are called *β -sheets*, and the irregular structures attaching them are *loops*. At this point, the polypeptide (chain of amino acid monomers) is big enough to combine the *α -helices*, *β -sheets*, and *loops* to complete the overall folding and generate domains, the *tertiary structure*. Here, the whole protein adopts a three-dimensional shape that looks random and irregular; however, the forces between amino acids stabilize and create this layout to perform the protein's specific function. Sometimes a single polypeptide (monomer or subunit) is insufficient to carry out its activity. They need to combine multiple subunits to adopt the active conformation, giving the *quaternary structure*. For example, the proliferating cell nuclear antigen (PCNA) is a homotrimer composed of three identical monomers (*Figure 1.2, right-side*) (Dieckman et al., 2012). Another well-known example is hemoglobin, a heterotetramer

constituted by two α and two β subunits (four monomers in total) (Robert J. Ouellette, 2018).

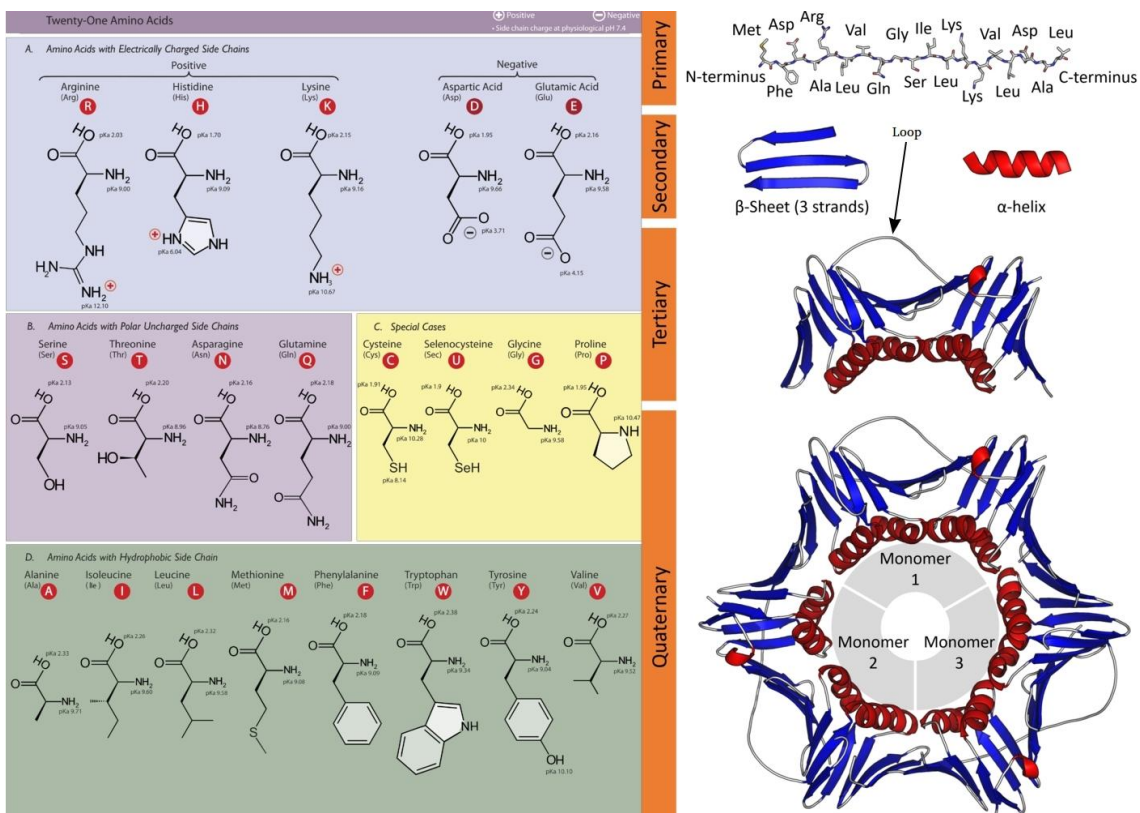


Figure 1.2. (left-side) The 21 proteinogenic α -amino acids (including selenocysteine) are classified according to their chemical properties. Charges are at physiological pH (7.4). Image created by [Dan Cojocari](#), licensed by [CC BY-SA 3.0](#) (source: https://commons.wikimedia.org/wiki/File:Amino_acids.png). **(right-side)** Summary of protein structure (primary, secondary, tertiary, and quaternary) using the example of PCNA. (PDB: 1AXC). Modification from the original image created by [Thomas Shafee](#), licensed by [CC BY 4.0](#) (source: [https://commons.wikimedia.org/wiki/File:Protein_structure_\(full\).png](https://commons.wikimedia.org/wiki/File:Protein_structure_(full).png))

The 3D surface of the whole protein is full of bumps and pockets. Most of them are just a consequence of the protein folding, but when a pocket favors the binding with another molecule with high specificity, we call this region *binding site* (BS). The molecule that binds to this area is usually named *ligand*, which can be of any kind, such as DNA, RNA, other proteins, hormones, lipids, small compounds, and ions. Then, drugs are ligands used for therapeutic purposes.

2.2. Protein-ligand binding

2.2.1. Binding sites

Protein BSs are usually concave cavities where the (cognate) ligand tends to allocate with high specificity. This event triggers a cascade of conformational changes in the protein that can alter, induce, or block its function or modify the ligand. There are different types of BSs depending on the effect that their binding produces in the complex.

Orthosteric sites are pockets where ligands bind to perform protein endogenous action. Specifically, in enzymes, which are proteins that catalyze chemical reactions, this cavity is called the *active site*. When the ligand (substrate) binds there, it induces a chemical reaction to obtain a new product (Wilson, 2010). For example, kinases are enzymes that catalyze the donation of a phosphate group from adenosine triphosphate (ATP) to produce adenosine diphosphate (ADP) and phosphorylate the target-protein molecule.

In some proteins, the binding of the ligand produces a conformational change that modifies the BS, opening a new pocket that was not present in the apo form (protein with the ligand unbound), so they are not easily detectable. These *cryptic binding sites* or *cryptic pockets* (Cimermanic et al., 2016) are only detected in holo form (when the ligand is bound).

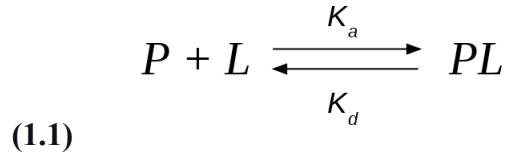
Apart, *allosteric sites* have regulatory functions. The binding of the ligand can increase or reduce the regular activity of the protein, usually due to a conformational change that affects protein dynamics (Srinivasan et al., 2014).

2.2.2. Binding kinetics

Binding events are dynamic. Along with time, the ligand will be temporarily within the BS, and later on, it will unbind, letting it free again to allocate another molecule.

Equation 1.1 describes this process, where P is the protein-free, L is the ligand-free, and PL is the protein with the BS occupied by the ligand. The binding constant or association constant (K_a) describes the strength of the binding of the ligand to the protein, and contrary, the dissociation constant (K_d) indicates its trend to separate

(Cleaves, 2011). Both are obtained from the free protein ($[P]$), free ligand ($[L]$), and complex ($[PL]$) concentrations (*Equations 1.2 and 1.3*).

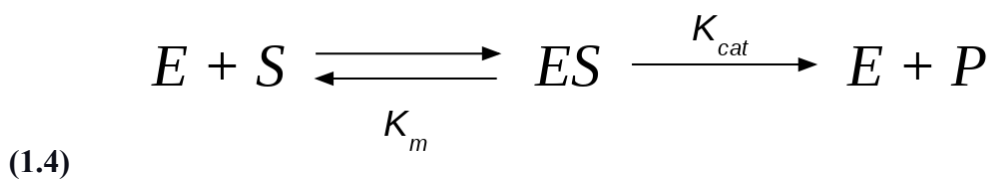


Equation 1.1: Ligand binding equilibrium model.

$$(1.2) \quad K_a = \frac{[PL]}{[P] \cdot [L]} \quad (1.3) \quad K_d = \frac{[P] \cdot [L]}{[PL]}$$

Equations 1.2 and 1.3. Description of association constant (K_a) and dissociation constant (K_d).

The paradigm slightly changes when the protein is an enzyme. In this case, the binding of the substrate releases the product (*Equation 1.4*) (Michaelis et al., 2011). Here the catalytic constant (K_{cat}) indicates the maximal number of molecules of substrate converted to product (Roskoski, 2015), and the Michaelis constant (K_m) describes the enzyme/substrate binding.



Equation 1.4. Enzyme kinetics model. E= enzyme; S= substrate; P= product; K_{cat} = catalytic constant; K_m = Michaelis constant.

2.2.3. Inhibition

Previously, we have talked about ligands in general. Within this diverse group of molecules, there are *inhibitors*, a ligand that decreases the regular function of the protein. Most inhibitors bind to the protein with non-covalent interactions, such as hydrogen bonds (HBs), electrostatic or hydrophobic forces. Here multiple weak bonds

in combination create a specific and firm binding. These are non-covalent *reversible inhibitors*, and there exist different types according to their effect on the protein.

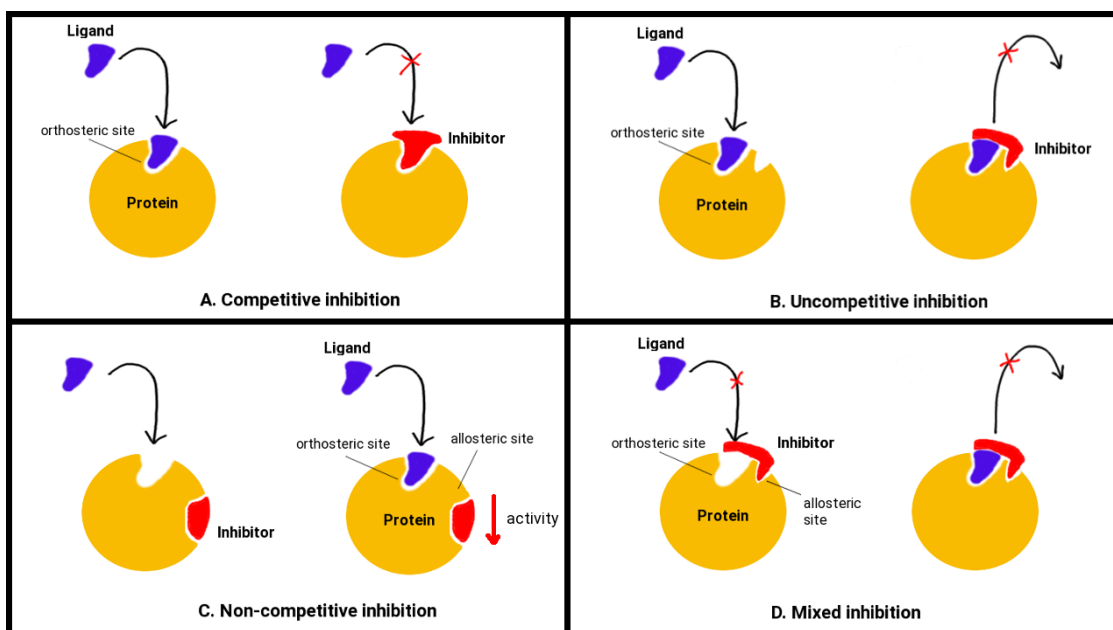


Figure 1.3. Representation of reversible inhibition types.

Competitive inhibitors (Figure 1.3A) are molecules with an affinity towards the orthosteric site of the biomolecule, preventing the union of the natural target when it is bound. Most of the inhibitors in this thesis will be competitive ones; however, three more kinds of inhibition need to be explained.

Uncompetitive inhibitors (Figure 1.3B) bind to the protein-ligand complex, difficulting the ligand release; therefore, it can only happen in the presence of the ligand. Differently, *non-competitive inhibitors* (Figure 1.3C) bind to an allosteric site, independently of whether the ligand binds or not, producing conformational changes that affect the activity of the protein but not the union of the ligand to the orthosteric BS. The last type is the *mixed inhibition* (Figure 1.3D), combining *competitive* and *uncompetitive*. Here, the ligand can bind to the free enzyme or the bound state, but it shows more affinity for one of both states.

Figure 1.4 shows kinetics for all reversible inhibition types. In competitive inhibition, the inhibition constant (K_i) is equal to the K_d of the inhibitor. The lower the K_i , the greater the binding affinity of the inhibitor, so indirectly, it indicates the strength of the

binding. Drugs with lower K_i require less concentration to perform the same effect; thus, *in vitro* assays try to estimate this constant to compare different inhibitors. Then, K_i can only be accurately reported as a binding constant when the mechanism of inhibition is identified.

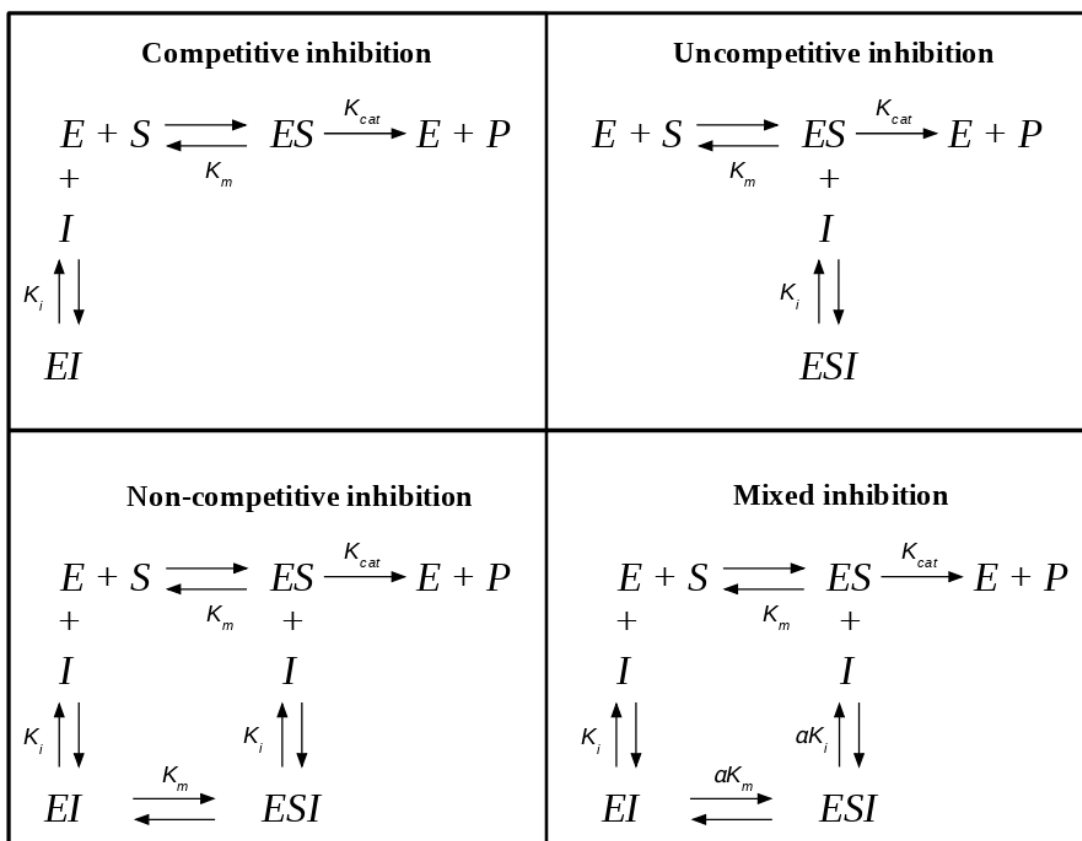


Figure 1.4. Kinetics schema for the different types of reversible enzyme inhibition. E= enzyme; S= substrate; P= product; I= inhibitor; K_{cat} = catalytic constant, K_m = Michaelis constant; K_i = inhibition constant. Notice that mixed inhibition has the equilibrium displaced towards E or ES state; then, this will depend on the size of a (large to uncompetitive, small to competitive).

Even though K_i and K_d are precise metrics to evaluate inhibitors' potency, they are not always easy to obtain directly (by isothermal titration calorimetry, fluorescence quenching, or surface plasmon resonance). A more straightforward but less accurate metric is used for cases that are not affordable to evaluate the K_d , called IC_{50} .

IC_{50} is the concentration of an inhibitor that halves the enzyme's activity. It is easier to compute than K_i but results are exclusively comparable under the same experimental

conditions (Yung-Chi & Prusoff, 1973). Therefore, this metric is beneficial for comparing a series of inhibitors in the same laboratory.

Sometimes the protein-ligand binding is reported in energy instead of K_i or K_d , independent of concentration values. In the following section, we will establish their relationship.

2.2.4. Relation between kinetics and binding free energies

When a ligand binds into a BS, the bound state is energetically favorable. The ligand ‘prefers’ to be within the protein cavity instead of floating around the solvent. Thermodynamically, this energetic reward or payment is called binding free energy or *Gibbs free energy* (ΔG), defined as the energy difference between the protein-ligand bound and unbound states. As shown in *Equation 1.1*, both states are connected by K_d (or the inverse, K_a) in equilibrium. Thus, ΔG can be related to K_d via *Equation 1.5* (Moore, 1973). Notice that K_d is a general term used for substrates; when the ligand is an inhibitor, this constant changes the name to K_i , which is the dissociation constant of the inhibitor.

$$(1.5) \quad \Delta G = -RT \ln \frac{K_d}{C^\circ}$$

Equation 1.5. Relation between binding free energy and equilibrium (dissociation) constant. ΔG = Gibbs free energy; R = ideal gas constant; T = temperature; C° = standard reference concentration (1 mol/L).

Along with this thesis, most of the computational techniques applied in drug discovery pursue the correct estimation of ΔG to precisely compare different ligands, so the relationship between binding kinetics and energetic terms is a must to understand the goal of these methods properly.

2.3. Early drug discovery

Delivering a completely safe, effective, and administrable drug to the market is an extensive and highly costly process (Van Norman, 2016; Wouters et al., 2020). Early

drug discovery (EDD) screenings are performed to obtain a drug candidate. All their steps are summarized in *Figure 1.5*.

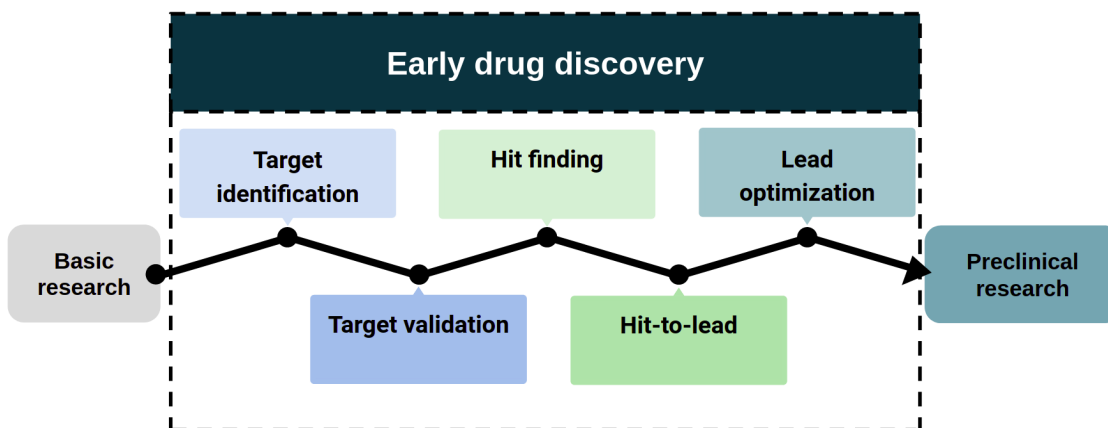


Figure 1.5. Schematic representation of EDD phases.

The discovery of any drug begins with intense basic research to identify the biomolecular mechanism behind the disease of interest. Often, this process starts in an academic-industrial collaboration environment, where multiple hypotheses of which biomolecular pathway should be activated or inhibited to solve malfunctions caused by the disease are evaluated (Emmerich et al., 2021; Everett, 2015; J. P. Hughes et al., 2011). This initial step of any drug discovery project is called *target identification*, which is crucial to know the effect of blocking or activating a specific biomolecule. Proper targets must be safe, effective, and druggable, among others. F.ex: bioinformatic and data mining strategies are novel methods that are helping in this step (Yongliang Yang, S James Adelstein, Amin I Kassis, 2009).

Once the target is identified, further demonstrations are required to ensure its involvement in the disease, and this step is named *target validation*. Multiple *in vitro* and *in vivo* techniques, such as expression profiles, functional analysis, biomarkers, and cell-based models, are conducted to complete the validation (Hans-Joachim Anders, 2007).

Then, in the *hit finding*, hundreds of series of low molecular weight (MW) compounds are screened to produce the activation or inhibition (depending on the aim) of the target of interest. Here the final goal is to obtain *hits*, small compounds with low but

demonstrable activity. High throughput screening (HTS) is the most common strategy, where large libraries of compounds are directly tested against the target using a complex assay system (Fox et al., 2006). These studies allow us to get which specific compounds to bind to the target and give information about which structures and functional groups are more prone to have activity against the target. Thanks to the advances in supercomputation, we can use this knowledge to find hit compounds virtually, which is known as *virtual screening* (VS). VS has significantly reduced the costs of testing experimentally expensive libraries of compounds, and it has become a standard tool in the EDD. For example, 2D-based machine learning techniques can virtually screen a 1 billion molecular library in 24 hours of modest supercomputing resources (Gentile et al., 2020; Ton et al., 2020).

Once a hit is identified, multiple modifications are added to this initial molecule to improve its potency. This step is called *hit-to-lead* (H2L). Usually, it starts with compounds in the range of micromolar (μM : 10^{-6}g/mol) of IC₅₀ to finally increase their affinity to the nanomolar (nM: 10^{-9}g/mol) range, raising its potency to 1000 folds, approximately.

Researchers would generate a potent compound with high affinity and selectivity against the target at the end of this step. However, this molecule needs to be optimized to ensure its absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties in the body. Large batteries of tests are performed, such as analysis of solubility and permeability, microsomal stability assays, and examination of CYP450 inhibition, among others. The final goal is to get a compound able to reach the biomolecular target without being degraded or becoming toxic along the path (van de Waterbeemd & Gifford, 2003). Luckily, at the end of this long process, a drug candidate is obtained to continue preclinical research (*in vivo* animal models).

This thesis is focused on the *hit finding* and *hit-to-lead* phases of the EDD under a computed-based point of view. The following sections will introduce some of the computational-based methods used to model protein-ligand interactions and predict which compounds are more prone to hit the target of interest or increase the affinity against it.

3. Molecular modeling

Since the mid-nineteenth century, multiple advances have been made in organic chemistry and structural theory. Decades of research that started from rationalizing formulas, two-dimensional and three-dimensional draws, and designing ‘favor configurations’ have resulted in what today is known as *molecular modeling* (Williams, 1996). Nowadays, this is a general term to refer to any theoretical model or computational technique to predict molecule behavior (Genheden et al., 2017). In the past, the most straightforward computations could be done by hand, but when the system’s complexity increases, computers are mandatory tools, which is the trend when modeling biomolecules. In this field, molecular modeling offers an extensive toolkit of highly versatile techniques that can be applied in multiple steps of the drug discovery process. This thesis will review a few of the most common methodologies and tools for hit finding and hit-to-lead phases.

3.1. Molecular mechanics and force fields

This section will present *physics-based* models to describe atoms, providing the environment to mimic the behavior of molecules in computational modeling techniques.

When dealing with fully atomistic models, we can roughly classify them as molecular mechanics (MM) or quantum mechanics (QM), depending on the theoretical level of detail; however, there also exist hybrid models (QM/MM). MM defines atoms as particles using classical mechanics (Newtonian mechanics), while QM includes a high description level, including explicit electron particles. Due to the high complexity of QM calculations, they are costly in computation. Thus, they are usually applied in particular cases, such as descriptions of enzymatic catalytic processes (Náray-Szabó et al., 2013).

MM considers atoms as solid spheres linked by covalent bonds modeled as springs. Each sphere is placed in a coordinate space (internal or Cartesian), provided with a Van Der Waals (VDW) radius (representing the closest approach between atoms) and electrostatic charges. All terms are combined through a potential energy function obtained from the sum of bonded and non-bonded energy terms (*Equation 1.6*).

$$(1.6) \quad E = E_{\text{bonded}} + E_{\text{nonbonded}}$$

Equation 1.6. The simplified potential energy function of MM models.

The first term comprises intramolecular or internal interactions between covalently bound atoms, and it englobes four types of sub-terms: bonded, angles, dihedrals, and improper dihedrals. All individual sub-terms are expanded in *Equation 1.7*. Bonds, angles, and improper dihedrals sub-terms are measured as the harmonic potential deviation from the equilibrium value (bond length or angles). In contrast, the dihedral component is represented by a sum of cosine functions with multiplicities. This last term models steric interactions between atoms separated by three covalent bonds, known as 1,4 pairs.

$$(1.7) \quad E_{\text{bonded}} = \sum_{\text{bonds}} K_b (r - r_{eq})^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_{eq})^2 +$$

$$\sum_{\text{impr. dihedrals}} K_\varphi (\varphi - \varphi_{eq})^2 +$$

$$\sum_{\text{dihedrals}, n} \sum_{\phi, n} K_{\phi, n} [1 + \cos(n\phi - \delta_n)]$$

Equation 1.7. Expanded bonded energy term. K_b = bond force constant; r = bond length; r_{eq} = equilibrium bond length; K_θ = angle force constant, θ = angle; θ_{eq} = equilibrium angle; K_φ = improper dihedral force constant, φ = improper dihedral; φ_{eq} = improper equilibrium dihedral; ϕ = dihedral; $K_{\phi, n}$ = dihedral amplitude; n = dihedral multiplicity; δ_n = dihedral phase.

Two terms constitute the non-bonded energy (*Equation 1.8*). A Coulomb (electrostatic) potential models the former. At the same time, the latter applies a Leonard-Jones (L-J) potential, which is attractive at high distances and becomes repulsive when distances are

too short (Vanommeslaeghe et al., 2014). *Figure 1.6* illustrates all energetic terms graphically and separately.

$$(1.8) \quad E_{\text{non-bonded}} = \sum_{\text{nonbonded pairs } ij} \frac{q_i q_j}{4\pi D r_{ij}} + \sum_{\text{nonbonded pairs } ij} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]$$

Equation 1.8. Expanded nonbonded energy term. $q_i q_j$ = partial charges; $\frac{1}{4\pi D}$ = Coulomb constant; ϵ_{ij} = L-J well depth; $\sigma_i \sigma_j$ = L-J radius; $r_i r_j$ = atom-atom distance

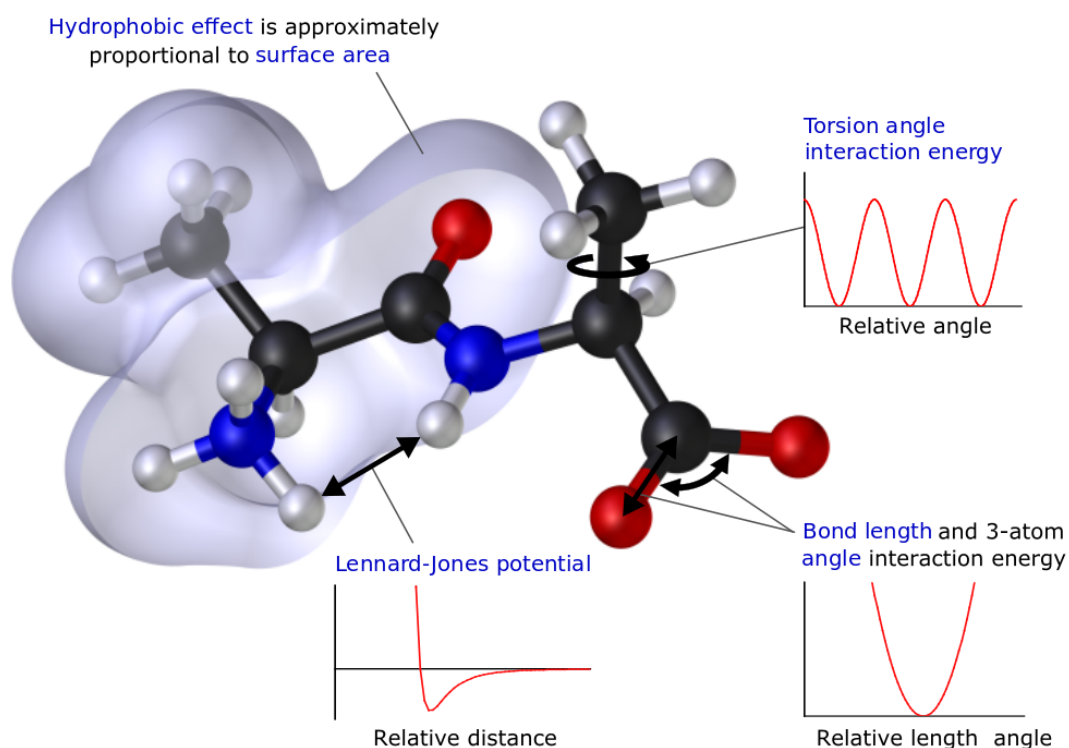


Figure 1.6. Illustration of interactions present in molecular mechanics model. Potential energy functions are also shown. Licensed by [CC BY-SA 3.0](https://creativecommons.org/licenses/by-sa/3.0/) (source: https://commons.wikimedia.org/wiki/File:MM_PEF_3_small.svg)

Force fields contain the set of constants, distances, charges, and angles needed to compute all previously mentioned energies. Their design is based on experimental values or calculations that must be extensively parameterized and validated. There are dozens of force fields, and each tries to aim its use into a concrete application field. Among the most common force fields used to parameterize biomolecular environments we can find *AMBER* (Maier et al., 2015), *CHARMM* (Patel & Brooks, 2004; Vanommeslaeghe et al., 2009), *OPLS-AA* (Jorgensen et al., 1996; Shivakumar et al., 2012), *GAFF* (J. Wang et al., 2004), and *GROMOS* (Oostenbrink et al., 2004).

In molecular simulations, biomolecules and organic compounds are surrounded by the solvent, and its influence also needs to be considered. There are two different ways to model solvent molecules: explicitly and implicitly. Water molecules are directly included in the model in explicit solvents, showing a more realistic approximation where interactions between water molecules and the solute (protein, DNA, organic compounds) can be caught. *TIP3P* (Price & Brooks, 2004) and *SPC* (Berendsen et al., 1987) are the most extensively used explicit models. Even though these methods are more accurate, adding such a significant amount of molecules to the model is computationally expensive. In this sense, implicit models are more straightforward approaches that treat the solvent as a continuous homogeneous polarizable medium, directly included by adding the term to the potential energy function. This term is called solvation energy, expanded in *Equation 1.9*. The polar term is the energy of distributing the charges of the solute, and the non-polar term is the unfavorable energy needed to create a cavity to allocate the volume and shape of the solvate (excluded solvent) and their favorable attractive energy of VDW interactions with the solvent (Levy et al., 2003). *OBC* (Onufriev et al., 2004), *surface generalized Born model* (SGBNP), and its variable dielectric version (VDGBNP) (Zhu et al., 2007) are some examples of implicit solvent models.

$$(1.9) \quad \Delta G_{solv} = \Delta G_{pol} + \Delta G_{np}$$

Equation 1.9. Solvation energy function. ΔG_{solv} = solvation energy; ΔG_{pol} = polar term; ΔG_{np} = non-polar term.

Combining force field parameters with solvation models allows the calculation of the system's potential energy for a specific set of atom coordinates, which is fundamental in molecular modeling. The following sections will show some of the most relevant molecular modeling techniques.

3.2. Virtual Screening methods

Virtual screening (VS) methods, similarly to HTS, are usually used in the early stages of the drug discovery process to enrich libraries with more active compounds. They are fast enough to perform a quick superficial screening of libraries with thousands of compounds in just a few hours, substantially reducing experimental time and costs. In general terms, VS techniques are classified into two major groups: *ligand-based* or *receptor-based* (Gimeno et al., 2019).

Ligand-based VS methods stand on the premise that compounds with similar properties to the active ones will be more prone to be active. They only need ligands and do not involve large biomolecules in their calculations; thus, they are significantly faster. Fingerprint-based (Cereto-Massagué et al., 2015), 3D-shape similarity (Kumar & Zhang, 2018), electrostatic potential similarity, and ligand-based pharmacophores (Wermuth et al., 1998) are examples of groups of ligand-based techniques.

The previous methods assist the quick and dirty filtering with millions of compounds. Eventually, this screening will lead to highly similar compounds, and hopefully, some of them will also be active. However, ligand-based studies can fail when facing targets with complex BS or systems with poor binders' information. At this point, it is essential to increase computation expenses and include the receptor in calculations. Protein-ligand *docking* is the most popular method from *receptor-based VS*.

3.2.1. Docking

Molecular docking is the gold-standard method used in VS. The aim of these techniques is to predict the ligand's best binding mode within the receptor's BS, which must be included within a user-selected grid. *Figure 1.7* represents a general view of the method.

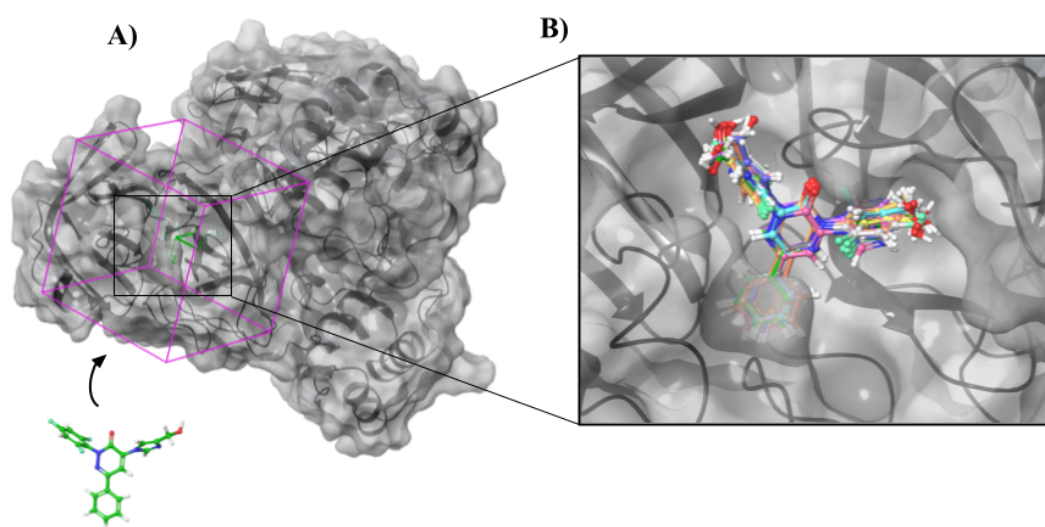


Figure 1.7. Illustration of docking method. **A)** Full view of the receptor, highlighting the grid (pink square) where the ligand (green compound) will be docked. **B)** Zoom into the BS, showing a superimposition of resultant docked poses.

Most of the docking methods follow the same pipeline. First, they generate multiple protein-ligand poses, and then they assess which of them fits the best by applying scoring functions (Meng et al., 2011). This approach is more straightforward in rigid methods, considering receptor and ligand as frozen structures. These cases try to adapt the given ligand conformation in the target's BS, exploring multiple poses that fit the available space. Early versions of *DOCK* (Kuntz et al., 1982), *FLOG* (Miller et al., 1994), or even protein-protein docking tools such as *FTDOCK* (Gabb et al., 1997) use this rigid approach. This way of thinking is faster but highly limited, as protein-ligand complexes are dynamic systems, and potentially, this can be the key to the binding process. In the ideal scenario, they should consider both receptor and ligand as flexible entities. However, including total flexibility to the receptor is a costly computation that some software avoids, keeping only the receptor rigid while the ligand is treated as a flexible structure. *AutoDock* (Morris et al., 1998), *GOLD* (Jones et al., 1997), *FlexX* (Rarey et al., 1996), and *Glide* (Friesner et al., 2004a; Halgren et al., 2004) are examples of this class of tools, showing a good trade-off between accuracy and time-consumption (S.-Y. Huang, 2018). Closer to including flexibility in the receptor, some docking techniques allow incorporating multiple receptor conformations, joining all of them and

attempting to fit the ligand into the ensembled receptor instead of a single conformation (S.-Y. Huang & Zou, 2007). *FlexXE* (Claußen et al., 2001), and a later version of *DOCK* (Knegtel et al., 1997) are some examples. A different approach that tries to add flexibility to receptors is *induced-fit* docking. They partially include this flexibility by sampling the receptor's side-chains but fixing backbone atoms. *Induced-fit Glide* (Sherman et al., 2006) and *AutoDock4.0* (Morris et al., 2009) are two representative software of this group.

As has been already mentioned, all these methods determine their predictions by applying scoring functions to rank poses, with the principal goal of discerning active from inactive ligands. Some use energetic scores based on force fields; others try to fit empirical data or knowledge-based scorers that have been designed to reproduce experimental structures. Lastly, some establish consensus functions attempting to correct every score's errors (Kitchen et al., 2004).

3.3. Molecular Dynamics

In contrast with the previous method, molecular dynamics (MD) is a complex technique that provides motion to the system to reach its equilibrium properties and gather information related to its kinetics.

Since their first application in the late 1950s (Alder & Wainwright, 1957), molecular dynamics has become one of the most visible and popular simulation methods in recent years (Hollingsworth & Dror, 2018). The theoretical basis behind MD is more straightforward than one could think. Any biomolecular system is surrounded by the solvent, where each atom exerts forces onto others. By solving Newton's equations of motion, new atoms' positions can be predicted as a function of time. In each time step, usually in the range of femtoseconds (fs), all system atoms are potentially propagated, creating 'movies' that show the system's motion. The high degree of precision and flexibility they offer allows us to see complex biomolecular phenomena that are not even observable experimentally, such as significant conformational changes of proteins like allosteric effects, translocations, or even folding pathways, which can be replicated through long MDs (H. Chen et al., 2018; Jang et al., 2002; Renault et al., 2019). One of the main problems of these techniques is the demand for time and computation

resources, usually requiring potent graphics processing units (GPU) or multiple central processing units (CPU). Most biochemical events happen in the range of nanoseconds, microseconds, or even milliseconds. Thus, as these simulations advance femtosecond by femtosecond, extensive computing processes can take several days.

In contrast with VS methods, MD simulations are fascinating to perform H2L, exploring a few modified versions of the hit compound and increasing its potency. Here, the goal is to simulate full binding events to estimate the ligand's affinities (for more details, visit *Section 3.5*). *GROMACS* (Abraham et al., 2015), *NAMD* (Nelson et al., 1996), *Desmond* (Bowers et al., 2006), *AMBER* (Case et al., 2005), *CHARMM* (Brooks et al., 2009), or *OpenMM* (Eastman et al., 2017) are among the most commonly used MD packages. Even though all these methods are relevant and extensively used, we have not applied them in this thesis as they are out of scope due to their high computational cost.

3.4. Monte Carlo Simulations

Monte Carlo comes from a city in Monaco famous for its casinos. This group of methods has obtained this name because of the *roulette* game, an elementary generator of random numbers ("Monte Carlo Method," 2008). The discovery of MC methods was attributed to Stanislaw Ulam and John von Neumann (Metropolis & Ulam, 1949), two researchers that participated in the Manhattan Project to develop the atomic bomb. During an illness period, Ulam played hundreds of solitaire games, coming up with a method to estimate winning probabilities. In 1946, he joined the von Neumann team, and they started to apply this innovative approach to neutron diffusion in fissionable material (Eckhardt, 1987). In this case, they run all their calculations in ENIAC, the first programmable computer. These first studies exemplify how probabilistic and stochastic approaches can solve purely deterministic problems.

The same premise applies to molecular simulations. Herein, multiple protein-ligand configurations are generated from random perturbations of the system (Fichthorn & Weinberg, 1991). These perturbations translate and rotate a few groups of atoms, like ligands. In some methods, perturbations are expanded to the receptor, displacing backbone atoms or even amino acid side-chains. Resultant configurations must be

revised according to a given thermodynamic condition, for example, energy (Paquet & Viktor, 2015). Each transition is assessed by applying a probability-based function such as a Metropolis criterion (Y. Chen & Roux, 2015), accepting or rejecting the new protein-ligand configuration. *MCPRO* (Cabeza de Vaca et al., 2018), *ProtoMS* (*ProtoMS*, n.d.), *Faunus* (Lund et al., 2008), and *PELE* (Borrelli et al., 2005) are prominent examples of MC simulation techniques.

Figure 1.8 illustrates the differences between MD and MC methods in sampling the potential energy landscape. Rather than providing information on time evolution, the last methods collect relevant time-independent conformations that could also be obtained with long MD simulations. In this thesis, we will use *PELE*, the in-house MC software from our group. In the following section, we will focus our explanation on this software.

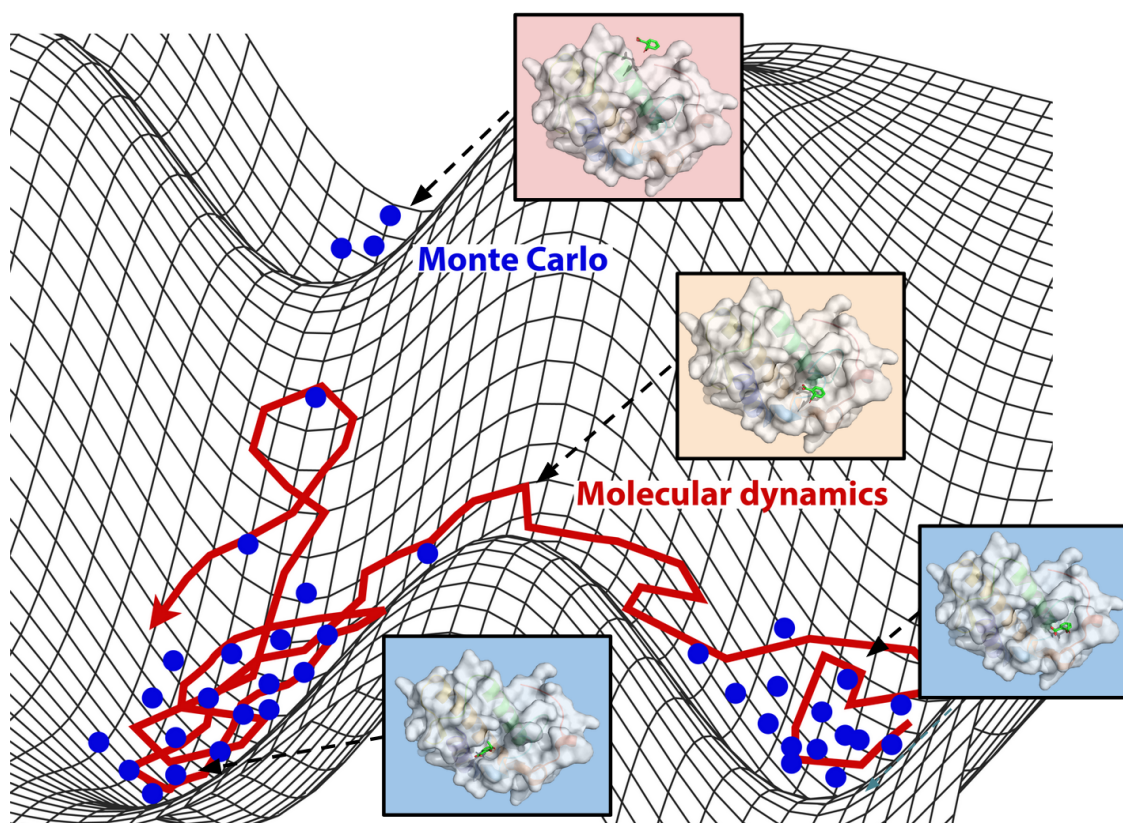


Figure 1.8. Representation of a potential energy surface sampling around the conformational space in MC and MD simulations. Structures show representative binding modes throughout the profile. A modified version of the licensed image by CC BY 4.0.

(**continuation**)(source:https://commons.wikimedia.org/wiki/File:Sampling_in_Monte_Carlo_and_molecular_dynamics.png). Structures generated with PyMOL (Schrödinger & DeLano, 2018).

3.4.1. PELE

PELE (protein energy landscape exploration) was published in 2005 (Borrelli et al., 2005), becoming a novel MC method combining steered perturbations with protein side-chain prediction algorithms and minimization cycles.

The innovation of PELE lies in performing complex movements to increase the acceptance probability instead of doing smooth and simple perturbations as most traditional MC methods do. Thus, contrary to others, PELE includes protein structure prediction algorithms, helping to keep a high acceptance ratio. *Figure 1.9* shows a basic illustration of a PELE step.

Each MC step starts with the *perturbation* phase by randomly rotating and translating the ligand. This ligand is not considered a rigid structure; instead, a fixed core is defined as linked to rotatable side-chains. The ligand displacement is restricted within a user-centered *box* of variable size, limiting its perturbation within this space. Then, multiple protein-ligand conformations are randomly proposed, choosing the one with the lowest energy. Alternatively, the *steering* flag can be set to use the same translation direction in successive steps, increasing the sampling of low probable events. The ligand is perturbed, so the turn of protein structure is after it. Here an *anisotropic network model* (ANM) (Atilgan et al., 2001) is employed on alpha carbons, giving elasticity to the backbone. This is done by applying a minimization with the alpha-carbons constrained to a given “normal mode” combination to favor the reorganization of backbone atoms.

Once the perturbation phase has finished, the *relaxation* stage begins. Side-chain prediction techniques are applied around the ligand location (typically 6 Å) to optimize interactions (such as π -stacking, HBs, electrostatic) by using Jacobson’s et al. algorithm (Jacobson et al., 2002). Then, the whole system is relaxed with minimization to increase the acceptance probability, computing and comparing the final energy with the initial pose to accept/reject the step. When the energy of the new configuration is lower than the original one, the step will be immediately accepted (*Equation 1.10*). However, if this

is not the case, a *Metropolis criterion* (Metropolis & Ulam, 1949) is employed to compare the result with a random number (*Equation 1.11*). By doing so, non-favorable configurations can also be accepted.

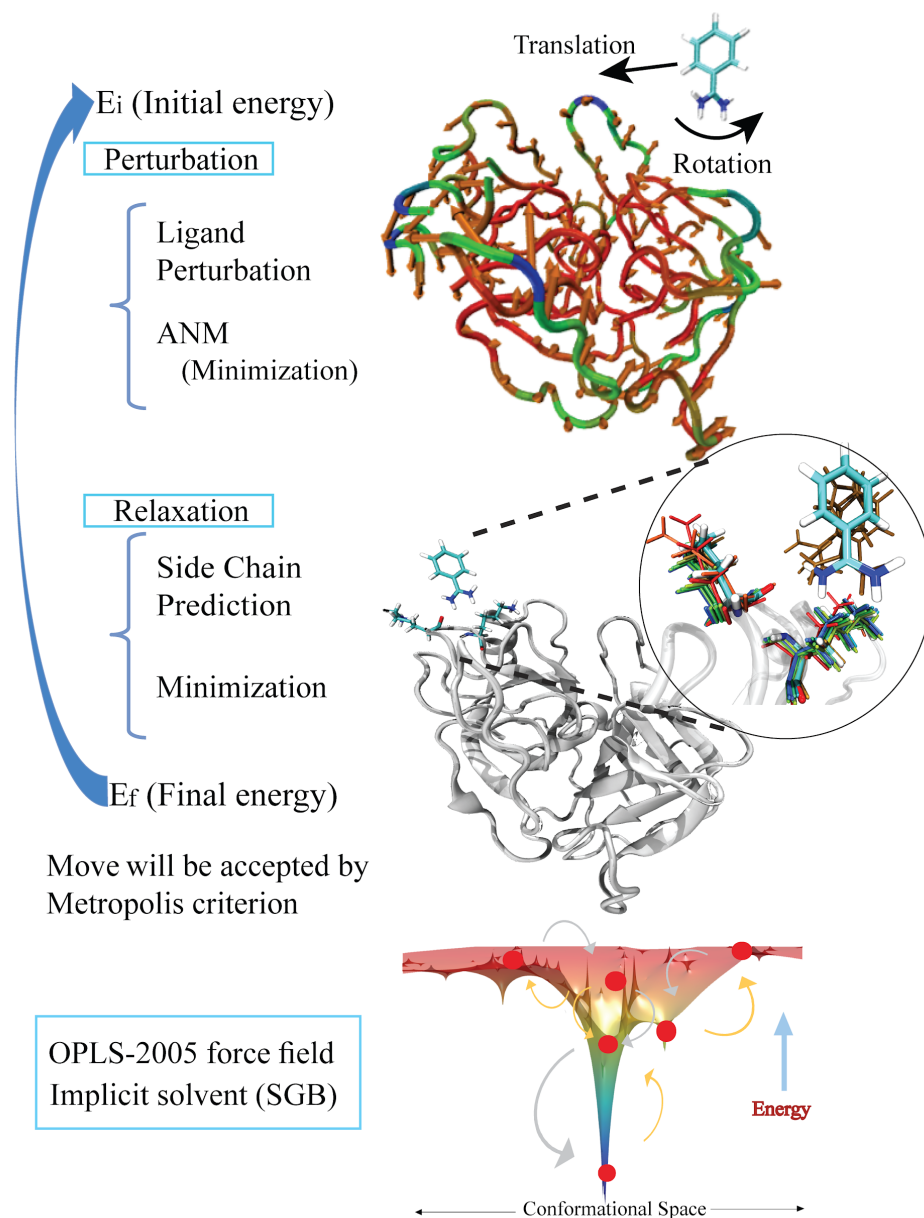


Figure 1.9. Schematic representation of the different stages within a PELE step. Image author: Image author: Ryoji Takahashi.

Classically, PELE could only use OPLS2005 (Banks et al., 2005; Kaminski et al., 2001) and AMBER99sbBSC0 (Pérez et al., 2007) force fields, but in the latest version, the OpenForceField initiative (Qiu et al., n.d.) has been incorporated to the list. OBC

(Onufriev et al., 2004) and VDGBNP (Zhu et al., 2007) are the two implicit solvent models that can be used; however, a few essential explicit water molecules can be placed in the simulation.

$$(1.10) \quad \Delta V < 0 \qquad (1.11) \quad e^{(-\Delta V/K_{\beta}T)} < R$$

Equations 1.10 and 1.11. The probability of acceptance within the PELE algorithm. First and second criterion. ΔV = difference of potential energy between final and initial state; K_{β} = Boltzmann constant; T = simulation's temperature; R = random number [0,1].

In the last few years, multiple external packages and improvements to the PELE algorithm have been incorporated. In 2017 Lecina et al. developed AdaptivePELE, a method to enhance simulation sampling and reduce the computation demand based on clustering and spawning algorithms (Lecina et al., 2017). In 2019 and 2020, Gilabert et al. created *PELE-MSM* to estimate absolute binding free energies by applying Markov State Models (Gilabert et al., 2019, 2020). Then, in the same year, a new algorithm to perturb explicit waters within PELE MC steps was released by Municoy et al. called *aquaPELE* (Municoy et al., 2020). They are examples of newly developed methods around this software.

PELE has been extensively used in multiple fields, such as enzyme engineering studies (Alejaldre et al., 2020; Gentil et al., 2020; Martínez-Martínez et al., 2018; Roda et al., 2020; Santiago et al., 2018), drug discovery of small molecules (Carlson et al., 2016; Gilabert et al., 2018; Grebner et al., 2016; Saen-Oon et al., 2019), or even to predict protein-protein interactions (Amengual-Rigo et al., 2020).

3.4.2. AquaPELE

In September of 2020, Municoy et al. published the aquaPELE algorithm. The strategy behind this new implementation is to introduce a new waterMC step routine in the PELE perturbation phase to explore explicit water molecules within the BS. Perturbable waters can only move within a user-defined water region (WR), constant in all the simulations. Additionally, users can specify the number of waters to move and define subsets to be perturbed together in each MC step.

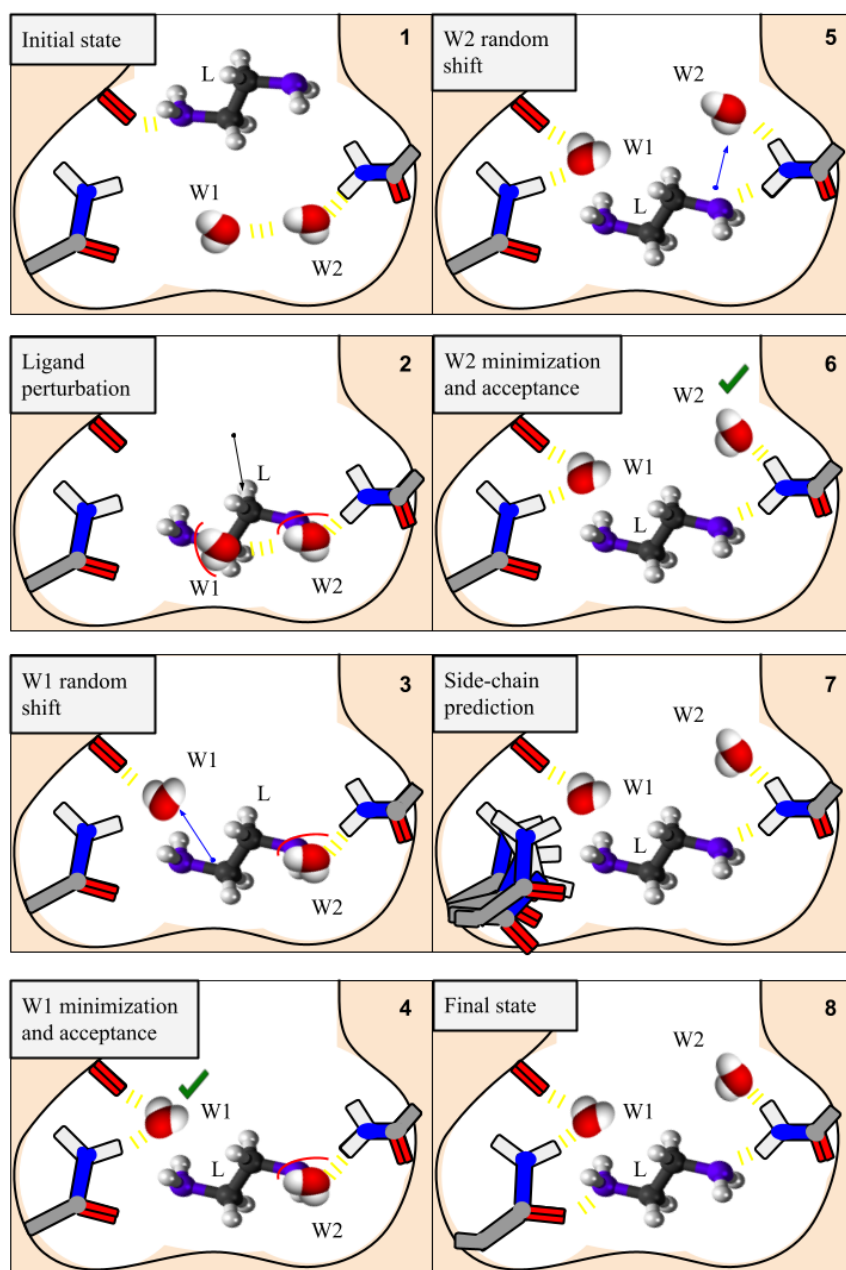


Figure 1.10. Illustration of aquAPELE MC step where two water molecules (W1 and W2) are perturbed. First (panel 2), the ligand (L) is translated and rotated from its original position (black arrow), ignoring clashes (red lines) with W1 and W2. In the following steps (panel 3-6), water molecules are perturbed (blue arrows), minimized, and accepted (green tick), ending with the side-chain prediction (panel 7) that will lead to the final state (panel 8). Notice that a few protein amino acids have been drawn within the protein cavity: two glutamine side-chains and the oxygen of a backbone carboxylic group. Besides, hydrogen bond interactions are represented as yellow dashed lines.

The general view of the content within an aquaPELE step is summarized in *Figure 1.10*, where the central core of the new algorithm is represented on panels 3-6. Here, there are two explicit waters (*W1* and *W2*) within the subset to be perturbed. In the ligand perturbation (panel 2), the ligand (*L*) will be randomly translated and rotated, ignoring clashes with water molecules. In the same way, the energy computation does not include these waters, and ligand conformations are evaluated independently from water positions. After the ligand perturbation, the waterMC step is applied, starting by individually perturbing the *W1* similarly to the ligand (panel 3). Once a new position is found, the local region around the water molecule is minimized through a truncated Newton algorithm to get the optimal orientation. This new configuration is assessed by employing a Metropolis criterion to choose whether to keep the pose (panel 4) or move back to the initial state. If more than one perturbable water is defined, the same perturbation-minimization-Metropolis procedure is executed individually for each extra molecule (*W2* in *Figure 1.10*, panels 5-6). Finally, after the relaxation stage, the step is completed, including side-chain prediction and the whole system minimization (panels 7-8).

AquaPELE has been extensively tested in multiple hydrate systems, including its ability to relocate interfacial water molecules that lead to improvements in free energy, showing a perfect fitting between experimental observations and reported results (Municoy et al., 2020).

3.5. Estimating binding free energies

We have reviewed molecular modeling techniques to reproduce ligand-protein configurations by geometrical fitting such as docking or system thermodynamics like in MD. The information that all these methods provide is interesting to understand the key elements that contribute to the correct binding mode of the ligand or their dynamics, but sometimes it is not enough when comparing different candidates. At this point, we need an accurate metric to estimate the binding strength or inhibition potency, then most of the previous techniques are used to predict binding free energies.

3.5.1. Docking scores

Docking aims to identify correct protein-ligand configurations for the given system quickly. Then, they are extensively used to enrich libraries of compounds by taking the top score ligands. As mentioned in *Section 3.2.1*, several scoring functions exist to assess the generated configurations. Even though they are good at predicting binding modes, they perform better in one system than in others. Their more superficial estimations generally fail to distinguish between ligands with no significant differences in free binding energies (Gilson & Zhou, 2007).

Some have been designed to identify nanomolar compounds instead of micromolar (the usual goal of VS) or even work with specific target groups. Hence, their performance depends on how they have been trained or optimized (Cole et al., 2005).

3.5.2. End-point methods

In the trade-off between computation cost and accuracy, this group of techniques scales a step forward by introducing protein-ligand flexibility. The idea behind end-point free energy estimations is to use energetic differences between the bound and unbound states. *Molecular mechanics generalized Born surface area* and *Poisson-Boltzmann surface area* (MMGBSA and MMPBSA) are widely used techniques within this group (Kollman et al., 2000). Usually, short MD or MC simulations are needed to sample and estimate the dynamic properties of the system. For each state, the energy is computed by applying *Equation 1.12*, finally estimating the free whole system's energy, subtracting the mean states of unbound ligand and unbound receptor to their bound complex (*Equation 1.13*).

All states can be obtained from a single complex-ligand simulation by deleting the appropriate atoms to reproduce the ligand-free and protein-free states. Not so common, by running three individual simulations (only ligand, only receptor, and ligand-receptor) (Genheden & Ryde, 2015).

$$(1.12) \quad G = E_{bonded} + E_{electrostatic} + E_{vdw} + G_{polar} + G_{nonpolar} - TS$$

Equation 1.12. G = state free energy; E_{bond} = energy bonded terms; $E_{electrostatic}$ = energy electrostatic terms; E_{vdw} = energy VDW terms; G_{polar} = polar solvation energy; $G_{nonpolar}$ = non-polar solvation energy; T = temperature; S = entropy estimation.

Linear interaction energy (LIE) is another strategy that, contrary to MMGBSA/MMPBSA, computes the interaction energies between the ligand and its environment. Only simulations with the complex and the free ligand are performed (Åqvist et al., 1994).

Regarding computational cost, end-point methods are more expensive than docking tools but less costly than *alchemical methods* (explained in the next section).

$$(1.13) \quad \Delta G_{binding} = \langle G_{RL} - G_L - G_R \rangle$$

Equation 1.13. $\Delta G_{binding}$ = binding free energy; G_{RL} = free energy of ligand-receptor state; G_L = free energy of free ligand state; G_R = free energy of free receptor state.

3.5.3. Alchemical methods

Thermodynamically, absolute binding free energy is the difference between two states: bound and unbound (ligand and protein in the solvent). This energy is independent of their path, only depending on the initial and final stages. Thus, computing rigorous ΔG requires converged MD or MC simulations for each state, which is an extraordinarily long process. Luckily, when comparing congeneric series of ligands, they usually show slight differences between them (a few atoms or small moieties), and for these cases, computing the absolute ΔG is not needed. Instead, *relative binding free energies* ($\Delta\Delta G$) can estimate differences between similar ligands (Gilson & Zhou, 2007).

Kirkwood (Kirkwood, 1935) and Zwanzig (Zwanzig, 1954) demonstrated that the $\Delta\Delta G$ could be computed with *alchemical methods*. *Thermodynamics integration* (TI) (Kirkwood, 1935) and *free energy perturbation* (FEP) (Zwanzig, 1954) are the two main techniques within this group. *Figure 1.11* summarizes the thermodynamic cycles that fundamentals these methods, relying on the idea that the conversion of a ligand A

to a ligand B can be done by connecting multiple non-physical transition microstates. As both ligands' initial and final states should be similar, the alchemical transition (horizontal) would be faster than simulating the unbound to bound transition (vertical).

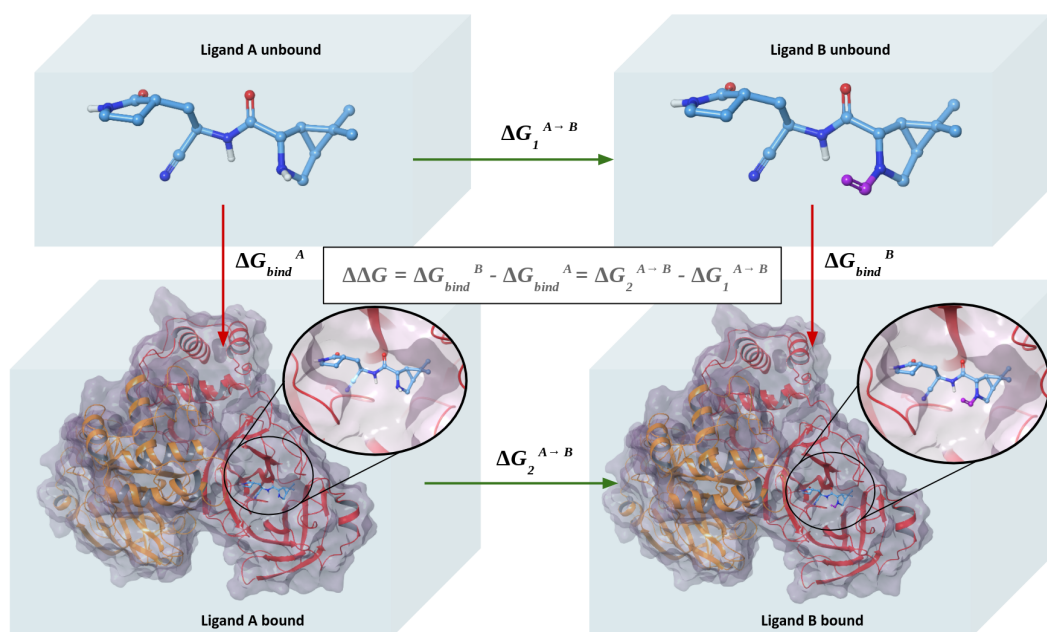


Figure 1.11. Illustration of the thermodynamic cycle employed in alchemical methods. Blue boxes around ligands and complexes represent the solvent. Arrow colors emphasize the time consumption of the process, being green fast and red slow. Then, the relative binding free energy ($\Delta\Delta G$) can be estimated by computing the difference of binding free energy between $\Delta G_2^{A\rightarrow B}$ and $\Delta G_1^{A\rightarrow B}$ (horizontally), which tends to be faster than doing it following the vertical path.

These *alchemical* microstates are associated with *lambda* (λ) coefficients that go from λ zero (ligand A) to λ one (ligand B) with intermediate states (f.e $\lambda=0.1$, $\lambda=0.2$, $\lambda=0.3$,...) where the properties from one ligand are slowly transformed to another (Cournia et al., 2017).

Jorgensen group was the first to apply FEP onto MC simulations (Cabeza de Vaca et al., 2019; Jorgensen & Ravimohan, 1985) to compute relative binding free energies. Since that moment, FEP calculations have improved considerably. In 2015, a new protocol called FEP+ (L. Wang et al., 2015) was published. They primarily tested more than 200

ligands, determining good performance and errors around 1.0 kcal/mol, becoming one of the gold-standard methods used for H2L.

Although the recent expansion of FEP techniques, they show some limitations. For instance, they require high-quality input structures, usually coming from co-crystallized ligands of the series of interest. Additionally, they can only transform one ligand into another similar, avoiding adding big groups of atoms or even modifying the net charge. The number of computational resources and time can also be an obstacle. They usually use GPUs instead of CPUs and require a moderately long simulation time (MDs of 5-20 nanoseconds length per lambda), often not enough to converge in systems with large-scale protein movements (Cournia et al., 2017), requiring more time for complex rearrangements. Thanks to recent advances in improving sampling protocols with FEP+, one perturbation would necessitate around 7.5-9h in 4 x GPUs GTX 1080Ti (30-36h/GPU) in a middle-size system (Fratev & Sirimulla, 2019). However, the high expenses in computational resources and licenses are still unaffordable for a big part of the community. Then, even though they succeed in predicting $\Delta\Delta G$, all these limitations restrict their use to particular cases.

3.6. Machine learning

Machine learning (ML) is a game-changer that has presented simple solutions to complex problems in many fields. These methods aim at turning computer behavior more similar to human brains; they analyze large amounts of data to learn and perform any task autonomously. This data usually contains information that can be processed and interpreted for this algorithm, detecting patterns and relations that humans cannot even notice. Living in the *big data* era, we can now access the large volume of public bioactivity data that are particularly valuable for these techniques (Richter & Ecker, 2015). Public repositories such as PubChem (Kim et al., 2020), ZINC (Irwin & Shoichet, 2005), DrugBank (Wishart et al., 2006), or ChEMBL (Gaulton et al., 2012) offer especially valuable data. By exploiting this information, the drug discovery field has also benefited from the development and application of ML techniques (Dara et al., 2021; Vamathevan et al., 2019).

The fine ML model selection will depend on multiple parameters: the dataset size, performance, complexity, and the nature of the problem to be solved (Metwalli, 2020; Valdarrama, 2021). They are roughly classified as unsupervised or supervised learning. The first refers to clustering-based methods that do not require assigned tags in the data, while the latter demands continuous values (regressors) or categorical labels (classifiers). Examples of simple ML algorithms are Support Vector Machines, Naive Bayes, or Random Forest. Deep learning (DL) and Artificial Neural Networks (ANN) offer a higher complexity level within the ML field. When using layers of neurons, we refer to ANN, which mathematically activates or inactivates them to “think” like human brains. When these networks incorporate multiple layers (more than three), we refer to DL (Mishra & Gupta, 2017). DL and ANN are probably the most popular subbranches of ML. However, they require more computational resources and more significant amounts of data than simpler ML models.

Any ML methodology must follow the following steps: 1) collect data; 2) read numeric descriptors; 3) select the best variables; 4) train, and 5) validate the model (Fernandez-Lozano et al., 2016). All input data (e.g., images, molecules, human behavior) must be converted to mathematical values readable by algorithms. Molecular information can be represented differently depending on the complexity level. Simple physicochemical descriptors (number of atoms, MW, etc.), 2D structures that encode chemical, size, and shape information, or complex protein-ligand interaction maps are examples of data that can be used as molecular descriptors (Carracedo-Reboredo et al., 2021).

Drug discovery and molecular modeling techniques have incorporated ML algorithms in recent years. Some docking tools include ML-based models, trained to combine energy-based predictions, geometry, and chemical information in a single scoring function (Kinnings et al., 2011). Alternatively, they can be trained to design target-specific inhibitors (Ekins et al., 2017). Moreover, the application of DL to quantitative structure-activity relationship (QSAR) studies has notably accelerated docking-based virtual screenings to screen ultra-large datasets with more than 1 billion compounds (Gentile et al., 2020).

Other examples of DL models are *generative models*. This group of methods can learn from large drug databases, understand the chemical space, and generate *de novo* drug-like compounds (Bian & Xie, 2021). General models will create non-specific chemical entities. However, the model can be fine-tuned through transfer learning to focus the generation on the desired hit-target complex (Chenjing Cai, Shiwei Wang, Youjun Xu, Weilin Zhang, Ke Tang, Qi Ouyang, Luhua Lai, and Jianfeng Pei, 2020). For example, *Variational autoencoders* (VAE) (Kingma & Welling, 2013) are generative models where a neuron layer encodes (encoder) compounds information in a continuous latent space. Then, another network decodes (decoder) the data to retrieve new drug-like molecules different from the trained ones (Joo et al., 2020). A schematic illustration of its structure is shown in *Figure 1.12*. After training, this latent space can be manipulated, enabling transfer learning inputs to reward any desired property.

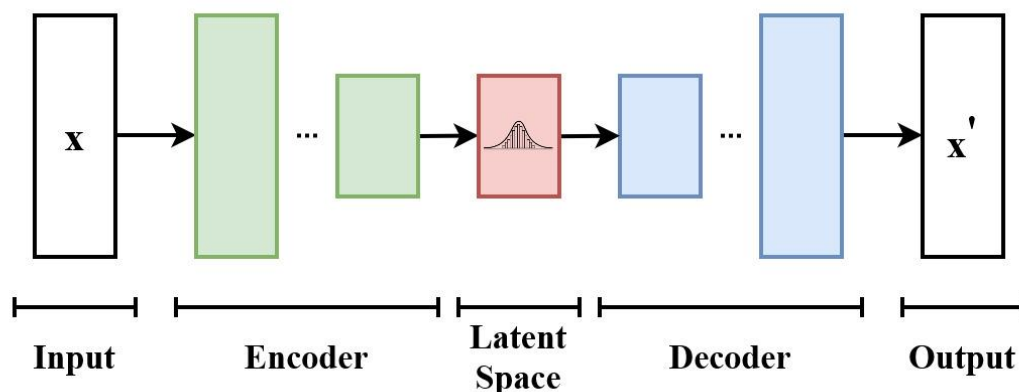


Figure 1.12. VAE structure. Image created by EugenioTL under a license CC BY 4.0
(source: https://commons.wikimedia.org/wiki/File:VAE_Basic.jpg)

AlphaFold (Jumper et al., 2021; Senior et al., 2020) and RoseTTAFold (Baek et al., 2021) are two famous DL applications that solved one of the most extensive problems in computational biology of the century, predicting protein folding from its primary amino acid sequence.

ML has demonstrated to be an excellent versatile problem-solver in many fields, offering a vast toolbox of robust algorithms. Moderately incorporating them in real drug discovery frameworks will offer new advantages in designing novel compounds.

3.7. Ligand growing

In H2L studies, most efforts rely on optimizing the ligand potency (J. P. Hughes et al., 2011). Medicinal chemists propose modifications of the initial hit to gain extra interactions with the receptor. These hits are usually ‘*grown*’ by including new R-groups or decorators, helping to reach HBs, electrostatic or hydrophobic interactions that favor the new lead compound (Brown & Boström, 2018).

Ligand growing strategies are prevalent within the fragment-based drug design (FBDD). The initial scaffold is a hit compound of low molecular MW with residual potency, so-called *fragments*. These compounds are linked, merged, or grown to increase their potency to the nM range. Here it is crucial to obtain experimental 3D information, commonly coming from X-rays, to choose the proper building blocks for FBDD (Erlanson et al., 2016; Murray & Rees, 2009).

As mentioned in this thesis, molecular modeling technologies provide computational solutions to drug design problems. This fact is not different when exploring R-groups around a molecular *core* or *scaffold*. Here, instead of screening large libraries with millions of compounds, *fragment libraries* are frequently used; contrary to regular high throughput virtual screening (HTVS), the average size for a fragment library is far smaller (usually <1000 fragments) (Trevizani et al., 2017). A wide variety of techniques can be applied to handle this problem. Moreover, researchers typically customize libraries’ content according to their purpose. They can include fragments to favor a specific synthetic pathway or be enriched with certain functional groups (carboxylic acids, amines), or a property (electrophilicity, hydrophobicity), or optimizing libraries against a concrete target such as kinases (Kidd et al., 2018). Thus, libraries’ right design and application onto the proper hit are critical in fragment growing screenings.

A quick and straightforward strategy is combining different reactive fragments around the scaffold X-ray crystal structure, building hundreds of new molecules to dock them later. BOMB (Jorgensen, 2009) or CombiGlide from Schrödinger (an extension from Glide) can handle this approach (Schrödinger, 2019).

The fragment selection can be unguided or directed by structural-based rules, such as pharmacophoric approaches, or by applying geometrical constraints to the scaffold (Schulz et al., 2018). Similar strategies such as the abovementioned generative models have become more prevalent in recent years (Vamathevan et al., 2019). These methods are beneficial to enriching fragment libraries with more prone to bind compounds, but all of them are usually assessed by docking at the end of the day.

An additional strategy to create new ligands is applying *de novo* algorithms that start from a seed core/scaffold already placed in the BS instead of building molecules from scratch. Plenty of programs offer this kind of strategy. Among the most widely used software, we find *LUDI* (Böhm, 1992), *LigBuilder 2.0* (Yuan et al., 2011), and *Autogrow 3.0* (Durrant et al., 2013). Others focus their efforts on tracking the synthetic feasibility, such as *SYNOPSIS* (Vinkers et al., 2003) and *NAOMINext* (Sommer et al., 2019). Most of them consider a rigid representation of the BS, which severely affects the generation of high-quality structural models and binding energy predictions for the grown derivatives (Schneider & Fechner, 2005).

Decorating these chemical scaffolds is not an easy task. Despite having rich and accurate 3D structures, the role of the receptor motion and the solvent are vital factors that can interfere with the predictions of this group of methods. The typical case that exemplifies this phenomenon is the epidermal growth factor receptor (EGFR). The first-generation ligand *gefitinib* (Rawluk & Waller, 2018) shows good inhibition levels, but after decorating it, they improved it by creating the second-generation ligand *lapatinib* (Voigtlaender et al., 2018). Looking at 3D structures, the bound conformation of lapatinib was audibly distinct from gefitinib. Unexpectedly, the addition of a bulky group led to the opening of a cryptic sub-cavity, not detectable in the unbound state.

As exemplified above, incorporating receptor flexibility can be essential when expanding R-groups. Several methods have been developed with this goal in mind, *LEA3D* (Douquet et al., 2005), *SkelGen* (Dean et al., 2006), and *OpenGrowth* (Chéron et al., 2016), where ligand and local side chains are sampled to take into account the dynamics of the BS.

All previously mentioned methods show a common factor. They can handle, propose or select R-group modifications according to some rules. Despite doing so, they usually require to be coupled to molecular docking or simulations to generate proper geometries or affinities estimations. Even though rigid/flexible docking or end-point approximations have not been designed for this purpose, they can be applied afterward to predict binding modes and binding energies. More rigorously, alchemical methods can mutate and expand small R-groups, providing accurate and reliable relative binding free energies. However, as we have seen in *Section 3.5.3*, these techniques show some limitations regarding computational cost and can only handle a series of similar ligands.

Thus, after reviewing various ligand growing approaches, there is an empty gap in the literature between the expensive alchemical free energy methods and the quick standard docking-based techniques. We hypothesize that the PELE MC method could offer an excellent accuracy-cost balance for exploring middle-size libraries (100-1.000) of compounds. With this idea in mind, during my Master's thesis, we started the development of a novel fragment growing strategy couple to PELE, called *FragPELE*, whose first fully functional version was released during the first year of the Ph.D.

Chapter 2. Objectives

This thesis aims to develop and apply new computational strategies focused on improving the accuracy-time consumption trade-off for hit-to-lead studies. We also aimed to collaborate in real drug design projects involving both stages, VS and H2L, put our methods to test, and design new strategies to face prospective drug discovery projects. These objectives could be split into smaller ones:

- I. *Complete the development of FragPELE; our novel ligand growing method focused on hit-to-lead studies. The technique must be validated through retrospective studies to fill the empty gap between fast docking-based techniques and the expensive alchemical methods.*
- II. *Use and improve the previous method by detecting and solving errors, adding additional functionalities, automatized, user-friendly protocols, and coupling them with fragment libraries.*
- III. *Collaborate in real industrial and academic drug design projects, proposing new virtual screening strategies and also applying our recently developed ligand growing software. If it is possible, contribute to the finding of a new hit or lead compound.*

The following chapters will show the main results obtained from this thesis, also describing all the methodology employed and the newly developed algorithms.

Chapter 3 will first introduce *FragPELE*, the new ligand growing method designed to tackle hit-to-lead studies. It will follow the validation studies of this initial software, comparing its performance with other gold-standard methods (Glide SP, InducedFit-Glide, MMGBSA, and FEP+). Second, we will assess the ability of *FragPELE* to displace buried waters and score compounds in hydrated systems, combining this technology with the recently released *aquaPELE* (Municoy et al., 2020).

In *Chapter 4*, we will show the methods used and results of the *SilicoDerm* project, in this case, an industrial collaboration with *Almirall* company. For confidentiality

purposes, compounds and structures of this section will be anonymized, focusing the report on the methodology.

Chapter 3. FragPELE: dynamic ligand growing within a binding site

H2L studies rely on hit optimization to improve binding affinities for their biological target. As discussed in Chapter 1 (*Section 3.5 and 3.7*), providing fast and accurate answers to induced-fit protein-ligand binding remains challenging. Thus, we introduce FragPELE, a promising strategy to explore R-group expansion while tracking BS induced-fit effects with acceptable time consumption. The development of this technique was initialized in my Master's thesis, but it was finally published in the JCIM journal (Perez et al., 2020) in my first Ph.D. year. *Section 1* will include and adapt some passages quoted from this article.

1. FragPELE algorithm

1.1. The method

FragPELE is a ligand growing algorithm that slightly expands variable-sized fragments onto the user-selected site. Protein motion and side-chains are sampled during the process to adapt the cavity to the newly added moiety. This effect is produced by employing a Python layer that orchestrates on top of PELE software (for more information, visit *Section 3.4.1* from *Chapter 1*) by combining ProDy (Bakan et al., 2011), BioPython (Cock et al., 2009), and pandas (McKinney & Others, 2011) external libraries.

The design of the algorithm relies on the idea of automatically growing one or several R-groups onto a well-defined protein-hit complex (preferably coming from an X-ray structure). Based on AdaptivePELE (Lecina et al., 2017), this method also links consecutive PELE simulations, called *growing steps* (GS). Similarly but simpler to alchemical methods (Cournia et al., 2017), GS represents transition *microstates* where the original ligand is slowly transformed into another. At the same time, efficient system

rearrangements are done by incorporating normal modes to displace the backbone, side-chain sampling, and minimization procedures.

A global view of the strategy followed by the FragPELE algorithm is shown in *Figure 3.1*, which involves several parts: preparation, fragment linkage, fragment reduction, fragment growing, and sampling/scoring.

Preparation. As mentioned, FragPELE intends to grow fragments onto the same hit, scaffold, or core molecule. Users must provide PDB files containing reliable protein-ligand complexes, previously solving protonation and any structural issue (missing atoms, loops, multiple occupancies). Ions, cofactors, and explicit water molecules can be added as part of the receptor. In the same way, users must include fragment molecules to grow (fully protonated) in separate PDB files. Finally, the user can fill a tabular input file with instructions on which fragments to connect to which growing site and which atoms. In the initial version, the considered growing sites are only the heavy atoms bound to terminal hydrogen that will be replaced for an entire fragment.

Fragment linkage. This stage links the given fragment with the chemical scaffold onto the user-specified growing site. The previously mentioned tabular file contains the pairs of atoms to connect (one from the scaffold and the other from the fragment). This linkage is performed by obtaining the bond vectors formed between the heavy atom and its bound hydrogen for both fragment and scaffold to join them later and superimpose its coordinates (see *Figure 3.2, A to B*). Subsequently, hydrogen atoms will be erased to create a covalent bond. If any of the chosen heavy atoms present more than one hydrogen bond, the user can select which one to use, controlling then the enantiomeric form of the new ligand. If not specified, the algorithm automatically selects the hydrogen of the scaffold that will produce fewer clashes with the protein. Alternatively, for these cases where the user wants to grow something on a heavy atom without any hydrogen bond, this atom can be easily replaced manually.

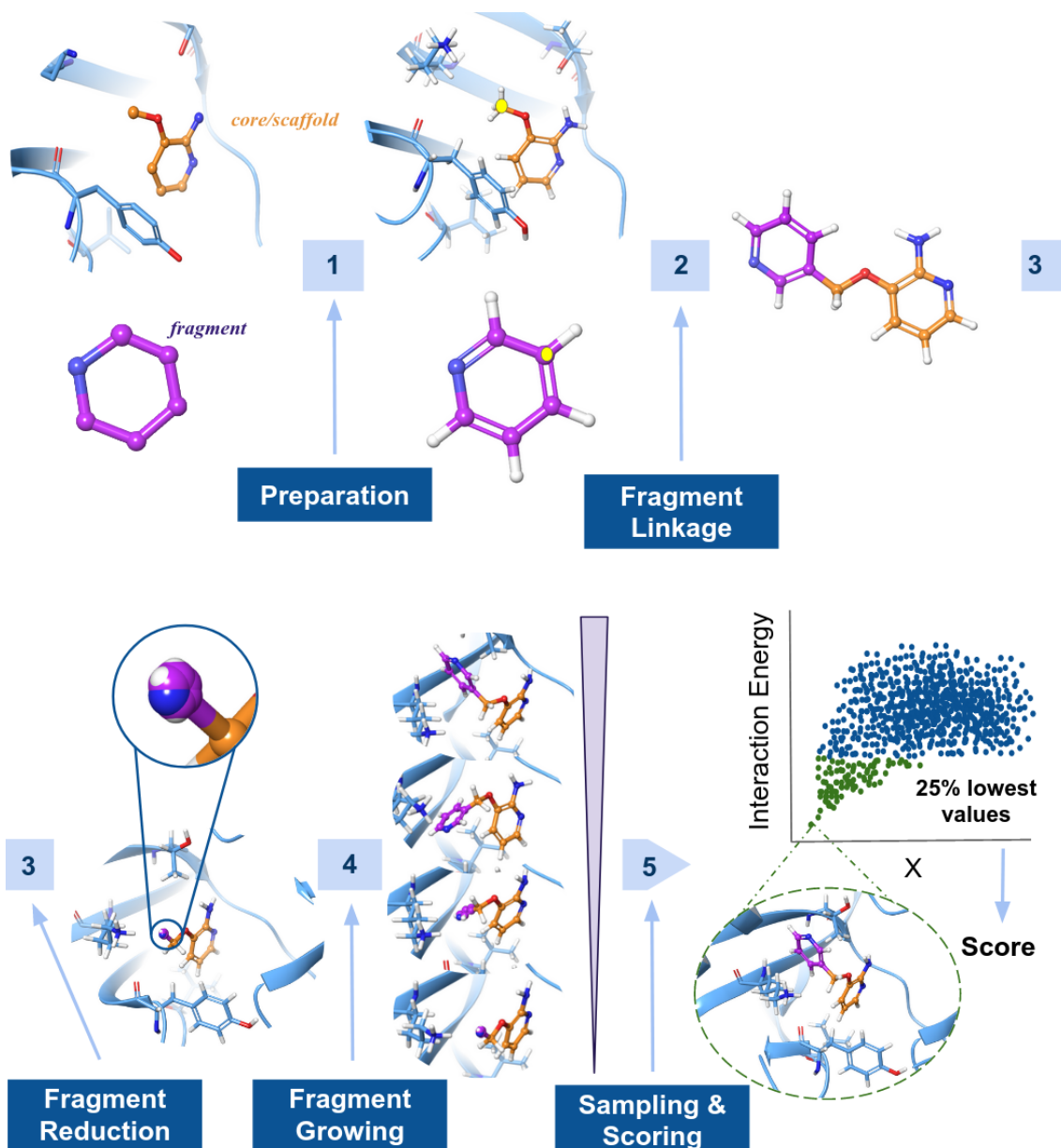


Figure 3.1. Overview of FragPELE pipeline. The yellow circle underlines the heavy atoms connected in the fragment linkage. Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

It will not be surprising that the abrupt addition of the fragment would lead to intramolecular clashes with the core of the molecule. Under this scenario, the algorithm includes an intramolecular clashes detector that rotates 10 degrees along the new bond axis until locating a more favorable position without collisions. If it does not find any favorable solution, the resolution is increased to 1 degree, raising an error after turning the 360°.

The new fragment is linked in the input complex file at the end of this stage. Besides, in another PDB file, the new ligand is stored to use the topology later to calculate the parameters of the force field.

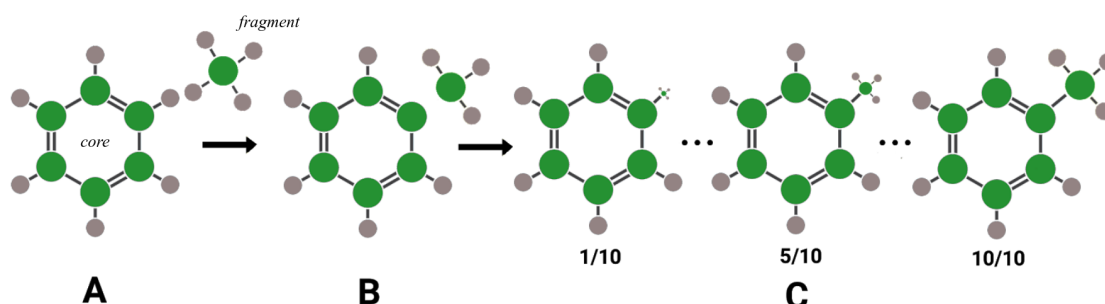


Figure 3.2. Representation of the growth of a methyl onto a benzene core. Carbons are depicted in green and hydrogens in gray. Initially, (A) the fragment and the core are separated as two isolated entities; (B) the fragment is placed by aligning and erasing the hydrogens that create the new bond. Finally, (C) the fragment is miniaturized and ready to be grown along with the fragment growing phase. Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

Fragment reduction. This step reduces some force field parameters (FFP) for the fragment's atoms and its geometry in the protein-ligand complex. Going into detail, the specific FFP being reduced are the L-J radii (σ), partial charges (q), and the equilibrium bonding distance (r_{eq}) for each bond with fragment atoms (see *Section 3.1* from *Chapter 1*). The calculation is differently applied for each parameter. The initial L-J radii (σ_o) and partial charges (q_o) are computed following *Equations 3.1* and *3.2*. Notice that this reduction in non-bonded parameters aims to keep the whole fragment with a hydrogen-like volume and distribute its charge (qH) among the fragment atoms. For the initial equilibrium bonding distances (r_o), *Equation 3.3* is applied for all fragment's bonds, except the one which connects with the scaffold (r_{link}), using *Equation 3.4* instead.

$$(3.1) \quad \sigma_o = \frac{\sigma}{L+1} \qquad (3.2) \quad q_o = \frac{qH}{(L+1)N}$$

Equations 3.1 and 3.2. Obtention of initial non-bonding parameters. σ_o = initial L-J radii; q_o = initial partial charge; σ = original L-J radii; qH = hydrogen replaced charge; L = total GS to be performed; N = number of fragment's atoms.

$$(3.3) \quad r_o = \frac{r}{L+1} \qquad (3.4) \quad r_{link} = rH + \frac{r-rH}{L+1}$$

Equations 3.3 and 3.4. obtention of initial bonding parameters. r_o = initial equilibrium bonding distance; r = original equilibrium bonding distance; L = total GS; r_{link} = equilibrium bonding distance involving scaffold-fragment linking atoms; rH = equilibrium distance between hydrogen replaced and scaffold atom.

The distance between the miniaturized fragment and the core atoms is expanded through all equations compared to other bonds to avoid intramolecular clashes when initializing the simulation. At the same time, the charge of the replaced hydrogen atom is spread and proportionally reduced by the number of GS to imitate the electrostatics of the original hydrogen atom. Notice that this charge is spread and reduced to avoid artifacts. Accordingly, a default value of 10 GS is chosen, which provides a good balance between smooth and fast-growing, and avoids large perturbations in the hydrogen-like volume caused by too low or too high L . Angles and dihedrals parameters are kept constant.

Before starting the next stage, the fragment atom coordinates are geometrically modified by reducing all bond distances (vector modules), also following *Equation 3.3*.

Fragment growing. By combining the concepts of enhanced sampling introduced by AdaptivePELE (Lecina et al., 2017) and alchemical perturbative methods (Zwanzig, 1954), the new fragment is grown in a series of GS. Multiple parallel PELE simulations are executed at each GS after progressively increasing the FFP (miniaturized in the previous step). An initial GS 0 is executed before modifying the FFP from the *fragment reduction* stage to adapt the system. Afterward, they are linearly expanded in the next GS, applying *Equations 3.5 and 3.6*. *Figure 3.3* shows a visual example of this schema.

$$(3.5) \quad \lambda = \frac{1}{L-S+1} \qquad (3.6) \quad X_{step} = \lambda X$$

Equations 3.5 and 3.6. FFP linear increase. L = total GS; S = current GS; X_{step} = any FFP for the current GS; X = any original FFP.

After each GS, the resulting structures are clustered by protein-ligand contact maps (using a distance threshold between any atom from an amino acid and the ligand of 3 Å) using a *k-means* algorithm to enhance the variability of poses. Five clusters are generated by default, and the poses with the lowest interaction energy are spawned to initialize the next GS. PELE calculates the interaction energy values by subtracting the energy of the ligand and the energy of the receptor, both isolated, to the energy of its complex: $E(PL) - E(L) - E(P)$. MM models (see *Section 3.1* from *Chapter 1*) are the basis of this energy function, considering bonding, non-bonding, and solvation terms (*Equation 3.7*). In summary, the *fragment growing* is an iterative process where the output of the previous GS is the input for the next one, chaining then successive simulation rounds until the size of the fragment reaches the original values.

$$(3.7) \quad E = E_{bond} + E_{angle} + E_{torsion} + \\ E_{improper\ torsion} + E_{vdw} + E_{ele} + \\ \Delta G_{solv,pol} + \Delta G_{solv,penalty} + E_{constraints}$$

Equation 3.7. PELE energy function. E = energy; ΔG = free energy.

Sampling simulation. Once the fragment is fully grown (at the end of the last GS), a more extended simulation of 20 MC steps is executed to better explore the new protein-ligand conformations and score them. This general score is the average of the 25 percent poses with the lowest interaction energies from the whole simulation.

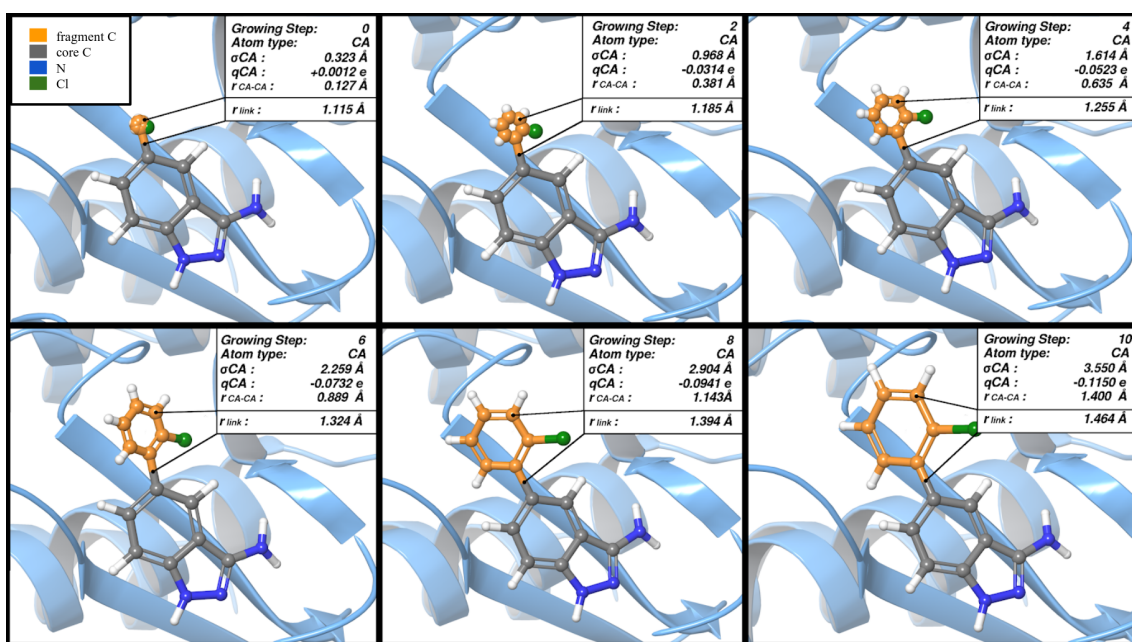


Figure 3.3. Example of chlorophenyl fragment growing with FragPELE onto an amino-indazole in 10 GS. The inset shows structures and FFP values for each panel's aromatic carbon (CA). σ = L-J radii; q = partial charge; r = equilibrium bonding distance. Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

FragPELE uses OPLS2005 (Banks et al., 2005; Kaminski et al., 2001) force field and VDGBNP (Zhu et al., 2007) implicit solvent model. The standard protocol consists of 10 GS (+ 1 of equilibration, called GS 0), containing each 47 independent MC simulations of 6 PELE steps (282 poses/GS) and a final sampling simulation of 20 steps (940 poses). These simulations are configured with low translations (0.05-0.10 Å) and rotations (0.02-0.05 radians), allowing only displacements within a spherical box of 4 Å around the initial center of mass of the ligand to perturb the system slightly. Additionally, positional constraints onto explicit waters (when needed) are mandatory in this version.

Two main tests have been carried out to validate the method. First, we evaluated the capacity to reproduce X-ray crystal structures accurately. Second, we also studied the ability of FragPELE to predict experimental binding affinities for congeneric series of compounds.

1.2. Structural validation

This structural benchmark evaluates whether FragPELE can reproduce native X-ray crystal structures when expanding fragments in different scenarios. For this purpose, a total of thirteen systems were studied; one was explicitly considered to capture cryptic sub-pockets opening.

System preparation. All structures with missing side-chains were corrected using the 3D builder from Maestro (Schrödinger, 2018), followed by energy minimization of these residues with OPLS2005 (Jorgensen et al., 1996) force field and the implicit SGB solvent (Gallicchio et al., 2002). Systems were prepared with Protein Preparation Wizard (Sastry et al., 2013) from Schrödinger, including analysis and H-bond optimization using PROPKA (Olsson et al., 2011) at pH 7.

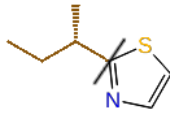
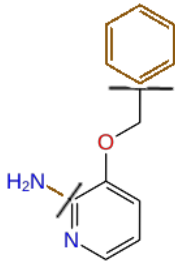
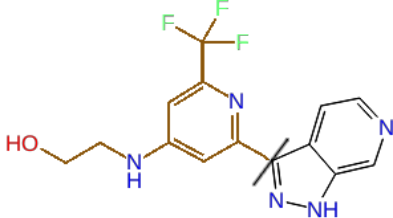
1.2.1. Self-growing

The initial validation consisted of simply growing a part of a molecule that had been previously removed from original crystals. We called this concept *self-growing*.

Systems and methods. Simulations were initialized from the X-ray with the remaining scaffold (after deleting the fragment part) situated in the BS. Later, the subtracted moiety will be grown again to reproduce the original pose. To achieve this goal, three well-known systems with available crystallographic data were chosen from Steinbrecher et al. FEP+ Benchmark studies (Steinbrecher et al., 2015): (a) *Major Urinary Protein* (MUP-I) in complex with *sec-butyl-thiazoline* (PDB code: 1I06) (Timm et al., 2001), (b) *p38 α kinase* or *mitogen-activated protein kinase 14* (MAPK14) co-crystallized with *3-(benzyloxy)pyridin-2-amine* (PDB code: 1W7H) (Hartshorn et al., 2005), (c) Bacterial DNA ligase in complex with an azaindazol (PDB code: 4CC6) (Howard et al., 2013).

Fragments were deleted and replaced by a hydrogen atom by using Maestro software (Schrödinger, 2018). In *Table 3.1*, 2D structures of all ligands are depicted, highlighting the fragment (deleted) part for each one. In DNA ligase and p38, two tests were performed, one keeping structural waters and the other without them, while in MUP-I, we could only test with no waters due to the hydrophobicity of its BS.

Table 3.1. Summary of the systems used in the self-growing. The fragments to be grown with a FragPELE simulation are brown, while the scaffold remains black. Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

System	PDB ID	Resolution (Å)	Ligand
MUP-I	1I06	1.9	
p38 kinase	1W7H	2.2	
DNA Ligase	4CC6	2.01	

Standard FragPELE simulations were run, and results were assessed by comparing the root mean square deviation (RMSD) of the ligand heavy atoms for the lowest binding energy pose with its original crystal structure (pre-aligning alpha carbons). Three different groups of heavy atoms were employed in RMSD calculations: core, fragment, and whole molecule atoms.

Results. RMSD results are summarized in *Table 3.2*. All three systems showed values under 2 Å, retrieving the native-like conformation for all molecule parts, fragments, and core. However, close attention must be paid to hydrated BS. When we do not include explicit waters in the calculation, the RMSD values slightly increase. Results on p38 suggest that the ligand tends to occupy the steric space left by the removed waters,

despite no further interactions being gained (Figure 3.4). Lastly, results from DNA ligase (4CC6) (Figure 3.5) demonstrate, at least in this simple test, that FragPELE can recover native interactions when growing large fragments (14 heavy atoms size). Here the trifluoride moiety of the fragment interacts with the surrounding hydrophobic residues quickly.

Table 3.2. Self-growing results in the core, fragment, and total heavy atom RMSD. Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.


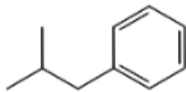
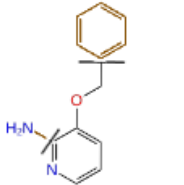
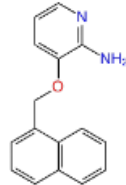
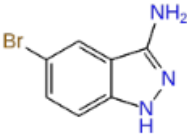
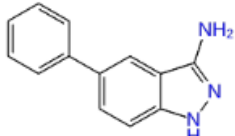
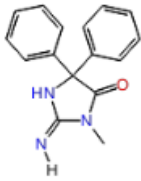
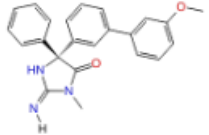
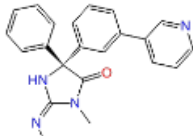
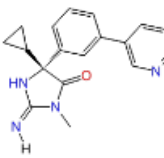
	System	PDB code(s)	Waters in simulation	RMSD core (Å)	RMSD fragment (Å)	Total RMSD (Å)	Figure
Self-growing	MUP-I	1I06	No	1.58	1.37	1.49	-
	p38	1W7H	No	2.53	2.46	2.51	3.4-upper
			Yes	1.96	1.54	1.80	3.4-bottom
	DNA ligase	4CC6	No	2.28	2.04	2.15	-
			Yes	1.23	1.59	1.46	3.5

1.2.2. Cross-growing

The second structural validation consisted of growing one or more fragments onto a crystallographic protein-scaffold complex and trying to reproduce the poses of a second X-ray containing the full-size new ligand (sharing the same scaffold). We named this concept *cross-growing*.

Systems and methods. In this case, we used four systems, from which we had at least two different crystal structures: one co-crystallized with the core ligand and at least a second one that could be generated by expanding an R-group onto the first. These four targets are *T4 lysozyme*, *p38 kinase*, *tyrosine-protein kinase JAK-II* (all three obtained from (Steinbrecher et al., 2015)), and the *beta-secretase I* (BACE). Table 3.3 depicts the ligands for all systems. Most systems do not include explicit water molecules, except BACE for the transition from 4DJU to 4DJW to gain a specific water-ligand interaction through the fragment grown.

Table 3.3. Systems used in cross-growing. We reproduce the ligands by growing fragments onto X-ray structures with co-crystallized scaffolds. In p38 kinase, the scaffold was generated by removing the brown parts of the molecule (from the self-growing benchmark, see *Table 3.1*). Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

System	PDB ID initial	Resolution (Å) initial	PDB ID to reproduce	Resolution (Å) reference	Scaffold	Ligand to reproduce
Lysozyme	181L	1,8	184L	1,8		
p38 kinase	1W7H	2,2	1WBW	2,41		
JAK2	3E62	1,92	3E63	1,9		
BACE	4DJU	1,8	4DJV	1,73		
			4DJW	1,9		
	4DJX	1,5	4DJY	1,86		

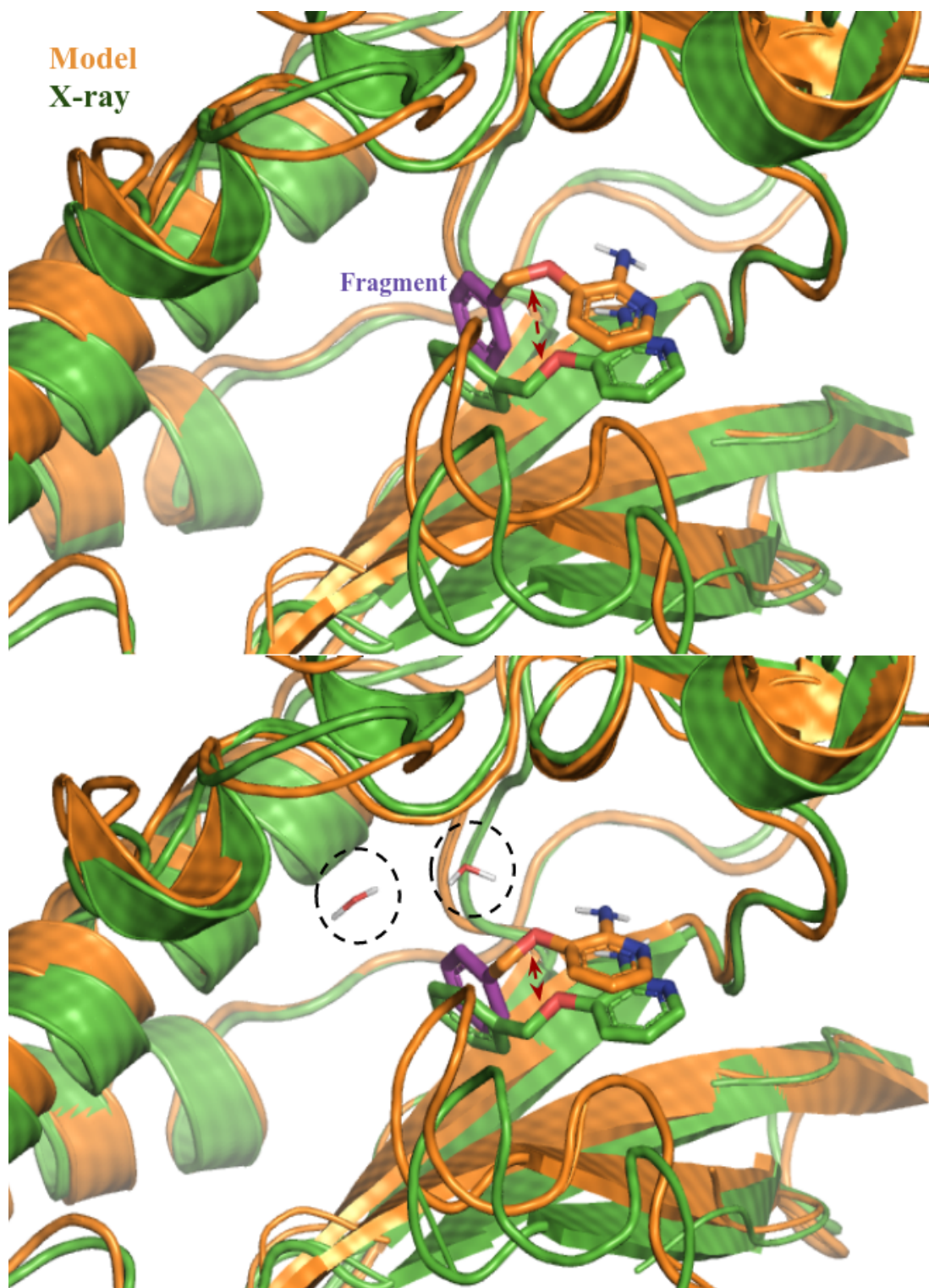


Figure 3.4. Growth of a phenyl with (below) and without (above) waters to reproduce PDB 1W7H. The distance between the model and the X-ray is reduced when including water molecules (red arrows). Created with PyMOL (Schrödinger & DeLano, 2018). Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

Identical to the previous analysis, standard FragPELE simulations were run, and RMSD values were computed onto the second X-rays to assess the quality of the results. Also, critical protein-ligand interactions of the native structure were compared with the ones found in the structures retrieved from FragPELE simulations.

Results. RMSD results are seen in *Table 3.4*. All values fall below 2 Å, excluding p38 kinase. A closer look at the lowest interaction energies structures for this system (*Figure 3.6*) revealed that naphthyl moiety rotated almost 180°, increasing the RMSD of this fragment. As illustrated in *Figure 3.6*, the addition of the fragment produces a displacement of K53 towards an essential pi-cation interaction. Despite the high RMSD value of 2.69 Å against 1WBW, K53 still conserves the specific lysine-glutamine lock of kinases.

Table 3.4. Cross-growing results in the core, fragment, and total heavy atom RMSD. Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

	Protein	PDB code(s)	Waters in simulation	RMSD core (Å)	RMSD fragment (Å)	Total RMSD (Å)	Figure
Cross-growing	Lysozyme	181LI to 184L	No	0.34	1.1	0.75	-
	p38	1W7H to 1WBW	No	1.38	3.46	2.69	3.6
	JAK-II	3E62 to 3E63	No	0.47	1.67	1.08	3.7
	BACE	4DJU to 4DJV	No	0.79	0.96	0.84	-
		4DJU to 4DJW	Yes	1.02	0.41	0.92	3.8
		4DJX to 4DJY	No	1	1.34	1.05	-

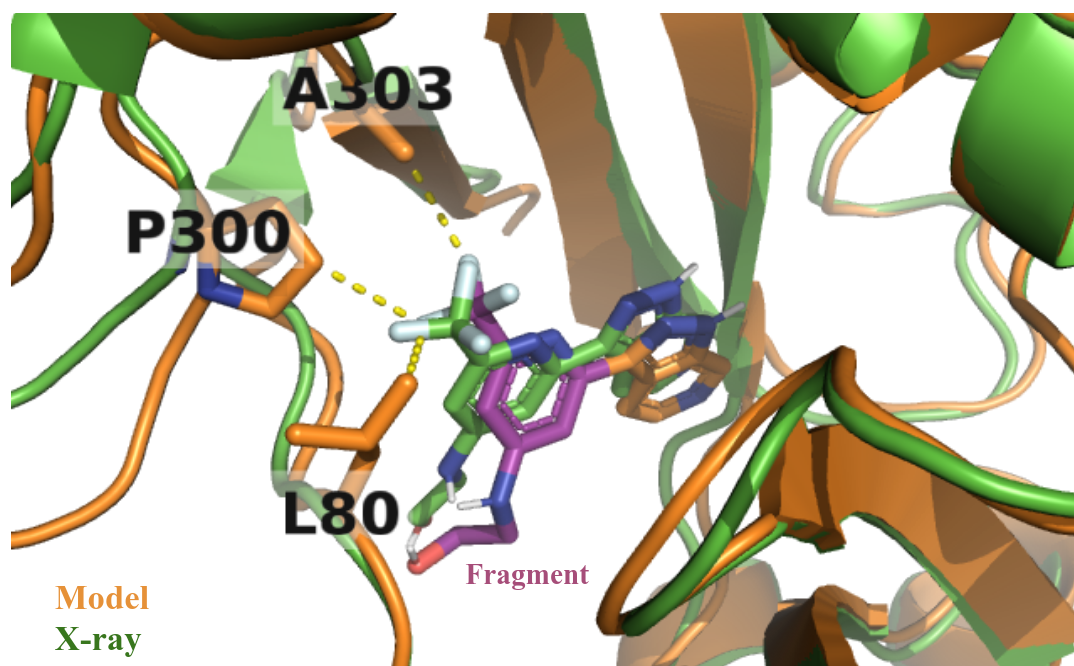


Figure 3.5. Growing of the fragment onto the core to reproduce PDB 4CC6. Yellow dashed lines highlight the interactions between the surrounding residues of the receptor and the trifluoride of the fragment. Created with PyMOL (Schrödinger & DeLano, 2018). Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

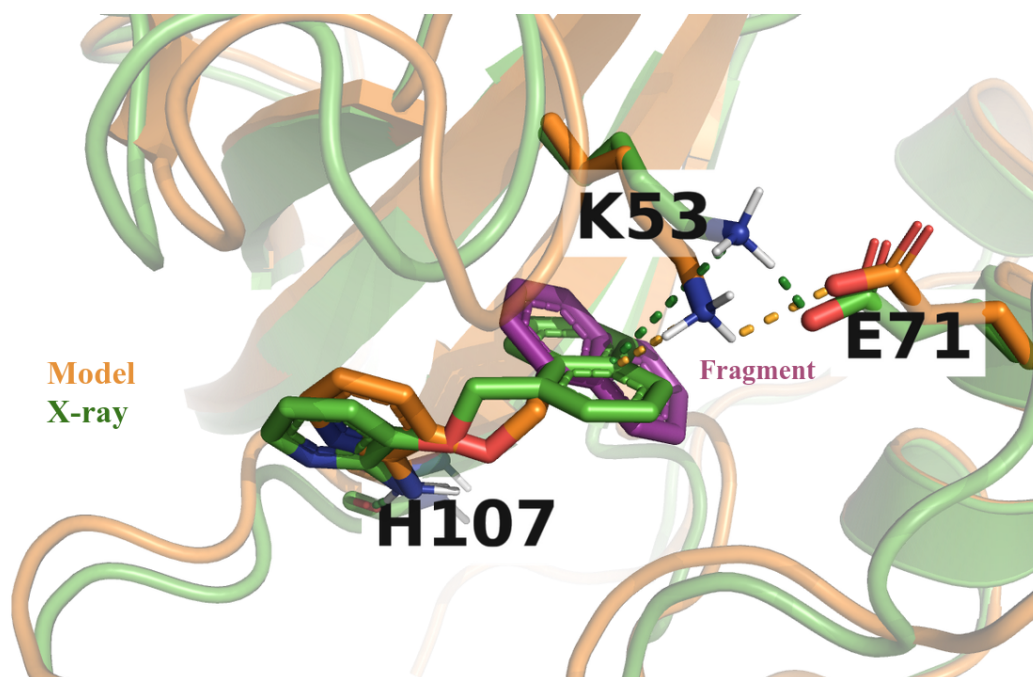


Figure 3.6. Lowest interaction energy structure from the growing of naphthyl (purple) and amino fragments from the core of 1W7H (orange) to reproduce the crystal structure

(**continuation**) 1WBW (in green). The addition of the naphthyl displaces K53 to allow the pi-cation interaction, favoring the contact between K53 and E71 (present in the crystal). Subsequently, the amino fragment's addition creates a second interaction with H107. Thus, all native interactions present in 1WBW were recovered. Created with PyMOL (Schrödinger & DeLano, 2018). Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

For JAK-II, when growing a phenyl onto the scaffold of 3E62 (*Figure 3.7*), the side-chain prediction algorithm relocated the aspartic acid to accommodate the six-membered ring, showing a full-ligand RMSD of 1.08 Å against the X-ray. Finally, in the BACE system, for the transition from 4DJV to 4DJW, we could recover a crucial interaction between the fragment and crystallographic structural water (*Figure 3.8*). Gaining this interaction locks the conformation of the ligand and favors enthalpically the position of the explicit water.

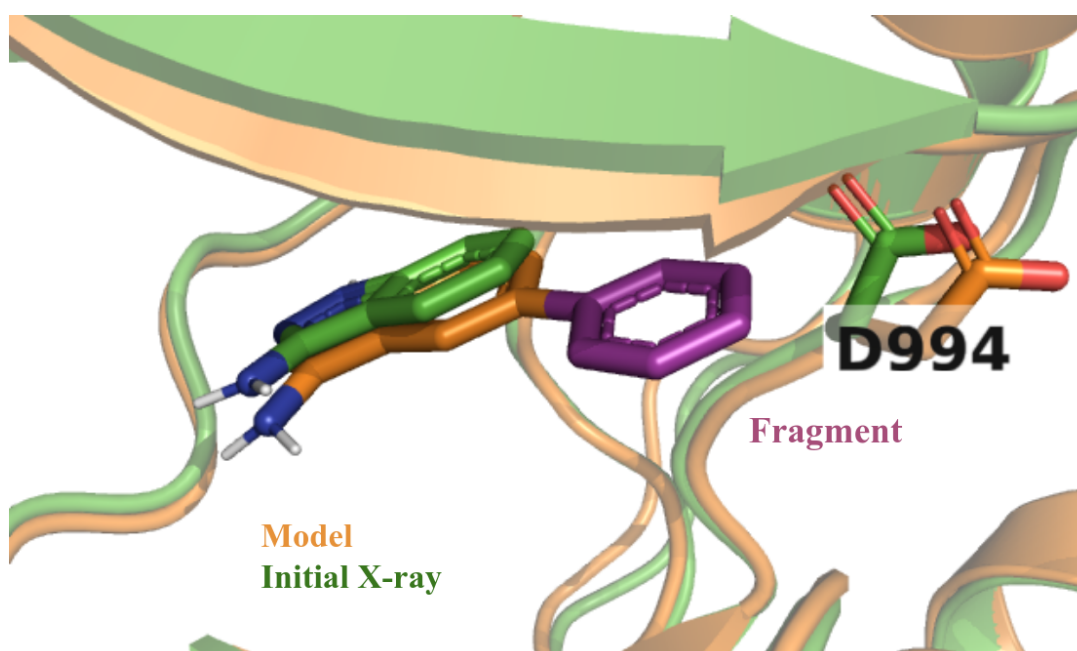


Figure 3.7. Growth of a phenyl onto the initial crystal 3E62, where D994 accommodates the insertion of the fragment. Created with PyMOL (Schrödinger & DeLano, 2018). Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

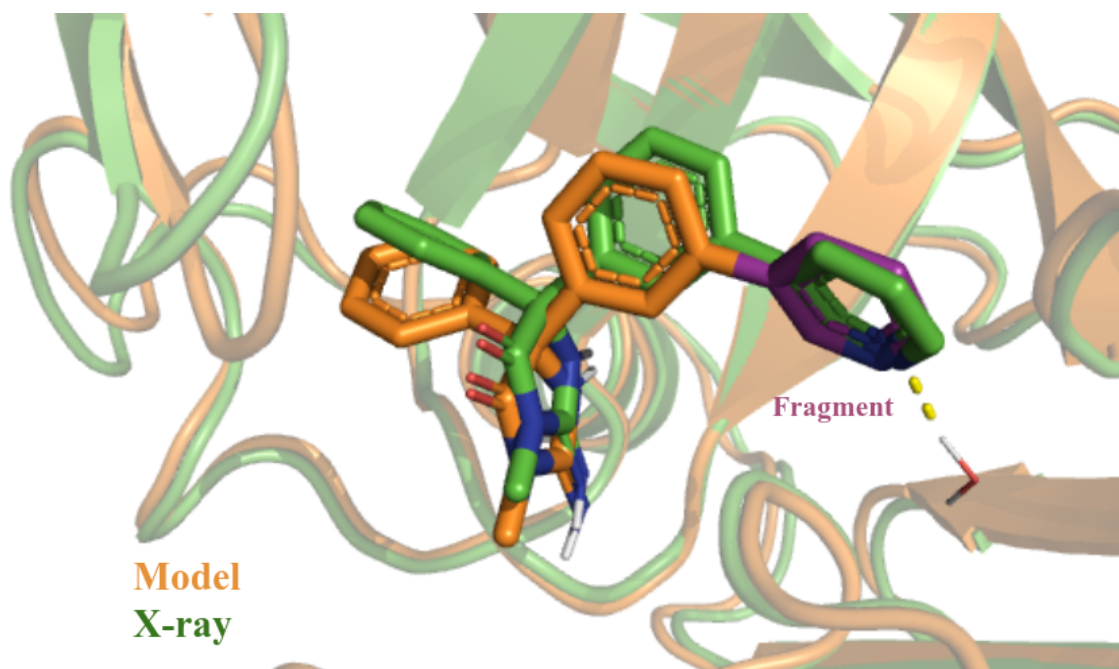


Figure 3.8. Model (orange) of 4DJW (green) by growing a fragment onto 4DJU. Notice how the model reproduces the interaction with the water molecule present in both crystal and model. Created with PyMOL (Schrödinger & DeLano, 2018). Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

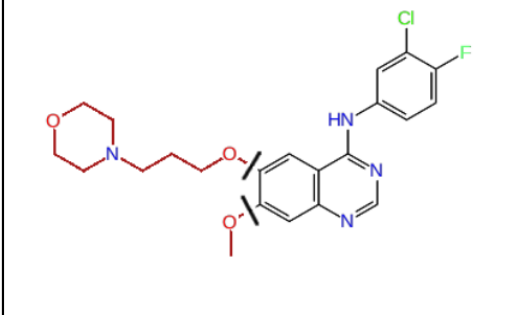
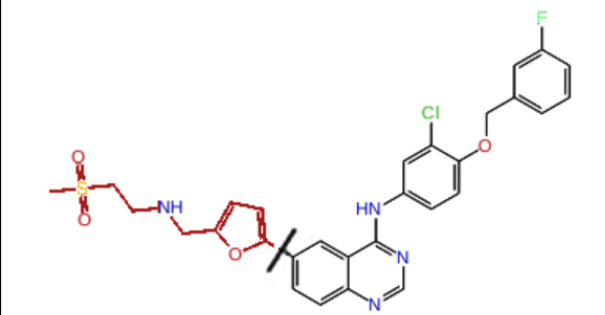
1.2.3. Cryptic sub-pockets

Our next structural consisted of checking the ability of FragPELE to identify hidden cavities close to the orthosteric site. This concept is called *cryptic sub-pockets*, as they only open up in the presence of ligands with particular R-groups. *Lapatinib* and *gefitinib* drugs are two popular inhibitors of the epidermal growth factor receptor (EGFR), which plays an essential role in carcinogenesis (*Expression of EGFR in Cancer - Summary - the Human Protein Atlas*, n.d.; Uhlén et al., 2015). Specifically, lapatinib is known to show a slower dissociation rate than gefitinib due to the differences in how they bind (Bilancia et al., 2007). At the same time, we sought to target a novel-induced cavity next to the ATP BS.

Systems and methods. Simulations were focused on the development of the second-generation inhibitor lapatinib (PDB code: 1XKK) (Wood et al., 2004) by extending the first-generation inhibitor gefitinib (PDB code: 4WKQ). A solvent-exposed R-group is generally used to modify the ADME properties of the drug,

which are not relevant for the present study (see *Table 3.5*). Thus, to accelerate the simulation time, we decided to delete this region from the gefitinib scaffold with too many rotatable bonds, given that sampling all possible conformations would take much unnecessary computational time.

Table 3.5. Data of the system used in the cryptic sub-pockets benchmark. Brown colored solvent-exposed moieties were replaced for hydrogen atoms and were not included in the simulations. Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

System	ID initial	Resolution (Å) initial	ID final	Resolution (Å) final
EGFR	PDB: 4WQK	1,85	1XKK	2,4
Ligand initial		Ligand final		
				

To assess results, we focus the analysis on two side-chains lining the sub-pocket region, *M766* and *F856*. As observed in crystal structures, the former must move aside to let enough space to place the bulky fluoro-phenyl, and the latter is stabilizing a crucial pi-pi interaction with this fragment.

Results. In *Figure 3.9*, we show FragPELE resultant binding mode. It is observed how the ligand displaces *M766* and simultaneously orients *F856* closer to the grown fragment to reproduce the pi-pi stacking interaction.

As seen in *Figure 3.10*, all amino acids initially lining close to the fragment in lapatinib X-ray are also present in the model. The pi-pi stack between *F856* is not represented because of the flip of 180° of the ring (*Figure 3.11*), which interposes the fluorine atom; however, centroid distances (5.7 Å) and angles (54°) are close to suitable conditions for this kind of interaction. Notice in *Figure 3.11* that the allocation of the new fragment is

mainly done by side-chain displacements and not for large movements in the backbone (alpha-helix in this case).

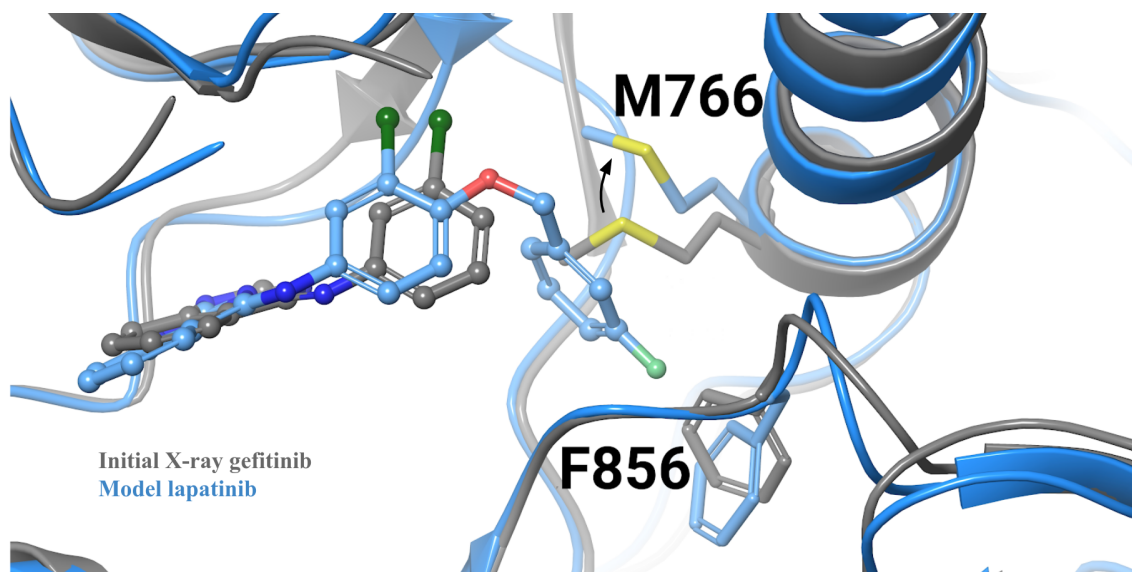


Figure 3.9. Generation of lapatinib from gefitinib. Model onto the initial protein-ligand complex. M766 moves aside, placing the fragment and F856 re-orient to catch a possible interaction with the aromatic group. Created with Maestro (Schrödinger, 2018). Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

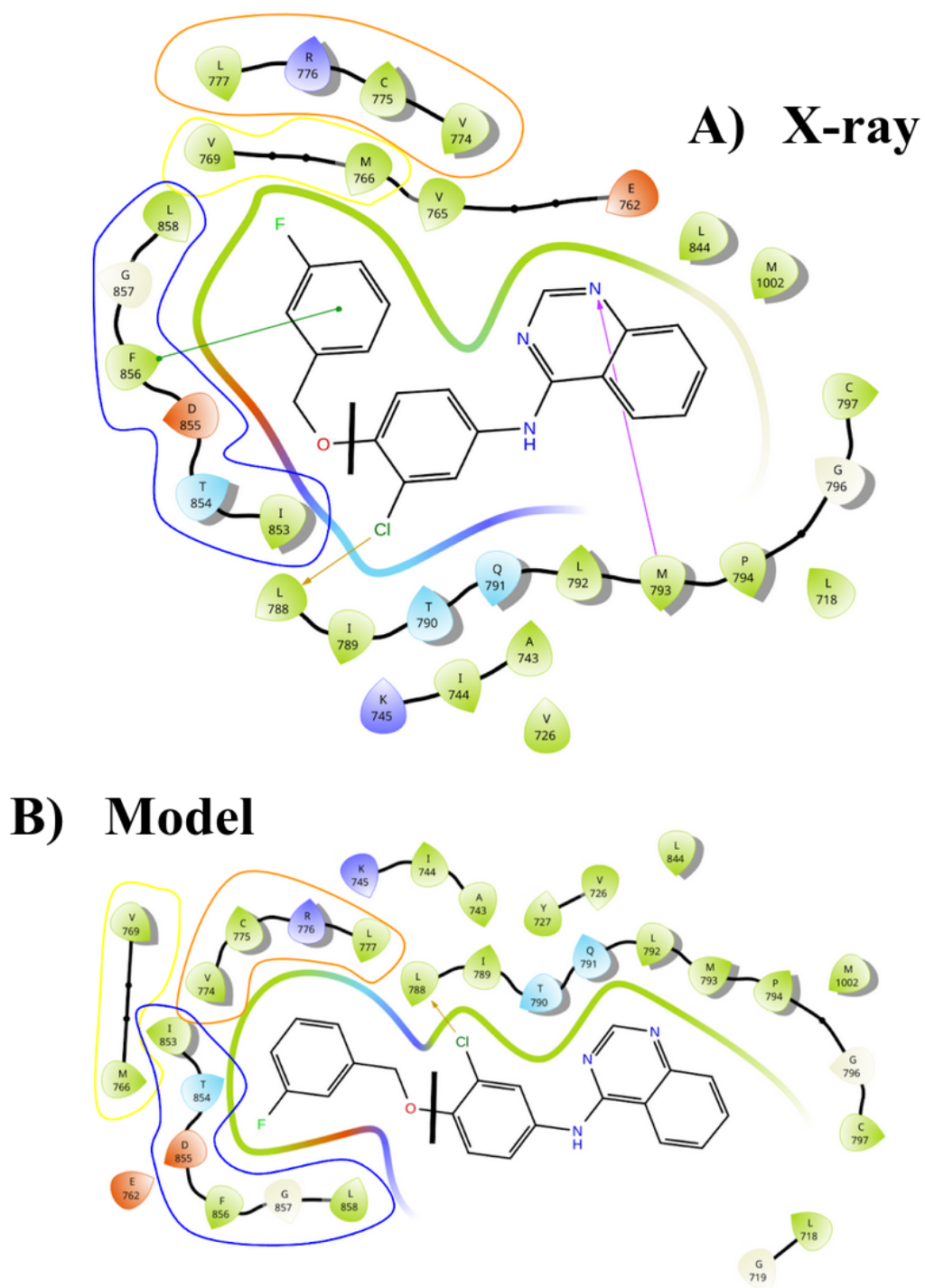


Figure 3.10. Interaction diagrams of lapatinib crystal (A) and the model built with FragPELE (B). Amino acids at 4.5 Å are represented in different colors: green for hydrophobic, glycines in gray, positively charged in dark-blue, negatively charged in red, and polar amino acids in blue. Arrows and lines show interactions: green for pi-pi stacking, H-bonds in purple, and halogen bonds in orange. Groups of amino acid lining fragments were colored to compare A and B. Notice that the model fragment is flipped

(continuation) 180° in comparison with the X-ray. Created with Interaction Diagram of Maestro (Schrödinger, 2018). Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

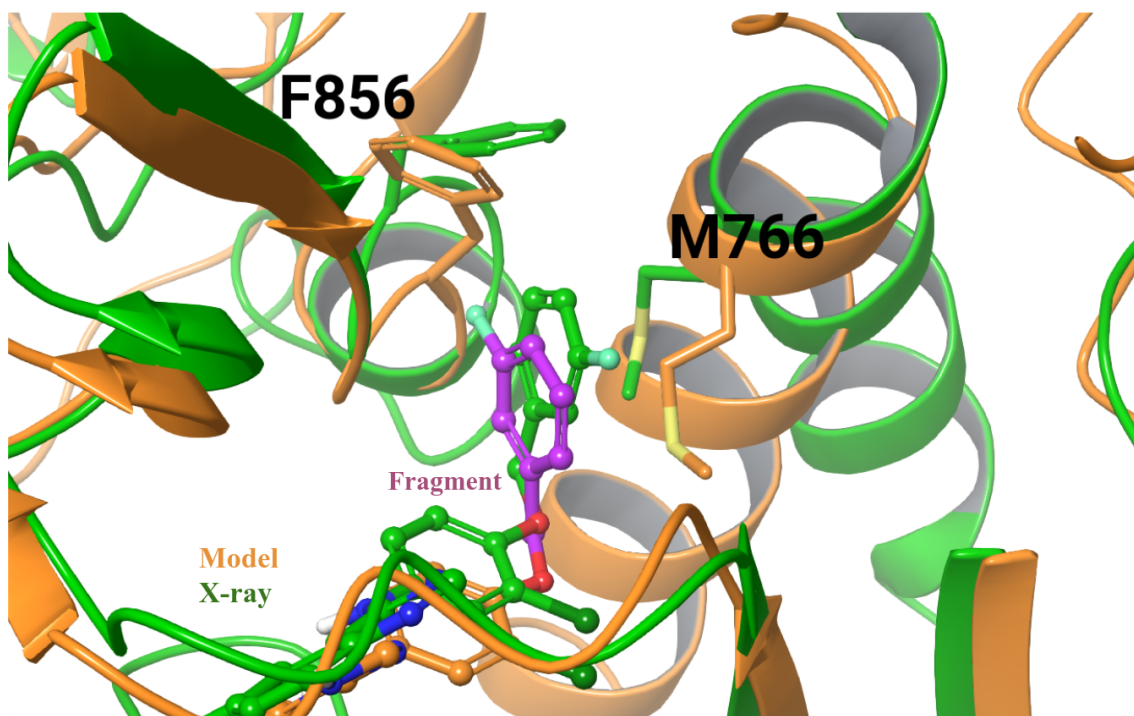


Figure 3.11. Modeling of 4WKQ (lapatinib). This figure complements the information given in *Figure 3.10*. Created with Maestro (Schrödinger, 2018). Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

1.2.4. Growing bulky R-groups

In the final test, we tried to evaluate the reliability when growing a fragment that decreases the binding affinity of its precursor. This study was performed to verify that FragPELE does not create false positives when growing a bulky R-group.

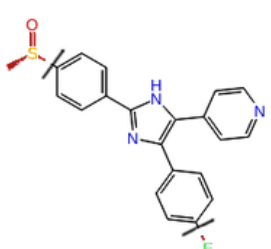
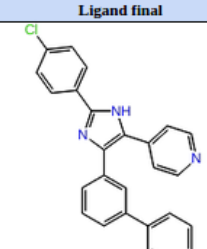
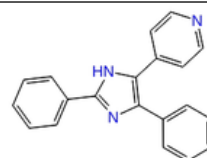
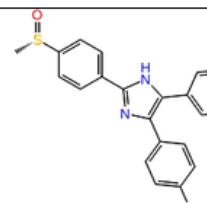
Systems and methods. We tested growing a small series of MAPK p38 inhibitors whose IC50s were available for all analogs (*Table 3.6*), but only one had crystal structure available (1A9U) (Z. Wang et al., 1998). The experimental series was obtained from ChEMBL (Gaulton et al., 2012) (ChEMBL71403 (Chang et al., 2001), ChEMBL69929 (Liverton et al., 1999)), where we checked that the members had a wide range of IC50 values. Two fragments confer higher potency (low nM), whereas the other presents lower affinity (low μM) than the reference. This system deals with a

prototypical ATP-competitive kinase inhibitor, anchoring a heterocycle to the central hinge residue (in this case, *M109*) through an H-bond (Howard et al., 2014).

We hypothesize that the less potent ligands within the series have too bulky R-group, which cannot fit. To verify this, we retrieved structures from FragPELE simulations to compare them to the initial pose of the only analog whose single X-ray structure is available.

Results. Predicted binding poses are shown in *Figure 3.12*, where it is observed that the addition of cyclohexane promoted the displacement of the ligand outside the cavity (yellow model of *Figure 3.12*), breaking the canonical hinge interaction with *M109* (a backbone interaction). The other two ligands kept this contact, which is well known to contribute to a substantial part of the binding affinity. We suggest that the cyclohexane ring is too big to fit into the ATP BS correctly, causing a decrease in binding potency due to these steric effects. However, as no X-ray structure is provided, this fact cannot be ultimately proven.

Table 3.6. Summary of the systems used to rationalize binding affinities. In brown, it represents the part of the ligand that has been deleted. Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

System	ID initial	Resolution (Å)	ID final	Ligand initial	Ligand final	IC50
p38 kinase	PDB: 1A9U	2,5	CHEMBL71403			10µM
			CHEMBL69929			28nM
			PDB: 1A9U			48nM

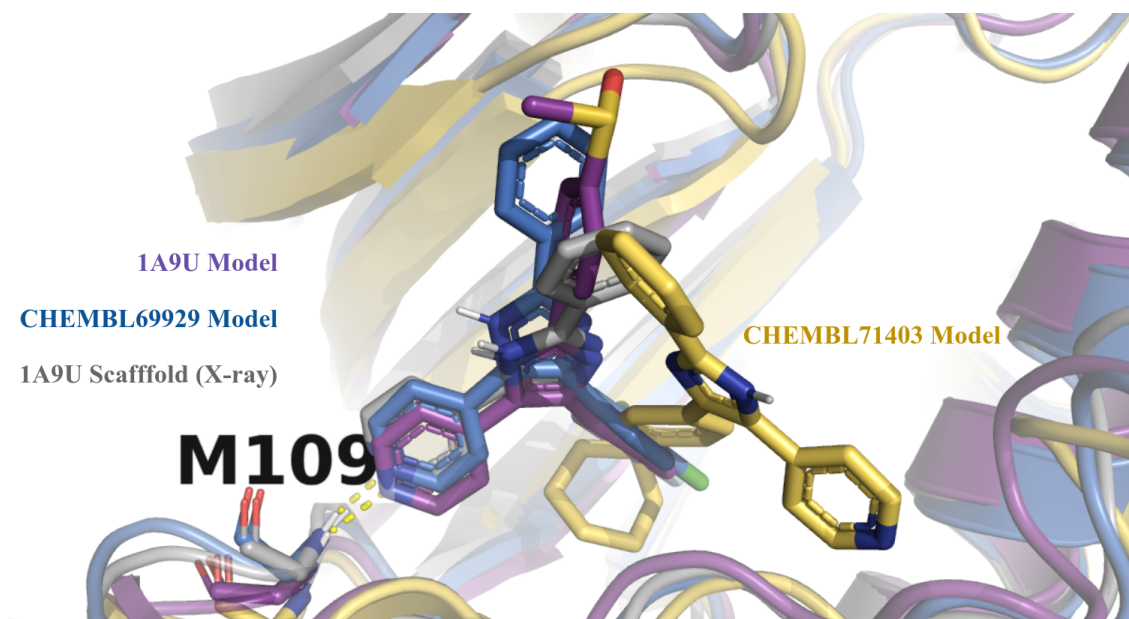


Figure 3.12. Growing of several fragments to create three different inhibitors of p38 type II. 1A9U (purple) and CHEMBL69929 (blue) are binders with significantly higher affinity, and CHEMBL71403 (yellow) is the lowest affinity one. In gray, it represents the position of the scaffold pose, derived from crystal 1A9U after deleting the fluorine and methanesulfonyl groups. Created with PyMOL (Schrödinger & DeLano, 2018). Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

1.2.5. Protein motion

Although it was not our primary goal, our analysis focused on protein motion. To know the perturbation level of protein atoms, we compared retrieved structures from FragPELE simulations with the original crystals. We have obtained low RMSD values for those systems with lower mobility (MUP-1, DNA ligase, BACE, and lysozyme) and higher ones in kinases (p38 and JAK-II), which have more flexibility (*Table 3.7*). However, all protein RMSD were relatively low (between 1.16 and 4.67Å), so we could consider that mainly side-chain rearrangements were fundamental to reproducing ligand poses.

Table 3.7. Heavy atom RMSDs of the protein (backbone + side-chains) for the models against crystals. Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

	System	PDB code(s)	Waters in simulation	Protein RMSD (Å)
Self-growing	MUP-I	1i06	No	1.56
	p38	1w7h	No	4.67
			Yes	3.31
	DNA ligase	4cc6	No	1.68
			Yes	1.45
Cross-growing	Lysozyme	1811 to 1841	No	2.91
	p38	1w7h to 1wbw	No	3.28
	JAK-II	3e62 to 3e63	No	3.87
	BACE	4dju to 4djv	No	1.55
		4dju to 4djw	Yes	1.52
		4dju to 4djy	No	1.16

1.3. Growing and scoring

As a final step, we assessed whether the interaction energies generated for the grown R-groups could be used to rank molecules in a H2L stage. PELE's interaction energy usually discriminates against similar-sized ligands to the same target. Thus, we hypothesize that they may work when scoring ligands with a common structural scaffold; we also expect differences in entropic terms to be minor compared to the change in enthalpy.

The chosen benchmark involves the FEP+ original study (Steinbrecher et al., 2015), which allowed us to compare our technique with Glide, MMGBSA, and FEP+. The structural validation indicated that FragPELE could accurately predict the ligand-bound geometry within a BS after R-group growth, even in cases where significant heavy atom

gains are involved. Hence, this study assessed the accuracy of FragPELE at growing and scoring fragments and if the software falls midway between the accurate but expensive FEP and the cheap but sometimes inexact docking algorithms.

Systems and methods. We evaluated our software performance at growing and scoring fragments in five systems on the previously mentioned benchmark (Steinbrecher et al., 2015): T4 Lysozyme, DNA ligase, MUP-I, JAK-II, and p38. These systems were carefully chosen for variable fragment sizes, BS characteristics, and MW. We discarded those cases where the molecule was not amenable to R-group growing methodologies, such as molecules constituted by single rings or alchemical transformations of heavy atoms of the core. To see the ligands list visit *Appendix A*. Standard Induced-Fit Glide (Schrödinger, 2018) calculations with an OPLS3 force field (Harder et al., 2016) were run, and our results were compared to FEP+, Glide SP docking, and MMGBSA directly provided in the benchmark paper (Steinbrecher et al., 2015). FragPELE simulations were run following the same protocol as in the structural benchmark.

Results were evaluated by computing the coefficient of determination (R^2) between the predicted and the experimental values. This metric will determine the proportion of variability in the predicted values explained from the experimental ones, giving the goodness of fit between both. Therefore, we would know whether an estimator is helpful in ranking the ligands independently of its ability to predict absolute free energies. They range from 0 to 1, one meaning perfect fitness between variables.

Results. A complete view of the correlation results is illustrated in *Table 3.8*, and individual correlation plots are shown in *Appendix B* section. The correlations obtained with FragPELE are moderately worse than FEP+ for lysozyme, DNA ligase, and JAK-II systems. Our method contrives FEP+ in MUP-I and, surprisingly, outperforms FEP+ in the p38 results. Moreover, our results are substantially better than Glide SP scores for almost all systems, as the latter does not account for side-chain flexibility. However, for p38 and lysozyme, which have a low MW correlation, Glide (SP and Induced-Fit) and MMGBSA perform poorly, while FragPELE and FEP+ obtain good correlation values. Finally, all methods struggle for JAK-II (0.32 MW correlation), with only FEP+ achieving an acceptable correlation (0.64).

One-tailed paired t-tests (threshold %5) were computed using the SciPy python package (Virtanen et al., 2020), taking R^2 for each system as observables. Results confirmed that FragPELE predictions higher correlate with experimental values than GlideSP, Glide Induced-fit or either MMGBSA (default and flexible) (p -values 0.046, 0.0048, 0.048, $0.01 < 0.05$, respectively), but non-significant differences were reported in FEP+ (p -value = 0.726 > 0.05). When applying left-sided t-test comparing FragPELE with FEP+, non-significant differences were observed (p -value = 0.274 > 0.05), suggesting that FEP+ does not outperform FragPELE.

Table 3.8. Coefficient of determinations between experimental data and different scoring approaches from FEP+ benchmark of Steinbrecher et al., FragPELE, and Induced-Fit Glide. Reprinted (adapted) with permission from (Perez et al., 2020). Copyright 2020 American Chemical Society.

System	PDB code	R^2 FragPELE	R^2 FEP+*	R^2 Glide SP**	R^2 Glide Induced - fit	R^2 MMGBSA default**	R^2 MMGBSA flexible**	R^2 MW
Lysozyme	181L	0.64	0.79	0.32	0.28	0.40	0.3	0.32
DNA ligase	4CC5	0.88	0.98	0.36	0.75	0.01	0.36	0.92
MUP-I	1I06	0.96	0.94	0.92	0.84	0.86	0.75	0.93
JAK-II	3E62	0.48	0.64	0.50	0.19	0.50	0.21	0.32
p38	1W7H	0.87	0.69	0.09	0.50	0.01	0	0.63

* R^2 recomputed with predicted values from Steinbrecher et.al.

** R^2 directly extracted from original FEP+ benchmark.

Regarding computing time, FragPELE spends an average of one hour per fragment on 48 Intel Xeon Platinum 8160 processors and can be quickly executed on any CPU cluster. Thus, its computational cost falls midway between FEP and docking but still tracking for the dynamics of the ligand-protein system.

We could validate the FragPELE method to perform dynamic fragment growing through the structural and scoring benchmarks, assessing its potential for predicting binding geometries and affinities. A further discussion about the technique will be expanded in its correspondent section in *the Discussion* chapter.

Additionally, this initial version still presents some limitations, such as only expanding from hydrogen, not modifying FFP solvent parameters along with fragment growing stage, and only R-group addition through single bonding. After their application in several drug design projects, errors were solved, and new functionalities arose. More details regarding this topic are discussed in *Appendix C*.

2. Growing on hydrated sites

The effect of water molecules in hydrated binding sites can play a crucial role in defining cavity interactions. Hence, they must be considered in drug discovery studies (Ichihara et al., 2014). Herein, waters usually coordinate key protein-protein or protein-ligand HBs, which are fundamental to stabilizing the substrate binding.

Several hit-to-lead campaigns aim to modify hit compounds to displace water molecules and gain these indispensable new interactions, improving ligands' potency (Abel et al., 2008; Barillari et al., 2007; J. M. Chen et al., 1998; Michel et al., 2009). In the previous section (*Section 1, Chapter 3*), FragPELE has been presented as a promising ligand growing tool to predict binding geometries and score the newly grown fragments accurately; however, growing fragments on hydrated sites was still an unfinished business. The method performed worse when not considering explicit waters from hydrated systems, with the limitation of fixing them through position constraints when included. The recent release of aquaPELE (Municoy et al., 2020) (*Section 3.4.2 of Chapter 1*), a new PELE version that includes MC steps following a mixed implicit/explicit solvent model, could be the ideal solution for FragPELE simulations to face hydrated systems, allowing the displacement and relocation of explicit water molecules while the ligand is expanded. Therefore, we combined both tools and assessed their ability to displace or keep explicit water molecules, enhancing the prediction when facing hydrated BS. Similarly to FragPELE's benchmarks, we divided them into two parts: structurally and scoring assessments. Additionally, all the work described within this section has been in collaboration with Ignasi Puch during his Master thesis.

2.1. Structural validation with explicit waters

This structural study aims to evaluate the ability to relocate buried explicit water molecules when growing fragments on hydrate cavities, assessing the predicted poses for ligand and displaced water molecules.

2.1.1. Methods

Benchmark design and preparation. Ten structural tests were designed, where the growth of a fragment onto an initial X-ray structure led to the second crystallographic structure. For these tests, five varied and structurally well-defined receptors with hydrated BS were selected: *heat shock protein 90-alpha* (HSP90), *bromodomain-containing protein 4* (BRD4), *TATA-Box binding protein associated factor 1* (TAF1), *sialic acid-binding periplasmic protein* (SiaP), and the *checkpoint kinase 1* (CHK1). *Figure 3.13* shows the set of chosen systems for this structural benchmark. *A-D* and *O-R* (Kung et al., 2011; Woodhead et al., 2010) are binders of HSP90, *E-F* (Nittinger et al., 2019) of BRD4, *G-I* (Nittinger et al., 2019) of TAF1, *J-L* (Darby et al., 2019) of SiaP, and *M-N* (Foloppe et al., 2006) of CHK1. Systems *A* to *N* were considered positive tests (waters abandon the cavity).

We also aimed to prove that not all water molecules abandon the cavity after growing a fragment on the initial scaffold. Unfortunately, in the literature, we could not find any co-crystallized system where the growth of a fragment does not displace explicit waters. Consequently, the last group (*O-R*) was virtually designed and rationalized as the control group. In the *O-P* test, we proposed the addition of hydrophilic hydroxyl fragments onto 3RLQ (*A*), trying to stabilize the water molecule (*A248*) of the X-ray instead of the methyl group from 3RLR (*B*). Then, in *Q-R*, we performed a self-growing exercise to see if both water molecules (*A1* and *A3*) stayed in place, interacting with the fragment after re-growing it.

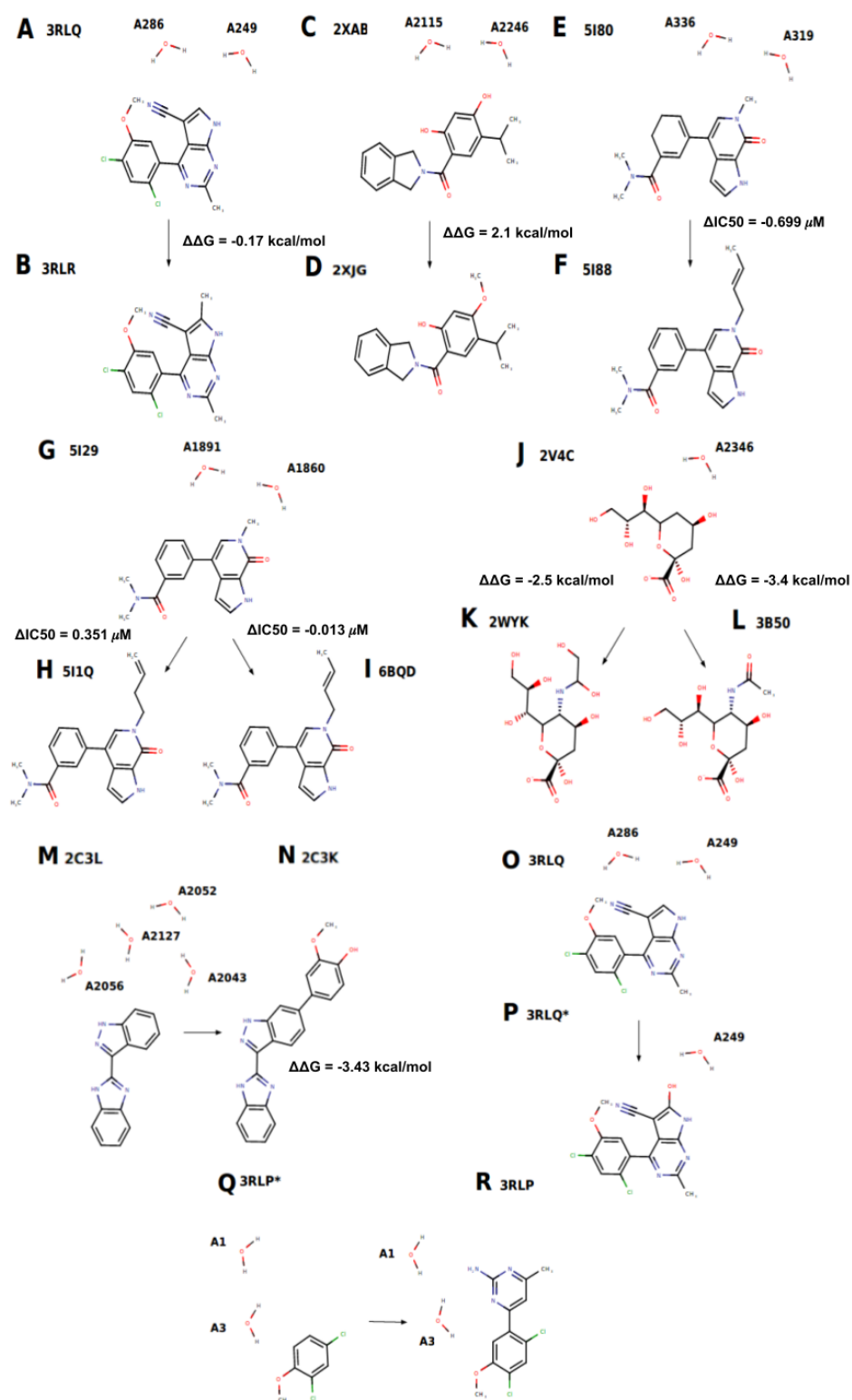


Figure 3.13. 2D illustration for the systems considered in the structural benchmark (A to R). Next to each label, the PDB code of the X-ray crystal is indicated. One or more new compounds are created from the initial structure by growing a single fragment. Arrows

(continuation) represent these transitions, showing the differences in free energy or IC50s between each pair. Notice that most of the explicit water molecules (labeled according to the PDB file information) are entirely displaced by the fragment grown (*A-N*), except for the two last systems (*O-R*). *P* and *Q*, highlighted with an asterisk (*), are modified versions of the X-ray.

For each pair of protein-ligand complexes (indicated with arrows in *Figure 3.13*), the initial X-ray was utilized as input structure for aquaPELE and FragPELE simulations. Crystals with missing side-chains were introduced with the 3D Builder tool of Maestro, running then a local minimization on these residues with the OPLS2005 force field (Banks et al., 2005; Kaminski et al., 2001) and implicit solvent SGB (Gallicchio et al., 2002). Systems were prepared with Protein Preparation Wizard (Sastry et al., 2013), deleting all explicit water molecules further than 5 Å of any ligand atom. The definition of protonation states and its H-bonds optimization was done with PROPKA (Olsson et al., 2011) at the crystallization pH for each system.

First, a control simulation with aquaPELE was run to evaluate the algorithm's reliability in finding hydration sites and good ligand positions. When preparing the waters of these complexes, solvent-exposed explicit waters were removed, keeping only buried ones stabilizing crucial protein-protein interactions or the ligand-protein interface. We fixed these water molecules whose location is well-known, except those prone to be displaced when growing a bulky fragment onto the scaffold, defining them as waters to be perturbed. Perturbable waters were manually moved to a new random location at ~2.5 Å of distance from their original site to avoid any bias toward initial structures. Water regions were defined case by case, centering the spherical box close to the growing site. The radii were big enough to let waters explore the whole BS and partially reach solvent-exposed regions. With this configuration, we expected that perturbable waters could find the most suitable place to stay within the cavity or the solvent when any favorable location is not found. *Figure 3.14* shows an illustrative example of the water selection strategy.

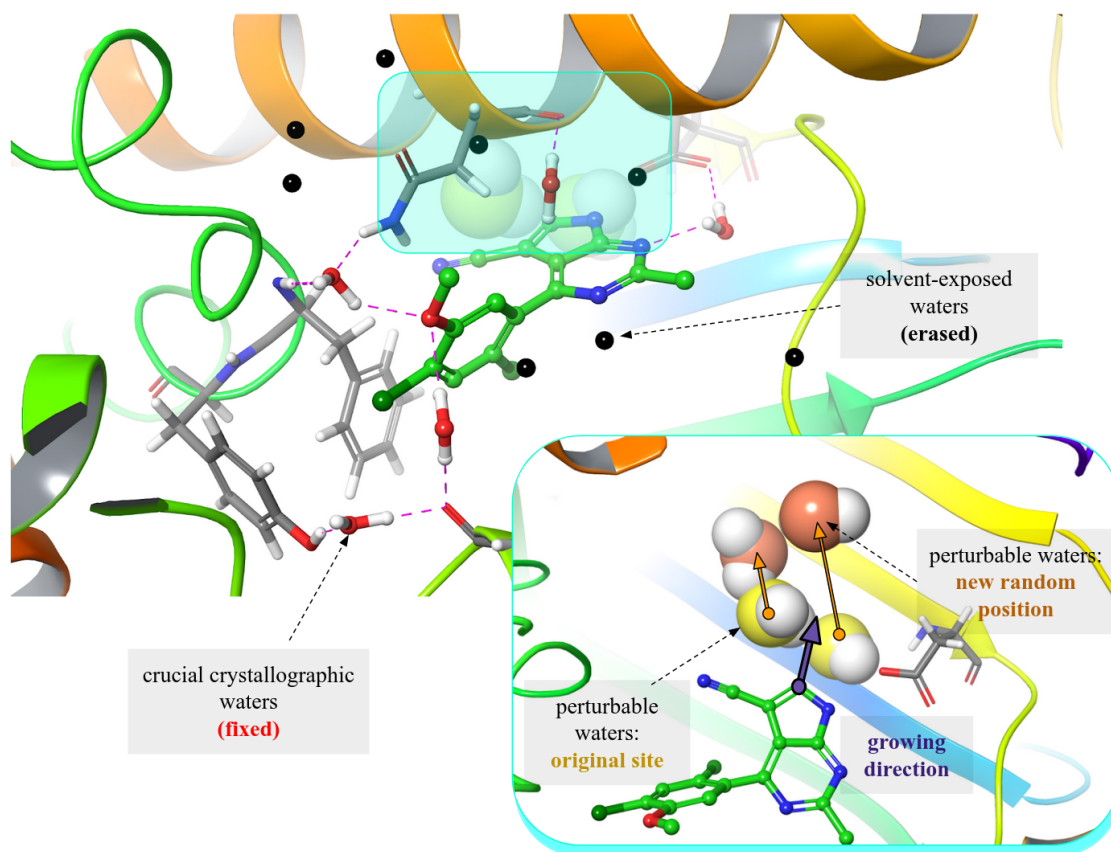


Figure 3.14. Water selection example (PDB code: 3RLQ). Solvent-exposed and non-relevant water molecules (in black) will be deleted from the system, keeping and fixing the ones that stabilize ligand-protein or protein-protein interactions (in red). The explicit waters around the growing direction (purple arrow) will be perturbed in aquaPELE simulation (yellow spheres); thus, to avoid biases, they had to be displaced (orange arrows) to a new random position (orange spheres). Created with Maestro (Schrödinger, 2018).

Finally, both techniques (aquaPELE and FragPELE) were combined with growing the desired fragment in each pair of initial-final crystal structures. We used the same initial structure and configuration as the previous simulation to compare the results.

Simulation configuration. AquaPELE simulations were configured to explore the variable ligand conformations and hydrate sites within the receptor BS. We performed 47 parallel simulations of 400 MC steps (18.800 total steps). Ligand's translations and rotations were respectively set to 0.25-0.5 Å and 0.1-0.05 radians, defining its perturbation box centered on the ligand's center of mass with a radius of 4 Å (similarly

to the standard FragPELE configuration). Fix explicit water molecules were determined case by case considering the characteristics of each receptor's cavity (structural role, protein-ligand coordination, water networks). Table 3.9 shows which explicit waters were fixed for each system.

Table 3.9. Fixed waters for each system, accordingly identified as in PDB files.

Systems	PDB scaffold	Fixed waters IDs
HSP90	3RLQ	1, 237, 243, 258, 313
	2XAB	2118, 2155
	3RLP*	6, 230, 240, 244, 261, 274, 292, 297, 310, 311, 336, 370, 372
BRD4	5I80	352, 364
TAF1	5I29	1832, 1882
SiaP	2V4C	2016, 2037, 2088, 2109, 2110, 2198, 2226, 2228, 2251, 2253, 2343, 2344, 2345
CHK	2C3L	-

Regarding aquaPELE configuration, water regions were set according to the geometry of each BS (centered in the growing site and big enough to reach solvent-exposed regions and the entire cavity). Herein, two explicit water molecules were perturbed, the minimum amount recommended to permit interaction networks between them. We also applied the best-reported water perturbation parameters of aquaPELE (Municoy et al., 2020) (temperature of acceptance of 5000K, 100000 steric trials, and randomly alternating translations of 4-2 Å).

The same aquaPELE configuration was used in the simulations of FragPELE 3.0 (see *Appendix C*). The growth was done in 10 GS of 6 MC steps, using the *Soft-core* protocol (defined in *Appendix C* section), with strong minimizations in the early first half of the phase (and then relaxed). The sampling simulation consisted of 400 MC steps, comparable with the initial control simulation.

Both aquaPELE and FragPELE simulations were run utilizing the OPLS2005 force field (Banks et al., 2005; Kaminski et al., 2001). Additionally, the implicit part of the mix solvent model used for aquaPELE was parameterized by employing the variant dielectric version of the SGB implicit solvent model (VDGBNP) (Zhu et al., 2007).

Simulation analysis. PELE simulations provide thousands of different accepted protein-ligand conformations, including multiple positions for the perturbed water molecules. Thus, a ligand and water clustering analysis has been performed on the protein-ligand 3D structures retrieved from simulations to find the essential binding modes and hydration sites. This analysis strategy has been stated in *Municoy et al.'s* work (Municoy et al., 2020); however, we will shortly summarize it in the following lines. Initially, the three-dimensional space is discretized with a clustering mean-shift algorithm from the Scikit-learn library (Pedregosa et al., 2011), which defines the coordinates of the heavy atoms as centroids. The algorithm computes the weighted mean given a set of points $N(x)$ and a λ window size (Equation 3.8), where $K(x_i - x)$ is a flat window function with a fixed bandwidth with the form stated in Equation 3.9. This bandwidth will define the distance between clusters.

$$(3.8) \quad m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) x_i}{\sum_{x_i \in N(x)} K(x_i - x)}$$

Equation 3.8. Mean-shift computation.

$$(3.9) \quad K(x) = \begin{cases} 1 & \text{if } x \leq \lambda \\ 0 & \text{if } x > \lambda \end{cases}$$

Equation 3.9. Definition of $K(x)$ function.

First, the clustering of the ligand binding modes was performed with the aforementioned mean-shift algorithm, using, in this case, a bandwidth of 2.5 Å, assigning then a specific cluster to every snapshot of the simulation. Second, to define the hydration sites of each ligand pose, a new round of clustering was applied to water molecules for each of the defined ligand clusters. Similarly, these molecules were

iteratively clustered with the mean-shift algorithm but defined a smaller bandwidth of 1.5 Å to imitate the atomic radii of a water molecule. Cluster densities (ρ) are calculated by comparing the most populated cluster to the population of each cluster with *Equation 3.10*, where N_i is the number of times that a water molecule landed in cluster i . In contrast, N_{max} is the maximum number of times a water molecule has landed in any cluster. This definition makes the density always have values between 0 and 1. Moreover, the ligand heavy atoms' RMSD against the reference X-ray (see *Figure 3.13*) was computed for each pose in the simulation, together with interaction energies.

$$(3.10) \quad \rho = N_i / N_{max}$$

Equation 3.10. Water density equation.

This study evaluated both the ability to predict ligand poses and hydrated sites. We selected the most populated ligand cluster of each aquaPELE or FragPELE to compare their RMSD values and the density of water molecules with the reference X-ray.

2.1.2. Results

Ligand poses. Energy profiles (*Figure 1D* from *Appendix D*) have revealed that the most populated clusters are generally the ones with the lowest RMSDs, as is summarized in *Figure 3.15*. These clusters coincide with the lowest energetic ones, proving that PELE's energy function can correctly score X-ray-like conformations. The only exception to the rule is system *D*, where the two most populated clusters are very close in terms of energy and RMSD.

Structurally, most of the selected clusters kept the mean of RMSDs around 1 Å, except in systems *D*, *F*, and *N*, which had higher values but below 2.5 Å in all of them. These higher RMSD conformations are not surprising when facing cavities with parts of the ligand close to solvent-exposed regions, as in systems *D* and *F*, where part of the scaffold has been slightly moved to the solvent as a consequence of the inner fragment expansion. Contrary, the explanation for the high RMSDs in the system *N* is the allocation of the growth fragment in a large empty pocket. It can quickly flip and adopt multiple conformations, giving; as a result, these high and wide ranges of values.

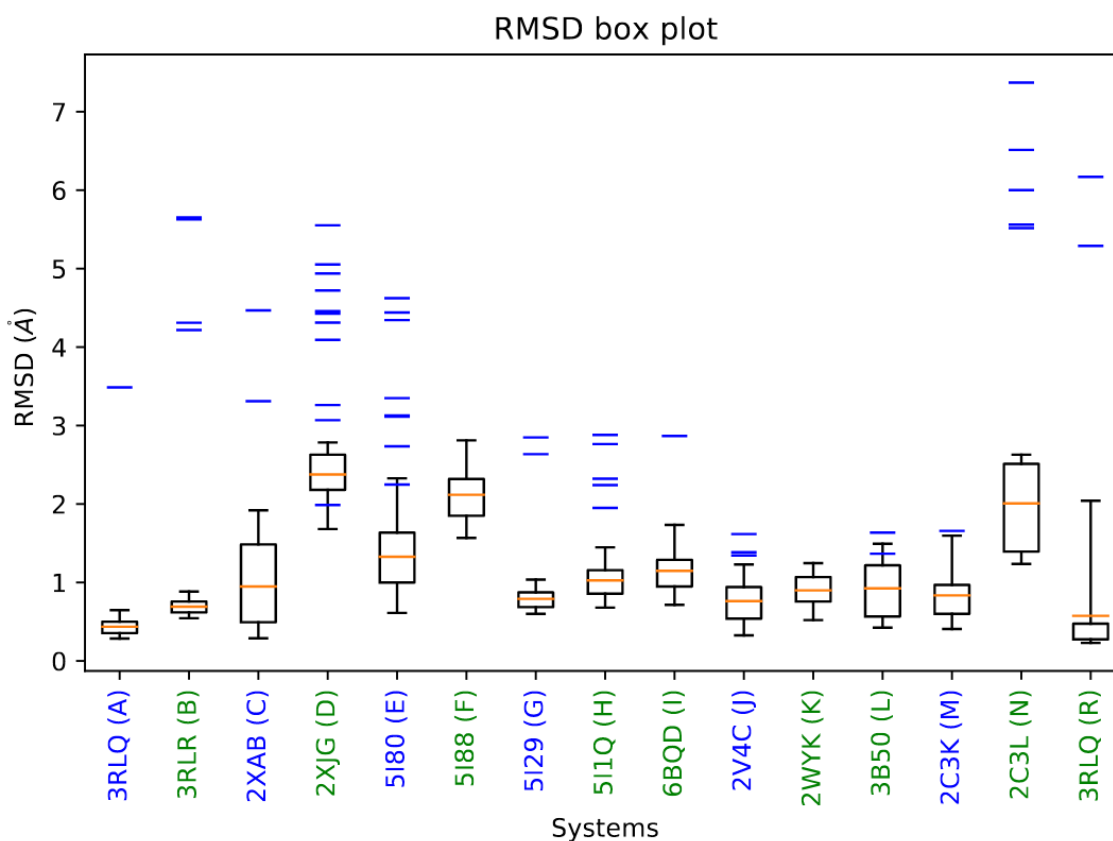


Figure 3.15. Ligand clusters RMSD distribution. The boxplot represents the most populated cluster, while blue lines show the mean RMSD values for the other clusters. PDB codes and their assigned letters correspond to the panels in *Figure 3.13*. The blue-labeled identify the results from only aquaPELE simulations and green-labeled the combination of aquaPELE and FragPELE methods.

Hydrated sites. The waters of the most populated ligand cluster resulting from the above analysis were clustered. Water densities were calculated using *Equation 3.10*, and distances from their original site were also obtained. All results are shown in *Table 3.10*.

In the HSP (1) (panel *A* of *Figure 3.13*), we pursue the displacement of two water molecules, A249 and A286, that were not present in the growth ligand (panel *B* of *Figure 3.13*). In this case, a water region of 10 Å radii had to be set to let the water molecules migrate to the solvent. The water clusters in the first aquaPELE simulation predict the most populated cluster halfway between the positions of the two X-ray locations of the water molecules (*Table 3.10*). Even though the hydration site was predicted correctly, finding a density of 1.0 (100% of the time, there is a water molecule

within), we observed only one water cluster instead of two. This fact is explained due to the moderately large bandwidth set in the clustering analysis ($\lambda = 1.5 \text{ \AA}$), which led to obtaining only one site, but whether this value is reduced, two sites arise. To explain these results, we checked electron density maps (*Appendix E, Figure E1*). We observed three close contacts of A249 with amino acids, while A286 only interacted with one of them. Additionally, the electronic cloud is bigger in A249 than in A286. This information hypothesizes that A249 is a more stable hydrated site than A286, which seems more transitory (*Appendix E, Figure E1*).

When the methyl group was expanded, the cluster was displaced, getting closer to the A249 position ($\Delta r_{cc} = 1.13 \text{ \AA}$), and its density shrank too ($\Delta \rho = -0.45$), predicting, therefore, an apparent reduction in the water molecule's presence and getting rid of the hypothesized lowly stable water molecule (A286) (*Table 3.10*).

In the second HSP90 system, the same radii size of 10 \AA was set. Here, aquaPELE simulations precisely predict the position of both explicit waters, showing short distances between crystallographic and predicted pose and high densities ($\rho = 1.0$ and 0.68) (*Table 3.10*). Only one water molecule was completely displaced when growing the fragment, while A2115 stayed. In 2XJG X-ray, the space of A2246 is occupied by the growth fragment, whereas the displacement of A2115 is not evident (*Appendix E, Figure E2*).

BRD4 required an expansion of the water region to 10.5 \AA in order to reach the solvent. Its BS shows a complex network of water molecules buried in the growth direction, composed of four waters (A320, A331, A333, and A360) (*Figure E3, Appendix E*). AquaPELE densities for the water sites of interest (A319 and A336) were lowered by 0.67 and 0.13 (*Table 3.10*). The algorithm could also predict the previously mentioned network, with densities of 1.0 , 0.1 , 0.85 , and 0.23 , respectively (*Figure E3, Appendix E*); it revealed lower densities in A319 and A336 than in the HSP systems (*Table 3.10*). Moreover, the growth of the fragment vanished both molecules and predicted at the same time the new distribution of the waters' network (A305, A341, A350, A351, and A414 with $r_{cc} = 1.03, 0.58, 0.38, 0.93, \text{ and } 1.24 \text{ \AA}$ and $\rho = 1.0, 0.14, 0.22, 0.24 \text{ and } 0.19$ respectively) (*Figure E3, Appendix E*).

Table 3.10. Structural comparison between X-ray hydrated sites and simulated water clusters. Checkmark (✓) indicates a distance below 1 Å between the average coordinates of the oxygen atom (from the water molecule) and its crystallographic pose ($\Delta r_{cc} < 1$ Å). A single cluster was detected for both explicit waters when one density extends over two different water IDs (i.e., HSP90 (1), A249, and A286: $\rho = 1.0$). Notice that the last column shows the difference in density ($\Delta\rho$) in contrast with the *only aquaPELE* simulation (ρ).

Systems	PDB scaffold	Water ID	ρ	$\Delta r_{cc} < 1$ Å	PDB ligand	$\Delta\rho$	
HSP90 (1)	3RLQ	A249	1.0	1.84	3RLR	-0.45	
		A286		1.25		-	
HSP90 (2)	2XAB	A2246	1.0	✓	2XJG	-1.0	
		A2115	0.68	✓		-0.03	
BRD4	5I80	A319	0.67	✓	5I88	-0.67	
		A336	0.13	✓		-0.13	
TAF1	5I29	A1891	0.09	✓	5I1Q	-0.06	
						-0.44	
		A1860	0.49	✓	6BQD	-0.09	
						-0.36	
SiaP WT	2V4C	A2346	0.07	✓	3B50	-0.07	
				✓	2WYK	-0.07	
CHK1	2C3L	A2056	0.25	✓	2C3K	-0.25	
		A2127	1.0	1.3		-0.91	
		A2052		2.24			
		A2043	-	-		-	
Control	HSP90 (1)	3RLQ	A249	1.0	1.84	-	+ 0.03
			A286	1.0	1.25	0.0	
	HSP90 (3)	-	A1	1.0	✓	3RLP	-0.03
			A3	0.80	✓		+0.20

Likewise, in TAF1, we also encountered a network of water molecules near the growing area (*Figure E4* of *Appendix E*), obtaining a reduced density of the waters of interest ($\rho = 0.09$ and 0.49). By utilizing radii of 10 Å, the simulation predicted two other water clusters belonging to the cited crystallographic water network: A1833 ($\rho = 1.0$ and $r_{cc} = 0.62$ Å) and A1952 ($\rho = 0.26$ and $r_{cc} = 0.64$ Å). When growing the first fragment the

cluster density sank to negligible values ($\rho = 0.03$ and 0.05) and three crystallographic water locations were predicted at once: A1815, 1859 and 1867 with $\rho = 0.08$, 1.0 and 0.4 , and $r_{cc} = 0.57$, 0.88 , and 0.84 Å respectively (*Figure E4 of Appendix E*). The growth of the second fragment displaced A1891 while the other persisted with a residual density of 0.13 . Again, the three same explicit water poses were predicted (the names are changed to A1808, A1812, and A1817), obtaining densities of 0.08 , 1.0 , and 0.4 ; and r_{cc} of 0.21 , 0.74 , and 0.63 Å, respectively (*Figure E4 of Appendix E*).

SiaP WT was considered one of the most challenging systems due to its holo-structure, making it difficult to access the bulk solvent. Moreover, its BS was extremely hydrophilic, possessing many water molecules surrounding the ligand (*Figure E5 of Appendix E*). The radii of the water region had to be set to 12 Å to reach the solvent, being the largest one in all simulations. Even though the goal was only to move one water molecule, two were added to form HBs networks. AquaPELE results showed a low-density cluster ($\rho = 0.07$), probably due to the high number of hydrated sites around the ligand. However, six water sites within the BS were predicted (A2011, A2017, A2021, A2113, A2192, and A2212 with $\rho = 0.60$, 0.13 , 1.0 , 0.07 , 0.33 , and 0.05 ; and $r_{cc} = 0.73$, 0.70 , 0.48 , 0.75 , 0.46 , and 0.74 Å respectively) (*Figure E5 of Appendix E*). The growth of the first fragment fully displaced the perturbable water A2346, finding a network of three water molecules around it (A328, A347, and A362 with $\rho = 1.0$, 0.32 , and 0.68 ; and $r_{cc} = 0.56$, 0.61 , and 0.36 Å) (*Figure E5 of Appendix E*). Similarly, when growing the second fragment, the residual cluster disappeared, and the position of the four water molecules network was also predicted (A2009, A2019, A2104, and A2227 with $\rho = 1.0$, 0.23 , 0.22 , and 0.30 ; and $r_{cc} = 1.49$, 0.51 , 1.01 , and 1.33 Å, respectively) (*Figure E5 of Appendix E*).

The last case, CHK1, presented a set of relevant properties that make it a challenging system: an exposed BS, the largest fragment, and the most significant number of molecules to perturb (4). Due to the wide-open BS, the water region radii had to be set to 11 Å. AquaPELE simulations predicted two relevant water clusters: one close to A2056 ($\rho = 0.25$, $r_{cc} = 0.18$ Å) and another in between of A2056 and A2127 ($\rho = 1.0$, and $r_{cc} = 2.24$ and 1.3 Å and respectively) (*Figure E6, Appendix E*). The growth of the big-sized fragment made the cluster less populated disappear and remarkably reduced

the density of the second one to 0.09, pushing the clusters with higher populations to solvent-exposed areas.

Control systems. The first control study consisted of using the HSP (1) system (X-ray 3RLQ) to expand a hydrophilic fragment (hydroxyl) instead of the already tested hydrophobic methyl moiety. As the aquaPELE simulation was already computed in the previous-mentioned test, we only executed the growing part using identical configuration parameters. Results correctly identify a cluster of waters close to the A249, keeping the same density value as we found in the aquaPELE simulation. Hence, the explicit water stayed and was not displaced. Additionally, at 1.21 Å of distance from A286, a new cluster of water molecules appeared, but its density was residual ($\rho = 0.03$). This result revealed that our method could displace or not the water molecules depending on the physicochemical properties of the grown fragment.

In the second control, we performed a self-growing exercise to assess whether our algorithm can preserve the position of water molecules (and reproduce the X-ray structure) after growing a fragment close to them. Here, we also use a water region with a radius of 10 Å. AquaPELE simulations identified two hydrated sites tight to A1 ($r_{cc} = 0.26$ Å) and A3 ($r_{cc} = 0.46$ Å) with $\rho = 1.0$ and 0.8, respectively. Interestingly, the same sites were suitably recognized when growing the fragment, but the cluster's density close to A3 was boosted to 0.97 (A1 $r_{cc} = 0.36$ Å; A3 $r_{cc} = 0.19$ Å). Thus, the allocation of explicit waters of the original X-ray (3RLQ) was nicely reproduced.

2.2. Growing and scoring on hydrated systems

2.2.1. Methods

Benchmark design and preparation. In this second part, we performed a growing and scoring exercise on three congeneric series of HSP90 inhibitors with known experimental binding affinities. However, only a few of them had available X-ray structures. *Figure 3.16* defined a common structural scaffold (S1, S2, and S3) for each series, and then, to assemble the complete set of molecules, we grew different substituents on R1, R2, R3, and R4 sites (*Tables 3.11*). Scaffold structures were constructed from an X-ray that contained the shared substructure by replacing R-groups

with hydrogen atoms. S1 series was initialized from 3RLP X-ray (compound *Kung-7*), S2 series from 3RLQ (compound *Kung-13*), and S3 from 2XAB. Systems were prepared identically than in the structural section (visit *Section 2.1.1* of *Chapter 3*).

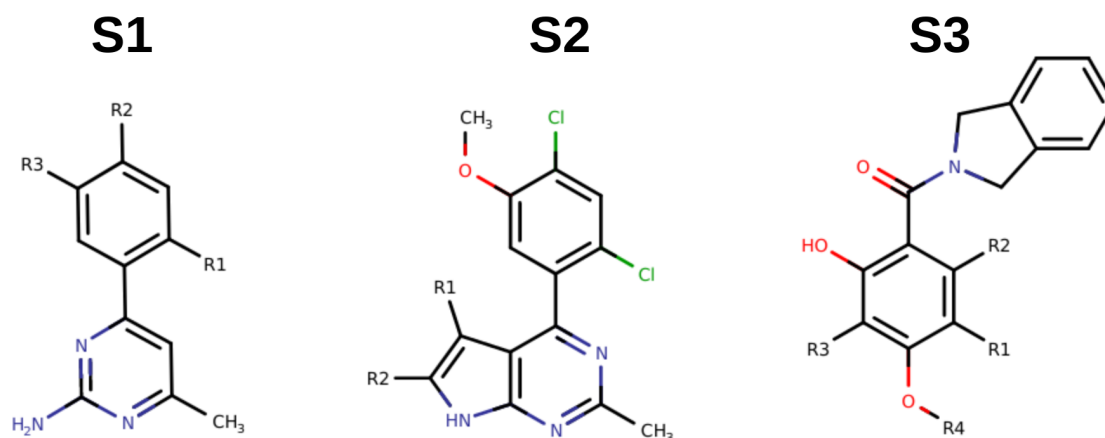


Figure 3.16. HSP90 common scaffolds for the three series of inhibitors used in the scoring benchmark. S1 and S2 series were obtained from Kung et al., and S3 from Woodhead et al. studies (Kung et al., 2011; Woodhead et al., 2010). The content of R1, R2, R3, and R4 moieties for each series are specified in *Tables 3.11*.

Simulation configuration. To properly analyze the effect of the explicit waters on FragPELE predictions, three different simulation conditions were employed: no explicit waters (NW), with fixed waters (FW), and also perturbing the water molecules close to the region of growth with the aquaPELE algorithm (W). The same criteria from the previous section were utilized to choose the explicit waters to keep fixed.

AquaPELE simulations were prepared following the exact configuration of the structural section by selecting two perturbable water molecules, but, in this case, as we were not interested in structures, simulations started from their crystallographic pose. When more than one fragment has to be grown to construct the entire compound, we run intermediate growing simulations to add R-groups until only one substituent is missing. These intermediate simulations pretend to be faster than regular FragPELE growing, aiming to grow the R-group quickly with a reduced sampling (the latest simulation will be the explorative one). Here, the fragment is placed at 33% of its final size, starting in the second GS and finalizing at the fifth. Ligand translation and

rotations (perturbation) are suppressed, allowing only side-chain prediction and energetic minimizations to keep the scaffold in a similar position to the original one but permitting at the same time the expansion of substituent. Compounds that shared the same R-group (and scaffold) were joined together to run solely one intermediate simulation and reduce the number of calculations. The growth of the last R-group was performed following the standard FragPELE protocol (explained in the previous section) but lowering the number of MC sampling steps to 50 (instead of 400). Our interest is to explore the energies and not the hydrated sites. Consequently, based on the first FragPELE scoring benchmark, we consider that a sampling of 2350 MC steps (50 x 47 parallel simulations) produced enough data to get valid predictions in a reasonable computation time.

In the S1 series, chlorides and oxymethyl fragments were grown onto different positions of the scaffold's benzene ring (visit *Figure 3.16* and *Tables 3.11*). Waters surrounding the other rings were fixed when involving the protocols W and FW. Differently, in protocol W, explicit waters close to the growing regions were defined as perturbable. In the S2 series, we focused the fragment's growth on another region, close to the pyrrolo(2,3-d)pyrimidine ring. Therefore, waters neighboring this area were mobile while the waters close to the benzene ring were fixed. Lastly, in S3, just a single water molecule that coordinated a ligand-protein interaction was fixed. The complete list of fixed and perturbable water molecules for each system is reported in *Table 3.12*.

Tables 3.11. Calculated ΔG from experimental K_i 's from Kung et al. and Woodhead et al. studies (Kung et al., 2011; Woodhead et al., 2010). Values were obtained applying the formula $\Delta G = RT \cdot \ln(K_i)$, where $T=298K$.

S1				
Compound	R1	R2	R3	ΔG (kcal/mol)
Kung-4	OCH ₃	H	H	-6.95
Kung-5	OCH ₃	Cl	H	-7.59
Kung-6	Cl	Cl	H	-8.80
Kung-7	Cl	Cl	OCH ₃	-9.34

S2			
Compound	R1	R2	ΔG (kcal/mol)
Kung-10	COCH ₃	H	-9.13
Kung-11	F	H	-9.13
Kung-12	Cl	H	-9.13
Kung-13	CN	H	-10.08
Kung-14	H	CH ₃	-7.77
Kung-15	Cl	CH ₃	-9.54
Kung-16	CN	CH ₃	-10.25
Kung-17	CN	CH ₂ CH ₃	-10.66

S3					
Compound	R1	R2	R3	R4	ΔG (kcal/mol)
Woodhead-2	isopropyl	H	H	H	-12.63
Woodhead-5	isopropyl	H	H	CH ₃	-10.66
Woodhead-6	CH ₂ CH ₃	H	H	H	-11.85
Woodhead-7	cyclopropyl	H	H	H	-11.44
Woodhead-8	sec-butyl	H	H	H	-11.61
Woodhead-9	<i>tert</i> -butyl	H	H	H	-12.26
Woodhead-10	Cl	H	H	H	-11.44
Woodhead-11	isopropyl	H	F	H	ND
Woodhead-12	isopropyl	F	H	H	ND

Tables 3.12. Fixed and perturbable waters for each series (S1, S2, and S3).

Series	Fixed waters IDs	Perturbable waters IDs
S1	A1, A3, A4, A244	A6, A261
S2	A237, A258, A313	A249, A286
S3	A2118	A2115, A2246

Score calculations. Simulation results were analyzed following the same clustering strategy as in the structural *section* (described above). We computed six scoring criteria based on PELE's interaction energies to determine which one correlates the best with experimental energies. On the one hand, the cluster selection was based on the lowest energy value: (1) the percentile 5 of interaction energies (P5), (2) the percentile 25 (P25), and the mean interaction energy (MBE). We used the reported energy value as a scorer in all of them. On the other hand, we based the clustering selection on population instead of energies, using the same energies calculations: the percentile 5 (POP5), the percentile 25 (POP25), and mean interaction energy (POPMBE).

2.2.2. Results

Scoring results from structural free energies. Systems from the structural benchmark not only reported X-rays but changes in binding energies (free energies) and IC50 values were included (*Figure 3.13*). Even though it was not the primary intent of these simulations, we drew on them to assess whether our scores could correctly predict or not their energetic differences.

The six scores were computed for the growth of 5 fragments (A to B, C to D, J to K, J to L, and M to N, as labeled in *Figure 3.13*) from the simulations that combined FragPELE with the aquaPELE algorithm. BRD4 and TAF1 were discarded from the analysis to make the results comparable, as they reported changes in IC50 instead of $\Delta\Delta G$. Correlations between experimental free energies and the predicted values are summarized in *Table 3.13*. Results showed an adequate predictive power for all scoring methods, with R^2 above 0.8. Nevertheless, the sample size of these predictions is relatively small, and consequently, we can not extract reliable conclusions. In the next

section, we will extend these symbolic results on a more extensive set and compare the calculations with or without aquaPELE and how it enhances by using merely the implicit solvent approach.

Table 3.13. Correlation between experimental free energies from the structural set and predicted data (six scores, see *Score calculations* of *Methods* section). These values were computed with an $n=5$.

Scoring method	r	R^2
P5	0.91*	0.83*
P25	0.99*	0.98*
MBE	0.96*	0.92*
POP5	0.91*	0.83*
POP25	0.93*	0.86*
POPMBE	0.91*	0.83*

* p -value < 0.05 (t-test). Significant association between experimental and predicted values.

Scoring results from the S1, S2, and S3 series. The correlation between experimental and predicted free energies for the S1, S2, and S3 series by employing the three experimental conditions (W, FW, and NW) is summarized in *Tables 3.14*.

S1 series reported the best results, with Pearson's correlation close to 1. Notably, incorporating water molecules into the simulation (W or FW) improved the quality of the predictions compared with the NW protocol. The best correlations were obtained by integrating FragPELE with aquaPELE and scoring with the POP5; still, P5, P25, and POP25 showed similarly good performances (*Tables 3.14, S1*).

In the S2 series, the same pattern can be observed: protocols including explicit waters (W and FW) outperform the ones without them (NW) (*Tables 3.14, S2*). The best score is reported in the FW protocol (scoring method MBE). However, r values can quickly drop, indicating a high variability depending on how the cluster is selected, and actually, the significance of the correlation can be affected. In this case, choosing the scorer cluster by population reported a lower and non-significant correlation than using

energetic criteria (Table 3.14, S2). Contrary, W protocol results are more stable than FW and NW since the correlation variance is also higher.

Tables 3.14. Correlation between experimental calculated free energies and predicted data by applying the six scores (see *Score calculations* section) in the S1, S2, and S3 congeneric series. Protocols used: W) aquaPELE+FragPELE, FW) FragPELE with explicit waters fixed, NW) FragPELE without explicit waters (only implicit solvent).

S1 (n=4)			
Scoring method	r (W)	r (FW)	r (NW)
P5	0.98*	0.96*	0.71
P25	0.97*	0.93	0.72
MBE	0.92	0.91	0.75
POP5	0.99*	0.96*	0.76
POP25	0.97*	0.93	0.78
POPMBE	0.92	0.91	0.76
Original	0.55	0.35	0.52

S2 (n=8)			
Scoring method	r (W)	r (FW)	r (NW)
P5	0.78*	0.75*	0.65
P25	0.72*	0.83*	0.58
MBE	0.78*	0.84*	0.80*
POP5	0.76*	0.61	0.63
POP25	0.73*	0.63	0.50
POPMBE	0.73*	0.61	0.48
Original	0.78*	0.82*	0.77*

S3 (n=7)			
Scoring method	r (W)	r (FW)	r (NW)
P5	0.58	0.39	0.24
P25	0.46	0.29	0.31
MBE	0.48	0.30	0.27
POP5	0.58	0.32	0.31
POP25	0.44	0.20	0.31
POPMBE	0.50	0.16	0.22
Original	0.51	0.26	0.27

* **p-value < 0.05 (t-test)**. Significant association between experimental and predicted values.

The S3 series showed the lowest correlation compared with the previous two but followed the same trend: perturbing water molecules retrieve the best results. As only one water molecule was fixed, differences between perturbing (W) and not perturbing them (FW and NW) are more prominent than in the previous two series (*Tables 3.14*). However, in this case, including aquaPELE has not been enough to correlate predicted and experimental values accurately (*Tables 3.14, S3*).

Looking at *Figure 3.17*, one can quickly observe the general view of the results depending on the scoring method and the protocol employed. According to the results mentioned above, it is clear that in hydrated systems, the best strategy is fixing crucial waters and perturbing the ones closer to the region of growth (W). Additionally, this protocol is the most stable in terms of scoring (*Figure 3.17*). When looking at the deviation depending on the scoring method used, the W protocol showed a standard deviation of only 0.03. In contrast, the other two protocols, FW and NW, have a higher dispersion (0.06 and 0.04, respectively). FW results are slightly worse than W, but both are audibly outperforming NW (*Figure 3.17*); therefore, these results highlight the importance of keeping track of explicit water molecules in molecular simulations and their effect on the prediction's trustworthiness.

Regarding scoring methods, both P5 and MBE reported the highest averages computed by combining all protocols (*Figure 3.17*). Still, when perturbing water molecules, the best way to compute the score is using the 5th percentile of interaction energies and selecting the cluster by any of both criteria, population or energy, as the most populated cluster coincides with the one with the lowest energies. This rule is not followed when waters are not perturbed, and subsequently, the correlation drops when selecting clusters by population (POP25 and POPMBE) (*Figure 3.17*). In those cases, it is recommended to pick clusters by energy instead.

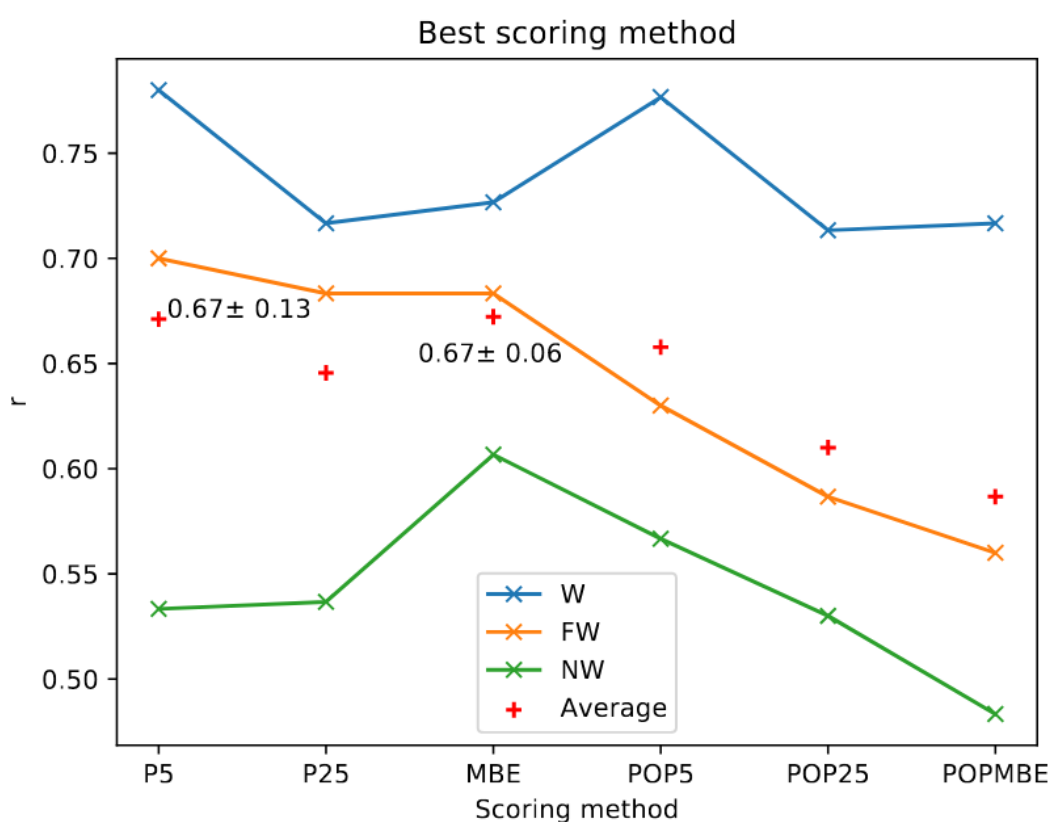


Figure 3.17. Average Pearson's correlation (r) for each protocol type (W, FW, and NW) and scoring methods (P5, P25, MBE, POP5, POP25, and POPMBE). Notice that in P5 and MBE (the two highest), the average \pm standard deviation is indicated. No error bars are shown to help the graph's visualization.

To confirm that the combination of techniques improves our correlations equally to the previous benchmark, we computed one-tailed paired t-tests (threshold 5%) with the SciPy python package (Virtanen et al., 2020) using the Pearson correlations for each

system as samples. This test confirmed that the inclusion of water molecules, combining FragPELE with aquaPELE, results in a higher correlation with experimental values than using simply FragPELE with implicit solvent models (p -value = 0.029 < 0.05).

With the results of this study, integrating both in-house FragPELE and aquaPELE methods, we can better characterize ligand potencies in a less labor-intensive and time-consuming strategy than using them individually.

Chapter 4. SilicoDerm: an industrial drug discovery project

In 2018 started SilicoDerm, a collaborative project between the pharmaceutical company Almirall and our team, the Electronic and Atomic Protein Modeling (EAPM) group from Barcelona Supercomputing Center (BSC). The main goal of this research was to develop and optimize new VS techniques to identify potent compounds to treat dermatological inflammatory diseases.

In January of 2019, I got involved in the project to pursue one of the main objectives of this thesis: participate in an actual drug design project devising new VS strategies and test the recently developed ligand growing software.

Due to the confidentiality of this project, we cannot publish identifiers, names of the targets, or 2D or 3D images that could provide information on the structure of any molecule studied. Thus, anonymized labels will be used to correctly identify receptors, ligands, or any confidential data.

1. Precedents

This project is based on studying two different targets from the kinases family, labeled as *Kinase 1* and *Kinase 2*. Recognizing specific low MW binders for these receptors is still challenging compared with other targets of the same family. In VS campaigns, researchers screen millions of compounds, and after several filters, experts visually inspect and manually choose the desired molecules for further *in vitro* testing. This project aims at a new protocol to optimize VS methods by applying advanced induced-fit simulation techniques on a subset of compounds (~1000) pre-selected from regular docking-based HTVS. After this process, we expect to enrich the activity range of the top 50-100 compounds. I joined this already running project by applying FragPELE in a H2L exercise in *Kinase 1*. Later on, we participated in the second target, *Kinase 2*, from scratch.

2. FragPELE on Kinase 1

Complementing the initial screening performed on *Kinase 1*, we used FragPELE to optimize the compound Q1-694. This high-affinity candidate came from a previous study performed by Almirall, where its experimental affinity ($IC_{50}=6nM$) and binding mode were determined. Thus, our collaborators provided as starting structure a co-crystallized X-ray to later grow several molecular decorators with FragPELE and try to increase its affinity against *Kinase 1*.

2.1. Methods

FragPELE validation on X-ray. Initially, we structurally validated FragPELE (version 1.0) for the current target by performing a self-growing onto Q1-694 crystal. The crystal was prepared by executing Protein Preparation Wizard from the Schrödinger package (Sastry et al., 2013), including missing side-chains and preserving explicit water molecules around the ligand BS (5 Å beyond the ligand center of mass). Hydrogens were incorporated, and H-bonds optimized with PROPKA (Olsson et al., 2011) at 7 pH. The general structure of the ATP BS of any kinase can be observed in *Figure 4.1*. Most of the inhibitors are designed to settle this site, maintaining the HBs with the hinge region between the two lobes of the kinase (D. Huang et al., 2010; Kedika & Udugamasooriya, 2018).

In our case, to perform the self-growing exercise, we selected as a scaffold the warhead region of Q1-694 that was interacting with the hinge region, which is key to stabilizing the binding mode. Then, a fragment of 9 heavy atoms was removed to grow it. FragPELE simulations were configured following the standard protocol: 10 GS, containing 47 independent MC simulations of 6 PELE steps, box size of 4 Å radii, translations of 0.10-0.05 Å, and rotations of 0.05-0.02 radians, and a final sampling simulation of 20 PELE steps.

RMSD of the ligand heavy atoms for the lowest binding energy pose was compared against the X-ray structure. Fragment, scaffold, and total RMSD values were 1.13 Å, 2.14 Å, and 1.82 Å, respectively, not showing significant structural differences between model and crystal. Additionally, all hinge and fragment interactions were conserved

after in our model, so we decided to move ahead and perform a second retrospective study to assess FragPELE's scores.

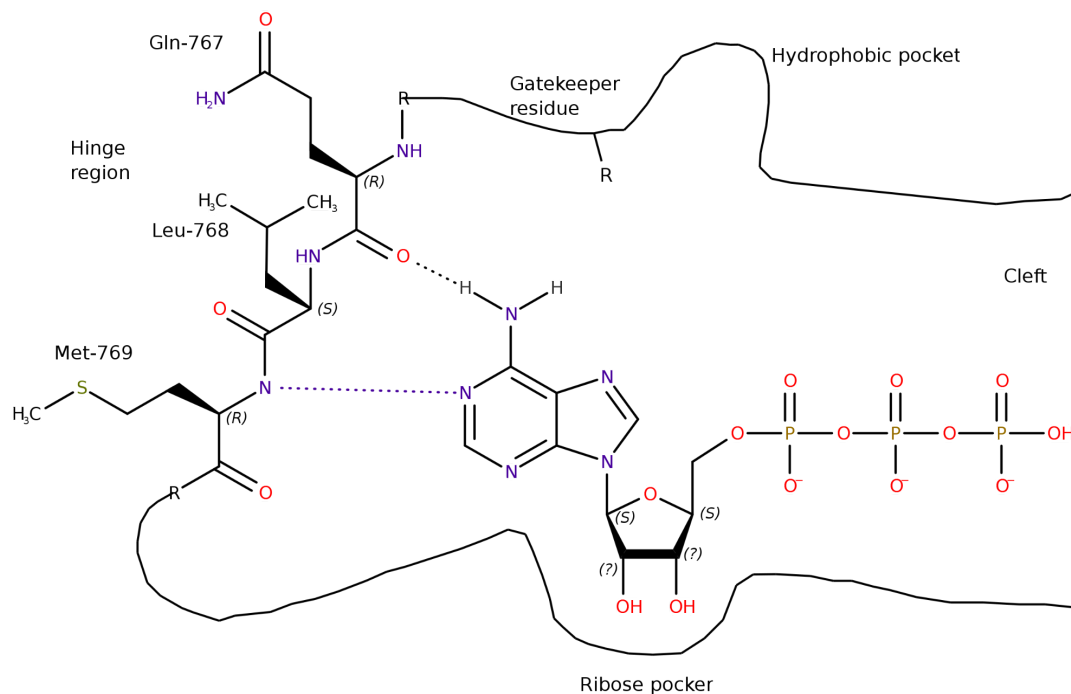


Figure 4.1. Binding of ATP to kinase EGFR. Dotted lines showed the ATP-hinge interaction. Image licensed by [CC BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0/) (source: https://commons.wikimedia.org/wiki/File:Binding_of_ATP_to_kinase_active_site_of_EGFR.svg)

Experimental series for the retrospective study. The experimental series consisted of 28 ligands from Almirall with known IC₅₀ values. According to the distribution of their R-groups, compounds could be classified into three different subseries. These schematic views can be seen in *Figure 4.2*, where they all share a common scaffold region. However, *series 2* included two functional groups as part of the scaffold, one not present in *series 1* (“Group 2”) and the other not shown in *series 3* (“Group 1”). Growing sites were set in different regions depending on the series, as stated with the *R* symbol in *Figure 4.2*. IC₅₀ values were converted to pIC₅₀ (logarithmic scale) to compare with energy values. The range of activities of the two first sets was relatively narrow: 7.28-7.51 in the first and 7.92-9-14 in the second, and slightly larger in the third one: 6.07-8.55 (with two of them inactive) in the third. A different view of these distributions

can be observed in *Figure 4.3*. Obtaining good correlations considering these tiny ranges was a challenging task when treating them separately.

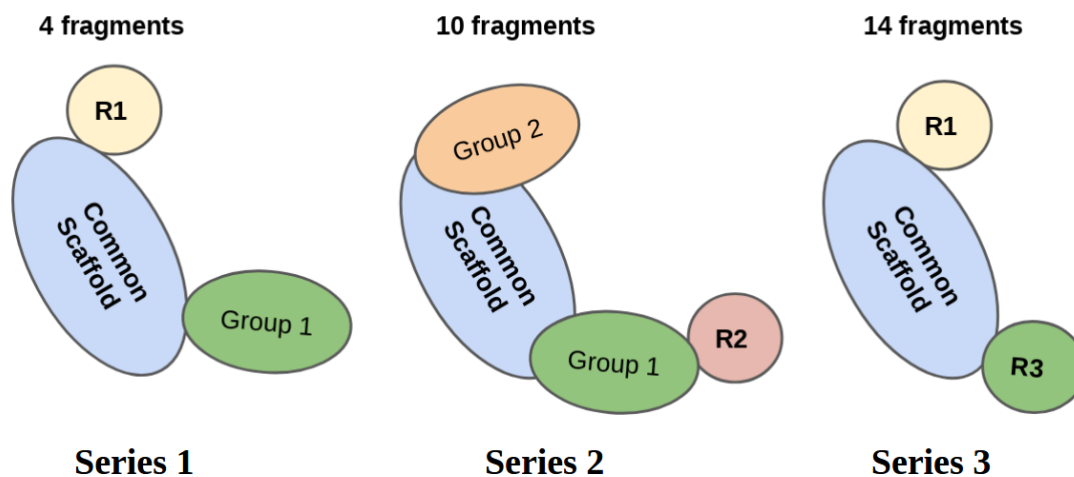


Figure 4.2. Schematic representation of the fragment's distribution in Almirall's congeneric series. Notice that the growing site in each series was identified with the *R*.

pIC50 values distribution

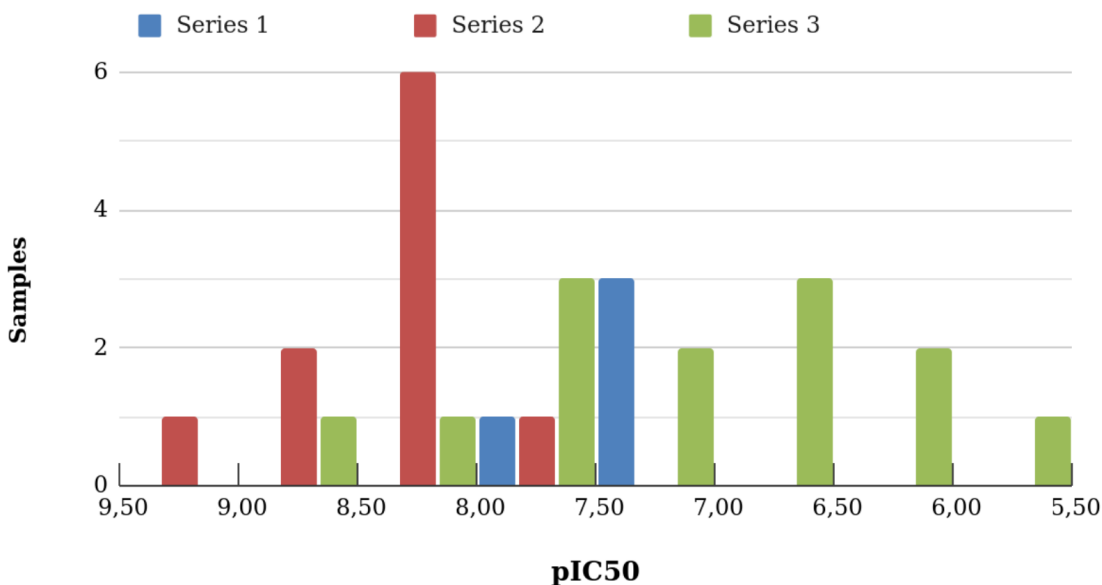


Figure 4.3. Distribution of experimental pIC50 values for the three sets of inhibitors for *Kinase 1*.

Ligand Q1-694 followed the identical distribution of R-groups as the series 1; then, the scaffold pose of this X-ray was employed to grow the different fragments onto it.

FragPELE simulations. We propose multiple initial states for *FragPELE* simulations: (1) from X-ray with explicit waters, (2) from the lowest binding energy pose with explicit waters, (3) from X-ray without explicit waters, (4) from the lowest binding energy pose without explicit waters. We run three replicas changing the random number generator seed for each initial condition, accounting for 12 simulations per ligand. To generate the initial structures, we first deleted the R1 region from the Q1-694 X-ray to create the structures coming from the X-ray. Second, we run a short PELE simulation of 20 MC steps with low translation and rotations to minimize the previous structure, picking the lowest interaction energy pose. This simulation was done to rectify the crystallographic binding mode, as the distance to reach the essential interaction with the hinge was too large. The resultant pose with the interaction recovered was the input structure in *series 1* (initial states 2 and 4) and 3 (by removing *Group 1*). Notice that *series 3* contained two different growing sites; therefore, two successive *FragPELE* simulations should be executed, employing the structure recovered from the first simulation as input for the second one.

We run standard *FragPELE* simulations configured with low translation and rotations of the scaffold region. Fragments of the R1 growing site were grown, building all ligands from series 1. The lowest interaction energy pose obtained for the ligand whose R1 group matched with *Group 2* was used as the scaffold pose for the second series and the X-ray (which also contained *Group 2*). Due to the closeness between the fixed waters and the R2 growing site, we had to erase these molecules from the model to avoid fragment-water collisions (limitation of *FragPELE* 1.0). To determine the effect of water molecules in the system, we compared results with and without explicit waters for series 1, where we observed no differences ($R^2=0.996$). Thus, we moved forward without water molecules (6 simulations per ligand).

2.2. Results

We did not observe significant differences between simulations beginning from X-ray or minimized poses, so we computed the mean and standard deviation for the six replicas of each ligand. Results were compared and correlated with experimental and plotted in *Figure 4.4*. The figure shows that Pearson correlation coefficients were low or even

negative when computing them individually for series with tiny ranges (*series 1 and 2*). In *series 3*, where the range of activities is more extended than the others, we got a better correlation (however, not significant due to the poor number of samples). We determined that *series 2* were overscored due to larger fragments, some of them including positive charges, making them non-comparable.

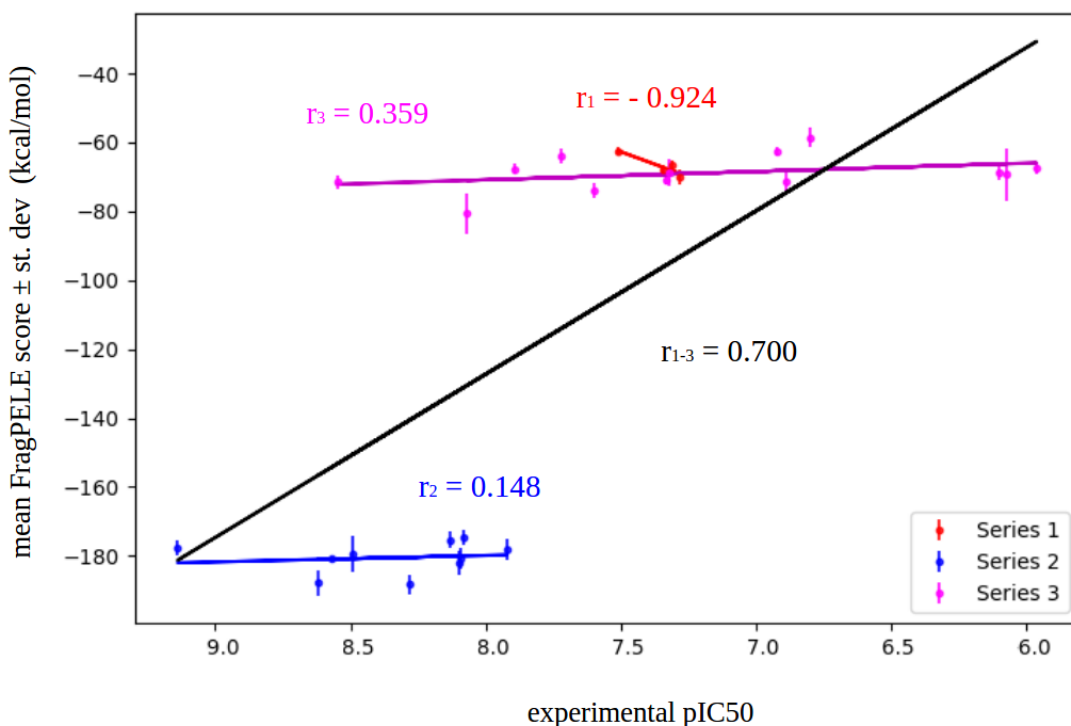


Figure 4.4. Scatter plot comparing FragPELE scores against experimental pIC50 values for the three series of ligands for the Kinase 1.

Furthermore, inactive compounds were also scored as low-affinity compounds, showing values of $-64,5 \pm 1,5$ and $-69,4 \pm 2,97$ kcal/mol. Finding false positives is quite familiar with FragPELE protocol, where ligands barely move and cannot abandon the BS. However, this fact is compensated mainly by the lower score values.

Considering the results obtained in *series 3*, we apply the protocol in a prospective study. Herein, 56 fragments proposed by Almirall were grown onto the R1 site of the X-ray scaffold structure. Results were compared with the score obtained in the X-ray self-growing simulations, as the experimental pIC50 was known (IC50 of 6.1nM; pIC50 of 8.19). Then, all scores obtained were normalized and interpolated according to the self-growing score, computing the *predicted pIC50* (Equation 4.1). A global view of the

results is seen in *Figure 4.5*. Notice that positively charged tautomers of *f1*, *f1S*, and *f1R*, showed overscored results, revealing consistency with the aforementioned validation.

$$(4.1) \text{ Predicted } pIC50 = Xray \text{ } pIC50 \cdot \frac{\text{Score Fragment}}{\text{Score Xray}}$$

The top 10 compounds were selected and sent to Almirall to consider their synthesis. Four of them were synthesized and experimentally assayed, showing IC50 values of 1100nM for *f1*, 26nM for *f6*, 120nM for *17*, and 1.5nM for *96*, being this last one four times more active than the original hit.

Predictive values (mean all simulations \pm SEM)

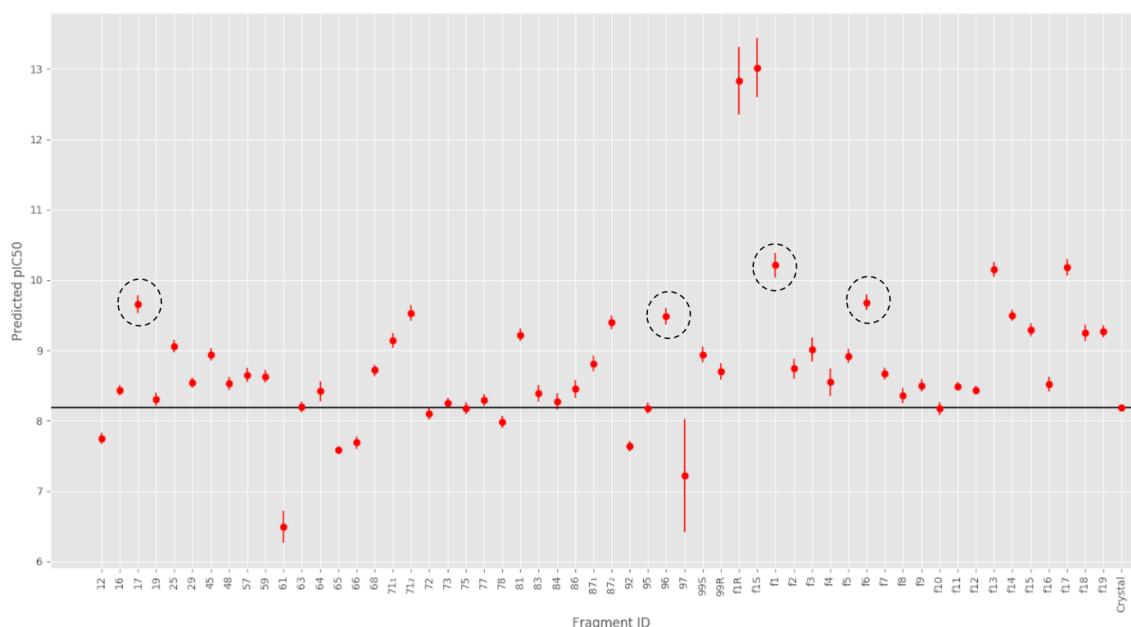


Figure 4.5. Predicted pIC50 values (score) based on FragPELE scores for each proposed fragment. The black line delimits the score of the crystal self-growing. Dotted circles indicate the compounds selected to synthesize. Fragments f1R and f1S were positively charged tautomeric forms of the fragment f1, and consequently, these values are overscored. Notice that error bars represent the standard error of the mean (SEM).

3. Study of Kinase 2

From what we learned from the *Kinase 1* study, where we basically involved an induced fit protocol on top of Glide docked results, we required a new strategy to optimize the screening against the new target *Kinase 2*. Unlike *Kinase 1*, in *Kinase 2*, there is plenty of public information available, including X-ray structures and experimental inhibition data. Two public co-crystallized protein-ligand X-rays were selected to proceed with this study, labeling them *Xray 1* and *Xray 2*. The former has the activation loop missing, while the latter includes this part but shows a slightly narrower ATP binding site. Some clues about these structural differences can be observed in *Figure 4.6*. The following section will expand the new target-specific VS optimized protocol developed for *Kinase 2*.

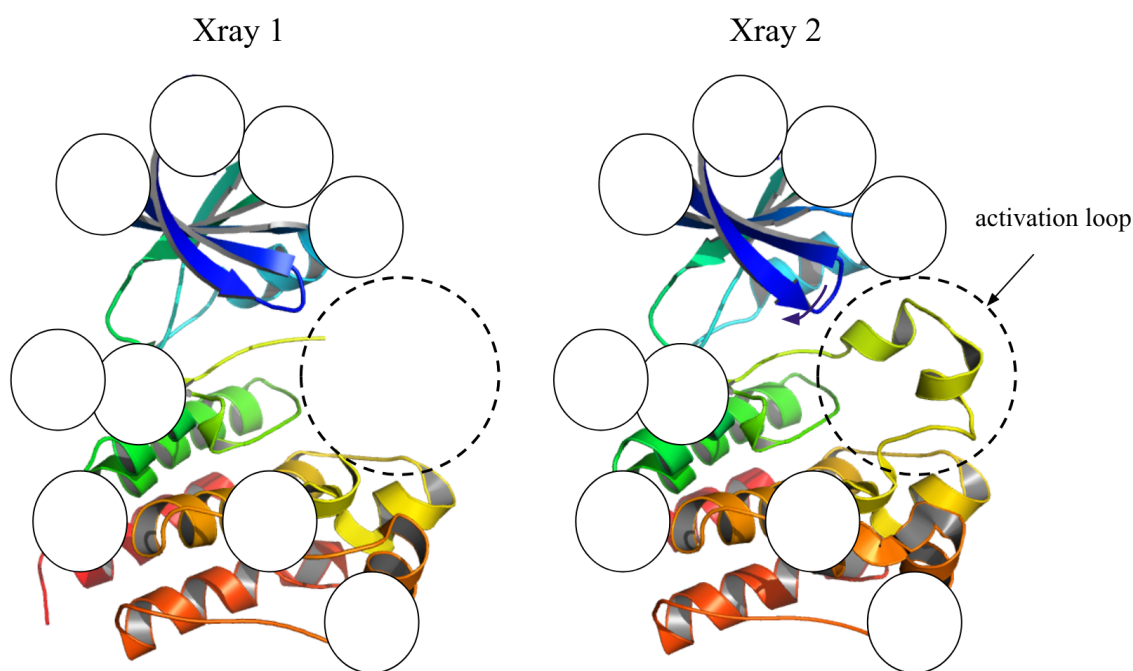


Figure 4.6. Xray 1 and Xray 2 structural differences. Overlapping white balls have been included for difficulting the Xray's identification. Created with PyMOL (Schrödinger & DeLano, 2018).

3.1. Machine learning on optimizing virtual screening pipelines

This exercise's final goal is to optimize VS pipelines by enriching our predictions through machine learning methods to identify higher affinity hit compounds than regular VS approaches. Herein, we developed a slightly different approach from the exercise performed for *Kinase 1*. In this case, taking advantage of all available knowledge and experimental data for the *Kinase 2*, we thought of developing a target-specific machine-learning-based analysis tool to classify between low (or inactive) and highly active compounds. The design of this VS pipeline consisted of several steps that we will explain in detail in the following lines.

3.1.1. Methods

Data collection. Six sets were obtained from ATP BS assays on *Kinase 2*, reported in ChEMBL (Gaulton et al., 2012, 2016) database (version April 2020). They were selected trying to reduce as much as possible the differences between experimental IC₅₀ values (picking experiments from the same laboratory, with similar conditions) in order to minimize the errors associated with their values. Due to the lack of data from similar sources, we were forced to mix at least two different conditions. After cleaning duplicates, we compiled 302 compounds that were enriched with 24 different ligands supplied by Amirall. Compounds with non-exact IC₅₀, labeled with > 20 μM or > 50 μM in the assays, were assumed inactive. Therefore, the final set was formed by 20 inactive and 306 actives, and its pIC₅₀ distribution is shown in *Figure 4.7*.

Systems preparation. Xray 1 and Xray 2 were prepared with Protein Preparation Wizard (Sastry et al., 2013), analyzing and optimizing H-bonds with PROPKA (Olsson et al., 2011) at pH 7. All explicit waters and ions were deleted from the system. Given that the hinge amino acids compose a hydrogen acceptor-donor-acceptor pattern, we require ligands showing the complementary sequence (acceptor-donor-acceptor). Thus, Epik's (Greenwood et al., 2010) algorithm from the Schrödinger package was used to forge all possible ligands stereoisomers in solution, and protonation states with two consecutive donor, or acceptor atoms, placed in the same direction, were discarded to allow the correct hydrogen-bonding pattern to interact with the hinge. After applying this filter, the state with the lowest energetic penalty was picked for each ligand.

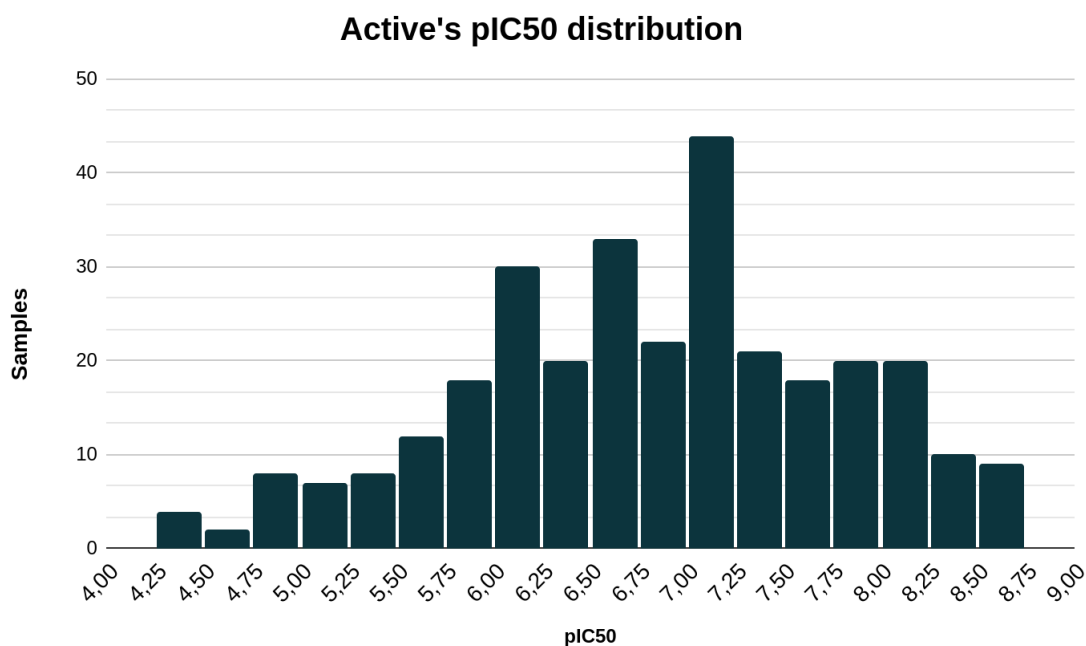


Figure 4.7. pIC50 values distribution of the 306 active compounds.

Initial screening. Ligands were docked using Glide SP (Friesner et al., 2004b; Schrödinger, 2018) onto both receptors. The center of both grids was set in the central amino acid of the hinge, with 20 Å of diameter (default value). Additionally, we imposed constraints on the HBs of the hinge. Two dockings were executed following two criteria; fitting at least one constraint (1-HB-constrained) or two (2-HB-constrained). The output was configured to write only one pose (from 5). The results between 1-HB-constrained and 2-HB-constrained groups were compared, selecting the pose with the lowest Glide score for each protein-ligand complex.

PELE refinement. Structures from docking's results interacted with the kinase hinge due to the docking constraints. Then, the previously selected poses were used as the initial one for a short induced-fit simulation with AdaptivePELE (Lecina et al., 2017). The configuration of this protocol consists of 3 epochs of 100 PELE steps, only allowing movements of the ligand within a small box of 4 Å radii from the ligand center of mass. Translations and rotations were set to 1-0.5 Å and 0.25-0.1, respectively. After each epoch, poses are clusterized by ligand RMSD values following a k-means algorithm. The structures are spawned inversely proportional to the population of each cluster, selecting more structures from the less populated clusters to initialize the new

epoch. Besides, an epsilon of 0.25 towards interaction energies was set (the lowest interaction energies spawn 25% of the poses of the new epoch). This cluster and spawning strategy enhances the best poses but also explores potential new (less populated) binding poses. This protocol required 32 processor units, running then in each epoch 32 independent PELE simulations (one per processor).

Filtering. At the end of the simulation, thousands of new ligand-protein conformations are generated. We hypothesized that its correct analysis is a clue step to understanding the binding properties of each ligand. Therefore, we followed several rounds of filtering and a final clustering step.

First, HBs were computed using a modified *MDtraj* (McGibbon et al., 2015) hydrogen bond identification algorithm based on *Baker-Hubbard* criteria (Baker & Hubbard, 1984). This algorithm sets three criteria to identify HBs: a cutoff angle $\theta > 120^\circ$, a distance $< 2.5\text{\AA}$ between hydrogen donor and acceptor atoms, and be present in at least 10% of the snapshots of the trajectory. As we aimed to compute hydrogen bond frequencies, we excluded the last criteria and independently applied the other two to each step of trajectories. Afterward, we filtered out all poses that did not fit at least two of the three key HBs with the hinge. After that, we applied an energetic selection to keep only the poses with the 25% lowest interaction energies values. Finally, a *k-means* clustering by RMSD of the ligand (bandwidth of 2.5\AA) is employed in the remaining structures. The pose with the lowest interaction energy is selected as a representative structure of the cluster. However, the lowest interaction energy is selected when more than one cluster per simulation is obtained. A plot example of this filtering process can be seen in *Figure 4.8*.

Metrics collection. From all the previous analyses, we thought of several properties that could be obtained for each protein-ligand complex to define or provide information on the system. Thus, we collected them as descriptors for each system. From the selected docking pose, we used the *glide score*. From the simulation, we took: the total number of accepted PELE steps, the mean total energy and interaction energy for all the steps of the simulation, the number of poses fitting 2HBs with the hinge, the size of the selected cluster, the minimum and the mean interaction energy of the poses within the selected cluster, the mean amount of successive steps interacting with the central hydrogen bond

of the hinge; named as *central mean resilience*. Then, we could also extract relevant information from the ligand itself: MW, number of heavy atoms, and number of rotatable bonds.

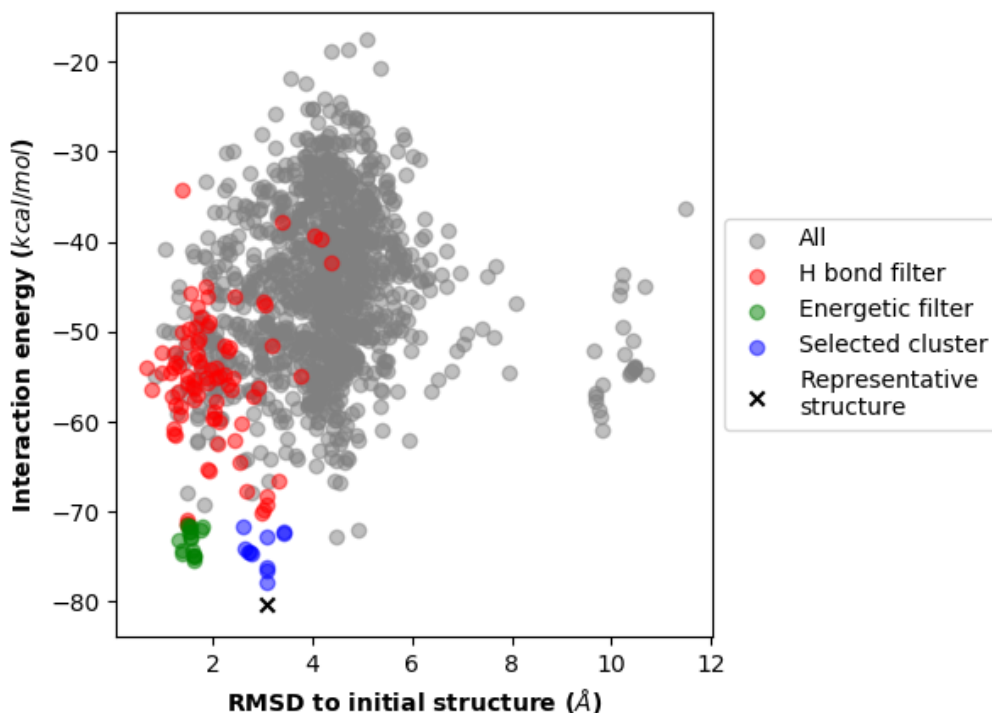


Figure 4.8. Scatter plot showing the results of the filtering of an AdaptivePELE simulation. Each dot represents a single pose, and colors illustrate the poses selected after applying filtering layers. Notice that upper filtering criteria englobe the ones below (F.ex: green dots also have passed the H bond filter).

Machine learning pipeline. Due to the size of the available experimental data and the inaccuracies of the IC₅₀ values, we decided to build a straightforward classification model to distinguish between high and low activity ligands for the *Kinase 2*. We used Pandas (McKinney & Others, 2011) and Scikit-learn (Pedregosa et al., 2011) python packages to design the ML pipeline. Therefore, we use all the descriptors mentioned above for the entire dataset of compounds (ChEMBL and Almirall sets) to train and test the model. Ligands with pIC₅₀ \geq 6.5 were classified as *highly active* (174 ligands) and the others as *low active* (132 ligands).

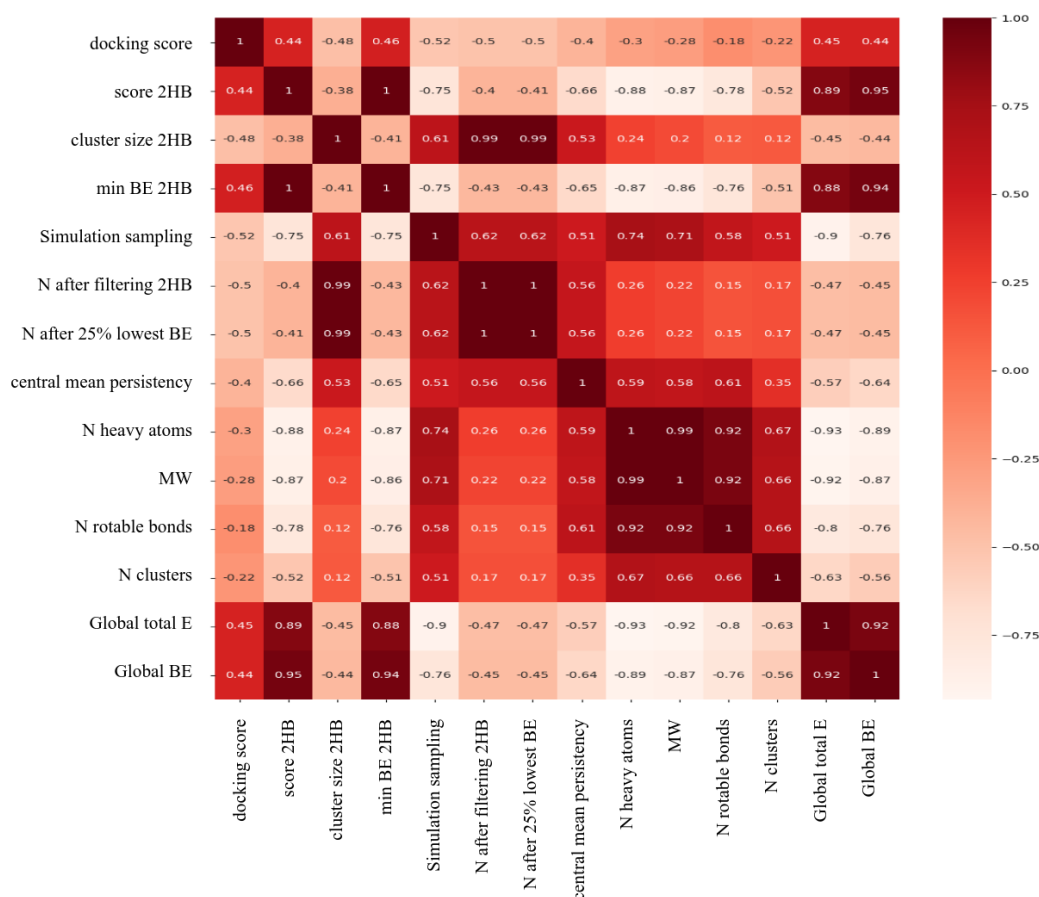


Figure 4.9. Correlation matrix between features in Xray 1 model. The numbers within squares represent Pearson's coefficients.

Before starting the training of any ML classifier, datasets were previously cleaned by transforming any *NaN* value into zeroes. Additionally, a *feature ranking* analysis is needed to know which of the available descriptors/features work better to classify between the two groups. We first tested the most common normalization algorithms (Normalizer, StandardScaler, RobustScaler, and MinMaxScaler) onto a fixed classifier model. We used the RandomForestRegressor, which can rank the features according to their importance; thus, we can obtain the ranking of features depending on the normalization procedure applied. Additionally, all features were correlated against all, building a pairwise correlation matrix of Pearson's coefficients (*Figure 4.9*). The highly correlated feature groups (>0.7 or <-0.7) must be filtered to reduce the model noise by picking only one of them per group. Then, after filtering them, in *Xray 1* and *Xray 2*

datasets, we selected and ranked features shown in *Figure 4.10* and *Figure 4.11*, respectively.

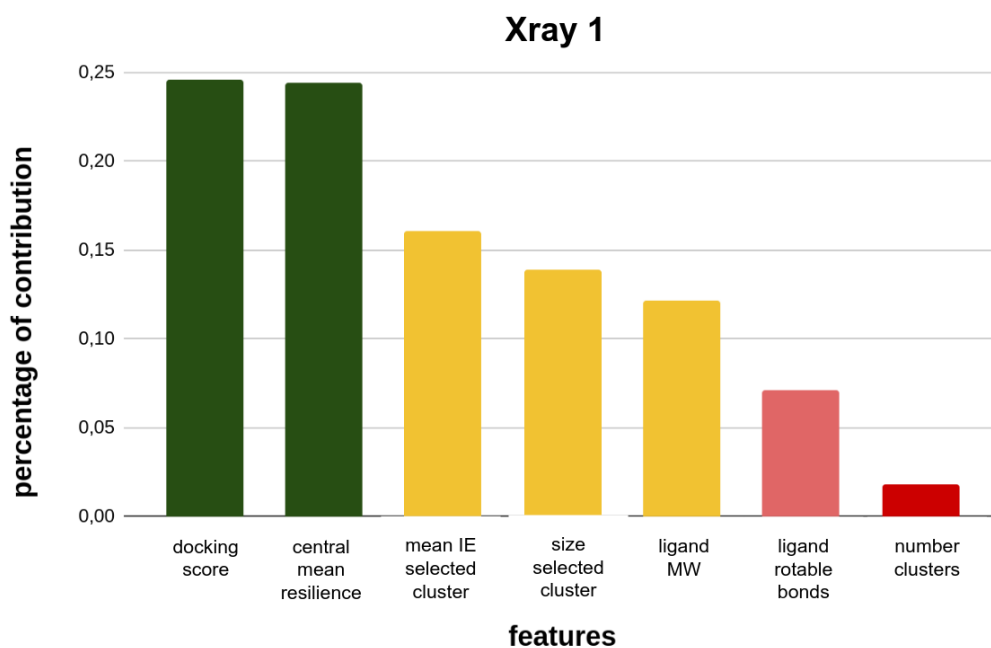


Figure 4.10. Feature ranking for *Xray 1* dataset.

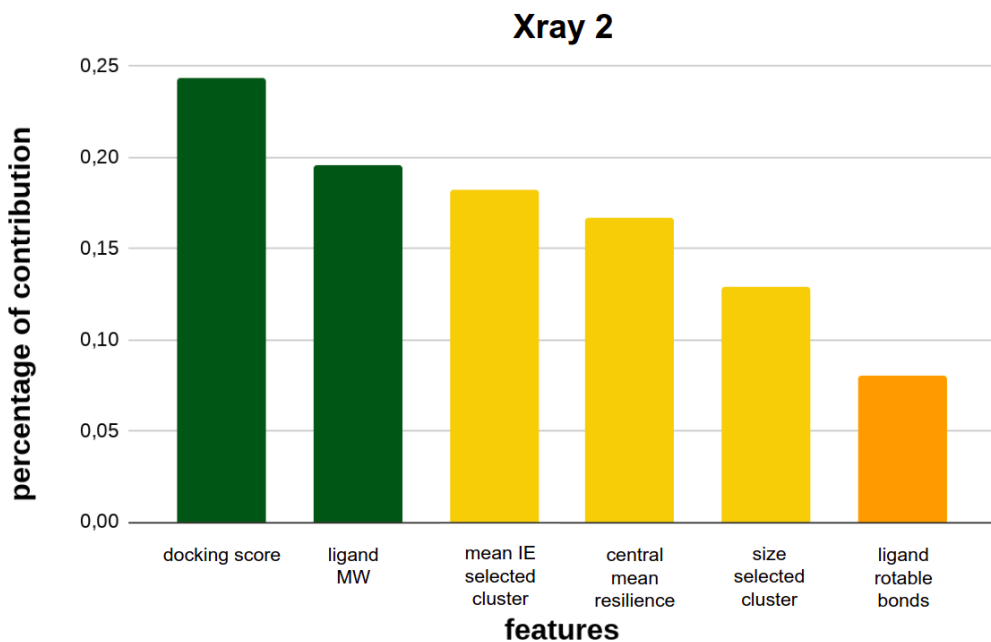


Figure 4.11. Feature ranking for *Xray 2* dataset.

To optimize and train the model, the whole dataset must be split into training (75%) and testing (25%) subsets. The former fits the model to predict labeled values, learn from the data, and optimize its parameters. The latter is needed to assess the trained model and provide an unbiased evaluation. Therefore, this subset must be only used in the definitive deployed model. The training accuracy can be improved through *cross-validation* (15-fold cross-validation). The set is split N times (15 in our case) into different subsets and then evaluated in each iteration. The sub-test set is changed in every iteration, and subsequently, the whole training set is used to test the model performance. In the end, the average value between all iterations is computed to assess the model's performance. A schematic view of this split pipeline is represented in *Figure 4.12*.

First, we picked normalization and classification models that best fit our testing set. We executed several training rounds by testing all normalization methods versus all classifiers and assessing each combination with cross-validation. For each result, we computed the mean accuracy (fraction of right classified samples), precision [$true\ positives / (true\ positives + false\ positives)$], recall [$true\ positives / (true\ positives + false\ negatives)$], and f1-score [$2 \cdot (precision * recall) / (precision + recall)$].

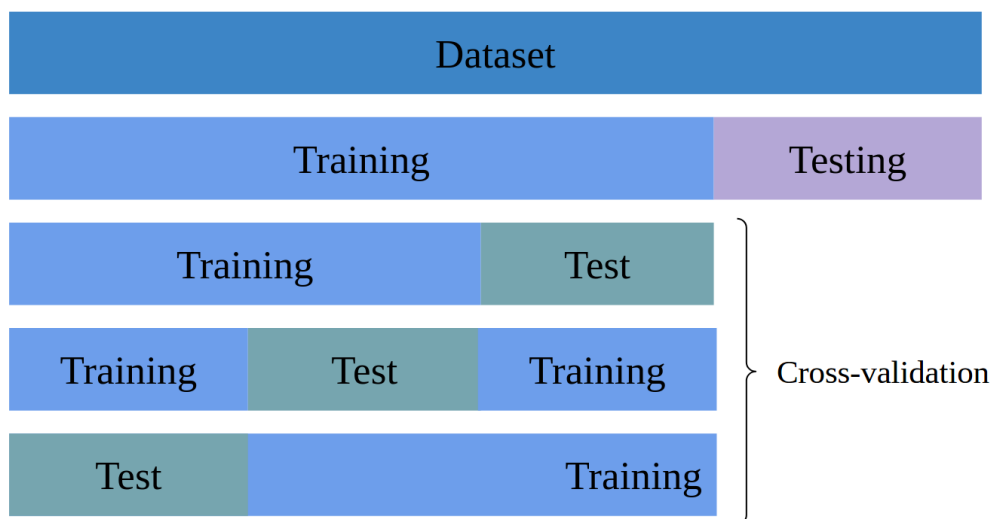


Figure 4.12. Schematic representation of the dataset splitting in ML pipelines, specifically when applying cross-validation.

The combination with the best *f1-score* average in the cross-validation was selected to proceed with the hyperparameter tuning of the model by utilizing a *grid search* approach. This type of algorithm performs an exhaustive search of a subset of specified hyperparameters lists, testing all combinations through cross-validation and automatically selecting the one with the best scores. Thus, we could obtain a completely optimized and trained classifier model for a given set.

In order to select the best number of features for *Xray 1* and *Xray 2* models, we run the previous pipeline iteratively by removing the lowest-ranked feature. For each round, we computed the 15-fold cross-validation *f1-scores* mean, picking then, for each *Xray*, the model with the highest number of features with a mean *f1-score* higher than 0.6. *Xray 1*. Results are reported in *Table F1* and *Figure F1*, as well *Xray 2* results are shown in *Table F2* and *Figure F2* of *Appendix F*.

Thus, we selected the best model with the top 4 ranked features for *Xray 1* and the top 5 for *Xray 2*. Both models used K-Nearest Neighbor (KNN) classifiers. However, *Xray 1* was normalized with the *StandardScaler* method, while the *Xray 2* model employed the *MinMaxScaler*, being both models trained similarly. Finally, they were assessed onto the training set, computing confusion matrices, accuracies, precisions, recalls, and *f1-scores*.

After selecting all models, we also wanted to check the predictive power of those models with a mean *f1-score* higher than 0.6 that was previously discarded. These test results were included in *Tables F1* and *F2*. Herein we can observe that even though a docking-based trained model retrieved *f1-scores* higher than 0.6 in the cross-validation exercise, this value drops when testing in a different set, suspecting that single-feature trained models can lead to overfitting (*Table F1*).

Lastly, a simpler model based on Morgan fingerprints (circular fingerprints) was trained following the previously defined pipeline. In contrast with the other models, these fingerprints store the 2D structure of the ligand in a series of binary digits (Rogers & Hahn, 2010). Then, the model was trained by utilizing these features to distinguish between high-active and low-active compounds. In this case, the best-selected classifier was a *Random Forest*.

3.1.2. Results

Cross-validation and test results are summarized in *Table 4.1*. Surprisingly, the best model results were obtained based on Morgan fingerprints with a high f1-score of 0.86. *Xray 1* model outperformed the *Xray 2*; however, it showed an f1-score of 0.68, clearly below the straightforward fingerprint-based model. As fingerprints only consider the ligand 2D structure, we suspected high overfitting in the model, making it non-generalizable to any external set slightly different from the trained one. Almirall provided us with a completely new small set of 16 unique compounds with known activities to prove this theory. Then, we followed the same simulation pipeline and applied the three previously trained models to the new set. Importantly, the fingerprint-based model dropped its performance significantly, while the other models kept similar values (*Table 4.2*). In addition, and somehow surprisingly, we observed a correlation between the probability of belonging to the high-active group (computed by the model) and experimental pIC50s in *Xray 1* and 2 models, showing R² of 0.17 and 0.12, respectively. Even though these are not splendid values, in the *Xray 1*, a clear outlier is breaking the correlation. When removed from the computation, it remarkably increases the R² to 0.48 (see *Figure 4.13*).

Table 4.1. Cross-validation and test results for the selected machine learning model.

Model	Cross-validation (mean ± standard error)				Test			
	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
Morgan FP	0.82±0.024	0.84±0.026	0.78±0.027	0.79±0.028	0.87	0.89	0.85	0.86
Xray 1	0.65±0.028	0.62±0.036	0.62±0.030	0.61±0.034	0.68	0.68	0.68	0.68
Xray 2	0.65±0.027	0.65±0.028	0.65±0.028	0.64±0.028	0.68	0.62	0.64	0.63

The small size of the external set made us suspect that the results could be serendipitous. Thus, we decided to observe if the correlation of the *Xray 1* model was also present in the test set. As *the data collection* section mentioned, we categorized those compounds with IC50 > 20µM or > 50µM as inactive. In order to include these compounds in the calculation and not induce an extreme bimodal dispersion of pIC50s,

we assigned them a value of 3 (mM range). Additionally, we compute the correlation between the four individual features and pIC50s. *Figure 4.14* shows the results of this reanalysis. None of the individual descriptors correlated with experimental values; however, some showed residual classification power. For instance, the mean IE and the size of the selected cluster could roughly distinguish some inactive from high-active candidates. Even though none of these descriptors correlated, the combined model's score improved the classification and correlation power with an R^2 of 0.218 ($p = 0.0491 < 0.05$).

Table 4.2. Almirall's internal set classification results.

Model	Accuracy	Precision	Recall	F1-score
Morgan FP	0,56	0,5	0,57	0,53
Xray 1	0,75	0,75	0,75	0,75
Xray 2	0,75	0,75	0,75	0,75

Moving to a prospective scenario, Almirall's collaborators executed a VS task on *Xray 1*. From 7 million compounds of the ZINC library, they were filtered by drug-like and non-toxicity features, getting 1.8 million compounds. Afterward, they ran the first run of dockings of these compounds with HTVS Glide on the *Xray 1* to preselect 64.894 ligands, and later, they executed a second run with SP Glide applying H-bonds constraints, selecting only those hits with docking scores below -8. In the end, they provided us with these 785 commercially available ligands to run our optimization protocol. We executed dockings and PELE refinement and collected all the metrics to run the machine learning analysis, providing a complete summary containing each descriptor and machine learning model probabilities. Almirall picked 32 compounds from these results, but only 23 were available for the experimental assays. They experimentally assessed IC50 values through ADP-GLO™ kinase assays, obtaining two hit compounds; one with an IC50 of 4μM (MW ~ 340g/mol) and the other of 140nM (MW~ 267g/mol). The enzymatic activities for the top 10 compounds are reported in *Table 4.3*. Additionally, a general view of all individual descriptors for *Xray 1* simulations that defined active and inactive compounds can be seen in *Appendix F*,

Figure F3. Q2-866 does not stand out in terms of energy (docking score, mean IE) but shows a higher ranking in cluster size (poses interacting with the hinge) and mean resilience interaction with the kinase hinge. Contrary, all Q2-672 descriptors fall around the average. Interestingly, following the cluster-like classification and non-linear behavior of the KNN algorithm, the ML model assigned a high probability of being active.

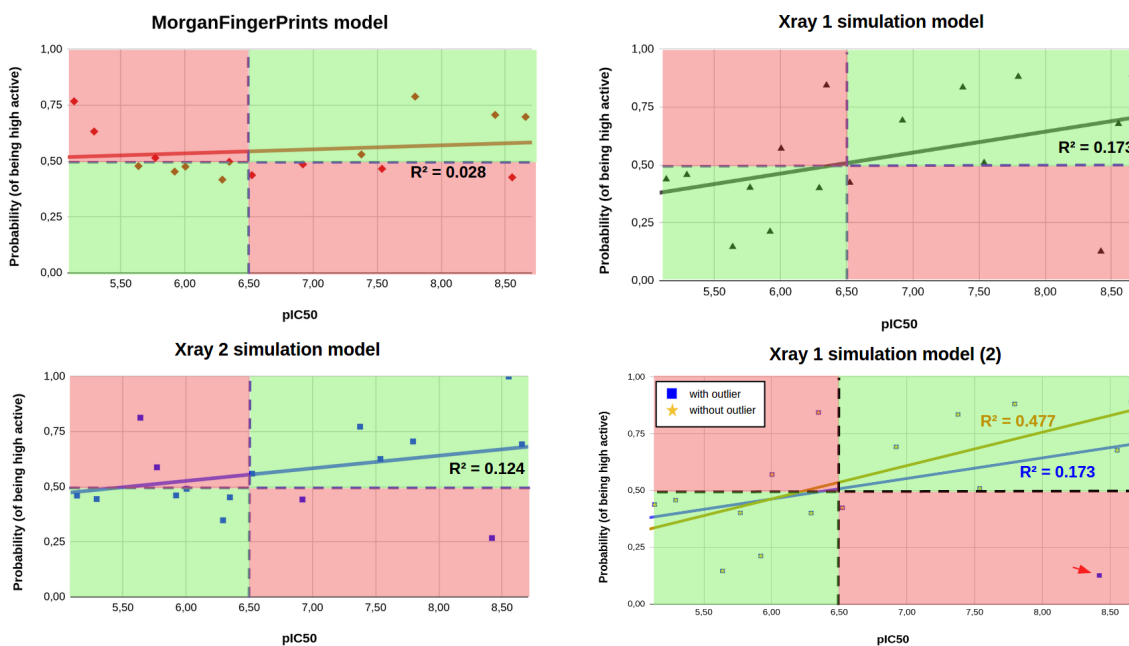


Figure 4.13. External set testing results. Correlation between the probability of belonging to the ‘high active’ group and pIC50. The red quadrant represents the wrong-classified samples, and the green the right-classified ones. Notice that the bottom-right graph tries to show the effect in the correlation of erasing a single point.

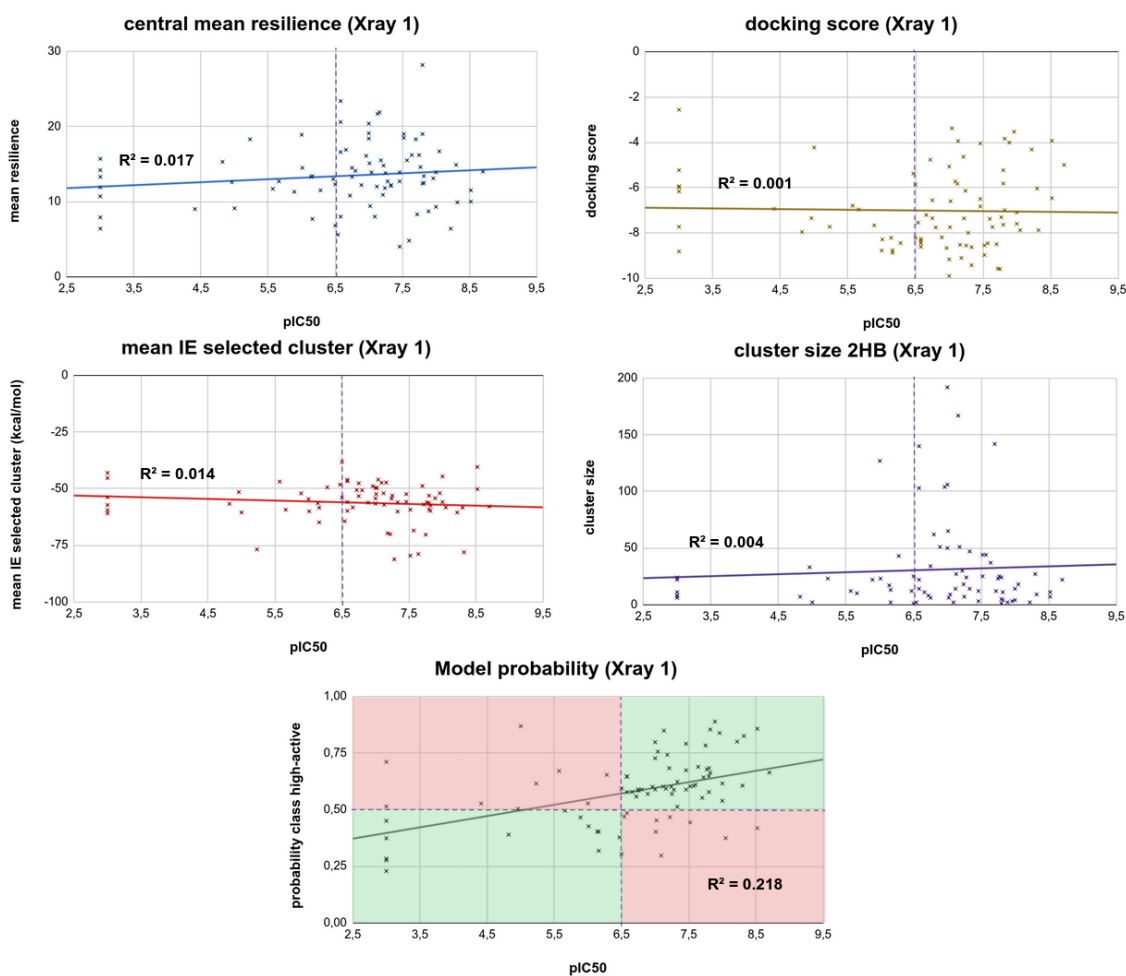


Figure 4.14. Scatter plots of the test set results for the four individual descriptors (top) and the model probability (bottom) against experimental pIC50 values. The red quadrant represents the wrong-classified samples and the green the right-classified ones. Horizontal lines set the limit between both classes.

Table 4.3. Reported enzymatic assays result for ADP-GLO™ assay by Almirall.

Compound	enzymatic activity				Xray 1 model probabilities
	%inh 1μM	%inh 10μM	%inh 40μM	IC50 (μM)	
Q2-866	43	88	100	0.140	0.809
Q2-672	33	80	95	4.1	1
Q2-784	1	13	-2	>40	0.836
Q2-953	-3	11	24	>40	0.434
Q2-198	2	10	18	>40	0.407
Q2-028	-3	3	7	>40	0.582
Q2-772	-5	2	-4	>40	1
Q2-420	-16	-2	17	>40	0.443
Q2-749	-3	-5	-11	>40	0.829
Q2-831	2	-6	13	>40	0.634

3.2. Hit-to-lead with FragPELE

To end SilicoDerm, we performed a H2L study on *Kinase 2* by applying FragPELE code. Our collaborators supplied a new X-ray structure co-crystallized with a hit compound to optimize through fragment growing. Herein, Almirall designed a new series of ligands that contained a modified version of the original scaffold X-ray. Three different growing positions were proposed, and for each position, a series of fragments containing a specific functional group, having then three sets: 18 pyridine, 22 alkyl, and 19 aromatic fragments. A general schema of this scenario is represented in *Figure 4.15*.

3.2.1. Methods

The original X-ray structure was prepared with Protein Preparation Wizard (Sastry et al., 2013), including the hydrogens and optimizing their orientation with PROPKA (Olsson et al., 2011) at 7 pH. Explicit water molecules were removed from the system. Then, the new scaffold was manually forged by modifying the X-ray ligand with the 3D builder from Maestro (Schrödinger, 2018), replacing two chemical structures with others similar to the original ones (see *Figure 4.15*). After this change, we executed the

same PELE refinement simulation protocol we used in the VS study (see the *previous section*) to readapt the system to the new scaffold. The pose with the lowest interaction energy was retrieved as the starting core complex in FragPELE simulations.

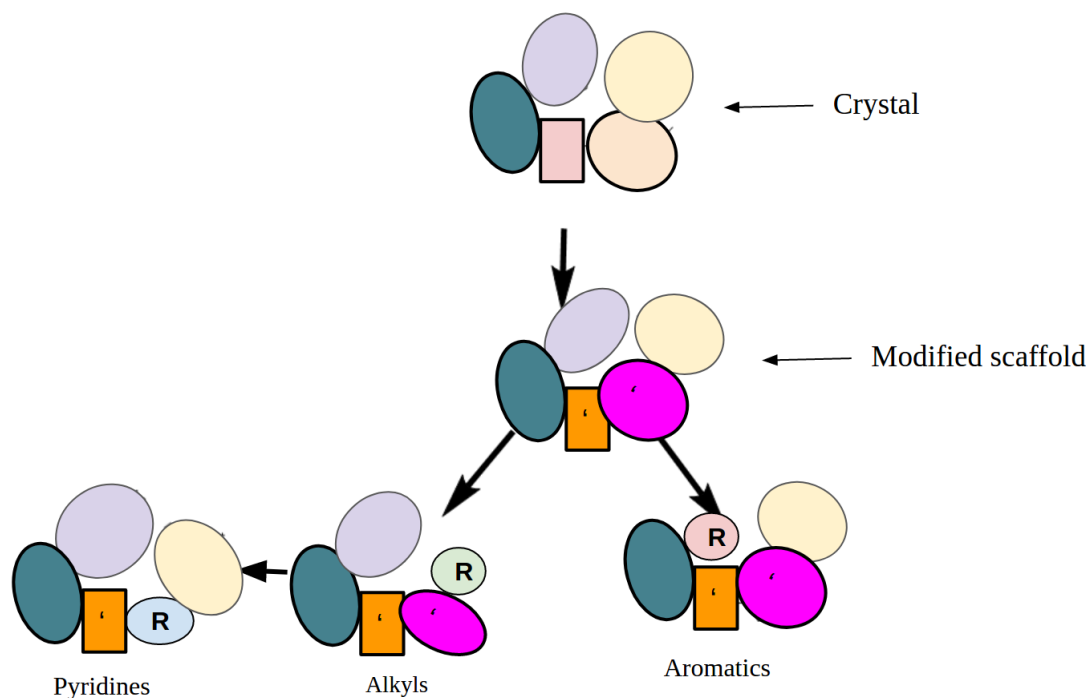


Figure 4.15. Schematic view of the ligand shapes for the H2L study in *Kinase 2*. For anonymization purposes, square and circular shapes hide the chemical structures. Each color represents a static chemical structure, while *Rs* identify the different growing sites.

Fragments were prepared with the Epik algorithm (Greenwood et al., 2010) of Schrödinger's LigPrep software, creating all possible tautomer and enantiomeric forms for each compound.

Many of the generated fragments contained one or more chiral centers. As errors were reported when growing fragments with several enantiomeric forms, we developed FragPELE 3.0.0 (see *Appendix C*) to face this study. Simulations began from a λ value of 0.25, meaning we instantly place a fragment with 25% of its final size to reduce computation. The growing stage was executed utilizing the new *Softcore-like* mode in 6 GS, increasing the minimization convergence criteria to 0.01 RMS in the first half of the growing stage and relaxing it to 0.1 in the second half. For those fragments containing chiral centers, dihedrals constraints were applied in the first half of the growing to keep the enantiomer state invariable in the early growing process. Besides,

based on other projects' experience, the sampling of the final simulation was extended to 50 PELE steps instead of 20.

Alkyl and aromatic series were constructed in a single growing simulation. In contrast, pyrimidines required two consecutive runs (the inner fragment and the constant terminal group). The lowest interaction energy pose from the first simulation was picked as input for the second one. The FragPELE simulations took around 1 hour and 30 minutes in 48 processor units per fragment.

3.2.2. Results

Based on the previous simulation results, Almirall synthesized 35 new compounds not reported in the literature as *Kinase 2* inhibitors. Tables 4.4, 4.5, and 4.6 show the reported inhibition activities and FragPELE scores for the alkyl, aromatic, and pyrimidines series.

Table 4.4. Enzymatic inhibitions and FragPELE scores for the Aklyl series. Rows with scores > -80 kcal/mol are colored in darker gray. False-positives or false-negatives are in red.

Compound (state)	FragPELE score (kcal/mol)	Enzymatic inhibition, IC50 (nM)
11-Alkyl (1)	-103.00	0.7
4-Alkyl (1)	-94.68	7.5
19-Alkyl (4)	-86.39	33
13-Alkyl	-71.89	430
1-Alkyl (2)	-65.65	130
22-Alkyl (2)	-65.17	130
17-Alkyl	-64.72	550
9-Alkyl (1)	-63.94	13
3-Alkyl	-63.64	330
10-Alkyl	-62.76	11
5-Alkyl	-62.76	400
8-Alkyl	-62.58	210
21-Alkyl	-62.19	440
7-Alkyl	-62.00	32

Table 4.5. Enzymatic inhibitions and FragPELE scores for the Aromatic series. Rows with scores > -80 kcal/mol are colored in darker gray. False-positives or false-negatives are in red.

Compound (state)	FragPELE score (kcal/mol)	Enzymatic inhibition, IC50 (nM)
16-Ar (1)	-106.71	1.2
13-Ar	-106.22	2.1
17-Ar	-104.54	180
12-Ar (1)	-104.13	14
1-Ar (1)	-102.42	0.69
11-Ar	-101.99	3.9
16-Ar (2)	-101.46	1.2
7-Ar	-100.66	2
8-Ar	-99.77	4.4
3-Ar	-99.21	42
2-Ar	-98.94	3
19-Ar (2)	-97.89	13
14-Ar	-96.21	14
18-Ar	-95.16	61
9-Ar (1)	-94.10	44
10-Ar (2)	-92.05	0.9
4-Ar	-80.49	180

Table 4.6. Enzymatic inhibitions and FragPELE scores for the Pyridine series. Rows with scores > -80 kcal/mol are colored in darker gray. False-positives or false-negatives are in red.

Compound (state)	FragPELE score (kcal/mol)	Enzymatic inhibition, IC50 (nM)
18-Pir	-102,58	4,2
1-Pir	-101,05	690
2-Pir	-95,71	440
16-Pir (1)	-93,61	1,6
9-Pir (2)	-92,98	1,7
17-Pir	-91,38	5,9
6-Pir	-73,95	1,9
10-Pir (1)	-71,32	1,7
15-Pir	<i>intramolecular clashes</i>	910

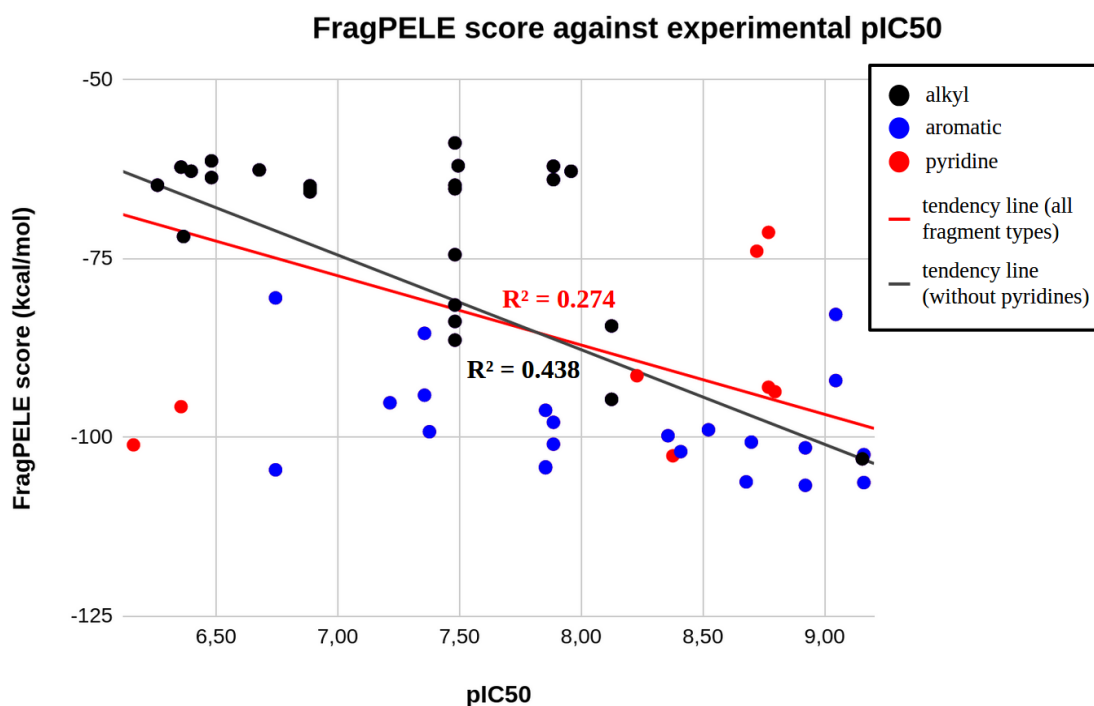


Figure 4.16. Correlation between experimental pIC50 and FragPELE scores for the different series of fragments.

In the alkyl and aromatic series, the top-scored compounds (< -80 kcal/mol) showed inhibition values lower than 50 nM (excepting 3 cases). At the same time, the compounds with lower scores (> -80 kcal/mol) usually reported lower activity values. However, when moving to the pyridines series, the percentage of false negatives and false positives is significantly increased. For further evaluation, FragPELE scores were correlated with the pIC50 of the experimentally tested compounds, showing an overall R^2 of 0.274. However, it rises to 0.438 when neglecting the pyridine series (*Figure 4.16*).

Chapter 5. Discussion

1. Total flexibility to predict binding poses with FragPELE

The results reported in this thesis suggested that applying our new algorithm to PELE has created a dynamic ligand growing tool that accommodates the receptor to the expanded molecular moiety. Self-growing and cross-growing results, which implied the reconfiguration of the receptor to the new molecule, showed good correlations between crystallographic and predicted geometries (*Tables 3.2 and 3.3*). Even though the model could not perfectly fit the reference structure in some cases, the native-like interactions were recovered (*Figure 3.6*).

Our first structural benchmark highlighted that our method only retrieved low RMSD values when water molecules were fixed on hydrated systems (*Table 3.2*). Consequently, facing prospective studies required previous knowledge of the positions of these waters. In this scenario, our method does not permit dynamically expanding R-groups on regions with explicit waters. Subsequently, these waters had to be deleted from these systems (reducing the accuracy of our predictions and losing the ability to relocate them after expanding the fragment). Crucial explicit waters can quickly be resolved by visually inspecting X-ray structures or using specific software such as WaterMap (Abel et al., 2008), which can provide insights into the water positions on these systems. Thanks to the recent development of aquaPELE, we could integrate both techniques and solve FragPELE obstacles (discussed below).

Reported results on the EGFR system suggested that FragPELE seems to be a promising tool for opening cryptic sub-cavities on binding sites through fragment growing. In this case, the repositioning of side-chains combined with a short displacement of backbone atoms was enough to produce the induced-fit effect to allocate the expanded moiety. In more complex systems which require larger backbone displacements, in principle, the protein perturbation performed within the PELE algorithm would be capable of readapting the receptor conformation and opening closed cavities. However, modeling the conformational changes produced by the ligand binding and the subsequent cryptic

pocket opening is still challenging. Opening cryptic cavities on apo structures is usually computed with long standard MDs (μs timescale) (Ode et al., 2012). Despite the considerable time consumption, it does not guarantee correctly modeling the holo conformation (Oleinikovas et al., 2016). In this case, we do not require to simulate the whole apo-holo transition, as FragPELE will start growing from bound conformations. In this sense, our study is more focused on small sub-cavities close to the native BS. Regardless of our anecdotic success, we must be cautious with our results, and further efforts should be made to widely benchmark and confirm FragPELE's technology.

After all these retrospective tests growing R-groups onto well-characterized protein-ligand complexes, which shared a common scaffold, one could think that FragPELE does always keep the scaffold of the molecule highly constrained and stationary along with simulations. The study performed on MAPK p38 inhibitors analogs hints that our strategy will not constantly adapt the receptor to give space to the new fragment, rearranging the whole ligand and even losing canonical interactions when needed (*Figure 3.12*). We cannot thoroughly check the quality of the predicted binding mode due to the lack of structural data. However, interestingly, the only model with a completely different ligand pose matches the one with lower affinity (*Figure 3.12*).

Reproducing X-ray structures is an important feature that can help medicinal chemists to see the effect of adding new chemical groups to the hit compound and guide the design of the lead compound. However, structural information cannot quickly determine which fragment will be more suitable for the studied BS when analyzing hundreds of modifications. The following section will discuss this topic based on our results.

2. Binding affinities and scoring

Developing appropriate scoring functions to rank ligand's affinities remains a challenging task. This thesis applied PELE's interaction energy as the leading scorer in more than 200 structurally diverse ligands distributed in eight targets of different families. Our first scoring benchmark (*Chapter 3, Section 1.3*) indicated that FEP+ outperformed FragPELE slightly regarding accuracy when ranking systems where the

MW correlation is lower than 0.5. However, both perform similarly when the gain of further interactions (additional mass) between fragment and receptor accounts for the affinity change. These results could indicate that our more straightforward technique accurately describes enthalpic contributions. We remind here that our simple score is the mean of the lowest 25% force field interaction energies poses (see *Chapter 3, Section 1.1*). A significant source of error might come from the lack of explicit solvent and entropic effects (both explicitly considered in FEP techniques).

Our interaction-energy-based score is different from FEP; we do not aim at computing absolute differences of binding free energies. PELE's energy values are overestimated as they were not designed for this purpose. Nevertheless, this score seems to help in evaluating relative differences in binding free energies in congeneric series of ligands sharing a common scaffold. Following this line, the results reported in the SilicoDerm project (*Figure 4.15*) also support this idea where we retrieved good predictions ($R^2 = 0.44$) in the *Kinase 2* when moving out the set of internal pyrimidines. Furthermore, the study performed in *Kinase 1* showed a clear overscoring on charged fragments (series 2 in *Figure 4.4*), making their scores not comparable to neutral compounds and reporting then similar limitations of other methods such as alchemical methods.

3. Effect of explicit waters

The proper study of protein-ligand interactions usually relies on water molecules, and therefore achieving reliable computational models depends on including explicit/implicit solvent models. Our first structural benchmark on FragPELE using the implicit VDGBNP model denoted lower correlation with crystallographic data when not fixing explicit water molecules on hydrate systems. In order to reduce our method limitations, we designed a new benchmark focused on hydrated systems. Herein, the introduction of aquaPELE on FragPELE reported low ligand heavy-atoms RMSD distributions (around 1 Å) in 6 out of 8 systems (*Figure 3.15*) and an excellent ability to diminish the presence of explicit waters after expanding fragments known to displace them (*Table 3.10*). Referring to this last point, finding crystallographic structures where the water molecule stands after adding a fragment pointing to a buried cavity was impossible for us. Consequently, we were forced to rationally design negative controls

to exemplify that the algorithm only will move out the unfavorable and keep the energetically stable waters.

One of the most important goals of this second study was to determine if adding the mixed implicit/explicit solvent model of the aquaPELE algorithm to FragPELE simulations could improve binding affinity predictions. In this scenario, we also faced the challenging expansion of multiple R-groups in two, three, and four different sites. Consequently, we successfully applied a new, faster, and more plain simulation to add intermediate R-groups, and in parallel, we also extended simulations testing the behavior of FragPELE's 3.0 protocol. Herein we could apply the new *Softcore-like growing* mode, reducing the number of growing steps and extending the sampling simulation. Results obtained in two out of three congeneric series showed a sufficient predictive power (*Figure 3.17*) (NW, POP5 scorer, averaged $r > 0.63$), except for the most chemically diverse series, where we grew fragments in 4 different positions and as one could expect correlations dropped. Indeed, when introducing the aquaPELE algorithm, the correlations between experimental and predicted energies notably improved in all three series, especially in series 3 (*Figure 3.17*), passing from low correlated to acceptable values.

Even though this hybrid approach could potentially enhance binding affinity and water displacement predictions, it is also true that the current method lacks metrics to rank the entropic contributions and apply proper energy corrections. Here more efforts are needed to compute water entropic terms from PELE MC simulations. Furthermore, in contrast with other methods, aquaPELE follows an approach where explicit water is not spawned or deleted from the system, and therefore they must go somewhere within the box. In this sense, defining box sizes long enough to reach solvent-exposed regions seems a reasonable and realistic strategy to discard these unfavorable water molecules. Currently, setting these parameters is still a task for the user. It depends on the properties of the BS; in the future, automatizing the box selection could be an exciting feature to add.

4. **FragPELE: growing and sampling**

FragPELE is a ligand growing approach coupled to PELE, and subsequently, along with the simulation, several MC steps (samples of protein-ligand poses) are taking place. In this context, our method uses PELE simulations to recreate protein-ligand states in two different key processes: the fragment growing and the sampling simulation. First, we will extend the discussion to the former.

In the fragment growing step, the FFP of the ligand are modified to create non-realistic physical states to convert the original scaffold onto another ligand with a fragment attached. We called GS to these steps, and given its parallelism with alchemical methods, we also referred to a lambda (λ) value associated with these microstates. We want to clearly state that differently from other alchemical techniques such as FEP, FragPELE's intermediate microstates do not pretend to sample and track the energetic path along with the multiple states (from ligand A to ligand B). Contrary, our method follows a more straightforward approach consisting of just a few PELE steps (6 by default in the latest implementation) to create stable poses quickly and, whenever possible, accommodate the receptor to the new extended ligand. Therefore, when *perturbing* (jumping to the next GS step) systems, the most crucial fact is to avoid inducing microstates with highly unbalanced energy terms that could produce atomic clashes and structures with wrong chemical configurations. After extensively using FragPELE, we proposed up to three different growing protocols (visit *Appendix C*). In this context, in the hit-to-lead campaign of the Kinase 2 in the SilicoDerm project, the application of the new *Softcore-like* protocol improved the ratio of correct simulations in fragments containing chiral centers. Moreover, following this research line, FragPELE code prepares the ground for further implementations of more accurate alchemical transitions coupled with PELE software.

As we have highlighted several times in this thesis, the first version of FragPELE followed a simple scoring function: the mean interaction energy of the 25% lowest energy samples. According to our first results, this calculation could be enough to rank a congeneric series of ligands. However, a MC simulation can easily generate thousands of protein-ligand conformations. Depending on the system's complexity, this simple

energy-based cutoff can capture very different binding modes and, therefore, neglect relevant information.

Including water molecules in our second benchmark led to a higher degree of variety of conformations, given that multiple water positions can be detected for each protein-ligand conformation. Being aware of this possible source of error, in aquaPELE simulations, it is recommended to use a clustering-based analysis (already reported in aquaPELE paper) (Municoy et al., 2020). By doing so, one can efficiently identify different binding modes (see *Appendix D*) and then compute energies of similar protein-ligand conformations, not only energy-filtering-based. Then, we could quickly obtain groups of similar poses and collect population and energy data from them by applying this strategy. As a general rule, for the studied systems, clusters with lower interaction energies (or even more population) correspond to structures closer to the crystallographic pose (*Appendix D*). Thus, as discussed in *section 2.2.2 of Chapter 3*, we explored different methods to identify top clusters (population or energy) and scoring methods (P5, P25, or mean). Most of the scorers followed a similar trend; however, following the structural data, better results are retrieved when selecting clusters based on interaction energy and picking only the 5% of samples with lower values (*Figure 3.17*). Finally, as expected, a slightly better experimental-predicted correlation in the studied systems was obtained using clustering-based than the whole-simulation analysis employed in the original score of FragPELE (*Tables 3.11*).

5. Importance of automatizing in drug design studies

Repeating a hundred times fundamental tasks, such as moving files, can become an absolute nightmare without an automatic way to do it. Luckily, no immense efforts/knowledge is needed to move thousands of files in a row, but things get more complicated when facing molecular modeling simulations in drug discovery projects. Following this line, running many such jobs is impossible if non-efforts are made in automatization. Even though the long time it takes for developers to create tools automatable (a feature that is usually taken for granted), in our opinion, not enough importance is attributed in some papers. Thus, when designing any new software, developers must consider this fact. One can create the best modeling tool, but when

users cannot easily automate their usage, it becomes a pain to set everything up to run it, which will ruin the method's applicability.

It is not strictly stressed along with the thesis (as this is not scientific content by itself), but considerable development labors have been made to improve the automatization of FragPELE from the original version 1.0 to 3.0. Given any library of fragments, the software is prepared to set up all the files and execute all simulations independently. As a result of this implementation, we could quickly grow more than a hundred fragments to face future drug discovery projects (such as SilicoDerm). Additionally, FragPELE code allows being run through PELE Platform (https://nostrumbiodiscovery.github.io/pele_platform/index.html) software, a recently developed external wrapper to improve even more the user-friendliness of all PELE-associated technologies.

In conclusion, the developing tasks are done to automatize further and adapt all codes to make them parallelizable and easy to use. This task has “invisibly” increased the method's efficacy, granting its implementation to quickly assess multiple fragments in drug discovery projects.

6. Prospective hit-to-lead with FragPELE

Regarding current H2L campaigns, our cooperation with Almirall in the SilicoDerm project has been an ideal scenario to test FragPELE's predictions blindly. This FragPELE-guided analysis in *Kinase 1* resulted in synthesizing 4 compounds from the top 10 scored. Herein, 4 of the 4 tested compounds were hits, but only one showed better activity than the original hit. Truchon and Bayly published the most recent study that we found to value our results in 2007. They reported a hit discovery ratio between 1% and 40% of prospective VS campaigns (Truchon & Bayly, 2007), and unfortunately, we could not find updated numbers focused on hit-to-lead campaigns. However, as we only had experimental data for these 4, our results could be considered an anecdotic success of FragPELE; however, showing good predictive behavior in H2L campaigns.

When designing scientific software, researchers initially benchmark and test the developed methods with just well-known datasets. Usually, when this software starts to

be massively used in many different scenarios, novel problems and weaknesses arise. In fact, in the H2L campaign of the *Kinase 2*, we had to deal with this situation. Growing fragments with chiral centers resulted in simulations with mixed R/S enantiomeric states. This property of FragPELE could be explored to be used as a selector of the most-probable enantiomeric state; instead, we decided to correct the method to keep the original state and assess the results based on the user-defined input.

Consequently, new functions and protocols were implemented in FragPELE (visit *FragPELE 3.0.0* of *Appendix C*) to confront this issue and solve this limitation. The number of GS was diminished to 6 (instead of 10) due to the augmented computation time (especially the first half of the growing phase) produced by a stronger minimization. Additionally, the sampling simulation was extended to 50 steps taking advantage of the reduction of GS to further explore the system with the full-sized ligand in a similar amount of computation.

Overall, the predictive power reported in this study is consistent with the previous results from the more extensive congeneric series of the kinase family in the original FragPELE paper (*Figure 4.16, B4 and B5*), stepping up the goodness of the method to face hit-to-lead campaigns. However, the two-steps growth in the pyridine series has raised a weakness in the technique, as it has not been designed for this goal: replacing non-terminal groups by expanding an intermediate fragment can critically sink the predictions (*Figure 4.16*). Creating new techniques to directly add similar fragments given a reference structure and optimizing the RMSD to imitate the original volume occupied by the molecule would be an exciting feature that would reduce the computation time and be useful for scaffold hopping.

Finally, applying FragPELE to a real-world problem has been an enriching experience that prepares methods to face scenarios that, sometimes, in a fully-academic environment, would have never been tuned. Many unexpected errors, ideas, or new automatization pipelines come to users' minds. For this reason, covering all the needs that the program should have, makes the software solid and reliable to face future drug discovery projects. As a note for the reader, FragPELE is currently being used in (multiple) different industrial projects by Nostrum Biodiscovery and has even successfully been used in a cloud (Amazon Cloud) implementation.

7. Optimizing virtual screening pipelines through machine learning algorithms

VS campaigns usually rely on computationally testing millions of compounds by employing fast methods such as molecular docking to identify, at least, a few compounds with low MW and low affinity (mM- μ M range). Luckily, some targets have multiple hits already detected. All this information around the binding mode, such as critical interactions, shape, physicochemical properties, or any other relevant data, can be helpful in designing more selective and potent compounds. Therefore, following the H2L context, we proposed optimizing VS pipelines by taking advantage of all this available data to obtain more potent inhibitors than standard methods.

Even though the 2D information of all known inhibitors is useful in QSAR studies can also be limiting in highly dynamic targets, such as kinases. Accordingly, in this thesis, we proposed integrating molecular simulations data to simple machine learning classifiers to retrieve compounds with high activity on the *Kinase 2*. Subsequently, including simulated data could enhance pipelines' performance.

Other groups have already successfully tried this kind of approach in the past years, integrating fingerprints and MD in the ABL1 kinase (Spyrakis et al., 2015) by clustering relevant conformations and classifying the compounds by linear discriminant analysis. However, this work was more focused on enhancing the hit discovery rates and not improving the affinity of the compounds. To this end, instead of creating the usual two classes for VS, active/inactive, we decided to classify compounds in high/low activity (adding the inactive compounds in the low activity group), expecting to yield hits with higher potency.

By applying our pipeline, we could determine that combining multiple types of descriptors (ligand-based, docking scores, and simulation-based) could increase the classification power than using single scores in most drug discovery projects (*Table F1* and *F2*). It also seems to show a higher applicability domain and to be less prone to suffer from overfitting than other 2D-based methods such as FingerPrints (*Table 4.2* and *Figure 4.13*).

After all this retrospective analysis, the technique was employed in a prospective study, and as a result, from 23 assayed compounds, 2 were active (8.7%). Considering the normal potency range of the hit discovery phase (100nM-5 μ M) (J. P. Hughes et al., 2011), both compounds fall within. However, the 140nM is close to the upper limit. Given their small size, the further optimization of these molecules is likely feasible, showing the potential to become real drug candidates.

In the VS executed with Glide by Almirall collaborators, these hits were ranked in positions 441 (140nM) and 688 (4.1 μ M), while our predictions reported high probabilities of being highly active (*Table 4.3* and *Figure F3*). However, one must be aware that the selection of compounds by Almirall was not purely based on the machine learning scores. They evaluated all individual metrics, selecting then the compounds also based on their expertise, price, and market availability. Considering this fact and the lack of data of all the non-assayed compounds, thoroughly evaluating its ability on these results would be pure speculation. Distinctly, we can ensure that computed features and machine learning-based scores assisted Almirall in selecting these compounds.

More groups are starting to use machine learning-based techniques to enrich virtual screening pipelines (Gupta & Zhou, 2021); not surprising given the success observed in other close fields, such as protein structure predictions with AlphaFold (Jumper et al., 2021). Given the increase of knowledge and the better accessibility to data, in our opinion, implementing strategies including machine learning algorithms could guide the future of VS.

8. The humanity behind machine learning

Generally speaking, machine learning algorithms are sometimes considered magic, programs that automatically can be trained and quickly score any given data, but this is far from reality. Algorithms learn from multiple tries and errors in the training stage. Human intervention must smartly prepare the data, select the most suitable algorithm, and accurately parameterize the machine learning pipeline. Curating the input data or selecting suitable descriptors are crucial steps to make a difference. Most of them are

purely dependent on the data, but others are not. In this section, we want to briefly discuss all the different human decisions that could have changed the outcome of our pipeline.

Firstly, we decided to build a two-classes classifier instead of a multi-class or a regressor. Machine learning algorithms require data to be trained appropriately. Due to the lack of information (very common in the drug discovery field), we picked the simplest model but guaranteed that the data would be enough. Second, we set an arbitrary threshold ($pIC_{50} = 6.5$) to distinguish between low and high active compounds. This value falls close to the median of our distribution, cutting it into approximately two halves of similar size. This decision can lead to misclassifying compounds around the limit due to the intrinsic experimental errors and differences between laboratories.

For this reason, when the amount of data is enough, it is recommended to include an intermediate class in between. Unfortunately, this was not our case, so we had assumed this possible source of error. Third, we collected just a few ligand-based properties, docking scores, and descriptors from PELE simulations that we thought were relevant to distinguishing low from high active compounds. However, there is an almost infinite number of metrics that we could have included in the model. Fourth, the selection of the normalization method, classification model, and feature selection are also steps susceptible to human intervention and the evaluation metric (f1-score, accuracy). There is no clear roadmap in the field, so we decided to create a list with the most common normalizers and classifiers and test all of them, selecting the one that retrieved a higher outcome.

In conclusion, the human effort behind any machine learning pipeline is sometimes vast and invisible. Data analysts experts in machine learning have further knowledge in the field. They know how to easily prepare every step to select the most suitable algorithm and train the models. For non-expert users (and sometimes also for experts), several rounds of trial and error must be done to parameterize all the pipelines and avoid falling into common pitfalls (f.ex. do not normalize, nor split the data). Due to these errors, the models obtained are usually overfitted, being more than good within the testing set but completely useless when applied to different sets. Researchers must be cautious with

this kind of results, and in case of doubt, it is recommendable testing the model in a completely different external set to provide some insights into the natural quality of the model.

9. Strengths, limitations and opportunities of the current techniques

In most of this thesis, we tackled the design of a computational tool to model the effect of expanding decorators on a molecular scaffold. The truth is that in the previous decade, many other programs have been designed to shed some light on the ligand growing field, such as *Autogrow3.0*, *LigBuilder2.0*, *OpenGrowth* or *NAOMINext*, among others (Chéron et al., 2016; Durrant et al., 2013; Sommer et al., 2019; Yuan et al., 2011). Most of them put efforts into quickly predicting which of the infinite decorators available in the chemical space best fits the target system and subsequently show more weaknesses in their energetic predictions (usually docking-based).

When virtually facing H2L campaigns, the quality of the predictions is critical, but the speed of computation is also a limiting feature. Docking tools allow for fastly testing from hundreds-thousands to millions of compounds, so they become the ideal software to use when there is limited information about which chemical groups are more prone to bind. Their predictions tend to fail more than in highly demanding computation techniques, and in this sense, alchemical methods such as FEP reported higher accuracies in predicting relative binding free energies. Even though several efforts have been recently made to speed up the computation time through GPU-based implementations (H. Chen et al., 2020), the calculation is still highly demanding. This time scale permits testing just a few modifications in a reasonable amount of time, which can be highly limiting when screening a small library of R-groups. One could think that end-point methods such as MMGBSA should fall between fast docking and the expensive FEP, but counterintuitively their performance is not higher than the former and below the latter (*Table 3.8*). With this concern, this thesis developed a novel ligand growing methodology called FragPELE. Structural and scoring benchmark results show potential dual use of the technique to accurately reproduce protein-bound native-like geometries and provide reasonable free binding energies predictions for the

studied systems. As discussed in *Chapter 3*, FragPELE calculations take around 1h per fragment on 48 computing cores (about one processor in Mare Nostrum IV supercomputer) with the designed protocol. This time can fluctuate from 20 minutes to 3 hours, depending on the size of the whole complex, the number of atoms in the fragment, or also the effort needed to adapt the protein to the new fragment; however, this time scale is still significantly lower than the computationally intensive FEP.

Even though the positive and goodness of the results reported with FragPELE, the type of targets and the chemical diversity proved is too small yet to get strong statements compared with other popular tools such as Glide or FEP+. In contrast to alchemical methods, FragPELE should be able to extensively screen around 1000 modifications in a week in a CPUs cluster of 288 processor units, which is not a big deal with the current high-performance computing resources available. This trade-off makes our method a suitable tool to test R-groups from a small/middle size library of fragments in an acceptable time scale and possibly fill this empty gap in the literature. The more people use the software, the better it will become, and along the time, it will also increase its reliability.

It is essential to be aware that FragPELE allows only partially to explore the chemical space. Combining N fragments, with N growth sites of the scaffold, with N linking atoms of the fragment gives, as a result, infinite new molecules. Therefore, an accurate preselection of better fragment candidates, preferably done by expert medicinal chemists, is still necessary. Further work must be done to smartly preselect which modifications will have higher probabilities of improving the initial compound. For example, a brilliant strategy would be using any of the previously mentioned docking-based methods to pick the best-scored modifications and our method later. Another option could be to design slightly less accurate but faster versions of the current algorithm to filter and reduce the number of molecules to apply FragPELE simulations later. In fact, in collaboration with Laura Malo from Guallar's lab, we have already developed the first version of this algorithm, named *FragHop*. The method was successfully applied to pre-select 5-10 fragments (from a library of thousands) per subpocket in a wide multiple-pocket BS. Combining it with new generative machine

learning models would be an innovative pipeline to automatically compose and assess novel chemical structures.

FragPELE's methodology would open a new window of opportunities applicable to other molecules and not just small drugs. For instance, the same strategy could be easily extended to side chains to automatically mutate and dynamically allocate new amino acids within protein binding sites. This application could be helpful in the enzyme engineering field and the design of short peptidic inhibitors. Additionally, as abovementioned, putting more computation efforts into analyzing and trying to reduce the perturbation of the system in the transitory alchemical microstates would result in a novel MC-based FEP method, with hopefully higher correlated predictions. Given all the potential alternative uses of the method, we encourage future scientists to do more research on this technique to make the most of it.

Furthermore, in the final stage of the thesis, we also approached the design of a new VS method based on machine learning pipelines focused on enriching the activity of the predicted compounds. This technique offers the opportunity to profit from the available knowledge and optimize the VS towards the target of interest. Its application into commercially available compounds helped identify two true inhibitors. However, the results of this commercial set highlighted a drop in the classification power of the method (*Table 4.3*). The model was prepared to distinguish between high and low active, including only 20 inactive compounds in the set. Therefore, the training was mainly done with active compounds and probably falling on a non-realistic active/inactive balance. Even though the model was applied on an enriched set (prone to be active), most of them should be non-active, and subsequently, the algorithm failed to detect them. Still, finding 2 active from 23 tested ligands is a high hit discovery rate that can be considered a remarkable success of the method.

Another limitation of this technique relies on the target applicability. On the one hand, we require a well-studied target, with many reported inhibitors needed to train the model. On the other hand, we would need good quality datasets. Databases like ChEMBL (Gaulton et al., 2012) only provide an acceptable amount of data in IC50s, usually in various laboratories and under different experimental conditions. In this scenario, discretizing the data and classifying them is an excellent approach to reducing

errors due to the experimental conditions. Additionally, most of these compounds are just known to inhibit the target, but the mechanism is still unknown (do they bind in the same cavity? or are they allosteric inhibitors?). All these experimental uncertainties add noise to models, negatively affecting the predictions.

For this reason, wasting time cleaning the data as much as possible is a must to improve the outcome of any trained model. Moreover, for future scenarios, enriching the training with a large set of decoys/inactive compounds and training a three-classes classifier (high, low, inactive) could prevent false positives when applying the model on a set of compounds with a shallow hit rate. Adding descriptors to the model could be another source of improvement. For example, adding more ligand-based descriptors such as logP values or individual descriptors from Glide docking (internal energy, ligand efficiency, glide lipo), or even trying to gather more descriptors from PELE simulations (distribution of solvent accessible surface areas), collecting then different sets of metrics depending on the inhibition mechanism of the target. By refining all these steps, the method should be easily extrapolated to other well-known target families.

Chapter 6. Conclusions

The conclusions extracted from this thesis are:

- FragPELE showed good correlations with crystallographic data and ligand affinities in multi-type benchmarks in many targets. FragPELE's free energy predictions performed similarly to FEP and outperformed Glide, filling the empty gap in the ligand growing literature between the accurate but computationally expensive alchemical methods and the fast but less exact docking.
- The original version showed some limitations solved along with the development of this thesis, and new functions have been made available to users.
- AquaPELE algorithm has solved FragPELE's impediments on hydrated growing regions, showing better performances when combining both techniques. Additionally, ligand poses' clusterization can help identify hydrated sites and scoring tasks.
- Applying FragPELE in the SilicoDerm project, in a drug discovery campaign in collaboration with Almirall company, has been an exhaustive test with various chemical structures. Our predictions in Kinase 1 ended up synthesizing a compound with a gain of the potency of 4 folds. In Kinase 2, the synthesis of 35 compounds assisted by FragPELE revealed good correlations with experimental predictions only when expanding terminal R-groups and losing performance when modifying scaffolds. Developing alternative methodologies to address scaffold hopping could provide more insights into these predictions.
- The innovative usage of machine learning tools to combine multiple-source data (ligand-based properties and molecular simulation) assisted in finding two hit compounds. We believe that this methodology is promising to be extended to other targets after further refinements.

References

- Abel, R., Young, T., Farid, R., Berne, B. J., & Friesner, R. A. (2008). Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *Journal of the American Chemical Society*, *130*(9), 2817–2831. <https://doi.org/10.1021/ja0771033>
- Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1-2*, 19–25. <https://doi.org/10.1016/j.softx.2015.06.001>
- Alder, B. J., & Wainwright, T. E. (1957). Phase Transition for a Hard Sphere System. *The Journal of Chemical Physics*, *27*(5), 1208–1209. <https://doi.org/10.1063/1.1743957>
- Alejaldre, L., Lemay-St-Denis, C., Perez Lopez, C., Sancho Jodar, F., Guallar, V., & Pelletier, J. N. (2020). Known Evolutionary Paths Are Accessible to Engineered β -Lactamases Having Altered Protein Motions at the Timescale of Catalytic Turnover. *Frontiers in Molecular Biosciences*, *7*, 599298. <https://doi.org/10.3389/fmolb.2020.599298>
- Amengual-Rigo, P., Carrillo, J., Blanco, J., & Guallar, V. (2020). Predicting Antibody Neutralization Efficacy in Hypermutated Epitopes Using Monte Carlo Simulations. *Polymers*, *12*(10). <https://doi.org/10.3390/polym12102392>
- Åqvist, J., Medina, C., & Samuelsson, J.-E. (1994). A new method for predicting binding affinity in computer-aided drug design. *Protein Engineering, Design & Selection: PEDS*, *7*(3), 385–391. <https://doi.org/10.1093/protein/7.3.385>
- Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., & Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal*, *80*(1), 505–515. [https://doi.org/10.1016/S0006-3495\(01\)76033-X](https://doi.org/10.1016/S0006-3495(01)76033-X)
- Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., ... Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373*(6557), 871–876. <https://doi.org/10.1126/science.abj8754>
- Bakan, A., Meireles, L. M., & Bahar, I. (2011). ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics*, *27*(11), 1575–1577. <https://doi.org/10.1093/bioinformatics/btr168>
- Baker, E. N., & Hubbard, R. E. (1984). Hydrogen bonding in globular proteins. In *Progress in Biophysics and*

-
- Molecular Biology* (Vol. 44, Issue 2, pp. 97–179). [https://doi.org/10.1016/0079-6107\(84\)90007-5](https://doi.org/10.1016/0079-6107(84)90007-5)
- Banks, J. L., Beard, H. S., Cao, Y., Cho, A. E., Damm, W., Farid, R., Felts, A. K., Halgren, T. A., Mainz, D. T., Maple, J. R., Murphy, R., Philipp, D. M., Repasky, M. P., Zhang, L. Y., Berne, B. J., Friesner, R. A., Gallicchio, E., & Levy, R. M. (2005). Integrated Modeling Program, Applied Chemical Theory (IMPACT). In *Journal of Computational Chemistry* (Vol. 26, Issue 16, pp. 1752–1780). <https://doi.org/10.1002/jcc.20292>
- Barillari, C., Taylor, J., Viner, R., & Essex, J. W. (2007). Classification of water molecules in protein binding sites. *Journal of the American Chemical Society*, *129*(9), 2577–2587. <https://doi.org/10.1021/ja066980q>
- Berendsen, H. J. C., Grigera, J. R., & Straatsma, T. P. (1987). The missing term in effective pair potentials. *The Journal of Physical Chemistry*, *91*(24), 6269–6271. <https://doi.org/10.1021/j100308a038>
- Bian, Y., & Xie, X.-Q. (2021). Generative chemistry: drug discovery with deep learning generative models. *Journal of Molecular Modeling*, *27*(3), 1–18. <https://doi.org/10.1007/s00894-021-04674-8>
- Bilancia, D., Rosati, G., Dinota, A., Germano, D., Romano, R., & Manzione, L. (2007). Lapatinib in breast cancer. *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO*, *18 Suppl 6*, vi26–vi30. <https://doi.org/10.1093/annonc/mdm220>
- Borrelli, K. W., Vitalis, A., Alcantara, R., & Guallar, V. (2005). PELE: Protein Energy Landscape Exploration. A Novel Monte Carlo Based Technique. *Journal of Chemical Theory and Computation*, *1*(6), 1304–1311. <https://doi.org/10.1021/ct0501811>
- Bowers, K. J., Chow, D. E., Xu, H., Dror, R. O., Eastwood, M. P., Gregersen, B. A., Klepeis, J. L., Kolossvary, I., Moraes, M. A., Sacerdoti, F. D., Salmon, J. K., Shan, Y., & Shaw, D. E. (2006). Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. *SC '06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, 43–43. <https://doi.org/10.1109/SC.2006.54>
- Brooks, B. R., Brooks, C. L., 3rd, Mackerell, A. D., Jr, Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., ... Karplus, M. (2009). CHARMM: the biomolecular simulation program. *Journal of Computational Chemistry*, *30*(10), 1545–1614. <https://doi.org/10.1002/jcc.21287>
- Brown, D. G., & Boström, J. (2018). Where Do Recent Small Molecule Clinical Development Candidates Come From? *Journal of Medicinal Chemistry*, *61*(21), 9442–9468. <https://doi.org/10.1021/acs.jmedchem.8b00675>
- Burton, A., Castaño, A., Bruno, M., Riley, S., Schumacher, J., Sultan, M. B., Tai, S. S., Judge, D. P., Patel, J. K., & Kelly, J. W. (2021). Drug Discovery and Development in Rare Diseases: Taking a Closer Look at the Tafamidis Story. *Drug Design, Development and Therapy*, *15*, 1225–1243. <https://doi.org/10.2147/DDDT.S289772>
- Böhm, H.-J. (1992). The computer program LUDI: A new method for the de novo design of enzyme inhibitors. In *Journal of Computer-Aided Molecular Design* (Vol. 6, Issue 1, pp. 61–78). <https://doi.org/10.1007/bf00124387>

- Cabeza de Vaca, I., Qian, Y., Vilseck, J. Z., Tirado-Rives, J., & Jorgensen, W. L. (2018). Enhanced Monte Carlo Methods for Modeling Proteins Including Computation of Absolute Free Energies of Binding. *Journal of Chemical Theory and Computation*, *14*(6), 3279–3288. <https://doi.org/10.1021/acs.jctc.8b00031>
- Cabeza de Vaca, I., Zarzuela, R., Tirado-Rives, J., & Jorgensen, W. L. (2019). Robust Free Energy Perturbation Protocols for Creating Molecules in Solution. *Journal of Chemical Theory and Computation*, *15*(7), 3941–3948. <https://doi.org/10.1021/acs.jctc.9b00213>
- Carlson, H. A., Smith, R. D., Damm-Ganamet, K. L., Stuckey, J. A., Ahmed, A., Convery, M. A., Somers, D. O., Kranz, M., Elkins, P. A., Cui, G., Peishoff, C. E., Lambert, M. H., & Dunbar, J. B., Jr. (2016). CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. *Journal of Chemical Information and Modeling*, *56*(6), 1063–1077. <https://doi.org/10.1021/acs.jcim.5b00523>
- Carracedo-Reboredo, P., Liñares-Blanco, J., Rodríguez-Fernández, N., Cedrón, F., J. Novoa, F., Carballal, A., Maojo, V., Pazos, A., & Fernandez-Lozano, C. (2021). A review on machine learning approaches and trends in drug discovery. *Computational and Structural Biotechnology Journal*, *19*, 4538–4558. <https://doi.org/10.1016/j.csbj.2021.08.011>
- Case, D. A., Cheatham, T. E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K. M., Jr, Onufriev, A., Simmerling, C., Wang, B., & Woods, R. J. (2005). The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, *26*(16), 1668–1688. <https://doi.org/10.1002/jcc.20290>
- Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., & Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. *Methods*, *71*, 58–63. <https://doi.org/10.1016/j.ymeth.2014.08.005>
- Chang, L. L., Sidler, K. L., Cascieri, M. A., de Laszlo, S., Koch, G., Li, B., MacCoss, M., Mantlo, N., O’Keefe, S., Pang, M., Rolando, A., & Hagmann, W. K. (2001). Substituted Imidazoles as Glucagon Receptor Antagonists. In *Bioorganic & Medicinal Chemistry Letters* (Vol. 11, Issue 18, pp. 2549–2553). [https://doi.org/10.1016/S0960-894X\(01\)00498-X](https://doi.org/10.1016/S0960-894X(01)00498-X)
- Chen, H., Li, L., Zhang, T., Qiao, Z., Tang, J., & Zhou, J. (2018). Protein Translocation through a MoS₂ Nanopore: A Molecular Dynamics Study. *Journal of Physical Chemistry C*, *122*(4), 2070–2080. <https://doi.org/10.1021/acs.jpcc.7b07842>
- Chen, H., Maia, J. D. C., Radak, B. K., Hardy, D. J., Cai, W., Chipot, C., & Tajkhorshid, E. (2020). Boosting Free-Energy Perturbation Calculations with GPU-Accelerated NAMD. *Journal of Chemical Information and Modeling*, *60*(11), 5301–5307. <https://doi.org/10.1021/acs.jcim.0c00745>
- Chenjing Cai, Shiwei Wang, Youjun Xu, Weilin Zhang, Ke Tang, Qi Ouyang, Luhua Lai, and Jianfeng Pei. (2020). Transfer Learning for Drug Discovery. *J. Med. Chem.*, *63*, 8683–8694.

- <https://doi.org/10.1021/acs.jmedchem.9b02147.s002>
- Chen, J. M., Xu, S. L., Wawrzak, Z., Basarab, G. S., & Jordan, D. B. (1998). Structure-based design of potent inhibitors of scytalone dehydratase: displacement of a water molecule from the active site. *Biochemistry*, *37*(51), 17735–17744. <https://doi.org/10.1021/bi981848r>
- Chen, Y., & Roux, B. (2015). Generalized Metropolis acceptance criterion for hybrid non-equilibrium molecular dynamics-Monte Carlo simulations. *The Journal of Chemical Physics*, *142*(2), 024101. <https://doi.org/10.1063/1.4904889>
- Chéron, N., Jasty, N., & Shakhnovich, E. I. (2016). OpenGrowth: An Automated and Rational Algorithm for Finding New Protein Ligands. *Journal of Medicinal Chemistry*, *59*(9), 4171–4188. <https://doi.org/10.1021/acs.jmedchem.5b00886>
- Cimermancic, P., Weinkam, P., Rettenmaier, T. J., Bichmann, L., Keedy, D. A., Woldeyes, R. A., Schneidman-Duhovny, D., Demerdash, O. N., Mitchell, J. C., Wells, J. A., Fraser, J. S., & Sali, A. (2016). CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. *Journal of Molecular Biology*, *428*(4), 709–719. <https://doi.org/10.1016/j.jmb.2016.01.029>
- Claußen, H., Buning, C., Rarey, M., & Lengauer, T. (2001). FlexE: efficient molecular docking considering protein structure variations. Edited by J. Thornton. In *Journal of Molecular Biology* (Vol. 308, Issue 2, pp. 377–395). <https://doi.org/10.1006/jmbi.2001.4551>
- Cleaves, H. J. (2011). Affinity Constant. In M. Gargaud, R. Amils, J. C. Quintanilla, H. J. (jim) Cleaves, W. M. Irvine, D. L. Pinti, & M. Viso (Eds.), *Encyclopedia of Astrobiology* (pp. 23–24). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-11274-4_40
- Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. <https://doi.org/10.1093/bioinformatics/btp163>
- Cole, J. C., Murray, C. W., Nissink, J. W. M., Taylor, R. D., & Taylor, R. (2005). Comparing protein-ligand docking programs is difficult. *Proteins*, *60*(3), 325–332. <https://doi.org/10.1002/prot.20497>
- Cournia, Z., Allen, B., & Sherman, W. (2017). Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *Journal of Chemical Information and Modeling*, *57*(12), 2911–2937. <https://doi.org/10.1021/acs.jcim.7b00564>
- Crick, F. (1970). Central Dogma of Molecular Biology. *Nature*, *227*(5258), 561–563. <https://doi.org/10.1038/227561a0>
- Dara, S., Dhamecherla, S., Jadav, S. S., Babu, C. M., & Ahsan, M. J. (2021). Machine Learning in Drug Discovery:

- A Review. *Artificial Intelligence Review*, 1–53. <https://doi.org/10.1007/s10462-021-10058-4>
- Darby, J. F., Hopkins, A. P., Shimizu, S., Roberts, S. M., Brannigan, J. A., Turkenburg, J. P., Thomas, G. H., Hubbard, R. E., & Fischer, M. (2019). Water Networks Can Determine the Affinity of Ligand Binding to Proteins. *Journal of the American Chemical Society*, *141*(40), 15818–15826. <https://doi.org/10.1021/jacs.9b06275>
- Dean, P. M., Firth-Clark, S., Harris, W., Kirton, S. B., & Todorov, N. P. (2006). SkelGen: a general tool for structure-based de novo ligand design. *Expert Opinion on Drug Discovery*, *1*(2), 179–189. <https://doi.org/10.1517/17460441.1.2.179>
- Dieckman, L. M., Freudenthal, B. D., & Washington, M. T. (2012). PCNA structure and function: insights from structures of PCNA complexes and post-translationally modified PCNA. *Sub-Cellular Biochemistry*, *62*, 281–299. https://doi.org/10.1007/978-94-007-4572-8_15
- Douguet, D., Munier-Lehmann, H., Labesse, G., & Pochet, S. (2005). LEA3D: a computer-aided ligand design for structure-based drug design. *Journal of Medicinal Chemistry*, *48*(7), 2457–2468. <https://doi.org/10.1021/jm0492296>
- Durrant, J. D., Lindert, S., & McCammon, J. A. (2013). AutoGrow 3.0: an improved algorithm for chemically tractable, semi-automated protein inhibitor design. *Journal of Molecular Graphics & Modelling*, *44*, 104–112. <https://doi.org/10.1016/j.jmgm.2013.05.006>
- Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., & Pande, V. S. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology*, *13*(7), e1005659. <https://doi.org/10.1371/journal.pcbi.1005659>
- Eckhardt, R. (1987). Stan ulam, john von neumann, and the monte carlo method. *Los Alamos Science / Los Alamos Scientific Laboratory*, *15*(30), 131–136. [https://books.google.com/books?hl=en&lr=&id=JTEZAQAIAAJ&oi=fnd&pg=PA131&dq=Eckhardt+R+\(1987\)+Stan+Ulam+John+von+Neumann+and+the+Monte+Carlo+method+Los+Alamos+Science+15\(131-136\)+30&ots=qy_itCH3kF&sig=kOs3GFi3YaItgmh16zM33L_m5vU](https://books.google.com/books?hl=en&lr=&id=JTEZAQAIAAJ&oi=fnd&pg=PA131&dq=Eckhardt+R+(1987)+Stan+Ulam+John+von+Neumann+and+the+Monte+Carlo+method+Los+Alamos+Science+15(131-136)+30&ots=qy_itCH3kF&sig=kOs3GFi3YaItgmh16zM33L_m5vU)
- Ekins, S., Godbole, A. A., Kéri, G., Orfi, L., Pato, J., Bhat, R. S., Verma, R., Bradley, E. K., & Nagaraja, V. (2017). Machine learning and docking models for Mycobacterium tuberculosis topoisomerase I. *Tuberculosis*, *103*, 52–60. <https://doi.org/10.1016/j.tube.2017.01.005>
- Emmerich, C. H., Gamboa, L. M., Hofmann, M. C. J., Bonin-Andresen, M., Arbach, O., Schendel, P., Gerlach, B., Hempel, K., Bespalov, A., Dirnagl, U., & Parnham, M. J. (2021). Improving target assessment in biomedical research: the GOT-IT recommendations. *Nature Reviews. Drug Discovery*, *20*(1), 64–81.

- <https://doi.org/10.1038/s41573-020-0087-3>
- Erlanson, D. A., Fesik, S. W., Hubbard, R. E., Jahnke, W., & Jhoti, H. (2016). Twenty years on: the impact of fragments on drug discovery. In *Nature Reviews Drug Discovery* (Vol. 15, Issue 9, pp. 605–619). <https://doi.org/10.1038/nrd.2016.109>
- Everett, J. R. (2015). Academic drug discovery: current status and prospects. *Expert Opinion on Drug Discovery*, 10(9). <https://doi.org/10.1517/17460441.2015.1059816>
- Expression of EGFR in cancer - summary - the human protein atlas*. (n.d.). Retrieved July 15, 2021, from <https://www.proteinatlas.org/ENSG00000146648-EGFR/pathology>
- Fernandez-Lozano, C., Gestal, M., Munteanu, C. R., Dorado, J., & Pazos, A. (2016). A methodology for the design of experiments in computational intelligence with multiple regression models. *PeerJ*, 4, e2721. <https://doi.org/10.7717/peerj.2721>
- Fichthorn, K. A., & Weinberg, W. H. (1991). Theoretical foundations of dynamical Monte Carlo simulations. *The Journal of Chemical Physics*, 95(2), 1090–1096. <https://doi.org/10.1063/1.461138>
- Foloppe, N., Fisher, L. M., Francis, G., Howes, R., Kierstan, P., & Potter, A. (2006). Identification of a buried pocket for potent and selective inhibition of Chk1: prediction and verification. *Bioorganic & Medicinal Chemistry*, 14(6), 1792–1804. <https://doi.org/10.1016/j.bmc.2005.10.022>
- Fox, S., Farr-Jones, S., Sopchak, L., Boggs, A., Nicely, H. W., Khoury, R., & Biro, M. (2006). High-throughput screening: update on practices and success. *Journal of Biomolecular Screening*, 11(7), 864–869. <https://doi.org/10.1177/1087057106292473>
- Fratev, F., & Sirimulla, S. (2019). An Improved Free Energy Perturbation FEP+ Sampling Protocol for Flexible Ligand-Binding Domains. *Scientific Reports*, 9(1), 16829. <https://doi.org/10.1038/s41598-019-53133-1>
- Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., & Shenkin, P. S. (2004a). Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. In *Journal of Medicinal Chemistry* (Vol. 47, Issue 7, pp. 1739–1749). <https://doi.org/10.1021/jm0306430>
- Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., & Shenkin, P. S. (2004b). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7), 1739–1749. <https://doi.org/10.1021/jm0306430>
- Gabb, H. A., Jackson, R. M., & Sternberg, M. J. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *Journal of Molecular Biology*, 272(1), 106–120. <https://doi.org/10.1006/jmbi.1997.1203>

- Gallicchio, E., Zhang, L. Y., & Levy, R. M. (2002). The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. *Journal of Computational Chemistry*, 23(5), 517–529. <https://doi.org/10.1002/jcc.10045>
- Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., & Overington, J. P. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(Database issue), D1100–D1107. <https://doi.org/10.1093/nar/gkr777>
- Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M. P., Overington, J. P., Papadatos, G., Smit, I., & Leach, A. R. (2016). The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1), D945–D954. <https://doi.org/10.1093/nar/gkw1074>
- Genheden, S., Reymer, A., Saenz-Méndez, P., & Eriksson, L. A. (2017). Chapter 1: Computational Chemistry and Molecular Modelling Basics. In *Computational Tools for Chemical Biology* (pp. 1–38). <https://doi.org/10.1039/9781788010139-00001>
- Genheden, S., & Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery*, 10(5), 449–461. <https://doi.org/10.1517/17460441.2015.1032936>
- Gentile, F., Agrawal, V., Hsing, M., Ton, A.-T., Ban, F., Norinder, U., Gleave, M. E., & Cherkasov, A. (2020). Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. In *ACS Central Science* (Vol. 6, Issue 6, pp. 939–949). <https://doi.org/10.1021/acscentsci.0c00229>
- Gentil, S., Rousselot-Pailley, P., Sancho, F., Robert, V., Mekmouche, Y., Guallar, V., Tron, T., & Le Goff, A. (2020). Efficiency of Site-Specific Clicked Laccase-Carbon Nanotubes Biocathodes towards O Reduction. *Chemistry*, 26(21), 4798–4804. <https://doi.org/10.1002/chem.201905234>
- Gilabert, J. F., Gracia Carmona, O., Hogner, A., & Guallar, V. (2020). Combining Monte Carlo and Molecular Dynamics Simulations for Enhanced Binding Free Energy Estimation through Markov State Models. *Journal of Chemical Information and Modeling*, 60(11), 5529–5539. <https://doi.org/10.1021/acs.jcim.0c00406>
- Gilabert, J. F., Grebner, C., Soler, D., Lecina, D., Municoy, M., Carmona, O. G., Soliva, R., Packer, M. J., Hughes, S. J., Tyrchan, C., Hogner, A., & Guallar, V. (2019). PELE-MSM: A Monte Carlo Based Protocol for the Estimation of Absolute Binding Free Energies. In *Journal of Chemical Theory and Computation* (Vol. 15, Issue 11, pp. 6243–6253). <https://doi.org/10.1021/acs.jctc.9b00753>
- Gilabert, J. F., Lecina, D., Estrada, J., & Guallar, V. (2018). Monte Carlo Techniques for Drug Design: The Success Case of PELE. In F. L. Gervasio & V. Spiwok (Eds.), *Biomolecular Simulations in Structure-Based Drug Discovery* (Vol. 5, pp. 87–103). Wiley-VCH Verlag GmbH & Co. KGaA. <https://doi.org/10.1002/9783527806836.ch5>

- Gilson, M. K., & Zhou, H.-X. (2007). Calculation of protein-ligand binding affinities. *Annual Review of Biophysics and Biomolecular Structure*, 36, 21–42. <https://doi.org/10.1146/annurev.biophys.36.040306.132550>
- Jimeno, A., Ojeda-Montes, M. J., Tomás-Hernández, S., Cereto-Massagué, A., Beltrán-Debón, R., Mulero, M., Pujadas, G., & Garcia-Vallvé, S. (2019). The Light and Dark Sides of Virtual Screening: What Is There to Know? *International Journal of Molecular Sciences*, 20(6). <https://doi.org/10.3390/ijms20061375>
- Grebner, C., Iegre, J., Ulander, J., Edman, K., Hogner, A., & Tyrchan, C. (2016). Binding Mode and Induced Fit Predictions for Prospective Computational Drug Design. *Journal of Chemical Information and Modeling*, 56(4), 774–787. <https://doi.org/10.1021/acs.jcim.5b00744>
- Greenwood, J. R., Calkins, D., Sullivan, A. P., & Shelley, J. C. (2010). Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *Journal of Computer-Aided Molecular Design*, 24(6-7), 591–604. <https://doi.org/10.1007/s10822-010-9349-1>
- Gupta, A., & Zhou, H.-X. (2021). Machine Learning-Enabled Pipeline for Large-Scale Virtual Drug Screening. *Journal of Chemical Information and Modeling*, 61(9), 4236–4244. <https://doi.org/10.1021/acs.jcim.1c00710>
- Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Thomas Pollard, W., & Banks, J. L. (2004). Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. In *Journal of Medicinal Chemistry* (Vol. 47, Issue 7, pp. 1750–1759). <https://doi.org/10.1021/jm030644s>
- Hans-Joachim Anders, V. V. (2007). Identifying and validating novel targets with in vivo disease models: Guidelines for study design. *Drug Discovery Today*, 12(11-12), 446–451. <https://doi.org/10.1016/j.drudis.2007.04.001>
- Harder, E., Damm, W., Maple, J., Wu, C., Reboul, M., Xiang, J. Y., Wang, L., Lupyan, D., Dahlgren, M. K., Knight, J. L., Kaus, J. W., Cerutti, D. S., Krilov, G., Jorgensen, W. L., Abel, R., & Friesner, R. A. (2016). OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *Journal of Chemical Theory and Computation*, 12(1), 281–296. <https://doi.org/10.1021/acs.jctc.5b00864>
- Hartshorn, M. J., Murray, C. W., Cleasby, A., Frederickson, M., Tickle, I. J., & Jhoti, H. (2005). Fragment-based lead discovery using X-ray crystallography. *Journal of Medicinal Chemistry*, 48(2), 403–413. <https://doi.org/10.1021/jm0495778>
- 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>
- Howard, S., Amin, N., Benowitz, A. B., Chiarparin, E., Cui, H., Deng, X., Heightman, T. D., Holmes, D. J., Hopkins, A., Huang, J., Jin, Q., Kretsoulas, C., Martin, A. C. L., Massey, F., McCloskey, L., Mortenson, P. N., Pathuri, P., Tisi, D., & Williams, P. A. (2013). Fragment-based discovery of 6-azaindazoles as inhibitors of bacterial DNA ligase. *ACS Medicinal Chemistry Letters*, 4(12), 1208–1212. <https://doi.org/10.1021/ml4003277>

- Howard, S., Amin, N., Benowitz, A. B., Chiarparin, E., Cui, H., Deng, X., Heightman, T. D., Holmes, D. J., Hopkins, A., Huang, J., Jin, Q., Kreatsoulas, C., Martin, A. C. L., Massey, F., McCloskey, L., Mortenson, P. N., Pathuri, P., Tisi, D., & Williams, P. A. (2014). *Fragment-Based Discovery of 6 Azaindazoles As Inhibitors of Bacterial DNA Ligase*. <https://doi.org/10.2210/pdb4cc6/pdb>
- Huang, D., Zhou, T., Lafleur, K., Nevado, C., & Caflisch, A. (2010). Kinase selectivity potential for inhibitors targeting the ATP binding site: a network analysis. *Bioinformatics*, *26*(2), 198–204. <https://doi.org/10.1093/bioinformatics/btp650>
- Huang, S.-Y. (2018). Comprehensive assessment of flexible-ligand docking algorithms: current effectiveness and challenges. *Briefings in Bioinformatics*, *19*(5), 982–994. <https://doi.org/10.1093/bib/bbx030>
- Huang, S.-Y., & Zou, X. (2007). Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. *Proteins*, *66*(2), 399–421. <https://doi.org/10.1002/prot.21214>
- Hughes, A. B. (Ed.). (2009). *Amino acids, peptides and proteins in organic chemistry: Origins and synthesis of amino acids*. Wiley-VCH Verlag. <https://doi.org/10.1002/9783527631766>
- Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. *British Journal of Pharmacology*, *162*(6), 1239. <https://doi.org/10.1111/j.1476-5381.2010.01127.x>
- Ichihara, O., Shimada, Y., & Yoshidome, D. (2014). The importance of hydration thermodynamics in fragment-to-lead optimization. *ChemMedChem*, *9*(12), 2708–2717. <https://doi.org/10.1002/cmdc.201402207>
- Irwin, J. J., & Shoichet, B. K. (2005). ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. In *Journal of Chemical Information and Modeling* (Vol. 45, Issue 1, pp. 177–182). <https://doi.org/10.1021/ci049714+>
- Jacobson, M. P., Kaminski, G. A., Friesner, R. A., & Rapp, C. S. (2002). Force Field Validation Using Protein Side Chain Prediction. In *The Journal of Physical Chemistry B* (Vol. 106, Issue 44, pp. 11673–11680). <https://doi.org/10.1021/jp021564n>
- Jang, H., Hall, C. K., & Zhou, Y. (2002). Protein folding pathways and kinetics: molecular dynamics simulations of beta-strand motifs. *Biophysical Journal*, *83*(2), 819–835. [https://doi.org/10.1016/S0006-3495\(02\)75211-9](https://doi.org/10.1016/S0006-3495(02)75211-9)
- Jones, G., Willett, P., Glen, R. C., Leach, A. R., & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, *267*(3), 727–748. <https://doi.org/10.1006/jmbi.1996.0897>
- Joo, S., Kim, M. S., Yang, J., & Park, J. (2020). Generative Model for Proposing Drug Candidates Satisfying Anticancer Properties Using a Conditional Variational Autoencoder. *ACS Omega*, *5*(30), 18642–18650. <https://doi.org/10.1021/acsomega.0c01149>
- Jorgensen, W. L. (2009). Efficient drug lead discovery and optimization. *Accounts of Chemical Research*, *42*(6),

- 724–733. <https://doi.org/10.1021/ar800236t>
- Jorgensen, W. L., Maxwell, D. S., & Tirado-Rives, J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society*, *118*(45), 11225–11236. <https://doi.org/10.1021/ja9621760>
- Jorgensen, W. L., & Ravimohan, C. (1985). Monte Carlo simulation of differences in free energies of hydration. *The Journal of Chemical Physics*, *83*(6), 3050–3054. <https://doi.org/10.1063/1.449208>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., & Jorgensen, W. L. (2001). Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides†. In *The Journal of Physical Chemistry B* (Vol. 105, Issue 28, pp. 6474–6487). <https://doi.org/10.1021/jp003919d>
- Kedika, S. R., & Udugamasooriya, D. G. (2018). Converting a weaker ATP-binding site inhibitor into a potent hetero-bivalent ligand by tethering to a unique peptide sequence derived from the same kinase. *Organic & Biomolecular Chemistry*, *16*(35), 6443–6449. <https://doi.org/10.1039/c8ob01406j>
- Kidd, S. L., Osberger, T. J., Mateu, N., Sore, H. F., & Spring, D. R. (2018). Recent Applications of Diversity-Oriented Synthesis Toward Novel, 3-Dimensional Fragment Collections. *Frontiers in Chemistry*, *6*, 460. <https://doi.org/10.3389/fchem.2018.00460>
- Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2020). PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research*, *49*(D1), D1388–D1395. <https://doi.org/10.1093/nar/gkaa971>
- Kingma, D. P., & Welling, M. (2013). *Auto-Encoding Variational Bayes*. <http://arxiv.org/abs/1312.6114>
- Kinnings, S. L., Liu, N., Tonge, P. J., Jackson, R. M., Xie, L., & Bourne, P. E. (2011). A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *Journal of Chemical Information and Modeling*, *51*(2), 408–419. <https://doi.org/10.1021/ci100369f>
- Kirkwood, J. G. (1935). Statistical Mechanics of Fluid Mixtures. *The Journal of Chemical Physics*, *3*(5), 300–313. <https://doi.org/10.1063/1.1749657>
- Kitchen, D. B., Decornez, H., Furr, J. R., & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews. Drug Discovery*, *3*(11), 935–949. <https://doi.org/10.1038/nrd1549>

- Knegtel, R. M. A., Kuntz, I. D., & Oshiro, C. M. (1997). Molecular docking to ensembles of protein structures 1 Edited by B. Honig. In *Journal of Molecular Biology* (Vol. 266, Issue 2, pp. 424–440). <https://doi.org/10.1006/jmbi.1996.0776>
- Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A., & Cheatham, T. E., 3rd. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of Chemical Research*, 33(12), 889–897. <https://doi.org/10.1021/ar000033j>
- Kumar, A., & Zhang, K. Y. J. (2018). Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. In *Frontiers in Chemistry* (Vol. 6). <https://doi.org/10.3389/fchem.2018.00315>
- Kung, P.-P., Sinnema, P.-J., Richardson, P., Hickey, M. J., Gajiwala, K. S., Wang, F., Huang, B., McClellan, G., Wang, J., Maegley, K., Bergqvist, S., Mehta, P. P., & Kania, R. (2011). Design strategies to target crystallographic waters applied to the Hsp90 molecular chaperone. *Bioorganic & Medicinal Chemistry Letters*, 21(12), 3557–3562. <https://doi.org/10.1016/j.bmcl.2011.04.130>
- Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., & Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology*, 161(2), 269–288. [https://doi.org/10.1016/0022-2836\(82\)90153-x](https://doi.org/10.1016/0022-2836(82)90153-x)
- Lecina, D., Gilbert, J. F., & Guallar, V. (2017). Adaptive simulations, towards interactive protein-ligand modeling. *Scientific Reports*, 7(1), 1–11. <https://doi.org/10.1038/s41598-017-08445-5>
- Lee, T.-S., Lin, Z., Allen, B. K., Lin, C., Radak, B. K., Tao, Y., Tsai, H.-C., Sherman, W., & York, D. M. (2020). Improved Alchemical Free Energy Calculations with Optimized Smoothstep Softcore Potentials. *Journal of Chemical Theory and Computation*, 16(9), 5512–5525. <https://doi.org/10.1021/acs.jctc.0c00237>
- Levy, R. M., Zhang, L. Y., Gallicchio, E., & Felts, A. K. (2003). On the nonpolar hydration free energy of proteins: surface area and continuum solvent models for the solute-solvent interaction energy. *Journal of the American Chemical Society*, 125(31), 9523–9530. <https://doi.org/10.1021/ja029833a>
- Libretexts. (2015, June 3). *Bond Lengths and Energies*. Libretexts. [https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_\(Physical_and_Theoretical_Chemistry\)/Chemical_Bonding/Fundamentals_of_Chemical_Bonding/Chemical_Bonds/Bond_Lengths_and_Energies](https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Chemical_Bonding/Fundamentals_of_Chemical_Bonding/Chemical_Bonds/Bond_Lengths_and_Energies)
- Lim, V. T., Hahn, D. F., Tresadern, G., Bayly, C. I., & Mobley, D. L. (2020). Benchmark assessment of molecular geometries and energies from small molecule force fields. *F1000Research*, 9. <https://doi.org/10.12688/f1000research.27141.1>
- Liverton, N. J., Butcher, J. W., Claiborne, C. F., Claremon, D. A., Libby, B. E., Nguyen, K. T., Pitznerberger, S. M.,

- Selnick, H. G., Smith, G. R., Tebben, A., Vacca, J. P., Varga, S. L., Agarwal, L., Dancheck, K., Forsyth, A. J., Fletcher, D. S., Frantz, B., Hanlon, W. A., Harper, C. F., ... O'Keefe, S. J. (1999). Design and synthesis of potent, selective, and orally bioavailable tetrasubstituted imidazole inhibitors of p38 mitogen-activated protein kinase. *Journal of Medicinal Chemistry*, *42*(12), 2180–2190. <https://doi.org/10.1021/jm9805236>
- Lund, M., Trulsson, M., & Persson, B. (2008). Faunus: An object oriented framework for molecular simulation. *Source Code for Biology and Medicine*, *3*, 1. <https://doi.org/10.1186/1751-0473-3-1>
- Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., & Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, *11*(8), 3696–3713. <https://doi.org/10.1021/acs.jctc.5b00255>
- Martínez-Martínez, M., Coscolín, C., Santiago, G., Chow, J., Stogios, P. J., Bargiela, R., Gertler, C., Navarro-Fernández, J., Bollinger, A., Thies, S., Méndez-García, C., Popovic, A., Brown, G., Chernikova, T. N., García-Moyano, A., Bjerga, G. E. K., Pérez-García, P., Hai, T., Del Pozo, M. V., ... The Inmare Consortium. (2018). Determinants and Prediction of Esterase Substrate Promiscuity Patterns. *ACS Chemical Biology*, *13*(1), 225–234. <https://doi.org/10.1021/acscchembio.7b00996>
- McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., Schwantes, C. R., Wang, L.-P., Lane, T. J., & Pande, V. S. (2015). MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal*, *109*(8), 1528–1532. <https://doi.org/10.1016/j.bpj.2015.08.015>
- McKinney, W., & Others. (2011). pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, *14*(9), 1–9. https://www.dlr.de/sc/portaldata/15/resources/dokumente/pyhpc2011/submissions/pyhpc2011_submission_9.pdf
- Meng, X.-Y., Zhang, H.-X., Mezei, M., & Cui, M. (2011). Molecular docking: a powerful approach for structure-based drug discovery. *Current Computer-Aided Drug Design*, *7*(2), 146–157. <https://doi.org/10.2174/157340911795677602>
- Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, *44*(247), 335–341. <https://doi.org/10.1080/01621459.1949.10483310>
- Metwalli, S. A. (2020, September 21). *How to Choose the Right Machine Learning Algorithm for Your Application*. Towards Data Science. <https://towardsdatascience.com/how-to-choose-the-right-machine-learning-algorithm-for-your-application-1e36c32400b9>
- Michaelis, L., Menten, M. L., Johnson, K. A., & Goody, R. S. (2011). The original Michaelis constant: translation of the 1913 Michaelis-Menten paper. *Biochemistry*, *50*(39), 8264–8269. <https://doi.org/10.1021/bi201284u>
- Michel, J., Tirado-Rives, J., & Jorgensen, W. L. (2009). Energetics of displacing water molecules from protein

- binding sites: consequences for ligand optimization. *Journal of the American Chemical Society*, 131(42), 15403–15411. <https://doi.org/10.1021/ja906058w>
- Miller, M. D., Kearsley, S. K., Underwood, D. J., & Sheridan, R. P. (1994). FLOG: A system to select ?quasi-flexible? ligands complementary to a receptor of known three-dimensional structure. In *Journal of Computer-Aided Molecular Design* (Vol. 8, Issue 2, pp. 153–174). <https://doi.org/10.1007/bf00119865>
- Mishra, C., & Gupta, D. L. (2017). Deep Machine Learning and Neural Networks: An Overview. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 6(2), 66–73. <https://doi.org/10.11591/ijai.v6.i2.pp66-73>
- Monte Carlo Method. (2008). In *The Concise Encyclopedia of Statistics* (pp. 359–360). Springer New York. https://doi.org/10.1007/978-0-387-32833-1_270
- Moore, W. J. (1973). Physical chemistry. *Journal of Molecular Structure*, 17(2), 434–435. [https://doi.org/10.1016/0022-2860\(73\)85191-9](https://doi.org/10.1016/0022-2860(73)85191-9)
- Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., & Olson, A. J. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. In *Journal of Computational Chemistry* (Vol. 19, Issue 14, pp. 1639–1662). [https://doi.org/10.1002/\(sici\)1096-987x\(19981115\)19:14<1639::aid-jcc10>3.0.co;2-b](https://doi.org/10.1002/(sici)1096-987x(19981115)19:14<1639::aid-jcc10>3.0.co;2-b)
- Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, 30(16), 2785–2791. <https://doi.org/10.1002/jcc.21256>
- Municoy, M., Roda, S., Soler, D., Soutullo, A., & Guallar, V. (2020). aquAPELE: A Monte Carlo-Based Algorithm to Sample the Effects of Buried Water Molecules in Proteins. *Journal of Chemical Theory and Computation*, 16(12), 7655–7670. <https://doi.org/10.1021/acs.jctc.0c00925>
- Murray, C. W., & Rees, D. C. (2009). The rise of fragment-based drug discovery. In *Nature Chemistry* (Vol. 1, Issue 3, pp. 187–192). <https://doi.org/10.1038/nchem.217>
- Nelson, M. T., Humphrey, W., Gursoy, A., Dalke, A., Kalé, L. V., Skeel, R. D., & Schulten, K. (1996). NAMD: a Parallel, Object-Oriented Molecular Dynamics Program. *The International Journal of Supercomputer Applications and High Performance Computing*, 10(4), 251–268. <https://doi.org/10.1177/109434209601000401>
- Nittinger, E., Gibbons, P., Eigenbrot, C., Davies, D. R., Maurer, B., Yu, C. L., Kiefer, J. R., Kuglstatler, A., Murray, J., Ortwine, D. F., Tang, Y., & Tsui, V. (2019). Water molecules in protein–ligand interfaces. Evaluation of software tools and SAR comparison. In *Journal of Computer-Aided Molecular Design* (Vol. 33, Issue 3, pp. 307–330). <https://doi.org/10.1007/s10822-019-00187-y>
- Ode, H., Nakashima, M., Kitamura, S., Sugiura, W., & Sato, H. (2012). Molecular dynamics simulation in virus research. *Frontiers in Microbiology*, 0. <https://doi.org/10.3389/fmicb.2012.00258>

- Oleinikovas, V., Saladino, G., Cossins, B. P., & Gervasio, F. L. (2016). *Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations*. <https://doi.org/10.1021/jacs.6b05425>
- Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M., & Jensen, J. H. (2011). PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *Journal of Chemical Theory and Computation*, 7(2), 525–537. <https://doi.org/10.1021/ct100578z>
- Onufriev, A., Bashford, D., & Case, D. A. (2004). Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Structure, Function, and Bioinformatics*. <https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.20033>
- Oostenbrink, C., Villa, A., Mark, A. E., & Van Gunsteren, W. F. (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. In *Journal of Computational Chemistry* (Vol. 25, Issue 13, pp. 1656–1676). <https://doi.org/10.1002/jcc.20090>
- Paquet, E., & Viktor, H. L. (2015). Molecular dynamics, monte carlo simulations, and langevin dynamics: a computational review. *BioMed Research International*, 2015, 183918. <https://doi.org/10.1155/2015/183918>
- Patel, S., & Brooks, C. L., 3rd. (2004). CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *Journal of Computational Chemistry*, 25(1), 1–15. <https://doi.org/10.1002/jcc.10355>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research: JMLR*, 12(85), 2825–2830. <http://jmlr.org/papers/v12/pedregosa11a.html>
- Pérez, A., Marchán, I., Svozil, D., Spöner, J., Cheatham, T. E., Loughton, C. A., & Orozco, M. (2007). Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. In *Biophysical Journal* (Vol. 92, Issue 11, pp. 3817–3829). <https://doi.org/10.1529/biophysj.106.097782>
- Perez, C., Soler, D., Soliva, R., & Guallar, V. (2020). FragPELE: Dynamic Ligand Growing within a Binding Site. A Novel Tool for Hit-To-Lead Drug Design. *Journal of Chemical Information and Modeling*, 60(3), 1728–1736. <https://doi.org/10.1021/acs.jcim.9b00938>
- Petsko, G. A., & Ringe, D. (2004). *Protein Structure and Function*. New Science Press. https://books.google.com/books/about/Protein_Structure_and_Function.html?hl=&id=2yRDWkHhN9QC
- Price, D. J., & Brooks, C. L., 3rd. (2004). A modified TIP3P water potential for simulation with Ewald summation. *The Journal of Chemical Physics*, 121(20), 10096–10103. <https://doi.org/10.1063/1.1808117>
- ProtoMS. (n.d.). Retrieved June 18, 2021, from <http://www.essexgroup.soton.ac.uk/ProtoMS/index.html>
- Qiu, Y., Smith, D., Boothroyd, S., Jang, H., Wagner, J., Bannan, C. C., Gokey, T., Lim, V. T., Stern, C., Rizzi, A.,

-
- Lucas, X., Tjanaka, B., Shirts, M. R., Gilson, M., Chodera, J., Bayly, C. I., Mobley, D., & Wang, L.-P. (n.d.). *Development and Benchmarking of Open Force Field v1.0.0, the Parsley Small Molecule Force Field*. <https://doi.org/10.26434/chemrxiv.13082561.v2>
- Rarey, M., Kramer, B., Lengauer, T., & Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology*, *261*(3), 470–489. <https://doi.org/10.1006/jmbi.1996.0477>
- Rawluk, J., & Waller, C. F. (2018). Gefitinib. In *Recent Results in Cancer Research* (pp. 235–246). https://doi.org/10.1007/978-3-319-91442-8_16
- Renault, P., Louet, M., Marie, J., Labesse, G., & Floquet, N. (2019). Molecular Dynamics Simulations of the Allosteric Modulation of the Adenosine A2A Receptor by a Mini-G Protein. *Scientific Reports*, *9*(1), 5495. <https://doi.org/10.1038/s41598-019-41980-x>
- Report to Congressional Requesters: New Drug Development—Science, Business, Regulatory, And Intellectual Property Issues Cited as Hampering Drug Development Efforts. (2007). In *Biotechnology Law Report* (Vol. 26, Issue 1, pp. 82–95). <https://doi.org/10.1089/blr.2006.9996>
- Richter, L., & Ecker, G. F. (2015). Medicinal chemistry in the era of big data. *Drug Discovery Today. Technologies*, *14*, 37–41. <https://doi.org/10.1016/j.ddtec.2015.06.001>
- Robert J. Ouellette, J. D. R. (2018). *Amino Acids, Peptides, and Proteins* (J. D. R. Robert J. Ouellette (Ed.); pp. 929–971). Academic Press. <https://doi.org/10.1016/B978-0-12-812838-1.50029-3>
- Roda, S., Santiago, G., & Guallar, V. (2020). Mapping enzyme-substrate interactions: its potential to study the mechanism of enzymes. *Advances in Protein Chemistry and Structural Biology*, *122*, 1–31. <https://doi.org/10.1016/bs.apcsb.2020.06.001>
- Rogers, D., & Hahn, M. (2010). Extended-Connectivity Fingerprints. In *Journal of Chemical Information and Modeling* (Vol. 50, Issue 5, pp. 742–754). <https://doi.org/10.1021/ci100050t>
- Roskoski, R. (2015). Michaelis-Menten Kinetics☆. In *Reference Module in Biomedical Sciences*. Elsevier. <https://doi.org/10.1016/B978-0-12-801238-3.05143-6>
- Saen-Oon, S., Lozoya, E., Segarra, V., Guallar, V., & Soliva, R. (2019). Atomistic simulations shed new light on the activation mechanisms of ROR γ and classify it as Type III nuclear hormone receptor regarding ligand-binding paths. *Scientific Reports*, *9*(1), 17249. <https://doi.org/10.1038/s41598-019-52319-x>
- Santiago, G., Martínez-Martínez, M., Alonso, S., Bargiela, R., Coscolín, C., Golyshin, P. N., Guallar, V., & Ferrer, M. (2018). Rational Engineering of Multiple Active Sites in an Ester Hydrolase. *Biochemistry*, *57*(15), 2245–2255. <https://doi.org/10.1021/acs.biochem.8b00274>
- Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R., & Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular*

- Design*, 27(3), 221–234. <https://doi.org/10.1007/s10822-013-9644-8>
- Schneider, G., & Fechner, U. (2005). Computer-based de novo design of drug-like molecules. *Nature Reviews. Drug Discovery*, 4(8), 649–663. <https://doi.org/10.1038/nrd1799>
- Schrödinger. (2018). *Maestro* (Version 2018-4) [Computer software].
- Schrödinger. (2019). *CombiGlide* (Version 2019-3) [Computer software].
- Schrödinger, L., & DeLano, W. (2018). *PyMOL* (Version 2.4) [Computer software]. <http://www.pymol.org/pymol>
- Schulz, R., Atef, A., Becker, D., Gottschalk, F., Tauber, C., Wagner, S., Arkona, C., Abdel-Hafez, A. A., Farag, H. H., Rademann, J., & Wolber, G. (2018). Phenylthiomethyl Ketone-Based Fragments Show Selective and Irreversible Inhibition of Enteroviral 3C Proteases. *Journal of Medicinal Chemistry*, 61(3), 1218–1230. <https://doi.org/10.1021/acs.jmedchem.7b01440>
- Sehnal, D., Bittrich, S., Deshpande, M., Svobodová, R., Berka, K., Bazgier, V., Velankar, S., Burley, S. K., Koča, J., & Rose, A. S. (2021). Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Research*, 49(W1), W431–W437. <https://doi.org/10.1093/nar/gkab314>
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Židek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792), 706–710. <https://doi.org/10.1038/s41586-019-1923-7>
- Sherman, W., Beard, H. S., & Farid, R. (2006). Use of an induced fit receptor structure in virtual screening. *Chemical Biology & Drug Design*, 67(1), 83–84. <https://doi.org/10.1111/j.1747-0285.2005.00327.x>
- Shivakumar, D., Harder, E., Damm, W., Friesner, R. A., & Sherman, W. (2012). Improving the Prediction of Absolute Solvation Free Energies Using the Next Generation OPLS Force Field. *Journal of Chemical Theory and Computation*, 8(8), 2553–2558. <https://doi.org/10.1021/ct300203w>
- Sommer, K., Flachsenberg, F., & Rarey, M. (2019). NAOMInext - Synthetically feasible fragment growing in a structure-based design context. *European Journal of Medicinal Chemistry*, 163, 747–762. <https://doi.org/10.1016/j.ejmech.2018.11.075>
- Spyrakakis, F., Benedetti, P., Decherchi, S., Rocchia, W., Cavalli, A., Alcaro, S., Ortuso, F., Baroni, M., & Cruciani, G. (2015). A Pipeline To Enhance Ligand Virtual Screening: Integrating Molecular Dynamics and Fingerprints for Ligand and Proteins. *Journal of Chemical Information and Modeling*, 55(10), 2256–2274. <https://doi.org/10.1021/acs.jcim.5b00169>
- Srinivasan, B., Forouhar, F., Shukla, A., Sampangi, C., Kulkarni, S., Abashidze, M., Seetharaman, J., Lew, S., Mao, L., Acton, T. B., Xiao, R., Everett, J. K., Montelione, G. T., Tong, L., & Balaram, H. (2014). Allosteric regulation and substrate activation in cytosolic nucleotidase II from *Legionella pneumophila*. *The FEBS Journal*,

- 281(6), 1613–1628. <https://doi.org/10.1111/febs.12727>
- Steinbrecher, T. B., Dahlgren, M., Cappel, D., Lin, T., Wang, L., Krilov, G., Abel, R., Friesner, R., & Sherman, W. (2015). Accurate Binding Free Energy Predictions in Fragment Optimization. *Journal of Chemical Information and Modeling*, 55(11), 2411–2420. <https://doi.org/10.1021/acs.jcim.5b00538>
- Timm, D. E., Baker, L. J., Mueller, H., Zidek, L., & Novotny, M. V. (2001). Structural basis of pheromone binding to mouse major urinary protein (MUP-I). *Protein Science: A Publication of the Protein Society*, 10(5), 997–1004. <https://doi.org/10.1110/ps.52201>
- Ton, A.-T., Gentile, F., Hsing, M., Ban, F., & Cherkasov, A. (2020). Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds. *Molecular Informatics*, 39(8), e2000028. <https://doi.org/10.1002/minf.202000028>
- Trevizani, R., Custódio, F. L., Dos Santos, K. B., & Dardenne, L. E. (2017). Critical Features of Fragment Libraries for Protein Structure Prediction. *PLoS One*, 12(1), e0170131. <https://doi.org/10.1371/journal.pone.0170131>
- Truchon, J.-F., & Bayly, C. I. (2007). Evaluating virtual screening methods: good and bad metrics for the “early recognition” problem. *Journal of Chemical Information and Modeling*, 47(2), 488–508. <https://doi.org/10.1021/ci600426e>
- Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigartyo, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., ... Pontén, F. (2015). Proteomics. Tissue-based map of the human proteome. *Science*, 347(6220), 1260419. <https://doi.org/10.1126/science.1260419>
- Valdarrama, S. (2021, July 13). *Considerations when choosing a machine learning model*. Towards Data Science. <https://towardsdatascience.com/considerations-when-choosing-a-machine-learning-model-aa31f52c27f3>
- Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews. Drug Discovery*, 18(6), 463–477. <https://doi.org/10.1038/s41573-019-0024-5>
- van de Waterbeemd, H., & Gifford, E. (2003). ADMET in silico modelling: towards prediction paradise? *Nature Reviews. Drug Discovery*, 2(3), 192–204. <https://doi.org/10.1038/nrd1032>
- Van Norman, G. A. (2016). Drugs, Devices, and the FDA: Part 1: An Overview of Approval Processes for Drugs. *JACC. Basic to Translational Science*, 1(3), 170–179. <https://doi.org/10.1016/j.jacbts.2016.03.002>
- Vanommeslaeghe, K., Guvench, O., & MacKerell, A. D., Jr. (2014). Molecular mechanics. *Current Pharmaceutical Design*, 20(20), 3281–3292. <https://doi.org/10.2174/13816128113199990600>
- Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., & Mackerell, A. D. (2009). CHARMM general force field: A force field for drug-like molecules

- compatible with the CHARMM all-atom additive biological force fields. In *Journal of Computational Chemistry* (p. NA – NA). <https://doi.org/10.1002/jcc.21367>
- Vinkers, H. M., de Jonge, M. R., Daeyaert, F. F. D., Heeres, J., Koymans, L. M. H., van Lenthe, J. H., Lewi, P. J., Timmerman, H., Van Aken, K., & Janssen, P. A. J. (2003). SYNOPSIS: SYNthesize and OPTimize System in Silico. *Journal of Medicinal Chemistry*, *46*(13), 2765–2773. <https://doi.org/10.1021/jm030809x>
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. <https://doi.org/10.1038/s41592-019-0686-2>
- Voigtlaender, M., Schneider-Merck, T., & Trepel, M. (2018). Lapatinib. *Recent Results in Cancer Research. Fortschritte Der Krebsforschung. Progres Dans Les Recherches Sur Le Cancer*, *211*, 19–44. https://doi.org/10.1007/978-3-319-91442-8_2
- Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., & Case, D. A. (2004). Development and testing of a general amber force field. *Journal of Computational Chemistry*, *25*(9), 1157–1174. <https://doi.org/10.1002/jcc.20035>
- Wang, L., Wu, Y., Deng, Y., Kim, B., Pierce, L., Krilov, G., Lupyan, D., Robinson, S., Dahlgren, M. K., Greenwood, J., Romero, D. L., Masse, C., Knight, J. L., Steinbrecher, T., Beuming, T., Damm, W., Harder, E., Sherman, W., Brewer, M., ... Abel, R. (2015). Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society*, *137*(7), 2695–2703. <https://doi.org/10.1021/ja512751q>
- Wang, Z., Canagarajah, B. J., Boehm, J. C., Kassisà, S., Cobb, M. H., Young, P. R., Abdel-Meguid, S., Adams, J. L., & Goldsmith, E. J. (1998). Structural basis of inhibitor selectivity in MAP kinases. In *Structure* (Vol. 6, Issue 9, pp. 1117–1128). [https://doi.org/10.1016/s0969-2126\(98\)00113-0](https://doi.org/10.1016/s0969-2126(98)00113-0)
- Wermuth, C. G., Ganellin, C. R., Lindberg, P., & Mitscher, L. A. (1998). Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). *Pure and Applied Chemistry. Chimie Pure et Appliquee*, *70*(5), 1129–1143. <https://doi.org/10.1351/pac199870051129>
- Williams, J. R. (1996). Organic Chemistry. Sixth Edition By T. W. Graham Solomons. John Wiley & Sons, New York. 1995. xxvii 1290 pp. 18.5 × 26 cm. ISBN 0-471-01342-0. \$90.95. In *Journal of Medicinal Chemistry* (Vol. 39, Issue 18, pp. 3602–3602). <https://doi.org/10.1021/jm960425p>
- Wilson, K. (2010). Enzymes. In *Principles and Techniques of Biochemistry and Molecular Biology* (pp. 581–624). Cambridge University Press. <https://doi.org/10.1017/CBO9780511841477.016>
- Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., & Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*,

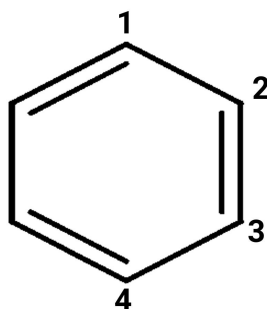
- 34(Database issue), D668–D672. <https://doi.org/10.1093/nar/gkj067>
- Wood, E. R., Truesdale, A. T., McDonald, O. B., Yuan, D., Hassell, A., Dickerson, S. H., Ellis, B., Pennisi, C., Horne, E., Lackey, K., Alligood, K. J., Rusnak, D. W., Gilmer, T. M., & Shewchuk, L. (2004). A unique structure for epidermal growth factor receptor bound to GW572016 (Lapatinib): relationships among protein conformation, inhibitor off-rate, and receptor activity in tumor cells. *Cancer Research*, *64*(18), 6652–6659. <https://doi.org/10.1158/0008-5472.CAN-04-1168>
- Woodhead, A. J., Angove, H., Carr, M. G., Chessari, G., Congreve, M., Coyle, J. E., Cosme, J., Graham, B., Day, P. J., Downham, R., Fazal, L., Feltell, R., Figueroa, E., Frederickson, M., Lewis, J., McMenamin, R., Murray, C. W., O'Brien, M. A., Parra, L., ... Woolford, A. J.-A. (2010). Discovery of (2,4-dihydroxy-5-isopropylphenyl)-[5-(4-methylpiperazin-1-ylmethyl)-1,3-dihydroisoindol-2-yl]methanone (AT13387), a novel inhibitor of the molecular chaperone Hsp90 by fragment based drug design. *Journal of Medicinal Chemistry*, *53*(16), 5956–5969. <https://doi.org/10.1021/jm100060b>
- Wouters, O. J., McKee, M., & Luyten, J. (2020). Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. In *JAMA* (Vol. 323, Issue 9, p. 844). <https://doi.org/10.1001/jama.2020.1166>
- Yongliang Yang, S James Adelstein, Amin I Kassis. (2009). Target discovery from data mining approaches. *Drug Discovery Today*, *14*(3-4), 147–154. <https://doi.org/10.1016/j.drudis.2008.12.005>
- Yuan, Y., Pei, J., & Lai, L. (2011). LigBuilder 2: a practical de novo drug design approach. *Journal of Chemical Information and Modeling*, *51*(5), 1083–1091. <https://doi.org/10.1021/ci100350u>
- Yung-Chi, C., & Prusoff, W. H. (1973). Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochemical Pharmacology*, *22*(23), 3099–3108. [https://doi.org/10.1016/0006-2952\(73\)90196-2](https://doi.org/10.1016/0006-2952(73)90196-2)
- Zhu, K., Shirts, M. R., & Friesner, R. A. (2007). Improved methods for side chain and loop predictions via the Protein Local Optimization Program: Variable dielectric model for implicitly improving the treatment of polarization effects. *Journal of Chemical Theory and Computation*, *3*(6), 2108–2119. <https://doi.org/10.1021/ct700166f>
- Zwanzig, R. W. (1954). High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics*, *22*(8), 1420–1426. <https://doi.org/10.1063/1.1740409>

Appendices

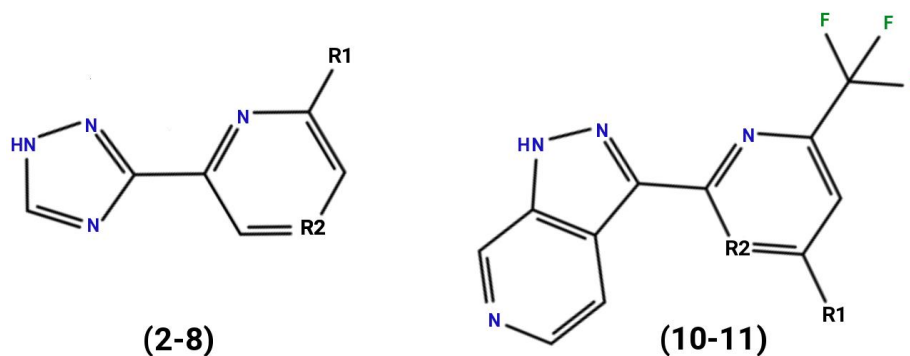
Appendix A

Experimental values from Steinbrecher et al. (Steinbrecher et al., 2015), used for the Growing and Scoring benchmark of FragPELE.

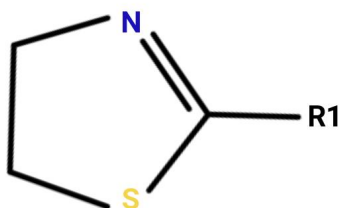
Table A1. Experimental values for T4 Lysozyme (181L)



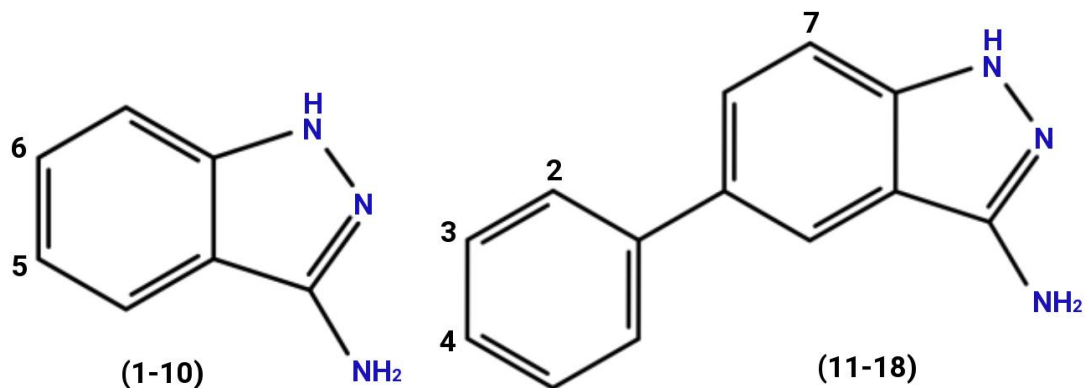
Ligand	R1	$\Delta G^{\circ}_{\text{Exp}}$ [kcal/mol]
LYS.1	butyl	-6.66
LYS.3	methyl	-5.48
LYS.4	1,2-methyl	-4.57
LYS.5	1,3-methyl	-4.63
LYS.6	1,4-methyl	-4.72
LYS.7	1-ethyl,2-methyl	-4.53
LYS.8	1-ethyl,3-methyl	-5.08
LYS.9	1-ethyl,4-methyl	-5.38
LYS.10	propyl	-6.51
LYS.11	isobutyl	-6.44
LYS.12	ethyl	-5.72

Table A2. Experimental data for DNA ligase (4CC5 from 2 to 8 and 4CC6 from 10 to 11)

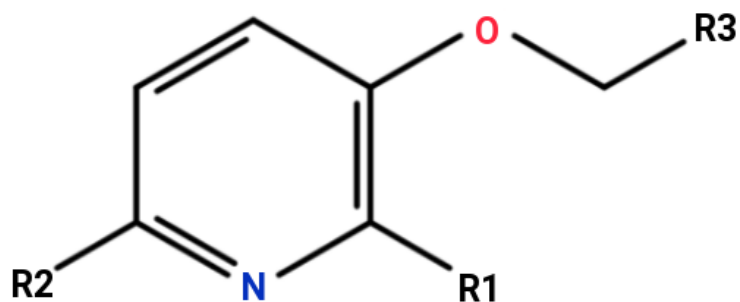
Ligand	R1	R2	$\Delta G^{\circ}_{\text{Exp}}$ [kcal/mol]
DNA.2	Cl	CH	-6.5
DNA.4	OMe	CH	-5.3
DNA.5	Me	CH	-5.25
DNA.6	cyclo-butyl	CH	-5.83
DNA.7	CF ₂ CH ₃	CH	-6.38
DNA.8	CF ₃	CH	-6.54
DNA.10	NHCH ₂ CHOH	CH	-9.05
DNA.11	NHCH ₂ CHOH	N	-10.67

Table A3. Experimental values for MUP-I (1106)

Ligand	R1	$\Delta G^{\circ}_{\text{Exp}}$ [kcal/mol]
MUP.1	S-sec-butyl	-8.42
MUP.2	R-sec-butyl	-8.42
MUP.3	iso-butyl	-9.09
MUP.4	n-propyl	-8.2
MUP.5	iso-propyl	-7.85
MUP.6	ethyl	-6.96
MUP.7	methyl	-5.64

Table A4. Experimental results for JAK-II (3E62 from 1 to 10 and 3E63 from 11 to 18)

Ligand	R1	$\Delta G^{\circ}_{\text{Exp}}$ [kcal/mol]
JAK.1	5-Br	-5.99
JAK.2	5-Phenyl	-7.91
JAK.3	6-Phenyl	-5.24
JAK.4	5-(3-Pyrazolyl)	-6.86
JAK.5	5-(4-Pyridyl)	-7.79
JAK.6	5-[4-(3,5-Dimethyl)isoxazolyl]	-6.2
JAK.7	5-[5-(2,4-Dimethyl)thiazolyl]	-8.86
JAK.8	5-(2-Chlorophenyl)	-8.77
JAK.9	5-(3-Chlorophenyl)	-8.36
JAK.10	5-[6-Chloro(2-pyridyl)]	-7.14
JAK.11	3-SO ₂ NHtBu	-7.43
JAK.12	4-SO ₂ NHtBu	-9.7
JAK.13	4-CONHtBu	-7.79
JAK.14	4-NHCONHtBu	-7.38
JAK.15	4-SO ₂ Ethyl	-8.33
JAK.16	4-SO ₂ NHtBu, 7-Cl	-8.07
JAK.17	4-SO ₂ NHtBu, 7-Me	-8.34
JAK.18	4-SO ₂ NHtBu, 2-Cl	-10.11

Table A5. Experimental results for p38 (1W7H)

Ligand	R1	R2	R3	$\Delta G^{\circ}_{\text{Exp}}$ [kcal/mol]
P38.1	NH ₂	H	phenyl	-3.94
P38.2	H	H	phenyl	-4.09
P38.3	NH ₂	H	2,6-dichloro-phenyl	-5.4
P38.4	NH ₂	H	1-naphthyl	-5.94
P38.5	H	ethanol-amine	2,6-dichloro-phenyl	-5.04
P38.6	H	1,1-dimethyl-ethanolamine	2,6-dichloro-phenyl	-6.3

Appendix B

Scatter plots of experimental values vs. energy predictions in Growing and Scoring benchmark of FragPELE.

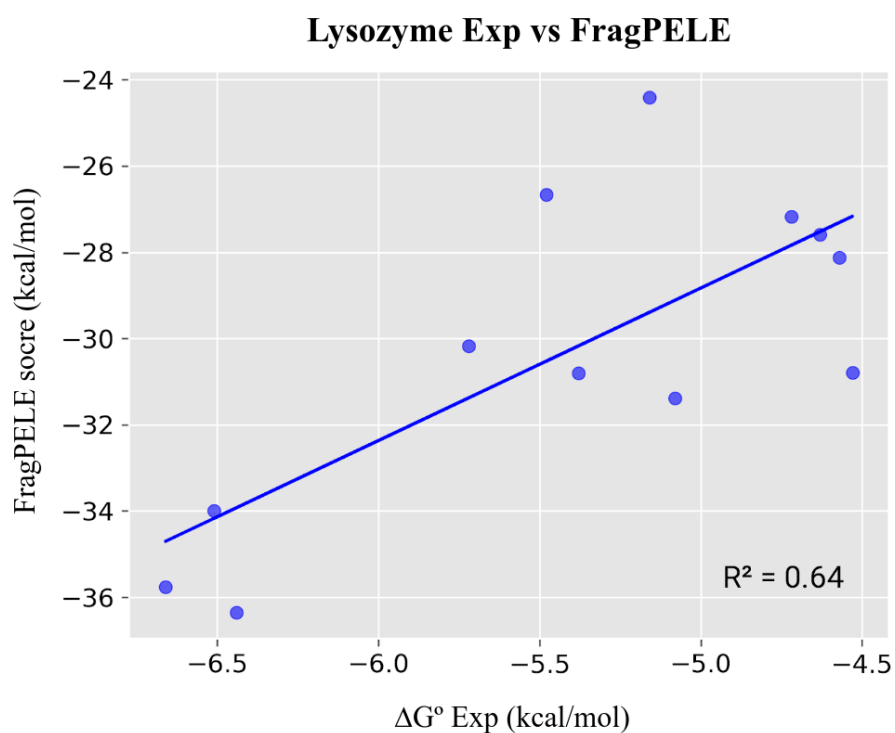


Figure B1. Correlation between experimental ΔG° values and FragPELE results for Lysozyme.

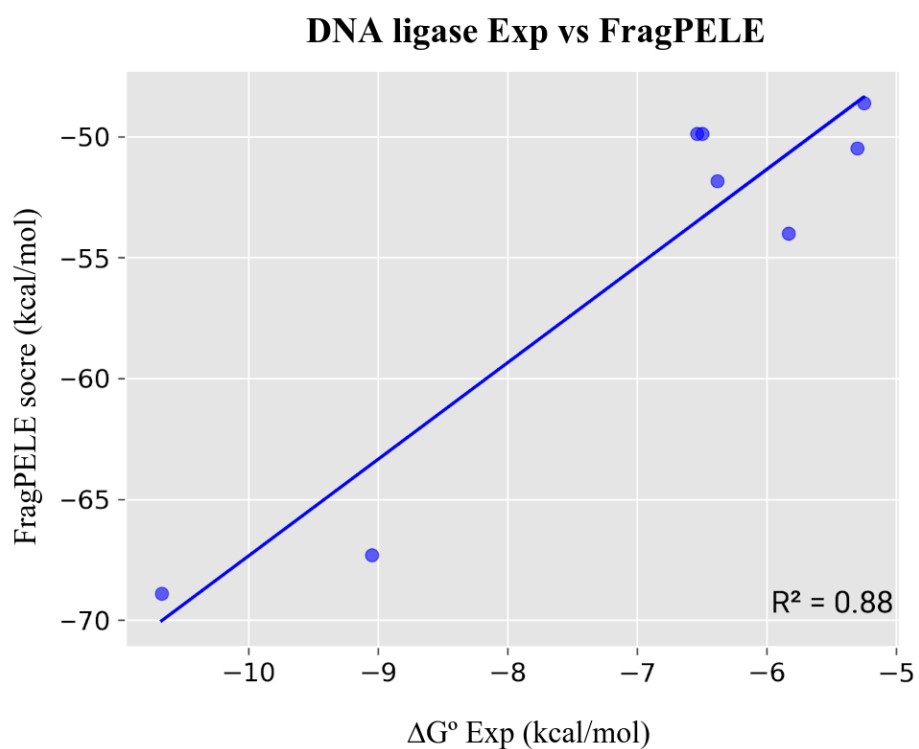


Figure B2. Correlation between experimental ΔG° values and FragPELE results for DNA ligase

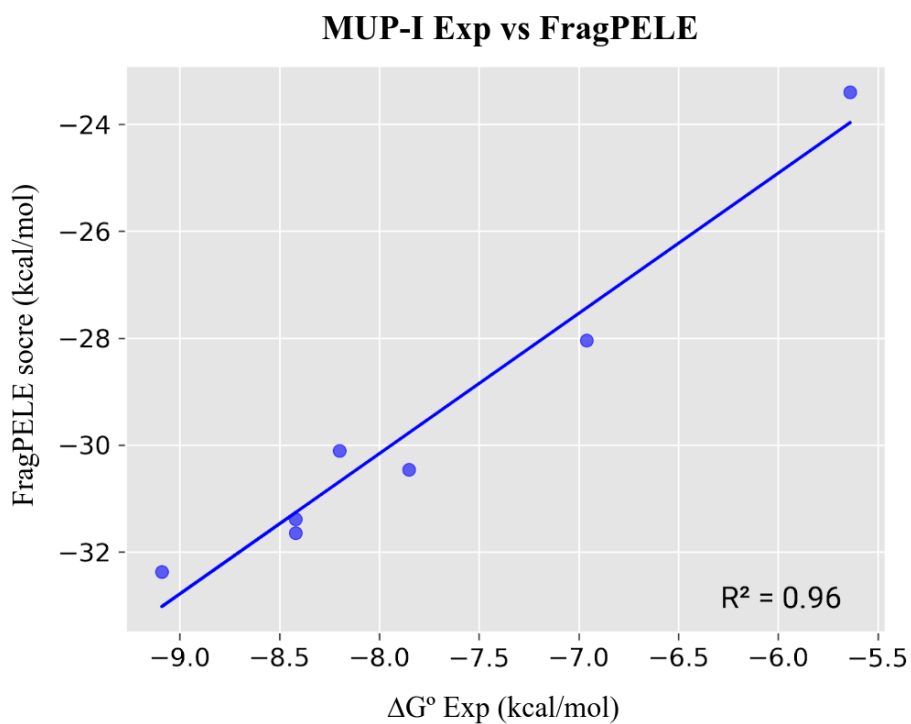


Figure B3. Correlation between experimental ΔG° values and FragPELE results for MUP-I.

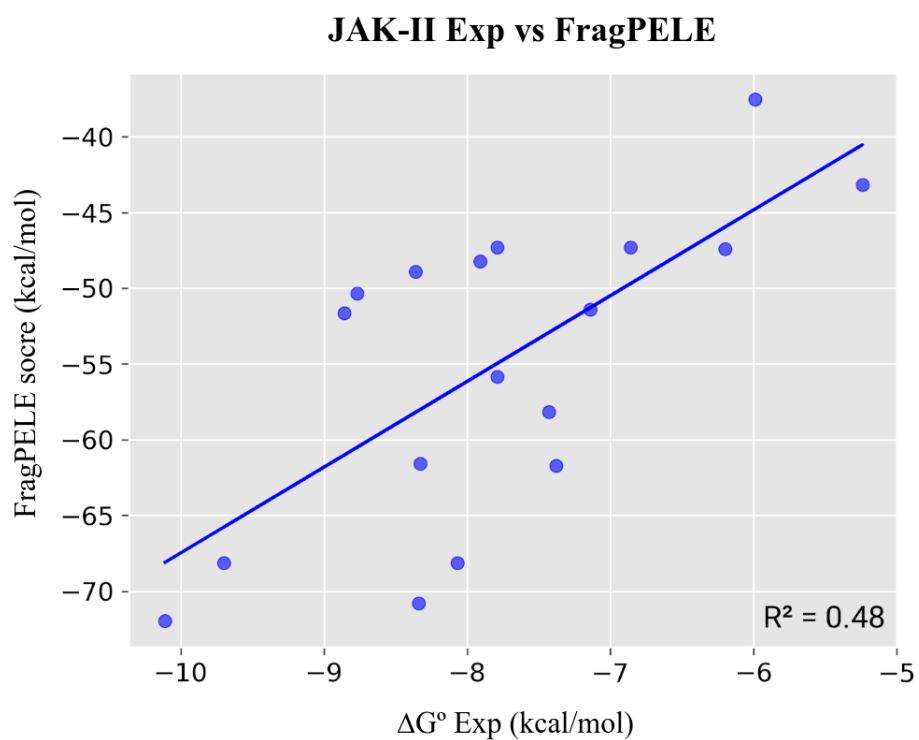


Figure B4. Correlation between experimental ΔG° values and FragPELE results for JAK-II.

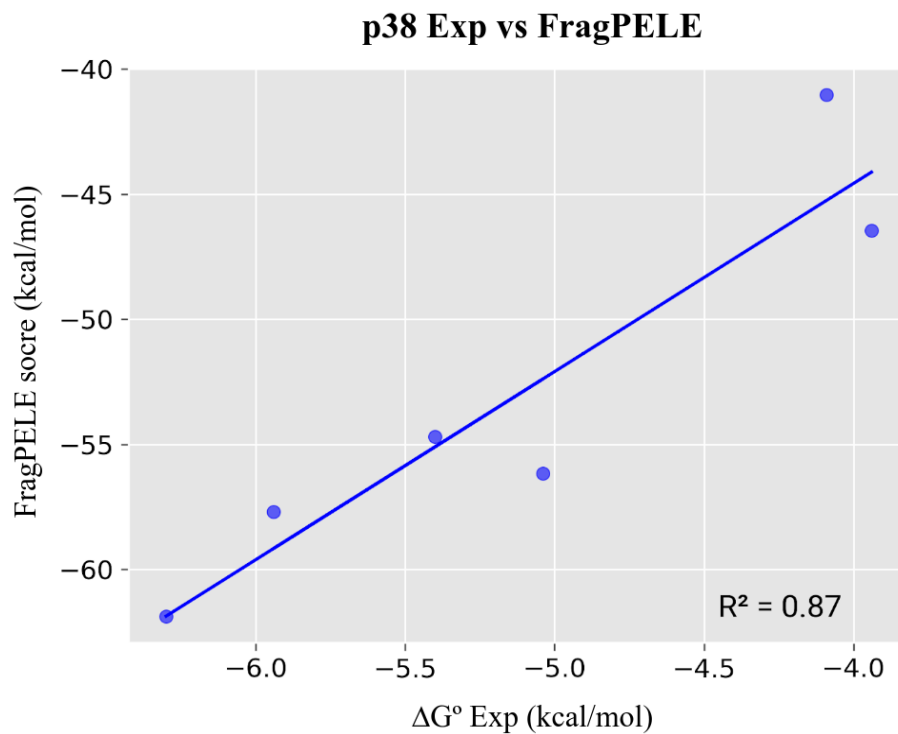


Figure B5. Correlation between experimental ΔG° values and FragPELE results for p38.

Appendix C

Changes in FragPELE protocol after its first published version

As mentioned in *Section 1* from *Chapter 3*, the initial version of FragPELE showed several limitations that have been solved throughout their use in drug discovery projects. Here we will summarize which modifications the algorithm has suffered in their major releases.

FragPELE 2.0

When growing fragments in some systems, users reported clashing simulations because of intramolecular contacts between the scaffold and fragment part. Thus, we propose some changes to tackle this issue:

Inclusion of parameters modification in the fragment growing stage. Initially, the algorithm only changed σ (VDW radii), partial charges, and bonding equilibrium distance. In this version, ϵ (VDW well deep) and solvent parameters (atomic radii to calculate surface and solvent accessible surface areas, alpha and gamma parameters for nonpolar models) are also grown, editing all non-bonding parameters.

Addition of 'allLinear' growing procedure. Instead of distributing the hydrogen charge among fragment atoms in the GS0, this charge is treated following the same equation for all FFP, being X_o the initial value for any FFP and X the final value:

$$X_o = \frac{X}{L+1}$$

Set the final equilibrium distance of the scaffold-fragment linking bond. Instead of expanding while growing the bonding distance of the linking bond, the new algorithm directly places the final distance to move away from the fragment (to avoid non-desired intramolecular contacts). Both previous changes are illustrated in *Figure C1*.

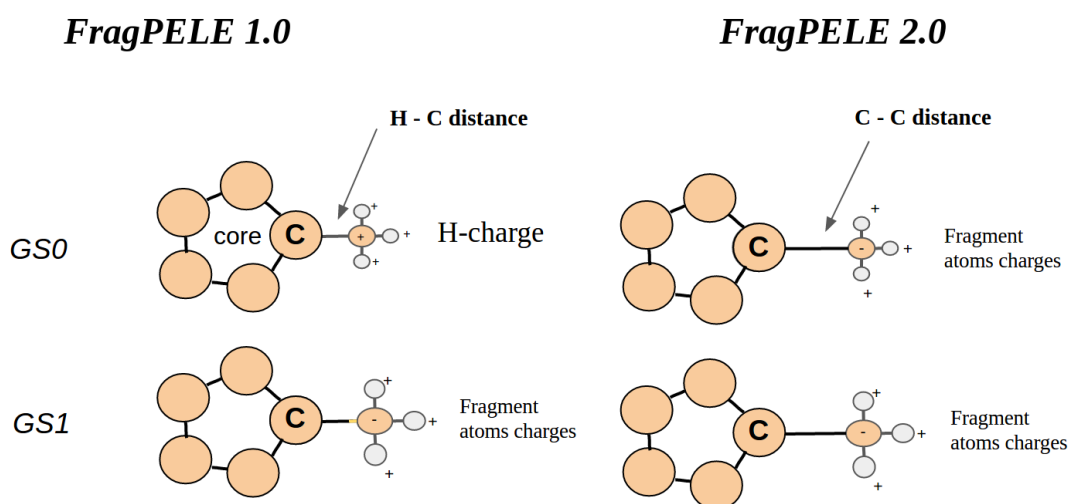


Figure C1. Schematic representation of the FFP treatment in the transition from GS0 to GS1 in FragPELE1.0 and FragPELE2.0. Differences in size between GS0 and GS1 illustrate the growth of bond lengths and VDW radius.

“Grow” the scaffold atoms in synchrony with the fragment. Adding a fragment to the core molecule can change the environment close to these atoms, significantly affecting the distribution of charges. Then, the algorithm reads the FFP of the core atoms in the initial structure (without the fragment) and compares them with the final one (including the fragment). Then, these scaffold parameters will be slightly transformed from the initial to the final state while growing.

By applying these modifications, the number of successful simulations was substantially increased without affecting the performance (re-tested with MUP-I congeneric series showing an R^2 of 0.94) (see *Section 1.3* from *Chapter 3*).

Other new functionalities were developed to make the software more user-friendly:

Growing onto heavy atoms instead of hydrogens. Users can grow fragments onto a heavy atom without requiring a hydrogen atom bound to it. The algorithm automatically replaces the linked atom (element independent) for the whole fragment, setting a new distance extracted from a library of bonding distances (Libretexts, 2015).

Bond-like selection. Following the same trend, users can pick any pair of atoms involved in a covalent bond (excepting cyclic regions) to be replaced for the new

scaffold-fragment linking bond. Consequently, the branch of atoms after the selected bond will be immediately deleted (see *Figure C2*). Notice that this procedure also can be employed to select fragment bonds.

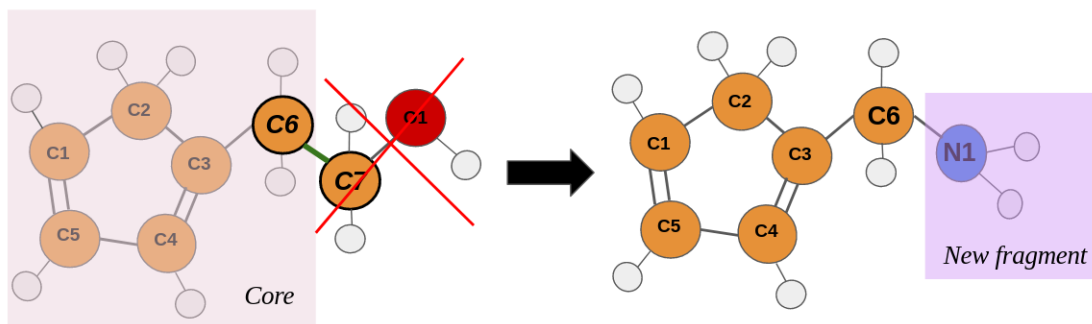


Figure C2. Schematic illustration of bond-like selection in FragPELE 2.0. After the selected bond, the new amino fragment replaces all connected atoms (in green).

Coupling fragments through double-triple bonding. The algorithm detects the selected bond type of the fragment atom, assigning the same type (single, double, or triple) to the newly formed covalent bond. For every extra bond, a free-hydrogen atom bonded to the scaffold linker is erased to not overcome the maximum number of bonds accepted by the atom. If no free-hydrogens are bonded to this atom, the algorithm will raise an exception. Additionally, the new bonding distance will be corrected according to their type for the proper generation of FFP.

Automatic preparation of libraries of fragments with LibPrep. This new software aims to facilitate the processing of molecules to run FragPELE towards a library of fragments. It requires an SDF file containing one or multiple 3D fragments (prepared), and it will automatically prepare all the files to grow them onto the desired scaffold atom. Code available in: https://github.com/carlesperez94/lib_prep

FragPELE 3.0

After extensively using the new version of FragPELE in multiple studies, we found some weaknesses. In this case, a few simulations have shown terminal atoms (normally hydrogens) adopting incorrect dihedrals conformations. Additionally, we detected

mixed enantiomers within the same sampling simulation when expanding fragments with chiral centers. We hypothesize that this situation could be originated due to instabilities produced by the partial charges of the miniaturized fragment in the very early stages of the growing procedure when the van der Waals radii are reduced, and there is not enough repulsion to wrong configurations as well chirality inversions. To address this problem, we proposed a few modifications:

The addition of a 'Softcore-like' growing procedure. Inspired by the *softcore* potentials applied in FEP (Lee et al., 2020), we keep null charges in the first half of the fragment growing phase, adding them in the second half. Moreover, charges are incorporated exponentially while the other terms are increased linearly (see *Figure C3*). Take in mind that the charges of the scaffold atoms will be fixed until modifying the charges of the fragment.

Growing of FFP parameters

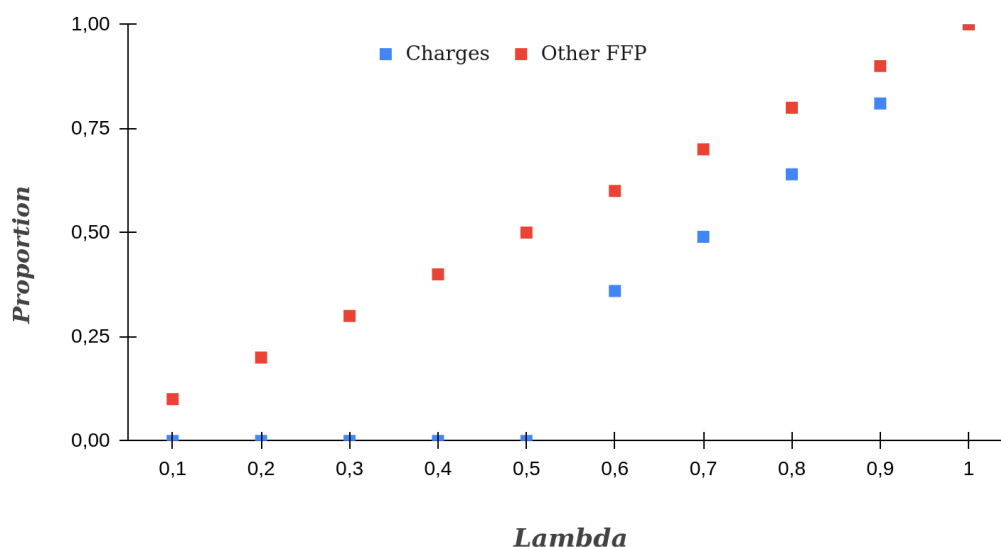


Figure C3. Graphical example to illustrate the FFP parameter changes in the softcore-like growing protocol of FragPELE.

Strengthen minimization in early growing steps. We suspected that the standard minimization performed in each growing step could not readapt the geometry of wrongly positioned atoms. Thus, the convergence criteria of the Truncated Newton minimizer (minimum RSM) of PELE in the first half of the growing process is

strengthened from the default value of 0.25 to 0.01 (kcal/mol/Å)² and then relaxed to 0.1 (kcal/mol/Å)² in the second half.

By combining both previous strategies, wrong dihedrals conformations were corrected. However, a third update was required to solve chirality issues:

Include harmonic constraints on dihedrals involved in chiral centers. Constraints were set in minimizations to keep fixed dihedrals only in the first half of the growing stage.

After applying all these changes, we repeated the MUP-I series to re-assess the predictive score, showing an R² of 0.95, also increasing the time per fragment to around 15 minutes as a consequence of increasing the minimization convergence criteria. However, we think spending more time per step is worth it if we reduce the amount of wrong miniaturized fragment conformations. Besides, extra new functionalities have been incorporated in FragPELE 3.0:

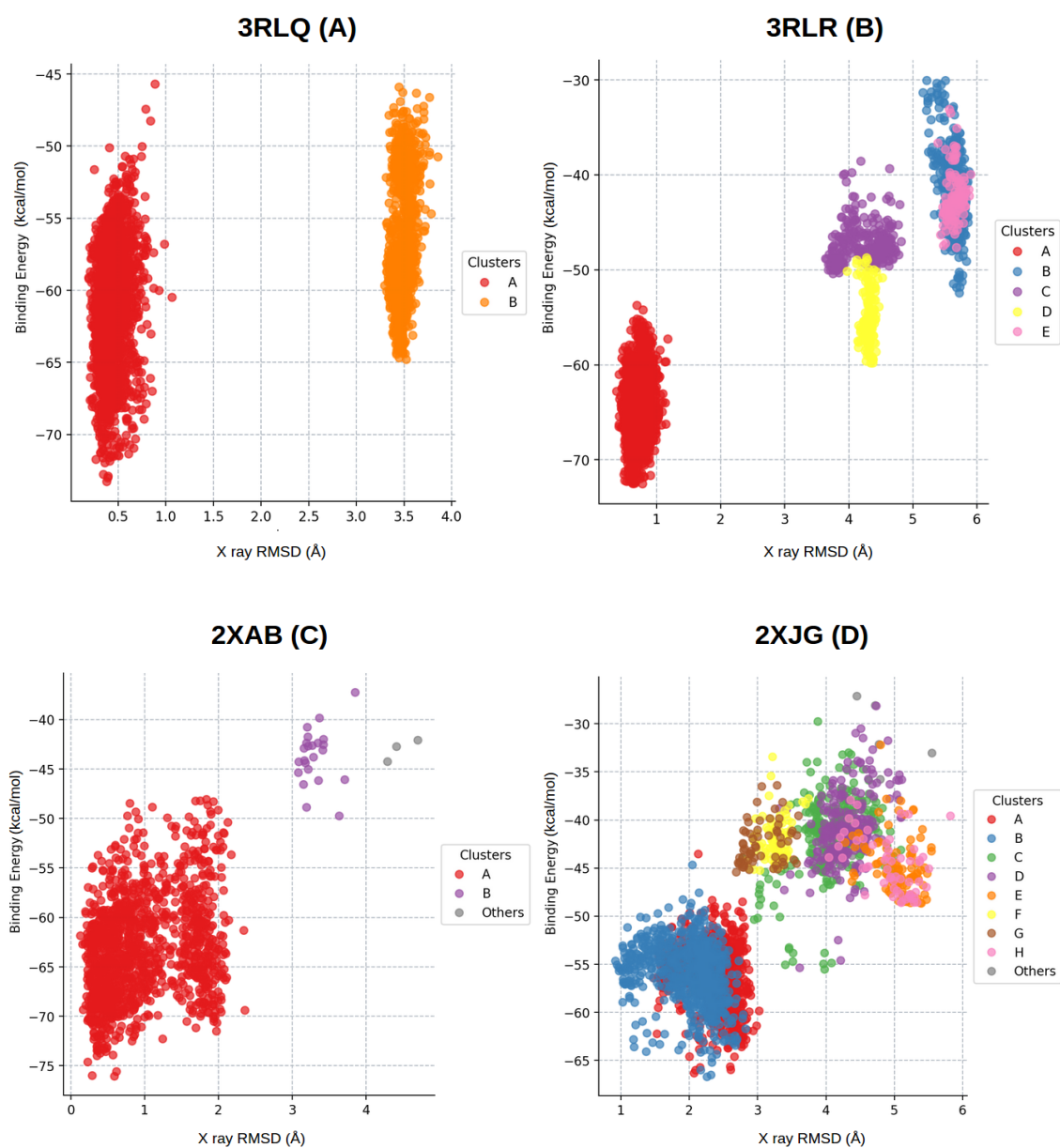
Starting simulations from any (0 to 1) lambda value. This change has been designed to save computation time in systems with open cavities or free space in the growing direction. Users can start simulations from any lambda value between 0 to 1, initializing from the GS closest to the defined value. Consequently, the initial size of the fragment will be proportional to the assigned lambda value.

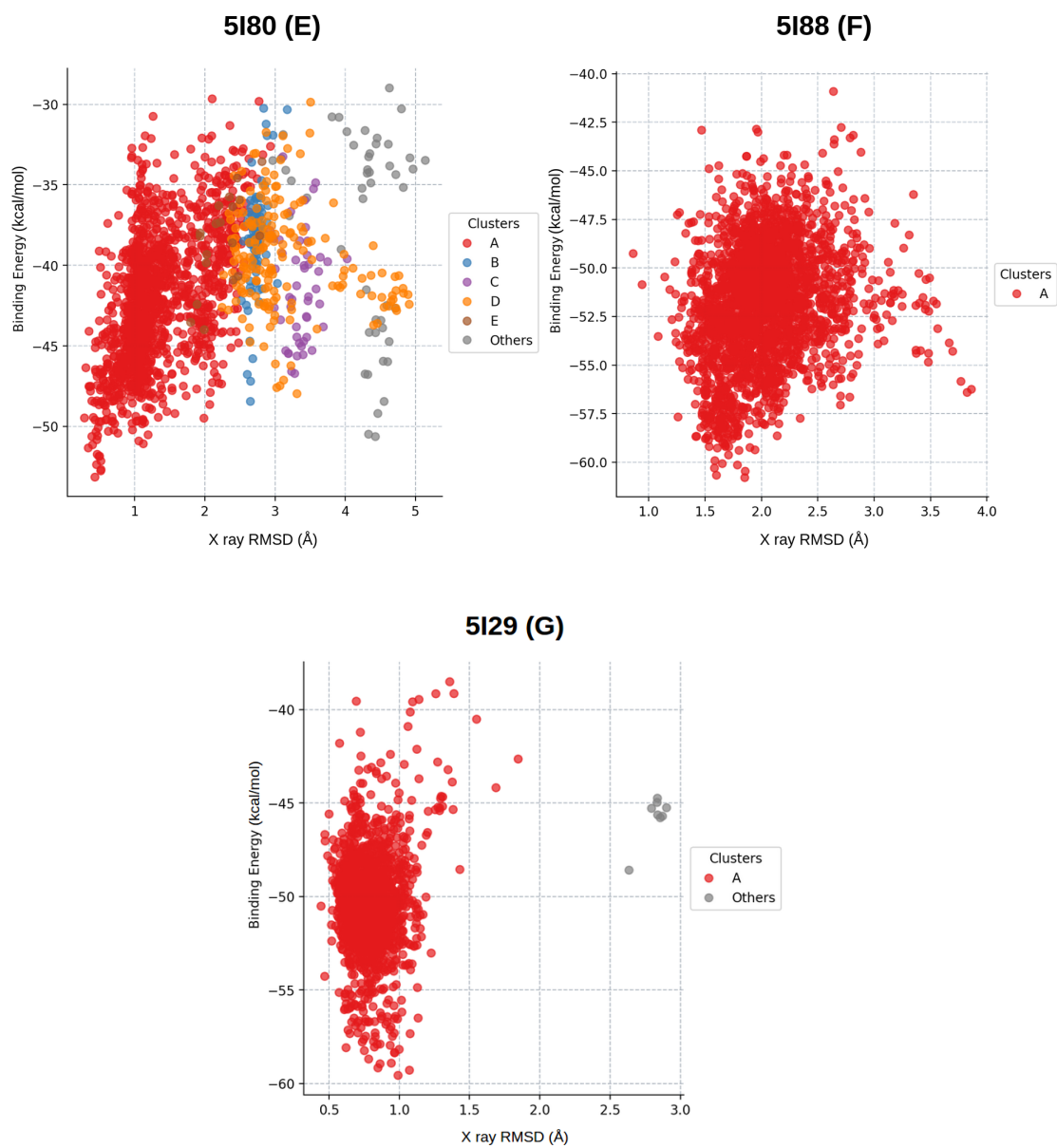
Compatibility with OpenForceField. FragPELE allows using the new Open Force Field initiative (Lim et al., 2020), which was recently included in PELE too.

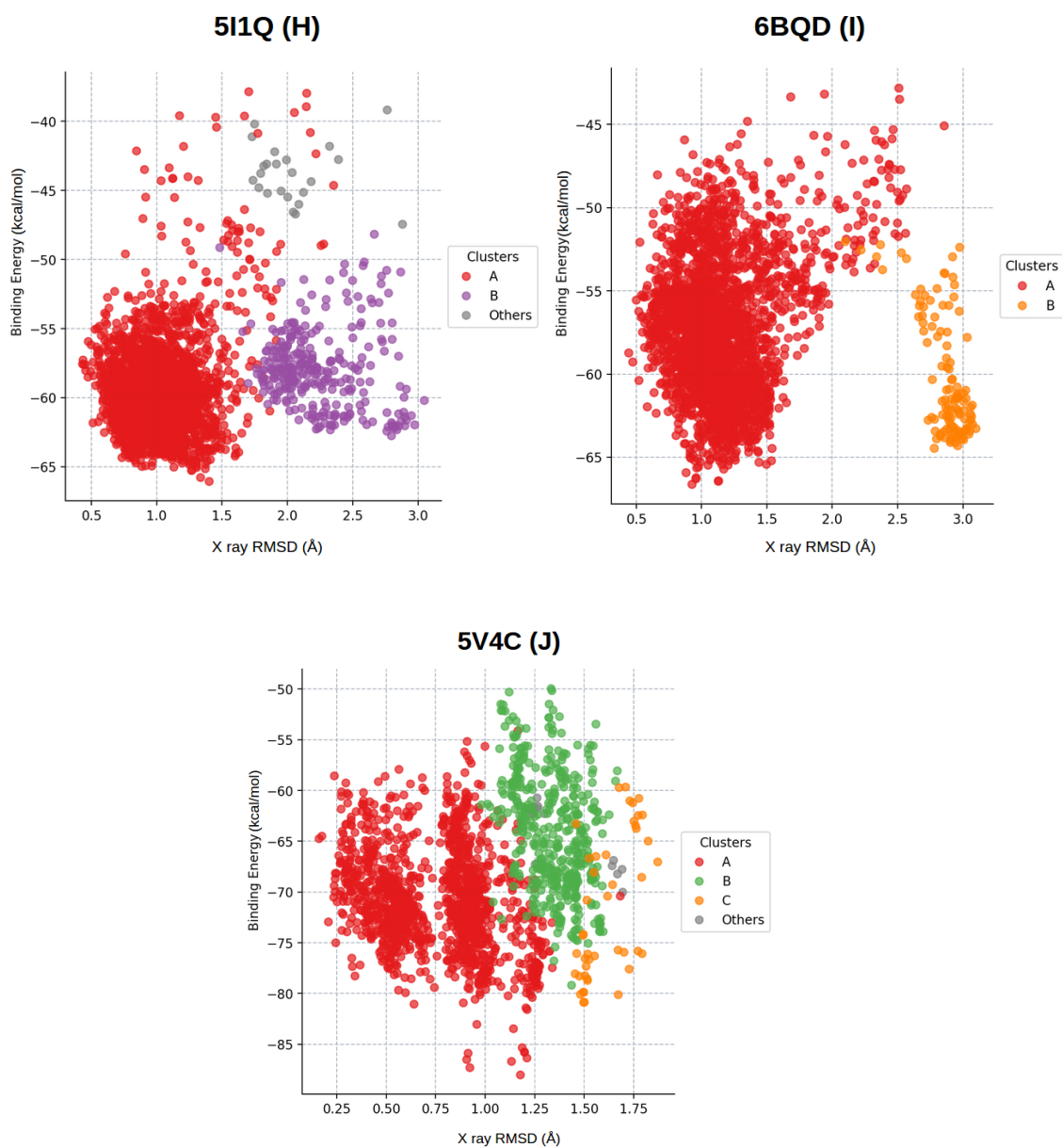
Covalent FragPELE. Fragments can be grown now onto amino acid side-chains or any covalently bound ligand. Users must specify the protein chain and the residue number where they want to attach the fragment. This functionality opens a new door to fastly screening of covalently bound ligands (previously assuming the reactivity of a specific warhead) and even mutate side-chains of amino acids (useful in protein-protein and enzyme engineering studies with PELE).

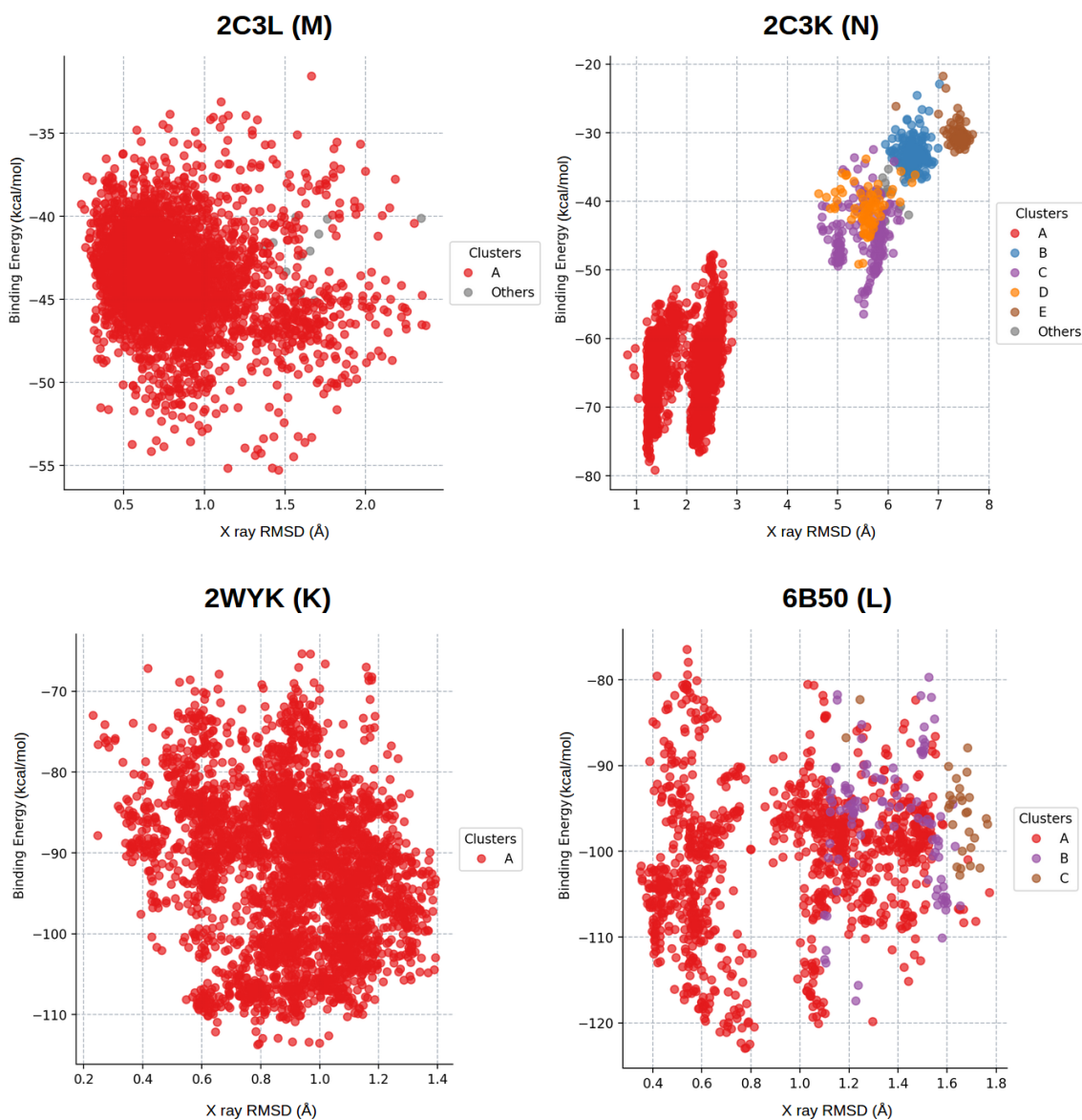
Appendix D

Energy profiles of Growing on hydrated systems study
(combining aquaPELE and FragPELE)









Figures D1. Energy profiles from the structural validation with explicit waters (*Section 2.1 of Chapter 3*). PDB codes and their assigned letters refer to the ones in *Figure 3.13*. Clusters have been colored and sorted alphabetically from the most populated cluster to the least, being the red the most populated cluster.

Appendix E

Supplementary figures of structural validation of AquaPELE and FragPELE algorithm.

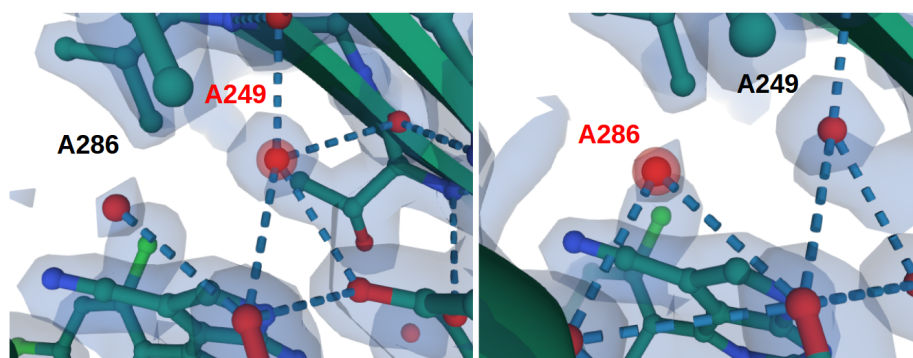


Figure E1. Electron density maps of 3RLQ X-ray. (left-side) Image centered in A249 and (right-side) centered in A286. Blue dashed lines show atom-atom interactions, while blue clouds around atoms represent the density map. Image created with Mol*Viewer (Sehnal et al., 2021).

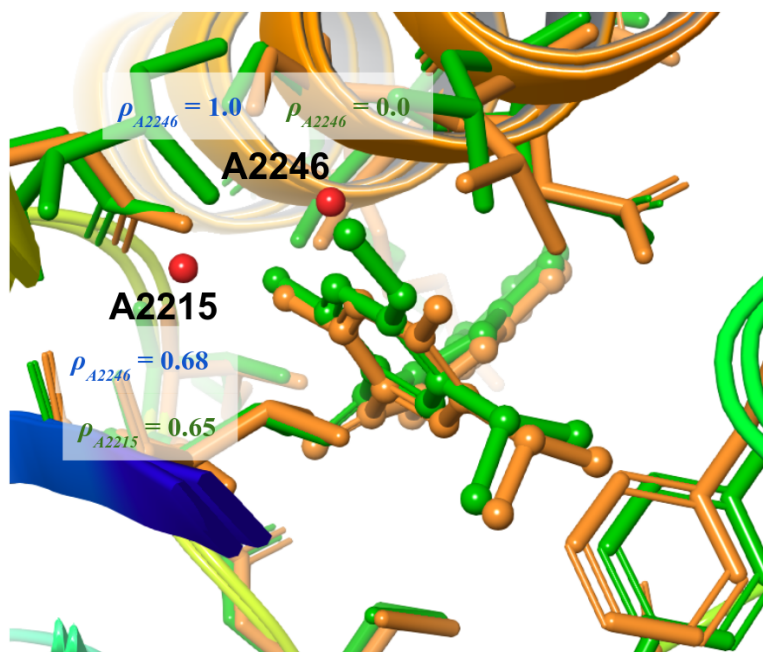


Figure E2. HSP90 (2). Explicit displaceable waters (in red) of the X-ray 2XAB (in orange), superimposed onto 2XJG (in green). A2246 is close to the growth methyl, while A2215 is quite far. Density results for each water cluster are also shown. Created

(**continuation**) with Maestro (Schrödinger, 2018). For each water position, blue labels show the water cluster densities of the aquaPELE simulation, and the green label shows the same densities after combining aquaPELE with FragPELE.

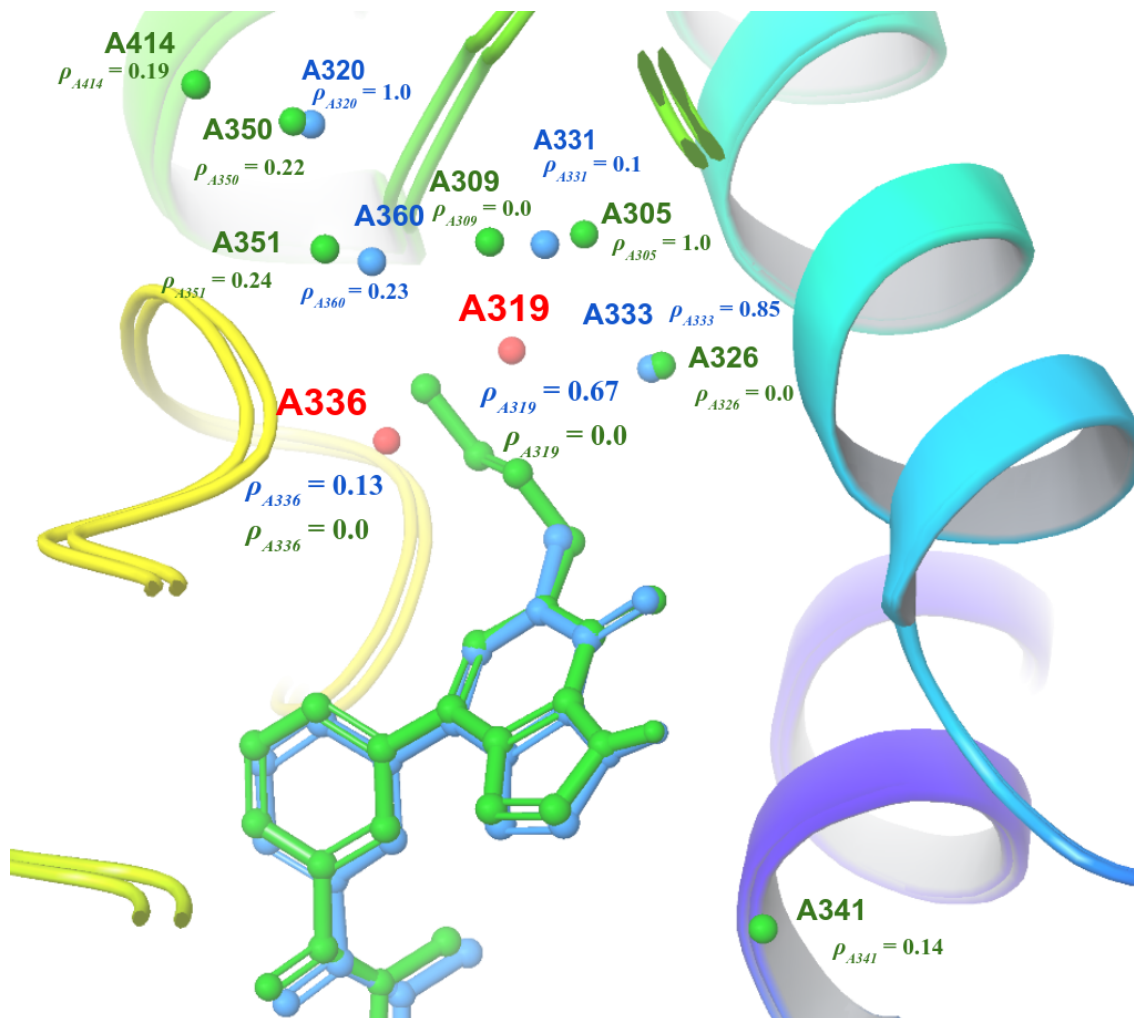


Figure E3. BRD4. Explicit displaceable waters (in red) of the X-ray 5I80 (in blue), superimposed onto 5I88 (in green). Crystallographic water positions are also shown and labeled according to the X-ray information. Notice that there is a buried complex network of water molecules close to the growing direction. For each water position, blue labels show the water cluster densities of the aquaPELE simulation, and the green label shows the same densities after combining aquaPELE with FragPELE. Created with Maestro (Schrödinger, 2018).

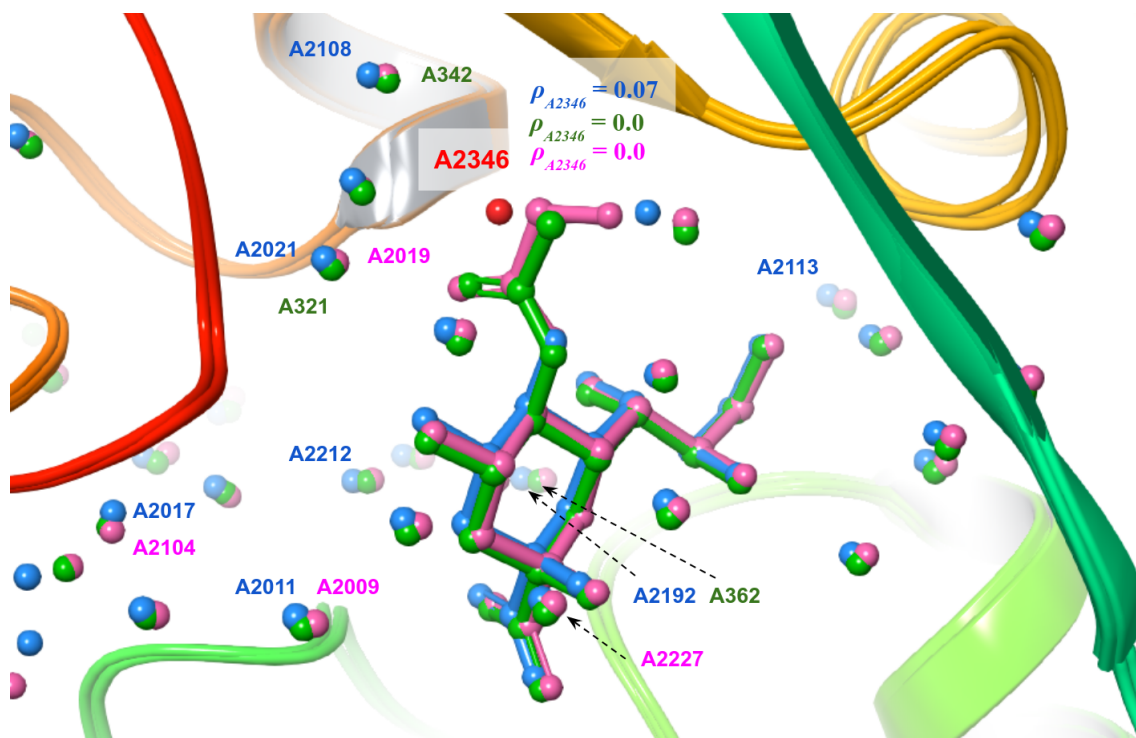


Figure E5. SiaP WT. Explicit displaceable waters (in red) of the X-ray 2V4C (in blue), superimposed onto 3B50 (in green) and 2WYK (in pink). Crystallographic water positions are also shown. Due to many water molecules, only the molecules detected in the clustering analysis and the perturbable water density were labeled. Created with Maestro (Schrödinger, 2018). For each water position, blue labels show the water clusters of the aquaPELE simulation, and the green labels show the same densities after combining aquaPELE with FragPELE to grow the first fragment (3B50), and pink for the second fragment (2WYK).

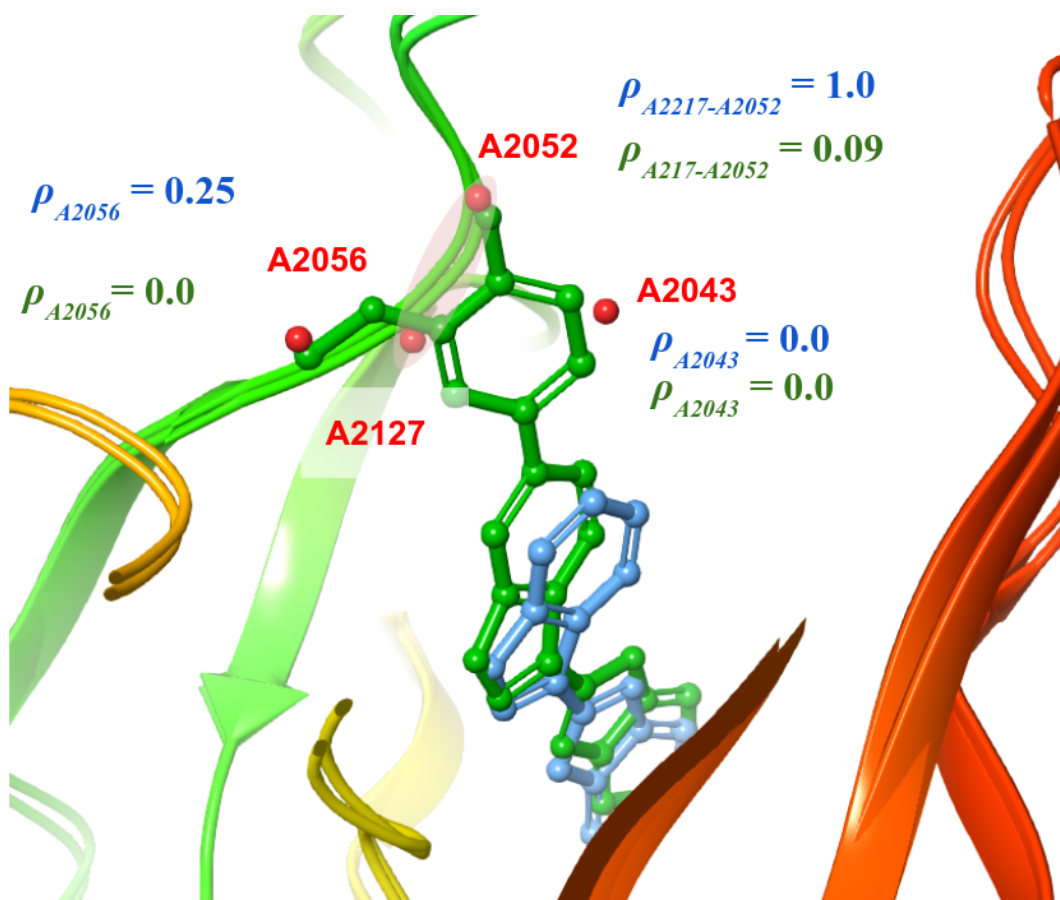


Figure E6. CHK1. Explicit displaceable waters (in red) of the X-ray 2C3L (in blue), superimposed onto 2C3K (in green). Notice that the fragment size is the largest one of the whole set, implying the displacement of 4 water molecules. For each water position, blue labels show the water cluster densities of the aquaPELE simulation, and the green labels show the same densities after combining aquaPELE with FragPELE. Only two major clusters were detected, one of them between A2052 and A2127. Created with Maestro (Schrödinger, 2018).

Appendix F

Supplementary figures of the VS pipeline based on ML methods

Table F1. Mean F1-scores, standard deviation (SD), and test results for the 15 fold-CV in the Xray 1 machine learning model. Note: the test was only computed when the mean F1-score was higher than 0.60 (green colored).

Xray 1	F1 score	standard deviation	standard error	Test
5 features	0,562	0,127	0,0327	nan
4 features	0,608	0,130	0,0335	0,68
3 features	0,478	0,113	0,0292	nan
2 features	0,637	0,122	0,0314	0,65
1 feature (docking)	0,606	0,120	0,031	0,46
1 feature (MET resilience)	0,55	0,120	0,0309	nan

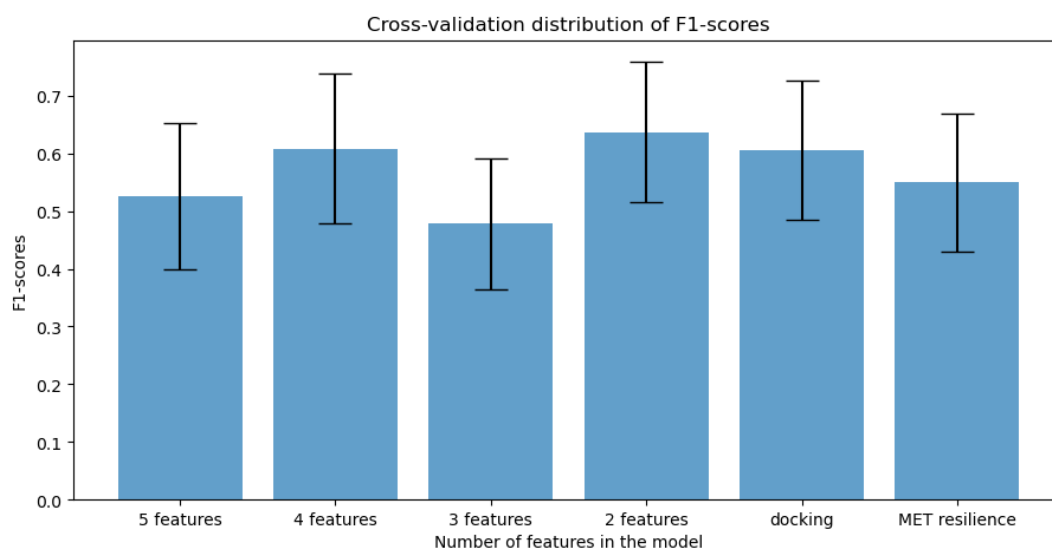


Figure F1. Distribution of the mean F1-scores \pm standard deviation for 15-folds cross-validation results in the Xray 1 machine learning model.

Table F2. Mean F1-scores, standard deviation (SD), and test results for the 15 fold-CV in the Xray 2 machine learning model. Note: test was only computed when the mean F1-score was higher than 0.60 (green colored).

Xray 2	F1 score	standard deviation	standard error	Test
5 features	0,645	0,108	0,0278	0,63
4 features	0,636	0,121	0,0312	0,67
3 features	0,641	0,110	0,0284	0,64
2 features	0,581	0,174	0,0448	nan
1 feature (docking)	0,567	0,143	0,0369	nan
1 feature (MET resilience)	0,495	0,130	0,0335	nan

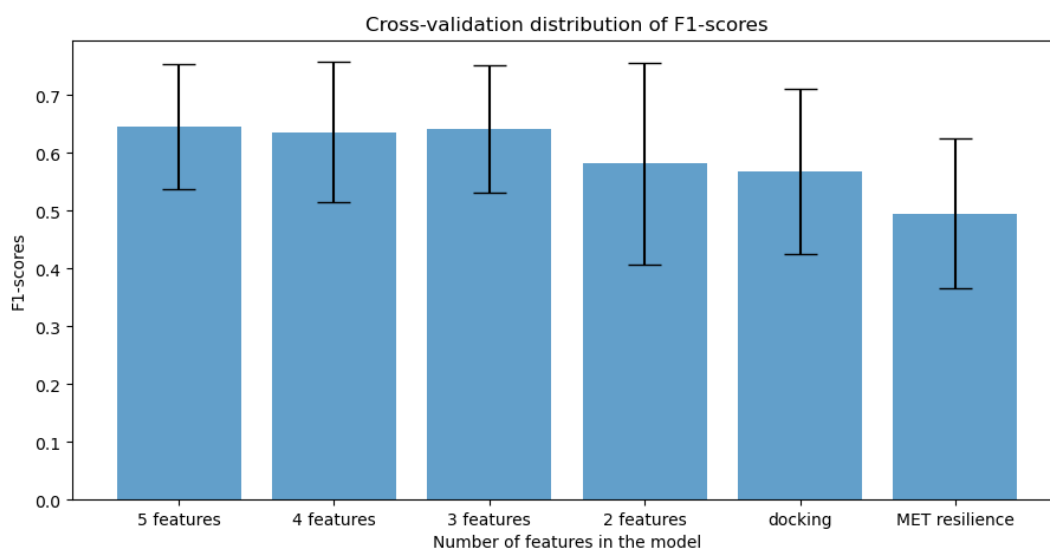


Figure F2. Distribution of the mean F1-scores \pm standard deviation for 15-folds cross-validation results in the Xray 2 machine learning model.

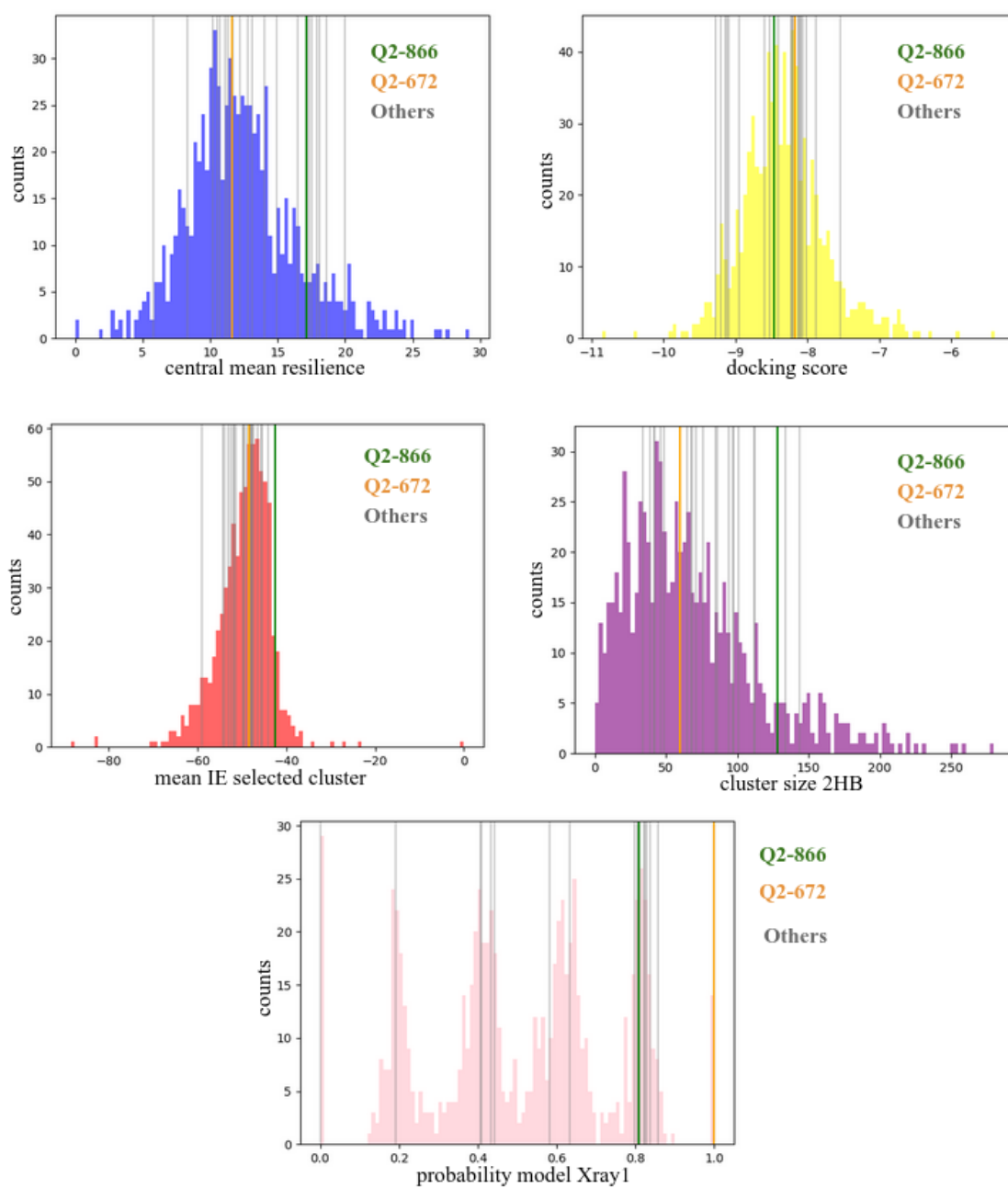


Figure F3. Individual features and probability distribution for commercial set results in Xray 1. Vertical lines show the position of the 23 compounds experimentally tested. Green and orange lines correspond to active compounds and gray to inactive.