ESCI upf.
School of International Studies

MEMÒRIA DEL TREBALL DE FI DE GRAU DEL GRAU (ESCI-UPF)

ALCHEMICAL-PELE:
alchemical simulations tool for the calculation of relative binding energies between ligands

AUTOR/A: Miguel Borge	**NIA: 106403**
GRAU Bioinformatics	
CURS ACADÈMIC: Tercero	
DATA: 21/06/2023	
TUTOR/S: Martí Municoy	

ESCI upf. School of International Studies

Bachelor's Degree in Bioinformatics (UPF-UPC-UB)

Final Grade Project

ALCHEMICAL-PELE: alchemical simulations tool for the calculation of relative binding energies between ligands

Miguel Borge

Scientific director: Martí Municoy[1]

[1]IT Department, Address: Av. de Josep Tarradellas, 8-10, 3-2, 08029 Barcelona

Abstract

Motivation: Drug discovery involves the intricate task of identifying and optimizing therapeutically active compounds, a process expedited by the advent of computational methodologies. Despite advancements, achieving precise binding free energy calculations remains challenging due to the expansive degrees of freedom inherent in protein-ligand complexes. To overcome these challenges, Alchemical-PELE employs Monte Carlo (MC) simulations for alchemical transformations in protein-ligand complexes. This enables the effective exploration of conformational space and accelerates the surpassing of energetic barriers. Moreover, the software calculates relative binding free energies (RBFE) between ligands, offering an efficient, precise computational framework for protein-ligand thermodynamics. This capability makes Alchemical-PELE a potent tool for identifying potential drug candidates.

Results: The computational workflow was applied to two protein systems, generating a comprehensive ranking of ligands based on calculated relative binding free energies. In the MCL1 system, the results were particularly encouraging, exhibiting a promising correlation between our computed values and the existing experimental data. However, the exploration of the Tyk2 system presented some challenges, specifically highlighting the cumulative impact of errors during the ligand ranking process. These findings elucidate the need for future enhancements to our methodology, prompting a continued pursuit of precision in this computational endeavor.

1 Introduction

The development of new therapeutic agents is a crucial aspect of modern medicine, yet the process of drug discovery is notoriously challenging and resource-intensive. Central to this endeavor is the identification and optimization of chemical compounds with potential therapeutic activity (Paul et al., 2010). In recent years, computational methods have emerged as an indispensable component of drug discovery, allowing researchers to rapidly explore vast chemical spaces and pinpoint promising drug candidates (Schneider, G. & Fechner, U., 2005). The growing interest in computational methods for drug discovery has led to the development of various techniques for estimating protein-ligand binding affinities, such as molecular docking (Kitchen D.B. et al., 2004), molecular dynamics simulations (Karplus, M., &

Kuriyan, J., 2005), and free energy perturbation (FEP) methods (Zwanzig, R. W., 1954; Kollman, P.A. 1993). These methods have advanced our understanding of the thermodynamics and kinetics of protein-ligand interactions and contributed to the identification of potential drug candidates. In particular, FEP methods (alchemical simulations) have gained popularity because of their ability to accurately estimate free energy changes and have become an essential part of the toolbox of computer-aided drug design (Chodera, JD. et al, 2011). These approaches rely on a series of alchemical transformations between different ligand states, allowing for the exploration of various potential binding modes and affinities (Wang L. et al., 2015). Recent studies have showcased the success of alchemical simulations in identifying promising drug candidates, such as the optimization of inhibitors for BACE1 (M. Ciordia et al., 2016) and the evaluation of variants of the m396 antibody in terms of conformational stability and binding affinity to SARS-CoV-1 and SARS-CoV-2 spike proteins (Zhu, F. et al., 2022).

These simulations can be broadly categorized into different types, such as dual-topology and single-topology approaches (Fig 1), based on their underlying theoretical framework. The single topology approach involves having only one site for any location shared between the end states, and then using “dummy” atoms (non-interacting atoms used to facilitate transformations) to account for any unique sites. As the transformation occurs, these dummy atoms are transformed into fully interacting atoms, while the shared site atom is directly transformed into a new atom type. On the other hand, the dual topology approach is characterized by the fact that shared sites between states do not share atoms. In this approach, no atom changes its type; instead, the interactions with the surrounding system are altered. The dual topology requires more dummy atoms, which in turn necessitates more CPU power and additional intermediates. However, it has a significant advantage in that these dummy atoms can simultaneously explore more conformational space while being decoupled. In either topology approach, it may be necessary to modify bonded terms, particularly for angle and dihedral terms. Due to the time scale of these motions, these terms converge quickly, resulting in small variances but large energy changes. For simple bonded parameters, such as harmonic spring constants and equilibrium bond lengths, linear changes are perfectly acceptable. Computing constrained bonds can be challenging since correction terms are needed due to the complete lack of phase space overlap. Each method has its own advantages and drawbacks, with the selection ultimately being dictated by the specific requirements and constraints of the simulation in question (Ou-Yang SS, et. al, 2012). In Figure 1, we can observe how a hybrid molecule can represent both a methanol (left) and an ethane (right) molecule simultaneously by using dummy atoms (D) and the different approaches, single at the top and dual at the bottom.

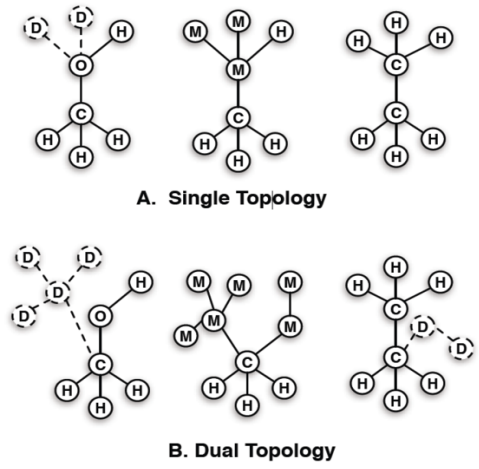

Figure. 1. Single and dual hybrid topology approaches of methanol and ethane. Extracted from Alchemistry.org

This approach employs Monte Carlo (MC) simulations to study alchemical transformations in protein-ligand complexes and determine the RBFE between ligands. MC simulations, focused on randomly sampling molecular configurations to investigate conformational space and calculate binding energies, allow for the analysis of interactions and binding affinities between proteins and ligands in various contexts, including biological

systems. In contrast, molecular dynamics simulations examine the motion and temporal evolution of molecules, providing details about molecular trajectories and fluctuations. Both approaches offer complementary perspectives in the study of protein-ligand interactions, with MC simulations being ideal for exploring conformational space with higher freedom and estimating binding energies, while molecular dynamics provide insights into the behavior and evolution of the involved molecules (Karplus, M. & McCammon, J. 2002).

Introducing the AlchemicalPELE software suite, the foundation of this study, is a necessity as we transition from theoretical concepts to practical applications. The central philosophy of AlchemicalPELE revolves around a quick MC method that expertly balances accuracy and computational cost. The software achieves this balance by implementing MC simulations for efficient exploration of the conformational space in protein-ligand complexes. This stochastic process effectively overcomes energetic barriers and explores a broader conformational space, providing an ensemble of diverse and representative system states. Alongside MC simulations, AlchemicalPELE carries out alchemical transformations to calculate relative binding free energies (RBFE) between ligands. Using the Bennett Acceptance Ratio (BAR) method, the software ensures precise calculation of RBFE values, contributing to the overall accuracy of the system. By synergizing the efficiency of MC simulations and the precision of alchemical transformations, AlchemicalPELE provides a cost-effective solution to the challenges faced in computational chemistry. As a result, it stands as a valuable tool in the realm of computational drug discovery, promoting the efficient identification and optimization of potential therapeutic compounds.

1.1 Objectives

The Alchemical-PELE tool is designed to be a user-friendly platform for fast relative binding energy calculations between a target and a ligand library, aiding in the identification of potential pharmaceuticals, automating the entire process of performing the different necessary alchemical simulations. This tool incorporates state-of-the-art computational methods and leverages MC simulations for alchemical estimations, effectively streamlining the drug discovery process. By integrating these advanced techniques, Alchemical-PELE contributes an innovative approach in studying protein-ligand interactions and accurately predicting protein-ligand affinities, greatly aiding in the quicker identification and optimization of novel drug candidates

2 Methods

This research introduces a novel workflow to perform alchemical simulations for relative binding free energy (RBFE) calculations within the framework of the PELE simulation software and the NBD Suite. This workflow processes a protein structure and a ligand library in standard formats, while leveraging the power of a computing cluster to run efficiently. The workflow execution sequence is controlled through an input.yaml file, which guides the various functional blocks in a simplified and easy-to-use manner. These functional blocks are a compilation from various sources: some already existed in the NBD suite, others are implementations of open-source tools, and finally, there are blocks that have been specifically developed for our tool.The following sections will delve into the details of each functional block, paying special attention to those blocks that have been specifically developed for this tool.

2.1 Topology extractor

The Topology Extractor serves as the first functional block in our workflow. Its core role is to convert protein structures and ligand libraries from any standard computational chemistry format into the Protein Data Bank (PDB) format. The PDB format is chosen because it is the preferred format for simulation in the PELE software. This conversion ensures that all subsequent processes function

smoothly, thereby enhancing the overall efficiency of our workflow.

2.2 PDB preprocessor

The next step in our workflow is the PDB Preprocessor. Its primary role is to inspect and refine the PDB files produced by the Topology Extractor, ensuring their suitability for simulation in the PELE software. It accomplishes this by scanning each PDB file to identify and correct common errors, including misplaced atoms, incorrect bond definitions, and missing entries. This meticulous preparation guarantees that the PDB files are optimally formatted for the next steps. By ensuring the integrity and accuracy of the data at this stage, the PDB Preprocessor significantly enhances the reliability of PELE simulations, serving as a cornerstone in the overall effectiveness of our workflow.

2.3 Docking ligands

The next functional block in our methodology incorporates the use of rDock, a prominent open-source program designed for molecular docking and scoring. Within our workflow, rDock's principal function is to execute iterative simulations involving the protein and the library of ligands, thus accurately elucidating potential ligand binding sites within the protein's structure. During each iteration, rDock situates the ligand within the protein's binding cavity, subsequently simulating interactions to determine the ligand's optimal conformation and orientation. This iterative mechanism facilitates the identification of the most energetically favorable ligand binding positions, a crucial aspect in comprehending protein-ligand interactions.

It is important to note that if the position of a reference ligand is already established relative to the utilized ligand library, this rDock stage could be bypassed, hence streamlining the overall workflow without affecting the integrity of the resulting data. In scenarios where the reference ligand's position remains unknown, this rDock phase remains invaluable. It produces detailed PDB files containing precise coordinates of optimal ligand positions within the protein, thus acting as a robust foundation for subsequent steps within our workflow. This stage significantly contributes to the overall robustness and effectiveness of our methodological approach.

2.4 Ligand pairing

Ligand pairing is an essential step in lead optimization, and it can be facilitated by the use of the open source tool called Lead Optimization Mapper (LOMAP) (Liu S. et. al. 2013). LOMAP plays a pivotal role in addressing the computational challenge associated with a large library of compounds, where the potential pairwise transformations could result in a significant number of required simulations. For example, a library of 100 ligands might require approximately 5000 simulations.

LOMAP automates the scheduling of free energy calculations and optimizes the selection process for the most suitable simulations to study the various compounds. To achieve this, LOMAP generates a molecular graph that represents the relationships between the molecules. Each compound is represented by a node in the molecular graph, and the relative binding free energy (RBFE) calculations between two adjacent compounds are denoted by edges. The selection of edges is based on a similarity score, which assesses the feasibility of each calculation. This score takes into account the size of the maximum common substructure (MCS) shared by two molecules and gives preference to matches that conserve ring systems, while treating all heavy atoms equivalently.

By assembling this graph of planned RBFE calculations, LOMAP reduces the number of edges while ensuring that each node participates cyclically. The result is a connected graph with minimized connections, streamlining the entire lead optimization process. This graph-based approach enhances efficiency, minimizes errors, and helps identify potentially erroneous calculations caused by poor convergence. Overall, LOMAP significantly

contributes to the robustness and effectiveness of the lead optimization methodology.

Alchemical-PELE, a computational framework for investigating protein-ligand interactions and performing relative free energy calculations, incorporates advancements in the field. It integrates various techniques and builds upon the use of LOMAP to optimize the statistical architecture of simulation experiments. Alchemical-PELE utilizes a reference protein crystal structure with an associated ligand and a separate library of ligands. Using the LOMAP method, Alchemical-PELE determines the simulations (transformations) that should be conducted between the ligands in the library. The LOMAP python implementation efficiently generates a graph that represents the chemical space of compounds and their feasible transformations. This graph-based approach assists researchers in prioritizing promising simulations for lead compound optimization, reducing the time and resources required to evaluate the potential of new drug candidates.

2.5 PELE alchemizer

Once the calculations are planned, the next step is to obtain the hybrid topology and the necessary files to run the simulations in PELE. To do this the following methods have been implemented in the Peleffy tool.

2.5.1 Generate hybrid topology

It first uses a topology construction method based on a single-topology and explicit MCS approach. This method is selected instead of dual topology because it can provide several advantages, including increased computational efficiency resulting from the use of fewer dummy atoms, greater simplicity in setting up the system and implementing the transformation process, enhanced compatibility with certain simulation software or codes that may favor single topology, the ability to maintain continuity of atom identity throughout transformations, and the possibility of achieving smoother transitions between states, facilitating a more seamless exploration of the conformational landscape (Truhlar, D.G. et. al, 2008).

Hybrid single-topology alchemical simulations have demonstrated success in various studies, such as the optimization of inhibitors targeting the influenza neuraminidase protein (Williams-Noonan et al., 2021). To perform these simulations in Alchemical-PELE, the hybrid single-topology structures must first be generated. This is achieved using the Peleffy (PELE Force Field Yielder) tool, which processes the input ligands and protein and generates the templates and hybrid topologies required for running the simulation in PELE, the software responsible for executing the simulation (Borrelli, K.W. et al., 2005) and the hybrid structure.

To obtain these files, Peleffy generates all the templates to be able to use the FEP method, which involves gradually transforming the ligand from its initial state to its final state using a scaling factor known as lambda. This approach enables the calculation of the relative binding free energies (RBFE) along the transformation path and provides insights into the thermodynamic properties of the ligand-protein complex. This gradual alchemical transformation with lambda is achieved through modifying the parameters at different epochs of the simulation, changing the various properties of the atoms of the ligand molecule so that in the initial and final states, the hybrid structure is characterized with the same properties as the two compounds whose RBFE is being studied (Boresch S. et al., 2003; Hansen, N., & Van Gunsteren, W.F. , 20 14; Mey, ASJS et. al, 2020).

For example, in the case of the transformation between 1,4-dichlorobenzene, A, (Fig 2. Left) and phenol, B, (Fig 2. Right), the common substructure, M, is the benzene ring (Fig 2. Middle). Both 1,4-dichlorobenzene and phenol have a benzene ring with different substituents. We set up two transformations: A → M and B → M, based on a deletion-only mode. 1,4-dichlorobenzene has two chlorine atoms on the benzene ring, while phenol has a hydroxyl group. During the transformation, we

replace one chlorine atom with a hydroxyl group and the other chlorine atom with a hydrogen atom. Initially, the additional hydrogen atom of the hydroxyl group is represented as a dummy atom in the hybrid topology. As the transformation progresses, the dummy atom gradually acquires the properties and interactions of the real atoms they are meant to represent. The chlorine atom being replaced by the hydroxyl group gradually transforms into an oxygen atom, and the additional hydrogen atom appears as part of the hydroxyl group. Similarly, the chlorine atom being replaced by a hydrogen atom transforms into a hydrogen atom. At the end of the transformation, the dummy atoms have fully converted into the real atoms of the hydroxyl group and hydrogen atom in phenol. The output generated is a molecule with the hybrid topology which can then be used to replace the ligand in the original input .pdb file, generating a file with the protein complex and the hybrid structure necessary for each simulation.

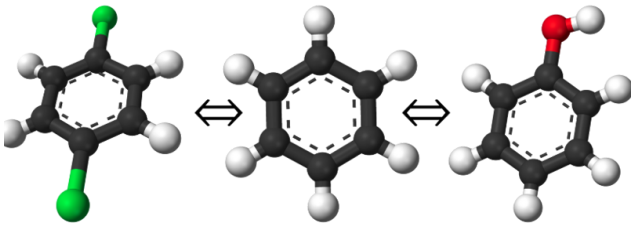

Figure. 2. From left to right: A the 1,4-dichlorobenzene; M, the MCS is a benzene ring; B, the phenol.

2.5.2 Constructing a path of intermediate states (lambdas) and PELE input files

Alchemical simulations involve the careful study of the relative binding energy between two compounds, with phase space overlap emerging as a critical factor. Phase space, a multidimensional space encompassing all possible configurations of a system considering both positions and momenta, directly affects the ease of transition between two states. Insufficient phase space overlap often leads to significant errors in free energy calculations due to the distinct configurations of the two states. Addressing this issue involves devising a thermodynamic path with improved phase space overlap between states. The definition of this path can be visualized as setting the alchemical path where forces on an atom or a group of atoms are modified, removed, or added. Such a path may incorporate unphysical, intermediate states that do not necessarily need to be experimentally observable or have chemical sense. The integration of more intermediate states into the pathway increases resemblance between the states, consequently increasing the likelihood of sufficient phase space overlap. This process, however, also increases the computational cost as more intermediate states are included. The challenge of selecting the "correct" or "good" path notwithstanding, constructing thermodynamic paths with enhanced phase space overlap aids in reducing errors in free energy calculations, thereby enabling more accurate relative binding energy calculations between two compounds.

The construction of a path of intermediate states, also known as lambdas, and the preparation of PELE input files is a process involving a high degree of flexibility, serving the unique requirements of the system under examination. Different protocols can be chosen in the input.yaml file, allowing the adjustment of lambdas as required, along with the fine-tuning of parameters like Coulomb charges (from molecule 1 and molecule 2), van der Waals (VDW) parameters, and bonded interactions.

The Linear Alchemical Potential formula (Eq. 1) provides a standard method to adjust these parameters during the transformation. It ensures a smooth transition of these parameters by considering the sum of the chemically modified potentials of the two end states, U1 and U2, in addition to the portions of potential unaffected by the alchemical transformation.

$$U(\lambda, \vec{q}) = (1 - \lambda)U_0(\vec{q}) + \lambda U_1(\vec{q}) + U_{\text{unaffected}}(\vec{q})$$

Equation 1. Linear Alchemical Potential formula

The sequence of modifications holds importance for maintaining system stability. It usually involves the detachment of the Coulombic parameters of molecule 1's exclusive atoms first, followed by the van der Waals parameters. Then the bonded parameters, exclusive van der Waals, and Coulombic parameters for molecule 2 are applied. This method, referred to as the soft-core potential, prevents charges and L-J interactions from being simultaneously switched off, thus maintaining system stability during the transformation process (Pitera, J.W., & van Gunsteren, W.F., 2002; Steinbrecher T. et al., 2007).

Such a protocol provides ample flexibility in designing the simulation and offers a means to align the process closely with the specific needs of the system under investigation. The ability to choose the protocol, adjust the lambdas, and change the parameters based on the individual requirements of the transformation makes this approach an efficient and customizable option for executing alchemical transformations in drug discovery pipelines.

2.6 Topology merger

Following the strategic selection of ligand pairings, the next functional block of our workflow tackles the incorporation of the optimal ligand positions, as determined by the previous docking step. Each chosen ligand is then substituted by a corresponding hybrid structure, meticulously maintaining the original orientation and position within the protein target. This process results in a PDB file comprising the target protein and the hybrid ligand, correctly positioned. This block is integral for facilitating accurate and efficient alchemical simulations by ensuring that the initial positioning of the hybrid ligand aligns with the best-determined locations from the docking stage.

2.7 Topology truncator

Advancing further into the workflow, the subsequent functional block encompasses a protein truncation process. A protein truncator simplifies the system by removing the non-essential parts of the protein structure, focusing primarily on the region of the protein that interacts with the ligand (also known as the binding site or active site). This significantly enhances computational efficiency by streamlining the system for a more manageable and quicker simulation. However, for studies requiring higher precision, this step can be bypassed to retain the full integrity of the protein structure.

2.8 PELE minimizer

Upon the application of the protein truncation process, the workflow proceeds to an integral stage of energy minimization. This step, although necessary only when truncation is performed, plays a vital role in maintaining the structural accuracy of the protein model. PELE's built-in algorithms come into play here, iteratively adjusting the atomic positions in the structure to lower the overall potential energy of the system. In doing so, the protein structure attains a state of lower energy, thereby closely resembling a stable, naturally-occurring structure.

The minimization process accounts for various energy components such as bond stretching, angle bending, and non-bonded interactions (van der Waals and electrostatic). This relaxation stage helps eliminate any steric clashes or unfavorable conformations that might have been introduced during the truncation process. Thus, while this energy minimization step is indispensable when truncation is used, it ensures the structural integrity, biological relevance, and the optimization of the active site for the subsequent steps in the workflow. As such, it significantly enhances the reliability of subsequent simulations, leading to more accurate and meaningful results.

2.9 PELE simulation

Upon completion of the preparatory phases, the simulation block of the workflow commences. This segment features the execution of RBFE calculations, which are each run independently for each transformation between two ligands. Such transformations are composed of a sequence of

small simulations (each lambda), or an epoch of the simulation. These epochs, while interdependent, operate sequentially, each initiating with the trajectory output from the prior epoch, ultimately yielding the complete transformation. These simulations are facilitated by the PELE software, which allows for efficient exploration of the protein's energy landscape through its Monte Carlo stochastic approach. PELE's energy function, crucial for assessing molecular interactions, is an amalgamation of three key components: molecular mechanics energy, solvation energy, and user-imposed constraints. The simulation block operates as follows:

1. Perturbations: small changes are implemented to the initial structure, including atom positions, bond rotations, and amino acid side chain orientations. This is done to efficiently explore the conformational space of the protein or protein-ligand complex.
2. Energy minimization: an optimization algorithm, such as stochastic gradient descent, is employed post each perturbation to locate and establish the local energy minimum in the conformational space.
3. Conformation acceptance or rejection: the Metropolis-Hastings algorithm forms the acceptance criterion, which determines the adoption or rejection of the new conformation, considering the energy differences and system temperature. Accepting new conformations allows for efficient exploration of the conformational space and avoids getting trapped in local energy minima.
4. Iterations: the preceding steps are iteratively executed to generate a series of conformations that accurately represent the dynamics of the protein or protein-ligand complex.

2.9.1 Alchemical work calculation

An integral subcomponent within the simulation block is the calculation of alchemical work. This procedure recalculates energies for all the accepted conformations from the PELE simulations, encompassing their respective lambda values as well as the preceding and following lambdas. The key outcome of this recalculated energy evaluation is the so-called alchemical work (forward and reverse work). It is obtained by observing the interaction energy difference between the conformation at its respective lambda and the preceding or following lambda.

To comprehend this concept, it's critical to clarify what the interaction energy is. The interaction energy represents the energetic relationship between the ligand and the protein, calculated using the equation: $E_{interaction} = E_{complex} - (E_{ligand} + E_{protein})$. Here, $E_{complex}$ is the total energy of the protein-ligand complex, E_{ligand} is the energy of the isolated ligand, and $E_{protein}$ is the energy of the protein devoid of the ligand. Therefore, the interaction energy embodies the energy variation between the ligand and protein within the context of their complex system.

Given the MC nature of the PELE simulations, a noteworthy consideration is made for accepted steps following a series of non-accepted ones. If an accepted conformation follows several rejected steps, this accepted conformation is counted an equivalent number of times as the preceding rejections. This rationale is based on the understanding that if a ligand does not relocate despite several perturbation attempts, it likely indicates the system has found an energetically favorable state. Upon completion of this process, we obtain a distribution of alchemical work values. The distribution comprises as many values as there were steps in the preceding simulation, hence effectively encapsulating the trajectory of the alchemical transformation. This detailed representation provides valuable insights into the transformation process and aids in the accurate estimation of the binding free energies.

2.9.2 Analyzing the simulations

The simulation block includes a subcomponent that uses the open-source library PyMBAR (Michael Shirts et. al, 2022) to derive RBFE from the alchemical work distributions. It is important to note that energies are first scaled by the factor 1/KbT, to ensure the correct physical units are applied and to achieve comparability before the use of PyMBAR methods.

Following the appropriate simulations, Alchemical-PELE carries them out, subsequently analyzing the results utilizing the BAR method to estimate binding free energies (Bennett C.H, 1976; Shirts, M.R., & Chodera, J.D., 2008). This method has been selected due to their ability to provide accurate and efficient estimates of free energy differences between different ligand states (Gapsys V, et al 2019; Schindler CEM, et al., 2020). The BAR-based methods are widely adopted techniques for estimating free energy differences in alchemical simulations. They offer improved accuracy and precision compared to other methods, such as thermodynamic integration (TI) and other approaches (Shirts MR, Pande VS, 2005; Ytreberg FM, et. al 2006 ; Klimovich PV et al., 2015). Calculating binding free energies is essential for understanding the thermodynamics of protein-ligand interactions and identifying potential candidates with optimal binding affinities (Wang, L. et al., 2015). By employing these methods, Alchemical-PELE provides an efficient computational framework for evaluating these essential thermodynamic properties, ultimately aiding in the identification of promising drugs for further development.

The BAR method is one of the pivotal techniques utilized in our process. The primary concept of this method is the computation of the free energy difference through the overlap of probability distributions for the initial and final states during a transformation, a concept rooted in statistical mechanics. These states are weighted according to the Boltzmann distribution, a probability distribution function that describes the states of a system in thermal equilibrium.

BAR relies fundamentally on the Helmholtz free energy difference equation for an NVT (constant Number of particles, Volume, and Temperature) ensemble (Eq. 2). This free energy difference, represented by ΔA, is linked directly with the ratio of probabilities of the two states, made evident through their partition functions. This expression allows us to write the free energy difference as a logarithmic ratio of the expectation values in states j and i for the numerator and the denominator, respectively. In this context, α(q) is a function that must be greater than 0 for all q.

$$\Delta A_{ij} = -k_B T \ln \frac{Q_j}{Q_i} = k_B T \ln \frac{\left\langle \alpha(\vec{q}) \exp[-\beta U_i(\vec{q})] \right\rangle_j}{\left\langle \alpha(\vec{q}) \exp[-\beta U_j(\vec{q})] \right\rangle_i}$$

Equation 2. Free energy difference equation

Bennett used the principle of variational calculus to derive a value for α(q) that minimizes the free energy variance. The outcome is an implicit function of free energy (Eq. 3). This is known as the full BAR equation and needs to be solved numerically. This equation, known as the full BAR equation, necessitates numerical solutions. Though the comprehensive derivation is available in Bennett's paper, this equation can also be derived via a maximum likelihood approach. There are also several equivalent expressions for this implicit equation.

In summary, the BAR method endeavors to find an optimal way to use information from both states to enhance the estimation of free energy. By applying statistics to the exact relationship between the energy difference distributions derived from the two states, the BAR method reduces calculation variability, enhancing the accuracy of the final RBFE estimates.

$$\sum_{i=1}^{n_i} \frac{1}{1+\exp(\ln(n_i/n_j)+\beta\Delta U_{ij}-\beta\Delta A))} - \sum_{j=1}^{n_j} \frac{1}{1+\exp(\ln(n_j/n_i)-\beta\Delta U_{ji}+\beta\Delta A))} = 0$$

Equation 3. Implicit function of free energy. The BAR equation

It is critical to recognize the relationship between the increase in free energy, represented by ΔA, and the Gibbs free energy change, ΔG, which is a standard measure in thermodynamics. Essentially, the Gibbs free energy change offers a measure of the maximum reversible work that a system can perform at constant temperature and pressure, excluding any work performed due to volume changes. This corresponds to the non-expansion work in a system and is directly related to the ΔA calculated in the BAR method, thereby connecting the statistical mechanics approach of the BAR method to classical thermodynamic quantities. This connection provides a reliable pathway to translate our calculated free energy differences into physically meaningful quantities that can guide the design of potent drug candidates.

Simultaneously, Exponential Averaging is also applied to derive the RBFE values. The computations for RBFE are carried out between successive lambdas, producing an RBFE value for each stage of the transformation. These individual RBFE values are then summed to yield the total RBFE for the full alchemical transformation, allowing for a comprehensive understanding of the transformation's energy profile and the overall binding free energy.

2.9 Rank ligands

Once all ligand transformations and subsequent RBFE calculations are completed, we are ready to rank the ligands based on a selected reference among them. To do this, we utilize the graph structure that encapsulates the simulations (Fig. 3). At this stage, we assign a weight to the edges of the graph, representing the simulation error. This error is derived from the error propagation method applied to each of the individual lambda simulations, quantifying the uncertainty in RBFE calculations due to simulation errors at each lambda state. Once the weights have been assigned, we can start to rank the ligands by traversing the graph from the reference ligand to all others. The aim is to add up the RBFE values along the path from the reference to each ligand, thus yielding the RBFE of each ligand compared to the reference.

This traversal is carried out using a minimum spanning tree algorithm. The initial reference ligand is randomly chosen among them. However, after traversing the graph once, we identify the ligand that has the most negative energy in relation to the initially chosen reference. This identified ligand becomes the new reference, and we repeat the process. Therefore, the final reference ligand is the most energetically favorable one among all the ligands, and all other ligands are assigned positive energy values relative to this final reference ligand. This provides a comprehensive ranking of all ligands based on their energetic favorability.

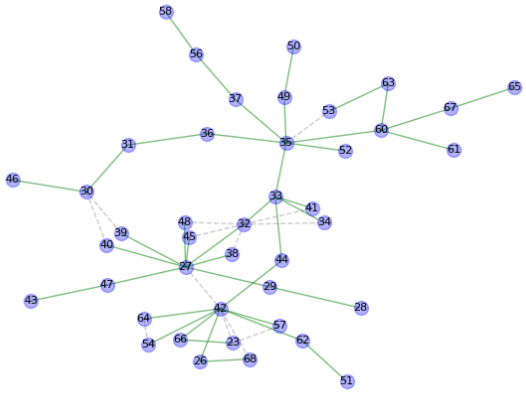

Figure. 3. Visualization of the transformation graph for the MCL1 system. The green solid lines represent the pathway traversed to derive the binding free energy for each ligand, while the dashed lines depict untraveled paths within the graph.

The initial scripts used in the early stages of this project are available for review and utilization in an accessible GitHub repository Alchemical-PELE. However, it is important to note that these individual functionalities have since been refined and integrated within the private repository of Nostrum Biodiscovery, thereby significantly enhancing their efficiency and usability.

3 Results and Discussion

In our analysis, the main objective was to derive a correlation between the relative binding free energies obtained computationally and experimental data. The comparison of these two datasets was aimed at assessing the predictive capacity of our method in ranking ligands. To validate the efficacy and precision of our method, we extensively tested our proposed workflow on the MCL1 and Tyk2 systems, each presenting a different challenge due to their respective structural and chemical characteristics. These systems, being well-characterized through thorough experimental investigation, served as robust benchmarks to evaluate our workflow's ability to calculate relative binding free energies and rank ligands. The different structures have been obtained from the GitHub repository protLig_benchmark (Gapsys V et. al 2014).

During the experiments on these systems, we adhered to a standardized simulation protocol. The protocol employed the OpenFF-2.0.0 forcefield (Boothroyd S et. al 2022) and the OBC solvent model (Onufriev, A et. al 2004). The simulations were performed using 20 CPUs, with each lambda epoch in the alchemical transformation process composed of 20 PELE steps. The alchemical transformation was systematically segmented into sequential epochs, each corresponding to specific lambda values: 0.0, 0.2, 0.4, 0.6, 0.8, and 1.0. We initiated the transformation process focusing on the unique Coulombic parameters of the first molecule, followed by a similar progression for the bonded and van der Waals parameters. Finally, we applied the same lambda progression to the Coulombic parameters of the unique atoms in the second molecule. This organized and systematic approach to the alchemical transformation helped maintain system stability throughout the process, culminating in a protocol comprising a total of 16 epochs.

3.1 Evaluation of RBFE with Alchemical-PELE

The MCL1 system was examined using a diverse library of 42 ligands and the Tyk2 system with a focused set of 15 ligands. These ligands underwent a variety of transformations (62 for MCL1 and 23 for Tyk2) following our standardized simulation protocol. Each transformation allowed us to calculate the relative binding free energy for each ligand pair, thereby generating a detailed understanding of the energetic landscape of these protein-ligand systems.

During our analysis of these systems, we discovered that errors can accumulate as we traverse the transformation graph to rank the ligands. The impact of these errors varied between the systems. In the case of the MCL1 system, the effect was minor, resulting in a correlation coefficient (R) of 0.63 between computed binding free energies and experimental data. However, the Tyk2 system experienced a more prominent effect. The initial correlation of 0.40 for Tyk2 decreased to 0.25 after ligand ranking (Fig. 4).

The correlation between experimental data and our calculated results in the MCL1 system supports the effectiveness of our method in estimating ligand binding energies. This validation strengthens the credibility of our approach and encourages further refinement and wider application. In contrast, the evaluation of the Tyk2 system posed challenges. While this system initially demonstrated an acceptable level of correlation in the RBFE values, the accumulated errors during the graph traversal negatively impacted the final correlation. This highlights the importance of improving our strategy to handle errors during the alchemical transformation process, aiming to enhance the robustness of ligand ranking.

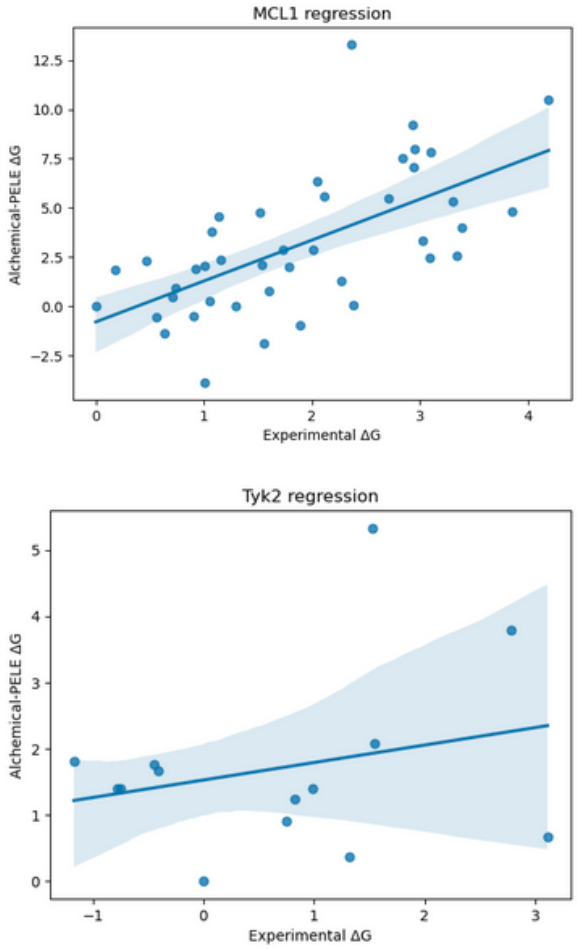

Figure. 4. Scatter plots illustrating the correlation between calculated and experimental RBFE values for the MCL1 and Tyk2 systems, respectively

The information of the results for each ligand, including experimental values and the corresponding ligand structures, is compiled in the tables of each system in the Supplementary Material. This integral part of our study presents detailed calculations and outcomes in a structured format while also providing visual context for the interactions of the ligands within the protein-ligand complexes. In this way, the overall understanding of our study is enhanced.

In evaluating the performance of our method, we compared it with several reputable free energy perturbation tools mentioned in the existing literature (Gapsys V et al., 2019). Notably, our tool exhibited marginally superior performance in the MCL1 system, surpassing the average outcome of FEP+ (Schrödinger, 2021) by 0.1 points. FEP+ is currently recognized as one of the most proficient tools in this field of study. However, the scenario was different in the Tyk2 system, where both FEP+ and various protocols from GROMACS (Mark James et al., 2015) outperformed our technique, achieving a higher score of 0.2 to 0.3 points. These variations in performance across different molecular systems highlight the inherent complexities in protein-ligand simulations and underscore the ongoing pursuit of consistently accurate methodologies across a wide range of systems

4 Conclusions

This study has afforded a valuable insight into the efficacy of our novel computational methodology for calculating relative binding free energies and ranking ligands. Through the application of this workflow to two distinct systems, MCL1 and Tyk2 we have been able to highlight its strengths and limitations. The exploration of these systems with our computational method has shed light on its potential as a tool for calculating relative binding energies. While additional work is needed to minimize error accumulation, we believe our approach holds significant promise in the realm of ligand binding affinity prediction.

Looking ahead, two key areas for future work have emerged. Firstly, due to the typical absence of experimental values for comparison, a method for determining the convergence of alchemical work distributions will be developed. This study between distributions would indicate whether the lambda protocol is accurate, or whether adjustments are required, potentially through the addition or removal of lambdas.

Secondly, another approach to explore is simulating the transformation of the ligand in the solvent outside the protein and subtracting these internal ligand energy values from the relative binding

energy calculation. Both approaches seek to enhance the robustness and accuracy of our method, ultimately aiming to continue the validation and refinement of this promising approach.

The advancements made in this study lay the groundwork for further refinement and application of this computational method, bolstering our ongoing commitment to improving the prediction of ligand binding affinities.

Acknowledgements

I extend my deepest gratitude to my mentor, Martí Municoy, for his invaluable guidance and support throughout this project. I also appreciate Nostrum Biodiscovery for the enlightening professional experience they provided. Lastly, I acknowledge the unwavering support of my parents, instrumental in my academic journey, and my brother, who has always been the first reader of all my writings.

Funding

This study was financially supported by Nostrum Biodiscovery, which also provided invaluable resources for this project. The computational costs and the provision of the necessary computational cluster were fully covered by the company. Their substantial support greatly facilitated the execution of this project, enhancing the feasibility of extensive simulations and data analyses.

References

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., & Schacht, A. L. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature Reviews Drug Discovery, 9(3), 203-214. doi: 10.1038/nrd3078

Schneider, G., & Fechner, U. (2005). Computer-based de novo design of drug-like molecules. Nature Reviews Drug Discovery, 4(8), 649-663. doi:10.1038/nrd1799

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. doi:10.1038/nrd1549

Karplus, M., & Kuriyan, J. (2005). Molecular dynamics and protein function. Proceedings of the National Academy of Sciences of the United States of America, 102(19), 6679-6685. doi:10.1073/pnas.0408930102

Zwanzig, R. W. (1954). High-temperature equation of state by a perturbation method. I. Nonpolar gasses. The Journal of Chemical Physics, 22(8), 1420-1426. doi:10.1063/1.1740409

Kollman, P. A. (1993). Free energy calculations: applications to chemical and biochemical phenomena. Chemical Reviews, 93(7), 2395-2417. doi:10.1021/cr00023a004

Chodera JD, Mobley DL, Shirts MR, Dixon RW, Branson K, Pande VS. Alchemical free energy methods for drug discovery: progress and challenges. Curr Opin Struct Biol. 2011 Apr;21(2):150-60. doi: 10.1016/j.sbi.2011.01.011. Epub 2011 Feb 23. PMID: 21349700; PMCID: PMC3085996.

Wang, L., Wu, Y., Deng, Y., Kim, B., Pierce, L., Krilov, G., ... & 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. doi:10.1021/ja512751q

Boothroyd S, Behara PK, Madin O, Hahn D, Jang H, Gapsys V, et al. Development and Benchmarking of Open Force Field 2.0.0 — the Sage Small Molecule Force Field. ChemRxiv. Cambridge: Cambridge Open Engage; 2022

Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. Proteins 2004, 55, 383−394.

Ciordia M, Pérez-Benito L, Delgado F, Trabanco AA, Tresadern G. Application of Free Energy Perturbation for the Design of BACE1 Inhibitors. J Chem Inf Model. 2016 Sep 26;56(9):1856-71. doi: 10.1021/acs.jcim.6b00220. Epub 2016 Aug 24. PMID: 27500414.

Zhu, F., Bourguet, F.A., Bennett, W.F.D. et al. Large-scale application of free energy perturbation calculations for antibody design. Sci Rep 12, 12489 (2022). https://doi.org/10.1038/s41598-022-14443-z

Ou-Yang SS, Lu JY, Kong XQ, Liang ZJ, Luo C, Jiang H. Computational drug discovery. Acta Pharmacol Sin. 2012 Sep;33(9):1131-40. doi: 10.1038/aps.2012.109. Epub 2012 Aug 27. PMID: 22922346; PMCID: PMC4003107.

Boresch, S., Tettinger, F., Leitgeb, M., & Karplus, M. (2003). Absolute binding free energies: a quantitative approach for their calculation. The Journal of Physical Chemistry B, 107(35), 9535-9551. doi:10.1021/jp0217839

Hansen, N., & Van Gunsteren, W. F. (2014). Practical aspects of free-energy calculations: a review. Journal of Chemical Theory and Computation, 10(7), 2632-2647. DOI: 10.1021/ct500161f

Williams-Noonan, Billy J, Elizabeth Yuriev and David K. Chalmers. “Relative Binding Free Energy Predictions for Inhibitors of Tetrameric Influenza Virus Neuraminidase.” (2021).

Borrelli KW, Vitalis A, Alcantara R, Guallar V. PELE: Protein Energy Landscape Exploration. A Novel Monte Carlo Based Technique. Journal of Chemical Theory and Computation. 2005 Nov;1(6):1304-1311. DOI: 10.1021/ct0501811. PMID: 26631674.

Mey ASJS, Allen BK, Macdonald HEB, Chodera JD, Hahn DF, Kuhn M, Michel J, Mobley DL, Naden LN, Prasad S, Rizzi A, Scheen J, Shirts MR, Tresadern G, Xu H. Best Practices for Alchemical Free Energy Calculations [Article v1.0]. Living J Comput Mol Sci. 2020;2(1):18378. doi: 10.33011/livecoms.2.1.18378. PMID: 34458687; PMCID: PMC8388617.

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

Liu S, Wu Y, Lin T, Abel R, Redmann JP, Summa CM, Jaber VR, Lim NM, Mobley DL. Lead optimization mapper: automating free energy calculations for lead optimization. J Comput Aided Mol Des. 2013 Sep;27(9):755-70. doi: 10.1007/s10822-013-9678-y. Epub 2013 Sep 26. PMID: 24072356; PMCID: PMC3837551.

Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. Journal of Computational Physics, 22(2), 245-268. doi:10.1016/0021-9991(76)90078-4

Shirts, M. R., & Chodera, J. D. (2008). Statistically optimal analysis of samples from multiple equilibrium states. The Journal of Chemical Physics, 129(12), 124105. doi:10.1063/1.2978177

Schindler CEM, Baumann H, Blum A, Böse D, Buchstaller HP, Burgdorf L, Cappel D, Chekler E, Czodrowski P, Dorsch D, Eguida MKI, Follows B, Fuchß T, Grädler U, Gunera J, Johnson T, Jorand Lebrun C, Karra S, Klein M, Knehans T, Koetzner L, Krier M, Leiendecker M, Leuthner B, Li L, Mochalkin I, Musil D, Neagu C, Rippmann F, Schiemann K, Schulz R, Steinbrecher T, Tanzer EM, Unzue Lopez A, Viacava Follis A, Wegener A, Kuhn D. Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. J Chem Inf Model. 2020 Nov 23;60(11):5457-5474. doi: 10.1021/acs.jcim.0c00900. Epub 2020 Sep 3. PMID: 32813975.

Gapsys V, Pérez-Benito L, Aldeghi M, Seeliger D, van Vlijmen H, Tresadern G, de Groot BL. Large scale relative protein ligand binding affinities using non-equilibrium alchemy. Chem Sci. 2019 Dec 2;11(4):1140-1152. doi: 10.1039/c9sc03754c. PMID: 34084371; PMCID: PMC8145179.

Michael Shirts, Kyle Beauchamp, Levi Naden, John Chodera, Jaime Rodríguez-Guerra, Stefano Martiniani, Chaya Stern, Mike Henry, Josh Fass, Richard Gowers, Robert T. McGibbon, Bradley Dice, Chris Jones, David Dotson, & Tucker Burgin. (2022). choderalab/pymbar: 3.1.1 (3.1.1). Zenodo. https://doi.org/10.5281/zenodo.7383197

Shirts MR, Pande VS. Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration. J Chem Phys. 2005 Apr 8;122(14):144107. doi: 10.1063/1.1873592. PMID: 15847516.

Ytreberg FM, Swendsen RH, Zuckerman DM. Comparison of free energy methods for molecular systems. J Chem Phys. 2006 Nov 14;125(18):184114. doi: 10.1063/1.2378907. PMID: 17115745.

Klimovich, P. V., Shirts, M. R., & Mobley, D. L. (2015). Guidelines for the analysis of free energy calculations. Journal of Computer-Aided Molecular Design, 29(5), 397-411. doi: 10.1007/s10822-015-9840-9

Mobley DL, Klimovich PV. Perspective: Alchemical free energy calculations for drug discovery. J Chem Phys. 2012 Dec 21;137(23):230901. doi: 10.1063/1.4769292. PMID: 23267463; PMCID: PMC3537745.

Truhlar, D.G. Chipot, C., Pohorille, A., Eds. Free Energy Calculations: Theory and Applications in Chemistry and Biology. Theor Chem Account 121, 105–106 (2008). https://doi.org/10.1007/s00214-008-0449-0

Pitera, J.W., & van Gunsteren, W.F. (2002). A Comparison of Non-Bonded Scaling Approaches for Free Energy Calculations. Molecular Simulation, 28, 45 - 65.

Steinbrecher T, Mobley DL, Case DA. Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations. J Chem Phys. 2007 Dec 7;127(21):214108. doi: 10.1063/1.2799191. PMID: 18067350.

Banks, Jay L. and Beard, Hege S. and Cao, Yixiang and Cho, Art E. and Damm, Wolfgang and Farid, Ramy and Felts, Anthony K. and Halgren, Thomas A. and Mainz, Daniel T. and Maple, Jon R. and Murphy, Robert and Philipp, Dean M. and Repasky, Matthew P. and Zhang, Linda Y. and Berne, Bruce J. and Friesner, Richard A. and Gallicchio, Emilio and Levy, Ronald M. "Integrated Modeling Program, Applied Chemical Theory (IMPACT)" J. Comput.Chem.26(16),pp.1752-1780,2005.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2742605/

Onufriev, A., D. Bashford and D. A. Case, "Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model", PROTEINS, 55, pp. 383–394, 2004. http://onlinelibrary.wiley.com/doi/10.1002/prot.20033/abstract

Schrödinger Release 2023-2: FEP+, Schrödinger, New York, NY, 2021.

Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, Erik Lindahl, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers,SoftwareX,Volumes1–2, 2015, Pages 19-25, ISSN 2352-7110, https://doi.org/10.1016/j.softx.2015.06.001.